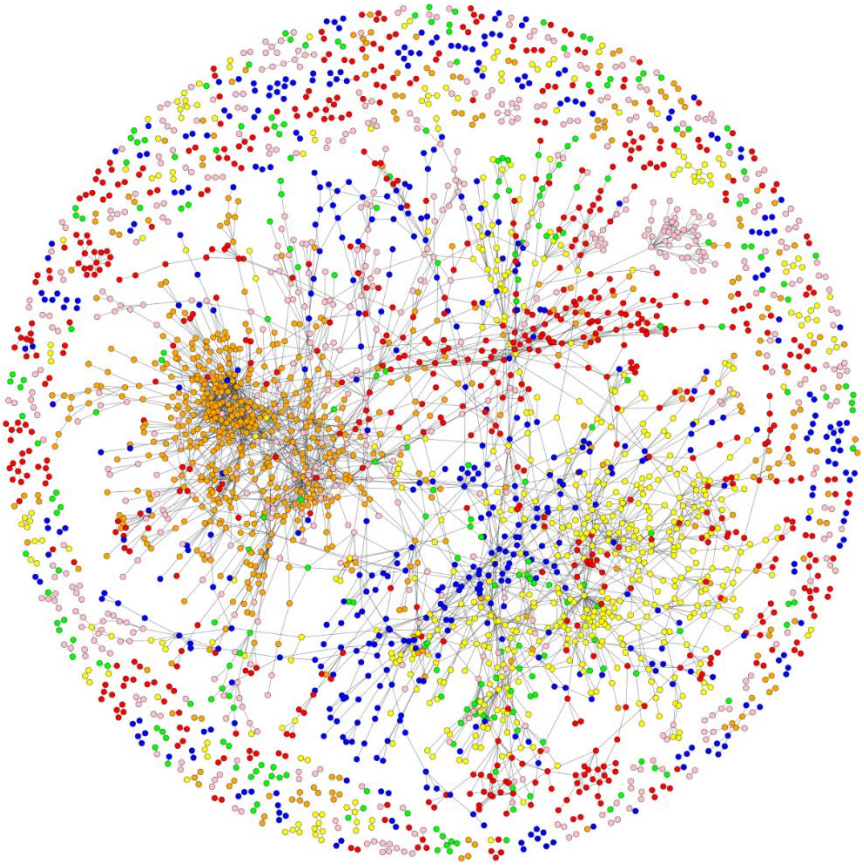
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Social Network Analysis

Assignment 2



Name : Ilias Dimos

AM : f2822102

Professor : Katia Papakonstantinopoulou

Table of Context

1) DBLP co-authorship graph

Question

You will first manipulate the raw data with the programming language of your choice to filter out all records that are not related to the five (5) conferences listed above, e.g, CIKM, or are older than 5 years. Then, you will create a total of 5 .csv files, one for each of the last 5 years, using the following format:
from,to,weight
author1,author2,5
author1,author3,2
Each .csv file should describe the weighted undirected co-authorship graph for the respective year, e.g., in the example above author1 has co-authored 5 papers with author2, and 2 with author3. Having created the .csv files it will be trivial to use them and create the respective igraph graphs.

Answer

In order to achieve the proper format of the raw data file we have created the Jupyter Notebook called **"Data_Transformation_SNA."**
After the proper insertion of the data, we had to filter the data in order to take only a data frame with the five conferences we are interested in. The next step was to filter the dataset furthermore in order to split it into five sub datasets based on the year of the publish of the books.
It was important to check for missing values because they were creating problems in the functions, we used to find the pairs of the authors. We managed to delete only one missing value from the dataset of 2018 in the WWW conference.
After that we took only the authors of each dataset, we split them using the comma separator and transformed the data frames into lists. This step was necessary in order for the function that creates the pairs to work properly.
After the correct computation of the pairs our final output can be shown in the figure 1 below :

```
[(('Philip S. Yu', 'Sihong Xie'), 6),
 (('Huan Liu 0001', 'Jiliang Tang'), 6),
 (('Quanzhi Li', 'Rui Fang'), 6),
 (('Quanzhi Li', 'Sameena Shah'), 6),
 (('Rui Fang', 'Sameena Shah'), 6),
 (('Chun-Ta Lu', 'Philip S. Yu'), 5)]
```

Figure 1 : Example of the Pair Computation of the authors in 2016

We can observe that the function computes how many times the two authors have been found together.
Finally with the proper commands and regular expressions we managed :

1) To remove first the parenthesis.
2) Split the authors into three columns named "from" "to" and "weight" based on the assignment's instructions.
3) And remove the quotes from the author names.

The last step was to remove any white spaces may have occurred from the data transformation and save the 5 csv's into our working directory in order to load them into the R programming Language to answer the rest of the assignment's questions. Also, it is very important to mention that there were 40 rows of the initial dataset that were completely destroyed and could not be fixed so we dropped them.

2) Average degree over time

Question

Your next task is to create plots that visualize the 5-year evolution of different metrics for the graph. More specifically, you will create plots for:

• Number of vertices

• Number of edges

• Diameter of the graph

• Average degree (simple, not weighted)

What do you notice for each of the 5 above metrics? Are there significant fluctuations during these five years?

Answer

Number of vertices

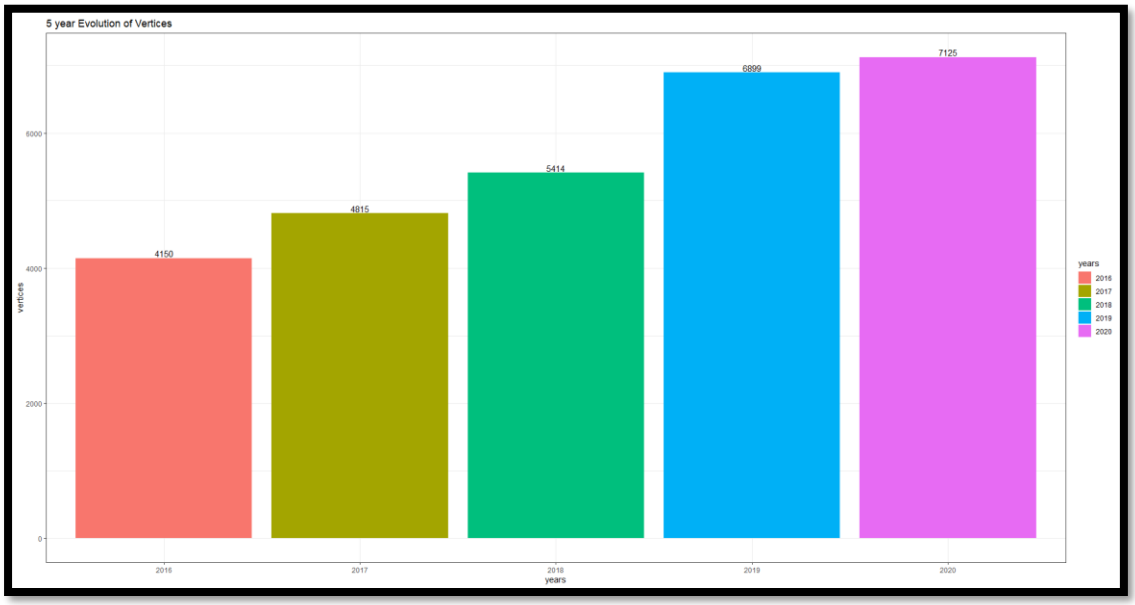


Figure 2 Evolution of Vertices from 2016 to 2020

The number of the vertices has an increasing rate over the years from 2016 to 2020 meaning that more authors contribute to the conferences we are studying.

Number of Edges

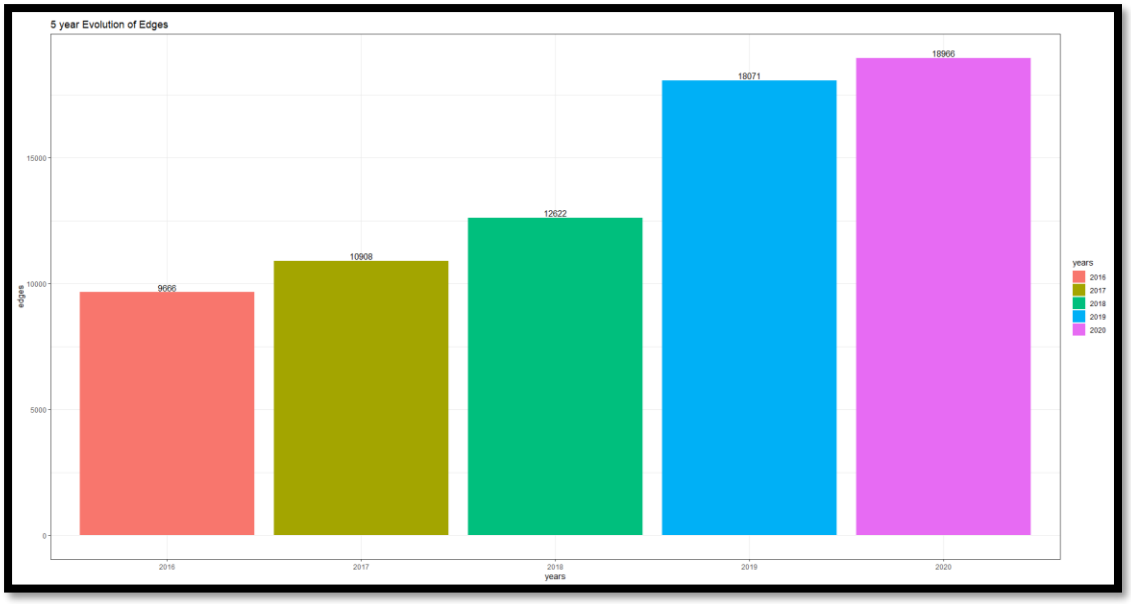


Figure 3 Evolution of Edges from 2016 to 2020

The number of edges also have an increasing rate over these years meaning that over these years the authors tend to connect with each other.

Diameter of the graph

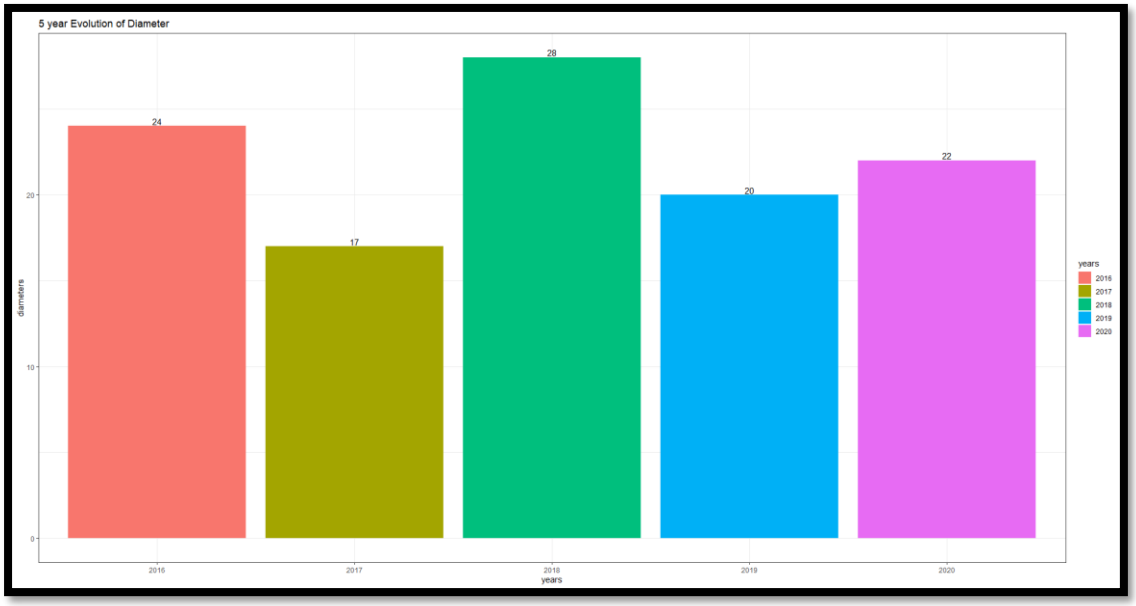


Figure 4 Evolution of Diameter from 2016 to 2020

The diameter was 24 in 2016 , decreased to 17 in 2017 and reached its peak in 2018 with 28. The decrease occurs because in this period the paths between the authors were shorter than the last year. Also, we have to mention that the paths between the authors were significantly bigger in 2018.

Average degree (simple, not weighted)

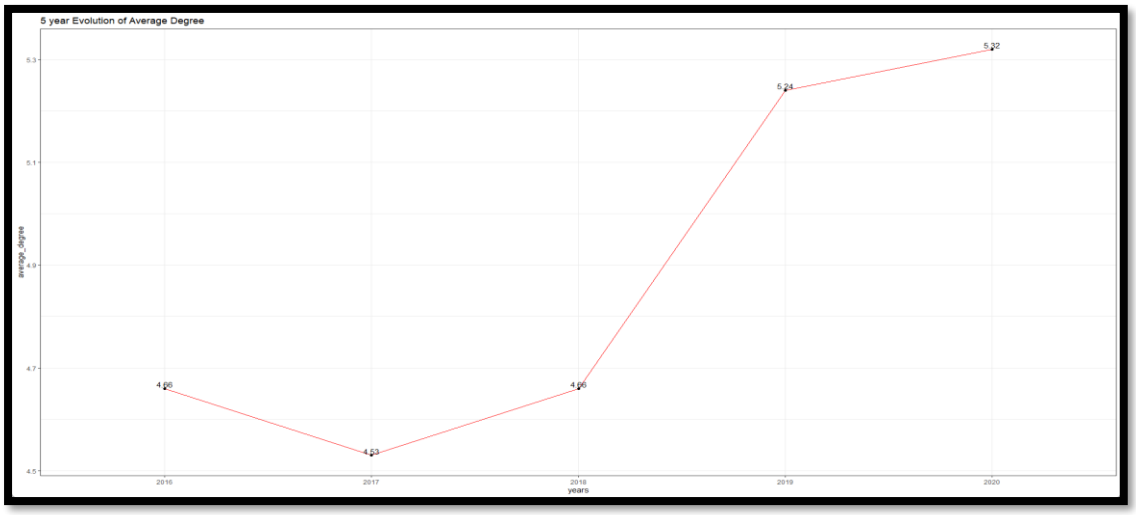


Figure 5 Evolution of Average Degree from 2016 to 2020

In figure 5 we can observe that the average degree starts from 4.66 in 2016, drops to 4.53 in 2017, and from 2018 to 2020 there is a significantly big increase reaching the maximum value of 5.32 meaning that every node of the network has approximately 5.32 neighbors.

3) Important nodes

Next, you will write to code to create and print data frames for the 5-year evolution of the top-10 authors with regard to:

• *Degree (simple, not weighted)*

• *PageRank*

Again, provide short comments on your findings. Do you notice variations on the top-10 lists for the different years?

Answer

Degree (simple, not weighted)

	Degree Score 2016
Philip S. Yu	46
Jiawei Han 0001	41
Hui Xiong 0001	39
Jieping Ye	32
Naren Ramakrishnan	32
Yi Chang 0001	31
Jiebo Luo	29
Rayid Ghani	28
Chang-Tien Lu	25
Yannis Kotidis	25

	Degree Score 2017
Philip S. Yu	44
Jiawei Han 0001	42
Hui Xiong 0001	38
Claudio Rossi 0003	32
Yi Chang 0001	32
Clemens Mewald	31
Heng-Tze Cheng	31
Martin Wicke	31
Mustafa Ispir	31
Zakaria Haque	31

	Degree Score 2018
Philip S. Yu	70
Jiawei Han 0001	37
Kun Gai	35
Wenwu Zhu 0001	28
Chao Zhang 0014	27
Jing Gao 0004	27
Jure Leskovec	27
Xing Xie 0001	26
Enhong Chen	25
Haifeng Chen	25

	Degree Score 2019
Philip S. Yu	69
Weinan Zhang 0001	59
Hui Xiong 0001	49
Jieping Ye	41
Jie Tang 0001	39
Jiawei Han 0001	37
Enhong Chen	36
Yong Li 0008	36
Jian Pei	35
Jingren Zhou	35

	Degree Score 2020
Jiawei Han 0001	69
Hongxia Yang	43
Hui Xiong 0001	42
Xiuqiang He	41
Ji Zhang	40
Peng Cui 0001	39
Christos Faloutsos	38
Wei Wang 0010	38
Jieping Ye	37
Jiliang Tang	35

Figure 6 Top 10 Degree Scores from 2016 to 2020

In figure 6 we can observe that in 2016 the authors with the highest degree score are approximatelly the same with 2019 with the Philip S. Yu being the number 1. In 2020 the degree score rating has completely changed and new authors are appearing. In general we observe that the degree is increasing over these 5 years meaining that the specefic authors are gainning a lot neigheibors (co-authors in our case).

PageRank

	PageRank Score 2016
Philip S. Yu	0.0017288334
Hui Xiong 0001	0.0014581015
Jiawei Han 0001	0.0014119510
Jiebo Luo	0.0013099364
Jieping Ye	0.0010027077
Yi Chang 0001	0.0009601005
Hanghang Tong	0.0009272920
Christos Faloutsos	0.0009216757
Maarten de Rijke	0.0009158533
Jiliang Tang	0.0009155034

	PageRank Score 2017
Philip S. Yu	0.0014558956
Jiawei Han 0001	0.0013585699
Hui Xiong 0001	0.0010997688
Jure Leskovec	0.0010681579
Jiebo Luo	0.0009454158
Hanghang Tong	0.0009285808
Jiliang Tang	0.0007750644
Yi Chang 0001	0.0007711858
Chao Zhang 0014	0.0007510406
Ingmar Weber	0.0007208090

	PageRank Score 2018
Philip S. Yu	0.0019809631
Jiawei Han 0001	0.0009301987
Jure Leskovec	0.0008753490
Wenwu Zhu 0001	0.0007842984
Chao Zhang 0014	0.0006775310
Xing Xie 0001	0.0006263373
Jing Gao 0004	0.0006259877
Martin Ester	0.0006201636
Yiqun Liu 0001	0.0006143691
Kun Gai	0.0006129884

	PageRank Score 2019
Philip S. Yu	0.0015871036
Hui Xiong 0001	0.0009633261
Weinan Zhang 0001	0.0008767308
Jieping Ye	0.0007255196
Hanghang Tong	0.0007021244
Jiawei Han 0001	0.0006855583
Peng Cui 0001	0.0006574207
Jie Tang 0001	0.0006517701
Enhong Chen	0.0006377621
Gerhard Weikum	0.0006257373

	PageRank Score 2020
Jiawei Han 0001	0.0010753255
Hui Xiong 0001	0.0007594661
Hongxia Yang	0.0007284981
Elke A. Rundensteiner	0.0006983864
Yong Li 0008	0.0006821198
Jieping Ye	0.0006800497
Peng Cui 0001	0.0006533883
Xiuqiang He	0.0006465968
Ji-Rong Wen	0.0006450074
Jiliang Tang	0.0006423610

Figure 7 Top 10 PageRank Scores from 2016 to 2020

In figure 7 We observe that in 2016 to 2017 the authors with the biggiest PageRank apear to be the same with 2 new entries in 2017. As the years go by the PageRank changes , reaching in 2020 were the top 10 authors are complete different in comparison with the previous years.

4) Communities

<u>Question</u>

Your final task is to perform community detection on the mention graphs. Try applying fast greedy clustering, infomap clustering, and louvain clustering on the 5 undirected co-authorship graphs. Are you able to get results with all methods? Include a short comment on your report regarding the performance of the 3 algorithms.

Then, pick one of the three methods as well as a random author that appears in all 5 graphs and write code to detect the evolution of the communities this user belongs to. Do you spot similarities in the communities?

Finally, you will create a visualization of the graph using a different color for each community. Make sure to have a look at the sizes of the communities and filter out all nodes that belong to very small or very large communities, in order to create a meaningful and aesthetically pleasing visualization.

Answer

In order to choose the best algorithm to continue we measured the speed time of each one of them. In the below table we can see the run times of each cluster algorithm.

Cluster Algorithm	Run Time
fast greedy clustering	0.393976 secs
infomap clustering	6.037862 secs
louvain clustering	0.4449248 secs

Based on the run time we see that the fast greedy clustering is the fastest method among the other two so we will proceed with this one.

To proceed to the other sub questions, we need to choose one author that appears in the 5 years. We choose the author named **Christos Faloutsos.**

The specific author has been appeared in those 5 years with the fast greedy clustering algorithm. In the table above we can see the ID of each community that the specific author belongs to from the years 2016 to 2020.

Years	Author	Community ID
2016	Christos Faloutsos	6
2017	Christos Faloutsos	51
2018	Christos Faloutsos	203
2019	Christos Faloutsos	4
2020	Christos Faloutsos	2

Now that we have found the IDs of the communities, we need to find their sizes in order make a plot to show the evolution of them among the 5 years. In the Table below we can observe the sizes of each one of the five communities.

Years	Community ID	Size of the Community
2016	6	78
2017	51	120
2018	203	23
2019	4	64
2020	2	316

In the figure 8 below we can see a visualization of the evolution of the community's the user belongs to

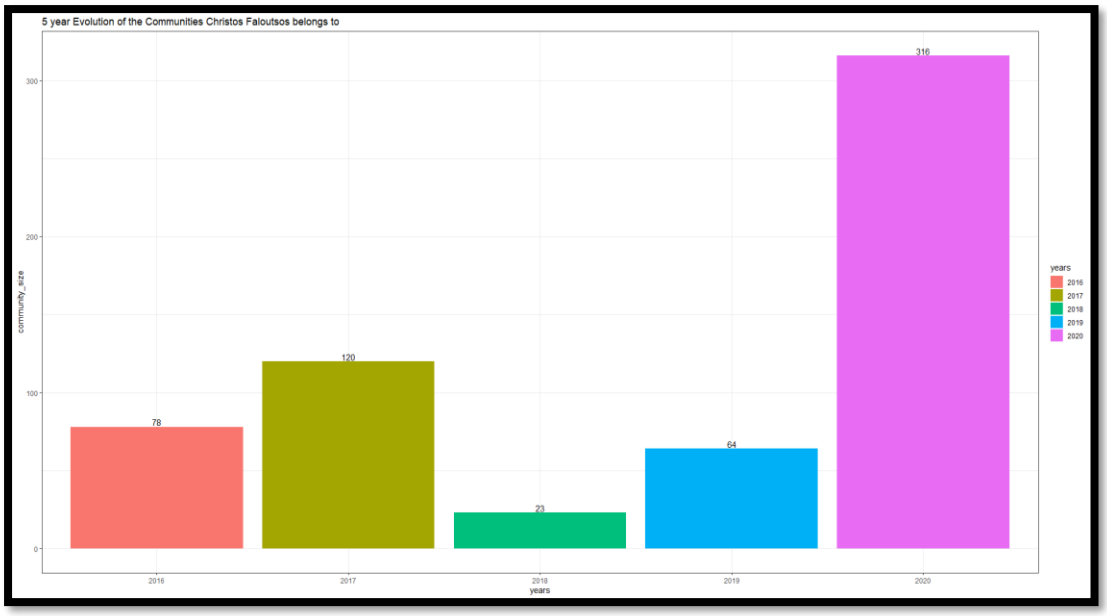


Figure 8 : Evolution of the Communities Size of the Christos Faloutsos

Our selected author in 2018 seems that had collaborated with the least number of authors compared to 2020 where he had collaborated with 316 other authors.

The next step is to find similarities between the specific communities. In order to find them we will print the number of the common authors between them in every year. The results for every year combination can be observed in the table below.

Years	Common Authors	Number of Common Authors
2016-2017	Bryan Hooi, Chengxi Zang, Christos Faloutsos, Kijung Shin, Peng Cui 0001, Shiqiang Yang, Tianyang Zhang, Wenwu Zhu 0001, Meng Jiang 0001, Yu Zheng 0004	10
2017-2018	Bryan Hooi , Christos Faloutsos	2
2018-2019	Bryan Hooi, Christos Faloutsos, Dongha Lee, Hwanjo Yu	4
2019-2020	Andrey Kan, Aston Zhang, Carl Yang, Chanyoung Park ,Christos Faloutsos, Dongha Lee, Donghyun Kim, Huajie Shao, Hwanjo Yu , Jiaming Shen, Jiawei Han 0001, Jingbo Shang, Junyoung Hwang, Namyong Park, Qi Zhu 0008, Shengzhong Liu, Shuochao Yao, Tarek F. Abdelzaher, Tianshi Wang, Tong Zhao, Xin Luna Dong, Zecheng Zhang, Zongyi Wang, SeongKu Kang, Liyuan Liu	25

We can say that in the period 2019 – 2020 we had the most collaborations of the authors and we had fewer collaborations in 2017-2018 period.

Finally, we need to visualize the graphs using different color for each community . In order to plot a meaningful result, we will filter out all nodes that belong to very small or very large communities. We have decided to plot only the nodes with size between 20 and 50.

Disclaimer

It is very important to mention that R by default uses only 8 colors to color the different communities so many of them that have the same color does not mean that are the same or that they have same structure.

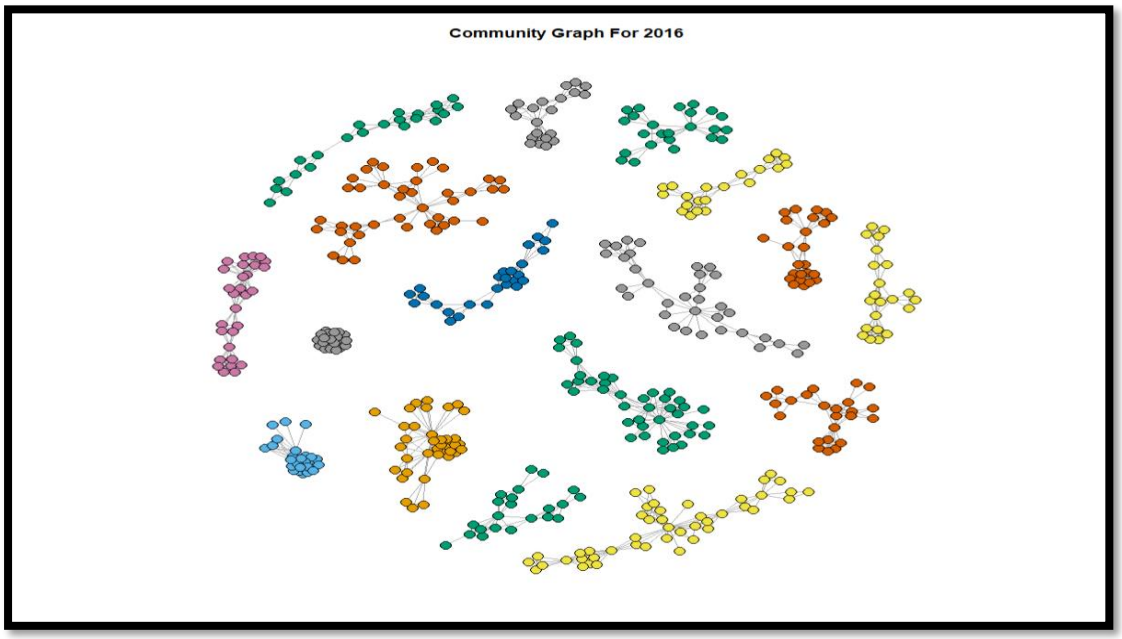


Figure 9 Community Graph for 2016

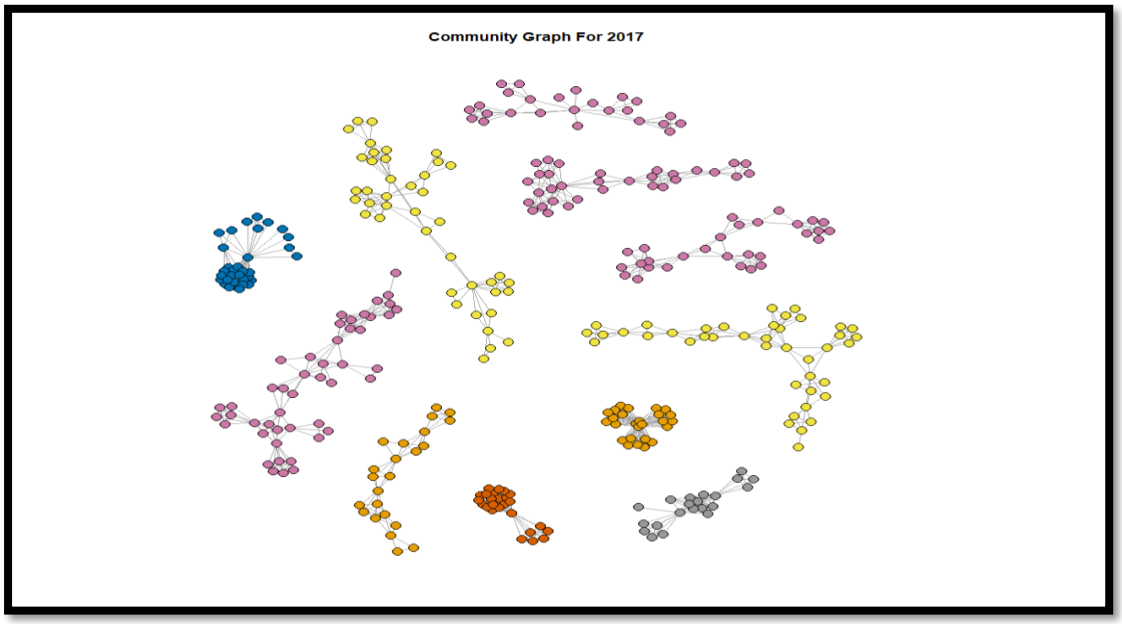


Figure 10 Community Graph for 2017

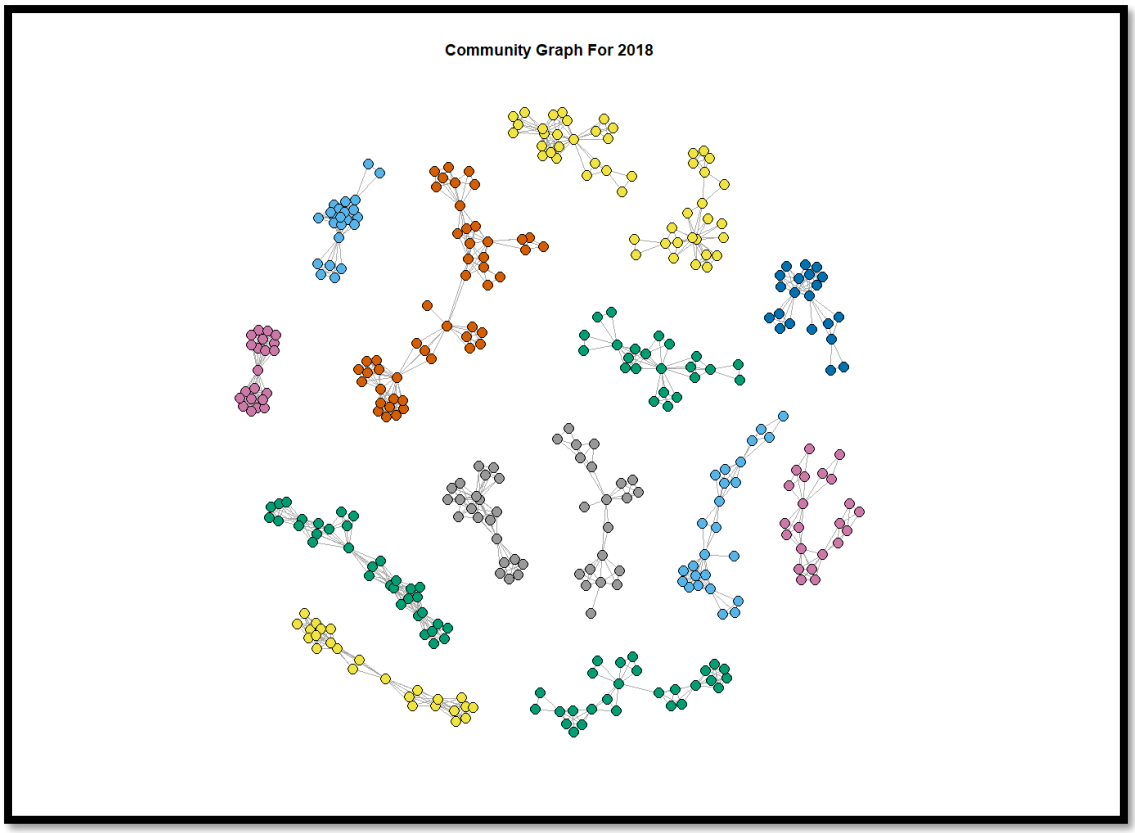


Figure 11 Community Graph for 2018

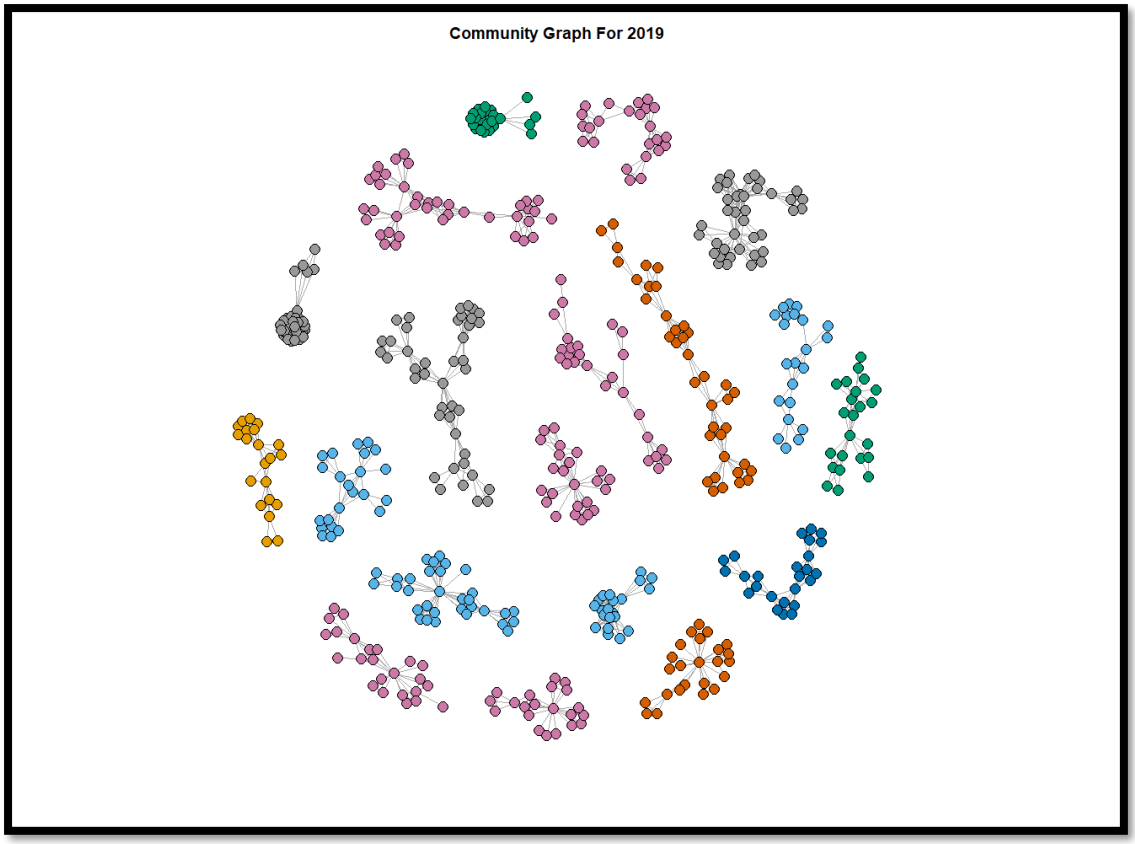


Figure 12 Community Graph for 2019

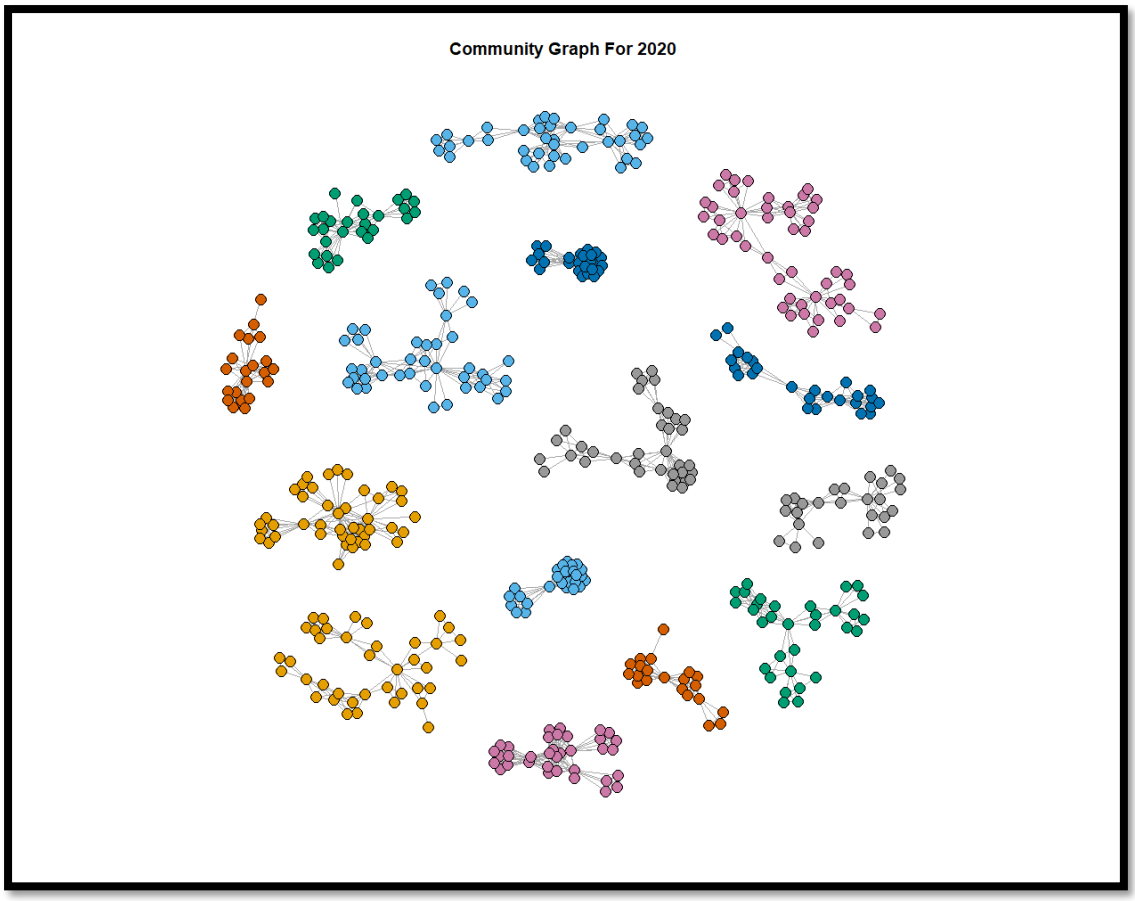


Figure 13 Community Graph for 2020