The CDM framework

ECONOMICS OF INNOVATION AND NEW TECHNOLOGY, 2017
VOL. 26, NOS. 1–2, 21–41
http://dx.doi.org/10.1080/10438599.2016.1202520

The CDM framework: knowledge recombination from an evolutionary viewpoint

Ad Notten[a], Jacques Mairesse[a,b,c] and Bart Verspagen[a,d]

[a]UNU-MERIT, Maastricht, The Netherlands; [b]CREST (ParisTech-ENSAE), Paris, France; [c]NBER, Cambridge, MA, USA; [d]School of Business and Economics, Maastricht University, Maastricht, The Netherlands

ABSTRACT

Since Crepon, Duguet, and Mairesse published their ground-breaking article ('Research, Innovation and Productivity: An Econometric Analysis at the Firm Level', 1998), the CDM framework has known a large diffusion, despite being published in a non-indexed journal. The present study is an exploration of the spread and recombination of 'knowledge' in the 'CDM universe', comprising all papers in Scopus indexed journals citing CDM or/and CDM cited papers. We assess first the speed and range of diffusion of CDM and investigate next the 'origins' and further 'genealogical' make up of the knowledge recombinations within the CDM universe. We find that CDM is cited by a growing number of papers, which spread over a variety of fields, and that it compares very well with the most cited comparable articles in indexed journals in its domain of research. We further find that the CDM universe is mainly constituted of three large clusters and for each of them we are able to identify knowledge paths going from the CDM and earlier cited papers to the subsequent main citing papers. We intend to provide a detailed interpretation of these findings in future work.

ARTICLE HISTORY
Received 30 August 2015
Accepted 6 June 2016

KEYWORDS
Science of science studies; knowledge diffusion; citation analysis; network analysis; cluster analysis

JEL CLASSIFICATION
B25; D83; D85

1. Conceptual background

1.1. Science of science's studies

'Science of science' studies have a long history. Authors such as De Solla Price (1963), Crane (1972), Merton (1973), and Cole (1983) have investigated different aspects of the process of scientific research and science as a whole, adding to our understanding. More recently Knorr Cetina (1999), Whitley (2000) and Becher and Trowler (2001) have further contributed to this literature. While a large number of these science of science studies have made use of bibliographic data, there has been some discussion on whether analysis of such data can provide a complete satisfactory picture of scientific interactions and knowledge flows (Garfield 2009). In support of the bibliographic approach it was stressed that co-citation data as a whole, including citations from conference proceedings, special issues of journals and other seminal and pertinent works, can give useful insights into the 'hidden life' of scientists and science (Fagerberg and Verspagen 2009). In this paper, we are specifically interested in the process of diffusion of knowledge itself, for which we think a bibliometric analysis is particularly appropriate, and we abstract from the social connections through which this knowledge flows. In this analysis, we are thus considering publications as vectors of knowledge and their citations as indicators of knowledge recombination.

CONTACT Ad Notten notten@merit.unu.edu UNU-MERIT, Boschstraat 24, Maastricht 6211AX, The Netherlands

The science of science literature has also brought the issues surrounding the definition of scientific disciplines to the fore. Changes in the definition of scientific disciplines are evidenced at its most dramatic when revolutionary developments take place and are acknowledged by the wider scientific community, whether they are real-world or purely theoretic. These 'breakthroughs', which are often themselves the culmination of a disciplinary area and at the same time the start of a new 'specialism' or 'field' re-define the boundaries of science and scientific disciplines (Mazloumian et al. 2011; Heinze et al. 2013).

1.2. Invisible colleges

The idea of disciplinary boundaries and the communities that live within have been examined by a number of authors both in the academic sphere (Wagner 2008) and in a more general setting (Lave and Wenger 1991). This idea has also been absorbed by the management literature in a quest for finding an answer to the closely related question of 'distant' search by firms for external knowledge by 'boundary spanning individuals' (Tushman 1977; Cohen and Levinthal 1990; Nonaka and Takeuchi 1995; Rosenkopf and Nerkar 2001). In an academic setting communities that initiate specialisms are not uniformly part of a specific professional association or any other institutional grouping, such as a department or dedicated research group, and can be associated with less organizationally constrained 'social circles' or 'invisible colleges' (Crane 1972; Becher and Trowler 2001).

The notion of invisible colleges, specifically in the context of community detection and disciplinary specialisms, has also been a basis of previous studies which did not turn their attention directly at the knowledge contained in a breakthrough, but at the influence of noted individuals within invisible colleges (see for instance Verspagen and Werker 2004). These studies focus on eminent or productive scientists who, according to Crane's 1972 book and corroborated by later works (Azoulay, Graff Zivin, and Wang 2010), have a reach far beyond the borders of the invisible college of which they are an integral part. As such these star scientists and their work contribute to the re-definition of the boundaries of a scientific field.

In addition to the personal and social linkages, invisible colleges also bound by common inspiration and knowledge linkages, in particular by a common scientific language (David 1998; Cowan, David, and Foray 2000). Although all these diverse linkages can be 'hidden', they do define invisible colleges as a whole contributing to the development of their unique identity and of their own specific literature (Fagerberg and Verspagen 2009; Bhupatiraju et al. 2012).

1.3. Knowledge diffusion and recombination, and present investigation

We start from the concept that knowledge communities are concentrated in invisible colleges. These communities, although not necessarily bound by an organizational boundary, are usually confined within the conceptual or cognitive boundaries of 'like-minded' people, perhaps better called specialisms or fields. These fields may then in turn be the definition of a 'knowledge boundary' which might overlap with the social or organizational boundaries (Polanyi 1962). However, the nature of knowledge is such that it can only grow sustainably and creatively when recombined with other, different, strands of knowledge.[1] We can first think of the recombination of 'local' knowledge taking place between research groups that work closely together, in the same or near-by location or in the same research niche.[2] Distant search for knowledge resulting in non-local knowledge recombination is the other possibility. Here we find researchers reaching out across geographical, institutional or disciplinary boundaries to search for knowledge of interest to their work. These researchers may initially look for knowledge that sticks out of their knowledge landscape. This can for instance be highly cited papers, papers by well-known authors in the field of interest or it might be a scientific breakthrough (either major or micro), which through distant search for knowledge subsequently spreads out across its initial boundaries. Although locally searched knowledge is more convenient to access, the knowledge base is in danger of rapid depletion if it is not replenished with new knowledge. The possibilities for re-combination within a closed knowledge base are limited and there is a necessity of renewal to avert its 'extinction', using Darwinian terminology. This replenishment can only be effected by searching for knowledge outside of its original boundaries.

In this paper, we explore as an example of such process of knowledge evolution the case of Crepon, Duguet, and Mairesse's (1998) article 'Research, Innovation and Productivity: An Econometric Analysis at the Firm Level', 'CDM' for short. CDM cites a number of previous papers, which can also be cited by the papers citing CDM. Similarly, along the years the papers citing CDM later can themselves cite the ones citing it earlier. They thus constitute what we shall call the 'CDM universe' of papers linked by a network of cross-citations. The process of diffusion and recombination of knowledge can be described and analyzed in two dimensions:

(1) A horizontal plane, represented by communities or clusters that are possibly overlapping. These communities can consist of scientists from the same discipline or similar sub-disciplines, or sharing similar affiliations, or linked by co-authorships, etc.
(2) A vertical plane within these communities, represented by linear or near-linear paths of knowledge recombinations, which in our case can be identified by forward directed citations among papers and which possibly branch out like a tree in each community.

These horizontal clusters and vertical directed knowledge paths within clusters provide us with a method to build a 'genealogical' or evolutionary tree of knowledge for a given field or sub-field of scientific literature. This is precisely what we intend to do in the case of the relatively small and specific body of literature closely surrounding CDM.

2. Data description, comparative assessment and disciplinary spread of CDM

In this section, we explain how we proceed to collect the bibliographic data we use in this study. We will also briefly record two preliminary investigations, which we think appropriate to set up the stage. We present more at length the methods and main results of our core analysis in Sections 4 and 5 respectively.

2.1. Data

The departure point for the collection of data for the CDM paper is not only its publication in the journal 'Economics of Innovation and New Technology', but also the NBER and CREST working paper series in which it appeared as well.[3–4] It is important in the context of this study to point out that the Routledge published journal 'Economics of Innovation and New Technology' is not indexed in the Thomson Reuters WoS – SSCI and it is only since 2008 in Elsevier's Scopus. This has repercussions, likely to be significant but impossible to assess precisely, on the generation of citation and reference links within both citation indexes and, as we shall see, an

Box 1. Three datasets:

i. For our comparative study we opt for employing Web of Science. One of the defining parameters for the comparison is to consider articles within the same broad subject area as CDM. The dataset used in Section 2.2 thus select all the articles that have the same topical focus as CDM without citing it necessarily of course. Note that the subject area classification is constructed on the basis of the source's (e.g. journal) topical focus and is not on a classification of the articles themselves.[7]
ii. For exploring the disciplinary spread of CDM we need to classify as well as possible the scientific discipline or subject area in which article are published. Although both Web of Science and Scopus citation indexes allow such classification, Web of Science offers a finer grained version including about 250 subject areas, which is for our study especially relevant as CDM's original specific domain is relatively confined. The dataset we used in Section 2.3 thus select all articles that have CDM as a cited reference.
iii. In order to construct a directed acyclical network which includes as many nodes as possible we found preferable to use Scopus to construct the dataset we use in Section 4. See endnote 6.

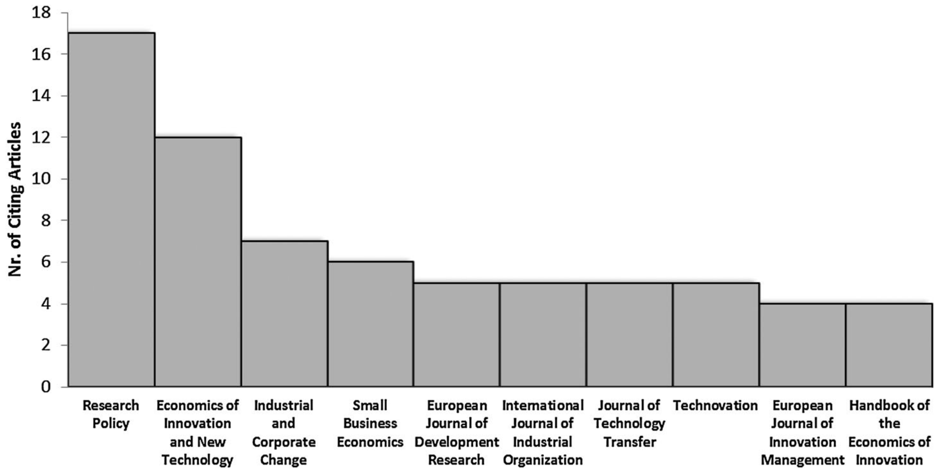


Figure 1. Number of references to CDM in the ten journals with most citations to CDM since its publication in 1998. Source: Scopus, 6 September 2013, authors own calculations.

impact on the early visibility of CDM and ensuing flows of citations.[5] Early on in our study we investigated a number of possible sources of evidence for the diffusion of the CDM model.[6] In practice, however, the quality of the data accessible from the citation indexes of Web of Science and Scopus is such that it leaves little room for competition. We thus relied on Web of Science and Scopus to extract the bibliographic data which we found best for our preliminary assessments and main analysis. We explain in Box 1 the three datasets we thus constructed.

As a first look using Scopus, we examined the distribution of the articles across the journals and other media sources. We show in Figure 1 the top ten citation sources from a total of 101 referring journals and books according to the number of articles citing CDM. We see, which is not surprising, that CDM has been primarily cited by articles published in specialized journals with a strong focus on economics of research and innovation, in particular Research Policy and Economics of Innovation and New Technology.

2.2. Comparative assessment

To have a picture of the academic landscape of CDM and be able to view it in a comparative perspective, we have gathered a pool of 'similar' articles. We have defined these articles by the presence in their abstract, author keywords or titles of the three topical keywords of 'research or R&D, innovation and productivity which make up CDM title. To collect them we have then used them as 'wildcards', as well as a limited number of synonyms and homonyms as 'wildcards' to search for matches in the article's metadata the citation index of Web of Science. In light of an assessment focusing on yearly citation counts, it is also appropriate to search for articles within a limited time interval surrounding the CDM publication year. As such we chose to look at a five year time interval of articles published between 1 January 1996, and 31 December 2000. We have further narrowed the search for articles published in journals classified in the subject areas of Economics and Management, keeping in mind that this is not exactly equivalent to classifying the articles themselves (see endnote 7). This resulted in a dataset of 390 articles indexed in Web of Science, to which we added CDM (which is not indexed in Web of Science).[8] Note that except CDM this dataset is made up of articles from indexed journals only and is not complete in the sense that it does not cover the whole of the peer-reviewed economic and management journals.[9]

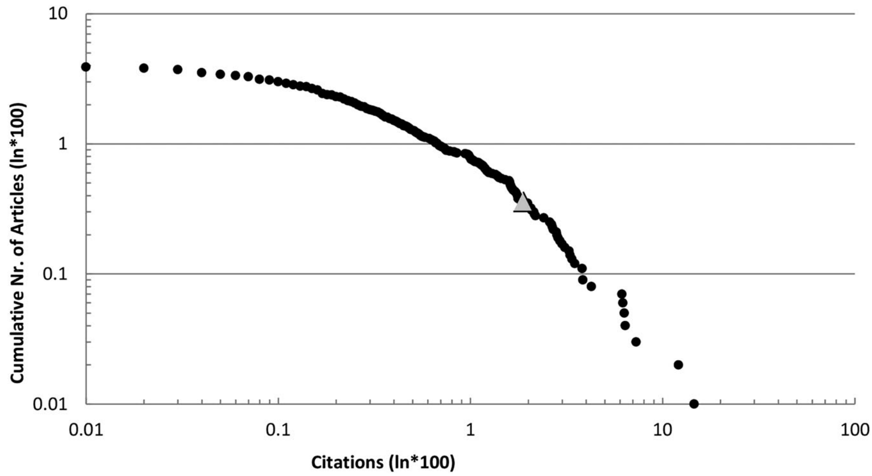


Figure 2. Log-log cumulative distribution of the number of articles ranked by increasing number of total citations received since publication, for the pool of 390 articles 'similar' to CDM, with CDM indicated by the triangle.
Source: Web of Science, authors own calculations. Log cumulative number of articles multiplied by 100 on the *y*-axis and log number of citations per article multiplied by 100 on the *x*-axis (after adding 1 to the number of citations to enable the use of log scales).

We show in Figure 2 the log-log cumulative distribution of the number of articles ranked by increasing number of total citations received over a 15 to 20 year period (from their date of publication between 1996 and 2000 to March 2015). We observe that it is a power-law function satisfying the 'Pareto principle' or 80–20 rule, where 20% of the articles gather an approximate 80% of the citations. We also see that the CDM article falls well within this 20% group of highly cited articles. To focus we chose to compare CDM to the set of 38 articles taken at random among the 97 articles of this 20% highly cited group (excluding CDM), and constructed a Table giving the yearly distribution of accumulated citations they have received.[10] We are interested to see whether CDM can be distinguished from these other 38 articles.

We first discover that CDM is in good company since 11 of the 38 articles are not published in a journal from the first quartile sub-group of journals with highest impact factor, and 3 of these 11 are not even published in journals from the second quartile sub-group of journals. We also see that CDM received its first citation only in 2002, five years after its publication, and is thus a very slow starter in terms of gathering citations when compared to the others. We observe however, that after 2002 the growth in cumulated citations received by CDM is quite rapid and also sustained, contrary to the growth in cumulated citations of a good proportion of the 38 articles which seems to peak around 2011. Part of these characteristics, in particular the slow start, could result from the fact that CDM has been published in a non-indexed journal. We can imagine that if it had been, it would have started gathering citations earlier and possibly more.[11]

2.3. Disciplinary spread

The conception and measurement of knowledge diffusion across space and over time has its roots in the models of contagion and evolving social networks (Watts and Strogatz 1998; Dodds and Watts 2004). Knowledge, somewhat like a 'virus', can spread progressively through a network. In our case CDM is the virus and the network is made up of articles published in different years and journals, which are 'infected' by the fact they cite CDM.[12]

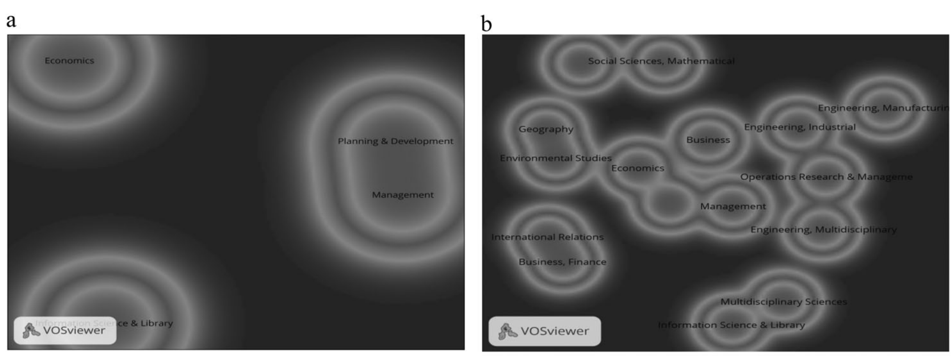
Table 1. The spread of CDM across scientific disciplines.

Time intervals	*T*	*T*+1	*T*+2	*T*+3	No. of times 'infected'	Total no. of articles
Disciplines "infected"						
Economics	x	x	x	x	4	79
Management	x	x	x	x	4	69
Planning and development	x	x	x	x	4	25
Engineering (Industrial)		x	x	x	3	14
Operational research and management science		x	x	x	3	10
Geography		x	x	x	3	7
Environmental studies		x	x	x	3	6
International relations		x	x	x	3	2
Business		x	x	x	3	29
Information and library science	x			x	2	2
Social sciences (Mathematical methods)			x	x	2	8
Engineering (Multidisciplinary)			x	x	2	3
Ethics			x		1	1
Public administration			x		1	1
Urban studies			x		1	1
Multidisciplinary sciences				x	1	1
Agricultural economics and policy				x	1	1
Engineering (Manufacturing)				x	1	1
Total	4	9	14	15	42	260

Source: Web Of Science, authors own calculations. X denotes the presence of one or more articles in the specific discipline in column one, in the time period indicated.

It is instructive to map the disciplinary spread of CDM outside of its original field of economics. To do so we pulled together all CDM citing articles from the Web of Science, taking advantage of its detailed classification of journals in about 250 subject areas (Shapira, Youtie, and Porter 2010). We find that CDM is cited over the years between its publication and 2013 in 260 articles, which are distributed over 18 subject areas, as shown in Table 1 (first and last columns). We see also that the CDM citing articles are for 30% in Economics, 27% in Management, 11% in Business, and 32% in the other 15 disciplines.

To examine this disciplinary spread over time, we also grouped the citing articles by disciplines for three year periods: *T* (2002–2004), *T*+1 (2005–2007), *T*+2 (2008–2010) and *T*+3 (2011–2013).[13] For a visualization we relied on the VOSViewer software to construct heat-maps showing the number of times a particular discipline is counted in each of the four periods. Both Table 1 and Figures 3(a) and 3(b), as well as the more detailed Figures A1–A4 in the Appendix, reveal a clear diffusion of CDM to a large set of subject categories in each time interval T to T+3 and with higher intensity for almost each of the disciplines 'infected'.



Subject Area network of CDM citing articles at T (2002-2004)

Subject Area network of CDM citing articles at T+3 (2011-2013)

Figure 3. (a) Subject Area network of CDM citing articles at T (2002–2004). (b) Subject Area network of CDM citing articles at T+3 (2011–2013).

3. Directed acyclical graphs, community detection algorithms and main path methods

In the first part of our study we have been able to confirm that CDM, although published in a non-indexed journal, is among the most cited comparable articles in its specific domain of research. We have also shown that its diffusion does not seem to slow down yet and has largely spread across a number of disciplines besides economics. We can now focus on what is our main purpose: the detection of clusters or communities in what we call the CDM network or 'universe' and the identification of main paths of diffusion and recombination of knowledge in this universe and its main clusters. In this Section, we give an overview of the methods we are using in both cases. However, since these methods can rely on the fact that the CDM universe constitutes a directed acyclical network, we first recall some properties of such networks. We will in the next Section present our findings.

3.1. Directed acyclical graphs

The CDM universe is defined by the collection of all the articles citing CDM but also the articles cited by CDM. Indeed both upstream knowledge sources and downstream knowledge diffusion, relative to intermediate events or to current terminal points (also called 'sinks') are important to consider for a thorough analysis. CDM itself is not explicitly part of its network, but remains implicitly present since it is directly linked to all the articles of its universe. As opposed to Bhupatiraju et al. (2012) we start with a network in which the 'cognitive division' or distance is not known beforehand. However papers in different scientific fields or sub-fields can be part of the CDM network, since it is also made up of all the papers cited by CDM and by all CDM citing papers. The CDM universe can also be considered as the directed acyclical network or graph (DAG) consisting of precedence relations between a given set of nodes, in our case citations between articles in a given scientific domain. Since all the articles in the CDM universe are 'forward' citing, the relation between them is unidirectional and as such there can be no cycles in the network. However, there can be multiple source articles (nodes with no incoming edges) as well as multiple 'sink' articles (nodes with no outgoing edges) as shown in the example in Figure 4.

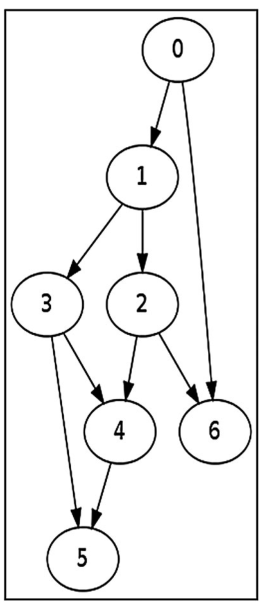


Figure 4. Simple example of a directed acyclical graph.

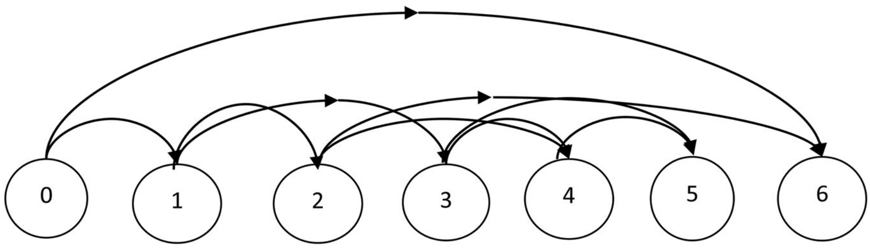


Figure 5. A topological representation of the directed acyclical graph of Figure 4.

The specific characteristics of acyclical networks allow a number of manipulations which are not possible with a non-acyclical network. In particular, next to its acyclic nature, the nodes in a DAG usually have a date or 'time-stamp' that is compatible with the edges. These characteristics make it possible to sort and order the network in consecutive steps, and to employ community detection algorithms. Figure 5 shows a possible 'topological' representation of the graph of Figure 4 where we see the source (0) at the left and the two sinks (5 and 6) at the right, and the intermediate nodes ranked in between.

3.2. Community detection algorithms

The development and use of community detection algorithms has proliferated since the Girvan and Newman's (2002) method based on edge betweenness (or centrality) was published. An update of that algorithm used a measure of 'modularity' as a goodness-of-fit criterion, in which 'modularity' is defined as the number of edges (citation links) within a given cluster of nodes (articles), minus the expected number of edges found in a similar network where edge placement is random.[14] Although still a popular method, the modularity maximization approach has the deficiency of trying to merge small clusters and split the larger ones (Lancichinetti and Fortunato 2012). We adopt here a new method, called OSLOM2 for Order Statistics Local Optimization Method, which is more appropriate for directed acyclical networks (Lancichinetti et al. 2011). This method identifies a cluster solution by calculating the probabilities of clusters as that of finding the same clusters in a random network with identical features, and by maximizing these probabilities.[15]

In addition to the OSLOM2 algorithm, we rely on a number of visualization methods contained in the social network analysis software packages Pajek and Gephi.

Box 2. A community detection algorithm adapted to directed acyclical graphs: OSLOM2

Unlike the earlier modularity-based algorithms, OSLOM2 can also be used on directed acyclical graphs. A solution based on the definition of two uniformly distributed random variables r_{in} and r_{out} is employed in this case. The first variable is based on the probability that node i has outgoing edges ending on nodes from a given subgraph G, while the second variable is based on the probability that node i has incoming edges originating from nodes from the same subgraph. The final score of node i is given by the product of $r_{\text{in}} * r_{\text{out.}}$ The algorithm next calculates the distribution of this product and determines the significance of its order statistics as in the generalized probability equation, stated as follows (Lancichinetti et al. 2011):

$$p(k_i^{\text{in}}|i,\ C,\ G) = A\frac{2^{-k_i^{\text{in}}}}{k_i^{\text{out}}!k_i^{\text{in}}!(m_c^{\text{out}} - k_c^{\text{in}})!(M*/2)!}. \tag{2}$$

Here C is a given subgraph of graph $G\backslash[C \cup \{i\}]$ while node $i \notin C$. Furthermore, the degree of subgraph C is m_c, the degree of node i is k_i, while all other nodes have a total degree of M. When separating these quantities as in- or out- degrees (relevant to C) we can write these as: k_i^{out}, k_i^{in}, m_c^{out} and m_c^{in}, while $M*$ is the internal degree of graph $G\backslash[C \cup \{i\}]$. And finally, A is the normalization factor.

3.3. Main path analysis methods

In order to explore the knowledge diffusion process within clusters or communities we use two main path analysis methods based on the topological ordering of directed acyclical networks. Among the multiple methods measuring the importance of nodes and edges in such networks, we investigated in detail the applicability of two: the Critical Path Method or CPM of Kelley and Walker (1959), and the Search Path Count or SPLC method pioneered by Hummon and Doreian (1989, 1990), which has been recently used by Verspagen (2007), Bhupatiraju et al. (2012) and Martinelli and Nomaler (2014). The CPM is a very general method used to plan, organize and schedule sequences of interdependent tasks in complex activities such as logistics and transport. The SPLC method identifies a 'main path' throughout a network, or through different sub-networks or clusters, by prioritizing nodes with a high in-degree centrality and high out-degrees to other high centrality nodes. It does so by assigning weights to each edge in the network that are appropriately based on the number of preceding and succeeding nodes.

4. Results: CDM clusters and knowledge main paths

In this Section we present our findings in what we called the horizontal and vertical planes of our analysis: first on the clustering of the CDM universe and next on the main knowledge paths for the CDM universe as a whole and within each of the three main identified clusters. Most of these results are shown in the text as graphs in Figures 6, 7, 8(a) and 8(b) with accompanying explanatory Tables. They are also complemented by similar Figures A5, A6 and A7 in Appendix 2.

4.1. Three main clusters

Applying the OSLOM2 method to the CDM universe we arrive at a clustering solution that we find clearly preferable to the ones obtained with the Girvan and Newman's modularity maximization approaches. A visualization of this solution, performed using the software packadge Gephi, is shown in Figure 6. It is also documented in detail in the Table of the internet annex http://www.merit.unu.edu/TheCDMFramework-InternetAnnex/, which for each of the 192 articles of the CDM universe indicates their cluster and their bibliographic reference (title, first authors, journal publication year and labels from 1 to 192).[16]

Figure 6 gives us the sense of the size of the clusters, of how they overlap and are interconnected. The cluster by far largest is cluster 2 in dark grey with 84 nodes. The two other significant clusters are clusters 1 and 3 in medium grey and light grey with respectively 32 and 26 nodes. Looking at the labels of the nodes, recorded in the internet annex Table, we can gather that the cluster 1 is dominated by papers on innovation and firm performance, particularly in the United Kingdom and Ireland, while cluster 3 leans towards productivity issues. The preponderant cluster 2 is more heterogeneous and seems to incorporate the complete breadth of CDM related works. The small clusters (nodes 141, 116, 15 and nodes 167, 168, 189) in darker grey have a specific focus on education and on competition and regulation. The overlapping nodes in black are technical or some miscellaneous intermediate papers. The homeless nodes in white include some papers that might be placed topically in one of the other clusters; however, their references and citations did not make it overly clear where they belong.

In the case of the three large clusters we observe that the recurrence of similar co-authorships is rather frequent. Such recurrence is more likely to appear the smaller the size and time dimension of the cluster. This is what we see in cluster 1 particularly, as well as in cluster 3, but less so in the nearly three times larger cluster 2. We thus remain confident that our cluster solution rests primarily on knowledge or cognitive proximity, not social or geographic proximity (Nooteboom et al. 2007).

In fact common characteristics as social relations (co-authorships) and geographic relations (affiliations), that we observe between articles within clusters, appear also when we focus our attention, as

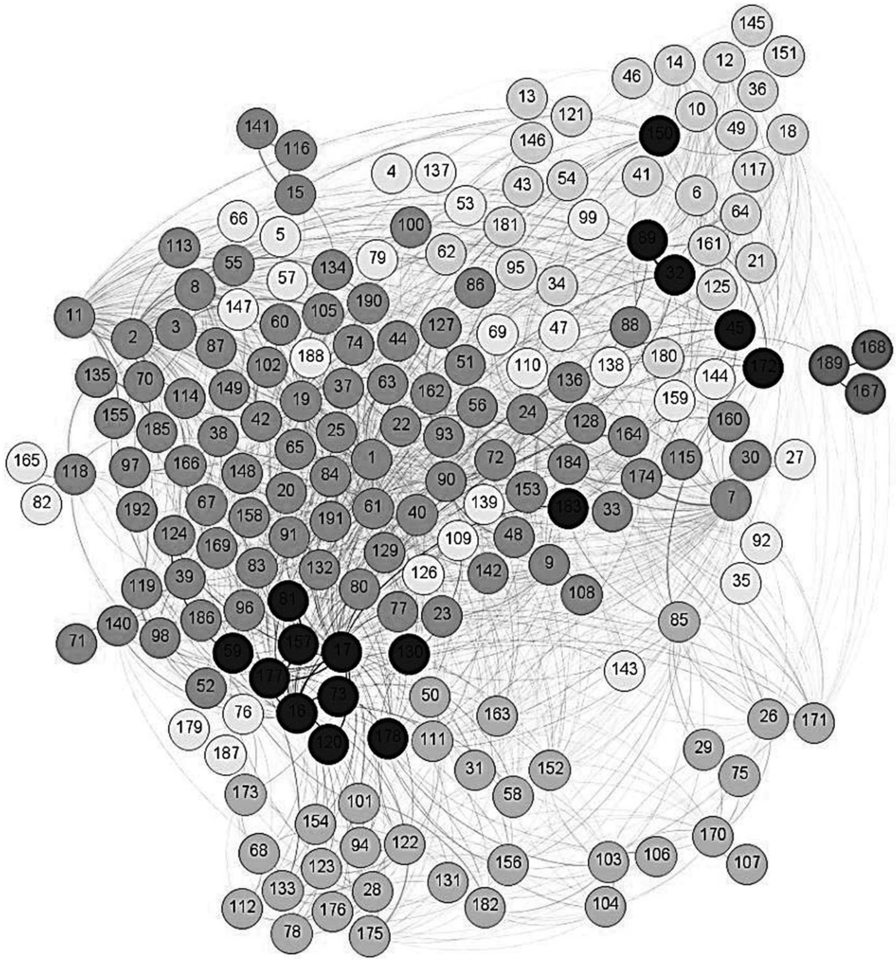


Figure 6. OSLOM2 clustering solution for the CDM universe.

we do below, on the associated main paths solutions (see Figures 8a and 8b, and in Figures A6 and A7 in the Appendix). These characteristics play here the same role as they do in the 'social contagion' model of Dodds and Watts (2004), in which the spread of information is dependent on multiple 'exposures' or on reaching certain 'thresholds'. In this model an actor or agent passes a threshold when he decides to use a certain piece of information based on 'information use decisions' made earlier by other actors in his network or community. This is especially relevant in the cases of clusters 2 and 3 for which we see that multiple thresholds are passed in terms of new authorship and geographic affiliations, contrary to cluster 1 for which these thresholds remain in place.

4.2. Main path for CDM universe as a whole

Although we have identified three distinct and sizeable clusters in the CDM universe, it is nonetheless instructive to compare the two knowledge main paths we find when we consider it as a whole, using the Critical Path Method (CPM) and the Search Path Count Method (SPLC). In Figure 7 where the nodes of the different clusters are denoted differently, we show the critical path obtained with the CPM method. We see that the three main clusters as calculated by OSLOM2 are incorporated in this critical path, with the squares representing the nodes from cluster 1, the stars the nodes from cluster 2, and the triangles the nodes from cluster 3. Interestingly we observe that these clusters

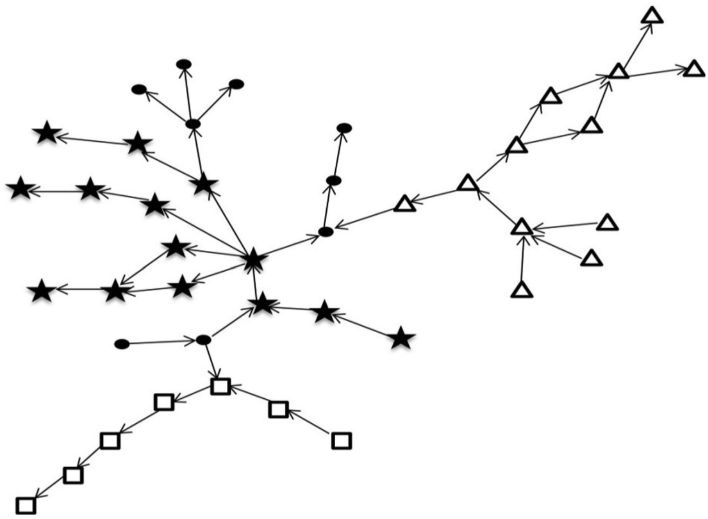


Figure 7. Critical path for the CDM universe as a whole with the nodes of the three largest clusters indicated by their shape and the knowledge flows indicated by the arrows on the edges.

make up three distinct parts of the path, and are separated by the black dots indicating the overlapping nodes, or knowledge 'bridge' or 'intermediary' nodes.

The findings provided by the SPLC method are not as telling as those we obtain with the CPM method. The order, label and cluster of the selected nodes in the main path, as calculated by this method, are recorded in Table 2. This Table is more informative than the Figure A5 in the Appendix 2, which is similar but less helpful than Figure 7 for the CPM main path. We observe in particular that this main path cuts right across clusters 1 and 2 with no real distinction between them and that it ignores cluster 3. At the difference of the CPM method, the results given by the SPLC method are unclear and not supportive of the OSLOM2 clustering solution.

4.3. Main paths for the largest CDM cluster

Beyond the comforting picture of the main knowledge path of the CDM universe as a whole provided by the CPM method, it remains most appropriate to look for the knowledge main paths within each of three well identified clusters separately. Since we find now that both the CPM

Table 2. SPLC main path for the CDM universe.

Sort order	Label	Cluster	Time-stamp
0 (source)	The common structure of statistical model of truncation, sample selection and limited dependent variables and a simple estimator for such models/Heckman	Overlapping	1976
1	Sample selection bias as a specification error/Heckman	Overlapping	1979
2	Firm-level innovation and productivity-Is there a common story across countries?/Janz, Lööf and Peters	Cluster 2	2003
3	Innovation and productivity across four European countries/Griffith, Huergo, Mairesse and Peters	Cluster 2	2006
4	Innovation and productivity in SMEs: Empirical evidence for Italy/Hall, Lotti and Mairesse	Cluster 2	2009
5	External interaction, innovation and productivity: An application of the innovation value chain to Ireland/Doran and O'Leary	Cluster 1	2011
6 (sink)	Are differing forms of innovation complements or substitutes?/Doran	Cluster 1	2012
7 (sink)	Regulation and firm perception, eco-innovation and firm performance/Doran and Ryan	Cluster 1	2012
8 (sink)	The Effects of national and international interaction on innovation: evidence from the Irish CIS: 2004–2006/Doran, Jordan and O'Leary	Cluster 1	2012

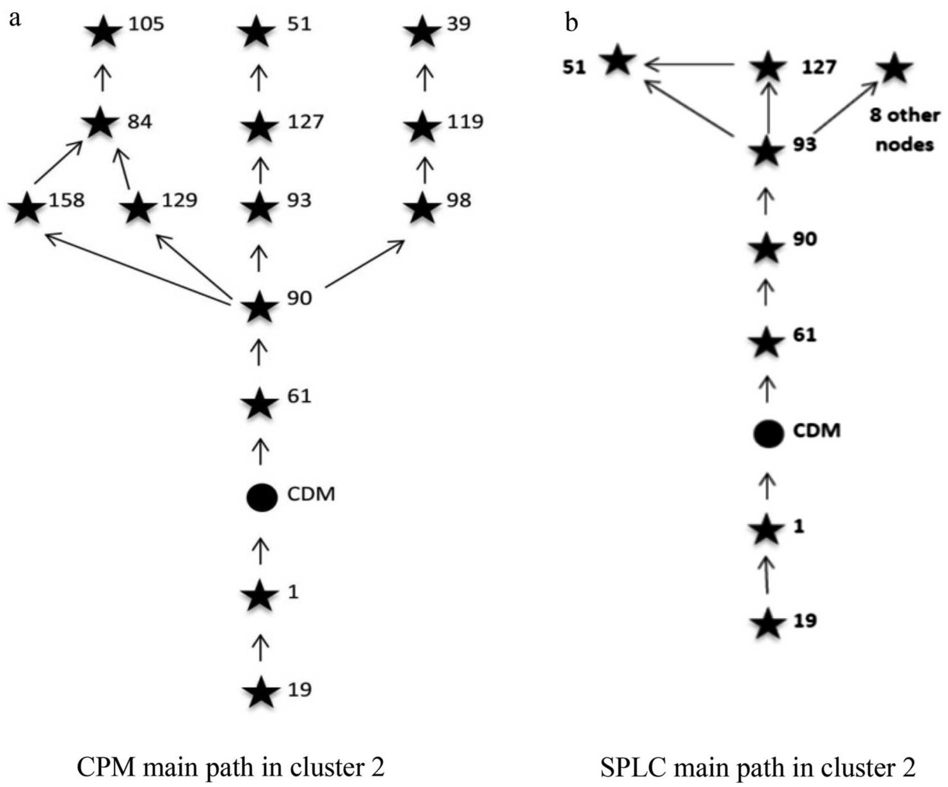


Figure 8. (a) CPM main path in cluster 2. (b) SPLC main path in cluster 2.

Legend Figure 8a	
19: Who Does R&D and Who Patents / Bound, Cummins, Griliches, Hall and Jaffe	1984
1: A reprise of size and R&D / Cohen and Klepper	1996
CDM	1998
61: Firm-level innovation and productivity-Is there a common story across countries? / Janz, Lööf and Peters	2003
90: Innovation and productivity across four European countries / Griffith, Huergo, Mairesse and Peters	2006
158: Technological innovation and productivity in late-transition Estonia: Econometric evidence from innovation surveys / Masso and Vahter	2008
129: Northern and southern innovativity: A comparison across European and Latin American countries / Raffo, Lhuillery and Miotti	2008
93: Innovation and productivity in SMEs: Empirical evidence for Italy / Hall, Lotti and Mairesse	2009
98: Innovation success of non-R&D-performers: Substituting technology by management in SMEs / Rammer, Czarnitzki and Spielkamp	2009
84: Innovation and economic development / Fagerberg, Srholec and Verspagen	2010
127: New product introduction and product tenure: What effects on firm growth? / Cucculelli and Ermini	2012
119: Marketing and organizational innovations in entrepreneurial innovation processes and their relation to market structure and firm characteristics / Schubert	2010
105: Innovation, research and development, and productivity in Chile / Alvarez, Bravo-Ortega and Navarro	2011
51: Drivers to firm innovation and their effects on performance: An international comparison / Fernandes, Ferreira and Raposo	2013
39: Cooperation with public research institutions and success in innovation: Evidence from France and Germany / Robin and Schubert	2013

Legend Figure 8b	
19: Who Does R&D and Who Patents / Bound, Cummins, Griliches, Hall and Jaffe	1984
1: A reprise of size and R&D / Cohen and Klepper	1996
CDM	1998
61: Firm-level innovation and productivity-Is there a common story across countries? / Janz, Lööf and Peters	2003
90: Innovation and productivity across four European countries / Griffith, Huergo, Mairesse and Peters	2006
93: Innovation and productivity in SMEs: Empirical evidence for Italy / Hall, Lotti and Mairesse	2009
127: New product introduction and product tenure: What effects on firm growth? / Cucculelli and Ermini	2012
51: Drivers to firm innovation and their effects on performance: An international comparison / Fernandes, Ferreira and Raposo	2013
44: Do eco-innovations harm productivity growth through crowding out? Results of an extended CDM model for Italy / Marin	2014
52: Effectiveness and efficiency of SME innovation policy / Foreman-Peck	2013
56: Evidence on the impact of R&D and ICT investments on innovation and productivity in Italian firms / Hall, Lotti and Mairesse	2013
86: Innovation and economic performance: The case of greek SMEs / Beneki, Giannias and Moustakas	2012
162: The effects of government financial support on business innovation in South Korea / Kim and Lee	2011
169: The impact of information and communication technologies: an insight at micro-level on one Italian region / Brasini and Freo	2012
190: Using innovation surveys for econometric analysis / Mairesse and Mohnen	2010
192: Young firms and innovation: A microeconometric analysis / Pellegrino, Piva and Vivarelli	2012

and SPLC methods lead to results that are quite similar for each cluster considered individually, we consider here the results found for our larger cluster 2, and only display for the smaller clusters 1 and 2 the main paths obtained by the SPLC method in Figures A6 and A7 in Appendix 2. We thus comment here the two Figures 8(a) and 8(b) and their detailed legends, which show respectively the CPM and SPLC main paths for cluster 2. We observe that the selection and ordering of the nodes along the main paths for the two methods are not only quite consistent but also rather close, to the contrary of our conclusion for the CDM universe overall. We see that out of 14 nodes for CPM and 15 for SPLC, the seven nodes 19, 1, 61, 90, 93, 127 and 51 are present on both main paths with the same consecutive ordering. We observe that the 2009 paper of Hall, Lotti and Mairesse (node 90) plays a key role as a disseminator of knowledge to several new branches in the SPLC ordering, while this role is played by the 2006 paper of Griffiths, Huergo, Mairesse and Peters (node 93) in the CPM ordering. In both cases we also find that the two originators before CDM for the core literature within the CDM universe are the 1984 article of Bound, Cummins, Griliches, Hall and Jaffe (node 19) and the 1996 article of Cohen and Klepper (node 1).

5. In the guise of a conclusion

We have started our exploratory study with an aperçu of science of science's studies and the concept of 'invisible college'. Our results clearly confirm our premise that the authors of the articles citing CDM and/or CDM cited articles, which form CDM universe, constitute such colleges.

We have indeed been able to build, in the main part of our analysis, an evolutionary tree encompassing both the horizontal and vertical dimensions of knowledge diffusion and recombination processes. We find that we can clearly identify three large clusters or communities, covering 75% of the CDM universe (respectively 44%, 17% and 14%), and that for each of them we can determine a meaningful knowledge path – which we could not obtain for the CDM universe as a whole.

In particular, we observe not only that the dynamics of knowledge recombination are very distinct for these three large communities, but that they also correspond to what has been called 'academic tribes', differentiated by institutional links and social and geographic proximity. More work would be necessary to better assess what these characteristics are, and how they interact with the knowledge diffusion dynamics, facilitating or impeding it, and orienting it. Further research is also needed to interpret better our present results in their cognitive dimension. This would imply a fine-grained look at the components of the new knowledge that are actually actually diffused and recombined in the CDM universe. Are they and to what extent mainly data related, factual or theoretical, practical or methodological?.

We have also touched upon a few complementary analyses that will need to be developed, such as the comparison with a sample of articles, contemporary to CDM and with similar focus. Likewise in our implementation of the methods we use to identify clusters and main paths in the CDM universe, we have encountered some issues, which will deserve broader consideration in wider settings (e.g. on universe possibly grounded on more articles than one only, with more previous references, and over a longer period). In particular, is it the case, and why, that the CPM and SPLC methods provide similar genealogical paths for well defined directed acyclical sub-networks, while CPM seems to work better for encompassing heterogeneous networks?

One conclusion, with no need for further analysis appears, clearly when we consider the list of authors of the largest cluster of the CDM universe. Nearly 20% of the articles making up this cluster have been written by a handful of authors, in particular B. Hall, J. Mairesse and P. Mohnen, often with younger coauthors. Much of their research has been shaped by discussing and working with the late Zvi Griliches, and by participating in the seminars and other activities of the NBER Productivity Program that he had launched in 1979 and directed until his death in 1999. The CDM framework and the lines of research it has stimulated find deeper roots in Zvi Griliches' research accomplishment and leadership. They are one testimony among others to his long-lasting influence.

Notes

1. For example see Nonaka (1994) and Hargadon (2003) for the information management context.
2. The discussion revolving around the issue whether and how proximity influences innovation, and specifically geographic, social and cognitive distance, has already produced a number of influential papers (Boschma 2005; Nooteboom et al. 2007; Broström 2010). Although this paper can be seen as a case study on the influence of cognitive proximity or knowledge distance, we would like to make a distinction between cognitive proximity in terms of topical (dis-)similarity and cognitive proximity in terms of influence or inspiration from intellectual predecessors. We should also note that while slightly reminiscent of the average degrees of separation within collaboration networks issue, as exemplified by Erdos' collaboration network (Newman 2001), we are not studying co-authorships in this context. It goes without saying of course that both topical similarity and co-authorships exist in our 'invisible college'.
3. NBER Working Paper number 6696, August 1998.
4. CREST Working Paper 1998, n°33.
5. The downward bias in terms of the number of citations received by CDM is likely to affect our two preliminary investigations of its diffusion relative to other similar articles published in indexed journals, and across different disciplinary fields. However, the impact on our central analysis is probably very small, since it is not based on the citations received by CDM, but on the citations of the papers citing CDM between themselves.
6. The EINT version is cited 216 times in RePEc, while the NBER version is cited 272 times and the CREST version 176 times (in EconPapers.repec.org as of 25-11-2013). The second source we consulted was Google Scholar (data collected through A. Harzing's 'Publish or Perish' software: http://www.harzing.com/pop.htm) in which we found some 1400 citations or mentions. It should be stressed that the REPEC and Google Scholar references/citation counts are not de-duplicated and cleaned. The third source was the Econlit bibliographic database, which interestingly only returns 56 citations for the NBER working paper. The fourth and fifth sources are the citations found in the Web of Science and Scopus. Here we find 136 records for the WoS (date of retrieval: 24 September 2013) and 180 records for Scopus (date of retrieval: 6 September 2013). At a later stage (date of retrieval: 18 February 2014), we updated the Scopus data finding 211 citations, which after cleaning gave 208 unique citations.
7. Classifying the articles would be possible using lexical-semantic techniques (as in Wang, Notten, and Surpatean 2013) but it remains a costly endeavour and we did not think it worthwhile in our case.
8. Retrieval date: 12 March 2015.
9. Adding non-indexed articles like we do for CDM would be in principle possible but would be difficult and costly. It would involve not only to select the proper non-indexed journals and relevant articles in these journals, but also to gather the citations they received in both these non-indexed journals and the indexed ones, without benefiting of the tools of the citation index of Web of Science.
10. See the Internet annex (http://www.merit.unu.edu/TheCDMFramework-InternetAnnex/).
11. Articles from journals which are indexed in a citation index are, on average, of higher quality when measured according to the H-Index of the journal (Harzing and van der Wal 2009). Hence, indexed journals are generally perceived to be of higher quality and as such enjoy a wider exposure. This leads to the hypothesis that articles in non-indexed journals, which do not receive the status and attention afforded to the indexed journals, gather citations later and at a slower pace than their indexed peers.
12. Based on Garfield's (1980) work and inspired by a contagion model similar to Dodds and Watts (2004). See also Van Eck and Waltman 2007.
13. As already mentioned, CDM has not received any citation in 1999 the first year of its publication, nor during the three following years 2000–2002.
14. See Newman and Girvan 2004; Newman 2006; and Blondel et al., 2012.
15. The program of OSLOM2 is retrievable at: http://www.oslom.org/software.htm.
16. The different clusters are represented in different shades of black, gray and white in Figure 6 in the text and in colors in the internet annex (http://www.merit.unu.edu/TheCDMFramework-InternetAnnex/).

Acknowledgement

The authors have benefitted from comments and suggestions of two referees and the participants to the three CDM workshops in London, Paris and Maastricht, in particular Bronwyn Hall, Hans Lööf, Pierre Mohnen and Bettina Peters.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCiD

Ad Notten http://orcid.org/0000-0002-3852-7954

References

Azoulay, P., J. Graff Zivin, and J. Wang. 2010. "Superstar Extinction." *Quarterly Journal of Economics* 125: 549–589.

Becher, T., and P. Trowler. 2001. *Academic Tribes and Territories: Intellectual Enquiry and the Culture of Disciplines*. Milton Keynes: Society for Research into Higher Education & Open University Press.

Bhupatiraju, S., Ö. Nomaler, G. Triulzi, and B. Verspagen. 2012. "Knowledge Flows: Analyzing the Core Literature of Innovation, Entrepreneurship and Science and Technology Studies." *Research Policy* 41: 1205–1218.

Blondel, V. C., J. -L. Guillaume, R. Lambiotte, and E. Lefebvre. 2012. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics* 2008 (10): P10008.

Boschma, R. 2005. "Proximity and Innovation: A Critical Assessment." *Regional Studies* 39 (1): 61–74.

Broström, A. 2010. "Working with Distant Researchers: Distance and Content in University – Industry Interaction." *Research Policy* 39 (10): 1311–1320.

Cohen, W. M., and D. A. Levinthal. 1990. "Absorptive Capacity: A New Perspective on Learning and Innovation." *Administrative Science Quarterly* 35 (1): 128–152.

Cole, S. 1983. "The Hierarchy of the Sciences?" *American Journal of Sociology* 89 (1): 111–139.

Cowan, R., P. David, and D. Foray. 2000. "The Explicit Economics of Knowledge Codification and Tacitness." *Industrial and Corporate Change* 9 (2): 211–253.

Crane, D. 1972. *Invisible Colleges: Diffusion of Knowledge in Scientific Communities*. Chicago, IL: University of Chicago Press.

Crepon, B., E. Duguet, and J. Mairesse. 1998. "Research, Innovation and Productivity: An Econometric Analysis at the Firm Level." *Economics of Innovation and New Technology* 7 (2): 115–158.

David, P. 1998. "Communication Norms and the Collective Cognitive Performance of Invisible Colleges." In *Creation and Transfer of Knowledge*, edited by G. Barba Navaretti, 115–163. Berlin: Springer.

De Solla Price, D. J. 1963. *Little Science, Big Science*. New York: Columbia University Press.

Dodds, P., and D. Watts. 2004. "Universal Behavior in a Generalized Model of Contagion." *Physical Review Letters* 92 (21): 218701.

Fagerberg, J., and B. Verspagen. 2009. "Innovation Studies: The Emerging Structure of A new Scientific Field." *Research Policy* 38: 218–233.

Garfield, E. 1980. "The Epidemiology of Knowledge and the Spread of Scientific Information." *Current Contents.* 35: 5–10.

Garfield, E. 2009. "From the Science of Science to Scientometrics." *Journal of Informetrics* 3: 173–179.

Girvan, M., and M. E. J. Newman. 2002. "Community Structure in Social and Biological Networks." *Proceedings of the National Academy of Sciences* 99: 7821–7826.

Hargadon, A. 2003. *The Surprising Truth How Companies Innovate*. Boston, MA: Harvard Business School Press.

Harzing, A. W. K., and R. van der Wal. 2009. "A Google Scholar H-Index for Journals: An Alternative Metric to Measure Journal Impact in Economics & Business?" *Journal of the American Society for Information Science and Technology* 60 (1): 41–46.

Heinze, T., R. Heidler, R. H. Heiberger, and J. Riebling. 2013. "New Patterns of Scientific Growth: How Research Expanded After the Invention of Scanning Tunneling Microscopy and the Discovery of Buckminsterfullerenes." *Journal of the American Society for Information Science and Technology* 64 (4): 829–843.

Hummon, N. P., and P. Doreian. 1989. "Connectivity in a Citation Network: The Development of DNA Theory." *Social Networks* 11 (1): 39–63.

Hummon, N. P., and P. Doreian. 1990. "Computational Methods for Social Network Analysis." *Social Networks* 12: 273–288.

Kelley, J., and M. Walker. 1959. "Critical-Path Planning and Scheduling." *Proceedings of IRE-AIEE-ACM '59 (Eastern)*, 160–173.

Knorr Cetina, K. 1999. *Epistemic Cultures. How the Sciences Make Knowledge*. Cambridge: Harvard University Press.

Lancichinetti, A., and S. Fortunato. 2012. "Limits of Modularity Maximization in Community Detection." *Physical Review E* 84 (6): 066112.

Lancichinetti, A., F. Radicchi, J. J. Ramasco, and S. Fortunato. 2011. "Finding Statistically Significant Communities in Networks." *PLoS ONE* 6 (4): e18961.

Lave, J., and E. Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.

Martinelli, A., and Ö. Nomaler. 2014. "Measuring Knowledge Persistence: A Genetic Approach to Patent Citation Networks." *Journal of Evolutionary Economics* 24 (3): 623–652.

Mazloumian, A., Y.-H. Eom, D. Helbing, S. Lozano, and S. Fortunato. 2011. "How Citation Boosts Promote Scientific Paradigm Shifts and Nobel Prizes." Plos One. doi:10.1371/journal.pone.0018975.

Merton, R. K. 1973. *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago, IL: University of Chicago Press.

Newman, M. E. J. 2001. "The Structure of Scientific Collaboration Networks." *Proceedings of the National Academy of Sciences,* 98 (2): 404–409.

Newman, M. E. J. 2006. "Modularity and Community Structure in Networks." *Proceedings of the National Academy of Sciences* 103 (23): 8577–8582.

Newman, M. E. J., and M. Girvan. 2004. “Finding and Evaluating Community Structure in Networks.” *Physical Review E* 69 (2): 026113.

Nonaka, I. 1994. “A Dynamic Theory of Organizational Knowledge Creation.” *Organization Science* 5: 14–37.

Nonaka, I., and H. Takeuchi. 1995. *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford: Oxford University Press.

Nooteboom, B., W. P. M. Van Haverbeke, G. M. Duijsters, V. A. Gilsing, and A. V. d.Oord. 2007. “Optimal Cognitive Distance and Absorptive Capacity.” *Research Policy* 36: 1016–1034.

Polanyi, M. 1962. *Personal Knowledge*. London: Routledge and Kegan Paul.

Rosenkopf, L., and A. Nerkar. 2001. “Beyond Local Search.” *Strategic Management Journal* 22 (4): 287–306.

Shapira, P., J. Youtie, and A. L. Porter. 2010. “The Emergence of Social Science Research in Nanotechnology.” *Scientometrics* 85 (2): 595–611.

Tushman, M. 1977. “Special Boundary Roles in the Innovation Process.” *Administrative Science Quarterly* 22 (4): 587–605.

Van Eck, N. J., and L. Waltman. 2007. “VOS: A New Method for Visualizing Similarities Between Objects.” In *Advances in Data Analysis: Proceedings of the 30th Annual Conference of the German Classification Society*, edited by H-J. Lenz, and R. Decker, 299–306. Heidelberg: Springer.

Verspagen, B. 2007. “Mapping Technological Trajectories as Patent Citation Networks: A Study on the History of Fuel Cell Research.” *Advances in Complex Systems* 10: 93–115.

Verspagen, B., and C. Werker. 2004. “Keith Pavitt and the Invisible College of the Economics of Technology and Innovation.” *Research Policy* 33: 1419–1431.

Wagner, C. 2008. *The New Invisible College*. Washington, DC: Brookings Institution Press.

Wang, L., A. Notten, and A. Surpatean. 2013. “Interdisciplinarity of Nano Research Fields: A Keyword Mining Approach.” *Scientometrics* 94 (3): 877–892.

Watts, D., and S. Strogatz. 1998. “Collective Dynamics of ‘Small-World’ Networks.” *Nature* 393 (6684): 440–442.

Whitley, R. 2000. *The Intellectual and Social Organization of the Sciences*. Oxford: Oxford University Press.

Appendix 1. *The Spread of CDM across scientific disciplines*

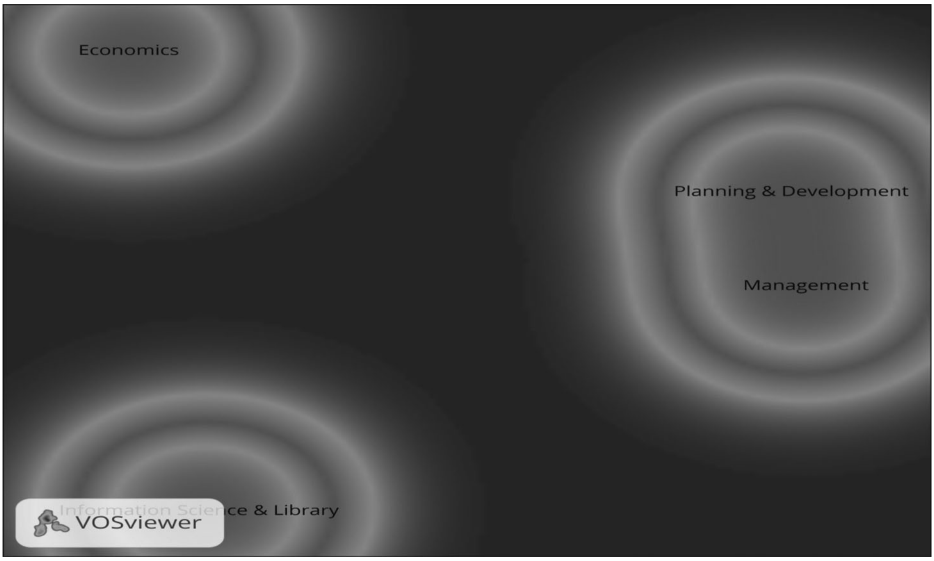


Figure A1. Subject Area network of CDM citing articles at *T* (2002–2004).

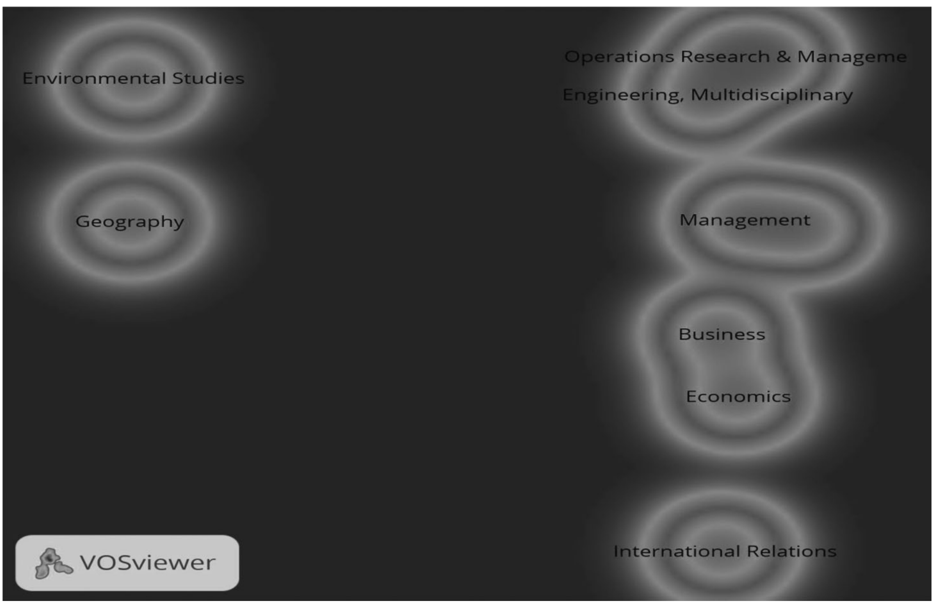


Figure A2. Subject Area network of CDM citing articles at *T*+1 (2005–2007).

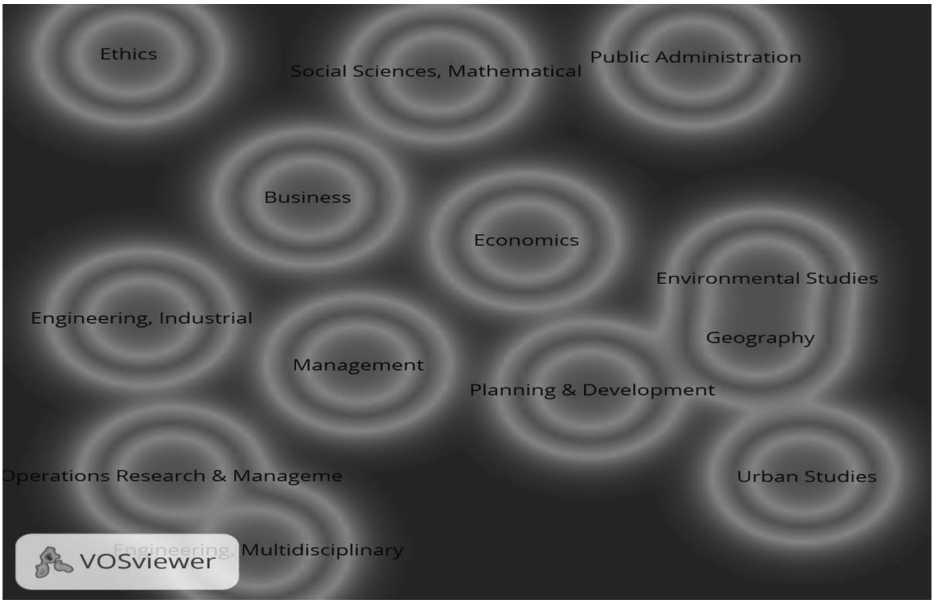


Figure A3. Subject Area network of CDM citing articles at T+2 (2008–2010).

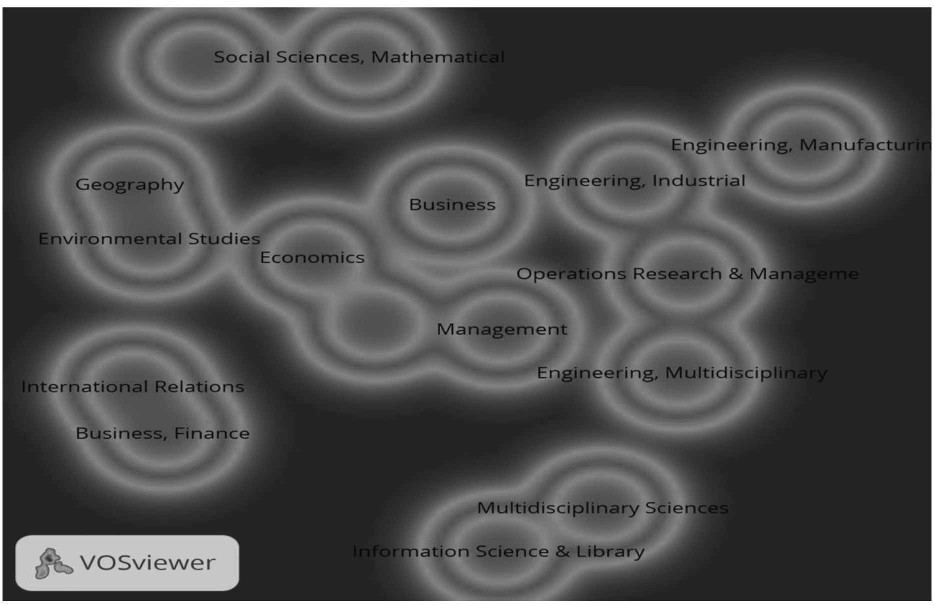


Figure A4. Subject Area network of CDM citing articles at T+3 (2011–2013).

Appendix 2. *SPLC Main Path and Hierarchies Clusters 1, 2 and 3*

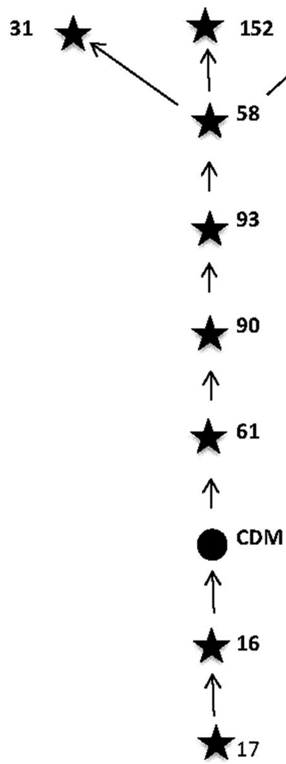


Figure A5. SPLC main path for the CDM universe as a whole.

Legend Figure A5	
17: The common structure of statistical model of truncation, sample selection and limited dependent variables and a simple estimator for such models / Heckman	1976
16: Sample selection bias as a specification error/Heckman	1979
61: Firm-level innovation and productivity-Is there a common story across countries? /Janz, Lööf and Peters	2003
90: Innovation and productivity across four European countries/Griffith, Huergo, Mairesse and Peters	2006
93: Innovation and productivity in SMEs: Empirical evidence for Italy/Hall, Lotti and Mairesse	2009
58: External interaction, innovation and productivity: An application of the innovation value chain to Ireland/ Doran and O'Leary	2011
31: Are differing forms of innovation complements or substitutes? / Doran	2012
163: Regulation and firm perception, eco-innovation and firm performance/Doran and Ryan	2012
152: The Effects of National and International Interaction on Innovation: Evidence from the Irish CIS: 2004–2006/Doran, Jordan and O'Leary	2012

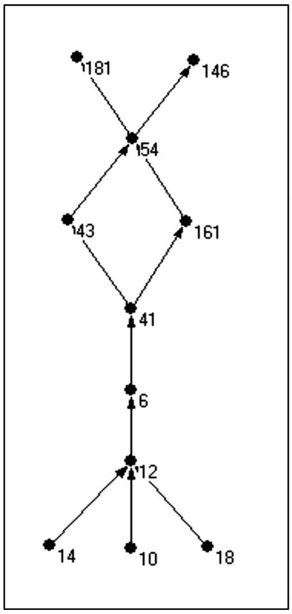


Figure A6. SPLC main path in cluster 3.

Legend Figure A6	
14: Recherche-développement et performances des entreprises: Une étude économétrique sur données individuelles/ Cuneo and Mairesse	1985
10: Productivity and R&D at the Firm Level in French Manufacturing/Cuneo and Mairesse	1984
18: Time-series and Cross-sectional Estimates on Panel Data: Why are They Different and Why Should They Be Equal?/ Mairesse	1990
12: R&D productivity: A survey of econometric studies at the firm level/Mairesse and Sassenou	1991
6: Exploring the Relationship between R&D and Productivity in French Manufacturing Firms/Hall and Mairesse	1995
41: Demand and innovation in productivity growth/Pianta and Crespi	2008
43: Diversity in innovation and productivity in Europe/Crespi and Pianta	2009
161: The economic impact of technological and organizational innovations. A firm-level analysis/Evangelista and Vezzani	2010
54: Engines of growth. Innovation and productivity in industry groups / Bogliacino and Pianta	2011
181: The relation between firm size and R&D productivity in different technological regimes / Revilla and Fernandez	2012
146: Profits, R&D, and innovation-a model and a test / Bogliacino and Pianta	2013

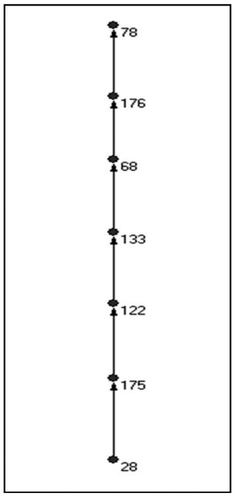


Figure A7 SPLC main path in cluster 1

Legend Figure A7	
28: An ex ante evaluation framework for the regional benefits of publicly supported R&D projects / Roper, Hewitt-Dundas and Love	2004
175: The innovation decision: An economic analysis / Du, Love and Roper	2007
122: Modelling the innovation value chain / Roper, Du and Love	2008
133: Output additionality of public support for innovation: Evidence for Irish manufacturing plants / Hewitt-Dundas and Roper	2010
68: From knowledge to added value: A comparative, panel-data analysis of the innovation value chain in Irish and Swiss manufacturing firms / Roper and Arvanitis	2012
176: The innovation value chain in new technology-based firms: Evidence from the U.K. / Ganotakis and Love	2012
78: Information systems, inter-functional collaboration and innovation in Taiwanese high-tech manufacturing firms / Ganotakis, Hsieh and Love	2013