

Available online at www.sciencedirect.com**SciVerse ScienceDirect**

Procedia - Social and Behavioral Sciences 22 (2011) 37 – 48

Procedia
 Social and Behavioral Sciences

Life-Cycles and Mutual Effects of Scientific Communities

Václav Belák, Marcel Karnstedt, Conor Hayes

Digital Enterprise Research Institute (DERI), National University of Ireland, Galway
 first.last@deri.org

Abstract

Cross-community effects on the behaviour of individuals and communities themselves can be observed in a wide range of applications. While previous work has tried to explain and analyse such phenomena, there is still a great potential for increasing the quality and accuracy of this analysis. In this work, we propose a general framework consisting of several different techniques to analyse and explain cross-community effects and the underlying dynamics. The proposed methodology works with arbitrary community algorithms, incorporates meta-data to improve the overall quality and expressiveness of the analysis and identifies particular phenomena in an automated manner. We illustrate the benefits and strengths of our approach by exposing in-depth details of cross-community effects between two closely related and well established areas of scientific research. This work focuses on techniques for understanding, defining and eventually predicting typical life-cycles and events in the context of cross-community dynamics. © 2011 Published by Elsevier Ltd. Open access under [CC BY-NC-ND license](https://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: community life-cycle; cross-community dynamics; co-citation analysis; SNA; graph mining

1. Introduction

Claims for scientific progress are often assessed using relatively static citation measures. However, the analysis of the life-cycle of a community provides much greater explanatory power for the progress and potential of a scientific field—for the community itself and external evaluators such as tenure committees, funding agencies, venture capitalists and industry. While previous work has examined scientific networks through co-citation and textual analysis, there is relatively little work on analysing dynamics and cross-community effects, particularly where closely related communities are competing for scientific, funding and industrial capital. In previous work (Karnstedt & Hayes, 2009), we proposed a general road-map for the cross-community analysis of scientific communities. In this work, we elaborate this idea further and present the results and insights we gained from an empirical study of two scientific disciplines.

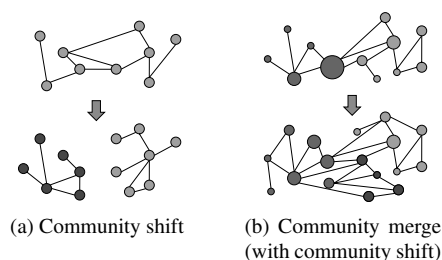


Figure 1: Possible phenomena: community shift and community merge

Thomas Kuhn (1996) introduced the idea of paradigm shift into the lexicon of scientific discourse as a means of explaining how new theories overturn existing theories within a scientific field. Kuhn's analysis tends to focus on well-known dramatic shifts in science such as the juncture between Newtonian and relativistic physics. In this paper, we explore whether similar behaviours occur in what might be termed 'normal science'. A paradigm shift as Kuhn describes it is something very significant, close to a "revolution in science". We investigate closely related phenomena, but of less dramatic importance and with smaller influence on the research field. Thus, in this work we use the term *community shift*. A particularly dominant and huge community shift is what we eventually understand as a paradigm shift. Figure 1a illustrates what we call a community shift. The upper part shows co-citation relations between different authors at a specific point in time. Over time, a sub-community detaches from its original community (lower part of the figure). This means that authors from either communities tend not to be cited together any more and over time the sub-community splits off from the original community. Figure 1b illustrates the opposite of this phenomenon, which we call a *community merge*. Over time, the communities approach each other, represented by more and more edges between members. This can lead to closely related communities or even to a merger into one larger community. For some communities, we may observe a combination of community shift and merge, indicated by the different colours of the nodes. Another interesting phenomenon described by Kuhn is a *paradigm articulation*, which refers to the maturation process of a community resulting in different groups specialising on sub-topics. Naturally, such an effect cannot be analysed solely on the basis of the network structure. In this case, the benefits of enriching the topological analysis by topical analysis become obvious. Consequently, we call such a phenomenon *community specialisation*. Similarly, in a *community topic change* a community may remain structurally stable over time while changing its topic.

In the course of the existence of a community, several distinctive events may be characterised. Apart from the obvious ones like its birth, growth, contraction, or death, others more specialised events may be defined. The focus of this work is on the events which are characteristic for scientific communities in context of their mutual effects, e.g. a community shift, merge, or topic change. The observation and understanding of such events may lead eventually to the notion of community life-cycles, which describe classes of observed dynamics of scientific communities.

As an illustrative example, we analyse the interactions over time of two closely related areas of scientific research using author-based co-citation network analysis (see, e.g., Gmür, 2003), supplemented by automatic extraction and investigation of the topics and expertise that form the core of each community. The research fields we chose to examine are Information Retrieval (IR) and Semantic Web (SW). IR provides techniques that are a standard part of Web and document search engines with well-defined methodological and evaluation techniques. SW research, on the other hand, focuses on improving Web search by developing standards and techniques for structuring the data on the Web. A witty commentary on both suggests that IR is driven by a problem whereas SW is driven by a solution (Baeza-Yates et al., 2008). Both are well-established communities focused on accurate, scalable search methods on the Web. Yet in terms of methodology and research culture, both communities are in many ways orthogonal, with large cores of each discipline indifferent to each other's work. As such, the two general research fields promise to reveal a wide range of interesting cross-community effects, influences and interactions. However, our main goal is to develop methods and techniques to analyse cross-community effects between arbitrary research fields. In summary, our contributions are:

- We propose techniques to enable scalable analysis of cross-community dynamics. The methodology is not limited to one particular community-detection method and is suited for different relations between individuals.
- We combine topological analysis with topical analysis by incorporating automatically extracted meta-data, both for enriching the actual analysis as well as enabling new methods of assessing the clustering quality.
- We discuss and evaluate different methods and measurements to automatically determine community overlap, community relations and specific interesting phenomena.
- We further analyse and discuss *life-cycle measurements* and their suitability for identification and explanation.

We present a general, flexible and extendible framework for analysing cross-community dynamics and highlight the suitability of different methods and techniques in this context. Although several recent works deal with the dynamics of communities, we are not aware of any work that discusses general methods for identification and analysis of cross-community effects in the context of life-cycles of communities. Our overall goal is to understand, define and eventually predict typical life-cycles and states of communities in the context of cross-community effects.

In Section 2, we give a brief summary of the most important related work. Our set of techniques and methods is described in Section 3, we describe our data in Section 4. We present results gained with these methods in Section 5. Section 6 concludes and indicates future work. Note that we use the terms communities and clusters interchangeably.

2. Related Work

Recently, the idea of analysing the dynamics of communities, in contrast to static community analysis, has gained attention. Most dynamic community analysis uses snapshots of the underlying network graph from different points in time. Communities found in these snapshots are compared over time and the development of communities is deduced and investigated. Greene et al. (2010b) are very close to our idea of understanding the life-cycles of communities, but do not investigate a similarly rich set of indicators and automated methods. While Palla et al. (2007) also investigate the time dependence of communities and community evolution, they focus on overlapping communities. Moreover, the community dynamics model presented in that work is based solely on structural features and cannot address the problem of analysis of evolution of scientific discourses because it lacks the content dimension. Therefore, we develop the previous research towards generating a general understanding of typical community life-cycles, the underlying mechanisms and reasons for phenomena aligned with Kuhn's observations. Tantipathananandh et al. (2007) also deal with the dynamics of communities, but propose an interesting alternative approach based on a graph-colouring problem. In contrast to our work, they focus on the dynamics of single individuals and how they switch between communities over time. However, an extension to the micro level of single community members is planned for our future work.

A crucial problem with the snapshot-based approach is the choice of the underlying time periods, which can have significant influence on the observed structures and behaviours. See Delvenne et al. (2010) for a discussion of this problem. A related problem is the question on how to track each community from one snapshot to the next, which is required to determine the community's evolution. This becomes even more complicated with the use of static community detection approaches, where communities for each snapshot are determined independently from other snapshots. A promising alternative is to use evolutionary clustering (Chi et al., 2007). In this approach, the community detection at a certain point in time is influenced by the community structure in former times. We designed our framework to be adaptive and extensible in both directions. While in this work we focus on existing static community-detection methods, we plan to investigate evolutionary clustering in future work.

3. Methodology

In this section, we motivate and explain the techniques used in our work, which are based on following requirements: (1) we expect the dynamics of the data set to be represented by snapshots of several consecutive time-steps; (2) an algorithm to detect communities in the network in each time step is available; (3) nodes (i.e., authors) have to be uniquely identified *among all* time-steps; and (4) for topical analysis, meta-data (i.e., topics) has been assigned to nodes in the network.

3.1. Community Detection and Tracking

We identified communities in each period using three popular community-detection algorithms, denoted as *Infomap* (Rosvall & Bergstrom, 2008), *Louvain* (Blondel et al., 2008) and *WT* (Wakita & Tsurumi, 2007). Whereas WT and Louvain are both based on modularity optimisation (Newman, 2010, p. 373), i.e., the topological feature of clustering, the Infomap reveals the community structure according to the information flow in the network modelled as a random walk. The topological approach inspects the co-citation structure from a rather static point of view, whereas Infomap reveals community structure by compressing a description of information flows on the network. The underlying co-citation network can be interpreted in both ways.

These algorithms were chosen because they are able to operate over weighted networks, they scale up to the size of the analysed network and for each an implementation is publicly available. Moreover, they produce non-overlapping communities. Therefore it is possible to easily visualise them. However, the requirements listed before can be fulfilled by a wide range of community-detection algorithms. Supporting overlapping communities is possible as well, but would require some modifications of the measures that we will now present.

We track a community over time by means of the highest overlap measured by the Jaccard coefficient (Palla et al., 2007; Greene et al., 2010b). The i -th community mined in time t , i.e., c_i^t , is matched according to the highest Jaccard coefficient value among all communities C^{t+1} mined in time $t + 1$:

$$\text{match}(c_i^t) = \arg \max_{c_j^{t+1} \in C^{t+1}} \frac{|c_i^t \cap c_j^{t+1}|}{|c_i^t \cup c_j^{t+1}|} \quad (1)$$

If two communities c_i^t and c_k^t have the maximal overlap with the same subsequent community c_j^{t+1} , the matching is again determined by the Jaccard coefficient value. The community that has the higher overlap with c_j^{t+1} is matched, the other community is then matched to the subsequent community with the second-highest overlap.

We identify other types of interesting relations between communities such as important ancestors or descendants. These relations can be defined as a modification of the Jaccard coefficient, where the overlap is relative to either the latter or the former community:

$$\text{ancestor}(c_i^t, c_j^{t+1}) = \frac{|c_i^t \cap c_j^{t+1}|}{|c_j^{t+1}|} \quad (2)$$

$$\text{descendant}(c_i^t, c_j^{t+1}) = \frac{|c_i^t \cap c_j^{t+1}|}{|c_i^t|} \quad (3)$$

3.2. Topic detection

In order to identify community topics and to “name” communities, we mined keywords using NLP techniques (Bordea, 2010) from the abstracts or full-texts for almost 70% of the underlying articles. In addition, author-provided keywords for 10% of the articles were extracted. All keywords were tokenised, stemmed (Porter, 1980) and ranked by a TF-IDF measure, but with keywords assigned to authors instead of documents. To determine the keywords for a given time period, we assigned the keywords from each document published in that time to all its co-authors. As a result, each author a was described by a bag-of-words vector k_a^t for each time period t . In content analysis, the overall cluster topic is usually characterised by its centroid. The topical centroids are thus derived from the keywords of all cluster members, according to the standard formula for centroid computation:

$$\text{centroid}(c) = \frac{\sum_{a \in c} k_a}{|c|} \quad (4)$$

Centroids were used for both computation and interpretation purposes. The measures discussed in the next section are based on (dis)similarity obtained as a standard cosine distance between two centroids. The total number of keywords, i.e., the dimension of k_a^t vectors, usually exceeded several thousands. Therefore, the interpretation of a cluster topic was derived from only its 20 highest-ranked keywords. However, this led sometimes to very rare keywords to be ranked highest, while the general yet informative keywords were discarded by the TF-IDF ranking. Hence, in addition to the 20 highest-ranked keywords, we also considered the 20 most frequent keywords (TF) of the cluster. We will refer to the union of these two sets of keywords as *characterising keywords*.

3.3. Community Life-Cycle Measures

The purpose of community life-cycle measures is to measure and explain the state and the evolution of the community along several in-depth dimensions. From the structural perspective, communities are described by *size* \mathcal{S} , normalized *group betweenness* \mathcal{B} (Everett & Borgatti, 1999), *author entropy* \mathcal{A} , and *relative density* ρ . All these measures have been successfully used before in the literature and promised to be particularly informative. Author entropy \mathcal{A} has been defined and explained by Hayes et al. (2007):

$$\mathcal{A}(c^{t+1}) = -\frac{1}{\log |C_o^{t+1}|} \sum_{c_o \in C_o^{t+1}} \frac{|c^t \cap c_o|}{|c^t \cap A^{t+1}|} \log \frac{|c^t \cap c_o|}{|c^t \cap A^{t+1}|}, \quad (5)$$

where A^{t+1} is the set of all authors in time $t + 1$ and C_o^{t+1} is the set of communities in time $t + 1$ containing authors of c^t . \mathcal{A} measures how much the authors of c^t are dispersed among other communities in a subsequent time-step. If the

authors are equally dispersed among subsequent communities, \mathcal{A} will approach 1, whereas if all the authors remain in the same community, \mathcal{A} will approach 0. Relative density ρ is defined as the ratio between intra-community degree and its total degree:

$$\rho(c) = \frac{\sum_{e \in E_{c_i}} w(e)}{\sum_{e \in E_{c_a}} w(e)}, \tag{6}$$

where E_{c_i} is the set of all internal edges of cluster c , E_{c_a} is the set of all edges incident to community c and w is a function assigning a weight to each edge. ρ is a *local* measure of community quality. As a community is usually understood as a subgraph with more intra- than inter-cluster edges, we chose this measure to investigate the level to which the community is structurally shaped. In case of self-referential communities, i.e., those ones without any edge to any other community, ρ will be 1, whereas a very open and ill-shaped community will have values near to 0.

For topic analysis, we use *topic stability* \mathcal{T} and *cluster content ratio* \mathcal{H} . Hayes et al. (2007) introduced \mathcal{T} as *topic drift* and \mathcal{H} as *cluster quality*. \mathcal{T} is the cosine distance between centroids of the same cluster in two subsequent time periods. \mathcal{H} is the ratio of intra- to inter-cluster similarity:

$$\mathcal{H}(c) = \frac{I}{\mathcal{E}} = \frac{\frac{1}{|c|} \sum_{a \in c} \cos(a, \text{centroid}(c))}{\cos(\text{centroid}(c), \text{centroid}(A))}, \tag{7}$$

where I is the average similarity between the cluster’s authors and its centroid and \mathcal{E} is the similarity between the cluster’s centroid and the centroid of the entire network (A is the set of all authors).

Interestingly, the cluster content ratio \mathcal{H} can be used to assess the quality of a clustering. Drawing on the observations of assortative mixing in citation networks (Newman, 2010, p. 220), i.e. papers in such networks tend to cite papers in the same field, we expect the clusters to be topically coherent, which can be assessed by \mathcal{H} . The average value of \mathcal{H} can thus be used to indicate whether the community-detection algorithm can uncover communities associated to coherent and distinct topics. Similarly, modularity Q was employed to measure the quality of communities from the structural point of view. Figure 2 compares the values per year of the \mathcal{H} (topic cohesion) and Q (community cohesion) for each algorithm. Surprisingly, the two algorithms based on modularity maximisation uncovered communities with lower modularity than Infomap. On the other hand, Louvain shows the highest average cluster content ratio of $\langle \langle \mathcal{H}_o \rangle \rangle \doteq 2$ (but with a variance of $\sigma_{\mathcal{H}_o}^2 = 4.33$). Both Louvain and Infomap communities generally had a high cluster content ratio with $\langle \langle \mathcal{H} \rangle \rangle > 1$ for every year, which suggests that these communities were associated with coherent topics. The high overall average cluster content ratio of Louvain communities shows that the members of these communities were twice as similar to each other in terms of content as to the rest of the network. Since the performance of WT is dominated by the other two algorithms, we do not discuss communities detected by it in the results section.

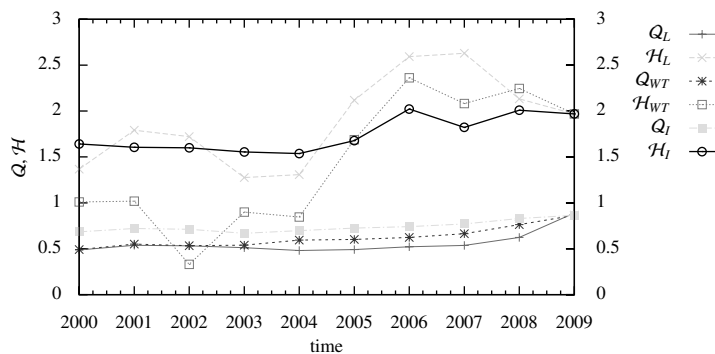


Figure 2: Modularity and average cluster content ratio per year. Subscripts L , WT , and I denote Louvain, WT, and Infomap algorithms, respectively.

3.4. Community Topic Evolution Measures

Events in a community life-cycle that are characterised by changes in both its structure and *topic* are particularly interesting as their analysis promises to shed some light on the emergence of new research fields or the mutual impact of communities. Therefore it is necessary to take into account the content dimension. In order to reveal these events and measure the significance of the underlying dynamics, we combined the *ancestor* (Eq. 2), *descendant* (Eq. 3) and author entropy \mathcal{A} (Eq. 5) measures with a measure of topic change *dissim* (Eq. 8, 9, and 10), where *dissim* is a complement to cosine distance. As a range of each measure remains within $[0, 1]$, it is possible to threshold less significant cases.

In case of a *community shift*, we are interested in a newly emerged community that is topically significantly different from its ancestor. On the contrary, a *community shift/merge* can be defined as a community merging with a topically different one. The merge can be expressed as a *descendant* relationship, especially if $\text{descendant}(c_i^t, c_j^{t+1}) \rightarrow 1$. Thus, the formulae used for community shift \mathcal{P}_S and community shift/merge $\mathcal{P}_{S/M}$ for $i \neq j$ are:

$$\mathcal{P}_S(c_i^t, c_j^{t+1}) = \text{dissim}(c_i^t, c_j^{t+1}) \times \text{ancestor}(c_i^t, c_j^{t+1}) \quad (8)$$

$$\mathcal{P}_{S/M}(c_i^t, c_j^{t+1}) = \text{dissim}(c_i^t, c_j^{t+1}) \times \text{descendant}(c_i^t, c_j^{t+1}) \quad (9)$$

\mathcal{P}_S and $\mathcal{P}_{S/M}$ express the relation between two communities regarding structural similarity and topical difference. The more c_i^t (c_j^{t+1}) is formed by members of c_j^{t+1} (c_i^t) and the more both communities differ in topic, the higher are these values. We expect a significant difference between the sizes of communities c_i^t and c_j^{t+1} , i.e., in a community shift c_j^{t+1} is usually smaller than its ancestor c_i^t , and in a community shift/merge c_i^t is usually smaller than its descendant c_j^{t+1} . Thus, we use a threshold of 0.5, which means that for the maximum value of $\text{dissim}(c_i^t, c_j^{t+1}) = 1$ the phenomena are only detected with at least 50% membership overlap.

The *community topic change* measure \mathcal{P}_C is used to detect cases where a community changes its topic while preserving its structure:

$$\mathcal{P}_C(c_i^t) = \text{dissim}(c_i^t, c_i^{t+1}) \times (1 - \mathcal{A}(c_i^{t+1})), \quad (10)$$

where \mathcal{A} is the author entropy as defined in Eq. 5. \mathcal{P}_C measures the change of the topic of the cluster c_i between subsequent time-steps. It discriminates cases where the cluster has rather weak structure, since then the entropy will be high. As we observed an entropy greater than zero in most cases, we chose a threshold of 0.3. The average entropy and its variance for Louvain clusterings were $\langle\langle \mathcal{A}_L \rangle\rangle \doteq 0.4$, $\sigma_{\mathcal{A}_L}^2 \doteq 0.1$. For Infomap clusterings these values were $\langle\langle \mathcal{A}_I \rangle\rangle \doteq 0.44$, $\sigma_{\mathcal{A}_I}^2 \doteq 0.14$. To further filter out cases of very small communities, with all the above measures we consider only communities with a minimal overlap of 5 authors.

An event detected by any of the community topic evolution measures is classed as having *inter-discipline dynamics* if both SW- and IR-related communities are involved. We mined the most frequent keywords from publications in both research fields in order to assign each analysed community to either SW or IR.

4. Data Set

In order to build networks of actors from the SW and IR fields, we first picked a set of major conferences from each field, see Table 2a. If we want to identify young and evolving communities, conferences are better suited than journals, particularly in computer science. In addition, we considered all co-located workshops and all work from Tim Berners-Lee, as the generally understood founder of SW. We selected all publications available for these venues (from year 2000 onwards) from DBLP¹ as seeds. Then, we used crawlers to fetch citation information from appropriate Web sources. But, to gather clean data, we used only citing works that we found again in the DBLP data. The total number of included articles is 39314. For the topical analysis, we were able to scrape 22975 abstracts and 3740 full texts. From these we mined 263742 keywords, extended by 18313 author-provided keywords.

Using the crawled citation information, we built a co-citation network of these authors. A sequence of time-slices of the network was obtained by a sliding time-window. In co-citation analysis, time-slices are often overlapping (Up-ham & Small, 2010; Small, 2006) in order to smooth volatility of the dynamic network. Thus the window size

¹<http://www.informatik.uni-trier.de/~ley/db/index.html>

Table 1: Details of the used data set

(a) Used conferences			(b) Statistics		
field	abbreviation	conference name	time slot	authors	edges
SW	ISWC	International Semantic Web Conference	2000–2002	1459	66039
SW	ESWC	Extended (former “European”) Semantic Web Conference	2001–2003	1906	87520
SW	ASWC	Asian Semantic Web Conference	2002–2004	2211	107499
IR	SIGIR	Special Interest Group on Information Retrieval conference	2003–2005	2468	120471
IR	ECIR	European Conference on Information Retrieval	2004–2006	2776	141093
IR	CIKM	International Conference on Information and Knowledge Management	2005–2007	3062	134132
IR	TREC	Text Retrieval Conference	2006–2008	3002	102928
			2007–2009	2190	44461
			2008–2009	1113	13340
			2009	83	159
			average	2027	81764.2

controls the resolution of the analysis, while overlap length determines the stability of the communities across time. We experimented with window sizes 1, 2, and 3 years with 0, 1, and 2 years overlap, respectively. We observed that non-overlapping slices do not result in stable community structure. Further, while the smaller overlapping windows yielded stable structure of communities, many cross-community events were observable only in the three-year windows overlapping by two years. Thus, we decided to use these for further analysis, as they promise to represent a good trade-off between clustering stability and resolution. The question of the best window sizes and overlap is a non-trivial one and is the subject of recent research (Delvenne et al., 2010; Sulo et al., 2010). A more rigorous investigation for our data is out of the scope of this work, but on our future agenda. Table 2b shows the number of nodes (authors) and edges in the resulting graphs. The weight of an edge refers to the number of co-citations in each time period. A co-citation between an author A_1 of document D_1 and another author A_2 of document D_2 occurs if we find a third document D_3 citing D_1 and D_2 , where D_1 and D_2 are published in the inspected time period. We used the networks weighted in that way as input for the Louvain algorithm, because it does not support floating-point weights. As a normalised weighting scheme usually produces better results, we used CoCit scores (Gmür, 2003) as input for Infomap.

5. Results

The aim of the results presented in this section is to assess the suitability of the introduced methodology to detect and explain cross-community phenomena like community shift or topic change. We present a selection of the most interesting cases identified by community topic evolution measures and use the life-cycle measures to explain them. For the sake of brevity, we present only a summary of the key topics of the community instead of a tagcloud. Likewise we refer to the time-slots only by the beginning of the interval, e.g. just 2000, instead of 2000–2002. Note that the life-cycle measures \mathcal{A} and \mathcal{T} are always computed with respect to the previous time step, e.g., \mathcal{A} in time t measures the level of dispersion of authors forming the community in time $t - 1$. For the sake of brevity, only the most informative values of the measures are presented and we point the reader to the supplementary information containing a comprehensive list of the values and a comparative analysis discussing their relevance to each class of the cross-community phenomena.²

5.1. Community Shift

The emergence of Louvain community 26 in 2006 has been identified as an inter-discipline community shift with $\mathcal{P}_S \doteq 0.62$. As Figure 3 illustrates, it was formed by community 6 “web information retrieval” (by 80%) and by community 5 “semantic web” (20%). Right after its emergence, the characterising keywords suggest the focus on interdisciplinary topics like “navigation”, “personalization” and “semantic web”. Under a massive influence of community 15 “semantic web and IR” in 2007, community 26 changed its topic ($\mathcal{P}_C \doteq 0.71$) towards “semantic

²<http://belak.net/doc/2010/procedia-si.pdf>

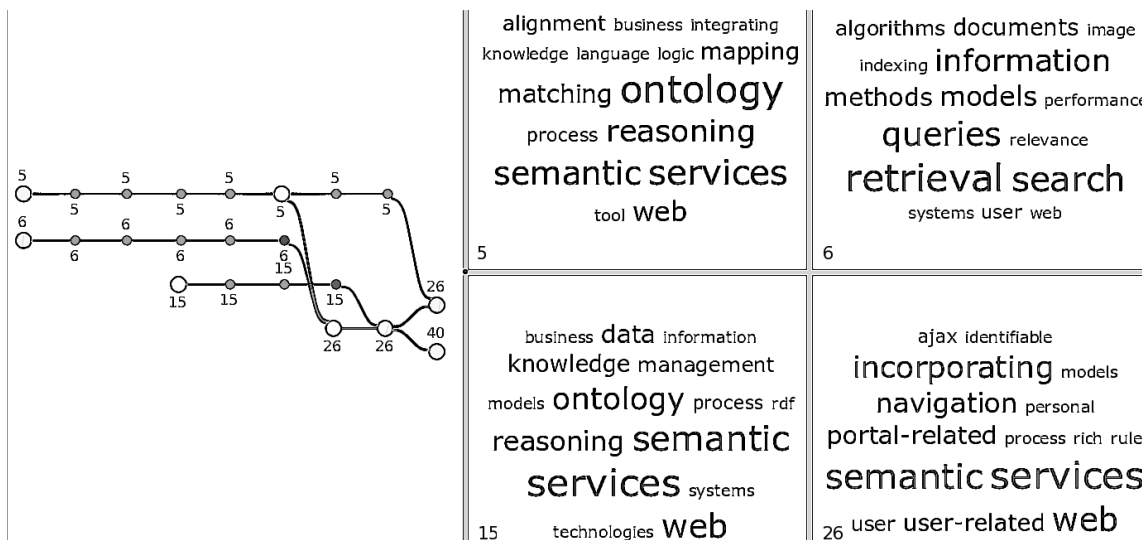


Figure 3: Main ancestors of community 26: 5 “semantic web”, 6 “web IR”, 15 “semantic web and IR”

web and business processes”, as illustrated in Figure 4 (figures based on output from TextLuas software package for visualisation of dynamic clusters of terms and objects (Greene et al., 2010a)). This influence is noticeable in Figure 5b, where the community is positioned nearer to community 15, in contrast to its original position in the bottom-right corner of Figure 5a. The change of topic towards more SW-related themes in 2007 is expressed by low topic stability $\mathcal{T} \doteq 0.29$ in 2007, while we investigated a rise to $\mathcal{T} \doteq 0.65$ in 2008. This shows the topical stabilisation.

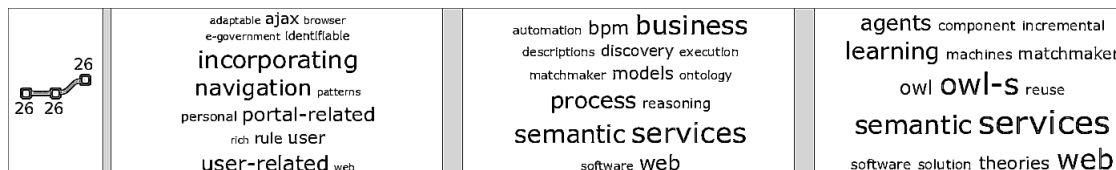


Figure 4: Evolution of topics of community 26. Note the sudden change of topics from navigation-related topics towards semantic web, which corresponds with the influence of community 15.

This pattern repeated in other shifts as well. First, emerging communities show a low topic stability, which means they significantly changed with respect to the previous time-step. As the communities are growing, their topics stabilise as well. Thus, community size and topic stability are useful for identifying shifts. Communities that first seem to be topically weak grow and grasp their own topical identity later on.

Shifts may also be interpreted as a *community specialisation*, when new communities with more specialised topics emerge, while the original community becomes eventually smaller. This is the case of Infomap community 9, which started with core concepts around SW. As several specialised communities split off in subsequent time steps, it concretised its topic towards “semantic web services”. While the topic stability \mathcal{T} had been very high since the beginning (see Figure 6), the size \mathcal{S} of the community plummeted since 2003, while the cluster content ratio \mathcal{H} started to rise at the same time. Between 2001 and 2004, we identified two community shifts towards more specific SW-related topics. \mathcal{S} , \mathcal{T} and \mathcal{H} provide valuable insights into these shifts, as they support the hypothesis of specialisation: the topically stable community (high \mathcal{T}) started to contract (diminishing \mathcal{S}), while several distinct communities shifted, which was accompanied by rising content cluster ratio \mathcal{H} . Other measures like relative density ρ , betweenness \mathcal{B} or author entropy \mathcal{A} did not seem to provide any further insights in this case.

One of the communities shifted from community 9 was community 99 “semantic desktop and personalization”,



Figure 5: Snapshots of the network depicting communities 6 “information retrieval” (pink), 5 “semantic web” (red), 15 “semantic web and information retrieval” (violet) and their descendant community 26 (green). Nodes represent authors, whose logarithmically scaled site betweenness is denoted by the node size. The position of the nodes is determined by a force-based layout (Fruchterman & Reingold, 1991). Note that occasional node overlaps are an artifact of rendering and each author occurs in each snapshot exactly once.

which emerged ($\mathcal{P}_S \doteq 0.54$) in 2003. The low topic stability \mathcal{T} and relative density ρ and high author entropy \mathcal{A} at the beginning suggest that the cluster was not very well defined during the first two time steps (see Figure 6). But, in 2005 this changed significantly. Since then, community 99 showed high topic stability and relative density. We assume that this is not a coincidence, because in 2006 the main EU project on social-semantic desktop NEPOMUK³ started. This is a similar pattern to the one observed in Louvain community 26 discussed above.

The community shift measure \mathcal{P}_S proved to be a useful measure to reveal phenomena like shifts and specialisation. For their explanation and interpretation, particularly topic stability \mathcal{T} , cluster content ratio \mathcal{H} , size \mathcal{S} , author entropy \mathcal{A} , and relative density ρ were helpful, in contrast to group betweenness \mathcal{B} .

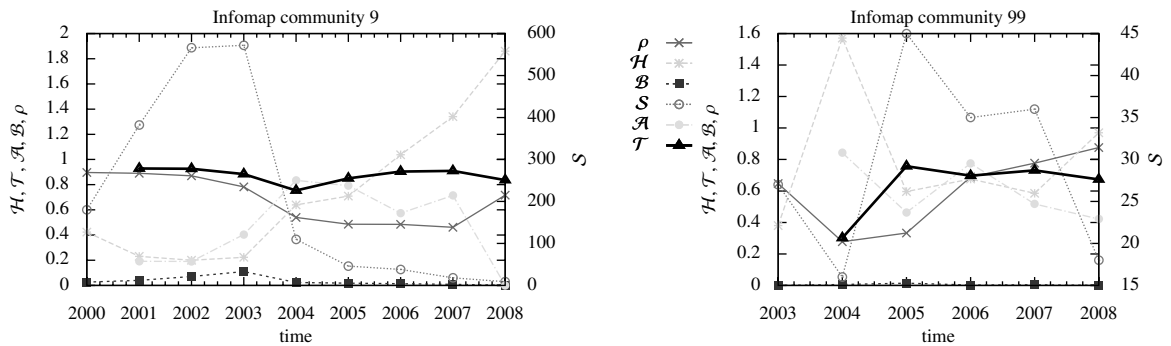


Figure 6: Life-cycle measures of Infomap communities 9 and 99

5.2. Community Shift/Merge

We assume this type of inter-community dynamics to be rather rare, as we identified only one shift/merge with absolute overlap of 11 authors and $\mathcal{P}_{S/M} \doteq 0.7$ between Infomap communities 86 and 0. Both communities were concerned with IR-related topics in general, while each had its specific theme: 86 being focused on “development”, “engine” and “system”, whereas 0 being focused on “question answering”. The merge was almost total as 90.9% authors from community 86 moved to 0 in 2003. Relative density $\rho \doteq 0.47$ and cluster quality $\mathcal{H} \doteq 1.91$ suggest that community 86 was topically coherent, but structurally rather weak. In spite of the strong topic, the community 86 thus dissolved to its related community 0.

The rareness of identified shift/merge phenomena suggests that a special *community split* measure should be investigated. The introduced community shift/merge measure may impose a too strong requirement, as it requires the whole topically distinct community to merge with another one. Likely, rather small groups separate from larger communities and merge with topically distinct ones.

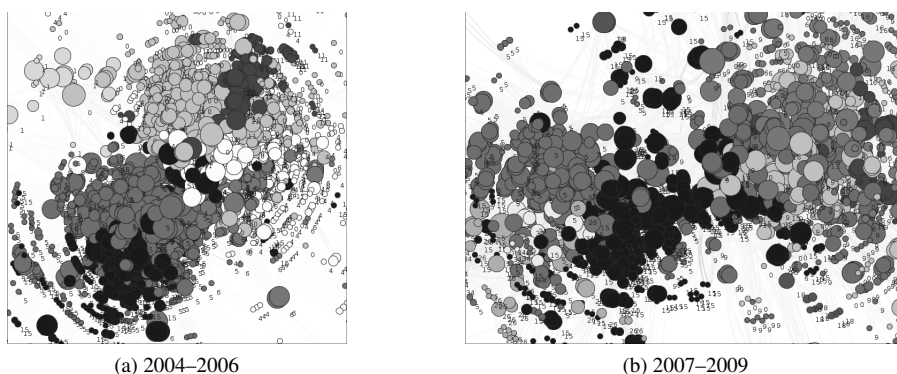


Figure 7: Snapshots of the network illustrating community 5 “semantic web” (red—left side), “information retrieval” communities 0, 4, 6 and 9 (grey, beige, pink and red—right side, respectively) and their intermediary community 15 (violet).

5.3. Community Topic Change

Louvain community 15 was detected by both community shift \mathcal{P}_S and community topic change \mathcal{P}_C measures. This community first emerged as a descendant of community 4 “information retrieval” with a specific topic “cross-language IR”, which has been detected as a shift ($\mathcal{P}_S \doteq 0.55$). In 2003, this community was under a massive influence of community 5 “semantic web”, with 53.1% of its members coming from that community. In the same year, the change of community topic ($\mathcal{P}_C \doteq 0.31$) occurred and the community began to focus mainly on SW—until 2005, when it was characterised by IR-related keywords as well. As Figure 7a depicts, it consisted of two parts in the beginning, which then merged and the whole community moved right between the SW and IR communities (see Figure 7b). Since 2004, there was not any IR-related keyword among the characterising keywords of community 5. Therefore, whereas community 5 kept its focus on the core SW-related topics, it largely participated in the formation of a new interdisciplinary community. This community, despite of still being focused mainly on SW-related themes, has functioned since then as a mutual intermediary between SW and IR communities. This hypothesis is supported by the above average betweenness \mathcal{B} in 2007, especially in contrast with the below average value in 2004 (see Table 2). Note that even though the betweenness of community 0 in 2007 was even higher ($\mathcal{B} \doteq 0.28$), this community was concentrated on core IR topics, and thus may not be perceived as an intermediary community between IR and SW disciplines. This again highlights the benefits of including content in the analysis. Further analysis of the ancestors of community 15 led to the conclusion that, despite its relations to IR-related communities 4, 0, 6 and 9, it had been mainly formed by the semantic web community 5—especially in 2003 (by 53.1%), 2004 (by 38.3%) and 2005 (by 27.5%). Therefore, efforts to establish this interdisciplinary collaboration originated mainly in the SW discipline.

Though the community topic change measure \mathcal{P}_C helped to identify interesting and relevant events, these changes were only one aspect of the involved cross-community dynamics. Further interpretation using other measures was inevitable. This was particularly the case for Louvain community 15, which was detected by more than just one community topic evolution measure. But, the intermediary character of this community was revealed by deeper analysis backed by visualisation and life-cycle measures. Group betweenness \mathcal{B} and *ancestor* measures were very helpful to gain a deeper understanding of changes of the community topic. Other life-cycle measures like \mathcal{T} , ρ , \mathcal{A} , or \mathcal{H} were not very informative in this context.

6. Conclusion and Future Work

We presented a general methodology for analysing community dynamics, uniquely combining *topological* and *topical* analysis and supported by special visualisation techniques. In this light, we focused on cross-community effects and tried to explore typical life-cycles of scientific communities. Exemplary, the methodology was applied to the co-citation network of scientists from two related research fields, IR and SW. Three *community topic evolution*

³<http://nepomuk.semanticdesktop.org>

Table 2: Sizes S and group betweenness, resp. its average, of community 15, resp. all clusters in 2004 and 2007

(a) 2004–2006				(b) 2007–2009			
cluster	S	\mathcal{B} , resp. $\langle\langle\mathcal{B}\rangle\rangle$	$\sigma_{\mathcal{B}}^2$	cluster	S	\mathcal{B} , resp. $\langle\langle\mathcal{B}\rangle\rangle$	$\sigma_{\mathcal{B}}^2$
c_{15}	444	0.0387	–	c_{15}	445	0.2629	–
all clusters	2776	0.0421	0.0045	all clusters	2190	0.0378	0.0063

measures tailored for identifying phenomena like community shift, merge/shift and change of topic were proposed and successfully assessed. We further proposed *life-cycle measures* characterising the states and evolution of communities. We identified several directions for future work, which we briefly summarise in the following.

Often, very strong shifts with $\mathcal{P}_S \rightarrow 1$ were associated with newly emerged communities, which disappeared in the next time step. To track the evolution of those communities, a larger data set that is not constrained to the a-priori chosen research fields is needed. Further, due to the reasons discussed before, the community shift/merge measure $\mathcal{P}_{S/M}$ has to be revised and extended by measures for actual *community split*. The community matching process may be improved by additionally employing the topical similarity. This would, however, conflict with the current measure for community topic change, as changes of community topic would be discarded early in the process of community matching. We observed that Louvain is particularly suited for coarse-grained quick insights, whereas Infomap is suited for fine-grained deeper analysis. This also suggests to investigate more community-detection algorithms, particularly ones producing overlapping communities, and evolutionary clustering methods.

The many cases we found for some of the investigated phenomena suggest that they are part of a usual community life-cycle. The evolution of scientific communities seems to be shaped by common mechanisms that can be identified by (a combination) of some of the proposed life-cycle measurements. Community size and cluster content ratio are particularly indicative for community shifts and community specialisation. Relative density, author entropy and topic stability are good candidates for automated analysis of the stabilisation process of new communities. Group betweenness and *ancestor* relation, together with topical analysis, have shown to be informative for detecting intermediary communities. We understand an appropriate combination of topological and topical analysis as inevitably for the accurate detection, understanding and prediction of community life-cycles.

Since the *ancestor* and *descendant* relations are defined on any two *sets* of nodes, and keyword vectors used for obtaining the similarity of clusters can be generalized to vectors of *features*, one goal for future is to generalize and apply the proposed methodology to other conceivable social communities, e.g. in discussion fora or blogosphere. With increasing availability of data describing usage of scholarly literature, we further intend to explore the possibilities of integration of usage-based measures into our framework (Bollen et al., 2009). As the presented analysis focused solely on the cross-community phenomena on the macro level, we consider an analysis and modelling of the micro level dynamics of authors and their mutual influence (Friedkin, 1998), e.g. identifying “discourse leaders” or recommending co-authorship, as a natural step in the future research.

Acknowledgement

We thank to our colleagues Georgeta Bordea and Dr. Paul Buitelaar from the Unit for Natural Language Processing for their help in extracting and processing the meta-data. We also thank to the reviewers of the manuscript for their helpful and inspiring comments. The material presented in this work is based upon works jointly supported by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2) and under Grant No. 08/SRC/I1407 (Clique: Graph & Network Analysis Cluster).

References

- Baeza-Yates, R., Mika, P., & Zaragoza, H. (2008). Search, web 2.0, and the semantic web. *IEEE Intelligent Systems*, 23, (pp. 80–82).
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, P10008.
- Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS one*, 4.

- Bordea, G. (2010). The semantic web: research and applications. chapter Concept extraction applied to the task of expert finding. (pp. 451–456). Springer.
- Chi, Y., Song, X., Zhou, D., Hino, K., & Tseng, B. L. (2007). Evolutionary spectral clustering by incorporating temporal smoothness. In *KDD '07* (pp. 153–162).
- Delvenne, J.-C., Yaliraki, S. N., & Barahona, M. (2010). Stability of graph communities across time scales. (pp. 12755–12760). volume 107.
- Everett, M., & Borgatti, S. (1999). The centrality of groups and classes. *The Journal of Mathematical Sociology*, 23, (pp. 181–201).
- Friedkin, N. (1998). *A structural theory of social influence*. Cambridge Univ. Press.
- Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Softw. Pract. Exper.*, 21, (pp. 1129–1164).
- Gmür, M. (2003). Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics*, 57, (pp. 27–57).
- Greene, D., Archambault, D., & Cunningham, P. (2010a). TextLuas: tracking and visualizing document and term clusters in dynamic text data. <http://mlg.ucd.ie/textluas>.
- Greene, D., Doyle, D., & Cunningham, P. (2010b). Tracking the evolution of communities in dynamic social networks. In *ASONAM'10*.
- Hayes, C., Avesani, P., & Bojars, U. (2007). *LNAI: Advances in Web Mining*. chapter An Analysis of Bloggers, Topics and Tags for a Blog Recommender System. Springer.
- Karnstedt, M., & Hayes, C. (2009). Towards cross-community effects in scientific communities. In *KDML'09* (pp. 52–55).
- Kuhn, T. S. (1996). *The Structure of Scientific Revolutions*. University Of Chicago Press.
- Newman, M. (2010). *Networks: an introduction*. Oxford Univ. Press.
- Palla, G., Barabási, A.-L., & Vicsek, T. (2007). Quantifying social group evolution. *Nature*, 446, (pp. 664–667).
- Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14, (pp. 130–137).
- Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *National Academy of Sciences USA*, 105, (pp. 1118–1123).
- Small, H. (2006). Tracking and predicting growth areas in science. *Scientometrics*, 68, (pp. 595–610).
- Sulo, R., Berger-Wolf, T., & Grossman, R. (2010). Meaningful selection of temporal resolution for dynamic networks. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs MLG '10* (pp. 127–136). New York, NY, USA: ACM.
- Tantipathananandh, C., Berger-Wolf, T., & Kempe, D. (2007). A framework for community identification in dynamic social networks. In *KDD '07* (pp. 717–726).
- Upham, S., & Small, H. (2010). Emerging research fronts in science and technology: patterns of new knowledge development. *Scientometrics*, 83, (pp. 15–38).
- Wakita, K., & Tsurumi, T. (2007). Finding community structure in a mega-scale social networking service. In *WWW/Internet'07* (pp. 153–162).