Fall 8-22-2017

Diffusion of Big Data in Indian Scientific Literature: Study of Research Productivity and Scientific Collaboration

Punit Kumar Singh
Banaras Hindu University, punitbhu@gmail.com

Ajay P. Singh Prof.
Banaras Hindu University, apsingh_73@yahoo.co.in

Diffusion of Big Data in Indian Scientific Literature: Study of Research Productivity and Scientific Collaboration

By

Punit Kumar Singh

Research Scholar,
Department of Library and Information Science
Banaras Hindu University, Varanasi-221005

&

Dr. Ajay P. Singh

Associate Professor,
Department of Library and Information Science
Banaras Hindu University, Varanasi-221005

Abstract:

Purpose: Big data, a buzzword of the present time, is a term used for extremly large data sets generated from the digital process which is not possible to analyze by traditional methods. These data sets are produced by digital devices such as smart phones, remote sensing, camera, microphones, RFID etc. The literature on big data is growing exponentially since 2011. Big data is tending to establish as a very important research field. This paper aims to explore the evolution, growth and scientific collaboration of the Indian publications in the field of big data.

Design/methodology/approach: A survey approach is used in the study while data for the study is collected from Scopus database for the year 2001 to 2015. Bibliometric analysis, visualization and mapping software are used to present the current status, growth trends and collaboration in big data research to examine its diffusion in Indian scientific literature.

Findings: We found that the big data research in India is gaining momentum and its diffusion and adoption is increasing tremendously. Conference and seminars are used to do social connect and interaction within the research community. The collaboration at institution level is found usual while collaboration at international level is low. Application of big data in health sciences and life sciences is yet to be explored in comparison to the social sciences and physical sciences.

Originality/ Value: This paper presents the growth, trends and collaboration in big data literature by the use of sophisticated bibliometric software and visualization software.

Keyword: Scientometrics, Big Data, Network Analysis, Visualization

Paper Type: Research Paper

INTRODUCTION:

Data is the recorded i.e. measured, collected, reported and analyzed facts which can be visualized. The data set is the collection of the data in a database table having different variables in columns and the particular value in the rows. These data sets are analyzed to constitute

information. Various types of large data sets are ubiquitous at the present time, for example, social network message flow data, meteorological data, location data, audio and video recordings, software logs, user logs, etc. It also includes the data of social networking site, social bookmarking, personal data blogs, posts, etc. These data sets are produced by digital devices such as smart phones, remote sensing, camera, microphones, RFID, etc. Big data is a term used for extremly large data sets generated from the digital process which is not possible to analyze by traditional methods.

Big data is basically a field of study in computer and information science while in last five years, it has recognized as a multidisciplinary subject due to a remarkable increase in big data literature in scholarly publication across various academic disciplines including management, health sciences, business, and information systems. It has created a new insight for business, government, education and social acts by the emergence of real-time; user generated information and communication (Frizzo-Barker, Chow-White, Mozafari, & Ha, 2016).

BIG DATA:

Big data is a buzzword of the present time. According to the characteristics, many definitions of the big data are popular at present like "3Vs" of big data e.g. volume, velocity, and variety (Laney, 2001); "4Vs" of big data which includes Veracity (What is big data? 2016) along with above; and "5Vs" of big data which have volume, velocity, variety, veracity and value (Marr, 2015). Gartner IT Glossary defines "Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation." (Gartner IT Glossary, 2016) Volume defines the large amount of data generated (Marr, 2015); Variety represents the diversified nature of the structured and unstructured data; Velocity refers the speed of the data generated and the speed of the flow of data; while Veracity explains the integrated and trustworthy of data for an organization and Value measures the usefulness of data for an intended purpose. It reveals the big data is too large, too rapid and too variable to process by the existing tools and techniques.

Data is an integral part of many disciplines as well social lives. Users of social networks such as Facebook, Twitter, and Instagram produce an enormous stream of different types of information every day such as music, pictures, text, etc. These data are helpful in data-driven decision making in the organizations (Chow-White & Green Jr, 2013). The multidisciplinary characteristics of big data as explained by Wu, X. et.al (2014), "Big Data concern large volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities and seeks to explore complex and evolving relationships among data." (Wu, Zhu, Wu, & Ding, 2014)

LITERATURE REVIEW:

An attempt is made to review the related literature relevant to the research area including Big Data, scientific collaboration, co-authorship, scientometric analysis, and social network analysis to obtain in-depth knowledge of the research problem.

Plenty of research publications on big data in various disciplines in last five years is found while only three research publications are counted on bibliometric analysis of this research field. The triple helix analysis (Park, 2014), Keywords co-occurrence mapping (Zhu, Liu, He, Shi, & Pang, 2015), and scientometric mapping (Singh, Banshal, Singhal, & Uddin, 2015) of big data literature is done by authors for limited period. Singh, et al. (2015) has done a scientometric analysis of the big data literature over a period of 2010 to 2014 by collecting the data from Scopus and WoK databases. The diffusion of big data literature in Indian research output till 2015 is explored by analysis of the realistic growth trends in the literature, a detailed picture of the scientific collaboration, and its multidisciplinary character which is presented in this study with help of a mixture of bibliometric analysis, network analysis, cluster analysis and visualization of big data literature to provide a more detailed and robust roadmap for further research in this field.

METHODOLOGY

Diffusion is the process by which an innovation is communicated through certain channels over time among the members of a social (Rogers, 1995). Diffusion research, especially in the case of diffusion of knowledge/ innovations, is the study to explain how, why and at what rate new ideas and technologies spreads among a group of peoples, institutions, and countries. The key elements of diffusion of innovations are innovation, adopters, communication channels, time, and social systems. The present article is based on these key elements for the study of diffusion of big data in Indian research community. Following methods are used to for the study of diffusion according to the key elements:

Innovation	:	Growth rate of big data publication (Research Productivity),
Adopters	:	Growth rate of new authors
Communication	:	Multidisciplinary character, document type distribution, keyword co-occurrence,
Time	:	Study involves time period of 1996-2015
Social Systems	:	Role modelers, social network analysis, co-authorship analysis, institutional and scientific collaboration

Big data is a fast growing multidisciplinary research field which has attracted the researchers across many disciplines to explore the feasibilities in and impact of this emerging field. The present study is based on Scopus database, the largest abstract and citation database of peer-reviewed literature e.g. scientific journals, books and conference proceedings (About Scopus, 2016). The keyword “big PRE/1 data” is used in ‘article title, abstract and keywords’ search to collect all the big data literature having the keywords big data included with one word in middle e.g. big sensory data, big learning data, etc. published till 2015. The whole 16,513 documents

retrieved at this step which is then refined to 785 documents after limiting to "India" in country/territory (as on September, 2016). Datasets are collected in .ris and .csv formats for each year as well as for each subject area defined by Scopus database.

Subject growth trends, collaboration clusters, interrelations, key research topics, research gaps, etc. are identified by using rigorous bibliometric tools. The systematic mapping and network analysis of the field help to illustrate the publications evolution over time graphically and identify areas of current research interests and potential directions for future research in big data in India (Fahimnia, Sarkis, & Davarzani, 2015). Processing of the co-citation and co-occurrence data for network mapping and visualization are done with the help of BibExcel (Bibexcel, 2016) bibliometric analysis tool which is also used to prepare the input data for a detailed network analysis while tabulation and the graphical representation are done through MsExcel. These studies require reformatting of the RIS file into some of different formats and hence producing several file types. An OUT-file needs first to be created to enable data for analysis in Bibexcel (Bibexcel, 2016). Pajek, (Batagelj, & Marvar, 1998) the statistical program for network analysis and VOSviewer, (van Eck & Waltman, 2010) a program for constructing and viewing bibliometric maps of authors or journals based on co-citation data or to construct and view maps of keywords based on co-occurrence data, is used for network analysis, cluster analysis and distance–based mapping.

FINDINGS:

Research Productivity Analysis:

The measurement of research productivity can be done by analysis of publication volume and its growth rate. The bibliometric analysis is the defined way to analyse the research productivity in a particular research field. We used the bibliometric analysis to explore the diffusion of big data in scientific research output of India till 2015.

Diffusion of Big Data Knowledge in India:

As observed from the Scopus database, the emergence of the big data in Indian scientific literature is started during 1996-2011. An exponential growth of big data papers published by Indians is observed during 2012 to 2015 as the literature approximately tripled in every next year. The number of Indian publications on big data increased from 18 in 2012 to 511 in 2015 i.e. more than 28 times increase (Fig.1) is a remarkable growth which is enough to prove the big data as a priority research field in India. The institute wise ranking of diffusion of the big data for more than five publications is observed in Fig. 2. The VIT University, Chennai with 31 papers is top ranked institute publishing on big data followed by Sathyabama University with 21 articles which is also a considerable growth in very short period.

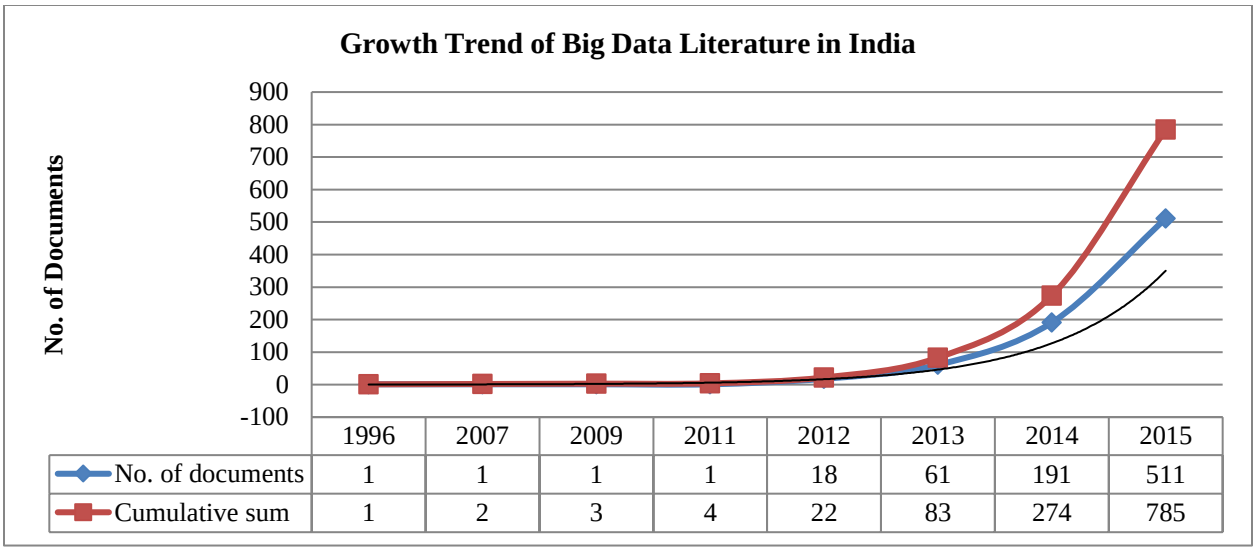

Fig.1 Growth Trend of Big Data Literature in Indian Scientific Publications Till 2015

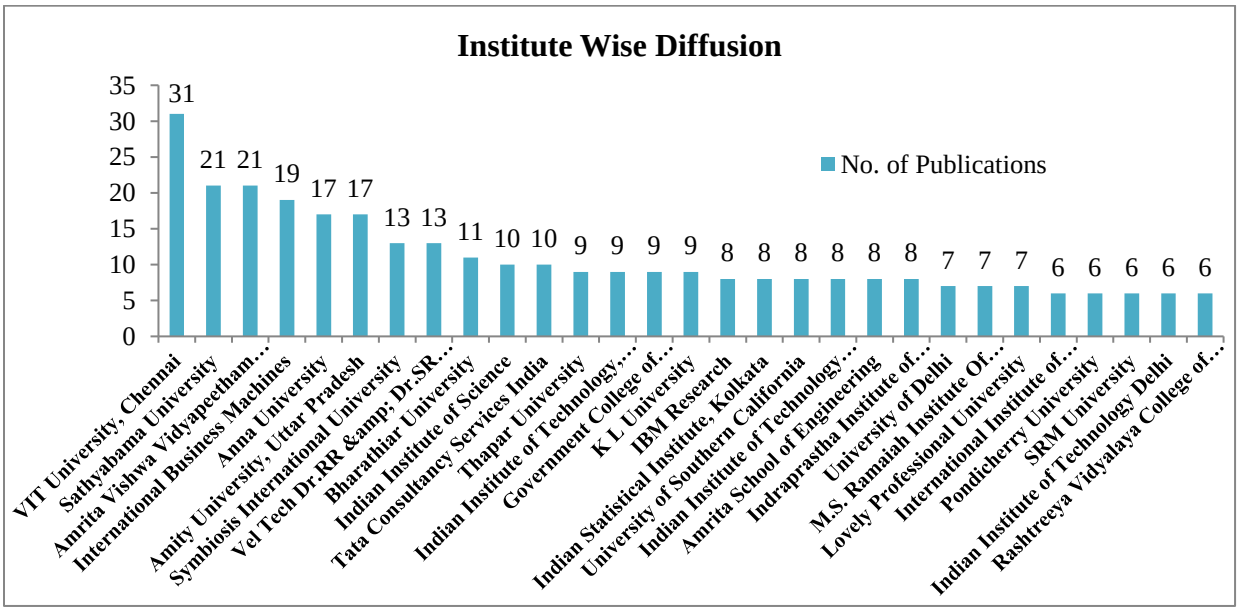

Fig 2 Institute Wise Diffusion of Big Data in Indian Scientific Literature Till 2015

Diffusion of Big Data in Other Disciplines than Computer Science: (Multidisciplinary Nature)

Scopus database arranges publications into four basic subject areas which are health sciences, life sciences, social sciences, and physical sciences further each subject area includes other subjects. Each publication is placed into different subjects according to its content. A study has done to find out the multi-disciplinary character of Indian publications on big data so that the

diffusion of big data in other subjects than computer science can be evaluated. According to table 1, it is clear that the big data is mainly a subfield of physical sciences which include the computer science while it is used in all subject areas of Scopus since 2013. Subject wise classification of the big data literature shows that big data is heavily used in the computer science, engineering, business management, social sciences and decision sciences as seen in figure 3.

Year of Publication	Health Sciences	Life Sciences	Social Sciences	Physical Sciences
Before 2012	-	-	-	4
2012	-	-	-	18
2013	2	3	7	56
2014	8	7	27	182
2015	18	25	49	500

Table 1 Distribution of Big Data Articles According to the Scopus Subject Areas

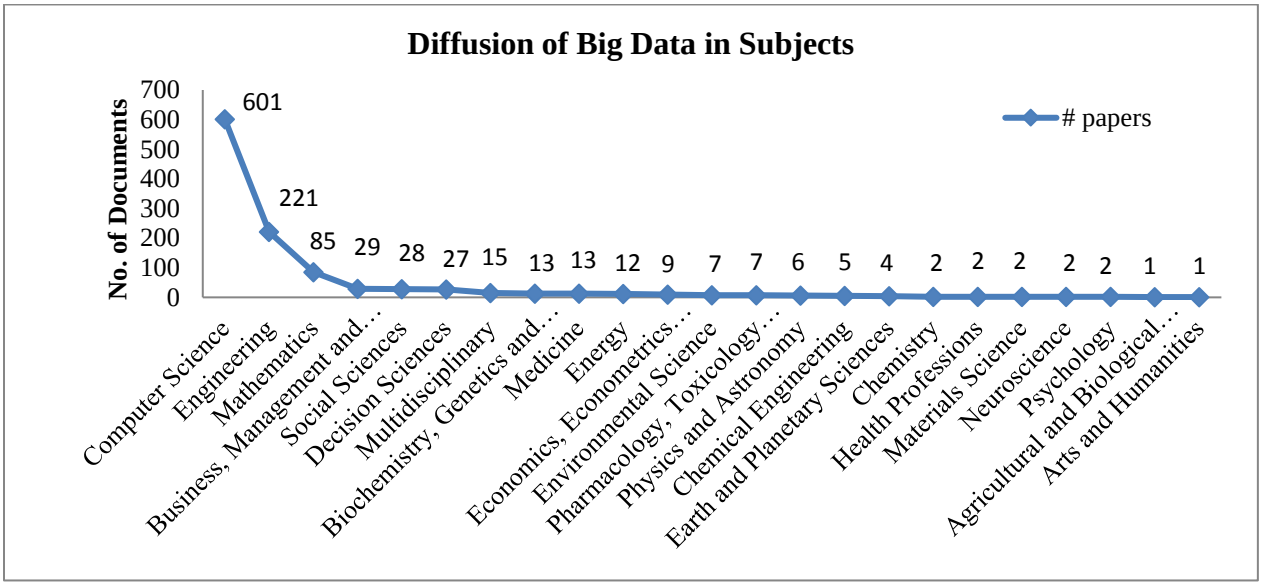

Fig. 3 Diffusion of Big Data in Different Subjects Till 2015; Source: Scopus database

In order to represent a clearer view on multi-disciplinary character of big data in Indian publications, we studied the keyword co-occurrence analysis of the whole dataset on big data as well as the dataset of four subject areas e.g. health sciences, life sciences, physical sciences and social sciences (fig 4).VOSviewer software (van Eck & Waltman, 2013) for its ability to provide easy-to-interpret graphical representations of bibliometric maps is used for the density visualization of the keyword matrix for which data is extracted and cleaned by BibExcel. The density view immediately reveals the general structure of the map. The color of a point in these distance-based maps depends on the number of items in the neighborhood of the point and on the importance of the neighboring items. The density view is particularly useful to get an overview of the general structure of a map and to draw attention to the most important areas in a map (van

Eck & Waltman, 2010). The big data co-occurred in every subject area i.e. multidisciplinary nature as shown in the fig 4 (A, B, C, D, & E). It is obvious to mention that according to the fig 4, health sciences is the discipline in which the research gap is observed while social science is the favoured research field after physical sciences to do big data research. It means the big data is applied more to solve social problems rather than the health issues. It is also noted from the keyword co-occurrence analysis that although big data is used in every research subject more or less, it is basically more nearer to computer science and its application in different areas are to be explored.

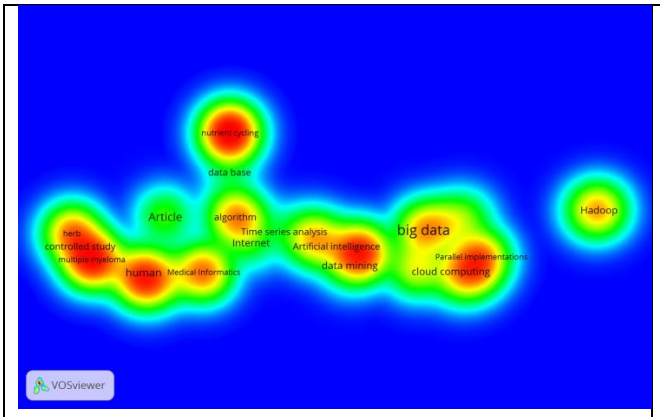

Fig 4A Keyword Co-Occurrence Density Network of Big Data in Health Sciences

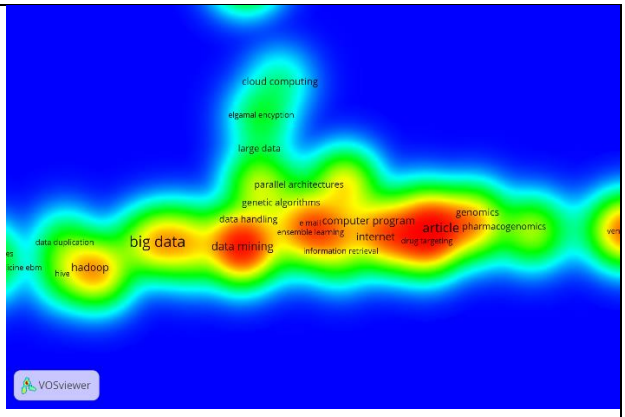

Fig 4B Keyword Co-Occurrence Density Network of Big Data in Life Sciences

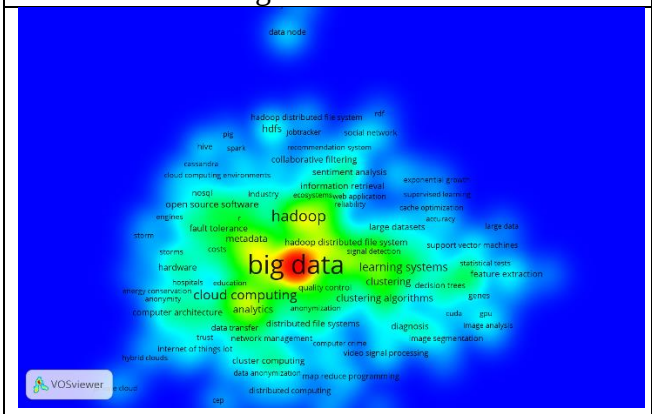

Fig 4C Keyword Co-Occurrence Density Network of Big Data in Physical Sciences

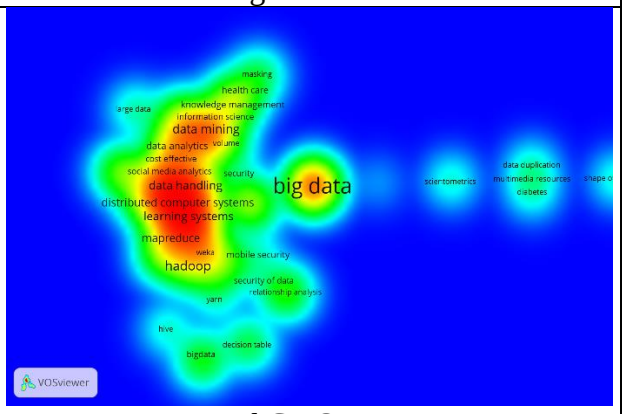

Fig 4D Keyword Co-Occurrence Density Network of Big Data in Social Sciences

Fig 4EKeyword Co-Occurrence Density Network of Big Data in Indian scientific publication till 2015.Sources: data cleaned from BibExcel software while density visualization of network is done by VOSviewer.

Fig 4 Keyword Co-Occurrence Analysis of Indian Research Output on Big Data

Adoption of Big Data in Indian Researchers:

We have attempted to study the diffusion of big data as a research discipline in the Indian research community by finding out the adoption rate of the big data in research community using a more refined approach by identifying the new authors collaborating each year (Darvish & Tonta, 2016).We recognized the "new authors" as those who published first time a big data paper as indexed in Scopus database. The new adopters in big data literature are counted by refining and cleaning the data for author with affiliation in each year through BibExcel software since 1996i.e. beginning of the big data research in India as per Scopus database. Each new Indian author participated in big data research in the following year is counted as a ''new adopter'' and added to the count of previous years.

At the beginning of the big data research in India, the number of unique authors was just 12 before 2012 whereas it rose to 189 in 2015 (Table 2 & Fig. 4). Adoption rate is rather slow during the period of 1996-2011 whereas a 'tipping point' is noted in 2012 when the number of new authors jumped from 12 in 1996-2011 to 65 in 2012 i.e. more than fivefold increase. The average number of new adopters during the period of 2012-2015 rose to 138 which is more than eleven fold increase of the total new adopters during 1996-2011. We observed the exponential growth in the adoption of the big data in scientific and research community in India during the period of 2012-2015 (Fig. 5) while the cumulative increase in the adoption of the big data in research community is scored to 564 till 2015 (Table2). The average rate of cumulative growth percentage in the adoption is 55.9% during the period of 2012-2015.

Year of Publications	No. of New Adopters	Cumulative Adopters	Rate of Cumulative Growth (%)
Before 2012 (1996-2011)	12	12	0
2012	65	77	84.41
2013	148	225	65.78
2014	150	375	40.00

2015	189	564	33.51

Table 2 Number of New Adopters and Cumulative Adopters of Big Data Till 2015

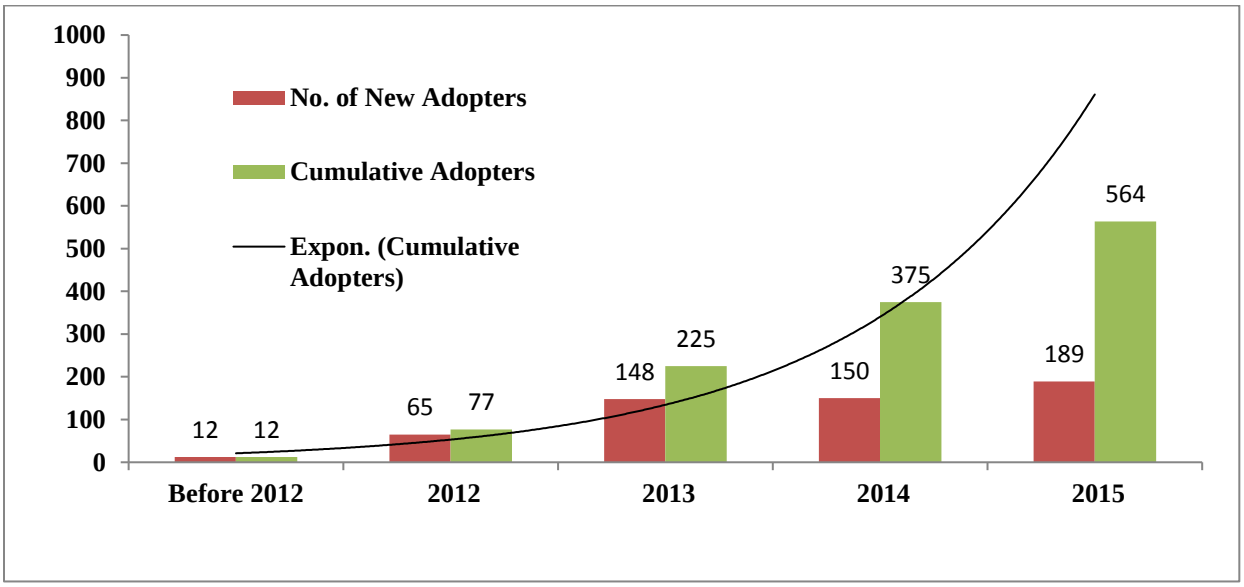

Fig. 5Growth of Adoption of the Big Data Based on the Cumulative Authors

Document Type Distribution:

Figure 6 depicts the diffusion of big data in Indian research publication in respect of document type distribution. It is clear from the figure that the conference paper is the most favoured document type followed by the journal articles.

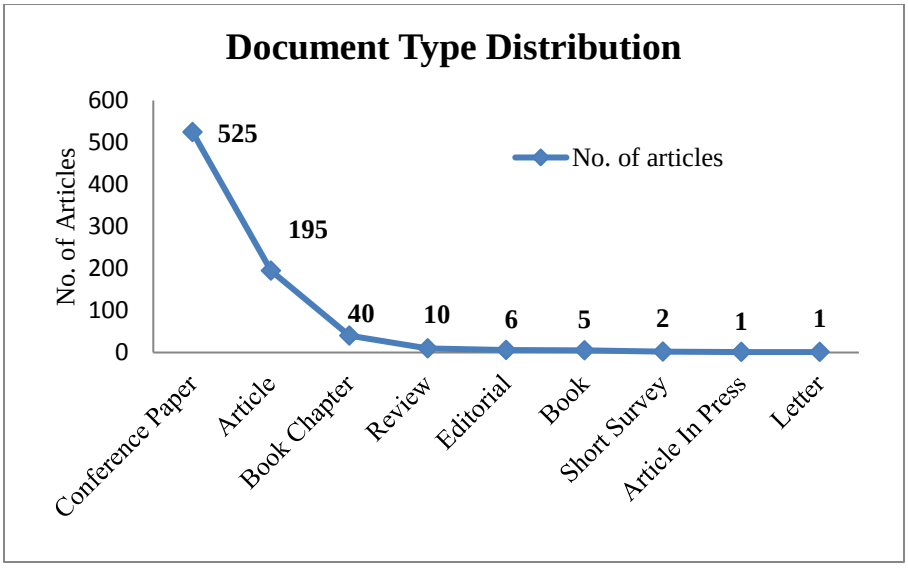

Fig 6Document Type Distribution of Big Data Publications

Scientific Collaboration:

Ardanuy (1012) stated that collaboration allows researchers to share techniques and is an excellent way to transfer knowledge, especially tacit knowledge. At present the scientific research is the matter of collaborative efforts of research groups while the publication of collaborative work shows signs of providing a greater number of citations (Ardanuy, 2012). Group of scientists collaborate at local, national and international levels to complete large projects. Big data research is multidisciplinary research which requires collaborations from different research fields (Sarwar & Hassan, 2015). In this paper we have tried to find out the status of scientific collaboration at author, institution/ national and international level in order to explore the diffusion of big data research output at local, national and international level.

Authorship Analysis:

Table 3 shows that out of 785 publication of big data, only 88 publications found solo authored while 697 papers have collaborative authorship. Out of 697 collaborative papers, 536 papers have national level of collaboration while 161 articles have at least one author from abroad. Favoured authorship in big data research is two or three authors. Collaboration in big data research is frequently seen which is due to the multi-disciplinary or inter-disciplinary aspects of big data.

No. of Authors	No. of Publications
Single	88
Two	297
Three	210
Four	112
Five	39
Six	18
More than 6	21

Attribute	No. of Publication
Solo Authorship	88
Collaborative Authorship	697
National collaboration	536
International collaboration	161

Table 3 Authorship analysis of big data research output of India

Co-Authorship Analysis:

In Co-Authorship analysis, the analysis of the social and professional network of authors (nodes) formed by co-authoring of articles together (edges) is performed to investigate the macro and micro characteristics in research collaboration. The social network analysis (SNA) uses a well-developed set of mathematical algorithms for the analysis and visualization of networks (Wasserman & Faust, 1994). SNA is a sociological approach to discover the topological properties of a network (Kumar, 2015). We anlysed the co-authorship network of big data research output of India at both micro and macro level in order to find out the level of the research collaboration and the role modelers in the research field. At micro level of the structure analysis the centralities (degree, closeness, and betweenness) of the top 10 author (table 4; & fig 7) are analysed while at macro level structure analysis the clusters, centralities and clustering co-efficient of the whole network are analysed as shown in table 4(Yan, Ding, & Zhu, 2010). We used BibExcel (Persson, Danell, & Schneider, 2009) to clean the data for network analysis which

is used by Pajek to do network mapping and visualization with help of network, partition and vectors.

Micro Level Structure Analysis:

Micro level network analysis involves the analysis of degree of collaboration of the individual author within the social network. Centrality measures like degree, weighted degree, closeness, and betweenness are the analysis method to measure the collaboration of the authors at micro level. Degree centrality measure is used to find the most active and most visible actor in the network in order to detect most collaborative authors. Closeness centrality measure is used to detect the actors that are closest to all others in the network in order to find out authors with extensive collaborative scope while betweenness centrality is used to detect actors who lies on a great number of shortest path in the network in order to find out the brokers and connectors i.e. role modelers with interdisciplinary approach (Yan et al., 2010).

Table 4 shows the ranking of the top 10 authors according to the number of publications, degree centrality, weighted degree centrality, closeness centrality, and the betweenness centrality. The most prolific Indian researcher in big data according to number of publications is V. Vijayakumar followed by P. Raj and A. Kumar while A. Kumar superseded to all as most collaborative author. A. Kumar, H. Kaur and R. Chauhan are the authors with extensive collaborative scope in respect of highest closeness centrality and may act as role modelers with interdisciplinary aspect in respect of highest betweenness centrality. Overall, A. Kumar is the most influential person in the network with a high level of centralities.

Rank	Author (Publications)	Author (Degree Centrality)	Author (Weighted Degree Centrality)	Author (Betweenness Centrality)	Author (Closeness Centrality)
1	Vijayakumar, V. (11)	Kumar, A. (24)	Kumar, A. (24)	Kumar, A. (0.0052)	Kumar, A. (0.0261)
2	Raj, P. (9)	Olivier, J. (16)	Vijayakumar, V. (23)	Kaur, H. (0.0038)	Chauhan, R. (0.0251)
3	Gupta, A. (9)	Loranger, J. (16)	Amudhavel, J. (20)	Chauhan, R. (0.0035)	Kaur, H. (0.0239)
4	Kumar, A. (8)	Mikolajczak, A. (16)	Dhavachelvan, P. (19)	Singh, S. (0.0032)	Dhavachelvan, P. (0.0230)
5	Kumar, P. (7)	Lemauviel-Lavenant, S. (16)	Olivier, J. (16)	Singh, J. (0.0031)	Singh, J. (0.0218)
6	Singh, S. (7)	Dhavachelvan, P. (16)	Loranger, J. (16)	Kumar, P. (0.0019)	Kumar, P. (0.0216)
7	Vasudevan, S.K. (5)	King, J. (16)	Mikolajczak, A. (16)	Pandey, S. (0.0017)	King, J. (0.0211)
8	Pal, A. (5)	Jolivet, C. (16)	Lemauviel-Lavenant, S. (16)	Srivastava, S. (0.0015)	Abbasi, T. (0.0211)
9	Simmhan, Y. (5)	Abbasi, T. (16)	King, J. (16)	Agarwal, S. (0.0009)	Nair, P.R. (0.0211)
10	Sharma, S. (5)	Amiaud, B. (16)	Jolivet, C. (16)	Sinha, R. (0.0009)	Fiala, M. (0.0211)

Table 4Ranking of authors according to publications, degrees and centralities

Figure7A represents general visualization of co-authorship network of big data research output of 865 Indian authors till 2015 with first partition and first & second vector through Kamada Kawai (free) network mapping graph by Pajek software. This general visualization of the co-authorship structure elaborated the research groups with different colours while the size of the nodes is according to the number of publications (Kamada & Kawai, 1989). In order to optimize the clusters of the co-authorship network shown in the figure 7A, the Kamada Kawai graph with the optimized inside clusters of the network is represented in figure 7B which has produced 8 clusters with different colours. The distance in nodes and clusters is according to the strength of the network both among the clusters and within the clusters. Table 5A represents the details of different clusters and its output. Again to find out the clearer picture of the clusters and research groups, the network with optimized clusters is mapped with layer in Y direction as shown in figure 7C. V. Vijayakumar, A. Kumar, A. Gupta, S. Singh and P. Kumar are the top 5 collaborative authors as clear from figure 7C. They are the central authors of the whole network which indicates that they are the most influential person in the network and may act as the role modelers in the big data research field in future. From table 4 and figure 7C, it is concluded that the V. Vijayakumar is the most prolific author followed by P. Raj while A. Kumar is the most collaborative author followed by V. Vijayakumar as well as A. Kumar, H. Kaur and R. Chauhan are the authors with extensive collaborative scope. It is interesting to observe that H. Kaur and R. Chauhan are not visualized in the fig 7C even having higher centralities than V. Vijayakumar due to the less number of publications but high collaborative aspects.

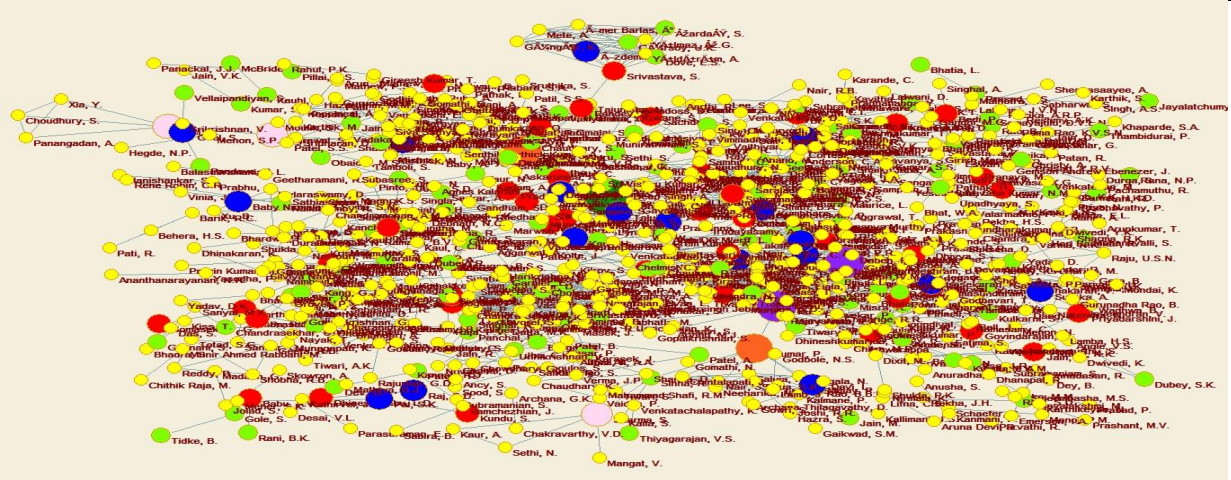

Fig 7A Co-Authorship network of big data research output of India till 2015

(Visualization of network with first partition and first & second vector through Kamada Kawai (free) network mapping graph by Pajek software)

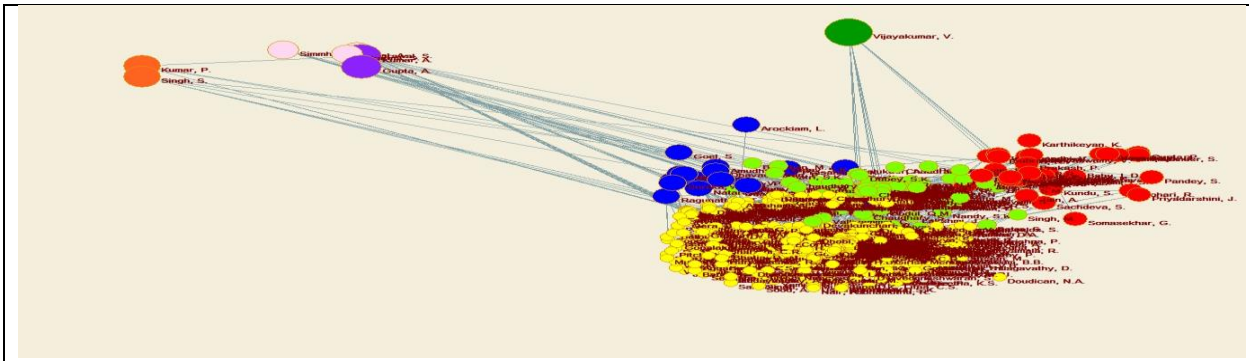

Fig 7B Co-Authorship network of big data research output of India till 2015
(Visualization of network with first partition and first & second vector through Kamada Kawai (optimized inside clusters) network mapping graph by Pajek software)

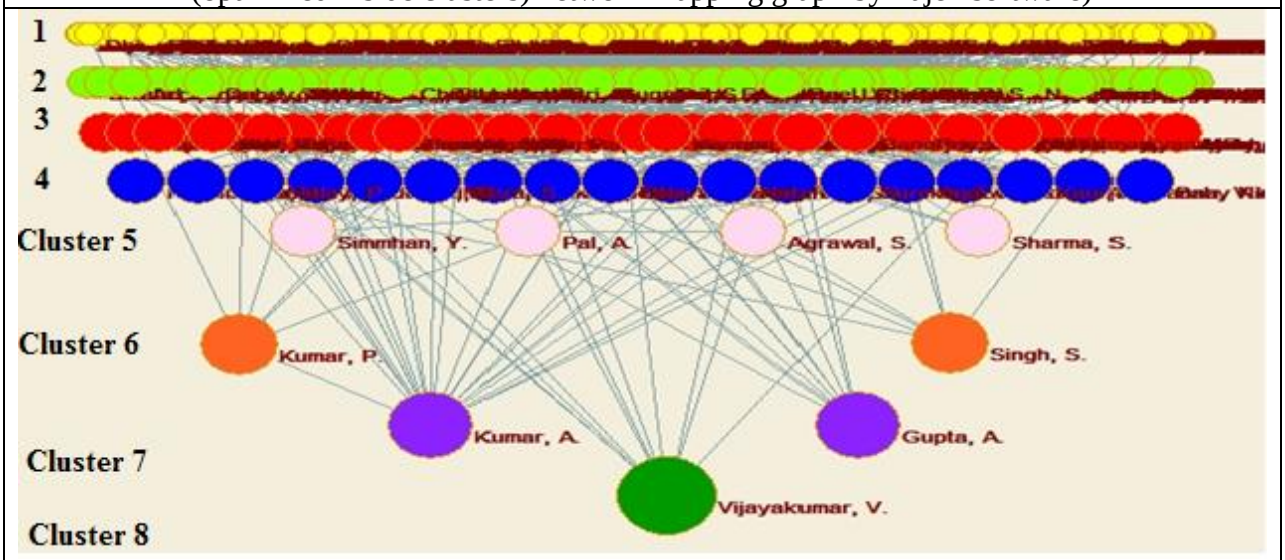

Fig 7C Co-Authorship network of big data research output of India till 2015
(Visualization of network with first partition and first & second vector through Kamada Kawai (optimized inside clusters) network mapping graph with layer in Y direction by Pajek software)

Macro Level Structure Analysis:

Using standard network centralization indices e.g. degree, closeness, betweenness, network clustering coefficient etc. calculated with Pajek (Mrvar & Batagelj, 2016), we tried to explore the strength and weakness of the network structure (Velden & Lagoze, 2008). The macro level of the network analysis involves component, distance and cluster, and degree distribution of the whole network in general. Component analysis is used to detect the degree of network scattering useful for comparison across discipline while distance and cluster study is used to observe the density and organization of a network in order to detect the collaboration pattern of the fields. We observed the weak collaboration pattern as the cluster 1 having the highest number of authors with weakest link strength while cluster 8 having only one author with strongest link strength concluded from Table 5A and figure 7C. Degree distribution is used to detect the

structure of a network for stratifying the authors according to the degree (Yan et al., 2010). Table 5B represents the results of calculation of standard network centralization indices for whole network and concludes a weak collaboration within the network. The closeness centrality for this network cannot be calculated due to weak network strength.

Clusters	No. of Authors
Cluster 1	672
Cluster 2	125
Cluster 3	41
Cluster 4	18
Cluster 5	4
Cluster 6	2
Cluster 7	2
Cluster 8	1

Table 5A Details of Clusters and Its Outputs

Indices	Output
Vertices (nodes)	865
Edges	2355
Degree Centrality (All)	0.02385091
Betweenness Centrality (All)	0.00512401
Network Clustering Co-efficient	0.91248817

Table 5B Details of Network Centralization Indices and Its Outputs

International Collaboration:

International research collaboration is defined as the share of articles published together with at least one author from another country anywhere in the world. Individual interest, government policy, motivation of scientists, and bilateral agreement between institutions are the main factors of the international collaboration. International collaboration among scientists may affected by different factors viz. size, economic and political policies of country as well as different aspects of migration and mobility of individuals (Sarwar & Hassan, 2015), cost-savings, the growing importance of interdisciplinary fields and geographical, economic or cultural interests(Wang, Thijs, & Gla, 2015; Katz, & Martin, 1997).

Table 6 represents ranking of top 10 countries including India according to centrality measures (i.e. degree, weighted degree, closeness, betweenness) (Freeman, 1978) for international collaboration in Indian big data research output. The total number of coauthored papers with at least one foreign author is 161 papers. Indian researchers have collaborated with researchers of 36 countries, out of which United States, Netherlands and Canada are the top 3 main actors in the collaboration according to the degree centrality while weighted degree centrality includes the United Kingdom in these (see fig 8). According to the closeness centrality and betweenness centrality, United States, Netherlands and Canada are the top 3 countries in the network which have authors with extensive collaborative scope with Indians. Overall from table 6 and fig 8, it can be generalized that United States, Netherlands, Canada, United Kingdom and South Africa are the top 5 countries which has remarkable collaboration with India in big data research while United States is seems to be most collaborative country for Indian researcher for big data research.

Rank	**Degree Centrality**		**Weighted Degree Centrality**		**Closeness Centrality**		**Betweenness Centrality**	
1	India	37	India	161	India	1	India	0.6301
2	United States	25	United States	106	United States	0.7551	United States	0.0999
3	Netherlands	19	United Kingdom	39	Netherlands	0.6727	Canada	0.025

4	Canada	18	Netherlands	31	Canada	0.6607	Netherlands	0.0173
5	South Africa	16	Canada	27	South Africa	0.6379	South Africa	0.0091
6	Greece	15	Australia	23	Greece	0.6271	Switzerland	0.0086
7	Australia	15	Greece	22	Australia	0.6271	France	0.0082
8	Switzerland	15	South Africa	21	Switzerland	0.6271	Greece	0.0051
9	United Kingdom	13	Switzerland	21	United Kingdom	0.6066	Australia	0.0051
10	France	12	France	18	France	0.5968	Portugal	0.0035

Table 6Centrality measure for international collaboration in Indian big data research output

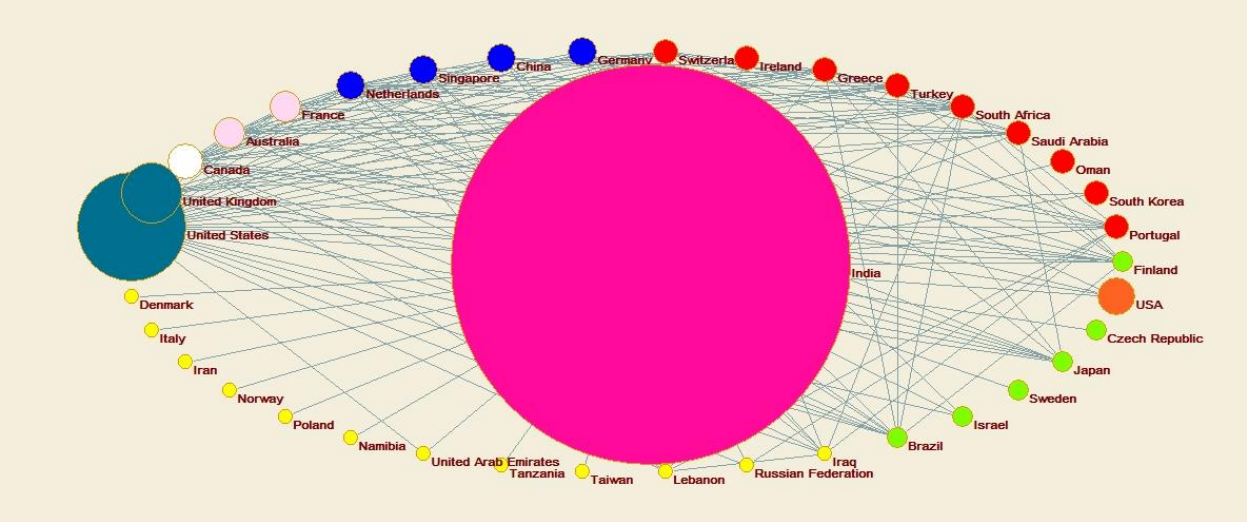

Fig 8 Circular visualization with first partition and first vector of international collaboration of Indian big data research Source: Pajek

Figure 8 shows the visualization of collaboration network of foreign countries with India in big data research in circular graph of network with first partition and first vector. The colour of the nodes reflects different vectors while the size of the nodes is according to the number of publications. It is clear from the figure that the most favoured countries to collaborate in big data research by Indians are United States and United Kingdom while Indian researches used to collaborate with 36 countries in very short period.

Institutional Collaboration:

The ranking of the top 10 institutes according to its frequency of existence is shown in table 7 which clears that the VIT University is the top ranked institution in India which has diffused big data research in India followed by Indian Institute of Technology. During this analysis, each institute with different campuses and places are treated as one and whole. It means the Indian Institute of Technology with different campuses is aggregated as Indian Institute of Technology.

Institute	Frequency (N=1346)	% of frequency
VIT University	59	4.383
Indian Institute of Technology	23	1.708
Sathyabama University	22	1.634
Anna University	19	1.411

Amrita Vishwa Vidyapeetham	14	1.040
Amity University	13	0.965
National Institute of Technology	12	0.891
Indian Statistical Institute	10	0.742
IBM Systems and Technology Group	9	0.668
Indian Institute of Science	9	0.668

Table 7 Ranking of Institutions

The institutional collaboration within India and abroad is shown in figure 9 with both the network visualization and the density visualization in order to produce clear picture of the social network of diffusion of big data in Indian scientific literature. Each colour represents a cluster of collaborated institutes with a particular Indian research institute. It is observed from the study that VIT University, Indian Institute of Technology, Sathyabama Univeristy, Amity University, and National Institute of Technology are the highly collaborative institutes at national level in big data research while Indian Institute of Technology and Amrita University has collaborated at international level and shows the affinity in international collaboration.

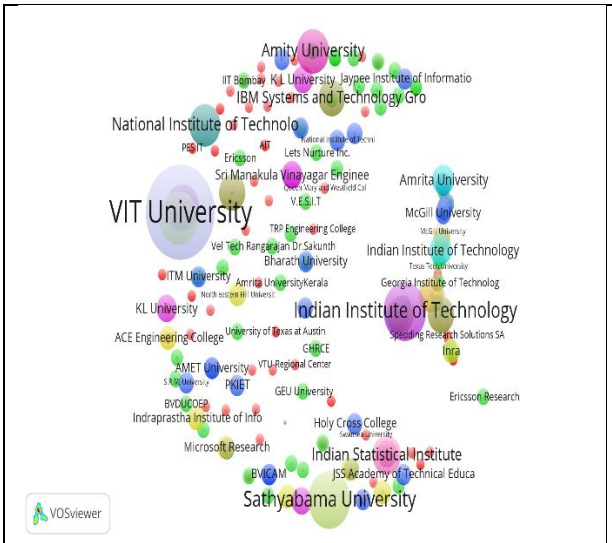

Fig 9A Network visualization of institutional collaboration

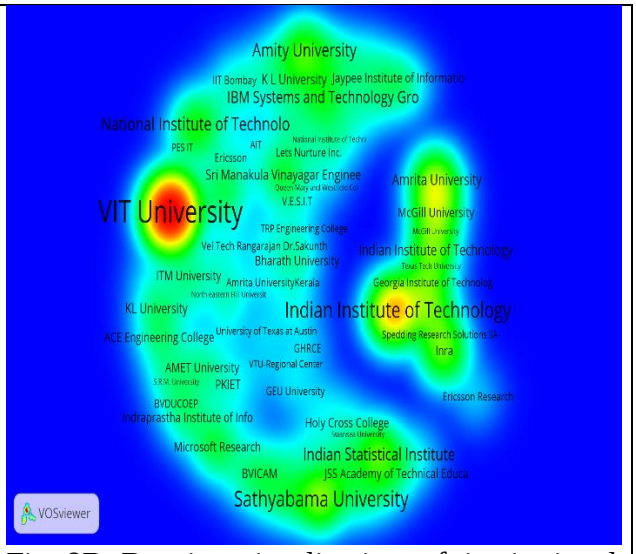

Fig 9B Density visualization of institutional collaboration

Fig 9 Visualization of institutional collaboration of big data research output of India.
Source: VOSviewer

Conclusions:

We have constructed and visualized the social and intellectual endeavor of big data research output of India by using the integrated analysis of social network analysis (SNA), co-occurrence analysis, cluster analysis and frequency analysis of words, and research productivity analysis with bibliometrics. Big data research in India is growing exponentially in recent years. The document type distribution indicates that the conferences and seminars are organized frequently on the big data to discuss and interact with the research groups, so why the conference publications are top ranked for publishing papers. However, we also noted that the density, the

degree centrality, and the betweenness centrality of the whole network were all very low, which indicated that the network was not strongly connected and the collaborative network in the field of big data research in India was very loose. But, It is also observed that the big data research in India is growing very fast so, in future, a remarkable quantity of collaborative work with strong social connect should be observed (Hou et al., 2008).

It is also observed that the most of the prolific authors are seen in the higher collaborative clusters which explain that the prolific authors and role modelers are collaborating more in order to diffuse the big data R&D in India. Institutions like IITs, NITs, Amrita University are collaborating with abroad to represent the quality research in big data at international level as well as national level while VIT University, Sathyabama University, and Amity University are intensively involved in big data research at national level. Researchers from 36 countries are also involved in the diffusion of the big data in Indian research output with

Diffusion of big data knowledge is gaining momentum at present. R & D in Big data continues to flourish due to both micro and macro level collaborations among researchers from different disciplines. Research gaps in big data are observed in health sciences and life sciences. The research output of this study presented with the help of research productivity and SNA analysis of big data research in India will not only help the decision makers to understand the multidisciplinary character of big data but also help and guide to develop funding mechanisms accordingly (Darvish & Tonta, 2016).

REFERENCES:

About Scopus. (2016).Retrieved from: https://www.elsevier.com/solutions/scopus

Ardanuy, J. (2012). Scientific collaboration in Library and Information Science viewed through the Web of Knowledge: The Spanish case. *Scientometrics*, *90*(3), 877–890. http://doi.org/10.1007/s11192-011-0552-1

Batagelj, V., & Marvr, A. (1998). Pajek – program for large network analysis. *Connections*, 47–57. http://doi.org/10.1.1.27.9156

BibExcel. (2016). Retrieved from: http://www8.umu.se/inforsk/Bibexcel

Chow-White, P. A., & Green Jr, S. E. (2013). Data mining difference in the age of big data: Communication and the social shaping of genome technologies from 1998 to 2007. *International Journal of Communication*, *7*(1), 556–583. JOUR. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84874973808&partnerID=40&md5=0273ab9e59b3a6b6aad6c143fab61dba

Darvish, H., & Tonta, Y. (2016). Diffusion of nanotechnology knowledge in Turkey and its network structure. *Scientometrics*, *107*(2), 569–592. http://doi.org/10.1007/s11192-016-1854-0

Eck, N. J. Van, & Waltman, L. (2013). VOSviewer Manual. *1 January 2013*, (January), 1–28. Retrieved from http://www.vosviewer.com/documentation/Manual_VOSviewer_1.5.4.pdf

Fahimnia, B., Sarkis, J., & Davarzani, H. (2015). Int . J . Production Economics Green supply chain management : A review and bibliometric analysis. *Intern. Journal of Production Economics*, *162*, 101–114. http://doi.org/10.1016/j.ijpe.2015.01.003

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239. http://doi.org/10.1016/0378-8733(78)90021-7

Frizzo-Barker, J., Chow-White, P. A., Mozafari, M., & Ha, D. (2016). An empirical study of the rise of big data in business scholarship. *International Journal of Information Management*, *36*(3), 403–413. http://doi.org/10.1016/j.ijinfomgt.2016.01.006

Gartner - IT Glossary. Big Data defintion. (available at: http://www.gartner.com/it-glossary/big-data/) (accessed on: 08.03.16)

Hou, H., Kretschmer, H., Liu, Z., Ou, H. A. H., Retschmer, H. I. K., & Iu, Z. E. L. (2008). The structure of scientific collaboration networks in Scientometrics. *Scientometrics* , *75*(2), 189–202. http://doi.org/10.1007/s11192-007-1771-3

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, *31*(1), 7–15. http://doi.org/10.1016/0020-0190(89)90102-6

Kumar, S. (2015). Co-authorship networks: a review of the literature. *Aslib Journal of Information Management*, *67*(1), 55–73.

Laney, D. (2001). 3D Data Management: Controlling Data Volume,Velocity, and Variety. *Application Delivery Strategies*, *949*(February 2001), 4. http://doi.org/10.1016/j.infsof.2008.09.005

Marr, Bernard, 2015. Why only one of the 5 Vs of big data really matters? (available at: http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters) (accessed on: 08.03.16)

Mrvar, A., & Batagelj, V. (2016). Analysis and visualization of large networks with program package Pajek. *Complex Adaptive Systems Modeling*, *4*(1), 6. http://doi.org/10.1186/s40294-016-0017-8

Park, H. W. (2014). An interview with Loet Leydesdorff: The past, present, and future of the triple helix in the age of big data. *Scientometrics*, *99*(1), 199–202. JOUR. http://doi.org/10.1007/s11192-013-1123-4

Persson, O., Danell, R., & Schneider, J. W. (2009). How to use Bibexcel for various types of bibliometric analysis. *Celebrating Scholarly Communication Studies: A Festschrift for Olle Persson at His 60th Birthday*, 9–24. Retrieved from http://lup.lub.lu.se/record/1458990/file/1458992.pdf#page=11

Rogers, E. M. (1995). *Diffusion of innovations*. *Macmillian Publishing Co.* http://doi.org/citeulike-article-id:126680

Sarwar, R., & Hassan, S.-U. (2015). A bibliometric assessment of scientific productivity and international collaboration of the Islamic World in science and technology (S & T) areas. *Scientometrics*, *105*(2), 1059–1077. http://doi.org/10.1007/s11192-015-1718-z

Singh, V. K., Banshal, S. K., Singhal, K., & Uddin, A. (2015). Scientometric mapping of research on "Big Data." *Scientometrics*, *105*(2), 727–741. JOUR. http://doi.org/10.1007/s11192-015-1729-9

V. Batagelj, a. M. (1998). Pajek – program for large network analysis. *Connections*, 47–57. http://doi.org/10.1.1.27.9156

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. http://doi.org/10.1007/s11192-009-0146-3

Velden, T., & Lagoze, C. (2008). Patterns of Collaboration in Co-authorship Networks in Chemistry - Mesoscopic Analysis and Interpretation.

Wang, L., Thijs, B., & Glänzel, W. (2015). Characteristics of international collaboration in sport sciences publications and its influence on citation impact. *Scientometrics*, *105*(2), 843–862. http://doi.org/10.1007/s11192-015-1735-y

Wasserman, S., & Faust, K. (1994). *Social Network Analysis, Methods and Applications*. Cambridge University Press, New York.

What is Big Data?. *Villanova University*. http://www.villanovau.com/resources/bi/what-is-big-data/#.Vt7sA3pSJV0

Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, *26*(1), 97–107. JOUR. http://doi.org/10.1109/TKDE.2013.109

Yan, E., Ding, Y., & Zhu, Q. (2010). Mapping library and information science in China: A coauthorship network analysis. *Scientometrics*, *83*(1), 115–131. http://doi.org/10.1007/s11192-009-0027-9

Zhu, L., Liu, X., He, S., Shi, J., & Pang, M. (2015). Keywords co-occurrence mapping knowledge domain research base on the theory of Big Data in oil and gas industry. *Scientometrics*, *105*(1), 249–260. JOUR. http://doi.org/10.1007/s11192-015-1658-7