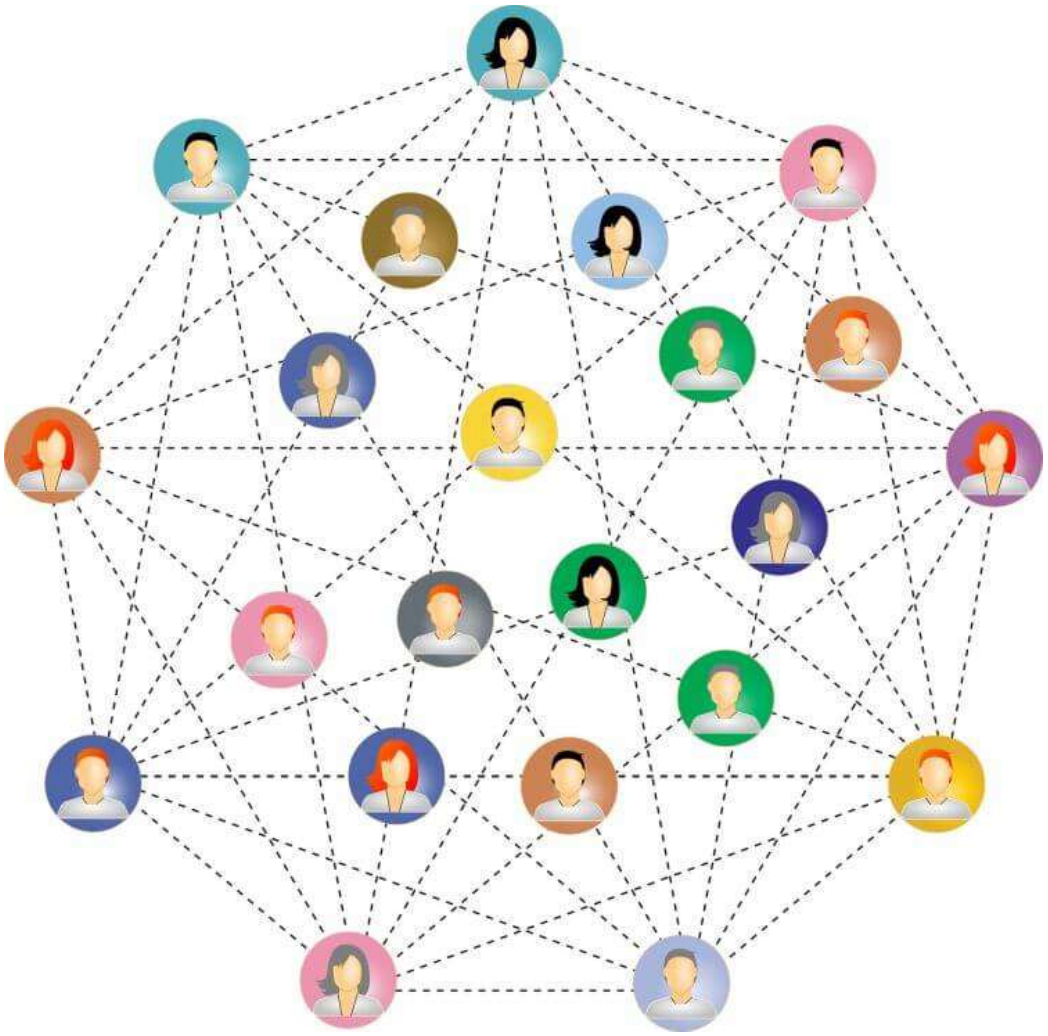


From raw data to temporal graph structure exploration

Georgia Vlassi p2822001 | Social Networking Analysis | 20/06/2021

1. DBLP co-authorship graph

For our analysis and implementations, we should download the dataset from the following link https://hive.di.uoa.gr/network-analysis/files/authors.csv.gz. The file contains records of authors, their articles, the year and place where they presented these articles and a list of the co-authors of each article.

Initially, by using the Unix commands, we find the records of articles that presented in the following conferences CIKM, KDD, ICWSM, WWW and IEEE BigData and we load them in a new comma delimited file. We use the new file to separate it in 5 sub files depending on the year the conference took place. As there were not the conferences listed above for year 2021, the 5 subfiles contains records for years 2016 to 2020 respectively.

The 5 new files are imported to Jupyter and we use Python language for the transformations. Reading the data, we summarize the NA values at the file of each year if any and we remove these values. Having cleaned the datasets, we convert them in list and we split their values by comma. In this way, we have a sorted list of lists for each year, which is easier to read the authors in pairs. By using the libraries ` collections` and `itertools` we count how many times a pair of authors exists in the file and we keep the counter, which will be used later as weight. Subsequently, the above files containing two columns, the pair of authors and a counter, are separated further, so as to have the three requested columns. The pair of authors are separated to From and To columns and the counter to Weight column. The final 5 files with the requested structure are exported to comma delimited files.

The above datasets are used as input in R Studio. The first column of each file, named `X` can be removed by setting it to NULL, as it indicates an index. We create an igraph graph from each data frame. The function used to create the igraph contains two parameters. The first one is the data frame, which indicates the edges and the vertices of the graph. The second parameter is a logical scalar, which indicates that our graph is undirected. Finally, the final column of each data frame named `Weight` is set as an edge attribute for the graph. Having created 5 weighted undirected graph, we can start our analysis.

2. Average degree over time

Having created our igraph graphs for the years 2016 to 2020, we can examine the 5-year evolution of different metrics. Observing the Figure 1 and Figure 2 there is an exponential growth in the number of vertices and edges from 2016 to 2020. Especially, in 2020 we have the maximum number of vertices and edges. This indicates that in this network we have the most authors and the greater number of relationships among them.

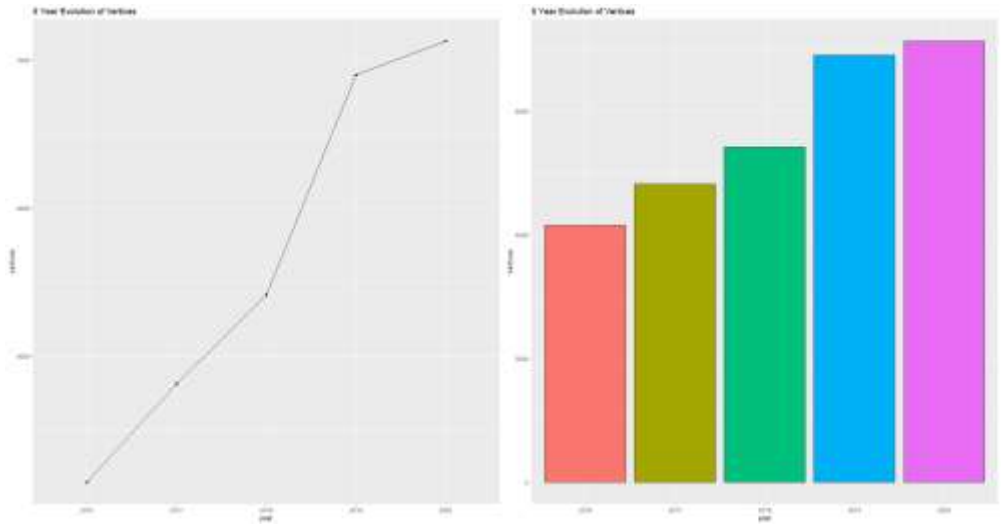


Figure 1 The 5 Year Evolution of Vertices

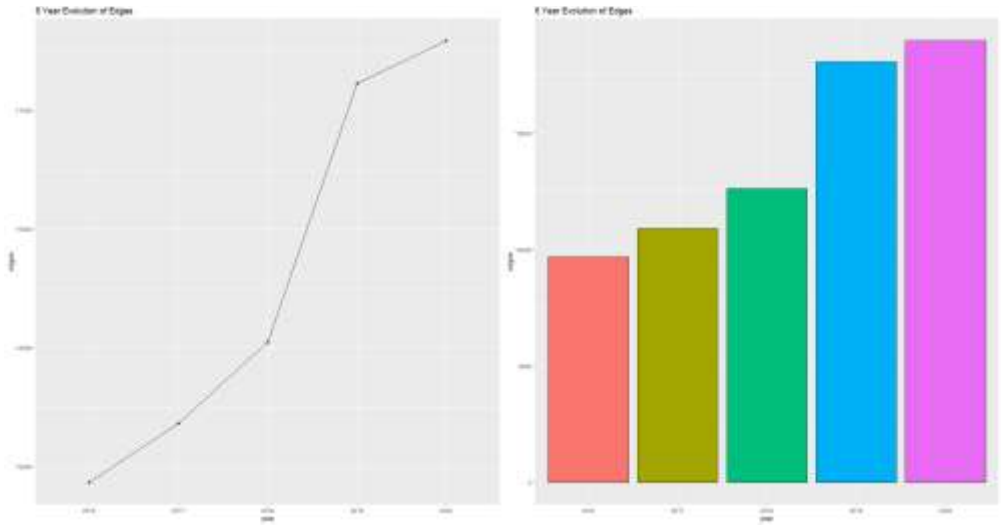


Figure 2 The 5 Year Evolution of Edges

Regarding the Figure 3, we observe that there is a significant fluctuation. This attributed to the fact that in years 2017 and 2019 the largest distance which must be traversed in order to travel from one author to another is smaller than the other years. Concerning the Figure 4, the small fluctuation in years 2016 and 2017 indicates that the connections among authors have been decreased in 2017. Moreover, we can assume that fewer authors cooperated in the papers that have published.

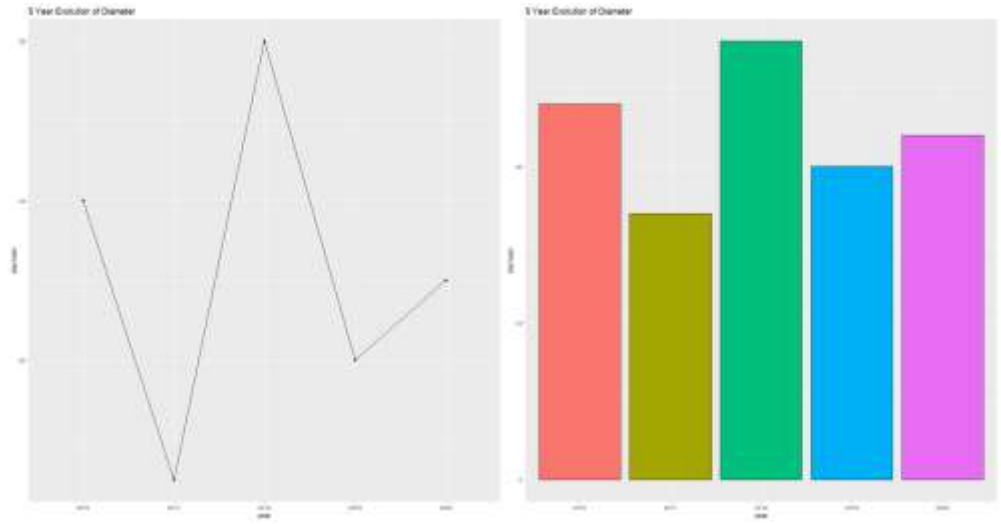


Figure 3 The 5 Year Evolution of Diameter

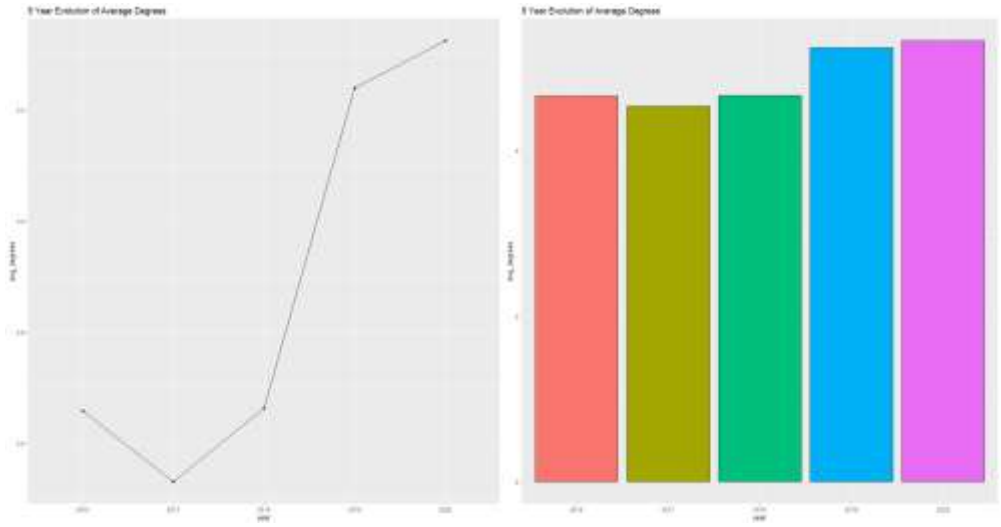


Figure 4 The 5 Year Evolution of Average Degrees

3. Important nodes

In the below tables are presented the top 10 authors based on the degree. In R Studio the outcomes are data frames. To calculate these authors, the degree is used to detect the connections that every author has through the network. Subsequently, we sort them in descending order and we print the first 10.

The top 10 authors based on Degree for year 2016:

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

The top 10 authors based on Degree for year 2017:

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

The top 10 authors based on Degree for year 2018:

Author	Degree
Philip S. Yu	70
Jiawei Han 0001	37
Kun Gai	35
Wenwu Zhu 0001	28
Jing Gao 0004	27

Chao Zhang 0014	27
Jure Leskovec	27
Xing Xie 0001	26
Enhong Chen	25
Qi Liu 0003	25

The top 10 authors based on Degree for year 2019:

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

The top 10 authors based on Degree for year 2020:

Author	Degree
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
Ruiming Tang	35

Observing the above tables, we conclude that only the 30% of authors are the same in the list of top 10 authors for the 5 years that are examined. These authors are Philip S. Yu, Jiawei Han 0001 and Hui Xiong 0001, who appeared in the 4 out of 5 years. Moreover, author Philip S. Yu appears consistently in the first place, followed by author Jiawei Han 0001 in the second place, apart from the year 2020, where he is in the first place.

In the second part of the 5-year evolution of the top 10 authors, our scope is to rank them based on the PageRank. The miniCRAN and magrittr packages should be loaded. The

first one is used to build a graph of package dependencies and the latter includes the %>% pipes, which are useful to sort the ranking in descending order and to convert it to a matrix.

The top 10 authors based on PageRank for year 2016:

Author	PageRank
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

The top 10 authors based on PageRank for year 2017:

Author	PageRank
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

The top 10 authors based on deg PageRank for year 2018:

Author	PageRank
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

The top 10 authors based on PageRank for year 2019:

Author	PageRank
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

The top 10 authors based on PageRank for year 2020:

Author	PageRank
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

Same with the first part, where the top 10 authors of each year where calculated based on the degree, so in the second part where the PageRank is concerned, the authors Philip S. Yu, Hui Xiong 0001 and Jiawei Han 0001 appear steadily in the top 10 authors and almost at the same top 3 places. The remaining places are filled with different authors than the ones of based on the lists of degree. We conclude that they are important and valuable authors with a high influence at the other authors and because of them there are lots of connections.

4. Communities

In the final task we should perform community detection on the five mention graphs. Initially, we try to apply fast greedy clustering, infomap clustering and louvain clustering on the 5 undirected co-authorship graphs. Although, all the methods completed successfully, the fast greedy clustering and infomap clustering are more inefficient, because the computing time needed for the creation of the graph communities is extremely high. Specifically, the time that infomap clustering need to complete is 11 times larger than the other two methods. As a result, we choose the louvain clustering for further investigation, which calculates the communities rapidly.

From the intersection of the louvain clustering applied for each year, we compute all the common authors at these communities. The author named ` Shaoping Ma` randomly selected to examine the 5 year evolution of these communities. The criterion which will be examined is to find if there are differences in the number of vertices. Observing the Figure 5, we conclude that the number of vertices of each community, where the author belongs, is not fixed, because there are fluctuations. More specific, in years 2017 and 2019 the number of vertices is downgraded, which is also confirmed by the calculation of diameter in the second question.

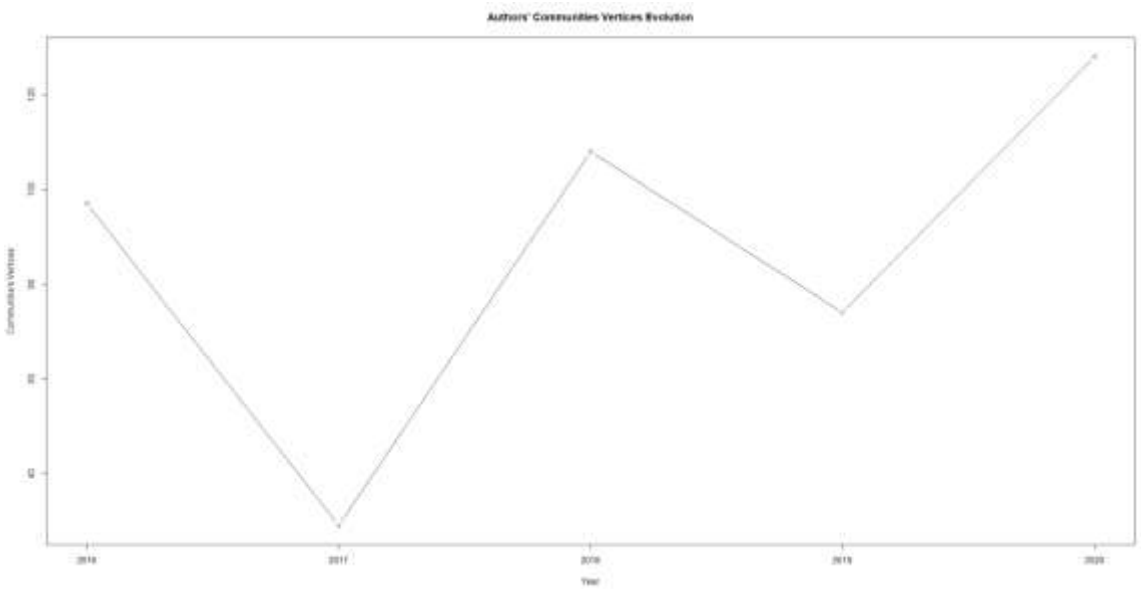
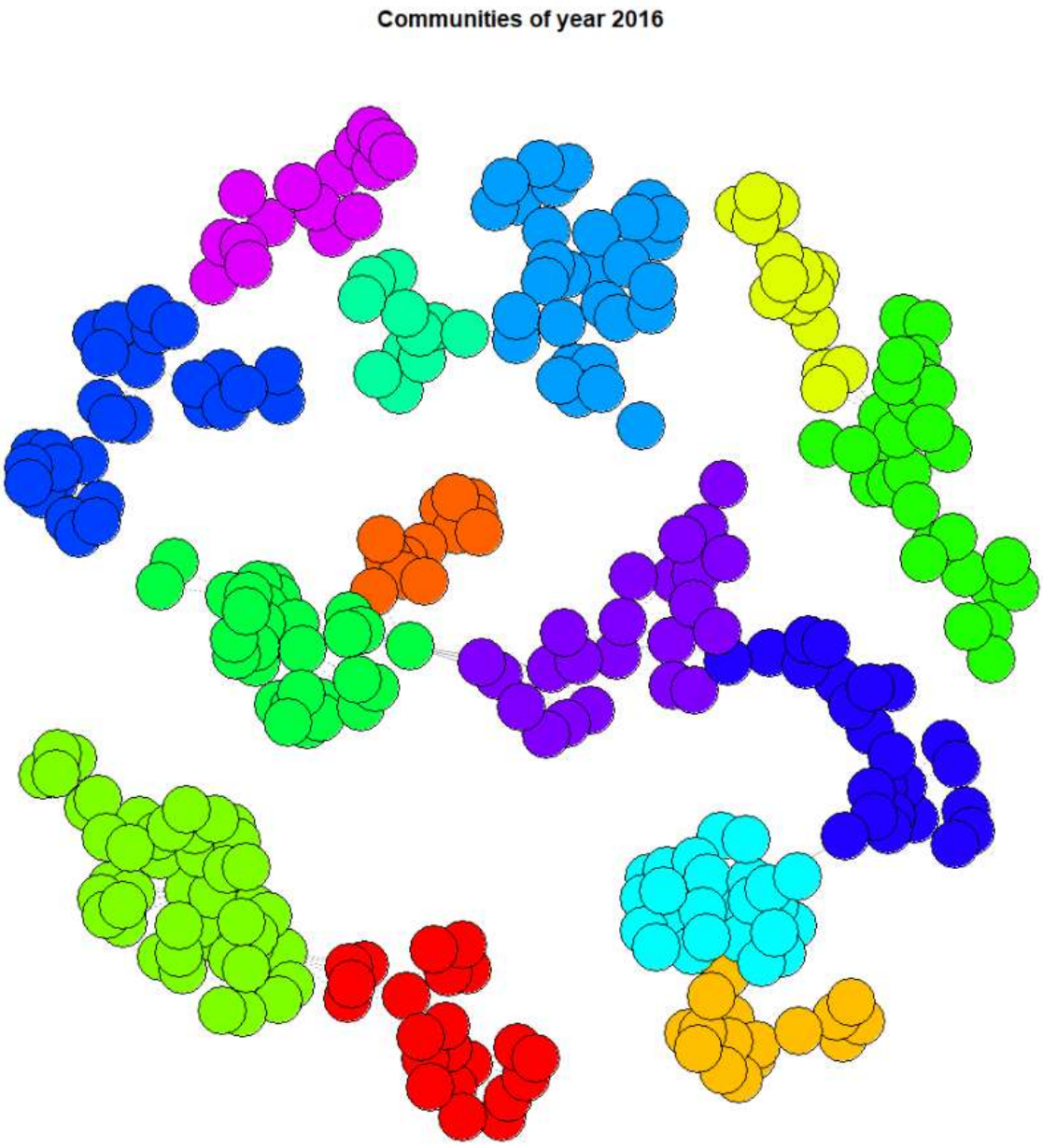
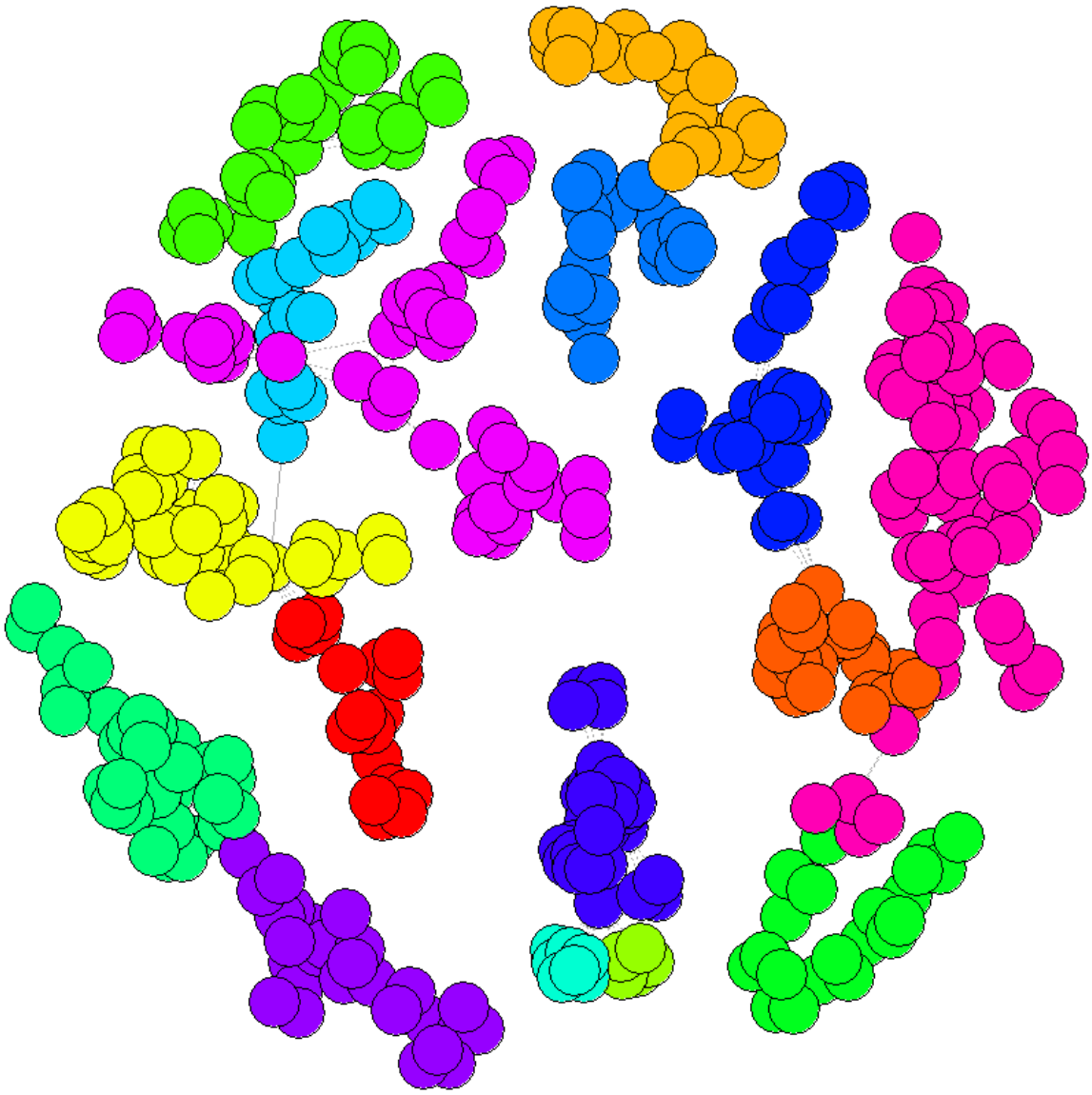


Figure 5 Authors' Communities Vertices Evolution

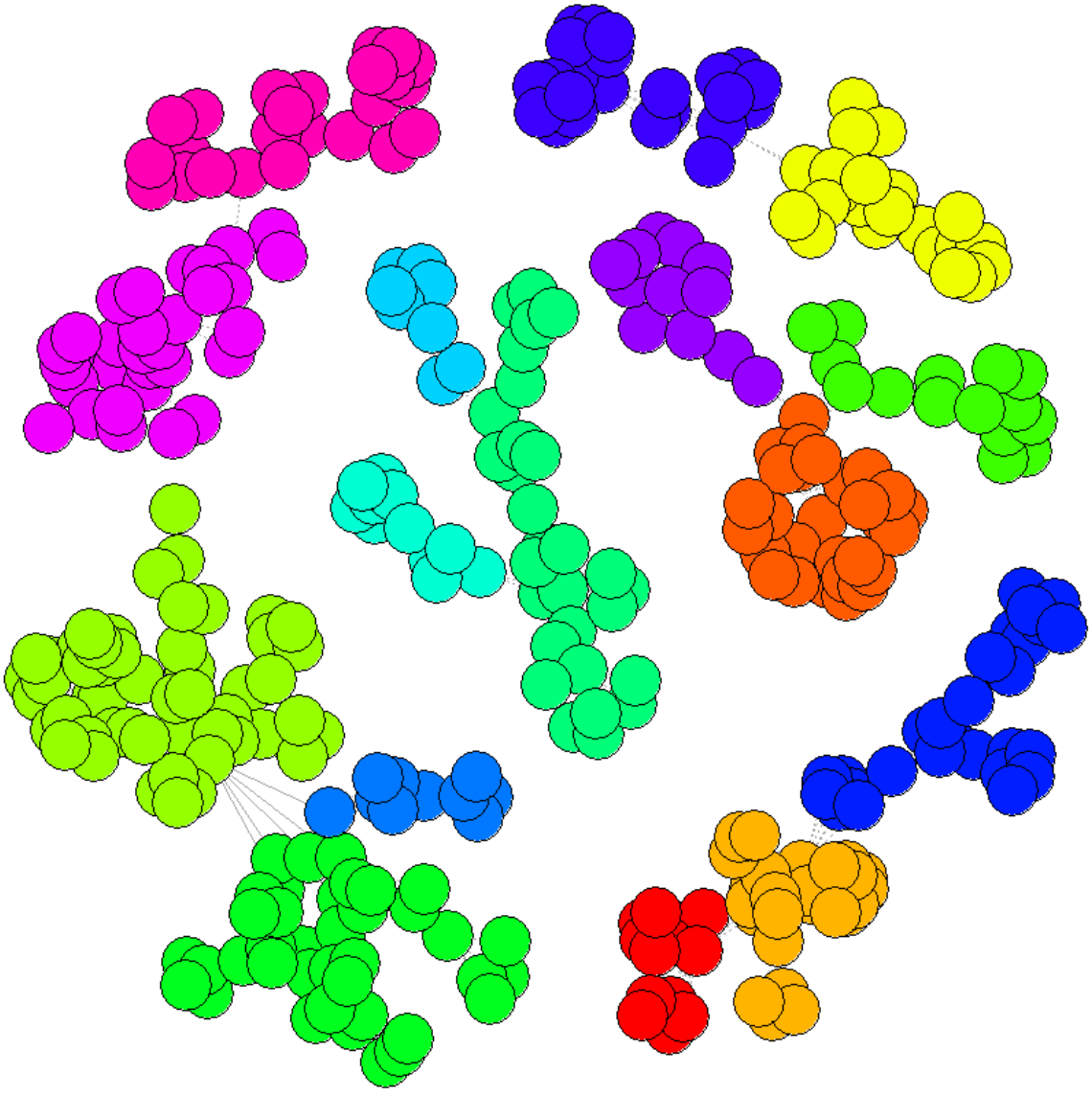
Having completed the analysis of the communities, we can visualize them. In order to avoid densely plots, only the communities with size greater than 40 and less than 80 are kept for the following visualizations.



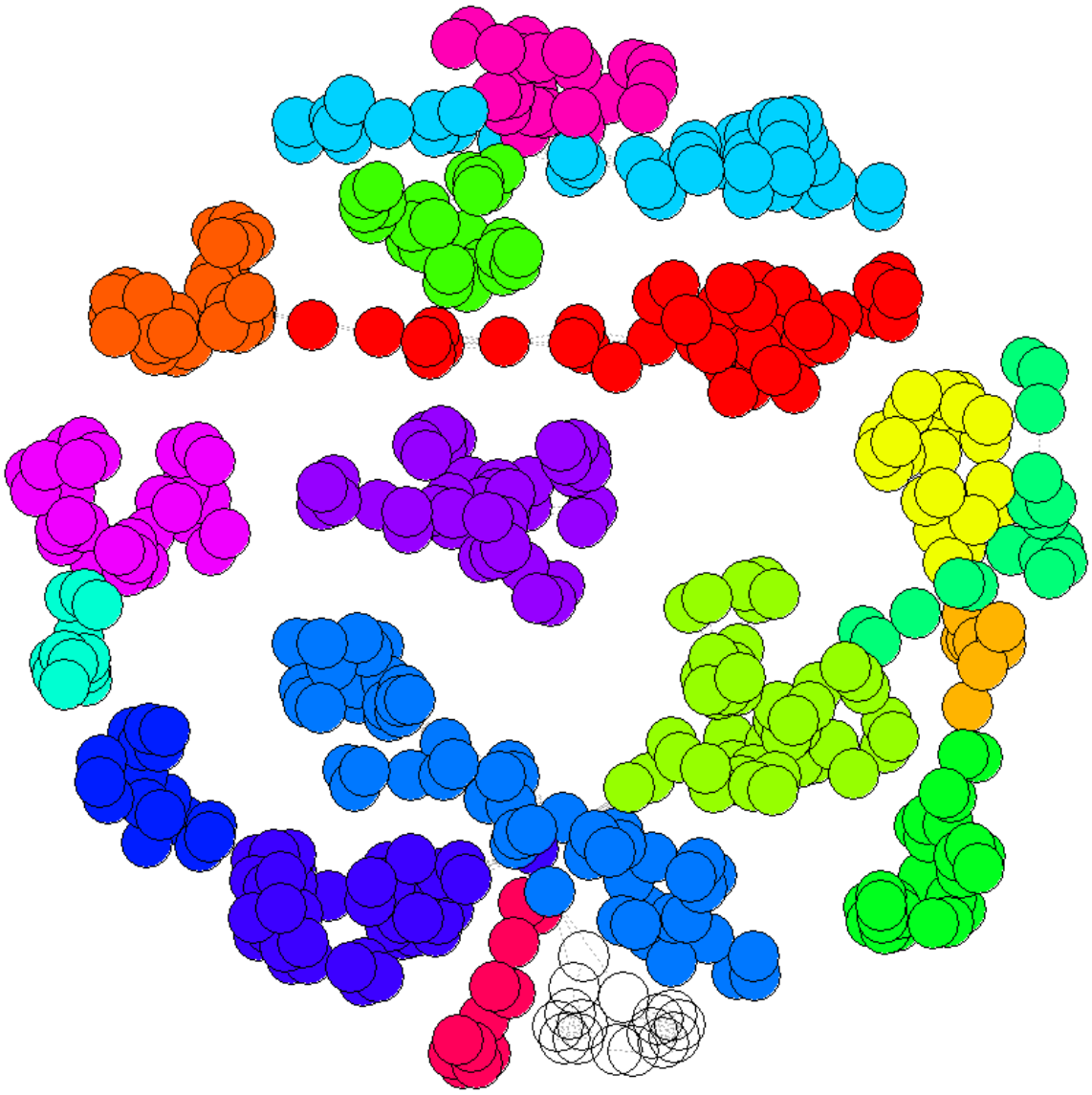
Communities of year 2017

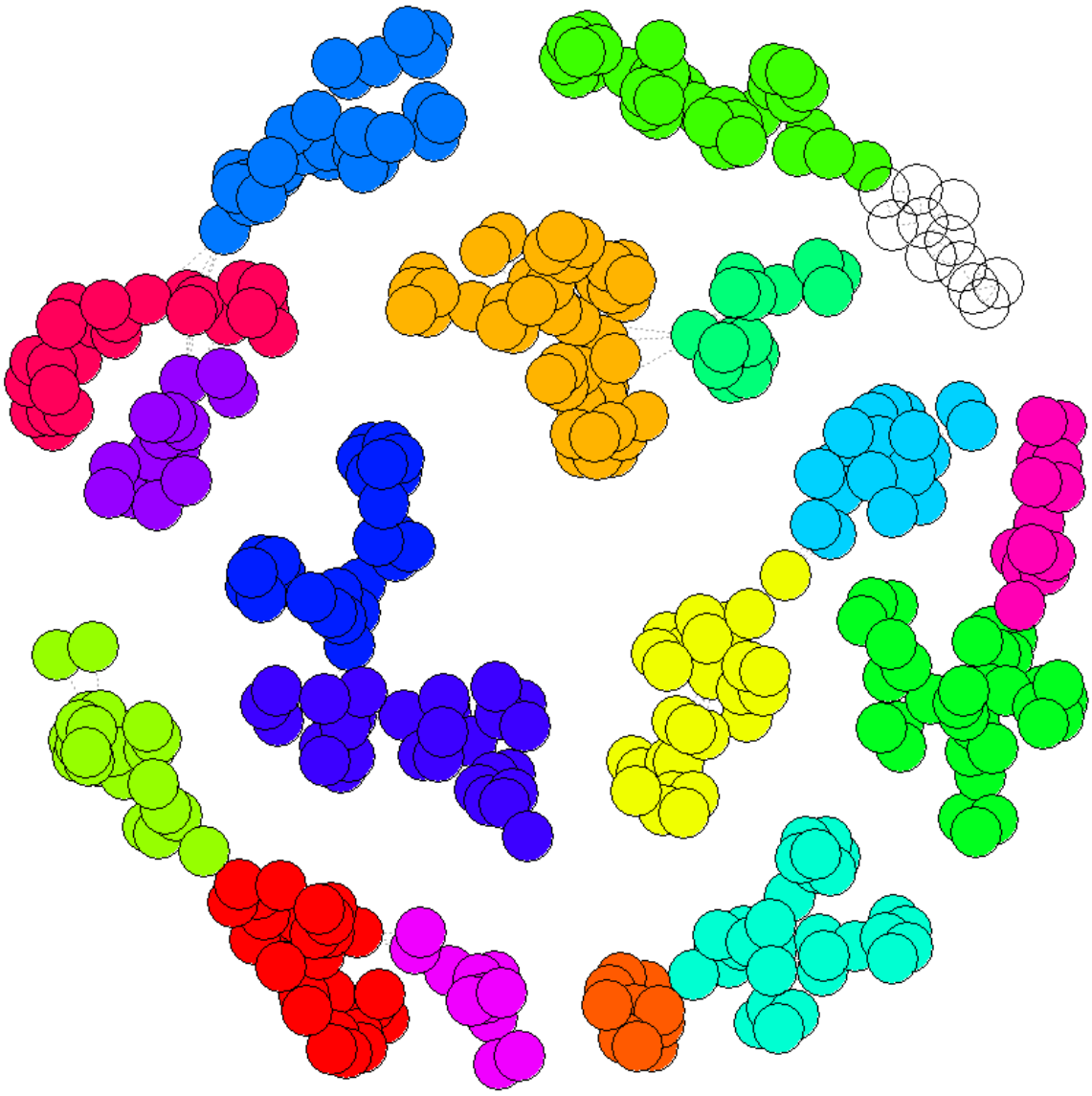


Communities of year 2018



Communities of year 2019




Communities of year 2020