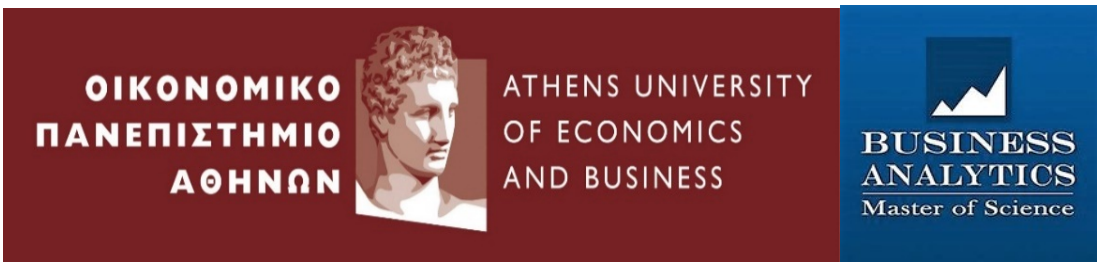


Department: Department of Management Science and Technology

M.Sc.: Business Analytics (Full-Time)

Course: Social Network Analysis

Professor: Katia Papakonstantinopoulou

Deliverable: Second Project

Student Name: Evangelos Lakkas-Pyknis

Registration Number: f2822306

E-mail: eva.lakkaspyknis@aueb.gr

Περιεχόμενα

Data Preparation

To handle the original large file I used the Unix command line in order to filter the original data and keep only the records that are not older than 5 years (so keep the records in the period 2016-2020) and at the same time correspond to the 5 conferences of interest referred in the example of the project (KDD, ICWSM, WWW, IEEE and CIKM)[1]. The code of the Unix command line can be seen in the .txt file of the compressed directory. With this code, I generated 5 .csv files that contain papers for the rereferred conferences for each year in the period 2016-2020. After that, I processed each one of the 5 .csv files using Python (this code can be seen in the Jupyter Notebook in the submitted compressed directory) to bring these files in the asked edge list weighted format (3 columns named from, to, and weight).

Average Degree over time

After writing some code in R using the igraph library to calculate the metrics of the number of vertices and edges, the diameter of the graph, and the average (not weighted) degree for each graph that corresponds to one of the years in the period 2016-2020 I created the following graphs to summarize this information.

As we can see from the line charts below the number of vertices (depicted with the blue line) we observe significant fluctuations and generally an increasing trend over the years except for the year 2018 where we observe a sudden decrease in the number of vertices and for this year this metric meets a global minimum before starting to increase again in the next years until reaching its maximum for the year 2020. In simpler words, the number of vertices corresponds to the number of academic authors of the papers that were submitted in the conferences that interest us. The number of edges (depicted with the red line) presents the same behavior as the one analyzed before for the number of vertices, but we can see that this metric is more volatile as it decreases from 2017 to 2018 and increases from 2018 to 2019 in a very intense way. Again, in simpler words, we can interpret this evolution as the evolution of cooperation between scientists for these five years.

[1] Note: during this filtering, I kept only the records that in the third column had values identical to the ones in the parenthesis. So records that correspond to conferences with similar but not identical names such as KDD Cup, WWW (Alternate Track Papers & Posters), or WWW (Companion Volume) have been excluded. If the filtering was implemented in another way obviously the results of the data preprocessing and thus of the network analysis would be different but because the project description did not provide more specific guidelines for this issue I adopted this rationale which seemed reasonable.

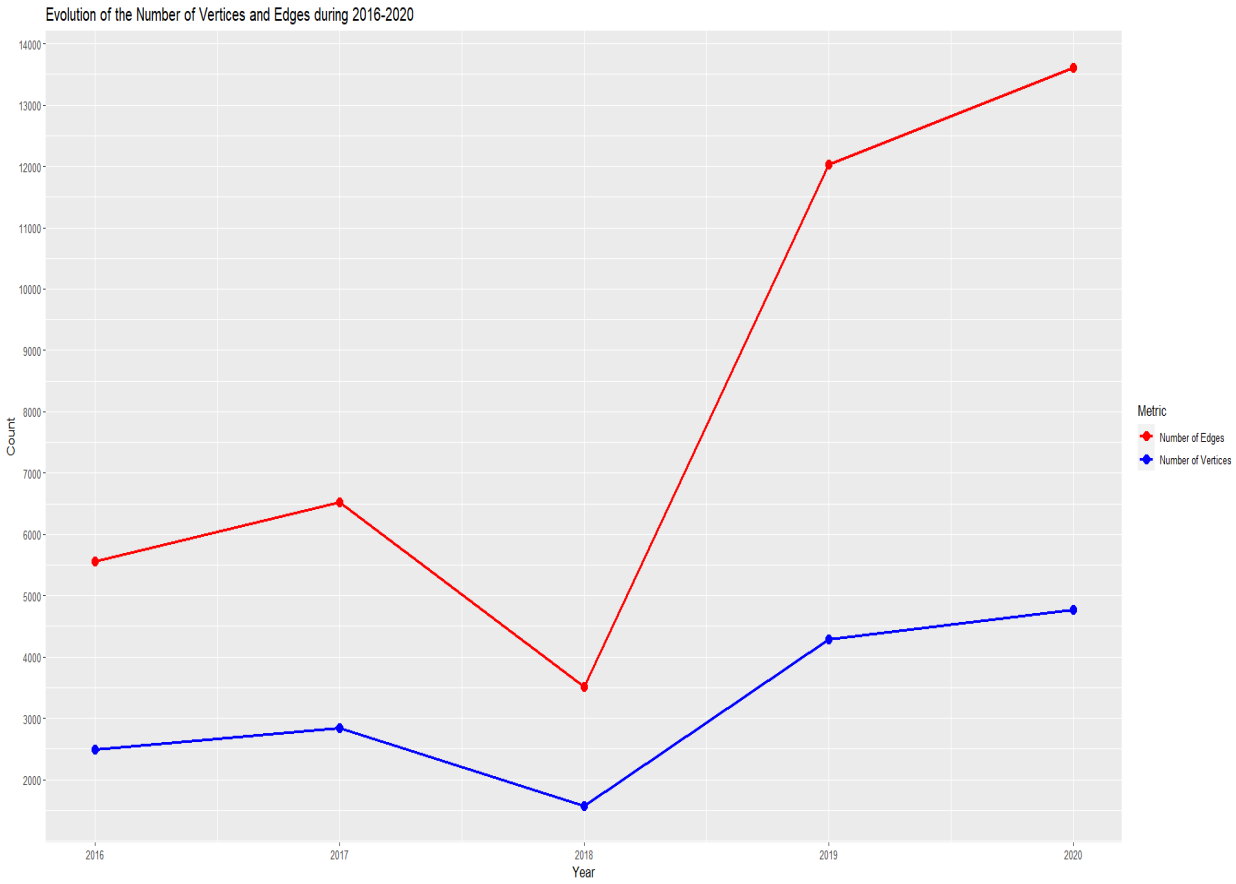


Figure 1: Number of vertices (blue line) and edges (red line) evolution through 2016-2020

All in all, based on these two metrics I conclude that the scientific activity in the conferences of interest presents an increasing trend with an exception for the year 2018.

In the following diagram we can see how the length of the knowledge graph diameter varies for each year over time. Here a quadratic trend is more clearly present as the diameter gets smaller for the first two years and increases linearly for the last two years. I can say that the fluctuation over this metric is significant because the length of the longest-shortest path decreased from 20 to 10 in just two years and then it surpassed 20 for the last year. This means that the knowledge graph gets more compact (as the distance between the two furthest vertices gets smaller) in the first two years and then expands again and becomes more sparse for the last two (as the distance between the two furthest vertices gets larger).

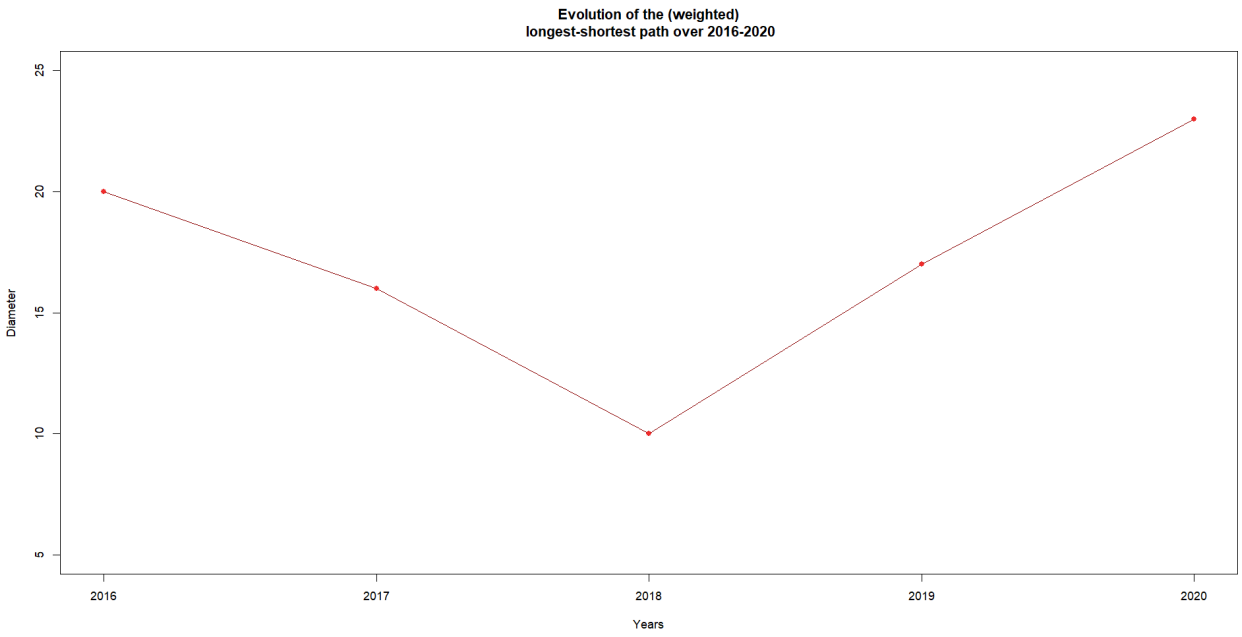


Figure 2: Diameter evolution over 2016-2020

As for the average (not weighted) degree we see that there is an increase for each year, except from the year 2018. The average simple degree does not appear very significant fluctuations and changes between consecutive years (especially when compared with the previously mentioned metrics) with the main exception being the change between 2018 and 2019 where the average degree makes a jump of approximately 1.5 units. So, the year 2020 presents the greatest average (unweighted) degree centrality for the knowledge graph.

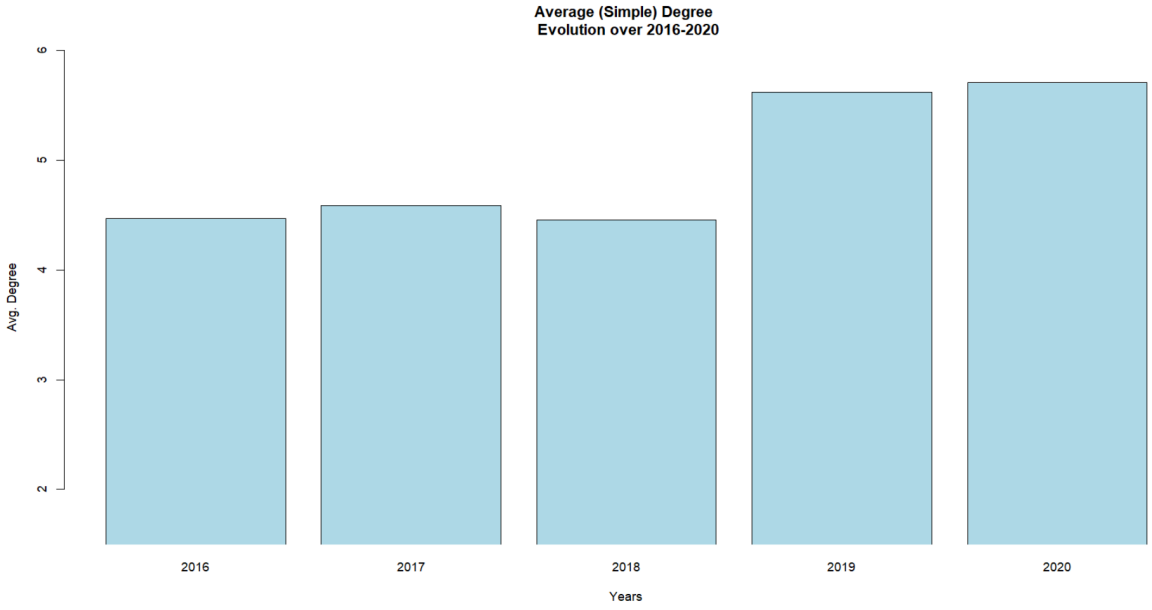


Figure 3: Average degree evolution

Important nodes

Top 10 authors with the highest degree per year

In the following tables we see a ranking of the authors based on their degree for each year. We notice that there is a clear change for the degree values through the years as there is an increase in the number of these values with the only exception being the 2017 to 2018 where the simple degree is decreased significantly but this decrease is followed by a rapid increase in the last two years. This alteration degree can be seen in the degree of the most important node for each year (2016: 41, 2017: 42, 2018: 27, 2019: 59, 2020: 62), of the least important node for each year (2016: 23, 2017: 26, 2018: 15, 2019: 34, 2020: 34) or if we compare the degree of the authors that stand in similar ranking positions for different years (i.e. see how the degree of the 4th most important vertex changes in the period 2016-2020). This evolution of the simple degree is consistent with the conclusions made from the bar chart in the previous task. As for the variation of the names of the authors in this top 10 list through the years we notice that it is not very large as there are authors that are present in this list in most or in all the years (such as Philip S. Yu or Hui Xiong 0001) and others that have a more temporal presence (such as Naren Ramakrishnan or Clemens Mewald)

Vertex (author)	Simple Degree 2016
Jiawei Han 0001	41
Hui Xiong 0001	37
Jieping Ye	32
Naren Ramakrishnan	29
Philip S. Yu	29
Yi Chang 0001	27
Rayid Ghani	25
Jiliang Tang	24
Hanghang Tong	23
Christos Faloutsos	23

Vertex (author)	Simple Degree 2017
Jiawei Han 0001	42
Hui Xiong 0001	36
Jieping Ye	31
Clemens Mewald	31
Mustafa Ispir	31
Martin Wicke	31
Zakaria Haque	31
Yi Chang 0001	28
Philip S. Yu	26
Jure Leskovec	26

Vertex (author)	Simple Degree 2018
Kun Gai	27
Philip S. Yu	19
Jiawei Han 0001	19
Martin Ester	19
Yiqun Liu 0001	18
Weinan Zhang 0001	17
Chao Zhang 0014	16
Maarten de Rijke	16
Liqin Zhao	15
Guorui Zhou	15

Vertex (author)	Simple Degree 2019
Weinan Zhang 0001	59
Hui Xiong 0001	49
Philip S. Yu	41
Jie Tang 0001	39
Jieping Ye	38
Yong Li 0008	36
Enhong Chen	36
Jingren Zhou	35
Jian Pei	35
Peng Cui 0001	34

Vertex (author)	Simple Degree 2020
Jiawei Han 0001	62
Hongxia Yang	43
Hui Xiong 0001	42
Xiuqiang He	41
Peng Cui 0001	39
Wei Wang 0010	38
Jieping Ye	37
Ruiming Tang	35
Jiliang Tang	35
Weinan Zhang 0001	34

Top 10 authors with the highest PageRank per year

In the following tables, we see a ranking of the authors based on their PageRank value for each year. By looking at these lists we can see that the PageRank metric has not a significant change through the 2016-2020 period, but this makes sense because PageRank has a standardization that simple degree does not (it takes values inside the range [0,1]) so it is rarer to spot a sudden change. Also, we can

observe that the authors that participate in these lists now have a steadier position for example Jiawei Han 0001 is the most important node according to PageRank for 3 out of the 5 years of interest. Similarly, Hui Xiong 0001 is steady in the top-3 for all the years except 2018 and Philip S. Yu is at the top-5 for 4 out of 5 of the years of interest.

Vertex (author)	PageRank 2016
Jiawei Han 0001	0.002357022
Hui Xiong 0001	0.002299988
Philip S. Yu	0.001819734
Jieping Ye	0.001646684
Hanghang Tong	0.001641615
Jiliang Tang	0.001576373
Christos Faloutsos	0.001538713
Maarten de Rijke	0.001525197
Huan Liu 0001	0.001437596
Yi Chang 0001	0.001408196

Vertex (author)	PageRank 2017
Jiawei Han 0001	0.002316291
Jure Leskovec	0.001810342
Hui Xiong 0001	0.001736838
Hanghang Tong	0.001422009
Philip S. Yu	0.001381459
Jiliang Tang	0.001357499
Chao Zhang 0014	0.001275487
Ingmar Weber	0.001221646
Yi Chang 0001	0.001205447
Markus Strohmaier	0.001122112

Vertex (author)	PageRank 2018
Yiqun Liu 0001	0.002054649
Maarten de Rijke	0.001891094
Philip S. Yu	0.001854036
Martin Ester	0.001728587
Jiawei Han 0001	0.001619384
Jure Leskovec	0.001548055
Shaoping Ma	0.001477192
Kun Gai	0.001454144
Huan Liu 0001	0.001429280
Evangelos Kanoulas	0.001410145

Vertex (author)	PageRank 2019
Hui Xiong 0001	0.0015496383
Weinan Zhang 0001	0.0014137658

Philip S. Yu	0.0013821797
Jieping Ye	0.0010727883
Peng Cui 0001	0.0010648325
Jie Tang 0001	0.0010579478
Gerhard Weikum	0.0010086359
Enhong Chen	0.0010017304
Jingren Zhou	0.0009979482
Liang Zhao 0002	0.0009958217

Vertex (author)	PageRank 2020
Jiawei Han 0001	0.0013753404
Hui Xiong 0001	0.0011450645
Hongxia Yang	0.0011246219
Yong Li 0008	0.0010245403
Jieping Ye	0.0010232387
Peng Cui 0001	0.0009989795
Jiliang Tang	0.0009633205
Ji-Rong Wen	0.0009629965
Xiuqiang He	0.0009575825
Ruiming Tang	0.0009041654

Communities

Fast Greedy, Infomap & Louvain clustering methods comparison

I was able to gain results from all the asked community detection methods in a very short period of time. Below I am making some comments on the results from each method.

Fast Greedy Clustering:

This algorithm detected a moderate number of communities across all years, with the number of communities ranging from 272 to 551. It achieved high modularity scores consistently, indicating a good community structure. The modularity scores were above 0.95 for all years.

Infomap Clustering:

Infomap identified the highest number of communities (compared to the two other algorithms) for all the years, with the number ranging from 297 to 692. Despite detecting the most communities, modularity scores were slightly lower compared to Fast Greedy and Louvain clustering methods, ranging from 0.917 to 0.961. This suggests that while Infomap finds more communities but the overall community structure might be less cohesive.

Louvain Clustering:

Louvain detected a number of communities similar to Fast Greedy but slightly higher in most cases, ranging from 274 to 552. Louvain achieved the highest modularity

scores consistently across all years, ranging from 0.962 to 0.981. This indicates that Louvain clustering provides the most robust community structure, balancing the number of communities with high modularity and thus it seems to be the best choice for these data.

Evolution of Philip S. Yu's communities

Based on the previous conclusion I implemented the Louvain algorithm because it had the best performance on my data to detect the evolution of the communities that Philip S. Yu belongs through the years. If we examine this evolution in the long run (for all five years in the period 2016-2020) we observe that Philip S. Yu has not any authors that accompany him in his communities for all the five years. However, if we look at a shorter time and we can spot some similarities for consecutive years. For example for the period 2016-17 Philip S. Yu participates in communities that in both years include the following authors (the authors that are in both communities can also be named as common neighbors): Jiawei Zhang 0001, Chenwei Zhang, Chun-Ta Lu, Lifang He 0001 and Yuanhua Lv. For the period 2017-18 this specific2018-19 there are 10 common neighbors author has only one common neighbor named Lei Zheng 0001, for and for 2019-20 his common neighbors are: Yangyong Zhu, Guixiang Ma, Yizhu Jiao, Jiawei Zhang, Bai Wang and Yun Xiong. To quantify this community similarity I used Jaccard Similarity as a metric. Jaccard similarity is defined in the following manner.

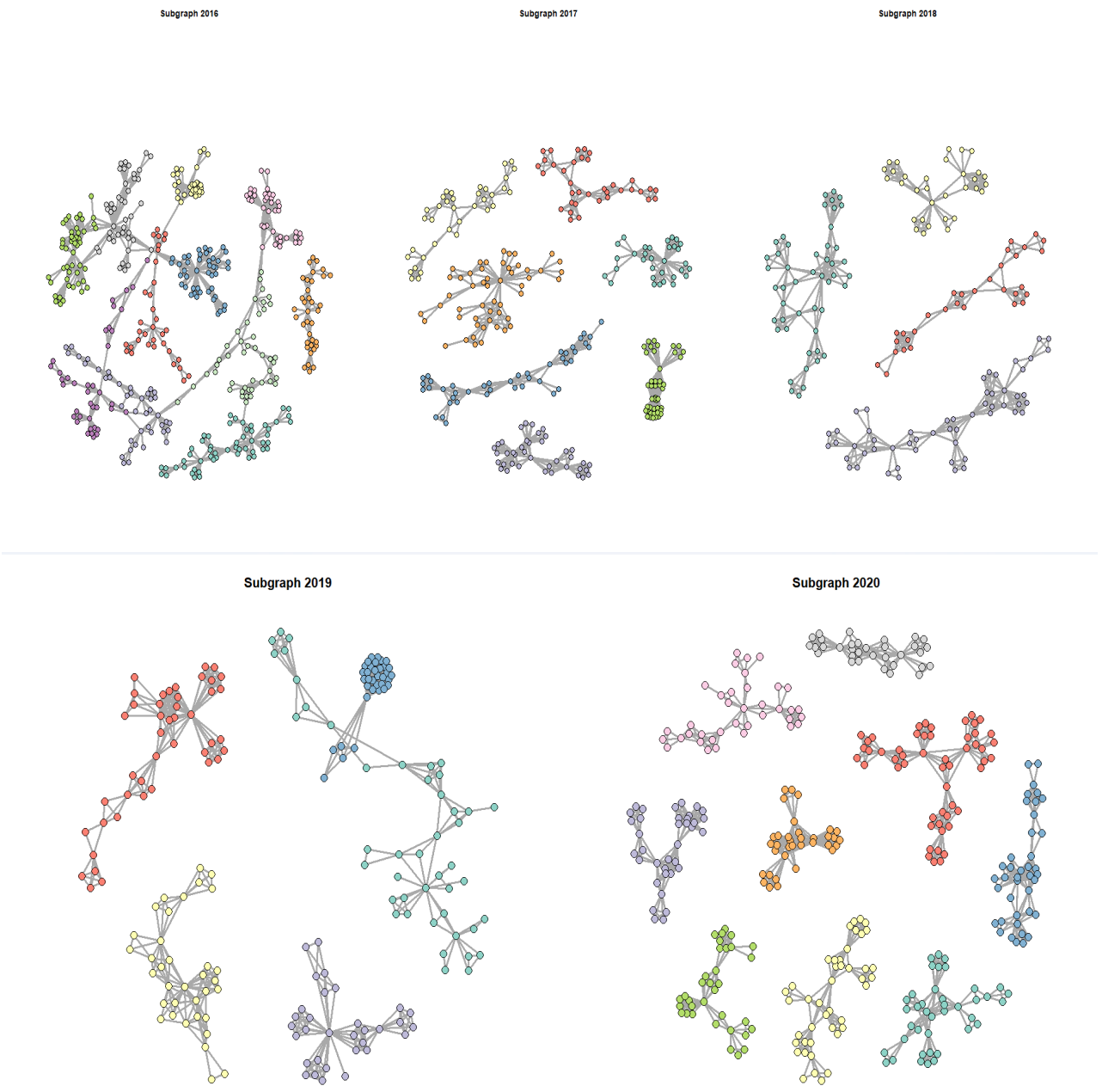
$$Jaccard\ Similarity(Community\ A, Community\ B) = \frac{Intersect(Community\ A, Community\ B)}{Union(Community\ A, Community\ B)}$$

This metric gave me the following results for the similarity of communities that Philp Yu belongs in consecutive years. We observe that the highest community similarity is met at 2018-19 and the low at 2017-18.

Time Period	Jaccard Similarity
2016-17	0.05
2017-18	0.02
2018-19	0.08
2019-20	0.04

Communities' visualization

Below 5 are shown which depict the communities for each year with each community having a distinct color. The communities presented are formed from the results of the Louvain algorithm and the communities with less than 30 and more than 55 nodes have been excluded. Louvain algorithm has been selected as more reliable due to having the highest modularity.



As we see 2020 has the most communities and 2018 the least.