Identifying geographic clusters: A network analytic approach

Roberto Catini [a,*], Dmytro Karamshuk [b], Orion Penner [a], Massimo Riccaboni [a,c]

[a] LIME Research Unit, IMT Institute for Advanced Studies, Lucca, Italy
[b] Department of Informatics, King's College, London, UK
[c] KU Leuven, Leuven, Belgium

ARTICLE INFO

ABSTRACT

In recent years there has been a growing interest in the role of networks and clusters in the global economy. Despite being a popular research topic in economics, sociology and urban studies, geographical clustering of human activity has often been studied by means of predetermined geographical units, such as administrative divisions and metropolitan areas. This approach is intrinsically time invariant and it does not allow one to differentiate between different activities. Our goal in this paper is to present a new methodology for identifying clusters, that can be applied to different empirical settings. We use a graph approach based on k-shell decomposition to analyze world biomedical research clusters based on PubMed scientific publications. We identify research institutions and locate their activities in geographical clusters. Leading areas of scientific production and their top performing research institutions are consistently identified at different geographic scales.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Recent advancements in our understanding of the evolution of scientific and technological systems stem from the theoretical conceptualization, and empirical investigation, of the dynamic relationship between the notions of innovation cluster and network. The idea of agglomeration – or local concentration – of related activities is hardly new, it was indeed raised by Alfred Marshall in his seminal work (Marshall, 1920) and subsequently developed by other scholars such as Jacobs (1961) and Perroux (1950), to name a few. The revival of the cluster idea among economic geographers, sociologists and economists of innovation, follows from the influential work of **47**, Porter (1990) and the global fame of the Silicon Valley (Saxenian, 1994). Scholars have contributed a multiplicity of interpretations of the original concept, thus resulting in a certain degree of theoretical and empirical confusion (Martin and Sunley, 2003; Maskell and Kebir, 2006; Maskell and Malmberg, 2007). Since the notion of cluster is only vaguely defined in terms of geographical scale and network properties, the cluster idea has been operationalized in different ways. The lack of consensus on a precise definition of clusters compounded with difficulties in identifying the location of inventive activities, has led

many economists to proxy innovation clusters with administrative regions.[1] Two elements are core in most cluster definitions. First, clusters are constituted by related activities. Second, clusters are geographically proximate groups of interlinked individuals and organizations. This common ground notwithstanding, boundaries of network communities and geographical areas are not clearly identified. Therefore, most of the available evidence has used standard industrial classification systems and administrative regions, or zip codes, to investigate localized knowledge spillovers and agglomeration economies. Apart from a few recent contributions (Ter Wal and Boschma, 2011; Balland et al., 2013; Delgado et al., 2013), by taking regional and industrial boundaries as fixed and pre-determined, most of empirical studies falls short in describing the evolution and plasticity of clusters both in geographical and technological terms. Regional clusters of innovative activities depend on the networks that arise from linkages among co-located inventors. Learning by interacting is a crucial force pulling inventors into clusters and a crucial factor of success for innovative clusters. Therefore, the local presence of bundles of related knowledge bases is constantly evolving. As a consequence, clusters change their location, size and performances over time and tend not to fit static administrative boundaries (Cerina et al., 2013).

---

* Corresponding author. Tel.: +39 3282313451.
  E-mail addresses: roberto.catini@imtlucca.it
(R. Catini), dmytro.karamshuk@kcl.ac.uk (D. Karamshuk), orion.penner@imtlucca.it
(O. Penner), m.riccaboni@imtlucca.it (M. Riccaboni).

[1] By the same token, the tendency to analyze the size of cities by means of administrative urban areas has been recently criticized (Rozenfeld et al., 2011).

As noted elsewhere, a full theory of clusters should answer to the questions of "what, how, why as well as when, where and who", focusing in particular on causal relationships (the "why" question) (Maskell and Kebir, 2006; Press, 2006). To achieve these goals quality data is required, as are analytical tools to study the evolution of clusters (Delgado et al., 2013), i.e. the emergence of new clusters, how they grow and decline, and how they shift into new fields (Menzel and Fornahl, 2010; Boschma and Frenken, 2006). To analyze cluster life cycles data should be: (a) referring to specific activities; (b) geo-localized at the finest possible level of resolution (i.e. street address); (c) time resolved to know when, exactly, a given activity has been performed; and (d) individual and institution names should be disambiguated to properly identify the economic agents who have been active in the cluster over time. The increasing availability of geo-referenced micro-data allows for the implementation of a bottom-up approach for clustering related activities in space and time. There is a rapidly growing literature in geography (Feser and Sweeney, 2000; Marcon and Puech, 2003, 2009), spatial statistics (Diggle and Chetwynd, 1991) and economics (Duranton and Overman, 2005), that has advanced the identification of clusters of various phenomena using a series of spatial point-process and distance-based techniques. More recently, new and rigorous methods have been developed to identify bundles of related activities (Delgado et al., 2013, 2012). Our work complements previous contributions in this field by developing a bottom-up network approach to map and trace the evolution of clusters over time. Most of the literature so far used zip code centroids to analyze the spatial distribution of activities based on geographical concentration measures.[2]

In this paper we introduce a new method based on network-analytic tools instead of spatial density measures. Our approach takes into account the relational patterns among co-located activities. Moreover, we develop a computationally efficient solution and apply it to a large database of street-level records. We focus on the geographical loci of scientific production, as opposed to economic production or human population density. Thus, because agglomeration patterns for R&D and production differ, we used the geographic distributions of scientific output instead of more conventional variables such as industry employment, plant output, or population statistics. For innovation clusters, patents and scientific publications are typically used in cluster identification because scientific and technological outputs tend to correspond to the locations in which inventors innovate. As in Furman et al. (2005), we use scientific publications to locate biomedical research activities. Specifically we use the PubMed database of approximately 23 million bibliographic records corresponding to peer-review publications in the biomedical sciences. To avoid relying on predetermined administrative boundaries, such as NUTS3 regions or Metropolitan Statistical Areas (MSAs), and to allow for the examination of the innovation landscape within cities, we employ and develop a variety of novel methods. To observe the geography of biomedical research activity on the finest scale possible we have exploited existing geocoding APIs to arrive at high resolution (street level) coordinates for a significant portion of all biomedical publications. To street level geocoded data, we apply an aggregation algorithm drawing inspiration from previous complex networks research. This approach works in a straightforward manner: treating all geocoded papers as nodes, we connect nodes (papers) that are less than a certain distance apart. This is repeated until there are no further nodes within the certain distance to link to. At the completion of this algorithm we arrive at a set of large geographic clusters, akin to MSAs, that capture the bulk of biomedical

production in a geographic area. To better understand the institutional make up of the clusters produced by the algorithm we develop an (institutional) disambiguation algorithm that allows for the identification of institutions from noisy addresses. The high resolution geocoding also allows for deeper analysis of the internal structure of clusters. Specifically, we carry out a "k-core" analysis in which low degree nodes are recursively removed from the graph, exposing the internal structure of a cluster.

A similar approach has been recently adopted by Alcácer and Zhao (2012) who use a clustering identification algorithm to analyze the role of internal cluster relationships in the global semiconductor industry. Our method differs from the one they implemented in four aspects. First, we use online mapping services (such as Yahoo or Google) to localize researchers. The same approach can be applied to extract location information from address fields in any data source including patents, publications, research projects and firm locations. Second, our method does not critically depend of the choice of parameters, such as the neighborhood radius and the contour threshold. Our method works at a finer level of resolution, based on the intuitive notion of walking distance between inventors. Third, apart from localizing the inventor, our algorithmic approach also provides a disambiguated name of the affiliation of the inventor. Thus, we are able to position firms and research institutions in clusters. Fourth, we defined clusters by the actual distribution of inventor locations following a spatially-embedded graph approach. By doing that, our method applies for any definition of local graph links. While we use a specific empirical context (global biomedical research) for illustrative purposes, the insights and methodologies are general and our methodology can be applied in other empirical settings. Our approach is particularly valuable in all cases in which information is available about the relationships between individuals and/or firms. In those cases, our methodology can take as an input the geo-referenced network of collaborations in which only a subset of co-localized nodes are connected.

The manuscript is organized in the following manner. First we carefully explain the geocoding and affiliation disambiguation approaches that generate the extremely high resolution geographical and institutional data required for our analysis. Then we describe in full the graph based algorithm we employ to identify the relevant clusters and cores of biomedical scientific production around the world. Following that we present the results of our method, showing the three largest clusters produced by our algorithm and probing their internal core structure with complex network analysis techniques, and then showing the structural change of the San Francisco cluster resulting from a particular development thereby illustrating the potentiality of our approach in analyzing the temporal evolution of clusters. We follow with general observations and trends arising from our analysis and conclude by discussing the avenues of future enquiry opened by our approach.

2. Preparing affiliation and location data

To capture and map knowledge production in the biomedical research field, we develop a novel approach to collect and cluster publicly-available knowledge in the life sciences. Namely, we draw on the National Library of Medicine's PubMed database. PubMed contains bibliographic records of over 23 million publications in the biomedical sciences, with full coverage dating back to 1966. The main advantage of using PubMed data is that it is open access and a number of other high value data sets have already been extracted from it, for example the Authority dataset that provides fully disambiguated author names (Torvik et al., 2005). An additional benefit of using of an open access dataset is it allows other scholars to

---

[2] One exception being (Feser and Sweeney, 2000) in which street addresses are also considered but only for a limited number of records.

verify and extend our methodology. Moreover, PubMed provides a detailed classification of the content of publications, based on the Unified Medical Language System (UMLS). The PubMed dataset does not, however, provide extensive information about the location at which the research was carried out: only the affiliation and address for each paper's corresponding author is provided. While practices vary across the disciplines and sub-disciplines of biomedical science, it is generally the case that the corresponding author is either the junior scientist that carried out the majority of the research, or the senior scientist (Principle Investigator) in charge of the research group in which the research was carried out. In either instance, in the majority of cases it seems a reasonable proxy that the corresponding author's affiliation corresponds to the location at which the majority of work for any given publication took place.

We emphasize that the methods described below may be applied to a wide variety of data sets, as long as there is a sufficiently descriptive address associated with each piece of data. The criteria for sufficiently descriptive differs slightly between the geocoding approach and the affiliation disambiguation algorithm. In the case of geocoding, the address must contain information beyond simply a city name. A full street address, for example, is certainly descriptive enough to result in a high resolution geocoding. Similarly, department name, or university name, or company name (with a city name for large multinationals) are generally enough to produce a high quality geocoding. In the case of the affiliation disambiguation algorithm the address (or other data in the record) must contain some hints as to the institution of origin, be it university, department, company, etc. To be concrete, examples of data to which both approaches may be applied include Thomson Reuters Web of Science and Elsevier Scopus publication databases, the European Patent Office (EPO) patent database and the Orbis global firm database by Bureau Van Dijk. On the other hand, an example of data ill suited for both approaches is United States Patent Office (USPTO) inventor address data, in which only the city name and state are included.

2.1. PubMed data geocoding

Despite the lack of structured geographic information in the PubMed dataset (the only available "country" field is empty for the majority of publication records), missing details on authors' locations can be effectively mined from affiliation strings. Provided in free text format, affiliation strings usually include information about authors' organizations and their detailed addresses. To map this textual data to a structured geographic information we exploit an out-of-box solution provided by the Yahoo Geocoding API.[3] The API is provided as a web-service which for an input affiliation string returns, in JSON format, a list of potential location matching. Each location record includes, among other data, a pair of latitude and longitude coordinates and a level of confidence (called "quality") that the suggested location is correct. In contrast, administrative information (e.g., country, region, city) is often absent or inaccurate in the API responses. Therefore, we manually map latitude and longitude pairs associated with each affiliation string to geographic administrative areas using GADM dataset.[4] As a result, each publication record in our dataset is assigned one or several location tags with geographic coordinates and corresponding administrative information. However, it is worth noting that in analysis presented later in this paper we use only the geographical coordinates of each publication, making no use of the administrative unit information.

Table 1
General statistics on the geolocation. In this context, a geocoding is said to be accurate if its quality is better than the city level.

Number of geocoded publications	9,754,647
Number of publications with "accurate geocoding"	6,163,880
Number of distinct locations	149,951
Number of distinct locations with "accurate geocoding"	130,738

Despite the fact that the proposed geocoding technique can effectively handle unstructured affiliation information from the PubMed dataset, the output of the geocoder requires some cleaning before it can be used for further analysis. Firstly, we discard all PubMed records with empty affiliation strings. Further, the geocoded data is filtered by the "quality" value associated with each location match. Particularly, we omit all location records with accuracy coarser than the street level. In Table 1 we report the resulting statistics. We note, that the filtered dataset accounts for the vast majority (63%) of publications in the original dataset. In this table "distinct locations" refer to unique [longitude, latitude] coordinates returned by the geocoding at which at least one publication exists. This is less than the number of papers because many papers get aggregated into the same coordinate, for example because they come from the exact same department. We note that we observe significant variation country to country in the rate of accurate geocoding. For the USA we see 84.5%, Japan 76.3%, the United Kingdom 70.0%, all quite solid. But on the other hand several countries are much lower, Italy and Korea coming in below 35% and Russia much below that. In cases where we observe poor geocoding translation and transliteration appear to be the main bottlenecks and future improvements to the approach are required.

2.2. Disambiguation of authors' affiliations

To characterize the spectrum of research activities concentrated in a given geographic region we need to systematize information about authors' affiliations in our data set. The major challenge here lies in extracting organization names from free form text, which constitutes the affiliation field in the PubMed data set and, further, identifying organizations from the extracted names. This problem, typically called 'disambiguation', is non-trivial due to a number of reasons which, among others, include various spelling discrepancies, different ordering of words, use of abbreviation etc. In this current work, we extract and disambiguate organization names from affiliation strings by applying the approach proposed in Jonnalagadda and Topham (2010). We justify our choice by the fact that the method proposed by Jonnalagadda and Topham (2010) performs reasonably well when applied to the PubMed data set whilst not requiring any information sources other than the affiliation strings. At this point in time it is worthwhile pointing out that there has been extensive previous work on author name disambiguation in publication databases (Torvik et al., 2005; Schulz et al., 2014) and patent databases (Lai et al., 2011). While fundamental goals of affiliation disambiguation and author disambiguation are similar, the techniques that need to be employed in each case are rather different. In the affiliation case the primary challenge is the diversity of unique strings that can refer to the same institution and there is minimal metadata available. In the author disambiguation case a lot of metadata is available and there is minimal variation in strings (i.e. John Smith may appear as J. Smith or J.M. Smith) but the major challenge is determining which J. Smith papers belong to which J. Smith.

This algorithm proceeds in two steps. First, it extracts the part of the affiliation string corresponding to organization name. This is achieved by classifying words and, further, phrases composed of words in three different classes of information: address, institution and department information. To classify individual words

---

we exploit a set of dictionaries, including a thesaurus of geographic names, organization keywords and person names. Once classes are assigned to individual words, the algorithm proceeds to classifying phrases (we split a string to phrases by commas) where a set of decision rules is applied. For example, a phrase is defined as address information if it contains no organization keywords but a geographic name possibly mixed with numbers (e.g., building number, zip code). In contrast, we say that a phrase defines an institution name if it consists of an organization keyword or an abbreviation mixed with a geographic or a person name, but not a number. For example, in the following affiliation string "Department of Chemical and Biochemical Engineering, The Advanced Centre for Biochemical Engineering, University College London, Torrington Place, London WC1E 7JE, UK" the algorithm is able to recognize "Torrington Place", "London WC1E 7JE", "UK" as address-related phrases and "Department of Chemical and Biochemical Engineering", "The Advanced Centre for Biochemical Engineering", "University College London" as organization description phrases. To distinguish between department and institution name we use a heuristic assumption that the one containing geographic or a person name is most probably an institution name, e.g., "University College London", and otherwise – name of a department, e.g., "Department of Chemical and Biochemical Engineering", "The Advanced Centre for Biochemical Engineering". Manual validation of the results suggests that 95% of institution names can be extracted correctly.

In the second step, the algorithm identifies various organization names that correspond to a single institution. The algorithm can spot and tolerate a reasonable number of discrepancies in the way organization names are written, including various spelling mistakes, missing words, different ordering of the words, etc. To this aim, we firstly tokenize all previously extracted institution names and build a dictionary of disambiguated words. This allows us to capture insignificant spelling differences between similar words, e.g., plural endings, missing apostrophes, missing letters etc. Particularly, we identify and unify the words that are not more than n characters different from each other, where n is defined as 80% of the word length. Further, we compare the phrases by computing the Levenshtein distance defined on the words level: we assume that each word in a phrase is a symbol and define the cost of a word removal as a length of the word and the cost of word replacement as a summary length of both words. We say that two phrases are similar if the Levenshtein distance between them is not larger than 4. As a result, we get the list of "normalized" institution names one of which assigned to each publication record. We note here, that the manual validation of the disambiguation results revealed a high level of accuracy, i.e., 72% of institution names were disambiguated correctly.

3. The clustering algorithm

To uncover the clusters underlying urban biomedical research productivity we start at the micro level, employing a geographical aggregation algorithm. Our algorithm is inspired by the City Clustering Algorithm (CCA) originally introduced by Rozenfeld et al. (2008) to construct cities without the use of administrative subdivisions. As originally implemented it exploited the population distribution, typically obtained from census data.

Two versions of the CCA algorithm have been developed: a discrete one designed on coarse grained data (e.g., population index defined on a homogeneous grid representing the region of interest) (Rozenfeld et al., 2011) and a continuous one, which considers two points on the map as neighbors if the distance between them is below a critical distance ℓ. In our case we will employ the
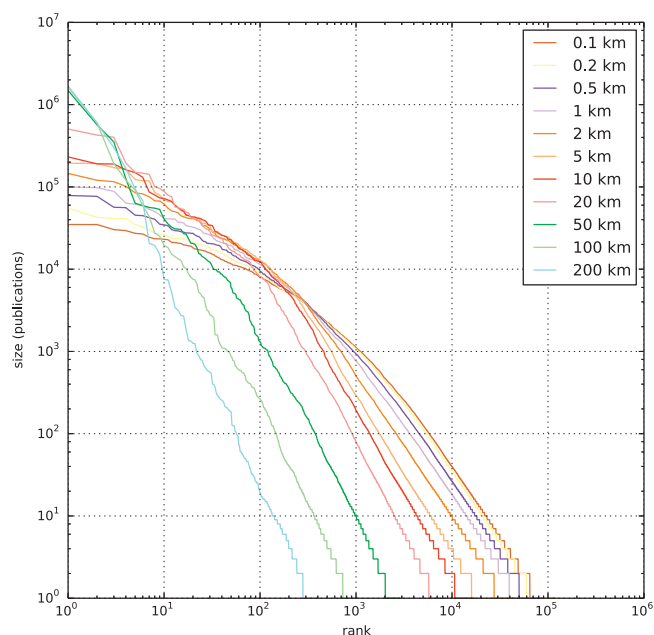

Fig. 1. The dependence of the size of a cluster on its rank. For all the distances a power-law behavior is exhibited in the tail.

continuous version because we are using the geocoded affiliations, and are not constrained by any coarse-graining procedure.

The actual clustering procedure is performed by repeating the following steps until all points are assigned to clusters:

- take one arbitrary location and assign it to a new cluster;
- find all locations closer than ℓ to the previous point and assign them to the same cluster;
- recursively add locations closer than ℓ to at least one location already in the cluster until there are no new locations within ℓ of any added location.

We have run the algorithm on a wide range of ℓ values to select the best value for defining clusters; the results can be found in Fig. 1.

A reasonable distance ℓ has been chosen heuristically as 1 km, and approximately corresponds to a "walking distance". We use the Authority data set (Torvik et al., 2005) and our geolocated data to verify if it is really the case that scientists collaborate more if they are at walking distance from each-other. First, we assigned to each author an "home address" computing the most frequent location appearing in all his/her publications. Then we extracted from the PubMed data a sample of 1 million publications and then took all the possible pairs of coauthors (i.e. we extracted all the possible 2-combinations without repetition, for each publication with at least two authors). For each pair of authors, we then computed their geographical distance according to their home addresses.

The results of this analysis is plotted in Fig. 2.[5] Fig. 2 shows that 1 km is a reasonable length scale for cluster generation as a sharp decrease in collaboration is observed at that distance.

To test the robustness of this distance we have repeated the complete analysis for 0.5 km and 2 km, obtaining similar results; however this choice is by no means definitive and will be the subject of further analysis. In Appendix A we report the maps of the main biomedical clusters in the US, Europe and Japan. A complete

---

[5] The collaborations at zero distance are discarded, since they represent a very large fraction of the total and would have impeded the readability of the chart.
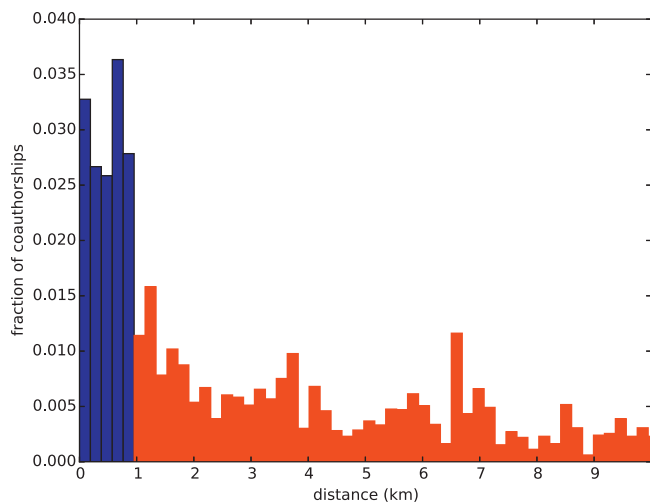
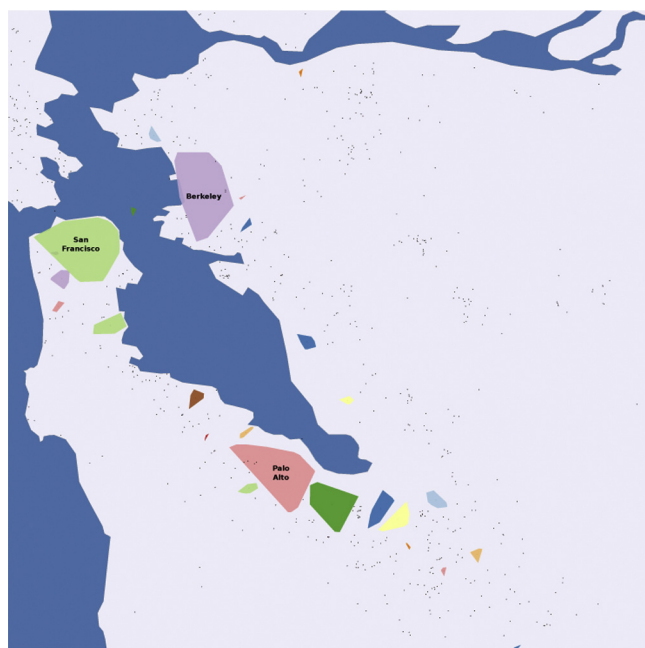**Fig. 2.** Frequency of collaboration observed at a given distance.

Fig. 3. The result of the CCA in the San Francisco Bay Area, for ℓ=1 km. Only clusters with at least a hundred publications are shown as colored polygons, other locations are plotted as black dots. Note that with this critical distance the stretch of sea between San Francisco and Berkeley segments the two high productive areas in separate clusters. (For interpretation of the references to color in this text, the reader is referred to the web version of the article.)

zoomable world map of all biomedical clusters is available at http://goo.gl/X4YUyG

In this way clusters of publications can be detected independently from the choice of the starting points, and for the most part they correspond to cities. Note that the correspondence between these components and the administrative definitions (e.g. the metropolitan statistical areas of the U.S.) is not bijective because some publications may be isolated geographically due to physical features such as the coastline and will be considered as a different cluster (see the San Francisco case in Fig. 3).

Since a worldwide system that defines regional boundaries does not exists, we have carried out a comparison of our clustering strategy with the following traditional administrative subdivisions: Metropolitan Statistical Areas and FIPS county codes (United States), and NUTS at level 3 (European Union).

Table 2
Comparison with MSAs. The list is sorted by the size of the matching component and truncated at the 10th entry.

Metropolitan statistical area	Publications	Fraction in biggest intersection
Boston–Cambridge–Newton	219,885	91%
New York–Newark–Jersey City	150,321	54%
Washington–Arlington–Alexandria	108,146	55%
Philadelphia–Camden–Wilmington	94,834	68%
Ann Arbor	87,113	93%
Houston–The Woodlands–Sugar Land	74,165	72%
Seattle–Tacoma–Bellevue	64,067	86%
Los Angeles–Long Beach–Anaheim	63,718	42%
San Francisco–Oakland–Hayward	53,716	45%
Baltimore–Columbia–Towson	52,281	52%

To perform the comparison each publication has been matched to the appropriate administrative region and to the corresponding connected component in our graph. Then, for each region, the biggest intersection (in terms of number of publications) with a network component has been selected. For the Metropolitan Statistical Areas (MSAs) each of our components intersects at most one (see Table 2). The fraction of papers in the intersection it is almost unitary (e.g. in the Boston – Cambridge case) there is a bijective correspondence. In the other cases our algorithm has detected multiple components (typically two main ones), and the fraction is approximately half: for example, in Washington–Arlington–Alexandria case, Bethesda is in a component different from downtown Washington. In these cases which is the "right" clustering can be debated, but we note that our method has the advantage of using actual research output instead of the distribution of the population.

When FIPS county codes and NUTS-3 levels are considered, our method has the advantage not to split clusters that span multiple administrative area, such as Boston-Cambridge in the US and London in Europe.

4. k-Shell decomposition

To further understand the complex structure of innovation clusters generated in the previous step we construct the graph of all publications, in which an edge is drawn between two vertices if the distance between them is less than ℓ.[6] Note that the connected components emerging in this graph correspond to the clusters identified by the algorithm described in the previous paragraph.

The cores of the different components can be revealed using the k-shell decomposition, a method which has been previously applied in complex networks to study the topology of the Internet (Carmi et al., 2007; Kitsak et al., 2010). The idea is to use a percolative process to assign each node in the graph (which in our case is a publication) a measure of its connectivity.

The actual method starts by assigning all isolated nodes a zero index. Then all nodes with exactly one neighbor (i.e. no other publication can be found within ℓ) are assigned the index 1 and removed from the graph. This step is repeated until no other nodes with one neighbor remain in the graph. At each further step the index k is increased by one and the nodes with degree k are removed iteratively, until the graph is empty.

Naturally, there is a maximum shell index k_{max} such that once all shells with $k \leq k_{max}$ are removed, the graph is empty. In our case k_{max} can be interpreted as the size (in term of the number of papers) of the area of most intense biomedical research, because

---

[6] Alternative definitions for local connectivity can be implemented as well. For instance, one can draw a link if the technological and geographical distance are both below a given threshold. Another possibility is to consider co-authorship links when the address field is available for all co-authors.
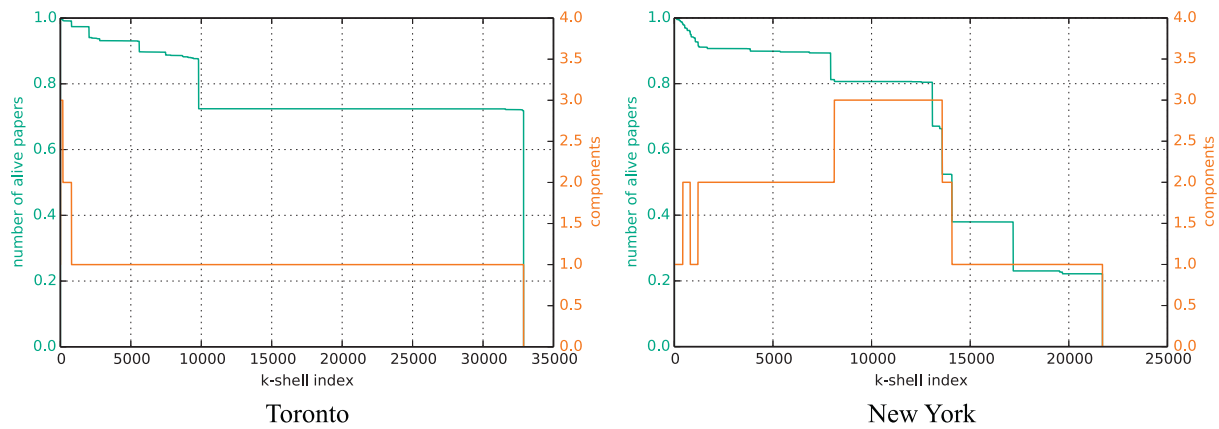
Toronto   New York

Fig. 4. A comparison between a monocore and a multicore city. The alive papers at small k are those still in the graph at the step k of the k-shell algorithm.

it is in practice determined by the total production of very tightly spaced buildings.

Looking at the dynamics of this k-shell process between 0 and k_{max} two main types of removal steps can be clearly distinguished:

- a vertex at the geographical frontier is dropped because the low density of papers in the neighborhood is lower than required to be in the k-shell; in this case two parts of the same city become disconnected.
- a significant number of vertices are dropped at once with a sharp reduction in the fraction of publications remaining in the graph; in this case the vertices are typically positioned very close to each other and the corresponding geographical region becomes empty after the next removal step.

For different cities the behavior can be strikingly different, as we see in Fig. 4. In some cities, like Toronto, the dynamics are concentrated at the very beginning of the k-shell decomposition (low k indices) and the fraction of the papers in the last shell is on the order of 1. We shall characterize such cities as "monocore". On the other hand some cities, like New York, experience sizeable drops in the fraction of the alive papers through out the transition from k = 0 to k = k_{max} and the city is split in different components at various stages. We characterize these cities as "multicore" and each of the cities we discuss in the final subsection fall into this category.

We can further characterize the core structure of cities by observing how the city breaks down as k increases. In that way we can define a functional "core" as a (large) set of publications that are geographically close to each other and represent a significant portion of the city's productivity. For the analysis below we define a core as such a group of publications with the same k-index that together are at least one tenth (1/10) the size (publications) of the k_{max} core for that city.

Using the journal impact factor as a proxy for publication impact it is possible to estimate the quality of publications arising from each city and each core. In this way we can study the heterogeneity of research quality within cities. To do this we use the Thomson Reuters Journal Citation Report (JCR) Impact Factor for 2012. A journal's 2012 Impact Factor is the average number of citations accumulated in 2012 by a paper published in the previous two years (2010–2011). To estimate the average quality of research within a core we simply average the Impact Factors of each publication in that core. This quantity can be thought of as the expected number of citations an "average" paper from the given core will accumulate in its first two years. It is a coarse measure as the individual papers will receive more or less citations than the journal average and it is possible that the output of specific cities or cores changes significantly over time. However, we are averaging over thousands
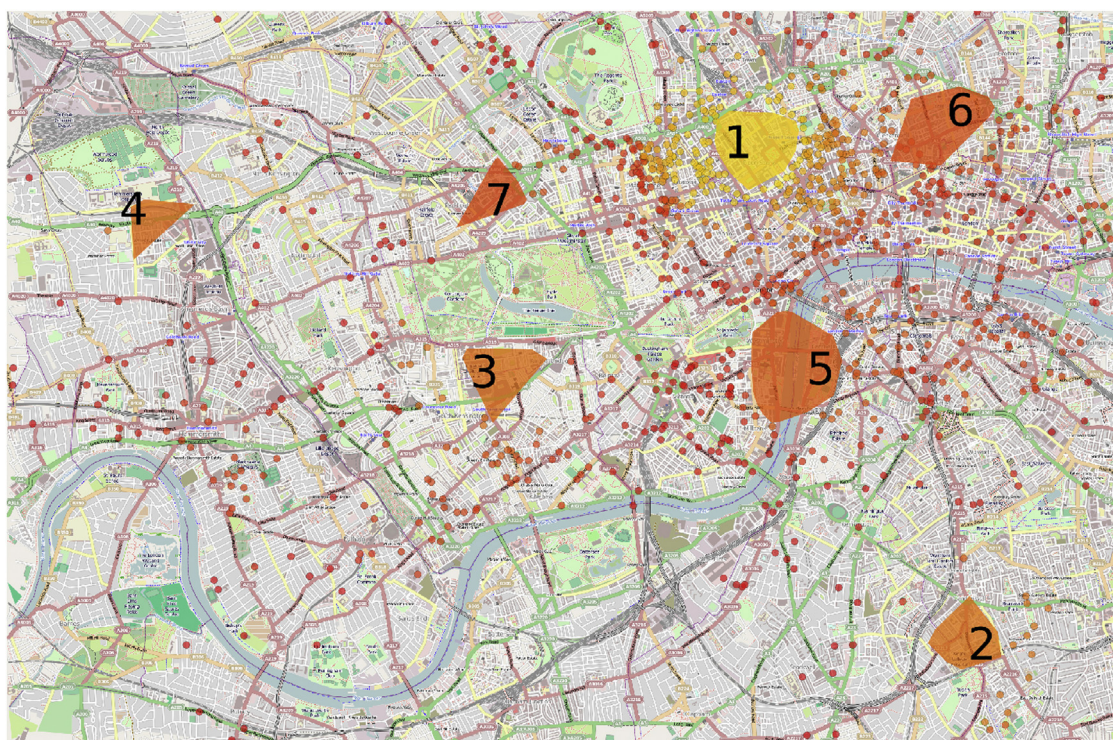
and tens of thousands of papers for each core so the average Impact Factor is a reasonable measure of impact for qualitative comparison. We have also tested with the 5-year Impact factor and Impact Factors from earlier years and there is no change in the qualitative results.

5. Results

5.1. The structure of the leading clusters

In this section we present results of our analysis for three major cities: London, Boston, and Tokyo. These cities are three of the largest according to the number of unique addresses discovered in our geocoding process and are among the top producers in terms of total publications. For each city, we present our results in three ways. First we show a descriptive map of core and k-shell structure of the city. In these maps each core is depicted as a polygon whose color is proportional to the number of papers published within the core. The color scale runs from red, to orange, to yellow, with the saturation point being approximately 42,000 papers as seen below in Boston. In addition to the polygonal cores, all other points that produced a publication are also indicated by small dot, with color proportional to the point's k-shell. Each core is also labeled with the core's rank in terms of total publications produced. The second way we present our results is in a table containing summary statistics for each core. Using the rank appearing in both the map and the table one can see a variety of properties for each core. These are the k-shell, total number of publications and their average Impact Factor as well as the main affiliation produced by our affiliation disambiguation approach. It is worth noting that the affiliations appearing in these tables have only been minimally altered from the direct output of our algorithm. Changes were only made in cases where the affiliation was simply too long, a good example being "Royal Postgraduate Medical School" in London has been abbreviated as "RPMS". The final figure we provide describing our city level results is a graph that tracks three quantities as a function of increasing k-shell. The first quantity is the total number of papers at that k-shell or higher. The second quantity is the number of distinct clusters present at a given k-shell. The third is the average Impact Factor of papers at that k-shell or higher. It is our conjecture that the evolution of these three curves can be used to better quantitatively characterize the structure of the cluster.

Turning our attention to the London cluster found in Fig. 5.1 we observe seven main cores. These cores seem to generally capture the geography of biomedical research within the city, hitting the main institutions. In comparison to the other top clusters London is particularly dominated by the first core, 27.3% of the papers lie in the first core as opposed by 8.6% in the second. While the most

City 1: London

99729 Publications	Maximum K-Shell: 27288
Average IF: 4.74	Core Threshold: 2729 Papers

Core #1		Core #2	
Affiliation(s):	University College London	Affiliation(s):	King's College London
Publications:	27291	Publications:	8538
K-Shell:	27288	K-Shell:	8537
Average IF:	5.08	Average IF:	3.98

Core #3		Core #4	
Affiliation(s):	Imperial College London	Affiliation(s):	RPMS, Hammersmith Hospital
Publications:	7645	Publications:	6874
K-Shell:	7644	K-Shell:	6873
Average IF:	5.37	Average IF:	5.19

Core #5		Core #6	
Affiliation(s):	St Thomas' Hospital	Affiliation(s):	Moorfields Eye Hospital
Publications:	5329	Publications:	4000
K-Shell:	5268	K-Shell:	3994
Average IF:	3.93	Average IF:	3.38

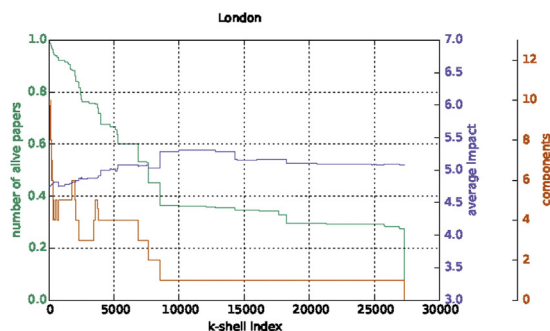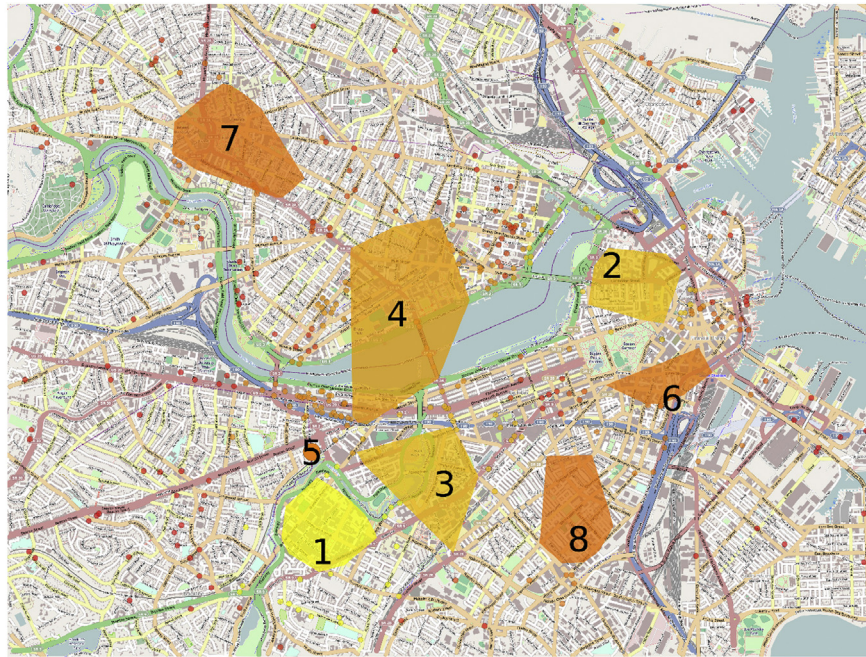| Core #7	
Affiliation(s):	St Mary's Hospital
Publications:	3772
K-Shell:	3769
Average IF:	4.24

Fig. 5. (a) London, (b) Boston and (c) Tokyo. (For interpretation of the references to color in this text, the reader is referred to the web version of the article.)

common affiliation (as determined by our affiliation disambiguation algorithm) in the first core is University College London, there is also a significant number of London School of Hygiene and Tropical Medicine, UCL Institute of Neurology, and University College Hospital affiliations. In reporting these affiliations, a limitation of the disambiguation algorithm is clear as UCL Institute of Neurology is indeed a part of UCL. However, given the general difficulty of institutional disambiguation the results are strong overall. In terms of IF

City 2: Boston

167958 Publications	Maximum K-Shell: 41138
Average IF: 6.18	Core Threshold: 4114 Papers

Core #1		Core #2	
Affiliation(s):	Harvard Medical School	Affiliation(s):	Massachusetts General Hospital
Publications:	41140	Publications:	24754
K-Shell:	41138	K-Shell:	24752
Average IF:	6.42	Average IF:	5.41

Core #3		Core #4	
Affiliation(s):	Brigham and Women's Hospital	Affiliation(s):	Massachusetts Institute of Technology
Publications:	23728	Publications:	17236
K-Shell:	23723	K-Shell:	14444
Average IF:	6.29	Average IF:	7.30

Core #5		Core #6	
Affiliation(s):	Beth israel Deaconness Medical Center	Affiliation(s):	Tufts University School of Medicine
Publications:	11921	Publications:	9785
K-Shell:	17848	K-Shell:	9989
Average IF:	5.88	Average IF:	5.09

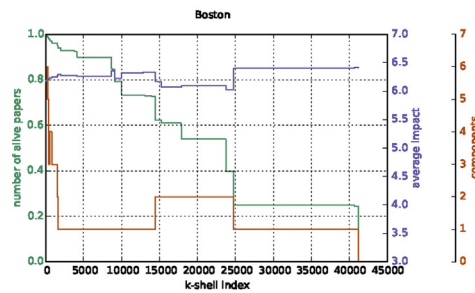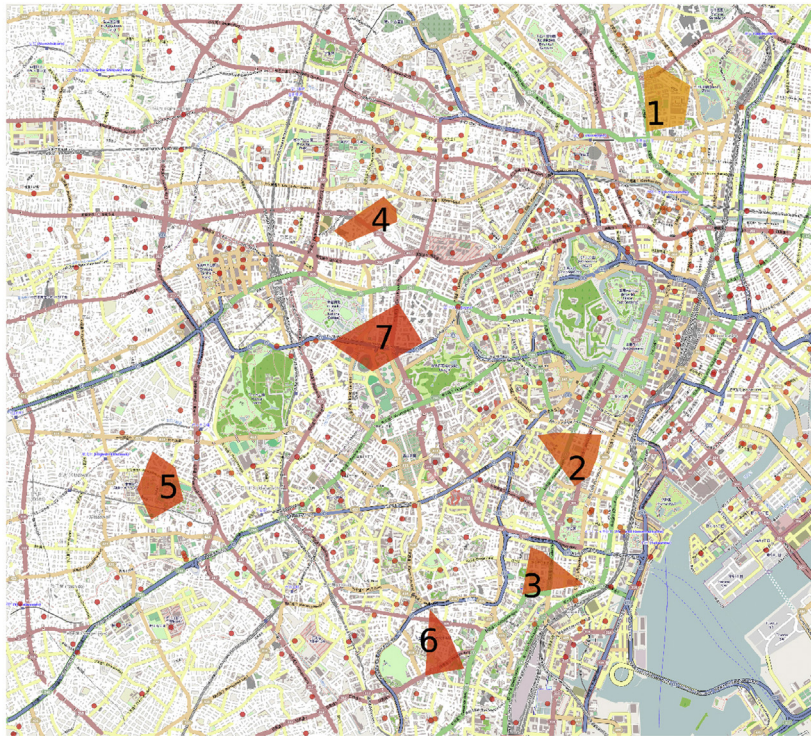| Core #7		Core #8	
Affiliation(s):	Harvard University	Affiliation(s):	Boston University School of Medicine
Publications:	9095	Publications:	8651
K-Shell:	9056	K-Shell:	8648
Average IF:	8.27	Average IF:	4.70

Fig. 5. (*Continued*)

a degree of heterogeneity is seen across London's cores. The large research Universities (UCL and Imperial College) display slightly higher average IF. The k-shell structure of London is largely dominated by the first core with the peripheral cores falling out quickly.

The Boston cluster, found in Fig. 5.2, spans two US counties, Suffolk and Middlesex and of all clusters identified in our analysis it is the largest in terms of papers produced. This cluster has been previously identified as top performing in other analysis (Owen-Smith et al., 2002). The cores identified capture the main points of production within the area: Longwoods Medical Area, Massachusetts General Hospital, Massachusetts Institute of Technology (MIT), Harvard, etc. At the same time cores #3 and #5 represent a direction for further improvement of our geocoding, as some portion of papers originating in the Longwoods Medical Area were placed at the centroids of two zip codes (02215 and 02115). Aggregating those two with the first core we note again that the Longwoods Medical Area accounts for approximately half of the publications in the entire Boston region. Looking for additional

City 3: Tokyo

61609 Publications	Maximum K-Shell: 14860
Average IF: 3.05	Core Threshold: 1486 Papers

Core #1		Core #2	
Affiliation(s):	University of Tokyo	Affiliation(s):	Jikei University School of Medicine
Publications:	14861	Publications:	4760
K-Shell:	14860	K-Shell:	4759
Average IF:	4.47	Average IF:	1.76

Core #3		Core #4	
Affiliation(s):	Keio University	Affiliation(s):	Tokyo Women's Medical University
Publications:	4049	Publications:	3850
K-Shell:	4048	K-Shell:	3849
Average IF:	1.76	Average IF:	2.47

Core #5		Core #6	
Affiliation(s):	University of Tokyo	Affiliation(s):	Kitasato University
Publications:	2798	Publications:	1999
K-Shell:	2797	K-Shell:	1998
Average IF:	3.03	Average IF:	3.48

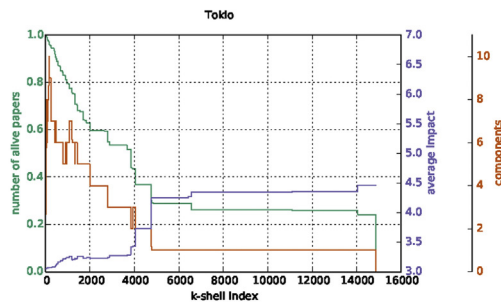| Core #7	
Affiliation(s):	Keio University School of Medicine
Publications:	1690
K-Shell:	1686
Average IF:	3.66

Fig. 5. (*Continued*)

affiliations within the first core we note, in addition to Harvard Medical School, Brigham and Women's Hospital, Beth Israel Deaconness Medical Center, Children's Hospital Boston, and Dana-Farber Cancer Institute. The second core is dominated by Massachusetts General Hospital. Core #4 is largely MIT, but in jumping the Charles River also covers Boston University's main campus and that is reflected by the disambiguated affiliations returned by our algorithm. Interestingly, in core #7 the presence of the Broad Institute was also detected. In terms of IF this cluster is somewhat heterogeneous, perhaps along lines of academic

Fig. 6. The San Francisco cluster up to year 1998 (top) and from 1999 to present (bottom).

prestige. From the k-shell analysis the fact there are several significant drops in the curve indicates that Boston is, indeed, an area with many distinct centers of production.

For the Tokyo cluster the algorithm seems to have hit the main loci of production, as shown in Fig. 5.3. The first core corresponds to the University of Tokyo main campus. In the second core an additional significant affiliation produced by the disambiguation is Toranomon Hospital. Core #5 is not an error, but actually the Komaba campus of the University of Tokyo. Within the fifth core the disambiguation algorithm also correctly identified Tokai University, which is present there though its Yoyogi campus. Within Tokyo we observe a large variation in both the productivity of cores, as well as their average IF. Indeed in the case of Tokyo the range of average IFs is very wide, but does not seem to correlate directly with productivity. The k-shell analysis correctly represents the city

as one dominated by a single core, the University of Tokyo main campus.

5.2. The dynamics of innovation clusters: the case of San Francisco

One of the main advantages of our approach it that is allows for the tracking of clusters and cores over time, in a variety of dimensions including size, shape, and content (specialization). To demonstrate this potential we consider the temporal evolution of the cluster centered on the city of San Francisco (light green in Fig. 3). In Fig. 6 one can see the cores of biomedical productivity before, and after, the year 1998. The year 1998 was selected as the separation point since it was the year in which the Mission Bay area redevelopment project started. Up to 1998 the biomedical

research activities in San Francisco were concentrated in the two main campuses of the University of California at San Francisco: the Mount Zion area, which stretches up to the California Pacific Medical Center and the Parnassus Heights campus followed by the San Francisco Veterans Affairs Medical Center. After 1998 the cluster extends and cores appear in the southern part of San Francisco. The Parnassus Medical Center underwent rapid development, becoming the main core of biomedical production immediately followed by the newly established Mission Bay campus. UCSF's Mission Bay Campus is the largest ongoing biomedical construction project in the world. Such an important change in the structure of the San Francisco biomedical area would have gone unnoticed by traditional approaches to the identification of innovation clusters. As illustrated by this example of San Francisco, our method allows for the tracking of the evolution of a cluster and its cores in space, ecology and technological content/specialization.

6. Final discussion

The agglomeration of innovative activities is the center of attention for a widespread community of researcher in various disciplines, and the cluster idea is one of the most popular topics amongst policy makers and local economic practitioners. Despite its widespread appeal there is still a lack of consensus on the appropriate geographical entities for empirical analysis and policy intervention.

In this paper we discussed some significant limitations of the common practice of relying on static and exogenously defined administrative regions to analyze innovation clusters, issues exacerbated in cross-country studies when regions are defined according to distinct national criteria. These limits are particularly restrictive when the emphasis is placed on the dynamic evolution of clusters over time.

We provide a new methodology for endogenously identifying clusters, based on the location of activities (activity-based clustering) and apply it to the case of research institutions involved in scientific and technological production. Our method exploits a network approach to identify the relevant production clusters. In this respect, it allows for the identification of local concentrations of related inventive activities, offering some clear advantages in terms of precision, flexibility and applicability to the analysis of clusters in time, across countries and data sources. Moreover, our method is computationally efficient and relies upon publicly available data and software. To the best of our knowledge, ours is the only method which allows to identify clusters based on the existence of local bundles of related geo-referenced activities rather than assuming that knowledge interdependencies always exist within administrative borders.

However, our approach is not without limitations that should be addressed in future work. First, geocoding should be improved, especially as far as country coverage is concerned, by testing alternative solutions for geo-referencing (i.e. Google places API or other similar services). Second, disambiguation can be refined by using repositories of names of individuals and institutions. Third, a better mathematical approach for characterizing the structure of clusters should be developed for the purposes of comparison across clusters

and with random distributions of R&D activities. Despite current limitations, we are confident that recent developments in open data policies and data-driven economic and geographical analyses will facilitate the development of better tools for dynamic content-based analysis of clusters.
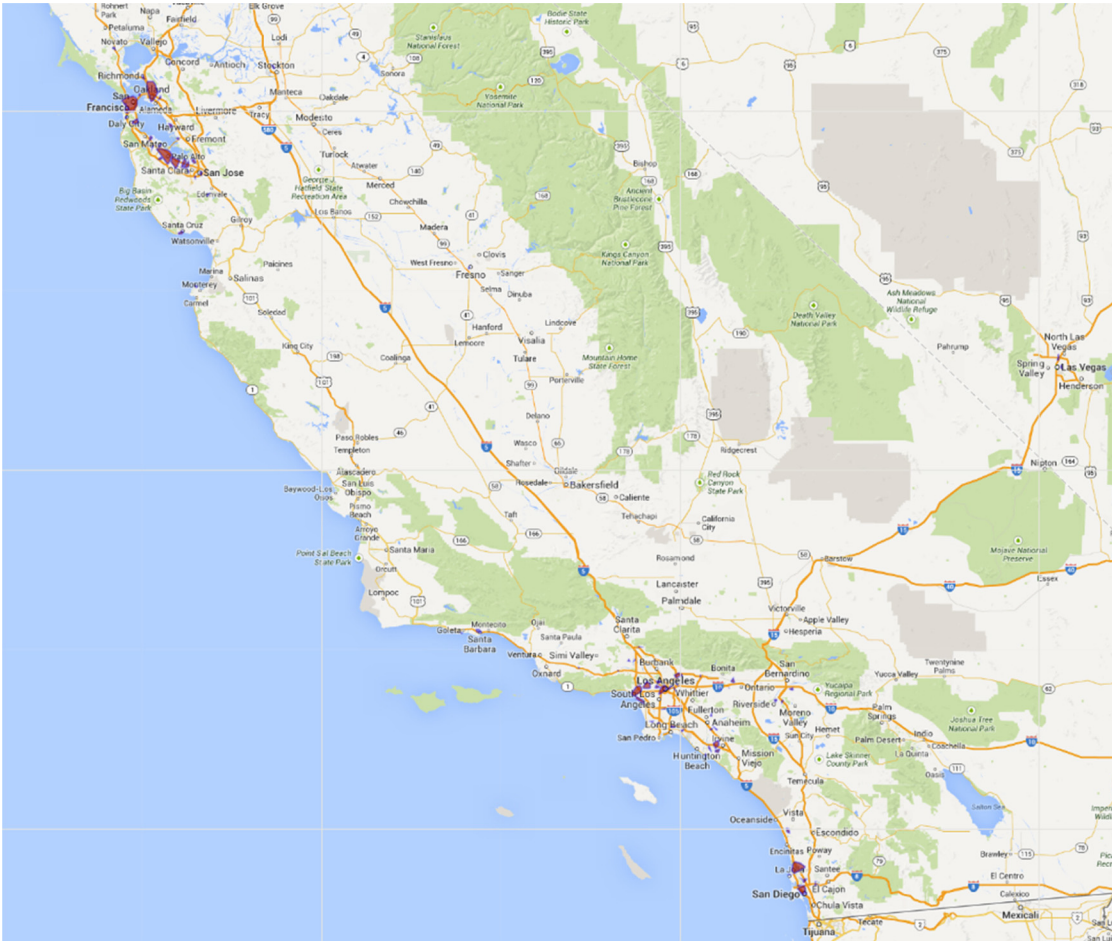
Further work in this area could move in a number of directions. First, different criteria can be used to draw links among nodes within a given distance. Namely, by using alternative data sources that provides multiple affiliations for patents and scientific papers, one could consider the number of co-authored papers as a proxy for real local relationships. In the current implementation of our method we assume that all researches active at a given distance are potentially connected. In future work the set of local connections could be restricted based on three criteria: (1) the analysis can be confined to the network made by actual links between co-located individuals such as collaborative or input–output relationships; (2) the set of potential collaboration at a given point in time can be limited by excluding individuals who are no more active at a given location and (3) only relationships among individuals working in related fields could be considered. Links between scientific and technological fields can be identified based on some population-based measure of relatedness of research areas within firms and in scientific and technological outputs. One more research opportunity is to analyze the evolution of clusters in time. This is essential to fully develop an evolutionary theory of cluster dynamics (Boschma and Frenken, 2006) and to empirically test the predictions about cluster life cycles (Press, 2006; Menzel and Fornahl, 2010). Additionally, the relationship between clusters and productivity deserves further scrutiny. In the present work, we have identified a link between the average impact factor of scientific production and its location in the inner core of top biomedical clusters. This should be considered, however, only as preliminary evidence of the possible presence of locational advantages. More sound evidence should be produced in the future, based on a proper definition of innovative clusters. Another possibility for future work is to use our method to better investigate the dynamics of firm location decisions. Since we cluster both institution names and geographical areas, it is possible to dynamically investigate the impact of firm and individual mobility on regional performances. In this respect, one of the main advantages of our method is that it can be applied to multiple data sources on scientific publications, patents and the location of firms and other research organizations. This allows for the testing of causal relationships between time-dependent location-specific advantages and firm performances. The range of new possibilities to better analyze the relationship between innovation networks and geographical clusters on the empirical ground can also contribute to the design of better policies and a more precise assessment of their effectiveness.
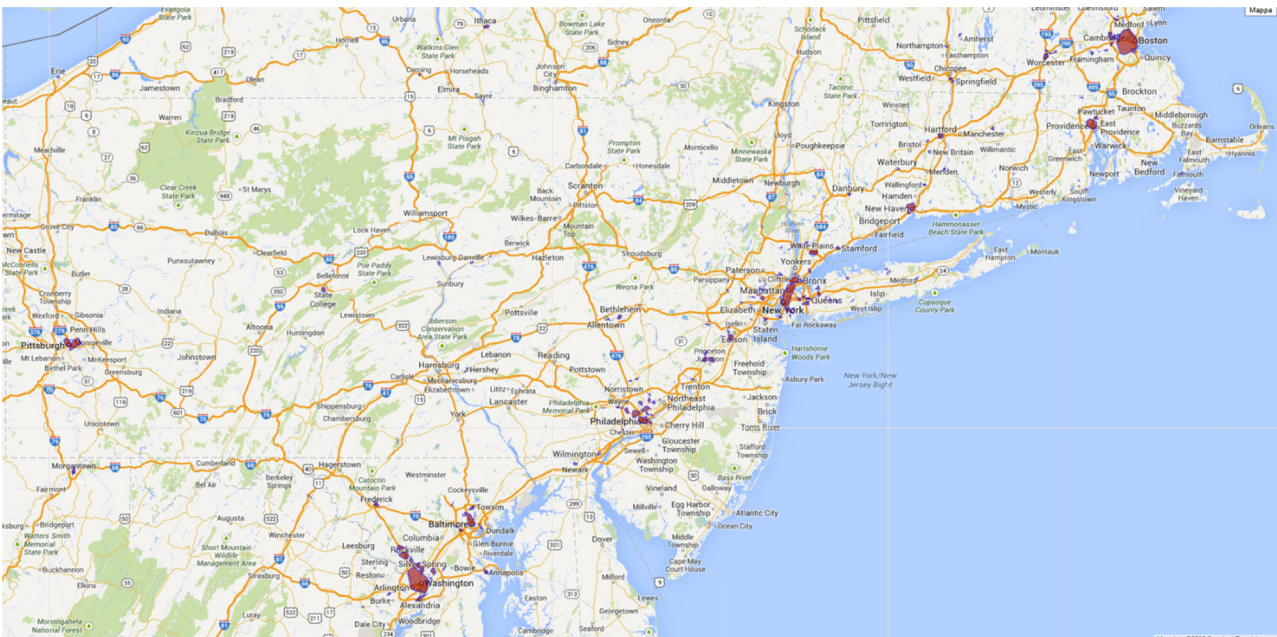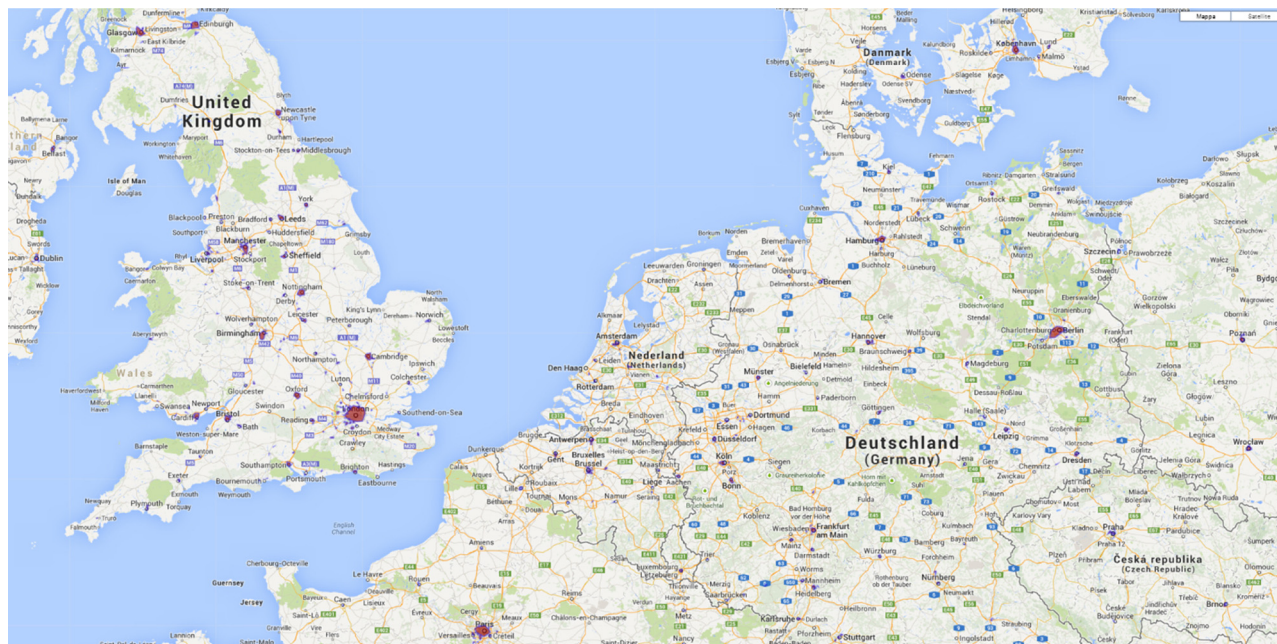
Appendix A. Appendix

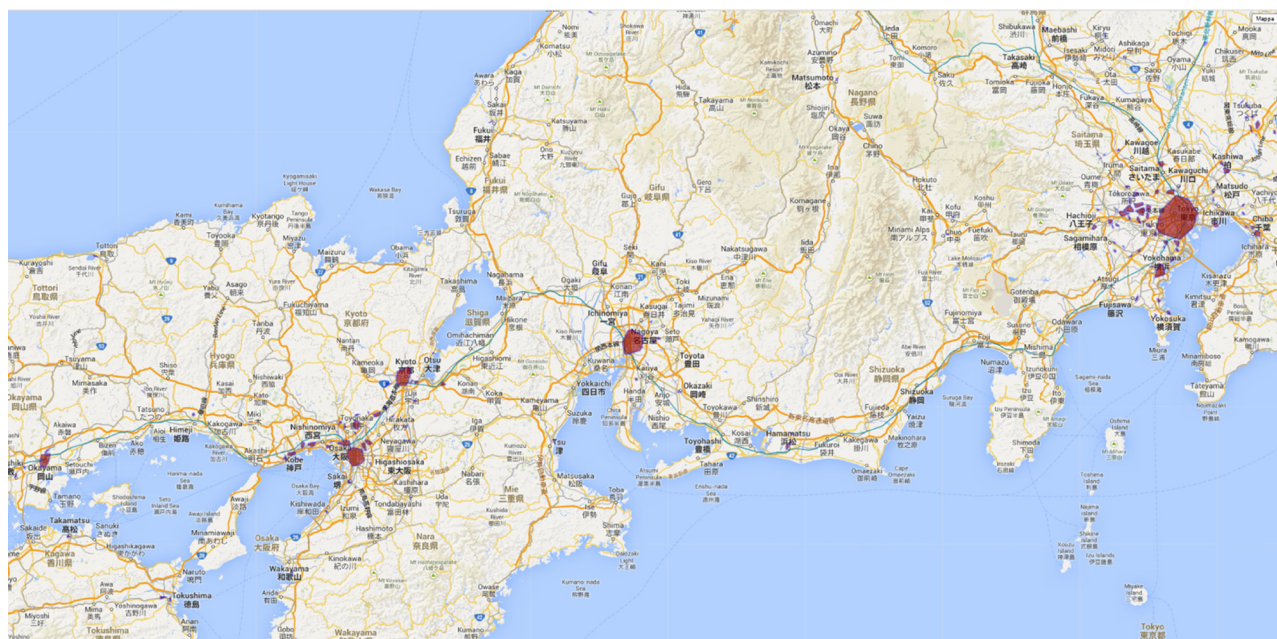## A.1. Major biomedical clusters, US west cost

A.2. Major biomedical clusters, US east cost

A.3. Major biomedical clusters in Northern Europe

A.4. Main biomedical clusters in Japan

References

Alcácer, J., Zhao, M., 2012. Local r&d strategies and multilocation firms: the role of internal linkages. Manage. Sci. 58 (4), 734–753.

Balland, P.-A., Boschma, R., Frenken, K., 2013. Proximity and Innovation: From Statics to Dynamics, Technical Report. Utrecht University, Section of Economic Geography.

Boschma, R.A., Frenken, K., 2006. Why is economic geography not an evolutionary science? Towards an evolutionary economic geography. J. Econ. Geogr. 6, 273–302.

Carmi, S., Havlin, S., Kirkpatrick, S., Shavitt, Y., Shir, E., 2007. A model of internet topology using k-shell decomposition. Proc. Natl. Acad. Sci. 104 (27), 11150–11154.

Cerina, F., Chessa, A., Pammolli, F., Riccaboni, M., 2014. Network communities within and across borders. Sci. Rep. 4, 4546.

Delgado, M., Porter, M.E., Stern, S., 2012. Clusters, Convergence, and Economic Performance, Technical Report. National Bureau of Economic Research.

Delgado, M., Porter, M.E., Stern, S., 2013. Defining Clusters of Related Industries, Technical Report. mimeo.

Diggle, P.J., Chetwynd, A.G., 1991. Second-order analysis of spatial clustering for inhomogeneous populations. Biometrics 47 (3), 1155–1163.

Duranton, G., Overman, H.G., 2005. Testing for localization using micro-geographic data. Rev. Econ. Stud. 72 (4), 1077–1106.

Feser, E.J., Sweeney, S.H., 2000. A test for the coincident economic and spatial clustering of business enterprises. J. Geogr. Syst. 2 (4), 349–373.

Furman, J.L., Kyle, M.K., Cockburn, I., Henderson, R.M., 2005. Public & private spillovers, location and the productivity of pharmaceutical research. Ann. Econ. Stat. 79/80, 165–188.

Jacobs, J., 1961. The Death and Life of Great American cities. Random House, New York.

Jonnalagadda, S., Topham, P., 2010. Nemo: extraction and normalization of organization names from PubMed affiliation strings. J. Biomed. Discov. Collab. 5, 50.

Kistak, M., Riccaboni, M., Havlin, S., Pammolli, F., Stanley, H.E., 2005. Scale-free models for the structure of business firm networks. Phys. Rev. E 81 (4), 036117.

Lai, R., D'Amour, A., Yu, A., Sun, Y., Torvik, V., Fleming, L., 2011. Disambiguation and Co-authorship Networks of the US Patent Inventor Database, vol. 2138. Harvard Institute for Quantitative Social Science, Cambridge, MA.

Marcon, E., Puech, F., 2003. Evaluating the geographic concentration of industries using distance-based methods. J. Econ. Geogr. 3 (4), 409–428.

Marcon, E., Puech, F., 2009. Measures of the geographic concentration of industries: improving distance-based methods. J. Econ. Geogr., 1–18.

Marshall, A., 1920. Principles of Economics. MacMillan, London.

Martin, R., Sunley, P., 2003. Deconstructing clusters: chaotic concept or policy panacea? J. Econ. Geogr. 3 (1), 5–35.

Maskell, P., Kebir, L., 2006. What qualifies as a cluster theory? In: Clusters and regional development: Critical reflections and explorations., pp. 30.

Maskell, P., Malmberg, A., 2007. Myopia, knowledge development and cluster evolution. J. Econ. Geogr. 7, 603–618.

Menzel, M.-P., Fornahl, D., 2010. Cluster life cycles-dimensions and rationales of cluster evolution. Ind. Corp. Change 19 (1), 205–238.

Owen-Smith, J., Riccaboni, M., Pammolli, F., Powell, W.W., 2002. A comparison of us and European university-industry relations in the life sciences. Manag. Sci. 48 (1), 24–43.

Perroux, F., 1950. Economic space: theory and applications. Q. J. Econ. 64 (1), 89–104.

Porter, M., 1990. The competitive advantage of nations. Harv. Bus. Rev. 68 (2), 73–93.

Press, K., 2006. A Life Cycle for Clusters? The Dynamics of Agglomeration, Change, and Adaption. Springer.

Rozenfeld, H.D., Rybski, D., Andrade, J.S., Batty, M., Stanley, H.E., Makse, H.A., 2008. Laws of population growth. Proc. Natl. Acad. Sci. 105 (48), 18702–18707.

Rozenfeld, H.D., Rybski, D., Gabaix, X., Makse, H.A., 2011. The area and population of cities: new insights from a different perspective on cities. Am. Econ. Rev. 101 (5), 2205–2225.

Saxenian, A., 1994. Regional Advantage. Culture and Competition in Silicon Valley and Route, vol. 128. Harvard University Press, Boston, MA.

Schulz, C., Mazloumian, A., Petersen, A.M., Penner, O., Helbing, D., 2014. Exploiting citation networks for large-scale author name disambiguation. EPJ Data Sci. 3 (1), 1–14, http://dx.doi.org/10.1140/epjds/s13688-014-0011-3

Ter Wal, A.L.J., Boschma, R., 2011. Co-evolution of firms, industries and networks in space. Reg. Stud. 45 (7), 919–933.

Torvik, V.I., Weeber, M., Swanson, D.R., Smalheiser, N.R., 2005. A probabilistic similarity metric for medline records: a model for author name disambiguation. J. Am. Soc. Inf. Sci. Technol. 56 (2), 140–158.