ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS

DEPARTMENT OF MANAGEMENT, SCIENCE & TECHNOLOGY

MSc BUSINESS ANALYTICS

“Assignment for Course: Social Network Analysis”

Full Name: STAMATIOS SIDERIS

Register Number: f2822113

Athens,2022

Contents

Introduction

The scope of the analysis is to create a graph database based on data that are available here: https://hive.di.uoa.gr/network-analysis/files/authors.csv.gz and analyze the relationships between authors and the papers they collaborated in.

Preparation of Dataset

Firstly, we clean the dataset using R in file “cleaning_data_code.R”. As the csv is very large, we use code that loads and reads the file per row. We keep rows that refer to the events CIKM, KDD, ICWSM, WWW, IEEE BigData performed in years 2016 to 2020. We also exclude papers that include only 1 author as the scope of the assignment is to inspect the relationships between different authors. Finally, we create 5 new csv files, one for each year and store them to folder “Cleaned data”.

Moreover, we need to convert the csv files to a form for/to/weight where the weight is the number of papers 2 authors have co-authored. To do so, we imported the files to python, see the Jupiter notebook “from_to_weight.ipynb”. Using the function included in the code, we are able to convert the csv files to appropriate format and create 5 new csv which could be found in folder “From_To_Weight_data”.

5-year evolution of metrics

We import the final csv files to R and create the graphs for each year using the library igraph.

Firstly, we calculate the vertices of each year. The number of vertices increases each year, see figure 1, which shows that more authors co-author every year. The fluctuation becomes more intense from year 2018 to year 2019.

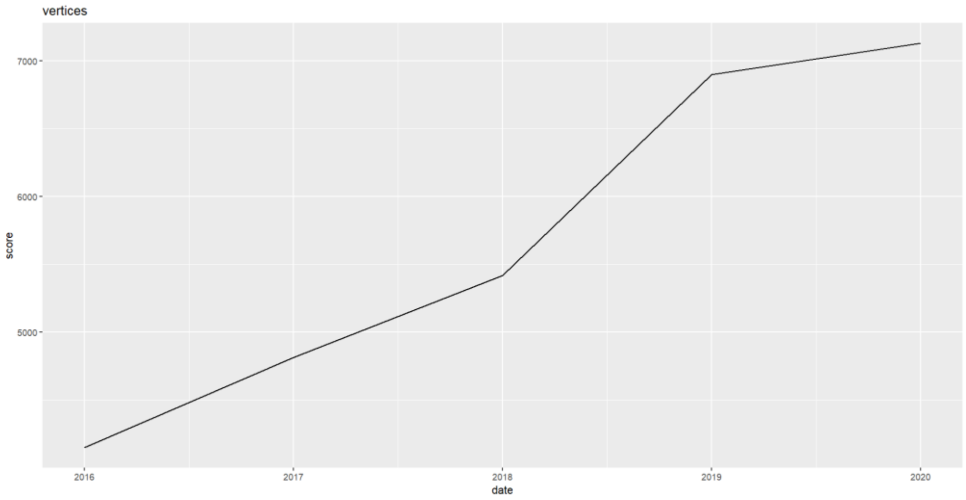


Figure 1 Number of vertices per year

Secondly, we calculate the number of edges. The number of edges increases each year, see figure 2, as more authors write papers and so the collaborations between them increase. The fluctuation becomes more intense from year 2018 to year 2019.

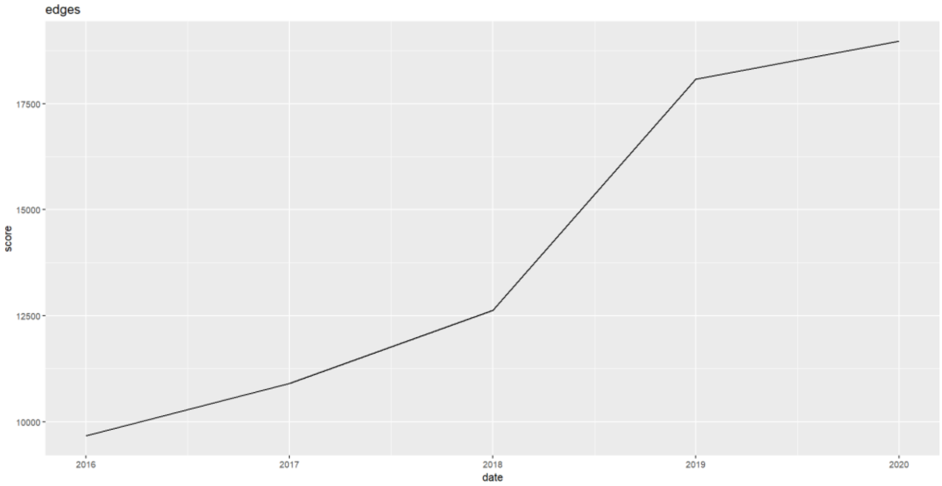


Figure 2 Number of edges per year

Thirdly, we calculate the diameters. Year 2017 has the shortest distance between the two most distant nodes while year 2018 the longest, see figure 3. There is significant fluctuation in all years.

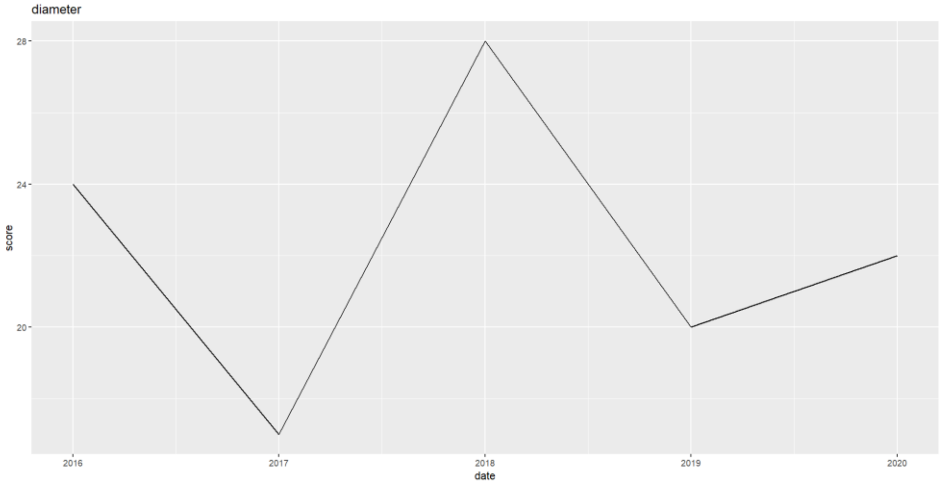


Figure 3 Diameter per year

Finally, we calculate the average degree. Year 2020 has the highest average number of edges per node while year 2017 the lowest, see figure 4. Although a decrease is observed from year 2016 to year 2017, a steady increase is observed in the next years leading to new highs.

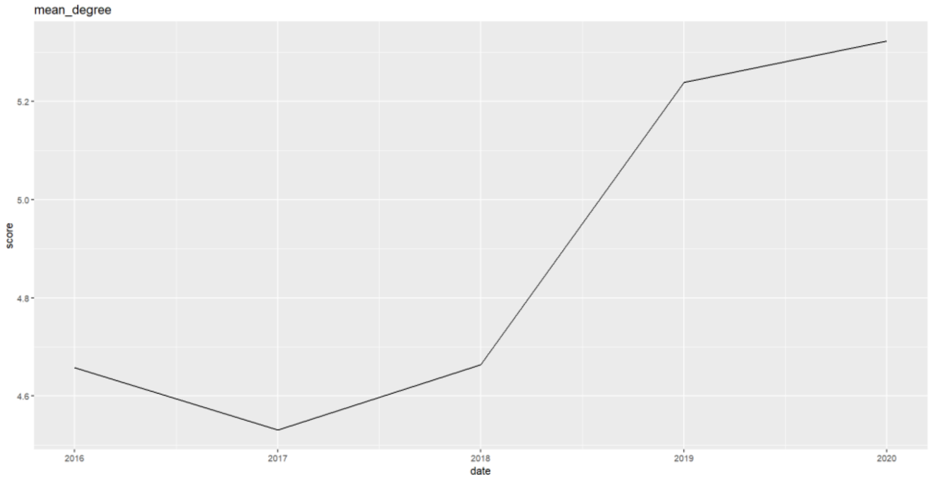


Figure 4 Average Degree per year

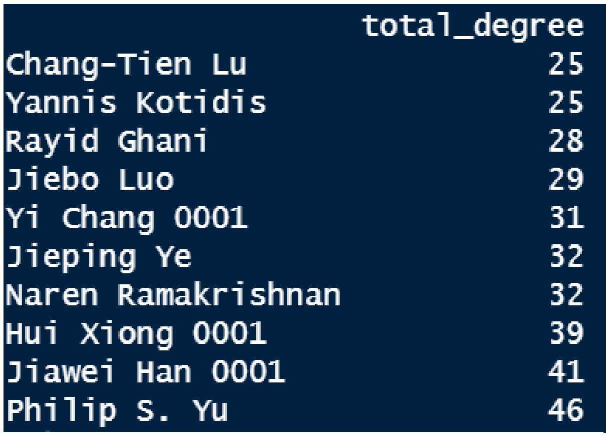
Important Nodes

The scope is to find the 10 most important nodes based on degree and page rank.

Based on degree, the highest degree is observed in 2018 equal to 70 and the lowest in 2016 and 2018 equal to 25. The top-10 changes a lot each year, but the top 3 authors mostly remain the same with “Hui Xiong 0001”, “Jiawei Han 0001” and “Philip S. Yu” to participate every year.

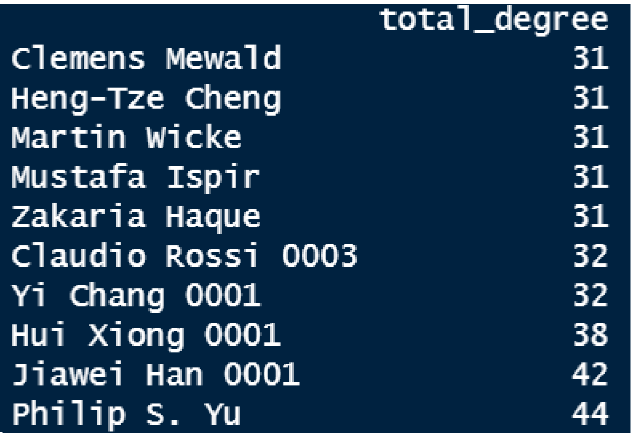
Based on page rank, the highest page rank is observed in 2018 equal to 0.0019 and the lowest in 2018 again equal to 0.00061. The top-10 changes a lot each year, but the top 3 authors mostly remain the same with “Hui Xiong 0001”, “Jiawei Han 0001” and “Philip S. Yu” to participate almost every year.

Based on degree the most important nodes are:



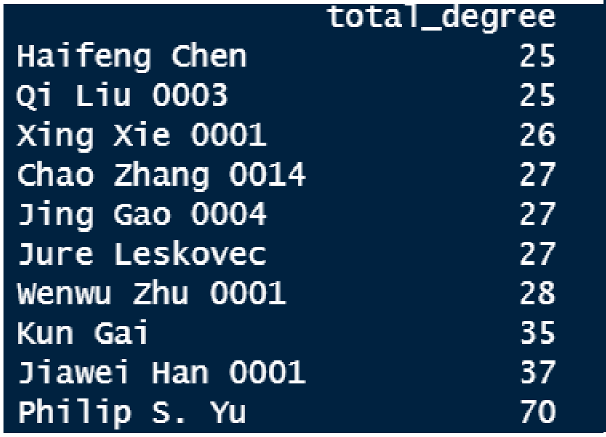
```
                    total_degree
Chang-Tien Lu                 25
Yannis Kotidis                25
Rayid Ghani                   28
Jiebo Luo                     29
Yi Chang 0001                 31
Jieping Ye                    32
Naren Ramakrishnan            32
Hui Xiong 0001                39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

Figure 5 Top-10 Nodes by degree in 2016



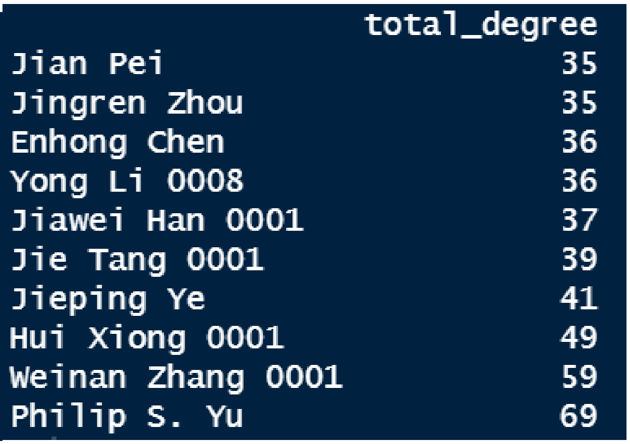
```
                    total_degree
Clemens Mewald                31
Heng-Tze Cheng                31
Martin Wicke                  31
Mustafa Ispir                 31
Zakaria Haque                 31
Claudio Rossi 0003            32
Yi Chang 0001                 32
Hui Xiong 0001                38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

Figure 6 Top-10 Nodes by degree in 2017



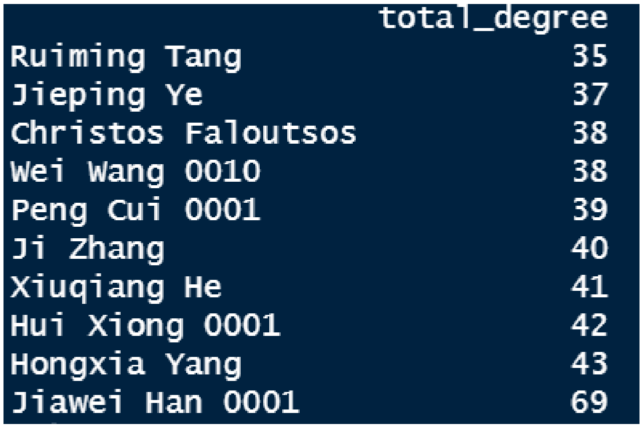
```
                    total_degree
Haifeng Chen                  25
Qi Liu 0003                   25
Xing Xie 0001                 26
Chao Zhang 0014               27
Jing Gao 0004                 27
Jure Leskovec                 27
Wenwu Zhu 0001                28
Kun Gai                       35
Jiawei Han 0001               37
Philip S. Yu                  70
```

Figure 7 Top-10 Nodes by degree in 2018



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

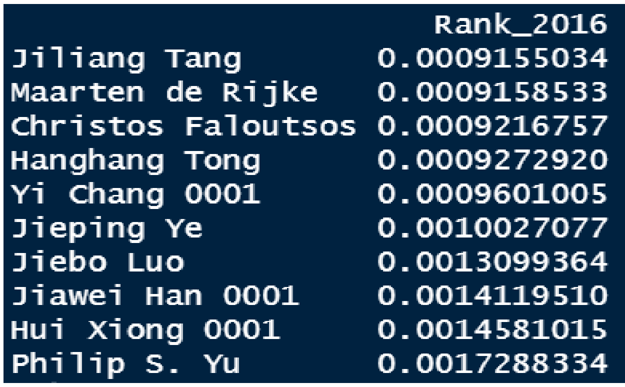
Figure 8 Top-10 Nodes by degree in 2019



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

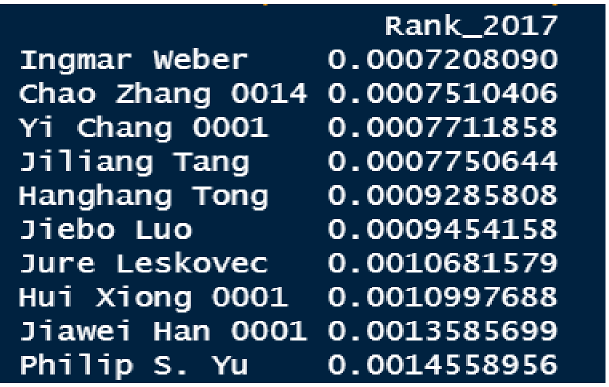
Figure 9 Top-10 Nodes by degree in 2020

Based on page rank the most important nodes are:



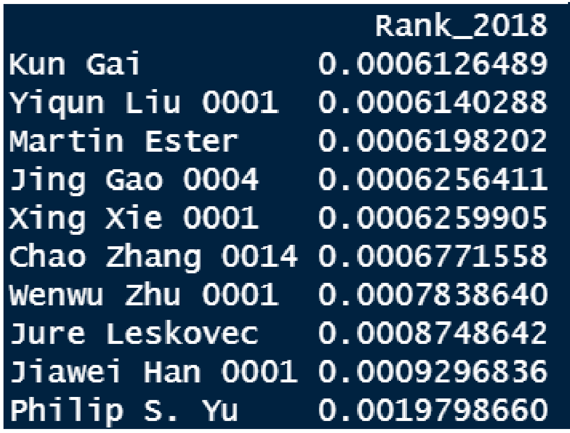
```
                      Rank_2016
Jiliang Tang       0.0009155034
Maarten de Rijke   0.0009158533
Christos Faloutsos 0.0009216757
Hanghang Tong      0.0009272920
Yi Chang 0001      0.0009601005
Jieping Ye         0.0010027077
Jiebo Luo          0.0013099364
Jiawei Han 0001    0.0014119510
Hui Xiong 0001     0.0014581015
Philip S. Yu       0.0017288334
```

Figure 10 Top-10 Nodes by page rank in 2016



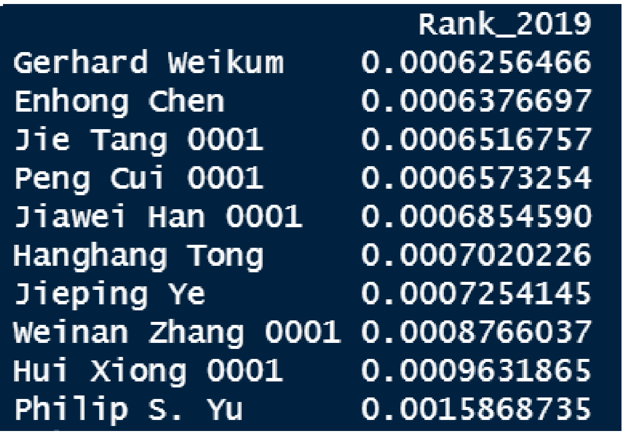
```
                   Rank_2017
Ingmar Weber    0.0007208090
Chao Zhang 0014 0.0007510406
Yi Chang 0001   0.0007711858
Jiliang Tang    0.0007750644
Hanghang Tong   0.0009285808
Jiebo Luo       0.0009454158
Jure Leskovec   0.0010681579
Hui Xiong 0001  0.0010997688
Jiawei Han 0001 0.0013585699
Philip S. Yu    0.0014558956
```

Figure 11 Top-10 Nodes by page rank in 2017



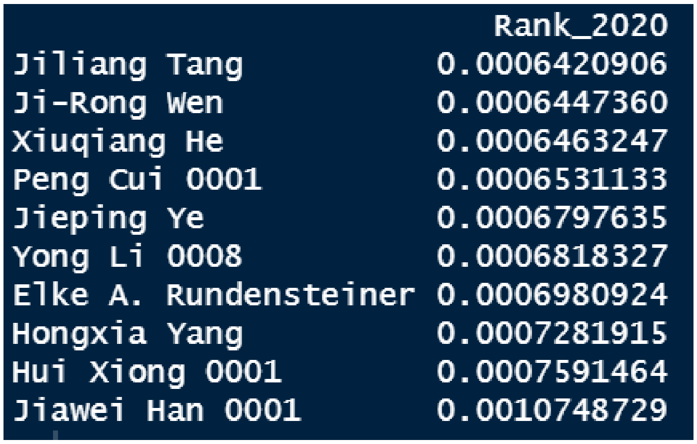
```
                      Rank_2018
Kun Gai            0.0006126489
Yiqun Liu 0001     0.0006140288
Martin Ester       0.0006198202
Jing Gao 0004      0.0006256411
Xing Xie 0001      0.0006259905
Chao Zhang 0014    0.0006771558
Wenwu Zhu 0001     0.0007838640
Jure Leskovec      0.0008748642
Jiawei Han 0001    0.0009296836
Philip S. Yu       0.0019798660
```

Figure 12 Top-10 Nodes by page rank in 2018



```
                        Rank_2019
Gerhard Weikum       0.0006256466
Enhong Chen          0.0006376697
Jie Tang 0001        0.0006516757
Peng Cui 0001        0.0006573254
Jiawei Han 0001      0.0006854590
Hanghang Tong        0.0007020226
Jieping Ye           0.0007254145
Weinan Zhang 0001    0.0008766037
Hui Xiong 0001       0.0009631865
Philip S. Yu         0.0015868735
```

Figure 13 Top-10 Nodes by page rank in 2019



```
                            Rank_2020
Jiliang Tang             0.0006420906
Ji-Rong Wen              0.0006447360
Xiuqiang He              0.0006463247
Peng Cui 0001            0.0006531133
Jieping Ye               0.0006797635
Yong Li 0008             0.0006818327
Elke A. Rundensteiner    0.0006980924
Hongxia Yang             0.0007281915
Hui Xiong 0001           0.0007591464
Jiawei Han 0001          0.0010748729
```

Figure 14 Top-10 Nodes by page rank in 2020

Communities

We use the algorithms greedy clustering, infomap clustering and Louvain clustering to detect the communities of each year. All the methods return results while infomap was a lot slower than the other 2 methods as it took 5.5 secs to run while the others 0.1 secs both.

We choose “Jiawei Han 0001” as an author that presents in all years and the Louvain method that we used before. The size of the community increases from 2016 to 2017 reaching 121 nodes, then drops in 2018 and 2019 to 86 and 69 Nodes respectively, to finally increase and reach a maximum for the 5 years equal to 124 Nodes. Comparing the similarity of nodes of each year to its forthcoming year, 2016 and 2017 have 14 similar nodes, 2017 and 2018 have 15 similar nodes, 2018 and 2019 have 25 similar nodes and 2019 and 2020 have 14 similar nodes.

Finally, we plot the communities recognized by the Louvain algorithm for each year and for a different color for each community in plots 15,16,17,18,19. For visualization reasons, we filter out communities with extreme sizes compared to the rest of the year.

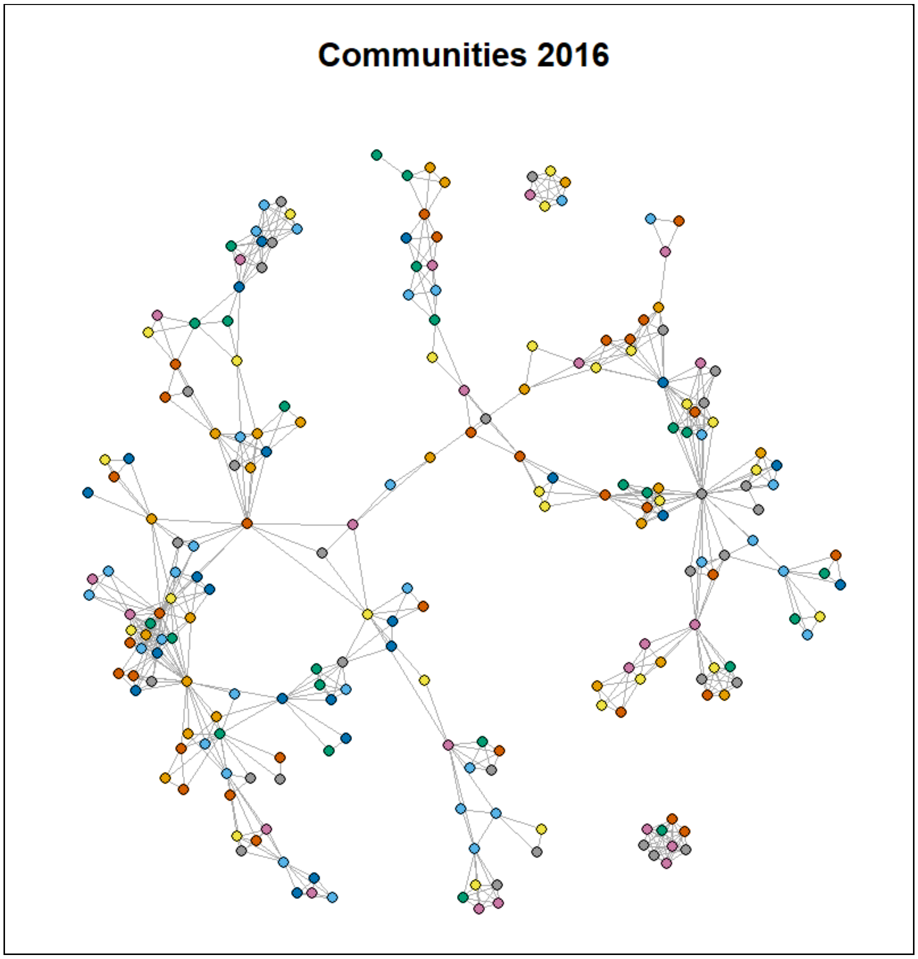


Figure 15 The communities of year 2016

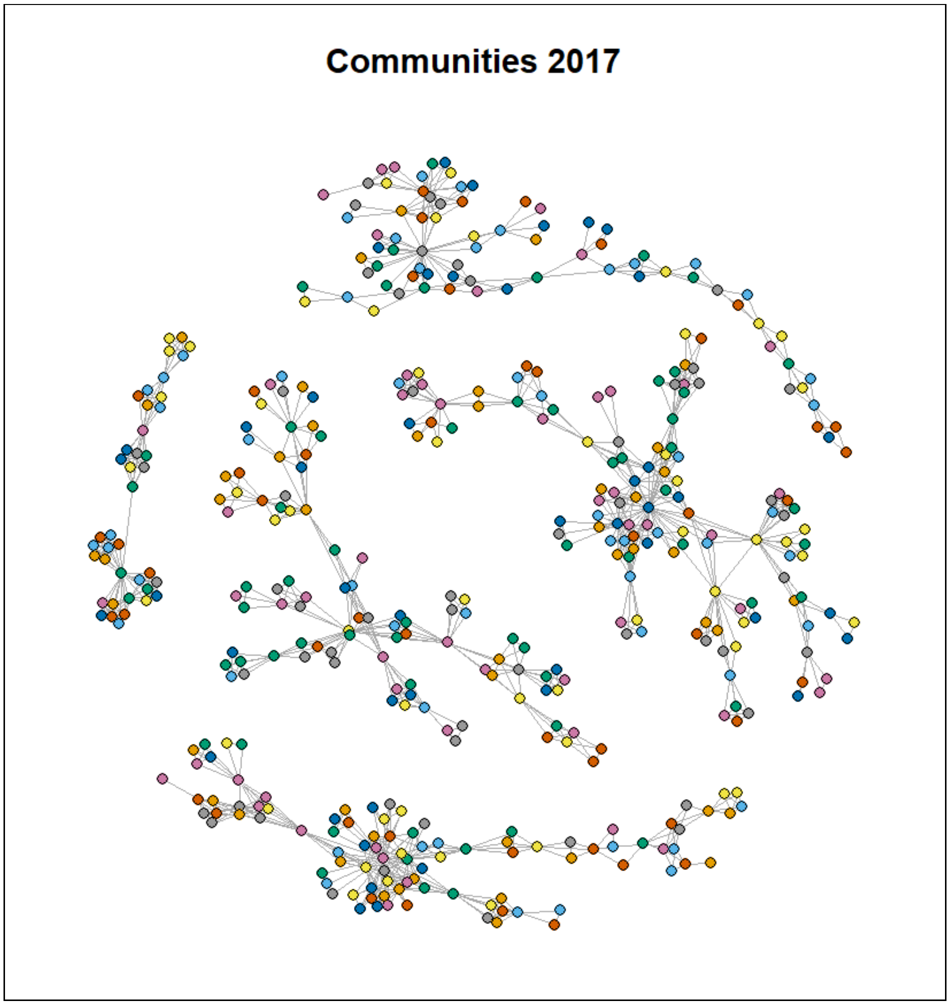


Figure 16 The communities of year 2017

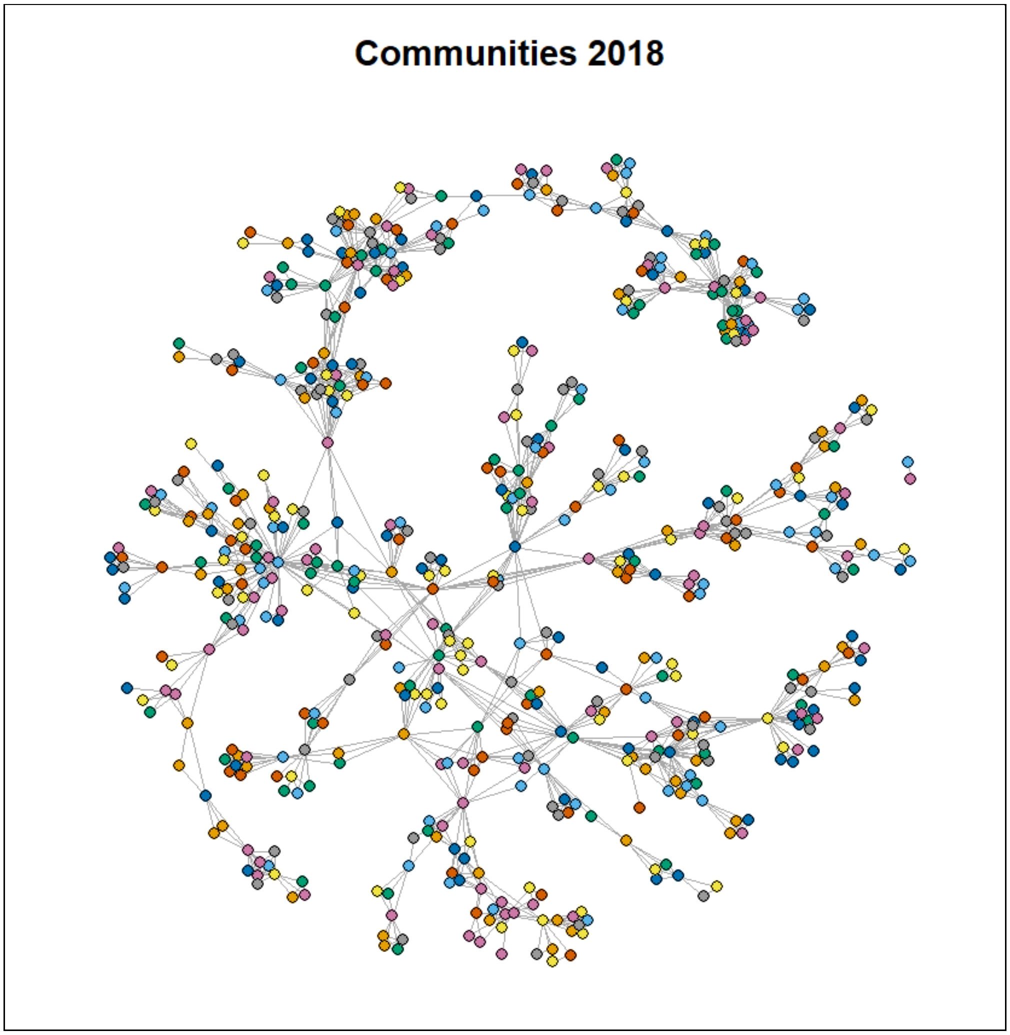


Figure 17 The communities of year 2018

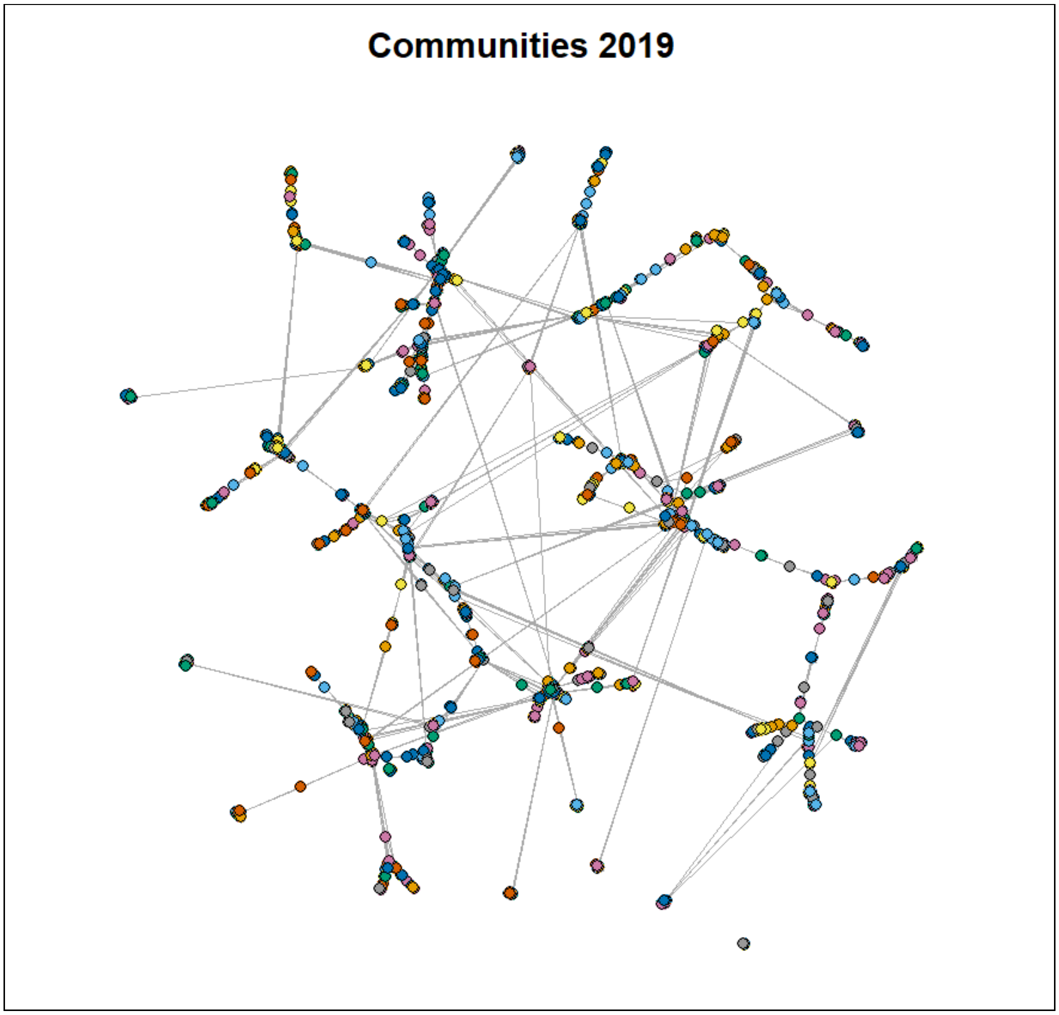


Figure 18 The communities of year 2019

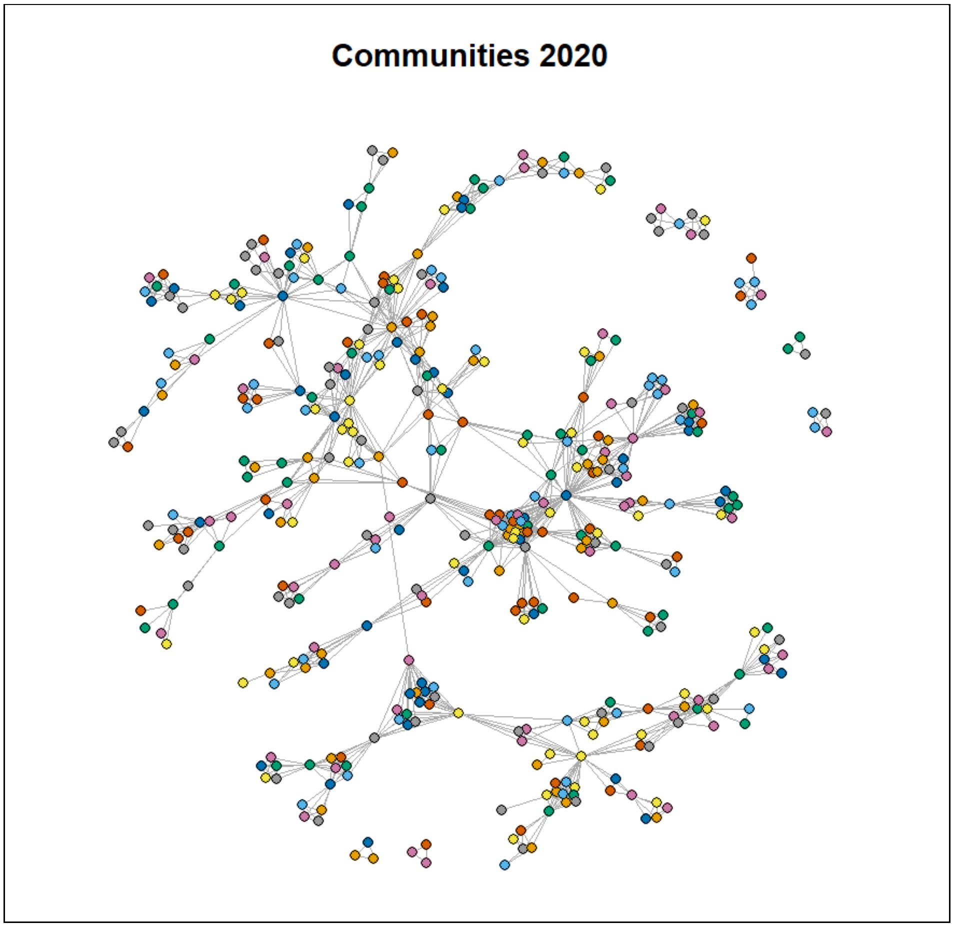


Figure 19 The communities of year 2020