Is the European R&D network homogeneous? Distinguishing relevant network communities using graph theoretic and spatial interaction modeling approaches

Michael J. Barber and Thomas Scherngell*

*Corresponding author,
Foresight and Policy Development Department,
Austrian Institute of Technology (AIT), Vienna, Austria
e-mail: thomas.scherngell@ait.ac.at

December 2010

Abstract. Interactions between firms, universities, and research organizations are crucial for successful innovation in the modern knowledge-based economy. Systems of such interactions constitute R&D networks, which may be meaningful segmented using recent methods for identifying communities, subnetworks whose members are more tightly linked to one another than to other members of the network. In this paper we identify such communities in the European R&D network using data on joint research projects funded by the fifth European Framework Programme. We characterize the identified communities according to their thematic orientation and spatial structure. By means of a Poisson spatial interaction model, we estimate the impact of various separation factors – such as geographical distance – on the variation of cross-region collaboration activities in a given community. The European coverage is achieved by using data on 255 NUTS-2 regions of the 25 pre-2007 EU member-states, Norway, and Switzerland. The results demonstrate that European R&D networks are not homogeneous, showing distinct, relevant substructures characterized by thematically homogeneous and spatially heterogeneous community groups.

JEL Classification: O38, L14, R15

Keywords: R&D networks, network communities, spatial interaction models, EU Framework Programmes

1 Introduction

Today it is widely believed that interaction between firms, universities and research organizations is crucial for successful innovation in the knowledge-based economy, in particular in knowledge-intensive industries. This gives rise to the notion of *R&D networks,* defined as a set of organizations performing joint R&D, for instance in the form of collaborative research projects, joint conferences and workshops, or shared R&D resources in the form of labor and capital (see, for instance, Powell and Grodal 2005). From a policy perspective, when acknowledging, *first*, that R&D networks are crucial for innovation and, *second*, that innovation is crucial for sustained economic growth (see Romer 1990), it seems elemental that modern STI policies emphasize supporting and fostering linkages between innovating actors. The principal European example of such STI policy instruments are the European Framework Programmes (FPs), which support pre-competitive R&D projects, creating a pan-European network of actors performing joint R&D.

Therefore, the investigation of the structure and dynamics of R&D networks is of great current interest, both in a scientific and in a policy context, and currently receives much attention in theoretical and empirical research of different scientific disciplines (see Ozman 2009). Here, we can distinguish between empirical research focusing on knowledge transfer in formalized joint research activities, as given by joint R&D projects or joint publications, and empirical studies using networks as measured by different indicators, such as co-patenting or patent citations, to trace knowledge flows or knowledge spillovers between organizations, regions, or countries (see Ejermo and Karlsson 2006).

There are two major approaches taken to analyse R&D networks: a *regional science* or *geography of innovation* perspective and a *social network analysis* perspective. In a regional science or geography of innovation context, the investigation of the geographical dimension of R&D collaborations is the central research objective. This follows from the assumption that geographical space is crucial for the localization of R&D collaborations and knowledge flows. The pioneering empirical study of Jaffe et al. (1993) provides evidence for the localization hypothesis of knowledge diffusion processes, in general confirmed by more recent empirical studies using different indicators and new spatial econometric techniques (see, for instance, Maurseth and Verspagen 2002, Fischer, Scherngell and Jansenberger 2006, Maggioni 2007, Hoekman et al. 2009, Scherngell and Barber 2009 and 2010). In a social network analysis context, the focus shifts to the analysis of network structures and dynamics using the mathematics of graph theory[1], under the assumption that structural relations are often more important for understanding observed behaviors than are attributes of the actors (see, for instance, Zucker and Darby 1998a and 1998b, Singh 2005, Thompson 2006, Vicente et al. 2010). Ter Wal and Boschma (2009) provide an overview of the increasing importance of social network analysis techniques in the fields of regional science and economic geography.

In this study, we combine the two research traditions by taking a social network analysis perspective when identifying substructures of European R&D networks constituted under the FPs, followed by taking a regional science perspective when analyzing the geographical dimension of identified substructures. In this context, previous work of Breschi and Cusmano (2004) and empirical studies by Scherngell and Barber (2009 and 2010) are central starting points for the current study. Breschi and Cusmano (2004) employ a social network perspective to analyze R&D collaborations with the objective of unveiling the texture of the European Research Area (ERA) using data on joint research projects of the fifth EU Framework Programme (FP), while Scherngell and Barber (2009 and 2010) focus on the geography of R&D collaborations across European regions.

However, results of these previous empirical works may differ across relevant substructures or communities of the whole FP network. Stated informally, a community is a subnetwork whose members are more tightly linked to one another than to other members of the network. A variety of approaches have been taken to explore this concept (see Fortunato 2010 for a useful review). Since network edges often indicate relationships of interest, detecting community groups can be used to partition the network vertices into meaningful sets, enabling quantitative investigation of relevant

[1] Graph theory is the study of mathematical structures consisting of a set of *vertices* (i.e. nodes) connected by a set of *edges* (i.e. links). This provides a precise, formal representation of networks.

subnetworks. Properties of the subnetworks may differ from the aggregate properties of the network as a whole, e.g., modules in the World Wide Web are sets of topically related web pages.

The objectives of the current study are: *first*, to detect communities in European R&D networks; *second*, to describe the spatial patterns of the identified communities; and, *third*, to identify determinants of the observed spatial patterns. We use data on joint research projects funded by the European Framework Programmes to capture European R&D networks. The identification of thematically distinct communities in these networks is realized using graph theoretic techniques described by Barber and Clark (2009). Further, we employ spatial analysis techniques to identify and describe spatial patterns of identified FP communities at a regional level. By means of a Poisson spatial interaction model, we estimate the impact of various separation factors on cross-region collaboration activities in a given community. In particular, we focus on how geographical distance impacts cross-region collaboration intensities across different FP communities. The results demonstrate that European R&D networks are not homogeneous, instead showing distinct, relevant substructures characterized by thematically homogeneous and spatially heterogeneous communities.

The research approach applied in this study is significant, both in a scientific as well as in a European policy context. It proposes a new way of looking into R&D network structures in Europe, combining a social network analysis with a geography of innovation perspective. As noted by Autant-Bernard (2007a), the geographical dimension of innovation and knowledge diffusion deserves closer attention by analyzing such phenomena as R&D collaborations. Such analyses are also of crucial interest for European STI policy, in particular for the integration and cohesion objective outlined in the concept of the European Research Area (ERA): improved coherence of the European research landscape and the removal of barriers to knowledge diffusion in a European system of innovation (see CEC 2007). Of course, insight into the status of integration in different thematic areas is a particularly valuable new view on this topic.

Further, the analysis provides important policy implications. By lending crucial insight into real-world topical structures of R&D networks constituted under earlier FPs, the analysis can inform the design of future FPs. Complementarily, a rich picture for regional policy actors is provided at the regional level on leading European regions with respect to cooperative research activities in specific thematic areas.

The paper is organized as follows. Section 2 presents the theoretical background and outlines the main hypotheses for the empirical study. Section 3 describes the data, posed in terms of networks and collaboration matrices. Section 4 describes the identified communities according to their thematic orientation, while Section 5 unveils the spatial distribution of the identified community groups. Section 6 briefly introduces the Poisson spatial interaction perspective to identifying determinants of the observed spatial community patterns, and presents the estimation results. Section 7 concludes with a summary of the main results, some policy implications and a short outlook.

2 Background and Main Hypotheses

R&D Networks inducing knowledge transfer between firms, universities and research organizations are considered to be crucial for successful innovation in the knowledge-based economy in general, and in knowledge-intensive industries in particular. In fact, we face a considerable increase—and we have done so for decades—in the number of inter-organizational R&D collaboration (Hagedoorn and van Kranenburg, 2003). The main reasons for this have been alleged to include the increasing need to access external knowledge – characterized by complementarity and tacitness – and the high degree of strategic flexibility in collaborative agreements (Kogut 1988, Teece 1992). Another reason may be the growing complexity of technology and the existence of converging technologies (see Pavitt 2005). In particular, firms have expanded their knowledge bases into a wider range of technologies (Granstrand 1998), increasing the need for distinct types of knowledge, so firms must learn how to integrate new knowledge into existing products or production processes (Cowan 2004). It may be difficult to develop this knowledge alone or acquire it via the market. The importance of R&D networks for innovation is also stressed by the various systems of innovation concepts that focus on interactions between different actors in a specific region, country or sector (see Lundvall 1992, among others). The main argument is that the sources of innovation are often distributed between firms, universities, suppliers and customers, giving rise to the notion of networks being the locus of innovation. Networks create incentives for interactive organizational learning, leading to faster knowledge diffusion within the innovation system and stimulating the creation of new knowledge or new combinations of existing knowledge.

The EU follows this view in its science and technology policy, mainly reflected in the concept of the European Research Area (ERA), whose aim is to improve coherence of the European research landscape and remove barriers for knowledge diffusion in a European system of innovation (see CEC 2007). The cornerstone of corresponding EU policy instruments is formed by the Framework Programmes (FPs) on Research and Technological Development. By means of this policy initiative, the EU has co-funded thousands of trans-national collaborative R&D projects. The main objectives of the instrument from a European technology policy view are to integrate national and regional research communities and to coordinate national research policies. Empirical studies such as the one of Breschi and Cusmano (2004) provide evidence for the establishment of a pan-European network of firms, universities, public research organizations, consultants and government institutions performing joint research funded by the FPs (see Roediger-Schluga and Barber 2006 for a comprehensive discussion of the EU FPs).

Previous empirical studies usually focused on complete FPs to describe networks of European R&D cooperation as captured by data on joint FP projects. However, empirical results of these studies may differ across relevant, thematically distinct community groups of the whole FP networks, and these differences may be of crucial interest in a European policy context. Stated informally, a community is a portion of the network whose members are more tightly linked to one another than to other members of the network. Precise formulation of the problem presents two main challenges. *First*, the notion of communities is somewhat vague, requiring a definition to be provided for what formally constitutes a community. *Second*, community solutions must also be practically realizable for networks of real-world scientific or policy interest. The interplay between these challenges allows a variety of community definitions and community identification algorithms suited to networks of different sizes (for useful overviews, see Fortunato and Castellano 2008, Fortunato 2010, and Porter et al 2009).

Meaningful communities have been identified in many networks of diverse character, corresponding to specialized research areas in co-authorship networks, topically related pages on the World Wide Web, and functional modules in cellular or genetic networks, amongst many others. Following the pioneering work of Girvan and Newman (2002) and Newman and Girvan (2004), many researchers, particularly in statistical physics, have investigated methods for detecting communities in large networks. Similarly, we hypothesize *first* that the European FP5 network consists of relevant, thematically distinct subnetworks that show distinct thematic and spatial characteristics.

Second, we hypothesize that geographic localization effects of knowledge flows are significantly smaller within identified communities than for the whole FP5 network, since the transfer of tacit knowledge may be easier in thematically relatively homogenous community groups. As mentioned above, the geography of innovation literature argues that knowledge flows among knowledge producing agents may be geographically bounded, since important parts of new knowledge have some degree of tacitness. Though the cost of transmitting codified knowledge may be invariant to distance, presumably the cost of transmitting non-codified knowledge across geographic space rises with geographic distance (see Jaffe et al. 1993, Audretsch and Feldman 1996). Scherngell and Barber (2009) provide evidence for the geographical localization of FP5 networks. In this study, we assume that localization effects decrease for an identified, thematically homogenous community. Due to a more homogeneous thematic focus of a community, the transfer of non-codified knowledge may not be as costly as would be the case for thematically more dispersed actors.

3 Empirical setting and Data

Our core data set to capture collaborative activities in Europe is the EUPRO database, which presently comprises data on funded research projects of the EU FPs (complete for FP1-FP6) and all participating organizations. It contains systematic information on the participating organizations including the full name, the full address, the type of the organization, and, where appropriate and possible, the organizational subentity involved in the project. For a full description of the EUPRO database and its contents, see Roediger-Schluga and Barber (2008)[2].

Constructing FP5 research networks

The study at hand draws on information concerning joint R&D projects funded in FP5[3]. Using the EUPRO database, we construct a graph or network containing the collaborative projects from FP5 and all organizations that are participants in those projects. An organization is linked to a project if and only if the organization is a member of the project. Since an edge never exists between two organizations or two projects, the network is bipartite. The network edges are unweighted; in principle, the edges could be assigned weights to reflect the strength of the participation, but the data needed to assign such network weights is not available.

Previous investigations of the FPs often have made use of one-mode projection networks (Almendral et al. 2007, Barber et al. 2006, Breschi and Cusmano 2004, Roediger-Schluga and Barber 2008), especially for the organizations. While the projection networks can be useful, the construction of the projections intrinsically loses information available in the bipartite networks, which can lead to incorrect community structures (Guimerà et al. 2007). In the present work, we thus focus exclusively on representation of FP5 as a bipartite network.

Detecting communities in European collaboration networks

Community identification in networks is the assignment of the network vertices to a smaller number of clusters. These clusters are hopefully relevant, and thus, drawing on the context of social networks, called communities. Recent community identification methods are based on analyzing the network structure, identifying communities as groups of vertices that are internally strongly connected but only weakly connected to

[2] The version of the EUPRO database used for this study contains information on 61,169 projects funded from FP1 to FP6, yielding 323,638 participations by 60.034 organizations (status: December 2010).

[3] FP5 had a total budget of 13.7 billion EUR and ran from 1998-2002 (CORDIS 1998). See Scherngell and Barber (2009) and CORDIS (1998) for further details on FP5.

the rest of the network. In empirical networks, vertices within communities are often found to be usefully related by content: edges reflect underlying processes relevant to the entities corresponding to vertices, so communities consist of entities with similar properties.

Community identification methods have been developed that are efficient enough to be suitable for large networks containing thousands or millions of vertices and edges. One such method is the label propagation algorithm (LPA) of Raghavan et al (2007). Each vertex is assigned a label; a community is the set of all vertices with a particular label. The vertices are initialized with distinct labels, thus beginning with all vertices in distinct communities. Vertices are repeatedly updated, replacing their labels with ones that better match the labels of their neighbors. Within tightly interlinked subnetworks, common labels reinforce one another, encouraging uniform labels to be adopted. In contrast, weak linking between tightly interlinked subnetworks means that relatively few neighbors will differ in labels, hindering the propagation of labels between the subnetworks. These two properties accord with the above idea of community, so the LPA proves to be quite effective in practice (Leung et al 2009).

Two properties of community solutions found by LPA warrant comment. *First*, since each vertex has a single label, the communities are disjoint; no vertex belongs to two communities. *Second*, community solutions are not generally unique; more than one label may be satisfactory for a vertex. Both of these properties suggest that some portion of the vertices may fit well in more than one community, so some care should be taken in interpreting specific community memberships. In this work, we consider statistical properties of the communities, which are more robust against reassignment of a few labels.

In this work, we make use of modest extensions to the LPA (Barber and Clark 2009). The specifics of the algorithms are detailed in Appendix C. Since we investigate bipartite networks, the communities will include vertices from the two parts of the network, i.e. communities will contain both projects and organizations.

Observing spatial collaboration patterns of communities across European Regions

To analyze the spatial patterns of the identified communities we first geocode each organization to a specific European region. We use a concordance scheme provided by Eurostat between postal codes and NUTS regions to trace the specific NUTS-2 region of an organization. The European coverage is achieved by using 255 NUTS-2 regions (NUTS revision 2003) drawn from the 25 pre-2007 EU member-states, Norway and Switzerland. The detailed list of regions is given in Appendix A[4]. Next we construct a region-by-region collaboration matrix $\boldsymbol{P}^{(c)}$ for each community *c*, aggregating collaborative activities at the organizational level to the regional level, giving the observed number of R&D collaborations $p_{ij}^{(c)}$ between two regions *i* and *j* (*i*, *j*, = 1, …, *n*) for each community *c*.

Following Scherngell and Barber (2009), we use a full counting method. For a project with three participating organizations in three different regions – say regions *a*, *b*, and *c* – we count three links: from region *a* to region *b*, from *b* to *c* and from *a* to *c*. When all three participants are located in one region we count three intraregional links. We exclude self loops to eliminate spurious self collaborations. The resulting regional collaboration matrix $\boldsymbol{P}^{(c)}$ then contains the collaboration intensities $p_{ij}^{(c)}$ between all (*i*, *j*)-region pairs for community *c*. The *n*-by-*n* matrix for each community is symmetric by construction ($p_{ij}^{(c)} = p_{ji}^{(c)}$).

4 Community structure in European R&D networks

This section differentiates the identified communities by developing community-specific profiles. Using the label propagation approach described in the previous section, we identified 3482 network communities. The communities vary greatly in size, as measured either by the number of organizations in the community or by the number

[4] We follow previous similar empirical work and rely on a NUTS2 disaggregation of the European territory (see Fischer et al. 2006, LeSage et al. 2007, Scherngell and Barber 2009 and 2010). The NUTS2 level provides the basis for the provision of structural funds by the EU, as well as for the evaluation of regional growth processes across Europe (see Fischer et al. 2009)

of projects in the community. Most (2878) 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 will consider eight of these nine largest communities (Barber, Fischer and Scherngell 2010); the ninth is of different character than the others, focusing on international cooperation rather than R&D. We do not consider the remaining smaller communities; while we thus exclude many communities, we are able to account for the majority of R&D cooperations in greater detail.

Figure 1: Community groups in the network of FP5 R&D cooperation

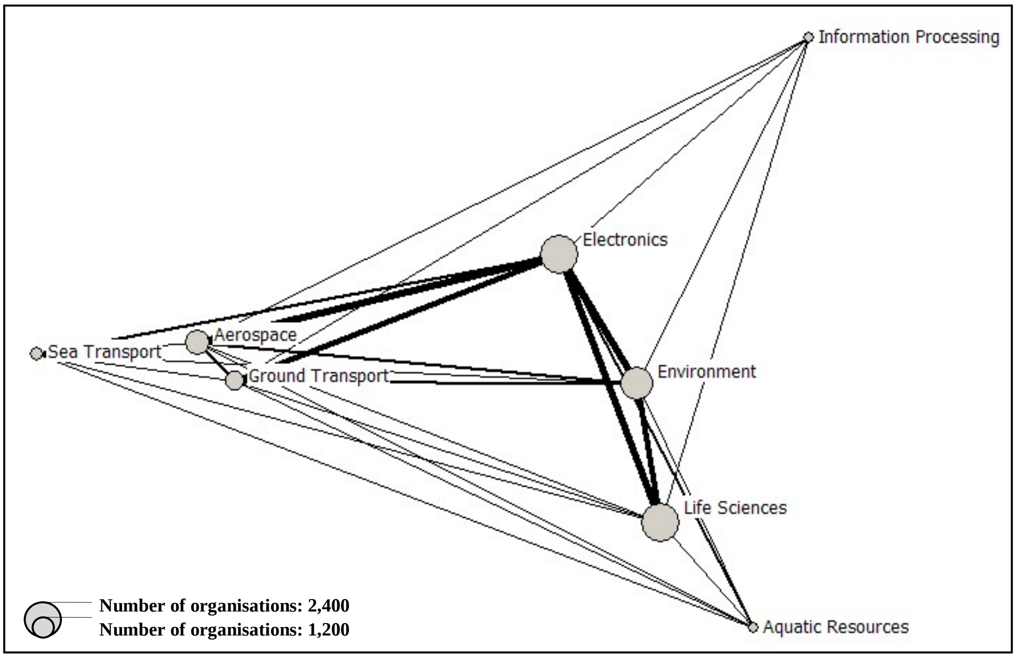


Figure 1 visualizes the network of key FP5 communities. We manually assign names to the communities based on consideration of their constituent projects and organizations (see below). We determine the position for the communities using methods from spectral graph analysis, so that communities that are strongly interconnected are positioned nearer to each other (for a practical overview see Higham and Kibble 2004). The node size corresponds to the number of organizations of the respective community, with the widths of the connection links corresponding to the number of inter-community project participations. In addition, Table 1 provides some summary statistics on the identified communities.

The *Life Sciences* and the *Electronics* communities have the greatest number of organizations. Due to the strong inter-community links, the *Electronics* community appears to have the highest collaboration intensity with other communities, i.e. competences relevant to this field are used intensively in other fields. The *Life Sciences* community shows a strong connection to the third largest community, *Environment*. The three transport-related communities are positioned near one another, i.e. they show relatively high inter-community collaboration intensity. The largest of these is *Aerospace*, and shows a stronger interaction with *Ground Transport* than with *Sea Transport*. The community *Aquatic Resources* has the strongest connection to *Environment*, while *Information Processing* shows comparably low collaboration intensities to all other communities.

The largest community (2,366 organizations), *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 (2,307 organizations), *Electronics*, focuses principally on information technology and electronics, with projects in related fields dealing with materials science, often related to integrated circuits; projects on algorithms, data mining, and mathematics; and a definite subset of projects concerning atomic, molecular, nuclear, and solid state physics. The third largest community (1,855 organizations), *Environment*, is focused on environment 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 (1,146 organizations), *Aerospace*, is focused on aerospace, aeronautics and related topics, including materials science, manufacturing, fluid mechanics, and various energy topics. The next (686 organizations), *Ground Transport*, is focused on land transport, with the projects dominated by railroad and, especially, automotive topics; notable subtopics include manufacturing, fuel systems, concrete, and pollution. The smallest transportation community (218 organizations), *Sea Transport*, focuses specifically on sea transport; virtually all project titles are shipping-related. The remaining communities, *Aquatic Resources* and *Information Processing*, are the smallest and most uniform thematically. Their thematic contents are fisheries and statistics.

Table 1: Summary statistics on FP5 communities

	Aerospace	Aquatic Resources	Electronics	Environ -ment	Ground Transport	Information Processing	Life Sciences	Sea Transport
Number of organizations	1,146	81	2,307	1,855	686	40	2,366	218
Number of participation	13,870	451	30,456	23,155	5,251	226	33,178	2,978
Number of projects	576	69	1447	971	374	20	1468	73
Average number of partners	24.206	11.136	26.403	24.965	15.309	11.300	28.046	27.321
Skewness of number of partners	4.263	1.169	5.132	4.512	6.739	1.097	4.749	1.718

5 Spatial Structure of communities in European R&D networks

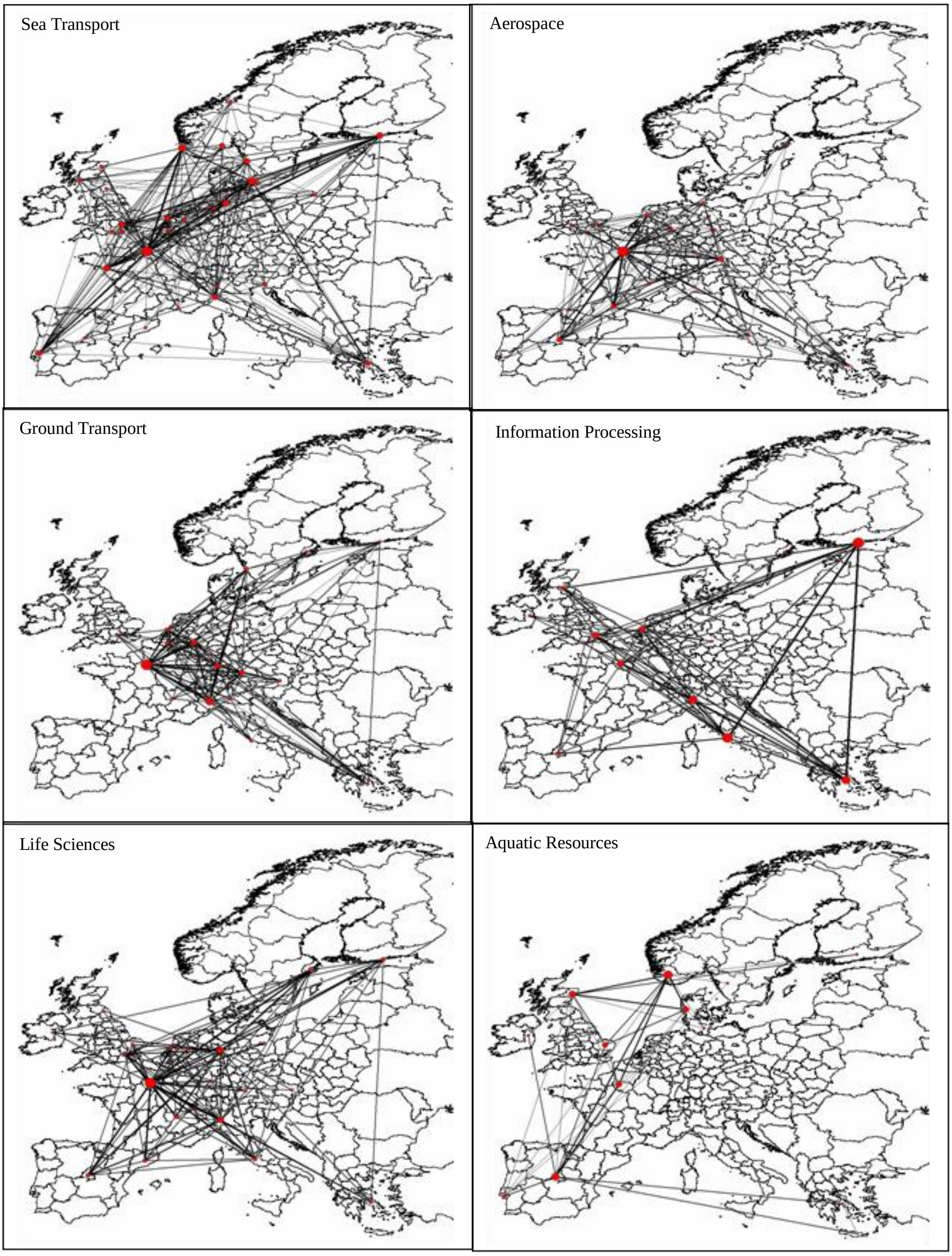
We next consider the spatial distribution of the eight FP5 communities. In Figure 2, we illustrate the spatial networks of the communities by aggregating individual observations on the organizations of a community to the regional level. Note that the region-by-region networks are undirected graphs from a network analysis perspective. The nodes represent regions; their size is relative to the number of organizations in the region that belong to the community.

The spatial network maps in Figure 2 reveal considerable differences among the collaboration patterns of the eight FP5 communities. One immediate result is that the region Île-de-France takes an important position in all communities. Furthermore, the visualization clearly reveals 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 the three transport related communities, the directions of the highest collaboration flows from Île-de-France differ markedly. For the

Sea Transport community we observe intensive collaborations to important sea ports in the north (Zuid Holland, Agder og Rogaland, Danmark, Hamburg) and the south (Liguria, Lisboa, Attiki), while, for the *Ground Transport* community, collaborations to 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. 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 Danmark and Etelä-Suomi (131 collaborations). In the community *Aquatic Resources* the regions Danmark and Agder og Rogaland (Norway) show the highest collaboration intensity, not only between them (21 collaborations) but also to other regions, while for the community *Information Processing* we identify Etelä-Suomi as the central region, featuring intensive collaboration with Attiki, Lazio and Lombardia.

To complement the maps shown in Figure 2, the numbers of project participations by organizations in each region for each community are also of interest; we tabulate the most active participants in Appendix B. This provides insight into which regions are most active for each community, in contrast to which regions are best connected, as described above. Interestingly, well connected regions may markedly differ from the most active regions.

Figure 2: Spatial patterns of eight FP5 communities



Sea Transport

Aerospace

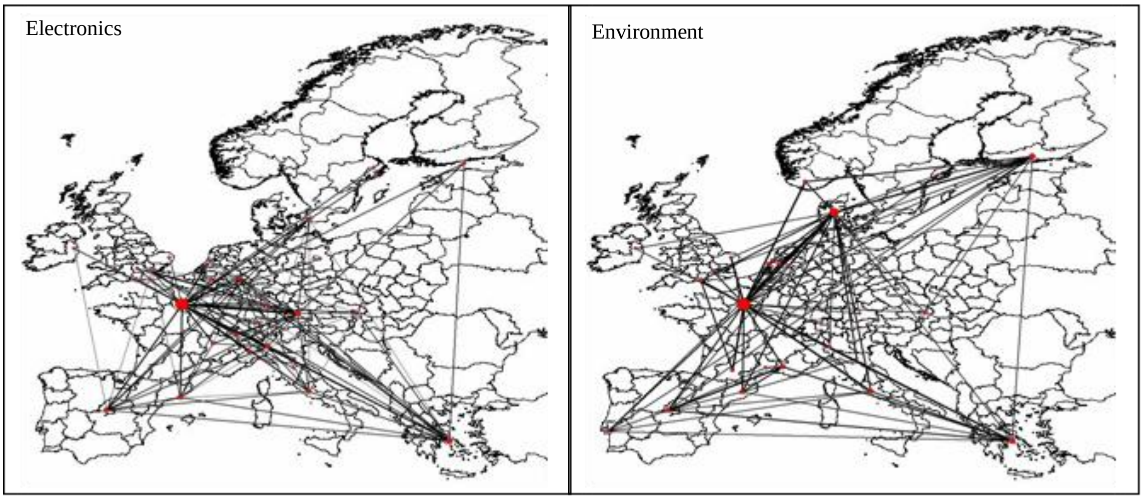
Ground Transport

Information Processing

Life Sciences

Aquatic Resources

Fig. 2 ctd.



Electronics

Environment

6 Identifying determinants of spatial community patterns

Our objective in this paper is not only to detect communities in European FP networks and describe their spatial configurations, but also to investigate determinants that influence the spatial community patterns. In particular, whether the influence of geographical distance differs across communities is of crucial importance in the context of an aspired European Research Area. Thus, we measure separation effects on the constitution of cross-region R&D collaborations in all detected communities. The spatial interaction model of the type used by Scherngell and Barber (2009 and 2010) in a similar context serves again as an appropriate basis. Spatial interaction models incorporate a function characterizing the origin *i* of interaction, a function characterizing the destination *j* of interaction and a function characterizing the separation between two regions *i* and *j*. The model is characterized by a formal distinction implicit in the definitions of origin and destination functions on the one hand, and separation functions on the other (see, for example, Sen and Smith 1995). Origin and destination functions are described using weighted origin and destination variables, respectively, while the separation functions are postulated to be explicit functions of numerical separation variables. The general model in our case is given by

$$P_{ij}^{(c)} = A_i \; B_j \; S_{ij} \qquad i, j = 1,\ldots, n \qquad (1)$$

with

$$A_i = A(a_i, \alpha_1) = a_i^{\alpha_1} \qquad i, j = 1,\ldots, n \qquad (2)$$

$$B_j = B(b_j, \alpha_2) = b_j^{\alpha_2} \qquad i, j = 1,\ldots, n \qquad (3)$$

$$S_{ij} = \exp\left[\sum_{k=1}^{K} \beta_k \; d_{ij}^{(k)}\right] \qquad i, j = 1,\ldots, n \qquad (4)$$

where $P_{ij}^{(c)}$ denotes a stochastic dependent variable that is realized by the number of observed collaboration flows $p_{ij}^{(c)}$ between region i and region j for each community c[5]. A_i denotes the origin function, B_j denotes the destination function, while S_{ij} represents a separation function. The a_i and b_j are measured in terms of the number of organizations participating in EU FP5 projects in the regions i and j, while α_1 and α_2 are scalar parameters to be estimated. Note that due to the symmetry of the origin and destination variables, we have a special case with α_1=α_2, i.e. numerical results for α_1 and α_2 should be equal up to numerical precision. The $d_{ij}^{(k)}$ are K separation measures, the β_k are corresponding parameters to be estimated that will show the relative strengths of the separation measures. We rely on separation measures used in similar studies (see, for instance, Fischer, Scherngell, and Jansenberger 2006; Scherngell and Barber 2009). We can group these separation variables into three categories:

(i) Variables accounting for *spatial effects*: $d_{ij}^{(1)}$ denotes geographical distance between two regions i and j as measured by the great circle distance between the economic centers of the regions, while $d_{ij}^{(2)}$ is a dummy variable that controls for neighboring region effects. We set $d_{ij}^{(2)}$ to one if two organizations are located in neighboring regions and zero otherwise, where neighboring regions are defined to share a common border.

(ii) Variables accounting for *institutional and cultural effects*: $d_{ij}^{(3)}$ is a country border dummy variable that takes a value of zero if two regions i and j are located in the

[5] Note that we do not exclude zero-flows or intraregional flows.

same country and one otherwise, while $d_{ij}^{(4)}$ is a language area dummy variable that takes a value of zero if two regions *i* and *j* are located in the same language area and one otherwise.

(iii) Variables accounting for *technological effects*: $d_{ij}^{(5)}$ measures technological distance by using regional patent data from the European Patent office (EPO). The variable is constructed (see Scherngell and Barber 2009) as a vector ***t***(*i*) that measures region *i*'s share of patenting in each of the technological subclasses of the International Patent Classification (IPC). Technological subclasses correspond to the third-digit level of the IPC systems. We use the Pearson correlation coefficient between the technological vectors of two regions *i* and *j* to define how close they are to each other in technological space. Though we focus on spatial, cultural and institutional effects in this study, we include technological distance, mainly as a control variable to allow for the possibility that geographical distance may just be a proxy for technological distance.

At this point, we are interested in estimating the parameters $\alpha_1 = \alpha_2$ and β_k for each community *c*. OLS estimation procedures are not appropriate for modeling research collaborations, due to their true integer nature and due to the assumption of non-normal errors. This suggests a Negative Binomial density distribution, i.e. a Poisson specification with heterogeneity, allowing for the overdispersion often observed for real world count data (see Cameron and Trivedi 1998). The Negative Binomial density distribution in our case is given by

$$f(P_{ij}^{(c)}) = \frac{\Gamma(p_{ij}^{(c)} + \delta^{-1})}{\Gamma(p_{ij}^{(c)} + 1)\Gamma(\delta^{-1})} \left(\frac{\delta^{-1}}{A_i\ B_j\ S_{ij} + \delta^{-1}} \right)^{\delta^{-1}} \left(\frac{A_i\ B_j\ S_{ij}}{A_i\ B_j\ S_{ij} + \delta^{-1}} \right)^{p_{ij}^{(c)}} \qquad (5)$$

where $\Gamma(\cdot)$ denotes the gamma function and δ is the dispersion parameter. Model estimation is done by Maximum Likelihood procedures (see Long and Freese 2001).

Table 2 presents the sample estimates of the spatial interaction models, with standard errors given in brackets. We use the Negative Binomial model specification as given by

Equation (5). The dispersion parameter δ is significant for all model versions, indicating that the Negative Binomial version is the right specification, i.e. the standard Poisson specification would be biased due to unobserved heterogeneity between the region pairs (see Scherngell and Barber 2009). The existence of unobserved heterogeneity that cannot be captured by the covariates leads to overdispersion and, thus, to biased model parameters for the standard Poisson model.

Table 2: Estimation Results of the Negative Binomial Spatial Interaction Models
[65,025 observations, asymptotic standard errors given in brackets]

	Negative Binomial spatial interaction models								
	Total FP5	**Life Sciences**	**Aquatic Resources**	**Electronics**	**Environment**	**Sea Transport**	**Ground Transport**	**Aerospace**	**Information Processing**
$\alpha_1 = \alpha_2$	0.706*** (0.003)	0.865*** (0.005)	0.777*** (0.024)	0.794*** (0.005)	0.659*** (0.005)	0.771*** (0.004)	1.055*** (0.010)	0.808*** (0.006)	1.202*** (0.008)
Geo [ß$_1$]	-0.278*** (0.008)	-0.110*** (0.011)	-0.072 (0.051)	-0.038** (0.012)	-0.036** (0.012)	-0.020 (0.038)	-0.224*** (0.020)	-0.103*** (0.017)	-0.017 (0.016)
Neig [ß$_2$]	0.184*** (0.036)	0.043 (0.051)	0.312 (0.248)	0.051 (0.052)	0.274*** (0.057)	0.201 (0.019)	0.033 (0.048)	0.253*** (0.013)	0.186 (0.283)
Count [ß$_3$]	-0.008 (0.023)	0.009 (0.0006)	-0.588*** (0.168)	-0.148*** (0.039)	0.119 (0.099)	-0.558*** (0.143)	-0.121 (0.081)	-0.342*** (0.055)	-0.141 (0.216)
Lang [ß$_4$]	-0.098*** (0.024)	-0.004 (0.034)	-1.118*** (0.152)	-0.002 (0.035)	-0.123*** (0.038)	0.326 (0.271)	-0.088 (0.057)	-0.023 (0.019)	-0.798*** (0.113)
Tech [ß$_5$]	-1.413*** (0.115)	-1.437*** (0.167)	-6.312*** (0.657)	-1.577*** (0.171)	-2.421*** (0.181)	-3.797*** (0.544)	-0.606** (0.283)	-2.511*** (0.257)	-1.533* (0.763)
Cons.	-2.539*** (0.128)	-7.042*** (0.175)	-8.557*** (0.654)	-6.197*** (0.179)	-4.148*** (0.185)	-6.175*** (0.545)	-10.124*** (0.297)	-5.303*** (0.260)	-16.851*** (0.855)
(δ)	1.047*** (0.009)	0.969*** (0.016)	2.835*** (0.743)	0.341*** (0.018)	0.530*** (0.013)	2.943*** (0.025)	2.044*** (0.044)	0.982*** (0.015)	6.115*** (0.526)
LL	-135,234.21	-65,657.63	-8,537.42	-75,200.45	-73,257.76	-18,829.43	-31,445.52	-54,124.21	-3,723.43
SS	6.523	5.212	4.979	5.001	4.213	5.176	6.712	5.732	5.174
M-R^2	0.173	0.224	0.128	0.196	0.155	0.133	0.251	0.171	0.249
BIC	-35,455.09	-37,935.68	-24,44.31	-36,647.48	-26,853.30	-4,169.42	-21,006.38	-22,283.32	-2,424.85

Notes: The dependent variable is the cross-region collaboration intensity between two regions *i* and *j* in a given community. The independent variables are defined as given in the text. LL denotes the log-likelihood, SS sum of squares, M-R^2 McFadden´s R-sqaured, BIC Bayesian Information Criterion. ***significant at the 0.001 significance level, **significant at the 0.01 significance level, *significant at the 0.05 significance level

The models produce quite interesting results in the context of the literature on European R&D networks on the one hand, and in the context of the literature on the geographic localization of knowledge flows on the other hand. The second column contains, for the purpose of comparison, the sample estimates for total FP5. The main conclusion of this model is that geographical distance between two organizations has a significant negative effect on the likelihood that they collaborate. However, technological distance between regions shows a larger negative effect on cross-region collaborative activities.

The impact of the different separation effects varies considerably across observed FP5 communities, both with respect to the magnitude of the estimates and to statistical significance. The most important result is that the negative effect of geographical distance is significantly weaker in any given FP5 community than for all FP5 collaborations taken as a whole. This indicates that geographical integration in European research is better developed in thematically more homogenous communities than between communities. In the *Aquatic Resources* community, the *Sea Transport* community and the *Information Processing* community, the effect of geographical distance is even insignificant, i.e. within these communities, there is no observable effect of geographical distance on the probability of collaboration between two organizations in Europe. The highest negative effect of geographical distance within a community is identified for the *Ground Transport* community (β_1 = -0.224).

While geographical distance effects are generally lower for the communities than for all FP5 collaborations, the neighboring region effects are even more variable. Neighboring regions effects cannot be identified for most communities, with the exception of the *Environment* community and the *Aerospace* community, which are subject to stronger neighboring region effects than the average of all FP5 collaborations, i.e. there is considerable significant spatial clustering of research collaborations in these communities at the regional level. Also institutional and cultural effects vary considerably across communities. The modeling results point to the existence of institutional barriers at the national level for collaboration in the *Aquatic Resources* community, the *Electronics* community, the *Sea Transport* community, and the *Aerospace* community, even though FP5 as a whole shows no such barriers. Language area effects are generally lower or insignificant, but the *Aquatic Resources* community and the *Information Processing* community are characterized by quite high negative language area effects, i.e. collaboration probability significantly decreases between organizations located in different language areas.

Concerning technological distance, we find that, in each community, the negative effect of technological distance is higher than for the whole FP network, except for *Ground Transport*; the collaboration probability with ‘technologically distant’ regions in a thematically homogenous community is lower than the average collaboration probability in FP5. For the outlier *Ground Transport,* one may speculate that the thematic area uses rather mature and/or widely used technologies prevalent in all regions, leading to a lower negative effect of technological distance. Additional background information on the composition and configuration of the communities would be needed for further interpretations of the sample estimates, which could be realized in only a descriptive way in the current study. Most importantly, the results demonstrate that separation effects for collaboration depend on the FP communities; this may provide a starting point for further research, in particular concerning the interpretation of the parameter estimates.

7 Concluding remarks

Using data on joint research projects funded by FP5, the objective of this study has been to detect and describe spatial patterns of communities in the European network of R&D cooperation and to identify determinants of the observed spatial community patterns. We have used techniques described by Barber and Clark (2009) to identify network communities, subnetworks whose members are more tightly linked to one another than to other members of the network. The determinants of the spatial patterns in eight of the largest identified communities are examined by means of Negative Binomial spatial interaction models, estimating how various separation factors—such as geographical distance—affect the variation of cross-region collaboration activities in a given community. The current study is to our knowledge the first one that combines community detection and spatial interaction modeling, applying this combination to study European R&D networks.

The study produced interesting results, both from a scientific point of view and in a European policy context. *First,* we detected relevant, thematically relatively homogenous FP5 communities, providing a new view on the R&D collaboration landscape in Europe. The largest communities identified are *Life Sciences*, *Electronics*, and *Environment*; these may contain further substructures of equal relevance. As communities become smaller, they also become more focused. We identified three transport-related communities: *Aerospace*, *Ground Transport*, and *Sea Transport*. The remaining communities, *Aquatic Resources* and *Information Processing*, are the smallest and most uniform thematically of those we have considered. *Second*, the spatial analysis of the large communities clearly reveals that the spatial configuration varies across communities. However, the region of Île-de-France plays a central role in each of the large communities. *Third*, the estimation results of the spatial interaction model show that the spatial integration of collaboration activities within the analyzed communities is more developed than for FP5 collaborations as a whole. The negative impact of geographical distance on the probability that two organizations collaborate is much lower when these organizations belong to the same community, while the negative impact of technological differences is generally more pronounced.

From a policy perspective, the identification and characterization of the spatial patterns of these thematically relevant substructures is of crucial interest. *First*, our analysis may serve as a starting point for analyzing the empirical thematic landscape of European R&D collaboration, which is of strategic interest for the design of future European policy programmes supporting collaborative R&D, in particular concerning the orientation of thematic foci. *Second*, the simple but essential spatial characterization of the large communities may serve as an important source of information for regional and national policy makers to identity their main peers for benchmarking exercises or stimulation of specific collaborations; this is tabulated in Appendix B. *Third*, in the context of the European policy goal of an integrated and coherent research area, the results indicate that the degree and evolution of integration may differ across technological areas and that specific technological characteristics should be considered when assessing progress towards that goal.

The study suggests several directions for future research. *First*, the interpretation of the spatial configuration of the largest identified communities was confined to the descriptive level, as in-depth interpretations of the different separation effects would require further background information about the actors involved in a specific community. Further work could focus on interpretation of separation effects, building on the results presented here. *Second*, the (spatial) evolution of the detected communities over time could be investigated, providing a deeper understanding on the dynamics of community formation and their spatial integration in the European R&D collaboration landscape. *Third*, while we have focused on large communities that cover the majority of the projects, there are thousands of smaller communities that we have not considered. Thus, strategies for analyzing these smaller communities could be explored, as could policy implications such as how to encourage integration of the small communities into the larger ones. Finally, alternative community identification methods could be used, for example to consider overlapping or hierarchical communities, accounting for the subthemes recognized in the larger communities.

Acknowledgements. This work has been funded by the FWF Austrian Science Fund (Project Nr. P21450) and the Portuguese FCT and POCI 2010 (project MAT/58321/2004) with participation of FEDER. We thank two anonymous referees for their critical comments and valuable suggestions.

References

Almendral, J. A., Oliveira, J. G., López, L., Sanjuán, M. A. F., and Mendes, J. F. F. (2007): The interplay of universities and industry through the FP5 network. *New Journal of Physics*, 9(6):183–98.

Audretsch, D.B. and Feldman, M.P. (1996): R&D spillovers and the geography of innovation and production, The American Economic Review 86, 630-640

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

Autant-Bernard, C., Billand, P., Frachisse, D. and Massard, N. (2007b): Social distance versus spatial distance in R&D cooperation: Empirical evidence from European collaboration choices in micro and nanotechnologies, *Papers in Regional Science* 86, 495-519

Barber, M. J. (2007): Modularity and community detection in bipartite networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 76(6):066102.

Barber, M. J. and Clark, J. W. (2009): Detecting network communities by propagating labels under constraints. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 80(2):026129.

Barber, M., Fischer, M.M., Scherngell. T. (2010): The Community Structure of R&D Cooperation in Europe. Evidence from a social networks perspective. Paper to be presented at the 50th European Congress of the Regional Science Association, 19-23 August 2010, Jönköping, Sweden

Barber, M. J., Krueger, A., Krueger, T., and Roediger-Schluga, T. (2006): Network of European Union-funded collaborative research and development projects. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 73(3):036132.

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

Breschi, S. and Lissoni, F. (2001): Knowledge spillovers and local innovation systems: A critical survey, *Industrial and Corporate Change* 10, 975-1005

Breschi S, Lissoni F (2009) Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. *Journal of Economic Geography* 9(4): 439-468

Cameron, A.C. and Trivedi, P.K. (1998): *Regression Analysis of Count Data*. Cambridge [UK] and New York [US], Cambridge University Press

CEC (Commission of the European Communities) (2007) Green paper 'The European Research Area: New Perspectives'. {SEC(2007) 412}, COM(2007)161 final, Brussels, 4 April 2007

Constantelou, A., Tsakanikas, A. and Caloghirou, Y. (2004): Inter-country technological linkages in European Framework Programmes: A spur to European integration? *International Journal of Technology Management* 27, 773 - 790

CORDIS (1998): Fifth Framework Programme (1998-2002). Available from http://cordis.europa.eu/fp5/about.htm

Cowan, R. (2004): Network models of innovation and knowledge diffusion, Unu-MERIT working paper series, 2004-016

Ejermo, O. and C. Karlsson (2006). Interregional Inventor Networks as Studied by Patent Coinventorships. *Research Policy* 35(3): 412-430.

Fischer, M.M (2001): Innovation, knowledge creation and systems of innovation, *The Annals of Regional Science* 35, 199-216

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

Fischer, M.M., Scherngell, T. and Reismann, M. (2009): Knowledge spillovers and total factor productivity: Evidence Using a Spatial Panel Data Model, *Geographical Analysis* 41, 204-220

Fortunato, S. (2010): Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

Fortunato, S. and Castellano, C. (2008): Community structure in graphs. In *Encyclopedia of Complexity and System Science*. Springer.

Gallie EP (2009) Is physical proximity necessary for knowledge spillovers within a cooperative technological network? The case of the French biotechnology sector. *Regional Studies* 43(1): 33-42

Girvan, M. and Newman, M. E. J. (2002): Community structure in social and biological networks. *PNAS*, 99(12):7821–7826.

Granstrand, O. (1998): Towards a theory of the technology-based Firm, *Research Policy* 27(5), 465-489

Guimerà, R., Sales-Pardo, M., and Amaral, L. A. N. (2007): Module identification in bipartite and directed networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 76(3):036102.

Hagedoorn, J. and van Kranenburg, H. (2003): Growth patterns in R&D partnerships: An exploratory statistical study, *International Journal of Industrial Organization* 21, 517–531

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

Hoekman, J., Frenken, K., Tijssen, R.J.W. (2010) Research collaboration at a distance : changing spatial patterns of scientific collaboration in Europe. *Research Policy* 39(5), 662-673.

Hoekman, J., Frenken, K., and van Oort, F. (2009): The geography of collaborative knowledge production in Europe, The Annals of Regional Science 43, 721-738

Jaffe, A.B., Trajtenberg, M. and Henderson, R. (1993): Geographic localization of knowledge spillovers as evidenced by patent citations, *Quarterly Journal of Economics* 108(3), 577-598

Katz, J.S. (1994): Geographical proximity and scientific collaboration, Scientometrics 31, 31-43

Kogut, B. (1988) Joint ventures: theoretical and empirical perspectives, *Strategic Management Journal*, 9, pp. 319–332.

LeSage, J., Fischer, M.M and Scherngell, T. (2007): Knowledge spillovers across Europe. Evidence from a Poisson spatial interaction model with spatial effects, *Papers in Regional Science* 86, 393–421

Leung, I. X. Y., Hui, P., Liò, P., and Crowcroft, J. (2009): Towards real-time community detection in large networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 79(6):066107.

Long, J.S. and Freese, J. (2001): *Regression Models for Categorical Dependent Variables Using Stata*. College Station [Texas], Stata Corporation

Lundvall BA (1992) *National innovation sy*stems. Pinter, London

Maggioni, M.A. and Uberti, T.E. (2007): Inter-regional knowledge flows in Europe: An econometric analysis. In Frenken K (ed.): *Applied Evolutionary Economics and Economic Geography*, Cheltenham, Edward Elgar, pp.230-255

Maggioni MA, Uberti TE (2009) Knowledge networks across Europe: which distance matters? *The Annals of Regional Science* 43(3): 691-720

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

Maurseth, P. B. and Verspagen, B. (2002): Knowledge spillovers in Europe: A patent citation analysis, *Scandinavian Journal of Economics* 104, 531-545

Newman, M. E. J. and Girvan, M. (2004): Finding and evaluating community structure in networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 69(2):026113.

OECD (1992): *Technology and the Economy: The Key Relationships*. Paris, OECD

Ozman, M. (2009),'Inter-firm Networks and Innovation: A Survey of Literature', *Economics of Innovation and New Technology* 18, 39-67

Pavitt, K. (2005) Innovation processes, in: J. Fagerberg, D. C. Mowery & R. R. Nelson (Eds), *The Oxford Handbook of Innovation*, pp. 86–114 (Oxford: Oxford University Press).

Ponds R., Van Oort, F.G., Frenken, K. (2007): The geographical and institutional proximity of scientific collaboration networks, *Papers in Regional Science* 86(3), 423-443

Ponds, R., Van Oort, F.G., Frenken, K. (2010) Innovation, spillovers and university-industry collaboration : an extended knowledge production function approach. *Journal of Economic Geography* 10(2), 231-255.

Porter, M. A., Onnela, J.-P., and Mucha, P. J. (2009): Communities in networks. *Notices of the American Mathematical Society*, 56(9):1082–1097+.

Raghavan, U. N., Albert, R., and Kumara, S. (2007): Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 76(3):036106.

Romer, P.M (1990) Endogenous technological change. *Journal of Political Economy,* 98: 71–102.

Roediger-Schluga, T. and Barber, M.J. (2006): The structure of R&D collaboration networks in the European Framework Programmes, Unu-MERIT working paper series, 2006-36

Roediger-Schluga, T. and Barber, M.J. (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, 321–347

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

Scherngell, T. and Barber, M. (2010): Distinct spatial characteristics of industrial and public research collaborations: Evidence from the 5th EU Framework Programme, *The Annals of Regional Science* [forthcoming]

Scherngell, T. and Hu, Y. (2010): Collaborative Knowledge Production in China. Regional Evidence from a Gravity Model Approach, Regional Studies [forthcoming]

Sen, A. and Smith, T.E. (1995): *Gravity Models of Spatial Interaction Behaviour*. Heidelberg, Berlin and New York, Springer

Singh, J. (2005). Collaborative Networks as Determinants of Knowledge Diffusion Patterns. *Management Science* 51(5): 756-770.

Teece, D. (1992) Competition, co-operation, and innovation. Organisational arrangements for regimes of rapid technological progress, *Journal of Economic Behaviour and Organization*, 18, pp. 1–25.

Ter Wal, A. L. J. and R. Boschma (2009). Applying social network analysis in economic geography: framing some key analytic issues. *The Annals of Regional Science* 43(3): 739-756.

Thompson, P. (2006). "Patent Citations and the Geography of Knowledge Spillovers: Evidence from Inventor- and Examiner-added Citations." *Review of Economics and Statistics* 88(2): 383-388.

Vicente, V., Balland, P.A. and Brossard, O. (2010) Getting into Networks and Clusters: Evidence from the Midi-Pyrenean Global Navigation Satellite Systems (GNSS) Collaboration Network, *Regional Studies,* in press, DOI: 10.1080/00343401003713340

Zucker, L. G., M. R. Darby and Armstrong J. (1998). Geographically Localized Knowledge: Spillovers or Markets? *Economic Inquiry* XXXVI: 65-86.

Zucker, L. G., M. R. Darby and Brewer M. (1998). Intellectual Human Capital and the Birth of U.S. Biotechnology Enterprises. *American Economic Review* 88(1): 290-306.

Appendix A

NUTS is an acronym of the French for the “nomenclature of territorial units for statistics", which is a hierarchical system of regions used by the statistical office of the European Community for the production of regional statistics. At the top of the hierarchy are NUTS-0 regions (countries) below which are NUTS-1 regions and then NUTS-2 regions. This study disaggregates Europe's territory into 255 NUTS-2 regions located in the EU-25 member states (except Cyprus and Malta) plus Norway and Switzerland. We exclude the Spanish North African territories of Ceuta y Melilla, the Portuguese non-continental territories Azores and Madeira, and the French Departments d'Outre-Mer Guadeloupe, Martinique, French Guayana and Reunion. Thus, we include the following NUTS 2 regions:

Austria: Burgenland, Kärnten, Niederösterreich, Oberösterreich, Salzburg, Steiermark, Tirol, Vorarlberg, Wien

Belgium: Prov. Antwerpen, Prov. Brabant-Wallon, Prov. Hainaut, Prov. Limburg (B), Prov. Liège, Prov. Luxembourg (B), Prov. Namur, Prov. Oost-Vlaanderen, Prov. Vlaams-Brabant, Prov. West-Vlaanderen, Région de Bruxelles-Capitale / Brussels Hoofdstedelijk Gewest

Czech Republic: Jihovýchod, Jihozápad, Moravskoslezsko, Praha, Severovýchod, Severozápad, Střední Morava, Střední Čechy

Denmark: Danmark

Estonia: Eesti

Finland: Åland, Etelä-Suomi, Itä-Suomi, Länsi-Suomi, Pohjois-Suomi

France: Alsace, Aquitaine, Auvergne, Basse-Normandie, Bourgogne, Bretagne, Centre, Champagne-Ardenne, Corse, Franche-Comté, Haute-Normandie, Île de France, Languedoc-Roussillon, Limousin, Lorraine, Midi-Pyrénées, Nord - Pas-de-Calais, Pays de la Loire, Picardie, Poitou-Charentes, Provence-Alpes-Côte d'Azur, Rhône-Alpes

Germany: Arnsberg, Berlin, Brandenburg, Braunschweig, Bremen, Chemnitz, Darmstadt, Dessau, Detmold, Dresden, Düsseldorf, Freiburg, Gießen, Halle, Hamburg, Hannover, Karlsruhe, Kassel, Koblenz, Köln, Leipzig, Lüneburg, Magdeburg, Mecklenburg-Vorpommern, Mittelfranken, Münster, Niederbayern, Oberbayern, Oberfranken, Oberpfalz, Rheinhessen-Pfalz, Saarland, Schleswig-Holstein, Schwaben, Stuttgart, Thüringen, Trier, Tübingen, Unterfranken, Weser-Ems

Greece: Anatoliki Makedonia, Thraki; Attiki; Ipeiros; Voreio Aigaio; Dytiki Ellada; Dytiki Makedonia; Thessalia; Ionia Nisia; Kentriki Makedonia; Kriti; Notio Aigaio; Peloponnisos; Sterea Ellada

Hungary: Dél-Alföld, Dél-Dunántúl, Észak-Alföld, Észak-Magyarország, Közép-Dunántúl, Közép-Magyarország, Nyugat-Dunántúl

Ireland: Border, Midland and Western; Southern and Eastern

Italy: Abruzzo, Basilicata, Calabria, Campania, Emilia-Romagna, Friuli-Venezia Giulia, Lazio, Liguria, Lombardia, Marche, Molise, Piemonte, Puglia, Sardegna, Sicilia, Toscana, Trentino-Alto Adige, Umbria, Valle d'Aosta/Vallée d'Aoste, Veneto

Latvia: Latvija

Lithuania: Lietuva

Luxembourg: Luxembourg (Grand-Duché)

Netherlands: Drenthe, Flevoland, Friesland, Gelderland, Groningen, Limburg (NL), Noord-Brabant, Noord-Holland, Overijssel, Utrecht, Zeeland, Zuid-Holland

Norway: Agder og Rogaland, Hedmark og Oppland, Nord-Norge, Oslo og Akershus, Sør-Østlandet, Trøndelag, Vestlandet

Poland: Dolnośląskie, Kujawsko-Pomorskie, Lubelskie, Lubuskie, Łódzkie, Mazowieckie, Małopolskie, Opolskie, Podkarpackie, Podlaskie, Pomorskie, Śląskie, Świętokrzyskie, Warmińsko-Mazurskie, Wielkopolskie, Zachodniopomorskie

Portugal: Alentejo, Algarve, Centro (P), Lisboa, Norte

Slovakia: Bratislavský kraj, Stredné Slovensko, Východné Slovensko, Západné Slovensko

Slovenia: Slovenija

Spain: Andalucía, Aragón, Cantabria, Castilla y León, Castilla-La Mancha, Cataluña, Comunidad Foral de Navarra, Comunidad Valenciana, Comunidad de Madrid, Extremadura, Galicia, Illes Balears, La Rioja, País Vasco, Principado de Asturias, Región de Murcia

Sweden: Mellersta Norrland, Norra Mellansverige, Småland med öarna, Stockholm, Sydsverige, Västsverige, Östra Mellansverige, Övre Norrland

Switzerland: Espace Mittelland, Nordwestschweiz, Ostschweiz, Région lémanique, Ticino, Zentralschweiz, Zürich

United Kingdom: Bedfordshire & Hertfordshire; Berkshire, Buckinghamshire & Oxfordshire; Cheshire; Cornwall & Isles of Scilly; Cumbria; Derbyshire & Nottinghamshire; Devon; Dorset & Somerset; East Anglia; East Riding & North Lincolnshire; East Wales; Eastern Scotland; Essex; Gloucestershire, Wiltshire & North Somerset; Greater Manchester; Hampshire & Isle of Wight; Herefordshire, Worcestershire & Warkwickshire; Highlands and Islands; Inner London; Kent; Lancashire; Leicestershire, Rutland and Northamptonshire; Lincolnshire; Merseyside; North Eastern Scotland; North Yorkshire; Northern Ireland; Northumberland and Tyne and Wear; Outer London; Shropshire & Staffordshire; South Western Scotland;

South Yorkshire; Surrey, East & West Sussex; Tees Valley & Durham; West Midlands; West Wales & The Valleys; West Yorkshire

Appendix B

We list here the most active regions for the eight communities considered in depth in this paper. For each community, we give the twenty regions with the highest number of participations in projects from the community. The number of participations is shown parenthetically. Regions are given in descending order of the number of participations.

Aerospace: Île de France (1232), Comunidad de Madrid (691), Oberbayern (581), Danmark (526), Noord-Holland (440), Köln (365), Attiki (320), Inner London (306), Lombardia (285), Greater Manchester (276), Bedfordshire & Hertfordshire (271), Etelä-Suomi (269), Campania (266), Midi-Pyrénées (248), Dytiki Ellada (247), Outer London (243), Lazio (241), Liguria (239), Hampshire & Isle of Wight (225), País Vasco (224)

Aquatic Resources: Agder og Rogaland (97), North Eastern Scotland (93), Danmark (91), Comunidad de Madrid (73), Flevoland (67), Noord-Holland (67), Hamburg (57), Algarve (55), Kriti (49), Attiki (47), Northern Ireland (39), Southern and Eastern (38), East Anglia (31), Andalucía (26), País Vasco (25), Galicia (24), Prov. West-Vlaanderen (22), Etelä-Suomi (21), Eastern Scotland (18), Vestlandet (17)

Electronics: Île de France (3537), Oberbayern (1390), Attiki (1182), Rhône-Alpes (1012), Comunidad de Madrid (863), Köln (831), Lombardia (768), Lazio (728), Zuid-Holland (578), Danmark (563), Berkshire, Buckinghamshire & Oxfordshire (559), Berlin (540), Région lémanique (531), Noord-Brabant (523), Inner London (519), Cataluña (509), Prov. Vlaams-Brabant (483), Southern and Eastern (471), Stuttgart (433), Outer London (430)

Environment: Île de France (1020), Danmark (782), Αττική / Attiki (627), Etelä-Suomi (580), Lazio (565), Zuid-Holland (526), Noord-Holland (479), Comunidad de Madrid (426), East Anglia (414), Lombardia (395), Southern and Eastern (378), Cataluña (373), Stockholm (357), Gelderland (355), Wien (350), Andalucía (326), Utrecht (306), Karlsruhe (305), Agder og Rogaland (295), Hampshire & Isle of Wight (294)

Ground Transport: Île de France (846), Stuttgart (698), Piemonte (587), Köln (385), Zuid-Holland (346), Lombardia (323), Oberbayern (293), Västsverige (290), Etelä-Suomi (226), Berkshire, Buckinghamshire & Oxfordshire (218), Kentriki Makedonia (200), Lazio (177), Hannover (175), País Vasco (168), Comunidad de Madrid

(144), Steiermark (141), Noord-Holland (127), Prov. Vlaams-Brabant (123), Rhône-Alpes (119), Darmstadt (118)

Information Processing: Eastern Scotland (40), Lombardia (21), Etelä-Suomi (20), Lazio (18), Zuid-Holland (16), Hampshire & Isle of Wight (14), Île de France (12), Attiki (11), Outer London (11), Stockholm (10), Sør-Østlandet (10), Danmark (7), Darmstadt (7), Southern and Eastern (7), Noord-Holland (5), Comunidad de Madrid (4), Essex (4), Limburg (NL) (4), Luxembourg (Grand-Duché) (4), Espace Mittelland (3)

Life Sciences: Île de France (1860), Danmark (1055), Gelderland (843), Outer London (703), Lombardia (658), East Anglia (637), Comunidad de Madrid (636), Inner London (605), Cataluña (569), Zuid-Holland (547), Utrecht (538), Lazio (529), Stockholm (521), Karlsruhe (519), Prov. Vlaams-Brabant (495), Rhône-Alpes (494), Southern and Eastern (481), Oberbayern (458), Région de Bruxelles-Capitale / Brussels Hoofdstedelijk Gewest (442), Eastern Scotland (396)

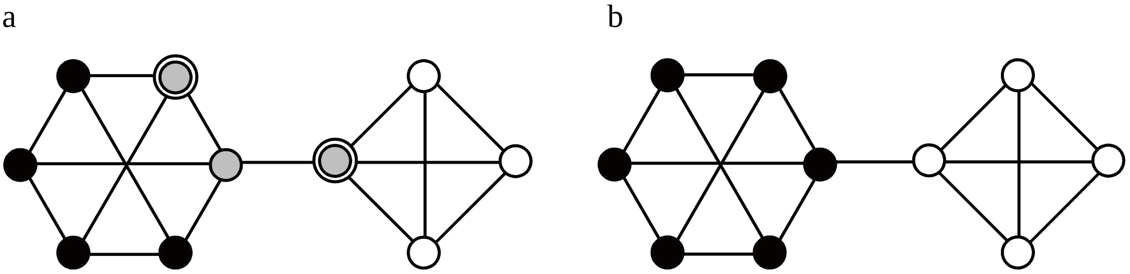
Sea Transport: Danmark (190), Liguria (144), Hamburg (137), Île de France (135), Outer London (115), South Western Scotland (105), Agder og Rogaland (99), Zuid-Holland (88), Attiki (76), Pays de la Loire (61), Bremen (58), Surrey, East & West Sussex (48), Västsverige (43), Comunidad de Madrid (40), Etelä-Suomi (36), Friuli-Venezia Giulia (35), Gelderland (35), Hampshire & Isle of Wight (33), Trøndelag (32), Région de Bruxelles-Capitale / Brussels Hoofdstedelijk Gewest (30)

Appendix C

Raghavan et al (2007) proposed a label propagation algorithm (LPA) for identifying communities in networks. Community membership is tracked by labels assigned to the graph vertices; a community is a set of all vertices with a particular label. Each vertex is assigned a single label, and thus belongs to a single community.

Call a label *satisfactory* for a vertex when no other label occurs more frequently among its neighbors. The core of the LPA is a process of replacing unsatisfactory labels with satisfactory ones, continuing until all vertices have satisfactory labels. This idea is illustrated in Figure C1 using a toy network with visually apparent community structure. In Figure C1a, there are three different labels, shown by the vertex shading. The black and white labels are all satisfactory for their vertices. Of the three gray labels, two are unsatisfactory for their vertices, shown by double borders on the vertices: one neighbors a single gray vertex and two black vertices, the other neighbors a single gray vertex and three white vertices. The third gray label is satisfactory: the vertex neighbors two gray vertices and two black vertices. In Figure C1b, all vertices have satisfactory labels.

Figure C1 Community Identification with Label Propagation



The algorithm begins from a state where all vertices have different labels (and thus are generally all unsatisfactory). Taken in random order, the vertices are considered to see whether their labels are satisfactory and updated to be satisfactory when not; if multiple labels would be satisfactory, one is chosen at random. For the example network shown in Figure C1a, the two vertices with gray labels must then be updated, one to have a black label, the other to have a white label; note that changing these two gray labels will cause the third gray label to become unsatisfactory. Multiple relabeling passes are made through the vertices, with the algorithm halting when all vertices have a satisfactory label, such as in Figure C1b.

The LPA offers a number of desirable qualities. As described above, it is conceptually simple, being readily understood and quickly implemented. The algorithm is efficient in practice. Each relabeling iteration through the vertices 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 needed to assign the final label to most of the vertices (typically over 95% of vertices in 5 iterations, see Raghavan et al. 2007, Leung et al. 2008).

The LPA defines communities procedurally, rather than as optimization of an objective function, and thus provides no intrinsic measure for the quality of communities found. To assess community quality, we can introduce an auxiliary measure, such as the popular modularity measure (Newman and Girvan 2004); in this work, more suitable is a version of modularity specialized to bipartite networks (Barber 2007). Using modularity, communities found using LPA are seen to be of high quality (Raghavan et al. 2007): label propagation is both fast and effective. Indeed, Leung et al. (2008) have proposed extensions to the label propagation algorithm that make it comparable to the best algorithms for community detection in quality and efficient enough to analyze very large networks.

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). LPAr is defined similarly to the original LPA presented above, but with additional randomness to allow the algorithm to avoid premature termination. In practice, this produces better communities as measured by modularity than does LPA. LPAb imposes constraints on the label propagation so that the algorithm identifies a local maximum in the bipartite modularity. The overall hybrid algorithm thus belongs to the recent class of algorithms based on modularity maximization (for a survey, see Fortunato 2010).