Proximity of firms to scientific production

Antonin Bergeaud, Arthur Guillouzouic

PROXIMITY OF FIRMS TO SCIENTIFIC PRODUCTION

Author(s): Antonin Bergeaud and Arthur Guillouzouic

Source: *Annals of Economics and Statistics*, March 2024, No. 153 (March 2024), pp. 105-134

Published by: GENES on behalf of ADRES

Stable URL: https://www.jstor.org/stable/10.2307/48767562

Annals of Economics and Statistics, Number 153, March 2024

PROXIMITY OF FIRMS TO SCIENTIFIC PRODUCTION

ANTONIN BERGEAUD[a] AND ARTHUR GUILLOUZOUIC[b]

Following Bergeaud et al. (2022), we construct a new measure of proximity between industrial sectors and public research laboratories. Using this measure, we explore the underlying network of knowledge linkages between scientific fields and industrial sectors in France. We show empirically that there exists a significant negative correlation between the geographical distance between firms and laboratories and their scientific proximity, suggesting strongly localized spillovers. Moreover, we uncover some important differences by field, stronger than when using standard patent-based measures of proximity.

JEL Codes: O32, O38, R12.
Keywords: Knowledge Spillovers, Technological Distance, Public Laboratories.

1. INTRODUCTION

Public research laboratories are often regarded as crucial building blocks in the advancement of new technologies. Since they are not primarily driven by immediate profit motives like the corporate sector, they play an essential role in generating the scientific knowledge then catalyzed by private innovation. There is compelling evidence in the literature that private firms benefit from academic research. For instance, studies by Azoulay et al. (2019) and Bergeaud et al. (2022) indicate that innovative companies respond to shifts in public research funding by enhancing their R&D effort and output, underscoring the existence of spillovers. However, tracking these knowledge transfers is challenging due to their varied nature, which can range from subcontracting and joint ventures to informal discussions and seminars (see Cohen et al., 2002; De Fuentes and Dutrénit, 2012 for reviews).

In this paper, we assess potential spillovers from French universities to private sectors and characterize their heterogeneity across various scientific domains. To do so, we rely on and generalize the metric of firms' proximity to science introduced by Bergeaud et al. (2022)-initially in the context of evaluating a public research funding program. In particular, we relate this measure of potential knowledge spillovers to industries' spatial concentration. We show a strong negative relationship between the spatial distance separating firms and research labs and their scientific proximity. Delving into scientific disciplines, we find that this pattern holds in most domains, but that magnitudes vary strongly.

The main novelty of the methodology introduced in Bergeaud et al. (2022) is its ability to position each industry within the "scientific space" using a parsimonious set of data on patenting and academic publications. We adopt their methodology and construct a proximity measure that quantifies the likelihood that a firm in industry i draws upon a paper produced by scientific laboratory l. This metric capitalizes on the vast, heterogeneous, and specialized spectrum of academic journals, shedding light on both the typical publication

We thank Émeric Henry and Clément Malgouyres for helpful comments and discussions at early stages of this project. We acknowledge financial support from the French Ministry of Higher Education and Research for this project.

[a]Department of Economics, HEC Paris and CEPR. bergeaud@hec.fr
[b]Paris School of Economics & IPP. arthur.guillouzouic@ipp.eu

outlets of researchers from laboratory l and the scientific sources that firms in industry i rely upon for innovation. Concretely, a public laboratory l and an industry i are deemed scientifically proximate if there exists an intersection between the set of academic journals where researchers from l predominantly publish and the journals frequently cited by patents owned by firms in i. One significant advantage of this measure is its ability to capture spillovers without requiring direct ties between academic research outputs and firms' patents, making it a more encompassing proximity metric between academia and industry than the one usually proposed by the literature.

Employing this methodology, we are able to comprehensively map the potential knowledge spillover network between every pair of laboratory and industry. This network serves as a valuable resource for researchers aiming to examine the transmission of shocks between a public laboratory and the private sector via the exchange of ideas. To the best of our knowledge, our study is the first to propose such an exhaustive overview across all scientific fields and industrial sectors. Analyzing the structure of this network, we find its organization quite intuitive. Two predominant poles emerge: on one side, laboratories specializing in biology, medicine, and immunology closely align with the manufacturing of chemicals and pharmaceutical products. Conversely, engineering and physics labs demonstrate connections with aeronautics as well as electronic and telecommunication equipment industries. Additionally, our findings uncover more nuanced relationships, especially highlighting the roles of mathematics, computer science, and social sciences.

The literature has emphasized a key empirical regularity regarding knowledge spillovers. Regardless of their form, they are mostly concentrated locally and therefore the scientific proximity is tied to the geographical proximity between laboratories and private industries (Abramovsky et al., 2007; Hausman, 2021; Jaffe, 1989). As our second contribution, we thus compare our measure of scientific proximity with the geographical proximity and show that on average, cities surrounding a laboratory tends to be more specialized in industries making use of the science produced by the laboratory. This finding supports the view that geographical proximity matters for knowledge spillovers to occur. However, we also show that there is a large heterogeneity across scientific fields. For example, we find that for materials science, energy, computer science and mathematics, when the distance increases by 1%, the concentration of exposed industries is lowered by 0.2 to 0.3%. In contrast, we do not find any significant association for chemistry or pharmacology.

We then confront our exposure with alternative measures of exposure, more standard in the literature because they are based on academic patents.[1] Specifically, we build a proximity measure comparing alternatively the patent classification IPC at either the 3-digits or the 4-digits level, and the Google embeddings[2] compositions of a public laboratory's patents on the one hand, and an industry on the other hand. We also use direct citations of papers produced by a laboratory as an alternative measure of proximity. We show that these alternative measures of spillovers produce noisier results and typically understate the influence of specific scientific fields, such as mathematics, that are of high importance to produce new technologies but are only rarely the object of a patent. Direct patent citations are an exception and show a strong association with industry concentration in all disciplines, suggesting that, albeit rare, these are a strong signal of local spillovers.

[1] See for instance Akcigit et al. (2021); Hausman (2021); Henderson et al. (1998); Trajtenberg et al. (1997).

[2] Srebrovic, 2019

Overall, our results confirm the relevance of this new measure of scientific proximity between public research and the private sector and pave the way for new research exploiting this network to better understand how firms draw their ideas and how government can best design their R&D and industrial policies. It is an easy to build procedure to capture spillovers, which allows overcoming the sparsity of direct citations found in patents.

Related literature.

The main contribution of our paper is to characterize the French innovation network by looking at scientific relationships between firms and laboratories. The relevance of an innovation network, whereby upstream discoveries influence downstream technologies, has been the focus of several studies. For example, Acemoglu et al. (2016) use citations between patents of different technological classes to map the innovation landscape in the US. However, the structure of their network only allows for citations between patents and therefore mostly concerns the private sector. Conversely, Fabrizio (2009) measures the proximity of firms to universities using the number of academic papers co-published by the two entities. Other papers have considered links from patents to academic articles by exploiting flows of citations (Cristelli et al., 2020). This allowed researchers to emphasize the fact that firms may differ in their reliance to science to develop their technologies and how this may influence the nature of their innovation (Ahmadpoor and Jones, 2017; Marx and Fuegi, 2020; Schnitzer and Watzinger, 2019).

While our approach also relies on such citations, we use a more indirect approach which provides a more complete picture of the network that links research laboratories and private industries. Indeed, our approach does not require existing links that are based on flows of citations, but rather exploits the diversity and specificity of academic journal. Thus, it aims at measuring the relevance of the science produced by a public lab for firms, rather than actual links between these entities. It departs from the two different ways through which the literature has approached this issue. One way has been to look at direct connections between private patents and research output of public laboratories (e.g. Azoulay et al., 2019). Another standard way has been to focus on academic patents, and rely on the proximity between the patent technological classes in which firms and universities apply (see for instance Akcigit et al., 2021; Hausman, 2021; Henderson et al., 1998; Trajtenberg et al., 1997). While patenting within academia has increased over time in France (Carayol and Carpentier, 2021), it is known to capture only a small part of all the knowledge produced by public labs (Agrawal and Henderson, 2002), making it worthwhile to consider all the scientific production of academic units in their potential spillovers.

Since we study the correlation between the scientific proximity and the geographical distance between laboratories and firms, our paper also relates to a rich literature that considers the spatial heterogeneity of innovative activities. Based on various data sources, this literature has highlighted that innovative actors—firms, research laboratories, and universities but also specialized venture capital funds—are geographically organized around clusters (e.g., Delgado et al., 2010; Hausman, 2021). As a result, the modern geography of innovation is characterized by local specialization hubs (Buzard et al., 2017; Egger and Loumeau, 2018) and by the existence of superstar cities that concentrate most innovation (Carlino et al., 2007; Gyourko et al., 2013). This organization is beneficial for innovation for at least two related reasons (Duranton and Puga, 2004). First, as explained previously, distance matters for knowledge spillovers to materialize (Audretsch and Feldman, 1996,

2004; Feldman and Kogler, 2010; Jaffe et al., 1993; Rosenthal and Strange, 2003) even though these spillovers can take many forms that are more or less identified (see e.g., Akcigit et al., 2021; Azoulay et al., 2019 and Aghion and Jaravel, 2015, for a review). The role of distance probably stems from the importance of interactions between scientists, engineers and technicians as a source of creativity (Lychagin et al., 2016) but also because subcontracting is an important channel of knowledge exchange between the public and private sector (Bergeaud et al., 2022). The second reason relates to other types of agglomeration effects, through which innovative activities benefit from being concentrated because of a specialized local labor market and amenities that are useful and valued by innovators (Carlino and Kerr, 2015; Combes and Gobillon, 2015). One natural way to look at these links is to use the network of citations across patents which has been shown to signal the existence of knowledge spillovers (Jaffe et al., 2000) and is often used to highlight flows of ideas (e.g. Aghion et al., 2021; Cotterlaz and Guillouzouic, 2020; Maurseth and Verspagen, 2002). Our approach does not require direct citations, and therefore allows us to capture very flexibly a full range of spillovers from universities to private firms.

Finally, our paper has implications for the funding of innovation activities. While the literature consensually establishes that public research funding stimulates private innovation (e.g. Azoulay et al., 2019; Bergeaud et al., 2022; Fleming et al., 2019; Hausman, 2021; Henderson et al., 1998), we show that there is significant heterogeneity in terms of the centrality of scientific fields. Such results are important in the design of an optimal R&D and industrial policy that factors in the differential sectoral impacts of funding specific public research.

The rest of the paper is organized as follows. Section 2 describes the construction of our proximity measure and presents some descriptive facts about the underlying network. Section 3 analyzes the correlation between the scientific proximity and the geographical distance across subjects. Section 4 concludes.

2. DATA AND CONSTRUCTION OF THE PROXIMITY MEASURE

2.1. *Baseline Scientific Proximity*

We start by constructing the proximity measure between public laboratories and firms as proposed in Bergeaud et al. (2022). The first source of data we use is scanR. scanR is a tool developed by the French ministry of research and innovation (MESRI) that gathers many different sources recording scientific activities of French public and private labs. It provides information on the universe of research papers published by all universities and public research centers in France, and on the journal $j \in \mathcal{J}$ in which they were published.

We further use the patCit database (Cristelli et al., 2020) to retrieve information on the set of papers cited by patents owned by French firms. The patCit project is a collaborative project which use natural language processing tools to retrieve citations included in the text and in the frontpage of all patent publications. The dataset consists in a rich network linking 7,718,253 patents and 3,338,231 distinct academic papers identified by their DOI. On average, a patent cites 3 academic papers (conditional on being in the dataset, i.e. on citing at least one academic paper). We considered any patent as long as the assignee has been identified as a French firm and matched to our data, regardless on the patent office. Finally, we use Crossref[3] to match these DOIs to bibliographical information about

[3]Crossref is a not-for-profit membership organization established to manage scholarly digital content

each article. Table A2 in the Appendix describes the number of papers cited by patents observed in our database (that is, with a defined subject and associated to a Siren).

The first difficulty we face is to assign academic papers to a specific laboratory. Indeed, a typical university is a collection of various research laboratories that work in different scientific fields all located in the same area.[4] There are many dependencies and overlaps across entities in the French public research system, which makes it difficult to assign papers over a period of time to one main stable structure. To circumvent this issue, we define a public research laboratory as a combination of a city c and a research domain d. We use a classification of research into 18 large domains[5] that are manual aggregations of the 352 subjects that are assigned by Crossref.[6]

This defines 1260 public laboratories $l \in \mathcal{L}$ that are located in 206 different cities. We further restrict the sample by removing journals that are too generalist and laboratories that are too small in terms of their number of publications. The procedure is described in Appendix A. Ultimately, our final sample counts 370 laboratories that are matched to firms in 145 5-digit industries (using the NACE classification).[7]

We then construct a measure of proximity as follows:

$$\text{prox}_i^{(l)} = \sum_{j \in \mathcal{J}} \eta_{l,j} \gamma_{j,i}, \tag{1}$$

where $\eta_{l,j}$ is the share that journal j represents in the publications by laboratory l between 2013 and 2020 and similarly, $\gamma_{j,i}$ is the share the patents owned by firms in industry i in overall patent citations of publications in journal j. In other words, this proximity measure combines the probability that a lab l produces knowledge in a given journal j with the probability that an industry i cites a paper in such journal j, and sums this over all journals. We consider patents first filed before 2018. These shares are such that:

$$\sum_{i \in \mathcal{I}} \gamma_{j,i} = 1 \text{ and } \sum_{j \in \mathcal{J}} \eta_{l,j} = 1$$

through Digital Object Identifiers (DOIs). It facilitates the registration of DOIs for academic publications and other research outputs to ensure consistent referencing and linking across different platforms. Beyond its core function of DOI registration, Crossref also provides tools and services for metadata retrieval and content tracking.

[4]Whenever several affiliations corresponding to several cities are found, we use a fractional count approach.

[5]These domains are Agriculture, Arts and Humanities, Business, Chemistry, Computer Science, Energy, Engineering, Environmental Science, Immunology and Microbiology, Materials Science, Mathematics, Medicine/Dentistry, Neuroscience/Psychology, Nursing/Paramedical, Pharmacology, Physics and Astronomy, Social Sciences and a last domains for all other fields. See Appendix A.2 for more details.

[6]This means that in some cases, we may merge different universities into a given entity if two different universities are located in the same city and work on the same domain. This will be essentially an issue in large cities such as Paris. We check that our results are not affected by removing the Paris area from the sample, see Section 3.4.

[7]The reduction in the number of labs might seem significant, as it represents 70% of the observations (though only 17% of the total number of papers). In reality, most of the laboratories that are removed have fewer than 10 papers (62%), and 25% have only one paper. These small laboratories likely correspond to smaller research groups or may simply result from errors in the affiliation or subject. They are over-represented in the subject categories "Arts and Humanities", "Others", "Energy", and "Nursing/Paramedical."

Interpretation.

Following the argument presented by Bergeaud et al. (2022), the metric $\text{prox}_i^{(l)}$ quantifies the degree to which a specific industry, denoted as i, draws its knowledge from a similar scientific domain as that typically produced by laboratory l. Conditional on publications in a given scientific journal being sufficiently homogeneous, this metric essentially represents the likelihood that a paper, when published by lab l, aligns with the knowledge requirements of industry i. Consider, for instance, a laboratory dedicated to nanotechnology research; it is poised to publish in a distinct set of journals that precisely demarcate its research focus. If a patent from a firm frequently cites ideas from these specific journals, then we would assign a strong proximity between the firm's industry and the given laboratory. It is then possible to gauge the potential impact or relevance of a lab's total output for industry i by simply multiplying this $\text{prox}_i^{(l)}$ by the lab's total number of published papers.

2.2. *Description of the Measure*

Table I identifies the 20 pairs of industry i and scientific domains d that have the highest average proximity. Formally, we calculate:

$$\text{prox}_i^{(d)} = \left(\sum_{l \in \mathcal{L}(d)} \text{N}_l \right)^{-1} \sum_{l \in \mathcal{L}(d)} \text{prox}_i^{(l)} \cdot \text{N}_l,$$

where $\mathcal{L}(d)$ is the subset of laboratories in $\mathcal{L}$ which corresponds to a domain d, and N_l is the number of papers from lab l. This formula simply calculated the average proximity of each scientific field d, weighting by the relative size of each laboratory. Therefore, it is simply a way to map the formula described in equation (1) at the industry $\times$ domain level. We can see that some industries appear several times, such as "Manufacture of perfumes and grooming preparations" and "Manufacture of air and spacecraft and related machinery", which results from their high centrality and reflects the fact that they are tightly connected to different scientific fields, as well as their size in the French economy. In terms of scientific fields, while some fields appear more frequently than others (e.g. "Materials science"), strong links are relatively well spread across fields as 11 different fields appear among the top 20 links. Other sensible links appear, for example laboratories in the field of agriculture are strongly connected with the manufacture of bread and pastry.

Network representation.

In order to better understand the proximity measure, we use it to define links in a network of connections. Indeed, the matrix of proximities $\text{prox}_i^{(l)}$ between industries and laboratories can be seen as a weighted directed graph summarizing the potential transfers of knowledge from laboratories to industries.[8] A first way to gauge the credibility of our proximity measure is to (visually) observe the extent to which labs of the same scientific field tend to cluster in space, meaning that they are linked with similar intensities to

[8]To spatialize the network, we use a standard force-based layout algorithm, ForceAtlas, which is built-in in the software Gephi.

TABLE I

TOP 20 INDUSTRIES AND SUBJECTS IN TERMS OF AVERAGE LAB PROXIMITIES

Subject	Industry		Av. proximity
Energy	Manufacture of scientific and technical instruments	2651B	0.439
Physics and Astronomy	Manufacture of air and spacecraft and related machinery	3030Z	0.260
Computer Science	Manufacture of aid to navigation equipment	2651A	0.221
Chemistry	Manufacture of perfumes and grooming preparations	2042Z	0.218
Immunology and Microbiology	Manufacture of pharmaceutical preparations	2120Z	0.196
Agriculture	Industrial manufacture of bread and fresh pastry	1071A	0.178
Pharmacology	Manufacture of perfumes and grooming preparations	2042Z	0.162
Immunology and Microbiology	Manufacture of perfumes and grooming preparations	2042Z	0.152
Materials Science	Manufacture of electronic components	2611Z	0.109
Pharmacology	Manufacture of pharmaceutical preparations	2120Z	0.105
Materials Science	Manufacture of aid to navigation equipment	2651A	0.102
Medecine/Dentistry	Manufacture of pharmaceutical preparations	2120Z	0.101
Computer Science	Manufacture of air and spacecraft and related machinery	3030Z	0.101
Materials Science	Manufacture of air and spacecraft and related machinery	3030Z	0.0980
Medecine/Dentistry	Manufacture of perfumes and grooming preparations	2042Z	0.0954
Engineering	Manufacture of air and spacecraft and related machinery	3030Z	0.0930
Physics and Astronomy	Manufacture of electronic components	2611Z	0.0837
Neuroscience/Psychology	Manufacture of perfumes and grooming preparations	2042Z	0.0725
Engineering	Manufacture of electricity distribution and control apparatus	2712Z	0.0720
Materials Science	Manufacture of glasses	3250B	0.0714

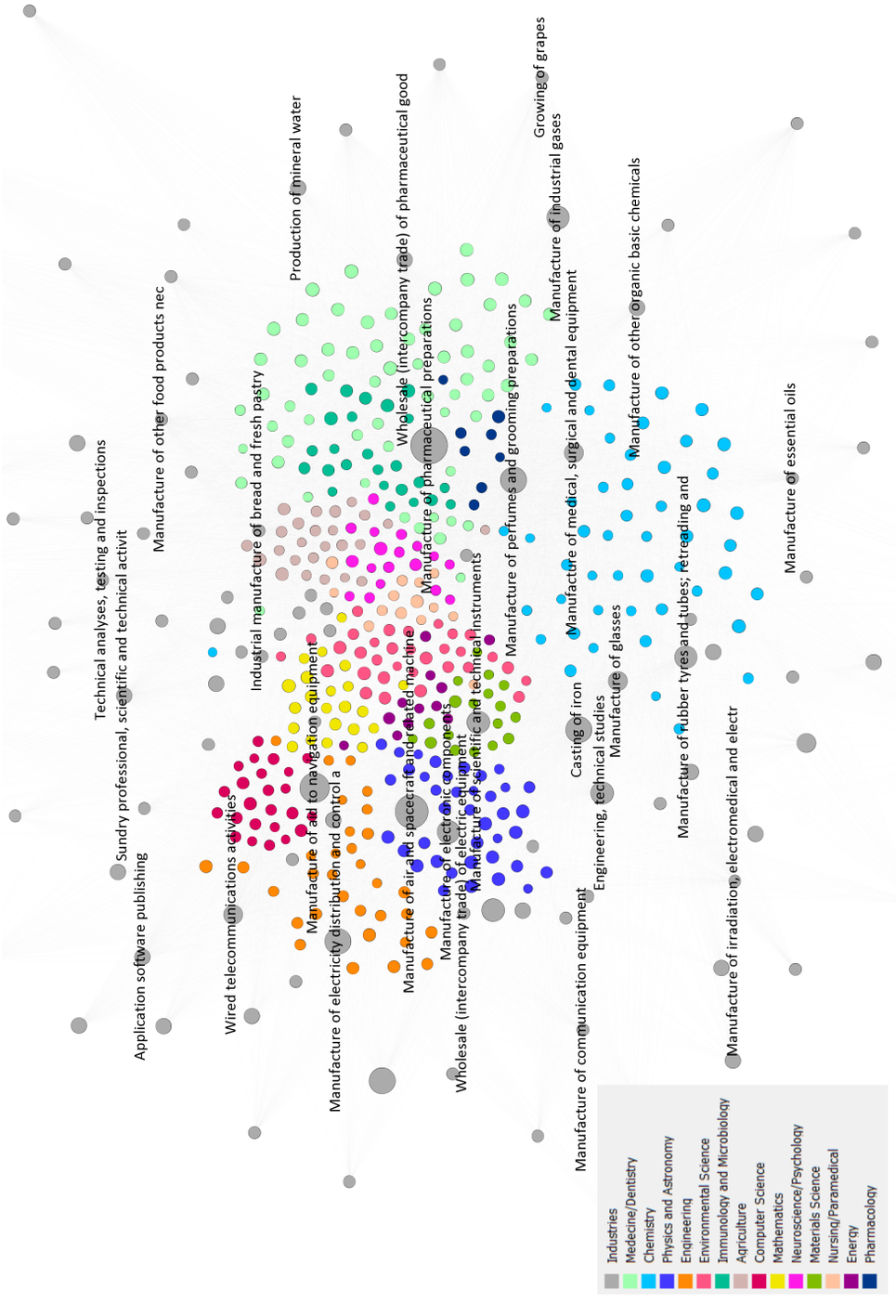
Notes: This table shows the top 20 links between subjects and industries in terms of average proximity across labs. Sources: scanR, patCit, Patstat.

similar industries.

Results are presented in Figure 1. Industries in the network are depicted as light gray dots, whose size reflects (not proportionally) their stock of patents. Similarly, labs are represented as colored dots according to their scientific subject, and the size of dots reflects (not proportionally) the number of papers they published. Only the labels of the industries benefiting from the largest overall proximities are displayed on the network for the sake of clarity.

We can already draw several conclusions from the relative position of labs in this graph. First, it appears very clearly that labs in a same subject (dots of a same color) tend to cluster in the same area, meaning that the composition of their proximities are relatively similar. It is important to note that this is not an a priori feature of the graph but results from the fact that these laboratories publish in similar academic journals. Second, the graph is generally oriented along a West–East axis: the eastern end contains mostly medical and health related labs, while the western end is mostly populated by engineering, physics and computer science labs. In the center of the network, clusters of labs in energy, environmental science, materials science and mathematics are closer to the engineering pole, while chemistry, agriculture and neuroscience specialties are closer to the medical eastern pole. Chemistry has a peculiar position, as it is much more spread out than other disciplines, and that some labs seem much closer to the medical pole, while others are closer to the center of the graph and to labs in physics.

Figure 1: Network representation of the scientific proximity measure



Notes: This Figure represents the directed network of scientific proximity links between 5-digits industry codes (NAF subclass) and French public labs (city × subject pairs). The labels of the industries ranking in the top 25% total proximities are displayed. Each grey dot represents an industry: the size of these dots reflects the patent stock of this industry. Each colored dot represents a public lab: each color is associated to a given subject (see legend), the size of dots reflects the number of papers published by the lab. The spatialization of the graph is obtained through of force-based layout (ForceAtlas in software gephi). Labels refer to the grey dot which is exactly at the center of the label. Sources: scanR, patCit, Patstat.

The position of industries relative to labs also appears quite intuitive. In the middle of the medical eastern pole of the network are three very large industries: the production of perfumes and toiletry, the production of pharmaceutical products, and the production of medical equipment. These are quite central within the medical pole, which reflects the fact that they draw knowledge from many different subjects: medicine, chemistry, microbiology and pharmacology. Some industries are closer to labs in one specific subject: the manufacture of industrial gases is tightly related to medicine, the growing of grapes is interestingly much closer to immunology and medicine than to agriculture, the manufacture of essential oils is closest to chemistry, the manufacture of bread and of food products are closest to labs in agriculture. In the engineering/physics western pole, we also observe very large industries being located in the center of the pole and therefore benefiting from physics, mathematics, computer science and engineering, such as navigation equipment, the aeronautical industry, and the electric industry. Finally, industries located close to the center of the network are interesting cases as well: the manufacture of glasses and of rubber are between the medical and the engineering poles, which probably benefit both from knowledge coming from engineering and from knowledge stemming from research in chemistry. The periphery of the network also includes generalist industry codes, such as technical analyses and testing, engineering and technical studies, manufacture of scientific instruments, etc.

Centrality and clustering measures.

To give more quantitative sense of the results previously described, we run a simple k-means clustering algorithm on the network and report the outcome in Table II where we used $k = 5$. k-means clustering aims at creating a partition of a network made of k clusters by minimizing the within cluster variance of the weighted adjacency matrix, where in our case, weights are proximities. This partition is constructed without any prior but as shown in Table II, the corresponding scientific fields composition of each cluster echoes what can be visually assessed in Figure 1.

In particular, the first cluster corresponds to the medical pole of the network, as it is composed almost entirely of labs in chemistry, immunology, medicine and pharmacology. The second cluster is centered around agriculture. The third cluster corresponds to what we called the engineering pole previously, and contains almost all the labs in physics, computer science, materials science and engineering. A fourth cluster contains labs in energy and environmental science. Finally, a fifth group emerges with miscellaneous fields. Overall, this exercise of k-means clustering confirms the interpretation which could be made from the visual inspection of Figure 1.

Finally, Table B1 in Appendix B presents various aggregate measures of centrality by subject. Consistently with Figure 1, we observe that chemistry and physics are the two fields with the highest level of centrality, which means that they are closely connected with a wider range of different industries. Indeed, chemistry will impact at the same time industries related to the manufacture of pharmaceutical products and manufacturing industries such as metallurgy.

TABLE II
CLUSTERING

Subject	Cluster	Nbr of labs in cluster	Share of subject in cluster
Cluster 1			
Chemistry	1	83	28.8%
Immunology and Microbiology	1	79	27.4%
Medicine/Dentistry	1	65	22.6%
Pharmacology	1	41	14.2%
Agriculture	1	12	4.2%
Neuroscience/Psychology	1	5	1.7%
Cluster 2			
Agriculture	2	36	85.7%
Nursing/Paramedical	2	5	11.9%
Environmental Science	2	1	2.4%
Cluster 3			
Physics and Astronomy	3	66	30.0%
Computer Science	3	63	28.6%
Materials Science	3	45	20.5%
Engineering	3	38	17.3%
Mathematics	3	8	3.6%
Cluster 4			
Energy	4	27	67.5%
Environmental Science	4	11	27.5%
Computer Science	4	1	2.5%
Chemistry	4	1	2.5%
Cluster 5			
Environmental Science	5	79	17.8%
Social Sciences	5	61	13.7%
Medicine/Dentistry	5	61	13.7%
Mathematics	5	53	11.9%
Engineering	5	45	10.1%
Neuroscience/Psychology	5	43	9.7%
Nursing/Paramedical	5	42	9.4%
Agriculture	5	28	6.3%
Materials Science	5	9	2.0%
Chemistry	5	8	1.8%

Notes: This table shows the results of the k-means clustering algorithm using 5 clusters on the lab–industry network. Only subjects with more than 1% of total cluster share are presented. Sources: scanR, patCit, Patstat.

3. SCIENTIFIC PROXIMITY AND GEOGRAPHICAL DISTANCE

3.1. *Aggregation of the proximity measure*

The central goal of the paper is to show how scientific proximity, as captured by the measure introduced above, depends on geographic distance. Such analysis requires a mapping between indsutries and spatial units that we build using a measure of the concentration of industry i in each city c.[9] We use the business register of establishments ("Répertoire SIRENE") and compute the share of plants in industry i located in city c, which we denote $w_{i,c}$.

From this, we construct a measure of the technological proximity between each pairs of city c and c' as:

$$A(c,c') = \sum_{l \in \mathcal{L}} \mathbb{1}[c(l)=c] \sum_i \text{prox}_i^{(l)} w_{i,c'},$$

where $c(l)$ denotes the city in which laboratory l is located. Said differently, we sum the total spillovers received by firms in all industry and located in city c' from all laboratories located in city c and weight this sum by the share of industry i in c'.

We also construct a similar measure but restricting to each scientific domain d:

$$A^{(d)}(c,c') = \sum_{l \in \mathcal{L}(d)} \mathbb{1}[c(l)=c] \sum_i \text{prox}_i^{(l)} w_{i,c'}.$$

This measure will be high for pairs of cities c, c' such that city c' has a high share of economic sectors which use the same science as the one produced by the laboratory located in city c. Of course, $A^{(d)}(c,c')$ is likely to reflect the fact that both technology intensive industries and public research laboratories are more likely to be concentrated in dense urban areas. This would increase the measure of the technological proximity between two cities c and c' but for reasons unrelated to actual knowledge spillovers from the university to the private firms.

We thus construct counterfactual weights $w_{i,c'}$ that would reflect the share of industry i in city c' if the spatial distribution of industries was random, keeping the same number of establishments in each city.[10] Using these weights, we construct an alternative measure $B^{(d)}(c,c')$ that we will use as a control variable. This is akin to the "dartboard" approach introduced by Ellison and Glaeser (1997), with the exception that we resample only the industries which have a non-zero scientific proximity.

For each city c', we construct the total exposure received from laboratories in a given scientific domain d by summing over the values of $A^{(d)}(c,c')$ for all c, where c denotes the location of a laboratory.[11]

This defines our dependent variable, which we denote $a_{c'}^{(d)}$, as the log of total proximity to scientific labs in domain d. We proceed similarly for the counterfactual $B^{(d)}(c,c')$

[9] We consider "intercommunalité" as our measure of city. These are a larger entities than "communes" which are on average very small (around 36,000 communes and 1250 intercommunalité).

[10] We only randomize the location of establishments that are in an industry with at least one positive proximity and kept other establishments is their actual location.

[11] In the baseline, we do not weight the observation in taking the sum. One alternative would be to weight this sum by the number of papers published by laboratories l located in city c and in field d to take into account that larger laboratories are likely to generate more spillovers. This only marginally impacts our result, see Figure B1 in Appendix B.

and use the corresponding logarithm of the unweighted sum $b_{c'}$ as a control variable. As explained above, this controls for the potentially endogenous location of some specific industries and laboratories. Finally, we also control for the logarithm of the population of city c', $p_{c'}$.

The independent variable of interest is the weighted average geographical distance with nearby laboratories working on subject d that we denotes $\delta_{c'}$. We only consider laboratories located less than 100km away as a baseline and use the number of papers they publish as weights. Section 3.4 explores how our results are affected by these choices. We therefore estimate the following model for all cities c' (around 1200 observations):

$$a_{c'}^{(d)} = \alpha_d b_{c'}^{(d)} + \beta_d \delta_{c'} + \gamma_d p_{c'} + \varepsilon_{c'}^{(d)}, \tag{2}$$

where $\varepsilon_{c'}^{(d)}$ is an error term which is allowed to be heteroskedastic.

We estimate equation (2) for each scientific domain d, and collect the estimated value of β_d as well as its standard errors. β_d captures the strength of the link between scientific and geographical proximity. The more negative β_d, the more industries are concentrated around labs with which they are technologically close.[12]

3.2. *Results*

Figure 2 plots the value of β_d for each scientific discipline d, when estimating regression (2). It shows that disciplines which have the highest concentration of technologically close industries located around them include many of the disciplines which appeared in the second cluster in the network analysis, namely materials science, computer science, mathematics, physics and to a lesser extent engineering. Thus, an important feature of our measure is its ability to detect concentration around disciplines which produce primarily basic knowledge and therefore probably issue few patents directly, although they publish knowledge in journals which are cited by private sector patents. Energy and environmental science, which were grouped in the fourth cluster in the network analysis, also appear to have a high extent of concentration of affected industries around them.

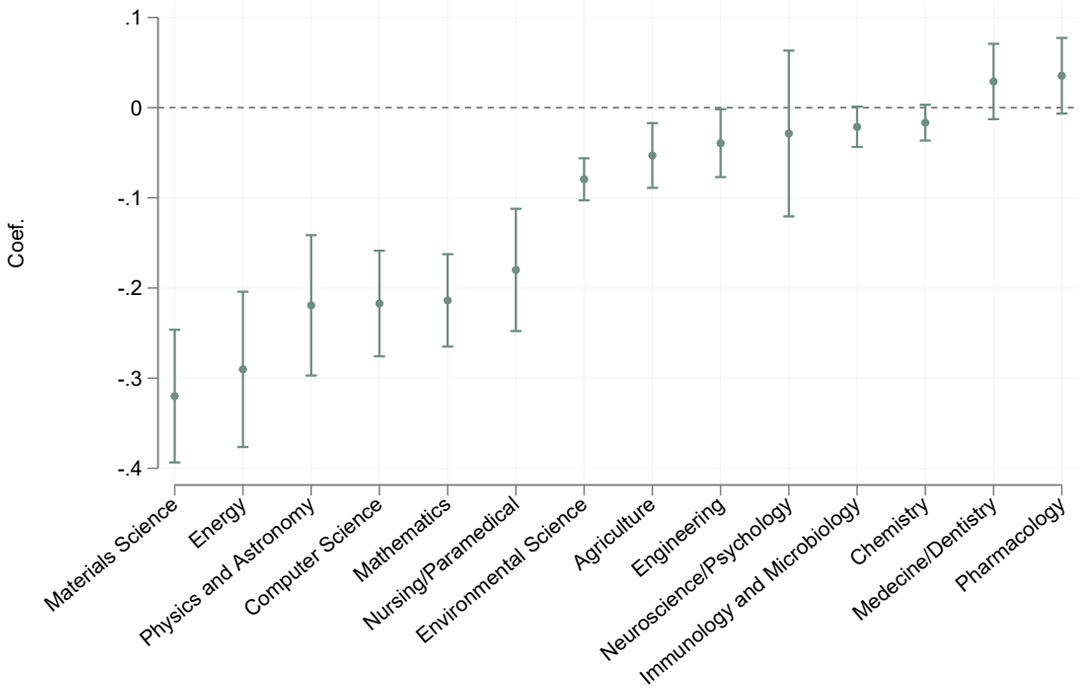
An important takeaway of this analysis is however that, with the notable exceptions of nursing research (which is fairly small discipline) and immunology/microbiology, labs in the medical and chemical areas (medicine and dentistry, neuroscience and psychology) are not located next to industries which are close technologically. This is in contrast with a part of the literature arguing that effects of university research on the private sector are particularly strong in these areas (Abramovsky and Simpson, 2011; Abramovsky et al., 2007; Azoulay et al., 2019). This contrast could stem from the specific geography of innovation in this industry in France or reflect the the fact that, while strong, these spillovers are not localized. In any case, reconciling our results with insights from the previous literature is an interesting avenue for further research.

3.3. *Comparing with other measures*

We now examine how the results change when we use alternative measures of proximity. The literature typically looks at more direct connections between private patents and

[12]To get a sense of what the regression is capturing, we map the value of $a_c^{(d)} - b_c^{(d)}$ for all cities c for the scientific field Materials Science in Figure B3 along with the location of laboratories in this same domain.

Figure 2: Correlation between scientific proximity and geographical distance, by field



Notes: This figure plots the coefficients and confidence intervals of a linear regression run by scientific field of equation (2). Sources: scanR, patCit, Patstat.

research output of public laboratories (e.g. Azoulay et al., 2019). One natural way to look at these links is to use the network of citations across patents which has been showed to signal the existence of knowledge spillovers (Jaffe et al., 2000) and is often used to highlight flows of ideas (e.g. Aghion et al., 2021; Cotterlaz and Guillouzouic, 2020; Maurseth and Verspagen, 2002).

In our case, however, constructing such measures of proximity between industries and laboratories is challenging for two main reasons. First, it requires identifying the list of patents published by researchers in each laboratories. Second, it is likely that these patents receive few citations from patents filed by French firms, making the citation network very sparse and noisy.

Regarding the first challenge, we select patents whose assignee in scanR is a French university. However, universities typically contain many different laboratories that work on various topics and it is impossible to match each patent to a specific laboratory. Conversely, a given laboratory can belong to different universities. Just as we did previously, we would like to define a laboratory as a pair of city and scientific domain. Yet, we cannot use the publication of academic articles to assign a domain as we only have information on the patents produced by the laboratory. One natural approach would be to use the IPC (International Patent Classification) classes that split patents into different categories based on the type of technology and techniques that they cover. However, IPC classes are very different in nature from scientific fields and are impossible to match with our list of 18 subjects. We therefore proceed differently and match patents to fields based on a list of keywords defining the scientific subjects and mentioned in the patents. More details are given in Appendix A.3. At the end of the procedure, we define 470 laboratories in 93 cities.

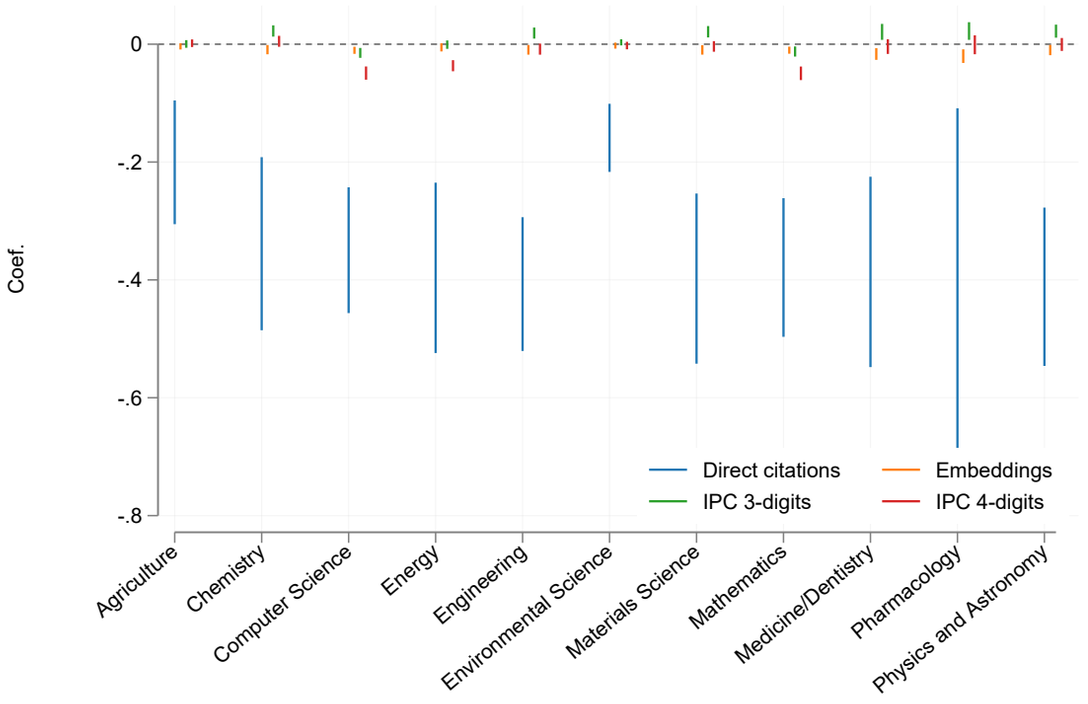
Regarding the second challenge, we start by calculating the number of citations received by each laboratory from patents filed by French firms. We found 1468 links between industries and laboratories (63 industries and 396 laboratories).

Given the scarcity of direct links, we complete our analysis by relying on more indirect alternative measures of proximity between industries and laboratories that are based on the similarity of their respective patent portfolios. We consider 3 such measures of proximity: first we use the correlation between the average embedding representation of all patents filed by each laboratory and each industry.[13] Second, we construct a distance based on the similarity between the sets of IPC classes that appear in all patents filed by industries and all patents filed by laboratories. Our proximity thus maximizes if two entities filed patents in the exact same sets of IPC classes with the exact same weights. We use 3-digit IPC classes and also experiment with 4-digit IPC classes.

In all cases, we follow the same methodology as in 2.1 but replace the value of prox by each of these new alternative measures.

[13]See Bergeaud et al. (2022) for more details on embedding representations of patents. Formally, the text of each patent is represented by a real vector of 64 dimensions constructed such that the dot product between the embedding of two different patents measure their similarity (in the sense that they are more likely to share the same technology classes, see Srebrovic, 2019). Our measure of proximity is thus simply the dot product between the average embedding vector of all patents filed by each firms in a given industry and the average embedding vector of all patent filed by each laboratory.

Figure 3: Correlation between alternative proximities and geographical distance, by field



Notes: This figure plots the coefficients and confidence intervals of a linear regression run by scientific field of equation (2) using 4 alternative measures of proximities as defined in Section 3.3. Sources: scanR, patCit, Patstat.

Results.

Results are presented in Figure 3 which, as in Figure 2, presents the coefficient β^d along with their 95% confident intervals for each domain—and for each of our 4 alternative measures. We can make several observations from this Figure. First, applying the same procedure as the one described in Section 3.1 only allows to identify a coefficient for 11 subjects. This highlights the broad coverage of the innovation network allowed by our main proximity measure which is one of its key advantages. Second, for the measures based on 3 or 4 digit IPC and the embedding proximity, the correlation with distance is always very small especially compared to the baseline. Third, the results using direct citation links is very strong in the sense that the correlation with distance is always negative, of a similar order of magnitude than the baseline but at the same time not very precisely estimated and without any significant differences across subjects. Moreover, this result is expected in the sense that the flows of citations between patents only transcribe the existence of concrete links between entities, which naturally increases the probability of colocation. We view these results as evidence that our baseline measure of proximity is better at capturing actual spillovers.

3.4. *Robustness*

In this section, we estimate equation (2) but with alternative constructions of the dependent variable a, and as a result of b, or on a restricted sample to assess the robustness of our results.

Our first robustness check is to change the criteria we apply to select journals based on how general they are. In the baseline model, we calculate for each journal the Herfindahl-Hirschmann Index (HHI) across these 18 scientific fields by counting the number of papers in each field. We remove any journal with a HHI lower than 0.5 which we consider as too generalist or multidisciplinary. In this section, we look at what happens if this threshold is reduced to 0.3 (allowing for more journals) or increased to 0.7. These tests are referred to as "Herf 30" and "Herf 70" respectively.

As another robustness test, we change the threshold distance value of 100km that we use to calculate the average distance $d_{c'}$. This threshold is changed to 50 and to 150 and the tests are respectively denoted "Dist 50" and "Dist 150".

We also change the sample by including journals that have only one relationship to an industry ("All links"), by including the R&D sector ("Inc. 72"), by having a minimum size of labs of 50 rather than 100 papers ("Labsize 50"), by considering the closest lab rather than the average ("Min dist"), and by removing the Paris area ("Remove paris").

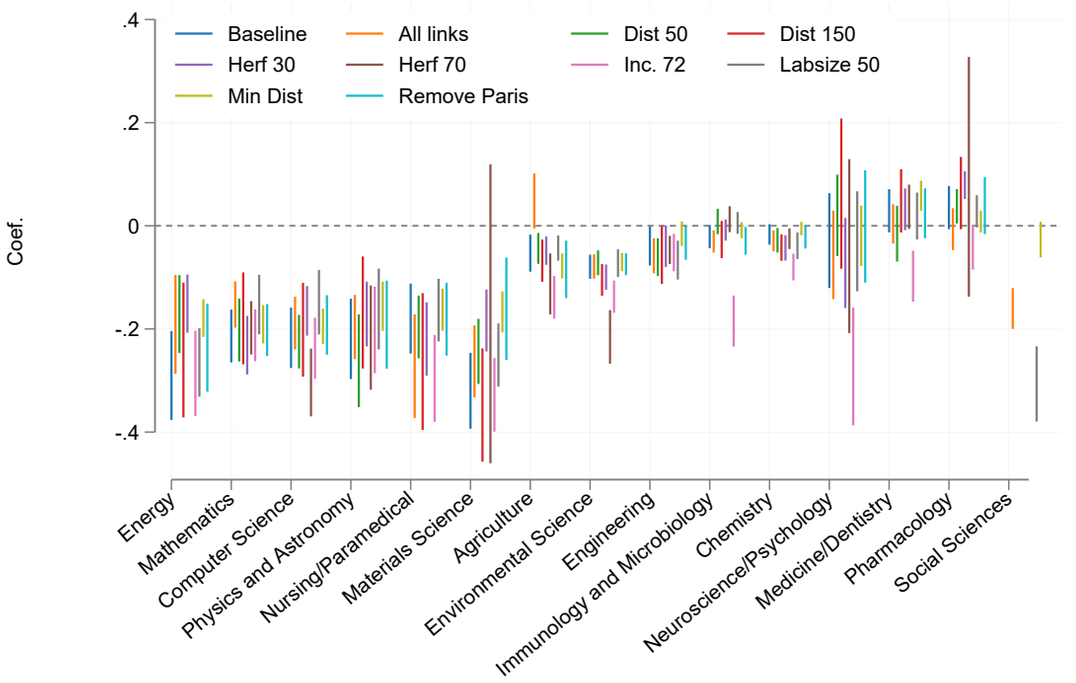
Figure 4 shows the results of submitting our baseline analysis to this battery of robustness tests. Figure B2 in Appendix B presents the results separately. Overall, a notable feature of this exercise is that the estimated coefficients are negative in the vast majority of the cases. The fields which display the greatest concentration around them (energy, mathematics, computer science, physics, nursing) are very robust to changes in the various thresholds we impose through the procedure, which supports the idea that our baseline result is not very sensitive. The rest of the fields tend to be associated with negative but smaller coefficients in absolute terms and often insignificant, confirming that there is little excess concentration of close industries around those labs. Finally, a coefficient associated to social sciences is only estimated under some of the scenarios, and takes very volatile values.

4. CONCLUSION

In this paper, exploiting the methodology proposed in Bergeaud et al. (2022), we construct a measure of scientific distance between public research laboratories and private firms in France. Based on this measure, we build the network linking industries and scientific fields. In the process, we document the existence of clusters in which scientific information circulates, around large themes such as medical applications or agriculture. An interesting feature of our results is to show the importance of fields, such as mathematics, that tend not to patent and yet produce knowledge that is used by firms. Such proximities cannot be captured by more traditional measures, in particular those based on academic patents.

The main contribution of the paper is to show a strong link between scientific and geographic proximity. Interestingly the strength of this correlation very much depends on the scientific field in which laboratories publish. It is very strong for material science or computer science, but rather loose for chemistry and medical science. This could be reflecting the fact that different channels underly spillovers depending on the academic subject. Indeed, some channels such as research subcontracting do not necessarily require the firm and the lab to be located close to each other, while others such as labor mobility

Figure 4: Robustness tests on the correlation between scientific proximity and geographical distance, by field



Notes: This figure plots the coefficients and confidence intervals of a linear regression run by scientific field of equation (2), for many different sample restrictions and sensitivity checks. Sources: scanR, patCit, Patstat.

are arguably more local. Explaining the source of this heterogeneity across fields will be an interesting avenue for future research.

APPENDIX A: DATA APPENDIX

A.1. *Data cleaning for baseline measure*

Our initial data comes from ScanR and contains 324,050 different papers (with a non empty doi and an author that is geolocated) published in 13,033 journals between 2013 and 2021 by a French public entity. A large share of these journals are small (for example proceedings of a conference) some are very specialized and other more generalist. We use the API provided by Crossref to assign each paper a scientific field. The main classification counts 352 subjects and we manually aggregate these subjects into a broader category of 18 groups: Agriculture, Arts and Humanities, Business, Chemistry, Computer Science, Energy, Engineering, Environmental Science, Immunology and Microbiology, Materials Science, Mathematics, Medicine/Dentistry, Neuroscience/Psychology, Nursing/Paramedical, Pharmacology, Physics and Astronomy, Social Sciences and an additional group for all other fields. The detailed composition of each of these 18 subjects is given in Section A.2.

From this, we proceed to a number of cleaning steps. The corresponding number of papers in each field is given in Table A1, starting from Step 1 which corresponds to the output from Crossref. We first calculate an index of specialization for each journal, defined as the Herfindahl-Hirschmann Index (HHI) across these 18 scientific fields. We remove any journal with a HHI lower than 0.5 which we consider as too generalist or multidisciplinary to help identifying a laboratory. This leaves 11,346 journals (Step 2).

On the private sector side, we start from 24,112 patents filed by 1,599 firms between 2000 and 2015. These patents cite 42,324 different papers that are published in 4,736 journals. We remove single relationships (industries cited only by 1 journal), generalist journals (from the HHI) and public sector as well as 2-digit industry 72 (R&D) and some specific industries such as holdings. At the end, our working database counts 3,239 journals and 179 industries (Step 3).

We further remove laboratories with less than 100 papers and subjects that have an aggregate proximity below 1.This in particular leaves out three subjects: Arts and Humanities, Business and Other. Merging all this together yields 12,305 relationships between 370 labs and 145 industries (Step 4).

A.2. *Building larger subjects*

This section describes the 18 scientific subjects used in the analysis by listing the Crossref fields included in each of them.

Agriculture:

"Agricultural and Biological Sciences (miscellaneous)", "Agronomy and Crop Science", "Animal Science and Zoology", "Aquatic Science", "Developmental Biology", "Ecological Modeling", "Ecology", "Equine", "Evolution", "Food Animals", "Food Science", "Forestry", "Agricultural and Biological Sciences", "Horticulture", "Insect Science", "General Agricultural and Biological Sciences", "General Veterinary", "Plant Science", "Small Animals", "Structural Biology", "Veterinary (miscellaneous)"

Arts and Humanities:

"Arts and Humanities (miscellaneous)", "Classics", "Cultural Studies", "General Arts and Humanities", "History and Philosophy of Science", "Library and Information Sciences", "Literature and Literary Theory", "Museology", "Philosophy", "Speech and Hearing", "Visual Arts and Performing Arts"

Business:

"Accounting", "Business", "Business and International Management", "Communication", "General Business", "Human Factors and Ergonomics", "Leadership and Management", "Leisure and Hospitality Management", "Management", "Management Science and Operations Research", "Management and Accounting", "Management and Accounting (miscellaneous)", "Management of Technology and Innovation", "Marketing", "Media Technology", "Organization", "Organizational Behavior and Human Resource Management", "Planning and Development", "Strategy and Management"

TABLE A1

NUMBER OF PAPERS IN SCANR DATABASE BY SUBJECTS AFTER EACH CLEANING STEP

Subject	Cleaning Steps			
	Step 1	Step 2	Step 3	Step 4
Agriculture	16,991	10,078	9,988	7,158
Arts and Humanity	1,108	855	-	-
Business	4,047	2,824	1,440	-
Chemistry	45,689	29,332	29,309	26,358
Computer Science	13,359	9,274	9,234	6,775
Energy	4,352	3,175	3,136	1,600
Engineering	24,600	13,739	13,659	10,332
Environmental Science	25,356	15,818	15,605	12,405
Immunology and Microbiology	22,527	16,115	16,087	12,686
Materials Science	12,407	4,497	4,426	2,537
Mathematics	12,511	9,609	9,529	7,201
Medicine/Dentist	65,024	59,677	59,463	55,552
Neuroscience/Psychology	9,689	6,891	6,824	5,017
Nursing/Paramedical	4,321	3,838	3,617	1,719
Others	1,856	871	754	-
Pharmacology	6,358	4,314	4,280	2,113
Physics and Astronomy	39,127	28,274	28,234	26,116
Social Sciences	13,716	11,119	10,523	8,281
Total	323,034	230,296	226,106	185,850

Notes: Number of academic papers published in journals by each of the 18 scientific fields after each of the cleaning steps described in Section A.1.

Chemistry:

"Analytical Chemistry", "Bioengineering", "Biochemistry", "Biochemistry (medical)", "Biophysics", "Catalysis", "Chemical Engineering (miscellaneous)", "Chemical Health and Safety", "Chemistry (miscellaneous)", "Clinical Biochemistry", "Coatings and Films", "Colloid and Surface Chemistry", "Electrochemistry", "Environmental C", "Environmental Chemistry"; "Filtration and Separation", "Fluid Flow and Transfer Processes", "General Biochemistry", "General Chemical Engineering", "General Chemistry", "Inorganic Chemistry", "Materials Chemistry", "Organic Chemistry", "Physical and Theoretical Chemistry", "Polymers and Plastics", "Process Chemistry and Technology", "Spectroscopy"

Computer Science:

"Artificial Intelligence", "Computer Graphics and Computer-Aided Design", "Computer Networks and Communications", "Computer Science (miscellaneous)", "Computer Science Applications", "Computer Vision and Pattern Recognition", "General Computer Science", "Human-Computer Interaction", "Modeling and Simulation", "Signal Processing", "Software", "Theoretical Computer Science", "Information Systems", "Information Systems and Management", "Management Information Systems"

Environmental Science:

"Atmospheric Science", "Computers in Earth Sciences", "Conservation", "Earth and Planetary Sciences (miscellaneous)", "Earth-Surface Processes", "Ecological Modeling", "Ecology", "Economic Geology", "Environmental Engineering", "Environmental Science (miscellaneous)", "General Earth and Planetary

TABLE A2

NUMBER OF PAPERS CITED IN THE PATENT DATABASE, BY SUBJECT

Subject	Nbr of papers
Agriculture	1,508
Arts and Humanity	113
Business	237
Chemistry	7,604
Computer Science	2,439
Energy	312
Engineering	3,385
Environmental Science	615
Immunology and Microbiology	6,513
Materials Science	908
Mathematics	329
Medicine/Dentist	8,073
Neuroscience/Psychology	789
Nursing/Paramedical	577
Others	51
Pharmacology	2,646
Physics and Astronomy	1,683
Social Sciences	173
Total	40,376

Notes: Number of academic papers by subject in the final database on the patent side, i.e. the patCit database merged with firm identifiers and subjects. The count is a fractional count (papers associated to several subjects are split).

Sciences", "General Environmental Science", "Geography", "Geology", "Geophysics", "Geotechnical Engineering and Engineering Geology", "Global and Planetary Change", "Nature and Landscape Conservation", "Ocean Engineering", "Oceanography", "Pollution", "Soil Science", "Sustainability and the Environment", "Water Science and Technology"

Energy:

"Energy (miscellaneous)", "Energy Engineering and Power Technology", "Fuel Technology", "General Energy", "Geochemistry and Petrology", "Nuclear Energy and Engineering", "Renewable Energy"

Engineering:

"Automotive Engineering", "Biomedical Engineering", "Architecture", "Building and Construction", "Civil and Structural Engineering", "Control and Optimization", "Control and Systems Engineering", "Electrical and Electronic Engineering", "Electronic", "Engineering (miscellaneous)", "General Engineering", "Hardware and Architecture", "Industrial and Manufacturing Engineering", "Instrumentation", "Mechanical Engineering", "Sensory Systems", "Surfaces", "Surfaces and Interfaces", "Transportation", "Waste Management and Disposal"

Immunology and Microbiology:

"Immunology", "Immunology and Allergy", "Immunology and Microbiology (miscellaneous)", "Epidemiology", "Genetics and Molecular Biology", "Genetics and Molecular Biology (miscellaneous)", "Ap-

plied Microbiology and Biotechnology", "Biotechnology", "Cell Biology", "Infectious Diseases", "Microbiology", "Microbiology (medical)", "Molecular Biology", "Molecular Medicine", "Parasitology", "Virology", "General Immunology and Microbiology"

Materials Science:

"Biomaterials", "General Materials Science", "Materials Science (miscellaneous)", "Metals and Alloys", "Optical and Magnetic Materials", "Mechanics of Materials", "Ceramics and Composites"

Mathematics:

"Algebra and Number Theory", "Analysis", "Applied Mathematics", "Computational Mathematics", "Computational Theory and Mathematics", "Discrete Mathematics and Combinatorics", "General Mathematics", "Geometry and Topology", "Logic", "Mathematics (miscellaneous)", "Numerical Analysis", "Probability and Uncertainty", "Statistics", "Statistics and Probability"

Medicine/Dentistry:

"Aging", "Anatomy" "Assessment and Diagnosis", "Cancer Research", "Cardiology and Cardiovascular Medicine", "Chiropractics", "Complementary and Manual Therapy", "Complementary and alternative medicine", "Critical Care", "Critical Care and Intensive Care Medicine", "Dermatology", "Diabetes and Metabolism", "Embryology", "Emergency Medicine", "Endocrine and Autonomic Systems", "Endocrinology", "Gastroenterology" "General Medicine", "Genetics", "Genetics (clinical)", "Genetics(clinical)", "Geriatrics and Gerontology", "Gerontology", "Health Informatics", "Hematology", "Hepatology", "Histology", "Internal Medicine", "Medical Laboratory Technology", "Medicine (miscellaneous)", "Monitoring", "Nephrology", "Neurology", "Neurology (clinical)", "Nuclear Medicine and imaging", "Oncology", "Oncology (nursing)", "Ophthalmology" "Oral Surgery" "Otorhinolaryngology", "Nutrition and Dietetics", "Obstetrics and Gynecology", "Pathology and Forensic Medicine", "Pediatrics", "Periodontics", "Podiatry", "Pulmonary and Respiratory Medicine", "Reproductive Medicine", "Rheumatology", "Surgery", "Transplantation", "Urology", "Dentistry (miscellaneous)", "General Dentistry", "Orthodontics"

Neuroscience/Psychology:

'Behavior and Systematics", "Behavioral Neuroscience", "Cellular and Molecular Neuroscience", "Cognitive Neuroscience", "Decision Sciences (miscellaneous)", "Developmental Neuroscience", "Fundamentals and skills", "General Decision Sciences", "General Neuroscience", "Neuropsychology and Physiological Psych", "Applied Psychology", "Biological Psychiatry", "Clinical Psychology", "Developmental and Educational Psychology" "Experimental and Cognitive Psychology", "General Psychology", "Neuropsychology and Physiological Psychology", "Neuroscience (miscellaneous)", "Psychiatric Mental Health", "Psychiatry and Mental health", "Psychology (miscellaneous)"

Nursing/Paramedical:

"Advanced and Specialized Nursing", "Critical Care Nursing", "Emergency Nursing", "General Nursing", "LPN and LVN", "Medical Surgical Nursing", "Nursing (miscellaneous)", "Occupational Therapy", "Perinatology", "Perinatology and Child Health", "Physical Therapy", "Orthopedics and Sports Medicine", "Physiology", "Optometry", "Physiology (medical)", "Rehabilitation", "Sports Therapy and Rehabilitation", "and Child Health"

Pharmacology:

"Anesthesiology and Pain Medicine", "Drug Discovery", "Drug Guides", "General Pharmacology", "Pharmaceutical Science", "Pharmacology", "Pharmacology (medical)", "Pharmacology (nursing)", "Pharmacy", "Toxicology", "Toxicology and Mutagenesis", "Toxicology and Pharmaceutics", "Toxicology and Pharmaceutics (miscellaneous)"

Physics and Astronomy:

"Acoustics and Ultrasonics", "Aerospace Engineering", "Astronomy and Astrophysics", "Atomic and Molecular Physics", "Computational Mechanics", "Condensed Matter Physics", "General Physics and Astronomy", "Mathematical Physics", "Nuclear and High Energy Physics", "Physics and Astronomy (miscellaneous)", "Radiation", "Radiological and Ultrasound Technology", "Radiology", "Space and Planetary Science", "Statistical and Nonlinear Physics", "Music"

Social Sciences:

"Anthropology", "Archeology", "Demography", "Education", "Family Practice", "Gender Studies", "General Social Sciences", "Community and Home Care", "Environmental and Occupational Health", "General Health Professions", "Health", "Health (social science)", "Health Information Management", "Health Policy", "Health Professions (miscellaneous)", "History", "Industrial relations", "Language and Linguistics", "Law", "Linguistics and Language", "Paleontology", "Policy and Law", "Political Science and International Relations", "Public Administration", "Public Health", "Religious studies", "Social Psychology", "Social Sciences (miscellaneous)", "Sociology and Political Science", "Tourism", "Urban Studies", "ethics and legal aspects", "Econometrics and Finance", "Econometrics and Finance (miscellaneous)", "Economics", "Economics and Econometrics", "Finance", "General Economics"

Others:

"Development", "Issues", "Life-span and Life-course Studies", "Maternity and Midwifery", "Reliability and Quality", "Risk", "Safety", "Safety Research", "Stratigraphy"

A.3. *Matching based on keywords*

This Appendix describes the list of keywords and bigrams that we assigned to each subject. These words are then used to match patents with subjects based on the "top terms" constructed by Google Patent and associated with each patent publication (see Srebrovic, 2019).

Agriculture:

"Agriculture", "Aquatic Science", "Developmental Biology", "Ecological Modeling", "Ecology", "Equine", "Evolution", "Food Animals", "Food Science", "Forestry", "Agricultural Sciences", "Horticulture", "Insect Science", "Biological Sciences", "General Veterinary", "Plant Science", "Small Animals", "Structural Biology", "Veterinary"

Arts and Humanities:

"Art', 'Humanities", "Classics", "Cultural Studies", "History", "Philosophy", "Library", "Information Sciences", "Literature", "Museology", "Speech", "Visual Arts", "Performing Arts"

Business:

"Accounting", "Business", "International Management", "General Business", "Ergonomics' ,'Leadership", "Leisure", "Hospitality Management", "Management", "Management Science", "Operations Research", "Marketing", "Media Technology", "Organization", "Organizational Behavior", "Human Resource", "Planning", "Strategy"

Chemistry:

"Analytical Chemistry", "Bioengineering", "Biochemistry", "Biophysics", "Catalysis", "Chemical Engineering", "Chemistry", "Clinical Biochemistry", "Coating", "Films' 'Surface Chemistry", "Electrochemistry", "Environmental Chemistry", "Filtration", "Separation", "Fluid Flow", "Transfer Processes", "Chemical Engineering", "Inorganic Chemistry", "Materials Chemistry", "Organic Chemistry", "Polymers", "Plastics", "Spectroscopy"

Computer Science:

"Artificial Intelligence", "AI", "Computer Graphics", "CAD", "Computer-Aided Design", "Computer Networks", "Computer Science", "Computer Vision", "Pattern Recognition", "Modeling", "Simulation", "Signal Processing", "Software", "Information Systems"

Environmental Science:

"Atmospheric", "Earth Sciences", "Conservation", "Earth", "Planetary", "Ecological", "Ecology", "Economic Geology", "Environmental Engineering", "Environmental Science,'Geography", "Geology", "Geophysics", "Geotechnical Engineering", "Engineering Geology", "Planetary Change", "Nature", "Landscape Conservation", "Ocean Engineering", "Oceanography", "Pollution", "Soil Science", "Sustainability", "Water"

Energy:

"Energy", "Power", "Fuel", "Geochemistry", "Petrology", "Nuclear Energy", "Renewable Energy"

Engineering:

"Automotive", "Biomedical", "Architecture", "Building", "Construction", "Civil Engineering", "Structural Engineering", "Control", "Optimization", "Electrical", "Electronic", "Hardware", "Industrial", "Manufacturing", "Instrumentation", "Sensor", "Surfaces", "Interfaces", "Transportation", "Waste"

Immunology and Microbiology:

"Immunology", "Allergy", "Microbiology", "Epidemiology", "Genetics", "Molecular Biology", "Microbiology", "Biotechnology", "Cell Biology", "Infectious", "Molecular", "Parasitology", "Virology"

materials science:

"Biomaterials", "Materials Science", "Metals,'Alloys", "Optical", "Magnetic"

Mathematics:

"Algebra", "Number Theory", "Analysis", "Applied Mathematics", "Computational Mathematics", "Combinatorics", "Geometry", "Topology", "Logic", "Numerical", "Probability", "Uncertainty", "Statistics", "Probability"

Medicine/Dentistry:

"Aging", "Anatomy", "Diagnosis", "Cancer", "Cardiology", "Cardiovascular Medicine", "Chiropractics", "Critical Care", "Intensive Care", "Dermatology", "Diabetes", "Metabolism", "Embryology", "Emergency", "Endocrine", "Endocrinology", "Gastroenterology", "Genetics", "Geriatrics", "Gerontology", "Health Informatics", "Hematology", "Hepatology", "Histology", "Internal Medicine", "Monitoring", "Nephrology", "Neurology", "Nuclear Medicine", "Oncology", "Ophthalmology", "Oral Surgery", "Otorhinolaryngology", "Nutrition", "Dietetics", "Pathology", "Forensic Medicine", "Pediatrics", "Periodontics", "Podiatry", "Pulmonary", "Respiratory Medicine", "Reproductive Medicine", "Rheumatology", "Surgery", "Transplantation", "Urology", "Dentistry", "Orthodontics"

Neuroscience/Psychology:

"Behavior", "Systematics", "Neuroscience", "Cognitive", "Decision", "Neuropsychology", "Physiological Psych", "Psychology", "Psychiatry", "Neuropsychology", "Neuroscience", "Pshychiatric", "Mental Health"

Nursing/Paramedical:

"Nursing", "Occupational Therapy", "Perinatology", "Physical Therapy", "Orthopedics", "Sports Medicine", "Physiology", "Optometry", "Rehabilitation", "Sports Therapy"

Pharmacology:

"Anesthesiology", "Pain", "Drug", "Pharmacology", "Pharmaceutical Science", "Pharmacology", "Pharmacy", "Toxicology", "Toxicology", "Mutagenesis", "Pharmaceutics"

Physics and Astronomy:

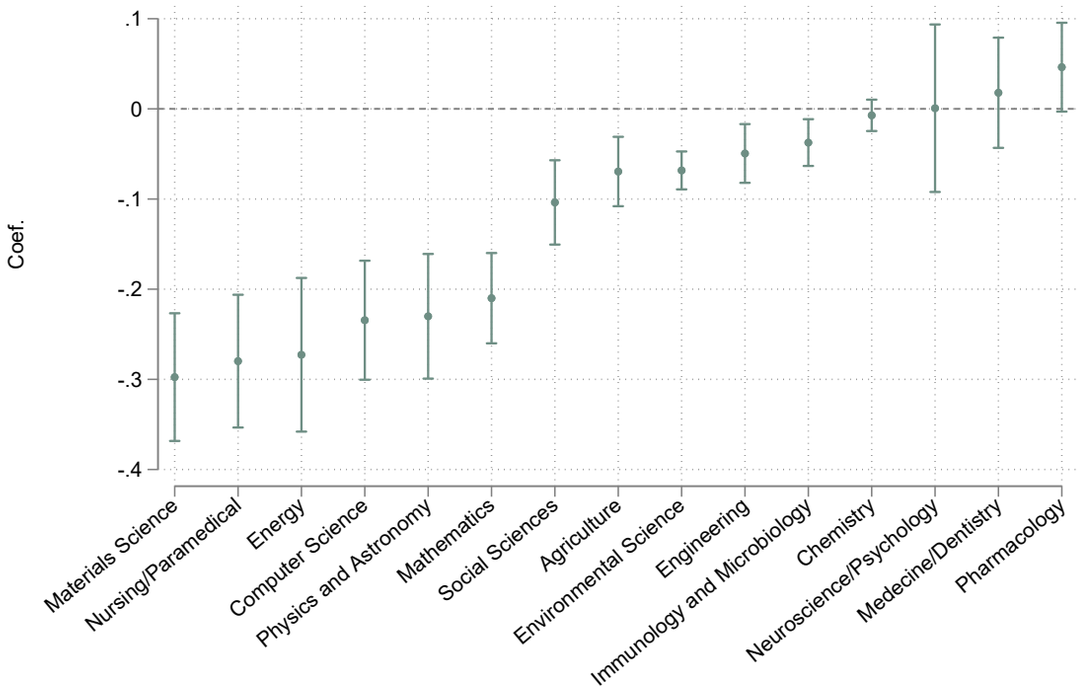
"Acoustics", "Ultrasonics", "Aerospace", "Astronomy", "Astrophysics", "Atomic", "Computational Mechanics", "Condensed Matter", "Radiation", "Radiological", "Ultrasound", "Radiology", "Space", "Planetary", "Music"

Social Sciences:

"Anthropology", "Archeology", "Demography", "Education", "Family", "Gender Studies", "Industrial relations", "Language", "Linguistics", "Law", "Paleontology", "Policy and Law", "Political Science", "International Relations", "Public Administration", "Public Health", "Religious studies", "Social Psychology", "Sociology", "Tourism", "Urban Studies", "ethics", "Econometrics", "Finance"

APPENDIX B: ADDITIONAL RESULTS

Figure B1: Exposure is weighted by the size of laboratories



Notes: This Figure replicates Figure 2 but the exposure A and its counterfactual B (see equation (2) are calculated by taking the weighted sum of exposure across all laboratories, weighting by the number of papers these laboratories published in each scientific field. See Section 3.1 for more details.

Figure B2: Detailed robustness results

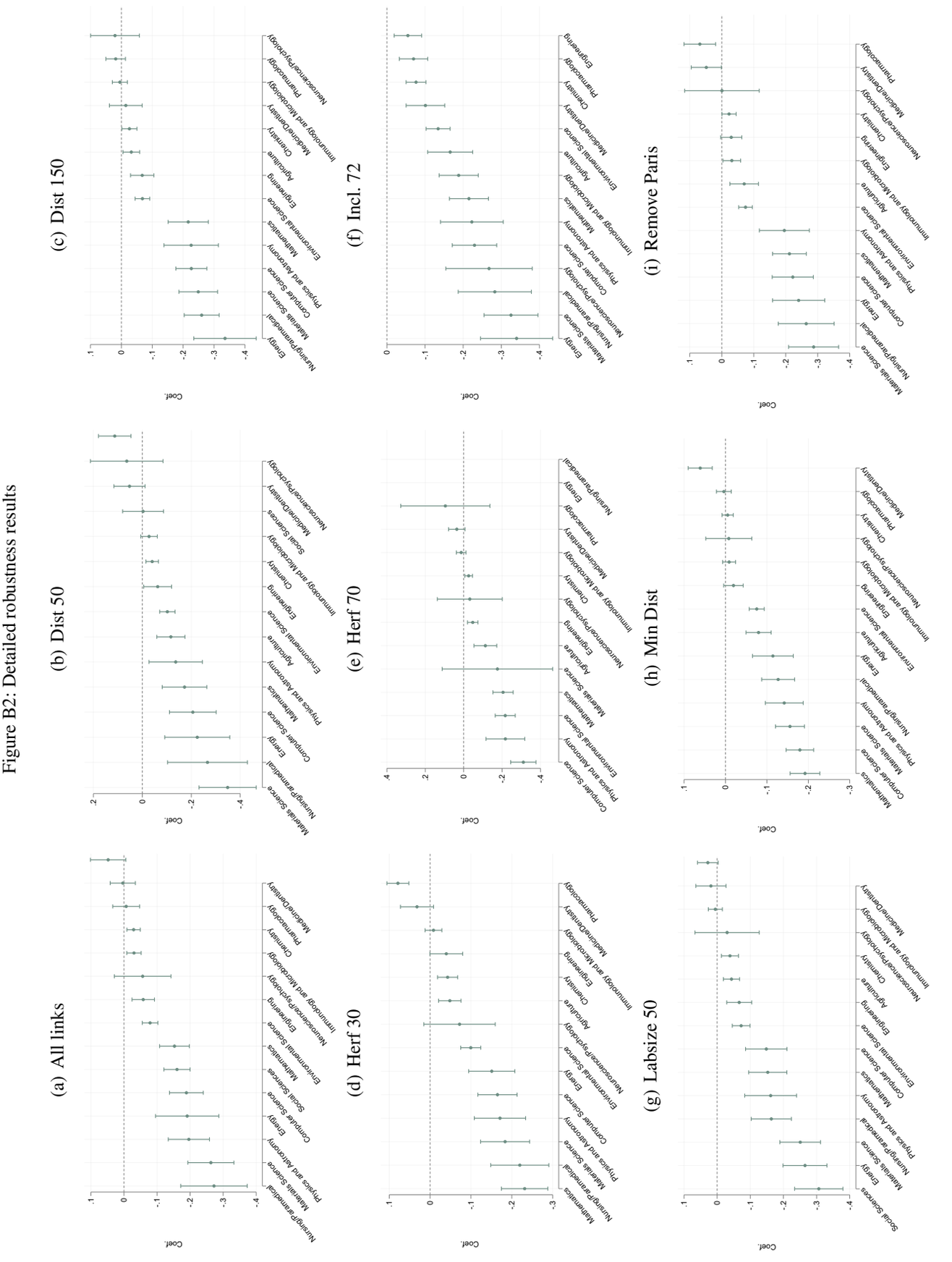


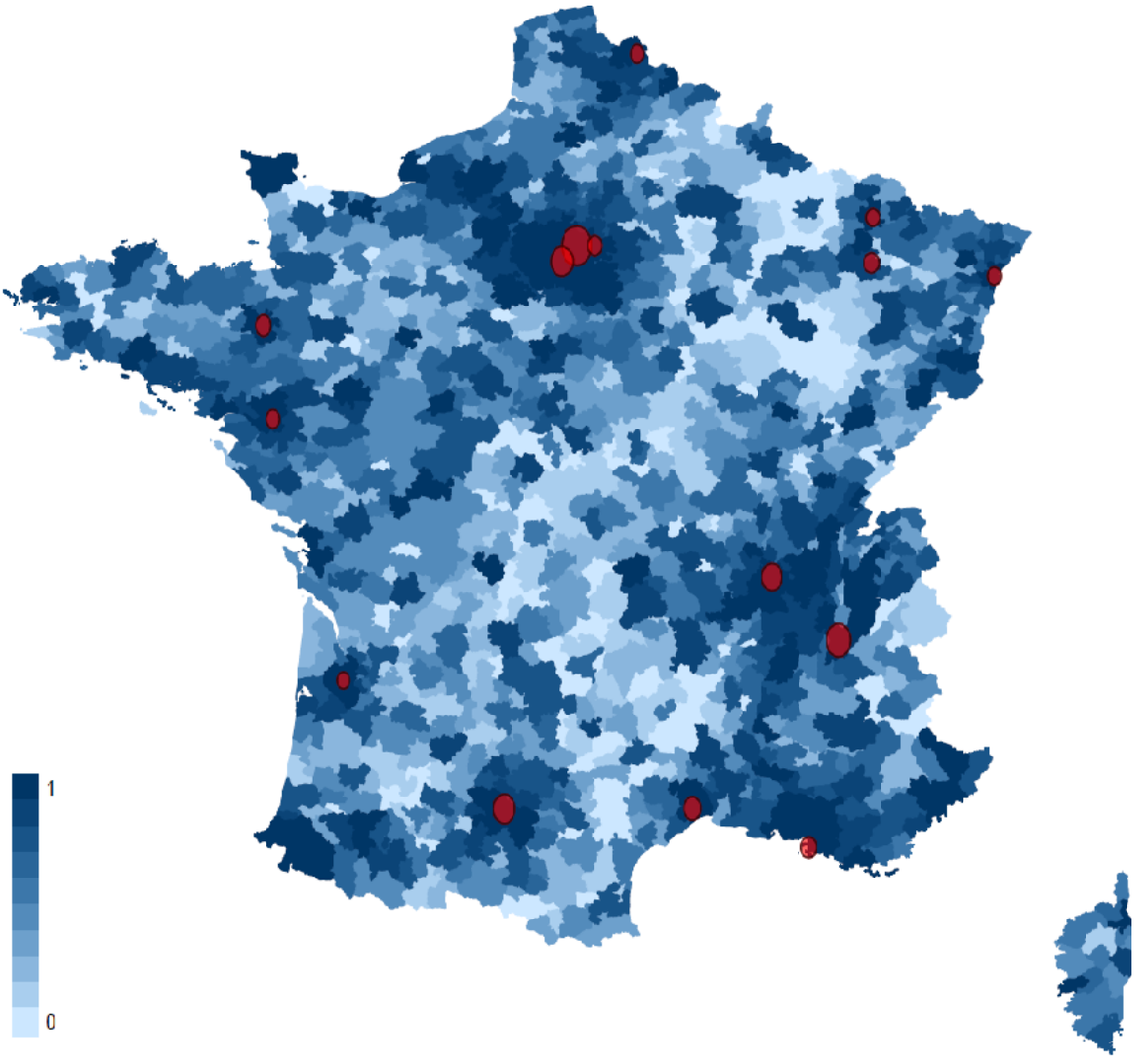
TABLE B1

CENTRALITY

Subject	Centrality		
	Closeness	Bonacich	Degree
Social Sciences	0.101	1.03	3.44
Nursing/Paramedical	0.118	1.08	7.36
Environmental Science	0.126	1.09	9.81
Mathematics	0.118	1.10	9.22
Neuroscience/Psychology	0.142	1.11	13.0
Medicine/Dentistry	0.166	1.16	35.1
Engineering	0.159	1.24	32.5
Agriculture	0.142	1.24	16.0
Energy	0.083	1.26	6.06
Pharmacology	0.158	1.26	23.1
Immunology and Microbiology	0.166	1.26	33.8
Computer Science	0.162	1.27	25.3
Materials Science	0.149	1.28	17.1
Physics and Astronomy	0.164	1.31	34.4
Chemistry	0.168	1.36	64.8

Notes: This table shows the average Closeness, Bonacich (Bonacich, 1987) and degree centrality of the proximity measure $\text{prox}_{l,i}$. The average is calculated across all laboratories l of a same subject, weighted by the number of papers published in the laboratory.

Figure B3: Spillover and location of laboratories in the field of Materials Science



Notes: This map shows the location of laboratories in Materials Science as well as the difference between the actual and counterfactual level of proximities by cities (standardized to range between 0 and 1). This corresponds to $a - b$ in equation (2). The size of the laboratories is proportional to their number of papers published. Sources: scanR, patCit, Patstat.

REFERENCES

L. Abramovsky and H. Simpson. Geographic proximity and firm–university innovation linkages: evidence from great britain. *Journal of Economic Geography*, 11(6):949–977, 2011.

L. Abramovsky, R. Harrison, and H. Simpson. University research and the location of business r&d. *Economic Journal*, 117(519):C114–C141, 2007.

D. Acemoglu, U. Akcigit, and W. R. Kerr. Innovation network. *Proceedings of the National Academy of Sciences*, 113(41):11483–11488, 2016.

P. Aghion and X. Jaravel. Knowledge spillovers, innovation and growth. *Economic Journal*, 125(583): 533–573, 2015.

P. Aghion, A. Bergeaud, T. Gigout, M. Lequien, and M. Melitz. Exporting ideas: How trade spills over to knowledge. CEP Discussion Paper 1960, 2021.

A. Agrawal and R. Henderson. Putting patents in context: Exploring knowledge transfer from mit. *Management Science*, 48(1):44–60, 2002.

M. Ahmadpoor and B. F. Jones. The dual frontier: Patented inventions and prior scientific advance. *Science*, 357(6351):583–587, 2017.

U. Akcigit, D. Hanley, and N. Serrano-Velarde. Back to basics: Basic research spillovers, innovation policy, and growth. *Review of Economic Studies*, 88(1):1–43, 2021.

D. B. Audretsch and M. P. Feldman. R&d spillovers and the geography of innovation and production. *American Economic Review*, 86(3):630–640, 1996.

D. B. Audretsch and M. P. Feldman. Knowledge spillovers and the geography of innovation. In *Handbook of regional and urban economics*, volume 4, pages 2713–2739. Elsevier, 2004.

P. Azoulay, J. S. Graff Zivin, D. Li, and B. N. Sampat. Public r&d investments and private-sector patenting: evidence from nih funding rules. *Review of Economic Studies*, 86(1):117–152, 2019.

A. Bergeaud, A. Guillouzouic, E. Henry, and C. Malgouyres. From public labs to private firms: Magnitude and channels of r&d spillovers. Discussion Paper 1882, Centre for Economic Performance, London School of Economics and Poiltical Science, 2022.

P. Bonacich. Power and centrality: A family of measures. *American journal of sociology*, 92(5):1170–1182, 1987.

K. Buzard, G. A. Carlino, R. M. Hunt, J. K. Carr, and T. E. Smith. The agglomeration of american r&d labs. *Journal of Urban Economics*, 101:14–26, 2017.

N. Carayol and E. Carpentier. The spread of academic invention: a nationwide case study on french data (1995–2012). *Journal of Technology Transfer*, pages 1–27, 2021.

G. Carlino and W. R. Kerr. Agglomeration and innovation. *Handbook of regional and urban economics*, 5: 349–404, 2015.

G. A. Carlino, S. Chatterjee, and R. M. Hunt. Urban density and the rate of invention. *Journal of Urban Economics*, 61(3):389–419, 2007.

W. M. Cohen, R. R. Nelson, and J. P. Walsh. Links and impacts: the influence of public research on industrial r&d. *Management science*, 48(1):1–23, 2002.

P.-P. Combes and L. Gobillon. The empirics of agglomeration economies. In *Handbook of regional and urban economics*, volume 5, pages 247–348. Elsevier, 2015.

P. Cotterlaz and A. Guillouzouic. The percolation of knowledge across space. *Available at SSRN 3010092*, 2020.

G. Cristelli, G. de Rassenfosse, K. Higham, and C. Verluise. The missing 15 percent of patent citations. Available at SSRN 3754772, 2020.

C. De Fuentes and G. Dutrénit. Best channels of academia-industry interaction for long-term benefit. *Research Policy*, 41(9):1666–1682, 2012. ISSN 0048-7333. . URL https://www.sciencedirect.com/science/article/pii/S0048733312000996.

M. Delgado, M. E. Porter, and S. Stern. Clusters and entrepreneurship. *Journal of Economic Geography*, 10(4):495–518, 2010.

G. Duranton and D. Puga. Micro-foundations of urban agglomeration economies. In J. V. Henderson and J. F. Thisse, editors, *Handbook of Regional and Urban Economics*, volume 4, chapter 48, pages 2063–2117. Elsevier, 1 edition, 2004. URL https://EconPapers.repec.org/RePEc:eee:regchp:4-48.

P. H. Egger and N. Loumeau. The economic geography of innovation. Technical Report DP13338, CEPR Discussion Paper, 2018.

G. Ellison and E. L. Glaeser. Geographic concentration in us manufacturing industries: a dartboard ap-

proach. *Journal of Political Economy*, 105(5):889–927, 1997.
K. R. Fabrizio. Absorptive capacity and the search for innovation. *Research Policy*, 38(2):255–267, 2009.
M. P. Feldman and D. F. Kogler. Stylized facts in the geography of innovation. In *Handbook of the Economics of Innovation*, volume 1, pages 381–410. Elsevier, 2010.
L. Fleming, H. Greene, G. Li, M. Marx, and D. Yao. Government-funded research increasingly fuels innovation. *Science*, 364(6446):1139–1141, 2019.
J. Gyourko, C. Mayer, and T. Sinai. Superstar cities. *American Economic Journal: Economic Policy*, 5(4): 167–99, 2013.
N. Hausman. University Innovation and Local Economic Growth. *Review of Economics and Statistics*, pages 1–46, 03 2021.
R. Henderson, A. B. Jaffe, and M. Trajtenberg. Universities as a source of commercial technology: a detailed analysis of university patenting, 1965–1988. *Review of Economics and Statistics*, 80(1):119–127, 1998.
A. B. Jaffe. Real effects of academic research. *American Economic Review*, pages 957–970, 1989.
A. B. Jaffe, M. Trajtenberg, and R. Henderson. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics*, 108(3):577–598, 1993.
A. B. Jaffe, M. Trajtenberg, and M. S. Fogarty. Knowledge spillovers and patent citations: Evidence from a survey of inventors. *American Economic Review*, 90(2):215–218, 2000.
S. Lychagin, J. Pinkse, M. E. Slade, and J. V. Reenen. Spillovers in space: does geography matter? *Journal of Industrial Economics*, 64(2):295–335, 2016.
M. Marx and A. Fuegi. Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal*, 41(9):1572–1594, 2020.
P. B. Maurseth and B. Verspagen. Knowledge spillovers in europe: a patent citations analysis. *Scandinavian Journal of Economics*, 104(4):531–545, 2002.
S. S. Rosenthal and W. C. Strange. Geography, industrial organization, and agglomeration. *Review of Economics and Statistics*, 85(2):377–393, 2003.
M. Schnitzer and M. Watzinger. Standing on the shoulders of science. Discussion Papers 13766, CEPR, 2019.
R. Srebrovic. "expanding your patent set with ml and bigquery". Available: https://cloud.google.com/blog/products/data-analytics/expanding-your-patent-set-with-ml-and-bigquery, 2019.
M. Trajtenberg, R. Henderson, and A. Jaffe. University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and New Technology*, 5(1):19–50, 1997.