Soc. Netw. Anal. Min. (2012) 2:69–95
DOI 10.1007/s13278-011-0030-z

REVIEW ARTICLE

Semantically interconnected social networks

Alessandro Cucchiarelli · Fulvio D'Antonio · Paola Velardi

Received: 13 May 2010 / Revised: 19 May 2011 / Accepted: 27 May 2011 / Published online: 29 June 2011

Abstract Social network analysis aims to identify collaborations and helps people organize themselves through community participation and information sharing. The primary sources for social network modelling are explicit relationships such as co-authoring, citations, friendship, etc. However, to enable the integration of on-line community information and to fully describe the content and structure of community sites, secondary sources of information, such as documents, e-mails, blogs and discussions, can be exploited. In this paper we describe a methodology and a battery of tools to automatically extract from documents the relevant topics shared among community members and to analyse the evolution of the network also in terms of emergence and decay of collaboration themes. Experiments are conducted on a scientific network funded by the European Community, the INTEROP network of excellence, and on the United Kingdom research community in medical image understanding and analysis.

Keywords Social networks · Semantic web · Natural language processing · Text analysis · Clustering · Computer-supported collaborative work

A. Cucchiarelli (✉) · F. D'Antonio
DIIGA, Università Politecnica delle Marche, Ancona, Italy
e-mail: a.cucchiarelli@diiga.univpm.it

F. D'Antonio
e-mail: dantonio@diiga.univpm.it

P. Velardi
DIS, 'Sapienza' University of Rome, Rome, Italy
e-mail: velardi@di.uniroma1.it

1 Introduction

Social networks (SN) are explicit representations of the relationships between individuals and groups in a community (Finin et al. 2005), such as friendship, co-authoring, citations, etc. They support both a visual and a mathematical analysis of human collaborations, measuring relationships and flows among people, groups, organizations, computers, web sites, and other information/knowledge processing entities. The availability of software tools and mathematical models (Wasserman and Faust 1994; Newman 2003; Scott 2000) has extended the use of SN analysis from social sciences to management consulting, financial exchanges, epidemiology and more.

Relationships among actors in traditional social network analysis (SNA) are modelled as a function of a quantifiable social interaction (e.g. co-authorships, citations, etc.) (Wasserman and Faust 1994). However, within a business, social or research community, network analysts are also strongly interested in the *communicative content* exchanged by the community members, not merely in the number of relationships. In this regard, it has recently been argued that a great potential benefit would arise from combining concepts and methods of social networks and the semantic web (Mika 2007). The semantic web (Berners-Lee et al. 2001) fosters a new generation of intelligent applications by providing resources and programmes with rich and effective ways to share information and knowledge on the web.

Lately, several papers (e.g. Finin et al. 2005; Jung and Euzenat 2007) and initiatives, like the SIOC "semantically interlinked online communities" project (Bojars et al. 2008), have proposed the integration of on-line community information through the use of ontologies, like the Friend Of A Friend (FOAF) ontology,[1] for describing personal profiles

[1] http://xmlns.com/foaf/spec.

and social networking information. Ontologies (Gruber 2003; Staab and Studer 2009) formalize knowledge in terms of domain entities, relationships between entities and axioms to enable logical reasoning. Through the use of ontologies and ontology languages, user-generated content can be expressed formally, and innovative semantic applications can be built on top of the existing social web.

Research efforts concerning the idea of the social web or semantic social network (SSN) focus primarily on the specification of ontology standards and ontology matching procedures, assuming a "Semantic Web-like" environment in which the documents shared among community members are annotated with the concepts of a domain ontology. This is a rather demanding postulate, since in real SNs the information exchanged between actors is mostly unstructured and we cannot expect that, given the variegated social and economical nature of many communities, these will eventually entrust knowledge engineers or simple users to manage the job of ontology building and semantic annotation. We therefore believe that the success of the SSN vision also depends on the availability of automated tools to extract and formalize information from unstructured documents, thus restricting human intervention to verification and post-editing.

In this paper we present a methodology and a battery of software applications to automatically learn and analyse the collaboration *content* in scientific communities. Text mining and clustering techniques are used to identify the relevant *research themes*. Themes, hereafter also called *topics,* are clusters of semantically related terms (or concepts) extracted from the documents shared among the community members and grouped according to some similarity criterion. We then apply traditional and novel social analysis measures to study the emergence of interest around certain themes, the evolution of collaborations around these themes, and to identify the potential for better cooperation. We call this process content-based social network analysis (CB-SN).

The idea of modelling the content of social relationships with clusters of terms is not entirely new. In Dhiraj and Gatica-Perez (2006) a method is proposed to discover "semantic clusters" in Google news, i.e. groups of people sharing the same topics of discussion. In McCallum et al. (2005) the authors propose the Author-Recipient-Topic model, a Bayesian network that captures topics and the directed social network of senders and recipients in a message-exchange context. In Zhou et al. (2006) the objective is reversed: they introduce consideration of social factors (e.g. the fact that two authors begin to cooperate) into traditional content analysis, in order to provide social justification to topic evolution in time; similarly, in Nallapati et al. (2008) topic models and citations are combined. Following a similar perspective is the work of Mei et al. (2008), in which content mining and co-authorship analysis are combined for *topical community discovery* in document collections. To learn topics, the latter three papers use variations of latent Dirichlet allocation (LDA), a generative probabilistic model for collection of discrete data such as text corpora (Blei et al. 2003).

Statistical methods for topic detection (and in particular LDA) require complex parameter estimation over a large training set; furthermore, there is no principled way to establish the number of topics to be learned. Another problem with state-of-the-art literature on topic detection is the bag-of-words model used for extracting content from documents. For example, in Dhiraj and Gatica-Perez (2006) one of the topics around which groups of people are created is "*said, bomb, police, London, attack*", an example from McCallum et al. (2005) is "*section, party, language, contract*, […]" and from Mei et al. (2008) is "*web, services, semantic, service, poor, ontology, rdf, management*". In many domains, the significance of topics could be more evident using key-phrases (e.g. "web" + "services" in the third example). As a matter of fact, this problem is pervasive in the term clustering literature (see, e.g. clusters in Tagarelli and Karypis 2008), regardless of the application. The authors are more concerned with the design of powerful clustering models than with the data used to feed these models. Unfortunately, it has been experimentally demonstrated (see, e.g. Kovacs et al. 2005) that clustering performance strongly depends on how well input data can be separated, which makes the selection of input features a central issue.

In our view, the content-based social analysis is more effective when the topics to detect have a strong semantic cohesion. A simple bag-of-words model seems rather inadequate at capturing the content of social communications, mainly in specialized domains where terminology is central (thematic blogs, research communities, networked companies, etc.) and which have higher applicative potential for social analysts.

Our paper can be considered as being split in two parts. The first part describes a method to identify topics of interest among community members based on a corpus of documents written by the members. The second part introduces a novel line of analysis by comparing the interest similarity between community members with their actual social network observed through co-authorship networks. This is used to analyse the evolution of interests and collaborations among the community members, and quantitative metrics are suggested to identify new opportunities for effective collaboration. A concrete case study is used to demonstrate that CB-SN analysis provides insight that standard, collaboration-based SN analysis is unable to provide.

The significant contribution of the paper lies in

- the completeness of the methodology, indicating the steps required right from mining the document corpus and social network to making useful inferences from the data;
- the novelty of the idea, mostly on modelling social relations both with the strength and the nature of relations.

Other specific contributions are

- a methodology for topic detection, whose novelty relies on an improved technique for feature extraction, and a new evaluation methodology used to adjust cluster granularity, using the metaphor of a librarian. The evaluation methodology allows it to choose the best clustering results in an ensemble obtained through different clustering algorithms and different values for the number of extracted topics;
- the definition of new measures to study the CB-SN (*network defragmentation* and *resilience*, related to the notion of bridgeness). A bipartite graph model between social actors and topics is also analysed.

We remark that our methodology is fully implemented and the relevant tools are either freely available as applications on our web sites, or as public web resources (like the clustering algorithms). Furthermore, the methodology is *domain independent*: capturing the research themes in different domains is supported by automated tools and requires minimal post-validation. Finally, unlike statistical models, our method for theme extraction requires *no training* or *parameter estimation*, though, as we have just remarked, some manual work is needed in order to validate the automatically learned "semantics" of a new domain.

The CB-SN model we propose enables a deeper (with respect to standard SN tools) and informative analysis of a community of interest, which could be used, for example, to improve the efficacy as well as monitor the effects of research coordination actions by funding bodies. This capability has been particularly useful to monitor a research group born during the European-Union funded INTEROP network of excellence (NoE), now continuing its mission within a permanent institution named "Virtual Laboratory on Enterprise Interoperability", the V-Lab.[2] In addition to the use of quantitative evaluation techniques, the availability of a real-world application domain has allowed us to verify the correspondence of the phenomena emerging from our CB-SN model with the reality of a well-established scientific community.

The paper is organized as follows: Sect. 2 describes the methodology and algorithms used for concept extraction and topic clustering. Section 3 presents the CB-SN model and measures and provides detailed examples of the type of analysis supported, applied to the cases of the INTEROP NoE and to the United Kingdom research community in medical image understanding and analysis (MIUA). Finally, Sect. 4 is dedicated to concluding remarks and the presentation of future work.

The methodology we propose is composed by several steps: terminology extraction, term similarity analysis, clustering, social network modelling. For each of these phases we faced the option of adopting a variety of existing methods, or a novel methodology. Analysing and comparing these options in detail would overload the paper (this task is deferred to dedicated papers, some of which have already been published) and divert from its main focus. We want to show that SNA can extend its scope and outcome by developing tools and metrics to analyse the diversity of topics of interest in a community and to identify new opportunities for effective collaboration. Having stated that, each section is nevertheless accompanied by a state of the art analysis and a synthetic justification of the adopted approach. In some cases, alternative solutions to a given problem have been taken into account, the major problem being how to compare them in a quantitative way. This is still an open research issue and we deal with it in Sects. 2.2 and 3.4.

2 Detecting domain topics in a social network

The objective of the CB-SN analysis is to detect the "hot" themes of collaboration within a network and monitor how these themes evolve and catalyse the interest of the network members. This is performed in three steps:

1. *Concept extraction*. In this step the concepts that play a relevant role in describing actors' relationships are extracted. This is performed by collecting the textual communications (e-mail, blogs, co-authored papers, documents of any type) shared among the community members and by extracting the set of relevant, domain-specific, concepts. The semantic similarity between concept pairs is then computed by analysing ontological and contextual (co-occurrence) relations in the domain. This leads to the definition of the concept feature vectors
2. *Topic detection*. In this step, a clustering algorithm is fed with the concept feature vectors to group concepts into *topics*. Topics are *clusters of semantically close concepts*, where both concepts and clusters depend on the specific set of documents that represent inter-actors communications over a given time interval, (e.g. the scientific publications produced by a research community over a given time span). As a result of this step,

[2] http://www.interop-vlab.eu.

a social relation between two given actors (persons or organizations) can be modelled in terms of the common themes of interest.

3. *Social network analysis*. Social network measures are used to model the collaboration content and its evolution over time. The collaboration strength is usually computed (Jamali and Abolhhassani 2006; Newman 2003) as the number of common activities between pairs of actors (e.g. the number of common papers in a research community). Instead, we build a network of actors connected by links whose weights are a function of *topic overlapping*. This network is what we call CB-SN model. Traditional and novel social network measures are applied to the study of a CB-SN, in order to analyse the dynamics of the agents' aggregation around the topics.

This processing steps are summarized in Fig. 1. Examples of input–output data in the figure belong to the INTEROP domain.

This section is dedicated to a description of the first two phases, concept extraction and topic detection. Its contribution is the following:

- The concept extraction method is new, even though it relies on tools described in other papers by the current authors;
- the computation of similarity vectors for subsequent topic clustering is based on the novel notion of *semantic co-occurrences*;
- the clustering algorithms that we use are available in the literature, but we propose a novel validation methodology to identify the best clustering results in an ensemble.

2.1 Concept extraction

The objective of the concept extraction phase is to identify the *emergent semantics* of a community, i.e. the concepts that better characterize the *content* of an actor's communications. Concepts are extracted from available texts exchanged among the members of the community. We refer to these texts as the *domain corpus*, represented by a set *D* of documents. Once the concepts have been identified, a similarity metric is defined to weight relations between concept pairs. Conceptual relations are derived from three sources: co-occurrence data extracted from the domain corpus, term definitions already available in a glossary or automatically extracted from the web, and a domain ontology or thesaurus. We first describe the concept extraction methodology and then the method to compute concept similarity.

Concept extraction and similarity computation are supported by the availability of three automated tools for terminology extraction (Sclano and Velardi 2007), automated glossary creation (Velardi et al. 2008a) and ontology learning (Velardi et al. 2007; Navigli and Velardi 2008). These tools have already been described and evaluated in our previously published work; therefore, here we do not go into detail about their functionalities.

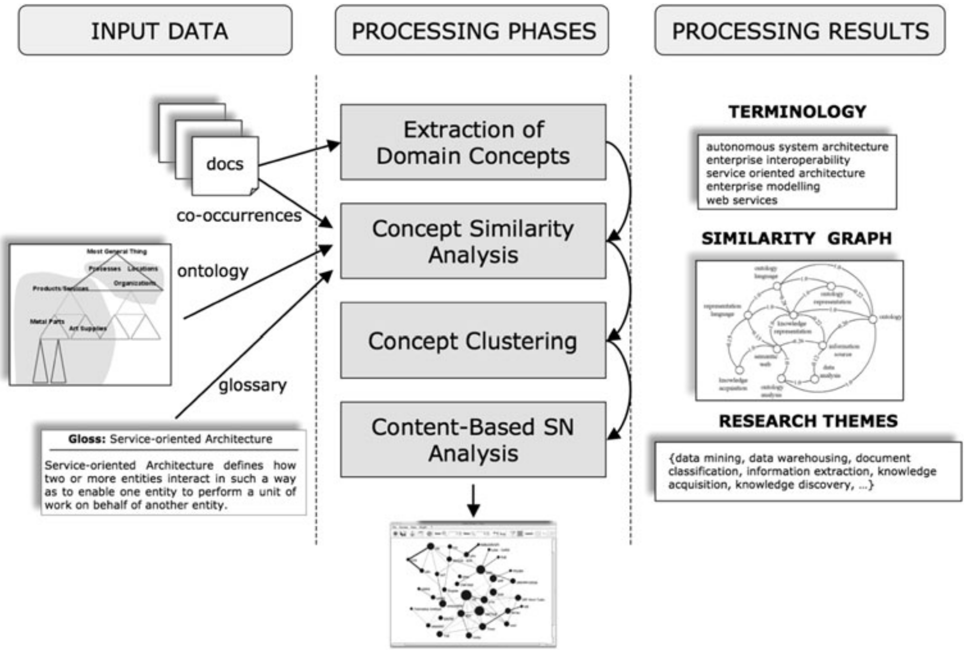

INPUT DATA
PROCESSING PHASES
PROCESSING RESULTS
docs
co-occurrences
Most General Thing
Processes
Locations
Organizations
Products Services
Metal Parts
Art Supplies
ontology
glossary
Gloss: Service-oriented Architecture
Service-oriented Architecture defines how two or more entities interact in such a way as to enable one entity to perform a unit of work on behalf of another entity.
Extraction of Domain Concepts
Concept Similarity Analysis
Concept Clustering
Content-Based SN Analysis
TERMINOLOGY
autonomous system architecture
enterprise interoperability
service oriented architecture
enterprise modelling
web services
SIMILARITY GRAPH
ontology language
ontology representation
representation language
knowledge representation
ontology
semantic web
information source
data analysis
knowledge acquisition
ontology analysis
RESEARCH THEMES
{data mining, data warehousing, document classification, information extraction, knowledge acquisition, knowledge discovery, ...}

Fig. 1 Processing phases for the themes of collaboration within a network

2.1.1 Concept identification

It has often been pointed out (Kang 2003; Nenadic et al. 2003; Hammouda and Kamel 2004) that terminological strings (e.g. multi-word sequences, or *key phrases*) are more informative features than are single words for representing the content of a document. Terminology is pervasive in specialized and technical domains: those with a higher applicative impact. Furthermore terminology reduces the problem of semantic ambiguity, a crucial one in language processing research. Regarding the cluster example in Mei et al. (2008), "*web, services, semantic, service, poor, ontology, rdf, management*", there is little doubt that the cluster would improve its significance as a research topic by merging some of its members in a unique concept, e.g. *web service, ontology management*. Key-phrase extraction is indeed a problem of optimized feature detection, which is almost ignored in topic clustering literature, although key-phrases have been used in related domains such as information retrieval (Zhong et al. 2004).

To extract terminological strings from a domain corpus, we used the TermExtractor system (Sclano and Velardi 2007), a freely accessible web application, developed by us. TermExtractor selects the concepts that are *consistently* and *specifically* used across the corpus documents, according to information-theoretic measures, statistical filters and structural text analysis. Several options are available to fit the needs of specific domains of analysis, among which singleton terms extraction, acronym detection, named entity analysis and single user or group validation. The tool allows the user to validate the resulting terminology in order to prune out possible remaining non-terminological string and terms not pertinent to the domain.

TermExtractor has been experimented in the large, within applications and by its many users, and showed to be one among the best available terminology tools. Its high quality in terminology extraction (precision may vary between 70 and 90% depending upon the domain and the many parameters) is at the basis of the subsequent steps of our CB-SN methodology.

2.1.2 Semantic similarity feature vectors creation

In natural language-related applications, similarity between words is typically measured using as its source either word co-occurrences (*distributional similarity*) or taxonomic and ontological relations (*ontological similarity*) (Resnik 1999; Bollegala et al. 2007; Hirst and Budanitsky 2001; Pedersen et al. 2007; Terra and Clarke 2003; Weeds and Weir 2006). However, both distributional and ontological similarity are useful for term clustering. An illustrative example is shown in Fig. 2.

The figure shows a set of terms occurring in the INTEROP domain and illustrates the general idea behind semantic-based topic clustering. Dashed lines indicate co-occurrence relations found in domain texts, whereas bold lines indicate semantic relations in the domain ontology (*is_a* links in the case of Fig. 3). The graph shows that terms of type collaboration tend to co-occur with terms of type organization management, thus suggesting the existence of a topic that we could name "collaboration + organization management", represented by all the terms in the figure. Our task of topic clustering consists precisely in extracting groups of co-occurring concepts, rather than terms. We call these semantic co-occurrences.

The first step towards semantic co-occurrences detection is the creation of an undirected graph, with nodes representing terminological strings (hereafter denoted also as domain concepts) extracted as described in Sect. 2.1.1, and edges connecting two terms t_i and t_j if any of the following conditions holds:

1. a direct relation exists in the domain ontology between the concepts expressed by t_i and t_j;
2. the term t_i occurs in the domain glossary definition of t_j;
3. the two terms co-occur in the document corpus.

Each edge has a weight equal to 1.0 if condition 1 or 2 holds, or equal to the Dice coefficient computed for t_i and t_j in case of condition 3. The Dice coefficient (Weeds and Weir 2006) is defined as follows:

$$\mathrm{Dice}(t_j, t_i) = \frac{2f(t_j, t_i)}{f(t_j) + f(t_i)}$$

where $f(t_j)$ and $f(t_i)$ are the occurrence frequency of t_j and t_i in the document corpus, and $f(t_j, t_i)$ is the co-occurrence frequency of the two terms. Whenever more than one condition is triggered to generate an edge, the maximum weight is chosen for that edge.

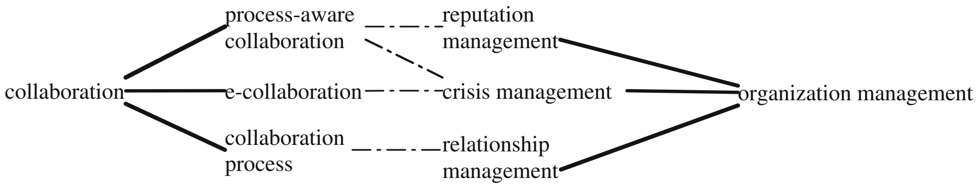


Fig. 2 Contextual and semantic similarity

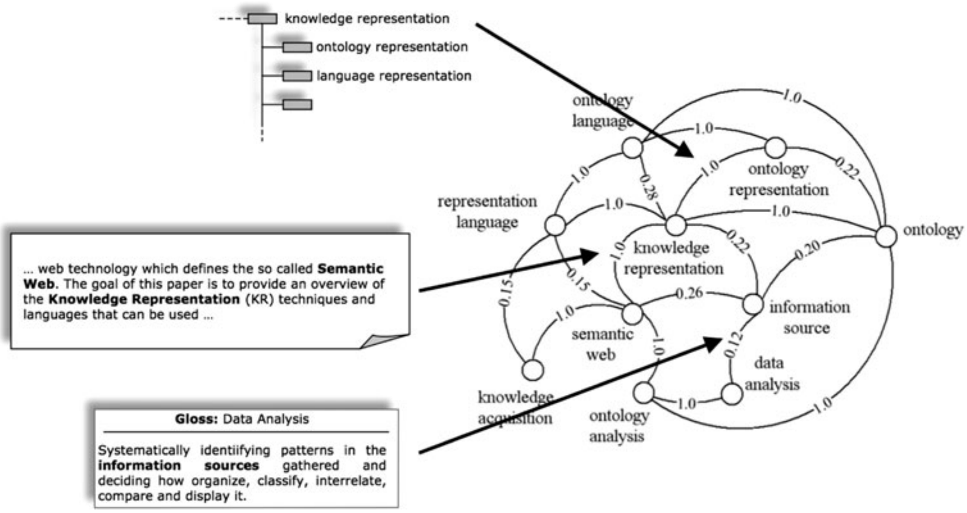

knowledge representation
ontology representation
language representation
... web technology which defines the so called **Semantic Web**. The goal of this paper is to provide an overview of the **Knowledge Representation** (KR) techniques and languages that can be used ...
Gloss: Data Analysis
Systematically identiifying patterns in the **information sources** gathered and deciding how organize, classify, interrelate, compare and display it.
ontology language
ontology representation
representation language
knowledge representation
ontology
information source
semantic web
data analysis
knowledge acquisition
ontology analysis
1.0
1.0
1.0
0.28
1.0
0.22
1.0
1.0
0.22
0.20
1.0
0.15
0.15
0.26
1.0
1.0
0.12
1.0
1.0

Fig. 3 An excerpt of the graph used to compute the semantic similarity between concept pairs

In Fig. 3 we show an excerpt of the graph obtained as described above for the INTEROP domain. Nodes in the graph were obtained by extracting domain terminology from the INTEROP document repository, a set of research papers collected by the project members.

The graph is the basic structure used to produce semantic similarity feature vectors as follows:

1. For each pair of concepts t_i and t_j, we perform a depth-first search and compute the set $P(t_i, t_j)$ of edge paths of length $\leq L$ (where L is the maximum path length) which connect the two concepts and that we call *semantic paths*.
2. We then determine the semantic similarity between t_i and t_j. Several measures exist in the literature (see Budanitsky and Hirst 2006 for an overview), although many of them either focus on ontological relations (e.g. the taxonomy hierarchy, such as Sussna 1993; Hirst and St-Onge 1998; Wu and Palmer 1994; Leacock and Chodorow 1998; Ponzetto and Strube 2007) or are based on a notion of information content, which, together with probability estimates, again requires taxonomical information about the concepts to compare (Resnik 1999; Jiang and Conrath 1997; Lin 1998). In contrast, here we propose an unconstrained graph-based measure which is independent of the relation edge type (thus allowing us to combine ontological and co-occurrence relations). We determine the similarity between t_i and t_j as the maximum score among all paths connecting the two concepts:

$$\mathrm{sim}(t_i, t_j) = \max_{p \in P(t_i, t_j)} \mathrm{score}(p)$$

where the score of a path p is given by the inverse of the exponential of its length $|p|$ (i.e. the number of its edges), multiplied by the weight of p, that we define as the product of the weights of its edges $e_1,\ldots,e|p|$:

$$\mathrm{score}(p) = \varepsilon^{-|p|} \prod_{i=1}^{i=|p|} w(e_i)$$

where $w(e_i)$ is the weight assigned to edge e_i in the graph G and ε is the Euler's number (see Ponzetto and Strube 2007 for the motivation of the exponentially decreasing function used). For example, according to the excerpt in Fig. 3, the set of paths P(*text mining*, *data analysis*) is shown in Fig. 4 (we set the maximum path length L to 4).

As a result, the semantic similarity between the two concepts is given by the highest scoring path, namely p_2, which provides a score of 0.049.

3. Finally, each domain concept t_i is associated with an n-dimensional similarity vector x_i, where n is the total number of extracted concepts, and the jth component of x_i is the normalized semantic similarity between concepts t_i and t_j:

$$x_{ij} = \frac{\mathrm{sim}(t_i, t_j)}{\max_{k=1,\ldots,|V|} \mathrm{sim}(t_i, t_k)}$$

In the following, we denote with X the matrix of similarity vectors, where $|X| = |V| = n$. An example of similarity vector x_i (showing only the components with highest similarity with t_i) is

```
activity diagram = (class diagram (1.000), process analysis (0.630),
                    software engineering (0.493), enterprise software
                    (0.488), deployment diagram (0.468), bpms paradigm
                    (0.467), workflow model (0.444), model driven
                    architecture (0.442), workflow management (0.418),
                    ...)
```

A detailed evaluation of the semantic similarity measure presented in this section and a comparison with other similar measures would divert us from the main theme of the paper; the interested reader can find more information about the measure in Navigli and Crisafulli (2010). We only mention here latent semantic indexing (LSI)

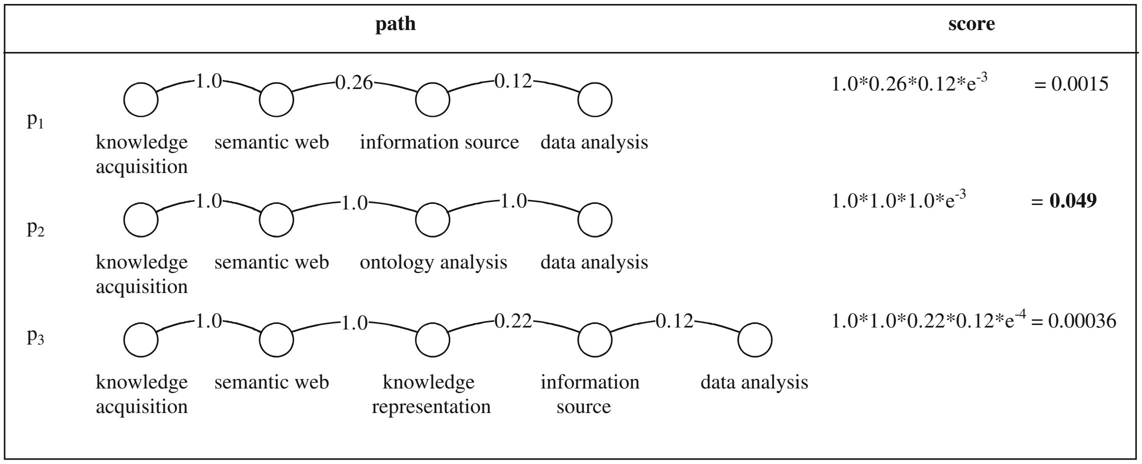


Fig. 4 The paths connecting text mining and data analysis, with their respective scores

(Landauer et al. 2007), one of the most popular methods not based on ontological knowledge used to detect similarity between words that do not actually co-occur. To justify our choice, we notice that the use of an explicit semantic model makes it possible to evaluate and assess detected term correlations, which is otherwise opaque with LSI. The patterns below, which were automatically extracted, provide a clear justification of a detected similarity value, as well as its source of evidence:

```
'business model' is 0.1547429395403686 related to 'process model':
[(business model → process model, 0.1547429395403686, type: dice)]
'business transaction' is 0.12180175491295145 related to 'mobile payment':
[(business transaction → transaction, 0.36787944117144233, type: ontology),
(transaction → mobile payment, 0.33109149705429813, type: glossary)]
```

In contrast, since latent similarity is derived mathematically, no justification is available for manual validation.

2.2 Topic detection

The second step towards the identification of the collaboration themes within a network is the detection of relevant topics. This is a *clustering* task: the objective is to organize concepts into groups, or clusters, so that the concepts within a group are more similar to each other than to concepts belonging to different clusters. The extracted topics are intended to represent the relevant *collaboration themes* in a SN.

There is a wide literature on clustering methods [see Jain et al. (1999) and Tan et al. (2006) for a survey and Russo (2007) for an in-depth summary of more recent techniques], but clustering remains an unsolved problem, that even leads to an impossibility theorem (Kleinberg 2002). At the level of existing algorithms, a scholar may quickly get lost in the "*bewildering forest of methods*" (Kleinberg 2002), optimizations and variants thereof, comparative experiments, etc. Even when focusing to text-related applications, such as document clustering and topic detection, an extremely large number of methods and variation over such methods have been used.[3] In topic clustering, the LDA method appears to have recently emerged (Mei et al. 2008; Zhou et al. 2006; Nallapati et al. 2008), but it has two major limitations as is pointed out in Ha-Tuc and Srinivasan (2008): one is *scalability*, as it is extremely expensive to run the model in large corpora (as we do); the other is the inability to model the key concept of *relevance* (how much a topic is directly related to the subject of a document or to a domain described through a corpus) that we use in the subsequent SNA (see Sect. 3.1). Thus, even when considering the specific properties of the application, the "*bewildering forest of methods*" does not thin out.

Clustering evaluation criteria are an open issue as well (Zhao and Karypis 2004): in Kovacs et al. (2005) it is experimentally shown that none of the proposed *internal clustering validity indices* reliably identifies the best clusters, unless these are clearly separated. The only objective way of evaluating a clustering result is *within a specific application* for which a gold standard is available. Unfortunately, there are no available standards for topic clustering.

To summarize, the literature does not offer any readily available solution, but at most "available" solutions, e.g. clustering test suites and implemented algorithm libraries.

To set a course on our clustering problem, we proceeded as follows:

1. First, we identified two available and popular clustering algorithms, namely K-means and repeated bisections. This allowed us to produce a collection of

[3] See http://www.iturls.com/English/TechHotspot/TH_DocCluster.asp for a list of text-related clustering applications.

clustering results with different settings of the clustering parameters (Sect. 2.2.1).

2. Second, we tried to clearly specify the desired properties of the clustering task at hand, given the application goals. This eventually led us to the definition of a validation methodology, which we named "*the efficient librarian*" criterion (Sect. 2.2.2).
3. Third, we used the validation methodology of step (2) to select the best clustering results.

The above procedure is then applied to our case of interest, the INTEROP research community, to detect the relevant network topics (Sect. 2.3).

2.2.1 Clustering algorithms

We employed two different clustering algorithms, for which widely used implementations are available (see Sect. 2.3.1):

- *K-means*. A well-known, widely used (even for text related applications) and fast clustering algorithm (Kanungo et al. 2002). It first defines k initial centroids, one for each cluster. Next, each vector in the data set is assigned to its closest centroid. The k centroids are then recalculated, vectors are reassigned and the procedure iterates until convergence. We use a standard k-means implementation with random seeding. For each value of k we perform 10 k-means clustering runs and take, as the final result, the one maximizing the efficient librarian metric proposed in Sect. 2.2.2;
- *Repeated bisections*. This algorithm first partitions the vectors into two groups based on the optimization of a particular criterion function; then one of these groups is selected and further partitioned, and so on. The procedure is repeated until a set of k clusters is output. The algorithm is claimed to produce high-quality clustering results by the authors (Zhao and Karypis 2005), though in a text-based application the results of this method are encouraging but not clear-cut (Purandare and Pedersen 2004).

The input to the clustering algorithms is a matrix of similarity vectors. In our experiments we used two different matrix of the same dimension for a comparison: X generated by using the semantic vectors and X' by using the LSI (see Sect. 2.1.2). Both matrices are sparse and high-dimensional, thus affecting the quality of the derived clusters, so we applied the singular value decomposition (SVD), the standard method used in LSI, to reduce the data sparseness problem (like in Kuhn et al. 2007).

The clustering process is articulated in four steps:

1. Computation of the similarity matrix X (or X').
2. Application of a rank reduction r to X (or X') by using the SVD method. Since X is a symmetric square $n \times n$ matrix ($n = |V|$), we obtain (Eckart and Young 1936) $X = U^r \Sigma^r U^{rT}$where U^r is a reduced $n \times r$ matrix, and $r \ll n$.
3. Use of the reduced matrix as input to our clustering algorithm.
4. Computation of a clustering C by applying the K-means (KM) or the repeated bisections (RB) algorithms to the set of feature vectors in U^r and requiring a number of output clusters k.

This procedure can be repeated by varying the two parameters k (number of clusters) and r (rank of the reduced matrix). Hereafter we indicate with C a set of clusters $\{C_h : h = 1,\ldots,k\}$, denoted also as clustering result, obtained by feeding the algorithm with either semantic vectors or with latent vectors, using either KM or RB clustering, and with some variable setting of the k and r parameters. We denote with $\wp$ the ensemble of clustering results obtained by varying the parameters. Our aim is to select the best clustering $C_{\text{BEST}} \in \wp$, according to the criteria described in the next section.

It has already been observed that cluster ensembles offer a solution to the challenges arising from the ill-posed nature of clustering (Domeniconi and Al-Razgan 2009), particularly the problem of parameter setting. However, current ensemble methods, when they come to combine the ensemble results in a consensus function, are again faced with the parameter setting problem.

2.2.2 The efficient librarian criterion for clustering evaluation

Given a set of parameter values, a clustering algorithm produces a result by using some *internal validity* measure to identify a cluster C_i that maximizes the intra-cluster similarity and minimizes the inter-cluster one. Many measures are defined in the literature (see Kovacs et al. 2005 for a comparative analysis), though none clearly emerged.

For different parameter sets, the best clustering output $C_{\text{BEST}} \in \wp$ is chosen either manually, or on the basis of some *external criterion*, which typically consists in validating the results against gold standard data sets. Such ideal data sets are not available for topic clustering, so we define a novel validity measure that is used a posteriori to select C_{BEST} among the set of clustering results $\wp$ coming from different parameter settings. Considering that clustering and clustering evaluation are not the main topics of this paper, in the following we only sketch out the defined criterion for clustering evaluation.

The measure is based on the definition of some desirable properties of the clusters we wish to obtain. To describe these properties, we use the metaphor of the *efficient librarian*.

The *efficient librarian* would like to define a suitable set of labelled shelves that satisfy

1. *Compactness*. Copies of the same book might be placed on different shelves, allowing for multiple classification. However, to reduce cost, the classification must be such as to minimize the need for multiple copies.
2. *Even dimensionality*. Books should be more or less evenly distributed in the various shelves.
3. *Generalization power*. If a set of new books arrives, it should not be necessary to add a new shelf with a new label, neither the properties (1) and (2) should vary significantly. The selected labels must be *predictive* of the library domain as much as possible.

Translating the librarian metaphor to the topic clustering domain, books are the set of *documents* shared among the members of a reference community and shelves are the *topics* (clusters) learned by a clustering algorithm.

Let D be the set of documents and V the collection of *domain concepts* (as defined in Sect. 2). Each document $d_i \in D$ is tagged with a subset of domain concepts $V_i \subseteq V$ (each weighted according to its document relevance) on the base of the occurrences of the corresponding terms in the document.

Measuring compactness. The *compactness* criterion requires that each document is described by few clusters (one, at the best), following the *Minimum Description Length* principle (Hansen and Yu 2001). Given a clustering result C, for any document d_i, we compute a *document vector* of k elements ($k = |C|$) representing the probability that the cluster $C_h \in C$ properly describes the document d_i. The probability is estimated by using the *normalized term frequency-inverse document frequency*, a standard measure (Baeza-Yates and Ribeiro-Neto 1999) of the relevance of a term t in a document d, applied to all terms $t_j \in C_h$ of the document d_i. We then compute the *normalized entropy* of each vector and, for a given clustering C, we can say that the best cluster with respect to the compactness is the one which minimizes mne(A,C), the *mean normalized entropy* of the matrix A of the vectors' elements defined for the set D.

Measuring even dimensionality. The *even dimensionality criterion* requires that each cluster $C_h \in C$ describes more or less the same number of documents: in other terms, that the probability distribution of the clusters in C, $p(C_h)$, is close to a uniform distribution.

We estimate $p(C_h)$ from the definition of $p(d_i|C_h)$ given by the Bayes theorem and the hypothesis of a uniform probability distribution for the documents in D (commonly used in Information Retrieval, see Fuhr 1992; Macherey et al. 2002; Chlia and De Wilde 2006). As for the compactness, we compute the *normalized entropy* of each C_h and the *mean normalized entropy* of C, ne(C). The best cluster with respect to the dimensionality is the one which maximizes ne(C).

To both maximize ne(C) and minimize mne(A,C), giving them equal relevance, we combine the two measures as follows:

$$c\&d(A, C) = \mathrm{ne}(C) - \mathrm{mne}(A, C) \tag{1}$$

where *c&d* stands for "*compactness and dimensionality*". Thus, the best clustering result C_{BEST} in the set $\wp$ is given by

$$C_{\mathrm{BEST}} = \arg\max_{C \in \wp}\{c\&d(A, C)\} \tag{2}$$

We note that C_{BEST} maximizes *c&d* in the set $\wp$; however, other clusterings (not in $\wp$) might exist for which *c&d* is higher.

Measuring generalization. Finally, we have to consider the *generalization criterion*. The method described in Sect. 2.1 takes a set of domain documents and extracts from them concepts, semantic chains and similarity vectors, which are the input to the clustering algorithms. The generalization criterion requires that, if we learn a clustering C with a given estimated value $c\&d(A,C)$ from a learning set D_1 of documents, and then we consider another set D_2 of domain documents not used for learning, then the value of $c\&d(A',C)$, where A' is a matrix obtained joining the matrices corresponding to D_1 and D_2, *should fall within the error bounds of the* $c\&d(A,C)$ *estimate*. We estimate the error bounds of the measure (1) by means of bootstrapped confidence intervals (see, e.g. Wood 2005).

2.3 Clustering experiments in the INTEROP domain

We used the methodology described in Sects. 2.1 and 2.2 to extract the relevant terms of the INTEROP community, to learn a set of clustering results, and finally to select the best result according to the efficient librarian criterion.

2.3.1 Experimental setup

We collected 1,452 full papers or abstracts authored by the about 400 INTEROP project members belonging to 46 organizations published from 2002 (before the start of the project, in 2003) until the end of the project, in 2007. Table 1 summarizes the results and data of the corpus analysis phase. The similarity vectors' computation was then applied to the extracted concepts, using semantic relations encoded in the INTEROP ontology[4], and the

[4] The INTEROP ontology can also be browsed at http://lcl.uniroma1.it/tav/choose.jsp.

Table 1 Summary data on corpus analysis

Number of analysed papers	1,452
Size of domain vocabulary V	1,426
Domain Ontology used	http://interopvlab.eu/backoffice/tav

co-occurrence relations extracted from the domain corpus and from the INTEROP glossary. Examples of similarity vectors in the INTEROP domain have already been provided in the related sections.

To run the experiments as described in the Sect. 2.2 and obtain our ensemble $\wp$ of candidate clusterings, we used

1. for SVD range reduction and LSI feature vectors, the Colt libraries for matrix computation in Java[5];
2. the CLUTO implementation of the repeated bisections clustering algorithm[6];
3. a standard implementation of the K-means algorithm available from the Weka tool suite[7].

Each experiment result C is labelled with a string composed as follows:

- SV/LV indicate that input feature vectors are semantic (SV) or latent (LV);
- KM/RB indicates the clustering algorithm used: KM stands for the K-means algorithm, and RB for repeated bisections; and
- k is an integer indicating the number of generated clusters

Hence, for example, *SV-RB-100* represents the clustering result obtained with semantic vectors, repeated bisections and $k = 100$. We experimented with $k \in \{50, 60, 70,\ldots,300\}$ and SVD range reduction with $r \in \{50, 75, \ldots, 200\}$, recall $n = |V|$.

Co-occurrence relations of Sect. 2.1 and the matrix A of Sect. 2.2.2 have been obtained using 80% of the documents in D (learning set), while the remaining 20% (test set) have been used to verify the confidence intervals in matrix A'. Therefore, clustering results are generated using only similarity vectors extracted from the learning set (however, we used the complete domain ontology to draw semantic relations in the graph G).

Finally, we artificially created three additional clustering outputs: the output named BASELINE is obtained by generating a cluster for each concept in V and hence $k = n$; the result named RANDOM is, as the name suggests, a random clustering of the terms in V; finally, MANUAL is a clustering which was semi-manually created by selecting an apparently "nice" clustering and adjusting it manually, based on our knowledge of the domain.

2.3.2 *Analysis of the data*

For each clustering result we evaluated compactness and dimensionality by computing the $c\&d(A,C)$ value, plotted in Fig. 5 (for sake of readability, the x-axis shows only some of the 80 clustering experiments we made). To evaluate clustering results according to the generalization criterion we estimated the confidence intervals at 95% using 10,000 samples from the learning set. Next, we computed $c\&d(A',C)$ on the complete set of documents. We evaluate the confidence intervals on A' rather than only on the test set, because according to the librarian criterion, we wish to verify that the existing shelves can properly classify the new books. In any case we have successfully verified that the confidence intervals are respected also by the test set alone. All the clustering results turned out to fall within the confidence interval. In Fig. 6 clustering results are ordered according to the difference $|c\&d(A,C)–c\&d(A',C)|$.

In Fig. 5 the best clustering according to the $c\&d$ criterion is *SV-RB*-50; however, looking the generated clusters at detail, there are evident commonalities among "close" results of the same type. Below, we show some of the generated clusters in *SV-RB*-50, where we manually assigned a label in bold to each cluster to suggest the subsumed research theme.

```
interaction C6:
{social interaction,inter-enterprise
collaboration,database interaction,business
interaction,interaction,multimedia interaction}

e-services C9:
{e-service,service-based architecture,e-finance,
web mining,e-banking,e-government,e-booking}

requirements C11:
{conformance requirement,engineering
requirement,business requirement, configuration
requirement,traceability,testing requirement,
class-based representation formalism,language
specification,integration requirement,production
cycle,organisation
requirement,requirement,testability,security
service,user requirement,tolerance,manufacturing
planning,consistency checking}

interop-problem C29:
{interoperability barrier,representation
interoperability,interoperability problem,
ERP interoperability,organisation
interoperability,Interoperability
measurement,financial institution,intra organisation
integration,knowledge interchange format,e-government
interoperability,enterprise integration
framework,organizational barriers,
enterprise integration}
```

Generally, RB performs better than KM and SV better than LV according to our criterion, but for some parameter settings, there are LV results that perform better than the corresponding SV clustering results (see the central

[5] http://acs.lbl.gov/~hoschek/colt.

[6] http://www.cs.umn.edu/~karypis/cluto.

[7] http://www.cs.waikato.ac.nz/ml/weka.

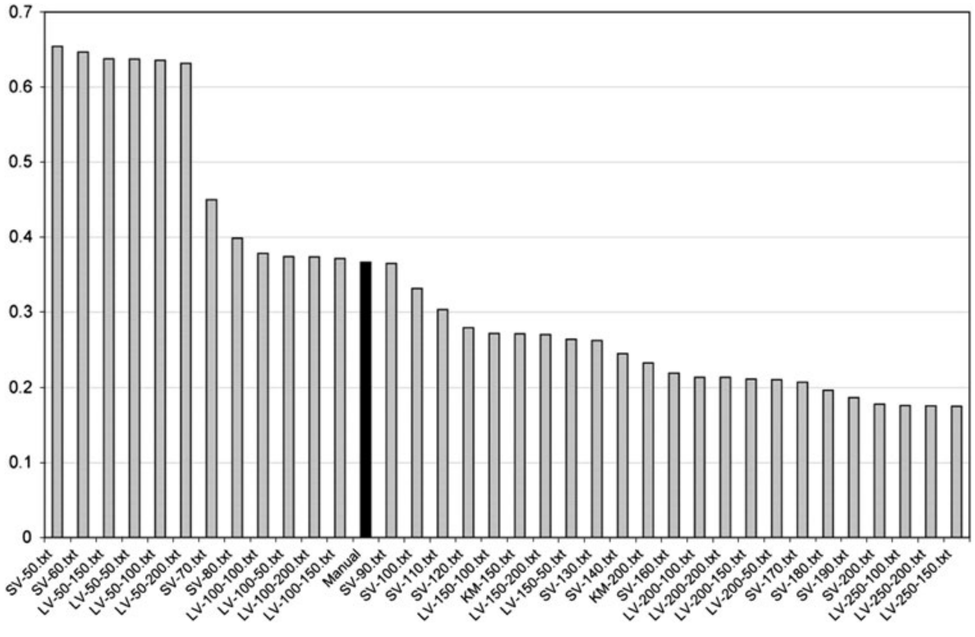


Fig. 5 Clustering results ordered according to compactness and even dimensionality

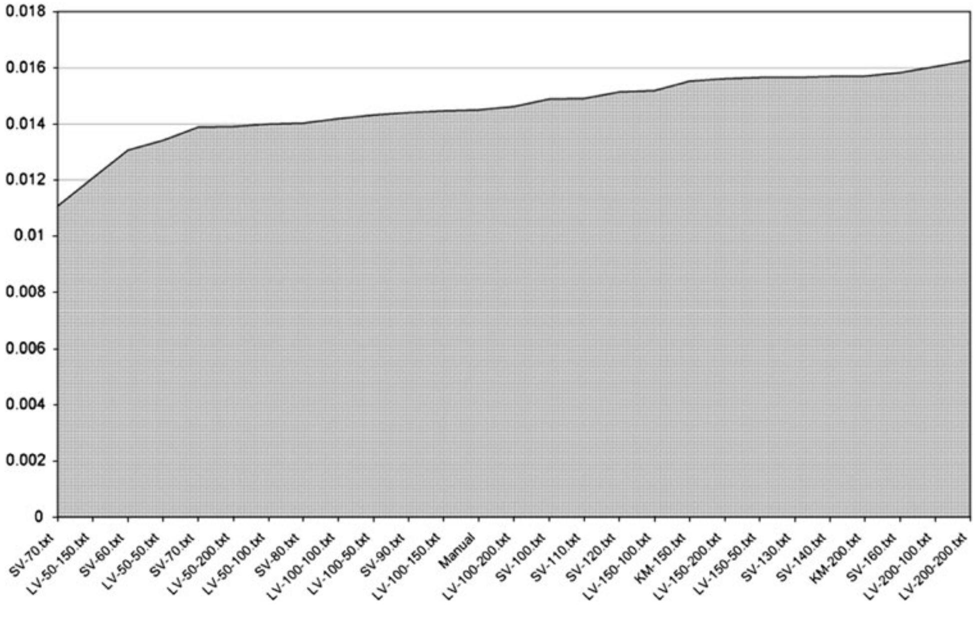


Fig. 6 Results ordered according to generalization

segment of the *x*-axis). It is very interesting to notice that MANUAL is not the best (in other applications manual clusters are considered to be the reference). In fact, humans tend to aggregate concepts according to semantic similarity, but in topic clustering co-occurrence relations (which are hardly captured by humans) play a relevant role: the goal of topic clustering is to find out how concepts, possibly from different areas, aggregate and contribute to define new research themes. For example, a human would find clusters C6, C9 and C11 very plausible, because of the intuitive semantic closeness of the cluster members, but he would not be able to imagine that, for example, *organizational barriers* and *enterprise integration* are related (C29). The connection arises from several papers in the corpus having these concepts logically connected (being the first a barrier for the second). Finally, the RANDOM and BASELINE experiments do not appear in the Figure, which only shows the top-ranked experiments. The BASELINE lies in position n.40, RANDOM in position n. 46.

Regarding Fig. 6, it appears that the dynamics of the *generalization* measure is limited: when computing $c\&d(A',C)$ for the complete document collection, almost all the clustering results produce a value falling within the confidence intervals of the $c\&d(A,C)$ estimate which was obtained with the reduced collection. Even when ordering the results according to the minimum distance between the *c&d* values (see previous section), the difference is minimal. This can be explained by the fact that both LSI and semantic similarity have an implicit generalization power: connection paths are established even between those pairs of terms for which a co-occurrence relation is not found in the learning set of documents (as intuitively illustrated in Fig. 2 for the case of semantic co-occurrence). In conclusion, the *c&d* criterion ((2) in Sect. 2.2.2) appears to be more helpful than the generalization criterion (at least if an ontology is available) for the purpose of selecting the best clustering result.

Finally, when comparing the best "semantic" and "latent" clusters (LV-50-150), it is blatantly clear that the main aggregation criterion is contextual relatedness, rather than semantic relatedness. However, this is not per se a drawback. From a qualitative point of view, the real advantage of semantic vectors/clusters is the fact that inter-concept relations can be inspected and justified.

2.4 Another experimental domain: the MIUA community

To test the validity of our method for the analysis of research communities based on social networks, which is described in Sect. 3, we considered a second application domain: the United Kingdom research community in MIUA. This is a multidisciplinary community of experts in computer science, engineering, physics, clinical practice and fundamental bioscience interested in image analysis applied to medicine and related biological sciences. We chose this community as a second test-bed of our analysis for three main reasons:

- It is a community of researchers focused on a narrow technical domain, who express their competences through publicly available papers;
- it is an open community, whose members change over time;
- its governance has goals similar to those of a NoE (to define the main topics of the research area, to let the people share the results of their research, to spread competencies throughout the community, thus facilitating the collaboration among researchers), but it is not mainly focused on short-term results.

While the first characteristic is in common, the others mark relevant differences between the INTEROP and the MIUA communities, thus giving us the opportunity to check our method of analysis in significantly different domains.

We applied the same methodology for the best cluster definition described in the previous sections to a corpus of 157 full papers, written by 372 researchers, collected from three editions of the MIUA community conference, held in 2006, 2007 and 2008. Both resources needed for concept extraction, the glossary and the ontology (see Sect. 2.1), are derived from the medical subject headings (MeSH) thesaurus[8], a resource maintained by the US National Library of Medicine and used for indexing articles from biomedical journals for the MEDLINE/PubMED database. Mainly focused on the biomedical field, MeSH contains also terms strictly related to image processing (like *computer-assisted image interpretation*, *computer-assisted image processing*,…), thus covering many of the terms in the MIUA papers. We collected the terms' definitions of MaSH into the MIUA glossary, while a simple ontology has been obtained by transforming all the hierarchical links among terms in the thesaurus into *is–a* chains.

3 Social network analysis

The previous sections have been dedicated to the presentation and the evaluation of a methodology to capture the relevant research themes within a scientific community. The results of this methodology are then used to perform a new type of SNA, the CB-SN analysis, which is described in this section.

To enable CB-SN analysis, we have used both well-established and novel SN measures. Furthermore, we have developed a visualization tool to support the analysis and to verify its efficacy in real domains. We refer here to the specific application to *research communities* (both the INTEROP European Union NoE) and the MIUA community), but the approach is general, as long as written material is available to model the relationships among the actors, and it is applicable to different domains, by choosing the most appropriate social network measures according to the phenomena we wish to investigate.

Thanks to the rich documentation available for INTEROP NoE, which is related not only to the goals of the initiative but also to its governance, we have been able to evaluate the efficacy of our analysis in this domain:

- with reference to the main network *monitoring objectives* (which are common to all NoEs and, in general, to research funding bodies); the kind of information

[8] http://www.nlm.nih.gov/mesh/.

required for network monitoring cannot be extracted using traditional SN models, except for the mere counting of common publications;

- through a *qualitative comparison* of automatically derived findings with known, manually derived information reported in the INTEROP NoE evaluation reports and in the INTEROP knowledge repository (Velardi et al. 2007);
- through a *quantitative comparison* between a partner's similarity measures obtained by automatic topic extraction and the ones based on terms that the partners have manually inserted in the INTEROP Competency Map to describe their interests.

Even though the same information is not available in the MIUA case, the similarity between the domains (both are research communities mainly aimed at facilitating knowledge sharing and contacts among researchers) led us to apply the same type of analysis, with results partially in common with the ones obtained for the INTEROP NoE and in part different, as we will show in the following sections.

This section is organized as follows: first, we describe how we model the CB-SNs (Sect. 3.1). Next, we introduce and motivate the SN measures that will be used to perform the networks study (Sect. 3.2). Then, we briefly introduce the graphic tool that we have developed and present several examples of analyses that can be conducted with our model (Sect. 3.3). Finally, Sect. 3.4 presents a quantitative evaluation of the INTEROP domain, and 3.5 summarizes our findings.

3.1 Modelling the content-based social network

In this section we describe the process used to model the INTEROP CB-SN. The same process has been applied to the MIUA domain, being the only differences in the nature of the network vertices (in the first case they represent the 46 INTEROP research groups, in the second case the 378 researchers of the MIUA community) and in the definition of the research topics (the set of cluster C_{BEST}).

The INTEROP CB-SN has been modelled as a graph $G_{\mathrm{SN}} = (V_{\mathrm{SN}}, E_{\mathrm{SN}}, w)$, whose vertices V_{SN} are the community research groups g_i and whose edges E_{SN} represent the content-based social relations between groups, weighted by a function w: $E_{\mathrm{SN}} \rightarrow [0, 1]$. In the following we describe how to populate the set E_{SN} and define the weight function w of a CB-SN G_{SN}.

First, we need to model the knowledge of a research group in terms of the *research topics* (i.e. the clusters C_h in C) that better characterize the group activity. With C we now denote for simplicity's sake the C_{BEST} selected according to (2) in Sect. 2.2.2. The objective is to associate with each research group g, on the basis of its publications grouped in a collection D of documents and on C, a k-dimensional vector whose hth component represents the *relevance* of topic C_h with respect to the group's scientific interests. For each document $d_i \in D$, we compute a k-dimensional vector w_i of elements z_{ih} (with $k = |C|$) such that

$$z_{ih} = \frac{l_{h,i}}{|C_h|} \sum_{j:x_j \in C_h} \mathrm{ntf} - \mathrm{idf}(t_j, d_i)$$

where x_j is the similarity vector associated with concept t_j (as defined in Sect. 2.2), $l_{h,i}$ is the number of concepts of C_h found in d_i, and ntf–idf() is the *normalized term frequency–inverse document frequency*, we have also used for *measuring compactness* in Sect. 2.2.2. Therefore, each z_{ih} in w_i estimates the overlap of d_i with the topic $C_h \in C$.

Then, in the INTEROP case, given a research group $g \in V_{\mathrm{SN}}$, we define a k-dimensional vector p_g, which is the centroid of all document vectors associated with the publications of the group g. In the MIUA case, the same vector is defined for each author. In both cases, if a document (a paper) is authored by different researchers (MIUA) or by members of different groups (INTEROP), it contributes to more than one centroid calculation. We determine the similarity between pairs of groups g, g' by the *cosine* function (Salton and McGill 1983):

$$\cos - \mathrm{sim}(g, g') = \cos(p_g, p_{g'}) = \frac{p_g \cdot p_{g'}}{|p_g||p_{g'}|}$$

For each pair of groups g, $g' \in V_{\mathrm{SN}}$, if cos-sim(g,g') > 0 we add an edge (g,g') to E_{SN} with a corresponding weight $w(g,g')$ = cos-sim(g,g'). We do not filter any edge at this stage (for example by defining a threshold to eliminate the low $w(g,g')$ value edges) leaving all the filtering actions to the network analyst in the phase of graph analysis through the GVI tool. As a result, we are now are able to characterize the scientific interests of research groups, as well as interest similarity among them. To perform an SN analysis of the G_{SN} network, we now need to define a set of appropriate social network measures.

3.2 Social network measures

In the field of SNA, different measures have been defined to delve into the networks' characteristics. They are generally focused on the density, dimension and structure of a network (mainly in terms of its relevant sub-elements), and on the role and centrality of its nodes (Scott 2000). For the purposes of the analyses we intend to carry out on research communities, described in Sect. 3.3; the following SNA measures, defined in Wasserman and Faust (1994), have been selected:

- *Degree centrality of a vertex* A measure of the connectivity of each node $g \in V_{SN}$:

  $$\mathrm{DC}(g) = \deg(g)$$

  where $\deg(g)$ is the degree of g, i.e. the number of incident edges.
- *Average degree centrality* A network global measure of the interconnections among the nodes:

  $$\mathrm{ADC} = \frac{\sum_{i=1}^{M} \mathrm{DC}(g_i)}{M(M-1)}$$

  where M is the number of the network nodes (i.e. $M = |V_{SN}|$).
- *Variance of degree centrality* A network global measure of the $\mathrm{DC}(g)$ dispersion:

  $$\mathrm{VDC} = \frac{\sum_{i=1}^{M} (\mathrm{DC}(g_i) - \mathrm{ADC})^2}{M}$$

- *Weighted degree centrality of a vertex* A measure of the connectivity of each node that takes into account the edges' weight:

  $$\mathrm{DC}_w(g) = \sum_{e \in E_g} w(e)$$

  where $E_g \subseteq E_{SN}$ is the set of edges incident to node g, and $w(e)$ is the weight of the edge e, as defined in the previous section.

We have used the above measures to investigate how to facilitate the governance of research networks in Sect. 3.3.1.

In addition, we have defined a new measure to trace the evolution of a network over time (Sect. 3.3.2) through the analysis of its *connected components*. We shall now introduce the concepts on which this measure is based.

Given two nodes of a network, a and b, b is reachable from a if a path from a to b exists. If the edges are not directed the *reachability* relation is symmetric. An undirected network is connected if, for each pair of its nodes a and b, b is reachable from a. If a network is not connected, a number of maximal connected sub-networks can be identified: such sub-networks are called *connected components* and NCC denotes their number in a given network.

Closely related to the connected components is the concept of *bridge*. A *bridge* is an edge in the network that joins two different connected components (Scott 2000). By removing a bridge, the number of connected components NCC of a network increases by one unit, as shown by the example of Fig. 7.

If we represent the set of connected components of the network after the removal of a bridge e as $\mathrm{NewCC}(e)$, the following measure can be defined for our network G_{SN}:

- *Bridge strength* The dimension, associated with each bridge e, of the smallest connected component in the graph obtained by removing e from G_{SN}.

  $$\mathrm{BS}(e) = \underset{cc \in \mathrm{NewCC}(e)}{\arg\min} |\mathrm{Nodes}(cc)|$$

where $\mathrm{Nodes}(cc)$ is the set of nodes in G_{SN} belonging to the connected components.

With respect to our social network, the $\mathrm{DC}(g)$ and $\mathrm{DC}_w(g)$ measure the potential for collaboration of each community member: in the first case, by considering only the presence of common interests between nodes and in the second, by taking into account also the similarity values. The ADC is a global measure of how far the members share the potential research interests (i.e. the cohesion of the research community), whilst the VDC measures their dispersion (i.e. how homogeneous the community members are with respect to the number of shared interests). Finally, $\mathrm{BS}(e)$ can be used to estimate how *"resilient"* the community is, i.e. formed of members that widely share research interests and that are not part of sub-communities loosely interconnected and focused on specific topics. This can be done by characterizing and tracing the evolution of those subnets that are coupled to the rest of the network through a single bridge over a period of time.

To analyse more in depth the relevant phenomena inside the research community modelled by the social network defined above, we enriched the graph components (i.e. nodes and edges) in G_{SN} with an associated data set including both the values of the SN measures defined in this section and other data useful for their characterization. For each node, this data set includes a unique identifier (the name of the INTEROP groups or the MIUA researchers), the number of researchers in each INTEROP group, the

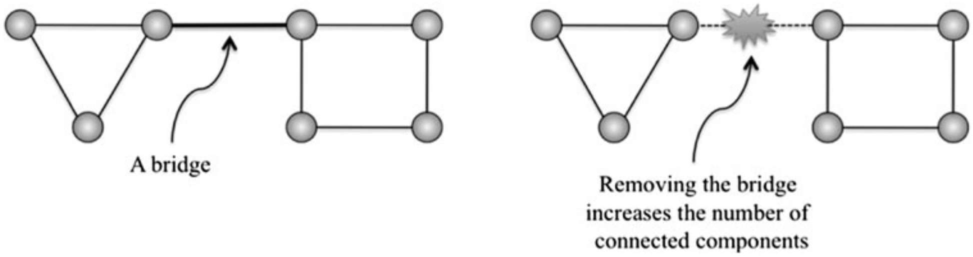

A bridge
Removing the bridge increases the number of connected components

Fig. 7 Bridges and connected components

total number of publications produced in a given time interval and the number of them co-authored by other members of the research community. In the same way each edge, along with the value of the similarity between the nodes it connects, is associated with the set of concepts that contribute to the similarity value.

This additional information has no direct influence on the network topology (in its basic formulation, only similarity relations between nodes have been considered), but it may be used for the analysis of the dependencies among the network structure and the data associated with its elements. In this way, for example, the correlation between the number of publications of a community member and the properties of its direct neighbours can be investigated, or a filter by topic can be applied to the similarity relations of a single member, as detailed later.

Before actually providing an in-depth example of CB-SN analysis using the model described so far, we briefly introduce a graphic tool we have developed to support visual analysis of a CB-SN.

3.3 Analysis of a research network

The purpose of this section is to show that our content-based SN model allows it to perform an analysis that is far more informative and useful than in traditional network models. The previously defined SN measures were applied to analyse different network configurations, each focused on a different aspect of the research community. They have been built by using various sets of members' publications, relations and attributes associated with nodes and edges. These measures can be used both to highlight any relevant phenomena emerging from the social structure of a research community and to analyse its evolution.

In order to analyse the networks produced, we built up Graph VIewer (GVI), a JUNG[9] library based software tool able to support the incremental analysis of a network by acting directly on its graphic representation. The core of GVI is the mechanism of incremental filtering: at each step of the analysis, some elements of the network (nodes and edges) can be selected according to some filter conditions on the values of the elements' attributes. The selected items can then be hidden, if not relevant for the analysis to be made, or highlighted, by changing their graphic representation (shape and colour). GVI allows both the creation of simple filters on a single attribute and their combination through AND/OR operators. Another relevant feature of the application is the capability to set the dimensions of the nodes and the edges according to the value of one of their attributes. For example, the thickness of the edges can be a function of the similarity value between the nodes (the thicker the line, the higher the similarity), or the size of the shape representing a node can be related to the associated DC measure (the greater the size, the higher the DC). This feature is extremely useful, because it gives a graphic representation of the distribution of relevant parameters selected by the analyst out of the entire set of network elements.

The examples of analysis provided in the next subsections are taken both from the INTEROP NoE and the MIUA community. As already remarked, in the first case detailed data on the community members and publications are available on the NoE collaboration platform; we were able to verify then the actual correspondence of any phenomenon highlighted in the following sections with respect to the reality. In the MIUA case, due to its different and less structured nature, such a correspondence can only be evaluated on the basis of the general knowledge of the characteristics of research communities.

In both cases, the results described in the following sections clearly show the ability of the CB-SN model to reveal the shared interests, the active and the potential collaborations among the members of the community and their evolution over time.

3.3.1 Facilitating governance of research networks

To monitor the evolution of Networks of Excellence like INTEROP and verify that collaboration targets are indeed met, precise types of indicators are usually manually extracted by the network governance committee from available data. We show here that these types of indicators can be automatically and more precisely extracted by our CB-SN model and tools.

Relevant types of information to support the governance of research networks are

- potential areas for collaboration between groups that share similar interests (but do not cooperate), to avoid fragmentation and duplication of results;
- popularity of research topics, to redirect, if necessary, the effort on potentially relevant but poorly covered topics;
- evolution of collaborations over time, to verify that governance actions actually produce progress towards research de-fragmentation.

This information is also useful to analyse the characteristics of a research community like MIUA, such as the competences, the shared research interests of its members and the level of collaboration among them (measured through the paper co-authorship).

In the following we show four possible types of analysis that can be carried out on the graphs representing the two communities:

[9] http://jung.sourceforge.net.

1. potential research collaboration between members (on a single topic or on all topics);
2. co-authorship relations;
3. real versus potential collaborations;
4. delving into members' competences.

The obtained results are conceptually similar for the two domains thus: for sake of brevity, we will show and comment the data of MIUA only for the analyses of types 3 and 4.

Potential research collaboration on all topics. To evaluate the fields of potential research collaboration between groups, the network of similarity was built by considering all the 1,452 publications written by the INTEROP community members, and the $\mathrm{DC}_w(g)$ for each research group was calculated. Figure 8 is a graph that shows the subnet of the global network obtained by selecting only the edges for which cos-sim(g_i,g_j) is equal to or greater than a given threshold (set to 0.97, but manually adjustable by the social analyst). We note that such a threshold is used for visualization purposes only: in this case to highlight the groups with the strongest similarity of research interests. In the figure, the dimension of the nodes and the thickness of the edges are related, respectively, to the $\mathrm{DC}_w(g)$ and the cos-sim values.

The biggest nodes represent the community members that have the highest potential in joint researches, whereas the thickest edges reveal the best potential partners. The GVI allows also the visualization of the topics involved in the similarity relations by clicking on the corresponding edges.

Potential research collaboration on a given topic. Another analysis we carried out was the selection of a topic and the visualization of all INTEROP groups with research interests involving that topic. Starting with the global network of the previous analysis we selected, as an example, the subnet of all nodes sharing the cluster *C9* of C_{BEST}.

C9: {e-service, service-based architecture, e-finance, web mining, e-banking, e-government, e-booking}.

The result is a complete graph (each node is connected with all the others) of 24 out of 46 research groups that share a common research interest on the topic (Fig. 9a). We then filtered the graph by selecting the edges with cos-sim(g_i,g_j) $\geq$ 0.97, to highlight a higher potential collaboration between groups involving that topic. Figure 9b shows the subnet of filtered potential collaborations where the thickest edges reveal the best potential links.

Co-authorship relations. We used the SNA approach to give an insight into the real research partnerships among the INTEROP groups. We modelled such relations through a "traditional" co-authorship network, where the edge between each pair of nodes has an associated weight that is the normalized number of papers co-authored by the members of the two groups. This value, $\mathrm{CP}_{\mathrm{norm}}(i,j)$, is defined as

$$\mathrm{CP}_{\mathrm{norm}}(i,j) = \frac{\mathrm{CP}(i,j)}{\min(P(i), P(j))}$$

where CP(i,j) is the number of publication co-authored by the members of groups i and j, $P(j)$ is the number of publications of group j and min($P(i)$,$P(j)$) is the smallest value between $P(i)$ and $P(j)$.

In this way $\mathrm{CP}_{\mathrm{norm}}(i,j) = 1$ expresses the condition of maximum possible co-authorship between two groups (i.e. one group has all its publications co-authored by the other). Figure 10 shows the network obtained by considering the papers co-authored by researchers belonging to different groups, in which the thickness of the edges is proportional to the $\mathrm{CP}_{\mathrm{norm}}(i,j)$ and the dimension of the nodes to the DC(g). In the figure, it is possible to see "at a glance"

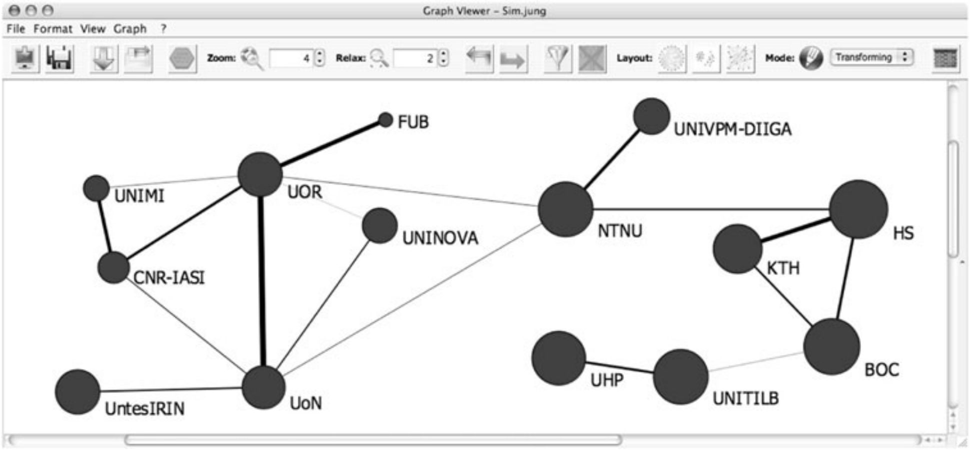


Fig. 8 Graphic representation of DC_w in a subnet of strongly related nodes

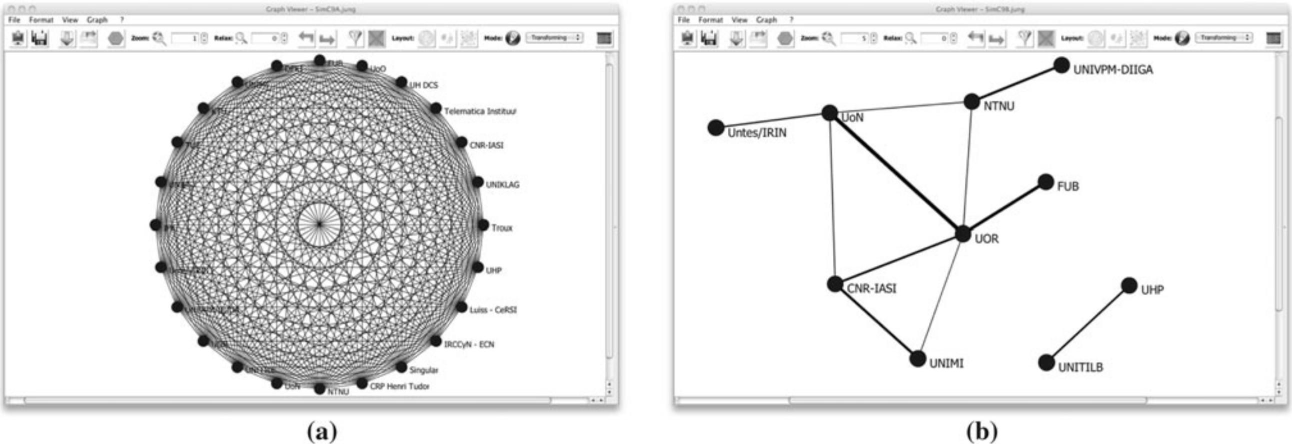


Fig. 9 A subnet of groups sharing the same research interest (**a**) and a highlight of the highest potential collaborations among them (**b**)

(biggest nodes) the groups that have the highest number of co-authorship relations with the others, and the pairs of groups that have a high number of papers co-authored (thickest edges), i.e. groups that have a strong collaboration in research activities.

Real versus potential collaborations. By combining co-authorship and interest similarity data (e.g. network types like those in Figs. 8 and 10), the network analyst can identify those groups that have many research topics in common, but do not cooperate. This kind of analysis has proven to be very useful in the diagnosis of the research networks, like INTEROP, where one of the main objectives was to improve collaboration and result sharing among partners. We conducted the analysis on a variant of the network used for potential research collaboration (Fig. 8) that was built by adding a second type of edge representing the number of papers co-authored by the groups' members. Figure 11 shows the network as visualized by GVI, after the application of a filter that selects the similarity edges having cos-sim(g_i,g_j) $\geq$ 0.97 and the co-authorship edges with *co-authored_papers* $\geq$ 3. These two thresholds have been experimentally selected (however, as previously remarked, thresholds are user adjustable) to focus the attention on the strongest correlations between groups. In the Figure, the curved lines are used for the co-authorship relation and the bent lines for the similarity relation. Furthermore, the line thickness is proportional to the value of the corresponding relation and the node dimension to the number of publications of the associated group.

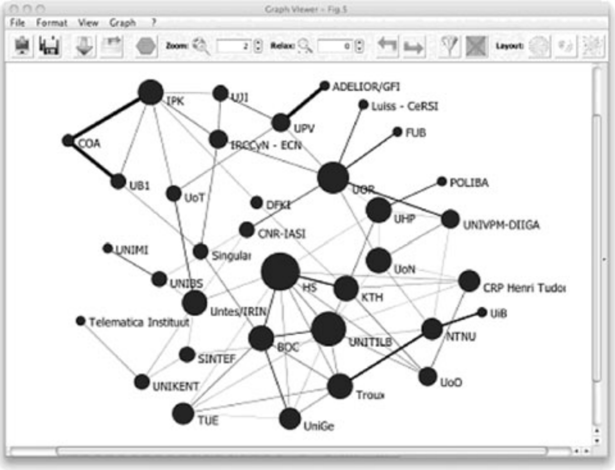


Fig. 10 The "traditional" co-authorship social network of INTEROP members

The graph clearly shows the groups which have many common interests (high similarity value) but few papers co-authored (CNR-IASI and UNIMI, HS and BOC, UNITILB and UHP,…), as well as the groups which have few common interests but many co-authored papers (KTH and UNITILB, KTH and UHP, UOR and UNIVPM-DIIGA,…).

In the latter case, we must remember that the absence of a similarity link between two nodes does not mean that the similarity value between them equals to 0.0, but that the value is lower than 0.97. If we remove the similarity threshold, the network of similarity is fully connected (46 nodes and 1,035 edges). Another element that justifies this apparent incongruence is that nodes in a graph represent research groups of variable dimensions. Large groups have variegated interests; therefore, they might have strong co-operations concerning only a small subset of their competences. For the activity of monitoring and stimulating the integration of the members of a research network like INTEROP, this analysis is highly interesting. It reveals the set of groups which have good potential to collaborate, but who probably do not have adequate knowledge of their common interests, and the set of groups who are very clever at maximizing the chances of a partnership.

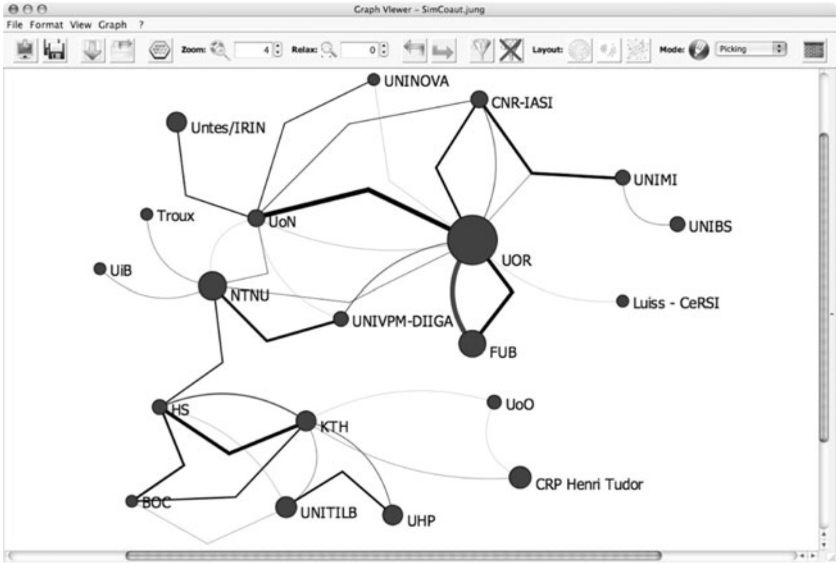


Fig. 11 The INTEROP co-authorship + similarity network

Another interesting phenomenon shown by the graph is the presence of a ‘clique’, a well-known concept in graph theory defined as “a subset of graph nodes in which every node is directly connected with the others in the subset, and that is not contained in any other clique” (Scott 2000). If we focus on similarity relations (the bent lines in the graph), for example, the groups KTH, HS and BOC form a clique. Each of them has some research topics in common with the others, so that the clique can be seen as a sort of research sub-group. On the contrary, observing the co-authorship relations, we see that UntesIRIN is not related to the other nodes, and this is strong evidence of a poor collaboration between groups whilst having a strong similarity in their research interests.

A similar analysis has been conducted on the MIUA community. In this case, as the first step of the graph processing, we have filtered out all the authors with few publications in the collection of the community conferences. By considering that the average value of papers written by a single author in the 3 years of the MIUA conferences is 1.5, and that the variance of the values is 1.7, we have filtered out all the authors with less then 3 published papers.

Figure 12 shows the graph obtained after the application of a second filter that selects the similarity edges having cos-sim(g_i,g_j) > 0.2 and the co-authorship edges with *co-authored_paper* > 1. These values have been experimentally selected to hide the nodes with the lowest similarity in research interests and to remove one-off co-authorships. Curved and bent lines, thickness of edges and node dimension assume the same meanings as in Fig. 11.

Like in the INTEROP graph, we can see authors having common interests but few co-authored papers (Crum and Taylor, Fox and Brady, Hill and Brady,…), as well as authors with few common research interests but more than one co-authored paper (Cootes and Twining, Cootes and Taylor) more able to take advantage of the collaboration opportunities. The Figure also shows a ‘clique’ of three authors, formed by Crum, Hill and Fox. All these findings can be used to characterize the research community as well as to define actions to stimulate the collaboration among the researchers.

Delving into members’ competences. The GVI capability to display and give support to the analysis of *bipartite graphs* (i.e. graphs having nodes belonging to two disjoint sets) gave the analyst the opportunity to delve deeply into the INTEROP groups’ competences. Consider the case in which she/he is interested in studying the impact of a specific topic, e.g. *information systems*. In C_{BEST} there is no single cluster which groups all the concepts related to *information systems* (e.g. *information system development, enterprise information system*, etc.).

It is rather obvious that there is no single way to group concepts together: our clustering algorithm uses co-occurrence and ontological relations as a cue for creating topics, but a social analyst may wish to investigate the relevance of a research area that is not necessarily represented in a compact way by a single cluster. To support this analysis, the GVI provides the capability of applying a filter to the network, to hide all the nodes associated with those topics not having any component concept with a name that contains *information_system* as a substring. The result is the bipartite graph shown in Fig. 13.

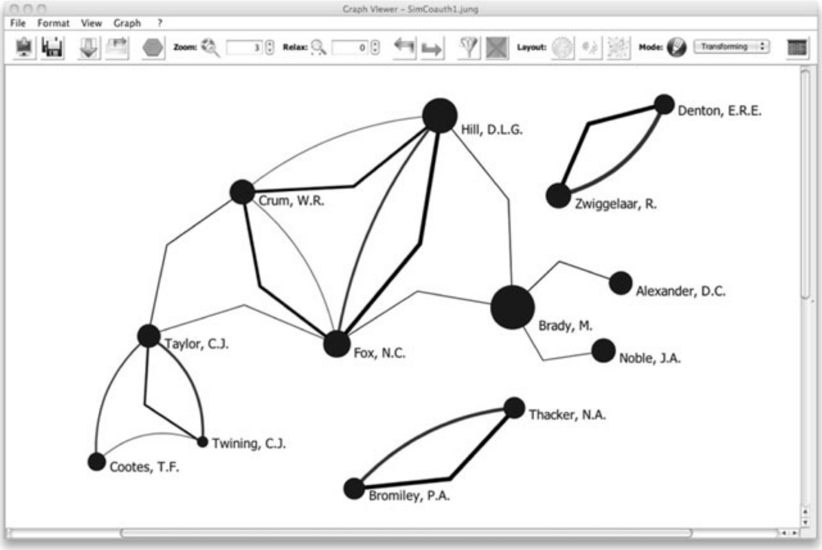
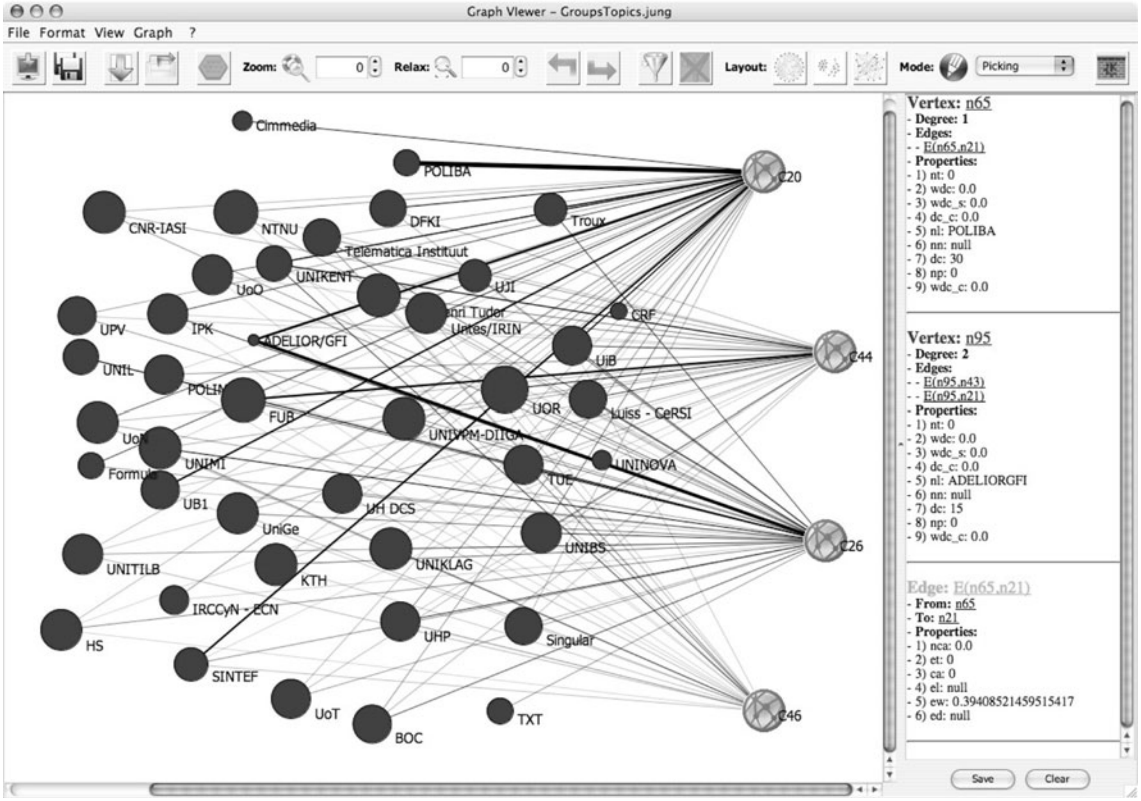


Fig. 12 The MIUA co-authorship + similarity network

The network contains only three topics, corresponding to clusters C20, C26, C44 and C46 (the nodes with different shapes on the right-hand side of the figure), the only ones including concepts related to *information system*:

- **C20**: {…,mobile information system, information system validation, information system language, information system analysis, information system interface, information system reengineering,…}



```
Vertex: n65
- Degree: 1
- Edges:
- - E(n65,n21)
- Properties:
- 1) nt: 0
- 2) wdc: 0.0
- 3) wdc_s: 0.0
- 4) dc_c: 0.0
- 5) nl: POLIBA
- 6) nn: null
- 7) dc: 30
- 8) np: 0
- 9) wdc_c: 0.0

Vertex: n95
- Degree: 2
- Edges:
- - E(n95,n43)
- - E(n95,n21)
- Properties:
- 1) nt: 0
- 2) wdc: 0.0
- 3) wdc_s: 0.0
- 4) dc_c: 0.0
- 5) nl: ADELIORGFI
- 6) nn: null
- 7) dc: 15
- 8) np: 0
- 9) wdc_c: 0.0

Edge: E(n65,n21)
- From: n65
- To: n21
- Properties:
- 1) nca: 0.0
- 2) et: 0
- 3) ca: 0
- 4) el: null
- 5) ew: 0.39408521459515417
- 6) ed: null
```

Fig. 13 The bipartite graph of the INTEROP groups involved in "information system" research

- **C26**: {…,information system specification, mobile web information system, information system integration, information system model, information system verification, information system, information system development, multilingual information system,…}
- **C44**: {…,cooperative information system,…}
- **C46**: {…,enterprise information system,…}

In this way, by using the GVI filtering capability, it was possible to highlight all the groups which have competences related to the *information_systems* field. By using the GVI options that make the thickness of the edges proportional to the value of the associated weights and the dimension of the nodes representing groups to their DC(*g*), one more relevant fact comes out by observing Fig. 13. Two groups (POLIBA and ADELIOR/GFI) have a strong competence in topic C20 (the weight of the edge between POLIBA and C20 is 0.39, as reported in the text pane on the right of the figure, which shows the characteristics of any graph element selected by the user), although ADELIOR/GFI has a narrower set of competences with respect to POLIBA (its node size is smaller, corresponding to a DC(*g*) of 15, the node's dc property in the text pane), and it is more focused on the topic.

The same analysis has been carried out on the MIUA data. The term *imaging* has been selected as an example of a general concept an analyst wishes to investigate with respect to its relevance for the community members. The term is not present in any of the clusters representing the topics of this domain, but four terms which can be considered as its specialization are included in the following clusters of C_{BEST}:

- **C2:** {…,echo-planar imaging,…}
- **C5:** {…,three-dimensional imaging,…}
- **C11:** {…,diagnostic imaging,…}
- **C43:** {…,radionuclide imaging,…}

Figure 14 shows the bipartite graph of relations between the members involved in research activity on *imaging*. The representation used here (thickness of the edges and dimension of the nodes) is the same as in Fig. 13. As for the INTEROP case, the figure shows some researchers that have a similar level of competence on a given topic (for example, Brady, Hill and Arridge on topic C43) but narrow or wide set of competences (Brady vs. Hill or Brady vs. Arridge,…), being more or less focused on the topic.

3.3.2 Network evolution over time

Shared research interests. By using the ADC and VDC measures it is possible to estimate the evolution of the

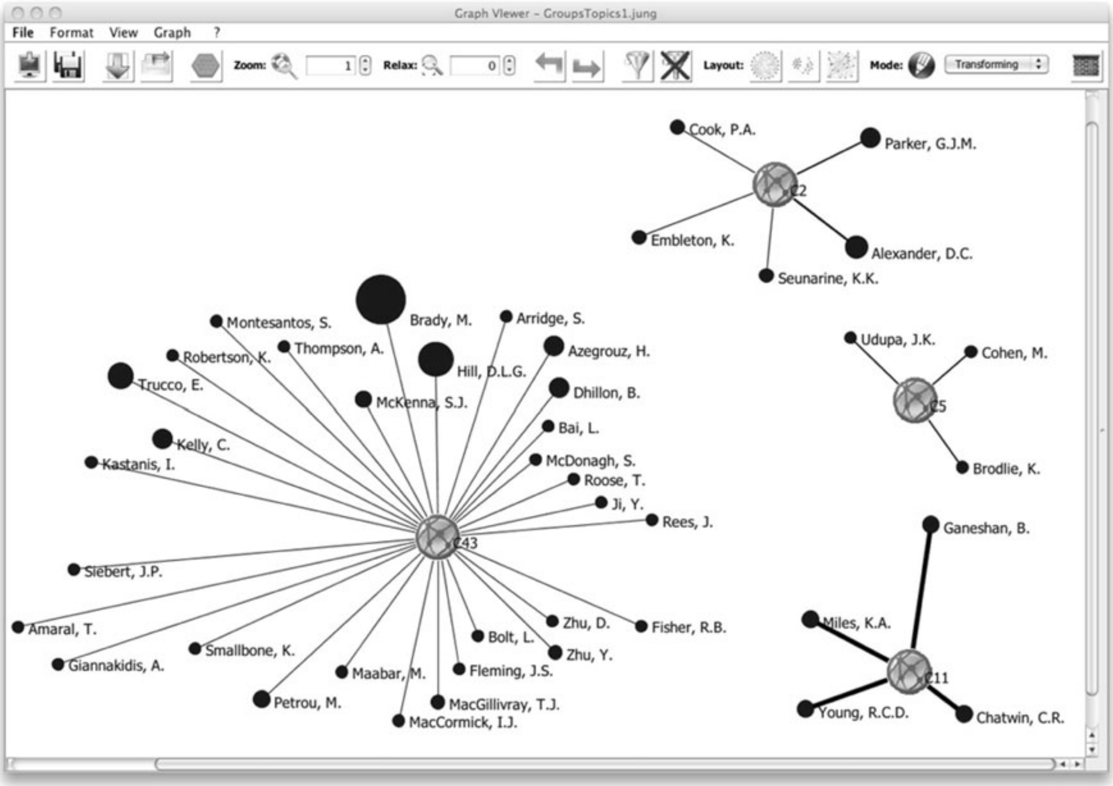


Fig. 14 The bipartite graph of the MIUA researchers involved in "imaging" research

research interests shared among the members of a research community. Topic popularity can be simply determined by counting the number of members sharing interest in each of the *k* topics define for the community, in given time intervals (e.g. during 1 year). Examples of this kind of analysis are not shown here because it is not based on SN measures (it is simple frequency plots), yet, topic popularity might be very useful for identifying, for example, "over-crowded" topics as well as poorly covered, and yet relevant, research themes. We illustrate hereafter SN measurements for analysing topic evolution over time.

Over time, a NoE of research groups working on the same domain for 3 years with the goal of reducing research fragmentation is expected to have increasing ADC values, because the shared research interests increase during the project, and decreasing VDC value, indicating a reduction of the shared interest dispersion. Any other combination of these measures' trends is a clear evidence of the non-optimal development of the NoE. Similarly, in a community like MIUA, increasing ADC and decreasing VDC values are symptoms of the ability of the researchers to share their interests and to focus themselves on the relevant topic of the research field. But these goals are more difficult to achieve, due to some relevant differences with respect to a NoE: the absence of strong governance, a long-term vision of the evolution of the research themes and the characteristic of the community to be 'open', i.e. formed by a number of researches that vary over time.

A qualitative analysis of the main aspects of a research community evolution, based on the ADC and VDC trends relation, is shown in Table 2.

For the INTEROP NoE, the two measures have been applied to four networks, obtained by grouping the 1,452 papers written by the community members into four incremental sets, each of which contains, respectively, the documents produced before the end of 2003, 2004, 2005 and up to the end of the project. Table 3 summarizes the obtained results.

As expected, the ADC value constantly increased throughout the project and the VDC decreased. Moreover, it is interesting to note that the highest increment of the ADC over a single period of time was reached at the end of the second year of the project, the one in which the preliminary results of the partners' joint activities were obtained.

For the MIUA community, we have conducted a similar analysis by grouping the 157 papers of the researchers into three incremental sets, related to the periods 2006, 2006–2007 and 2006–2007–2008 (the entire collection of documents). Table 4 shows the values of ADC and VDC related to the three corresponding networks.

While the ADC increases slightly, the VDC grows significantly, and this is strong evidence that only few stable members increase their shared interest over time, while many researchers give a one-off contribution to the community over the observed time period.

Table 2 ADC and VDC relation

		ADC	
		Increase	**Decrease**
VDC	**Increase**	*Few groups benefit from the global growth of shared interests. Social disparity increases in the community.*	*The reduction of shared interests involves few groups, and social disparity increases in the community.*
	Decrease	*The growth of shared interests involves many groups. Social disparity in the community decreases.*	*The reduction of shared interests is generalized, but social disparity decreases in the community.*

Table 3 INTEROP ADC and VDC evolution

Set	# of documents	ADC	VDC
1	595	0.832	133.194
2	859	0.914	76.775
3	1,127	0.956	41.130
4	1,452	1.000	0.041

Table 4 MIUA ADC and VDC evolution

Set	# of documents	ADC	VDC
1	50	0.008	49.393
2	97	0.033	405.441
3	157	0.070	1385.557

Table 5 INTEROP connected components evolution

Set	NCC	CC Dim
1	9	38,1,1,1,1,1,1,1,1
2	8	39,1,1,1,1,1,1,1
3	4	43,1,1,1
4	3	44,1,1

Research groups defragmentation. Another indicator to take into account for the analysis of a network over a period of time is the evolution of the number of connected components NCC (see Sect. 3.2). If the network grows appropriately, we expect that initially isolated groups (or clusters of groups) establish connections among them. This was one of the stated goals of INTEROP: to join groups coming from different research areas or from different fields (industry vs. academia). In Table 5, the NCC values referred to the four networks described in Table 3 are shown, together with the dimension of the various connected components (CC Dim.).

For example, the network associated with the first set of documents (those published before 2003) has 9 connected components: one with 38 nodes and eight made by a single node. The values in the table show the reduction of NCC over time due to an aggregation of the isolated groups into the larger component. Actually, when the INTEROP project was launched, most of the participating organizations had common research interests, but not all. As a practical example, the partners from Ancona (UPM) and Roma (RLS) were more oriented on research towards natural language processing and on information retrieval, initially an unshared theme in the INTEROP community. Throughout the project, a fruitful application of these techniques to interoperability problems has led to a better integration of the two organizations within the NoE, as well as to the emergence of NLP-related concepts among the "hot" INTEROP research themes.

The results of the same analysis on MIUA lead to completely different evaluation. As shown in Table 6, although NCC decreases from 268 to 71 over time (column *Total* in the Table), the majority of the connected components are formed by a single node (column Dim = *1*). This fact supports the conclusion drawn at the end of the analysis of shared research interests inside the community: over time, the cluster of members sharing interests increases (the dimension of the biggest CC changes from 89 to 294), but the set of researchers loosely related to this group remains relevant.

Research community resilience. A research community having members that share their research interests with many others is a "robust" community, in which the cohesion of the researchers is based on a wide network of relations. One of the main goals of a NoE is to spread knowledge about researchers' activities and interests, thus fostering potentially new collaborations and relations among the members. In some cases, the aggregation of the community begins with a relation between single members of different, well-established, small groups of researchers, focused on specific domains of interest. In this situation, the network modelling the community is connected (there are chains of interests that connect any pair of members), but the removal of a single link may split the network into two separate components. This phenomenon, strictly related to the presence of bridges (as defined in Sect. 3.2), is known as *network resilience*, and can be considered a measure of community robustness. In a NoE, this measure is expected to increase over time, as a consequence of sharing knowledge and interests.

Table 6 MIUA Connected Components evolution

Set	NCC			
	Total	Dim=1	Dim>1	
	Num	Num	Num	CC Dim
1	268	260	8	89,4,4,4,3,3,3,2
2	175	171	4	190,4,4,3
3	71	69	4	294,4,3,2

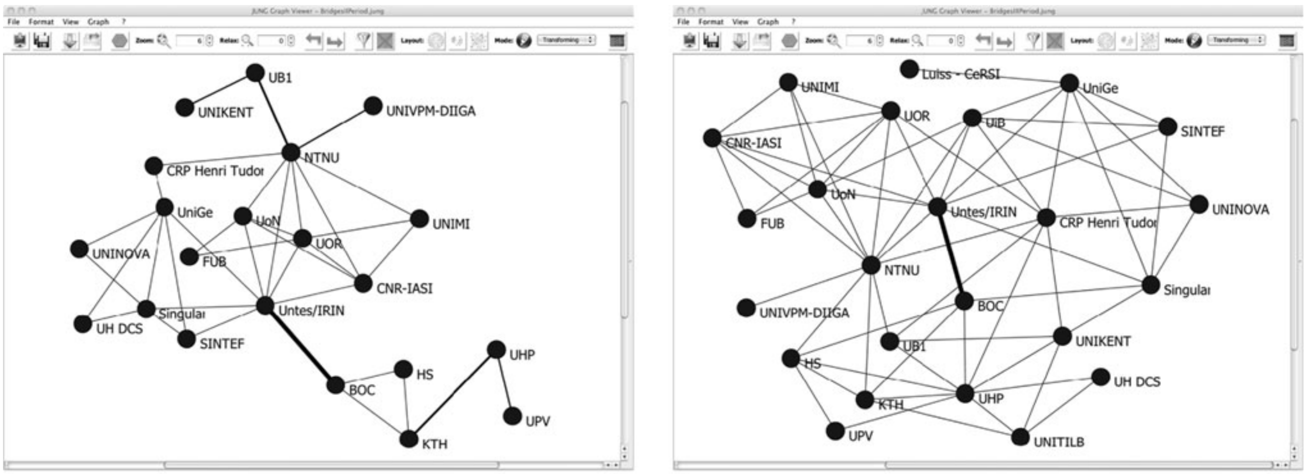
We evaluated the *resilience* of the INTEROP NoE by adding the BS(e) measure to the edges of the social networks described in the previous analysis. The networks were filtered by selecting only the edges with cos-sim(g_i,g_j) $\geq$ 0.85, in order to focus on the strongest potential collaborations. A total of six bridges were found in the network modelling the second period of the project: one with BS(e) = 5, two with BS(e) = *2* and the remaining with BS(e) = 1. Considering the first bridge, (the only one representing an interconnection between potential sub-communities formed by a significant number of research groups) we can see that it is no longer present in the following period. Figure 15 shows the GVI plots of the networks corresponding to the second (a) and third (b) year of the project. In Fig. 15a, in which the thickness of the edges is proportional to the BS(e) value, the thickest edges (between Untes/IRIN and BOC) is the bridge with BS(e) = 5, and the sub-communities it connects are clearly represented. In the following year (Fig. 15b), the edge, highlighted with a thicker line so as to be more easily located, has lost its previous role, because the shared potential interests among the researchers belonging to the original sub-communities have increased. This provides evidence that the NoE activities have strengthened the robustness of the community.

We did not draw the same conclusion in the MIUA case. The analysis of the bridges evolution in this community showed an increase in their number over time. We carried out a set of experiments with different thresholds on the edges' similarity values, all of them leading to the same results: the number of bridges increases over time. For example, the networks related to the second and third sets of data, both filtered by selecting only the edges with cos-sim(g_i,g_j) $\geq$ 0.85 show, respectively, 1 and 11 bridges with BS(e) > 1. Even in term of *resilience*, this community seems to be far from the expected evolution.

3.4 Quantitative evaluation

We have attempted to provide a measurable evaluation of the efficacy of the proposed methodology. This evaluation has only been conducted for the INTEROP project, and it is based on manually inserted information on partner's competences. Similar information is not available for the MIUA community.

During the project, the ontology was acquired by using the previously referred OntoLearn learning methodology



(a) (b)

Fig. 15 Bridges evolution over the 2004–2005 period

(Velardi et al. 2007), which was collectively evaluated by the project partners through web-based collaborative interfaces. The ontology was built based on a subset of the documents that we used in this experiment (deliverables and papers produced from the beginning of the project to mid 2006), plus other available documents in the interoperability domain (deliverables of related EU projects, conference proceedings,…) and used thereof to annotate INTEROP documents and to allow researchers to express their competencies. Partners were asked to access the INTEROP Competency Map (Velardi et al. 2008b) and select from the ontology the set of concepts describing at best their domain of expertise.[10]

We can then build an alternative content-based model of the NoE as in Sect. 3.1, but rather than creating a vector for each document, we can generate a vector for each researcher, using his/her selected concepts. The two SNs, one based on automatically extracted research themes, and the other based on the partner's stated concepts, can then be compared. Unfortunately, the comparison suffers some problems, namely

(a) the vocabulary V used in our experiment and the vocabulary C of concepts labels in the ontology do not coincide, but have about 65% overlapping, because the ontology was acquired before the publication of many papers by the network members and because it was based on other documents in the interoperability domain, not authored by network members (thus not belonging to the SN);
(b) the partners were invited to populate the Competency Map with their favourite terms, but they carried out the task with variable dedication: some selected 10–15 concepts, others just one. This problem is partly lessened by the fact that we model a research group, rather than a single researcher, thus collecting into a unique vector all the partner's selected terms.

On the other hand, as the Competency Map is the only available objective source of a partner's research themes, we decided, upon some data alignments, to perform a comparison, convinced that somehow we would be given insight into our model's performance.

We proceeded as follows: given a research group g, we first compared the set of terms T^g_{manual} expressing the expertise of g obtained by merging all interests manually inserted by a group's partners, with the set of terms $T^g_{\text{extracted}}$ obtained from a group member's publications. For a fair comparison, we eliminated those concepts not in $V \cap C$. We computed the term-set overlapping in two ways: in the first, we computed the simple concept overlap between correspondent vectors; in the second, we regarded two concepts $t_j \in T^g_{\text{manual}}$, $t_i \in T^g_{\text{extracted}}$ as overlapping iff t_j and t_i are separated by one single step of generalization/specialization in the ontology. This is the case of a t_j and a t_i directly linked through an *is–a* relation ($t_j \xrightarrow{\text{is-a}} t_i$), so being t_j a specialization of t_i. In the following, some generalizations taken from the ontology are shown:

$$\text{trust management} \xrightarrow{\text{is-a}} \text{quality management}$$

$$\text{scripting language} \xrightarrow{\text{is-a}} \text{formal language}$$

$$\text{application integration} \xrightarrow{\text{is-a}} \text{software integration}$$

$$\text{data integration} \xrightarrow{\text{is-a}} \text{information integration}$$

The average "simple" overlap between correspondent sets is 22%, while the average "extended" overlap

[10] This information is available on the INTEROP-Vlab KMap site http://interop-vlab.eu/backoffice/km.

Table 7 Overlap between manually specified and automatically extracted research group vectors, in descending order (first 12)

C	Average simple overlap	Average extended overlap
SV-50	0.830178	0.870670
SV-60	0.815879	0.868360
SV-70	0.811379	0.859122
SV-80	0.777383	0.849885
SV-90	0.776068	0.845266
SV-100	0.809903	0.829099
LV-50-50	0.769590	0.822171
SV-120	0.751866	0.816705
LV-50-100	0.765823	0.815668
SV-110	0.797152	0.814815
SV-130	0.750402	0.812065

increases up to 59.69%. The reason is that partners in general either selected very specific terms, or very general, but not both. Instead, their papers include both general (introductory) and specific terminology. The final number (nearly 60%) is very reasonable if we consider that the terminology of a paper includes both "central" (for the authors) concepts and applicative concepts, e.g. the problems a methodology intends to solve (for example, ontology representation helps to solve semantic conflicts).

For each group g we considered the k-dimensional vector p_g computed as explained in Sect. 3.1 and the correspondent vector p'_g, obtained from manually inserted terms. We then compared p_g and p'_g, when alternative clustering methodologies are used. Table 7 lists the result of the experiment, which was repeated with all "semantic" and "latent" clustering outcomes (see Sect. 2.3.2).

The table shows that, when ordering the clustering results according to the overlapping criterion, again SV results are better than LV, though the ordering does not coincide with that of Fig. 5 (but remarkably, the first two "winning" clustering results of Table 7 are the same as in Fig. 5, evidence that assigns some merit to our "wise librarian" methodology).

3.5 Summary of findings

Using different combinations of the SNA measures, as defined in Sect. 3.2, to analyse in depth the different types of phenomena related to a research community modelled as a social network, and supported by the GVI application to evaluate the community members involved in the conducted analyses at a glance, we have highlighted some relevant aspects of both the INTEROP NoE and the MIUA research community and their evolution over time.

We have been able to identify potential collaborations among the members on the basis of their competences, and to focus the attention on a subset of them, related to a specific sub-area of the research domain. Moreover, through the co-authorship relation, an analysis of the established joint research activities has been conducted, and the strongest partnerships as well as the more active members of the communities have been made clear. Then, by using evidence both of potential and real collaborations between them, the partnerships to be strengthened have been revealed. Last, it has been possible to discover how the competences of the members are distributed over the different topics of the research fields.

A deeper characterization was carried out through the analysis of the communities' evolution over time. It showed that

(a) for INTEROP the shared research interests increased and the social disparity decreased during the project, while for MIUA they both increased;
(b) there was a de-fragmentation of the INTEROP community over time while in MIUA it remained relevant;
(c) while INTEROP has become more "robust" with respect to its initial structure, MIUA has had an opposite trend.

As stated at the beginning of section, for the INTEROP case a great part of the described results has been assessed through a comparison with the information contained in the deliverables concerning the NoE monitoring activity. We also compared a network model in which research topics have been manually expressed by the INTEROP partners with the automatically extracted topics.

4 Concluding remarks and future work

In this paper, we presented a novel SNA methodology which deals with the semantic content of social relations, rather than their surface realization as in traditional SNA. The motivation behind this work lies in the fact that network analysts are typically interested in the communicative content exchanged by the community members, not merely in the number of relationships. Especially in the analysis of research networks, the use of semantics allows the discovering of topics shared by otherwise unrelated research entities, emerging themes together with their most active research entities, and so on. To the best of our knowledge, no similar methodology in the social network literature provides such a refined capability of analysis, as we showed in the previous section.

While our work builds of well-established techniques such as clustering and SNA, and on previous results by the authors on terminology and glossary extraction, the paper provides several novel contributions to the implementation and analysis of CB-SNs:

- We extend the notion of term co-occurrences to that of semantic co-occurrences, with the aid of a novel graph-based similarity measure which combines lexical co-occurrences with ontological information.
- We provide a novel criterion for evaluating the clustering results based on the intuitive notions of compactness, even dimensionality and generalization power.
- We experiment with traditional and novel SN measures, used to support the study and evolution of collaboration themes in a research network.

Our methodology has been fully implemented, including a visualization interface, which facilitates the study of the community by a social analyst. The detailed analysis of Sect. 3 demonstrates that actually the quality and richness of information that can be extracted thanks to our CB model has no comparison with the simple co-authorship model. This type of analysis has an evident practical impact, especially for research funding bodies willing to monitor the effect of their supporting actions.

Concerning the evaluation of our methodology, we remark that the literature on SNA does not provide formal, quantitative and standard criteria for performance appraisal, but usually discuss examples of the different SN measures application in specific domains. Accordingly, in this paper the efficacy of the content-based SN model has been supported by a qualitative evaluation of its utility when applied both to a small research community in MIUA, and to a well-studied case, the INTEROP NoE, for which an analysis has already been manually conducted and reported in the documents of the NoE governing committe[11]. These reports, along with available network databases, have been used to check whether the information extracted by our SNA tools matches the real characteristics of the NoE. This type of evaluation, though commonly adopted in the SN literature, is not entirely satisfactory; therefore, in the INTEROP case, we conducted another evaluation, based on available data on partner's research interests, defined by the partners themselves. This experiment, despite some inherent limitations, helped to provide a measurable assessment of the proposed methodology.

A single problem remains open in our approach: the selection of the best clustering (in the last step of topic detection, Sect. 2.2), which is a complex and still unresolved research challenge. However, our CB-SN analysis methodology is not strictly bonded with a specific clustering algorithm, and some rough tests we have conducted demonstrate that the use of different algorithms has no considerable impact on the outcomes and on the general findings. Instead, we verified that the relevance of the final result is positively influenced by the accurate selection of textual features (terminology extraction), by the use of semantics in combination with contextual evidence (both evaluation criteria adopted in this paper supported this claim) and by the modelling technique used for the social network creation along with the SNA measures chosen for the analysis.

Some interesting aspects and extensions of the proposed methodology have been left for future work. Among them are the ability to model the evolution of topics over time, by identifying how a topic evolves during the community's lifetime, and a more refined representation, with respect to the graphs of Figs. 13 and 14, of a bipartite graph with two types of nodes (topics and researchers) to show this evolution. This is a promising extension that we defer to future publications.

Acknowledgments The authors wish to thank Vincenzo Casini for his help in developing the GVI tool.

References

Baeza-Yates R, Ribeiro-Neto R (1999) Modern Information Retrieval. ACM Press Series/Addison Wesley, New York

Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Scientific American, May

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Machine Learn Res 3:993–1022

Bojars U, Breslin JG, Finn A, Decker S (2008) Using the semantic web for linking and reusing data across Web 2.0 communities. Web Seman Sci Services Agen World Wide Web 6(1):21–28

Bollegala D, Matsuo Y, Ishiuka M (2007) Measuring semantic similarity between words using web search engines. In: Proceedings of the 16th international conference on world wide web, Banff, Alberta

Budanitsky A, Hirst G (2006) Evaluating WordNet-based measures of semantic distance. Comput Linguist 32(1):13–47

Chlia M, De Wilde P (2006) Internet search: subdivision-based interactive query expansion and the soft semantic web. Appl Soft Comput 6(4):372–383

Dhiraj J, Gatica-Perez D (2006) Discovering groups of people in google news. In: Proceedings of the 1st ACM International workshop on human-centered multimedia (HCM). Santa Barbara, CA

Domeniconi C, Al-Razgan M (2009) Weighted cluster ensembles: methods and analysis. In: ACM transactions on knowledge discovery from data, vol 2, No. 4

Eckart C, Young G (1936) The approximation of one matrix by another of lower rank. Psychometrika 1:211–218

Finin T, Ding L, Zhou L, Joshi A (2005) Social networking on the semantic web. In: The learning organization, Emerald pub, New York, pp 418–435

Fuhr N (1992) Probabilistic models in information retrieval. Comp J 35(3):243–255

[11] http://interop-vlab.eu/ei_public_deliverables/interop-noe-deliverables/.

Gruber T (2003) It is what it does: the pragmatics of ontology. Invited presentation to the meeting of the CIDOC Conceptual Reference Model committee, Smithsonian Museum, Washington

Hammouda K, Kamel M (2004) Efficient phrase-based document indexing for web document clustering. IEEE Trans Knowl Data Eng (TKDE) 16:1279–1296

Hansen M, Yu B (2001) Model selection and the principle of minimum description length. J Am Stat Assoc 96:746–774

Ha-Tuc V, Srinivasan P (2008) Topic models and a revisit of text-related applications. In: Proceedings of conference on information and knowledge management, Napa Valley, CA, pp 25–32

Hirst G, Budanitsky A (2001) Lexical chains and semantic distance. In: Proceedings of EUROLAN-2001, Iasi, Romania

Hirst G, St-Onge D (1998) Lexical chains as representations of context for the detection and correction of malapropisms. In: Fellbaum C (ed) WordNet: an electronic lexical database. MIT Press, USA, pp 305–332

Jain K, Murty M, Flynn P (1999) Data clustering: a review. In: ACM computing surveys, vol 31, No. 3. pp 264–323

Jamali M, Abolhhassani H (2006) Different aspects of social network analysis. In: Proceedings of the 2006 IEEE-WIC-ACM international conference on web intelligence, Hong Kong, pp 66–72

Jiang J, Conrath D (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of international conference on research in computational linguistics, Taiwan

Jung J, Euzenat J (2007) Towards semantic social networks. In: Proceedings of the European semantic web conference (ESWC), Innsbruck, Austria, pp 267–280

Kang S (2003) Keyword-based document clustering. In: Proceedings of the 6th international workshop on information retrieval with Asian languages, vol 11. Japan, pp 132–137

Kanungo T, Mount DM, Netanyahu N, Piatko C, Silverman R, Wu AY (2002) An efficient k-means clustering algorithm: Analysis and implementation. IEEE Trans Pattern Anal Machine Intell 24:881–892

Kleinberg J (2002) An impossibility theorem for clustering. In: Advances in neural information processing systems 15: Proceedings of the 2002 conference. Bradford Books, pp 446–453

Kovacs F, Legany C, Babos A (2005) Cluster validity measurement techniques. In: Proceedings of 6th international symposium of Hungarian researchers on computational intelligence. Budapest, Hungary

Kuhn A, Ducasse S, Girba T (2007) Semantic clustering: identifying topics in source code. In: Journal of Information and software technology, vol 49, no. 3. pp 230–243

Landauer TK, McNamara DS, Dennis S, Kintsch W (eds) (2007) Handbook of latent semantic analysis, Lawrence Erlbaum Associates Inc., Mahwah

Leacock C, Chodorow M (1998) Combining local context and WordNet similarity for word sense identification. In: Fellbaum C (ed) WordNet: an electronic lexical database. MIT Press, USA, pp 265–283

Lin D (1998) An information-theoretic definition of similarity. In: Proceedings of the 15th international conference on machine learning. Madison, USA

Macherey W, Viechtbauer J, Ney H (2002) Probabilistic retrieval based on document representations. In: Proceedings of the international conference on spoken language processing, Denver, CO, pp 1481–1484

McCallum A, Corrada-Emmanuel A, Wang X (2005) Topic and role discovery in social networks. In: Proceedings of international joint conference on artificial intelligence (IJCAI), Edinburgh, pp 786–791

Mei Q, Cai D, Zhang D, Zhai C (2008) Topic modeling with network regularization. In: Proceedings of WWW 2008, April 21–25, 2008 Beijing, China

Mika P (2007) Social networks and the semantic web, series in semantic web and beyond, vol 5. Springer, Berlin

Nallapati R, Ahmed A, Xing E, Cohen WW (2008) Joint latent topic models for texts and citations. In: Proceedings of KDD 2008, August 24–27, 2008, las Vegas, Nevada, USA

Navigli R, Crisafulli G (2010) Inducing word senses to improve web search result clustering. In: Proceedings of the 2010 conference on empirical methods in natural language processing (EMNLP 2010), MIT Stata Center, Massachusets, pp 116–126

Navigli R, Velardi P (2008) From glossaries to ontologies: extracting semantic structure from textual definitions. Ontology learning and population: bridging the gap between text and knowledge. In: Buitelaar P, Cimiano P (eds) Series information for frontiers in artificial intelligence and applications, IOS Press, Amsterdam, pp 71–87

Nenadic G, Rice S, Spasic I, Ananiadou S, Sy B (2003) Selecting text features for gene name classification: from documents to terms. In: Proceedings of the ACL workshop on NLP in biomedicine, vol 13. Sapporo, Japan, pp 121–128

Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167–256

Pedersen T, Pakhomov SV, Patwardhan S, Chute CG (2007) Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 40(3):288–299

Ponzetto SP, Strube M (2007) Knowledge derived from Wikipedia for computing semantic relatedness. J Artificial Intell Res 30(1):181–212

Purandare A, Pedersen T (2004) Word sense discrimination by clustering contexts in vector and similarity spaces. In: Proceedings of the conference on computational natural language learning (CoNLL), May 6–7, 2004, Boston, MA, pp 41–48

Resnik P (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. J Artificial Intell Res 11:95–130

Russo V (2007) State of the art of clustering techniques: support vector methods and minimum Bregman information principle, Master Thesis, University of Napoli “Federico II”, Computer Science Dept

Salton G, Mcgill M (1983) An Introduction to modern information retrieval. McGraw-Hill, New York

Sclano F, Velardi P (2007) TermExtractor: a web application to learn the common terminology of Interest Groups and Research Communities. In: Proceedings of 9th conference on terminology and artificial intelligence (TIA 2007), Sophia Antinopolis

Scott J (2000) Social network analysis. SAGE Publications, Chennai

Staab S, Studer R (2009) Handbook on ontologies. Springer, Berlin

Sussna M (1993) Word sense disambiguation for free-text indexing using a massive semantic network. In: Proceedings of the second international conference on information and knowledge management, Washington, DC, USA, pp 67–74

Tagarelli AY, Karypis G (2008) A segment-based approach to clustering multi-topic documents. In: Proceedings of SIAM data mining conference text mining workshop, Atlanta, Georgia, USA

Tan P, Steinbach M, Kumar V (2006) Cluster analysis: basic concepts and algorithms. In: Introduction to data mining. Addison-Wensley, New York

Terra E, Clarke CL (2003) Frequency estimates for statistical word similarity measures. In: Proceedings of the 2003 Conference of the North American chapter of the ACL on HLT (NAACL ‘03), Morristown, NJ, pp 165–172

Velardi P, Cucchiarelli A, Petit M (2007) A taxonomy learning method and its application to characterize a scientific web community. IEEE Trans Data Knowl Eng (TDKE) 19(2):180–191

Velardi P, Navigli R, D’Amadio P (2008a) Mining the web to create specialized glossaries. IEEE Intell Syst 23:5

Velardi P, Cucchiarelli A, D'Antonio F (2008b) Monitoring the status of a reserach community through a knowledge map, web intelligence, agent systems. Int J 6(3):1–22

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, UK

Weeds J, Weir D (2006) Co-occurrence retrieval: a flexible framework for lexical distributional similarity. Comput Linguist 31(4):439–475

Wood M (2005) Bootstrapped confidence intervals as an approach to statistical inference. Organ Res Methods 8(4):454–470

Wu Z, Palmer M (1994) Verb semantics and lexical selection. In: Proceedings of 32nd annual meeting of the association for computational linguistics (ACL), Las Cruces, New Mexico, USA, pp 133–138

Zhao Y, Karypis G (2004) Empirical and theoretical comparisons of selected criterion functions for document clustering. Machine Learn 55(3):311–331

Zhao Y, Karypis G (2005) Hierarchical clustering algorithms for document datasets. Data Min Knowl Disc 10:141–168

Zhong M, Chen Z, Lin Y, Yao J (2004) Using classification and key phrases extraction for information retrieval. In: Proceedings of 5th World Congress on intelligent control and automation, June 15–19, 2004, Hangzhou, China

Zhou D, Ji X, Zha H, Giles CL (2006) Topic evolution and social interactions: how authors effect research. In: Proceedings of CIKM 2006, November 5–11, 2006, Arlington, Virginia, USA