Supplementary Information:
Characterizing production and consumption in Physics

Qian Zhang, Nicola Perra, Bruno Gonçalves, Fabio Ciulla, Alessandro Vespignani

February 18, 2013

Contents

1 Extracting Geographic Information

The database of Physical Review publications used in this paper consists of 463,348 articles, each of which is identified by a unique Digital Object Identifier (DOI). 83% of these articles (450,655) record the publishing year, the author(s) of the article, as well as the corresponding affiliation(s). An article may have more than one affiliation, and the database provides affiliation strings for each article. In total, we have 945,767 affiliation strings, and we aim to extract country and city information from the affiliation strings for each article.

We observe that an affiliation string likely stands for a single affiliation, roughly consisting of several comma separated fields:

`(SUB-INSTITUTE)*, (INSTITUTE), (OTHER INFORMATION)*, (CITY), (OTHER INFORMATION)*, (COUNTRY/STATE)`

where '`SUB-INSTITUTE`' means department, college, institute, laboratory within an institute, the asterisk refers to any repetition of the field (including zero), and '`OTHER INFORMATION`' usually means the province (or region) name, postal codes, or P. O. Box. For instance,

`PHYSICS DEPARTMENT, THE ROCKEFELLER UNIVERSITY, NEW YORK, NEW YORK`

`THE INSTITUTE FOR PHYSICAL SCIENCES, THE UNIVERSITY OF TEXAS AT DALLAS, P. O.BOX 688, RICHARDSON, TEXAS`

`PHYSICS DEPARTMENT, UNIVERSITY OF GUELPH, GUELPH, ONTARIO N1G 2W1, CANADA`

Figure. 1 shows the probability distribution of the number of comma separated fields for all affiliation strings. The mean value of such numbers is 4.33 and the standard deviation is 1.156. 86% of all affiliation strings have between 3 and 5 comma separated fields, while the percentage rises to 97% for those with less than 8 such fields (mean$\pm 3\sigma$). Therefore, we first assume that an affiliation string with no more than 7 comma separated fields represents a single affiliation, and the remaining ones may consist of multiple affiliations.

1.1 Parsing country names

We first extract country and U.S. state names from single affiliation strings. To find country names, we create a dataset of country names except U.S. from ISO 3166 country codes [1], and the name of U.S. states from Wikipedia [2]. For some historical country names in the 20th century (e.g., the Soviet Union, Yugoslavia, East Germany), we manually add them in the dataset. Besides, for some countries, we take into consideration the name variations, like full official names and the name in its official language, and possible abbreviations, e.g., U.S.S.R for the Soviet Union, People's Republic of China for China, Deutschland for Germany, etc.

Based on the above assumptions and observations, for an affiliation string with no more than 7 comma separated fields, we first search the field representing a country name, the process of which is called '*field match*'. For each field in an affiliation string, we eliminate the words with numbers 0-9, which may represent a postal code, and then try to match the field with any of the country name in our country name dataset.

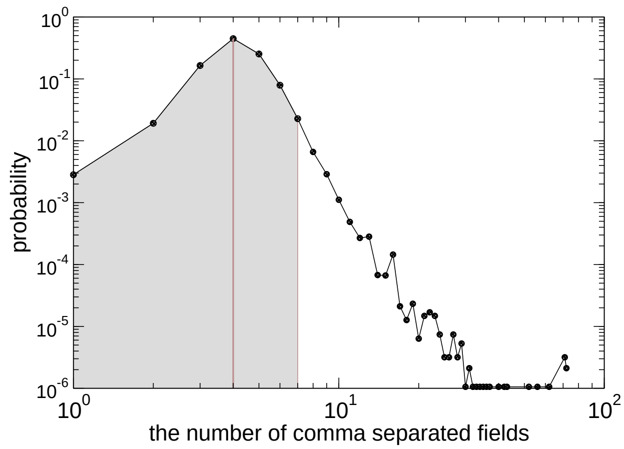


Figure 1: The probability distribution of the number of comma separated fields in an affiliation string. The mean value of such the number is 4.33 and the standard deviation is 1.156. The grey area in the plot represents the band with the width of 3 standard deviations, which implies that the most of affiliation strings consist of no more than 7 comma separated fields.

If there is no field match for an affiliation string, it is possible that either the author did not write a country name specifically but some other fields, like the institution name, include a country name (e.g., `RANDAL MORGAN LABORATORY OF PHYSICS, UNIVERSITY OF PENNSYLVANIA`), or the country name is mixed with other information in a field, like a city name or a non-numeric postal code (e.g., `MAX-PLANCK-INSTITUT FÜR MOLEKULARE PHYSIOLOGIE POSTFACH 500247 D-44202 DORTMUND GERMANY`). Moreover, for the affiliation strings with '*field match*' results, other fields in that string may also contain country names for multiple affiliation cases (e.g., `ARGONNE NATIONAL LABORATORY, ARGONNE, ILLINOIS 60439 AND OHIO STATE UNIVERSITY, COLUMBUS, OHIO`). For the kind of affiliation strings without field match results, we try to match the country name word by word in all fields in that affiliation strings, and for the ones with some field matched, we match the country names word by word in other fields. We call this process '*string match*'. If there is a single match from the above two steps, we assign the matched country name to this affiliation string, and classify it into affiliation strings with unique country name. If there are multiple country names matched, we set these affiliation strings aside for later processing.

The above two procedures of '*field match*' and '*string match*' give unique country name to 95.11% affiliation strings (899, 575 out of 945, 767), but 1.83% (17, 278 out of 945, 767) affiliation strings have no country name detected. The remaining 3% affiliation strings either contain more than one country name or have more than 8 fields which may represent multiple affiliations.

The next step is to focus on '*splitting the multiple affiliations*' into single records. The case of an affiliation string with multiple country names varies. For instance, it may represent one affiliation but include the country names with overlapped words (e.g., Mexico vs. New Mexico for *string match* procedure, like

THE UNIVERSITY OF NEW MEXICO, ALBUQUERQUE NEW MEXICO and Washington vs. Washington, D.C. for *field match* procedure, like THE GEORGE WASHINGTON UNIVERSITY, WASHINGTON, D.C.); or some country names may represent a city, a region or a street, (e.g., ST. JOHN'S UNIVERSITY, JAMAICA, NEW YORK); or the union states for some historical countries (e.g. FACULTY OF CIVIL ENGINEERING, UNIVERSITY OF BELGRADE, BULEVAR REVOLUCIJE 73, 11000 BEOGRAD, SRBIJA, YUGOSLAVIA). We go through this scenario first, and try to filter out affiliation strings of unique affiliation. We assume that two country names cannot appear in the neighbor fields or in the neighbor words. Thus, if we found two country names in neighboring fields, we consider the latter one as the real country name. But if two country names are in the same comma separated field, we determine the country name(s) based on their position. We assign an index to each of the words in that field according to the order of the words. If the number of words between the first indices of two country names is less than the number of the words of the longer country name, the country name with the larger length is the country name. For instance, in the above example THE UNIVERSITY OF NEW MEXICO, ALBUQUERQUE NEW MEXICO, we find two country names in the second field: NEW MEXICO and MEXICO with the word indices 2 and 3 respectively. The number of words between two indices is 1, which is smaller than the length of NEW MEXICO, so we determine NEW MEXICO is the country name for this affiliation.

After performing the multiple name checking described above, we consider the remaining affiliation strings consisting of multiple affiliations. We observe that the affiliation strings in this scenario usually contain elements implying multiplicity, like AND and semicolons. For example:

> THE RICE INSTITUTE, HOUSTON, TEXAS AND THE COLLEGE OF THE PACIFIC, STOCKTON, CALIFORNIA
>
> INSTITUTE FOR ADVANCED STUDY, PRINCETON, NEW JERSEY 08540 AND PHYSICS DEPARTMENT, CALIFORNIA INSTITUTE OF TECHNOLOGY, PASADENA, CALIFORNIA
>
> ISTITUTO DI FISICA DELL'UNIVERSITA, ROMA, ITALY; AND ISTITUTO NAZIONALE DI FISICA NUCLEARE, SEZIONE DI ROMA, ITALY

If there are semicolons in the affiliation strings, we split the affiliation strings by the position of the semicolon. However, if there is no semicolon, while there is an AND, we have to exclude the case like 'DEPARTMENT OF PHYSICS AND ASTRONOMY'. To do so, we observe that if an AND joins two affiliations, the country name usually should appear closely before the AND, so we split the string into two part by an AND if the last word position of the country name before AND is at most one word far from the AND (We allow one word between the country name and AND because of possible non-numeric postal codes.), and the AND does not join any two of the descriptive words of research subjects, which usually appear in the information of institute and sub-institute. We built a list of descriptive words by calculating the frequency of the word appearance in the first field of all affiliation strings. The top 20 frequently appeared descriptive words are listed in Table. 1.

For the affiliation strings with more than 7 fields, e.g.,

> CENTER FOR THEORETICAL PHYSICS, DEPARTMENT OF PHYSICS AND ASTRONOMY, UNIVERSITY OF TEXAS AT AUSTIN, TEXAS 79712; CENTER FOR ADVANCED STUDIES, DEPARTMENT OF PHYSICS AND ASTRONOMY, UNIVERSITY OF NEW MEXICO, ALBUQUERQUE, NEW MEXICO 97131;

Table 1: The top 20 descriptive words of research subjects.

word	frequency	word	frequency
PHYSICS	314266	RESEARCH	55692
SCIENCE	37345	THEORETICAL	32976
ASTRONOMY	32247	ENGINEERING	28179
MATERIALS	27572	PHYSIK	24083
CHEMISTRY	23821	FISICA	23649
FÍSICA	22711	PHYSIQUE	21928
NUCLEAR	21860	TECHNOLOGY	18769
SCIENCES	16999	APPLIED	16184
THEORETISCHE	12994	MATHEMATICS	10978
SOLID	10351	PHYSICAL	9194

`AND MAX-PLANCK-INSTITUT FÜR QUANTENOPTIK, D-8046 GARCHING BEI MUNCHEN, WEST GERMANY`

we first split it by semicolons but not by `AND`. The split substrings will be processed step by step from *field match* to *string match* and possibly *splitting multiple affiliations*, in the same way as an affiliation string with no more than 7 fields is processed.

It is worth to note that even after splitting process, some of the affiliation strings still contain more than one country name, like

`LOS ALAMOS NATIONAL LABORATORY, UNIVERSITY OF CALIFORNIA, LOS ALAMOS, NEW MEXICO`

for which the above steps give both California and New Mexico as its country names, or

`INSTITUTE FOR QUANTUM COMPUTING, UNIVERSITY OF WATERLOO, N2L 3G1, WATERLOO, ON, CANADA, ST. JEROME'S UNIVERSITY, N2L 3G3, WATERLOO, ON, CANADA, AND PERIMETER INSTITUTE FOR THEORETICAL PHYSICS, N2L 2Y5, WATERLOO, ON, CANADA`

of which the first substring after splitting by `AND` (`INSTITUTE FOR QUANTUM COMPUTING, UNIVERSITY OF WATERLOO, N2L 3G1, WATERLOO, ON, CANADA, ST. JEROME'S UNIVERSITY, N2L 3G3, WATERLOO, ON, CANADA`) still contains another affiliation and there is no more semicolon and `AND` to indicate the position to split. Figure. 1 shows that on average affiliation strings representing a single affiliation consist of four fields, therefore we split the affiliation (sub)strings of multiple country names but without any semicolon and `AND` at the position of the country names if the number of fields between two country names is not smaller than 4. Thus the final country names for the affiliation strings of the above two examples are 'New Mexico' and three 'Canada's respectively.

To double check the results obtained from the above procedures, we use Google geocoders from geopy toolbox [3] to get the country names searched by Google map, and call this step *Google geocoders checking*. Unfortunately, Google geocoders usually cannot code the affiliation strings with department information or even institution information. To avoid these exceptions, for the affiliation string with more than three fields, we send the last three fields as an address string to geocoders, and for others we input the whole string to geocoders. Google geocoders return a comma separated address string for each input. If the returned string is not empty, we match the country names, 2-letter or 3-letter abbreviations in our country name dataset with the returned result. Once the matched result represent the same country as we extracted, we say the country name we parsed for this affiliation string is validated. It should be noted that we do not use Google geocoders (or other geocoders like Yahoo or Bing) directly to search country names because to our best knowledge there is no evidence to guarantee the accuracy of the results from these APIs.Thus we perform this step of checking to get better accuracy.

Figure. 2 summarizes the above steps to extract country names from affiliation strings in a flow chart. As the result, the 3% of affiliation strings with multiple country names and more than 7 fields are finally split into 46, 353 new records. In the end, we obtain 963, 206 records of single affiliation, of which 97.68% (940, 896) have a country name validated with Google geocoders. Figure. 3 indicates that after 1940, we parsed validated country names for more than 95% of papers in each year. We use these affiliation strings with validated country names to build citation networks at the country level after 1940, and as the inputs to extract city names.

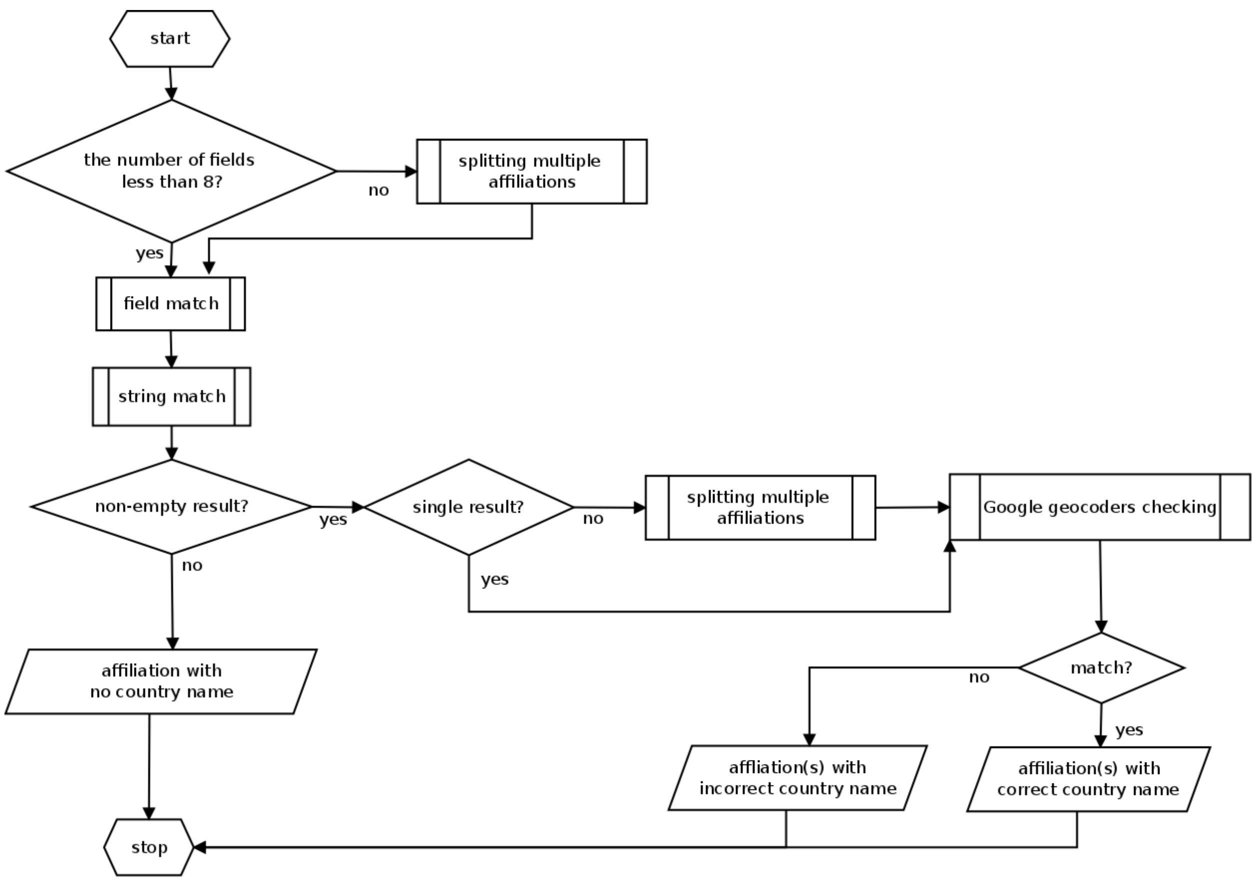


Figure 2: The flow chart of the procedure to extract country name(s) from affiliation strings.

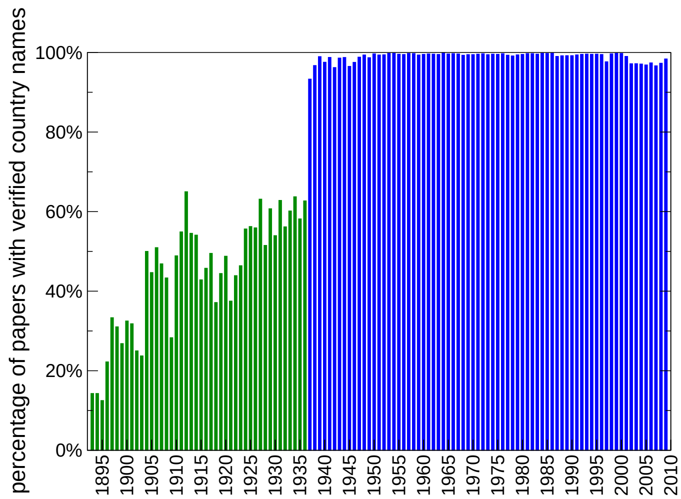


Figure 3: The percentage of papers (DOIs) with validated country names per year. The plot shows that after 1940 we obtain more than 95% of papers with verified country names (blue bars).

1.2 Parsing city names

We use the database of GeoNames to parse the name of cities in the affiliation strings with identified country names. GeoNames database includes geographical data such as names of villages, cities, and other types of places in various languages, elevation, population and others from various sources [4]. The variations of languages for geographic names allow us to identify city names written in languages other than English. Each record of places in the database also includes its country name and possibly the first level of administrative division (e.g., the states in the United States). We first filter records that represent cities (by the feature codes attribute in GeoNames data), and arrange cities by the names of countries and US states. For countries like the Soviet Union and Yugoslavia, we combine the cities of their former union countries; and for East Germany we simply use the cities in Germany.

The final results from the above section is a set of affiliation strings, each of which owns a unique country name, so we argue, that to our best effort, each affiliation string now only represents an institution and has one city name if any. Since each affiliation string now has a validated country name, we only use the city list of that country to avoid the same city name in different countries.

After cleaning the data, the first step to parse city names is '*field match*', as we performed to find country names. For each field, we delete words with numbers and try to match it with city names in filtered city dataset for that country. If there are matched city names, we list both the name and coordinates as outputs, otherwise we perform '*string match*' on the affiliation strings trying to match city names word by word.

As we did to validate country names, we use Google geocoders from geopy toolbox to check the correctness of the city names we extract from affiliation strings. The procedure is similar to that for the country names: the affiliation strings excluding the department level information are given as input to Google geocoders, and the non-empty Google searched results are saved for the next step of validation.The coordinates and city names given by Google geocoders for an affiliation string are based on the name of the institutions, and may be different from the name extracted and the coordinates of the city given in GeoName database. To determine if the extracted city name is correct, we simply calculate the geographic distance between the coordinates given by GeoNames database and the ones given by Google geocoders, and if the distance is less than 50km, we say the extracted result is matched with Google searched result. For the affiliation strings with multiple city names, we choose the one which has the shortest Vincenty's distance from the Google geocoded result.

In total, we have 92.6% ($871,345$ out of $940,896$) affiliation strings with validated city names. Figure. 4a shows the the percentage of papers (DOIs) with validated city names per year, from which one can observe that we obtain validated city names for more than 90% of papers after 1940, and for this reason we use data after that year to perform analysis at the city level in this paper. Figure. 4b displays the percentage of papers with validated city names to the total number of papers for each country after 1940. The abscissa is 60 country names ordered by the total number of papers for each country after 1940. These top 60 countries contribute 95% of the papers published in Physical Review journals after 1940, as shown by the cumulative distribution of the total number of papers for all countries (the red dot curve). From Figure. 4b we claim that for the most of major countries contributing to publications in Physical Review journals we have unbiased results of parsing city names.

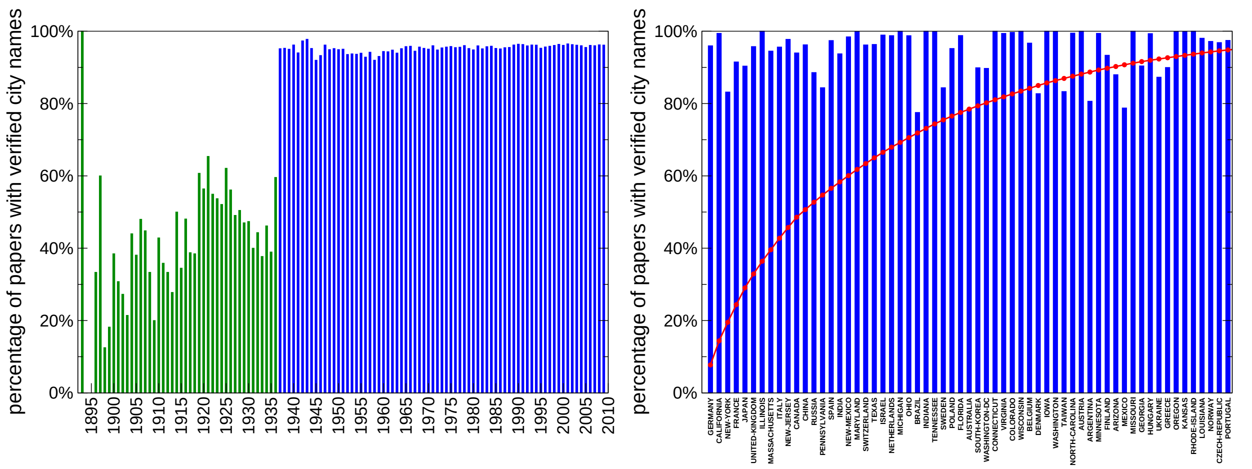
So far we have obtained geographic coordinates and city names for the affiliation strings from Google geocoders and GeoName database. However, different city names may represent the same city, geographically close cities or different administrative levels. For instance,

```
DEPARTMENT OF PHYSICS, BOSTON COLLEGE, BOSTON, MASSACHUSETTS 02467, USA

DEPARTMENT OF PHYSICS, BOSTON COLLEGE, CHESTNUT HILL, MASSACHUSETTS
```

Because Chestnut Hill is not a city in Massachusetts in GeoNames database, the city name extracted from these two affiliation strings for Boston College is Boston, while Google geocoders gives the city name of Newton. In this case, one cannot automatically determine which city this affiliation should be in. One possible way to solve such the problem is to project the coordinates into polygons of 'cities' in shapefiles for geographic information systems software. However, the existent shapefiles have different granularities for different countries. It may be unfair to compare the scientific products in different level of administrative units over different countries.

Therefore, we cluster cities according to their geographic coordinates into 'urban areas' or 'academic cities' in each country. For each country, we perform hierarchical/agglomerative clustering with the geographic distance matrix, of which the distances are calculated with Vincenty's formula. With the dendrogram produced from the clustering process, we cut off the branches from the maximum height value to lower ones until the distance between any point in a cluster and the centroid of the cluster is less than 25km (the maximum distance within the cluster is 50km) for all clusters. We call such clusters 'academic cities'. The final coordinates of an academic city is the centroid of all coordinates inside that cluster, and the academic city is named with the city name which has the most papers in that cluster. We notice that due to



(a) The percentage of papers with validated city names per year. (b) The percentage of papers with validated city names per country.

Figure 4: The percentage of papers (DOIs) with validated city names per year (a), and the percentage of papers (DOIs) with validated city names per country (b). (a) clearly shows that after 1940 we obtain more than 90% of papers with verified city names for each year (blue bars). In (b), the x-axis is top 60 countries ranked by the total number of papers after 1940 in each country. The red dot curve is the cumulative distribution function of the number of papers over countries after 1940. For the major contributing countries in terms of paper production, we have obtained more than 80% of papers with validated city names.

the differences between geographic areas in different countries, some cities are merged into one academic city and some other cities are split into two. For instance, Boston, Cambridge, Newton in Massachusetts are now clustered into one urban area with the name Boston; and Dubna in Moscow Oblast now becomes a separate academic city. Finally, we have a list of academic cities for each paper (DOI), and all the analysis we made at the city level in this paper refer to the unban areas or academic cities.

2 Building the citation networks

A citation network consists of a set of nodes (cities) and directed links representing citations that one paper written in one city is cited by a paper written in another one according to the references of the latter. For example, if a paper is written in node i cites one paper written in node j there is an edge from i to j, i.e., j receives a citation from i and i sends a citation to j. As shown in Figure (1) in the main text, a directed link from Ann Arbor to Rome and another link to Madrid are built since paper A, which is from Ann Arbor, Michigan, cites the paper B from Rome, Italy and Madrid, Spain. Because the paper A was also contributed by authors from another two cities: Los Alamos in New Mexico and New York City in New York, from each of these two cities, there is also a link to Rome and another to Madrid.

The weight of a link is defined as following. In a given time window, the total number of citations for the papers written in j received from papers written in a, is the weight of the link $(i \to j)$, and the total number of citations for those paper written in j sent to the papers written in k is the weight of the link $(j \to k)$. For instance, in time window t, there is one paper written in node j, which cited two papers written in node k and was cited by three papers written in node i, then there are $w_{i,j} = 3, w_{j,k} = 2$, and we add up such weight for all papers written in that node j and obtain the weights for links. For the paper written in multiple cities, say j_1, j_2, the weight will be counted equally, i.e., $w_{i,j_1} = w_{i,j_2}, w_{j_1,k} = w_{j_2,k}$. The time window we use in this paper is 1 year.

3 Basic properties of data and citation networks

We observe a significant growth of the published articles and the citations in recent 50 years, as shown in Figure. 5. Meanwhile, the percentage of papers contributed by authors in the United States has decreased from nearly 90% in early 1960's to current 36% (Figure. 6). Correspondingly, the number of cities contributing to publications in APS journals, as well as their internal interactions, has increased dramatically, as illustrated in Figure. 7 and Figure. 8.

In Table. 2 we report basic statistic properties for the city-to-city citation networks in selected years. Figure. 9a reports the cumulative distribution functions for in- and out-degree of the city-to-city citation networks in different years. The distributions are with behaviors close to power-law with the exponential cutoff. As the year increases, the range of values of k_{in} and k_{out} extends. We define the in/out-strength of node i as the total number of citations it sends/receives at that year. Figure. 9b displays the cumulative distribution function for in- and out-strength of the city-to-city citation networks in different years. The pattern of strength distributions is quite similar to the degree distributions.

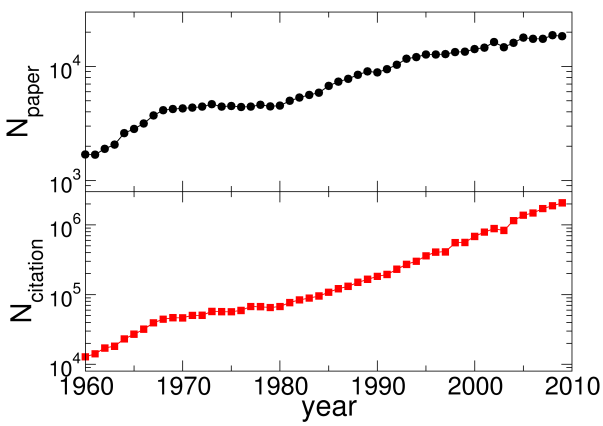


Figure 5: The number of papers (top) and the number of citations (bottom) as the function of time (1960-2009).

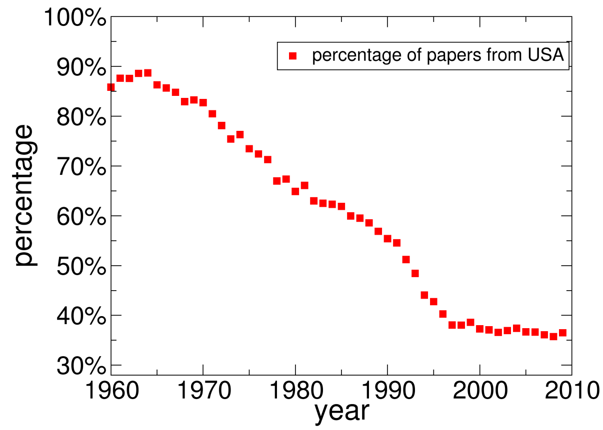


Figure 6: The percentage of papers contributed by authors from USA as the function of time (1960-2009).

Table 2: Summary of basic statistic features for city-to-city citation networks in different years.

year	V	E	k_{in}				k_{out}				S_{in}				S_{out}				w_{ij}			
			mean	std.	min	max	mean	std.	min	max	mean	std.	min	max	mean	std.	min	max	mean	std.	min	max
1960	222	2517	11.34	18.13	0	90	11.34	15.20	0	84	41.24	111.16	0	765	41.24	95.99	0	940	3.64	11.57	1	336
1970	438	9461	21.60	38.97	0	236	21.60	26.72	0	153	87.53	288.39	0	2893	87.53	198.54	0	1758	4.05	13.98	1	564
1980	635	17028	26.82	47.96	0	332	26.82	34.84	0	206	94.08	311.71	0	4182	94.08	213.94	0	2164	3.51	11.02	1	557
1990	897	43324	48.30	80.31	0	539	48.30	58.37	0	329	207.59	671.95	0	9125	207.59	459.34	0	4372	4.30	13.00	1	830
2000	1327	109438	82.47	126.79	0	754	82.47	102.83	0	556	801.76	2640.94	0	34768	801.76	2167.73	0	20862	9.72	29.71	1	1568
2009	1704	204747	120.16	178.22	0	968	120.16	151.16	0	822	3033.86	9230.21	0	104149	3033.86	8651.34	0	76044	25.25	75.12	1	3004

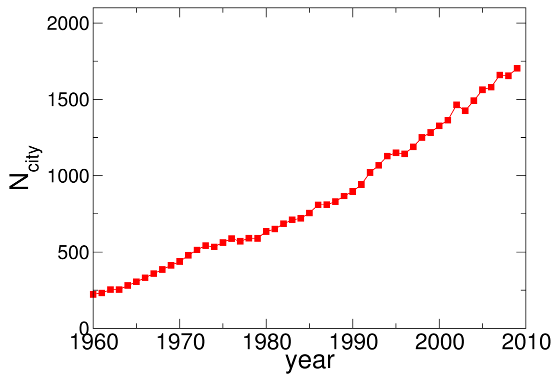


Figure 7: The number of nodes (cities) for city-to-city citation networks as the function of time (1960-2009).

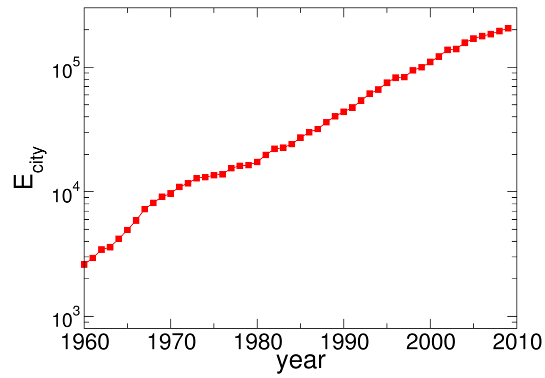
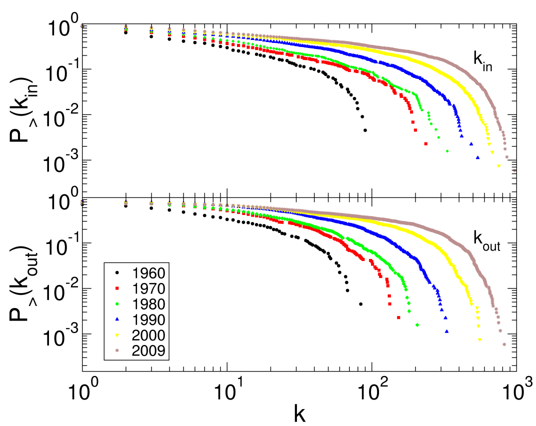
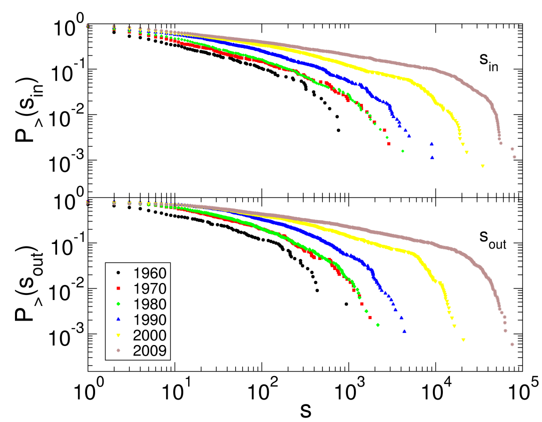


Figure 8: The number of links for city-to-city citation networks as the function of time (1960-2009).



(a) The cumulative distribution function of the degrees for citation networks at the city level.



(b) The cumulative distribution function of the strength for citation networks at the city level.

Figure 9: The cumulative distribution function of degree and strength for city-to-city citation networks in year 1960, 1970, 1980, 1990, 2000 and 2009.

4 Top producers/consumers and results from knowledge diffusion proxy

In Figure. 10 we show the cumulative distribution of the absolute citation unbalance $|\Delta s|$ for producers and consumers at the city level. Similar to the cumulative distributions of strength, the distributions are characterized with heavy tails, and the distributions have become broader as the time increases.

We list top 20 producers and consumers at the city level from 1985 to 2009 (Table. 3), from 1960 to 1980 (Table. 4). It is worth noting that the definition of unbalance Δs is from the difference between the number of citations sent and received, which cannot distinguish between cities with a large amount of production and consumption and those with less production and consumption.

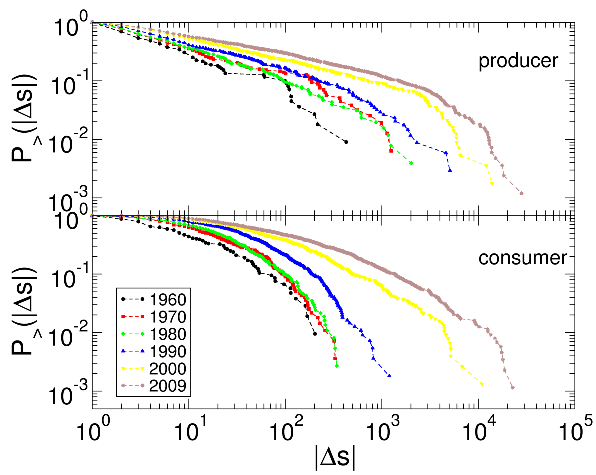


Figure 10: The cumulative distribution function of the citation unbalance for producers and consumers at the city level in year 1960, 1970, 1980, 1990, 2000 and 2009.

Table 3: Top 20 producers and consumers at the city level (1985-2009)

(a) Top 20 producer cities

rank	1985	1990	1995	2000	2005	2009
1	Piscataway	Piscataway	Piscataway	Boston	Boston	Boston
2	Boston	Boston	Boston	Piscataway	New York City	Berkeley
3	Berkeley	Palo Alto	Yorktown Heights	Los Angeles	Los Angeles	New Haven
4	Princeton	Yorktown Heights	Berkeley	Berkeley	Tallahassee	Suwon
5	Yorktown Heights	Berkeley	Los Angeles	Chicago	Palo Alto	Princeton
6	Ithaca	Princeton	Urbana	New York City	Berkeley	Piscataway
7	New York City	Ithaca	New York City	Lemont	Piscataway	Higashihiroshima
8	DC	New York City	Chicago	Urbana	Urbana	Prairie View
9	Palo Alto	San Diego	Ithaca	Philadelphia	Pavia	Los Angeles
10	Lemont	Philadelphia	Lemont	Princeton	West Lafayette	Lubbock
11	Los Angeles	Chicago	Princeton	West Lafayette	Ithaca	Palo Alto
12	Chicago	Santa Barbara	Palo Alto	Batavia	Rochester	Batavia
13	San Diego	Pittsburgh	Santa Barbara	Rochester	Honolulu	New York City
14	Seattle	Lemont	Philadelphia	Yorktown Heights	Batavia	Nashville
15	Rehovot	Los Angeles	Minneapolis	Palo Alto	Yorktown Heights	Bristol
16	New Haven	New Haven	San Diego	Dallas	Irvine	Rochester
17	Urbana	Orsay	Batavia	Tsukuba	Lemont	Urbana
18	Pittsburgh	Holmdel	Zurich	Waltham	Minneapolis	Daegu
19	Villigen	Stony Brook	Waltham	Madison	Philadelphia	Tallahassee
20	Waltham	Batavia	Madison	East Lansing	Boulder	Pittsburgh

(b) Top 20 consumer cities

rank	1985	1990	1995	2000	2005	2009
1	Stuttgart	Tokyo	Moscow	Beijing	Beijing	Athens
2	Toronto	Beijing	Beijing	Seoul	Barcelona	Gwangju
3	Gaithersburg	Tsukuba	Seoul	Lancaster	Coventry	Bratislava
4	Annandale	Tallahassee	East Lansing	Grenoble	Valencia	Vancouver
5	Bloomington	Vancouver	Lubbock	Dubna	Perugia	Madrid
6	Minneapolis	Grenoble	Montreal	Manhattan	Moscow	Berlin
7	Warsaw	Seoul	Tallahassee	Quito	Heidelberg	Trieste
8	Berlin	Kolkata	Davis	Suwon	London	Mainz
9	Vancouver	Charlottesville	Dallas	Stillwater	Dubna	Waco
10	Ames	Durham	Taipei	Santander	Riverside	Paris
11	West Lafayette	Buffalo	Berlin	Lawrence	Amsterdam	Valencia
12	Charlottesville	Warsaw	Tokyo	Kraków	Hefei	Coventry
13	Seoul	Tempe	Toyonaka	Marseille	Dresden	Moscow
14	Montreal	Berlin	Delhi	Tokyo	Bellaterra	Bellaterra
15	Trieste	Madrid	Trieste	Karlsruhe	Shanghai	Lanzhou
16	Kyoto	Sao Paulo	St Petersburg	Daegu	Evanston	Shanghai
17	Tokyo	Taipei	Dresden	Udine	Taipei	Sao Paulo
18	Varanasi	Brussels	Bologna	Oxford	Glasgow	Kolkata
19	Rio De Janeiro	Mainz	Munich	Moscow	Liverpool	Clermont
20	Ridgefield	Davis	Cambridge	Ruston	Bari	Hefei

Table 4: Top 20 producers and consumers at the city level (1960-1980)

(a) Top 20 producer cities

rank	1960	1965	1970	1975	1980
1	Boston	Princeton	Berkeley	Boston	Boston
2	Princeton	Berkeley	Boston	Berkeley	Princeton
3	Urbana	Boston	Princeton	Palo Alto	Piscataway
4	Oak Ridge	Piscataway	Chicago	Princeton	Berkeley
5	Piscataway	New York City	Piscataway	Piscataway	Palo Alto
6	New York City	Los Angeles	Palo Alto	Ithaca	Ithaca
7	Los Angeles	Los Alamos	Albany	Chicago	New York City
8	Los Alamos	Albany	San Diego	Oak Ridge	Chicago
9	Chicago	Ann Arbor	Madison	San Diego	San Diego
10	Ithaca	Pittsburgh	New York City	New Haven	Los Angeles
11	Rochester	Meyrin	Pittsburgh	Los Angeles	Stony Brook
12	DC	Waltham	Waltham	Urbana	New Haven
13	Madison	Urbana	Meyrin	Pittsburgh	Philadelphia
14	Bloomington	Cambridge	Ithaca	Batavia	Albany
15	Utrecht	Bloomington	Cambridge	Providence	Urbana
16	Durham	Lemont	Los Angeles	Albany	Albuquerque
17	London	Ithaca	Los Alamos	Durham	Waltham
18	Saskatoon	DC	New Haven	Rochester	Batavia
19	Sydney	Chicago	Livermore	Livermore	College Park
20	St Louis	Zurich	London	DC	Pittsburgh

(b) Top 20 consumer cities

rank	1960	1965	1970	1975	1980
1	Berkeley	West Lafayette	Evanston	Stony Brook	Austin
2	Palo Alto	Palo Alto	West Lafayette	Grenoble	Boulder
3	New Haven	Orsay	Austin	Columbus	Tokyo
4	Pittsburgh	College Park	Trieste	Stuttgart	Haifa
5	Waltham	Albuquerque	Columbus	Toronto	Toronto
6	San Diego	Livermore	Delhi	Austin	Bhubaneswar
7	Lemont	Delhi	Amherst	East Lansing	Rehovot
8	Livermore	Minneapolis	Rochester	Amherst	Ottawa
9	West Lafayette	Trieste	Milwaukee	Mumbai	Paris
10	Poughkeepsie	Providence	Baton Rouge	Denton	Santa Barbara
11	Evanston	Ames	Buffalo	Mexico City	Houston
12	Tallahassee	Rochester	Seattle	Munich	Golden
13	Columbus	Evanston	Salt Lake City	Paris	Stuttgart
14	Canberra	San Diego	Haifa	Honolulu	Kolkata
15	Yorktown Heights	Syracuse	Hoboken	Montreal	Toyonaka
16	Arlington	Rehovot	Lincoln	Orsay	Kyoto
17	Rome	Hoboken	Gainesville	Roskilde	Grenoble
18	Meyrin	Oxford	Tucson	Madison	Jülich
19	Ames	El Segundo	Bloomington	West Lafayette	Vancouver
20	Irvine	Milan	East Lansing	Rehovot	Kingston

5 Top ranked cities from scientific production ranking algorithm

We show the cumulative distribution of scientific production ranking scores for cities in selected years in Figure. 11. We notice that ranking scores are also characterized with heavy tail distributions. In addition, we also observe that both the maximum and minimum ranking scores has decreased with time, and the tail of the distribution becomes steeper in recent decades, which indicates the differences of ranking scores between top ranked cities have gradually shrunk.

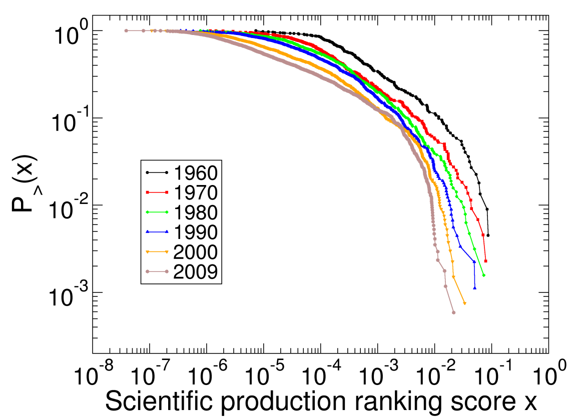


Figure 11: The cumulative distribution function of scientific production ranking scores for cities in year 1960, 1970, 1980, 1990, 2000 and 2009.

In Table. 5 and Table. 6, we report top 50 cities ranked from scientific production ranking algorithm from 1985 to 2009 and from 1960 to 1980 respectively.

Table 5: Top 50 cities from scientific production ranking algorithm (1985-2009)

rank	1985	1990	1995	2000	2005	2009
1	Piscataway	Piscataway	Boston	Boston	Boston	Boston
2	Boston	Boston	Piscataway	Berkeley	Los Angeles	Berkeley
3	Berkeley	Berkeley	Berkeley	Piscataway	Berkeley	Los Angeles
4	Palo Alto	Palo Alto	Los Angeles	Los Angeles	Orsay	Tokyo
5	New York City	Yorktown Heights	New York City	New York City	Tokyo	Orsay
6	Los Angeles	Los Angeles	Urbana	Chicago	Princeton	Chicago
7	Ithaca	New York City	Chicago	Urbana	Piscataway	Paris
8	Los Alamos	Los Alamos	Lemont	Rochester	Palo Alto	Princeton
9	Princeton	Princeton	Palo Alto	Batavia	New York City	Rome
10	Yorktown Heights	Urbana	Batavia	West Lafayette	Philadelphia	Piscataway
11	Lemont	Chicago	Philadelphia	Lemont	Urbana	London
12	Urbana	Philadelphia	Madison	Orsay	Santa Barbara	Urbana
13	Chicago	Ithaca	Rochester	East Lansing	Rome	Lemont
14	Philadelphia	Lemont	West Lafayette	Ann Arbor	Columbus	Philadelphia
15	Orsay	Orsay	Orsay	Tokyo	College Park	Oxford
16	DC	Santa Barbara	Princeton	College Station	New Haven	Santa Barbara
17	College Park	College Park	Los Alamos	Tsukuba	Lemont	New Haven
18	Oak Ridge	Oak Ridge	Rome	Philadelphia	Madison	Rochester
19	Santa Barbara	Livermore	Tsukuba	Palo Alto	Paris	Madison
20	Rochester	Batavia	Santa Barbara	Madison	San Diego	Columbus
21	Rehovot	Tokyo	Yorktown Heights	College Park	Chicago	College Park
22	San Diego	Rochester	College Station	Pittsburgh	Tsukuba	Batavia
23	Pittsburgh	San Diego	Pittsburgh	Rome	Oxford	Moscow
24	New Haven	Columbus	Ithaca	Princeton	Oak Ridge	East Lansing
25	Stony Brook	Madison	College Park	Los Alamos	Tallahassee	Palo Alto
26	Seattle	Pittsburgh	New Haven	New Haven	Rochester	Pittsburgh
27	Columbus	DC	Ann Arbor	Toyonaka	Beijing	San Diego
28	Boulder	Rehovot	Pisa	Durham	Pittsburgh	Ann Arbor
29	Paris	Stuttgart	Waltham	Columbus	Ames	Tsukuba
30	Livermore	Paris	East Lansing	Stony Brook	West Lafayette	Seoul
31	Madison	Minneapolis	Oak Ridge	Santa Barbara	Batavia	Pisa
32	Austin	Boulder	Tokyo	Albuquerque	Pisa	West Lafayette
33	Tokyo	New Haven	Stony Brook	Baltimore	Boulder	Padua
34	Jülich	West Lafayette	San Diego	Toronto	Padua	Dubna
35	Zurich	Stony Brook	Minneapolis	Pisa	London	Evanston
36	Batavia	Bloomington	Baltimore	Tallahassee	Montreal	Ames
37	Bloomington	Seattle	Padua	Waltham	Livermore	New York City
38	Minneapolis	Ann Arbor	Toronto	Ithaca	Los Alamos	Toronto
39	West Lafayette	Austin	Boulder	Moscow	Seoul	Oak Ridge
40	Ann Arbor	Zurich	Albuquerque	Montreal	East Lansing	Baltimore
41	East Lansing	Vancouver	Stuttgart	Padua	Moscow	Beijing
42	Stuttgart	Holmdel	Livermore	San Diego	Nashville	Karlsruhe
43	Evanston	Rome	DC	Ames	Ann Arbor	Taipei
44	Grenoble	Ames	Paris	Evanston	College Station	College Station
45	Syracuse	Waltham	Seattle	Meyrin	Vancouver	Meyrin
46	Providence	Albuquerque	Rehovot	Gainesville	Irvine	Los Alamos
47	Ames	Toyonaka	Durham	Honolulu	Taipei	Toyonaka
48	Albany	Albany	Toyonaka	Paris	Dallas	Liverpool
49	Waltham	Jülich	Columbus	Oak Ridge	Meyrin	Davis
50	Nashville	Grenoble	Dallas	Bloomington	Cincinnati	Amsterdam

Table 6: Top 50 cities from scientific production ranking algorithm (1960-1980)

rank	1960	1965	1970	1975	1980
1	Berkeley	Berkeley	Boston	Boston	Boston
2	Boston	Boston	Berkeley	Piscataway	Piscataway
3	New York City	Princeton	Piscataway	Berkeley	Berkeley
4	Princeton	Piscataway	Palo Alto	Palo Alto	Palo Alto
5	Chicago	New York City	Princeton	New York City	New York City
6	Piscataway	Chicago	New York City	Princeton	Princeton
7	Urbana	Los Angeles	Chicago	Ithaca	Los Angeles
8	Los Angeles	Urbana	Los Angeles	Los Angeles	Chicago
9	Ithaca	Palo Alto	Urbana	Chicago	Ithaca
10	Pittsburgh	Pittsburgh	Ithaca	Lemont	Lemont
11	Oak Ridge	Lemont	Pittsburgh	Urbana	Los Alamos
12	Los Alamos	DC	Lemont	Batavia	Philadelphia
13	DC	Ithaca	San Diego	Philadelphia	Urbana
14	Rochester	Los Alamos	Oak Ridge	Oak Ridge	Oak Ridge
15	Philadelphia	Albany	Philadelphia	Pittsburgh	College Park
16	Albany	Oak Ridge	DC	College Park	Batavia
17	Palo Alto	Philadelphia	Albany	DC	Orsay
18	Lemont	Waltham	New Haven	San Diego	Stony Brook
19	New Haven	New Haven	Waltham	Rochester	DC
20	Madison	Madison	College Park	Los Alamos	Pittsburgh
21	College Park	San Diego	Los Alamos	New Haven	Rochester
22	Bloomington	College Park	Madison	Madison	Yorktown Heights
23	Waltham	Rochester	Rochester	Waltham	New Haven
24	Ann Arbor	Ann Arbor	Ann Arbor	Stony Brook	San Diego
25	Minneapolis	Livermore	West Lafayette	Yorktown Heights	Rehovot
26	West Lafayette	West Lafayette	Livermore	Albany	Madison
27	Houston	Meyrin	Minneapolis	Orsay	Livermore
28	Syracuse	Seattle	Rehovot	Seattle	Seattle
29	Livermore	Minneapolis	Oxford	Providence	Waltham
30	Columbus	Rehovot	London	Livermore	Albany
31	Durham	Cleveland	Yorktown Heights	Rehovot	Evanston
32	St Louis	Yorktown Heights	Meyrin	Minneapolis	West Lafayette
33	Oxford	Oxford	Orsay	Evanston	Austin
34	Cleveland	London	Ames	Durham	Providence
35	Baltimore	Bloomington	Evanston	West Lafayette	Minneapolis
36	Seattle	Evanston	Seattle	Ames	Ann Arbor
37	Providence	Cambridge	Cleveland	London	Albuquerque
38	Rehovot	St Louis	Stony Brook	Ann Arbor	Paris
39	Ames	Syracuse	Cambridge	Cleveland	East Lansing
40	Cambridge	Ames	Providence	East Lansing	Bloomington
41	London	Detroit	Durham	Albuquerque	Cleveland
42	Ottawa	Columbus	Santa Barbara	Austin	College Station
43	Tokyo	Durham	Boulder	Oxford	Zurich
44	Meyrin	Orsay	Riverside	Santa Barbara	Oxford
45	Detroit	Houston	St Louis	St Louis	Ames
46	South Bend	Boulder	Hamburg	Boulder	London
47	Birmingham	Baltimore	Detroit	Columbus	Durham
48	Jerusalem	Tokyo	Columbus	Zurich	Boulder
49	San Diego	Paris	Syracuse	Cambridge	St Louis
50	Sydney	Rome	Bloomington	Rome	Columbus

6 Relation between research outputs and investment

In this section, we report the relation between research outputs (i.e., citations) and investment on scientific research. As discussed earlier, we parsed city information based on country information for each affiliation, therefore we can aggregate the number of citations for cities to their countries, and measure the relation between research outputs and investment on research in that country. In Figure. 12, we plot the correlation between the average number of citations received by each country in 1996-2009 and the average amount of gross domestic product (GDP) spent on research and development (R& D) (in current US dollars) in that country in that period. We also plot the correlation between the average number of citations received by one country in the same period and the average research population in that country within the same time window. The number of citations received approximately linearly scales with both quantities. Such findings are consistent with the results reported in [6], which studied the database of the Institute for Scientific Information (ISI). This similarity indicates, although APS dataset is limited, it is representative of the scientific production for major countries. The data of GDP, the fraction of GDP spent on R& D, and the research population are from The World Bank data [5].

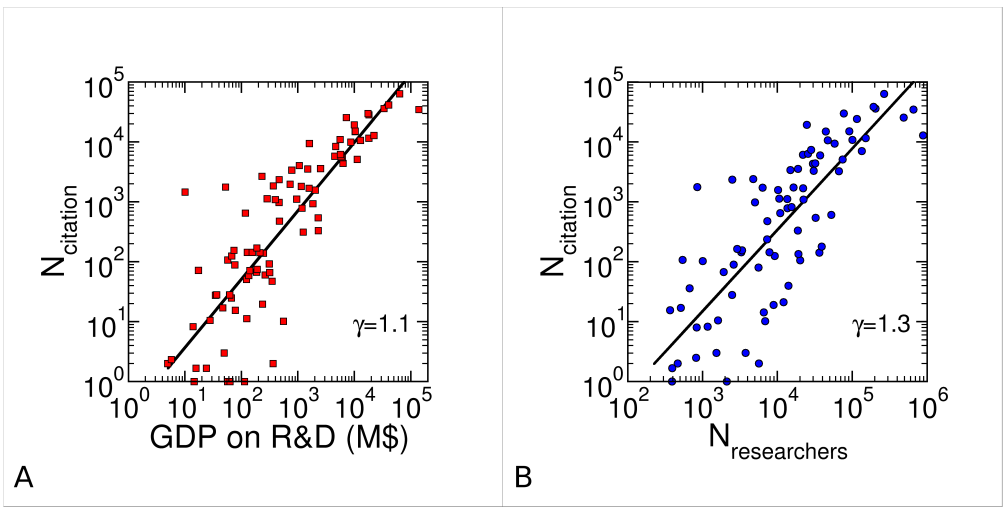


Figure 12: Relation between research outputs and the investment. (A) The average citations received by each country as a function of the average GDP on research and development (R& D) in million US dollars from 1996 to 2009. (B) The average citations received by each country as a function of the average research population in that country from 1996 to 2009. The solid black line shows the power-law fitting with the exponent 1.1 and 1.3 respectively.

References

[1] http://www.iso.org/iso/country_codes.htm.

[2] http://en.wikipedia.org/wiki/List_of_U.S._states.

[3] http://code.google.com/p/geopy/.

[4] http://www.geonames.org/.

[5] http://data.worldbank.org/.

[6] Raj Kumar Pan, Kimmo Kaski, and Santo Fortunato. World citation and collaboration networks: uncovering the role of geography in science. *Scientific Reports*, 2:902, 2012.