Michael J. Barber and Manfred M. Fischer and Thomas Scherngell

The Community Structure of R&D Cooperation in Europe. Evidence from a social network perspective

Article (Accepted for Publication)
(Refereed)

The Community Structure of R&D Cooperation in Europe[*].

Evidence from a social network perspective

Michael J. Barber[1], Manfred M. Fischer[2] and Thomas Scherngell[1]

[1]Foresight and Policy Development Department,
Austrian Institute of Technology, Vienna, Austria

[2]Institute for Economic Geography and GIScience,
Vienna University of Economics and Business, Vienna, Austria

Abstract. The focus of this paper is on pre-competitive R&D cooperation across Europe, as captured by R&D joint ventures funded by the European Commission in the time period 1998-2002, within the 5th Framework Program. The cooperations in this Framework Program give rise to a bipartite network with 72,745 network edges between 25,839 actors (representing organizations that include firms, universities, research organizations and public agencies) and 9,490 R&D projects. With this construction, participating actors are linked only through joint projects.

In this paper we describe the community identification problem based on the concept of modularity, and use the recently introduced label-propagation algorithm to identify communities in the network, and differentiate the identified communities by developing community-specific profiles using social network analysis and geographic visualization techniques. We expect the results to enrich our picture of the European Research Area by providing new insights into the global and local structures of R&D cooperation across Europe.

[*] Accepted for publication in *Geographical Analysis*

1 Introduction

Knowledge production takes place within a complex web of interactions among firms, universities and research institutions (see, for instance, Fischer et al. 2006, Autant-Bernard et al. 2007, Fritsch and Kauffeld-Monz 2010). Long viewed as a temporary, inherently unstable organisational arrangement, R&D networks have become the norm, rather than the exception, in modern innovation processes (Powell and Grodal 2005). In the recent past, regional, national and supranational Science, Technology and Innovation (STI) policies have emphasized supporting and fostering linkages between innovating actors (for a discussion of major international examples, see Caloghirou et al. 2002). At the European level, the main STI policy instruments are the European Framework Programmes (FPs) that promote an integrated European Research Area (ERA). The FPs support pre-competitive R&D projects, creating a pan-European network of actors performing joint R&D.

In this paper, we examine pre-competitive European[1] R&D cooperations from a social network perspective, which focuses not on the individual social actors, but on the broader interaction contexts within which the actors are embedded. The notion of a social network and the procedures of social network analysis have attracted considerable interest and curiosity from the social science community in recent years. Much of this interest can be attributed to the appealing focus of social network analysis on relationships among social actors, and on the patterns and implications of these relationships. The relationships may be of many kinds: economic, political, interactional, or affective, to mention a few. The focus on relations, and the patterns of relations, requires a set of procedures and analytical concepts that are distinct from methods of conventional statistics and data analysis.

As observed by Ter Wal and Boschma (2009, p. 793), “the potential of the application of network methodology to regional issues is far from exhausted.” Indeed, because networks are a natural and general way to represent and analyze relationships of all sorts, networks have been considered across the scientific spectrum, ranging from the social sciences to the natural sciences to pure mathematics. We hope to benefit from the great potential of this diverse literature, focusing initially on the possibilities offered by identifying communities in social networks. For regional science, in particular, community identification enables us to detect and investigate appropriate substructures of large social systems, such as Framework Programmes.

Social network analysis explicitly assumes that actors participate in social systems connecting them to other actors, whose relations comprise important influences on one another's behaviors. Central to network analysis are identifying, measuring, and testing hypotheses about the structural forms and substantive contents of relations among actors. This distinctive structural-relational emphasis sets social network analysis apart from individualistic, variable-centric traditions in the social sciences (Knoke and Young 2008).

The importance of social network analysis rests on two underlying assumptions. First, structural relations often are more important for understanding observed behaviors than are attributes of the actors. Second, social networks affect actors' perceptions, beliefs and actions through a variety of structural mechanisms that are socially constructed by relations among them. Direct contacts and more intensive interactions dispose actors to better information, greater awareness, and higher susceptibility to influencing or being influenced by others. Indirect relations through intermediaries also bring exposure to new ideas, and access to useful resources that may be acquired through interactions with others. Networks provide complex pathways for assisting or hindering flows of information and knowledge.

In this paper, the focus is on networks derived from R&D joint ventures funded by the European Commission in the time period 1998-2002, within the 5th Framework Programme (FP5). The Programme gives rise to a bipartite network with 72,745 edges existing between 9,490 projects and 25,839 actors representing formal organizations such as firms, universities and research organizations. With this construction, participating actors are linked through joint projects. The objective is to detect and describe the community structure of this network. Community detection may be loosely defined as partitioning the nodes or vertices into groups such that there is a higher density of links within them than between them. The definition is based on comparing intra-group density to inter-group sparseness. The popularity of density-based grouping is due to the likelihood that actors within communities share common properties and/or play similar roles within a network, and thus constitute a relevant subnetwork to consider in some detail. This is the motivation for analyzing network communities in general, and European R&D network communities in particular.

The usual approach for community detection in bipartite networks is to first construct a unipartite projection of one part of the network (i.e., a network of organizations by linking them when they cooperate in a project), and then to identify communities in that projection

using methods for unipartite networks. This unipartite projection can be illuminating, but intrinsically looses information. In this study, we use the recently introduced label-propagation algorithm (LPA) to explicitly account for the bipartite character of networks (see Raghavan et al. 2007, Barber and Clark 2009).

The paper is organized as follows. *Section 2* describes the community identification problem based on the concept of modularity. *Section 3* introduces the LPA to identify communities in the network under consideration. The LPA was originally presented operationally, with communities defined as the outcome of a specific procedure. In this paper, we consider an equivalent mathematical formulation, in which community solutions are understood in terms of optima of an objective function. *Section 4* differentiates the identified communities by developing community-specific profiles using social network analysis and geographic visualisation techniques. *Section 5* concludes with a summary of the main results, and a brief outlook.

2 The community-identification problem

A network of R&D cooperation can be viewed in several ways. One of the most useful views is as a graph consisting of vertices (nodes) and edges (links). A familiar representation is obtained by letting V be a set of vertices representing actors participating in FP5, and E be a set of vertex pairs or edges from $V \times V$, representing participation in a joint FP5 project[2]. The two sets together are a graph $G=(V, E)$. This is called a simple graph, because all pairs $\{u, v\} \in E$ are distinct and $\{u, u\} \notin E$.

Given a partition $V=V_1+V_2$ where no edges exist between pairs of elements within V_1 or V_2, then G is said to be bipartite. We can represent R&D cooperations as a bipartite graph, letting V_1 be a set of vertices representing actors participating in FP5, and V_2 be a set of vertices representing the projects funded in FP5, with an edge between two vertices if and only if one vertex is a project (and thus in V_2) and the other is an actor (and thus in V_1) that takes part in the project. The bipartite graph can be used to define the previously described graph of actors as a projection: define edges between actors when the actors are separated by a path of length two in the bipartite graph. The converse is not true. Thus the bipartite graph contains more information than the actor graph, and can be advantageous to use. In this paper, we focus principally on the bipartite network of actors and projects

We consider simple graphs on a large finite set $V=\{1, 2, \ldots, n\}$. The number of edges in a graph is denoted by m, and the number of edges incident on a vertex $i=1, \ldots, n$ is called the degree k_i. The connectivity pattern of a graph is encoded in the $n \times n$ adjacency matrix $\boldsymbol{A}$ with elements

$$A_{ij} = \begin{cases} 1 & \text{if } \{i, j\} \in E \\ 0 & \text{otherwise.} \end{cases} \qquad i, j=1, \ldots, n \qquad (1)$$

In many real world networks, the vertices vary widely in their degrees, reflecting a high level of order and structure. The degree distribution is highly skewed; many vertices with low degrees coexist with some vertices with high degrees. The distribution of edges may be both globally and locally heterogeneous, with a high concentration of edges within specific groups of nodes, and a low concentration between these groups. This feature of real world networks is called community structure.

A traditional approach to identifying community structure is simply to draw its network, positioning vertices close to one another when they are connected and farther apart when they are not, and identify the communities by eye. This approach works well for small-sized networks, and is viable for networks of tens or perhaps hundreds of vertices by means of computer-aided drawing using applications such as Pajek[3], UCINet[4], Graphviz[5], or Gephi[6]. However, the visual approach fails if the number of vertices is larger because the display becomes too cluttered. Drawing the 35,329 vertices of the FP5 R&D network as dots on a page would require the dots to be placed about 1 mm from each other, and we would still need to draw the edges that link them. Larger networks with millions or billions of vertices and edges are impossible to draw in practice. For all but the smallest networks, we must investigate statistical properties of the network connectivity patterns in order to "see" the community structure.

There is a plethora of ways to define the community-identification problem (for recent reviews, see Porter et al. 2009, Fortunato 2010). The most prominent formulation is based on the concept of modularity, a measure that evaluates the quality of a partition of a graph into subsets of vertices in comparison to a null model. Formally, the modularity Q is defined as

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - P_{ij} \right) \delta \left(g_i, g_j \right) \tag{2}$$

with the Kronecker delta term

$$\delta \left(g_i, g_j \right) = \begin{cases} 1 & \text{if } g_i = g_j \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where g_i and g_j denote the community groups to which vertices i and j are assigned, respectively, and P_{ij} denotes the probability in the null model that an edge exists between vertices i and j. Thus, the modularity Q is – up to a normalization constant – defined as the number of edges within communities minus those expected in the null model.

The standard choice of the null model is that proposed by Newman and Girvan (2004), and consists of a randomized version of the actual graph, where edges are rewired at random, under the constraint that each vertex i keeps its degree k_i. Assuming that P_{ij} may be written in the product form

$$P_{ij} = P_i \; P_j \,, \tag{4}$$

$$P_i = k_i \big/ \sqrt{2m} \;\text{ and }\; P_j = k_j \big/ \sqrt{2m} \,,$$

and, thus,

$$P_{ij} = \frac{k_i \; k_j}{2m}. \tag{5}$$

With this choice for P_{ij}, the modularity becomes

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ik} - \frac{k_i \; k_j}{2m} \right) \delta \left(g_i, g_j \right). \tag{6}$$

The goal now is to find a division of the vertices into communities such that the modularity Q is optimal. An exhaustive search for a decomposition is infeasible. Even for moderately large networks, far too many ways exist to decompose such networks into communities.

Here we study the bipartite character of the network in question. Bipartite networks have additional constraints that can be reflected in the null model. For bipartite graphs, the null model should be modified to reproduce the characteristic form of bipartite adjacency matrices (see Barber 2007 for more details)

$$\mathbf{A} = \begin{bmatrix} O_{n_1 \times n_1} & A_{n_1 \times n_2} \\ \left(A\right)_{n_2 \times n_1} & O_{n_2 \times n_2} \end{bmatrix} \tag{7}$$

where n_1 and n_2 denote the number of vertices in V_1 and V_2, respectively, and $n = n_1 + n_2$, and $O_{n_1 \times n_2}$ is the all-zero matrix with n_1 rows and n_2 columns. Using this null model, the following bipartite modularity Q_B is obtained:

$$Q_B = \frac{1}{2m} \sum_{i,j} \left(A_{ik} - \frac{2k_u d_v}{m} \right) \delta\left(g_i, g_j\right). \tag{8}$$

In Eq. (8), the degrees for the two parts of the network are handled separately as k_u and d_v. For the network that we consider here, k_u denotes the degree for vertices representing organizations (with k_u=0 for projects) and d_v denotes the degree for vertices representing projects (with d_v=0 for organizations).

3 A label-propagation algorithm for maximizing (bipartite) modularity

We detect network communities using an approach that builds on the label propagation algorithm (LPA) introduced by Raghavan et al. (2007). In the LPA, community assignments are described by labels assigned to the network vertices. Vertices are initially assigned unique labels; these labels may be numbers. Labels propagate dynamically between vertices, with the new label for a vertex assigned to match the most frequent label among the neighboring vertices. The relabeling is illustrated in Fig. 1, where a new label is assigned to the vertex marked with a question mark. The most frequent label among the neighbors is “2” and hence

the vertex also takes this label. Once a stable assignment of labels is obtained, network communities are taken to be sets of vertices bearing the same labels.

Figure 1 about here

We formalize the LPA following the presentation of Barber and Clark (2009), describing the LPA as an optimization problem. We introduce an objective function *H*, which is just the number of edges linking vertices with the same label (i.e., in the same community group *g*). This function can be expressed formally in terms of the adjacency matrix $\boldsymbol{A}$, giving

$$H = \frac{1}{2}\sum_{u,v} A_{uv}\ \delta(g_u, g_v), \tag{9}$$

where, as before, g_u and g_v denote the community groups (i.e. labels) to which vertices *u* and *v* are assigned. Label assignment corresponds to selecting a new community group g'_v for vertex *v* that maximizes *H* (i.e., a label *g* that occurs most frequently among the neighbors of *v*). Formally, this is

$$g'_v = \arg\max_{g} \sum_{u} A_{uv}\ \delta(g_u, g). \tag{10}$$

Multiple choices of *g* could produce a maximal *H*. In such a case, a specific label is selected by keeping the current label if it would satisfy Eq. (10), and otherwise taking a label at random that satisfies Eq. (10). This decision rule excludes non-terminating cycles where a vertex varies between different labels satisfying Eq. (10).

To put the label-update rule (10) into effect, we must also define an update schedule. A practical schedule, suggested by Raghavan et al. (2007), is to update the vertex labels asynchronously and in random order. Multiple updating passes are made through the vertices, continuing until all vertices have labels satisfying Eq. (10). This update schedule ensures termination of the search by eliminating cycles where two neighboring vertices continually exchange labels.

The LPA offers a number of desirable qualities. As previously described, it is conceptually simple, being readily understood and quickly implemented. The algorithm is efficient in practice. Each relabeling iteration through the vertices of a graph has a computational complexity linear in the number of edges in the graph. The total number of iterations is not a priori clear, but relatively few iterations are typically needed to assign the final label to most of the vertices (over 95% of vertices in five iterations; see Raghavan et al. 2007, Leung et al. 2008).

A significant drawback of the LPA is that the objective function H corresponds poorly to our conceptual understanding of communities. In fact, the global maximum in H is trivially obtained by assigning the same label to all vertices, providing no information at all about community structure. Thus, interesting community solutions must be located at local maxima in H, but H offers no mechanism for comparing the quality of the solutions. An auxiliary measure, such as the modularity Q, can be introduced to assess community quality. Using modularity, communities found using LPA are seen to be of high quality (Raghavan et al. 2007); label propagation is both fast and effective.

Barber and Clark (2009) have elucidated the connection between label propagation and modularity, showing that modularity can be maximized by propagating labels subject to additional constraints, and proposing several variations of the LPA. In this paper, we make use of a hybrid, two-stage label propagation scheme, consisting of the LPAr variant followed by the LPAb variant (see Barber and Clark 2009 for details). The LPAr is defined similarly to the original LPA, but with additional randomness to allow the algorithm to avoid premature termination. Instead of preferentially keeping the current label if it would satisfy Eq. (10), in the LPAr we always select randomly from those labels that satisfy Eq. (10). As this assignment could, in principle, prevent the algorithm from terminating, we consider the label propagation to be complete when no label changes in a pass through the vertices, rather than the more stringent condition that no label could change. This practice produces better communities as measured by Q or Q_B than does the LPA. The LPAb imposes constraints on the label propagation so that the algorithm identifies a local maximum in Q_B using a modified label update rule with the form

$$g'_v = \arg\max_g \sum_u \left(A_{uv} - \frac{2}{m} k_u d_v \right) \delta(g_u, g) . \qquad (11)$$

Update rule (11) can be implemented in such a fashion as to preserve the desirable properties of the LPA while imposing a clearer measure for community quality than that in Eq. (9).

4 Network communities and topical differentiation

In this section, we use the LPA algorithm to identify and differentiate communities for the European R&D cooperation network. We develop community-specific profiles to thematically characterize the network communities, and consider their spatial distribution. We identified 3,482 network communities. The communities vary greatly in size, as measured either by the number of organizations in a community or by the number of projects in a community (ranked by size in Fig. 2). Most (2,878) communities consist of just a single project with some or all of the participating organizations. In contrast, twenty or more projects are observed in just nine communities, but they contain over a third of the organizations and over half of the projects present in FP5. For the rest of this paper, we consider only eight of these nine largest communities (see Table 1 and Fig. 3); the ninth is of a different character than the others, focusing on international cooperation rather than R&D.

Figure 2 about here

Thematic differentiation and characterization of the network communities

Communities are identified using only the network structure, which arises from the processes by which projects form. To gain a better understanding of the nature of the communities, we examine the properties of the constituent organizations and projects. We focus particularly on three characteristics: (*i*) the standardized subject indices (sometimes also referred to as keywords) assigned to the projects by the EU, (*ii*) the project titles, and (*iii*) the identity of the organizations. By considering these three features, we find a strong thematic character for the communities. We summarize the community themes concisely in Table 1 and provide additional details here.

As a first step, we gain a basic understanding of the communities by examining their thematic orientation using standardized subject indices assigned to the projects in a community. There are 49 subject indices in total, ranging from *Aerospace Technology* to *Waste Management*; a

complete list of subject indices is given by CORDIS (2008). Absolute counts of projects with a particular subject index are uninformative, as the subject indices occur with different frequencies in FP5 projects. A more meaningful assessment is to compare the number of projects N_s in a community featuring a subject index S to the number $E[N_s]$ we would expect if the projects were chosen at random from FP5; differences in the values can be tested for statistical significance using a binomial test. In Table 1, we show the most strongly over-represented subject indices for each community, giving the values as a ratio $R_s = N_s / E[N_s]$ of actual occurrences (N_s) to expected occurrences of the index ($E[N_s]$). The subject indices are strongly suggestive of thematic differentiation between the communities, with communities apparently oriented toward the life sciences, transportation, electronics, and other topics.

Further insight into the communities is gained by examining the project titles, allowing a more specific characterization of their thematic character. Particularly for the larger communities, the titles suggest possible community substructures of a more specialized nature; we note the presence of such subnetworks, but do not pursue them further in this work. Using the standardized subject indices and the project titles, we assigned the names as shown in Table 1 to each community.

Table 1 about here

Larger communities show greater diversity in their substructure. The largest, *Life Sciences*, shows a broad selection of topics in biotechnology and the life sciences, including health, medicine, food, molecular biology, genetics, ecology, biochemistry, and epidemiology. The second largest, *Electronics*, focuses principally on information technology and electronics, with projects in related fields dealing with materials science, often related to integrated circuits; projects about algorithms, data mining, and mathematics, and a definite subset of projects with atomic, molecular, nuclear, and solid state physics. The third largest community, *Environment*, is focused on environmental topics, including environmental impact, environmental monitoring, environmental protection, and sustainability.

As communities become smaller, they also become more focused. We see, for example, three distinct transportation related communities. The largest of these, *Aerospace*, is focused on aerospace, aeronautics and related topics, including materials science, manufacturing, fluid mechanics, and various energy topics. The next, *Ground Transport*, has projects dominated

by railroad and, especially, automotive topics; notable subtopics include manufacturing, fuel systems, concrete, and pollution. The smallest transportation community, *Sea Transport*, is more specifically focused; virtually all project titles are shipping-related. The remaining communities, *Aquatic Resources* and *Information Processing*, are the smallest and thematically most uniform.

In Fig. 3, we visualize the network of key FP5 communities using a standard approach from spectral graph analysis, so that communities that show a relatively higher number of links between them are positioned nearer to each other. The vertices are positioned by taking the *x* and *y* coordinates to be the components of two eigenvectors of the normalized Laplacian matrix $\mathbf{L} = (l_{ij})_{n \times n}$ that is defined as

$$l_{ij} = \begin{cases} 1 & \text{if } i = j \text{ and } k_i \neq 0 \\ -\left[k_i \; k_j\right]^{-1/2} & \text{if } i \neq j \text{ and } k_i \text{ adjacent to } k_j \\ 0 & \text{otherwise} \end{cases} \qquad (12)$$

where k_i denotes the degree of vertex i. Matrix **L** can be written as

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2} \, \mathbf{A} \; \mathbf{D}^{-1/2} \qquad (13)$$

where **A** is the $n \times n$ adjacency matrix as defined in Section 2, **I** is the $n \times n$ identity matrix, and **D** is the $n \times n$ diagonal matrix with $D_{ij} = \lambda_i$ $(i = 1, ..., n)$ the *i*th eigenvalue of **L**. The relevant eigenvectors for network layout are those corresponding to the two smallest positive eigenvalues. The normalized Laplacian matrix is much studied in spectral graph analysis (Chung 1997), and is of great practical use in data clustering and visualization (Higham 2004, Seary and Richards 2003).

A node size corresponds to the number of organizations of the respective community. The *Life Sciences* and the *Electronics* community have the highest number of organizations. The *Electronics* community appears to have the highest collaboration intensity with other communities (i.e., knowledge produced in this field is used intensively in other fields). The *Life Sciences* community has a strong connection to the third largest community, *Environment*. The three transport-related communities are positioned on the left-hand side of

Figure 3; i.e., they show relatively high inter-community collaboration intensity. The largest of these is *Aerospace,* which is closer to *Ground Transport* than to *Sea Transport*. The community *Aquatic Resources* has the strongest connection to *Environment*, while *Information Processing* is far from all other communities.

Figure 3 about here

The structure of the network communities

Table 2 provides an overview of some measures that characterize the structure of the eight FP5 communities under consideration. We focus here on the partnerships present, and thus turn attention to the organizations projection graphs for the communities. Some differences in the network structure are worth noting. As indicated in the preceding subsection, the number of vertices, and thus the number of organizations, in a community is highest for the *Life Sciences* and *Electronics*. Though the number of organizations in these two communities is nearly equal, the number of edges is markedly higher in the *Life Sciences* community than in the *Electronics* community, leading to a higher density in the *Life Sciences* community. The average path length also varies across the eight communities. It is highest for the *Environment* community (2.797), though it has a lower number of vertices than the *Life Sciences* and the *Electronics* community; i.e., from a social network analysis perspective, the condition for diffusion of information is better in the latter two communities than in the *Environment* community. In all cases, the distribution of vertex degree (i.e., the number of partners) is skewed rightward. This skewness is highest for the *Ground Transport* community, with a value of 6.739. Compared to the other communities, *Ground Transport* features central hubs that are in many more projects than the other organizations, and are of great importance for the spread of information in the network.

Table 2 about here

Spatial patterns of the network communities

We next consider the spatial distribution of the eight FP5 communities. Figure 4 illustrates the projection of the communities onto NUTS-2 regions across Europe. The 255 regions cover the pre-2007 EU25 member states, as well as Norway and Switzerland. Note that the region-by-region community networks are undirected, weighted graphs from a network analysis

perspective. The nodes represent regions; their size is relative to their degree centrality, corresponding to the number of links connected to a region.

The spatial network maps in Fig. 4 reveal considerable differences in spatial collaboration patterns across eight FP5 communities. One important result is that the region Île-de-France takes an important position in all communities[7]. Furthermore, this visualization clearly discloses the different spatial patterns of the Transport related communities, *Aerospace*, *Ground Transport*, and *Sea Transport*. Though the region Île-de-France appears to be the central hub in all transport-related communities, the directions of the largest collaboration flows from Île-de-France differ markedly. For the *Sea Transport* community, we observe intensive collaborations with important sea ports in the north (Zuid Holland, Agder Rogeland, Denmark, Hamburg) and the south (Liguria, Lisbon, Athens), while for the *Ground Transport* community, collaborations with the east and south are dominant (Lombardia, Oberbayern, Stuttgart). In the *Aerospace* community, we can observe a strong localization of collaborations within France and its neighboring countries.

Figure 4 about here

In the largest community, *Life Sciences*, the highest number of collaborations is observed between the regions of Île-de-France and Piemonte (174), while the second largest community, *Electronics*, is characterized by a very high collaboration intensity between the regions of Île-de-France and Oberbayern (474 collaborations), followed by Île-de-France and Köln (265 collaborations), and Oberbayern and Köln (157 collaborations). In the *Environment* community, we find the strongest collaboration intensity between Denmark and Helsinki (131 collaborations). In the *Aquatic Resources* community, the regions Denmark and Agder Rogaland (Norway) show the highest collaboration intensity, not only between them (21 collaborations), but also with other regions, while for the *Information Processing* community, we identify Helsinki as the central region, featuring intensive collaboration with Athens, Lazio and Lombardia.

5 Summary and conclusions

In this paper, we employ recently developed methods to identify communities in European R&D networks using data from joint research projects funded by the European Framework

Programmes (FPs). The identification and characterization of thematically relevant substructures in these networks is of crucial importance in a European policy context. The present study complements earlier empirical work about the structure of R&D networks in Europe that neglect relevant substructures (see, for instance, Breschi and Cusmano 2004). To our knowledge, the current study is the first to apply the community detection methodology for identifying the relevant subnetworks in a regional science perspective.

Networks of R&D collaborations under the 5th Framework Programme give rise to a bipartite network, with 72,745 edges existing between 9,490 projects and 25,839 organizations which take part in them. With this construction, participating organizations are linked only through joint projects. The usual approach taken to identify communities in bipartite networks is to first construct a unipartite projection of one part of the network (i.e., a network of organizations by linking organizations when they cooperate in a project), and then detect communities in that projection using methods for unipartite networks. The unipartite projection can be illuminating, but intrinsically loses information because multiple bipartite networks with distinct connectivity patterns can give rise to the same projection (see Barber 2007). In this paper, we adopted a label propagation algorithm (LPA) for identifying community groups in this bipartite network. The LPA is designed for maximizing bipartite modularity that accounts for the bipartite character of the network (see Barber and Clark 2009). The advantages of the procedure are its conceptual simplicity, ease of implementation, and practical efficiency.

This study produces interesting results, both from a scientific point of view, and in a European policy context. We detect eight relevant, thematically relatively homogenous FP5 communities, providing a new view on the R&D collaboration landscape in Europe. The larger communities identified are *Life Sciences*, *Electronics*, and *Environment*. However, these communities may show further relevant substructures. As communities become smaller, they also become more focused. We identify three transport related communities: *Aerospace*, *Ground Transport*, and *Sea Transport*. The remaining communities, *Aquatic Resources* and *Information Processing*, thematically are the smallest and most uniform ones. Furthermore, results clearly reveal that the geographical distribution of the communities varies considerably. However, the region of Île-de-France plays a central role in each of the detected communities.

Further, we illustrate that the application of network analysis techniques has great potential in a regional science and spatial analysis context. In particular, the detection and investigation of substructures in social systems is of great relevance in regional science, and, thus, enhances our analytical toolbox for the spatial analysis of such social systems. By this, the study provides an important starting point for further employing and improving community detection algorithms for analyzing substructures of (spatial) R&D networks.

The general approach followed in this study may be extended and improved upon in several ways. Alternate community detection methods may be considered. More significantly, alternate definitions of what we mean by community may be considered, allowing investigation of hierarchical substructures of the communities, or community to overlap. Other methods from social network analysis may be used to characterize the network, and techniques from spatial analysis may be applied to characterize the network as a whole and its community structure.

Acknowledgments. The authors gratefully acknowledge the grant no P21450 provided by the Austrian Science Fund (FWF).

Endnotes

[1] R&D networks constituted under the heading of the FPs have recently attracted a number of empirical studies. Maggioni et al. (2007), and Scherngell and Barber (2009) focus on the geography of pre-competitive R&D networks across European regions by using data from joint research projects of FP5. Breschi and Cusmano (2004) employ a social network perspective to analyse R&D collaborations, with the objective to unveil the texture of the European Research Area (ERA); but their work predates the explosion of papers about network analysis following the seminal paper by Newman and Girvan (2004) introducing modularity.

[2] We use data from the EUPRO database, which comprises systematic information on funded research projects of the EU FPs (complete for FP1-FP6) and all participating organizations (see Roediger-Schluga and Barber 2008 for further details).

[3] http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[4] http://www.analytictech.com/ucinet/

[5] http://www.graphviz.org/

[6] http://gephi.org/

[7] We stress, however, that one cannot conclude from this finding that individual Parisian scholars or organizations are the most important for the communities.

References

Autant-Bernard C, Mairesse J and Massard N (2007) Spatial knowledge diffusion through collaborative networks, *Papers in Regional Science* 86(3), 341-350

Barber MJ (2007) Modularity and community detection in bipartite networks, *Physical Review E* 76(6), 066102

Barber MJ and Clark J (2009) Detecting network communities by propagating labels under constraints, *Physical Review E* 79(3), 032456

Breschi S and Cusmano L (2004) Unveiling the texture of a European research area: Emergence of oligarchic networks under EU Framework Programmes, *International Journal of Technology Management. Special Issue on Technology Alliances* 27(8), 747-772

Caloghirou Y, Vonortas NS and Ioannides S (2002) Science and technology policies towards research joint ventures, *Science and Public Policy* 29 (2), 82-94

Chung, F. R. K. (1997). *Spectral Graph Theory*. CBMS Regional Conference Series in Mathematics. American Mathematical Society, Providence, RI.

CORDIS (2008) List of CORDIS Subject Index Classification Codes, Available from http://cordis.europa.eu/guidance/sic-codes_en.html

Fischer MM, Scherngell T and Jansenberger EM (2006) The geography of knowledge spillovers between high-technology firms in Europe. Evidence from a spatial interaction modelling perspective, *Geographical Analysis* 38 (3), 288-309

Fortunato S (2010) Community detection in graphs, *Physics Reports* 75(3-5), 174-189

Fritsch, M and Kauffeld-Monz M (2010) The impact of network structure on knowledge transfer: an application of social network analysis in the context of regional innovation networks, *The Annals of Regional Science* 44(1), 21-38

Higham DJ and Kibble M (2004) A unified view of spectral clustering. Mathematics Research Report 02, University of Strathclyde

Knoke D and Young S (2008) *Social network analysis*. Sage Publications, Los Angeles, London, New Delhi and Singapore

Leung, IXY, Hui P, Liò P and Crowcroft J (2009) Towards real-time community detection in large networks, *Physical Review E* 79(6), 066107

Maggioni M.A, Nosvelli M and Uberti TA (2007) Space versus networks in the geography of innovation: A European analysis, *Papers in Regional Science* 86, 471-493

Newman MEJ and Girvan M (2004) Finding and evaluating community structure in networks, *Physical Review E* 69(2), 026113

Porter MA, Onnela J-K, and Mucha PJ (2009) Communities in networks, *Notices of the American Mathematical Society* 59(9), 1082-1097+

Powell WW and Grodal S (2005) Networks of innovators, In Fagerberg J, Mowery DC and Nelson RR (eds.), *The Oxford Handbook of Innovation*, Oxford, Oxford University Press, 56-85

Raghavan UN Albert R and Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks, *Physical Review E* 76(3), 036106

Roediger-Schluga T and Barber M (2008) R&D collaboration networks in the European Framework Programmes: Data processing, network construction and selected results, *International Journal of Foresight and Innovation Policy* 4(2), 321–347

Scherngell T and Barber MJ (2009) Spatial interaction modelling of cross-region R&D collaborations. Empirical evidence from the 5th EU Framework Programme, *Papers in Regional Science* 88(3), 531-546

Seary AJ and Richards WD (2003) Spectral methods for analyzing and visualizing networks: an introduction, In Breiger R, Carley K, and Pattison P (eds), *Dynamic Social Network Modeling and Analysis*, pp 209–228, Washington, D.C. The National Academies Press

Ter Wal, ALJ and Boschma R (2009) Applying social network analysis in economic geography: framing some key analytic issues, *The Annals of Regional Science* 43(4), 739-756

Fig. 1: Updating community assignment by propagating labels

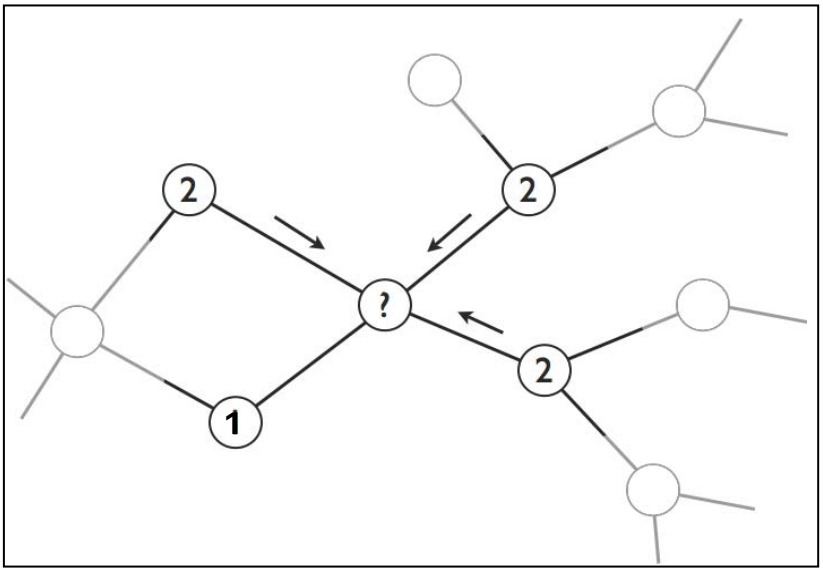


Fig. 2: Ranking of communities by number of organizations

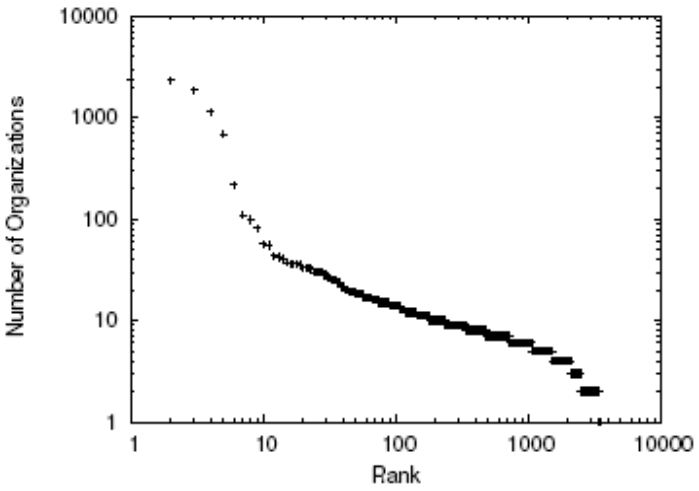
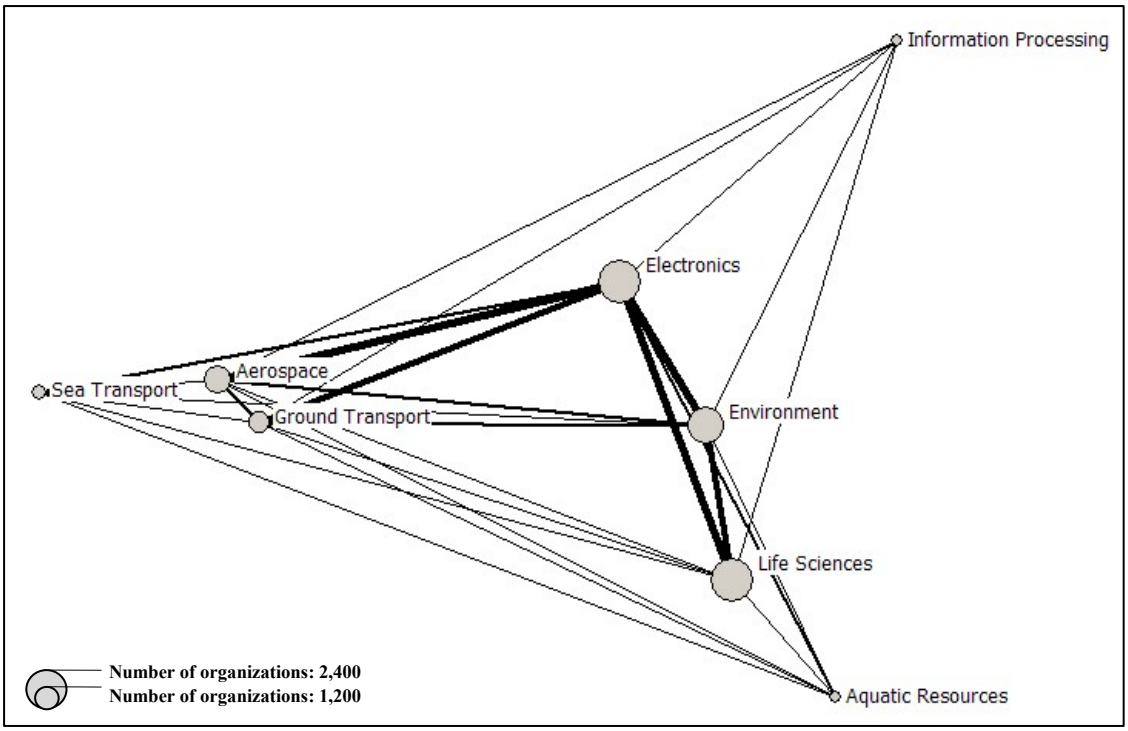


Fig. 3: Community structure in the network of R&D cooperation



Note: Vertex positions were determined using spectral graph analytic methods so that communities that are strongly interconnected are positioned nearer to each other. With these positions, the network was visualized using UCINet 6.303.

Fig. 4: Spatial patterns of the FP5 communities

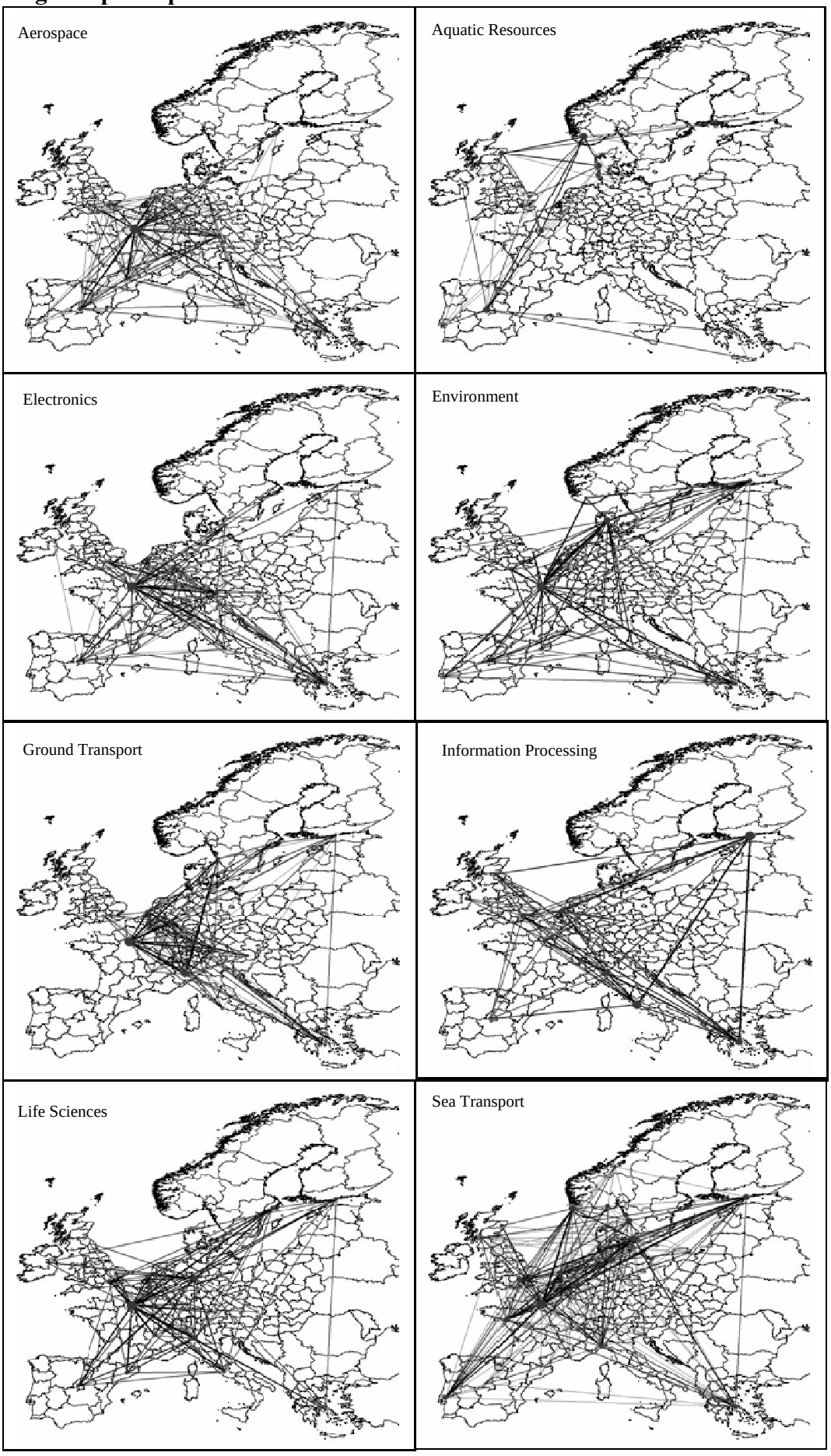


Table 1: Characterization of communities by thematic orientation as captured by $R_s = N_s / E[N_s]$, where N_s are the actual occurrences and $E[N_s]$ the expected occurrences of a specific subject index *s*

	$R_s > 5$	$5 \geq R_s > 3$	$3 \geq R_s > 1$
Aerospace	Aerospace Technology***	Energy Saving***; Energy Storage, Energy Transport***; Renewable Sources of Energy***; Transport***	Industrial Manufacture***; Information Processing, Information Systems***; Other Energy Topics**
Aquatic Resources	Agriculture***; Resources of the Sea, Fisheries***	Life Sciences***	Economic Aspects***; Environmental Protection***
Electronics	-	Electronics, Microelectronics***; Evaluation*; Telecommunications***	Education, Training***; Forecasting***; Information Processing, Information Systems***; Media***
Environment	Earth Sciences***; Meteorology***; Standards***	Forecasting***; Resources of the Sea, Fisheries***	Agriculture*; Environmental Protection***; Measurement Methods**; Regional Development*; Scientific Research*;
Ground Transport	Energy Storage, Energy Transport***	Fossil Fuels**	Energy Saving***; Environmental Protection*; Materials Technology*; Reference Materials*; Safety***
Information Processing	Electronics, Microelectronics***; Legislation, Regulations***; Mathematics, Statistics***; Policies***	-	Information Processing, Information Systems***
Life Sciences	-	Biotechnology***; Life Sciences***; Medicine, Health***; Regional Development***	Agriculture***; Food***; Policies***; Safety***; Scientific Research***; Social Aspects***; Waste Management***
Sea Transport	Transport***	Safety***	Environmental Protection***

Notes: Statistical difference tested using binomial tests whether N_s is different from $E[N_s]$
***significant at the 0.001 level, **significant at the 0.01 level, *significant at the 0.05 level

Table 2: Properties of the FP5 communities

	Aerospace	Aquatic Resources	Electronics	Environment	Ground Transport	Information Processing	Life Sciences	Sea Transport
Vertices *n*	1,146	81	2,307	1,855	686	40	2,366	218
Edges *m*	13,870	451	30,456	23,155	5,251	226	33,178	2,978
Average path length	2.669	2.199	2.732	2.797	2.549	1.731	2.713	2.030
Density	0.021	0.139	0.010	0.013	0.022	0.290	0.012	0.126
Skewness	4.263	1.169	5.132	4.512	6.739	1.097	4.749	1.718
Mean degree	24.206	11.136	26.403	24.965	15.309	11.300	28.046	27.321