Summer 6-15-2020

Authorship Pattern of 21st Century Data Science Research: A Scientometric Evaluation

Arindam Sarkar
Jadavpur University, infoarindam83@gmail.com

Ashok Pal, Dr.
Institute of Development Studies Kolkata, Kolkata-700064, India

Authorship Pattern of 21st Century Data Science Research: A Scientometric Evaluation

Arindam Sarkar[#$] and Dr. Ashok Pal*

[#]Department of Library and Information Science, Jadavpur University, Kolkata-700032, India,
**Institute of Development Studies Kolkata, Kolkata-700064, India,*
[$]Email: infoarindam83@gmail.com

Abstract

The present study tries to focus on the various facets of authorship pattern in data science during 2001-2018. Annual growth rate of articles, authorship pattern, author productivity rate, degree of collaboration, author collaboration network visualization and finally the application of Lotka's law are the major thrust of this research. The highest AGR 46.43% was noticed in the year 2016 followed by 39.53% in 2014 and 37.67% in 2015. The lowest AGR -20.75% was noticed in the year 2002. Only 21.83% articles were published by single author whereas 78.17% articles contributed by two or more than two authors. The lowest AAPP was 2.25 with highest PPA was 0.44 observed in the year 2015. On the other side, highest AAPP at 3.88 with lowest PPA at 0.25 is seen in the year 2002. The study reflects that overall Degree of Collaboration is 0.78 that indicates large number of collaboration among the authors. The highest Collaborative Index 5.06 is seen in the year 2001 and minimum Collaborative Index 2.63 is in the year 2015. Seventy authors with greatest total link strength have been represented through VOSviewer's author collaboration network. Finally it can be mentioned that the data set derived from this research largely follows Lotaka's law of author productivity.

Keywords: Data science; Authorship pattern; Author productivity; Degree of collaboration; Collaboration network; Lotka's law

1. INTRODUCTION

Twenty first century has witnessed remarkable growth in the field of science and technology and especially in the field of computer science and information technology this growth is humongous. Data science is such a developing field. It is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. Harvard Business Review in 2012 even called the jobs of this domain as the "Sexiest Job of the 21st Century"[1]. However, to measure the growth of research in this flourishing subject domain it is best to make a scientometric portrait of this subject and a major part of this scientometric analysis is the portrayal of authorship pattern. Growing trend of collaboration among researchers mainly in the science and technology domain is also a noteworthy feature of this present century. Authorship trend and collaborative research are important facets of scientometrics. This authorship pattern mainly deals with the kind of authors, nature and degree of collaboration among them and collaborative trend of authors[2].

The present study has been conducted to show the authorship pattern of the Data science research especially in the backdrop of 21st century.

2. LITERATURE REVIEW

Nalimov and Mulchenko[3] interpreted scientometrics as “the application of those quantitative methods which are dealing with the analysis of science viewed as an information process”. Khiste, Maske and Deshmukh[4] in their study has focused on big data as reflected in Jgate for the period from 2013–2017. Their result indicated that there were total 8930 articles on this subject domain during 2013 to 2017. United States of America and United Kingdom are the most attentive countries in the area of big data analytics. Liao et al.[5] focused on the bibliometric analysis and visualization of medical big data research They analysed a total of 988 references which were downloaded from the Science Citation Index Expanded and the Social Science Citation Index databases from Web of Science. The GraphPad Prism 5, VOSviewer and CiteSpace software are used for data analysis. Sarkar and Pal[6] in their scientometric study on data science represented the meticulous analysis of year and language wise distribution of publications, document type wise distribution of contributions, year wise citation analysis and country wise productivity in the field.

3. RESEARCH OBJECTIVES

The objectives of this study are:

i. To analyse the research productivity on data science in the 21st century.
ii. To show the annual growth rate of articles.
iii. To delineate the distribution of authorship pattern.
iv. To represent the authorship productivity along with the degree of collaboration over the study period.
v. To represent the collaborative index of articles.
vi. To delineate the author collaboration network.
vii. To show whether the author productivity follows Lotka’s law.

4. SCOPE AND LIMITATION

The study is restricted within a particular database, i.e. Scopus.com. In this study the documents on data science published from 2001 to 2018 have been collected.

5. METHODOLOGY

To find out the objectives of this study Scientometrics apparatus and techniques have been used. As a registered user of Scopus database by using a search string **TITLE (data science) AND PUBYEAR > 2000 AND PUBYEAR < 2019** [7]. Data have been collected on, April 5, 2019. After retrieval, data have been collected, consolidated, analysed and calculated using Microsoft Excel application. For portraying the authorship collaboration network VOSviewer software has been used.

6. RESULTS

6.1 Annual Growth Rate

"Annual growth rate (AGR) is the change in the value of a measurement over the period of a year"[8] To calculate AGR, the following formula (Velmurugan and Radhakrishnan's formula)[9] have been used:

AGR = End value – First value / First value * 100

Table 1 also shows the complete scenario of AGR from the year 2001 to 2018. It is observed that the highest AGR 46.43% was noticed in the year 2016 followed by 39.53% in 2014 and 37.67% 2015. The lowest AGR -20.75% was noticed in the year 2002.

Table 1. Year wise distribution of articles

Year	Number of Articles	Percentage (%)	Cumulative (%)	AGR (%)
2001	53	1.40	1.40	0
2002	42	1.11	2.51	-20.75
2003	47	1.23	3.74	11.90
2004	62	1.63	5.37	31.91
2005	84	2.21	7.58	2.18
2006	123	3.25	10.83	46.43
2007	105	2.77	13.60	-14.63
2008	106	2.79	16.39	0.95
2009	121	3.19	19.58	14.15
2010	162	4.28	23.86	33.88
2011	163	4.29	28.15	0.61
2012	167	4.40	32.55	2.45

2013	215	5.66	38.21	28.74
2014	300	7.92	46.13	39.53
2015	413	10.88	57.01	37.67
2016	453	11.95	68.96	9.68
2017	573	15.11	84.07	26.49
2018	604	15.93	100.00	5.41
Total	**3793**			

6.2 Authorship Pattern

Table-2 shows the authorship pattern of research contributions published on data science from the period of 2001 to2018.

Table 2. Distribution of authorship pattern

Year	Single Authorship	Multiple Authorship	Total Articles	Year	Single Authorship	Multiple Authorship	Total Articles
2001	20	33	53	2010	35	127	162
2002	11	31	42	2011	31	132	163
2003	11	36	47	2012	40	127	167
2004	16	46	62	2013	48	167	215
2005	18	66	84	2014	69	231	300
2006	38	85	123	2015	97	316	413
2007	25	80	105	2016	83	370	453
2008	23	83	106	2017	120	453	573
2009	22	99	121	2018	121	483	604

Total number of single authorship contributions: 828;
Total number of multiple authorship contributions: 2965 and;
Total number of articles: 3793

From the table-2 it becomes clear that only 21.83% articles were published by single author whereas 78.17% articles contributed by two or more than two authors.

6.3 Author Productivity

Table 3 shows scenario of average author per paper (AAPP) and productivity per author (PPA) in the selected time zone of this study. The formula for the AAPP and productivity per author are as follows.

Average author per paper (AAPP) = Number of authors / Number of papers
Productivity per author (PPA) = Number of papers / Number of authors

Table 3: Author productivity

Year	Papers	Authors	AAPP*	PPA*	Year	Papers	Authors	AAPP*	PPA*
2001	53	187	3.52	0.28	2010	162	537	3.31	0.30
2002	42	163	3.88	0.25	2011	163	501	3.07	0.32
2003	47	171	3.63	0.27	2012	167	529	3.16	0.31
2004	62	219	3.53	0.28	2013	215	613	2.85	0.35
2005	84	311	3.70	0.27	2014	300	687	2.29	0.44
2006	123	409	3.32	0.30	2015	413	930	2.25	0.44
2007	105	388	3.69	0.27	2016	453	1106	2.44	0.41
2008	106	373	3.51	0.28	2017	573	1390	2.42	0.41
2009	121	411	3.39	0.29	2018	604	1599	2.64	0.38

AAPP*= average author per paper, **PPA*=** productivity per author

From table 3, it is found that lowest AAPP was 2.25 with highest PPA was 0.44 in the year 2015. On the other side, highest AAPP at 3.88 with lowest PPA at 0.25 is seen in the year 2002.

6.4 Degree of Collaboration

Table 4 describes the degree of collaboration among the authors. In this study the Degree of Collaboration (C) of the contributors has been calculated using the Subramanyam[10] formula. The formula is as follows:

Degree of Collaboration (C) **=** Nm / Nm+Ns

Where,

C = Degree of Collaboration
Nm = Number of multiple authored paper
Ns = Number of single authored paper

Table 4. Degree of collaboration

Year	Single Authored Paper (Ns)	Multiple Authored Paper (Nm)	Total (Ns+Nm)	Degree of Collaboration (C)
2001	20	33	53	0.62
2002	11	31	42	0.73
2003	11	36	47	0.76
2004	16	46	62	0.74
2005	18	66	84	0.78
2006	38	85	123	0.69
2007	25	80	105	0.76
2008	23	83	106	0.78
2009	22	99	121	0.81
2010	35	127	162	0.78

2011	31	132	163	0.80
2012	40	127	167	0.76
2013	48	167	215	0.77
2014	69	231	300	0.77
2015	97	316	413	0.76
2016	83	370	453	0.81
2017	120	453	573	0.79
2018	121	483	604	0.79
Total	**828**	**2965**	**3793**	**0.78**

Above table (Table-4) shows the individual year wise Degree of Collaboration and as a whole (from the year 2001 to 2018) Degree of Collaboration. In this case overall Degree of Collaboration(C) = 0.78 that indicates large number of collaboration is found among the authors.

This table also reveals that the highest value of DC 0.81 is observed in the year 2009, 2016 and the lowest value of 0.62 is in the year 2001.

6.5 Collaborative Index of Articles

Collaborative Index (CI) of articles is the mean number of authors per joint paper. For this case, the single authored papers have been omitted. To determine the mean number of authors per jointly authored paper, the following Elango and Rajendran's formula[11] has been used.

Collaborative Index (CI) = Total number of authors / Total joint papers

Table 5: Collaborative Index (CI) of articles

Year	Multi-authored Papers	Total authors of multi-authored papers/articles	CI*	Year	Multi-authored Papers	Total authors of multi-authored papers/articles	CI*
2001	33	167	5.06	2010	127	502	3.95
2002	31	152	4.90	2011	132	470	3.56
2003	36	160	4.44	2012	127	489	3.85
2004	46	203	4.41	2013	167	565	3.38
2005	66	293	4.43	2014	231	618	2.67
2006	85	371	4.36	2015	316	833	2.63
2007	80	363	4.53	2016	370	1023	2.76
2008	83	350	4.21	2017	453	1270	2.80
2009	99	389	3.92	2018	483	1478	3.06

CI*= Collaborative Index

It can be observed from Table-5 that the highest CI 5.06 is in the year 2001 and minimum CI 2.63 is in the year 2015. Average CI is 3.82 during the selected period of time.

6.6 Author Collaboration Network

In the visualization presented in Figure 1, each circle represents a researcher. Large circles represent researchers that have many publications (Foster, I., Li,X., Zhang, Z. etc. for example are prominent from their size in the network.) . Small circles represent researchers with only a few publications. The closer two researchers are located to each other in the visualization, the more strongly they are related to each other based on bibliographic coupling. In other words, researchers that are located close to each other tend to cite the same publications, while researchers that are located far away from each other usually do not cite the same publications[12].

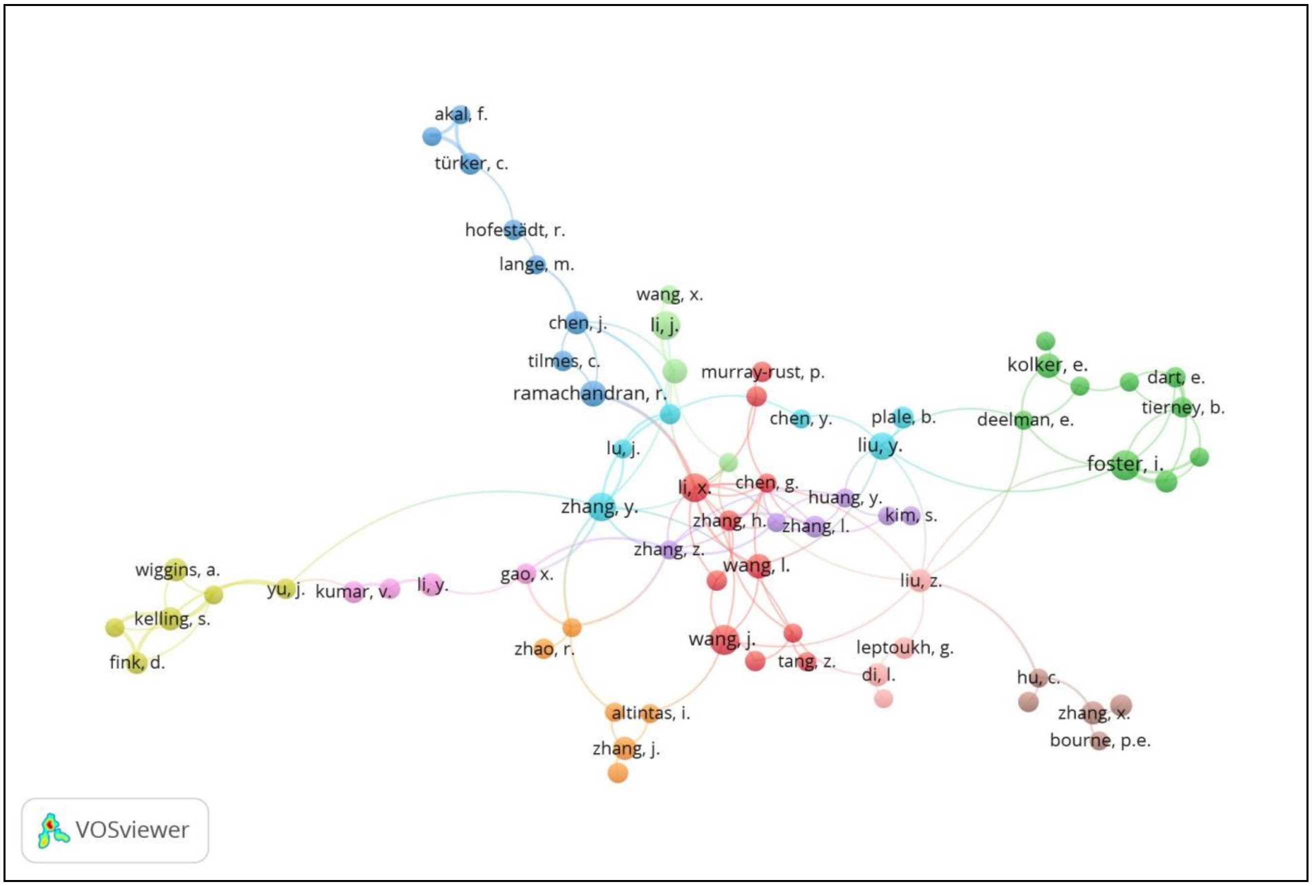

Figure 1. Author Collaboration Network

The threshold which has been set for the visual portrayal of author collaboration network is minimum of 5 documents of an author to be chosen. Out of the 10524 authors 123 only meet the threshold. Out of these 123 authors, 70 authors with greatest total link strength have been represented through this figure. After VOSviewer has calculated the total strength of the co-authorship links with other authors, 11 clusters have been formed in this network.

6.7 Application of Lotka's Law

Lotka conducted an experiment on the author productivity. The simplest equation to represent Lotka's law is: $x^a y = c$ where x stands for the contributions; y stands for the number of authors, and c is constant. Using the above equation, the value of c will be determined according to Sen's method[13].

Table 5. Verification of Lotka's law

Number of papers (x)	**Number of author (y) (observed)**	**Number of author (y) (expected) with the value a=**1.850
1	869	869
2	241	241
3	115	113
4	71	66
5	49	44
6	35	31
8	21	18
9	21	15
10	14	12
11	13	10
12	9	8
13	9	7
14	10	7
15	9	6
19	4	3
20	5	3
21	2	3

Taking in account the value of as given in the first row of the Table 5, we get
$1^a. 869 = c$ [as $1^a= 1$]
$869 = c$
Now, using the data of the second row (Table 5), we can find out the value of *a*.
$2^a. 241 = 869$
$2^a= 3.605$
$a \log 2 = \log 3.605$
$0.301 = 0.5569$
$a= 0.5569/0.301$
$a= 1.850$

Applying the value of *a* the expected values of *y* have been determined in Table 5. It may be observed from the table that the value of *y* is quite close to the actual values when calculated with *a* = 1.850. Therefore, it may be concluded that the data set derived from this study largely follows Lotka's law.

7. CONCLUSIONS

The present study tries to focus on the various facets of authorship pattern in data science during 2001-2018. The highest AGR 46.43% was noticed in the year 2016 followed by 39.53% in 2014 and 37.67% in 2015. The lowest AGR -20.75% was noticed in the year 2002. Only 21.83% articles were published by single author whereas 78.17% articles contributed by two or more than two authors. The lowest AAPP was 2.25 with highest PPA was 0.44 observed in the year 2015. On the other side, highest AAPP at 3.88 with lowest PPA at 0.25 is seen in the year 2002. The study reflects that overall Degree of Collaboration is 0.78 that indicates large number of collaboration among the authors. The highest Collaborative Index 5.06 is seen in the year 2001 and minimum Collaborative Index 2.63 is in the year 2015. Seventy authors with greatest total link strength have been represented through VOSviewer's author collaboration network. Finally it can be mentioned that the data set derived from this research largely follows Lotaka's law of author productivity.

REFERENCES

1. Data science. *Wikipedia.* https://en.wikipedia.org/wiki/Data_science (accessed on 12 April 2019).

2. Pillai, K. G. S. Authorship patterns in physics literature: An informetric study on citations in doctoral theses of the Indian Institute of Science. *Ann. of Lib. & Inf. Stu.*, 2007, 54 (2), 90-94. http://nopr.niscair.res.in/bitstream/123456789/3248/4/ALIS%2054%282%29%2090-94.pdf (accessed on 15 April 2019).

3. Nalimov, V.V. & Mulchenko, Z.M. Study of science development as an information process. *Scientometrics*, 1989, 15, 33-43.

4. Khiste, G. P.; Maske, D.B. & Deshmukh, R. K. Big data output in Jgate during 2013 to 2017: A bibliometrics analysis. *Int. J. of Scientific Research in Com. Sci., Eng. and Inf. Tech,* 2018 3(1), 1252-1257.

5. Liao, H.; Tang, M.; Luo, L.; Li, C.; Chiclana, F. & Zeng, X. J. A bibliometric analysis and visualization of medical big data research. *Sustainability*, 2018, 10(1), 166.

6. Sarkar, Arindam & Pal, Ashok. Where does data science research stand in the 21st century: Observation from the standpoint of a scientometric analysis. *Lib. Phi. and Pra. (e-journal),* 2019, 2561, 1-9. https://digitalcommons.unl.edu/libphilprac/2561/ (accessed on 22 April 2019).

7. Noruzi, Alireza. YouTube in scientific research: A bibliometric analysis. *Webology*, 2017, *14*(1), http://www.webology.org/2017/v14n1/editorial23.pdf

8. Annual growth rate. *Wikipedia.* https://en.wikipedia.org/wiki/Annual_growth_rate (accessed on 12 April 2019).

9. Velmurugan, C. & Radhakrishnan, N. Malaysian Journal of Library and Information Science: A scientometric profile. *J. Scientometric Res.*, 2016, **5**(1), 62-70. doi: 10.5530/jscires.5.1.9

10. Subramanyam, K. Bibliometric studies of research in collaboration: A review. *J. of Inf. Sc.,* 1983, 6 (1), 33-38.

11. Elango, B. & Rajendran, P. Authorship trends and collaboration pattern in the marine sciences literature: A scientometric study. *Int. J. Inf. Dissemination Technol.*, 2012, 2(3), 166- 169. http://www.ijidt.com/ index.php/ijidt/article/viewFile/91/91 (accessed on 24 April 2019).

12. Van Eck, N.J. & Waltman, L. Visualizing bibliometric networks. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact: Methods and practice*, 2014, 285–320.

13. Sen, B.K. Lotka's law: A view point. *Ann. Lib. & Inf. Stu*., 2010, 57(2), 166-67.