Winter 11-11-2019

Bibliometric Survey of Privacy of Social Media Network Data Publishing

Rupali Gangarde Ass. Prof.
Symbiosis Institute of Technology (SIT) affiliated to Symbiosis International (Deemed University), Pune, India., rupali.gangarde@sitpune.edu.in

Amit Sharma Dr.
Lovely Professional University, Punjab, India., amit.25076@lpu.co.in

Ambika Pawar Dr.
Symbiosis Institute of Technology (SIT) affiliated to Symbiosis International (Deemed University), Pune, India., ambikap@sitpune.edu.in

Bibliometric Survey of Privacy of Social Media Network Data Publishing

Rupali Gangarde[1], Dr. Amit Sharma[2], Dr. Ambika Pawar[3]

[1]Research Scholar at Department of CSE, Lovely Professional University, Punjab India

and Assistant Professor, Symbiosis Institute of Technology (SIT) affiliated to

Symbiosis International (Deemed University), Pune, India.

Email: rupali.41800492@lpu.in

rupali.gangarde@sitpune.edu.in

[2]Ph.D. Guide and Associate Professor, Lovely Professional University, Punjab, India.

Email: amit.25076@lpu.co.in

[3]Ph.D. Guide and Associate Professor, Symbiosis Institute of Technology (SIT) affiliated to Symbiosis International (Deemed University), Pune, India.

Email: ambikap@sitpune.edu.in

ABSTRACT

We are witness to see exponential growth of the social media network since the year 2002. Leading social media networking sites used by people are Twitter, Snapchats, Facebook, Google, and Instagram, etc. The latest global digital report (Chaffey and Ellis-Chadwick 2019) states that there exist more than 800 million current online social media users, and the number is still exploding day by day. Users share their day to day activities such as

their photos and locations etc. on social media platforms. This information gets consumed by third party users, like marketing companies, researchers, and government firms. Depending upon the purpose, there is a possibility of misuse of the user's personal & sensitive information. Users' sensitive information breaches can further utilized for building a personal profile of individual users and also lead to the unlawful tracing of the individual user, which is a major privacy threat. Thus it is essential to first anonymize users' information before sharing it with any of the third parties. Anonymization helps to prevent exposing sensitive information to the third party and avoids its misuse too. But anonymization leads to information loss, which indirectly affects the utility of data; hence, it is necessary to balance between data privacy and utility of data.

This research paper presents a bibliometric analysis of social media privacy and provides the exact scope for future research. The research objective is to analyze different research parameters and get insights into privacy in Social Media Network (OSN). The research paper provides visualization of the big picture of research carried on the privacy of the social media network from the year 2010 to 2019 (covers the span of 19 years). Research data is taken from different online sources such as Google Scholar, Scopus, and Research-gate. Result analysis has been carried out using open source tools such as Gephi and GPS Visualizer.

Maximum publications of privacy of the social media network are from articles and conferences affiliated to the Chinese Academy of Science, followed by the Massachusetts Institute of Technology. Social networking is a frequently used keyword by the researchers in the privacy of the online social media network. Major Contribution in this subject area is by the computer science research community, and the least research contribution is from art and science. This study will clearly give an understanding of contributions in the privacy of social media network by different organizations, types of contributions, more cited papers, Authors contributing more in this area, the number of patents in the area, and overall work done in the area of privacy of social media network.

Keywords: Bibliometric Analysis, Social Media Network, Privacy, Data publishing, Anonymization

1. INTRODUCTION

Online social media network connects the world by providing an online networking platform to users across the globe. Trillions of users' data get generated on social media application provider/owner's side. This data has great potential to do business in today's market. Users' data consists of sensitive information which needs to be protected. Different techniques can be applied to protect the privacy of data. Graph representation social media data is one of the approaches. In a graph-based approach, Social media data is represented in the form of nodes, edges/links, and attributes. For e.g., the Facebookdata set of California Institute of Technology has representation: Node denotes actor or user, an edge that links two actors. The user has a set of attributes like student/faculty status flag, the gender of users.

The privacy of social media data is achieved using anonymization techniques. According to (Cai & Xu, 2018), anonymization is carried out at different levels: node, link, and attribute level anonymization. (Hay, 2007, Cai, 2016 and Qian 2016) suggest node anonymization can be achieved by simple Naive anonymization techniques by replacing all nodes with alphabets and random numbers. (Cai, 2007) The perturbation technique has been used to achieve anonymization by simply adding and deleting nodes. This technique has drawbacks of infiltration, which can be solved by improved algorithms. Also, anonymity has been achieved by grouping similar structure nodes called autotroph means creating paragraphs of a user's with similar characteristics. Thus profiling & tracing of an individual user can be protected even though an adversaries get holds of these sub graphs.

Link anonymization can be achieved by a simple perturbation technique (Dwork, 2011 and Ji, 2016). Another technique to anonymize links is through random walk as presented by authors (Korolova, 2008 and Mittal, 2012). In this method, node traversed in a network by selecting a random node and performing a random walk. A new edge is added at the end of a random walk. Laplacian Noise method used to perform link anonymization (Dwork, 2006 and Krishnamurthy, 2008).

Attribute anonymization is also playing an important role in social media networks. But still not explored heavily and need focus in upcoming research as attributes consist of

sensitive information of users and need to be protected. Sensitive attribute information can be anonymized as it can be used by advertisers. (Samarthi, 2001 & Jiang, 2006) structural similarity follows k-anonymity for making k users anonymous; this protects user privacy.

Future research gaps & challenges

- All of the above methods have an edge, node, or both anonymization, but attribute anonymization remains unexplored for future research.
- The Existing privacy scheme has applied privacy to the K-anonymity level only (Siddula, 2019). There is scope to extend existing work by designing novel algorithms to achieve privacy to higher privacy levels such as l-diversity and T-closeness level.
- Existing research Zobo, 2017) addresses optimal inference attacks by applying sanitization methods like adding an attribute or perturbing (replace one attribute with another) to link and attributes. Present results can be improved by devising other privacy models like k-anonymity and differential privacy.
- To protect the structural similarity of the social media network (Siddula, 2019) proposed clustering approach for the node, link, and attribute anonymization. The privacy of the social media network can improve by applying higher privacy models like l-diversity.

2. INITIAL DATA COLLECTION

Publication database can be accessed using open and paid access (Sarmiento, & Nagi, 1999). A different publication such as Scopus, Mendeley, Google Scholar ScienceDirect, and Research-gate are popular and have a rich database. Scopus is an abstract and citation database and having various disciplines like Health, Life Science, Social Sciences, and physical sciences. The paper focuses on Scopus Database.

2.1 Keyword Phases

The search for the privacy of social media network data publishing performed in two phases. As shown in Table 1, first phase publications search for "privacy of social media network data publishing," and followed by "anonymization" OR "clustering" OR "generalization" as second phase search.

Table 1: Significant keyword search phases

Phases	Keywords
First Phase	"privacy" AND social media network data publishing
Second Phase	"anonymization" OR "clustering" OR "generalization"

Source: http://www.scopus.com (accessed on 21st Oct 2019)

2.2 Publications Count According to Language

The research paper is based on the Scopus database. Preliminary search results with keyword search phases generated a total of 792 publications. Then this study takes into account only the English language with 787 publications, as shown in (Table 2). So for research purposes, only English publications will be considered.

Table 2: Privacy of OSN Scopus documents publishing languages records

Sr. No.	Language	Number of Publications
1	English	787
2	Chinese	4
3	Polish	1
	Total	**792**

Source: http://www.scopus.com (accessed on 21st Oct 2019)

The researchers of the privacy of social media networks publish most of the research papers in articles and conferences. Table 3 summaries that articles and conference proceedings are a rich source of OSN privacy publications, as 71% of papers are published by researchers in both of them.

Table 3: Percentage of privacy of OSN Scopus documents

Publication type	Number of Publications	percentage of 792
Article	355	44.82%
Conference Paper	211	26.64%
Book	100	12.62%
Book Chapter	60	7.57%
Review	54	6.81%
Conference Review	4	0.50%
Note	2	25.00%
Undefined	6	0.75%

Source: http://www.scopus.com (accessed on 21st Oct 2019)

As shown in Figure 1. the total article published 355 papers, and the conference has 211 publications.

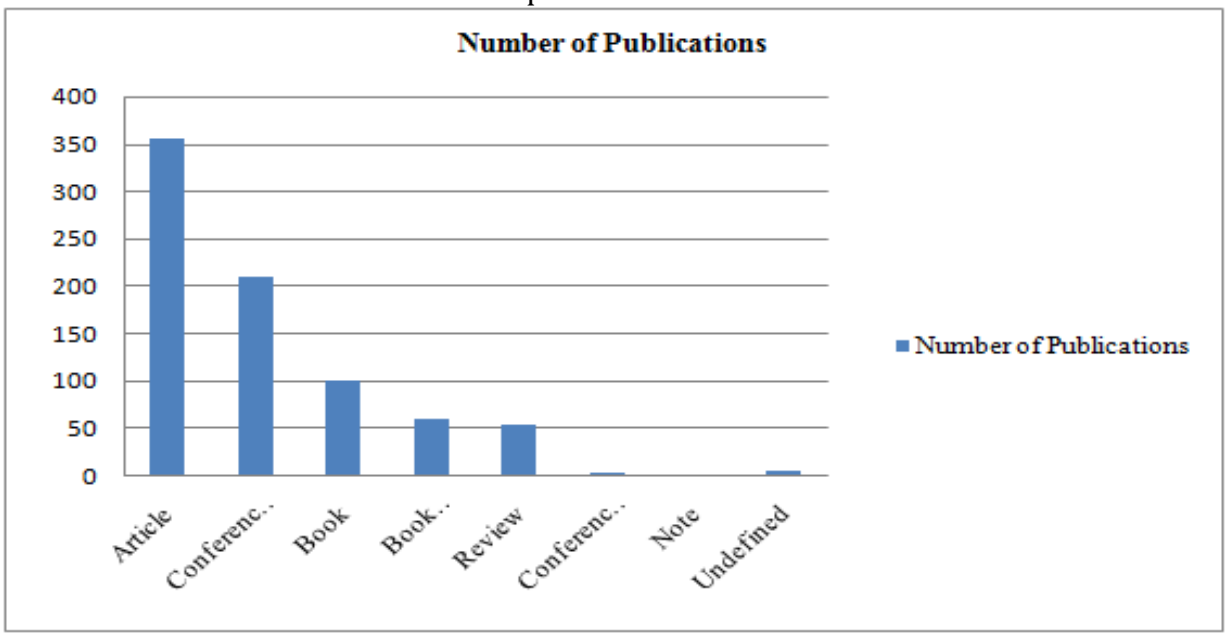


Figure 1: Different Source Types of publications in OSN
Source: http://www.scopus.com (accessed on 21st Oct 2019)

2.3 Publication trends per year

Documents retrieved are journals, conference proceedings, book series, books, and trade publications within the span of 2011 to 2019. Publication trends are shown in Table 4.

Table 4: Privacy of OSN Scopus documents publishing per year

Year	No of scopus publication count
2011	25
2012	33
2013	38
2014	63
2015	98
2016	101
2017	138
2018	155
2019	106

Source: http://www.scopus.com (accessed on 21st Oct 2019)

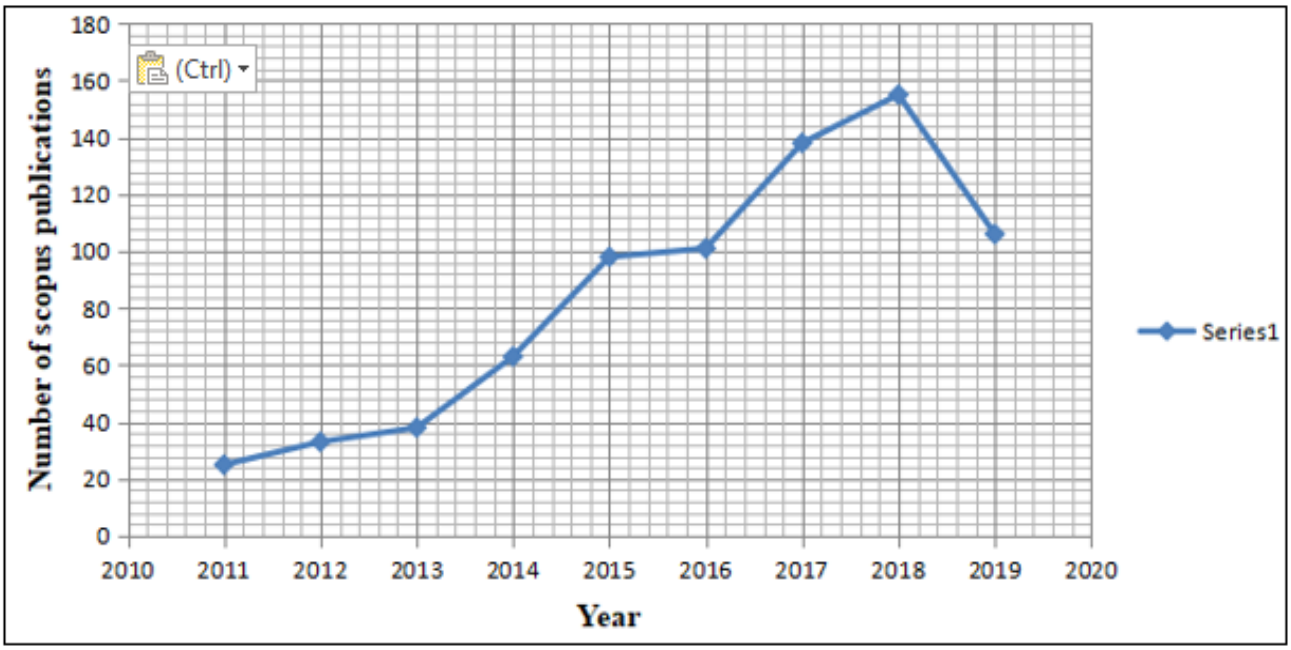


Figure 2: Yearly publishing trend in OSN privacy
Source: http://www.scopus.com (accessed on 21st Oct 2019)

As shown above Figure 2 visualizes the graph, which shows researchers have published a maximum number of Scopus publications in the year 2018 with 155 research publications,

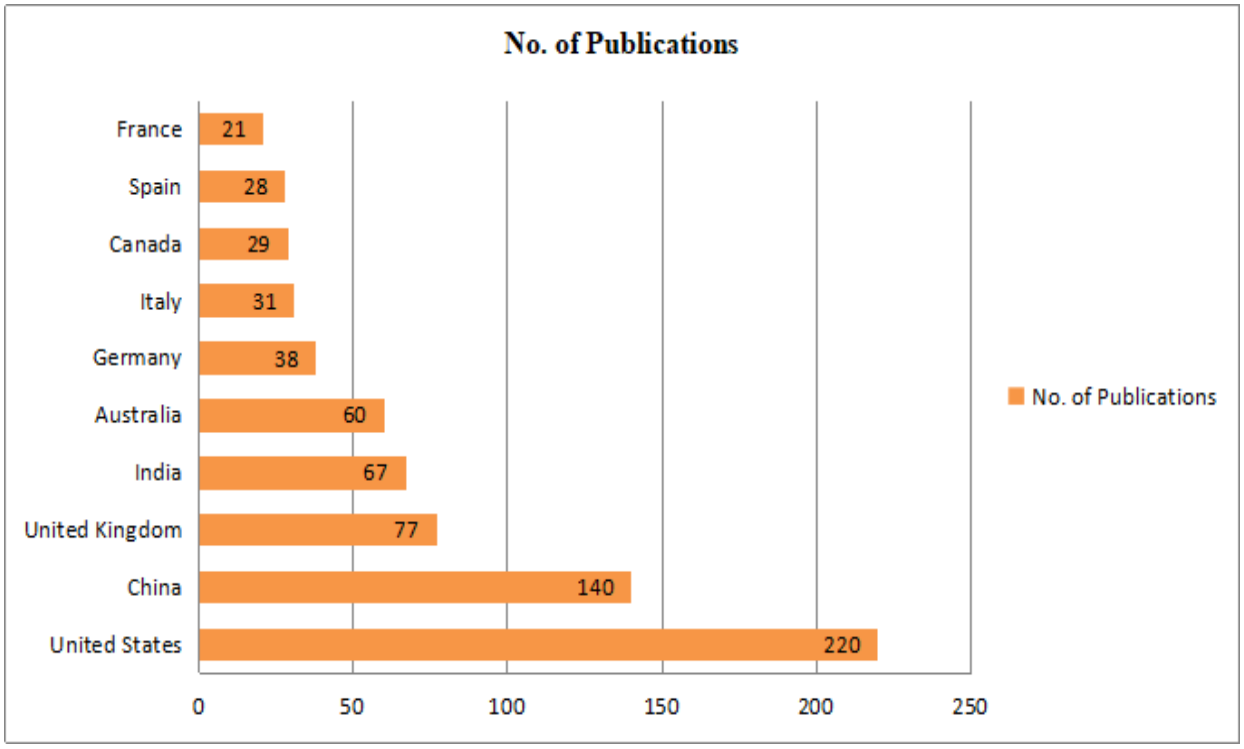
followed by 137 documents in 2017. 2019 is still in a way to increase research publication count.

3. BIBLIOMETRIC ANALYSIS

Bibliometric analysis of privacy social media network concept, performed by the following strategy:

- Survey of OSN research work in different countries and its citation
- Statistics of keywords appearing in publications, authors' affiliation.

Figure 3 gives a clear picture that the United States has led in the privacy of OSN publications (220) followed by China (140) and the United Kingdom with 77publications. France shows fewer publications count i.e., 21 in the research area privacy of the online



social media network.

Figure 3: Highest publishing countries on OSN privacy

Source: http://www.scopus.com (accessed on 21st Oct 2019)

Figure 4 shows the geographic locations of research work done in the area of Privacy of online social media network(OSN). The USA leads in a publication with 30.9%, China with 19.7%, followed by the United Kingdom and India 10.8% and 9.4%, respectively. Canada, Spain, and France have the least contribution in the area of privacy of OSN.

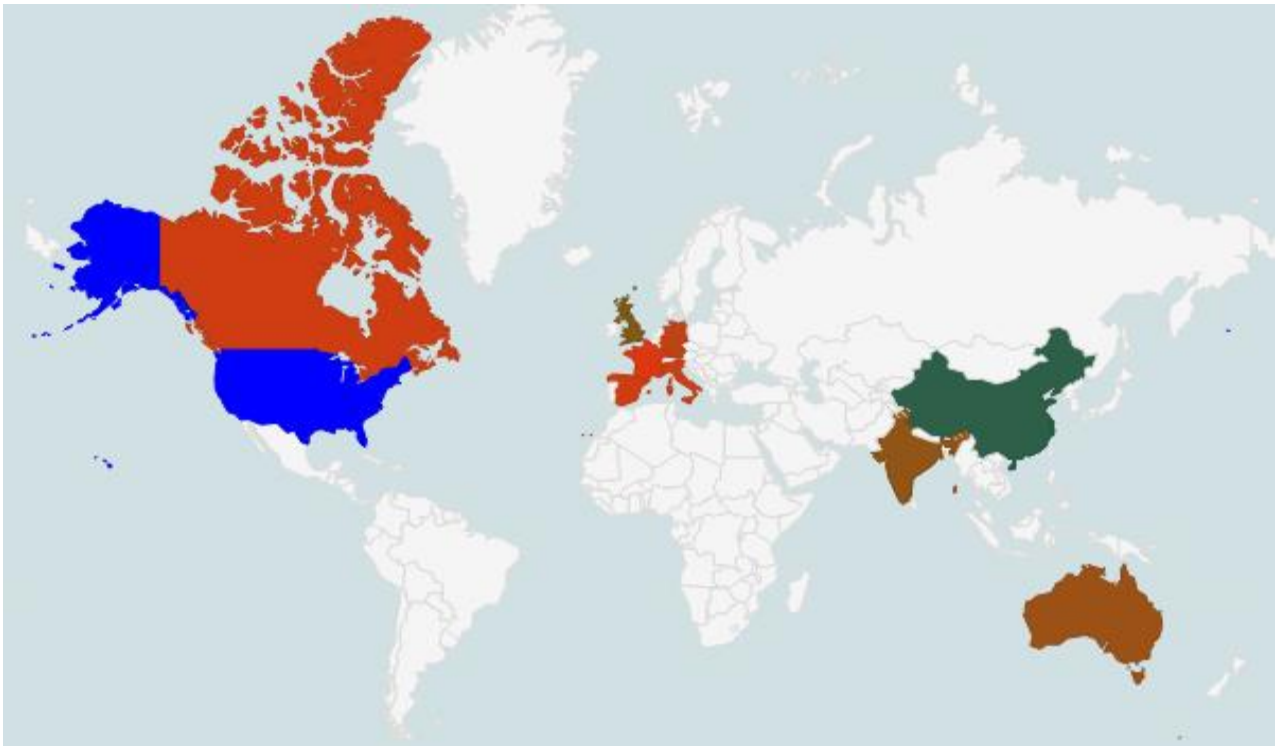


Figure 4: Highest publishing countries on OSN privacy

Source: http://www.scopus.com (accessed on 21st Oct 2019)

3.1 Initial keywords statistics

Keywords represent the interest of the researcher's area of interest. A proper search keyword directs researchers to significant areas. Table 4 shows the top 10 keywords that appear in research publications in the area of privacy of online social media networks. A total of 187 publications have used the top keyword "social networking" in the research publications, followed by the keyword "data privacy" used in 155 research publications. Anonymization keyword is also important as it appeared in the top 10 keyword lists.

Table 4: First ten keywords in OSN privacy

Keywords	Number of Publications
Social networking	187
Data Privacy	155
Data Mining	84
Privacy	82
Big Data	80
Social Media	52
Social Networks	50
Anonymization	49
Privacy Preserving	46
Artificial Intelligence	45

Source: http://www.scopus.com (accessed on 21st Oct 2019)

3.3 Network Analysis

The network of OSN authors presented graphically using "Gephi" open-source software. Gephi visualizes a social media network graphically. Different operations like navigation, manipulation, filtering, and clustering of a social media network data can be performed. Authors, citation authors have, authors' affiliation, publication title, and year, keywords used by authors. In figure 5, there are three essential data clusters for a cluster of author keywords and source title with 840 nodes and 1364 edges.

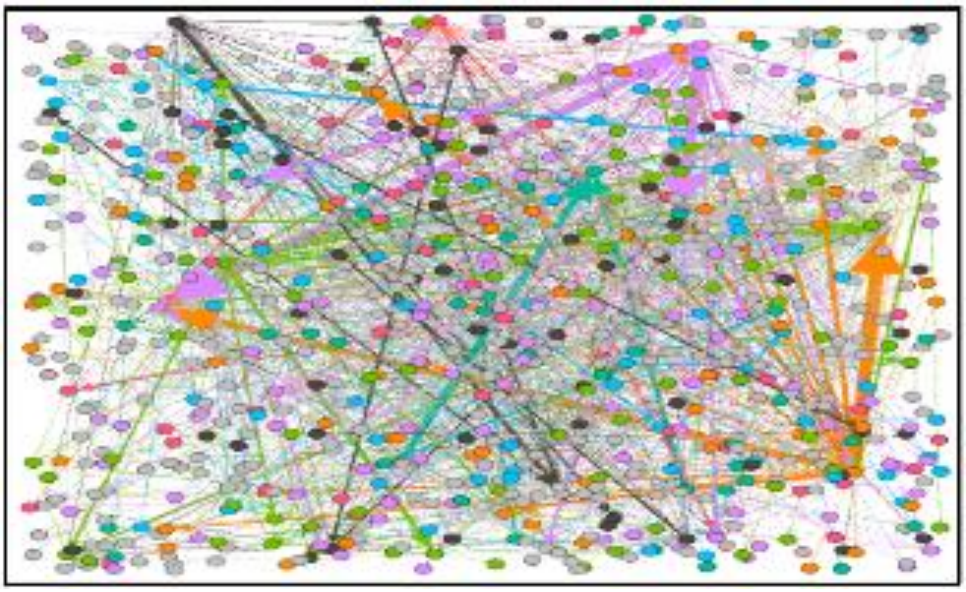


Figure 5: Author keyword and source title cluster
Source: http://www.scopus.com (accessed on 21st Oct 2019)

In figure 6 shows a cluster of publications title, their publication year. Different numbers of publications are mostly from 2010-2019. Figure 6 shows a cluster of networks with the author and keywords in their publications with 768 nodes and 393 edges.

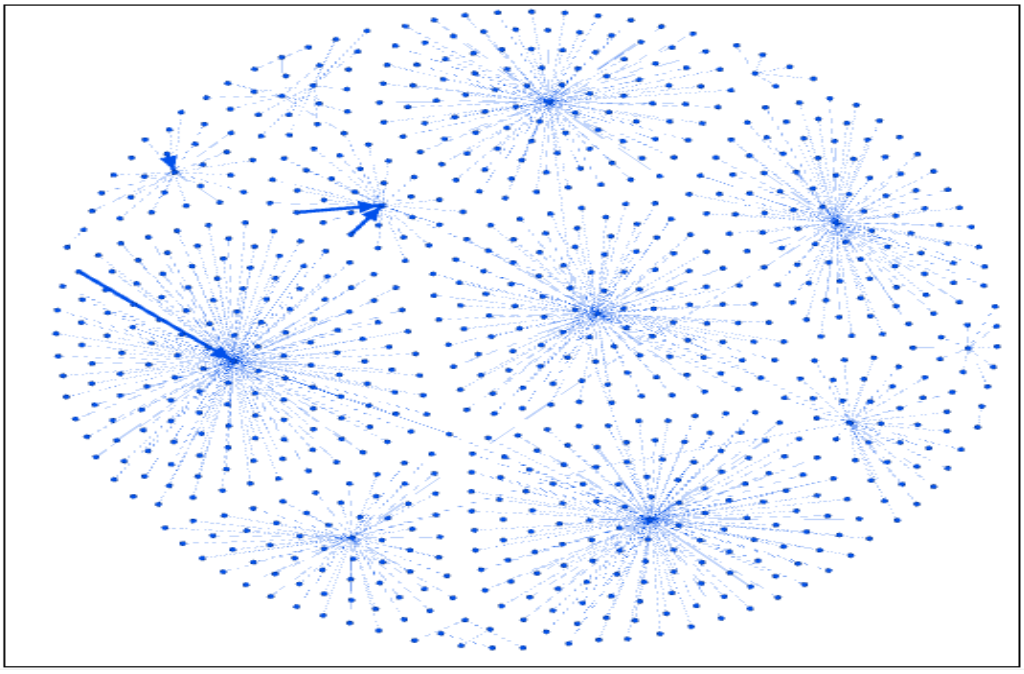


Figure 6: Publication title and publication year
Source: http://www.scopus.com (accessed on 21st Oct 2019)

Figure 7 shows a cluster of authors and keywords. Each author connected with different keywords.

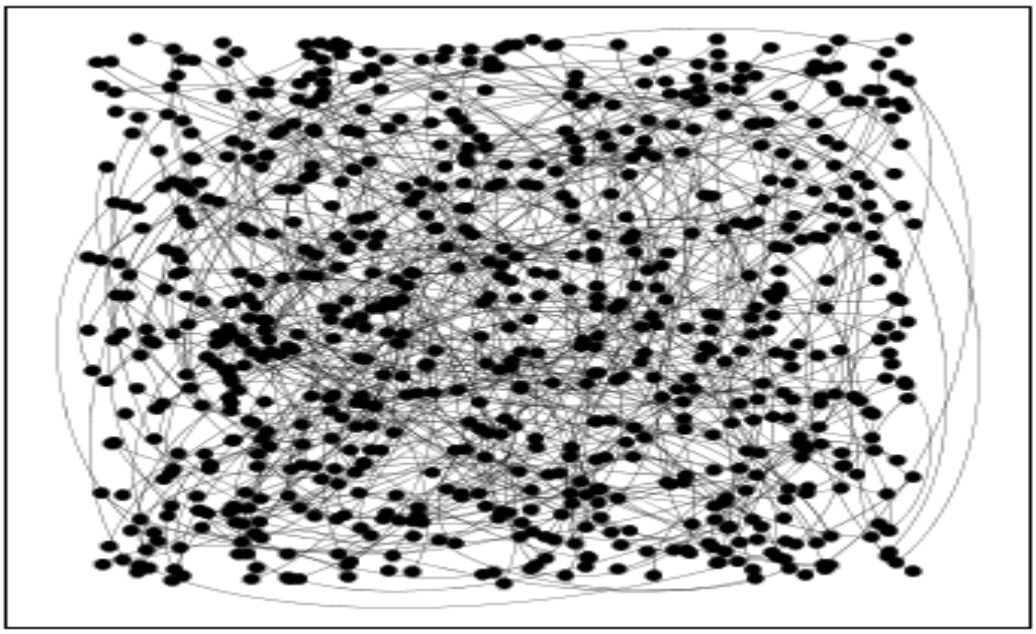


Figure 7: Cluster of authors and keywords
Source: http://www.scopus.com (accessed on 21st Oct 2019)

Figure 8 shows a network of a cluster of year-wise publications citations during 2010-2019.

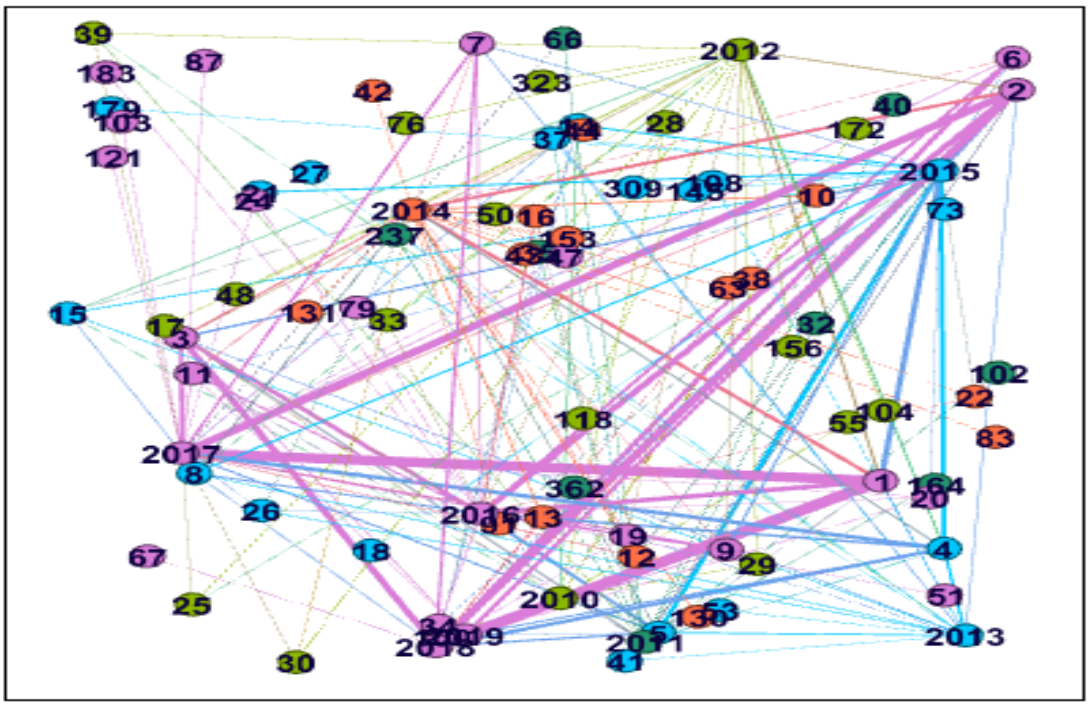


Figure 8: Cluster of publication citations by year from(2010-2019)

Figure 9 shows a cluster of affiliation of authors, the language of publication, and type of publication. ForceAtlas2 algorithm and modularity applied in a Gephi tool do analysis. It has 790 nodes and 1558 edges. Cluster is formed according to types of publications like conference, journal, book, book chapter.

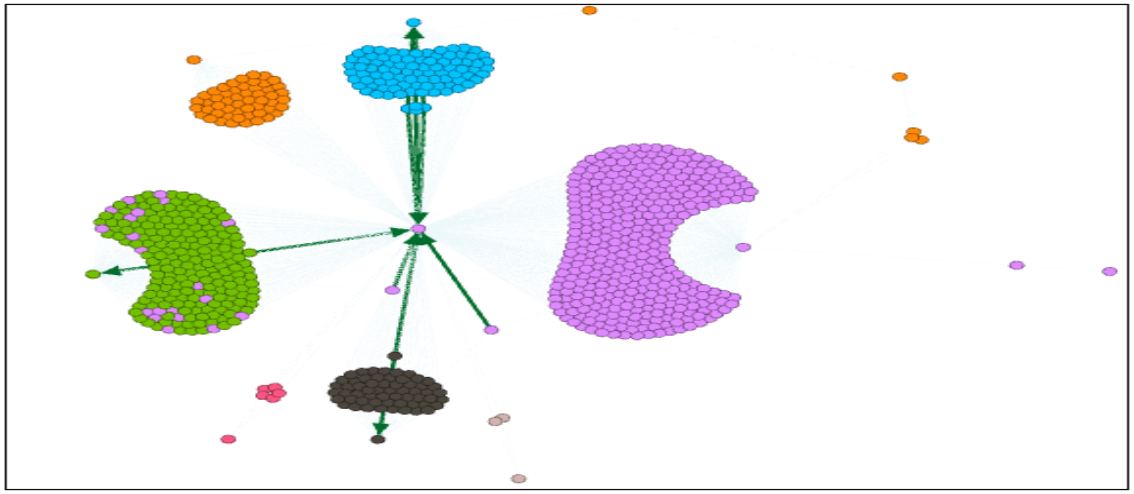


Figure 9: Analysis of affiliation, language, and type
Source: http://www.scopus.com (accessed on 21st Oct 2019)

Figure 10 shows the analysis of citations of publications per year during the span of 2010-2019. Maximum citations to the publication received during the year 2015, followed by 2014 and 2011.

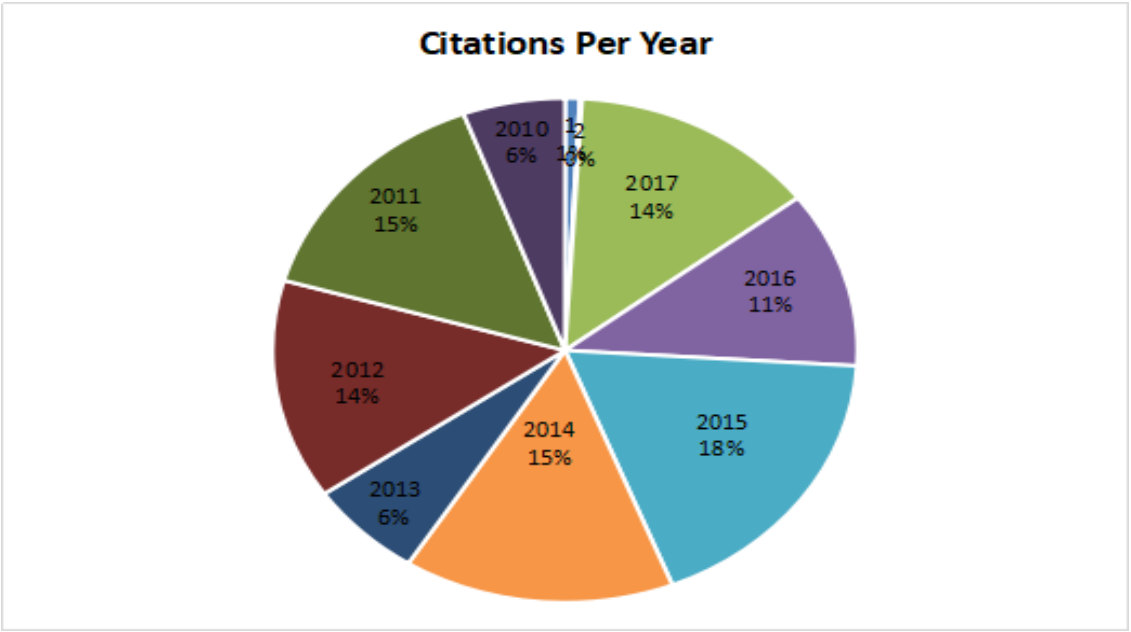


Figure 10: Analysis of citation per year

3.4 Subject Areas

Figure 11 shows the extracted privacy of social media network research publications is partitioned subject wise. According to the pie chart, the Computer science area carried maximum research and followed by Engineering and social sciences.

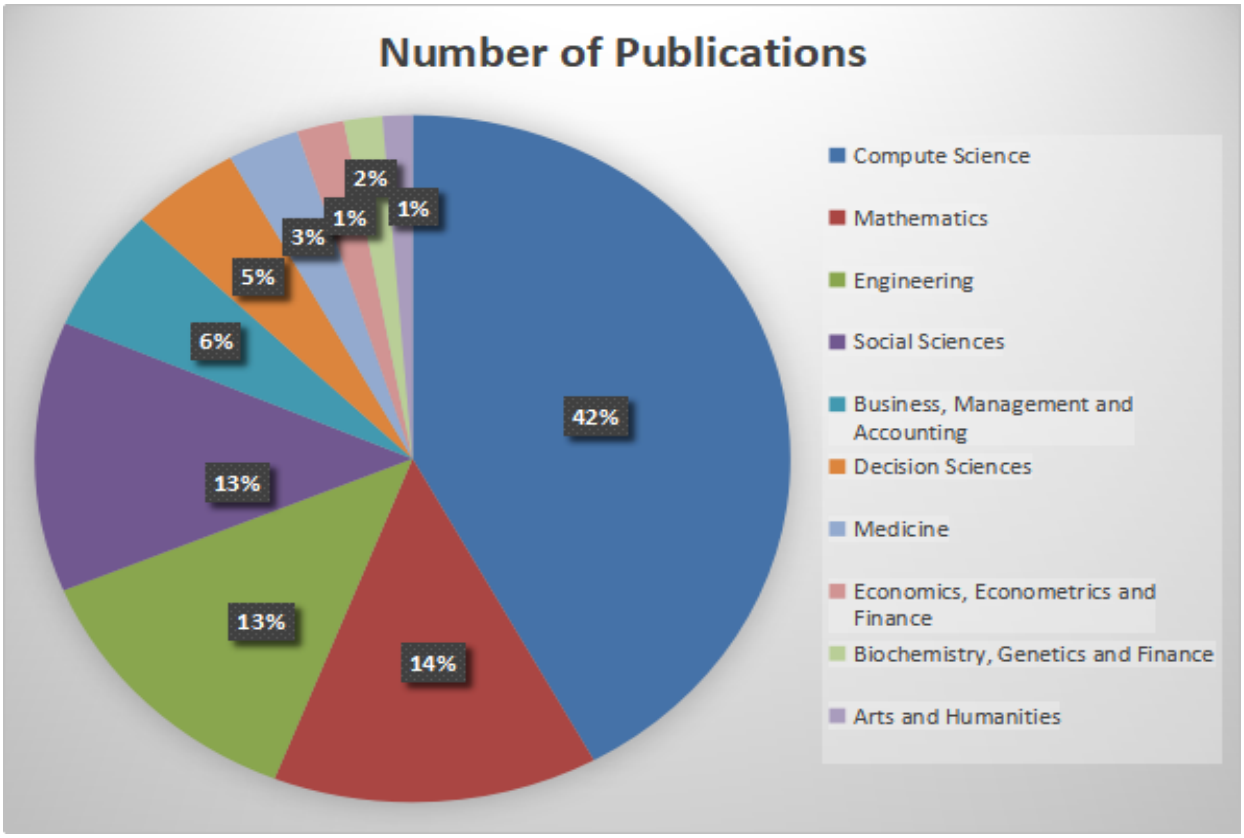


Figure 1:. Different subject areas of literature for OSN privacy
Source: http://www.scopus.com (accessed on 21st Oct 2019)

3.5 Affiliation Statistics for Privacy of OSN

Figure 12 shows the top 10 organizations/ Universities contributing to privacy in OSN research. It indicates the University of Illinois has the highest publications, but overall trends show that following universities near about equally working on the research topic, i.e. the privacy of OSN

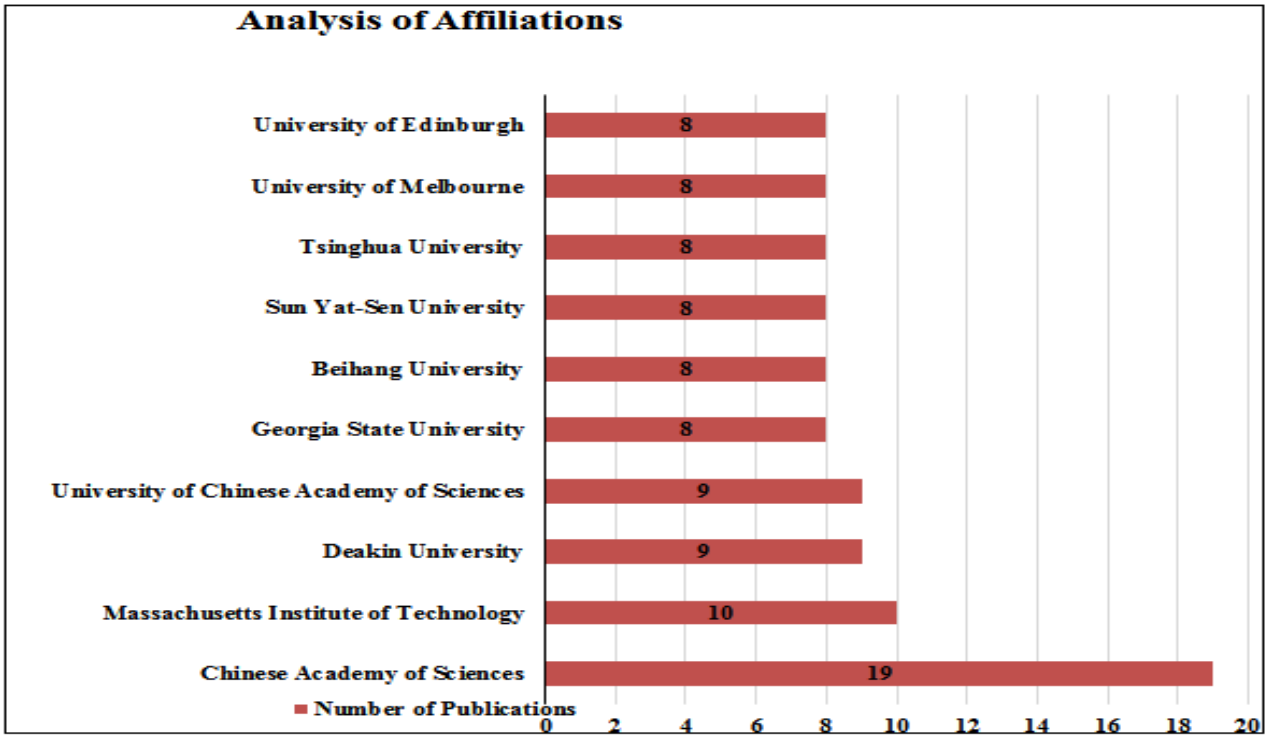


Figure 12: Affiliation statistics for OSN privacy

3.6 Analysis of types of Sources for Publications

Figure 13 shows analysis by different document source type. Journal source has a 51% contribution in the area of privacy of OSN.

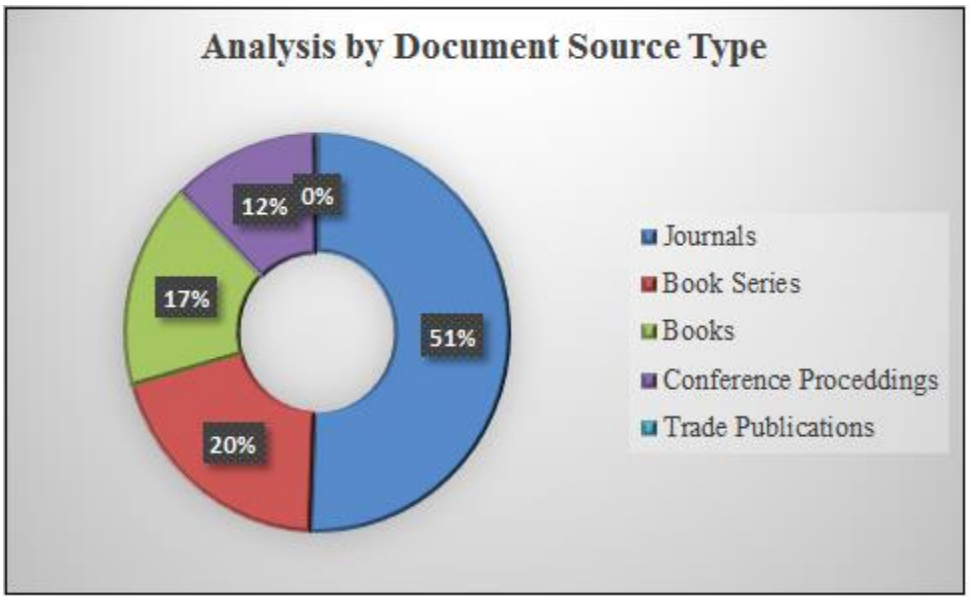


Figure 13: Types of Sources for Publications Publishing in Privacy of OSN
Source: http://www.scopus.com (accessed on 21st Oct 2019)

3.7 Authors Statistics for Privacy of OSN Publications

Figure 14 indicates the highest publishing top 10 authors in the privacy domain of OSN.

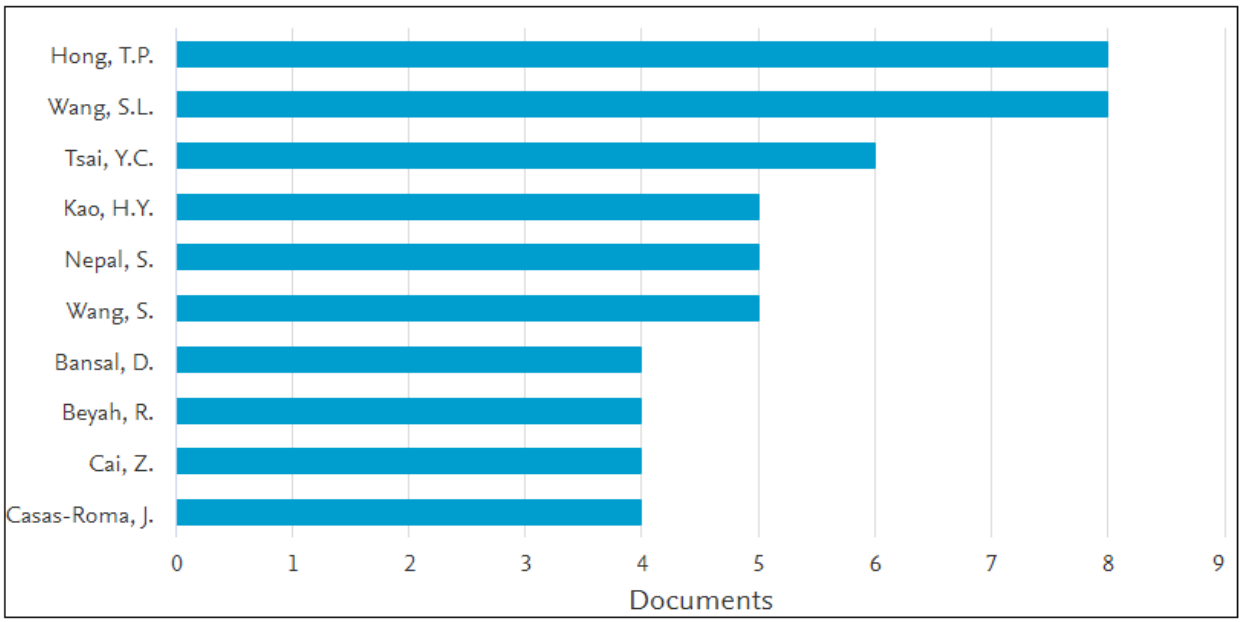


Figure 14: Top Authors Publishing in Privacy of OSN
Source: http://www.scopus.com (accessed on 21st Oct 2019)

3.7 Analysis of Funding sponsor

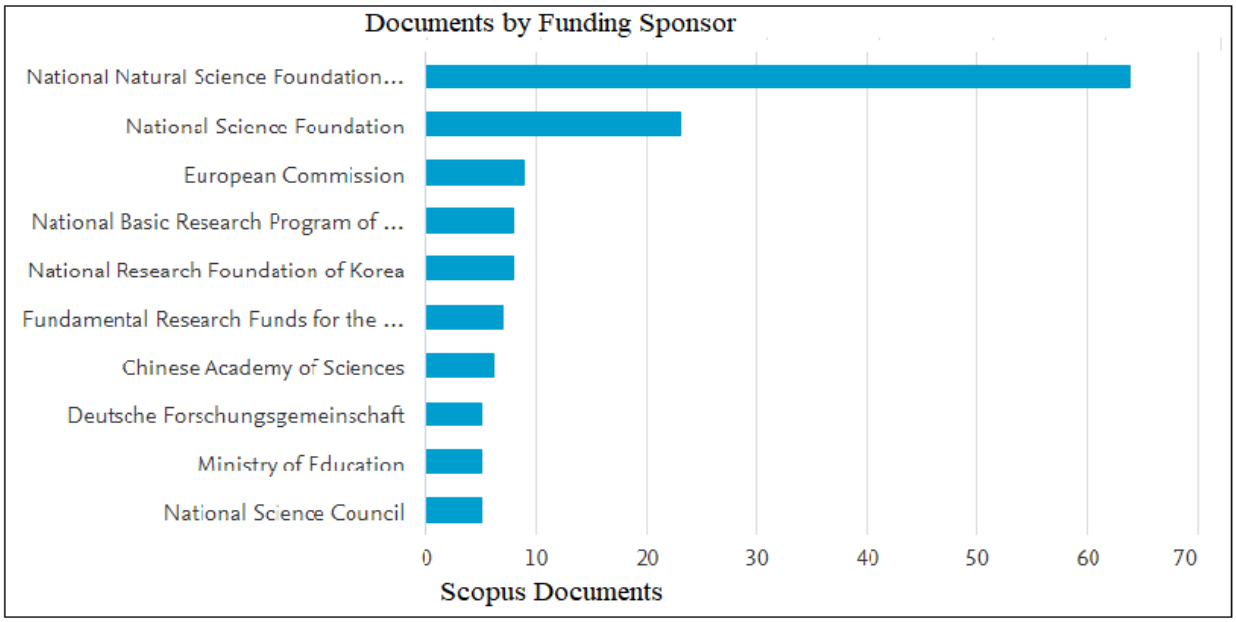


Figure 15: Scopus Documents by Funding Sponsor for Privacy of OSN
Source: http://www.scopus.com (accessed on 21st Oct 2019)

Table 5 shows total of 1341 patents are approved to privacy of OSN area from the year 2010-2019 and Patent office's that grant patents to inventors/applicants. The United States Patent and Trademark Office is topmost with 1278 patents.

Table 5: Patent offices and number of patents

Patent Office	Numbre of Patents
United States Patent and Trademark Office	1278
European Patent Office	34
Japan Patent Office	14
World Intellectual Property Organization	13
United Kigdom Intellectual Property Office	2

Source: http://www.scopus.com (accessed on 21st Oct 2019)

Figure 16 shows the number of patents per year. 2014 has a max, i.e. 15.0% patents among the 2010-2019 year span.

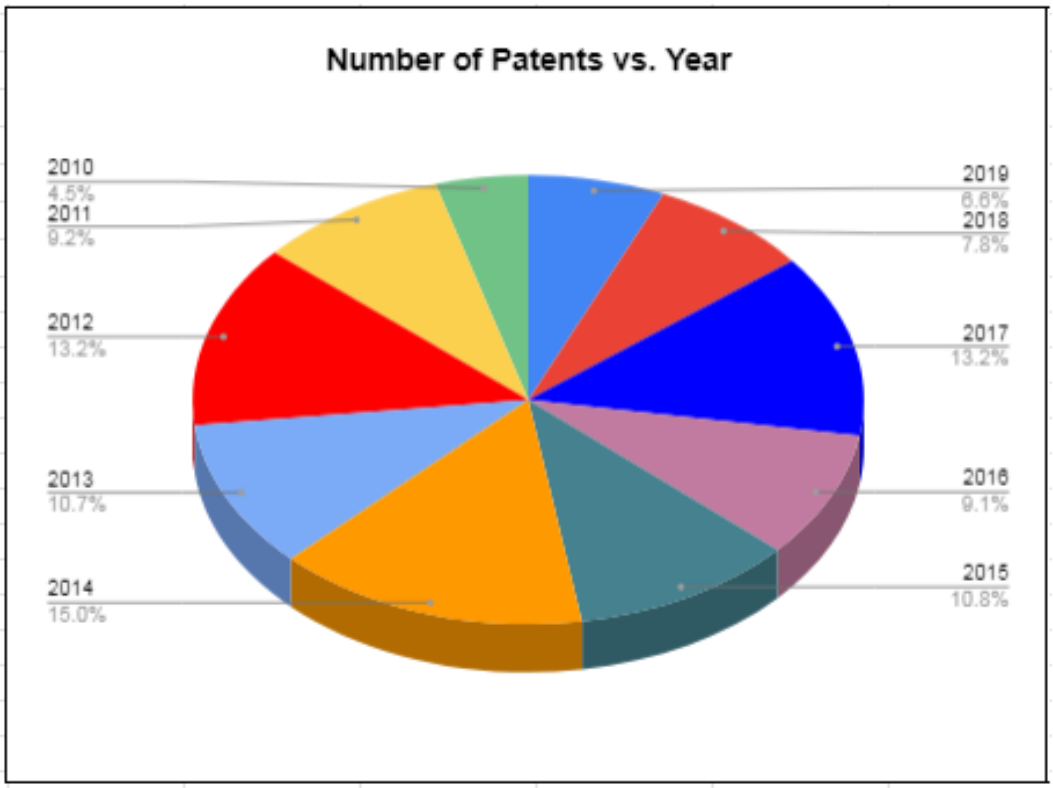


Figure 16: Scopus Documents by Funding Sponsor for Privacy of OSN
Source: http://www.scopus.com (accessed on 21st Oct 2019)

3.8 Citation Analysis of Research Area

Table 6 presents yearly citations obtained through publications in the area of privacy of social media networks. The overall citation count in the research area, with 799 documents, is 9537 to date. Table 7 shows a list of the first ten papers and citations received by these research papers to date.

Table 6: a Publication citation analysis of privacy of Social Media Network

Year	**<2010**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**Total**
No. of Citations	216	137	214	319	463	465	783	1105	1612	2146	2194	**9537**

Table 7: Top citation Publication analysis of privacy of Social Media Network

Publication Title	**<2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**Total**
Routes for breaching and protecting genetic privacy	5	36	27	37	26	22	148
Class based graph anonymization for social network data	60	13	18	9	15	10	125
Sharing graphs using differentially private graph models	23	16	14	12	20	18	103
Audience selection for on-line brand advertising: Privacy-friendly social network targeting	48	6	7	7	7	3	30

Collective Data-Sanitization for Preventing Sensitive Information Inference Attacks in Social Networks	0	0	0	12	30	28	70
Community-enhanced de-anonymization of online social networks	0	2	9	7	11	15	44
User k-anonymity for privacy-preserving data mining of query logs	14	9	8	4	2	2	25
Privacy-preserving social network data publication	0	0	1	8	15	9	33
De-anonymizing social networks and inferring private attributes.	0	0	1	8	15	9	33
The complexity of social network anonymization	7	6	6	4	0	4	20

REFERENCES

Chaffey, D., & Ellis-Chadwick, F. (2019). *Digital marketing*. Pearson UK.

Cai, Z., & Zheng, X. (2018). A private and efficient mechanism for data uploading in smart cyber-physical systems. *IEEE Transactions on Network Science and Engineering*.

Narayanan, A., & Shmatikov, V. (2009). De-anonymizing social networks. *arXiv preprint arXiv:0903.3276*.

Cai, Z., He, Z., Guan, X., & Li, Y. (2016). Collective data-sanitization for preventing sensitive information inference attacks in social networks. *IEEE Transactions on Dependable and Secure Computing*, *15*(4), 577-590.

Dwork, C. (2011). Differential privacy. *Encyclopedia of Cryptography and Security*, 338-340.

Qian, J., Li, X. Y., Zhang, C., & Chen, L. (2016, April). De-anonymizing social networks and inferring private attributes using knowledge graphs. In *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications* (pp. 1-9). IEEE.

Ji, S., Mittal, P., & Beyah, R. (2016). Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey. *IEEE Communications Surveys & Tutorials*, *19*(2), 1305-1326.

Korolova, A., Motwani, R., Nabar, S. U., & Xu, Y. (2008, October). Link privacy in social networks. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 289-298). ACM.

Mittal, P., Papamanthou, C., & Song, D. (2012). Preserving link privacy in social network based systems. *arXiv preprint arXiv:1208.6189*.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006, March). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference* (pp. 265-284). Springer, Berlin, Heidelberg.

Krishnamurthy, B., & Wills, C. E. (2008, August). Characterizing privacy in online social networks. In *Proceedings of the first workshop on Online social networks* (pp. 37-42). ACM.

Samarati, P. (2001). Protecting respondentΓs Privacy in Microdata release. *IEEE Transactions on Knowledge and Data Engineering*, *13*(6).

Jiang, W., & Clifton, C. (2006). A secure distributed framework for achieving k-anonymity. *The VLDB Journal—The International Journal on Very Large Data Bases*, *15*(4), 316-333.

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006, April). l-diversity: Privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)* (pp. 24-24). IEEE.

Goldberger, J., & Tassa, T. (2009, December). Efficient anonymizations with enhanced utility. In *2009 IEEE International Conference on Data Mining Workshops* (pp. 106-113). IEEE.

He, Z., Cai, Z., & Yu, J. (2017). Latent-data privacy preserving with customized data utility for social network data. *IEEE Transactions on Vehicular Technology*, *67*(1), 665-673.

Siddula, M., Li, Y., Cheng, X., Tian, Z., & Cai, Z. (2019). Anonymization in Online Social Networks Based on Enhanced Equi-Cardinal Clustering. *IEEE Transactions on Computational Social Systems*, *6*(4), 809-820.

Sarmiento, A. M., & Nagi, R. (1999). A review of integrated analysis of production-distribution systems. *IIE transactions*, *31*(11), 1061-1074.