Scientometrics
DOI 10.1007/s11192-016-2125-9

The research space: using career paths to predict the evolution of the research output of individuals, institutions, and nations

Miguel R. Guevara[1,2,4] · **Dominik Hartmann**[1,3] · **Manuel Aristarán**[1] · **Marcelo Mendoza**[4] · **César A. Hidalgo**[1]

Received: 7 April 2016

Abstract In recent years scholars have built maps of science by connecting the academic fields that cite each other, are cited together, or that cite a similar literature. But since scholars cannot always publish in the fields they cite, or that cite them, these science maps are only rough proxies for the potential of a scholar, organization, or country, to enter a new academic field. Here we use a large dataset of scholarly publications disambiguated at the individual level to create a map of science—or *research space*—where links connect pairs of fields based on the probability that an individual has published in both of them. We find that the research space is a significantly more accurate predictor of the fields that individuals and organizations will enter in the future than citation based science maps. At the country level, however, the research space and citations based science maps are equally accurate. These findings show that data on career trajectories—the set of fields that individuals have previously published in—provide more accurate predictors of future research output for more focalized units—such as individuals or organizations—than citation based science maps.

Keywords Maps of science · Research policy · Innovation policy · Career paths · Scientograms · RCA

Mathematics Subject Classification 68U35 · 94A17 · 05C90 · 91D30 · 68R10

Electronic supplementary material The online version of this article (doi:10.1007/s11192-016-2125-9) contains supplementary material, which is available to authorized users.

✉ Miguel R. Guevara
mguevara@mit.edu; miguel.guevara@upla.cl

[1] Macro Connections, The MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA

[2] Department of Computer Science, Universidad de Playa Ancha, Valparaiso, Chile

[3] Chair for Economics of Innovation, University of Hohenheim, Stuttgart, Germany

[4] Department of Informatics, Universidad Técnica Federico Santa María, Santiago, Chile

Published online: 17 September 2016

JEL Classification I230 · I280 · 94A17

Introduction

While most scientists are trained in one specialized academic field, their scholarly contributions usually involve multiple fields. In fact, 99.8 % of the 215,390 scholars that had a Google Scholar profile by May 24, 2014, and that received citations in at least ten different papers, had published in two or more academic fields (with fields defined according to the 308 categories in the SCImago classification of journals from Scopus). But trans-disciplinary efforts are not constrained to pairs of disciplines. In fact, 99.2 % of these scholars had also published in three or more fields, and 97.5 % of them in four or more. These numbers show that the work of most scholars is not constrained to a single academic discipline, but often spans at least a few of them.

But while most scholars do not publish in a single discipline, their contributions are nevertheless confined to a small set of highly related fields. Consider, for instance, the 24,125 scholars in our dataset (see Data and Methods) that have published at least two papers in "Molecular Biology." 46.6 % of these scholars also had published in "Clinical Biochemistry," but only 0.95 % of them also published in "Economics and Econometrics." Since the total number of scholars with at least two papers in "Clinical Biochemistry" (11,110) is similar to the number of scholars with at least two papers in "Economics and Econometrics" (10,479), the larger overlap of the first pair vis-à-vis the second, tells us that "Molecular Biology" is more related to "Clinical Biochemistry" than to "Economics and Econometrics."

But the structure of these academic overlaps is not theoretically surprising. Scholars are often trained in narrowly defined academic disciplines, and they spend most of their careers in relatively homogenous academic departments. This homogeneity in training also leads to relatively high levels of homogeneity in their social and professional networks, limiting the information available to a scholar in her immediate social network. One indicator of the social homogeneity of the social networks of scholars is the large number of marriages among scientists from the same academic discipline. Within fields marriages among scientists in computer science, chemistry, electrical engineering, microbiology and physics (according to the scientific classifications of the National Research Council and National Science Foundation) are as frequent as 56 % for women scientists in their first marriage, and 63 % for women scientist in their second marriage (compared to 14 and 32 % for males) (Fox 2005). Among women in the first marriage, 36 % marry a scholar within the same field of science. Thus, the professional and social institutions where scholars are embedded (Granovetter 1985) reduce the opportunity for scholars to develop the contacts, or skills they need to enter "distant" academic fields. As a result, the diversification paths followed by individuals, organizations, and countries, are constrained by the homogeneity of the social networks of scholars and their professional institutions. These various constraints should be reflected in the structure of the network connecting related academic fields.

But the prevalence of researchers publishing in multiple academic fields is good news for those looking to either predict the evolution of research production, or evaluate the potential of an organization to enter a particular academic field. In fact, the overlapping participation of scholars in related disciplines tells us about the possible career paths of scholars. Moreover, since research organizations, and national research efforts, are

composed of networks of scholars, the network of related academic disciplines—in terms of the ability of authors to publish in distinct fields of science—should be useful to predict the probability that a country or organization will enter a new academic field.

Here we leverage information on the observed career paths of more than two hundred thousand scholars to introduce the *research space*, a map connecting pairs of fields based on the probability that an author has published in both of them. We argue that this map captures implicit information about the skills, social networks, and institutions constraining the movement of scholars into different academic disciplines. We validate the predictive superiority of the research space by using Response Operator Characteristic curves (ROC curves) and show that the research space is a more accurate predictor of the future presence of an individual or organization in an academic field than citation based or knowledge flow science maps.

Mapping science through knowledge flows and career paths

In recent decades bibliometricians, information scientists, sociologists, physicists, and computer scientists, have created maps of science connecting fields that either cite each other, or that cite similar literature (Börner et al. 2012; Boyack et al. 2005; Leydesdorff and Rafols 2009; Waltman et al. 2010). These citation based maps of science are often interpreted as knowledge flow maps (Zhuge 2006) and tell us, for instance, if the knowledge developed in one field is used to produce knowledge in other fields. Ultimately, these maps help us categorize science and understand the trans-disciplinary impact of scholarly work. It must be noted that we are focused on maps where nodes represent fields of science. We are not focused on other type of maps that can be built on publications data, just like co-authorship networks, international collaborations, or topic evolving maps.

Most knowledge flow science maps use one of three methods: co-citation, direct citations, or bibliographic coupling. Co-citation networks (Boyack et al. 2005; Moya-Anegón et al. 2004; Small 1973, 1999) connect academic disciplines by looking at the reference section of a paper and connecting the areas of the papers that appear in the same list of references (i.e. they connect papers A and B, if paper C cites both of them) (Fig. 1a). Direct citation networks, on the other hand, (Boyack et al. 2005; Leydesdorff and Rafols 2009; Rosvall and Bergstrom 2008) connect academic disciplines when a paper from one discipline cites a paper from another discipline (Fig. 1b). Direct citation networks includes both, networks where scholars differentiate the source and target fields, and un-directed networks, where information on what field is citing, and what field is cited, is disregarded. Finally, bibliographic coupling networks (Börner et al. 2012; Boyack et al. 2005), connect pairs of disciplines when papers from different fields cite the same other papers (Fig. 1c).

Beyond citation-based maps, scholars have also used online searchers to connect academic disciplines. The Clickstream Science Map by (Bollen et al. 2009) connects academic disciplines based on the probability that a scholar who searched for a paper from one field, also searched for a paper from another field. In spirit, the clickstream map is similar to the networks created from co-citations or bibliographic coupling because it also focuses on knowledge flows. Yet since online searches are a more common expression of interest in a topic than a formal citation (the latter requires the costly process of publication), efforts like clickstream help leverage new datasets that are more dynamic than those based on citations.

But what are these science maps used for? One common use of knowledge flow maps is to categorize knowledge. The idea of knowledge categorization has a long tradition in bibliometrics, going back at least to the work of Paul Otlet, the creator of the Universal

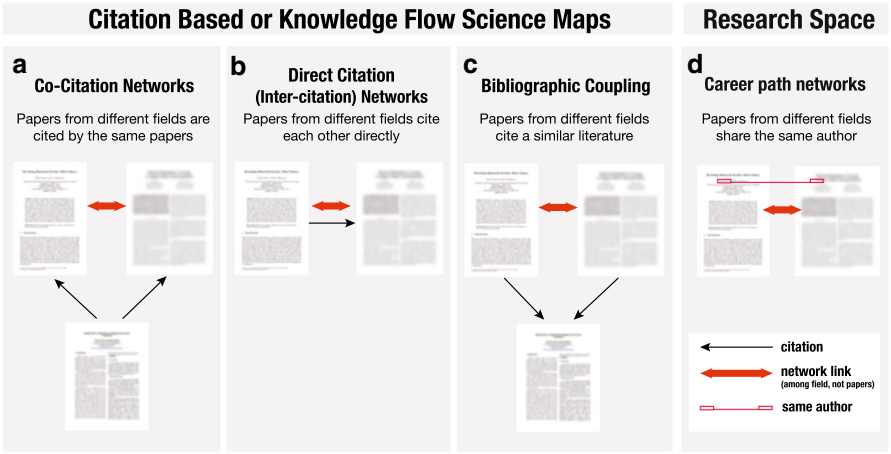

Fig. 1 Methods used to create science maps. Citation based or Knowledge flow Science maps include: Co-citation networks, that connect the academic fields of papers that appear in the same reference sections; Direct citations networks, that connect fields when papers from these fields cite each other; and Bibliographic coupling networks, that connect fields that cite a similar literature. The Research Space is not based on citations and connects fields when researchers are likely to have published in both of them

Decimal Classification, and Ramon Llull, the creator of the XIV century science tree. This idea, however, continues to be influential in recent projects, such as the consensus Map of Science (Klavans and Boyack 2009) or the UCSD Science Map and Classification System (Börner et al. 2012). The UCSD science map has been used to construct a classification of 554 research areas that some university libraries now use to understand the research production of their scholars. Another example of the use of science maps includes the cross-citation maps of Leydesdorff and Rafols (2009), who overlaid the research structure of universities (Rafols et al. 2010) to contextualize a university's research output.

Another use of science maps is as policy instruments. In a world where research budgets are constrained, and the probability of succeeding in a field is uncertain, science promotion agencies (like the N.S.F. in the U.S., the F.A.P.s in Brazil, or the C.N.R.S. in France) need to decide the amount of funds they will allocate to each field, including those where a country or institution may not have a presence and the probability of success is uncertain. Science maps can help estimate a field's strategic value, by helping administrators estimate the probability of success, and therefore the cost, of venturing into a new research area.

But research fields are not only connected by the knowledge flows that are expressed in citations. Since scholars around the world participate in multiple fields, information about the career trajectories of scholars (Fig. 1d) represents a viable alternative to knowledge flow maps. In fact, career trajectories have been used to create predictive maps in other areas of research. For instance, labor flows among industries have been used to study the stability of industrial clusters (Neffke and Henning 2013), and the labor mobility of displaced workers (Neffke et al. 2016). Labor flows among occupations have also been used to create online tools that help visualize the possible career paths of workers or the industrial evolution of cities ("DataViva" 2016).

Here, we use the career trajectories of hundreds of thousands of scholars to create a map of science—or research space—to predict the future research output of countries, organizations, and individuals. We find that for the most disaggregate units (individuals and

organizations) the research space is a more accurate predictor of the development of future research areas than knowledge flow based science maps.

Data and methods

Data

Research maps where links connect areas sharing authors are uncommon because most datasets on research production are not properly disambiguated at the author level (i.e. these datasets lack the ability to distinguish among authors with similar names). Here, we solve the disambiguation problem by looking only at data from authors who have created a profile in Google Scholar. We note that the Google Scholar dataset is not free of biases, as the adoption of Google Scholar is not uniform across academic fields, or age groups. So we interpret our results in the narrow context of the data used to produce them. These results are applicable only to the career trajectories that are observable in Google Scholar.

We filter this dataset by focusing only on scholars with less than fifty publications in each year, because those with more than fifty publications tend to have many publications that are miss-assigned and are not theirs (see Online Resource 1 for more details). Our filtered dataset contains 319,049 authors who have authored a total of 4745,774 publications indexed in 16,873 journals and proceedings between 1971 and 2014 (we note that in the introduction we have a smaller number of authors because there we considered only authors with at least ten papers that have received one citation).

We assign each publication to a research category based on the journal in which it was published using Scopus classification system provided by SCImago that includes 27 main areas of knowledge that are subdivided into 308 fine grained categories. In our dataset we use only the 2 categories for which at least one paper was found (For a complete list of categories see Online Resource 1).

We also aggregate the author level data by identifying the organization (i.e. the university or research institution) and country where the scholar participates in. We first identify organizations by matching the verified email provided in the Google Scholar profile of the author, and then, assign organizations to countries according to the list of institutions provided by the Webometrics Ranking of World Universities (Cybermetrics Lab 2015).

For comparisons we download the UCSD science map (Börner et al. 2012), which is a citation based science map based on bibliographic coupling (Fig. 1c) available for download at: http://sci.cns.iu.edu/ucsdmap/. When comparing with the UCSD science map we transform all of our papers to their classification, since in the same website, a one-way mapping from journals to their classification was available.

Constructing the research space

We begin the construction of our research space by defining the presence of a scientist s in academic field f. We define the presence of a scientist s in a field f at time T by taking the sum of the papers produced by scientist s in academic field f before time T, normalized by the number of co-authors she had on each paper p denoted by variable n_p and the number of fields of the journal where the paper was published m_p (since a single paper can be assigned

to multiple categories depending on the journal). Formally we define the matrix $\boldsymbol{X}_{sf}(T)$ as the summation over all papers $p(s, f, T)$ produced by scientist s in field f before time T as:

$$\boldsymbol{X}_{sf}(T) = \sum_{p(s,f,T)} \frac{1}{n_{p(s,f,T)} m_{p(s,f,T)}}$$

$\boldsymbol{X}_{sf}(T)$ is an indicator of the presence of a scientist in a field that controls for the number of co-authors with which a scientists has published and the number of fields in which a journal is classified. We then discretize $\boldsymbol{X}_{sf}(T)$ to remove scientists that have produced only a marginal contribution to field f (scientists that have only produced a small anecdotal participation in field f in an effort with many co-authors). We remove marginal contributions by creating the matrix $\boldsymbol{P}_{sf}(\mathrm{T})$, which is equal to one if the output $\boldsymbol{X}_{sf}(T)$ of scientist s in field f is larger than 0.1 [in a simple example for a scientist with only one paper in some field, 0.1 could represent a paper with other 9 co-authors ($n_p = 10$) in a journal indexed in only one field ($m_p = 1$); or a paper as solo author ($n_p = 1$) in a journal indexed in ten categories ($m_p = 10$)]. Formally, $\boldsymbol{P}_{sf}(\mathrm{T})$ is defined as:

$$\boldsymbol{P}_{sf}(T) = \begin{cases} 1 & \text{if } \boldsymbol{X}_{sf} > 0.1 \\ 0 & \text{otherwise} \end{cases}$$

We then calculate the number of authors that have participated in fields f and f' before time T by taking the inner product of $P_{sf}(\mathrm{T})$ with itself across all scientists. Formally, we define the matrix $M_{ff'}(\mathrm{T})$ as:

$$M_{ff'}(T) = \sum_{s} P_{sf}(T) P_{sf'}(T)$$

Finally, we define the proximity between fields f and f' denoted by variable $\phi_{ff'}$ by taking the probability that a scientist with presence in field f' also has presence in field f:

$$\phi_{ff'}(T) = \frac{M_{ff'}}{\sum_s P_{sf'}},$$

where $\sum_s P_{sf'}$ is the total number of scientists that have presence in field f'.

$\phi_{ff'}(T)$ is the adjacency matrix representing the research space expressed by the career trajectory of scientists in our dataset observed up to time T.

Figure 2 shows a network visualization of the research space ($\phi_{ff'}(2011)$) (i.e. using data from 1971 to 2010). Here nodes are research areas (in UCSD classification) and links connect research areas that are likely to share authors. Colors are assigned according to the main areas defined by the classification, and node sizes are proportional to the total number of papers produced in that area (for papers with multiple categories, we distribute their contribution equally among all of the categories available). Since most proximities are larger than zero, we visualize the network using only the strongest links, which are the links in the Minimum Spanning Tree (MST) and the links for which the conditional probability of sharing authors is larger than 21.2 % a threshold that allows to visualize a rich community structure. Furthermore, to simplify the visualization we take only the maximum of the probability between two areas, since the matrix of proximities is not symmetric (a similar visualization of the research space in SCImago classification is provided in the Online Resource 1).

Next, we compare the links in the research space with the UCSD bibliographic coupling science map using a scatter plot and a linear model (Fig. 3). Surprisingly, since we expect

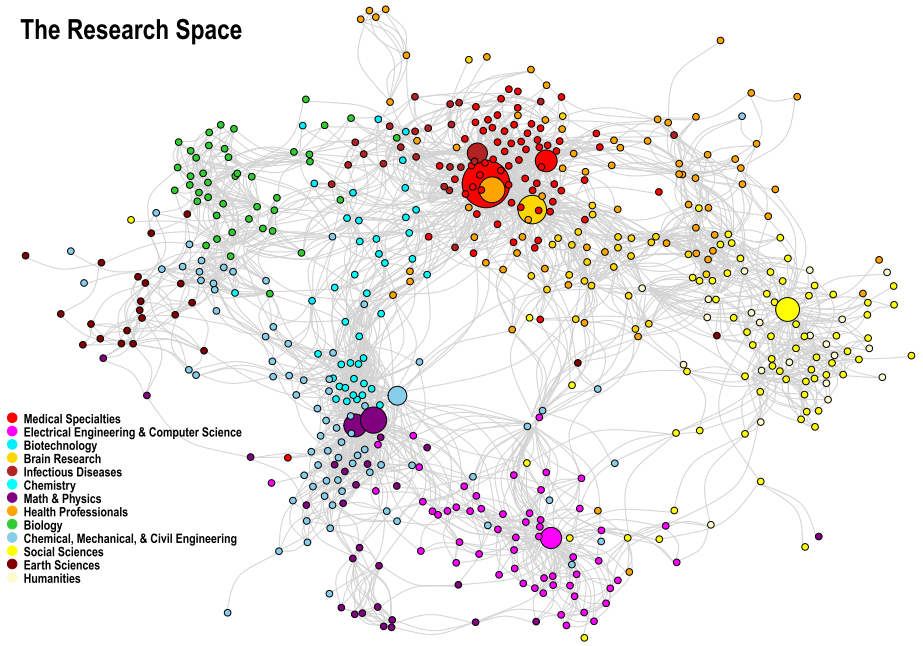

Fig. 2 The Research Space. *Nodes* represent research fields and *links* connect fields that are likely to share authors. The size of nodes is proportional to the number of papers published in that field

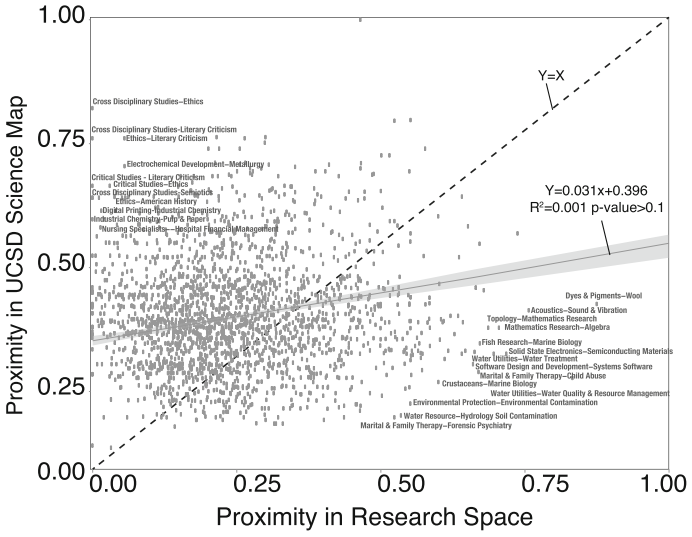

Fig. 3 Comparison between the links estimated for the research space and those reported for the UCSD science map. Since the matrix of proximities in the research space is by definition not symmetric, we use the maximum value of the link between each pair of areas. The observed low correlation is also true for the minimum, or average

fields that share authors to cite each other, we find a relatively low correlation ($R^2 = 0.001$) between the links in both maps. For instance, the proximity among "Crustaceans" and "Marine Biology", or "Environmental Protection" and "Water Treatment" in the research space is high, while the volume of citations among both of these pairs of fields in the UCSD science map is low. Conversely, "Cross Disciplinary Studies" with "Ethics", or "Electrochemical Development" and "Metallurgy" are pairs of fields that often cite each other, but share a relatively small number of co-authors. This orthogonally between both maps tells us that predictions made with either of them will likely be dissimilar since the UCSD map is capturing the relatedness or knowledge flows between fields, and the research space is capturing the *shared capacities* needed to produce science in different fields. In the context of tacit and explicit forms of knowledge (Collins 2010), the UCSD science map appears to capture the reutilization of explicit knowledge, whereas the research space appears to capture the reutilization of the tacit skills and capacities needed to produce explicit knowledge.

Using the research space to predict future research output

We next use the research space to predict the future presence of an individual, organization, or country in a research field. To make these predictions we define five possible states for individuals, organizations, or countries in a research field. These states are: inactive, active, nascent, intermediate, and developed. To define these states we compare the presence of an individual, organization, or country (s), in a research field (f), with the presence that we expect from that individual, organization, or country, based on its effective number of papers X_{sf}. If the effective number of papers produced by an individual, organization, or country (an entity s) in field f is larger than the effective number of papers we expected from an entity with that many total papers in that field, then we say that entity s is developed in the field f. Formally we define the level of development of an individual, organization, or country s in field f using the Revealed Comparative Advantage indicator (Balassa 1965) which is defined as:

$$\mathrm{RCA}_{sf} = \frac{\frac{X_{sf}}{\sum_f X_{sf}}}{\frac{\sum_s X_{sf}}{\sum_{sf} X_{sf}}}$$

The RCA and its normalized version, known in Scientometrics as the Activity Index (AI), have been widely used to analyze the research output of countries (Abramo and D'Angelo 2014; Cimini et al. 2014; Elhorst and Zigova 2014; Guevara and Mendoza 2013; Harzing and Giroud 2014). Here, we use RCA_{sf} to define the five discrete states that we use to characterize the diversification and evolution of the research output of individuals, organizations, and countries:

Inactive (with no papers in the field):	$0 = \mathrm{RCA}_{sf}$
Active (with papers in the field):	$0 < \mathrm{RCA}_{sf}$
Nascent (with a few papers in the field):	$0 < \mathrm{RCA}_{sf} < 0.5$
Intermediate (with less papers than expected in the field):	$0.5 \leq \mathrm{RCA}_{sf} < 1$
Developed (with more papers than expected in the field):	$1 \leq \mathrm{RCA}_{sf}$

We then predict the probability that individual, organization, or country, s will increase its level of development in field f by creating an indicator of the fraction of fields that are connected to field f and that are already developed by s. When we are evaluating transitions to a developed state (to $\mathrm{RCA}_{sf} > 1$), we define $\boldsymbol{U}_{sf}$ as a matrix that is equal to one when $\mathrm{RCA}_{sf} \geq 1$ and 0 otherwise. We then use this indicator to evaluate transitions from inactive to active for individuals, organizations, and countries; from nascent to developed for organizations and countries, and from intermediate to developed for organization and countries. When evaluating a transition, we define $\boldsymbol{U}_{sf} = 1$ when RCA_{sf} transition generates a true positive.

By using the $\boldsymbol{U}$ matrix we define the density of entity s on field f (ω_{sf}), which is our estimator of the probability that entity s will increase its level of activity in field f as:

$$\omega_{sf} = \frac{\sum_{f'} U_{sf'} \phi_{ff'}}{\sum_{f'} \phi_{ff'}}$$

Finally, to predict a transition of entity in field f between a pair of states (i.e. from inactive to active), we look at all fields that are in the initial state (i.e. inactive) and sort them by density (ω_{sf}). The prediction is that the field with higher density will transition to a higher state of development (e.g. from inactive to active), before the fields with lower densities. We apply this method to evaluate the three types of transitions considered, from inactive to active, from nascent to developed, and from intermediate to developed.

For the UCSD science map, we use the same algorithm, but replacing $\phi_{ff'}$ and $\phi_{f'f}$ by the links $\phi_{ff'}$ between fields made available in (Börner et al. 2012). The construction of the links of the UCSD science map is detailed in the supplementary material of (Börner et al. 2012).

Results

We now use the methodology described above to predict the future presence of an individual, organization, or country, in a field that he or she has not participated in. Our predictor of the fields that an individual, institution, or country will develop in the future is based on a ranked list of the densities of the inactive, nascent, or intermediate fields. For instance, for the first type of transition, from inactive to active, the rationale behind the prediction is that the higher the density ω of an inactivated scientific field the higher is the likelihood that this field will become active in the next period. As a simplified example: if an individual, institution or country is already publishing papers in biology and chemistry it is more likely that the unit will publish next period a paper in biochemistry, than a unit that publishes neither in biology nor chemistry.

To measure the accuracy of our predictions (i.e. comparing the ranked density list for both the Research Space and the UCSD map), we use the area under the Response Operator Characteristics curve (ROC curve). The ROC curve plots the true positive rate of a predictive algorithm (in the y-axis) against its false positive rate (x-axis). A random prediction, having the same rate of true positives and false positives, produces a ROC curve with an area of 0.5, so values between 0.5 and 1 represent the accuracy of the predictive method. The ROC curve is a standard statistic used to measure the accuracy of a predictive method and is related to the Mann–Whitney U test, which measures the probability that a true positive is ranked above a false positive. In our case, if the first fields in the density ranking of the inactive fields become active or developed (depending on the transition), then the

ROC curve will move one step up on the *y*-axis with the true positives and the area under the curve will increase. In contrast, when the first ranked fields do not become active or developed, the ROC curve moves one step to the right on the x-axis with the false positives and the area under the curve will decrease.

To make our predictions using the research space we construct our proximity matrix using only data from years prior to 2011 (i.e. from 1971 to 2010). We then look at the state (i.e. inactive, active, etc.) of individuals, organizations, and countries for each research field using data from 2008 to 2010 (see examples of overlay maps with the defined states in the Online Resource 1). Finally, we create a ranked list for each transition to predict the future level of development (i.e. from inactive to active) of each individual, organization, and country, observed between 2011 and 2013. In the remainder of the paper we study seven changes in the level of development of an entity in a field. Changes from inactive to active for individuals, institutions and countries, and changes from nascent to developed, and from intermediate to developed for organizations and countries For individuals we only look at transitions from inactive to active, since the nascent and intermediate levels do not make sense for individuals given their limited output (compared to organizations and countries). Also, it must be noted that marginal contributions of authors were already removed from the dataset (see section Constructing the Research Space).

Figure 4a–c, compare the accuracy achieved by the research space and the UCSD science map for the transition from inactive to active. Figure 4d–g compare the accuracy of the transitions from nascent to developed and from intermediate to developed, respectively. The distributions of areas under the ROC curve obtained for each transition and method are shown using boxplots (where the horizontal bar is the median, the red circle is the mean, the box contains the interquartile range, and the whiskers encompass more than 96 % of the sample). These boxplots describe the distribution for the areas under the ROC curve obtained, respectively, for 4850 individuals 730 organizations (including research

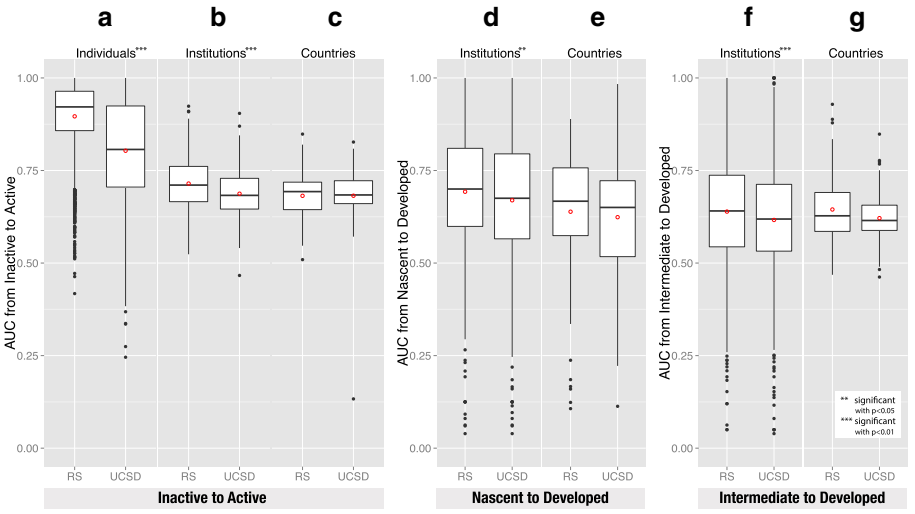

Fig. 4 The predictive power of the Research Space (RS) versus the UCSD science map. For each entity, a ROC curve is calculated across fields for the given transition. Each boxplot represents the distribution of AUCs. Higher values indicate higher predictive accuracy

institutions), and 77 countries. The inclusion criteria involved all entities satisfying the inequality

$$\sum_{T=T_0}^{T=T_0+\Delta T} \sum_{f} X_{sf}(T) \geq B\Delta T$$

with $B = 3$ for individuals, and $B = 30$ for countries and organizations. The inequality helps us focus on the most productive individuals, organizations, and countries.

We now focus on transitions from inactive to active [from having no papers in the field ($\mathrm{RCA}_{sf} = 0$) to having some ($\mathrm{RCA}_{sf} > 0$)]. For both individuals (Fig. 3a) and organizations (Fig. 3b) we find that the predictions made using the research space are significantly more accurate than the predictions made using the UCSD science map. The average area under the ROC curve for individuals (Fig. 3a) is 0.8963 for the research space and 0.8034 for the UCSD science map. This difference is highly statistically significant (ANOVA p value <0.001). Moreover, we note that for individuals the increase in AUC from 0.8 to 0.9 is not only statistically significant but also substantial increase in the number of true positives compared to false positives. In the case of the research space, having an AUC of 0.9, instead of 0.8, means that we expect to find 62 % of all true positives, instead of 40 %, after finding the first 10 % of false positives.

For organizations (Fig. 3b), the averaged accuracy is lower, but the research space is also significantly more accurate than the UCSD science map when it comes to predicting the future presence of an organization in a research field (averages are $\mathrm{AUC}_{\mathrm{research\_space}} = 0.7148$, $\mathrm{AUC}_{\mathrm{UCSD\_science\_map}} = 0.6873$, ANOVA p value <0.001). For countries, however, both methods are equally accurate (Fig. 3c averages are $\mathrm{AUC}_{\mathrm{research\_space}} = 0.6816$, $\mathrm{AUC}_{\mathrm{UCSD\_science\_map}} = 0.6819$, ANOVA p value >0.1), indicating that the increase in accuracy observed for the research space expressed itself for more disaggregate units (individuals and organizations).

Now, we focus on transitions from nascent to developed. These are transitions where a country, or organization, went from having a relatively small presence in a research field ($0 < \mathrm{RCA}_{sf} < 0.5$), to a presence that is larger than what is expected from their size and the size of the field ($\mathrm{RCA}_{sf} > 1$). Once again we find that for organizations (Fig. 3d) the predictions made using the research space are significantly more accurate than the predictions made using the UCSD science map when it comes to predicting the future development of a organization in a research field (averages are $\mathrm{AUC}_{\mathrm{research\_space}} = 0.6927$, $\mathrm{AUC}_{\mathrm{UCSD\_science\_map}} = 0.6696$, ANOVA p value <0.05). For countries, however, Fig. 3e both methods are equally accurate (averages are $\mathrm{AUC}_{\mathrm{research\_space}} = 0.6387$, $\mathrm{AUC}_{\mathrm{UCSD\_science\_map}} = 0.6239$, ANOVA p value >0.1), indicating that for transitions from nascent to developed the increase in accuracy observed for the research space is also expressed itself for more disaggregate units (individuals and organizations).

Finally, we look at the transitions from intermediate to developed. These are transitions where a country or organization, went from having a good-sized presence in a research field ($0.5 \leq \mathrm{RCA}_{sf} < 1$), to a presence that is larger than what is expected from their size and the size of the field ($\mathrm{RCA}_{sf} \geq 1$). Once again we find that for organizations (Fig. 3f) the predictions made using the research space are significantly more accurate than the predictions made using the UCSD science map. The average area under the ROC curve for organizations is 0.6390 for the research space and 0.6164 for the UCSD science map. This difference is highly statistically significant (ANOVA p value <0.01). For countries, however, Fig. 3g both methods are equally accurate (averages are $\mathrm{AUC}_{\mathrm{research\_space}} = 0.6447$, $\mathrm{AUC}_{\mathrm{UCSD\_science\_map}} = 0.6213$, ANOVA p value >0.05),

indicating that for transitions from intermediate to developed the increase in accuracy observed for the research space is also expressed itself for more disaggregate units (individuals and organizations).

Table 1 summarizes our results. Rows represent the levels of aggregation (individuals, organizations, and countries), and columns represent the transitions studied (inactive to active, nascent to developed, and intermediate to developed).

Discussion

Understanding the structure of research production is important for scientists, universities, and countries, because it can help them comprehend where they are and where they can go. In this paper we contributed to this literature by introducing the research space, a map of science where links connect pairs of fields if individuals are likely to publish in both of them. We used the research space to predict changes in the level of development of individuals, organizations, and countries, for research fields, finding that the research space is a significantly more accurate predictor of the evolution of research output for fine-grained units (individuals and organizations), than the UCSD citation based science map. Both maps, however, are of comparable accuracy when predicting the evolution of the research output of countries, indicating that the research space is particularly relevant for evaluating the research output of individuals and organizations.

Moreover, we found a low correlation between the research space and UCSD citation based science maps. This result shows that citing different fields and the capacity to publish in different fields are not the same.

The research space adds a new layer of information on the relatedness between scientific disciplines that is based more on the shared capacities and goals of scholars than on the information and knowledge flow between academic fields. This is in contrast with cross-disciplinary citations which are more likely to capture how scientific fields borrow methods (e.g. biologist citing statistician), and motivations (e.g. urban economist citing urban planner) from one other.

We suggest, therefore, the use of citation based maps to categorize and understand the knowledge relatedness and flows between scientific fields. The research space, based on the career paths and publishing behavior of authors, is more accurate predictor of the scientific fields into which more disaggregated units (i.e. individuals and institutions) move next. Therefore, it can be a useful tool for research policy and the evaluation of research portfolios.

Our results raise several questions for further research. The first methodological question is if our results would hold also a dynamic classification of science based on changes in paper's topics (Waltman et al. 2010). This analysis would require us to adapt and apply our methods to a dynamic classification of science, which requires advanced techniques in acquiring, storing and analyzing new information on publications such as using on-line machine learning algorithms.

A second qualitative question is the financial cost required to develop each particular research in an area. Simple intuition tells us that the costs required to develop a field vary enormously for different areas of research. Some research fields require large infrastructure investments, like the advanced facilities needed to perform cutting edge work in biology or the accelerators and reactors needed to make progress on particle or plasma physics. Other areas of research, like data science or economics, can be stimulated by opening more

Table 1 Summary results

Transition	Inactive to active ($RCA_{sf} = 0$ to $RCA_{sf} > 0$)		Nascent to developed ($0 < RCA_{sf} < 0.5$ to $RCA_{sf} \geq 1$)		Intermediate to developed ($0.5 \leq RCA_{if} < 1$ to $RCA_{sf} \geq 1$)	
Aggregation	Research space	UCSD science map	Research space	UCSD science map	Research space	UCSD science map
Individuals	AUC = 0.896***	AUC = 0.803	N/A	N/A	N/A	N/A
Organizations	AUC = 0.715***	AUC = 0.687	AUC = 0.693**	AUC = 0.670	AUC = 0.639***	AUC = 0.616
Countries	AUC = 0.682	AUC = 0.682	AUC = 0.639	AUC = 0.624	AUC = 0.645	AUC = 0.621

Levels of aggregation are represented in rows and transitions in columns. Values of Area Under the Curve (AUC) reported, correspond to the average for each transition in each aggregation level. Transitions to a developed state were not evaluated for Individuals, since those transitions are not meaningful because of the small number of publications

*** Significant with $p < 0.01$; ** significant with $p < 0.05$

positions for faculty, graduate students, and postdocs, since the infrastructure costs needed to perform research in these fields are modest compared to the ones needed to perform research in more capital intensive fields. In the future, a methodology to evaluate the potential of success of an individual or organization in a field, together with the costs needed to advance research in that direction, would help provide a tool that policy makers could use to strategize the development of research efforts. Our hope is that the methods advanced in this paper are a step in that direction.

Acknowledgments M.G and C.H were supported by the Massachusetts Institute of Technology MIT Media Lab Consortia and MIT Chile Seed Fund. M.G was supported by the Universidad de Playa Ancha, Chile (ING01-1516) and the Universidad Técnica Federico Santa María, Chile (PIIC). D.H was supported by the Marie Curie International Outgoing Fellowship within the EU 7th Framework Programme for Research and Technical Development: Connecting_EU!—PIOF-GA-2012-328828. M.M was supported by Basal Project FB-0821. C.H was supported by the Metaknowledge Network at the University of Chicago.

References

Abramo, G., & D'Angelo, C. A. (2014). How do you define and measure research productivity? *Scientometrics*. doi:10.1007/s11192-014-1269-8.

Balassa, B. (1965). Trade liberalisation and "revealed" comparative advantage1. *The Manchester School, 33*(2), 99–123. doi:10.1111/j.1467-9957.1965.tb00050.x.

Bollen, J., Van de Sompel, H., Hagberg, A., Bettencourt, L., Chute, R., Rodriguez, M. A., et al. (2009). Clickstream data yields high-resolution maps of science. *PLoS ONE, 4*(3), e4803. doi:10.1371/journal.pone.0004803.

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., et al. (2012). Design and update of a classification system: The UCSD map of science. *PLoS ONE, 7*(7), e39464. doi:10.1371/journal.pone.0039464.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics, 64*(3), 351–374. doi:10.1007/s11192-005-0255-6.

Cimini, G., Gabrielli, A., & Sylos Labini, F. (2014). The scientific competitiveness of nations. *PLoS ONE, 9*(12), e113470. doi:10.1371/journal.pone.0113470.

Collins, H. (2010). *Tacit and explicit knowledge*. Chicago: University of Chicago Press.

Cybermetrics Lab. (2015). About Us | Ranking Web of Universities. Retrieved February 25, 2016, from http://webometrics.info/en/About_Us.

DataViva. (2016). Retrieved February 3, 2016, from http://en.dataviva.info/.

Elhorst, J. P., & Zigova, K. (2014). Competition in research activity among economic departments: Evidence by negative spatial autocorrelation. *Geographical Analysis, 46*(2), 104–125. doi:10.1111/gean.12031.

Fox, M. F. (2005). Gender, family characteristics, and publication productivity among scientists. *Social Studies of Science, 35*(1), 131–150.

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology, 91*(3), 481–510.

Guevara, M., & Mendoza, M. (2013). Revealing comparative advantages in the backbone of science. In *Proceedings of the 2013 workshop on computational scientometrics: Theory and applications* (pp. 31–36). New York, NY: ACM. doi:10.1145/2508497.2508503.

Harzing, A.-W., & Giroud, A. (2014). The competitive advantage of nations: An application to academia. *Journal of Informetrics, 8*(1), 29–42. doi:10.1016/j.joi.2013.10.007.

Klavans, R., & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology, 60*(3), 455–476. doi:10.1002/asi.20991.

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology, 60*(2), 348–362. doi:10.1002/asi.20967.

Moya-Anegón, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., & Munoz-Fernández, F. J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics, 61*(1), 129–145. doi:10.1023/B:SCIE.0000037368.31217.34.

Neffke, F., & Henning, M. (2013). Skill relatedness and firm diversification. *Strategic Management Journal, 34*(3), 297–316.

Neffke, F., Otto, A., & Hidalgo, C. A. (2016). *The mobility of displaced workers: How the local industry mix affects job search strategies*. Retrieved from http://www.frankneffke.com/files/NeffkeOttoHidalgo_DisplacedWorkers.pdf.

Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology, 61*(9), 1871–1887. doi:10.1002/asi.21368.

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences, 105*(4), 1118–1123. doi:10.1073/pnas.0706851105.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*(4), 265–269.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science, 50*(9), 799–813. doi:10.1002/(SICI)1097-4571(1999)50:9<799:AID-ASI9>3.0.CO;2-G.

Waltman, L., van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics, 4*(4), 629–635. doi:10.1016/j.joi.2010.07.002.

Zhuge, H. (2006). Discovery of knowledge flow in science. *Communications of the ACM, 49*(5), 101–107. doi:10.1145/1125944.1125948.