June 2021

Comparing the United States' and China's Leading Roles in the Landscape of Science

CSET Data Brief

AUTHORS
Autumn Toney
Melissa Flagg

Introduction

Research output is a frequently used index to assess the global competition for leadership in science and technology (S&T). This competition—among countries as well as institutions—leads to challenging questions: Which countries are leading in publication output? Which institutions are producing the most influential research? Which organizations are investing the most in research?

These questions are frequently addressed with broad overviews of the research landscape, but a focused analysis on subsets of research provides more nuanced and accurate comparisons. While one entity may appear to dominate in a broad area of research, it might fall in the ranks when the research area is broken down into subsets.

To explore country-level research output at varying levels of granularity, we navigate the landscape via CSET's recently developed Map of Science.[1] Using this clustering of scientific research publications that is sourced from a massive database, we analyze scientific publication output at three different levels of aggregation: 1) research clusters, 2) research regions, and 3) research districts. We present a comparative analysis of U.S. and Chinese research publication outputs using these different aggregations of scientific research publications.

This work lays the foundation for further studies of the data to understand the nuances in scientific research and thorough analyses of research output competition.

Key findings include:

- Different levels of scientific research publication aggregation can lead to different analytical conclusions—specificity matters.
- In a granular view of global scientific research, the United States and China, combined, dominate almost two-thirds of the research publication output. However, the rest of the world leads in more than one-third of publication output.

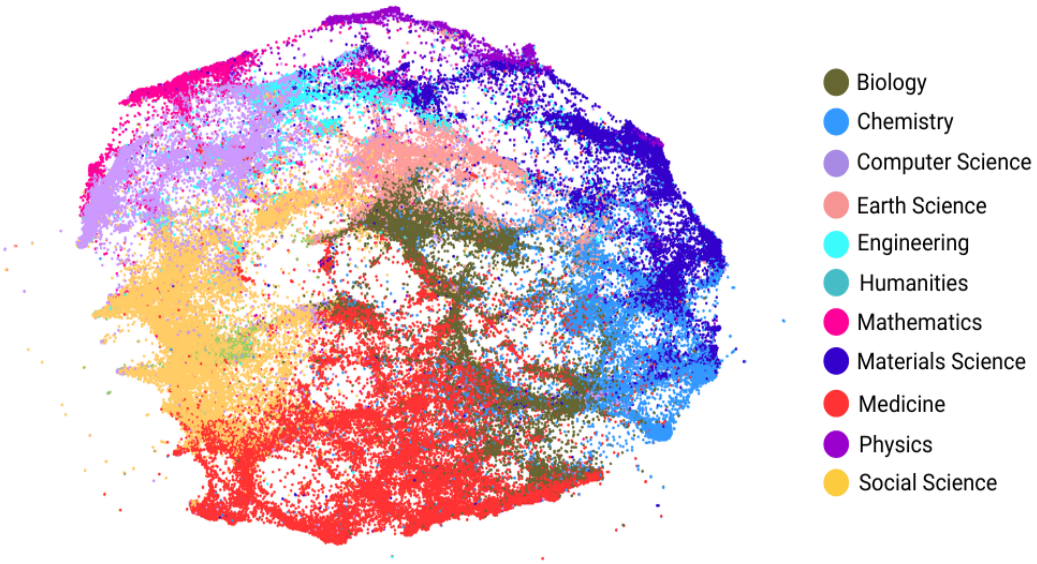
- In a general view of global scientific research, China leads research publication output in STEM fields, whereas the United States leads in medicine and social science, with no major competitors.

Findings

Our analysis begins with CSET's Map of Science, a visual representation of the structure of global scientific publications.[2] This approach starts with grouping nearly 105 million research publications into 126,915 research clusters, hereafter referred to as "RCs." Each RC represents a subset of research publications derived from direct citation links; publications in a given RC cite one another more than publications outside of their cluster. The Map of Science displays citation relationships by placing RCs that have stronger citation links closer together. Just as a physical map shows a geographic landscape, the Map of Science illustrates the current research landscape, allowing us to contextualize the scientific environment to inform decision making in S&T.

The RC-level view of the Map of Science is the most granular aggregation of scientific research publications that lets us answer questions about specific areas of research without analyzing research publications individually. We can color code the RCs within the Map of Science to identify various patterns and trends in research. In Figure 1, each RC (denoted by a point on the map) is colored by its broad area of research, providing a clear visualization of research by a general topic area.

Figure 1. Map of Science by Research

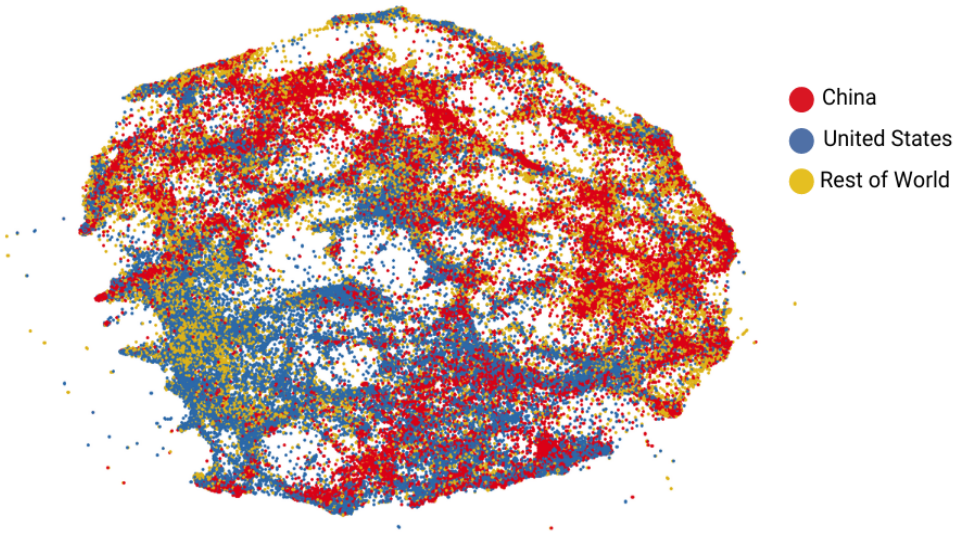


Source: CSET Map of Science.

For this analysis, we instead assign each RC a color based on the dominant country in terms of cluster publication count (shown in Figure 2). Publications are assigned to countries based on the physical address of the author's organization.[3] Therefore, multiple countries can be assigned to one publication if written by multiple authors from different organizations. Each RC location remains the same in Figure 1 and Figure 2.

We highlight the U.S.- and China-dominant RCs, and group all other countries (135 total) into a "rest of world" category. The United States dominates 37 percent of the RCs and China dominates 28 percent, leaving the rest of the world dominating in 34 percent of RCs.[4] This dominant country map highlights the areas of research in which the United States leads (e.g., social science) versus the areas of research in which China leads (e.g., materials science). However, the high amount of overlap hampers a general understanding of the areas of research in which the United States and China lead.

Figure 2. Map of Science by Dominant Country



Source: CSET Map of Science.

Research Regions

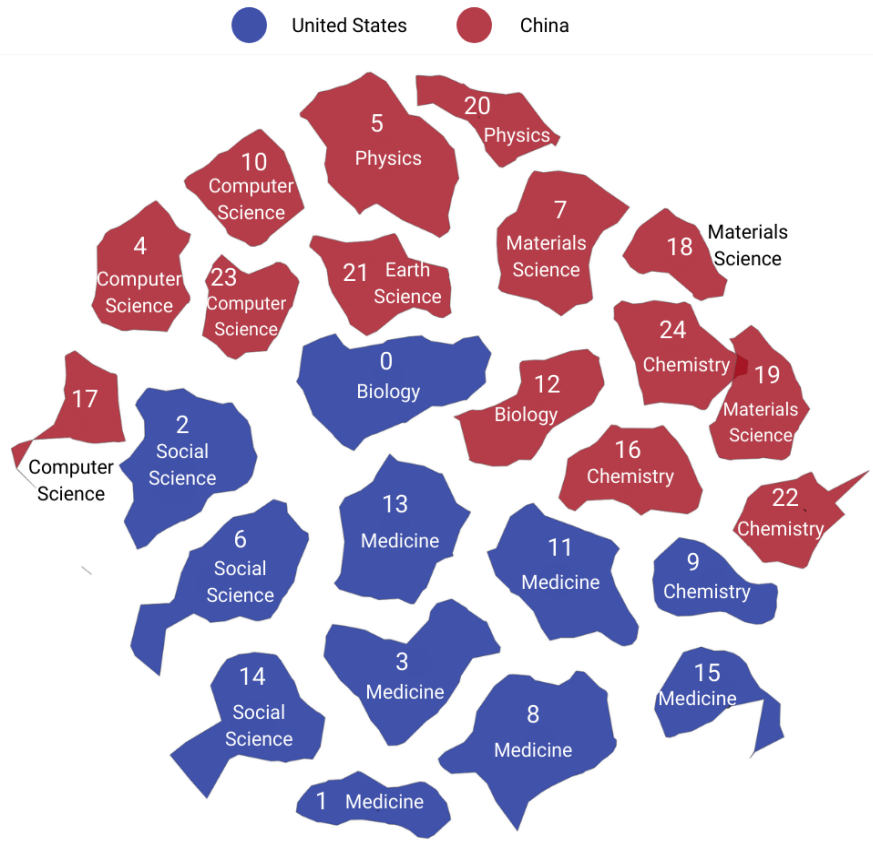
In order to understand the United States versus China competition in research publication output at a general level, we look at the most aggregated view of the Map of Science: research regions. We aggregate our 126,915 RCs into 25 research regions, generated by spatial clustering[5] since RC proximity represents relatedness in citation links. Similar to the RC clustering, regions contain RCs that cite one another more than RCs outside of their region. Spatially clustering RCs into regions maintains the original Map of Science layout and generates a cohesive aggregation of RCs. This high level of aggregation into research regions is explicitly related to a large grouping of similarly focused areas of research, and not related to geographical regions.

Each region is assigned a numeric ID (0-24),[6] as well as the most common broad research area from all its member RCs to provide a general understanding of the region's research area. Additional RC statistics and features (e.g., average yearly growth and Microsoft Academic Graph fields of study)[7] are aggregated on the region-level view, enabling general analysis. The aggregation from individual RCs to research regions provides the most general

overview of global competition in research publication output using the Map of Science.

Figure 3 displays the Map of Science aggregated into research regions, where each region is the color of its corresponding dominant country by publication counts. The only two countries that dominate at the research region level are the United States and China, with the United States being dominant in 11 of the regions, and China in 14. This view gives a clearer picture of which broad areas of research the United States and China dominate. In terms of publication output, China leads in STEM areas of research while the United States leads in medicine and social science.

Figure 3. Map of Science: Research Region View



Source: CSET Map of Science.

Switching from the RC-level Map of Science to the research region-level map, we see a different pattern in the country leaders of research publication output—only the United States and China appear. The difference between Figure 2 and Figure 3 highlights how specificity can change the analytical output. In Figure 1 we see that countries other than the United States and China lead in focused areas of research, but in aggregation their contributions are lost.

Tables 1 and 2 display aggregated details on each region, including the broad area of research, the top three fields of study,[8] the country concentration,[9] and the average three-year growth. From these details we can obtain a general comparison of U.S. and Chinese scientific research output.

Table 1. U.S. Dominant RC Regions

Region ID	Research Area	Top Three MAG Level 1 Fields of Study	% U.S. Papers	Avg. 3-year Growth
0	Biology	ecology, zoology, botany	17%	11%
1	Medicine	surgery, cardiology, anesthesia	24%	12%
2	Social Science	marketing, knowledge management, public relations	17%	7%
3	Medicine	internal medicine, physical therapy, endocrinology	24%	23%
6	Social Science	law, pedagogy, gender studies	25%	3%
8	Medicine	surgery, gastroenterology, immunology	22%	9%
9	Chemistry	organic chemistry, microbiology, biochemistry	20%	15%
11	Medicine	endocrinology, cell biology , immunology	22%	14%
13	Medicine	neuroscience, endocrinology, ophthalmology	24%	14%
14	Social Science	clinical psychology, nursing, psychiatry	32%	15%
15	Medicine	cancer research, oncology, pathology	23%	15%

Source: CSET Map of Science.

Table 2. China Dominant RC Regions

Region ID	Research Area	Top Three MAG Level 1 Fields of Study	% Chinese Papers	Avg. 3-year Growth
4	Computer Science	pure mathematics, computer network, algorithm	20%	9%
5	Physics	structural engineering, mechanics, astrophysics	21%	11%
7	Materials Science	metallurgy, composite material, geochemistry	28%	12%
10	Computer Science	control theory, mathematical analysis, control engineering	30%	11%
12	Biology	agronomy, botany, virology	19%	12%
16	Chemistry	food science, chromatography, biochemistry	23%	13%
17	Computer Science	computer security, distributed computing, combinatorics	17%	9%
18	Materials Science	condensed matter physics, optoelectronics, analytical chemistry	21%	12%
19	Materials Science	chemical engineering, composite material, inorganic chemistry	30%	18%
20	Physics	optics, optoelectronics, astrophysics	19%	8%
21	Earth Science	hydrology, climatology, remote sensing	20%	9%
22	Chemistry	crystallography, photochemistry, chromatography	24%	14%
23	Computer Science	computer vision, pattern recognition, algorithm	25%	16%
24	Chemistry	chemical engineering, composite material, waste management	26%	15%

Source: CSET Map of Science.

U.S. publication concentrations range from 17 percent to 32 percent, with the two research regions (14 and 6) with highest U.S. publication concentrations (32 percent and 25 percent) falling under the social science broad research area. Concentrations in China range from 19 percent to 30 percent, with the two research regions (19 and 23) that have highest Chinese publication concentrations (30 percent and 25 percent) falling under materials science and computer science, respectively. While the United States and China appear to dominate research regions, we find that their respective concentrations never exceed the majority (greater than 50 percent).

The United States is dominant in the highest-growing research region (3), which falls under medicine and includes RCs that contain COVID-19 related research. China is dominant in the

second highest growing research region (19), which falls under materials science and includes RCs that focus on chemical engineering, composite material, and nanotechnology.

Research Districts

The second-level view of aggregation for the Map of Science is the research district level. Each of the 25 research regions is broken down into 10 districts, resulting in 250 distinct research districts.[10] Similar to the research region view, each district is assigned a numerical ID,[11] labeled with the most common broad research area of its member RCs, and RC details and features are aggregated for analysis across districts.

Figure 4 displays the research district view, maintaining the region research area labels for comparison. The United States dominates in 122 research districts (48 percent), while China dominates in 127 (51 percent). With the district view, we can now see that several research regions have higher country competition (e.g., region 17) than others and 15 regions have diversity in the country dominating at the district-level. Specifically, we see that research region 5 (physics) has an India-dominated district and is the only district dominated by a country other than the United States or China. The India-dominated district falls under the materials science broad research area and focuses on mechanics and thermodynamics.

Figure 4. Map of Science: Research District View

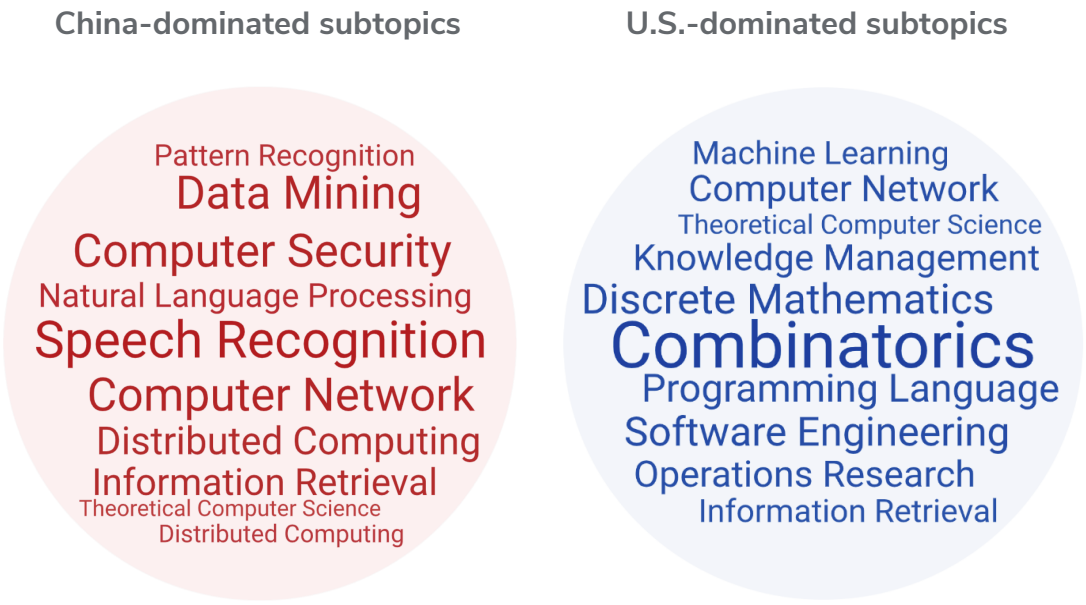
Source: CSET Map of Science.

We use research region 17 (computer science) as an example region to explore more deeply, since it has an even split between the U.S.-dominated districts and the China-dominated districts. Nine districts of research region 17 fall under the broad research area of computer science, with one falling under mathematics (district 1704).

Figure 5 displays the top subtopics of research[12] for the U.S.-dominated districts and the China-dominated districts within research region 17. Comparing the similarities of subtopics across both China and the United States, we find overlaps of *computer network*, *information retrieval*, and *theoretical computer science* in the subtopic research areas. Figure 5 also highlights the differences between the United States and China. We find that China has a strong focus in applied computer science, with topics

like natural language processing, speech recognition, and data mining, whereas the United States has a strong focus in foundational computer science, with topics such as combinatorics, discrete mathematics, and software engineering.

Figure 5. Leading Subtopics of Research for China-Dominated Districts and U.S.-Dominated Districts in Research Region 17 (computer science)



Source: CSET Map of Science.

Table 3 provides details on each district in research region 17, including the dominant country, the top field of study, the country concentration, and the average three-year growth. U.S. publication concentrations range from 14 percent to 22 percent, with research districts 1707 and 1704 having the highest U.S. publication concentrations (22 percent and 18 percent). China publication concentrations range from 15 percent to 28 percent, with research districts 1706 and 1703 having the highest Chinese publication concentrations (28 percent and 21 percent). China-dominated research districts lead in average three-year growth, with 1703 and 1706 as the top two highest growing districts.

Table 3. Research Region 17 District-Level Details

U.S. Dominated Districts

ID	Top MAG Field of Study	% U.S. Papers	Avg. 3-year Growth
1700	Software Engineering	14%	6%
1704	Combinatorics	18%	13%
1707	Knowledge Management	22%	5%
1708	Programming Language	15%	0.7%
1709	Operations Research	18%	10%

China Dominated Districts

ID	Top MAG Field of Study	% Chinese Papers	Avg. 3-year Growth
1701	Speech Recognition	15%	7%
1702	Computer Network	20%	7%
1703	Natural Lang. Processing	21%	15%
1705	Computer Security	19%	10%
1706	Data Mining	28%	16%

Source: CSET Map of Science.

Conclusion

Analyzing the scientific research publication landscape through different levels of aggregation provides context for a more granular analysis of the global competition in S&T. When visualizing this competition with a conceptual map, one needs to be able to understand the overall research terrain of the country before diving down to the specific roads through each district and region. This analysis presents the differences in global competition outcomes when viewing research publication data on varying levels of granularity.

When we compare country leaders from the research regions (the most general level), we default to a bilateral view where China leads and the United States follows closely behind. However, this view neglects the research output from the rest of the world, which we can clearly see in the research clusters (the most granular level). Specifically, at the most aggregate level (25 research regions) China leads 14 regions and the United States leads 11 regions. At a more granular level (250 research districts) China leads 127 districts, the United States leads 122 districts, and India leads one district. Finally at the most granular level (126,915 research clusters) the United States leads 46,837 clusters, countries other

than the United States and China lead 42,526 clusters, China leads 36,125 clusters.

By aggregating research publications, we find China dominating in the hard sciences and the United States in medicine and soft sciences, but when we view the same data in deeper levels of granularity, we see a more detailed picture of leadership in research output. By looking at clusters of publications that represent focused areas of research, we can highlight research areas where alliances may be beneficial. This encourages further analysis of the scientific literature to uncover specificities in focused areas of research.

Future CSET analyses will study different comparisons between U.S. and Chinese research publications such as:

- Exploring specific RC regions or RC districts to present a more detailed view of their compositions.
- Identifying the country leader by influential research publication counts specifically, as opposed to all publication counts in a given area of research, in order to compare research publication output qualitatively.
- Analyzing the subset of RCs that have experienced or are forecasted to experience extreme growth over the next three years as a metric of which country leads in the most “active” areas of research.

Equipped with increasingly detailed analyses, policy professionals can utilize data-driven reports for more targeted decision-making.

Authors

Autumn Toney is a data research analyst at CSET, where Melissa Flagg is a senior fellow.

Acknowledgments

For feedback and assistance, we would like to thank Catherine Aiken, Kevin Boyack, Shelton Fitch, Richard Klavans, Igor Mikolic-Torreira, and Lynne Weil.

Document Identifier: doi: 10.51593/20210020

Endnotes

[1] CSET's Map of Science clustering model is explained in detail in Ilya Rahkovsky et al., "AI Research Funding Portfolios and Extreme Growth," *Frontiers in Research Metrics and Analytics* 6 (2021): 11. CSET is currently developing an interactive tool for users to explore the Map of Science, as well as short explanations of how to use the Map of Science to answer specific questions.

[2] More details on this approach to mapping scientific literature: Richard Klavans and Kevin W. Boyack, "Research portfolio analysis and topic prominence," *Journal of Informetrics* 11, no. 4 (2017): 1158-1174, doi: 10.1016/j.joi.2017.10.002.

[3] 1,427 research clusters do not have a dominant country assigned; this excludes 1 percent of all clusters.

[4] Due to rounding percentages, the three values add to 99 percent.

[5] We use agglomerative hierarchical clustering, helpful explanation here: Tim Bock, "What is Hierarchical Clustering?," *Displayr Blog*, https://www.displayr.com/what-is-hierarchical-clustering/. The shape of the regions form naturally from the clustering process.

[6] IDs are randomly assigned and not based on any kind of ranking.

[7] Details on Microsoft Academic Graph fields of study: https://academic.microsoft.com/topics/topicGraphExplorer.

[8] MAG provides fields of study labels on research publications from a general category (Level 0) to the most granular category (Level 5). There are 19 Level 0 MAG fields of study and 292 Level 1 fields of study. We choose level 1 in order to provide details beyond the most general category. Full lists can be viewed at: https://academic.microsoft.com/topics/topicGraphExplorer.

[9] For a given country, country concentration is the percentage of RCs where that country leads in publication count out of the total number of RCs in a given RC region.

[10] Similar to RC regions, RC districts are generated via agglomerative hierarchical clustering and the shape of the districts form naturally.

[11] IDs are prefixed with their region ID and their suffix is a digit between zero and nine.

[12] Generated by MAG level 1 field of study.