Open Source Software Collaboration Networks on GitHub

Brandon Kramer[1,*], Gizem Korkmaz[1], Bayoán Santiago Calderón[1], and Carol Robbins[2]

[1]University of Virginia, USA
[2]National Science Foundation, USA
[*]kb7hp@virginia.edu

ABSTRACT

Please find our extended abstract with preliminary findings below.

Extended Abstract

Open Source Software (OSS), defined by Open Source Initiative[1], is computer software with its source code made available with a license in which the copyright holder provides the rights to study, change, and distribute the software to anyone and for any purpose. OSS is everywhere, both as specialized applications nurtured by devoted user communities, and as digital infrastructure underlying platforms used daily by millions. It is developed, maintained, and extended both within and outside of the private sector, through the contribution of independent developers as well as people from universities, government research institutions, businesses, and nonprofits. Examples include Linux operating system, Apache server, and R.

While the extent and impact of OSS is currently unknown, recent estimates suggest that its magnitude is significant. For example, Apache is estimated to hold the largest market share of domains (35%) and active websites (41%) as of November 2017[2]. The Apache server, developed with federal and state funds at the National Center for Supercomputing Applications at the University of Illinois, is estimated to be equivalent to between 1.3% and 8.7% of the stock of prepackaged software currently accounted for in U.S. private fixed investment[3].

Software developed in the business sector is generally accounted for in Gross Domestic Product (GDP) measures as an intangible asset or intellectual property product. Despite its ubiquity and extensive use, reliable measures of the scope and impact of OSS developed outside of the business sector are scarce[3]. Activities around OSS development, a vital component of science activity, are not well-measured in existing federal statistics on innovation. The creation and use of these modifiable software tools highlight an aspect of technology diffusion and flow that is not captured in science and technology indicators[4,5].

Collaboration and sharing are essential features of OSS projects; many are developed in free repositories and information embedded in these repositories, including the code, contributors, and development activity, is publicly available. Current and comprehensive survey data do not exist for the contributions of OSS. However, as OSS is disseminated online, a wealth of information is available in the code and headers of the software programs themselves. Availability of these data — information on interactions between projects as well as between contributors — brings the possibility of developing network analysis methods to study the impact of these projects and contributors, and the diffusion of OSS innovation through the OSS ecosystem. Leveraging techniques used in bibliometrics and patent analysis[6–8], measures of creation and use of OSS would complement existing science and technology indicators on peer-reviewed publications and patents that are calculated from databases covering scientific articles and patent documents[5,7]. We aim to show how these data can be used to shed light on OSS ecosystem – interactions between OSS developers, main actors, the structural features of these networks and how they change over time – that is currently not well understood[9–11].

In this paper, we use data from GitHub, the largest platform with 31 million users and developers worldwide, obtaining information about OSS projects. We collect 5.2 million project repositories, containing metadata such as author, license, commits (approved code edits), and lines of code. We focus on the interactions among projects and among developers in the OSS community and represent these interactions using networks of contributors (through collaborations between developers). We use network analysis methods to study collaboration patterns, its dynamics, and influential international actors. Illuminating the interactions in the OSS community using network analysis methods, the research questions that we explore include (i) What are the patterns of collaboration among OSS developers? (ii) What are the structural differences between scientific collaboration networks (e.g., coauthorship networks) and OSS developer networks? (iii) What are the dynamics of international OSS collaboration networks? (iv) What are the most influential countries in terms of network centrality and effort invested in OSS development?

Related Work

Open source software development is closely related to collaborative production in academic research. The recent trend in many disciplines (e.g., economics, sociology, physics, management, mathematics) is to capture the patterns of collaboration using a network: a web of collaborative interactions in which two researchers are linked if they work together on a project or coauthor a paper (e.g.,[12–17]). There exists a growing literature, both empirical and theoretical, on collaboration and coauthorship networks (e.g.,[13,14,18–22]). A review of these studies can be found at[23]. This work is also related to the studies that focus on the relationship between factors of productivity and the network properties. Some of these studies suggest that productivity is affected by both the number of links and network structure due to communication effectiveness and exchange and flow of information[15,24,25]. Moreover, the local network structure and the centrality of an individual in the network will affect the outcome[19,26–28]. Empirical studies have found correlation between the centrality measures in coauthorship networks and research performance of authors and institutions, e.g.,[29–34].

In recent years, international collaboration has become an increasingly important factor in shaping scientific productivity[35–37]. In fact, there are a number of number of notable examples of how the US government has engaged in international collaborations that serve to generate innovation in science and technology[38]. Leydesdorff and colleagues[39] shows that international collaboration positively impacts the impact of scholarly publications, but also finds that government funding has a negligible influence on citation impact. Yet, some still express concern that when US scientists engage in international collaboration it diminishes the US's advantage in the global science, technology and innovation industry[40]. Using international co-author networks, Wagner et al.[37] find this network is a self-organizing system largely independent from national, political or geographical systems, although when they conduct a country-specific analysis they see considerable variation. More specifically, this pattern is not necessarily a matter of geopolitical power as much as it is matters of being geographically-isolated (Japan, Australia) and/or being underdeveloped (India, Egypt). This means that international processes are governing in top-down fashion while national mechanisms are governing other countries from the bottom-up[37,39,41]. We are interested in exploring whether the patterns of international collaborations on OSS development differ from those on co-authorship, and who the major actors in the OSS ecosystem are.

Data and Methods

Keller et al. (2018)[9] describes the overall approach used here to explore data sources beyond surveys to improve and extend indicators of science and engineering activity and of innovation. This approach includes structured processes to discover, acquire, profile, clean, link, explore the fitness-for-use, and statistically analyze the data. Here we gather and use publicly available metadata about individual OSS projects on GitHub and their contributors and organizations, as well as information within the code.

We define the OSS universe as the repositories on GitHub that have an OSI-approved license[1]. OSS licenses define limitations of use and provide developers with rights over their work while promoting the dispersion of free, accessible code. OSS licenses allow software to be freely used, modified and shared. First collected a complete catalog of every public repository on GitHub that had an OSI-approved machine detectable license. Secondly, for each repository which was not a fork or a mirror (i.e., original content), we collected the commit data. The commit data include the user, when it was authored/committed, and lines added/deleted. This information allows us to construct a table for each user to repository (i.e., the base branch), when and how much direct contributions occurred. All the data was collected through the GitHub v4 GraphQL API for a time coverage since beginning of GitHub through August 15, 2019.

Information about users and organizations was collected through the GHTorrent project [42] which regularly collects public scope data from GitHub using the GitHub v3 RESTful API. The original dataset was comprised of 2,143,407 distinct contributors and 5,188,818 distinct repositories. The relevant tables included information about the users and organizations. Some variables included username, website, emails for organizations, and company and location for users. The location attributed to users for our purposes are derived from the free form user-provided field which differs from GitHub system which is enhanced by access to geolocation by IP addresses.

In this paper, we use a subset of this larger dataset to conduct analyses on (1) contributor networks with valid country codes attached to user logins and (2) country level contribution networks. The contributor networks are comprised of 717,231 users (34% of the total) that had location information attached to their user login information and 2,719,090 repositories that were linked to users that had valid country codes. These users and repos ranged from 2008 (GitHub's first year as an organization) to 2018. In order to construct contributor networks, we first projected bipartite matrices of contributors and repositories to a single-mode network where nodes are contributors and the ties between those nodes represent commits to the same repository. To gain a more nuanced understanding of network evolution over time, we ran various network descriptive measures for each individual year (2008, 2009, 2010, etc.) as well as descriptives for each cumulative intervals in the network's evolution (i.e., 2008, 2008-09, 2008-10, etc.). These network descriptives include the network's node count, edge count, weighted edge

count, the density, transitivity, modularity as well as the number of isolates, dyads and triads using the igraph package[43] in the open-source software program R (citation). While these measures were calculated using the default functions in igraph, it is of note that we used Clauset and colleagues' (2004) function to calculate modularity, which was then scaled to the size of the logged network size in accordance with previous sociological scholarship reporting longitudinal changes in network modularity.

We also constructed networks comprised of country-to-country collaborations. These networks essentially link the country codes to the contributor networks just described and then reduce those networks into country-to-country collaborations instead of individual GitHub users. As we did with the contributor networks, we ran descriptive network statistics for the cumulative networks from 2008-2018 using the igraph package[43], which are reported in Table 2. To better understand how countries engage in international collaboration in the context of open-source software, we have also added some network visuals that were produced in the Gephi open-source software[44].

Preliminary Results

We observe that the number of GitHub repositories with an OSS license has been increasing rapidly as shown in Figure 1. In 2008, there were around 3.5K repositories, which increased to 176K in 2013, and by 2018, there were 1.3M repositories. Although 87 OSI-approved licenses exist, the top 13 licenses contain >99% of OSS repositories on GitHub. BSD (Berkeley Software Development) and MIT licenses were developed in these universities[45,46]. GPL is the GNU Public License developed by Richard Stallman and the Free Software Foundation[47,48]. Artistic, PHP, Python, and Apache are project-specific licenses that conform to OSI standards . The five licenses (presented in Figure 1) together comprise about 93% of all OSS on GitHub. The most popular licenses are MIT, Apache, GPL and BSD. The MIT license that allows developers to use the code for any purpose is the most common OSS license (55%). The GPL license grants the ability to use the respective code under the stipulation that derivative work remains open source. The five licenses (presented in Figure 1) together comprise about 93% of all OSS on GitHub.

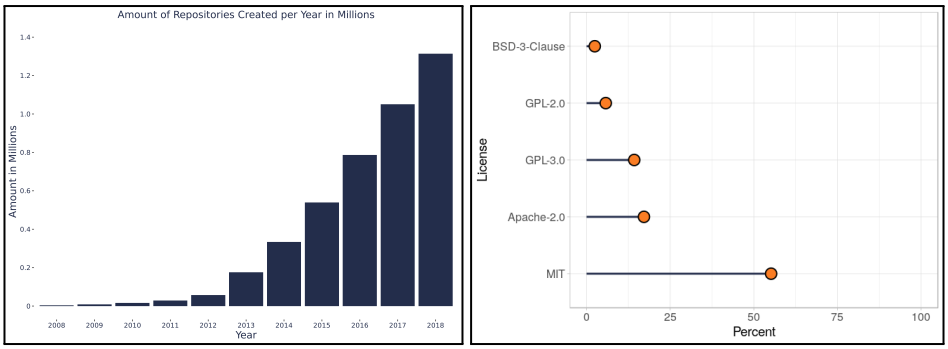

Figure 1. Number of OSS Repositories on GitHub with OSI-Approved Licenses.

Contributor Networks

We develop the contributor networks for the years 2008 – 2018 using the data described above. In order to understand changes in the evolution of the GitHub collaborator network, we graphed the raw totals of nodes, edges, weighted edges, total commits, isolates, dyads and the size of the largest k-core component over time (see Table 2).

Year	Nodes	Edges	Commits	Avg Deg	Density	Transitivity	Modularity
2008	5.3K	58.0K	3.9M	22	0.0040	0.49	0.42
2008-09	12.3K	195.5K	10.3M	32	0.0026	0.48	0.49
2008-10	24.5K	729.0K	19.0M	59	0.0024	0.59	0.58
2010-11	48.3K	232.2K	33.0M	96	0.0020	0.65	0.68
2010-12	90.6K	5.8M	53.7M	127	0.0014	0.65	0.72
2010-13	197.1K	12.8M	82.9M	130	0.0007	0.60	0.74
2010-14	372.6K	25.1M	123.0M	135	0.0004	0.52	0.75
2010-15	625.6K	56.8M	175.3M	182	0.0003	0.59	0.77
2010-16	977.0K	124.1M	240.7M	87	0.0003	0.74	0.78
2010-17	1.4M	178.9M	306.7M	97	0.0001	0.71	0.78
2010-18	2.0M	275.9M	370.8M	104	0.0001	0.67	0.79

Table 1. Network Descriptives for GitHub Contributor Networks for Years 2008 - 2018.

When we look at the cumulative network data (Figure 2), we see clear exponential growth trends in the raw number of nodes, edges, weighted edges, total commits, isolates and dyads over time. While there is still growth when examining the raw totals in the year-by-year graphs (Figure 2), we do see that this exponential growth is interrupted starting in 2016. While it is not entirely clear why there is a drop in edges/dyads in 2017, we are confident that this trend coincides with what actually happened on GitHub's platform given that the node count continues to rise at a consistent rate during this period.

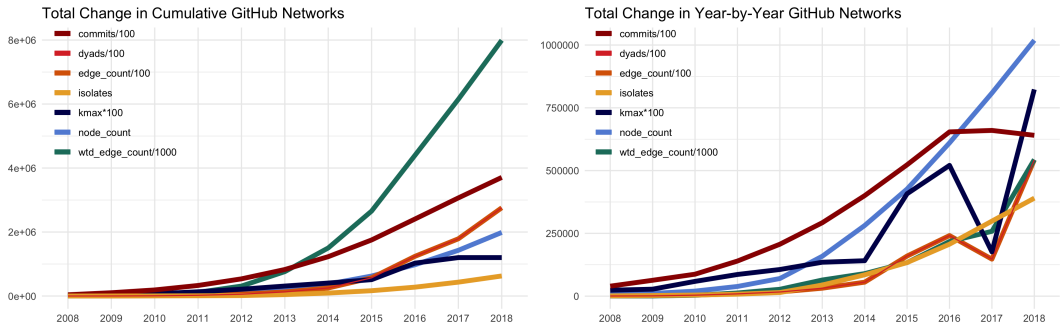

Figure 2. Total change in descriptive measures over time.

Moving to the trends in edges and dyads, we see that these growth rates follow fairly similar trends to node and isolate growth. The notable exception in the trends is that there is a marked increase in edges in 2015 and 2016. This increase in edge/dyad formation suggests that the overall network is becoming more densely connected than the number of new nodes entering the network. Given that this spike occurs 2-3 years after the rise in isolates discussed in the previous paragraph, it follows that an increase in edge/dyad formation is essentially a "lag effect" of these previously isolated users becoming structurally integrated into the network.

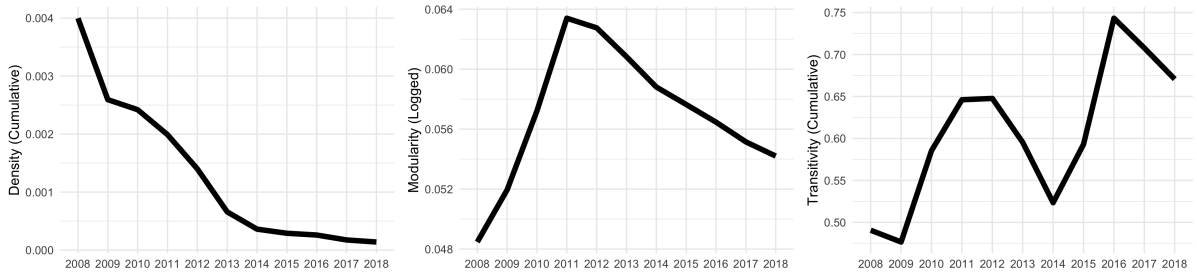

Figure 3. Descriptives for cumulative longitudinal network analysis (From left to right: A. Density, B. Modularity, C. Transitivity).

Figure 3 shows trends in descriptive statistics for the cumulative networks, including density, transitivity, and modularity. First, the density of the network follows a near linear trend downward from 0.004 in 2008 to 0.0001 in 2018 (Figure 3A). This trend suggests that the network, though sparsely connected at its onset, becomes even more disconnected over time as novel users join the network. In other words, the growth of users outpaces the number of links made between GitHub users over the course of the decade. Figure 3B documents changes in network modularity over time, which has been scaled for the size of the network over time akin to sociological work on the topic[49]. The scaled modularity shows a sharp increase over the first three years of the network's history before a dramatic decrease from 2011 to 2018. This trend suggests that GitHub's early network formation was characterized by more distinct communities until 2011 when there is a pronounced shift toward network integration, which aligns with the decrease in modularity after that point in time.

Transitivity follows an even more complex, non-linear pattern over time (Figure 3C), which initially increases from 2009 to 2011 before declining from 2012 to 2014, rising to its highest point in 2016 before declining again in 2017. As a reminder, transitivity is a clustering coefficient - calculating the number of closed triplets in a graph over the number of possible triplets (i.e. the number of open and closed triplets in the graph). The non-linear trend that arises in the GitHub networks is difficult to explain, but our best estimation is that the change in transitivity is a function of (1) activity happening near the core of the graph and (2) the "lag effect" of isolates entering the network only to later become structurally integrated into the core of the network.

In this section, we highlighted three main findings. First, while the overall growth rate of the GitHub collaboration network has slowed over time, there are still clear patterns of exponential growth in this network from 2008-2018. Alongside this

growth, the density of the networks steadily declines over time. In contrast, the modularity and transitivity of the network both follow non-linear trends that appear to arise from "lag effects" in user integration into the network. While there is marked increase in users joining GitHub in 2013, these isolate users do not become structurally integrated into the network until 2015. In contrast, GitHub users seem to be less likely to form ties in 2017, which leads to a drop in transitivity from 2016-2018. Overall, these analyses show the importance of detailing year-by-year changes in contributor networks to see more nuanced, non-linear network dynamics and how they may potentially link to specific historical events in the history of open-source software development.

International Collaborations

In this section, we study the contributions by countries, their collaborations, and the major actors in the international OSS ecosystem. When looking to the country level network data, we see similar a pronounced increase in the number of nodes, edges and commits from 2008-2018 (see Table 2). Overall, this network is quite comparable to Wagner and colleagues'[37] paper on international citation networks, suggesting that international collaboration on GitHub has a similar structure to collaboration in other scientific contexts.

Year	Nodes	Edges	Commits	Avg Deg	Avg Wtd Deg	Diameter	Density	Transitivity	Modularity
2008	72	755	1,298,765	21	9,646	2	0.295	0.782	0.00069
2008-09	87	1,285	3,618,888	30	41,753	2	0.344	0.786	0.00166
2008-10	105	1,901	6,956,764	36	131,363	3	0.348	0.796	0.00033
2010-11	130	2,937	14,154,516	45	388,002	3	0.350	0.794	0.00060
2010-12	149	4,087	25,709,139	55	915,872	3	0.370	0.763	0.00106
2010-13	183	5,377	43,073,956	59	1,740,054	3	0.323	0.763	0.00099
2010-14	199	6,572	66,559,436	66	2,991,750	3	0.334	0.787	0.00097
2010-15	213	8,102	96,808,561	76	4,655,610	3	0.359	0.789	0.00091
2010-16	219	9,560	132,972,139	87	6,971,384	3	0.400	0.801	0.00089
2010-17	224	10,819	167,733,172	97	9,229,399	3	0.433	0.798	0.00097
2010-18	230	11,964	198,614,447	104	10,892,990	3	0.454	0.814	0.00134

Table 2. Network Descriptives for International Collaboration Networks for Years 2008 - 2018.

Table 3 presents top 10 countries with highest contributions based on % shares of kilo (thousand) lines of code (KLOC) in Table 3. KLOC is used in software engineering as the measure of effort, which is then used to calculate the cost of development for the software[50]. We observe that US contributed to the 15.6% of the KLOC, followed by China with 2.8%, and 2.4% United Kingdom. While the GHTorrent data only offers a sample, our data seems to be fairly representative with recently completed projects on the development of open source software across the world[51]. We also summarize the network centrality measures of the countries in Table 3, and we provide the network visual for the subgraph with top 15 nodes in Figure 4.

	OSS Contributions and % shares						Network Centrality			
Country	KLOC Added	share %	Num. commits	share %	Num. Contributors	share %	Degree	Closeness	Betweenness	Clustering
United States	73.3M	15.6	87.6M	21.2	211K	9.8	202	0.97	1092	0.58
China	13.1M	2.8	9.8M	2.4	53K	2.5	186	0.90	176	0.68
United Kingdom	11.3M	2.4	13.9M	3.4	39K	1.8	192	0.93	293	0.64
Germany	8.7M	1.9	13.6M	3.3	38K	1.8	191	0.92	331	0.64
Canada	6.2M	1.3	7.9M	1.9	26K	1.2	194	0.94	571	0.62
India	6.1M	1.3	4.2M	1.0	37K	1.7	187	0.91	344	0.66
France	5.8M	1.2	7.1M	1.7	24K	1.1	190	0.92	441	0.65
Netherlands	4.2M	0.9	8.3M	2.0	13K	0.6	191	0.92	273	0.64
Brazil	3.9M	0.8	3.1M	0.7	23K	1.1	182	0.89	122	0.70
Japan	3.1M	0.7	4.5M	1.1	15K	0.7	184	0.89	163	0.68
Total	471.1M	–	413 M	–	2.1M	–				
NA	286.3M	60.8	195.3M	47.3	1.4M	66.5				

Table 3. OSS Contributions by Country and Network Centrality Measures. Top 10 countries are listed.

In our country to country network, we also calculated the modularity of the graph using Clauset et al.'s[52] algorithm. These results reveal that the country collaboration networks is comprised of two large communities with a number of smaller communities. In Figure 5, we visualize these communities on a map. While the largest community (in blue) consist mostly of English-speaking countries as well as other prominent OSS contributors like China, Brazil and Russia, the countries in red seem to cluster based on geographical and linguistic factors. For example, a number of Spanish speaking countries in South and Central America cluster together while a number of countries in North Africa and the Middle East also cluster together (in red). Future work involves looking further into understanding the level of contribution that these periphery countries contribute to OSS more broadly.

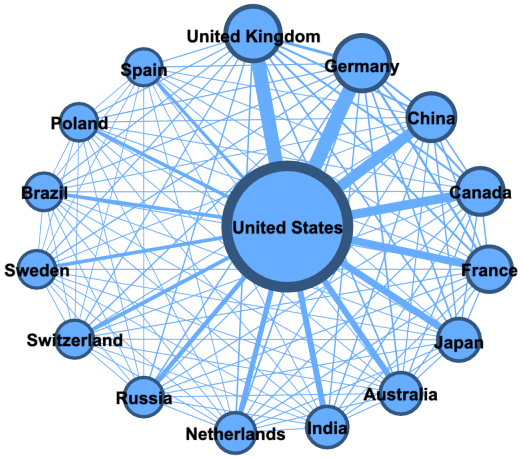

Figure 4. Network visual of top-15 collaborative countries on GitHub. In this graph, nodes represent countries while the weight of the tie corresponds to the number of times GitHub users from that country have contributed to the same GitHub repository.

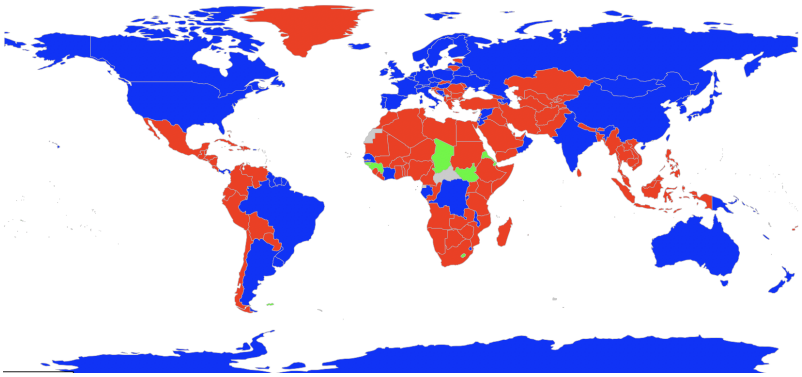

Figure 5. Worldmap of International OSS Collaborations. Colors represent the communities identified based on network modularity.

Acknowledgements

This material is based on work supported by U.S. Department of Agriculture (58-3AEU-7-0074). The authors acknowledge Research Computing at the University of Virginia for providing computational resources and technical support that have contributed to the results reported within this publication. We also acknowledge the Data Science for the Public Good Program participants Cong Cong, Calvin Isch, and Eliza Tobin.

References

1. Open Source Initiative. The open source definition (1998). https://opensource.org/osd.
2. Netcraft. Web server survey. https://news.netcraft.com/archives/2017/11/21/november-2017-web-server-survey.html (2017). Accessed: 2018-02-01.
3. Greenstein, S. & Nagle, F. Digital dark matter and the economic contribution of Apache. *Res. Policy* **43**, 623–631 (2014).
4. National Science Board. Science and engineering indicators (2018). NSB-2018-1. Alexandria, VA: National Science Foundation. https://www.nsf.gov/statistics/indicators/.
5. Hall, B. H. & Jaffe, A. B. Measuring science, technology, and innovation: A review. *Rep. prepared for Panel on Dev. Sci. Technol. Innov. Indic. for Futur.* (2012).
6. Rehn, C., Gornitzki, C., Larsson, A. & Wadskog, D. Bibliometric handbook for Karolinska Institutet (2014). https://kib.ki.se/sites/default/files/bibliometric_handbook_2014.pdf.

7. Science-Metrix. Bibliometrics and Patent Indicators for the Science and Engineering Indicators 2018. Technical Documentation (2018). http://www.science-metrix.com/en/methodology-report.

8. Osareh, F. Bibliometrics, citation analysis and co-citation analysis: A review of literature i. *Libri* **46**, 149–158 (1996).

9. Keller, S., Korkmaz, G., Robbins, C. & Shipp, S. Opportunities to observe and measure intangible inputs to innovation: Definitions, operationalization, and examples. *Proc. Natl. Acad. Sci. (PNAS)* **115**, 12638–12645 (2018).

10. Robbins, C. *et al.* The scope and impact of open source software as intangible capital: A framework for measurement with an application based on the use of R packages. In *NBER Conference on Research in Income and Wealth: Big Data for 21st Century Economic Statistics* (National Bureau of Economic Research, 2019). http://papers.nber.org/conf_papers/f111802.pdf.

11. Robbins, C. *et al.* Open source software as intangible capital: Measuring the cost and impact of free digital tools. In *The Sixth IMF Statistical Forum: Measuring Economic Welfare in the Digital Age: What and How?* (International Monetary Fund (IMF), 2018). https://www.imf.org/en/News/Seminars/Conferences/2018/04/06/6th-statistics-forum.

12. Goyal, S., Van Der Leij, M. J. & Moraga-González, J. L. Economics: An emerging small world. *J. political economy* **114**, 403–412 (2006).

13. Newman, M. E. Scientific collaboration networks. i. Network construction and fundamental results. *Phys. Rev. E* **64** (2001).

14. Bosquet, C. & Combes, P.-P. Do large departments make academics more productive? Agglomeration and peer effects in research (2013). CEPR Discussion Paper No. DP9401.

15. Moody, J. The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *Am. sociological review* **69**, 213–238 (2004).

16. Acedo, F. J., Barroso, C., Casanueva, C. & Galán, J. L. Co-authorship in management and organizational studies: An empirical and network analysis. *J. Manag. Stud.* **43**, 957–983 (2006).

17. Grossman, J. W. The evolution of the mathematical research collaboration graph. *Congr. Numerantium* 201–212 (2002).

18. Ductor, L. Does co-authorship lead to higher academic productivity? *Oxf. Bull. Econ. Stat.* **77**, 385–407 (2015).

19. Ductor, L., Fafchamps, M., Goyal, S. & van der Leij, M. J. Social networks and research output. *Rev. Econ. Stat.* **96**, 936–948 (2014).

20. Newman, M. E. The structure of scientific collaboration networks. *PNAS* **98**, 404–409 (2001).

21. Newman, M. E. Scientific collaboration networks. ii. Shortest paths, weighted networks, and centrality. *Phys. Rev. E* **64** (2001).

22. Newman, M. E. Coauthorship networks and patterns of scientific collaboration. *PNAS* **101**, 5200–5205 (2004).

23. Kumar, S. Co-authorship networks: a review of the literature. *Aslib J. Inf. Manag.* **67**, 55–73 (2015).

24. Burt, R. S. Structural holes versus network closure as social capital. In *Social capital*, 31–56 (Routledge, 2017).

25. Jackson, M. O. & Yariv, L. Diffusion of behavior and equilibrium properties in network games. *Am. Econ. Rev.* **97**, 92–98 (2007).

26. Menzel, H. & Katz, E. Social relations and innovation in the medical profession: The epidemiology of a new drug. *Public Opin. Q.* **19**, 337–352 (1955).

27. Calvó-Armengol, A. Job contact networks. *J. economic Theory* **115**, 191–206 (2004).

28. Banerjee, A., Chandrasekhar, A. G., Duflo, E. & Jackson, M. O. The diffusion of microfinance. *Science* **341**, 1236498 (2013).

29. Lee, S. & Bozeman, B. The impact of research collaboration on scientific productivity. *Soc. Stud. Sci.* **35**, 673–702 (2005).

30. Yan, E. & Ding, Y. Applying centrality measures to impact analysis: A coauthorship network analysis. *J. Assoc. for Inf. Sci. Technol.* **60**, 2107–2118 (2009).

31. Yan, E., Ding, Y. & Zhu, Q. Mapping library and information science in china: A coauthorship network analysis. *Scientometrics* (2010).

32. Ye, Q., Li, T. & Law, R. A coauthorship network analysis of tourism and hospitality research collaboration. *J. Hosp. & Tour. Res.* **37**, 51–76 (2013).

33. Uddin, S., Hossain, L. & Rasmussen, K. Network effects on scientific collaborations. *PloS one* **8**, e57546 (2013).

34. Abbasi, A., Altmann, J. & Hossain, L. Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *J. Informetrics* **5**, 594–607 (2011).

35. Wagner, C. S. & Leydesdorff, L. Network structure, self-organization, and the growth of international collaboration in science. *Res. policy* **34**, 1608–1618 (2005).

36. Wuchty, S., Jones, B. F. & Uzzi, B. The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007).

37. Wagner, C. S., Park, H. W. & Leydesdorff, L. The continuing growth of global cooperation networks in research: A conundrum for national governments. *PLoS One* **10**, e0131816 (2015).

38. Lyons, E. E. *et al.* How collaborating in international science helps america. *Sci Dipl* **5**, 2 (2016).

39. Leydesdorff, L., Bornmann, L. & Wagner, C. S. The relative influences of government funding and international collaboration on citation impact. *J. Assoc. for Inf. Sci. Technol.* **70**, 198–201 (2019).

40. on Prospering in the Global Economy of the 21st Century, C. *Rising above the gathering storm: Energizing and employing America for a brighter economic future* (National Academies Press Washington, DC, 2007).

41. Wagner, C. S., Whetsell, T. A. & Leydesdorff, L. Growth of international collaboration in science: revisiting six specialties. *Scientometrics* **110**, 1633–1652 (2017).

42. Gousios, G. & Spinellis, D. Ghtorrent: Github's data from a firehose. In *2012 9th IEEE Working Conference on Mining Software Repositories (MSR)*, 12–21 (IEEE, 2012).

43. Csardi, G., Nepusz, T. *et al.* The igraph software package for complex network research. *InterJournal, Complex Syst.* **1695**, 1–9 (2006).

44. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An open source software for exploring and manipulating networks. *Int. AAAI Conf. on Weblogs Soc. Media* (2009).

45. BSD Licenses. Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=BSD_licenses&oldid=888925685 (2019). [Online; accessed 26-January-2019].

46. MIT License. Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=MIT_License&oldid=881603526 (2019). [Online; accessed 26-January-2019].

47. Tozzi, C. & Zittrain, J. *For Fun and Profit: A History of the Free and Open Source Software Revolution* (MIT Press, 2017).

48. GNU General Public License. Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=GNU_General_Public_License&oldid=886725631 (2019). [Online; accessed 26-January-2019].

49. Shwed, U. & Bearman, P. S. The temporal structure of scientific consensus formation. *Am. sociological review* **75**, 817–840 (2010).

50. Boehm, B. W. *et al. Software Cost Estimation with COCOMO II (with CD-ROM)* (Prentice Hall, Upper Saddle River, NJ, USA, 2000), 1st edn.

51. Octoverse. The state of the Octoverse (2018).

52. Clauset, A., Newman, M. E. & Moore, C. Finding community structure in very large networks. *Phys. review E* **70**, 066111 (2004).