Exploring Innovation: Unveiling the Geographic Network of Scientific Advancements

Florian Comte[a,c] **and Thomas Trevisan**[b,c]

In the latest years the amount of ground breaking research has been decreasing, even thus the amount of papers published per year is generally increasing. In this setting, it's even more essential for young people who intend to grow in an important and innovative environment to know which universities, and in general which countries, are revolutionizing scientific fields the most. In our study, we provide an extensive analysis on a network of universities collaborations that published ground breaking papers on Nature during 2023. To find such universities and countries, we relied on widely used techniques like centrality measures and community detection algorithms, but also on less used ones like dictionary-based topic analysis. Comparing the results of these techniques we get that centrality measure provide results that are more in line with online rankings, while the text based analysis promotes unexpected countries. Generally, we get that United States and United Kingdom have a very high quality of research, as one would expect.

Research | Universities | NLP | Social network analysis | Scientific fields

Significance Statement

In recent years, there's been a big increase in scientific and technological knowledge, which should make it a great time for major breakthroughs. However, studies show that progress is slowing down in important fields of science. In this paper we aim at finding universities, and more generally countries, where the research is quality oriented, and where it is most likely to bring innovation. Using research articles from Nature, we built several networks (one for each science field we analyzed) of approximately 1000 nodes and 3000 edges each, to see where is the majority of breakthroughs taking place.To do so, we exploited centrality and community detection techniques on such networks, and compared the results with an approach based on papers content.

In recent decades, there has been a marked exponential expansion in the corpus of contemporary scientific and technological knowledge, creating conditions that should be ripe for substantial advances(1). Yet contrary to this view, studies suggest that progress is slowing in several major fields (2). In this perspective, the ratio of research relevant with respect to its volume is decreasing, and it is harder and harder for young talents that aim to innovate to find a research quality-oriented universities that can help reaching their goal. In this paper, we provide an extensive research that seeks for the most thriving university in the past year according to some of the major fields of science. In order to find innovative papers, we relied on data from the Nature journal, one of the best scientific journals in the world, under the assumption that in this journal are published papers mostly showing scientific breakthroughs. This assumption is most validated by Nature selection criteria, selecting papers that 'report original scientific research (the main results and conclusions must not have been published or submitted elsewhere) and are of outstanding scientific importance'. We considered papers from 2023 published in 7 wide fields of science, like: physics, mathematics and computing, ecology, genetics, microbiology, diseases and health care. The gathered data allowed for the construction of a multigraph of universities and their research collaborations, that has served to the application of widely used measures, such as centrality (3, 4)and community detection(5), along with less used ones like dictionary-based topic analysis(6, 7). As centrality measures we selected degree and betweenness centrality to both identify universities and countries that are prominent within the network and that are playing pivotal roles in maintaining efficient communication and collaboration between disparate academic institutions. Concerning community detection, we tested modularity of partitions created with the famous Louvain algorithm and compared them with a partition by country of belonging. Finally, dictionary-based methods were used to provide a different measure of prominence of countries, and therefore universities, in each and every field of science we analyzed. This dictionary-based approach provides us with a country score for every science field studied, based on a field score assigned to words found in the articles. In the following section, we present our results, where we analyze and compare the different measures we used to access the landscapes of innovation.

1. Results

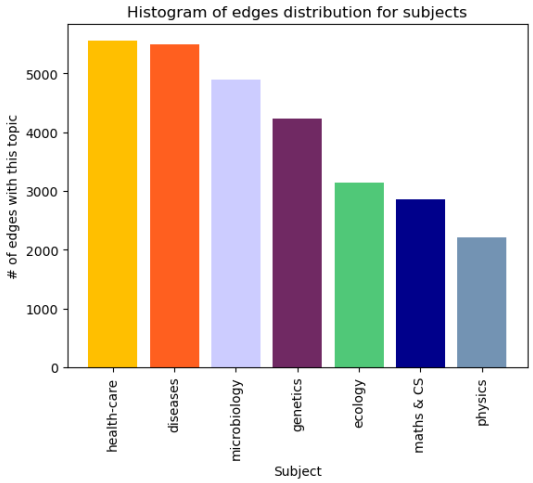
Taking advantage of data scraped from Nature, we built a multigraph that comprises edges from several scientific fields. The edges distribution (Section 3.A of notebook) indicates less links between universities in the mathematical fields such as physics and mathematics and computing,

Author affiliations: [a]Dipartimento di Ingegneria e Scienza dell'Informazione, University of Trento, 38122 Trento; [b]Faculté informatique et communications (IC), École polytechnique fédérale de Lausanne (EPFL) CH-1015 Lausanne; [c]Department of Applied Mathematics and Computer Science, Section for Human-Centerd Artificial Intelligence, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

Florian Comte took care of: Scraper, Network 2D and 3D Plotting, TR-TF wordclouds
Thomas Trevisan took care of: Network creation, Network analysis, TF-IDF wordclouds, Topic based dictionary analysis
Both authors contributed equally on: paper writing

while more links in life sciences fields such as healthcare, diseases and microbiology, as shown in Figure 1.

Fig. 1. Illustration of the number of collaborations for each field we are studying. Observing the histogram, it becomes evident that health-care, diseases and microbiology are the three prominent fields, followed by the clusters of ecology, genetics. In contrast, mathematics and physics exhibit a lower collaboration level. This leads us into thinking that the universities in field of physics and mathematics are collaborating less during 2023 than other fields or that the number of universities for each article in these fields are lower (or clustered, if it's always the same universities on a subject).



This can happen because mathematical fields are inherently less cooperative than life sciences fields, but it most likely to happen because, even thus the amount of articles per field is the same, there are less universities working and publishing in mathematical fields than in life sciences fields. The latter reason is also confirmed by the amount of nodes in the networks, being lower for the physics and mathematics and computing networks. A first measure of importance is the degree distribution for each derived network. By analyzing it, we found out our networks resemble a scale-free behaviour, having a degree exponent between 2 and 3 (except ecology, which has 3.1) (8) and satisfying the friendship paradox. In Fig 2 is reported the healthcare network according to degree, taken as example since it is the one with the most edges.

In table 1 we listed for every field the top universities according to their degree.

Another measure we used to assess university importance is the broadly used betweenness centrality. This measure allowed us to reward universities based on the diversity of links and based on the connectivity they provide to the network as a whole. These nodes work as bridges for the knowledge spread around different universities. Table 1 reports the top universities according to betweenness centrality.

We can notice that most of universities come from USA and UK. This means that these 2 countries are in general very considered for research collaborations, and probably very active in the fields in general.

It could still be the case that universities in these two countries have high betweenness because of the community they belong to. Indeed we verified that universities from the same country tend to cooperate more with one another, possibly leading to increase in centrality. Furthermore,

Fig. 2. 3D representation of the health-care biggest component. South Korea emerges as the predominant country with the largest degrees, indicating substantial collaborative activity. A discernible cluster on the right side suggests extensive intra-country collaborations within South Korea. We can also see that we don't have one outlier, but the dispersion seems to be very balanced.

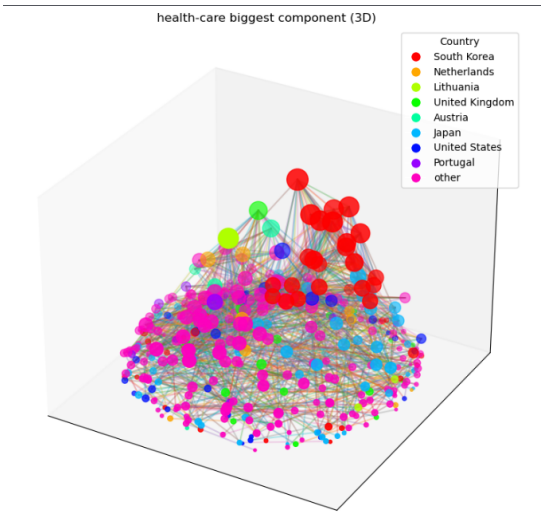


Table 1. This table represents the best universities according to their degree in the network and their betweenness centrality in the network. For every field, the first university according to degree and the first according to betweenness centrality are printed respectively. The two scores are not to be compared, since they use different scales.

Field	University	Degree	Betweenness
Healthcare	Seoul National University H.	**113**	0.017
	University of Pennsylvania	42	**0.090**
Diseases	University of Toronto	**128**	0.0
	University of Oxford	69	**0.004**
Microbiology	University of Liverpool	**190**	0.0
	Liverpool School of Tropical Med.	86	**0.043**
Genetics	Tampere University	**107**	0.002
	University of Southern California	41	**0.072**
Ecology	University of California	**73**	0.033
	University of Copenhagen	61	**0.079**
Maths and Cs	King Saud University	**152**	0.016
	Lebanese American University	138	**0.023**
Physics	University of California	**57**	0.078
	University of California	57	**0.078**

the top universities for centrality belong all to the biggest community, suggesting a wide network of cooperations.

To verify this, we also generated a clustered version of the networks, where nodes are countries and edges are a summation of edges between universities from those countries. This additional networks yield result in agreement: USA is the top central country according to most fields. The data about country centrality can be found in 2. We went on verifying if the countries these universities are from tend to use a language that represents the field they excel in. To do this, we trained a matrix of tf-idf scores per field for every word using articles that weren't used to build the network. The most important words per every field are shown in Fig 3. After training our tf-idf weights, we computed the text-based score of countries using papers that were selected to build

Table 2. This table represents the best countries according to their betweenness centrality and to their text-based score. For every field, the first country according to centrality and the first according to text-based score are printed respectively. The two scores are not to be compared, since they use different scales.

Field	Country	Centrality Score	Text-Score
Healthcare	USA	**0.254**	1.9
	Gambia	0.003	**15.8**
Diseases	USA	**0.110**	2.1
	Azerbaijan	0.0	**11.9**
Microbiology	USA	**0.113**	3.5
	Bulgaria	0.0	**23.8**
Genetics	France	**0.110**	2.4
	Malawi	0.01	**6.9**
Ecology	USA	**0.111**	3.1
	Mongolia	0.0	**31.0**
Maths and CS	USA	**0.159**	1.4
	Uzbekistan	0.0	**17.6**
Physics	USA	**0.097**	1.4
	Kosovo	0.0	**9.2**

the network. The results of countries that tend to use a very field-oriented language in their papers can be found in Table 2.

Fig. 3. Wordclouds for each subject (using sklearn). The wordclouds can help us identyifing the current focus of the research in each field, the one which is yielding more results. For example, we can see that in the mathematics and computing wordcloud there are words such as *sentiment* and *lstm*, suggesting that language models are indeed prominent in the field at the moment.

2. Discussion

In conclusion, our methods yields very different results. On the one hand, using network analysis we obtained top universities that somewhat agree with most used worldwide rankings, such as the World Universities Ranking by Times Higher Education. Therefore, we find among the top universities in our network universities in California for physics and universities in South Korea for healthcare. On the other hand, our text-based scoring resulted in the rise of less developed countries, such as Gambia for healthcare and Kosovo for physics. This is most likely due to a normalization factor that penalizes more countries with a lot of published papers. Overall, we got two very different measure about how much have universities and countries been involved in certain fields of science.

A. Future work. For future work, it could be interesting to compute a text-based score on single universities more than on countries. A task to leave for other work is also to find a better normalization factor for such score. Regarding the network, we think it could be interesting to study the behaviour of the networks over time, to see how collaboration patterns evolved. Going further, it would be interesting to relate those patterns to various situations and conditions of the countries: how does it relates to current wars? Does it depend on the amount of money spent on R&D? Has the modularity of countries increased over the years, and does it correlates to the level of globalization, and the pace globalization is going at? Lastly some other important journal could be added to the network, in order to make the analysis more representative of the actual situation.

3. Methods

A. Data Extraction (1.A in notebook).

A.1. Scraping using BeautifulSoup. In our research, we employed the BeautifulSoup library for web scraping on the nature.com website. The initial step involved extracting information about the authors of the articles. Subsequently, we performed preprocessing and extracted collaboration details for each article, saving the obtained data files. Simultaneously, we stored the text of the articles in separate folders for future processing. We also stored in json files the universities and collaborations.

A.2. Dataset Segmentation. We divided our article dataset into two halves. One half was used to generate the network (as explained in the previous sub section) and the other half was utilized to train a TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer. This approach ensures that the vectorizer is trained on papers not included in our network. The rationale behind this is to prevent the vectorizer from being biased towards the specific content of our network, allowing it to generalize better to a broader set of articles. Consequently, we can subsequently apply the learned tfidf values to the words in papers associated with the network articles, enhancing the overall robustness of our analysis.

B. Data Preprocessing (1.A. in notebook). Before constructing the networks, a comprehensive data preprocessing pipeline was implemented to ensure the integrity and relevance of the extracted information. The primary focus was on refining the representation of universities and their collaborations derived from nature.com articles. The key steps involved in this preprocessing phase are detailed below:

B.1. University Data Extraction. To compile a dataset of universities, the scraper meticulously examined each article, extracting affiliation addresses from authors. The extracted addresses, such as "Zhejiang University of Science and Technology, Hangzhou, China," were subject to a meticulous filtering process. Since nature.com encompasses a broad spectrum of contributors, including institutes and hospitals, the goal was to retain only affiliations associated with academic institutions.

B.2. Affiliation Address Parsing. Each extracted address underwent parsing, employing comma separation to isolate individual components. Keywords indicative of academic affiliations, including "University," "Politecnico," "Escuela," "École," "Universitat," "Università," among others, were used to filter and identify university names within the parsed components.

B.3. Mapping to Actual Addresses. The resulting university names were mapped back to their original addresses, forming associations like "Zhejiang University of Science and Technology" to its complete address, such as 'Zhejiang University of Science and Technology, Hangzhou, China.'

B.4. Clustering the university names. Given that a single university may have multiple addresses, it is necessary to organize them in groups, associating each university name with a list of corresponding exact addresses. To achieve this, a machine learning model was utilized. The model encoded university names, and a cosine similarity matrix was generated to assess the similarity between names. We were then able to create similar groups which contains all the exact addresses for a certain abstract university name.

B.5. Country Standardization. Incorporating a country standardization step, each specific country mentioned in the dataset was associated with a generic representation. This step facilitated subsequent analyses by allowing the use of abstract country names, ensuring consistency in the representation of country data.

The preprocessing phase culminated in a refined dataset, where universities were represented uniformly, collaboration data was structured, and country names were standardized. This prepared dataset laid the foundation for the subsequent construction of meaningful networks and facilitated a more accurate and insightful analysis of the collaborative landscape among academic institutions.

C. Network Generation.

C.1. Multigraph network (2.A. in notebook). We used the library NetworkX to generate a multigraph network. In this network, each node corresponds to a university along with its respective country. An edge in the network signifies a collaboration on a specific subject between two universities. The network is undirected.

Our approach involves creating individual edges for each subject-specific collaboration between universities. Additionally, we include an overarching edge between two universities, denoted by the subject attribute "All." The weight of this overall edge is set as the sum of the weights of all subject-specific collaborations between the two universities.

C.2. Cleaning the network (2.B. in notebook). We decided to remove isolated nodes. Isolated nodes refer to universities that do not engage in any collaborations within the given network context. We also removed the self loops because we want to study the collaborations between different universities.

C.3. Creation of subnetworks for each subject analyzed (3.B. in notebook). Having gained a clearer understanding of the data distribution across subjects, our next step involves building subnetworks. Each subject will be represented as an individual network, and concurrently, we will create a comprehensive subnetwork encompassing all connections between universities, regardless of the subject. This approach allows us to explore both subject-specific collaborations and the overarching collaborative landscape across all subjects.

D. Analytical Methods.

We will primarily leverage the NetworkX library to conduct a comprehensive analysis of the network. Below is a compilation of the specific methods we employed for network analysis, along with their respective purposes.

D.1. Basic Stats (4.B in notebook).

Definition Basic statistical analysis involving the calculation of fundamental network metrics, such as the count of nodes (universities) and edges (collaborative connections) in the subnetworks.

Purpose To provide foundational insights into the network's size, composition, and the top three universities actively engaged in collaborative efforts.

D.2. Degree Distributions (4.3 in notebook).

Definition An examination of the distribution of node degrees, where the degree represents the number of collaborations a university has. This is visually represented through both linear and log-log plots.

Purpose A deeper exploration of connectivity patterns, aiming to identify whether the network exhibits scale-freeness. Scale-freeness suggests the presence of a few highly connected universities, indicating potential hubs of influence.

D.3. Exponent (4.D in notebook).

Definition The determination of the power-law exponent involves utilizing the powerlaw package to fit a power-law distribution to the node degrees within the collaboration network. The power-law exponent serves as a quantitative measure, characterizing the degree distribution and offering insights into the network's resilience, robustness, and the potential presence of influential hubs.

Purpose This process unveils the underlying scale-free nature of the network, indicating the presence of a few highly connected nodes that significantly influence its structure.

D.4. 2D and 3D Plots (4.H in notebook).

Definition Visualization of the collaboration network in two and three dimensions, providing a graphical representation of the relationships and overall connectivity.

Purpose Offering an intuitive interpretation of the network structure, aiding in the identification of collaborative patterns and potential hubs.

D.5. Wordcloud (7. in notebook).

Definition A visual representation of the most frequently occurring words or terms in the articles, with word size indicating frequency.

Purpose Offering a visually appealing summary of key themes and topics within the collaboration network, providing a quick overview of prevalent research areas.

D.6. Community Detection (6. in notebook).

Definition Application of the Louvain algorithm for detecting communities within the network, with an analysis based on country divisions.

Purpose Identification of cohesive groups of universities with strong collaborative ties, allowing for a nuanced examination of regional or national collaboration patterns.

D.7. Betweenness Centrality (5. in notebook).

Definition Betweenness centrality measures the extent to which a university acts as a crucial intermediary in connecting other universities within the collaboration network. It quantifies the number of shortest paths passing through a specific university, highlighting its role as a bridge or mediator in facilitating collaborations.

Purpose This analysis aims to identify universities playing pivotal roles in maintaining efficient communication and collaboration between disparate academic institutions. It provides insights into potential influencers and key connectors within the network.

D.8. Dictionary-Based Topic Analysis (8. in notebook).

As an additional measurement, we provided a text-based approach to score countries based on the language it is used in their papers. To do this, we trained a tf-idf matrix of weights using articles that we saved for this task. These articles were not considered for the creation of the network. Using this set as a training set, we created a dictionary that, given a word, it provided how much that word was important in each of the fields taken in consideration. At first we tried an approach based on Term Frequency-Term Ratio, but not having good results, we relied on the TF-IDF algorithm provided by the sklearn library. Then, we applied the learnt weights in the dictionary to the articles we considered to create the network, obtaining for every article a score for every field, based on the sum of the scores of words used in such article. Now, if a word is found with no associated weight, we will not consider such word for the summation. We sum the article score to every country that contributed to that article, and normalize by dividing the summation for the amount of articles published by that country.

Definition Dictionary-based topic analysis involves employing a predefined set of keywords or terms to categorize the topics of articles.

Purpose The goal is to gain deeper insights into the thematic focus of collaborations among universities, facilitating a qualitative understanding of research trends and areas of emphasis.

1. TMA Fink, M Reeves, R Palma, RS Farr, Serendipity and strategy in rapid innovation. *Nat. Commun.* **8**, 2002 (2017).
2. N Bloom, CI Jones, J Van Reenen, M Webb, Are ideas getting harder to find? *Am. Econ. Rev.* **110**, 1104–44 (2020).
3. J Scott, *Network analysis: A handbook.* (Sage Publications), (1992).
4. S Wasserman, K Faust, Social network analysis: Methods and applications. (1994).
5. S Fortunato, D Hric, Community detection in networks: A user guide. *Phys. Reports* **659**, 1–44 (2016) Community detection in networks: A user guide.
6. SP Kasiviswanathan, P Melville, A Banerjee, V Sindhwani, Emerging topic detection using dictionary learning in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11. (Association for Computing Machinery, New York, NY, USA), p. 745–754 (2011).
7. L Guo, CJ Vargo, Z Pan, W Ding, P Ishwar, Big social data analytics in journalism and mass communication. *Journal. Mass Commun. Q.* **93**, 332–359 (2016).
8. AL Barabási, M Pósfai, *Network science.* (Cambridge University Press, Cambridge), (2016).