Supporting information for:

**The impact of air transport availability on research collaboration: A case study of four university campuses**   
Adam Ploszaj1,Xiaoran Yan2, Katy Börner2,3

1 Centre for European Regional and Local Studies EUROREG, University of Warsaw, Warsaw, Poland

2 Indiana Network Science Institute, Indiana University, Bloomington, Indiana, United States of America

3 School of Informatics, Computing, and Engineering, Indiana University, Bloomington, Indiana, United States of America

Contents

[1. Data Acquisition and Preparation 1](#_Toc529037041)

[1.1. WoS Publication Data 1](#_Toc529037042)

[1.2. Flight Traffic Data 5](#_Toc529037043)

[2. Expertise Profiles: Science Maps 8](#_Toc529037044)

[3. Code instructions for result replication 10](#_Toc529037045)

# Data Acquisition and Preparation

Two different types of data are used in this analysis: WoS publication data and flight traffic data. The acquisition, preprocessing, and matching of these datasets is detailed subsequently.

## WoS Publication Data

The IU Network Science Institute (IUNI) Clarivate Analytics (previously Thomson Reuters)’ Web of Science dataset (1900-2013)[[1]](#footnote-1) is used in this study. The data comprises publications from

* Science Citation Index Expanded
* Social Sciences Citation Index
* Arts &Humanities Citation Index
* Book Citation Index -- Science
* Book Citation Index -- Social Sciences & Humanities
* Conference Proceedings Citation Index -- Science & Technical

We define a relevant author as someone who has an affiliated address that contains the case insensitive search terms “Indiana Univ”, “Univ Michigan” or “Arizona State Univ”. There are 121,956 papers published in WoS that have at least one author with at least one affiliation at one of these three institutions. Only 4,995 papers associated with the three institutions are published before 2007, see Fig. 1. Therefore, we restrict our analysis to the years 2008-2013, leaving 116,961 papers in the dataset.



**Figure 1:** The histogram of paper counts per publication year

From the 111,961 papers published during 2008-2013, we discard 20,741 papers that have no affiliations. Next, we examined the number of papers per “document type” (“doctype” column), see histogram in Figure 2.

The “article” document type is the most widely used, we restrict the data to this subset of 74,211 out of 91,220 publications. We further restricted the set of papers to those with complete affiliation and deleted 10,590 papers with incomplete affiliations for all authors, leaving 63,621 total publications for this study.



**Figure 2:** The histogram of document types

For papers with more than 30 authors (355 out of 63,621 total papers), the first and last 15 authors were kept. For the analysis of collaborations, we only keep the first three addresses for authors with more affiliations (2,637 out of 344,665 total paper-author pairs were affected). The decision here is to maintain a manageable size of collaborations that is also meaningful. The number of authors per paper and number of affiliations per author is shown in Figure 3.

**Figure 3:** Histogram of author counts per paper (left) and affiliation counts per author (right)

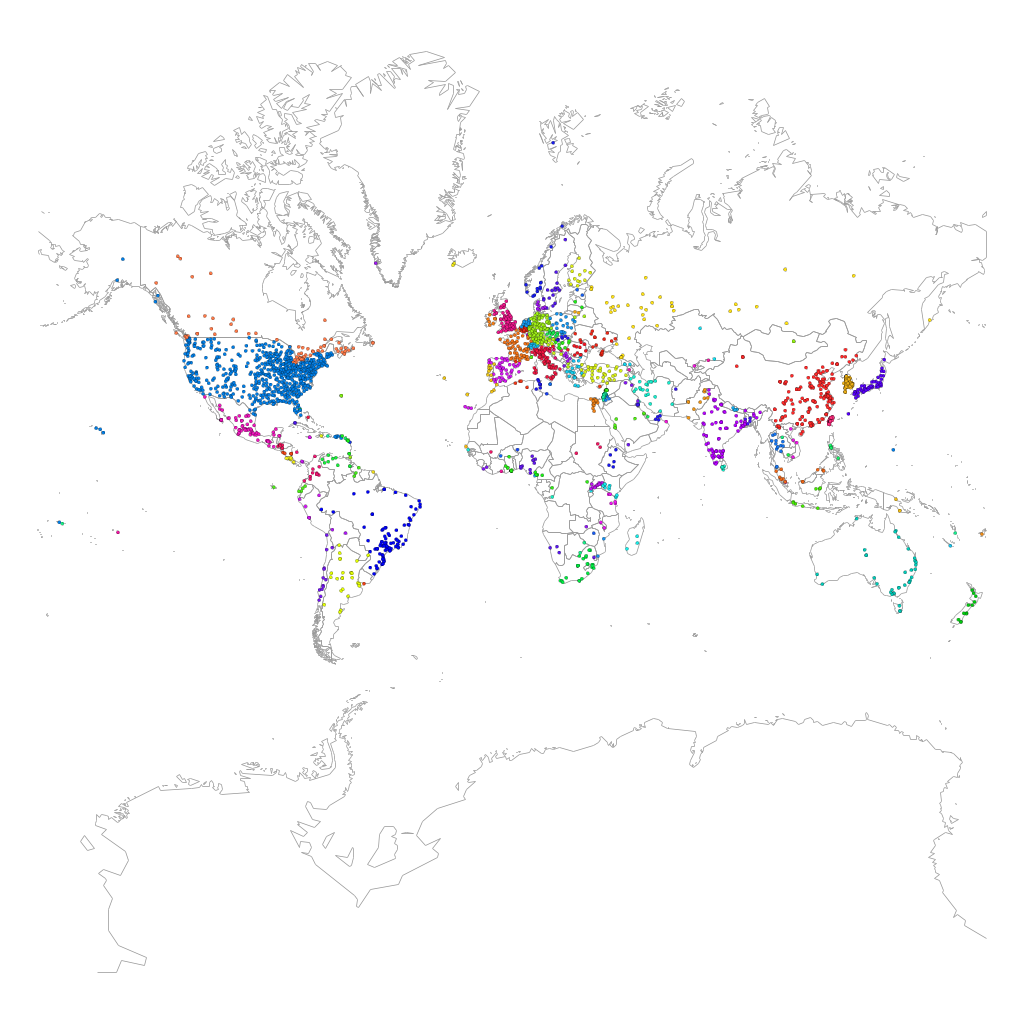
We choose to map addresses to city-level geolocations. We extracted city, state, and country combinations from the “city+state+country” tags provided by WoS. The Bing Geocoder in the Science of Science (Sci2) Tool was used to geocode the cities, following the workflow detailed in Section 4.7.3 of the Sci2 Wiki.[[2]](#footnote-2) With additional manual corrections for all geo-tags with more than 20 appearances as affiliations, all 6,874 unique “city+state+country” tags were successfully geo-located. The table is stored as “[cityListJoin.csv](https://github.com/everyxs/FlightCoauthor/blob/master/data/WoS/cityListJoin.csv)”, with following columns:

* Row id, full address, city, state, country, latitude, longitude, metro join id, standard name

To disambiguate city names, we aggregated cities into metro areas according to statistical areas in the US and EU (represented by the “metro join id” column). We manually added metro area mappings for major cities in other countries including Canada, China, India, Japan, etc. The resulting merge table is called “[cityListMerged.csv](https://github.com/everyxs/FlightCoauthor/blob/master/data/WoS/cityListMerged.csv)”, with following columns:

* Full address, latitude, longitude, standard name, country, number of affiliations, merged row id list

Cities from the previous table are merged into bigger metro/statistical areas. The mapping is given in the rowIDlist column. A total of 3,699 unique “city+state+country” strings are merged into other metro arears, resulting in 3,175 basic level geo-locations, as illustrated in Figure 4.



**Figure 4:** Merged cities and metro areas on the world map under this study.

Using the aforementioned geo-mapping, we can constructed co-authorship relationships between cities/metros. In this study, we focus on four university campuses and their ego-centric co-locations, namely, Arizona State University at Tempe (ASU), Indiana University Bloomington (IUB), Indiana University-Purdue University Indianapolis (IUPUI) and University of Michigan at Ann Arbor (UMICH).

To account for each city/metro’s background overall academic activity, we introduce an external data set of academic productivity (Mazloumian, Helbing, Lozano, Light & Börner, 2013), stored in the table “[locations.csv](https://github.com/everyxs/FlightCoauthor/blob/master/data/WoS/locations.csv)”, with following columns:

* Location id, location name, longitude, latitude, sum of papers, merged cities

Here the column “Sum of Papers” is used to capture the overall academic activity. We mapped the locations with at least five papers to the aforementioned city/metro level geolocations, merging locations that are within 10 kilometers.

To allocate the 63,621 publications to the four universities under investigation, we consider the following four cities/metros as the egos of the network analysis

* UMICH: Ann Arbor MI , United States
* ASU: Tempe AZ United States
* IUB: Bloomington, United States
* IUPUI: Indianapolis IN, United States

Each paper in our dataset that has at least one affiliation with any of the four city/metro will be included for the construction of the corresponding university-ego network. Notice that if there are other academic institutions in the aforementioned cities/metros they were merged with the ego.

## Flight Traffic Data

Global flight traffic data was downloaded from the web site OpenFlights[[3]](#footnote-3). We used two data files, with the first file being “[airports.csv](https://github.com/everyxs/FlightCoauthor/blob/master/data/WoS/airports.csv)”, with following columns:

* Airport id, name, city, country, IATA/FFA code, ICAO code, latitude, longitude, altitude, timezone

The second file is “[routes.csv](https://github.com/everyxs/FlightCoauthor/blob/master/data/WoS/routes.csv)”, with following columns:

* IATA code of the operating Airline, airline ID, IATA code of the source airport, airport ID of the source (mapped to “airports.csv”), IATA code of the destination airport, airport ID of the destination (mapped to “airports.csv”), number of stops during the flight, IATA equipment code (plane model)

To get a more accurate measure of flight traffic, we translated the IATA equipment code to number of passenger seats in corresponding plane models, , captured by the table “[planeWake.csv](https://github.com/everyxs/FlightCoauthor/blob/master/data/WoS/planeWake.csv)”, with following columns:

* IATA equipment code, manufacturer, type/model, wake, seats

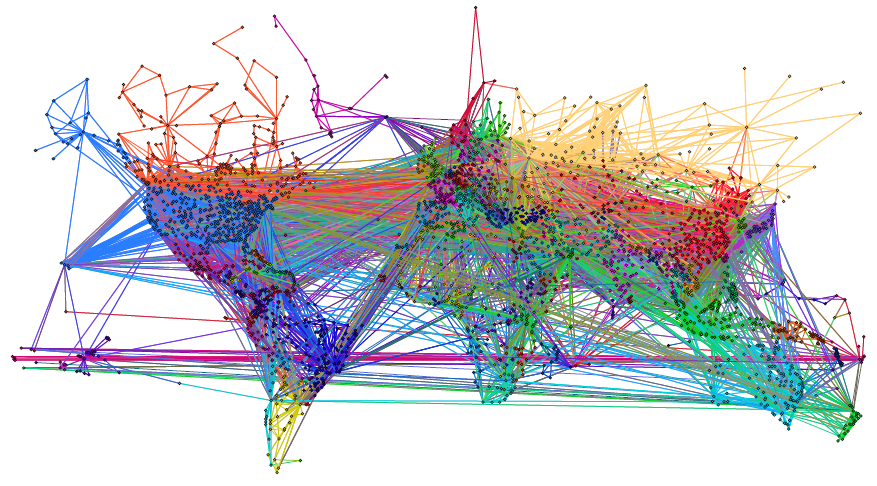
The table is based on the IATA wake table from internet[[4]](#footnote-4), and we manually input the number of typical passenger seats for popular plane models from Airbus and Boeing. For the other models, we used a general estimation that maps Wake “L” to 40 seats, “M” to 120 seats and “H” to 360 seats. When a flight listed multiple plane models, we took the arithmetic average.

We constructed a weighted flight network for flight routes by joining the first two tables, and a second weighted flight traffic network by also including the plane ‘Wake’ column. After filtering out disconnected airports, the resulting two networks have 3,253 nodes and 37,133 directed weighted edges, see Figure 1. Note that rather small airports, e.g., Monroe County Airport near Bloomington, are not covered in the dataset. Another observation is that almost all routes are one-way flights with 0 stops, and the flight traffic between two airports is usually symmetrical.

Following the city to metro mapping we did for the universities, we merge airports in the same city and metro area into a single node. The table “[airports\_AP\_joined\_new.csv](https://github.com/everyxs/FlightCoauthor/blob/master/data/WoS/airports_AP_joined_new.csv)” extends the original “[airports.csv](https://github.com/everyxs/FlightCoauthor/blob/master/data/WoS/airports.csv)” with metro mappings and is used as the input file. The resulting mapping of merged airports is stored in the table “[airportMergeMap.csv](https://github.com/everyxs/FlightCoauthor/blob/master/data/WoS/airportMergeMap.csv)”, with following columns：

* airport ID (mapped to “airports.csv”), name, metro join id (mapped to “cityListJoin.csv”), latitude, longitude, airport merge list (mapped to “airports.csv”)

Here, the airport merge is indicated by the “airport merge list” column, where the airport with the highest count of total seats is selected to represent all other airports in the same metro-area by combining all flights. Connecting the nodes with the merged airlines, we have a directed weighted network with 2,854 nodes and 33,113 edges. Another observation of the dataset is that almost all traffic flows are symmetric between airports. For computational simplicity, we thus treat the air traffic network as undirected. After filtering out disconnected airports, both the flight count and the seat number weighted networks have 2,854 nodes and 16,877 weighted edges, see Figure 5 and the [graphML](https://github.com/everyxs/FlightCoauthor/blob/master/data/WoS/airlinesMerged.graphml) file.



**Figure 5:** Air traffic data network (A) with merged cities and metro areas

In particular, the four egos are associated with the following three airports:

* UMICH: DTW (20.8 mi from Ann Arbor), 199 direct flights with 31384.2 seats
* ASU: PHX (4.3 mi from Tempe), 255 direct flights with 37238.5 seats
* IUB & IUPUI: IND (40.1 mi from IUB, 8.1mi from IUPUI), 62 direct flights with 7907.7 seats

To capture the total possible flight traffic flow from one city to another and up to four connecting flights, we aggregate all possible paths up to 3 stops. This is done by a weighted sum of powers of the direct flight adjacency matrix W:

Notice here the flow represents the number of flights/seats from node *i* to node *j* by taking connecting flight(s), and each powered term of is normalized by taking an element-wise inverse power. Using this weighted power series, we calculated six total flight flow networks with 0, 1, 2 and 3 stops for both flight networks, four for flight route counts, and another four with the additional number of seats per flight information.

Next, we map each metro-level geolocations in a given ego-centric collaboration network with one of the 2,854 airport locations based on geo-distance, associating each affiliation with the closest airport. See Figure 6 for a visualization of the alignment.



**Figure 6.** Aligned collaboration ego-network (bottom) and the global airline network (top).

By mapping affiliations and airports to the same city/metro level geo-locations, we constructed the following data table “[SelectCities[Edges].csv](https://github.com/everyxs/FlightCoauthor/blob/master/data/WoS/SelectCities%5BEdges%5D-full.csv)” for our data modeling and analysis, with following columns:

* Source id, target id, edge id, dist2air, mass, seatsXstop, lineXstop, GeoDist, CollabPaper

Here, each row represents a pair of city/metro level geo-locations that have affiliated authors that collaborated on papers of the extracted WoS dataset. The “SeatsXStop” and “LinesXStop” columns are the eight aforementioned total flight flow networks. The “CollabPaper” column counts the number of articles that such collaboration has happened in our dataset.

Aggregating all four ego networks, we have a total of 10,060 alters. By examining the combinations of Dist2Air+MASS+geodist 1,080 duplicated ego-alter pairs were identified. Excluding these duplicates leads to a final datasets with 8,980 observations. For details of the variables and analysis, please refer to the main text of the paper.

The other columns are calculated according to the following “[SelectCities[Nodes].csv](https://github.com/everyxs/FlightCoauthor/blob/master/data/WoS/SelectCities%5BNodes%5D.csv)” table, with following columns:

* Node id, name, latitude, longitude, country, Dist2Air, mass

where node id is mapped with the “Source-Target” pairs in the edge table. The “Mass” column in the edge table is simply the product of the overall academic activities of the two end nodes and the “GeoDist” column is calculated according to the geo coordinates of the nodes.

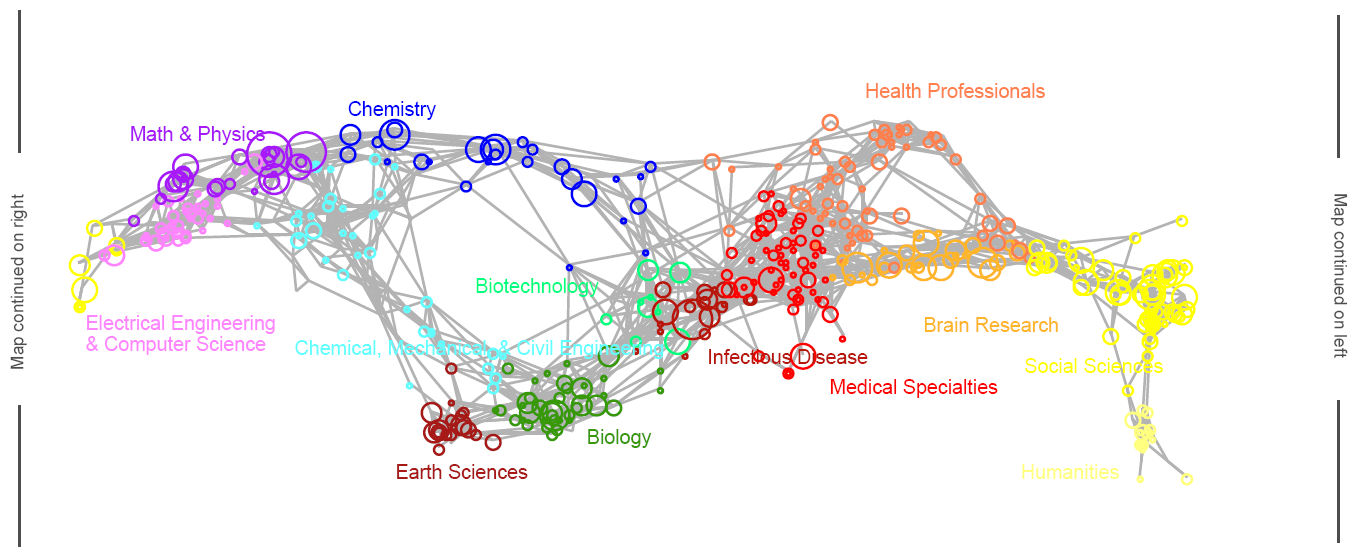
# Expertise Profiles: Science Maps

Global maps of science are generated through scientific analysis of large-scale scholarly data sets (e.g., publication, funding, and patent data) in an effort to connect and make sense of the bits and pieces of knowledge they contain, see the *Atlas of Science* (Börner, 2010) and the Places & Spaces: Mapping Science exhibit at <http://scimaps.org>. The UCSD Map of Science is the most comprehensive map in existence today (Börner et al, 2012). It organizes and visually represents 554 subdisciplines of science and their relationships to one another. Subdisciplines are grouped into 13 overarching disciplines that are color coded (red for medicine, green for biology, etc.) and labelled. Compared to the map of the world, disciplines resemble continents while subdisciplines can be seen as countries. Science maps can be used to gain overviews of “all-of-science” or of a specific subdiscipline. The default zoom level on each map will display the high-level aggregation of the data showing the 13 disciplines. Zooming in will break the disciplines into the 554 subdisciplines.

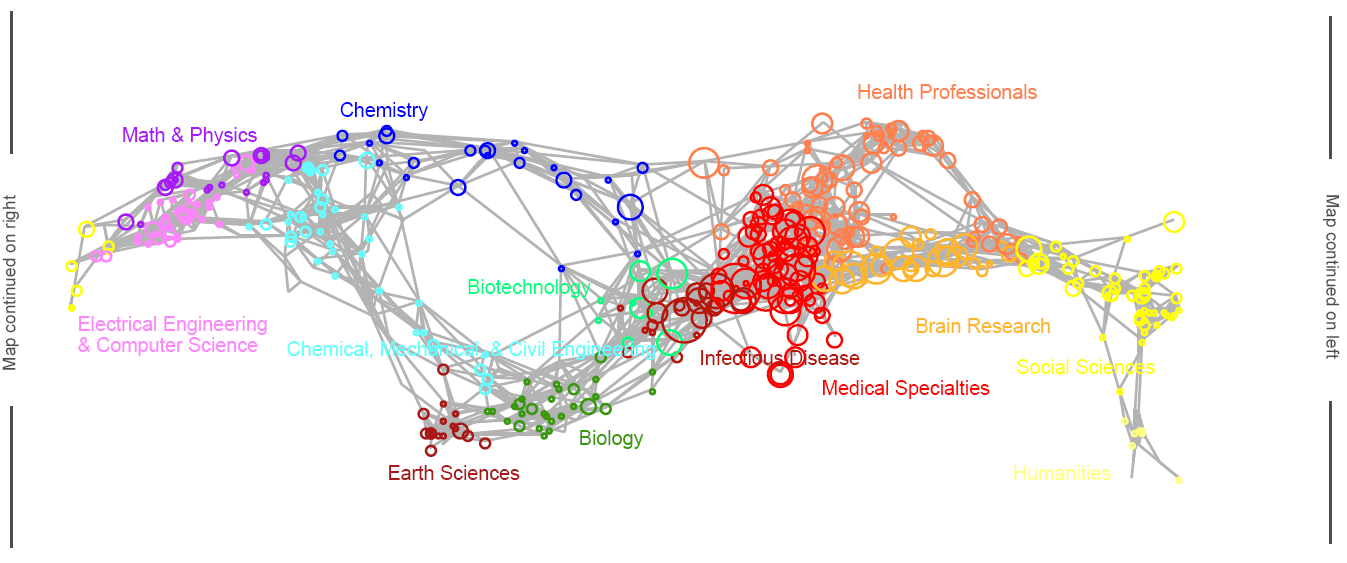
Using a journal name based or keyword based mapping process, data overlays can be computed. For example, expertise profiles for an institution are generated by reading a file with relevant publications, identifying unique journal names, and overlaying geometric symbols such as circles atop the subdiscipline(s) that are associated with each journal. Multiple journals might be associated with a subdiscipline requiring aggregation. Circle symbols are area size coded relative to the number matching journals and colored according to the respective discipline.

Figure 7 shows all three maps of science: IUB (A), IUPUI (B), ASU (C), UMICH (D). For IUB, there are 8,665 papers published in 2,688 unique journals and 2,407 out of the 2,688 journals were mapped to 457 subdisciplines; for IUPUI there are 10,265 papers published in 2,610 journals; 2,391 out of these 2,610 journals were mapped to 438 subdisciplines; for ASU there are 13,647 papers published in 3,596 journals and 3,178 out of 3,596 journals were mapped to 493 subdisciplines; and UMich has 40,380 publications in 5,744 journals and 4,986 out of 5,744 records were mapped to 522 subdisciplines.

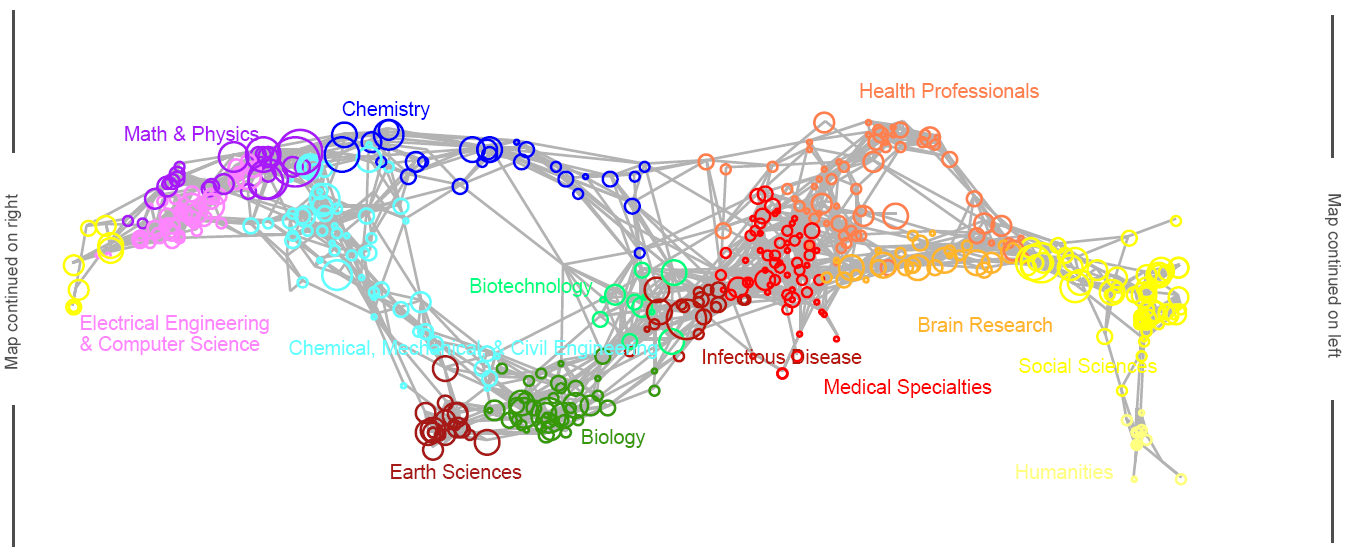
**A (IUB)**

****

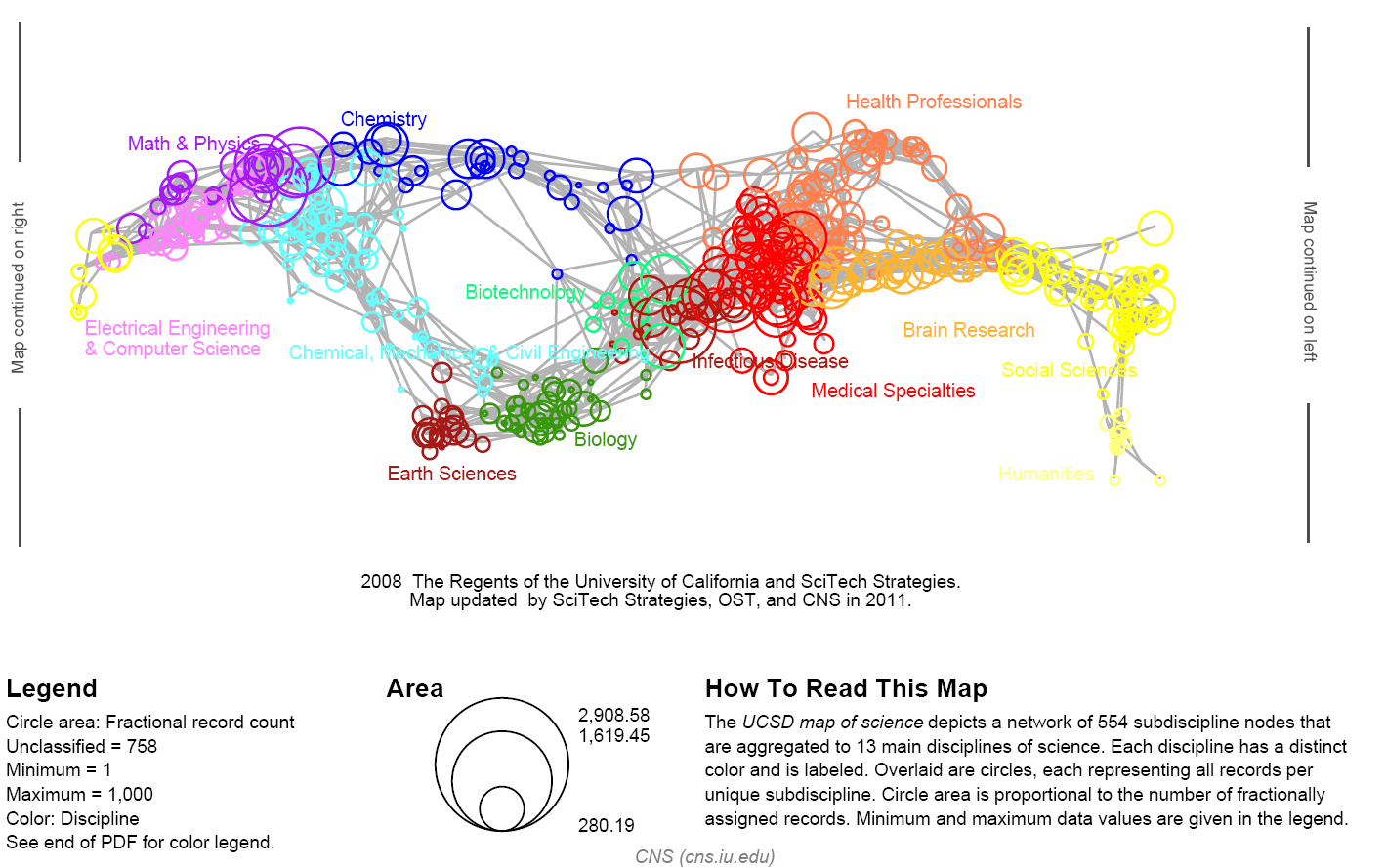
**B (IUPUI)**

****

**C (ASU)**

****

**D (UMICH)**

****

**Figure 7.** Science maps for all four universities showing differences in expertise profiles and productivity

As can be seen, the three universities have a rather different number of papers, published in smaller or larger numbers of unique journals. IUB has the smallest number of papers but comparably a rather large number of subdisciplines, see Table 1. As can be seen in Figure 7, IUPUI and UMich have a strong *Medial Specialties* focus—both have medical campus. All but IUPUI are strong in *Physics* and *Chemistry*.

**Table 1.** Number of papers, journals, and disciplines covered by the four universities

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Disciplines** | **IUB** | **IUPUI** | **ASU** | **UMICH** | **IUB** | **IUPUI** | **ASU** | **UMICH** |
| Biology | 661 | 76 | 995 | 1,218 | 7.6% | 0.7% | 7.3% | 3.2% |
| Biotechnology | 244 | 292 | 336 | 931 | 2.8% | 2.8% | 2.5% | 2.4% |
| Brain Research | 743 | 1,368 | 709 | 3,125 | 8.6% | 13.3% | 5.2% | 8.2% |
| Chemical, Mechanical, & Civil Engineering | 104 | 137 | 1,051 | 2,370 | 1.2% | 1.3% | 7.7% | 0.6% |
| Chemistry | 563 | 251 | 879 | 1,926 | 6.5% | 2.4% | 6.4% | 5.0% |
| Earth Sciences | 392 | 77 | 721 | 847 | 4.5% | 0.8% | 5.3% | 2.2% |
| Electrical Engineering & Computer Science | 279 | 153 | 975 | 1,522 | 3.2% | 1.5% | 7.1% | 4.0% |
| Health Professionals | 483 | 1,754 | 919 | 5,118 | 5.6% | 17.1% | 6.7% | 13.4% |
| Humanities | 105 | 11 | 50 | 147 | 1.2% | 0.1% | 0.4% | 0.4% |
| Infectious Diseases | 493 | 924 | 564 | 2,708 | 5.7% | 9.0% | 4.1% | 7.1% |
| Math & Physics | 1,322 | 206 | 1,712 | 4,353 | 15.3% | 2.0% | 12.5% | 11.4% |
| Medical Specialties | 517 | 3,681 | 468 | 8,314 | 6.0% | 35.9% | 3.4% | 21.7% |
| Social Sciences | 1,822 | 550 | 2,731 | 3,867 | 21.0% | 5.4% | 20.0% | 10.1% |
| Multiple Categories | 448 | 349 | 751 | 1,996 | 5.2% | 3.4% | 5.5% | 5.2% |
| Unclassified | 489 | 436 | 784 | 1,938 | 5.6% | 4.2% | 5.7% | 5.1% |
| **Total #papers** | **8,665** | **10,265** | **13,647** | **40,380** |  |  |  |  |
| **Total #journals** | **2,688** | **2,610** | **3,596** | **5,744** |  |  |  |  |
| **Total #subdisciplines** | **457** | **438** | **493** | **522** |  |  |  |  |

# Code instructions for result replication

The code for **data manipulation and data table generation** in written in JAVA. Source code and required input files are shared at https://github.com/everyxs/FlightCoauthor. The modeling and analysis is done in STATA and code and data is shared at the same GitHub repository under the folder [/STATA](https://github.com/everyxs/FlightCoauthor/tree/master/STATA).

The “FlightCoauthor.zip” contains a JAR package and a data folder set up for reproducing the data tables used in the main article. Java Runtime Environment (JRE) 7.0 or newer must be installed; all other required libraries are included in the JAR package. Once the zip content is extracted, run the JAR package by typing the following command under the extracted folder:

> Java –jar FlightCoaurhtor.jar

The data folder contains/produces the following files:

* Raw input files: “papers.csv”, “cityListJoin.csv”, “locations.csv”, “airports\_metro.csv”, “airports\_AP\_joined\_new.csv”, “routes.csv”, “planeWake.csv”, “legacyNodeTabel.csv”, “legacyEdgeTabel.csv”
* Intermediate data I/O file: “cityListMerged”, “massMatch0.csv”, “airportMergeMap.csv”, “collabEdges.csv”, “routesWeighted.csv”
* Final output data tables: “SelectCities[Nodes].csv”, “SelectCities[Edges].csv”

Since the Web of Science raw data is under Intellectual Property restrictions, we can only release a small sample in “papers.csv” for illustration purposes. As a result, the final edge table “SelectCities[Edges].csv” will have a very small “CollabPaper” column. For the actual “SelectCities[Edges].csv” table used in our analysis, please refer to “SelectCities[Edges]-full.csv”.

To recompile the sources files, Apache Commons lang3 and math3 Libraries are needed. The source code includes three classes, and their associated data files listed here:

* PrimaryCity.java, the main class where all other data is processed, including geo-coding, city merging, collaboration counting, etc.
  + Input files: “cityListJoin.csv”, “legacyNodeTabel.csv”, “locations.csv”, “papers.csv”, “collabEdges.csv”
  + Output files: “cityListMerged”, “massMatch0.csv”, “SelectCities[Edges].csv” (final edge data table)
* FlightNode.java, which handles airport merging and mapping with cities/metros.
  + Input files: “airports\_metro.csv”, “airports\_AP\_joined\_new.csv”, “massMatch0.csv”,
  + Output files: “airportMergeMap.csv”, “SelectCities[Nodes].csv” (final node data table)
* FlightEdge.java, which handles flight flow calculation and mapping with cities/metros.
  + Input files: “legacyEdgeTabel.csv”, “routes.csv”, “planeWake.csv”, “aiportMergeMap.csv”
  + Output files: “routesWeighted.csv”, “collabEdges.csv”

**Regression analyses** presented in the article can be replicated using STATA .do file: “AirSciColl\_replication.do”. STATA version 14 or newer is needed to run the code. The STATA code uses analytical dataset named “AirSciColl\_Analitical\_Dataset.csv” that combines data tables described above. For convenience, the analitical dataset is limited to main variables used in the analysis. The code presented in “AirSciColl\_replication.do” also allows reproduction of tables 1-7 and figure 3 from the main body of the article.

1. <http://iuni.iu.edu/resources/web-of-science> [↑](#footnote-ref-1)
2. <http://wiki.cns.iu.edu/pages/viewpage.action?pageId=1245861> [↑](#footnote-ref-2)
3. 2 <http://openflights.org/data.html>. [↑](#footnote-ref-3)
4. 9 http://www.flugzeuginfo.net/table\_accodes\_iata\_en.php [↑](#footnote-ref-4)