Supplementary Information - World citation and collaboration networks: uncovering the role of geography in science

Raj Kumar Pan,[1] Kimmo Kaski,[1] and Santo Fortunato[1,2,*]

[1] *Department of Biomedical Engineering and Computational Science, Aalto University School of Science, P.O. Box 12200, FI-00076, Finland*

[2] *Complex Networks and Systems Lagrange Laboratory, Institute for Scientific Interchange (ISI), Torino, Italy*

I. MATERIALS AND METHODS

A. Data description

For our study we used all publications in English in the databases of Science Citation Index Expanded, Social Sciences Citation Index, and Arts & Humanities Citation Index for the years 2003-2010. The database of the Institute for Scientific Information (ISI) Web of Science also includes publications in other major languages, but consists of a relatively small number of items, accounting for $< 5\%$ of total publications. For each publication in the database, we have the name of the journal in which it is published, the volume and page number of the publication, its year of publication, the names of the authors, the list of their affiliations and its references and other additional information. We used the list of references to construct the network of citations between papers. For each publication we extracted the city and country of authors institutions from the affiliation data. Whenever a publication has several authors, it is counted and assigned to each location. Note that we only have the list of authors and the list of affiliations for each paper, however there is no corresponding match between these two lists and hence the individual level author affiliation can not be used in our study. Further although the affiliations are being recorded with increasing consistency, their use still poses major challenges in uniquely and accurately identifying them. For this reason, we parsed the affiliations of all publications and have determined the geographic location only at the city and country level. We also we use the publicly available resources (`www.wikipedia.org` and `maps.google.com`) to disambiguate the names of the places in case there are multiple name variation, typos and name changes during the time period of study.

B. GDP

The gross domestic product (GDP) is the value of all final goods and services produced within a nation in a given year and is the primary indicators used to gauge the health and size of a country's economy. We consider the average GDP (in US dollars) of a country during 2003-2010. A nation's GDP at purchasing power parity (PPP) exchange rates is the sum value of all goods and services produced in the country valued at prices prevailing in the United States. This is the measure most economists prefer when looking at per-capita welfare and when comparing living conditions or use of resources across countries.

C. R&D spending

Expenditures for research and development are current and capital expenditures (both public and private) on creative work undertaken systematically to increase knowledge, including knowledge of humanity, culture, and society, and the use of knowledge for new applications. R&D covers basic research, applied research, and experimental development.

D. Number of researcher

Researchers in R&D are professionals engaged in the conception or creation of new knowledge, products, processes, methods, or systems and in the management of the projects concerned. Postgraduate PhD students engaged in R&D are included.

E. Statistics

To fit the data and calculate different estimates we use the following methods:

Estimation of standard errors. Bootstrapping is a distribution-free re-sampling method used to estimate the parameters of interest from the empirical data. We have used this method in order to calculate the standard error of the mean. Let $x_1, x_2, \ldots, x_n$ be the dataset with mean $\bar{x}$. The standard error is then calculated as follows [1]: (i) Draw N samples each of size n with replacement from the original data. (ii) For each of the N samples calculate the sample mean $\hat{x_1}, ..., \hat{x_N}$ (iii) The standard error is then given by, $SEE(\bar{x}) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\hat{x_i} - \bar{\hat{x_i}})^2}$, where $\bar{\hat{x_i}} = N^{-1}\sum_i^N \hat{x_i}$ is the mean of the N bootstrap sample. In this study we have used 10^4 bootstrapped samples, i.e., $N = 10^4$.

* To whom correspondence should be addressed; e-mail: santo.fortunato@aalto.fi

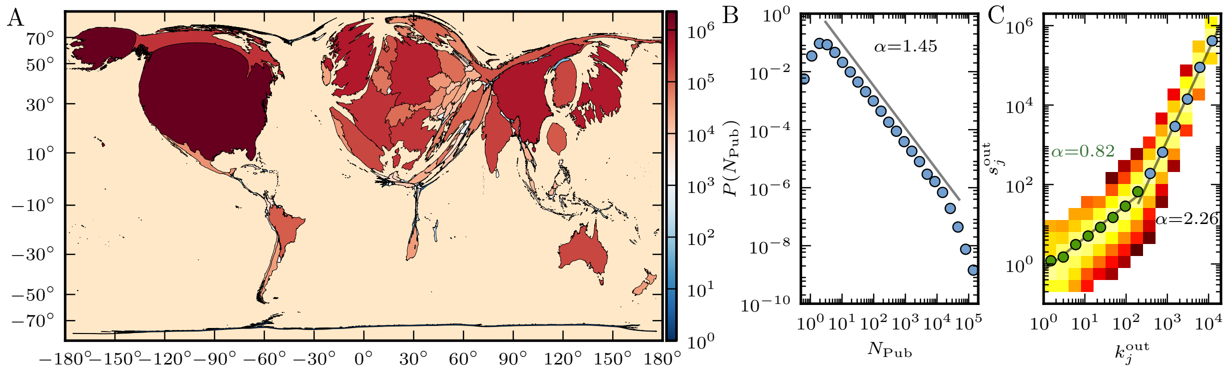

Figure S1. Research contribution in terms of number of publications. (A) Map of the country's research contribution, where the area of each country is scaled and deformed according to its number of publications. (B) The probability distribution function of the research contribution of cities in terms of their number of publications. The dashed line shows a power law scaling behavior with exponent 1.45 ± 0.01. (C) Node out-strength against its out-degree for city citation network. There are two distinct power law scaling regions, with scaling exponents 0.82 ± 0.04 and 2.26 ± 0.07 for low and high degree (> 200) nodes, respectively.

Estimation of significance difference. The above bootstrapping procedure however does not tell whether the difference in the means of two distributions is significant or not. In this case the re-sampling has to be performed according to an appropriate null hypothesis, whereas for standard errors the re-sampling procedure was unrestricted.

Let us consider two independent samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$, and suppose that we are interested in the difference in the population means, $\delta = \bar{x} - \bar{y}$. Consider that the null hypothesis is $H_0 : \bar{x} - \bar{y} = 0$. We create the bootstrap sample by choosing n elements without replacement from the pooled set $x_1, \ldots, x_n, y_1, \ldots, y_m$. The remaining m elements constitute the other sample. We then calculate the mean of both these samples and determine the difference between them, say $\hat{\delta}_i = \bar{\hat{x}}_i - \bar{\hat{y}}_i$. In analogous fashion N re-samples are made, and the bootstrap p value is defined as $p = \frac{(\#(\hat{\delta}_i \geq \delta,\ \forall i))+1}{N+1}$. In this study we have used $10^4 - 1$ bootstrapped samples.

Power-law exponent. We use maximum likelihood techniques to estimate the scaling exponent of power law distributions [2].

Regression Coefficient. We used the linear regression analysis to study the relationship between the corresponding variables. We determine the regression coefficient using the ordinary least squares. The error term of the regression coefficient represents the standard error of the estimate.

F. Map construction

Statistical data with embedded geographical information can be visualized with standard maps which are color coded by region. However these maps are sometimes hard to interpret as the statistical measures are often correlated with the other indicators. We have used a diffusion-based method to create different density-equalizing maps [3]. In this method we start with an inhomogeneous distribution of the research contribution (in terms of citations, say) and let the diffusion process evolve until a homogeneous equilibrium state is reached: the displacements are then reinterpreted to generate the cartogram.

II. RESULTS

We consider the research contribution of each country in terms of the number of publications $N_{\rm Pub}$, normalized by the number of participating countries in that publication. To visualize the results, we create a cartogram in which the geographic regions are deformed and rescaled in proportion to their relative research contribution [3]. We observed that the contribution of different countries in terms of publications is heterogeneous and varies over 6 order of magnitude. Fig. S1A shows that North America (32.4%), Europe(33.7%) and Asia(27.4%) have prominent contribution in terms of the number of publications. On the other hand, Africa, South America and Oceania contribute less than 7% of world's publications. Table. S1 shows the contribution, number of countries and cities in each continents. It also indicates the statistics of the top countries of each continent. It is evident that the United States are the leading country in the world both in terms of publications and citations to them. It is followed by China, United Kingdom, Japan, and Germany in terms of publications, whereas in terms of citations it is followed by United Kingdom, Germany, Japan, and China. We indicate the fraction of total publications $f_{\rm Pub}$, the fraction of total citations received $f_{\rm Cite}$ and the average number of citations per paper, for countries that received more than 0.005% of world citations. Countries are listed in decreasing order of the fraction of total citations received. The superscripts in $f_{\rm Pub}$ and

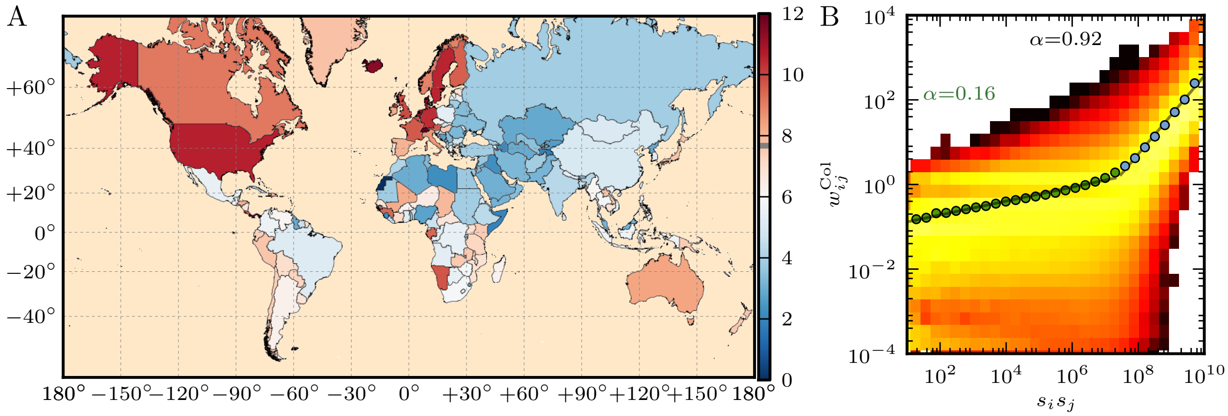

Figure S2. (A) Average number of citations of each country. World map where countries are color coded based on the average number of citations per publication. Most countries stay below the world's average of 7.67.(B) Weight of the links against the product of the strengths of the endpoints in the collaboration network of cities. There are two different scaling regions (2×10^7), with exponents 0.16 ± 0.01 and 0.92 ± 0.03.

f_{Cite} indicate the world ranking of that country according to the numbers of publications and citations, respectively. We then consider the contribution in terms of the number of publications at the level of cities. In Fig. S1B we plot the probability distribution of the cities' contributions in terms of their publications and observed that it follows a power law scaling behavior with exponent 1.45 ± 0.01. By plotting the out-degree against the out-strength, we find that there is power law scaling behavior with $\langle s^{\mathrm{out}} \rangle (k^{\mathrm{out}}) \propto (k^{\mathrm{out}})^{\alpha}$ (Fig. S1C). However, there are two distinct scaling regimes: for nodes with small k_i^{in} (< 200) the exponent is $\alpha = 0.82 \pm 04$, while for large k_i^{out} (≥ 200) the exponent is $\alpha = 2.26 \pm 0.07$. The super-linear behavior suggests that stronger links are more frequently connected to high out-degree nodes.

Next we consider the average number of citations per paper of each country and plot it on a colorpleth map (Fig. S2A). For calculating the average citation of a country we consider all its publications and count the total number of citations to all these articles during the period of 2003-2010. In the case where a publication has multiple affiliations from different countries, it is counted multiple times for the countries' averages, once for each of the affiliated countries. In Table S1, we have also given the average number of citations per paper of the top countries in each continent. The world average is 7.67. United States, Canada, Australia and most of the European countries have average number of citations larger than the world average. In Europe Switzerland leads the table, followed by Denmark and Netherlands. In contrast most of the countries from Asia stay below the world average, the only exception being Israel. Most of the countries in Africa and South America are below the world average as well. Other notable countries are Bermuda (16.97±5.95), Gambia (16.17±3.10), Panama (12.41± 0.68), Iceland (11.43± 0.71), Seychelles (11.11± 2.40), Guinea-Bissau (10.10± 0.97), Costa Rica (9.82± 0.93), and Austria (9.75± 0.09). For the collaboration network of cities we plot the weight of the links against the product of the strengths of the connecting nodes, expressing the expected weight of random collaborations (Fig. S2B). As for citations we find that $w_{ij}^{\mathrm{Col}} \propto (s_i s_j)^{\alpha}$, with two different scaling exponents. If $s_i s_j < 2 \times 10^7$, $\alpha = 0.16 \pm 0.01$ $(R = 0.11)$, whereas if $s_i s_j > 2 \times 10^7$ $\alpha = 0.92 \pm 0.03$ $(R = 1.18)$.

In Fig. S3A,B we plot the probability of existence of a link as a function of the product of strength of the end-points of the link. We found that as the product increases, both in the collaboration and the citation network the probability of link existence increases, as expected. In Fig. S3C,D we show the variation of the link weight against the distance between the end-points. We found that both in the collaboration and the citation network on the average the link weight decreases as a power-law with exponent 0.31 ± 0.01 and 0.22 ± 0.01, respectively. In this figure, while calculating the averages we have only considered the existing links between nodes. However, in the main text we have seen that the probability of link existence also decreases with distance. If we take this information while calculating the averages, i.e., we consider the non-existent links by assigning weight zero to them, we found that in both the collaboration and the citation network, the average link weight decreases with distance as a power law, with exponent 0.88 ± 0.01 and 0.51 ± 0.01, respectively (Fig. S3E,F). Note that this property is different from what has been observed in the mobile phone communication network, where it was shown that the weight of the existing links are independent of the distance, whereas the overall link weight decrease as a result of decreasing probability of having a link as the distance increases [4].

In the main paper we have considered the research performance of each country based on the number of citations. In addition, here we consider the performance of a country based on its number of publications. As before, in Fig. S4A, we plot the research contribution in terms of

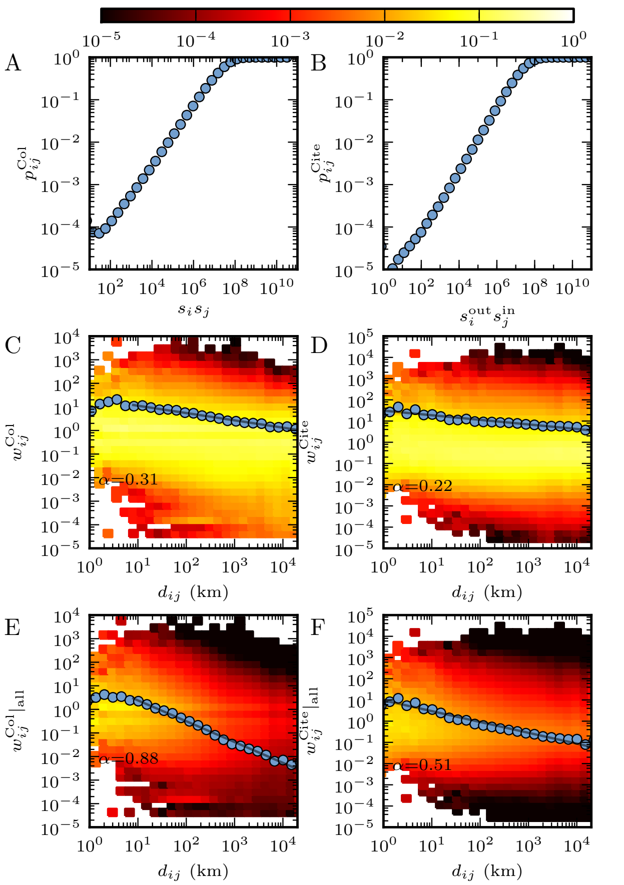

Figure S3. Gravity law in the world collaboration and citation networks.(A) Variation of the probability of existence of a link between two nodes as a function of the product of their strengths in the (A) collaboration network and (B) citation network of cities. Variation of the average link weight against the distance between the cities in the (C) collaboration network and (D) citation network. For each distance the average ratio is also shown. In this case only the existing links are considered while calculating the averages. The solid line indicates a power law behavior with exponent $\alpha = 0.31 \pm 0.01$ and 0.22 ± 0.01 respectively. Variation of the average link weight against the distance between the cities in the (E) collaboration network and (F) citation network. For each distance the average ratio is also shown. In this case all possible node pairs are considered in order to calculate the average, i.e., links that do not exist are considered with weight 0. The solid line indicates a power law behavior with exponent $\alpha = 0.88 \pm 0.02$ and 0.51 ± 0.02, respectively.

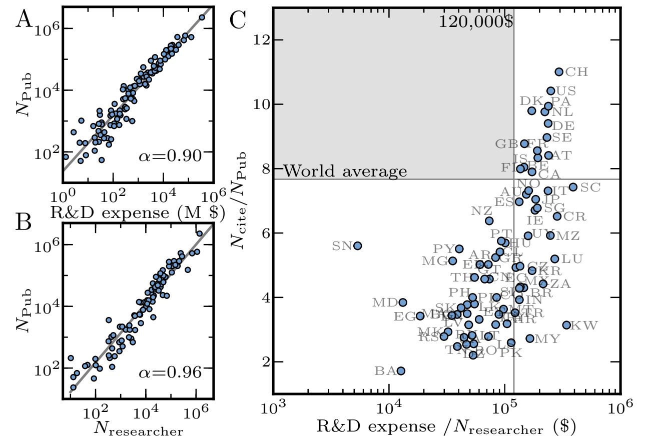

Figure S4. Relation between research contribution in terms of number of publications and funding. Country's number of publications against the (A) expenditure in research and development (in million dollars, and purchasing power parity), (B) number of researchers in that country. The solid line indicates a scaling with exponent 0.90 ± 0.03 and 0.96 ± 0.03, respectively. (C) The plot of average spending per researcher against the average number of citation per paper of that country. The average number of citations is now defined as the ratio of the normalized number of citations and normalized number of publications (see text). The horizontal line indicates world average, the vertical line indicates the spending of 120 000 \$ per researcher.

the number of publications $N_{\rm Pub}$ against the countries' R&D expenditure in terms of purchasing power parity (PPP). We found that this indicator also scale almost linearly with the spending. We next consider the dependence of research performance on the number of researchers in that country (Fig. S4B). The research contribution in terms of publications also scale linearly with the number of researchers in that county.

Finally as a measure of the average publication quality of a country we consider the ratio of the normalized number of citations and normalized number of publications of that country. This is an alternative measure of the average number of citations per paper we mentioned above, which is not normalized by the number of authors in a paper. In the previous measure each publication from a country (independently of the number of participating countries) gets equal weight while calculating the average. In this other measure, if there are n countries in a publication, each country would get $1/n$ as credit for that publication, so that publication would give a lower contribution to the average number of cites per paper than before. In Fig. S4C we plot the new quantity against the average spending per researcher of the country (R&D expenditure divided by the number of researchers). Although this plot is similar to the one in the main paper, there are certain differences in the average number of citations of some countries. For example, Italy, Spain, Norway are now below the world average. This means that the publications from these countries with international collaborators contribute significantly to the average impact of their scientific production.

In order to check whether the countries in Fig. 5C can be categorized into different groups based on the average spending per researcher and the average number of citations, we used two different clustering methods. The k-means clustering technique [5] partitions the data into k-mutually exclusive clusters. The aim here is to determine whether there are inherent clusters in Fig.5C and Fig S4C. For the k-means clustering method we need to

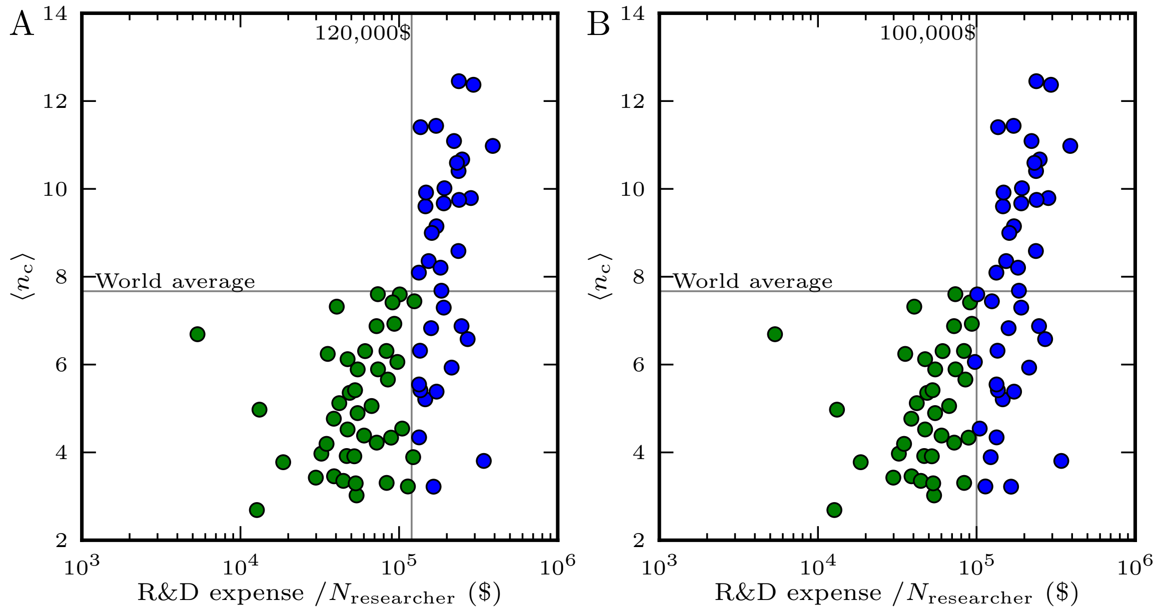

Figure S5. Data clustering. (A) Decomposition obtained using k-mean clustering with $k = 2$. (B) Decomposition obtained using mean shift clustering. Each cluster is indicated by a color.

specify the number k of clusters before starting the clustering process. The method consists in the minimization of an objective function expressing the sum of square distances between each data point and its *centroid*, i.e. a geometrical point whose position is also consistently determined by the minimization procedure: each centroid corresponds to one cluster. We can follow a procedure to minimize the objective function iteratively by finding a new set of cluster centroids that can lower the value of the objective function at each iteration. On using this method with $k = 2$, we found that the countries can be classified into two groups, one with average spending less than about 120,000 $ per researcher per year and other with average spending more than about 120,000 $ (Fig. S5). We also use a different method, the mean shift clustering algorithm [6] to determine the clusters in the data in Fig.5. This is a nonparametric clustering technique and does not require prior knowledge of the number of clusters. The mean-shift algorithm seeks local maxima of density of points in the feature space. This method also detects two different clusters, one with average spending less than about 100,000 $ per researcher per year and the other with average spending more than about 100,000 $. Thus, these two methods give slightly different thresholds however the results are qualitatively similar.

[1] B. Efron and R. Tibshirani, *An introduction to the bootstrap* (Chapman & Hall, New York, 1993).

[2] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Review **51**, 661 (2009).

[3] M. Gastner and M. Newman, Proc. Natl. Acad. Sci. U.S.A. **101**, 7499 (2004).

[4] G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel, J. Stat. Mech. **2009**, L07003 (2009).

[5] D. Sculley, in *Proceedings of the 19th international conference on World wide web*, WWW '10 (ACM, New York, NY, USA, 2010) pp. 1177–1178.

[6] D. Comaniciu and P. Meer, Pattern Analysis and Machine Intelligence, IEEE Transactions on **24**, 603 (2002).

Table S1: Research contribution of different continents and their top countries. The number of countries and cities in each continent are indicated by $N_{\text{Countries}}$ and N_{Cities}, respectively. Fraction of publications f_{Pub}, fraction of citations received f_{Cite} and the average number of citations per paper of each continent is also indicated. For top countries in each continent we list the fraction of publications f_{Pub}, fraction of citations received f_{Cite}, the average number of citations per paper. The superscript indicates the countries' rank in the world in terms of number of publications and number of citations. Only countries that receive more than 0.005% of all citations are shown.

Continent	$N_{\text{Countries}}$	N_{Cities}	f_{Pub} (in %)	f_{Cite} (in %)	Avg. Cites	Country name	f_{Pub} (in %)	f_{Cite} (in %)	Avg. Cites
Africa	57	749	1.32	0.65	5.00±0.05	South Africa	0.430^{33}	0.248^{37}	5.92±0.08
						Egypt	0.286^{38}	0.128^{40}	3.78±0.05
						Tunisia	0.100^{52}	0.036^{52}	3.33±0.13
						Nigeria	0.126^{50}	0.031^{56}	2.82±0.25
						Kenya	0.038^{65}	0.028^{58}	7.55±0.29
						Morocco	0.055^{60}	0.025^{60}	4.20±0.10
						Algeria	0.067^{54}	0.023^{62}	3.01±0.08
						Tanzania	0.019^{83}	0.014^{74}	7.27±0.29
						Uganda	0.018^{85}	0.014^{75}	7.04±0.27
						Cameroon	0.021^{77}	0.010^{80}	4.72±0.18
						Ethiopia	0.022^{75}	0.009^{81}	4.36±0.17
						Ghana	0.016^{86}	0.008^{85}	5.29±0.21
						Zimbabwe	0.011^{95}	0.007^{87}	5.92±0.28
						Malawi	0.009^{103}	0.006^{88}	7.11±0.32
						Senegal	0.008^{104}	0.006^{90}	6.72±0.29
						Botswana	0.010^{97}	0.005^{95}	5.46±0.54
						Gambia	0.003^{128}	0.005^{96}	15.87±3.02
						Cote d'Ivoire	0.006^{107}	0.005^{97}	7.20±0.41
Asia	49	3853	27.36	17.71	5.58±0.01	Japan	6.457^{4}	5.939^{4}	7.68±0.03
						China	7.216^{2}	4.304^{5}	5.05±0.02
						South Korea	2.509^{10}	1.582^{13}	5.38±0.04
						India	2.727^{9}	1.398^{15}	4.35±0.03
						Taiwan	1.671^{15}	1.037^{16}	5.19±0.04
						Israel	0.863^{22}	0.837^{20}	8.86±0.10
						Turkey	1.450^{17}	0.667^{22}	3.89±0.03
						Russia	1.875^{13}	0.622^{24}	3.92±0.05
						Singapore	0.521^{29}	0.461^{28}	7.29±0.08
						Iran	0.747^{23}	0.308^{31}	3.31±0.03
						Thailand	0.244^{41}	0.147^{38}	5.88±0.17
						Malaysia	0.195^{42}	0.069^{45}	3.22±0.07
						Pakistan	0.175^{44}	0.059^{48}	3.23±0.07
						Saudi Arabia	0.131^{49}	0.046^{50}	3.12±0.07
						Jordan	0.061^{58}	0.023^{61}	3.27±0.11
						Vietnam	0.037^{68}	0.020^{66}	5.59±0.30
						Indonesia	0.032^{70}	0.019^{67}	5.73±0.23
						Kuwait	0.044^{61}	0.018^{68}	3.80±0.16
						Bangladesh	0.041^{63}	0.018^{69}	4.74±0.17
						Lebanon	0.038^{66}	0.018^{70}	4.47±0.13
						UAE	0.041^{62}	0.018^{71}	4.03±0.14
						Philippines	0.035^{69}	0.017^{72}	6.20±0.30
						Cyprus	0.027^{73}	0.012^{76}	4.31±0.15
						Sri Lanka	0.022^{76}	0.011^{79}	4.90±0.18
						Armenia	0.026^{74}	0.009^{82}	6.18±0.31
						Oman	0.021^{79}	0.008^{86}	3.50±0.14
						Georgia	0.020^{82}	0.006^{89}	3.94±0.20
						Nepal	0.012^{92}	0.006^{91}	5.36±0.29
						Uzbekistan	0.020^{81}	0.006^{92}	3.50±0.17

Continued on next page

Table S1 – *Continued from previous page*

Continent	$N_{\mathrm{Countries}}$	N_{Cities}	f_{Pub} (in %)	f_{Cite} (in %)	Avg. Cites	Country name	f_{Pub} (in %)	f_{Cite} (in %)	Avg. Cites
Europe	47	6625	33.69	35.25	9.29±0.01	United Kingdom	6.509^{3}	7.453^{2}	9.91±0.04
						Germany	5.131^{5}	6.299^{3}	10.41±0.04
						France	3.611^{7}	4.034^{6}	9.67±0.04
						Italy	3.415^{8}	3.258^{8}	8.59±0.04
						Netherlands	1.829^{14}	2.331^{9}	11.08±0.07
						Spain	2.482^{11}	2.258^{11}	8.09±0.05
						Switzerland	1.114^{19}	1.600^{12}	12.38±0.09
						Sweden	1.227^{18}	1.436^{14}	10.59±0.09
						Belgium	0.923^{21}	1.004^{17}	10.02±0.08
						Denmark	0.655^{25}	0.838^{19}	11.45±0.12
						Finland	0.640^{26}	0.672^{21}	9.59±0.10
						Austria	0.595^{27}	0.653^{23}	9.75±0.11
						Poland	1.110^{20}	0.579^{25}	5.43±0.05
						Norway	0.506^{30}	0.483^{26}	8.98±0.10
						Greece	0.684^{24}	0.468^{27}	6.30±0.06
						Portugal	0.460^{32}	0.345^{30}	6.93±0.08
						Czech Republic	0.464^{31}	0.301^{32}	6.31±0.08
						Ireland	0.340^{36}	0.298^{33}	8.20±0.16
						Hungary	0.338^{37}	0.251^{36}	7.61±0.13
						Slovenia	0.171^{45}	0.096^{41}	5.41±0.09
						Ukraine	0.271^{40}	0.088^{42}	3.46±0.07
						Romania	0.274^{39}	0.079^{43}	3.30±0.09
						Slovakia	0.150^{48}	0.072^{44}	5.12±0.12
						Croatia	0.164^{47}	0.068^{46}	4.53±0.12
						Serbia	0.167^{46}	0.061^{47}	3.42±0.07
						Bulgaria	0.125^{51}	0.056^{49}	4.76±0.09
						Estonia	0.063^{57}	0.041^{51}	6.91±0.21
						Lithuania	0.095^{53}	0.035^{53}	3.91±0.13
						Iceland	0.030^{71}	0.031^{57}	11.46±0.63
						Belarus	0.064^{56}	0.020^{65}	3.37±0.10
						Latvia	0.021^{78}	0.009^{83}	5.34±0.43
						Luxembourg	0.012^{91}	0.008^{84}	6.58±0.34
						Moldova	0.011^{96}	0.006^{93}	4.94±0.31
North America	37	5346	32.40	42.33	10.36±0.02	United States	28.116^{1}	38.216^{1}	10.67±0.02
						Canada	3.616^{6}	3.728^{7}	9.15±0.05
						Mexico	0.523^{28}	0.292^{34}	5.57±0.10
						Puerto Rico	0.037^{67}	0.028^{59}	7.66±0.26
						Cuba	0.040^{64}	0.022^{63}	4.81±0.14
						Costa Rica	0.014^{88}	0.012^{77}	9.93±0.87
						Panama	0.009^{102}	0.012^{78}	12.43±0.81
Oceania	21	844	2.89	2.67	8.22±0.05	Australia	2.448^{12}	2.301^{10}	8.36±0.05
						New Zealand	0.425^{34}	0.354^{29}	7.60±0.10
South America	14	782	2.34	1.39	5.75±0.04	Brazil	1.551^{16}	0.871^{18}	5.21±0.04
						Argentina	0.399^{35}	0.261^{35}	6.31±0.10
						Chile	0.193^{43}	0.136^{39}	7.42±0.16
						Colombia	0.066^{55}	0.034^{54}	5.65±0.19
						Venezuela	0.060^{59}	0.034^{55}	6.11±0.28
						Uruguay	0.027^{72}	0.021^{64}	6.81±0.21
						Peru	0.019^{84}	0.014^{73}	7.68±0.30
						Ecuador	0.008^{105}	0.005^{94}	7.39±0.37