

RANDOM WALK IN THE ACADEMIC PUBLICATION

A data-driven approach

Ahmad Reza Ehyaei



Contents

Introduction	2
Papers	3
Affiliations	9
Authors	14

Introduction

Open Academic Graph (OAG) is a large knowledge graph unifying Microsoft Academic Graph (MAG) and AMiner. Two projects' major purpose is to develop a heterogeneous graph comprising scientific publication records and citation linkages between those articles, as well as authors, institutions, journals, conferences, and fields of study. After unifying the two datasets, the result can be found [here](#).

The dataset contains many tables, as:

- authors: includes some information about the authors, such as their names, affiliations, positions, and a summary of their publications.
- affiliations: contains the name of affiliation, type, and location by latitude and longitude.
- venues: shows information about where work was published.
- papers: this table has many fields and contains publication metadata such as title, publisher, issn, and...
- links: contains id tables for joining each table from MAG to AMiner.

If you have an Azure account, Microsoft has automated uploads of new versions of MAG to Azure Storage.

In this project, we can try to find new insights from the academic publication world!

Papers

The paper table contains information about paper's metadata. This table contains 208915369 rows. This is the biggest table in the academic graph datasets. In the below table, the fields are shown.

Table 1: Papers Table Schema

Field Name	Field Type	Description	Example
id	string/long	paper ID	53e9ab9eb7602d970354a97e
title	string	paper title	Data mining: concepts and techniques
authors.name	string	author name	Jiawei Han
author.org	string	author affiliation	Department of Computer Science, University of Illinois at Urbana-Champaign
author.org_id	long	author affiliation id	157725225
author.id	string/long	author ID	53f42f36dabfaedce54dcd0c
venue.id	string/long	paper venue ID	53e17f5b20f7dfbc07e8ac6e
venue.name	string	paper venue name	Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial
year	int	published year	2000
keywords	list of strings	keywords	["data mining", "structured data", "world wide web", "social network", "relational data"]
fos.name	string	paper fields of study	Web mining
fos.w	float	fields of study weight	0.6596908570000007
references	list of long	paper references	[4909282, 16018031, 16159250, 19838944, ...]
n_citation	int	citation number	40829
page_start	string	page start	11
page_end	string	page end	18
doc_type	string	paper type: journal, book title...	book
lang	string	detected language	en
publisher	string	publisher	Elsevier
volume	string	volume	10
issue	string	issue	29
issn	string	issn	0020-7136
isbn	string	isbn	1-55860-489-8
doi	string	doi	10.4114/ia.v10i29.873
pdf	string	pdf URL	//static.aminer.org/upload/pdf/1254/370/239/53e9ab9eb7602d970354a97e.pdf
url	list	external links	["http://dx.doi.org/10.4114/ia.v10i29.873", "http://polar.isi.uned.es/revista/index.php/ia/article/view/479"]
abstract	string	abstract	Our ability to generate...
indexed_abstract	string (JSON format)	indexed abstract	{"IndexLength": 164, "Inverte dIndex": {"Our": [0], "ability": [1], "to": [2, 7, ...]}}

Many questions concerning the Microsoft Academic Graph may be posed and searched. To begin, we attempt a statistical explorations of the data in order to obtain a better knowledge of the world of publications. We begin with a graph showing the annual publication count through time. The plot is shown below.

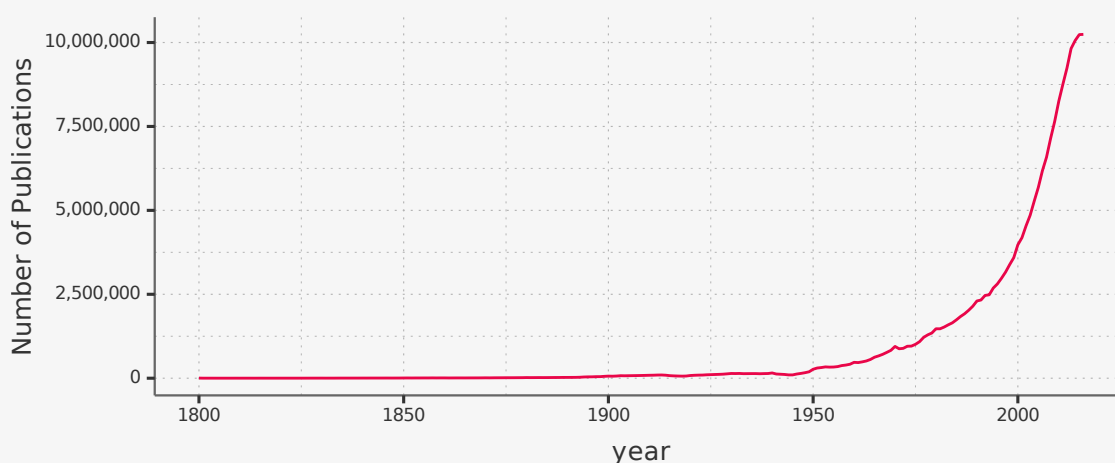


Figure 1: annual publication count through years

According to the plot, the number of publications has increased exponentially. The graph's vertical axis is scaled logarithmically in base 2 to make this clearer. In addition, the data is fitted using a regression line, and the findings reveal that the data follows an exponential model. The model slope coefficient indicates that every 25 years, the number of publications in the globe doubles.

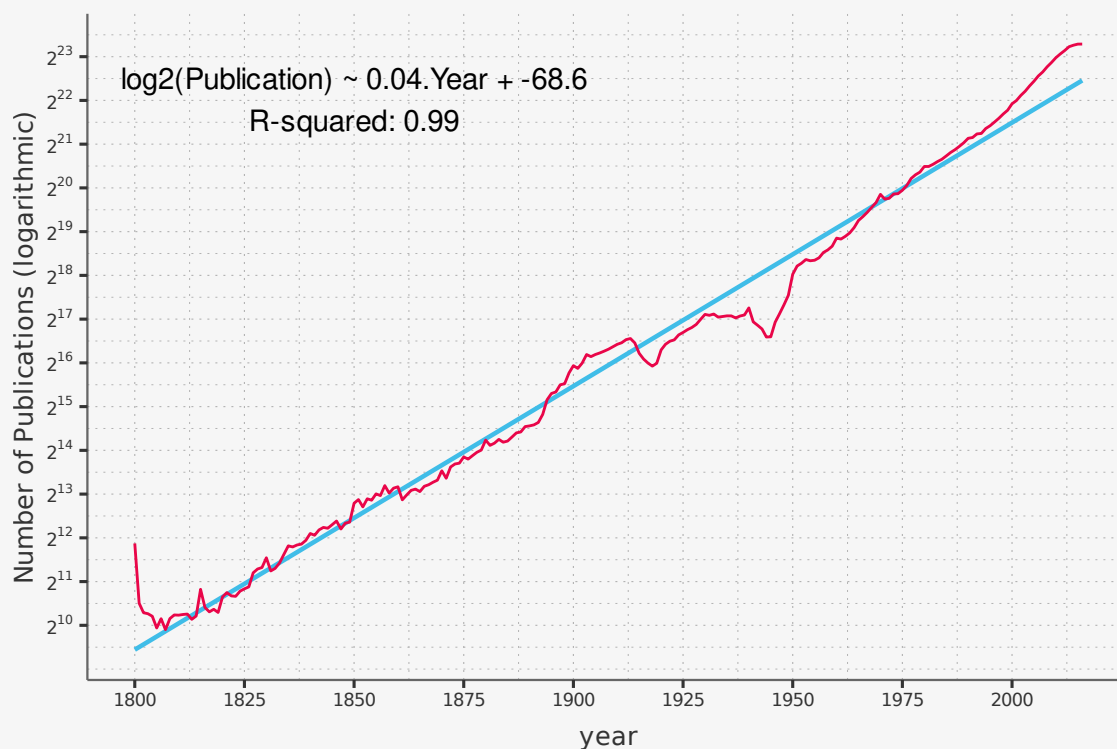


Figure 2: The behavior of annula publication's logarithmic growth

Table 2: Summary of Liear Model Log(Pubs) by Year.

(Intercept)	-68.623*** (0.662)
year	0.042*** (0.000)
Num.Obs.	217
R2	0.985
R2 Adj.	0.985
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001	

When you look at the annual publications, you see two sudden drops. The coincidence of this event with the world wars shows the effect of war on the decline of science.

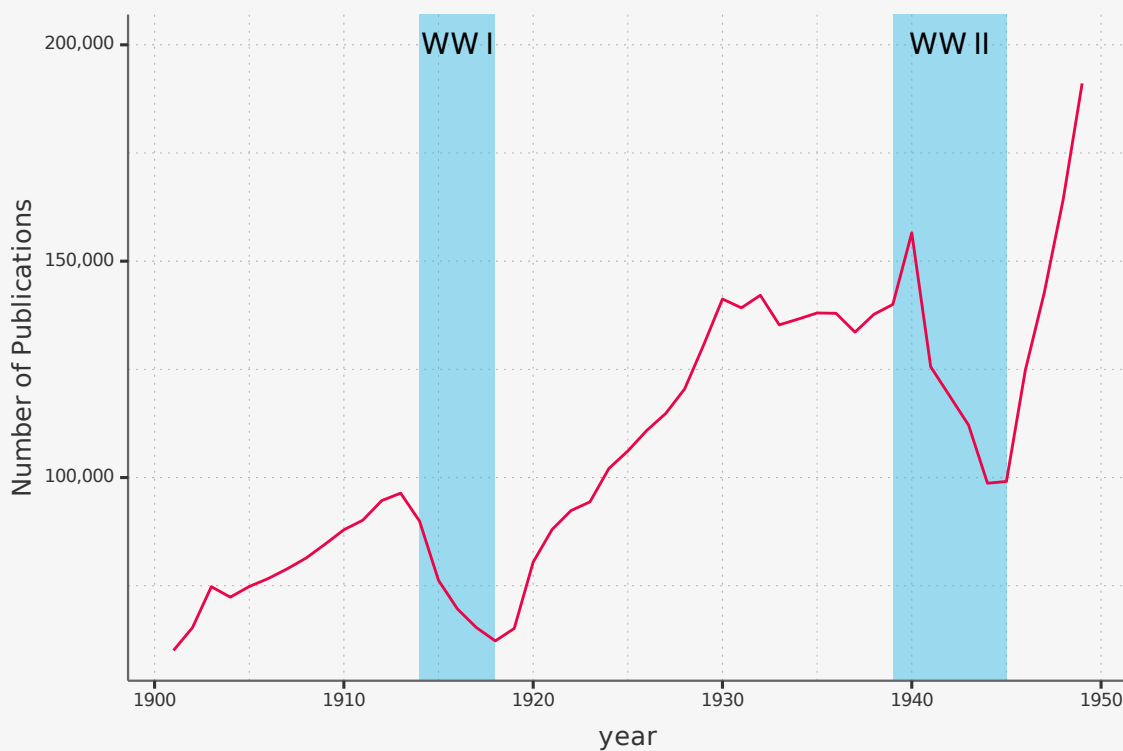


Figure 3: The effect of war on the growth of science

In the below table, we can view publications according to the type of document.

Document Frequency		
Type		Count
Journal		80,646,297
		75,300,276
Patent		45,298,165
Conference		4,250,431
BookChapter		2,312,088
Book		1,108,112
Total	—	208,915,369

As you can see, there are a lot of publications that aren't labeled.

The yearly number of publications is shown by document type in the graph below to illustrate the growth trend in further depth over the last fifty years (1967-2017).

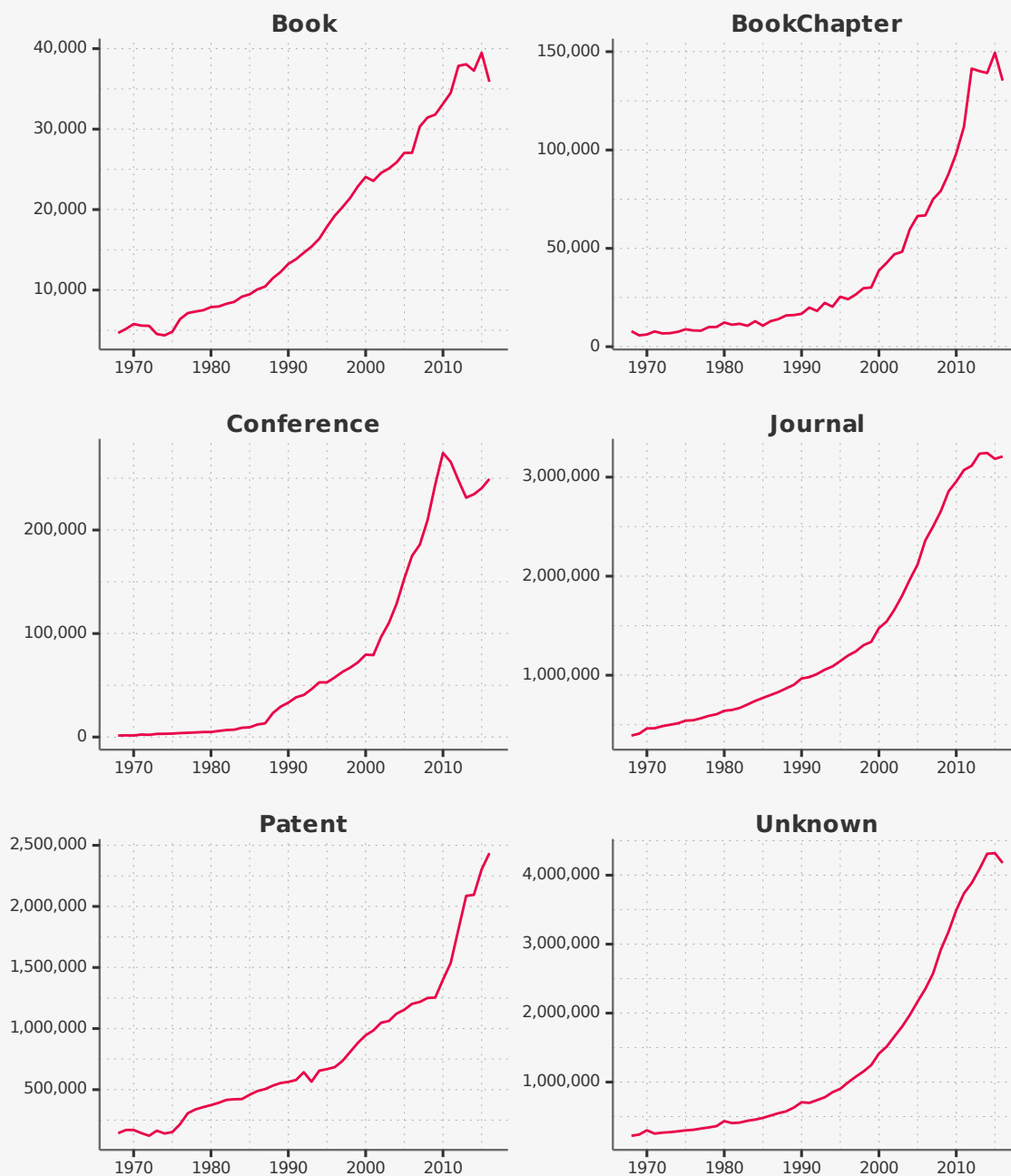


Figure 4: Annual publication number by document type

The conference publications appear to follow a slightly different pattern than the others.

The annual total number of citations was divided by the total number of articles to arrive at the impact factor. For each document, an annual chart was plotted to assess the effect of the published works.



Figure 5: Annual publication impact by document type

As can be observed, all types of documentation have a decreasing impact factor.

Another aspect of the data we looked into was the number of writers for each publication. The percentage of publications per number of writers is displayed in the bar plot below.

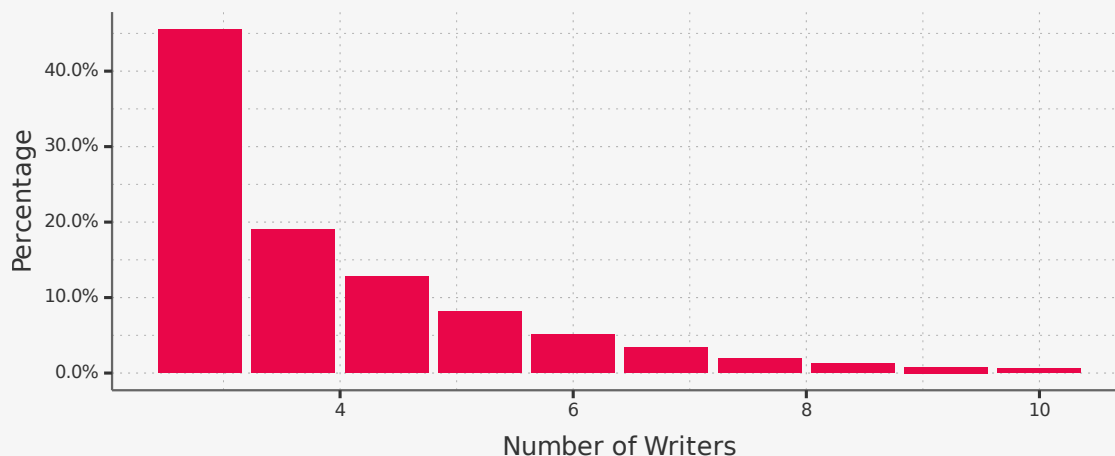


Figure 6: The percentage of the number of writers for publications

Researchers' willingness to collaborate has risen in recent years. To observe this statement in the data, we first compute the percentage of publications per year in terms of the number of writers, and then we build a heatmap of it.

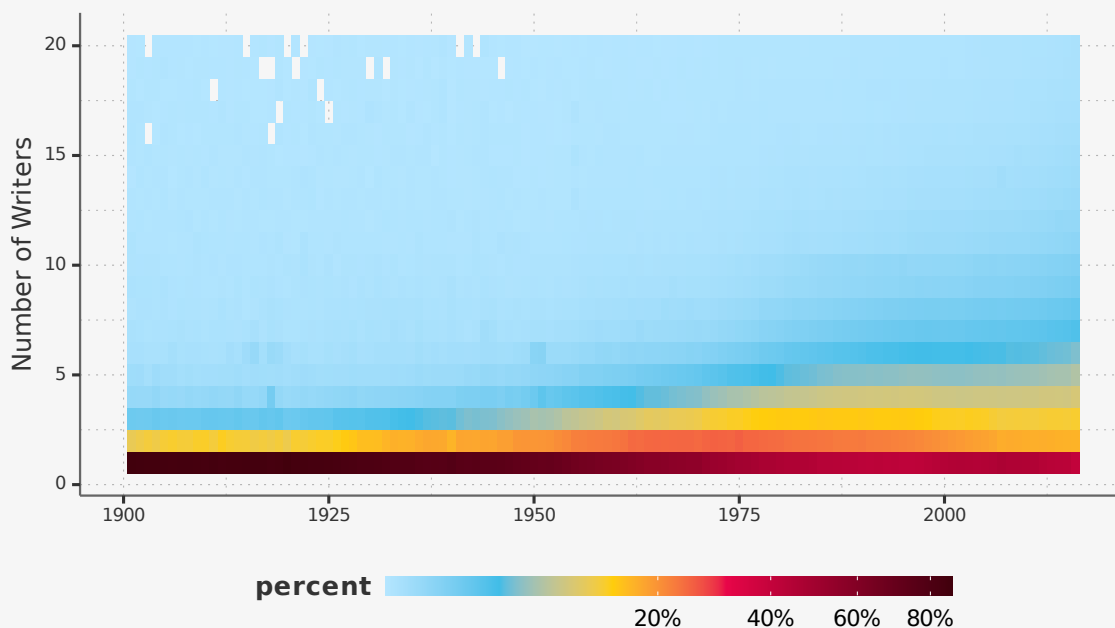


Figure 7: Annual number of writers heatmap

Affiliations

The affiliation table contains information about where the papers were submitted. This table contains 25708 rows. In the below table, the fields are shown.

Affiliation Table Schema			
Field Name	Field Type	Description	Example
id	string/long	affiliation id	136199984
DisplayName	string	affiliation name	Harvard University
NormalizedName	string	normalized affiliation name	harvard university
type	string	affiliation type	university
url	string	official page	http://www.harvard.edu/
WikiPage	string	Wikipedia page	http://en.wikipedia.org/wiki/Harvard_University
Latitude	string	latitude	42.37444
Longitude	string	longitude	-71.11694

To see the countries that are active in the field of research, we counted the number of research centers in each country. The heatmap below depicts the geographic distribution of research centers.

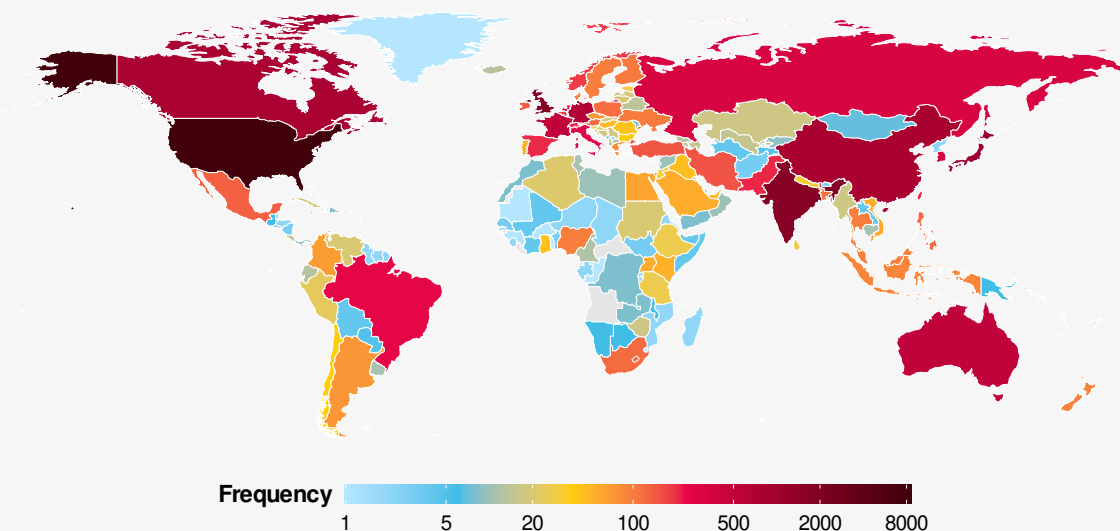


Figure 8: The distribution of academic organizations on the world map.

For more details, the table below shows the top ten countries with the most research centers.

Top 10 countries with the most academic organizations

country	Frequency
USA	8549
UK	1836
India	1767
Japan	1057
China	831
Canada	817
Germany	678
Australia	516
France	515
South Korea	338

The distribution of German research centers is displayed as a thermal diagram for further information.

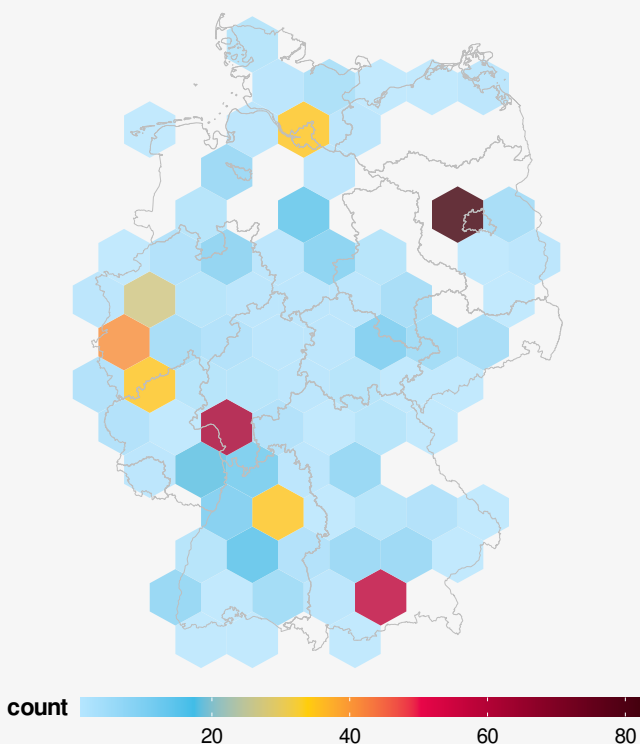


Figure 9: The location of academic organizations in Germany.

Each point on the chart reflects a research center's location. Each center's exact position on the map is

depicted in the scatter plot below.

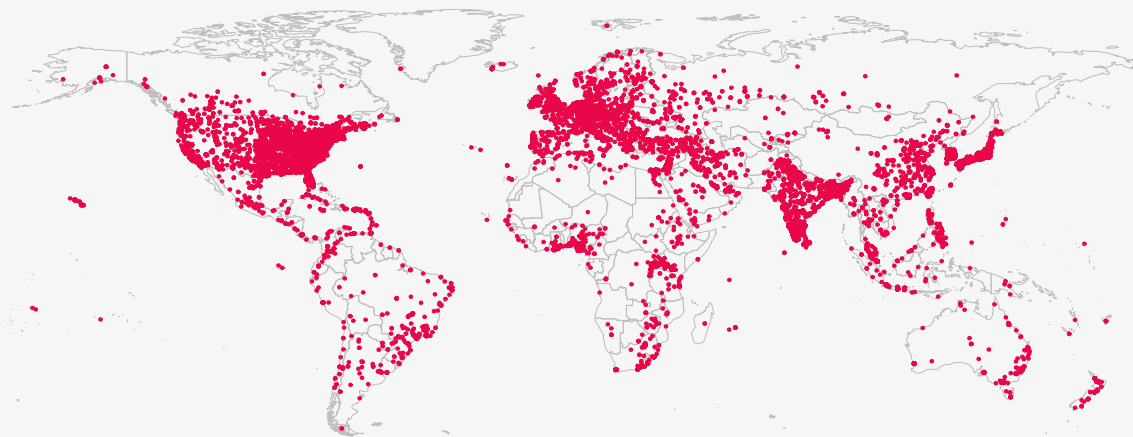


Figure 10: Location of academic organization on world map.

The total number of publications, total number of citations, and the impact factor for each institution were computed to determine the active research centers. The top 15 centers with the most publications are shown in the table below.

The Top 15 Organizations with the Most Publications.

Organization	Number of Publication	Number of Citation	Impact Factor
Chinese Academy of Sciences	2104240	26953600	12.8
Harvard University	1527420	70086023	45.9
Centre national de la recherche scientifique	1459893	28020818	19.2
University of Tokyo	1259705	22020022	17.5
Max Planck Society	1161691	31926466	27.5
CERN	1083666	19513077	18.0
Russian Academy of Sciences	1048923	6814819	6.5
Stanford University	1039212	40999300	39.5
National Institutes of Health	1025767	54469411	53.1
University of Michigan	982389	29504984	30.0
University of São Paulo	935371	8788652	9.4
Kyoto University	850419	15391665	18.1
Osaka University	849596	14316967	16.9
University of California, Los Angeles	813389	27580303	33.9
University of Washington	798856	32255361	40.4

Also included were the top research centers in terms of citations as well as the top research centers with the greatest impact factors.

The Top 15 Organizations with the Most Citations.

Organization	Number of Publication	Number of Citation	Impact Factor
Harvard University	1527420	70086023	45.9
National Institutes of Health	1025767	54469411	53.1
Stanford University	1039212	40999300	39.5
University of Washington	798856	32255361	40.4
Max Planck Society	1161691	31926466	27.5
University of Michigan	982389	29504984	30.0
Centre national de la recherche scientifique	1459893	28020818	19.2
University of California, Los Angeles	813389	27580303	33.9
Chinese Academy of Sciences	2104240	26953600	12.8
Massachusetts Institute of Technology	761634	26151027	34.3
Johns Hopkins University	707138	25169822	35.6
University of California, San Francisco	557786	24847360	44.5
University of California, Berkeley	715890	24248308	33.9
University of Pennsylvania	705037	24014870	34.1
University of California, San Diego	616412	23777136	38.6

The Top 15 Organizations with the Most Impacats and Over 1000 Publications.

Organization	Number of Publication	Number of Citation	Impact Factor
Celera Corporation	4609	3158398	685.3
SRA International	1009	374133	370.8
Wellcome Trust Sanger Institute	30396	5753162	189.3
J. Craig Venter Institute	11883	2149099	180.9
American Diabetes Association	1053	182241	173.1
Broad Institute	34194	5671132	165.9
Osiris Therapeutics, Inc.	1018	157242	154.5
deCODE genetics	4626	693394	149.9
Clinical Trial Service Unit	5061	719266	142.1
American Heart Association	5771	807924	140.0
American Cancer Society	9638	1276702	132.5
Affymetrix	5100	662813	130.0
Swiss Institute of Bioinformatics	6652	840026	126.3
Illumina	6178	778349	126.0
Élan	2842	334361	117.6

To see the statistical distribution of the impact factor of research centers, histogram it is plotted.

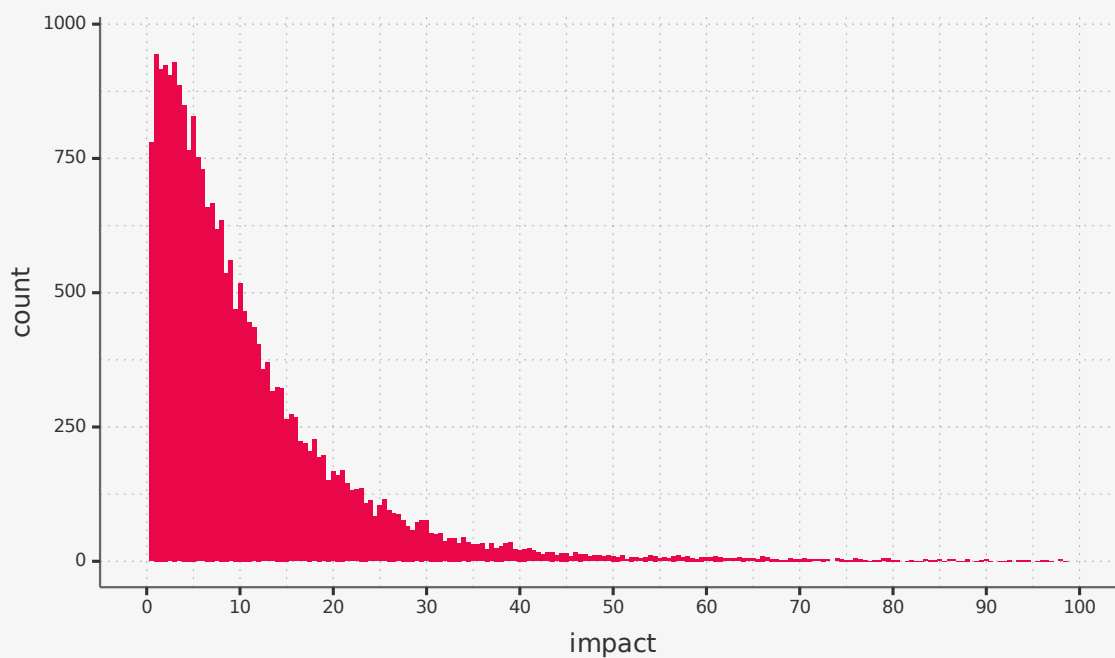


Figure 11: Histogram of Organization Imacts Less than 100 With a Binwidth of 0.5.

It is also possible to determine the degree of scientific impact for each country and display it on the globe as a heatmap.

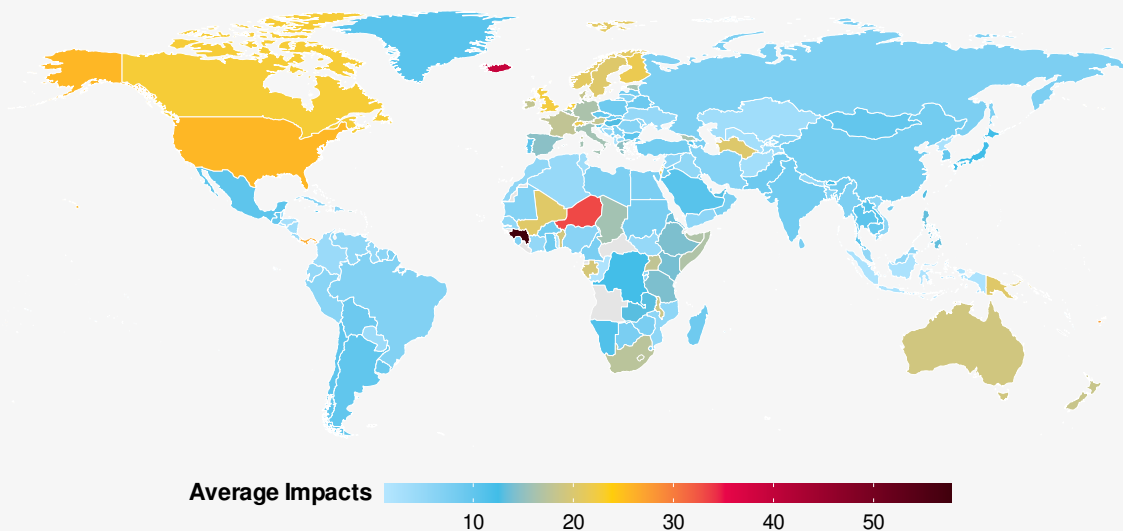


Figure 12: The distribution of acedemic organizations Impacts.

Authors

Fields for author information, academic history, and impact measures are included in the authors' table. This table contains 253144301 rows. The fields are listed in the table below.

Authors Table Schema

Field Name	Field Type	Description	Example
id	string/long	author id	53f42f36dabfaedce54dcd0c
name	string	author name	Jiawei Han
normalized_name	string	normalized author name	jiawei han
orgs	list of strings	author affiliations	["Department of Computer Science, University of Illinois at Urbana-Champaign"]
org	string	author organization	Department of Computer Science, University of Illinois at Urbana-Champaign
last_known_aff_id	long	last known affiliation	157725225
position	string	author position	professor
n_pubs	int	the number of author publications	1217
n_citation	int	author citation count	191526
h_index	int	author h-index	175
tags.t	string	research interests	"data mining"
tags.w	int	weight of interests	243
pubs.i	string/long	author paper id	53e9b9fbb7602d97045f7bb8
pubs.r	int	author order in the paper	0

The quantity of citations is used to quantify the impact of each research paper. The frequency of citations by writers is shown in the table below.

The Top 10 Frequency of Total Authors' Citations

Number of Citation	Frequency	Percent Share
0	159758931	63.1%
1	20144876	8.0%
2	11136954	4.4%
3	7359595	2.9%
4	5347158	2.1%
5	4136283	1.6%
6	3338694	1.3%
7	2791633	1.1%
8	2365922	0.9%
9	2049786	0.8%
10	1798441	0.7%
11	1597036	0.6%
12	1423916	0.6%
13	1288451	0.5%
14	1169026	0.5%

As can be seen, the research of many authors has not been used in the work of others.

A bar plot of the frequency of citations less than 100 has been displayed to observe the distribution of the number of citations.

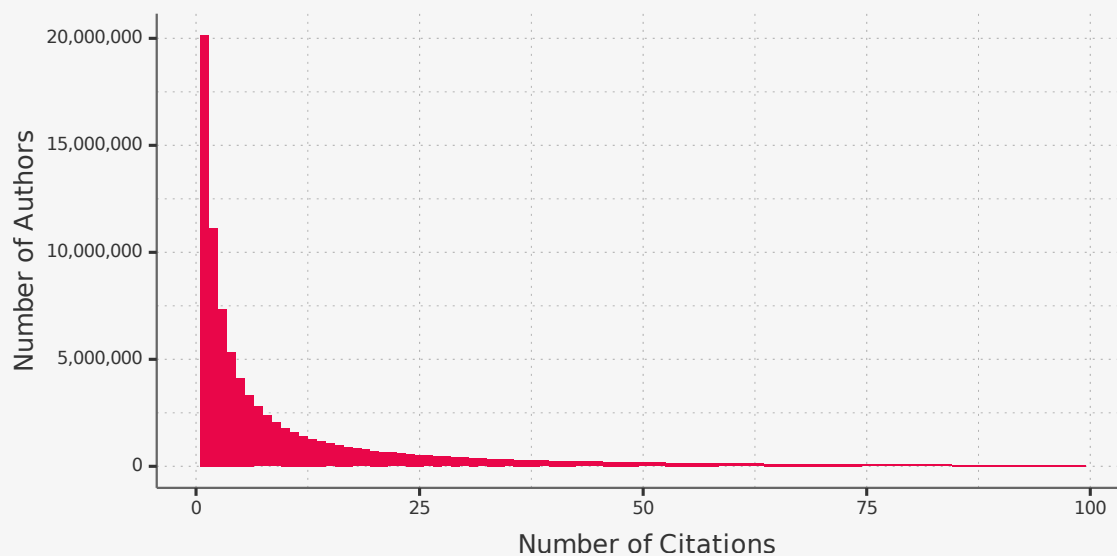


Figure 13: The Frequency of Authors by Number of Total Citations

To show the statistical behavior of writers with total citations of over 100, we used the logarithmic x-axis, the result of which can be seen below.

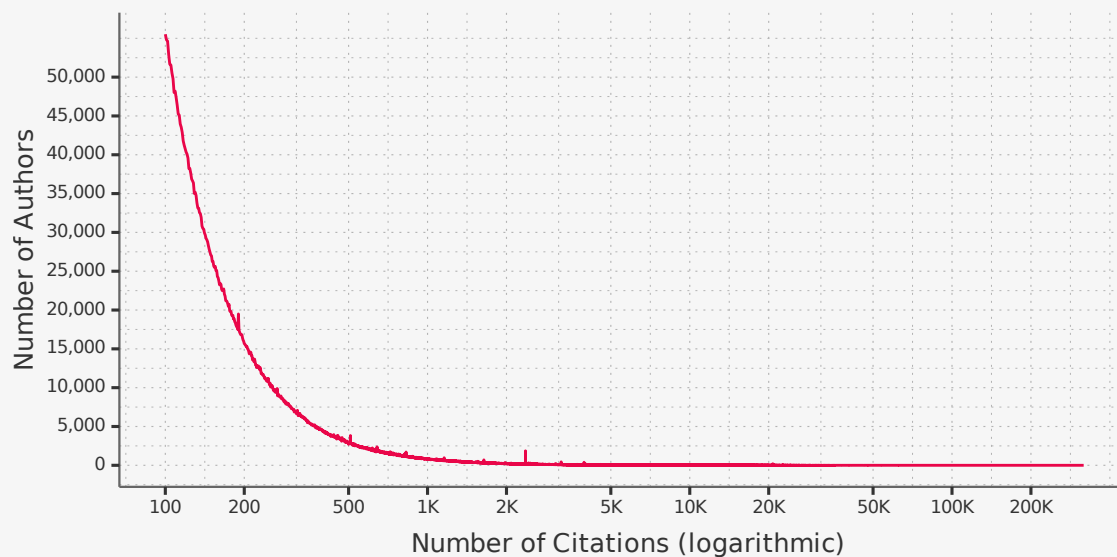


Figure 14: The Frequency of Authors by Number over Total Citations 100

To see how many people have more than specific value citations, we drew the tail Distribution of citations. For example, about a 3.6 million people have more than a 200 citations.

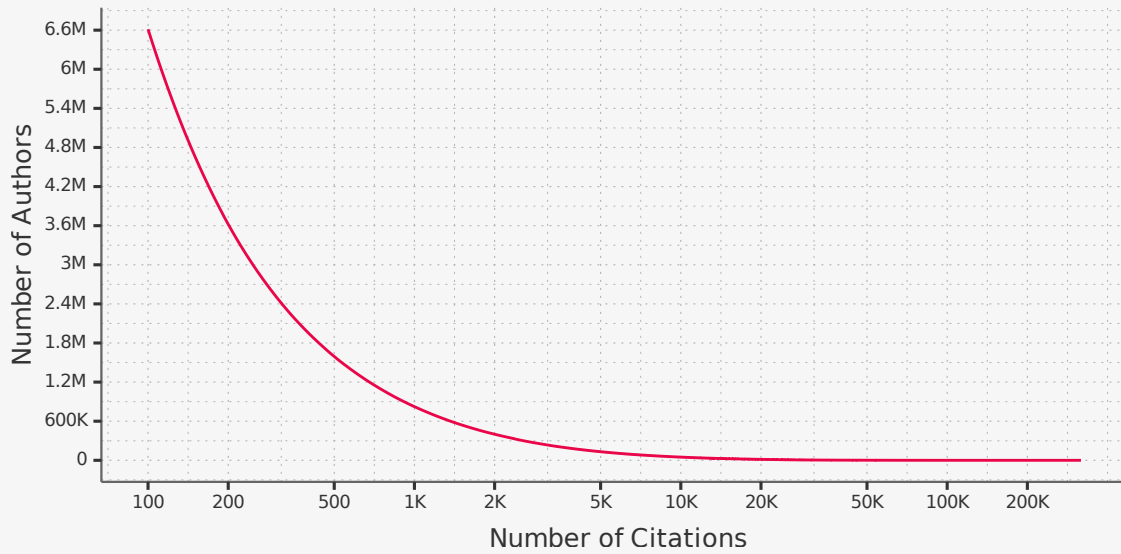


Figure 15: Tail Distribution of Authors over Total Citations 100

In the overall citation density function, there is an odd leap. A portion of the chart has been chosen for better visibility in the plot below.

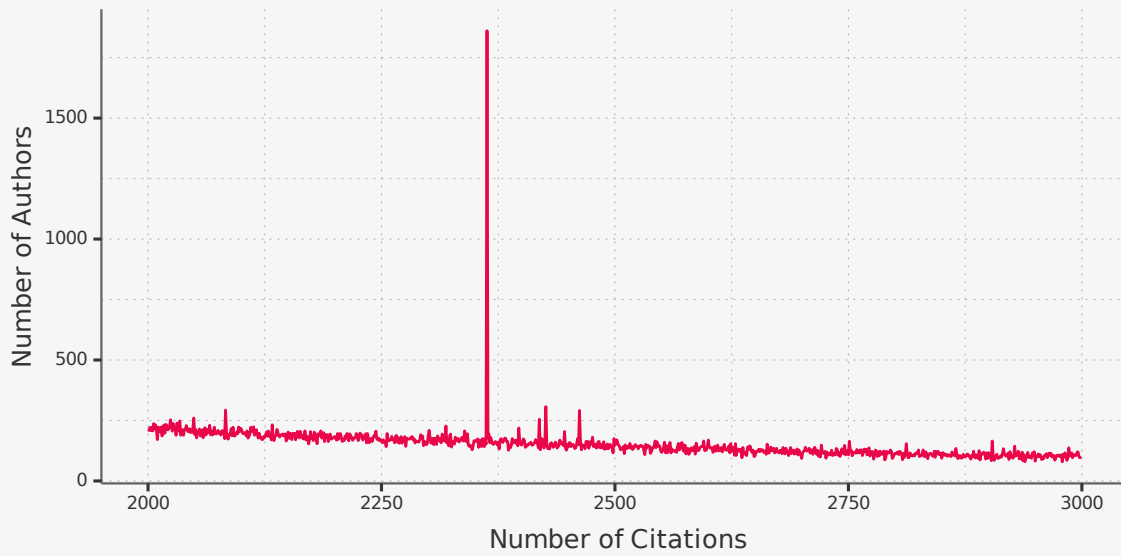


Figure 16: Strange Jump in Total Citations Density Function

The frequency of the number of articles by each author is shown in the table below.

The Top 15 Frequency of Total Number of Authors' Publications.

Number of Publication	Frequency	Percent Share
1	210702425	83.2%
2	18059438	7.1%
3	7251838	2.9%
4	3950886	1.6%
5	2477118	1.0%
6	1718516	0.7%
7	1248849	0.5%
8	954450	0.4%
9	748627	0.3%
10	605597	0.2%
11	498337	0.2%
12	418605	0.2%
13	354343	0.1%
14	305987	0.1%
15	264587	0.1%

The following table shows the frequency of the number of writers in relation to the number of articles. It's a logarithmic scale on the vertical axis.

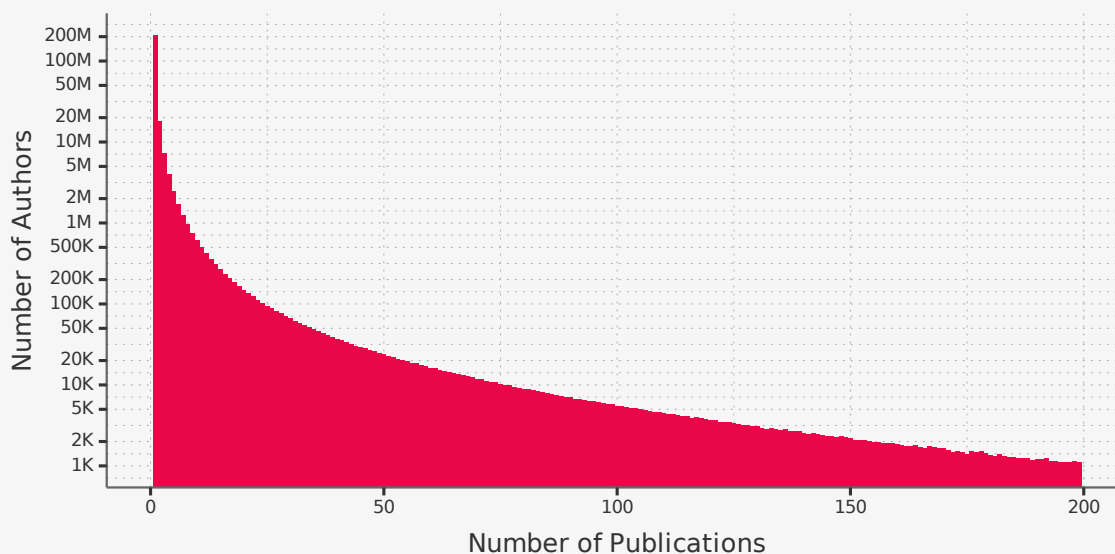
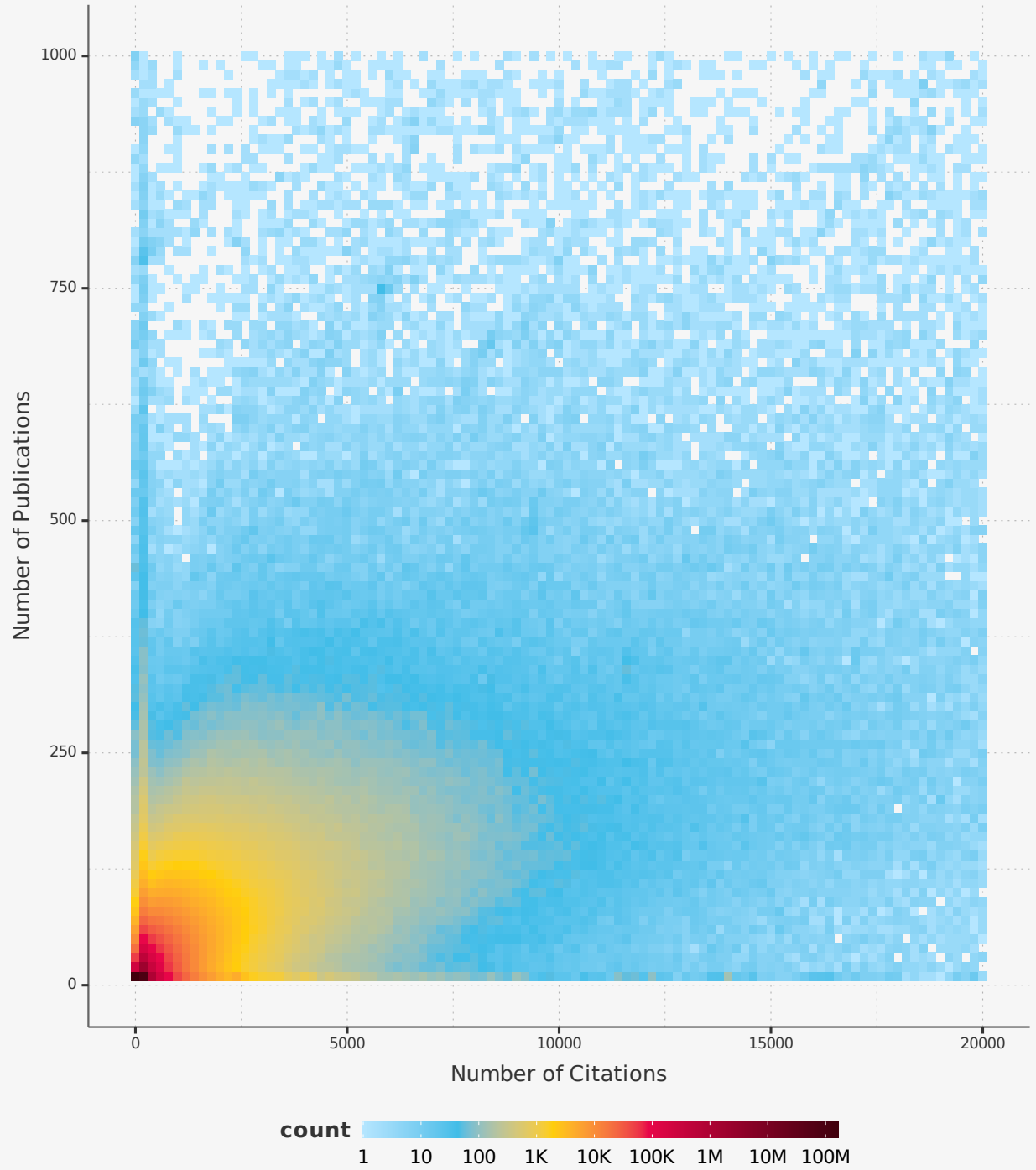


Figure 17: The Frequency of Authors by Number of Total Publications

In the following plot, the statistical behavior of the number of authors can be plotted in terms of the number of articles and the number of citations.



articles and 20 frequent tags. The percentage of papers submitted for each organization in each subject was then computed and shown in a heatmap.

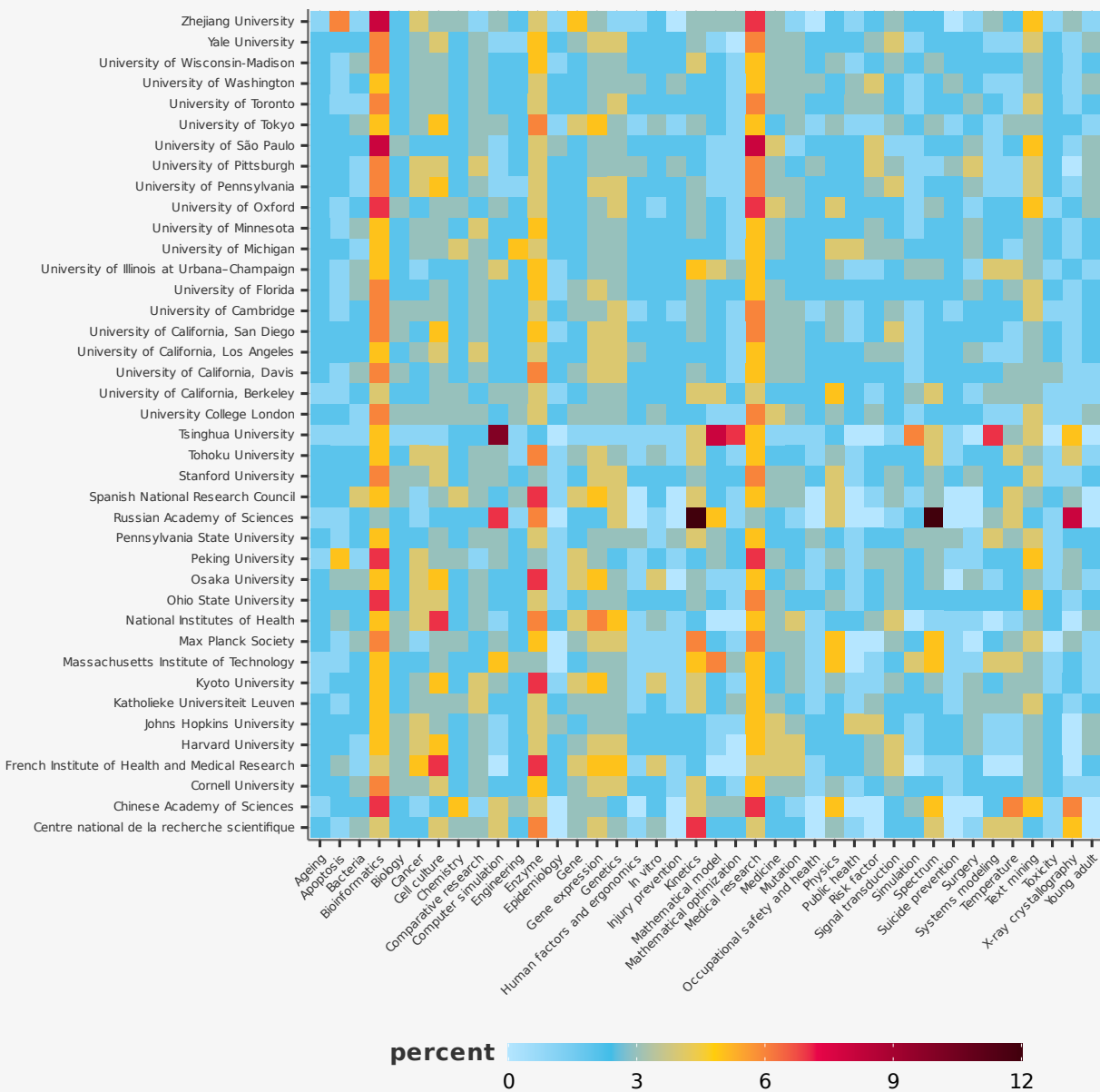


Figure 19: The percentage of publications by organizations and tags