SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
APPLIED PHYSICS
SCIENTIFIC DATA
COMPUTATIONAL SCIENCE
ECOLOGICAL NETWORKS

Received
6 November 2012

Accepted
10 December 2012

Published
30 January 2013

Correspondence and requests for materials should be addressed to D.H. (dhelbing@ethz.ch)

Global Multi-Level Analysis of the 'Scientific Food Web'

Amin Mazloumian[1], Dirk Helbing[1], Sergi Lozano[1,2], Robert P. Light[3] & Katy Börner[3]

[1]Chair of Sociology, in particular Modeling and Simulation, ETH Zürich, CH-8092 Zürich, Switzerland, [2]IPHES, Institut Català de Paleoecologia Humana i Evolució Social, C/Marcelí Domingo s/n, 43007 Tarragona, Spain, [3]Cyberinfrastructure for Network Science Center, School of Library and Information Science, Indiana University, 10th Street & Jordan Avenue, Bloomington, IN 47405, USA.

We introduce a network-based index analyzing excess scientific production and consumption to perform a comprehensive global analysis of scholarly knowledge production and diffusion on the level of continents, countries, and cities. Compared to measures of scientific production and consumption such as number of publications or citation rates, our network-based citation analysis offers a more differentiated picture of the 'ecosystem of science'. Quantifying knowledge flows between 2000 and 2009, we identify global sources and sinks of knowledge production. Our knowledge flow index reveals, where ideas are born and consumed, thereby defining a global 'scientific food web'. While Asia is quickly catching up in terms of publications and citation rates, we find that its dependence on knowledge consumption has further increased.

Paper and citation counts are the 'official currency' in science and are widely used to assess the productivity and impact of authors, institutions, and scientific fields[1–5]. Many academic rankings focus on numbers $P(t)$ of publications in leading journals and citations rates $C(t)$, i.e., on knowledge production and consumption over time t. Examples are rankings of people, institutions, cities, or journals[6–9]. They show that new powers such as China and Brazil have recently emerged on the global scientific landscape[10]. Extrapolating these trends, it seems that the USA and Europe might lose their academic leadership.

However, academic leadership requires one to be first to publish a paper and others to cite the ideas. Simple counts of publications and citations of an entity (be it an author, institution, city, geographic area, journal, or scientific field) do not reveal who cites whom (thereby consuming knowledge from others), and who is cited (i.e., who produces knowledge consumed by others). The network-based approach proposed here assumes the existence of a 'scientific food web' that interconnects academic entities via knowledge flows.

A network perspective is important, because in many complex systems (such as the scientific ecosystem), interaction effects can be more relevant for the resulting system behavior than the properties of the interacting entities themselves. For example, it has been shown that author teams manage to be more successful than single authors[11–14]. The social, network-based character of knowledge diffusion underlines this perspective as well[15–17].

Compared with other ecosystems[18,19], an entity in the *scientific* food web is considered to be particularly *successful* ('*fit*'), if its knowledge is *consumed* (cited) more than expected. The analogy to ecosystems is chosen here to pronounce the mutual interdependencies and synergy effects in knowledge creation, since the production of new knowledge is nourished by the previous existence of relevant knowledge sets and their recombination. This is in line with research that uses the concept of ecosystems to shed new light on financial markets[20] and the evolution of national economies[21].

In previous work, networks of scientific papers[22] were used to analyze the evolution of scientific fields[23], to study innovation diffusion[24,25] or clickstream patterns[26], and to model the emergence and development of scientific fields[27]. Moreover, knowledge diffusion has been mapped between 500 major U.S. academic institutions, using a 20-year dataset of 47, 073 *PNAS* papers[28]. Other research studied knowledge import patterns for the field of transportation[29]. Our current study goes significantly beyond this by proposing and validating a new network-based index measuring higher-than-expected knowledge flows, which can be consistently applied on multiple levels. We demonstrate this by evaluating 13 million papers to identify global trends of knowledge diffusion at the level of continents, countries, and cities.

Results

We have analyzed the 80 million citations between 13 million papers published in the time period 2000 to 2009, as recorded in Thomson Reuters Web of Science (WoS). As the interaction of geographic locations is of particular interest, we have geolocated the papers using the first authors' postal address. (The addresses of the other authors are often not available in this database.)

To measure the knowledge flow between geographic locations or areas, collectively referred to as entities i, we proceed as follows: Let C_{ij} be the number of citations, which papers produced by entity j receive from papers by entity i in the time period under consideration. Then, $C_j = \Sigma_i C_{ij}$ is the total number of citations that entity j receives from all entities. $R_i = \Sigma_j C_{ij}$ is the total number of references listed in papers produced by entity i, citing other papers in our dataset. $R = \Sigma_i R_i$ is the total number of references pointing to other papers. P_i denotes the number of papers produced by entity i and $P = \Sigma_i P_i$ the total number of papers generated during the time period of consideration.

In order to assess the significance of knowledge flows, we need some kind of baseline scenario to compare with. Let us assume all papers would have the same capacity to attract citations. In such a case, the references listed in papers of entity i would cite the papers published by entity j in a *proportional* way, and the expected number of citations from i to j would be $E(C_{ij}) = R_i P_j / P$. Consequently, the proposed network-based index

$$S_{ij} = \frac{C_{ij} - E(C_{ij})}{R_i} = \frac{C_{ij}}{R_i} - \frac{P_j}{P} \qquad (1)$$

measures the *excess citations* per reference (i.e., the relative surplus). The index quantifies the interactions between a finite number of papers, which are distributed over a fixed set of entities such as geolocations. Note that the above formula considers self-citations of entity i. The slightly modified network flow index

$$F_{ji} = \left(\frac{C_{ij}}{R_i - C_{ii}} - \frac{P_j}{P - P_i} \right) \frac{P_i}{P} \qquad (2)$$

with $F_{ii} = 0$ removes the effect of self-citations. Defining *(excess) knowledge flows* from entity j to i in this way, the weighting factor P_i/P takes into account the volume of papers contributing to them, and the formula has the favorable mathematical properties $-1 \leq F_{ij} < 1$ and $\Sigma_i F_{ij} = 0$ This makes the index values easy to interpret: A positive flow indicates a surplus, i.e., an entity is cited more often than expected. A negative flow indicates a deficit, i.e., the entity is cited relatively little compared to the number of papers it produces. A neutral knowledge flow is not necessarily a sign of academic inactivity, but indicates that an entity receives the number of citations expected on average.

We now define the index of *(scientific) fitness*

$$K_i = \sum_j F_{ij} \qquad (3)$$

by summing up the (excess) knowledge flows from an entity i to all other entities j. It measures how much the consumption of knowledge created by entity i exceeds the statistical expectation.

$-1 \leq K_i < 1$ and $\Sigma_i K_i = 0$. It becomes negative if entity j cites other publications more frequently than expected, while a positive value indicates that j is a net creator of knowledge. Entities with a negative overall knowledge flow are referred to as *'knowledge sinks'*, while those with positive knowledge flow are called *'knowledge sources'*. As entities with low academic activity rate around $K_i = 0$, fitness does not measure academic strength, but the likely ability to thrive, if the consumption of external knowledge would be reduced. In other words, scientific fitness, as defined above, measures the resilience to the reduction of external inputs of knowledge.

To assess the plausibility of our new indices, we create rankings of geolocations based on the number of papers P_i, the number of citations C_i, the number of citations per paper C_i/P_i, and the fitness K_i. In order to have reliable index values (based on the statistical law of large numbers), we consider only entities with more than $P_i = 500$ publications in the investigated time period (848 geolocations fit that criterion in both considered time periods).

Our analysis shows the following: (1) The number of papers P_i and citations C_i are largely determined by the size of a country or city (see Tables S1 [countries] and S2 [cities]). Industrial countries perform better, but emerging scientific powers such as China and Brazil are catching up quickly (see Fig. 1). (2) The number of citations per paper, C_i/P_i is particularly high in countries such as Switzerland and The Netherlands, while the average performance of big countries seems to be pulled down by a large number of poorly performing academic institutions (see Table S1). The related city ranking appears to be sensitive to particularities such as the research focus of an institution. (3) Rankings based on *fitness* for countries (see Table S1) *and* for cities (Table S2) confirm known knowledge-producing areas in the world (see Fig. 2).

A closer look at the knowledge flows between different areas of the world shows that, despite the phenomenal growth of scientific productivity in Asian countries (see Fig. 1), the dependence on knowledge produced in the North America and Europe has further

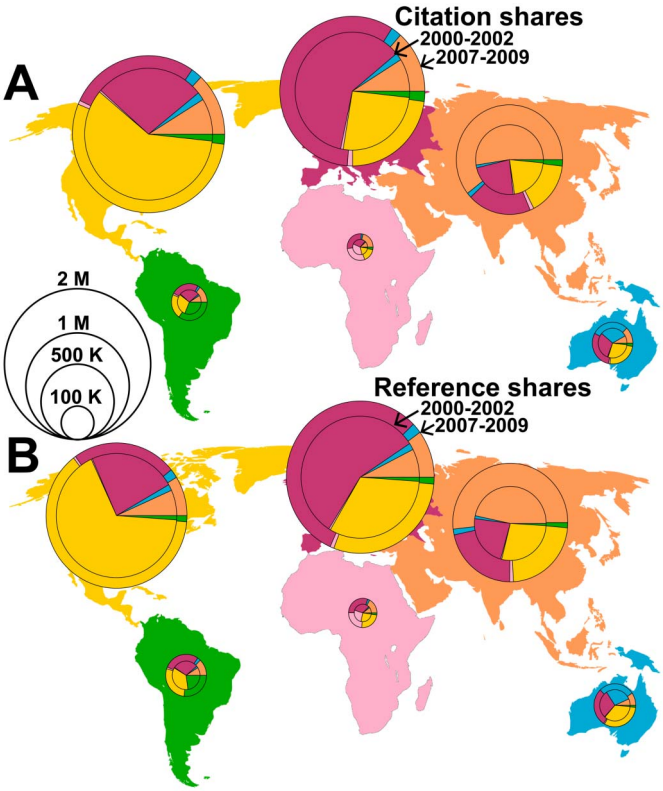

Figure 1 | World map of knowledge production and consumption in 6 major geographic areas of the world (North America, South America, Europe, Asia, Australia and Africa). Circle size reflects the number of papers P_i produced by the corresponding entities i. The inner circle is for 2000–2002, the outer one for 2007–2009. The size of the pies represents (A) the relative proportion of citations C_i that the entities earned in the 6 geographic areas, (B) similar for references R_i recorded in the Thomson Reuters Web of Science database. The number of papers and citations have increased over time in all geographic areas, but their shares of references and citations have changed. For example, Asia reaches higher shares recently, characterizing it as an emergent scientific power, which has become almost comparable to North America or Europe. Note that, in the three leading knowledge producing areas, the majority of references cites papers published in the same geographic area, i.e., proximity matters.

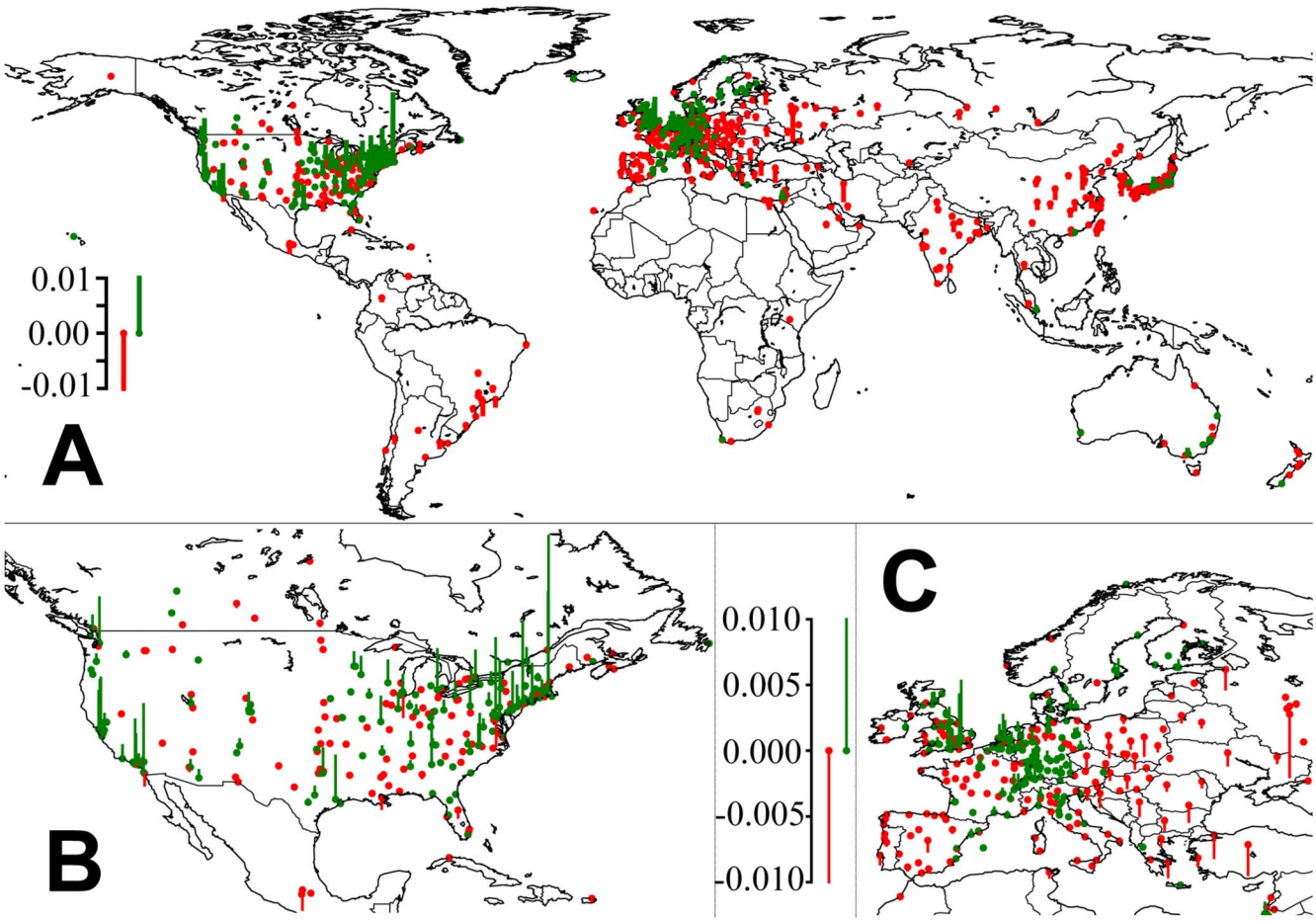

Figure 2 | World map of the greatest knowledge sources and sinks, based on our scientific fitness index. Green bars indicate that the number of citations received is over-proportional, red that the number of citations received is lower than expected (according to a homogeneous distribution of citations over all cities that have published more than 500 papers). It can be seen that most scientific activity occurs in the temperate zone. Moreover, areas of high fitness tend to be areas that are performing economically well (but the opposite does not hold).

increased (see Fig. 3). Note that the bars in Fig. 3 do not measure the distance between competitors, but the rate at which this distance changes. That is, as long as the bar is green, the distance between competitors grows, while the rate at which the distance increases shrinks, if the bar gets shorter.

Examining Fig. 3 more closely, the top row shows that North America is a major source of the knowledge that is consumed in Europe and Asia. Nevertheless, the excess knowledge flows from North America and Europe have decreased in the past decade. South America, in contrast, is improving its performance, while Africa's scientific activity is still on a low level (see also Fig. 1).

Discussion

In conclusion, our study addresses the fact that indices exclusively oriented at knowledge production or consumption (such as numbers of papers or citations), do not measure the most crucial property of the ecosystem of science, which is knowledge *exchange*. Given that the creation and diffusion of knowledge largely depend on social networks[30,31], and given the large relevance of network theory in many scientific areas, we believe that classical, *node-based indices* must be complemented by *network-based indices*. We, therefore, expect that there will be a whole new class of network-based scientific indices *besides* betweenness centrality[32] and PageRank[33] to characterize the scientific ecosystem in the future.

Here, we propose to measure scientific leadership by a network-based index that quantifies excess knowledge consumption by others. For knowledge leadership, the position of an entity in the scientific food web is crucial. It is important to understand whether a scientific entity is citing or being cited ('consumed'), and whether one is first to publish an idea or second. Our definition of knowledge flows captures the essence of this and has favorable mathematical properties. Our empirical analysis with *Web of Science* data reveals that the consumption of knowledge from North America, Europe, and China decreases relative to their production, while South America is improving its position.

The network-based *knowledge flow index* has the favorable property that it allows one to derive a related node-based index, which we call the *fitness* of a scientific entity. The corresponding *fitness ranking* is compatible with other science rankings of cities (see Fig. 4). However, our index has the advantage that everyone with access to citation data can measure it, since it does not require surveys or Web analytics. Notably, the same knowledge flow and fitness indices can also be applied to scientific fields and subdisciplines.

Methods

Dataset. Journal paper records for the years 2000–2009 were retrieved from Thomson Reuters' Web of Science (WoS) data. The dataset comprises about 13 million journal papers and 80 million citations between them (citations of non-WoS papers are not included in the dataset). While the WoS dataset has inherent limitations, e.g., does not include book or most conference publications and it is dominated by English publications, WoS is still considered to be the standard dataset to run global citation analyses accross all areas of science.

Determination of geographic location. All papers were geolocated with a success rate of 91.8%, using the Yahoo! geocoder in the Science of Science (Sci2) Tool. We found 47,333 unique geo-locations (latitude-longitude pairs) and 80 million citations

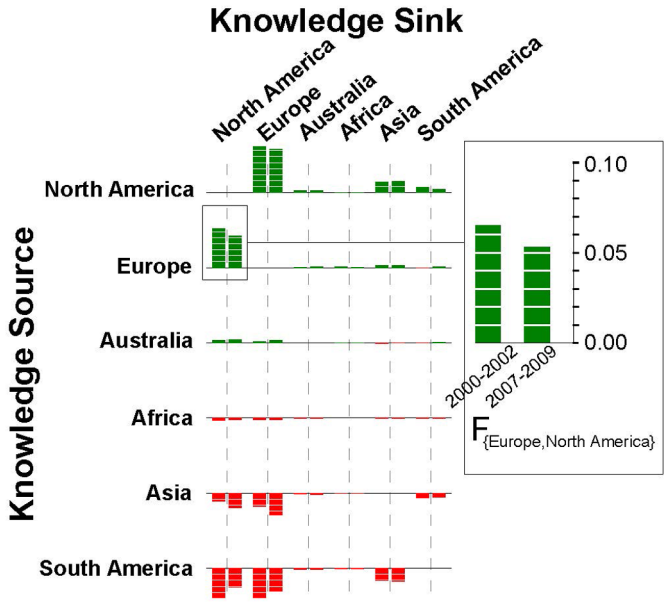

Figure 3 | Relative knowledge flows between major geographic areas. North America and Europe exceed the expected number of citations to all geographic areas by far. South American and Asian papers are less cited than expected. Note that the bars do not measure the distance between competitors, but the rate at which this distance changes. That is, as long as the bar is green, the distance between competitors grows, while the rate at which the distance increases shrinks, if the bar gets shorter. For example, South America has significantly reduced the pace at which the knowledge gap with regard to its competitors increases. Europe and North America are the two most strongly coupled knowledge-producing areas in the world. However, North America has considerably higher relative knowledge flows to South America and Asia than Europe.

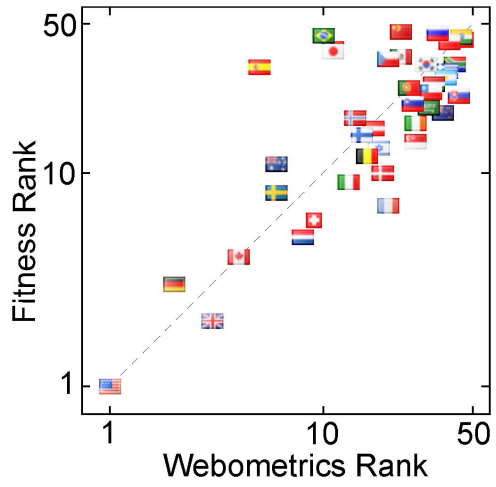

Figure 4 | Comparison of our fitness-based ranking of countries with the Webometrics ranking. The Webometrics rank measures the visibility of academic institutions in terms of web links to their web domains (see http://www.webometrics.info/Distribution_by_Country.asp). However, it does not measure relative citations flows. The Webometrics ranking of 2012 includes 45 countries, while our fitness values are determined for the 46 countries, which produced more than 1,000 Web of Science-listed papers per year. The figure shows the 40 countries, which appeared in both rankings. The United State is first in both rankings. The other countries are distributed around the diagonal, indicating that both rankings are pretty much consistent with each other.

between them. From these, we identified all science locations with more than 500 publication during the period under consideration. Self-citations were excluded.

Each paper was geolocated, using the address field of the corresponding author (C1), or reprint address fields (RP) when no corresponding author was specified. For the selected address fields, we used the state field (NP), if the country field (NU) pointed to the US, and the city field (NY) otherwise. 12 million journal records (91.8% of all publications) were successfully geo-located with 80,793 unique city/state-country pairs. The other 8.2% of the journal records either did not have the required address fields, or the Yahoo! Geocoder detected their address fields as invalid. The geocoder provided latitude and longitude values with 14 digit decimals.

We aggregated the neighboring locations by rounding latitude and longitude values to 2 digits. This resulted in 47,333 unique geo-locations (latitude-longitude pairs). Finally, the top-550 major geolocations were identified, using the number of raw citation and reference counts.

Identification of trends from geolocated records. To identify trends, the 10-year dataset was divided into 8 partially overlapping time slices: 2000–2002, 2001–2003 … 2007–2009. For each time slice, a network was extracted that comprises citations of papers published in the last year, received from papers published within the whole time slice. For example, time slice 2000–2002 considers all citations of recorded papers published in 2002, received from papers published in 2000–2002. Therefore, only early citations to papers are captured, i.e., those made shortly after publication. The total number of these early citations is 31 million (which amounts to 39% of the total number of citations from and to papers published within the whole 10-year time period).

1. Garfield, E. Citation indexes for science: A new dimension in documentation through association of ideas. *Science* **122**, 108–111 (1955).
2. Cronin, B. *The Citation Process: The Role and Significance of Citations in Scientific Communication* (Taylor Graham, London, 1984).
3. 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).
4. Radicchi, F., Fortunato, S., Markines, B. & Vespignani, A. Diffusion of scientific credits and the ranking of scientists. *Phys. Rev. E* **80**, 056103 (2009).
5. Petersen, A. M., Jung, W.-S., Yang, J.-S. & Stanley, H. E. Quantitative and empirical demonstration of the Matthew effect in a study of career longevity. *Proc Natl Acad Sci USA* **108**, 18–23 (2011).
6. Bornmann, L., de Moya Anegón, F. & Leydesdorff, L. The new excellence indicator in the world report of the SCImago institutions rankings 2011. *J. Informetrics* **6**, 333–335 (2012).
7. Bollen, J., Van de Sompel, H., Hagberg, A. & Chute, R. A principal component analysis of 39 scientific impact measures. *PLoS ONE* **4**, e6022 (2009).
8. Bornmann, L. & Leydesdorff, L. Which cities produce worldwide more excellent papers than can be expected? A new mapping approach-using Google Maps-based on statistical significance testing. *J. Am. Soc. for Information Science and Technology* **62**, 1954–1962 (2011).
9. Mazloumian, A., Eom, Y.-H., Helbing, D., Lozano, S. & Fortunato, S. How citation boosts promote scientific paradigm shifts and Nobel prizes. *PLoS ONE* **6**, e18975 (2011).
10. Zhou, P. & Leydesdorff, L. The Emergence of China as a leading nation in science. *Research Policy* **32**, 83–104 (2006).
11. Guimera, R., Uzzi, B., Spiro, J. & Amaral, L. A. N. Team assembly mechanisms determine collaboration network structure and team performance. *Science* **308**, 697–702 (2005).
12. Börner, K., Dall'Asta, L., Ke, W. & Vespignani, A. Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity* **10**, 58–67 (2005).
13. Wuchty, S., Jones, B. F. & Uzzi, B. The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007).
14. Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N. & Malone, T. W. Evidence for a collective intelligence factor in the performance of human groups. *Science* **330**, 686–688 (2010).
15. Valente, T. W. *Network Models of the Diffusion of Innovations* (Hampton Press, Cresskill, NJ, 1995).
16. Lazer *et al.* Life in the network: the coming age of computational social science. *Science* **323**, 721–723 (2009)
17. Christakis, N. A. & Fowler, J. H. *Connected* (Back Bay Books, 2011).
18. May, R. M. Ecology - The structure of food webs. *Nature* **301**, 566–568 (1983).
19. Gross *et al.* Generalized models reveal stabilizing factors in food webs. *Science* **325**, 747–750 (2009).
20. May, R. M. & Haldane, A. G. Systemic risk in banking ecosystems. *Nature* **469**, 351–355 (2011).
21. Hidalgo, C. A., Klinger, B., Barabasi, A.-L. & Hausmann, R. The product space conditions the development of nations. *Science* **317**, 482–487 (2007).
22. de Solla Price, D. J. Networks of scientific papers. *Science* **149**, 510–515 (1965).
23. Boyack, K. W., Börner, K. & Klavans, R. Mapping the structure and evolution of chemistry research. *Scientometrics* **79**, 45–60 (2009).
24. Goffman, W. & Newill, V. A. Generalization of epidemic theory: An application to the transmission of ideas. *Nature* **204**, 225–228 (1964).

25. Goffman, W. Mathematical approach to the spread of scientific ideas-the history of mast cell research. *Nature* **212**, 449–452 (1966).
26. Bollen *et al.* Clickstream data yields high-resolution maps of science. *PLoS ONE* **4**, e4803 (2009).
27. Bettencourt, L. M. A., Kaiser, D. I., Kaur, J., Castillo-Chavez, C. & Wojick, D. E. Population modeling of the emergence and development of scientific fields. *Scientometrics* **75**, 495–518 (2008).
28. Börner, K., Penumarthy, S., Meiss, M. & Ke, W. Mapping the diffusion of information among major U.S. research institutions. *Scientometrics* **68**, 415–426 (2006).
29. Wu, J. Geographical knowledge diffusion and spatial diversity citation rank. *Scientometrics* **94**, 181–201 (2013).
30. Newman, M. E. J. The structure of scientific collaboration networks. *Proc Natl Acad Sci USA* **98** 404–409 (2001).
31. Lee, C. & Cunnigham, P. *The geographic flow of music* (available at arXiv:1204.2677) (2012).
32. Leydesdorff, L. & Rafols, I. Indicators or the interdisciplinarity of journals: Diversity, centrality, and citations. *J. Informetrics* **5**, 87–100 (2011).
33. Chen, P., Xie, H., Maslov, S. & Redner, S. Finding scientific gems with Google's PageRank algorithm. *J. Informetrics* **1**, 8–15 (2007).

Acknowledgements

The authors would like to thank Vincent Larivire and Kevin W. Boyack for expert comments on the paper. This work started while KB was a Visiting Scientist at the ETH Zürich in 2011. We acknowledge Thomson ISI as required by the terms of use of our WoS data license. This work was partially supported by the Future and Emerging Technologies programme FP7- COSI-ICT of the European Commission through project QLectives (grant no.: 231200), the James S. McDonnell Foundation, and the National Institutes of Health under award NIH U01 GM098959.

Author contributions

A.M. and D.H. developed the new index. Data analysis has been performed by A.M. and R.P. A.M. produced the figures. The concept of the study was developed and the supervision was performed by K.B., D.H. and S.L. The manuscript was written jointly by all authors.

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: Mazloumian, A., Helbing, D., Lozano, S., Light, R.P. & Börner, K. Global Multi-Level Analysis of the 'Scientific Food Web'. *Sci. Rep.* **3**, 1167; DOI:10.1038/srep01167 (2013).