SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
COMPUTATIONAL SCIENCE
APPLIED PHYSICS
STATISTICS
APPLIED MATHEMATICS

Received
13 September 2012

Accepted
14 November 2012

Published
29 November 2012

Correspondence and requests for materials should be addressed to S.F. (santo.fortunato@aalto.fi)

# World citation and collaboration networks: uncovering the role of geography in science

Raj Kumar Pan[1], Kimmo Kaski[1] & 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.

**Modern information and communication technologies, especially the Internet, have diminished the role of spatial distances and territorial boundaries on the access and transmissibility of information. This has enabled scientists for closer collaboration and internationalization. Nevertheless, geography remains an important factor affecting the dynamics of science. Here we present a systematic analysis of citation and collaboration networks between cities and countries, by assigning papers to the geographic locations of their authors' affiliations. The citation flows as well as the collaboration strengths between cities decrease with the distance between them and follow gravity laws. In addition, the total research impact of a country grows linearly with the amount of national funding for research & development. However, the average impact reveals a peculiar threshold effect: the scientific output of a country may reach an impact larger than the world average only if the country invests more than about 100,000 USD per researcher annually.**

The strength of most interactions in nature typically decreases with the distance between objects or constituents. The most famous example is Newton's gravitational force, which is known to decay with the square of the distance between the masses. This principle holds also outside the realm of physical processes. Recent studies on mobile phone communication networks[1,2] and blogs[3] have revealed that the probability for a social tie to occur between agents decays with a power of their distance.

Likewise, scientific interactions are likely to take place between scholars localized in the same or nearby areas. Scientists tend to cluster in space, since the elaboration and progress of a project requires frequent discussions between collaborators that is hardly possible if they live far apart. Factors based on cultural, linguistic and institutional differences cause additional obstacles to long-distance cooperation[4]. Further, research funding is mostly allocated at the national level[5], thus favoring regional over international collaborations.

Nowadays, the Internet and the greater affordability of international transportation have enormously reduced distances between people, overcoming both geographic and cultural barriers[6–8]. This in turn has made scientific collaborations between distant scholars far easier than before[9–14]. Nevertheless, the role of geography in the creation and recognition of scientific output is not yet fully known. For example, How do scientific interactions depend on distance? Is collaboration concentrated within the perimeter of a university, of a city or of a country, as it used to be in the past, or has it become truly international, possibly due to the modern information and communication technologies?

Multi-authored collaborations serve as big opportunity for science[15], as one can integrate a wide range of competence and skill, to attack difficult problems, with an enhanced chance of success. Indeed, the last decades have witnessed the formation of larger and larger research teams[16,17]. In particular, multi-university collaborations have been growing at a fast pace and are more likely to lead to high impact publications[18], especially if they involve different countries[19,20]. On the other hand, there is also evidence of decreasing returns from large team size, likely from management inefficiencies, which limits the productivity arising from collaboration[21].

Geographic proximity is also likely to affect the process of giving and receiving credits for someone's work, expressed by paper citations. For most papers one expects to find a decaying probability of citation with distance, as new findings are typically more visible in the area where the authors operate. This is confirmed by a recent study[22]. In addition, collaboration patterns are likely to influence and be influenced by citations. While collaborating, scholars become more familiar with the scientific output of their co-authors, which then has a higher chance to be cited in the future. In turn, scholars citing frequently each other's work have strongly overlapping research interests, and are more likely to become co-authors sooner or later. Therefore citations and collaborations between distinct locations are likely to be correlated. However, it is crucial to assess how collaborative patterns affect citation flows, to be able to disentangle the actual impact of a publication (and, therefore, its merit) from credits coming through social networking. A geographic analysis of citation flows between cities is also useful to understand how quickly a new result gets recognized by the scientific community in different geographical areas, which may help to uncover how new scientific paradigms spread and get established[23].

Knowing how scientific interactions vary with distance is also valuable for practical reasons. To scholars, it might suggest how to choose collaborators in order to optimize the impact and visibility of their research. To institutions and governments, it might advice suitable allocations of funds for regional and international projects, in order to improve the scientific outcome for a given amount of resources. It is then not surprising that spatial scientometrics has acquired a prominent role during the last few years. There are a number of studies carried out exploiting the enhanced availability of citation data[24]. Yet there are other factors, namely funding, that also plays a crucial role in the development of a research project, as it not only contribute towards the direct and overhead costs of the research but also facilitates the cooperation and collaboration among researchers working in different locations and different fields[25]. Since both public and industrial resources are used to fund academic research, it is also natural to question the result and impact obtained with these resources[26,27].

We have performed the first comprehensive study of citation and collaborative interactions between different geographic locations. We used one of the world's largest citation databases to derive the citation and the collaboration network, i.e. weighted networks where nodes are cities and links are citations and collaborations between the corresponding cities (see Methods). The analysis of these networks[28–31] discloses the existence of gravity laws as well as non-trivial correlation between collaborations and citations. Finally, we explore the issue of the importance of funding to research and development in promoting high quality science, by studying the relationship between national expenditure, the number of publications and their impact in terms of number of citations for different countries.

## Results

The research contribution of each country in terms of the (normalized) number of citations received $N_{\mathrm{Cite}}$ is illustrated in the world map of Fig. 1A. Colored maps can be misleading as the value assigned to a large area gives an impression of a much greater impact of that color in the visualization. We thus created a cartogram, in which the geographic regions are deformed and rescaled in proportion to their relative research contribution[32]. The citation strengths of countries span over seven orders of magnitude. North America and Europe receive 42.3% and 35.3% of world's citations, respectively. In contrast, the contribution by Asia amounts to only 17.7% of world's citations while the total contribution of Africa, South America and Oceania is lower than 5%. In this ranking the United States is the leading country followed by the United Kingdom, Germany, Japan, and China. The corresponding world map in terms of countries' number of (normalized) publications is shown in the Supplementary Fig. S1 online. This heterogeneity suggests that a small number of countries have a substantial contribution to research while the rest has a negligible contribution. In Fig. S2 online we report the results for the average number of citations of each country.

In order to find out the quality of papers published by different countries we consider the number of citations of each of the papers written by that country. In Fig. 1B we plot the probability distribution of the number of citations of papers in the largest 20 countries. A paper is associated to a country if at least one of its affiliations is from that country. All these distributions are broad and vary over four orders of magnitude. When each distribution is rescaled by the average number of citations of papers of the respective country, all curves nicely collapse (Fig. 1C). This result suggests that the functional form of the citation distribution is the same in each country and that the difference between countries can be effectively summarized by the average number of citations. This type of universality holds at the level of scientific disciplines as well[33].

Next we consider the contribution at the level of cities. In Fig. 1D we plot the probability distribution of the cities' citations. The distribution is broad, spanning over five orders of magnitude, and it follows a power law decay with exponent 1.46 ± 0.03. This suggests a relationship with the population of the city, as the city size distribution obeys the Zipf law[34,35], i.e. decays as a power law (with exponent 2). The observed power law scaling relation might suggest a self-organization phenomena due to the agglomeration benefits in science. These advantages can be due to the ease of collaboration between groups working in similar fields, sharing of infrastructure and support, etc., which leads to efficient integration and transfer of information.

We now consider the weighted citation network between cities, where the nodes are the cities that are connected by weighted and directed links, indicating publications of one city citing publications of the others. The network has 18,199 nodes and 9,494,021 links including 14,447 self-links (i.e., citations within the same city). In Fig. 1D we plot the cumulative distribution of the weights of self-links and links between different nodes. Both these distributions are broad; however, the weights of self-links are more heterogeneous, revealing a bias towards self-citations. Next we calculate the number of incoming links, i.e., the in-degree $k_i^{\mathrm{in}}$ of each node $i$ and its in-strength, $s_i^{\mathrm{in}} = \sum_j w_{ji}^{\mathrm{Cite}}$, which equals the number $N_i^{\mathrm{Cite}}$ of (normalized) citations received. By plotting the in-degree against the in-strength, we find that there is a power law scaling behavior with $\langle s^{\mathrm{in}} \rangle (k^{\mathrm{in}}) \propto (k^{\mathrm{in}})^{\alpha}$ (Fig. 1E). However, there are two distinct scaling regimes: for nodes with small $k_i^{\mathrm{in}}$ ($< 200$) the exponent is $\alpha = 0.91 \pm 0.03$ (regression coefficient ± standard error of the estimate $R = 0.95 \pm 0.01$), while for large $k_i^{\mathrm{in}}$ ($\geq 200$) the exponent is $\alpha = 2.20 \pm 0.08$ ($R = 2.01 \pm 0.01$). The super-linear behavior suggests that stronger links are more frequently connected to high in-degree nodes. The out-strength of the nodes follows a similar relationship with the outdegree of the nodes (see Supplementary Fig. S1 online). Finally, we plot the weights of the links $w_{ij}^{\mathrm{Cite}}$ against the product of the node strength $s_i^{\mathrm{out}} s_j^{\mathrm{in}}$. The product $s_i^{\mathrm{out}} s_j^{\mathrm{in}}$ gives the weight of a link that is expected to occur by chance between $i$ and $j$ if all the papers would be citing each other at random. Even in this case there are two distinct scaling regions, $w_{ij}^{\mathrm{Cite}} \propto \left( s_i^{\mathrm{out}} s_j^{\mathrm{in}} \right)^{\alpha}$, where $\alpha = 0.13 \pm 0.01$ ($R = 0.19 \pm 0.0003$) if the product is less than $2 \times 10^7$, while for larger values of the product $\alpha = 0.99 \pm 0.01$ ($R = 1.07 \pm 0.001$). This suggests that the observed citation is as expected between high strength nodes, while it is much lower in case of cities with low strength.

Let us now consider the collaboration network at the city level, where the nodes are cities and weighted undirected links indicate the presence and frequency of collaborations between scholars of different cities. There are 18,199 nodes in the network and 1,256,718 undirected links including 14,954 self-links. The weight of the self-links indicates the amount of internal collaboration. The degree of a node $i$ indicates the number of other cities with which $i$ collaborates and its strength is indicative of, but not coincident with, the number of papers written by scholars of institutions in that city.

In Fig. 2A we plot the cumulative probability distribution of link weights. As for citations, the weights of self-links are more broadly distributed than the weights of the links between different cities, showing that scholars of a city collaborate more frequently with each other than with colleagues from any other city. The distributions of collaboration and citation streams between cities differ from their analogues in mobile phone communications and world trade, that show log-normal distributions[2,36]. Next, we consider the fraction of

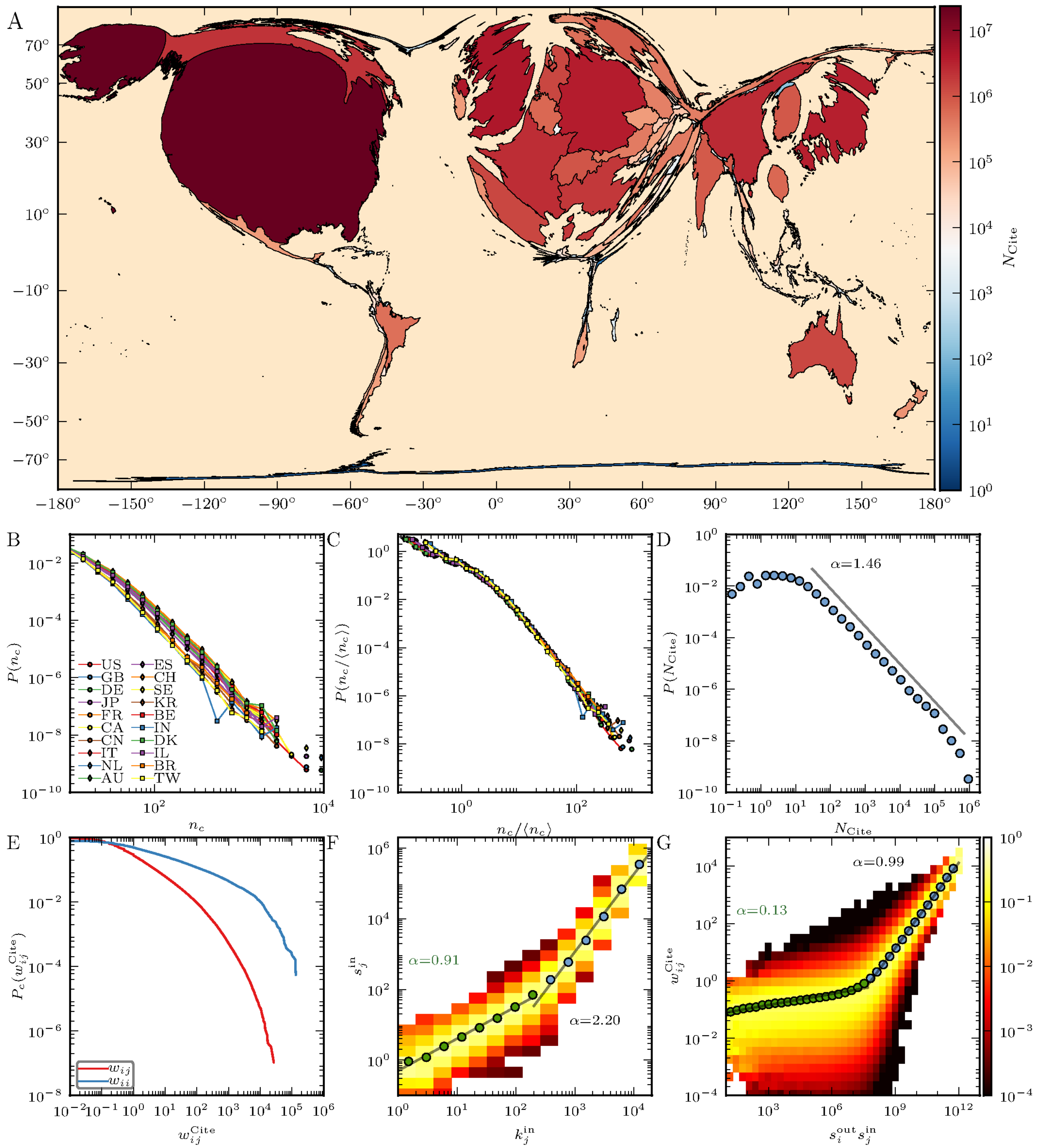

**Figure 1** | **Properties of the world citation network.** (A) Citation map of the world where the area of each country is scaled and deformed according to the number of citations received, which is also represented by the color of each country. (B) Citation distribution of papers of top 20 countries. If a paper is written by authors from multiple countries, the paper contributes to each country. (C) When the distributions in (B) are normalized by the average number of citations of each country, they fall on top of each other. (D) Probability distribution function of the number of citations received by each city. (E) Cumulative distribution function of the link weights $w_{ij}$ (excluding self-links) and self-links $w_{ii}$ in the citation network of cities. (F) Node in-strength against its in-degree for the city citation network. (G) Link weight against the product of the strengths of the connected nodes in the city citation network. For each plot we show the corresponding best-fit lines and power law exponents.

internal collaboration by calculating the ratio of the weight of the self-link to the strength of the node. By plotting $w_{ii}^{\rm Col}/s_i$ against the strength of the node $s_i$, we see that the ratio increases with $s_i$, indicating that as the city size increases most of its collaborations take place within the city (Fig. 2B). However, for small cities most of their papers are written with external collaborators. The node degree

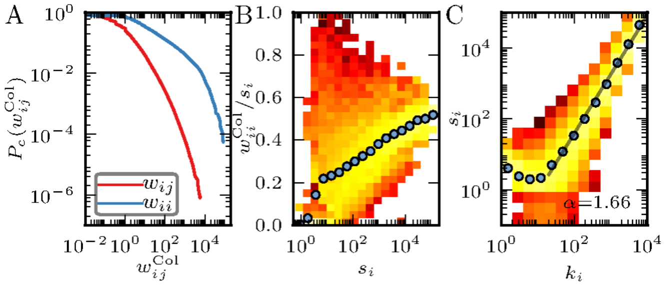

**Figure 2 | Properties of the world collaboration network.** (A) Cumulative probability distribution of the link weights in the collaboration network of cities. Self-links are shown separately. (B) Fraction of internal collaboration, indicated by the ratio of the weight $w_{ii}^{\mathrm{Col}}$ of the self-link and strength $s_i$ of a node, against $s_i$. (C) Strength of a node against its degree. The straight line indicates a power law behavior with exponent 1.66 ± 0.04. In these plots we use the same colorbar as in Fig. 1.

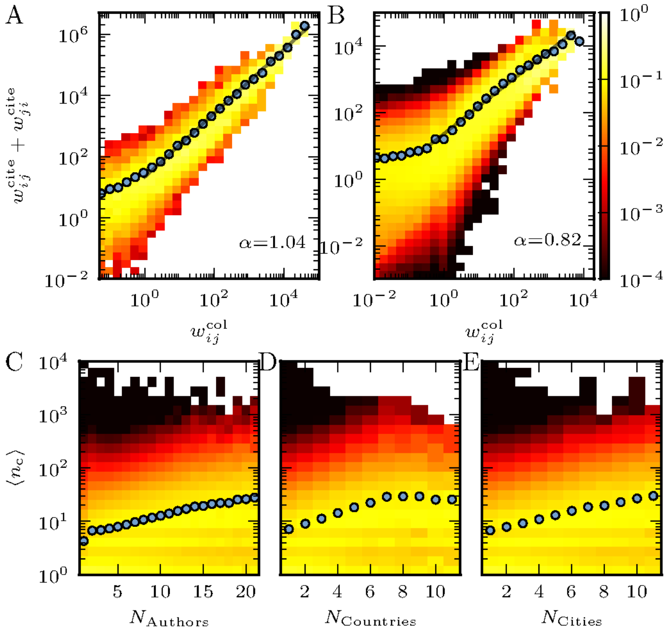

**Figure 3 | Correlation between the world citation and collaboration networks.** Weight of the links in the citation network against the corresponding links in the collaboration network at the (A) country level and (B) city level network. Power law scaling is shown by solid lines with exponents 1.04 ± 0.01 and 0.82 ± 0.02, respectively. Density plot of the number of citations of a publication against the number of (C) co-authors, (D) countries (E) cities in the affiliation. The circles indicate the average trend.

scales with its strength as $\langle s \rangle(k) \propto k^{\alpha}$, where $\alpha = 1.66 \pm 0.04$ ($R = 1.65 \pm 0.01$) (Fig. 2C). This super-linear scaling suggests that higher degree nodes are more frequently connected by stronger links.

Let us explore the relationship between the citation and the collaboration networks at both the country and the city level. At the country level the collaboration network comprises 226 nodes and 10,308 undirected links, including 219 self-links. In the citation network there are also 226 nodes but 28,869 directed links, including 215 self-links. In Fig. 3, we plot the weight of links of the collaboration network, $w_{ij}^{\mathrm{Col}}$ against the weight of the same links in the citation network, $w_{ij}^{\mathrm{Cite}} + w_{ji}^{\mathrm{Cite}}$. We find scaling $w_{ij}^{\mathrm{Col}} \propto \left(w_{ij}^{\mathrm{Col}} + w_{ji}^{\mathrm{Col}}\right)^{\alpha}$ where $\alpha = 1.04 \pm 0.01$ ($R = 1.08 \pm 0.008$) for countries (Fig. 3A), and $\alpha = 0.82 \pm 0.02$ ($R = 1.05 \pm 0.002$) for cities (Fig. 3B), i.e. the increase in collaboration is linearly related to the amount of citations exchanged between the two countries/cities.

We now consider the dependence of the number of citations of a paper on the number of coauthors of that paper and on the number of affiliations of its coauthors. It has been previously shown that papers published by teams often get more citations than single author papers[17,18]. Our results also show that the average number of cites of a publication increases with the number of co-authors of that publication (Fig. 3C). Furthermore, the average number of citations of a publication increases with the number of affiliated countries and cities of its authors (Fig. 3D and E). In order to separate the effect of the number of coauthors and different type of collaboration (internal, domestic and international) we grouped each paper based on its affiliations and number of coauthors. In Table 1, we consider papers with a given number of authors and categorize them according to whether all the affiliations listed in the paper are from a single city, from multiple cities in a single country or from different countries. For an equal number of authors, publications having multiple international affiliations get a statistically significant increment ($p < 10^{-4}$) in the number of citations with respect to publications with only domestic affiliations. Thus, crossing territorial boundaries also pays off in terms of scientific impact. In contrast, multiple domestic affiliations do not positively effect the number of citations when the number of authors in a publication is less than 6.

Next we consider the effect of geographical proximity on the citation and collaboration networks by determining the geographic location (latitude and longitude) of each place in the dataset[37] (see Methods). We found that the probability that there is a link between two cities in the collaboration network decreases as a power law as the distance between the two cities increases (Fig. 4A). The power law exponent is 0.57 ± 0.01. Our results are different from those obtained in Ref. 38, where it was found that the distribution of distances between co-authors decreases exponentially. Such difference might be due to the limited dataset used in Ref. 38, which included only papers published before 1990, and possibly also due to the recent advances in communication and transportation technologies.

Many spatially embedded networks have been observed to follow gravity laws[37], where the flow between two locations follows

$$T_{ij} \propto \frac{P_i P_j}{d_{ij}^{\alpha}}. \tag{1}$$

Here, $T_{ij}$ is the flow between nodes $i$ and $j$, $P_i$ and $P_j$ are the populations of nodes $i$ and $j$, respectively and $d_{ij}$ is the geodesic distance between $i$ and $j$, the value of exponent $\alpha$ being dependent of the system. For the collaboration network Eq. 1 becomes

$$w_{ij}^{\mathrm{Col}} \propto \frac{s_i s_j}{d_{ij}^{\alpha}}. \tag{2}$$

In Fig. 4B, we plot the ratio $w_{ij}^{\mathrm{Col}} \big/ (s_i s_j)$ against the distance $d_{ij}$ between all node pairs. We found that as the distance increases $\left\langle w_{ij}^{\mathrm{Col}} \big/ (s_i s_j) \right\rangle$ decreases as a power law with the exponent $\alpha = 1.16 \pm 0.03$ ($R = -0.97 \pm 0.002$), except at very short distances. As we have seen before, collaboration and citation between two places are correlated. Hence, we also look at the geographical proximity in the citation network. We found that the probability that there is a link between two cities in the citation network also decreases with distance as a power law (Fig. 4C). In this case the power law exponent is much lower (0.30 ± 0.01). The gravity law for the citation network reads

$$w_{ij}^{\mathrm{Cite}} \propto \frac{s_i^{\mathrm{out}} s_j^{\mathrm{in}}}{d_{ij}^{\alpha}}. \tag{3}$$

In Fig. 4D we plot $w_{ij}^{\mathrm{Cite}} \Big/ \left(s_{ij}^{\mathrm{out}} s_i^{\mathrm{in}}\right)$ against the distance between all the node pairs in the citation network. As for the collaboration network

Table 1 | Dependence of citations on collaboration. We categorize each paper by the number of authors and their affiliations. For each of these groups we indicate the fraction of papers that are in the group and the mean number of citations. The error represents the standard error of the mean, calculated using bootstrap sampling with repetition

| $N_{\text{Authors}}$ | $f_{\text{Papers}}$ (in %) | Single City | Multiple City | Multiple Countries |
|---|---|---|---|---|
| 1 | 13.03 | 4.25 ± 0.02 | 4.95 ± 0.12 | 5.24 ± 0.11 |
| 2 | 19.01 | 6.80 ± 0.02 | 6.11 ± 0.04 | 7.00 ± 0.05 |
| 3 | 18.34 | 6.92 ± 0.02 | 6.38 ± 0.03 | 7.30 ± 0.04 |
| 4 | 14.95 | 7.19 ± 0.02 | 7.02 ± 0.03 | 8.03 ± 0.04 |
| 5 | 11.10 | 7.62 ± 0.03 | 7.66 ± 0.03 | 8.79 ± 0.04 |
| 6 | 8.01 | 8.13 ± 0.04 | 8.52 ± 0.05 | 9.77 ± 0.05 |
| 7 | 5.20 | 8.85 ± 0.05 | 9.56 ± 0.07 | 10.90 ± 0.07 |
| 8 | 3.45 | 9.50 ± 0.07 | 10.67 ± 0.09 | 12.10 ± 0.10 |
| 9 | 2.22 | 10.23 ± 0.10 | 11.52 ± 0.12 | 13.17 ± 0.12 |
| 10 | 1.53 | 10.57 ± 0.12 | 12.45 ± 0.14 | 14.70 ± 0.15 |
| >10 | 3.17 | 13.82 ± 0.17 | 16.64 ± 0.16 | 21.37 ± 0.17 |

we found that $\left\langle w_{ij}^{\text{Cite}} \middle/ \left(s_i^{\text{out}} s_j^{\text{in}}\right) \right\rangle$ decreases with distance as a power law with the exponent $\alpha = 0.77 \pm 0.02$ ($R = -0.35 \pm 0.001$). The above analysis shows the existence of an important spatial component in both the citation and the collaboration network. It shows that both our collaborators and our citations typically come from our spatial neighborhood. Further, long distance collaborations as well as citations decrease as a power law of distance. The difference of the scaling exponents of the two networks suggests that two distant places are more likely to cite each other than collaborate. Additional results are shown in the Supplementary Fig. S3 online.

The research performance of each country is generally estimated on the basis of the number of publications and citations. Although these are straightforward measurements of research output, they depend on a wide spectrum of resources[39]. For instance, the number of researchers and facilities (instruments, laboratories, libraries and other resources) available are typically different in different countries. A key determinant is the funding available for research & development (R&D). To quantify the expenses in R&D of a country we consider the fraction of gross domestic product (GDP) that is spent on R&D. To get rid of economic inequalities in different countries we consider the R&D spending in terms of the purchasing power parity (PPP). In Fig. 5A, we plot the number of citations $N_{\text{Cite}}$ against the R&D expenditure and find that it scales linearly with funding. Such correlation is not surprising, but the scaling exponent is non-trivial. It suggests that it is not possible to perform or contribute substantially unless there is a corresponding amount of funding available for research. Moreover, the research contribution in terms of citations also scales linearly with the number of researchers in that country (Fig. 5B). This result is consistent with the fact that the R&D expenditure is correlated with the number of researchers. The number of publications of a country also shows similar scaling against R&D expenditure and number of researchers (Supplementary Fig. S4 online).

Finally as a measure of impact of a country's scientific output we consider the average number of citations to the publications of that

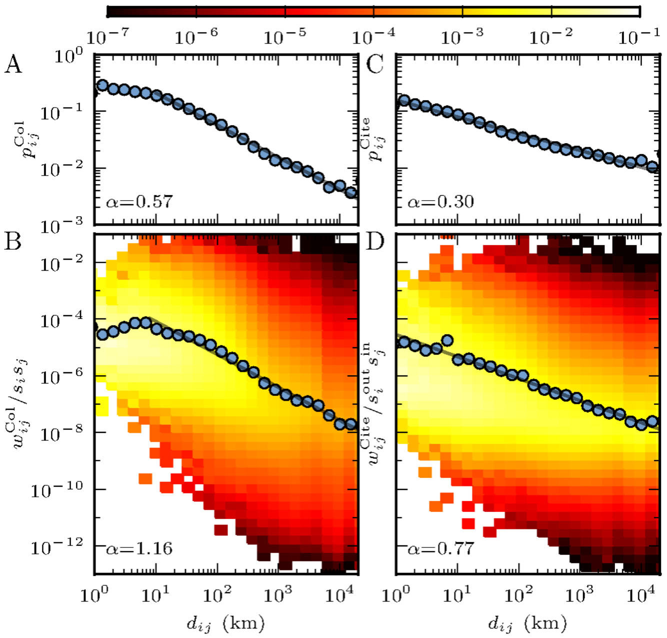

**Figure 4 | Effect of geographical proximity in the world collaboration and citation networks.** The probability of existence of a link as a function of the distance between two cities in the (A) collaboration network and (B) citation network. Distribution of the ratio of the link weight and product of the strengths of its endpoints in (C) collaboration network, $w_{ij}^{\text{Col}} \big/ s_i s_j$ and (D) citation network, $w_{ij}^{\text{Cite}} \big/ s_i^{\text{out}} s_j^{\text{in}}$ against the distance $d_{ij}$ between the cities. For each distance the average ratio is also shown. The solid line indicates a power law behavior with exponent $\alpha = 1.16 \pm 0.03$ and $0.77 \pm 0.02$ respectively.

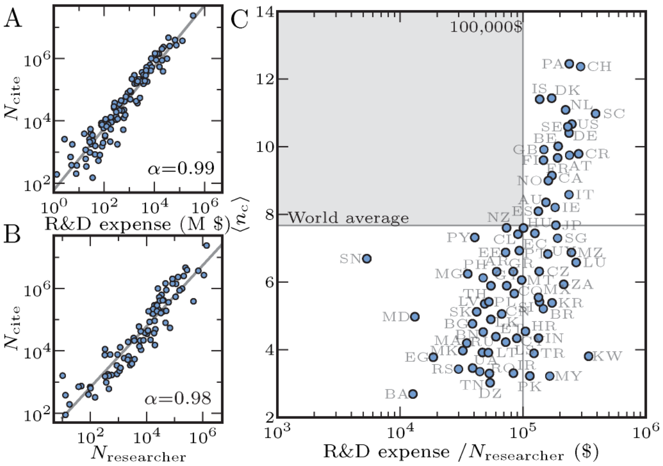

**Figure 5 | Relation between research outcome and funding.** Average number of citations per paper of a country against (A) the expenditure in research and development (in millions of dollars per year, and purchasing power parity) and (B) the number of researchers in that country. The solid line indicates power law scaling with exponent 0.99 ± 0.03 and 0.98 ± 0.04, respectively. (C) Average number of citations per paper of a country against the average spending per researcher. The horizontal line indicates the average number of citations over all papers of all countries, the vertical line indicates the threshold of about 100,000 $ per researcher per year.

country. In Fig. 5C we plot this number against the average spending per researcher per year (R&D expenditure divided by the number of researchers). The latter is not the average salary of researchers in that country, as it includes other expenditures such as infrastructure, bureaucracy, instruments, etc. This plot is much more scattered than the previous plots and does not show any definite correlation pattern. In order to identify groups of countries that behave similarly or show similar characteristics we use the $k$-mean clustering technique[40]. By using this clustering method with $k = 2$, we found that the countries can be classified into two groups, one with average spending less than about 100,000 $ per researcher per year and other with average spending more than about 100,000 $ (Fig. S5 online). Another clustering methods also gives qualitatively similar results. This separation in two groups, distinguished by the average spending per researcher per year (vertical line in the plot) also reveals another striking feature. If the average spending is less than about 100,000 $ (vertical line in the plot) per researcher per year we see an increase in the average number of citations with the spending. However if the average spending exceeds this limit, it becomes scattered and independent of funding. This figure shows that very rich countries like Kuwait and Luxembourg have high funding per researcher, still the average number of citations per paper is low. Countries like India, Brazil have high funding per researcher as well, but low average number of cites; this might mean they are investing more on infrastructure. Switzerland, Costa Rica, Panama, Germany, Austria, Netherlands, United States have high spending per researcher and their average number of citations is also high. If we display the number of cites per paper averaged over all countries (horizontal line), we see that there are no countries in the top left quadrant, i.e. it is not possible to do better than the world's average unless there is sufficient spending. Additional measures of a country's research performance and corresponding rankings are reported in the Supplementary Table S1 online.

## Discussion

Our thorough analysis of the world citation and collaboration networks has revealed that the effects of geography on the dynamics of science are relevant, despite the recent advances in communication and transportation. The occurrence of gravity laws for both citation and collaboration implies a preference by scientists to interact with peers in their geographic areas. However, long-distance interactions are not rare, as the interaction strength and probability are characterized by power law decays. Our work follows similar findings in mobile phone communication[1,2], social media[3] and international trade[41], reinforcing the belief that gravity laws hold in several different contexts, and that scientific interactions are not exceptional from this point of view. Thus, the gravity law is a fundamental relationship holding also in human dynamics.

Citation and collaboration streams between distinct locations are strongly correlated, with an approximately linear relation. An increase in the number of collaborations between two cities is then expected to be followed by a proportional increase in the flow of citations between the cities. This is justified from the fact the people/groups working in similar fields and subject area are more likely to cite as well as collaborate with each other, and also suggests a natural bias towards self-citation, of which we have provided strong quantitative evidence.

From the point of view of scientific impact, it pays off for a team to put together several institutions with a strong international participation. While part of this effect could be justified by the fact that having people from different locations facilitates the circulation of a work, which then becomes more visible and susceptible to be cited, the trend indicates that it is more likely to produce high quality work through international collaborations. It would be valuable to be able to disentangle the impact due to social networking from that due to the quality of the paper. Our findings pave the way for the first quantitative assessment of this issue. As a consequence, we expect to observe an increasing tendency to form large teams with members of many different countries in the future.

We also disclose a striking effect in the relationship between the national expenditure per researcher and the impact of the scientific output of a country. If the average spending per researcher per year is low, it is impossible for a country to do better than the world average, in terms of the average number of cites per paper. So there is a minimal funding quota that needs to be exceeded if a country wishes to have a scientific output of high average quality. Exceeding the threshold, however, does not guarantee success. This suggest that in science money acts as a kind of threshold motivator: if one does not pay people enough they will not be motivated and the outcomes of the research are poor; if people are paid sufficiently to take the issue of money off the table, internationally competitive findings are within reach. On the other hand, for conceptual and creative tasks, paying more than a certain threshold does not necessarily increase the output[42–44]. Further, our analysis reveals that at the country level funding has a positive linear impact on the research output both in terms of number of publications as well as citations. Thus, it is not possible for a country to increase its research output substantially without a sizeable increase in investments.

In the future we plan to study the role of cities' population, in particular on the distributions of citation and collaboration strengths along with their flows. It is well known that most characteristics of cities are strongly correlated to the size of their populations[45]. Furthermore, an analysis of the evolution of the world citation and collaboration networks would show how the spatial dimension of science dynamics has been affected by the progress of technology, internationalization and extreme events (e.g. wars, economic crises). This way one could infer how the scientific landscape has been shaping up in the last decades and how it is possible to create more efficient partnerships, via dedicated funding programs at the national and/or international level, and consequently a more productive and successful scholarly world.

## Methods

**Data description.** We have analyzed all publications (articles, reviews and editorial comments) written in English from 2003 till the end of 2010 included in the database of the Institute for Scientific Information (ISI) Web of Science. For each publication we extract the affiliations of the authors and the corresponding citations to that publication. We parsed the affiliations of all publications and have determined the geographic location at the city and country level. If there are multiple affiliations listed in a publication, the latter is associated with all represented cities and countries. After obtaining the locations we use the publicly available resources (www.wikipedia.org and maps.google.com) to determine their coordinates (latitude and longitude). Our dataset consists of 8,094,948 publications which have received 62,105,592 citations during the period 2003–2010. We were able to extract the geographical information from 8,092,314 publications. Affiliations refer to 226 countries and 37,750 cities. In order to get rid of anomalies due to any misclassification, we have only considered those places that have appeared in at least 5 publications during the period 2003–2010. This cutoff led us to 18,199 cities, producing 99.8% of the total publications and receiving 99.9% of total citations.

Country level information regarding expenditures for research and development (R&D) in terms of purchasing power parity (PPP) and number of researchers in R&D are obtained from the World Bank Data (databank.worldbank.org) for each year between 2003 till 2010. By aggregating these yearly datasets we determine the average of each of the above quantities for the period 2003–2010. The data of expenditure for R&D is available for 102 countries, the numbers of researchers for 89 countries and for 77 countries both datasets are available. Further details can be found in the Supplementary Methods online.

**Network construction.** We have analyzed the data at the country and the city level. As the publications and their affiliations form a bipartite graph, we construct the collaboration network between countries (cities) by projecting it onto the space of affiliations. In this collaboration network individual countries (cities) act as nodes, and links between them indicate that they have appeared in the same publication. If a paper is written by authors with $n$ affiliations, we put $\frac{1}{2}n \times (n-1)$ undirected links between each possible pair of collaborating countries (cities), with every link having weight $\frac{2}{n \times (n-1)}$. The total weight between any pair of nodes is the sum of all the weights over all the publications in the dataset. If there is a single affiliation in a publication then we put a self-link with weight 1.

In the citation network between countries (cities) nodes are papers which are linked if one paper cites the other. If a paper written by authors with $n$ affiliations cites a paper written by authors with $m$ affiliations we put $n \times m$ directed connections from each of the $n$ citing countries (cities) to each of the $m$ cited countries (cities), every link having weight $1/(nm)$. The total weight of a directed link between two countries (cities) is the sum of all the weights over all the citations in the dataset. Since there can be multiple affiliations from the same country (city) in a publication, there are self-loops both in the world citation and in the world collaboration networks.

**Great-circle distance.** The geodesic or the great-circle distance is the shortest distance between any two points on the earth measured along a path on the surface of the earth. Given the latitudes and longitudes of two points, we have used the Haversine formula to calculate the great-circle distance between them[46]. In these calculations, we considered the earth's radius to be 6372.8 KM.

1. Lambiotte, R. *et al.* Geographical dispersal of mobile communication networks. *Physica A* **387**, 5317–5325 (2008).
2. Krings, G., Calabrese, F., Ratti, C. & Blondel, V. D. Urban gravity: a model for inter-city telecommunication flows. *J. Stat. Mech.* **2009**, L07003 (2009).
3. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P. & Tomkins, A. Geographic routing in social networks. *Proc. Natl. Acad. Sci. USA* **102**, 11623–11628 (2005).
4. Okubo, Y. & Zitt, M. Searching for research integration across europe: a closer look at international and inter-regional collaboration in france. *Sci. Publ. Policy* **31**, 213–226 (2004).
5. Banchoff, T. Institutions, Inertia and European Union Research Policy. *J. Common. Mark. Stud* **40**, 1–21 (2002).
6. Cairncross, F. *The death of distance: How the communications revolution is changing our lives* (Harvard Business School Press, Boston, 2001).
7. Finholt, T. A. & Olson, G. M. From laboratories to collaborators: A new organizational form for scientific collaboration. *Psychol. Sci.* **8**, 28–36 (19970101).
8. Teasley, S. & Wolinsky, S. Scientific collaborations at a distance. *Science* **292**, 2254–2255 (2001).
9. Georghiou, L. Global cooperation in research. *Res. Policy* **27**, 611–626 (1998).
10. Rosenblat, T. S. & Mobius, M. M. Getting closer or drifting apart? *Q. J. Econ* **119**, 971–1009 (2004).
11. Havemann, F., Heinz, M. & Kretschmer, H. Collaboration and distances between german immunological institutes–a trend analysis. *J. Biomed. Discov. Collab.* **1**, 6 (2006).
12. Chandra, A., Hajra, K., Das, P. & Sen, P. Modelling temporal and spatial features of collaboration network. *Int. J. Mod. Phys. C* **18**, 1157–1172 (2007).
13. Agrawal, A. & Goldfarb, A. Restructuring research: Communication costs and the democratization of university innovation. *Am. Econ. Rev.* **98**, 1578–90 (2008).
14. Hennemann, S., Rybski, D. & Liefner, I. The myth of global science collaborationcollaboration patterns in epistemic communities. *J. Informetr* **6**, 217–225 (2012).
15. Katz, J. S. & Martin, B. R. What is research collaboration? *Res. Policy* **26**, 1–18 (1997).
16. Adams, J. D., Black, G. C., Clemmons, J. R. & Stephan, P. E. Scientific teams and institutional collaborations: Evidence from u.s. universities, 1981–1999. *Res. Policy* **34**, 259–285 (2005).
17. Wuchty, S., Jones, B. & Uzzi, B. The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007).
18. Jones, B., Wuchty, S. & Uzzi, B. Multi-university research teams: Shifting impact, geography, and stratification in science. *Science* **322**, 1259–1262 (2008).
19. Narin, F., Stevens, K. & Whitlow, E. S. Scientific cooperation in europe and the citation of multinationally authored papers. *Scientometrics* **21**, 313–323 (1991).
20. Glänzel, W., Schubert, A. & Czerwon, H. A bibliometric analysis of international scientific cooperation of the european union. *Scientometrics* **45**, 185–202 (1999).
21. Petersen, A., Riccaboni, M., Stanley, H. & Pammolli, F. Persistence and uncertainty in the academic career. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 5213–5218 (2012).
22. Börner, K., Penumarthy, S., Meiss, M. & Ke, W. Mapping the diffusion of scholarly knowledge among major u.s. research institutions. *Scientometrics* **68**, 415–426 (2006).
23. Pan, R. K., Sinha, S., Kaski, K. & Saramaki, J. The evolution of interdisciplinarity in physics research. *Sci. Rep.* **2**, 551 (2012).
24. Frenken, K., Hardeman, S. & Hoekman, J. Spatial scientometrics: Towards a cumulative research program. *J. Informetr* **3**, 222–232 (2009).
25. Lee, S. & Bozeman, B. The impact of research collaboration on scientific productivity. *Soc. Stud. Sci.* **35**, 673–702 (2005).
26. Arora, A., David, P. & Gambardella, A. Reputation and competence in publicly funded science: estimating the effects on research group productivity. *Annales d'Economie et de Statistique* 163–198 (1998).
27. Arora, A. & Gambardella, A. The impact of nsf support for basic research in economics. *Annales d'Economie et de Statistique* 91–117 (2005).
28. Caldarelli, G. *Scale-Free Networks* (Oxford University Press, Oxford, UK, 2007).
29. Barrat, A., Barthélemy, M. & Vespignani, A. *Dynamical processes on complex networks* (Cambridge University Press, Cambridge, UK, 2008).
30. Newman, M. *Networks: An Introduction* (Oxford University Press, Inc., New York, NY, USA, 2010).
31. Rosvall, M. & Bergstrom, C. T. Mapping change in large networks. *PLoS ONE* **5**, e8694 (2010).
32. Gastner, M. & Newman, M. Diffusion-based method for producing density-equalizing maps. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 7499 (2004).
33. Radicchi, F., Fortunato, S. & Castellano, C. Universality of citation distributions: Toward an objective measure of scientific impact. *Proc. Natl. Acad. Sci. USA* **105**, 17268–17272 (2008).
34. Zipf, G. K. *Human Behaviour and the Principle of Least Effort* (Addison Wesley, Cambridge, Massachusetts, 1949).
35. Gabaix, X. Zipf's law for cities: an explanation. *Q. J. Econ.* **114**, 739–767 (1999).
36. Bhattacharya, K., Mukherjee, G., Saramaki, J., Kaski, K. & Manna, S. S. The international trade network: weighted network analysis and modelling. *J. Stat. Mech.* **2008**, P02002 (2008).
37. Barthélemy, M. Spatial networks. *Phys. Rep.* **499**, 1–101 (2011).
38. Katz, J. S. Geographical proximity and scientific collaboration. *Scientometrics* **31**, 31–43 (1994).
39. Johnes, J. & Johnes, G. Research funding and performance in uk university departments of economics: a frontier analysis. *Econ. Educ. Rev.* **14**, 301–314 (1995).
40. Sculley, D. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, WWW. **10**, 1177–1178 (ACM, New York, NY, USA 2010).
41. Kaluza, P., Kölzsch, A., Gastner, M. & Blasius, B. The complex network of global cargo ship movements. *J. R. Soc. Interface* **7**, 1093–1103 (2010).
42. Adams, J. Inequity in social exchange. *Advances in experimental social psychology* **2**, 267–299 (1966).
43. Alderfer, C. *Existence, relatedness, and growth: Human needs in organizational settings.* (Free press, New York, NY, US, 1972).
44. Deci, E. & Ryan, R. *Intrinsic motivation and self-determination in human behavior* (Plenum Press, New York, 1985).
45. Bettencourt, L., Lobo, J., Helbing, D., Kuhnert, C. & West, G. B. Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7301 (2007).
46. Sinnott, R. Virtues of the haversine. *Sky Telescope* **68**, 158 (1984).

## Acknowledgments

Financial support from EUs FP7 FET-Open to ICTeCollective Project No. 238597, and from the Academy of Finland, the Finnish Center of Excellence program 2006–2011, Project No. 129670 are gratefully acknowledged. Certain data included herein are derived from the Science Citation Index Expanded, Social Science Citation Index and Arts & Humanities Citation Index, prepared by Thomson Reuters, Philadelphia, Pennsylvania, USA, Copyright Thomson Reuters, 2011.

## Author contributions

All authors designed the research and participated in the writing of the manuscript. RKP collected the data, analysed the data and performed the research.

## Additional information

**Supplementary information** accompanies this paper at http://www.nature.com/scientificreports

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Pan, R.K., Kaski, K. & Fortunato, S. World citation and collaboration networks: uncovering the role of geography in science. *Sci. Rep.* **2**, 902; DOI:10.1038/srep00902 (2012).