Project 2: From raw data to temporal graph structure exploration

Name: Panagiota Gkourioti

Student ID: P2822109

Course: Social Network Analysis – Spring 2022

Instructor: Katia Papakonstantinopoulou

Athens University of Economics and Business

M.Sc. Program in Business Analytics

Contents

1. DBLP co-authorship graph

DBLP (Digital Bibliography & Library Project) has a large collection of journal articles with its meta data, papers published in various national and international conferences, and other online publications in the field of computer science.

The original compressed file "authors.csv.gz" consists of 2.793.928 conference proceedings records and contains information about the **year** the paper was published, the **authors**, the **title** of the paper, and the **conference** where the paper was presented.

The purpose is to construct a co-authorship network where two authors are connected if they publish at least one paper together. We initially manipulate the given data and create 5 csv files, describing the weighted undirected co-authorship graph for the years 2016 to 2020. The weight depends on the frequency of co-authorship.

With the use of Python programming language, we firstly filter out all records that are not related to the conferences "CIKM", "KDD", "ICWSM", "WWW" and "IEEE BigData" or are older than 5 years. We then drop a record with missing values and end up with 8720 records.

The next step is to create a dictionary for each year containing the pairs of the authors who wrote papers together and the frequencies of each pair. We then convert them into data frames with columns "From", "To" and "Weight", representing the relations of the authors and their weight.

Finally, we export the data frames to csv files with 9666, 10908, 12622, 18071 and 18966 rows describing the weighted undirected co-authorship graph for the respective year between 2016 and 2020. Having created the .csv files, we will import them into R-Studio and use them to create the respective igraph graphs.

2. Average degree over time

After loading the csv files into R-Studio, we attempt to visualize the 5-year evolution of different metrics for the graph. Specifically, we create plots for the following metrics, which are displayed in Figure 2.1.:

- ✓ *Number of vertices*
- ✓ *Number of edges*
- ✓ *Diameter of the graph*
- ✓ *Average degree*

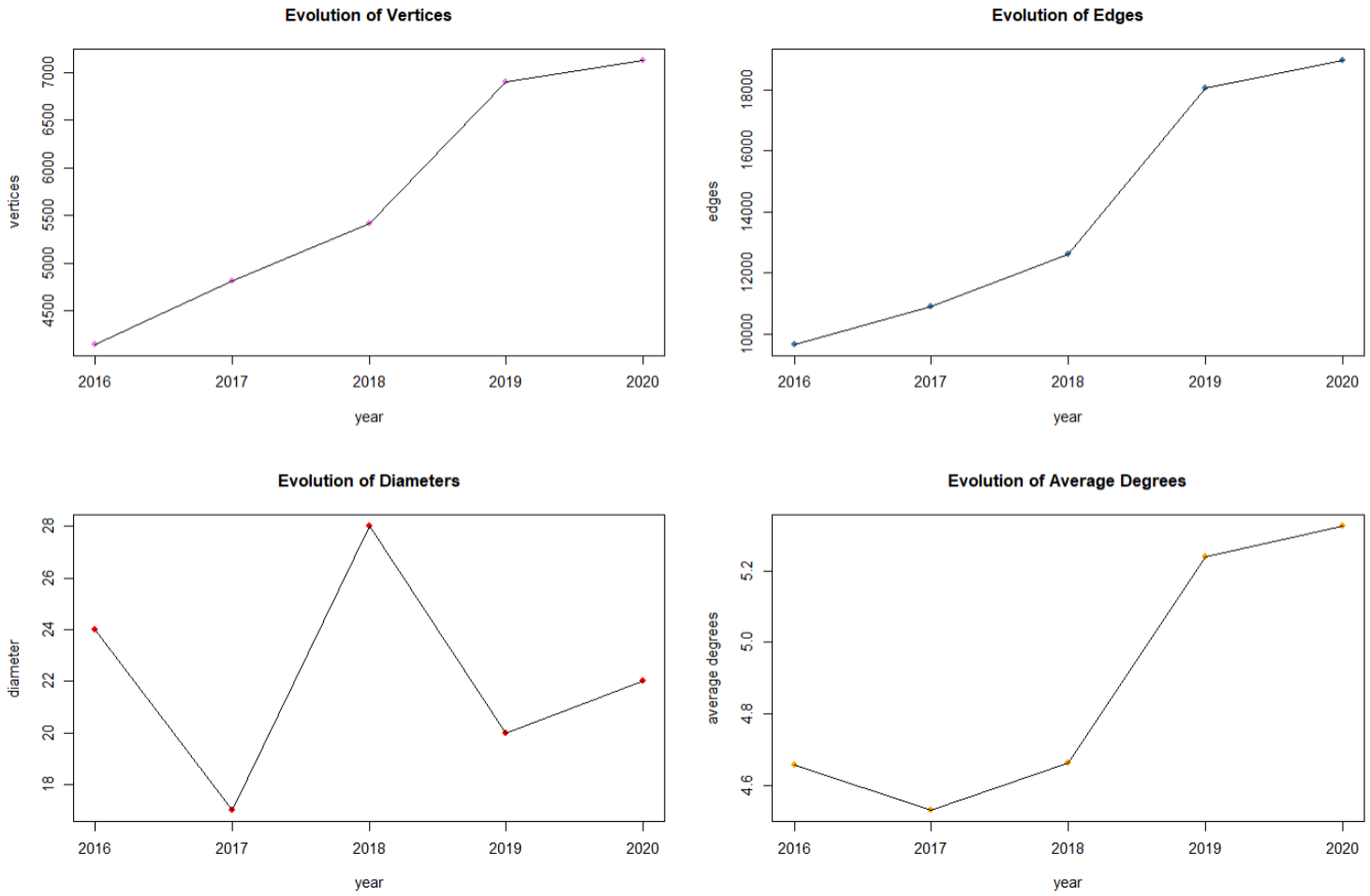


Figure 2.1.: Evolution of Vertices, Edges, Diameters and Average Degrees

By observing at the plots, we notice that there has been an increase in the number of vertices, edges and average degree of the networks over the years. This proves that more and more authors participate in writing papers over time as well as collaborate with each other to write them. On the other hand, the diameter of each graph, which represents the shortest distance between the two most distant nodes in the network, has some fluctuations. Specifically, the maximum authors collaborating to write a paper were 28, in 2018, while the minimum was 17, in 2017. Finally, by looking at the average degree plot, we confirm that in 2017, the average

number of authors collaborated to write papers was the lowest, while in 2020 it was the highest (above 5).

3. Important nodes

The following step of the analysis is to find the 10 most important authors for each year, taking into consideration two metrics, their degree and their PageRank values.

- The 5-year evolution of the top-10 authors with regard to **Degree** (simple, not weighted) for each year is presented below:

names	*top_degrees_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

names	*top_degrees_2017*
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

names	*top_degrees_2018*
Philip S. Yu	70
Jiawei Han 0001	37
Kun Gai	35
Wenwu Zhu 0001	28
Jure Leskovec	27
Jing Gao 0004	27
Chao Zhang 0014	27
Xing Xie 0001	26
Enhong Chen	25
Qi Liu 0003	25

names	*top_degrees_2019*
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

names	*top_degrees_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
Ruiming Tang	35

Table 3.1.: 5-year evolution of the top-10 authors with regard to Degree

We can observe that Philip S. Yu is the top author regarding the degree for years 2016-2019 and Jiawei Han 001 in 2020, meaning that they collaborated with many other authors every year. Moreover, it is highly likely that an author who has collaborated with others for writing papers in the first year, that they will do the same in the next years.

- The 5-year evolution of the top-10 authors with regard to **PageRank** for each year is presented below:

names	*page_rank_2016*
Philip S. Yu	0.001729
Hui Xiong 0001	0.001458
Jiawei Han 0001	0.001412
Jiebo Luo	0.001310
Jieping Ye	0.001003
Yi Chang 0001	0.000960
Hanghang Tong	0.000927
Christos Faloutsos	0.000922
Maarten de Rijke	0.000916
Jiliang Tang	0.000916

names	*page_rank_2017*
Philip S. Yu	0.001456
Jiawei Han 0001	0.001359
Hui Xiong 0001	0.001100
Jure Leskovec	0.001068
Jiebo Luo	0.000945
Hanghang Tong	0.000929
Jiliang Tang	0.000775
Yi Chang 0001	0.000771
Chao Zhang 0014	0.000751
Ingmar Weber	0.000721

names	*page_rank_2018*
Philip S. Yu	0.001981
Jiawei Han 0001	0.000930
Jure Leskovec	0.000875
Wenwu Zhu 0001	0.000784
Chao Zhang 0014	0.000678
Xing Xie 0001	0.000626
Jing Gao 0004	0.000626
Martin Ester	0.000620
Yiqun Liu 0001	0.000614
Kun Gai	0.000613

names	*page_rank_2019*
Philip S. Yu	0.001587
Hui Xiong 0001	0.000963
Weinan Zhang 0001	0.000877
Jieping Ye	0.000726
Hanghang Tong	0.000702
Jiawei Han 0001	0.000686
Peng Cui 0001	0.000657
Jie Tang 0001	0.000652
Enhong Chen	0.000638
Gerhard Weikum	0.000626

names	*page_rank_2020*
Jiawei Han 0001	0.001075
Hui Xiong 0001	0.000759
Hongxia Yang	0.000728
Elke A. Rundensteiner	0.000698
Yong Li 0008	0.000682
Jieping Ye	0.000680
Peng Cui 0001	0.000653
Xiuqiang He	0.000647
Ji-Rong Wen	0.000645
Jiliang Tang	0.000642

Table 3.2.: 5-year evolution of the top-10 authors with regard to PageRank

It can be observed that, regarding PageRank, the top authors are again Philip S. Yu and Jiawei Han 001.

4. Communities

For the final task, we will perform community detection on the 5 undirected co-authorship graphs. Specifically, we apply “fast greedy”, “infomap” and “louvain” clustering algorithms and compare their performance. By observing at the Tables 4.1., 4.2., it appears that “louvain” clustering algorithm performed better, compared to the other two algorithms, since it is more **time-efficient** and it has the highest **modularity** scores. “Louvain” algorithm maximizes a modularity score for each community, where the modularity quantifies the quality of an assignment of nodes to communities. This means evaluating how much more densely connected the nodes within a community are, compared to how connected they would be in a random network. Therefore, we choose this algorithm for community detection, as it is fast and can perform well at detecting communities in large networks.

fast_greedy	*infomap*	*louvain*
0.980	0.961	0.981
0.984	0.967	0.986
0.982	0.961	0.984
0.971	0.941	0.977
0.962	0.933	0.971

Table 4.1.: Modularity scores for each clustering algorithm

```
system.time(cluster_fast_greedy(graph_2016))
 user  system elapsed
 0.10    0.00    0.09
system.time(cluster_fast_greedy(graph_2017))
 user  system elapsed
 0.11    0.04    0.16
system.time(cluster_fast_greedy(graph_2018))
 user  system elapsed
 0.10    0.00    0.09
system.time(cluster_fast_greedy(graph_2019))
 user  system elapsed
 0.08    0.00    0.08
system.time(cluster_fast_greedy(graph_2020))
 user  system elapsed
 0.06    0.00    0.07
```

```
system.time(cluster_infomap(graph_2016))
 user  system elapsed
 1.14    0.00    1.14
system.time(cluster_infomap(graph_2017))
 user  system elapsed
 1.27    0.00    1.27
system.time(cluster_infomap(graph_2018))
 user  system elapsed
 1.48    0.00    1.48
system.time(cluster_infomap(graph_2019))
 user  system elapsed
 2.07    0.00    2.07
system.time(cluster_infomap(graph_2020))
 user  system elapsed
 3.10    0.00    3.09
```

```
system.time(cluster_louvain(graph_2016))
 user  system elapsed
 0.03    0.01    0.05
system.time(cluster_louvain(graph_2017))
 user  system elapsed
 0.03    0.00    0.03
system.time(cluster_louvain(graph_2018))
 user  system elapsed
 0.06    0.00    0.06
system.time(cluster_louvain(graph_2019))
 user  system elapsed
 0.08    0.00    0.08
system.time(cluster_louvain(graph_2020))
 user  system elapsed
 0.06    0.00    0.06
```

Table 4.2.: System time for each clustering algorithm run

Having picked the Louvain clustering algorithm, we randomly pick the author “Dawei Yin” that appears in all 5 graphs and find the communities the author was part of each year in order to observe how these five communities evolved over the five-year period. In Figure 4.1. the **evolution of the number of vertices** per community is displayed.

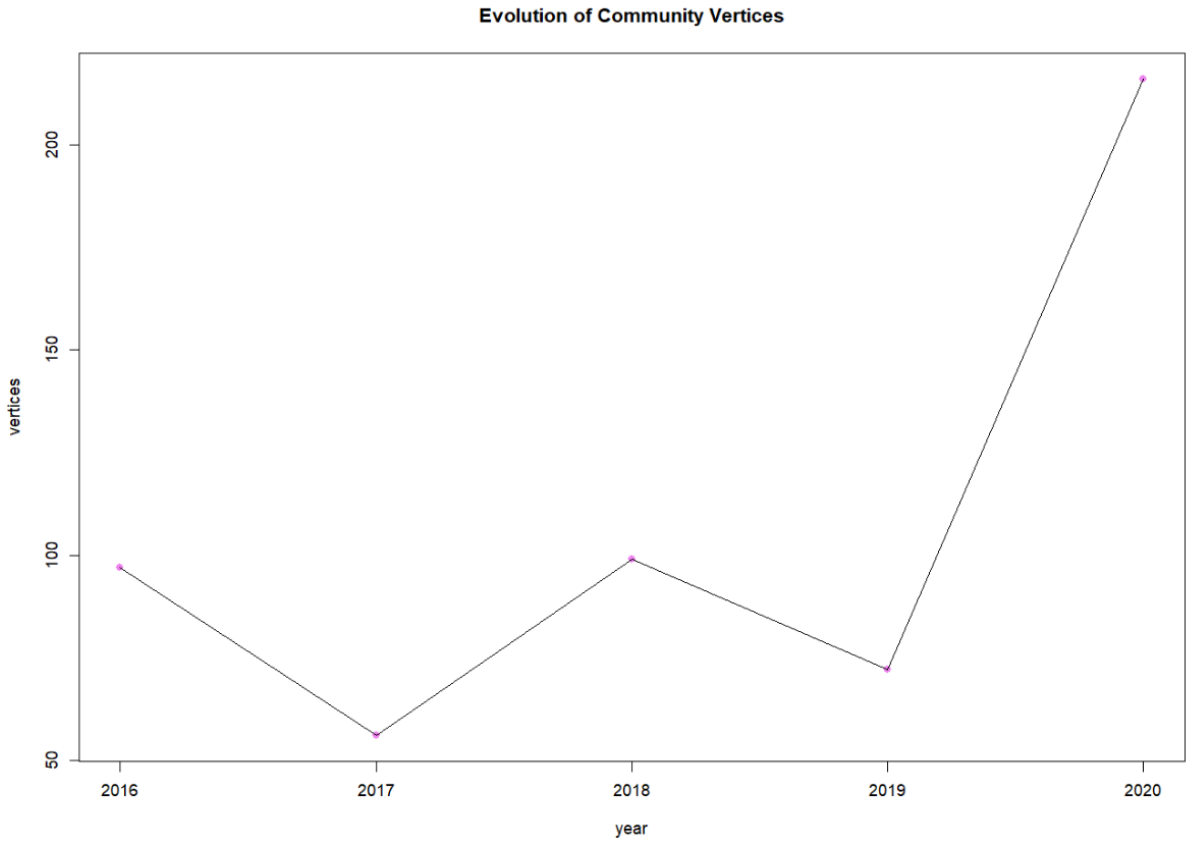


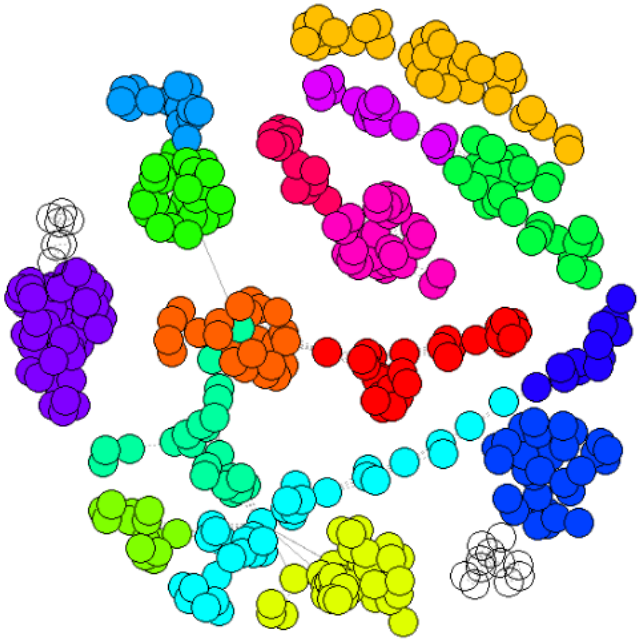
Figure 4.1.: Evolution of Community Vertices

It appears that, overall, there has been a big increase in the size of the communities to which “Dawei Yin” belonged. Namely, in 2017, the authors comprising the community were 56, while in 2020, they rose to 216.

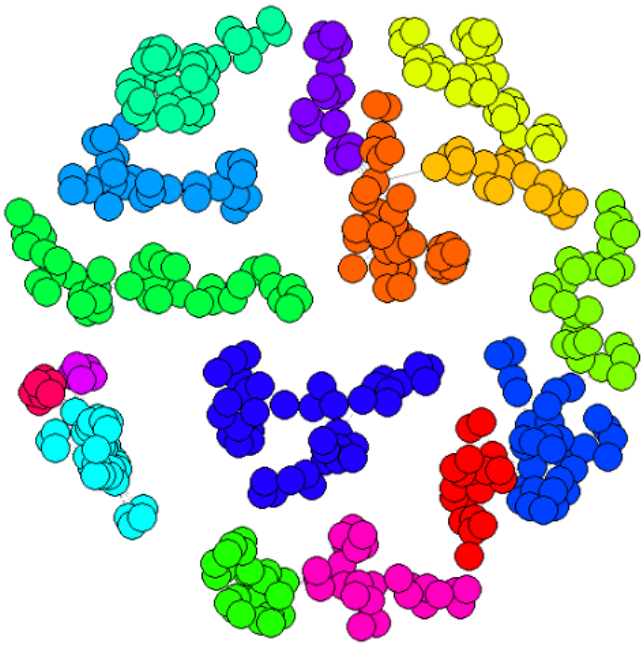
Subsequently, we search for **similarities** between these communities across the years and find out that only “Dawei Yin” and “Jiliang Tang” were the common authors in all 5 of them.

The final part of the analysis is to plot the mid-sized communities, consisting of 40 to 90 authors, with a different color depicting each community and an induced subgraph in order to create a meaningful and aesthetically pleasing visualization. The plots are displayed in Figure 4.2.

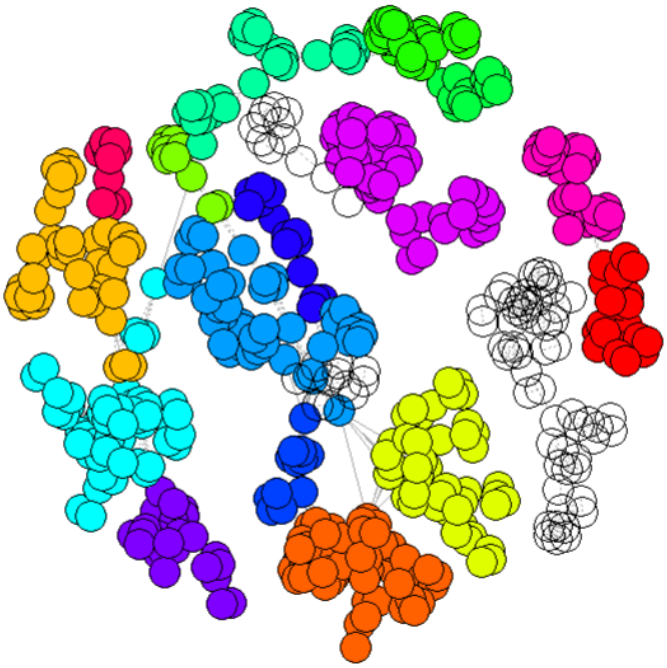
Communities of 2016



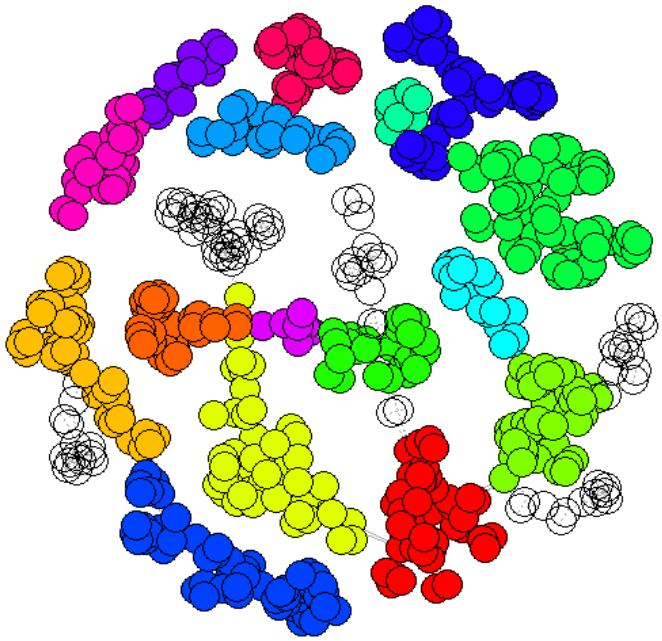
Communities of 2017



Communities of 2018



Communities of 2019



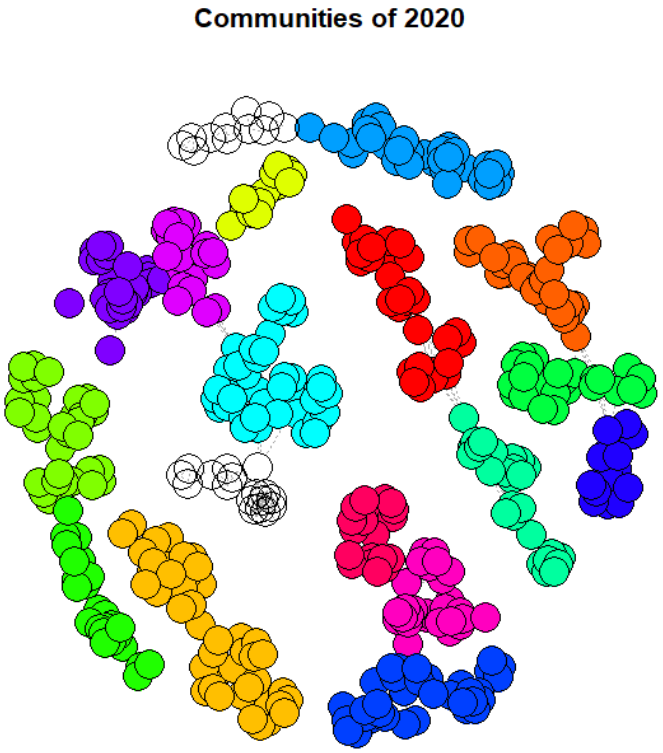


Figure 4.2.: Communities plots for each year