Characterizing scientific production and consumption in Physics

Qian Zhang¹, Nicola Perra¹, Bruno Gonçalves², Fabio Ciulla¹ & Alessandro Vespignani¹,²,³,⁴

¹Laboratory for the Modelling of Biological and Sociotechnical Systems, Northeastern University, Boston MA 02115 USA, ²Aix Marseille Université, CNRS, CPT, UMR 7332, 13288 Marseille, France, ³Institute for Scientific Interchange Foundation, Turin 10133, Italy, ⁴Institute for Quantitative Social Sciences at Harvard University, Cambridge, MA, 02138.

We analyze the entire publication database of the American Physical Society generating longitudinal (50 years) citation networks geolocalized at the level of single urban areas. We define the knowledge diffusion proxy, and scientific production ranking algorithms to capture the spatio-temporal dynamics of Physics knowledge worldwide. By using the knowledge diffusion proxy we identify the key cities in the production and consumption of knowledge in Physics as a function of time. The results from the scientific production ranking algorithm allow us to characterize the top cities for scholarly research in Physics. Although we focus on a single dataset concerning a specific field, the methodology presented here opens the path to comparative studies of the dynamics of knowledge across disciplines and research areas.

Over the last decade, the digitalization of publication datasets has propelled bibliographic studies allowing for the first time access to the geospatial distribution of millions of publications, and citations at different granularities⁶⁻⁸ (see Ref. 9 for a review). More precisely, authors’ name, affiliations, addresses, and references can be aggregated at different scales, and used to characterize publications and citations patterns of single papers¹⁰⁻¹¹, journals¹²⁻¹³, authors¹⁴⁻¹⁶, institutions¹⁷, cities¹⁸, or countries¹⁹. The sheer size of the datasets allows also system level analysis on research production and consumption²⁰, migration of authors²¹⁻²², and change in production in several regions of the world as a function of time²³, just to name a few examples. At the same time those analyses have spurred an intense research activity aimed at defining metrics able to capture the importance/ranking of authors, institutions, or even entire countries¹⁴⁻¹⁵,¹⁷,²³⁻²⁹. Whereas such large datasets are extremely useful in understanding scholarly networks and in charting the creation of knowledge, they are also pointing out the limits of our conceptual and modeling frameworks³⁰ and call for a deeper understanding of the dynamics ruling the diffusion and fruition of knowledge across the the social and geographical space.

In this paper we study citation patterns of articles published in the American Physical Society (APS) journals in a fifty-year time interval (1960–2009)³¹. Although in the early years of this period the dataset was obviously biased toward the scholarly activity within the USA, in the last twenty years only about 35% of the papers are produced in the USA. The same amount of production has been observed in databases that include multiple journals, and disciplines⁷⁻¹⁹. Indeed the journals of the APS are considered worldwide as reference publication venues that well represent the international research activity in Physics. Furthermore this dataset does not bundle different disciplines and publication languages, providing a homogeneous dataset concerning Physics scholarly research.

For each paper we geolocalize the institutions contained in the authors’ affiliations. In this way we are able to associate each paper in the database with specific urban areas. This defines a time resolved, geolocalized citation network including 2,307 cities around the world engaged in the production of scholarly work in the area of Physics. Following previous works⁸⁻¹⁷ we assume that the number of given or received citations is a proxy of knowledge consumption or production, respectively. More precisely, we assume that citations are the currency traded between parties in the knowledge exchange. Nodes that receive citations export their knowledge to others. Nodes that cite other works, import knowledge from others. According to this assumption we classify nodes considering the unbalance in their trade. Knowledge producers are nodes that are cited (export) more than they cite (import). On the contrary, we label as consumers nodes that cite (import) more than they are cited (export). Using this classification, we define the knowledge diffusion proxy algorithm to explore how scientific knowledge flows from producers to consumers. This tool explicitly assumes a systemic perspective of knowledge diffusion, highlighting the global structure of scientific production and consumption in Physics.

The temporal analysis reveals interesting patterns and the progressive delocalization of knowledge producers. In particular, we find that in the last twenty years the geographical distribution of knowledge production has drastically changed. A paramount example is the transition in the USA from a knowledge production localized...
around major urban areas in the east and west coast to a broad geographical distribution where a significant part of the knowledge production is now occurring also in the midwestern and southern states in the USA. Analogously, we observe the early 90s dominance of UK and Northern Europe to subside to an increase of production from France, Italy and several regions of Spain. Interestingly, the last decade shows that several of China’s urban areas are emerging as the largest knowledge consumers worldwide. The reasons underlying this phenomenon may be related to the significant growth of the economy and the research/development compartment in China in the early 21st century. This positive stimulus, pushed up also the scientific consumption with a large number of paper citing work from other world areas. Indeed, the increase of publications is associated to an increase of the citations unbalance, moving China to the top rank as consumers since the recent influx of its new papers has not yet had the time to accumulate citations.

Although the knowledge diffusion proxy provides a measure of knowledge production and consumption, it may be inadequate in providing a rank of the most authoritative cities for Physics research. Indeed, a key issue in appropriately ranking the knowledge production, is that not all citations have the same weight. Citations coming from authoritative nodes are heavier than others coming from less important nodes, thus defining a recursive diffusion of ranking of nodes in the citation network. In order to include this element in the ranking of cities we propose the scientific production ranking algorithm. This tool, inspired by the PageRank algorithm, allows us to define the rank of each node, as function of time, going beyond the knowledge diffusion proxy or simple local measures such as citation counts or h-index. In this algorithm the importance of each node diffuses through the citation links. The rank of a node is determined by the rank of the nodes that cite it, recursively, thus implicitly weighting differently citations from highly (lowly) ranked nodes. Also in this case we observe noticeable changes in the ranking of cities along the years. For instance the presence of both European and Asian cities in the top 100 list increases by 50% in the last 20 years. This findings suggest that the Internet, digitalization and accessibility of publications are creating a more levelled playing field where the dominance of specific area of the world is being progressively eroded to the advantage of a more widespread and complex knowledge production and consumption dynamic.

Results

We focus our analysis on the APS dataset. It contains all the papers published by the APS from 1893 to 2009. We consider only the last 50 years due to the incomplete geolocalization information available for the early years. During this period, the large majority of indexed papers, 97.47%, contain complete information such as authors name, journal of publication, day of publication, list of affiliations and list of citations to other articles published in APS journals. We geolocalized 96.97% of papers at urban area level with an accuracy of 98.5%. We refer the reader to the Methods section and to the Supplementary Information (SI) for the detailed description of the dataset and the techniques developed to geolocalize the affiliations.

In total, only 43% of papers has been produced inside the USA. Interestingly, over time this fraction has decreased. For example, in the 60’s it was 85.59%, while in the last 10 years decreased to just 36.67%. While one might assume that the APS dataset is biased toward the USA scientific community, the percentage of publications contributed by the USA in APS journals after 1990 is almost the same as in other publication datasets. These alternative datasets contain journals published all over the world and mix different scientific disciplines. This supports the idea that the APS journals are now attracting the worldwide physics scientific community independently of nationality, and fairly represent the world production and consumption of Physics. It is not possible to provide quantitative analysis of the evaluation of the ranking of the nodes and disentangle it by an actual change of the dynamic of knowledge production. For this reason, and in order to minimize any bias in the analysis we focus our analysis in the last 20 years of data.

In order to construct the geolocalized citation network we consider nodes (urban areas) and directed links representing the presence of citations from a paper with affiliation in one urban area to a paper with affiliation in another urban area. For example, if a paper written in node i cites one paper written in node j there is an link from i to j, i.e., j receives a citation from i and sends a citation to j. Each paper may have multiple affiliations and therefore citations have to be proportionally distributed between all the nodes of the papers. For this reason we weight each link in order to take into account the presence of multiple affiliations and multiple citations. In a given time window, the total number of citations for papers written in j received from papers written in i, is the weight of the link i → j, and the total number of citations for those papers written in j sent to the papers written in k is the weight of the link j → k. For instance, if in a time window t, there is one paper written in node j, which cite two papers written in node k and was cited by three papers written in node i, then w_{jk} = 2, w_{ij} = 3, and we add all such weights for each paper written in that node j and obtain the weights for links. For papers written in multiple cities, say j1, j2, the weight will be counted equally. The time window we use in this manuscript is one year. We show an example of the network construction in Figure 1.

In order to define main actors in the production and consumption of Physics, we consider citations as a currency of trade. This analogy allows us to immediately grasp the meaning and distinction between producers and consumers of scientific knowledge. Nodes that receive citations export their knowledge to the citing nodes. Instead, nodes that cite, papers produced from other nodes of the network, import knowledge from the cited nodes. Measuring the unbalance trade between citations, we define producers as cities that export more than they import, and consumers as cities that import more than they export. More precisely, we can measure the total knowledge imported by each urban area as \( \sum_{j} w_{ij} \) and the total export as \( \sum_{i} w_{ij} \) in a given year. Those measures however acquire specific meaning when considered relatively to the total trade of physics production.

Figure 1 | Projecting a paper citation relationship into a city-to-city citation network. (A) Paper A written by authors from Ann Arbor, Los Alamos and New York cites one paper B written by authors from Rome and Madrid and another paper C from Oxford and Princeton. (B) In a city-to-city citation network, directed links from Ann Arbor to Madrid, Rome, Oxford and Princeton are generated, and similarly Los Alamos and New York are connected to the above four cited cities.
knowledge worldwide in the same year; i.e. the total number of citations worldwide $S = \sum_{ij} w_{ij}$. The relative trade unbalance of each urban area $i$ is then:

$$\Delta S_i = \frac{\sum_j w_{ij} - \sum_j w_{ji}}{S}.$$  \hspace{1cm} (1)

A negative or positive value of this quantity indicates if the urban area $i$ is consumer or producer, respectively. In Figure 2-A we show the worldwide geographical distribution of producer (red) and consumer (blue) urban areas for the 1990 and 2009. Interestingly, during the 90s the production of Physics knowledge was highly localized in a few cities in the eastern and western coasts of the USA and in a few areas of Great Britain and Northern Europe. In 2009 the picture is completely different with many producer cities in central and southern parts of the USA, Europe and Japan. It is interesting to note that despite the fraction of papers produced in the USA is generally decreasing or stable, many more cities in the USA acquire the status of knowledge producers. This implies that the quality of knowledge production from the USA is increasing and thus attracting more citations. This makes it clear that the knowledge produced by an urban area can not be considered to be measured only by the raw number of papers. Citations are a more appropriate proxy that encodes the value of the products. They serve as an approximation of the actual flow of knowledge. The Figure 2-A also makes it clear that cities in China are playing the role of major consumers in both 1990 and 2009. We also observe that cities in other countries like Russia
and India consumed less in 2009 than 1990. In other words, in 2009 both the production and consumption of knowledge are less concentrated on specific places and generally spread more evenly geographically. In order to provide visual support to this conclusion we show in Figure 2-B the geographical distribution of producers and consumers inside the USA. From the two maps it is evident the drift of knowledge production from the two coastal areas in the USA to the midwest, central and southern states. Similarly, in Figure 2-C we plot the same information for western Europe. In 1990 only a few urban areas in Germany and France were clearly producers. By 2009 this dominance has been consistently eroded by Italy, Spain and a more widespread geographical distribution of producers in France, Germany and UK.

**Knowledge diffusion proxy.** The definition of producers and consumers is based on a local measure, that does not allow to capture all possible correlations and bounds between nodes that are not directly connected. This might result in a partial view and description of the system, especially when connectivity patterns are complex. Interestingly, a close analysis of each citation network, see Figure 3, clearly shows that citation patterns have indeed all the hallmarks of complex systems, especially in the last two decades. The system is self-organized, there is not a central authority that assigns citations and papers to cities, there is not a blueprint of system’s interactions, and as clearly shown from Figure 3-C the statistical characteristics of the system are described by heavy-tailed distributions. Not surprisingly, the level of complexity of the system has increased with time. In Figure 3-A we plot the most statistically significant connections of the citation network between cities inside the USA in 1960, 1990 and 2009. We filter links by using the backbone extraction algorithm which preserves the relevant connections of weighted networks while removing the least statistically significant ones. We visualize each filtered network by using a bundled representation of links. The direction of each weighted link goes from blue (citing) to red (cited). Similarly, in Figure 3-B, we visualize the most significant links between cities in Europe (European Union’s 27 countries, as well as Switzerland and Norway). It is clear from Figure 3-A that in 1960 the citation patterns inside the USA were limited to a few cities, and in Europe only a few cities were connected. Instead, in 1990 and 2009 we register an increase in the interactions among a larger number of cities. The observed temporal trend is well known and valid not just for Physics. Among many factors that have been advocated to explain this tendency we find the increase of the research system and the advance in technology that make collaboration and publishing easier.

In order to explicitly consider the complex flow of citations between producers and consumers, we propose the knowledge diffusion proxy algorithm (see Methods section for the formal definition). In this algorithm, producers inject citations in the system that flow along the edges of the network to finally reach consumer cities where the injected citations are finally absorbed. The algorithm allows charting the diffusion of knowledge, going beyond local measures. The entire topology of the networks is explored uncovering nontrivial correlations induced by global citation patterns. For instance, knowledge produced in a city may be consumed by another producer that in turn produces knowledge for other cities who are consumers. This points out that the actual consumer of knowledge is not just signalled by the unbalance of citations but in the overall topology of the production and consumption of knowledge in the whole network. Indeed, the final consumer of each injected citation may not be directly connected with the producer. Citations flow
However, the fast publication growth increased the unbalance written in its capital, a dominant city for scientific research in China. We assume that, thanks to this positive stimulus, many more papers were produced in China in the early 2000s, it is reasonable to expect that the research population in China will increase in the near future cities in China might be moving among the strongest producers if a fair number of papers start receiving enough citations, which obviously depends on the quality of the research carried out in the last years. This is the case of cities like Tokyo which has gradually approached the citation balance in recent years. For instance, Table 2 shows that in 1990 Tokyo, was among the top consumers. But by 2009, its contribution to citation consumption had become less significant as observed from the data in Table 4 and Table 1.

**Ranking cities.** Authors, departments, institutions, government, and many funding agencies are extremely interested in defining the most important sources of knowledge. The necessity to find objective measures of the importance of papers, authors, journals, and institutions leads to the definition of a wide variety of rankings. Measures such as impact factor, number of citations and h-index are commonly used to assess the importance of scientific production. However, these common indicators might fail to account for the actual importance and prestige associated to each publication. In order to overcome these limitations, many different measures have been proposed. Here we introduce the scientific production ranking algorithm (SPR), an iterative algorithm based on the notion of diffusioning scientific credits. It is analogous to PageRank (PR), CiteRank (CR), HITS, SARA, and others ranking metrics. In the algorithm each node receives a credit that is redistributed to its neighbours at the next iteration until the process converges in a stationary distribution of credit to all nodes (see Methods section for the formal definition). The credits diffuse following citations links along all possible paths, sometimes through intermediate cities. In Table 1, and Table 2 we report the rankings of Top 10 final consumers evaluated by the knowledge diffusion proxy for the Top 3 producers in 2009 and 1990 respectively. We also list the Top 10 neighbours according to the local citation unbalance. From these two tables, it is clear that the final rank of each consumer, obtained by our algorithm, can be extremely different from the ranking obtained by just considering local unbalances. For instance, in 2009 Bratislava and Mainz rank in top 10 consumers absorbing knowledge produced in Beijing. However, according to long-term measures of unbalance, these two cities are ranked out of top 10 (shown in bold in Table 1). Interestingly, even the Top consumer for New Haven, Berlin, also does not rank among the Top 10 neighbours according to the citation unbalance. These findings confirm that in order to uncover the complex set of relationships among cities, it is crucial to consider the entire structure of the network, going beyond simple local measures.

In Figure 4-A and Figure 4-B we visualize the results considering the Top four producers cities in 2009 in the USA and in Europe respectively. We show their Top ten consumers over 20 years as function of time. The size of each circle is proportional to how many times each injected citation is absorbed by that consumer. In the plot, vertical grey strips indicate that the city was not a producer during those years (e.g. Orsay in 2008). The results show that, on average, Beijing is the top consumer for all of these producers in the past 20 years. Since China registered a big economical growth and increment of research population in the early 2000, it is reasonable to assume that, thanks to this positive stimulus, many more papers were written in its capital, a dominant city for scientific research in China. However, the fast publication growth increased the unbalance between sent and received citations. Each paper published in a given city imports knowledge from the cited cities. Reaching a balance might require some time. Each city needs to accumulate citations back to export its knowledge to others cities. We can speculate that in the near future cities in China might be moving among the strongest producers if a fair number of papers start receiving enough citations, which obviously depends on the quality of the research carried out in the last years. This is the case of cities like Tokyo which has gradually approached the citation balance in recent years. For instance, Table 2 shows that in 1990, Tokyo, was among the top consumers. But by 2009, its contribution to citation consumption had become less significant as observed from the data in Table 4 and Table 1.

**Ranking cities.** Authors, departments, institutions, government, and many funding agencies are extremely interested in defining the most important sources of knowledge. The necessity to find objective measures of the importance of papers, authors, journals, and institutions leads to the definition of a wide variety of rankings. Measures such as impact factor, number of citations and h-index are commonly used to assess the importance of scientific production. However, these common indicators might fail to account for the actual importance and prestige associated to each publication. In order to overcome these limitations, many different measures have been proposed. Here we introduce the scientific production ranking algorithm (SPR), an iterative algorithm based on the notion of diffusioning scientific credits. It is analogous to PageRank (PR), CiteRank (CR), HITS, SARA, and others ranking metrics. In the algorithm each node receives a credit that is redistributed to its neighbours at the next iteration until the process converges in a stationary distribution of credit to all nodes (see Methods section for the formal definition). The credits diffuse following citations links along all possible paths, sometimes through intermediate cities. In Table 1, and Table 2 we report the rankings of Top 10 final consumers evaluated by the knowledge diffusion proxy for the Top 3 producers in 2009 and 1990 respectively. We also list the Top 10 neighbours according to the local citation unbalance. From these two tables, it is clear that the final rank of each consumer, obtained by our algorithm, can be extremely different from the ranking obtained by just considering local unbalances. For instance, in 2009 Bratislava and Mainz rank in top 10 consumers absorbing knowledge produced in Beijing. However, according to long-term measures of unbalance, these two cities are ranked out of top 10 (shown in bold in Table 1). Interestingly, even the Top consumer for New Haven, Berlin, also does not rank among the Top 10 neighbours according to the citation unbalance. These findings confirm that in order to uncover the complex set of relationships among cities, it is crucial to consider the entire structure of the network, going beyond simple local measures.

In Figure 4-A and Figure 4-B we visualize the results considering the Top four producers cities in 2009 in the USA and in Europe respectively. We show their Top ten consumers over 20 years as function of time. The size of each circle is proportional to how many times each injected citation is absorbed by that consumer. In the plot, vertical grey strips indicate that the city was not a producer during those years (e.g. Orsay in 2008). The results show that, on average, Beijing is the top consumer for all of these producers in the past 20 years. Since China registered a big economical growth and increment of research population in the early 2000, it is reasonable to assume that, thanks to this positive stimulus, many more papers were written in its capital, a dominant city for scientific research in China. However, the fast publication growth increased the unbalance between sent and received citations. Each paper published in a given city imports knowledge from the cited cities. Reaching a balance might require some time. Each city needs to accumulate citations back to export its knowledge to others cities. We can speculate that in the near future cities in China might be moving among the strongest producers if a fair number of papers start receiving enough citations, which obviously depends on the quality of the research carried out in the last years. This is the case of cities like Tokyo which has gradually approached the citation balance in recent years. For instance, Table 2 shows that in 1990, Tokyo, was among the top consumers. But by 2009, its contribution to citation consumption had become less significant as observed from the data in Table 4 and Table 1.

**Ranking cities.** Authors, departments, institutions, government, and many funding agencies are extremely interested in defining the most important sources of knowledge. The necessity to find objective measures of the importance of papers, authors, journals, and institutions leads to the definition of a wide variety of rankings. Measures such as impact factor, number of citations and h-index are commonly used to assess the importance of scientific production. However, these common indicators might fail to account for the actual importance and prestige associated to each publication. In order to overcome these limitations, many different measures have been proposed. Here we introduce the scientific production ranking algorithm (SPR), an iterative algorithm based on the notion of diffusioning scientific credits. It is analogous to PageRank (PR), CiteRank (CR), HITS, SARA, and others ranking metrics. In the algorithm each node receives a credit that is redistributed to its neighbours at the next iteration until the process converges in a stationary distribution of credit to all nodes (see Methods section for the formal definition). The credits diffuse following citations links along all possible paths, sometimes through intermediate cities. In Table 1, and Table 2 we report the rankings of Top 10 final consumers evaluated by the knowledge diffusion proxy for the Top 3 producers in 2009 and 1990 respectively. We also list the Top 10 neighbours according to the local citation unbalance. From these two tables, it is clear that the final rank of each consumer, obtained by our algorithm, can be extremely different from the ranking obtained by just considering local unbalances. For instance, in 2009 Bratislava and Mainz rank in top 10 consumers absorbing knowledge produced in Beijing. However, according to long-term measures of unbalance, these two cities are ranked out of top 10 (shown in bold in Table 1). Interestingly, even the Top consumer for New Haven, Berlin, also does not rank among the Top 10 neighbours according to the citation unbalance. These findings confirm that in order to uncover the complex set of relationships among cities, it is crucial to consider the entire structure of the network, going beyond simple local measures.

In Figure 4-A and Figure 4-B we visualize the results considering the Top four producers cities in 2009 in the USA and in Europe respectively. We show their Top ten consumers over 20 years as function of time. The size of each circle is proportional to how many times each injected citation is absorbed by that consumer. In the plot, vertical grey strips indicate that the city was not a producer during those years (e.g. Orsay in 2008). The results show that, on average, Beijing is the top consumer for all of these producers in the past 20 years. Since China registered a big economical growth and increment of research population in the early 2000, it is reasonable to assume that, thanks to this positive stimulus, many more papers were written in its capital, a dominant city for scientific research in China. However, the fast publication growth increased the unbalance between sent and received citations. Each paper published in a given city imports knowledge from the cited cities. Reaching a balance might require some time. Each city needs to accumulate citations back to export its knowledge to others cities. We can speculate that in the near future cities in China might be moving among the strongest producers if a fair number of papers start receiving enough citations, which obviously depends on the quality of the research carried out in the last years. This is the case of cities like Tokyo which has gradually approached the citation balance in recent years. For instance, Table 2 shows that in 1990, Tokyo, was among the top consumers. But by 2009, its contribution to citation consumption had become less significant as observed from the data in Table 4 and Table 1.
self-consistently, implying that not all links have the same importance. Any city in the network will be more prominent in rank if it receives citations from high-rank sources. This process ensures that the rank of each city is self-consistently determined not just by the raw number of citations but also if the citations come from highly ranked cities. In Figure 5 we show the Top 20 cities from 1990 to 2009. Interestingly, we clearly see the decline and rise of cities along the years as well as the steady leadership of Boston and Berkeley. This behaviour is clear in Figure 6-B where we show the rank for cities in USA in 1990 and 2009. Meanwhile, the ranking of cities in European and Asian countries like France, Italy and Japan has increased significantly, as shown in both Figure 5 and Figure 6-A. In Figure 6-C we focus on the geographical distribution of ranks for a selected set of European countries in 1990 and 2009. In Table 3 we provide a quantitative measure of the change in the landscape of the most highly ranked cities in the world by showing the percentage of cities in the top 100 ranks for different continents. In Figure 7, we compare the ranking obtained by our recursive algorithm with the ranking obtained by considering the total volume of publications produced in each city. Since we are considering only journals by the APS, the impact factor is consistent across all cities and does not include disproportionate effects that often happen when mixing disciplines or journal with varied readership. It is then natural to consider a ranking based on the raw productivity of each place. As we see in the figure though the two rankings, although obviously correlated, provide different results. A number of cities whose ranking, according to productivity, is in the Top 20 cities in the world, are ranked one order of magnitude lower by the SPR algorithm. Valuing the number of citations and their origin in the ranking of cities produces results often not consistent with the raw number of papers, signaling that in some places a large fraction of papers are not producing knowledge...
as they are not cited. We believe that the present algorithm may be considered as an appropriate way to rank scientific production taking properly into account the impact of papers as measured by citations.

**Discussion**

In this paper we study the scientific knowledge flows among cities as measured by papers and citations contained in APS journals. In order to make clear the meaning and difference between producers and consumers in the context of knowledge, we propose an economical analogy referring to citations as a traded currency between urban areas. We then study the flow of citations from producers to consumers with the knowledge production proxy algorithm. Finally, we rank the importance of cities as function of time using the scientific production ranking algorithm. This method, inspired by the PageRank, allows us to evaluate the importance of cities explicitly considering the complex nature of citation patterns. In our analysis we considered just scientific publications contained in the APS journals. We do not have information on citations received or assigned to papers outside this dataset. These limitations certainly affect the count of citations of each city, potentially creating biases in our results. However, our findings, while limited to a particular dataset, are aligned with different observations reported by other studies focused on other datasets and fields. For example, we identify major US cities (e.g. Boston and San Francisco areas), as the most important sources of Physics. Similar observations have been done by Börner et al. at the institution level considering papers published in the Proceedings of the National Academy of Sciences, by Mazloumian et al. at country and city level with Web of Science dataset, and by...
Methods

Dataset. The dataset of the American Physical Society journals, considering papers published between 1983 and 2009 of which 450, 655 papers include a list of affiliations11. Each of paper may have multiple affiliations. In total there are 945, 767 affiliation strings.

In order to geolocalize the articles, we parse the city names from the affiliation strings for each article. First, we process each affiliation string and try to match country or US state names from a list of known names and their variations in different languages. We crosscheck the results with Google Map API obtaining validated location information for 97.7% of affiliation strings, corresponding to 445, 223 articles. It is worth noticing that we do not use Google Map API (or other map APIs like Yahoo! or Bing) directly for geocoding because, to our best knowledge, there are no accuracy guarantees to these API results. For each affiliation string with an extracted country or state name, we also match the city name against GeoName database12 corresponding to its country or US state. 92.6% of affiliation strings with extracted city names are subsequently verified with Google Map API. Finally, a total of 425, 233 publication articles successfully pass the filters we describe here.

The dataset also provides 4, 710, 548 records of citations between articles published in APS journals. To build citation networks at the city level, we merge the citation of 425, 233 publication articles successfully pass the filters we describe here.

Knowledge diffusion proxy algorithm. This analysis tool is inspired by the dollar experiment, originally developed to characterized the flow of money in economic networks56. Formally, it is a biased random walk with sources and sinks where a citation diffuses in the network. The diffusion takes place on top of the network of net trade flows. Let us define w_i as the number of citation that node i gives to j and w_j as the opposite flow. We can define the antisymmetric matrix T_{ij} = w_j - w_i. The network of the net trade is defined by the matrix F with F_{ij} = T_{ij} = T_{ji} for all connected pairs (i,j) with T_{ij} < 0 and F_{ij} = 0 for all connected pairs (i,j) with T_{ij} ≥ 0. There are two types of nodes. Producers are nodes with a positive trade balance Δs_i = s_i^+ - s_i^- = \sum_j T_{ij} - \sum_i T_{ji}. Their strength-in is larger than their strength-out. On the other hand, consumers are nodes with a negative trade balance Δs. As top of this network a citation is injected in a producer city. The citation follows the outgoing edges with a probability proportional to their intensities, and the probability that the citation is absorbed in a consumer city j equals to \rho_{in}(j) / \rho_i. By repeating many times this process from each starting point (producers) we can build a matrix with elements ϵ_{ij} that measure how many times a citation injected in the producer city i is absorbed in a city consumer j.

\[ F_i = q z_i + (1 - q) \sum_j \frac{P_j}{\sum_i w_{ij}} w_{ij} + (1 - q) z_i \sum_j P_j A j (s^+) . \]

\[ P_i = q z_i + (1 - q) \sum_j \frac{P_j}{\sum_i w_{ij}} w_{ij} + (1 - q) z_i \sum_j P_j A j (s^+) . \]

\[ z_i = \sum_j \Delta s_j / \rho_j, \]

where \rho defines the generic paper and \rho_j the number of nodes who have written the paper. It is important to notice that \rho_j = 1 only if the i-th node wrote the paper \rho, otherwise it equals zero.

1. Narin, F. & Carpenter, M. P. National Publication and Citation Comparisons. *Journal of the American Society for Information Science, 26*, 80–93 (1975).

2. Frame, J. D., Narin, F. & Carpenter, M. P. The Distribution of World Science. *Social Studies of Science, 7*, 501–516 (1977).

3. May, R. M. The Scientific Wealth of Nations. *Science, 7*, 793–796 (1997).

4. Batty, M. The Geography of Scientific Citation. *Environ Plan A* 35, 761–765 (2003).

5. Leydesdorff, L. & Zhou, P. Are the contributions of China and Korea upsetting the world system of science? *Scientometrics, 63*, 617–630 (2005).

6. Fortunato, S. & Vespignani, F. Opening the box: comparing EU and US scientific output by scientific field. *Technological Forecasting & Social Change, 74*, 1334–1356 (2007).

7. Pan, R. K., Kaski, K. & Fortunato, S. World citation and collaboration networks: uncovering the role of geography in science. *Sci. Rep., 2*, 902 (2012).

---

Table 3 | Percentage of top 100 ranked cities in continents in 1990 and 2009

| Continent | 1990 | 2009 |
|-----------|------|------|
| Asia      | 4.0% | 11.0%|
| Europe    | 24.0%| 33.0%|
| N. America| 72.0%| 56.0%|
8. Mazloomian, A., Helbing, D., Lozano, S., Light, R. P. & Börner, K. Global multi-level analysis of the 'scientific food web'. *Sci. Rep.*, 3, 1167 (2013).

9. Frenken, K., Hardeman, S. & Hoekman, J. Spatial scientometrics: Towards a cumulative research program. *Journal of Informetrics*, 3, 222–232 (2009).

10. Redner, S. How popular is your paper? An empirical study of the citation distribution. *Eur. Phys. J. B*, 4, 131–134 (1998).

11. Chen, P., Xie, H., Maslov, S. & Redner, S. Finding scientific gems with Google’s PageRank algorithm. *Journal of Informetrics*, 1, 8–15 (2007).

12. Garfield, E. Citation Analysis as a Tool in Journal Evaluation. *Science*, 178, 471–479 (1972).

13. Bergstrom, C. Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries News*, 68, 314–316 (2007).

14. Hirsch, J. E. An index to quantify an individual’s scientific research output. *Proc. Natl. Acad. Sci.*, 102, 16569–16572 (2005).

15. Egghe, L. Theory and practise of the g-index. *Scientometrics*, 69, 131–152 (2006).

16. Hirsch, J. E. Does the h index have predictive power? *Proc. Natl. Acad. Sci.*, 104, 19193–19198 (2007).

17. Börner, K., Penumarthi, S., Meiss, M. & Ke, W. Mapping the Diffusion of Information Among Major U.S. Research Institutions. *Scientometrics*, 68, 415–426 (2006).

18. Bornmann, L., Leydesdorff, L., Walch-Solimena, C. & Ettl, C. “Mapping excellence in the geography of science: An approach based on Scopus data. *Journal of Informetrics*, 5 no. 4, 537–546 (2011).

19. King, D. K. The scientific impact of nations. *Nature*, 430, 311–316 (2004).

20. Adams, J. Collaborations: The rise of research networks. *Nature*, 490, 335–336 (2012).

21. Laudel, G. Studying the brain drain: Can bibliometric methods help? *Journal of Informetrics*, 6, 326–329 (2012).

22. Garfield, E. Citation Indexing. *Its Theory and Application in Science, Technology, and Humanities*. John Wiley & Sons Inc., 1979.

23. Egghe, L. & Rousseau, R. Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science. Elsevier Science Publishers, 1990.

24. Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46, no. 5, 604–632 (1999).

25. Walker, D., Xie, X., Yan, K.-K. & Maslov, S. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment* 2007, P06010 (2007).

26. Castillo, C., Donato, D. & Gionis, A. Estimating Number of Citations Using Author Reputation. in *String Processing and Information Retrieval*, 68, 215–237 (2003).

27. Noorden, R. V. Global mobility: Science on the move. *Nature*, 490, 326–329 (2012).

28. Egghe, L. & Rousseau, R. *Introduction to Informetrics*: Quantitative Methods in Library, Documentation and Information Science. Elsevier Science Publishers, 1990.

29. Radicchi, F., Fortunato, S., Markines, B. & Vespignani, A. Diffusion of scientific credits and the ranking of scientists. *Phys. Rev. E*, 80, 056103 (2009).

30. Scharnhorst, A., Börner, K. & van den Besselaar, P. eds., *Models of Science Dynamics: Encounters Between Complexity Theory and Information Sciences*. Springer-Verlag, (2012).

31. APS, Data sets for research. (2010). (accessed June 8 2010).

32. http://data.worldbank.org/. (2012). (accessed September 12 2012).

33. Brin, S. & Page, L. The anatomy of a large-scale hypertextual web search engine, *Comp. Net. ISDN Sys.*, 30, 107 (1998).

34. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science*, 286, 509 (1999).

35. Barrat, A., Barthélemy, M. & Vespignani, A. *Dynamical Processes on Complex Networks*. Cambridge University Press (2008).

36. Newman, M. Networks. *An Introduction*. Oxford University Press (2010).

37. Vespignani, A. Predicting the behavior of techno-social systems. *Science* 325, 425–428 (2009).

38. Vespignani, A. Modeling dynamical processes in complex socio-technical systems. *Nature Physics*, 8, 32–30 (2012).

39. Angeles Serrano, M., Boguñá, M. & Vespignani, A. Extracting the multiscale backbone of complex weighted networks. *Proc. Natl. Acad. Sci.*, 106, 6483–6488, April (2009).

40. Boyandin, I., Bertini, E. & Lalanne, D. Using flow maps to explore migrations over time, in *Proceedings of GeoSpatial Visual Analytics Workshop in conjunction with The 13th AGILE International Conference on Geographic Information Science (GeoVA)* (2010).

41. Adams, J. & Griliches, Z. Measuring science: An exploration. *Proc. Natl. Acad. Sci.*, 93, 12664–12670 (1996).

42. Rosenblat, T. S. & Mobius, M. G. Getting Closer or Drifting Apart? Quarterly *Journal of Economics*, 119, no. 3, 971–1009 (2004).

43. Havemann, F., Heinz, M. & Kretschmer, H. Collaboration and distances between German immunological institutes a trend analysis. *Journal of Biomedical Discovery and Collaboration*, 1, 6 (2006).

44. Agrawal, A. & Goldfar, A. Restructuring Research: Communication Costs and the Democratization of University Innovation. *American Economic Review*, 98, no. 4, 1578–1590 (2008).

45. GeoNames, GeoNames. http://www.geonames.org/, Retr. (2012). (accessed February 7 2012).

46. Angeles Serrano, M., Boguñá, M. & Vespignani, A. Patterns of dominant flows in the world trade web *J. Econ. Interact. Coord.*, 2, 111–124 (2007).

47. ESRI, ArcGIS Desktop: Release 9.3. Environmental Systems Research Institute, Redlands, CA (2010).

48. Core Team, R. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. (2012).

Acknowledgments

This work has been partially funded by NSF CCF-1101743 and NSF CMII-1125959 awards. We acknowledge the American Physical Society for providing the data about Physical Review’s journals.

Author contributions

A.V., N.P. and Q.Z. designed research, Q.Z., B.G. and F.C. parsed data, Q.Z., N.P. and A.V. analysed data. All authors wrote, reviewed and approved the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/journal/v4/n1640/srep01640.

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

License: This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/

How to cite this article: Zhang, Q., Perra, N., Gonçalves, B., Ciulla, F. & Vespignani, A. Characterizing scientific production and consumption in Physics. *Sci. Rep.*, 3, 1640; DOI:10.1038/srep01640 (2013).