Categorical and Geographical Separation in Science
Sienkiwicz, J.; Soja, K.; Holyst, J.A.; Sloot, P.M.A.

Published in:
Scientific Reports

DOI:
10.1038/s41598-018-26511-4

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
We study scientific collaboration at the level of universities. The scope of this study is to answer two fundamental questions: (i) can one indicate a category (i.e., a scientific discipline) that has the greatest impact on the rank of the university and (ii) do the best universities collaborate with the best ones only? Restricting ourselves to the 100 best universities from year 2009 we show how the number of publications in certain categories correlates with the university rank. Strikingly, the expected negative trend is not observed in all cases – for some categories even positive values are obtained. After applying Principal Component Analysis we observe clear categorical separation of scientific disciplines, dividing the papers into almost separate clusters connected to natural sciences, medicine and arts and humanities. Moreover, using complex networks analysis, we give hints that the scientific collaboration is still embedded in the physical space and the number of common papers decays with the geographical distance between them.

The idea of so-called science of science is not entirely new: 20th century is well known for its critical works of Kuhn, Popper, Lakatos and Feyerabend who tried to build models describing how science should work or, which is far more important, to show how it in fact does work. However it is only in recent times that, owing to the start of the era of overwhelming data, it is now possible to track this problem quantitatively. Several studies are on a journey to answer such intriguing questions like "Who is the best scientist?", “What makes the best university” etc.

There are at least three separate factors that can be regarded as key components of today's science and the way it is recognized: papers, citations and rankings. The last one is devoted rather to whole unities like universities or departments although recent studies consider it also in the scope of individuals. It has been argued that rankings still can be perceived as not enough deep measures “providing finalized, seemingly unrelated indicator values”. On the other hand it is well known that scientific impact is a multi-dimensional construct and that using a single measure is not advisable.

Nonetheless, rankings are clearly a derivative of the number of published papers. However apart from just raw numbers the quality of science comes often with two additional factors: specialization and collaboration. Interestingly the type of the scientific category can dramatically change both the way the paper is written and received, e.g., in the case of simple lexical factors as title length its impact on the acquired citations change significantly from one category to another. In the same manner it is possible to spot that the number of citations per paper can vary by several orders of magnitude and are highest in multidisciplinary sciences, general internal medicine, and biochemistry and lowest in literature, poetry, and dance. These studies can go even as deep as to fascinating notion of scientific meme propagating along the citation graph.

Collaboration has been in the scope of interest for a long time and it is generally considered that it leads to high impact publications. One of recognized factors affecting the level of collaboration is undoubtedly geographic proximity: usually one expects to find a decaying probability of citation as well as common papers with distance, however it can also be connected to such features as ethnicity or level of economic development.

In this study we perform an investigation for a selected group of 100 best universities to unravel how the scientific productivity measured in the number of published papers per scientific categories (e.g, physics, art etc) correlates with the rank of the university. Using Principal Component Analysis (PCA) we study whether scientific categories coming from different areas (natural science, humanities etc) tend to stick together. In the second part of the paper we examine the complex network of scientific collaboration among 100 best universities and study the properties of such a network using the concept of weight threshold.
Results

We use the QS World University Ranking and service Web of Science datasets to examine patterns of category and geographic separation (see Methods for details). The data describes 100 best universities in a form of two matrices $P_{ij}$ (100 universities by 181 categories) and $C_{ij}$ (100 by 100 universities). The first matrix contains information about the number of papers published by a specific university $i$ in a given scientific category $j$ while the second one stores the total number of common papers among universities $i$ and $j$ (regardless of the category).

The main text of this paper concerns absolute numbers of quantities $P_{ij}$ and $C_{ij}$, while the Supplementary Information contains some results for the scaled cases.

Rank–number correlations for categories. It is interesting to understand how the university rank correlates with the number of scientific publications and, which is even far more intriguing, to split these relations according to different scientific categories. Naively one would expect a strong negative correlation between these quantities as larger number of papers should be reflected in acquiring higher rank (thus smaller number). The results for our data analysis are shown in Tables 1, 2 and Fig. 1, where we plot correlation coefficient $\rho$ against the total number of papers $N$ published in the given category (an alternative and much more straightforward method would be to use regression analysis however, in this case, it brings unreliable results - see SI for details). In each case $\rho$ was obtained by taking one of the columns $j$ of matrix $P_{ui}$, ranking it and correlating with the university rank, thus calculating Spearman's rank correlation coefficient. The outcome clearly suggests that there are categories for which we observe even positive correlation coefficient. On the other hand, one has to take into account the fact that in these cases statistical significance of such results is usually very low ($p$-value $> 0.05$) as depicted in Fig. 1. When treated as a whole the data points give evidence of a log-linear relationship $\rho = a + b\log N$ (blue solid line in Fig. 1) between correlation coefficient and the number of papers with $a = 0.098 \pm 0.056$ ($p = 0.08$) and $b = -0.0415 \pm 0.0068$ ($p < 0.001$). A similar fit performed only for the highly significant categories (red solid line in Fig. 1) yields $a = 0.285 \pm 0.072$ ($p < 0.001$) and $b = -0.127 \pm 0.0081$ ($p = 0.13$). An insignificant value of $b$ in this case means that the level of correlations for the selected group of categories is in fact constant, contrary to the previous situation where we observe a significant decrease with $N$. It is worth to mention here that using not absolute but relative numbers of papers (i.e., divide by the total number of papers from a given university) leads to different results where positive correlations for certain categories are significant (see Fig. S1 in Supplementary Information). Interestingly, the category of Multidisciplinary Sciences seems to be unexpectedly robust, regardless of the method used (cf Fig. 1 and S1 in SI) it yields the highest correlation value, which might suggest that interdisciplinary research has a substantial influence on university ranking.

Categorical separation. As a next step of our analysis, we check the hypothesis of categorical separation of science. In order to test this assumption we perform a Principal Component Analysis (PCA) for matrix $P_{ij}$ where we restrict ourselves to those categories that were identified as highly correlated ones (see Fig. 1). Figure 2 presents the results of this PCA: the main panel (Fig. 2a) shows a 3D projection of the original 44 categories onto the first three principal components. As can be seen in Fig. 2d, the first three principal components explain around 75% of data variability. Each category was marked with a color connected to its OECD classification that contains six different areas: Natural Sciences, Engineering and Technology, Medical & Health Sciences, Agricultural Sciences and Humanities, marking with a different color the scientific category Multidisciplinary Sciences. The 3D plot suggests two separate bundles of categories — one connected to medical sciences combined with complementary natural sciences (such as Virology or Cell Biology) and the second identified as mainly social sciences and humanities. Interestingly, such core natural sciences like Physics and Mathematics tend to point in directions separated from these two bundles. The other intriguing fact is almost complete absence of agricultural and engineering sciences (except for one category) in this scheme. Another typical way often used to present the results of PCA is to show them in a form of so-called bi-plot, i.e., two dimensional projections of consecutive PCs. Figure 2b,c provides this additional information: the values of the first PC are if the same sign, while the 2nd PC differentiates between natural sciences and other. It is Fig. 2c that uncovers a very clear distinction among natural sciences, medical sciences and social sciences with humanities. This distinction comes also in a clear way from the cluster analysis — Fig. 2e provides results from k-means algorithm used in case of the outcomes from PCA. When searching for three clusters we restrict ourselves to those categories that were identified as highly correlated ones (see Fig. 1). Figure 2 presents weights from k-means algorithm used in case of the outcomes from PCA. When searching for three clusters we restrict ourselves to those categories that were identified as highly correlated ones (see Fig. 1).

Network analysis. Apart from the categorical point of view we can also consider university quality by analyzing the direct connections between universities $i$ and $j$ on the basis of the collaboration matrix $C_{ij}$, where the element $C_{ij}$ gives the number of common publications of institutions $i$ and $j$. The structure of such a collaboration network is depicted in Fig. 3a where each node (vertex) is a university and links (edges) show the connections between them. The width of each link corresponds to the number of common publications between the universities.

The algorithm used to obtain this structure is the following. Using 100 highest ranked universities, for each of them ($u_1, u_2, \ldots, u_{100}$) we search for its publications $P_{u1}, P_{u2}, \ldots, P_{u_{100}}$. Then, if among the co-authors of $P_i$ there is any that comes from either of the universities $u_2, \ldots, u_{100}$, a link of weight $w = 1$ between those universities (e.g., $u_1$ and $u_2$) is established. The weight is increased by one each time $u_2$ is found among the following publications of $u_1$. Finally the weight of the link between nodes $u_i$ and $u_j$ is just the number of their common publications (as seen in the database).

Weights probability distribution. In order to examine the fundamental properties of the weighted network of collaboration we need to compute link weight probability distribution function (PDF) which can give an idea about the diversity of number of publications between universities. Figure 3b presents link weight PDF, suggesting a fat-tail distribution where the majority of link weights can be found between $w = 1$ and $w = 10$. 

References

Weight threshold. In the following analysis we will use the concept of weight threshold depicted in Fig. 4, let us take the original network of 5 fully connected universities seen in Fig. 4a and assume now that we are interested in constructing an unweighted network that would take into account only the connections with weight higher than a certain threshold weight $w_T$. A possible outcome of this procedure is presented in Fig. 4b - all the links with $w < w_T$ are omitted and as a result we obtain a network where links indicate only connections between nodes (i.e., they do not have any value).

Using weight threshold as a parameter it is possible to obtain several unweighted networks - for each value of $w_T$ in the range $[w_{\text{min}}, w_{\text{max}}]$ we get a different network $N_T(w_T)$ whose structure is determined only by $w_T$. Then, for each of these networks it is possible to compute standard network quantities: (i) number of nodes $N$ that...
have a at least one link (i.e., nodes with degree \( k_i = 0 \) are not taken into account), (ii) Number of edges (links) \( E \) between the nodes, (iii) the average shortest path \( \langle l \rangle \), (iv) clustering coefficient \( C \), (v) assortativity coefficient \( r \) (vi) size \( S \) of largest connected component with number \( n \) of components (see Materials and Methods for details).

**Network observables as a function of weight threshold.** Figure 5 depicts the above described network parameters as a function of the weight threshold \( w_T \). First, as can be seen in Fig. 5a, the number of nodes \( N \) is a linearly decreasing function of the weight threshold \( w_T \). The number edges \( E \) decreases faster, following an exponential function (Fig. 5b). On the other hand the average shortest path \( \langle l \rangle \) (Fig. 5c) is a non-monotonic function of weight threshold, reaching its peak for \( w_T \approx 200 \). Clustering coefficient \( C \) (Fig. 5d) decreases with weight threshold up to the point \( w_T \approx 500 \) where it rapidly drops down to 0. The most interesting is the behavior of \( r (w_T) \) shown in Fig. 5e: the coefficient starts with \( r < 0 \), while for larger thresholds it crosses \( r = 0 \) and for \( w_T \approx 200 \) it takes its maximal value. Then once again it drops down below zero reaching \( r \approx −0.4 \) for \( w_T \) around 500.

<table>
<thead>
<tr>
<th>Category</th>
<th>( N )</th>
<th>( \rho )</th>
<th>Category</th>
<th>( N )</th>
<th>( \rho )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literary Theory and Criticism</td>
<td>560</td>
<td>−0.261**</td>
<td>Literature</td>
<td>4158</td>
<td>−0.189</td>
</tr>
<tr>
<td>Management</td>
<td>5410</td>
<td>−0.242*</td>
<td>Marine and Freshwater Biology</td>
<td>3182</td>
<td>−0.051</td>
</tr>
<tr>
<td>Materials Science</td>
<td>35196</td>
<td>−0.163</td>
<td>Mathematical and Computational Biology</td>
<td>4155</td>
<td>−0.510***</td>
</tr>
<tr>
<td>Mathematics</td>
<td>20834</td>
<td>−0.394***</td>
<td>Mechanics</td>
<td>7236</td>
<td>−0.028*</td>
</tr>
<tr>
<td>Medical Ethics</td>
<td>778</td>
<td>−0.249*</td>
<td>Medical Informatics</td>
<td>1845</td>
<td>−0.401***</td>
</tr>
<tr>
<td>Medical Laboratory Technology</td>
<td>1635</td>
<td>−0.240*</td>
<td>Medicine</td>
<td>28662</td>
<td>−0.393***</td>
</tr>
<tr>
<td>Medieval and Renaissance Studies</td>
<td>720</td>
<td>−0.294***</td>
<td>Metallurgy and Metallurgical Engineering</td>
<td>4295</td>
<td>−0.152</td>
</tr>
<tr>
<td>Meteorology and Atmospheric Sciences</td>
<td>6003</td>
<td>−0.314***</td>
<td>Microbiology</td>
<td>9708</td>
<td>−0.264**</td>
</tr>
<tr>
<td>Microscopy</td>
<td>407</td>
<td>−0.061</td>
<td>Mineralogy</td>
<td>1307</td>
<td>−0.221*</td>
</tr>
<tr>
<td>Mining and Mineral Processing</td>
<td>795</td>
<td>−0.095</td>
<td>Multidisciplinary Sciences</td>
<td>15175</td>
<td>−0.594***</td>
</tr>
<tr>
<td>Music</td>
<td>935</td>
<td>−0.203*</td>
<td>Mycology</td>
<td>588</td>
<td>−0.127</td>
</tr>
<tr>
<td>Nanoscience and Nanotechnology</td>
<td>12710</td>
<td>−0.232*</td>
<td>Neuroimaging</td>
<td>2247</td>
<td>−0.452***</td>
</tr>
<tr>
<td>Neurosciences</td>
<td>36120</td>
<td>−0.445***</td>
<td>Nuclear Science and Technology</td>
<td>3605</td>
<td>−0.186</td>
</tr>
<tr>
<td>Nursing</td>
<td>2923</td>
<td>−0.181</td>
<td>Nutrition and Dietetics</td>
<td>5111</td>
<td>−0.206*</td>
</tr>
<tr>
<td>Obstetrics and Gynecology</td>
<td>8228</td>
<td>−0.345***</td>
<td>Oceanography</td>
<td>2763</td>
<td>−0.159</td>
</tr>
<tr>
<td>Oncology</td>
<td>25768</td>
<td>−0.320***</td>
<td>Operations Research and Management Science</td>
<td>4088</td>
<td>−0.281**</td>
</tr>
<tr>
<td>Ophthalmology</td>
<td>5846</td>
<td>−0.346***</td>
<td>Optics</td>
<td>13579</td>
<td>−0.275**</td>
</tr>
<tr>
<td>Ornithology</td>
<td>408</td>
<td>−0.092</td>
<td>Orthopedics</td>
<td>4399</td>
<td>−0.198*</td>
</tr>
<tr>
<td>Otolaryngology</td>
<td>2320</td>
<td>−0.228*</td>
<td>Paleontology</td>
<td>1733</td>
<td>−0.179</td>
</tr>
<tr>
<td>Parasitology</td>
<td>2200</td>
<td>−0.261**</td>
<td>Pathology</td>
<td>7470</td>
<td>−0.334***</td>
</tr>
<tr>
<td>Pediatrics</td>
<td>9863</td>
<td>−0.317***</td>
<td>Peripheral Vascular Disease</td>
<td>14139</td>
<td>−0.294**</td>
</tr>
<tr>
<td>Pharmacology and Pharmacy</td>
<td>17978</td>
<td>−0.249*</td>
<td>Philosophy</td>
<td>2381</td>
<td>−0.192</td>
</tr>
<tr>
<td>Physics</td>
<td>96469</td>
<td>−0.374***</td>
<td>Physiology</td>
<td>9709</td>
<td>−0.293**</td>
</tr>
<tr>
<td>Planning and Development</td>
<td>1539</td>
<td>−0.343***</td>
<td>Plant Sciences</td>
<td>7240</td>
<td>0.036</td>
</tr>
<tr>
<td>Poetry</td>
<td>237</td>
<td>−0.167</td>
<td>Political Science</td>
<td>4627</td>
<td>−0.307**</td>
</tr>
<tr>
<td>Polymer Science</td>
<td>4909</td>
<td>−0.191</td>
<td>Psychiatry</td>
<td>20036</td>
<td>−0.338***</td>
</tr>
<tr>
<td>Psychology</td>
<td>36186</td>
<td>−0.272**</td>
<td>Public</td>
<td>18308</td>
<td>−0.305**</td>
</tr>
<tr>
<td>Public Administration</td>
<td>1046</td>
<td>−0.165</td>
<td>Radiology</td>
<td>12963</td>
<td>−0.323**</td>
</tr>
<tr>
<td>Rehabilitation</td>
<td>3833</td>
<td>−0.094</td>
<td>Religion</td>
<td>2140</td>
<td>−0.157</td>
</tr>
<tr>
<td>Remote Sensing</td>
<td>1367</td>
<td>−0.200*</td>
<td>Reproductive Biology</td>
<td>4315</td>
<td>−0.217*</td>
</tr>
<tr>
<td>Respiratory System</td>
<td>7071</td>
<td>−0.347***</td>
<td>Rheumatology</td>
<td>5928</td>
<td>−0.238*</td>
</tr>
<tr>
<td>Robotics</td>
<td>2447</td>
<td>−0.199*</td>
<td>Social Issues</td>
<td>1503</td>
<td>−0.347***</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>7206</td>
<td>−0.462***</td>
<td>Social Work</td>
<td>1035</td>
<td>−0.201*</td>
</tr>
<tr>
<td>Sociology</td>
<td>5477</td>
<td>−0.319**</td>
<td>Soil Science</td>
<td>1142</td>
<td>−0.015</td>
</tr>
<tr>
<td>Spectroscopy</td>
<td>3043</td>
<td>−0.249*</td>
<td>Sport Sciences</td>
<td>4172</td>
<td>−0.093</td>
</tr>
<tr>
<td>Statistics and Probability</td>
<td>6058</td>
<td>−0.517***</td>
<td>Substance Abuse</td>
<td>3324</td>
<td>−0.255*</td>
</tr>
<tr>
<td>Surgery</td>
<td>16669</td>
<td>−0.301**</td>
<td>Telecommunications</td>
<td>9861</td>
<td>−0.191</td>
</tr>
<tr>
<td>Theater</td>
<td>402</td>
<td>−0.164</td>
<td>Thermodynamics</td>
<td>2154</td>
<td>−0.197*</td>
</tr>
<tr>
<td>Toxicology</td>
<td>3923</td>
<td>−0.168</td>
<td>Transplantation</td>
<td>5870</td>
<td>−0.292**</td>
</tr>
<tr>
<td>Transportation</td>
<td>1158</td>
<td>−0.184</td>
<td>Transportation Science and Technology</td>
<td>1991</td>
<td>−0.089</td>
</tr>
<tr>
<td>Tropical Medicine</td>
<td>1714</td>
<td>−0.312**</td>
<td>Urban Studies</td>
<td>1044</td>
<td>−0.191</td>
</tr>
<tr>
<td>Urology and Nephrology</td>
<td>8348</td>
<td>−0.283**</td>
<td>Veterinary Sciences</td>
<td>5039</td>
<td>−0.063</td>
</tr>
<tr>
<td>Virology</td>
<td>5516</td>
<td>−0.336***</td>
<td>Water Resources</td>
<td>3716</td>
<td>−0.054</td>
</tr>
<tr>
<td>Zoology</td>
<td>6031</td>
<td>−0.176</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Correlation coefficients in categories (ctnd).
it increases toward zero for larger $w_T$. In the case of largest connected component $S$ (Fig. 5f) we observe a series of rapid decreases, e.g., for $w_T \approx 100$ where $S$ drops down by 20%. These results are quantitatively different from the ones obtained by randomly reshuffling the weights of the network (see SI for details).

Network visualisation. The above described non-trivial behavior of quantities $r$, $C$ and $\langle l \rangle$ and $S$ cannot be the sole cause of the relations presented in Fig. 3b although a high number of points with $w_T \approx 100$ can be responsible for some of these effects. It seems that there has to be another phenomenon leading to such an effect. Using R’s `igraph` package we visualize connections between universities and community structure (denoted by color) for different values of $w_T$. The results for $w_T = 100, 200, 300$ and $w_T = 400, 500, 1000$ are shown in Figs 6 and 7, providing an input for further analysis. For $w_T = 100$ (Fig. 6a) the network is still percolated, i.e., it is possible to reach any node from another one; over that value a separation occurs - Chinese, Australian and Singapore, Japanese, Danish and Swedish as well as Swiss universities all form separate clusters. This observation is connected with large loss of $S$ in Fig. 5f. The remaining giant cluster is built out of American, Canadian, British, Dutch, and German universities (Fig. 6b). This is the area where both average path length $\langle l \rangle$ and assorativity $r$ take their maximal values. For $w_T = 300$ we witness the separation between US and British universities and from now on (with small exceptions) different clusters can be described as connected to different countries (or even smaller administrative units as English and Scottish universities are separated). Further plots depict progressing decay of connections between the universities that form either star-like structures (Japanese, Canadian, English and American in Fig. 7a,b) or ultimately chains (Fig. 7c).

A possible explanation to this phenomenon is in the geographical distance between the universities. In fact, Fig. 8 supports partially this assumption. The number of publications between universities $i$ and $j$ can be fitted with a decreasing power-law function of the geographical distance between them. The gap around $d = 5000$ is most probably caused by the presence of continents. Similar results regarding the role of geographical distance in science were obtained in previous studies. On the other hand the error bars in Fig. 8 give evidence that for relatively short distances ($d \in [1; 300]$ km) the number common papers can be considered constant. This in turn would support the hypothesis of country-driven rather than geographically-driven collaboration. A lower than expected value of collaboration for shorter distances could also have its origin in the fact that usually there is lack of universities of the same scientific profile in the direct vicinity.

Conclusions
Our results indicate that even such fundamental and straightforward analysis as calculation of correlation coefficient between position of the university in the ranking and the ranked number of papers $N$ in this category (shown as X-axis). The colors reflect statistical significance of the measure (see legend) and category names are shown only for the most significant points ($p$-value < 0.001). Solid lines represent log-linear fits to all points (blue) and most significant points ($p$-value < 0.001, red). Shades surrounding the lines represent 95% confidence interval.

Figure 1. Correlations coefficients. Each data point represents a separate scientific category and gives the Spearman’s correlation coefficient $\rho$ between the rank of the university and the ranked number of papers $N$ in this category (shown as X-axis). The colors reflect statistical significance of the measure (see legend) and category names are shown only for the most significant points ($p$-value < 0.001). Solid lines represent log-linear fits to all points (blue) and most significant points ($p$-value < 0.001, red). Shades surrounding the lines represent 95% confidence interval.
The second part of the paper is devoted to network analysis of the collaboration among 100 best universities. We used the concept of weight threshold to obtain several slices of the original weighted network at different levels of collaboration intensity. Treating the threshold as a control parameter we were able to track such network.

Figure 2. Principal Component Analysis (PCA) of scientific category data. Given the number of papers each of the 100 universities published in 44 different scientific categories (chosen according to results obtained in Fig. 1) we perform Principal Component Analysis. Panel (a) presents the outcome for three most important principal components: each arrow represents the position of an original category (e.g., Physics, Multidisciplinary Sciences) in the new set coordinates. The colors of arrows are connected to the OECD classification29 (see legend). Panels (b) and (c) show the projection of PCA results onto, respectively, 2nd PC — 1st PC and 3rd PC — 2nd PC planes. Panel (d) presents the cumulative value of variance explained by the consecutive PCs. Panel (e) shows the outcomes of cluster analysis (k-means algorithm) for the results obtained by PCA (we set the number of clusters to 3).

Figure 3. (a) Representation of the university collaboration network. Each node is a university and links show the connections between them. The width of each link corresponds to the number of common publications between the nodes in question. (b) Link weight probability distribution function (PDF).

The second part of the paper is devoted to network analysis of the collaboration among 100 best universities. We used the concept of weight threshold to obtain several slices of the original weighted network at different levels of collaboration intensity. Treating the threshold as a control parameter we were able to track such network.
notice that one has to form a very specific query in order to get rid of severe mistakes. Table 3 shows an exemplary list of the search universities together with the exact search phrase that had to be used.

Given the masses of data the usage of unsupervised methods that require no manual classification of documents is the best choice to track the evolution of science. In this way such phenomena as convergence and divergence could also serve as an index for changing relations among disciplines. In particular, one may use it as indicator of the interest a certain scientific area gains over the years. It is possible to spot the emergence of certain trends in science and, in effect, react by for example establishing a new direction of research in the university.

When used for temporal data, our analysis of principal components basing on the number of published papers showing which particular links break up first. Although the geographical proximity is an important factor, the law does not forbid long-distance interactions. In this study we restricted ourselves to only top universities – ones that are “obviously” similar) they earn an additional dimension when treated as a dynamical process. Given the masses of data the usage unsupervised methods that require no manual classification of documents is the best choice to track the evolution of science. In this way such phenomena as convergence and divergence can be instantly spotted. When used for temporal data, our analysis of principal components basing on the number of published papers showing which particular links break up first. Although the geographical proximity is an important factor, the law does not forbid long-distance interactions. In this study we restricted ourselves to only top universities – ones that are “obviously” similar) they earn an additional dimension when treated as a dynamical process. Given the masses of data the usage unsupervised methods that require no manual classification of documents is the best choice to track the evolution of science. In this way such phenomena as convergence and divergence can be instantly spotted.

Geographical distances among the nodes of the network usually come in the form of Tinbergen’s gravity model. Manifestations of spatial embedding of networks are truly omnipotent, ranging from the original inter-country trade through inter-city telecommunication flows and online friendship to active protest-ers. In the case of scientific collaboration Pan et al. show a clear preference for researchers to seek partners in their geographical proximity, however underlining that the very form of the gravity model (i.e., a power law) does not forbid long-distance interactions. In this study we restricted ourselves to only top universities showing which particular links break up first. Although the geographical proximity is an important factor, the results clearly show that in the case of small distances the connections are not formed distance-wise but rather country-wise. Moreover it also seems that the choice of data handling method (absolute values vs. normalized one) can play a crucial role: the description as well as Figs S2 and S3 in the Supplementary Material reveal a strong clustering between continents for the normalized data.

Discussion

The problem of the role of scientific categories and relations among them has intrigued the greatest minds of the past century. Lately, Dias et al. have explicitly quoted Karl Popper’s The Nature of philosophical problems and their roots in science where this great philosopher had questioned the traditional identification of scientific disciplines, convinced instead that one should rather look at cognitive and social aspect thereof. Dias et al. follow this trail by comparing coincidences among disciplines retrieved by (i) classification given by experts, (ii) Jaccard-like coefficient for citations and (iii) language-based Jensen-Shannon measure of dissimilarity in articles’ abstracts. The same aspect, although in much more indirect way, has been lately addressed by one of us, arguing that scientific segregation is visible even while examining relations between text length (or emotional content) and citation patterns. While these considerations may seem to be academic (e.g., detecting similarities among disciplines that are “obviously” similar) they earn an additional dimension when treated as a dynamical process.

Methods

Dataset. We used two prominent data providers: QS World University Ranking and Web of Science service. The first dataset consisted of 100 best universities ranked in the year 2009. The second dataset was obtained by querying the database of years 2008–2009 for publications coming from one of the above mentioned universities and store information about so-called subject category (i.e., the scientific category) and affiliation of co-authors. The obtained matrices $P_{ij}$ (100 universities by 181 categories) and $C_{ij}$ (100 by 100 universities) that were created on-the-fly without physically saving partial data contain, respectively, 1363821 and 496684 papers.

Abbreviations. The seemingly straightforward procedure of querying for a specific university name encounters some problems that could have a strong impact on the further results. Web of Science has a set of abbreviations commonly used for searching such as Univ for “University” or Coll for “College”. Moreover it is essential to notice that one has to form a very specific query in order to get rid of severe mistakes. Table 3 shows an exemplary list of the search universities together with the exact search phrase that had to be used.
Ambiguity of queries. The ’Search’ field is a search key that we use to associate with the authors of the publications and it can consist of one of the operators: + stands for AND operator in Boolean logic and | stands for NOT operator in Boolean logic. These operators are used to clearly assess the origin of the publication. Table 2 shows that using just the names of universities from the list (first column) would lead in the case of number 98 to obtaining publications of both Technical University in Munich and University of Munich, instead of just the latter.

Figure 5. Comparison of collaboration networks observable as functions of weight threshold $w_T$: (a) number of nodes $N$, (b) number of edges $E$, (c) average shortest path $\langle l \rangle$, (d) clustering coefficient $C$, (e) assortativity coefficient $r$, (f) size of the largest connected component $S$ (red points) and number of components $n$ (grey points).
Network analysis. Clustering coefficient $C_i$ for node $i$ is defined as the number of existing links among its nearest neighbors $e_i$ (i.e., nodes to which it has links) divided by the total number of possible links among them $k_i(k_i - 1)/2$

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$

(1)
The total clustering coefficient for the whole network is calculated as the average over all \( C_i \).

Assortativity coefficient \( r \) defined by

\[
    r = \frac{\frac{1}{2} \sum_i j_i k_i - \left( \frac{1}{2E} \sum j_i + k_i \right)^2}{\frac{1}{2E} \left( \sum j_i^2 + k_i^2 \right) - \left( \frac{1}{2E} \sum j_i + k_i \right)^2}
\]

Figure 7. Snapshots of network topology for different thresholds: (a) \( w_T = 400 \), (b) \( w_T = 500 \) and (c) \( w_T = 1000 \). The colors of vertices correspond to the assignment from a community detection algorithm (fast greedy modularity optimization algorithm\(^47\)) and therefore they can change from one panel to another. Plots were created combining open-source packages igraph\(^31\) (nodes and links) and maps\(^48\) (world map) for R language\(^30\).
where \( i \) goes over all edges in the network. The coefficient is in the range \([-1; 1]\), \( r = 1 \) means that the highly connected nodes have the affinity to connect to other nodes with high \( k_i \) while \( r = -1 \) happens when highly connected nodes tend to link to nodes with very low \( k_i \).

Average shortest path \( \langle l \rangle \) is calculated as the average value of shortest distance (measured in the number of steps) between all pairs of nodes \( i, j \) in the network.

**References**


**Acknowledgements**

The work was supported by FP7 FET Open project Dynamically Changing Complex Networks-DynaNets EU Grant Agreement Number 233847. JAH and PMAS also acknowledge the Russian Science Foundation grant 14-21-00137. The work was partially supported as RENOIR Project by the European Union Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 691152 and by Ministry of Science and Higher Education (Poland), grant Nos W34/H2020/2016, 329025/PnH/2016 and National Science Centre, Poland Grant No. 2015/19/B/ST6/02612.

**Author Contributions**

J.A.H. and P.M.A.S. conceived the study, K.S. collected the data, K.S. and J.S. analyzed the data, J.S. wrote the manuscript. All authors reviewed the manuscript.

**Additional Information**

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-26511-4.

**Competing Interests:** The authors declare no competing interests.

**Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.