Maastricht University

UNITED NATIONS UNIVERSITY
UNU-MERIT

Working Paper Series

#2016-048

River deep, mountain high: Of long-run knowledge trajectories within and between innovation clusters

Önder Nomaler and Bart Verspagen

Maastricht Economic and social Research institute on Innovation and Technology (UNU-MERIT)
email: info@merit.unu.edu | website: http://www.merit.unu.edu

Maastricht Graduate School of Governance (MGSoG)
email: info-governance@maastrichtuniversity.nl | website: http://www.maastrichtuniversity.nl/governance

Boschstraat 24, 6211 AX Maastricht, The Netherlands
Tel: (31) (43) 388 44 00

UNU-MERIT Working Papers
ISSN 1871-9872

Maastricht Economic and social Research Institute on Innovation and Technology
UNU-MERIT

Maastricht Graduate School of Governance
MGSoG

UNU-MERIT Working Papers intend to disseminate preliminary results of research carried out at UNU-MERIT and MGSoG to stimulate discussion on the issues raised.

River Deep, Mountain High: Of Long-Run Knowledge Trajectories Within and Between Innovation Clusters[1]

Önder Nomaler
(Eindhoven University of Technology, ECIS & UNU-MERIT)
Bart Verspagen
(UNU-MERIT and Maastricht University, Department of Economics)

Version 3.2, (final, September 2016)

Abstract:

We bring together the topics of geographical clusters and technological trajectories, and shift the focus of the analysis of regional innovation to main technological trends rather than firms. We define a number of inventive clusters in the US space and show that long chains of citations mostly take place between these clusters. This is reminiscent of the idea of global pipelines of knowledge transfer that is found in the geographical literature. The deep citations are used to identify technological trajectories, which are the main directions along which incremental technological progress accumulates into larger changes. While the origin and destination of these trajectories are concentrated in space, the intermediate nodes travel long distances and cover many locations across the globe. We conclude by calling for more theoretical and empirical attention to the "deep rivers" that connect the "high mountains" of local knowledge production.

JEL Codes: O33, O31, R11
Keywords: patent citations; regional concentration of inventive activities; technological trajectories

[1] We thank participants at the 9th Summer Academy on 'Innovation and Uncertainty' in Jena, Germany, 26 July-8 August 2015, participants at the workshop on 'Evolutionary approaches informing research on entrepreneurship and regional development', Gothenburg, 9-11 December 2015, for comments to a presentation of the results presented in this paper, and two anonymous referees and the editor for very useful comments on a first draft.

1. Introduction

The concentration of specific kinds of economic activity in small and confined geographical spaces is a long-lasting theme in the economics, business and geographical literature (cf. Marshall, 1890; Porter, 2000; Dicken and Lloyd, 1990). The development of new knowledge motivated by economic incentives is one of these activities that seem to be geographically clustered in a strong way (e.g., Storper, 1993). The literature on clusters of innovative activity suggests that a knowledge-based theory is needed to explain the more general trend of industry agglomeration (Malmberg and Maskell, 2002), or that the knowledge creating capacity of local clusters is decisive for growth and the long-run survival of the cluster (Bathelt, 2005).

Traditionally, the explanation for the tendency clustering has been sought in the broad concept of agglomeration economies, i.e., the idea that by locating close together, firms may reduce production costs. Agglomeration economies comprise a wide variety of factors (Dicken and Lloyd, 1990), including the availability of raw materials, the availability of a pool of specialised labour, the possibilities for an advanced division of labour, and, specifically for innovation clusters, knowledge spillovers that are bounded by distance (Jaffe et al., 1993). The role of social networks and entrepreneurship has also been stressed as a factor explaining clustering of innovation activities (Sorensen, 2003).

An important topic in the literature on innovation clusters is which specific aspects of the nature of knowledge give rise to clustering. The tacit vs. codified nature of knowledge may explain an important part (Gertler, 2003). The part of knowledge that cannot be codified, i.e., which is tacit, can most easily be transferred over small distances (by face-to-face contact). This gives rise to agglomeration economies in knowledge transfer and knowledge spillovers. The degree of complexity of knowledge has also been suggested as a factor in innovation clustering (Sorensen et al., 2006).

The concentration of innovation activities in confined spaces may also lead to a particular interactive structure between firms and other actors (such as research institutes, universities, regional policymakers including investment agencies) that acts as a system. This introduces the notion of a regional innovation system (e.g., Morgan, 2004; Lau and Lo, 2015). The crucial aspect of these systems is that institutions (in the widest possible definition, i.e., including formal rules of the game as well as informal customs and habits) partly determine the performance of the system. This implies that there is a meso-structure to regional innovation systems, which surpasses the level of decision-making in individual firms.

At an abstract level, what these approaches have in common is that they stress the importance of knowledge flows for the generation of new knowledge. Scientists and engineers that are engaged in the inventive process use existing knowledge to generate new ideas. In other

words, knowledge is cumulative. The combination of the cumulative nature of knowledge and the localised nature of knowledge flows gives rise to the existence of geographical clusters of knowledge production.

But knowledge flows are not exclusively local. This recognition gave rise to the idea that firms use a combination of "local buzz and global pipelines" (Bathelt et al, 2004) for sourcing new knowledge as an input into their own innovation activities. The local buzz that firms find inside the cluster consists of easily accessible knowledge resources that are "just there", like Marshall's (1890) idea that the mysteries of trade are no longer mysteries but are "in the air". Local buzz is the typical factor behind the success of Silicon Valley, where engineers, entrepreneurs and venture capitalists meet in bars and restaurants (Saxenian, 1994). Global pipelines, on the other hand, are consciously constructed and managed gateways to knowledge in different (long distance) localities. Strategic alliances (Owen-Smith and Powel, 2004) have been analysed as a tool for building global pipelines. Bathelt et al. (2004) argue that the co-existence of a high level of local buzz and access to many global pipelines provides the best environment for economic success, both for individual firms, and for cluster regions. Bathelt and Glückler (2011) discuss the broader consequences of local buzz and global pipelines for the geography of innovation.

The literature on innovation clusters, regional innovation systems, or local buzz vs. global pipelines, takes the firm as the key unit for analysing knowledge dynamics. The knowledge that this literature is interested in is economically motivated, and this makes firm decisions a logical starting point for the analysis. The novel contribution of the current paper is that we want to shift this focus from the firm to technology itself. We are not so much interested in the evolution of the firm population that invests in knowledge and that operates in local clusters, but rather in the way that technology itself evolves, and how this evolution interacts with space. The way in which we undertake this endeavour is to draw on an existing literature that conceptualises technological evolution as trajectories, and which quantifies these trajectories by using patent citations data. We will briefly summarise the main ideas in this literature below.

Technological trajectories are accumulated chains of incremental innovations (we use patent data to measure these innovations) that display the dominant long-run developments in technology. Examples of technological trajectories include Moore's law, which defines technological progress in personal computers, or the specific types of internal combustion engines used in motor cars. The idea of technological trajectories is important for analysing the main avenues along which technological change has an impact on the economy and society at large, including the firms that are the subject of the geographical literatures that were briefly referred to above. But when adopting a firm perspective, as much of the literature does, the main directions of technological change remain largely obscure.

The results of our analysis point to the importance of knowledge flows between local innovative clusters, rather than within clusters, for the development of technological trajectories. Although we confirm the strongly concentrated nature of knowledge generation activities, and of patent citations (which are the main unit of observation in our analysis), our results show that the accumulated knowledge flows between rather than within clusters are responsible for the main directions of technological change (the trajectories). This emphasis on between-clusters flows is reminiscent of the idea of global pipelines, but it differs in one crucial aspect. Whereas the global pipelines that firms use to access knowledge from far-away locations are consciously managed and constructed, the trajectories that we find emerge from collective rather than individual action. No single firm exclusively shapes a technological trajectory (see, e.g., the empirical evidence in Verspagen, 2007).

The research here is a novel combination of existing research traditions that have so far not been combined at all. One overlap that can be observed between these literatures is the use of data on patents and patent citations, which is a (small) part of the geographical literature (e.g., Sorensen et al, 2006; Jaffe et al., 1993), and dominates the branch of the trajectories literature that attempt to quantify the main concepts in this tradition (e.g., Verspagen, 2007, Mina et al., 2007; Nomaler and Martinelli, 2014). Another overlap lies in the use of formal network methods, e.g., Owen-Smith and Powell (2004) in the analysis of local buzz and global pipelines, while a specific form of network theory is the bread and butter of quantifying technological trajectories.

Otherwise, the underlying units of analysis (relationships between firms and other organisation on the one hand vs. relationships between inventions on the other hand), and the disciplinary backgrounds (geography, business studies and economics on the one hand vs. economic history and technology analysis on the other hand) are very different between the fields of literature that we try to bridge. Fitting with the early stage of such an ambitious combination of literatures, our analysis will be mainly explorative in nature. We will try to operationalise various concepts and ideas from both literatures and bring them together in an empirical overview of the trends that are observed in our database. We leave a fuller development of truly integrative conceptualisations to a later stage, which we hope will be forthcoming on the basis of the interest in the empirical facts that we provide.

Ultimately, what these empirical facts suggest is that the geographical concentration of patent citations that has been an important topic of the literature so far, is typical of individual small technological steps, while the main directions of technological change (trajectories) that are comprised of many cumulative incremental steps have a much wider spatial reach than is suggested by the analysis of patent citations by economic geographers. The underlying reason for this broader geographical reach lies in between-cluster knowledge flows. The collective

nature of these between-cluster flows adds a new dimension to the understanding of global pipelines that are analysed in a part of the geographical literature.

The rest of the paper is organised as follows. In Section 2, we provide a short overview of the relevant literature on technological trajectories, including a discussion of the indicators used (patents and patent citations). Section 3 introduces the database, and Section 4 presents the methods used. Section 5 is the first one where we present novel empirical results. Here, 35 inventive clusters in the US are presented, which will form the main unit of analysis in the remainder of the paper. Our analysis focuses exclusively on the US, because for this country we have data available that are broken down to the relevant geographical unit (counties). Section 6 looks at the geographical distribution of citations, both direct (which is the usual indicator of knowledge flow between localities), and so-called deep citations, which are our way of identifying technological trajectories. Section 7 looks in-depth at some of these technological trajectories, in this case between the largest of the 35 US inventive clusters. The concluding section 8 discusses how the evidence provided in our analysis has implications for the spatial nature of knowledge flows and knowledge production, and suggests directions for future research.

2. Technological trajectories and patent citations

The novelty of our research lays in the application of the idea of technological trajectories to the idea of innovation clusters, and in particular the quantification of a geographical dimension of technological trajectories. The notion of technological trajectories stems from a number of authors (in particular Dosi, 1982; Sahal, 1981) who analyse the history of technology from a strongly economic perspective. The central idea is that the economic impact of innovation takes place through a combined process of radical breakthroughs, incremental innovations and diffusion. This generates "technological paradigms" and "technological trajectories" (Dosi, 1982). By a technological paradigm, Dosi refers to a "model and pattern of solution of selected technological problems, based on selected principles from the natural science and on selected material technologies." A paradigm is the set of technological opportunities that emerges from a radical breakthrough, such as the application of steam power to industrial processes, or the notion of mechanised calculations based on binary logic.

The paradigm develops along a number of specific trajectories, which are accumulations of incremental improvements of a basic design. These incremental innovations are endogenous reactions to the specific circumstances in which the technology develops. For example, when labour costs are high and an important part of total production costs, the trajectory will likely take a labour-saving nature. This endogeneity of the main technological trends also implies that technological trajectories may lock-in to a particular direction, and ignore technological possibilities that lie further away in technological space (e.g., Arthur, 2014). As an

evolutionary process, technological change does not optimise globally, but adapts to local circumstances.

An example of a technological trajectory is the famous "Moore's Law" that describes the technological development of microprocessors (the law states that the number of transistors in a single integrated circuit doubles every two years). How different technological trajectories co-exist within the same paradigm is illustrated well by the example of steam engines. In one particular environment, trains, a trajectory of lean but powerful high-pressure engines emerged, while in the case of Cornish metal ore mines a completely different trajectory of very large engines with relatively low pressure developed (Nuvolari and Verspagen, 2009). In summary, a technological paradigm is a set of radical breakthroughs that defines developments in the techno-economic domain for the long run, while the technological trajectory adapts the paradigm to local circumstances through a series of cumulative and incremental innovations. Although made up of incremental steps, technological trajectories represent big changes over long periods of time. It is this kind of change that we are interested in here

Verspagen (2007) and Mina et al. (2007), based on Hummon and Doreian (1989) pioneered a method to map technological trajectories using patent citations. Their approach focuses on a small pre-defined field (fuel cells in the case of Verspagen, 2007; and medical technology in the case of Mina et al., 2007), for which it identifies a number of citation paths that capture the largest amount of knowledge flows in the field. The current paper applies the same method, with further developments, to the phenomenon of spatial knowledge flows. Thus, our analysis makes the novel combination of mapping trajectories of knowledge in technological space, with trajectories of knowledge in geographical space. Moreover, instead of analysing the trajectories in a single technology field (which is the norm in the literature), we look at a much larger patent dataset that covers all trajectories in all technology fields in the period under consideration.

The quantitative analysis of technological trajectories is mostly done with patent data, in particular with patent citations. These data are also used in the geographical literature on the concentration of patenting activities. Jaffe et al. (1993) found that in the US, distance is inversely related to the probability that two patents are linked by a citation. They control for a range of factors such as technology class and time, by matching actual citation pairs by pairs of patents that are similar in terms of these other variables, but do not cite each other, and then find that distance across the two patents in the pair is smaller in the group of patent citations. This result was confirmed in many follow-up studies, including other geographical areas, e.g., Maurseth and Verspagen (2002) and Bottazi and Peri (2003) for Europe.

The use of patent statistics in geographical analysis requires careful interpretation. Griliches (1990) discusses many aspects of patents as indicators. Perhaps the most crucial aspect of his discussion is the fact that patents are indicators of invention rather than innovation. They show the technical possibilities, but do not guarantee commercial relevance. In fact, many patents that are granted are not used commercially (Giura et al, 2007). This is more of a limitation for studies that aim at analysing the economic success of innovative firms, than it is for our study. As already explained, our emphasis is on technological trajectories, which represent the main directions of technology. Inventions are the basic unit of these trajectories, and whether or not these inventions are actually commercialised is of secondary importance as compared to a research design where the emphasis is on firms. That individual patents are the incremental steps that accumulate into a trajectory is consistent with the finding by Giura et al. (2007) and Gambardella et al. (2008), who show that most patents have small economic value, and provide small technological steps, with only a few very infrequent outlier patents representing radical change and large economic value.

Griliches (1990) also stresses that the extent to which firms patent their inventions differs greatly between industries. For example, in pharmaceuticals, patents are of crucial importance, because a product that is not patented can easily be imitated, at lower costs, by competitors. In other industries, such as machinery, patents are less important, because competitiveness depends more on factors that are not described in the patent itself. Also, some inventions cannot (or could not) be patented, such as software in some jurisdictions. As a result of this, patents are very common in some industries, and not in others. This does affect our analysis, as the geographical patterns of patented inventions may differ from non-patented innovations.

This literature has also addressed the issue of whether patent citations can actually be seen as a measure of spillovers or knowledge flows. This is, again, mostly relevant for studies that take the firm (or innovative efforts by the firm) as the main object of study. In many of those studies, for example when analysing the nature and causes of clustering of innovative efforts, the factor of interest are the flows that firms receive, and use to generate new knowledge. The implicit assumption is that patent citations indicate such flows, from the patent that is cited to the patent that is citing. The fact that many citations are added by patent examiners (instead of by the inventor herself), may be a reason in itself why the citation does not indicate an actual technology flow, as it seems to suggest that the inventor of the citing patent did not know the cited patent. This has been investigated, for example, by Thompson (2006) and Criscuolo and Verspagen (2008). The conclusion seems to be that although citations are noisy indicators, the conclusion of a geographical bias in knowledge flows stands even when only inventor-citations are used.

For our purposes, when analysing technological trajectories, patent citations will be used as indicators of technological relatedness. The citing patent is related to the cited patent, even if the inventor did not make the citation, because it is the job of the patent examiner to judge the novelty of the patent, and citations are used for that purpose. It is this kind of technological relatedness that the trajectory approach uses to map the main directions of technological developments. We will use the term technology flows for this, even though we do not imply that a flow between two firms (or inventors) has necessarily taken place.

3. Data

We use the OECD REGPAT database, which contains patent-level information, with geographical information about the inventors and applicants of the patent. The patents in the REGPAT database are patents issued by the European Patent Office (EPO), or filed under the so-called Patent Cooperation Treaty (PCT), which allows just one application at one of the participating offices, and get patents in multiple jurisdictions. We focus on the US, in particular the part of the US located on the main North American continent, with the exception of Alaska. The geographical entities used in the REGPAT database are always administrative regions. In the US, the regions are counties. In other nations, the geographical entity tends to be much larger, which is why we focus on the US.

We generally count patents in a fractional way, i.e., when there is more than one inventor, the patent is assigned to all geographical entities (regions) that the inventors come from, using weights that are proportional to the number of times a region appears on the inventor list. The same fractional procedure is applied to citations, which have a citing and a cited inventor list.

We divide the total period for which we have data, which is 1973 – 2012, into sub periods of 3 years, and focus on the time span that starts with 1986-88 and ends with 2004-06. The EPO only started in 1979, and had relatively few patent applications in the early years (before 1986). From 2007-09, the number of patents declines due to a backlog in processing (our dates are priority dates, as close as possible to the date of invention). These trends, and other basic information about the data will be illustrated and discussed below in Figure 1.

4. Methods

We now proceed to summarise the methods. Many of the details are left for the annex.

4.1. Identifying clusters

The first step in our methodology is a workable definition to identify inventive clusters as the main geographical unit of our analysis. The geography literature, taking the firm as the unit of analysis, remains close to Porter's (2000, p. 16) definition of a cluster as "… a geographically proximate group of interconnected companies and associated institutions in a particular field, linked by commonalities and complementarities." Because we focus on patented inventions

instead of firms, the notions of interconnectedness, commonalities and complementarities are substituted by the intensity of knowledge flows, proxied by the number of citations. We define a cluster as a set of geographically close counties among which knowledge flows are particularly strong. The emphasis on knowledge flows in our procedure for identifying clusters not only stems from the common notion that knowledge is an input for producing new knowledge, but also from the idea of technological relatedness. Knowledge flows only between related inventions, and hence we also capture the presence of related activities in the cluster. High inventive activity is a necessary but not sufficient condition for strong knowledge flows.

The cluster identification procedure is motivated by a broad analogy to the idea of metropolitan spaces, in which flows of commuter movements are concentrated (in our case, knowledge flows represent the commuter movements). The clusters are groups of spatially contiguous counties with high patenting activity and intensive knowledge flows between them. We look at citations and patents in the period 2001-06. The citation matrix, which is directed, gives the number of citations between the US counties. We standardise all cells in this matrix as follows: $cs_{ij} = \left(c_{ij}/(p_i p_j)\right)/(c/p^2)$, where c is the number of citations, p the number of patents, the subscripts i and j indicate counties, and the absence of a subscript indicates an aggregation of counties.

The standardisation expresses the number of citations between a pair of counties relative to its expected value, if citations were completely random (within the US). In this context, "random citations" means that the relative frequency of citations between a pair of counties is equal to the product of the shares of patents in the two counties. In other words, the more patents there are on either side of the citations relationship, the higher the number of expected citations is. A value lower (higher) than 1 for the standardised citations number indicates less (more) citations than expected randomly. We then binarise the matrix by setting all cells that are larger than 1 to 1, and all other cells to 0. We further thin out the number of ones in this matrix by setting to 0 all cells for which the original number of citations was less than 1, and all cells for which the patents for either the row-county or the column-county is less than 1. The latter part of the procedure is intended to ignore all counties that have very little inventive activity.

In this matrix, we check the geography of all cells with a value of 1. If the row- and column counties for such a cell share a border (we use queen contiguity), we put these counties on a preliminary list. We plot the counties on the list on a map, and check for contiguous areas, which are defined as the clusters. There are 30 clusters that appear in this way, of which 16 have two counties, and the largest cluster has 30 counties. We increase the number of clusters to 35 by merging two clusters that are separated by a sea border, and by including a number of

single counties that have an exceptionally large number of patents, and high internal knowledge flows. This is described in detail in the annex.

4.2. Deep citations

The next step in the methods is to associate the clusters, or geographical space in general, to technological trajectories. As technological trajectories are defined as citation chains (this will be explained in more detail below), we need a way to associate these chains to geography. For this we will use what we call deep citations. A very simple example is where patent A in region 1 is cited by patent B in region 2, which is in turn cited by patent C in region 3. The knowledge flow is then A(1) → B(2) → C(3). In terms of the start- and endpoint of this example path, we see a "deep" knowledge flow from region 1 to region 3. However, because patents usually cite more than one other patent, we need a way to conceptualise the complex networks that arise in the real world.

We implement the deep citations idea by looking at all patents in the last 3 years of the period that we used to construct the clusters (2004-06), and chart their "ancestry" in terms of patents from 1986-88. We use a calculation that was pioneered by Martinelli and Nomaler (2014), and that is akin to genealogy. It considers cited patents as the "parents" of the citing patent, and attributes a "parenthood" share of $1/n$ to each cited patent, where n is the total number of citations made by the "child." By multiplying the direct ancestry shares of different generations, we trace the ancestry across generations. For example, in the genealogy of human reproduction, the share of each parent would be 1/2, the share of each grandparent would be 1/4, and the share of each great-grandparent 1/8, etc. For the patent case we continue tracing back generations as long as we have not yet reached a patent in the period 1986-88. Obviously, this uses both long and short citation paths, including direct citations between the two periods (i.e., paths of length 2, which do exist but are rare). Also, not all patents from 2004-06 have ancestry in 1986-88, and we simply ignore those patents that do not have this.

4.3. Technological trajectories in sandwich networks

The deep citations represent all citation paths that exist in geographical space, but technological trajectories are selective important pathways from this large set. The idea is that only those deep citation chains that embody the highest amount of knowledge flows are representative of the main trends in "technology space", i.e., the trajectories.

Our method for identifying the trajectories is based on the methods proposed by Hummon and Doreian (1989), Verspagen (2007) and Liu and Lin (2012). The patent citation network is directed (knowledge flows from the cited to the citing patent), and also a-cyclical (starting at one node of the network, a path can never return to that node). Two classes of nodes (patents)

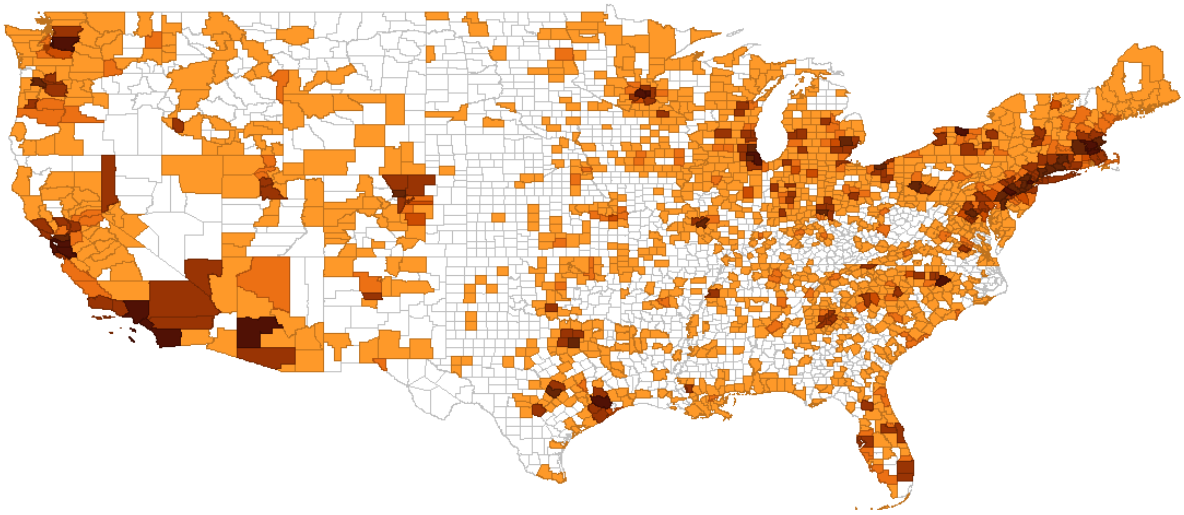
are of particular interest. A start-point is a patent that is cited, but does not cite any patents. An endpoint is a patent in 2004-06 that cites other patents, but is not cited itself.

For every citation in the network we calculate the so-called SPNP (Search Path Node Pair) indicator proposed by Hummon and Doreian. SPNP for the citation of patent p in patent q is counted as follows. First, count all patents in the network for which a path to p exists (including p itself). Then count all patents that can be reached from patent q (including q itself). SPNP is the multiplication of these two counts. It is the number of pairs that can be formed by the patents "upstream" and "downstream" the citation. Next, we identify for every start-point in the network the path (ultimately leading to an endpoint) that maximises the multiplication (sum of logs) of the SPNP values along the path. Such a path is called a main path.

While we will look at deep citations (the "prequel" to trajectories) at the scale of the entire US inventive space, the specific attributes of trajectories (main paths) are hard to summarise for the entire space, or even the entire list of 35 clusters. Therefore, we look at pairs of the clusters. We start with the full set of patents that forms the entire network of deep citation paths between two specific clusters, and find the main paths (trajectories) in this network. The technological trajectories that we find are therefore true geographical trajectories, as they run between two spatial units. For each pair of spatial clusters we extract the set of citations that connect the patents in period 2004-06 in the "to" cluster to patents in 1986-88 in the "from" cluster. This large network, which in-between the start- and endpoints of the citation paths contains many patents not invented in either the to- or from-cluster, is what we call the sandwich network for the specific pair of clusters that we are considering.

5. The US inventive landscape

The starting point of our analysis is the characterisation of the US patenting geography by spatial clusters, which embody the peaks in the technological landscape. Map 1 displays the number of patents per county in 2001-06. White areas indicate zero patents, for coloured areas the number of patents increases (roughly exponentially) with darkness. The clusters, identified by the procedure explained above, are displayed in Map 2. Table 1 lists the clusters and gives summary statistics.

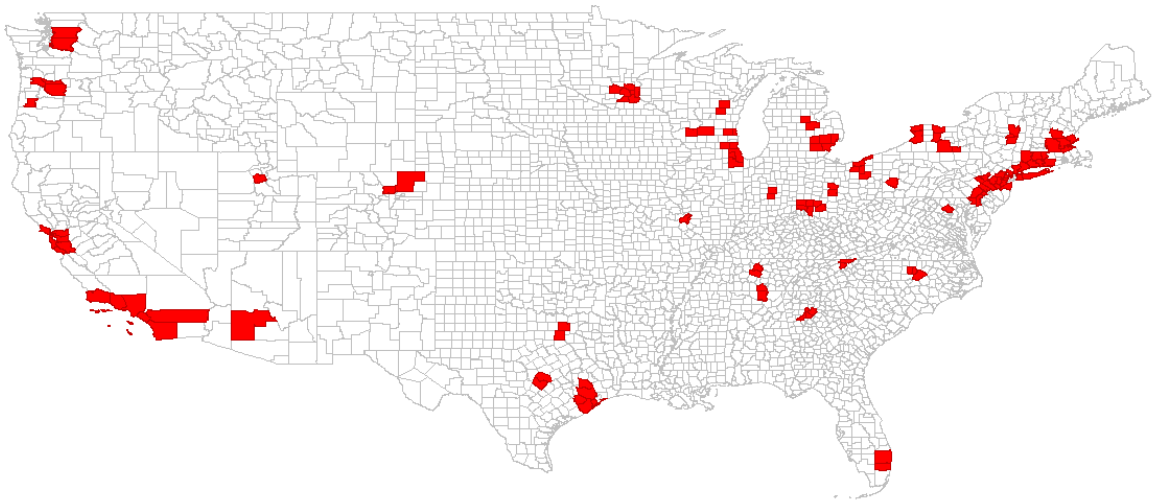


Map 1. Number of patents per county, 2001-03 and 2004-06 (white areas indicate zero patents, for coloured areas number of patents increases with darkness)

Comparing Map 2 and Map 1, we see that there is a large correlation between the number of patents in a county and cluster membership: almost all of the dark areas in Map 1 are also marked in Map 2. Note that this correlation is in no way obvious, as the cluster identification is primarily based on citation intensity, not number of patents (in fact, the number of patents is penalised, as the benchmarking of citation intensity divides by it). The number of patents in the table (2001-06 period) is distributed rather unequally over the 35 clusters. The largest cluster, i.e., the East coast, holds about 26% of the total, the smallest cluster (Nashville) just under 0.1%. The three largest regions account for just over half of the patents, the bottom-10 clusters for just over 4%. Thus, in terms of the sheer numbers, the inventive landscape in the US is very peaked, with a small number of leading clusters with very high inventive activity, a larger group of followers, and the far majority of counties (outside the clusters) contributing very little.

Main city/geographical name	Number of counties	Number of patents
East coast (Boston, New York City & Philadelphia)	34	25348
San Francisco	7	13862
Los Angeles & San Diego	6	11635
Phoenix	1	5649
Minneapolis	7	4331
Chicago	5	3547
Seattle	2	3534
Houston	5	2485
Detroit	5	2394
Durham/Chapel Hill	3	2332
Cincinnati	6	2010
Rochester	4	1743
Cleveland	5	1583
Dallas	2	1268
Atlanta	3	1250
Portland	3	1137
Austin	2	998
Montgomery/Washington	1	991
Pittsburgh	1	959
Madison, Wisconsin	2	930
Fort Lauderdale & Palm Beach	2	877
Boulder	2	845
Buffalo	2	815
Indianapolis	2	796
St. Louis	2	739
Benton	1	643
Kingsport	2	597
Salt Lake City	1	494
Schenectady &Albany	3	490
Milwaukee	2	451
Columbus	2	443
Appleton	2	390
Saginaw/Midland	2	356
Huntsville	2	226
Nashville	2	99

Table 1. Clusters and summary statistics



Map 2. 35 clusters, 2001-03 and 2004-06

6. The geography of direct and deep citations

By definition, patent citations are concentrated in the 35 clusters that were identified above. Thus, it is not surprising that direct citations are concentrated within these clusters. However, this is much less obvious for the deep citations (the stuff that trajectories are made of). Because of the indirect linkages that they embody, we may expect that the distribution of the deep citation chains is much less concentrated within single clusters than the direct citations.

As we are interested in comparing distributions of direct and deep citations, we need a benchmark to judge their geographical concentration. Like we did in the procedure that identified the 35 clusters, we construct such a benchmark by referring to the idea of random citations. In this case, it simply means that we break down the total number of cited patents in a specific cluster into the following categories: (i) from within the cluster itself, (ii) from other clusters (between), (iii) from US counties that are not part of a cluster, and (iv) non-US countries. The random chance that each of those categories is cited is simply their share in total patents, and this is our benchmark.

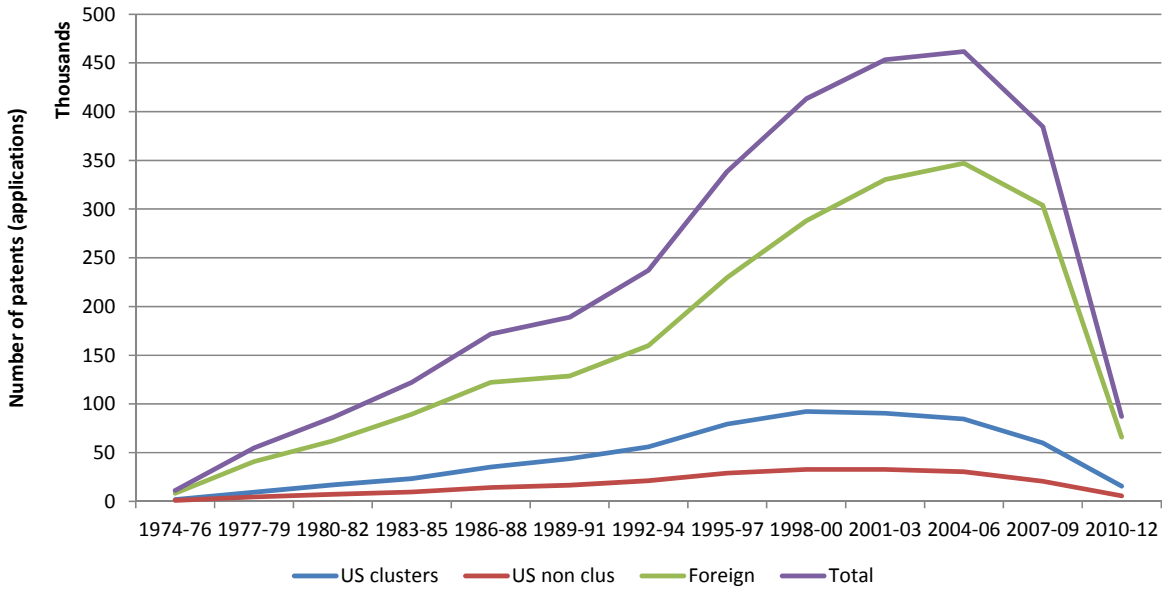


Figure 1. Number of patents in the database

Figure 1 displays the trends in the number of patents for the sub groups of the US cluster regions, the other US counties, and non-US patents. The figure illustrates how the benchmark matters. For example, the number of foreign patents is much larger than that of US patents (the share varies between 67 and 80% over the entire period), and hence the benchmark expects that foreign patents are the largest category in citations made by US patents. Within the US, the 35 clusters always take the majority of patents: 24% of the total in 1992-94 and 18% in 2004-06. In interpreting the trends in Figure 1, remember that the identification of the 35 US clusters was based on patents and citations in the sub period 2001-06 and that deep citations go back to the period 1986-88.

Figure 2 compares the distribution of direct citations and deep citations. The figure looks at citations made by (knowledge flowing into) the group of 35 US clusters in the period 2004-06. The top panel shows the distribution of these citations over the four categories. The sum of each colour of the bars is one.

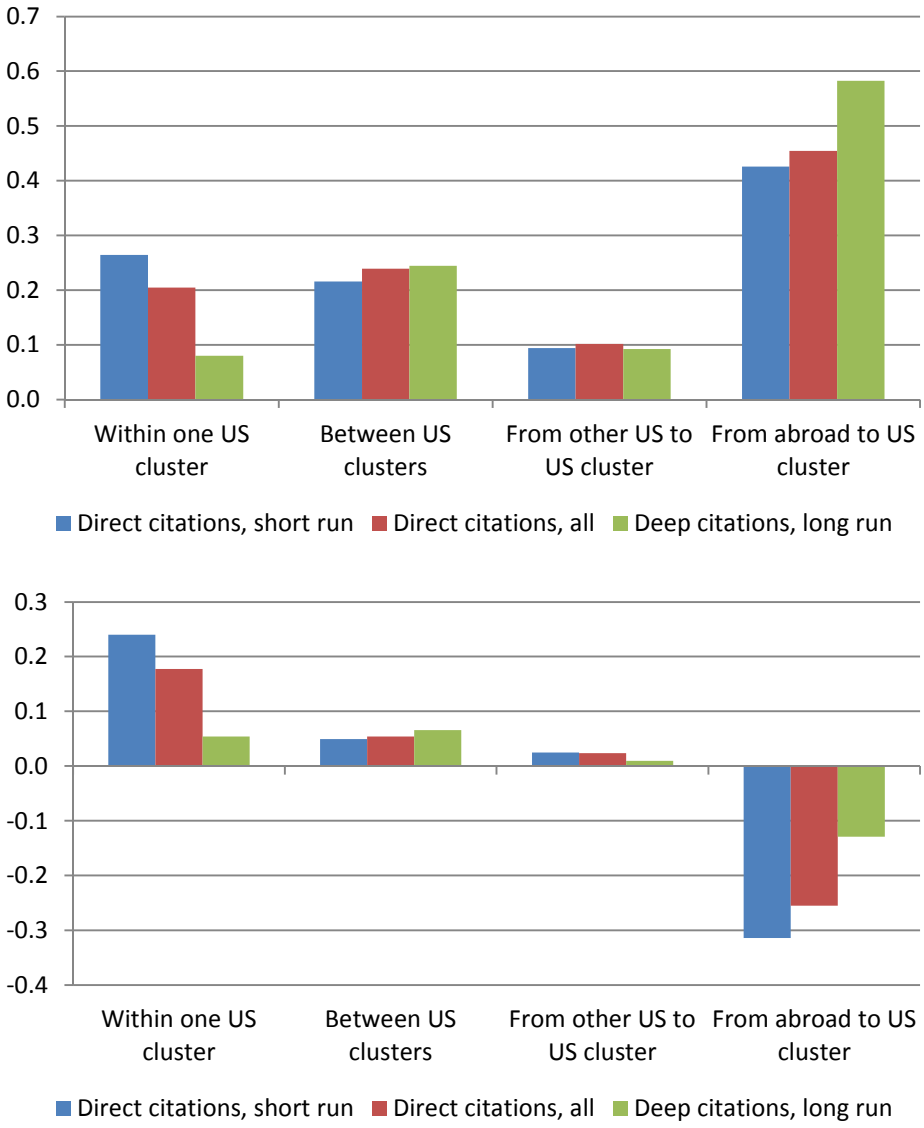


Figure 2. The distribution of direct and deep citations (top panel gives share of all citations, bottom panel is relative to the benchmark)

Direct citations in the short run are citations made in 2001-06, where both the citing and cited patent are from this period. This the same set of citations that were used to identify the clusters and hence we know that they are strongly spatially concentrated. The category of all direct citations includes the entire period 1986-2006, and hence also includes citations over longer time spans than the first category. We see in the top panel that all types of citations are clearly dominated by foreign sources, followed by US clusters (between), and the cluster itself (within). Other US regions are the smallest category.

The bottom panel of the figure invokes the benchmark, by subtracting from the values in the top panel the expected share of citations.[2] This clearly brings out the differences between the

[2] For deep citations, we also use the benchmark constructed on the basis of citations in the 1986-1988 period.

types of citations. Direct short-run citations are very much biased to the own cluster (within), much less so but still positively towards other US clusters (between) and US non-cluster regions, and negatively biased towards foreign patents. If we include longer-run direct citations, the bias for the own cluster remains but is weaker, while that of other US clusters increases slightly, and the bias to foreign patents become much less negative. The deep citations re-enforce this trend. Here the bias to the own cluster is still positive, but smallest among the three citation types. The bias towards foreign countries is still negative, but the absolute value is again smallest among the three types. However, the bias towards other US clusters (between) is largest and positive.

In conclusion, looking at deep citations, we find that these have a broader geographical spread than direct knowledge flows. Although deep citations are still biased to within-cluster flows, they are less so than direct citations. On the other hand, deep citations are also biased, and in a stronger way than direct citations towards flows between clusters. The technological trajectories that we are after are a subset of the deep citation chains analysed in this section, thus we conclude that the argument of strongly localised knowledge spillovers seems much less relevant for the case of technological trajectories than it is for the incremental changes associated with direct citations.

7. Technological trajectories

We now make the final step of our analysis and extract the actual technological trajectories from the entire set of deep citation chains. Because the category of flows between and within US clusters stands out in magnitude (Figure 2), we focus on this particular category. Results so far only consider the start- and endpoints of the deep citation chains. We now also look at the intermediate nodes on the deep citation chains, to be able to assess which are the main trajectories that cumulative knowledge development takes, and how these trajectories unfold in geographical space.

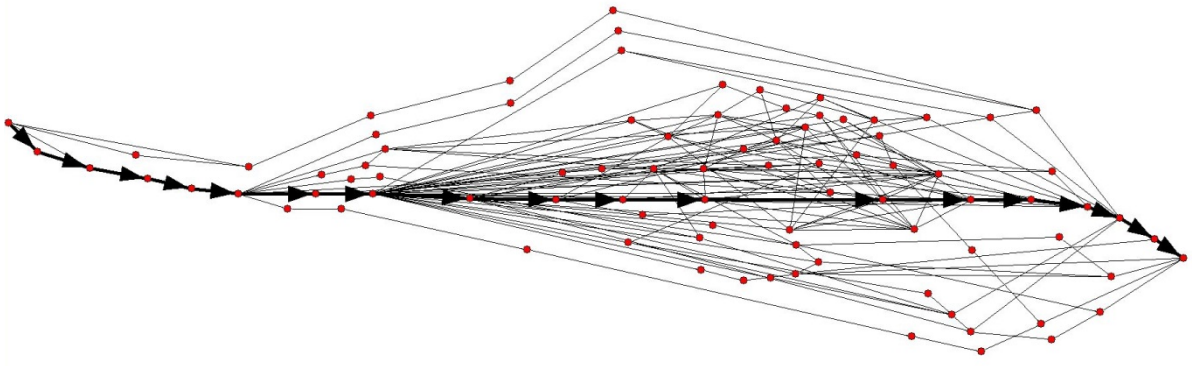


Figure 3. Top main path in the East coast-to-East coast sandwich network (nodes are patents, lines are citations; main path is indicated with arrows and bold lines)

Figure 3 shows an example of a main path (trajectory), in this case the one that was identified within the East coast-to-East coast sandwich as the main path with maximum multiplicative SPNP. The figure illustrates how the complexity of the underlying network of patent citations can be reduced to a main path or trajectory that highlights the main trends in technology space. The network in the figure is the network of all patents (and citations) in the sandwich that links the start- and endpoint of this main path. Many of the patents in this network, including those on the main path, are patents from a different geographical location than the East coast cluster itself. The main path itself, which consists of 19 patents, contains a 50% share of European inventors, a 30% share of inventors from the 35 US cluster regions, and a 10% share of inventors from US non-cluster regions.

The particular main path in the figure is a trajectory in pharmaceuticals. It starts with four patents on drugs against hypertension and heart disease, then three patents that broaden the range of diseases (including HIV, depression, migraine and psychosis), and finishes with a range of patents on anti-cancer drugs, in particular protein kinase inhibitors. The trajectory has many different firms, including big pharmaceutical firms such as DuPont, Merck, and Glaxo, specific gentech firms (AmGen), and smaller pharmaceutical firms (such as the Danish Neurosearch).

The sandwich networks are usually large networks, and hence we cannot look at all 1,225 pairs formed by the 35 clusters. Instead, as a sort of case study, we focus on the 16 pairs between the largest four clusters: East coast, San Francisco, Los Angeles/San Diego and Seattle. The 16 sandwich networks together comprise 23,060 main paths. We characterise these main paths by the share of patents from each of a set of 39 geographical entities (35 US clusters; US non-cluster regions; Europe; Japan; and rest of the world). The 35 US clusters together are the largest category, with a share slightly above one third. Europe follows in second place, with a share just under one third. The other three categories are markedly smaller, with US non-cluster counties as the largest (11%) of these three small categories.

In order to interpret the variety in the composition of the main paths in the sandwich networks, we undertake a cluster analysis, which classifies the 23,060 main paths into a small amount of groups. We use k-means cluster analysis, and settle for a division into 8 groups. Choosing 8 groups (k = 8) is somewhat arbitrary, but we considered alternative groups (k = 2..10). Less than 8 groups forces some of the interesting between-group heterogeneity that will be discussed below into a single group, while k = 9 of 10 does not add qualitatively different insights on the role of US cluster regions, which will be the main focus of our discussion. We have one large group (about one third of all main paths), which is group 2. Together with the next two largest groups (1 and 4), this group covers almost two thirds of all main paths. The remaining 5 groups vary between 9% and 5% of the total.

Figure 4 characterises the groups in terms of their composition. Remember that the full detail of the 35 US clusters, plus the 4 other categories, was used to classify the main paths (i.e., to form the groups). Figure 4 displays the mean scores in each of the 5 categories in the group, with the sample average subtracted. Hence a positive value indicates specialisation into this particular category.

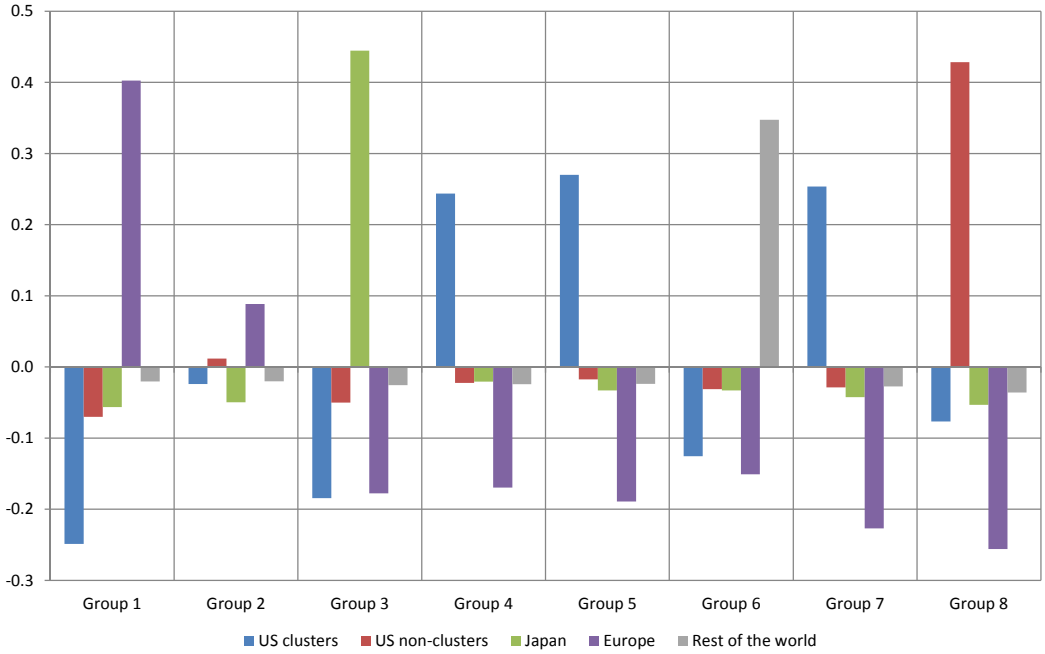


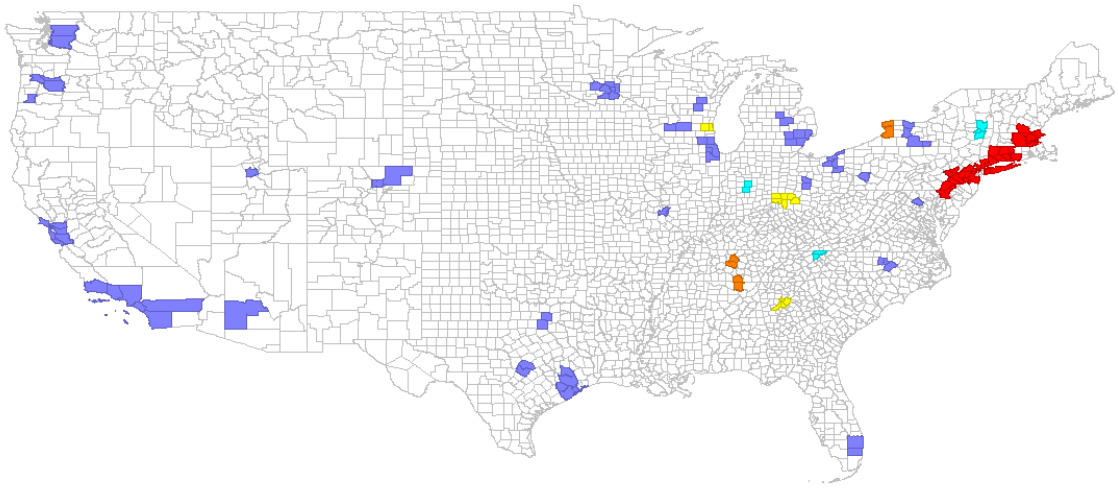
Figure 4. Group characteristics in the cluster analysis (deviations from total sample means)

Group 2 is the one that has the most homogenous distribution over the 5 categories. It is most clearly specialised in European contributions to the main paths, but much less so than group 1. It is also slightly specialised in US non-cluster regions, but much less so than group 8. The other groups, including groups 1 and 8, are specialised in only a single category. In group 1, which is one of the larger groups, this is Europe. In group 3, a small group, we have dominance of Japanese patents, in group 6 (also small) dominance of patents from the rest of the world. Groups 4, 5 and 7 are dominated by US cluster regions. Of this, group 4 is relatively large, the other two are small. Group 8 is dominated by US non-cluster counties.

The three groups that are specialised in US cluster regions (4, 5 and 7) can further be described by looking at which clusters play a role. We define an indicator that is akin to the revealed comparative advantage indicator to investigate this. It is defined as $\boldsymbol{R_{ij} = 100 \times S_{ij}/S_i}$, where *i* (1-35) indicates a cluster region, *j* (4, 5 or 7) is a group from the cluster analysis, $\boldsymbol{S_{ij}}$ is the share of cluster *i* in the total contributions of the 35 clusters to the main paths in the group *j*, and $\boldsymbol{S_i}$ is the share of cluster *i* in the contribution of the 35 clusters to all

(23,060) main paths. A value above (below) 100 indicates a relatively high (low) contribution of the cluster to the group.

Map 3 displays the profile (R values) of group 4, which is the largest of the three US cluster-based groups. The blue colour in the map indicates clusters that have a strongly below average contribution to this group ($R < 75$), cyan indicates $75 < R < 100$, i.e., a mildly below average contribution. The other colours indicate $R > 100$, with yellow $100 < R < 150$, orange $150 < R < 200$ and red $R > 200$.



Map 3. Specialisation profile of group 4

The map shows that group 4 is strongly influenced by the large East coast cluster. It is the only red cluster on the map. There are only 6 other clusters with $R > 100$ in this group, based in the East side of the country. The entire West coast and most of the Midwest and South have $R < 100$. We can conclude that this is a rather specialised group, in which main paths are strongly concentrated in relatively few clusters. Besides the East coast clusters, the clusters in this group with $R > 100$ are small clusters, i.e., they are low on the list in Table 1. The yellow clusters are Durham/Chapel Hill (ranked 10 in the table), Atlanta (rank 15), and Milwaukee (rank 30). The orange clusters are Buffalo (rank 23), Huntsville (rank 34) and Nashville (rank 35). It seems that these smaller clusters are strongly dependent on the large East coast cluster in terms of being present on the main paths that connect our large clusters in the US inventive system.

Similar maps for groups 5 and 7 confirm the impression from Map 3. Thus, the evidence suggests that these main paths themselves are clustered, instead of consisting of evenly distributed geographical patterns. The main trajectories of knowledge development are like selective rivers unfolding in space, rather than like the wind that blows broadly in a wide circle. The clusters of the US inventive landscape each play a particular role in these deep rivers of main technological trends.

8. Discussion and conclusions

We looked at the geography of technological change, taking a technology perspective rather than the usual perspective of organisations (firms). We use patents and patent citations as the indicator of technological change. Previous findings in the geographical literature have focused primarily on the strong geographical concentration of patent citations (Jaffe et al., 1993), which is in line with the idea that innovation takes place mostly in spatial clusters (cf. Malmberg and Maskell, 2002). Our analysis confirms this strong geographical concentration of patent citations, but in our technological perspective, we interpret it as applying mainly to incremental steps in technological space.

The more substantial part of our analysis looked at accumulated sequences of these incremental direct citations. We call these long-run citation chains deep citations, and we find that these deep citations are especially strong between US incentive clusters, even though within-cluster flows are also important in deep citations (but less so than between-cluster flows). The deep citation chains, by their accumulation of incremental steps, also cover longer technological distances, i.e., they represent cumulative change along which technology takes big steps in the long run. We use methods from a separate literature that uses network theory to map the major trajectories in technological space as the citation chains that attract most knowledge flows in the network of direct citations. By identifying those deep citation chains – the main paths – that capture the largest flows of knowledge, we are able to focus on the particular trajectories of knowledge development that embody the strongest long-run forces of technological change.

Our main finding is that technological trajectories develop between innovation clusters rather than exclusively within clusters. This is akin to the idea that firms in clusters construct and manage global pipelines for knowledge transfer in addition to the local buzz that they find in their own cluster (Bathelt et al., 2004). But contrary to the global pipelines, technological trajectories are a result of collective efforts of firms (and other organisations), and they develop in an evolutionary way as accumulated local change, without a top-down design process. In such an evolutionary process, the forces of change do not likely lead to an outcome that is fully optimal in the sense that economic theory usually assumes. Instead, we may see situations in which firms and other organisations are adapted to local circumstances, including the strategies of their competitors. This may well lead to a situation of lock-in, in which inventors in a particular combination of local clusters jointly focus on a particular technological direction, whereas technological solutions may also be found in other parts of technological space.

The combination between the regional cluster literature and the technological trajectories literature is a novel one. Our results suggest that the geographical dimension is important for the construction of technological trajectories. When a technological trajectory develops, it does so along a specific spatial trajectory. This spatial trajectory mostly consists of chains of patents from inventive clusters. In other words, the results from the technological cluster literature have a strong relevance for the analysis of technological trajectories. The way in which firms use local buzz and build global pipelines will have a strong influence on how trajectories develop. On the other hand, the development of the technological trajectories that the firm is interested in will also determine how it builds its global pipelines of knowledge transfer, and where it will seek local buzz.

This raises questions about firm behaviour in the field of technological choice. Which kind of technological resources do firms seek locally, and which ones are sources through global pipelines? Can we even generalise about the answers to these questions, or do the answers differ between geographical locations (clusters)? In general, we suggest that theoretical and empirical work on the geography of innovation takes into account the idea of technological trajectories to develop a more coherent framework for understanding the geographical dimension of technology. For example, the narrative often goes that many of the important trends in ICT come from Silicon Valley, but our results suggest that these trajectories have a much wider geographical base, and are influenced by the global location decisions of firms.

Such a combination of theories of geography and technology may also yield new insights into theorising about technology itself (e.g., Arthur, 2014). It is likely that technological specialisation will play a large role in such a theoretical framework. Because of computational constraints, we have been unable to investigate the role of this factor. However, it is clear that knowledge flows are dependent on technical relations between technology fields, and hence the technological specialisation pattern of a cluster will determine from where it can receive the major parts of its technology inflows.

The technological profile of an innovation cluster is, however, not an exogenous factor. Instead, it emerges historically as a result of interaction between local and external actors (inventors, firms, research institutes), and the progress of technological trajectories. Therefore, the technological specialisation pattern of inventive clusters will never be the ultimate explanation for the spatial concentration of deep citations. It is a factor that needs to be explained itself, interwoven with the explanation of the concentration of deep citations itself.

Reasoning in the tradition of the innovation clusters literature and the technological trajectories literature would suggest, in our view, that path dependency and lock-in play a role in this process. Innovation clusters depend on interaction (locally and over longer distances)

that re-enforces itself by repetition and adaptation. Actors in a local cluster develop routines that become highly specific to their situation. By evolutionary selection, these routines are optimised to become a local fitness maximum, which is specific to the local cluster. Knowledge exchange between local actors and with a selective number of actors outside the local cluster becomes a crucial part of the inputs into the cluster. Obviously, this is a dynamic process, in which change is the norm rather than static equilibrium. Similarly, progress along technological trajectories is cumulative and path dependent. Our analysis suggests that the explanation of regional innovation clusters, their interaction, and of technological trajectories may benefit from more theoretical and empirical linkages between the regional innovation literature and the idea of technological trajectories.

References

Arthur, W.B. (2014) Complexity and the Economy, Oxford: Oxford University Press.

Bathelt, H. (2005) Geographies of production: Growth regimes in spatial perspective 2 – Knowledge creation and growth in clusters. Progress in Human Geography, 29: 204-16.

Bathelt, H., Malmberg, A., Maskell, P. (2004) Clusters and knowledge: Local buzz, global pipelines and the process of knowledge creation. Progress in Human Geography, 28: 31-56.

Bathelt, H., Glückler, J. (2001) The Relational Economy: Geographies of Knowing and Learning, Oxford: Oxford University Press.

Bettencourt, L.M.A., Lobo, J., Strumsky, D., (2007) Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size. Research Policy, 36: 107–120.

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

Bottazzi, L., Peri, G. (2003) Innovation and spillovers in regions: evidence from European patent data. European Economic Review, 47: 687–710.

Criscuolo, P., Verspagen, B. (2008) Does it matter where patent citations come from? Inventor vs. examiner citations in European patent. Research Policy, 37: 1892-1908.

Dicken, P., Lloyd, P.E. (1990) Location in space: Theoretical perspectives in economic geography. Third edition, New York: Harper and Row.

Dosi, G. (1982) Technological paradigms and technological trajectories. Research Policy, 11: 147–162.

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

Gambardella, A., Harhoff, D., Verspagen, B. (2008) The value of European patents. European Management Review, 5: 69-84.

Gertler, M. (2003) Tacit knowledge and the economic geography of context, or the undefinable tacitness of being (there). Journal of Economic Geography, 3: 75-99

Giura, P., Mariani, M., Brusoni, S., Crespi, G. Francoz, D., Gambardella, A., Garcia-Fontes, W., Geuna, A., Gonzales, R., Harhoff, D. Hoisl, K., Le Bas, C., Luzzi, A., Magazzini, L. Nesta, L., Nomaler, Ö., Palomeras, N., Patel, P., Romanelli, M., Verspagen, B. (2007) Inventors and invention processes in Europe: Results from the PatVal-EU survey. Research Policy, 36: 1107-1127.

Griliches, Z. (1990) Patent statistics as economic indicators: A survey. Journal of Economic Literature, 28: 1661-707.

Hummon, N.P., Doreian, P. (1989) Connectivity in a citation network: The development of DNA theory. Social Networks, 11: 39-63

Lau, A. K.W., Lo, W. (2015) Regional innovation system, absorptive capacity and innovation performance: An empirical study. Technological Forecasting and Social Change, 92: 99-114.

Liu, J. S., Lu, L. Y. Y. (2012) An Integrated Approach for Main Path Analysis: Development of the Hirsch Index as an Example. Journal of the American Society for Information Science and Technology, 63:528-542.

Malmberg, A., Maskell, P. (2002) The elusive concept of localization economies: Towards a knowledge-based theory of spatial clustering. Environment and Planning A, 34: 429-49.

Mina, A., Ramlogan, R., Tampubolon, G., Metcalfe, J.S. (2007) Mapping evolutionary trajectories: Applications to the growth and transformation of medical knowledge. Research Policy, 36: 789-806.

Martinelli, A., Nomaler, Ö. (2014) Measuring knowledge persistence: a genetic approach to patent citation networks. Journal of Evolutionary Economics, 24: 623-652.

Marshall, A. (1890), Principles of Economics, London: MacMillan.

Maurseth, P.-B., Verspagen, B. (2002) Knowledge Spillovers in Europe. A Patent Citations Analysis. Scandinavian Journal of Economics, 104: 531-545.

Morgan, K. (2004) The exaggerated death of geography: localized learning, innovation and uneven development. Journal of Economic Geography, 4: pp. 3-21.

Nuvolari, A. Verspagen, B. (2009) Technical choice, innovation, and British steam engineering, 1800–50. Economic History Review, 62: 685-710.

Porter, M. (2000) Location, Competition, and Economic Development: Local Clusters in a Global Economy. Economic Development Quarterly, 14: 15-34.

Sahal, D. (1981) Patterns of Technological Innovation (Addison-Wesley).

Saxenian, A. (1994) Regional advantage, culture and competition in Silicon Valley and Route 128, Cambridge MA: Harvard University Press.

Sorenson, O. (2003) Social networks and industrial geography. Journal of Evolutionary Economics, 13: 513-27.

Sorenson, O., Rivkin, J. W., Fleming, L. (2006) Complexity, networks and knowledge flow. Research Policy, 35: 994-1017.

Storper, M. (1993) Regional worlds of production: Learning and innovation in the technology districts of France, Italy and the USA. Regional Studies, 27: 433-455.

Thompson, P. (2006) Patent citations and the geography of knowledge spillovers: evidence from inventor – and examiner – added citations. Review of Economics and Statistics, 88: 383–388.

Verspagen, B. (2007) Mapping Technological Trajectories as Patent Citation Networks: a Study on the History of Fuel Cell Research, Advances in Complex Systems, vol. 10: 93-115.

Annex A1. Method of defining the clusters and related descriptive information on citation flows

Clusters are defined on the basis of direct citations only between patents of non-zero US identity. We count patents and citations fractionally. Consider a patent *k*, applied for at time cohort *t2*, having M_k inventors out of which $N_k^i \leq M_k$ report addresses that belong to US county R^i. This patent is considered as an $\frac{N_k^i}{M_k}$ patent of county R_{t2}^i. Further assume that this patent cites S_k other patents, one of which is a patent applied for at time cohort *t1* having M_m inventors out of which $N_m^j \leq M_m$ report addresses that belong to US county R^j. This citation adds

$$ {}_{m}^{k}\boldsymbol{df}_{i,j}^{t1,t2} = \frac{1}{S_k}\frac{N_k^i}{M_k}\frac{N_m^j}{M_m} \qquad \text{EQ 1} $$

(fractional) units of direct knowledge flow from county R^j of time cohort *t1* to county R^i of time cohort *t2*. Accordingly, the direct total knowledge flow (i.e., fractionally-counted citations) from county R^j of time cohort *t1* to county R^i of time cohort *t2* is

$$ \boldsymbol{c}_{i,j}^{t1,t2} = \sum_{\forall m,k}({}_{m}^{k}\boldsymbol{df}_{i,j}^{t1,t2}), \qquad \text{EQ 2} $$

over all patents *k* applied for at time cohort *t2*, and all patents *m* applied for at time cohort *t1*.

A stylized illustration is based on the graph representation a hypothetical citation network depicted in Figure A2.1. This toy citation network has 14 patents belonging to various time cohorts (T1, T2, T3 and T4) and various counties (C1, C2 and C3). For illustrative convenience, we assume in this stylized example that a patent belongs 100% to a single county. The network nodes that represent the patents are labeled in a way that identifies the time cohort of application, the county of origin and the patent number. For example the label T2_C3_P7 indicates patent number 7 which was applied for at time cohort number 2 and invented by someone from county number 3. The nodes are organized on the graph diagram in such a way that patents of the same time cohort are vertically aligned, and the vertical position is inversely related to the index number of the county of origin.

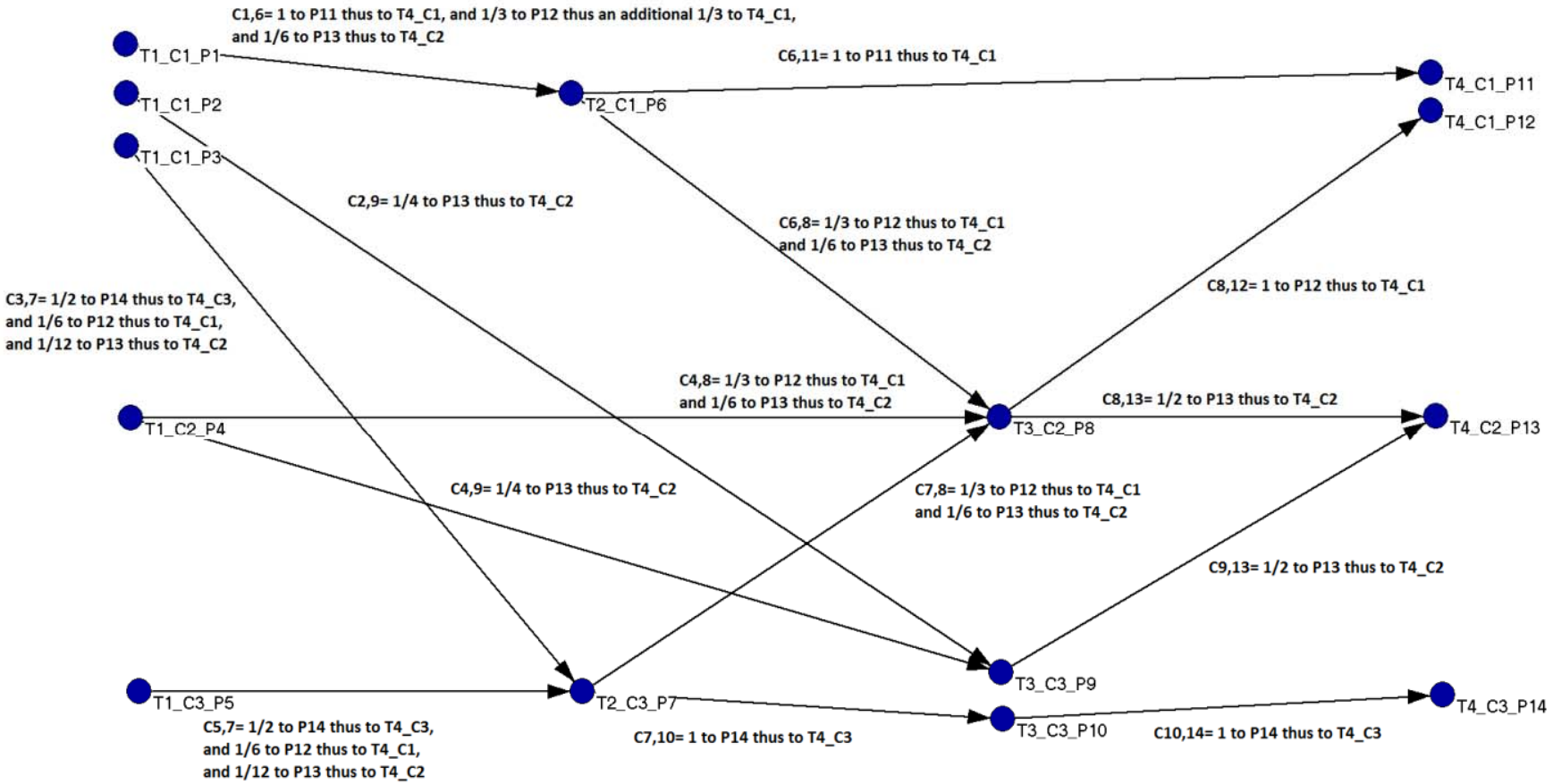


Figure A1.1. A hypothetical citation network of 14 patents

Table A1.1 shows the fractionally-counted direct citation flows ($_{m}^{k}df_{i,j}^{t1,t2}$ in Equation 1) implied by this hypothetical network. Cell values indicate knowledge flows from row entities to column entities. As an example, let us focus on patent m=P8 (i.e., T3_C2_P8) which has three citations, one to k=P4 (of T1_C2), one to k=P6 (of T2_C1), and one to k=P7 (of T2_C3). In our fractional system of accounting, each of these three citations imply a knowledge flow of 1/3=0.3333 to patent number 8, which are to be found on 4th 6th and the 7th rows of the 8th column of the numerical area of the table, which has the column label P8. Values in table A1.1 correspond to the variable $_{m}^{k}df_{i,j}^{t1,t2} = \frac{1}{S_k}$ defined with EQ1.

			T1					T2		T3			T4			
			C1			C2	C3	C1	C3	C2	C3		C1		C2	C3
			P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14
T1	C1	P1						1								
		P2									0.5					
		P3							0.5							
	C2	P4								0.333	0.5					
	C3	P5							0.5							
T2	C1	P6								0.333			1			
	C3	P7								0.333		1				
T3	C2	P8												1	0.5	
	C3	P9													0.5	
		P10														1
T4	C1	P11														
		P12														
	C2	P13														
	C3	P14														

Table A1.1. Fractional accounting of the direct citation flows in the stylized citation network depicted on Figure A1.1

		T1			T2		T3		T4		
		C1	C2	C3	C1	C3	C2	C3	C1	C2	C3
T1	C1				1	0.5		0.5			
	C2						0.33	0.5			
	C3					0.5					
T2	C1						0.33		1		
	C3						0.33	1			
T3	C2								1	0.5	
	C3									0.5	1
T4	C1										
	C2										
	C3										

Table A1.2. Direct citation flows on Figure A1.1 aggregated from the patent-level into time cohort-county combinations.

The numerical area of the table is in fact the citation matrix that underlies the graph in Figure A1.1, but column normalized. For N given patents, the citation matrix is the NxN square matrix $\mathbf{A}$, where $\mathbf{A}_{mk}$=1 if patent k cites patent m, and 0 otherwise. Thus, the number of citations made by a patent k is the sum of the kth column of $\mathbf{A}$. In the notation of EQ1, $s_k = \sum_{\forall m} \mathbf{A}_{mk}$.

Aggregation of the flows in Table A1.1 from the patent level to the level of the time cohort-county combinations, which is the operation implied by EQ2, leads to Table A1.2. Observe that the entries of Table A1.2 directly correspond to the variable $c_{i,j}^{t1,t2}$ defined earlier. For example, the value 0.5 which one finds on the 2nd row of the 5th column of the numerical area in Table A2.2 would correspond to $c_{C2,C3}^{T1,T3}$.

It can easily be observed that our hypothetical citation matrix indicates a higher number of short-run citations than citations of longer-run. That is, the majority of the citations go from patents of T4 to patens of T3, from patents of T3 to patens of T2, or from patents of T2 to patents of T1. Out of the 14 citations (which is only coincidentally equal to the number of patents), none span a time lag of 3 periods, and only four citations span a time lag of 2 periods (P11 citing P6, P8 citing P4, P9 citing P2 and P4), whereas the rest span only a lag of 1 time period. This is similar to what one observes generally in our dataset where the majority of the citations span a period of 3 to 6 years. For the clustering of the US counties, we consider the short-run citations of all patents of the time cohort 2004-2006. Accordingly, the two time cohorts considered for the clustering are *t1*= [2001, 2006] and *t2* = [2004, 2006].

For the two above-mentioned time cohorts, the dataset reports respectively 123,108 and 114,878 US patents, counted fractionally as explained above. Out of the 3,234 US counties, only 1,464 and 1,436 have more than 1.0 (fractionally-counted) patent in the two respective time cohorts. This implies 1,436x1,464 (approximately 2.1 million) pairs of US counties between which knowledge flows can potentially be observed. Out of this potential however, we observe only 38,980 pairs between which there is some flow, exhibiting a highly skewed distribution. The total knowledge flow from all US counties of time cohort *t1* to all US counties at cohort *t2* is about $c = \sum_{\forall i,j} c_{i,j}$ =24.5420 fractionally-computed citations. If we ignore the flows where the knowledge flow between the county pair is less than 1.0 fractionally-computed citation (i.e., flows only where c_{ij}>1), the number of county pairs drops to 3,464 and the total US flows to 15,435 fractionally-computed citations.

Our algorithm to cluster counties on the basis of the citation flows is based on our definition of the 'supra-normality' of citation flows, which in turn, is based on the expected number of citations. Given *c* citations actually observed between 2001-2006 patents of all US counties, if the number of citations made by a county was proportional to the county's share in all patents

of the destination period (i.e., $\frac{p_i^{t2}}{p^{t2}}$) and the number of citations received by a county was proportional to the county's share in all patents of the source period (i.e., $\frac{p_j^{t1}}{p^{t1}}$), then the expected number of citations between the period *t1* patents of county R^j and period *t2* patents of R^i is

$$\boldsymbol{E\left(c_{i,j}^{t1,t2}\right) = c\frac{p_j^{t1}}{p^{t1}}\frac{p_i^{t2}}{p^{t2}}} \qquad \textbf{EQ 3}$$

On the basis of this expected number, we define a "supra-normal knowledge flow" as the phenomenon where, for a given pair of counties *i* and *j*, the actual citation flow $c_{i,j}^{t1,t2}$ is higher than the expected flow $E\left(c_{i,j}^{t1,t2}\right)$. More formally, for the following metric

$$\boldsymbol{cs_{i,j}^{t1,t2} = \frac{c_{i,j}^{t1,t2}}{E\left(c_{i,j}^{t1,t2}\right)} = \frac{c_{i,j}^{t1,t2}}{p_j^{t1}p_i^{t2}}\frac{p^{t1}p^{t2}}{c}}, \qquad \textbf{EQ 4}$$

supra-normality of the citations between counties *i* and *j* is defined as the condition

$$\boldsymbol{cs_{i,j}^{t1,t2} > 1.} \qquad \textbf{EQ 5}$$

Figure A1.2 depicts the relation between the supra-normality of flows $cs_{i,j}^{t1,t2}$ and the joint sizes of the underlying county pair (i.e., $p_j^{t1}p_i^{t2}$, as implied multiplicatively by the expected flow equation above). Both magnitudes exhibit a highly skewed distribution over the county pairs. The larger/darker points belong to county pairs that are immediate neighbors to each other.

There are two highly interesting facts that catch attention. First, the frontier of supra-normality of citation flows is not determined by the immediately contiguous county pairs. However, further analysis (not fully reported here) indicates that the county pairs at the frontier are generally characterized by quite low geodesic distances between the underlying counties of the pair. Second, the lower right part of the chart indicates that there exists a large chunk of large-sized (in terms of patenting activity) county pairs which are characterized by less-than-expected (slightly-more-than-expected) citation flows and none (few) of these pairs share a common border. In contrast to the pairs at the frontier of supra-normal flows, further analysis (not fully reported here) indicates that the largely-sized county pairs in this chunk are generally characterized by high geodesic distances between the underlying counties of the pair (such as a county of the west-coast, and one of the east-coast).

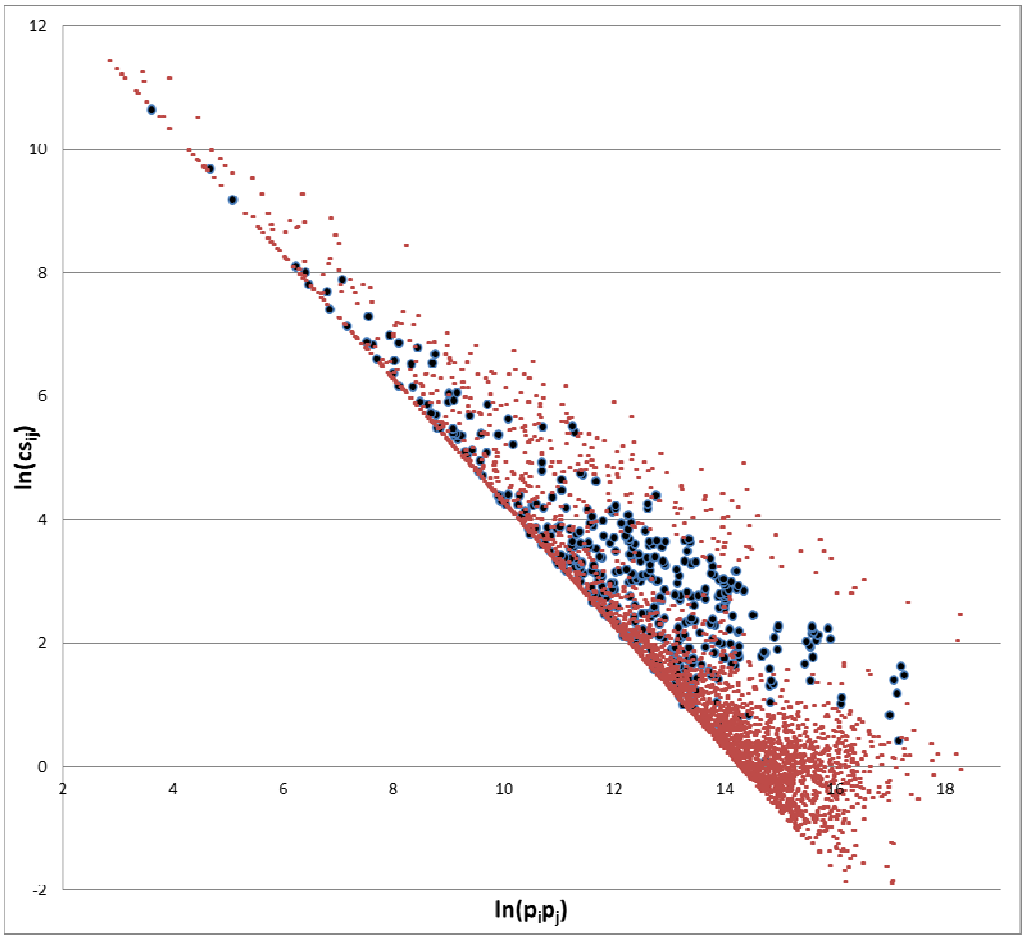


Figure A1.2 Supra-normality of flows between county pair vs. joint size of the county pair in terms of patenting activity[3].

The algorithm that the clusters of counties used in the analysis is entirely based on the magnitude of the variable $cs_{i,j}^{t1,t2} > 1$ and the contiguity relation between the US counties. Let us define matrix **fs**, respective elements of which are defined as

$$\boldsymbol{fs_{i,j}} = \begin{cases} \mathbf{1}, \boldsymbol{if}\ \boldsymbol{cs_{i,j}^{t1,t2}} > 1, and\ \boldsymbol{a_{i,j}} = \mathbf{1} \\ \mathbf{0}, \boldsymbol{otherwise} \end{cases} \qquad \text{EQ 6}$$

where $a_{i,j}$ (of the adjacency matrix **a**) is a binary variables which is equal to 1 if counties i and j are immediate neighbors and zero otherwise. Note that, although $fs_{i,j}$ is a pair-wise indicator, we do not need the $i \neq j$ restriction as an additional constraint to $fs_{i,j}$ =1 since for any i, $a_{i,i} = 0$ by definition (i.e, a county is not its own immediate neighbor).

[3] The observations shown are limited to the 3464 cases where the flow between the patents of the county pair is no less than 1.0 fractionally-computed citation (i.e., flows only where c_{ij}>1)

The clustering procedure starts by drawing a map of all counties, where the following rule applies for coloring the counties:

$$\boldsymbol{color_i} = \begin{cases} \boldsymbol{red, if\ \sum_{\forall j \neq i} fs_{i,j} \geq 1, or\ \sum_{\forall j \neq i} fs_{j,i} \geq 1} \\ \boldsymbol{white, otherwise} \end{cases} \qquad \text{EQ 7}$$

On such a map, one can visually identify clusters as a continuum of contiguous red counties. Formally, a cluster is a set of counties each of which is accessible from all other members of the cluster via at least one route that does not require passing over any border between a cluster member county and a non-cluster member county.

Mathematically, the operationalization of this definition requires the computation of a geodesic distance matrix **gd** on the basis of the matrix **fs** (defined above) used as the underlying adjacency matrix. Let us define a sequence of square matrices $\mathbf{tmp}^t$ each of the appropriate dimensions and initiate $\mathbf{tmp}^1 = \mathbf{fs}$. At each iteration of the following algorithm let the scalar $ms^t = \sum_{\forall i,j} tmp_{ij}^t$ (i.e., sum of all elements of the matrix). After setting t=1 we iterate on as (the operator · stands for matrix multiplication):

1) Compute matrix $\mathbf{smp}^{t+1} = \mathbf{fs} \cdot \mathbf{tmp}^t$
2) Construct matrix $\mathbf{tmp}^{t+1}$ such that

$$tmp_{ij}^{t+1} = \begin{cases} tmp_{ij}^t, if\ tmp_{ij}^t > 0 \\ t+1, if\ tmp_{ij}^t = 0\ and\ smp_{ij}^{t+1} > 0\ and\ i \neq j \\ 0, otherwise \end{cases}$$

3) If $ms^{t+1} = \sum_{\forall i,j} tmp_{ij}^{t+1} > 0$, then set t=t+1 and go to Step 1. Else set $\mathbf{gd} = \mathbf{tmp}^{t+1}$ and stop.

The diagonal elements of the matrix are all zero, which conforms to the definition of geodesic distance. Furthermore, for any county pair *i* and *j*, which are not mutually accessible to each other, $gd_{ij}=0$ as well. Element $gd_{ij}>0$ if and only if counties *i* and *j* are accessible to each other via trajectories that pass through at least gd_{ij} county borders. Given these properties (and the fact that matrix **gd** is not based on the geographical adjacency matrix but our alternative matrix **fs** which restrict the definition of adjacency from having a common border to not only having a border but also having supra-normal citation flows in between) we can identify the clusters as islands which are not accessible to each other.

For any given county *i*, all other counties *j* where $gd_{ij}>0$ belongs to one unique cluster together with county *i*. Thus, after setting Set $i=1$ and Cln=1, our algorithm which assigns US counties to their respective clusters works as follows:

1. If county *i* is not already assigned to any cluster:
    a. Assign county *i* to cluster *Cln*
    b. Also assign all counties *j* where $gd_{ij}>0$ to *Cln*

c. Set $i=i+1$, and $Cln= Cln+1$

2. If $i<=3,234$ go to[4] step 1, stop otherwise.

While this is the essence of our county-clustering into contiguous regions, it is obvious that the smallest setting in which a cluster can emerge out of our clustering method is a pair of counties. Nevertheless, one observes US counties which produce a high number of patents, yet enjoys no supra-normal flows of citations with any of its immediate neighbors.

Ten of these individual counties that do not make it to the initial list of clusters, but appear with a large number of patents in the period under consideration are added to the list of clusters. The largest number of patents, per county, on this new list of single counties is 5,649, the smallest number 415. The smallest number of patents in the initial list of clusters is 99 (average 2,626). Hence the 10 single counties that are added can be seen as large in terms of their inventive activity. Some of the 10 extra counties that are included share a border with one of the clusters. In this case, we add them to the cluster that they share a border with. This is the case for San Diego (joining the Los Angeles cluster), New London and New York (both joining the large East coast cluster). Finally, two of the extra counties border with each other, but not with another cluster. We add these two counties (Fort Lauderdale and Palm Beach) as a new cluster. The final list of clusters thus contains 35 members.

We also merge one of the small clusters (2 counties) to the large cluster with 30 counties. In this case, the small cluster is Long Island (counties Nassau and Suffolk), while the large cluster covers a large part of the East coast, stretching from Boston to Philadelphia. We add Long Island to the large cluster, because it is geographically very close, although it does not share a land border with the large cluster, and because it is arguably part of the New York urban environment that is, for the largest part, included in the East coast cluster.

[4] Note that 3,234 is the number of US counties in the dataset.

Annex A2. Genealogical accounting of deep citations

The framework in which we account for long-run knowledge flows through indirect (as well as direct) citation linkages bears a close analogy to genealogy, especially to the particular context of Mendelian genetics in its analytical operationalization. While we like to emphasize this analogy as our preferred theoretical foundation to our methods of accounting, the reader can as well perceive the method as a natural extension of the fractional accounting framework that underlies the analytical methodology of this paper, as generalized from direct citation flows to the accounting of knowledge flows through direct and indirect citation linkages.

For further illustration[5], let us go back to the hypothetical citation network depicted in Figure A1.1 and the related tables A1.1 and A1.2 which give the direct knowledge flows implied by the citation network in a matrix form.

Tables A2.1 and A2.2 are respectively analogous to Tables A1.1 and A1.2 yet they indicate both direct and indirect knowledge flows implied by the same network, as calculated according to our genealogical accounting method. The table entries (i.e., matrix elements) which are different than their counterparts in tables A1.1 and A1.2 are highlighted in the latter tables respectively. Let us discuss on the basis of a few illustrative examples.

As mentioned in annex A1, our hypothetical citation network entails 10 citation lags of 1 period, 4 lags of 2 periods and no citation lags of 3 periods. Thus an analysis that is limited by direct citations has nothing to offer in terms of the long-run trajectories (i.e., deep citations) that run between the early patents of T1 and the latest patents of T4.

Let us take the citation trajectory P3→P7→P8→P13 and think backwards in time. Since P13 makes two citations, the direct citation link P8→P13 is equivalent to a flow of ½ citations. Also given that P8 makes 3 citations, the flow through the trajectory P7→P8→P13 must be 1/2x1/3=1/6. And also given the two citations of P7, we calculate the (fractional) flow through from P3 to P13 through the trajectory P3→P7→P8→P13 as 1/2x1/3x1/2=1/12 which is approximately 0.083. Since P3→P7→P8→P13 is the only trajectory that connects P3 to P13, the 3rd row of the 13th column of the numerical area in table A2.1 is 0.083.

[5] For even further illustrations and explanations, see Martinelli & Nomaler (2014) which presents a variant of our accounting method in closer analogy to genealogy in its exposition.

			T1					T2		T3			T4			
			C1			C2	C3	C1	C3	C2	C3		C1		C2	C3
			P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14
T1	C1	P1						1		**0.333**			**1**	**0.333**	**0.167**	
		P2									0.5				**0.250**	
		P3							0.5	**0.167**		**0.5**		**0.167**	**0.083**	**0.5**
	C2	P4								0.333	0.5			**0.333**	**0.417**	
	C3	P5							0.5	**0.167**		**0.5**		**0.167**	**0.083**	**0.5**
T2	C1	P6								0.333			1	**0.333**	**0.167**	
	C3	P7								0.333		1		**0.333**	**0.167**	**1**
T3	C2	P8												1	0.5	
	C3	P9													0.5	
		P10														1
T4	C1	P11														
		P12														
	C2	P13														
	C3	P14														

Table A2.1. Genealogic/fractional accounting of the (direct and indirect) citation flows in the stylized citation network depicted on Figure A2.1. Cells are highlighted for all cases where cell value is different than the corresponding cell in table A2.1

Another pair, the deep citations between which span a lag of 3 periods is P4 and P13. Yet unlike the previous case of P3→P13, we find two trajectories between the latter pair: P4→P8-P13 and P4→P9-P13. Given the number of citations respectively made by P8 (3), P9 (2) and P13(2), we can calculate the flow through the former trajectory as 1/2x1/3=1/6 and the latter as 1/2x1/2=1/4. Therefore the total flow from P4 to P13 is calculated as 1/6+1/4=10/24 which is approximately equivalent to 0.417 citations. This value can be found in the 4th row of the 13th column of the numerical area in table A2.1.

Given our analogy to genealogy, this last case would be similar to a grandfather whose genetic footprint in his granddaughter is higher than ¼. In citation networks where patents typically make more than 2 citations, i.e., have more than 2 "parents" (contrary to the case of sexual reproduction in the animal kingdom) and also where one can often find several (overlapping and/or non-overlapping) trajectories between a given pair of related patents, the phenomenon highlighted by our last example is quite common.

After computing the flow between all connected pairs of patents via all existing trajectories, we populate the matrix **af** of all (direct and indirect) flows between patents. For our hypothetical citation network, the numeric part of Table A2.1 is the associated 14x14 matrix **af**. Analogously to Equation 2, we can aggregate the flows among patents to flows among time cohort-county combinations as

$$\boldsymbol{ac}_{i,j}^{t1,t2} = \sum_{\forall m,k}\left(\,{}_{m}^{k}\boldsymbol{af}_{i,j}^{t1,t2}\right), \qquad \text{EQ 8}$$

where indices *m* and *k* correspond to individual patent numbers, *i* and *j* to counties, *t1* and *t2* to time cohorts as in Equation 2. All results discussed in the main text are based on the values in the matrix **ac,** with t1= [1986, 1988] as the source and t2 = [2004, 2006] as the destination, which span a time period of 20 years.

		T1			T2		T3		T4		
		C1	C2	C3	C1	C3	C2	C3	C1	C2	C3
T1	C1				1	0.5	0.5	1	1.5	0.5	0.5
	C2						0.33	0.5	0.333	0.417	
	C3					0.5	0.17	0.5	0.167	0.083	0.5
T2	C1						0.33		1.333	0.167	
	C3						0.33	1	0.333	0.167	1
T3	C2								1	0.5	
	C3									0.5	1
T4	C1										
	C2										
	C3										

Table A2.2. Direct and indirect citation flows on Figure A2.1 aggregated from the patent-level into time cohort-county combinations.

Although useful for expositional purposes, our description of the computation of flows through deep citations (which is based on the enumeration of all trajectories between all possible pair of patents in question) is not necessarily the most efficient method for the according computations. The sequence of multiplications and summations can easily be formulated in matrix algebra and a practical computational method presents itself.

Recall the NxN citation matrix **A** where N is the number of patents in a given citation network. For any given *i*=1,2,…N and *j*=1,2,…,N, if patent *k* cites patent *m* $\mathbf{A}_{mk}$=1, and otherwise $\mathbf{A}_{mk}$=0. We indicated that the matrix of (fractionally-computed) direct citation flows **df** is essentially the column-normalized variant of the citation matrix, such that

$$\boldsymbol{df_{mk}} = \boldsymbol{A_{mk}} / \sum_{\forall p} \boldsymbol{A_{pk}}. \qquad \text{EQ 9}$$

One can find a direct analogy between matrix **A** and the intermediate requirements coefficient matrix in input/output economics. The related matrix algebra pioneered by Leontief maintains that given a matrix of direct usage coefficients **A**, the matrix product $\mathbf{A}^2$ gives the second degree indirect requirement coefficients, $\mathbf{A}^3$ the third degree indirect requirement coefficients and so infinitum.

Similarly, given a column-normalized citation matrix (**df**) that account fractionally for direct knowledge flows, the matrix product $\mathbf{df}^2$ accounts (fractionally) for the flows to all patents via all trajectories of exactly two citation links, $\mathbf{df}^3$ accounts (fractionally) for the flows via all

trajectories of exactly three citation links and so on. Given that a citation network is strictly acyclical, $\mathbf{df}^Z$ remains a non-zero matrix[6] up to the point where Z is less than or equal to the length of the longest trajectory in the network and becomes a zero matrix thereafter. Accordingly, given our earlier notation, the matrix of all citation flows is

$$\mathbf{af} = \mathbf{df} + \mathbf{df}^2 + \mathbf{df}^3 + \mathbf{df}^4 + \cdots + +\mathbf{df}^Z = \mathbf{I} - (\mathbf{I} - \mathbf{df})^{-1}. \qquad \text{EQ 10}$$

The right hand-side of this equation is the well-known identity in input-output economics that relates to the Leontief inverse. Although this is a highly efficient way of computing deep citation flows for citation networks of up to several thousand patents, even the computer power available nowadays is not quite sufficient to store a citation network of several million patents and their several million citation links in a square matrix and compute the matrix inverse. Therefore we have written a dedicated software (in the C++ language) which reads citation network data from an ASCII text file as node pairs, computes the direct and indirect flows, and (since the patent-level flow matrix is not directly necessary for our analysis), aggregates the flows among individual patents into flows among time cohort-county combinations and outputs the matrix **ac** defined by equation 8 (the matrix for which our toy example in this appendix is given in table A2.2).

[6] I.e., the sum of the all elements of the matrix is greater than zero.

Annex A3. Sandwich networks

Given two individual patents *m* and *k*, our definition of the corresponding network is the network that consists of all citation trajectories that connect *m* and *k* indirectly. For illustrative examples, let use resort again to the hypothetical citation network on Figure A1.1

Let us take patents P2 and P13, respectively, as source and sink of a potential sandwich. There exists only one trajectory that connects the two patents P2→P9→P13, which is thus the sandwich network between the selected patents. For the particular case of P1 and P13, the situation is similar: There is only a single trajectory, P1→P6→P8→P13 that connects the two selected patents. Let us also consider the sandwich between P3 and P13, which is also a single trajectory that visits P3→P7→P8→P13.

Now we are ready to aggregate our sandwich definition from the patent level to the time cohort-county combinations. P1, P2 and P3 are all patents that belong to county C1 at time cohort T1. Patent P13 is the only one that belongs to county C2 at time cohort T4. Therefore the sandwich network between C1_T1 and C2_T4 is the sub-network of the citation network on figure A1.1 which consists only of the citations P1→P6, P6→P8, P8→P13, P2→P9, P9→P13, P3→P7, and P7→P8.

Let us also consider the sandwich between P4 and P13. P4→P8→P13 and P4→P9→P13 are the two trajectories one can identify. Given that P4 is the only patent of C2_T1, we can conclude that the sandwich network between C2_T1 and C2_T4 (i.e., same county, different time periods) is the sub-network which consists of the citations P4→P8, P4→P9, P8→P13, and P9→P13.

It is interesting to note that citation links are not unique to sandwiches. The above examples clearly indicate that the citation link P8→P13 belongs to more than one sandwich, both at the level of individual patent pairs, and pairs of time cohort-county combinations. Also note that the computer algorithm that detects sandwich networks is not computationally expensive. Unlike in the illustrative example above, there is no need to enumerate all trajectories between the source and the sink patents. Our algorithm (also implemented in C++ the language), starting from all source patents, follows all forward citation links sequentially[7] and marks all patents visited in the process as "connected to the source", and then does the same backward in time. That is, starting from the sink patents, backward citations are followed sequentially[8] and all patents visited in the process are marked as "connected to the sink." The participants

[7] I.e., all citing patents of the source set, all patents that cite the citing patents of the source set, then all patents that cite the patents that cite the patents that cite the source set, and so on till the right truncation of the citation network is reached.

[8] I.e., all patents cited by the sink set, then all patents that are cited by the patents which are cited by the sink set and so on till the left truncation of the citation network is reached.

of the sandwich network are clearly the only patents which are marked from both sides as “connected to the source” and “connected to the sink” The sandwich network is ultimately extracted as the sub-network of our global citation network by removing all patents which are not visited from both directions as well as all citations made and received by the patents removed accordingly.

Annex A4. On the identification of the main paths (trajectories) in sandwich networks

The methods proposed by Hummon and Doreian (1989) and Verspagen (2007) are explained in detail in those papers. As also explained briefly in the main text, the main building block of the network-reduction procedure is a method that assigns an "importance" indicator (SPNP) to link (i.e., arc) in the citation network.

This indicator is essentially related to the number of trajectories in the entire network to which the link in question participates. Needless to say, any citation chain that links one patent to some other patent is considered a trajectory. For example consider a chain of citations P1→P2→P3→P4→P5→P6→P7 which is clearly a trajectory. However, despite constitution various sub-trajectories of the former, chains such as P2→P3→P4→P5→P6, P4→P5→P6→P7, P1→P2→P3→P4, …, are all considered by the SPNP metric as different trajectories, each of which add to the importance of any individual citation link that contributes to each.

Having computed the SPNP values for each citation link, Hummon and Doreian (1989), Verspagen (2007) and all the other studies that utilize this methodology, identify main paths (i.e., the most important trajectories in the network) with a simple local search algorithm which, for each start node[9] in the network, identifies a connected trajectory, by hopping from patent to patent, where, at each new patent arrived, moves on to the next patent which is connected to the current patent with the highest SPNP value among all patents that cite the current patent. Intuitively, this procedure aims at picking up trajectories that are composed of citation chains with as high SPNP values as possible in each of its constituent citations. And despite being a seemingly local search algorithm, the scope of the search is not completely local, since, by its construction, SPNP is a globally 'informed' indicator since it indicates the importance of the citation in terms of its position in the global scale of the network.

Not only does this local search algorithm make intuitive sense, but also published research that uses the method report findings that make technological sense. However, one cannot help but notice that what this local search algorithm is supposed to achieve exactly is not explicitly defined in terms of some well-defined optimality criteria. We address this issue here.

We compute the SPNP values in the standard way, yet we substitute the local search algorithm with a global optimizer. For each start node in the citation network, the algorithm enumerates all trajectories that emanate from the given node, for each trajectory, computes the multiplicative product (i.e., sum of the logarithm) of all SPNP values of the constituent citation links, and picks up the trajectory where this multiplicative product of the SPNP values is maximized.

[9] I.e., a patent which does not make any citation, mostly due the (left) truncation of the network.

Annex A5. On the classification and the geographical analysis of contributions to knowledge main paths

As mentioned in the main text, having identified the main paths of knowledge flows among the salient knowledge producing areas of the US, our next aim is to make an analysis in terms of the (global) geographic distribution of those who have contributed to these main paths.

Let us illustrate the issue at hand, again by using our hypothetical citation network on Figure A1.1 as an example. Accordingly let us consider the sandwich network between T1 incarnation of county C1 and the T4 incarnation of county C2. Let us decompose the contribution of all counties (independently of the time of contribution) to each main path that connects T1_C1 to T4_C2

	C1	C2	C3
P1→P6→P8→P13	50%	50%	0%
P2→P9→P13	0%	100%	0%
P3→P7→P8→P13	0%	50%	50%

The table above illustrates the respective contributions to each of the main paths. For the first trajectory between P1 and P13, lie two patents (P6 and P8) which respectively belong to counties C1 and C2. Thus in fractional terms, C1 must have contributed by 50% (i.e., one patent out of two contributions) and C2, another 50%. The second trajectory visits only one patent (P9) which implies 100% county C2 contribution to this main path. Finally, the contributions of P7 and P8 to the third main path respectively credit counties C2 and C3 for 50% contribution each.

In this toy example, each main path exhibits a significantly different pattern of contribution. Indeed, in such a small network it is simple to identify patterns to categorize different processes into a finite number of categories that make some sense. However, it becomes increasingly difficult to do so as the number of cases (i.e., main paths) to categorize and/or the number of counties that may potentially make a contribution increases. Unlike the 3 main paths – 3 counties case depicted in the example above, our sandwich networks typically feature several thousands of main paths (i.e., a high number of observations to classify) which may visit any combination of the approximately 5100 geographical entities (i.e., regions, counties) present in the REGPAT database (i.e., the geographical resolution of the analysis).

The appropriate tool for the job is clearly some sort of clustering algorithm , in particular 'k-means clustering'. This leaves us two related choices. One is the number of groups to impose on the algorithm (the k of the k-means clustering algorithm) and the second is the

geographical entities whose respective contribution to main paths is to be the basis of clustering. We take the 35 US clusters (as discussed in section 4) as well as the aggregate categories, US Non Cluster (i.e., all US counties that do not belong to any of the 35 clusters, as aggregated into one entity), Europe (i.e., all European NUTS2 regions aggregated), Japan (all Japanese provinces aggregated) and ROW (rest of the world as aggregated into one entity). This makes 39 geographical entities whose respective contribution to each of the 23,060 main paths we identify constitutes the 39 variables which were introduced into k-means the clustering algorithm of the software package MatLab. As discussed in the main text, the group size was chosen as k=8.

The UNU-MERIT Working Paper Series

2016-01 *Mexican manufacturing and its integration into global value chains* by Juan Carlos Castillo and Adam Szirmai
2016-02 *New variables for vocational secondary schooling: Patterns around the world from 1950-2010* by Alison Cathles
2016-03 *Institutional factors and people's preferences in social protection* by Franziska Gassmann, Pierre Mohnen & Vincenzo Vinci
2016-04 *A semi-endogenous growth model for developing countries with public factors, imported capital goods, and limited export demand* by Jan Simon Hallonsten and Thomas Ziesemer
2016-05 *Critical raw material strategies in different world regions* by Eva Barteková and René Kemp
2016-06 *On the value of foreign PhDs in the developing world: Training versus selection effects* by Helena Barnard, Robin Cowan and Moritz Müller
2016-07 Rejected Afghan asylum seekers in the Netherlands: Migration experiences, current situations and future aspirations
2016-08 *Determinants of innovation in Croatian SMEs: Comparison of service and manufacturing firms* by Ljiljana Bozic and Pierre Mohnen
2016-09 *Aid, institutions and economic growth: Heterogeneous parameters and heterogeneous donors* by Hassen Abda Wakoy
2016-10 *On the optimum timing of the global carbon-transition under conditions of extreme weather-related damages: further green paradoxical results* by Adriaan van Zon
2016-11 *Inclusive labour market: A role for a job guarantee* scheme by Saskia Klosse and Joan Muysken
2016-12 *Management standard certification and firm productivity: micro-evidence from Africa* by Micheline Goedhuys and Pierre Mohnen
2016-13 *The role of technological trajectories in catching-up-based development: An application to energy efficiency technologies* by Sheng Zhong and Bart Verspagen
2016-14 *The dynamics of vehicle energy efficiency: Evidence from the Massachusetts Vehicle Census* by Sheng Zhong
2016-15 *Structural decompositions of energy consumption, energy intensity, emissions and emission intensity - A sectoral perspective: empirical evidence from WIOD over 1995 to 2009* by Sheng Zhong
2016-16 *Structural transformation in Brazil, Russia, India, China and South Africa (BRICS)* by Wim Naudé, Adam Szirmai and Nobuya Haraguchi
2016-17 *Technological Innovation Systems and the wider context: A framework for developing countries* by Hans-Erik Edsand
2016-18 *Migration, occupation and education: Evidence from Ghana* by Clotilde Mahé and Wim Naudé
2016-19 *The impact of ex-ante subsidies to researchers on researcher's productivity: Evidence from a developing country* by Diego Aboal and Ezequiel Tacsir
2016-20 *Multinational enterprises and economic development in host countries: What we know and what we don't know* by Rajneesh Narula and André Pineli
2016-21 *International standards certification, institutional voids and exports from developing country firms* by Micheline Goedhuys and Leo Sleuwaegen

2016-22 *Public policy and mental health: What we can learn from the HIV movement* by David Scheerer, Zina Nimeh and Stefan Weinmann

2016-23 *A new indicator for innovation clusters* by George Christopoulos and Rene Wintjes

2016-24 *Including excluded groups: The slow racial transformation of the South African university system* by Helena Barnard, Robin Cowan, Alan Kirman and Moritz Müller

2016-25 *Fading hope and the rise in inequality in the United States* by Jo Ritzen and Klaus F. Zimmermann

2016-26 *Globalisation, technology and the labour market: A microeconometric analysis for Turkey* by Elena Meschi, Erol Taymaz and Marco Vivarelli

2016-27 *The affordability of the Sustainable Development Goals: A myth or reality?* By Patima Chongcharoentanawat, Kaleab Kebede Haile, Bart Kleine Deters, Tamara Antoinette Kool and Victor Osei Kwadwo

2016-28 *Mimetic behaviour and institutional persistence: a two-armed bandit experiment* by Stefania Innocenti and Robin Cowan

2016-29 *Determinants of citation impact: A comparative analysis of the Global South versus the Global North* by Hugo Confraria, Manuel Mira Godinho and Lili Wang

2016-30 *The effect of means-tested social transfers on labour supply: heads versus spouses - An empirical analysis of work disincentives in the Kyrgyz Republic*by by Franziska Gassmann and Lorena Zardo Trindade

2016-31 *The determinants of industrialisation in developing countries, 1960-2005* by Francesca Guadagno

2016-32 *The effects of productivity and benefits on unemployment: Breaking the link* by Alessio J. G. Brown, Britta Kohlbrecher, Christian Merkl and Dennis J. Snower

2016-33 *Social welfare benefits and their impacts on labour market participation among men and women in Mongolia* by Franziska Gassmann, Daphne François and Lorena Zardo Trindade

2016-34 *The role of innovation and management practices in determining firm productivity in developing economies* by Wiebke Bartz, Pierre Mohnen and Helena Schweiger

2016-35 *Millennium Development Goals (MDGs): Did they change social reality?* by Janyl Moldalieva, Arip Muttaqien, Choolwe Muzyamba, Davina Osei, Eli Stoykova and Nga Le Thi Quynh

2016-36 *Child labour in China* by Can Tang, Liqiu Zhao, Zhong Zhao

2016-37 *Arsenic contamination of drinking water and mental health* by Shyamal Chowdhury, Annabelle Krause and Klaus F. Zimmermann

2016-38 *Home sweet home? Macroeconomic conditions in home countries and the well-being of migrants* by Alpaslan Akay, Olivier Bargain and Klaus F. Zimmermann

2016-39 *How do collaboration and investments in knowledge management affect process innovation in services?* by Mona Ashok, Rajneesh Narula and Andrea Martinez-Noya

2016-40 *Natural disasters and human mobility* by Linguère Mously Mbaye and Klaus F. Zimmermann

2016-41 *The chips are down: The influence of family on children's trust formation* by Corrado Giulietti, Enrico Rettore and Sara Tonini

2016-42 *Diaspora economics: New perspectives* by A.F. Constant and K.F. Zimmermann

2016-43 *Entrepreneurial heterogeneity and the design of entrepreneurship policies for economic growth and inclusive development* by Elisa Calza and Micheline Goedhuys

2016-44 *Gini coefficients of education for 146 countries, 1950-2010* by Thomas Ziesemer

2016-45 *The impact of rainwater harvesting on household labor supply* by Raquel Tsukada Lehmann and Christian Lehmann

2016-46 *The impact of piped water supply on household* welfare by Raquel Tsukada and Degol Hailu

2016-47 *The impact of household labor-saving technologies along the family life cycle* by Raquel Tsukada and Arnaud Dupuy

2016-48 *River deep, mountain high: Of long-run knowledge trajectories within and between innovation clusters* by Önder Nomaler and Bart Verspagen