Course: Social Network Analysis

Project II

Student

Name: Loukopoulos Orestis

ID: f2822104

Program: Full Time 2021-2022

ATHENS, 2022

TASK 1

To begin with, we downloaded the CSV file from the given source, and we decompress it. Then, using UNIX we filtered out all the records that are not related to the following conferences: CIKM, KDD, ICWSM, WWW, IEEE. We had also to take only those papers that published in the last 5 years. We observed that for year 2021 we did not have any paper published in the given conferences. Hence, we decided to take the papers from 2016 to 2020, so as to be able to create the 5 CSV files, we have been asked for. I would also like to add that in order to be sure that the R-studio/Python reads all the lines of the filtered CSV, I count its rows using UNIX. Continuing, I used R-studio to load the filtered CSV (I faced some struggles loading into python, as many lines were skipped). As I load it in R-studio, I found out that two lines were problematic (issue with double quotes and comma). Because the number of problematic lines was so small, I fixed them manually. After fixing them I saved the cleaned filtered CSV internally and imported it to Python in order to create the 5 CSVs (i.e., one CSV for each year in a weighted edge list form). I used "Pandas" library as it is very useful for manipulating data. After creating the 5 files, I imported them again in R-studio and created a weighted undirected graph for each year using "igraph" library.

An explanation should be provided to the reader about the way and the order the coding files should be opened and read.

1) **UNIX_Code.txt** (It contains code for filtering the dataset)

2) **Cleaning-Dataset.R** (It contains code about reading the filtered dataset, cleaning some problematic observations, and extracting a new cleaned filtered dataset)

3) **Create_CSV_files.ipynb** (It contains code to transform the filtered dataset into a weighted edge list for every year. It extracts 5 CSV files)

4) **ORESTIS_LOUKOPOULOS-igraph.R** (Code for Task 1-Task 4)

An explanation should also be provided to the reade regarding the files attached to the zip file.

1) **file1.csv** (this dataset was created in UNIX and includes only the papers that have been published in the 5 conferences)

2) **file2.csv** (this dataset was created in UNIX and includes papers published in the 5 conferences -file1.csv- and only from 2016 to 2020)

3) **mydata.csv** (this dataset is the same as file2.csv but has been cleaned in R-studio)

4) **authors_2016.csv – authors_2020.csv** (the 5 CSV files which we have been asked to create -created in Python-)

TASK 2

Continuing, we created some plots in order to display 5-year evolution of different metrics for the graph. Starting with the evolution of the Number of Vertices during 2016-2020. From the following graph (Figure 1) it can be observed that the number of vertices was steadily increasing from 2016 to 2018, but in 2019 this number increased rapidly. Then for the last year (2020) it also increases but not that rapid.

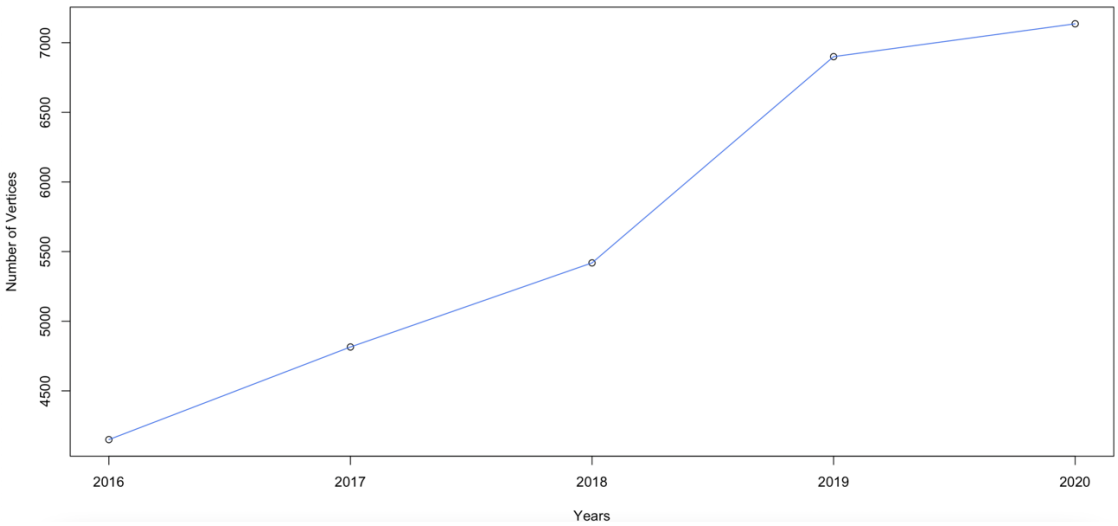

Figure 1: *Evolution of number of vertices from 2016 to 2020.*

Then, we examined the evolution of the number of edges through these years. From the following graph (Figure 2) it can be observed the same pattern as in vertices. A steadily increase from 2016 to 2018 and a rapid increase in 2019, then again, a milder increase in 2020.

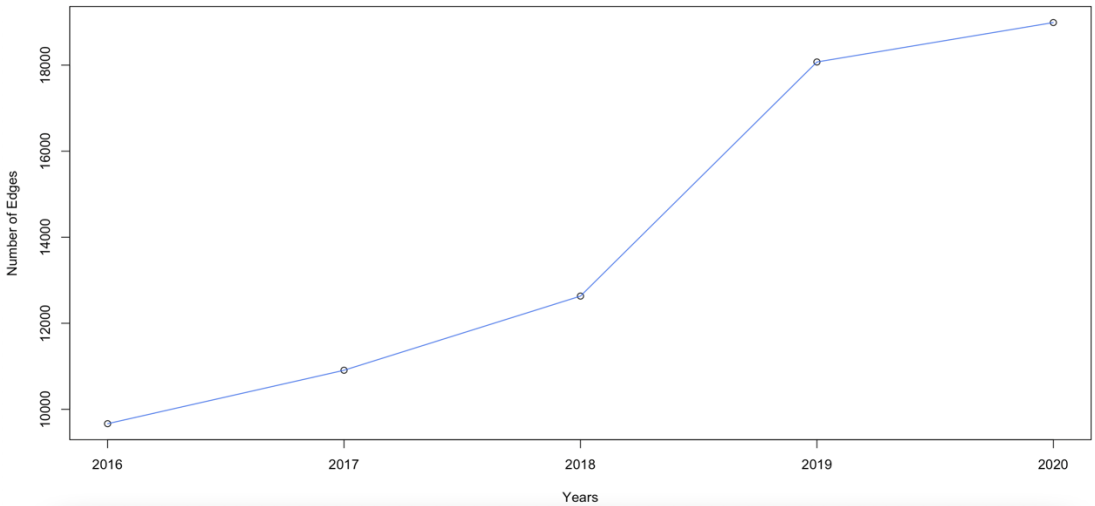

Figure 2: *Evolution of number of edges from 2016 to 2020.*

The evolution of diameter (Figure 3) follows a completely different pattern from the previous graph metrics. We can observe that there is a global minimum of diameter in 2017 then we have a rapid increase in 2018, with the diameter reaching its peak. Then in 2019 we have a rapid decrease and finally the diameter increases in 2020.

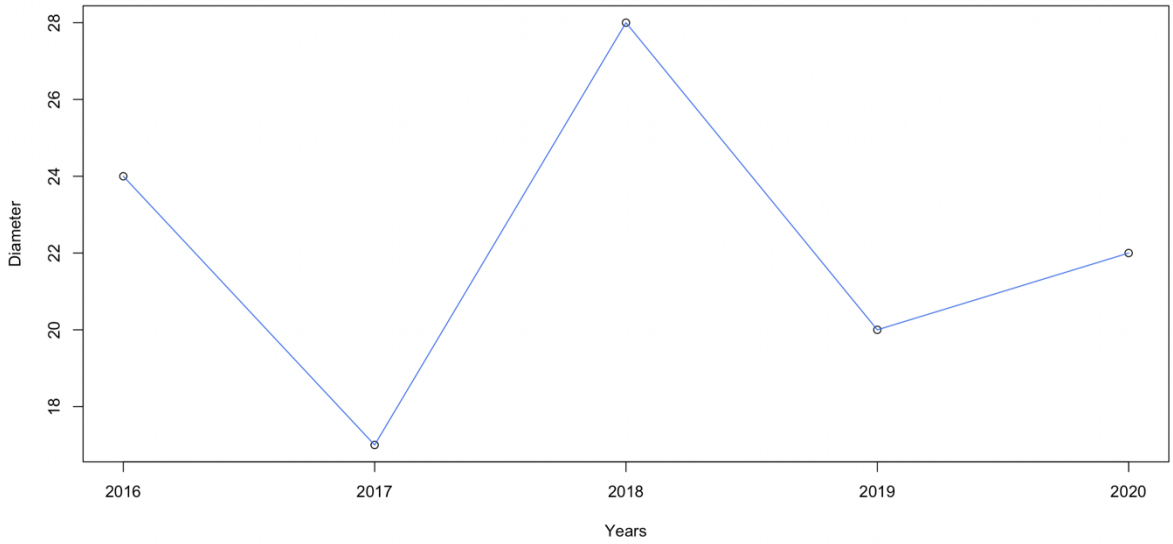

Figure 3: *Evolution of diameter from 2016 to 2020.*

As far as, the evolution of average degree (Figure 4) is concerned we can observe a decrease in 2017 while we have an increasing trend from 2018 to 2020, with a very rapid increase of average degree, again in 2019.

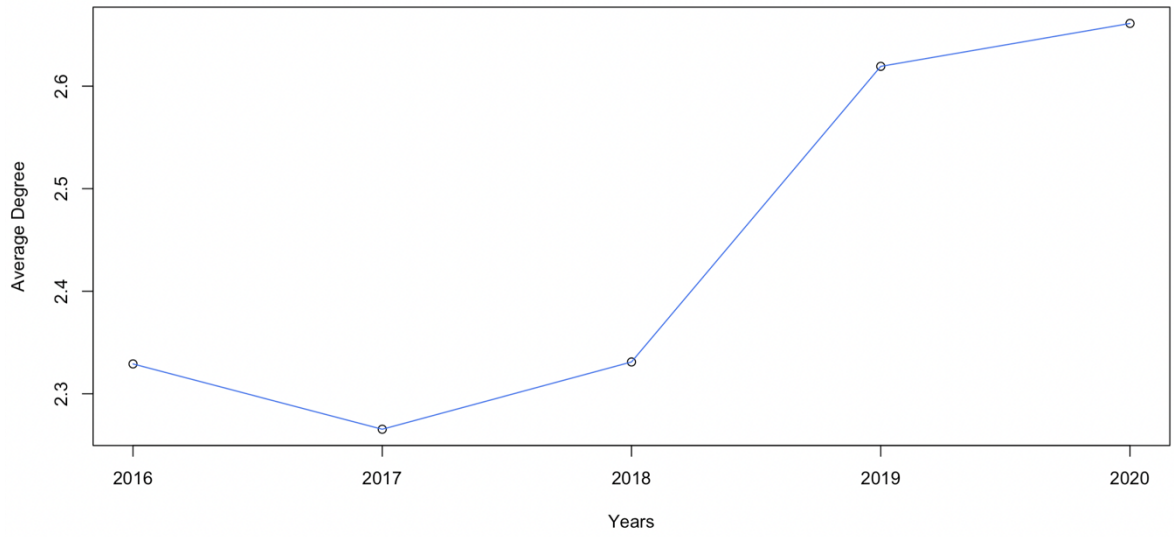

Figure 4: *Evolution of average degree from 2016 to 2020.*

TASK 3

Continuing, we created data frames in order to examine the evolution of the top 10 authors through the years, based on their degree and on their PageRank. First, let's take a look the rankings of the authors based on their degree.

2016

	Degree
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

2017

	Degree
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
Mustafa Ispir	31
Zakaria Haque	31
Martin Wicke	31

2018

	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
Qi Liu 0003	25
Haifeng Chen	25

2019

	Degree
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

2020

	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

Table 1: *Tables for top 10 authors based on their degree from 2016 to 2020*

From the above tables (Table 1) we can observe some variations in the authors appearing in the top 10 leaderboard. However, some authors such as "Philip S. Yu", "Jiawei Han 0001", "Hui Xiong 0001" are repairing through the years.

We will continue with the ranking of authors based on their PageRank through the years.

2016

	PageRank
Philip S. Yu	0.0017
Hui Xiong 0001	0.0015
Jiawei Han 0001	0.0014
Jiebo Luo	0.0013
Jieping Ye	0.001
Yi Chang 0001	0.001
Hanghang Tong	9e–04
Christos Faloutsos	9e–04
Maarten de Rijke	9e–04
Jiliang Tang	9e–04

2017

	PageRank
Philip S. Yu	0.0015
Jiawei Han 0001	0.0014
Hui Xiong 0001	0.0011
Jure Leskovec	0.0011
Jiebo Luo	9e–04
Hanghang Tong	9e–04
Jiliang Tang	8e–04
Yi Chang 0001	8e–04
Chao Zhang 0014	8e–04
Ingmar Weber	7e–04

2018

	PageRank
Philip S. Yu	0.002
Jiawei Han 0001	9e–04
Jure Leskovec	9e–04
Wenwu Zhu 0001	8e–04
Chao Zhang 0014	7e–04
Xing Xie 0001	6e–04
Jing Gao 0004	6e–04
Martin Ester	6e–04
Yiqun Liu 0001	6e–04
Kun Gai	6e–04

2019

	PageRank
Philip S. Yu	0.0016
Hui Xiong 0001	0.001
Weinan Zhang 0001	9e–04
Jieping Ye	7e–04
Hanghang Tong	7e–04
Jiawei Han 0001	7e–04
Peng Cui 0001	7e–04
Jie Tang 0001	7e–04
Enhong Chen	6e–04
Gerhard Weikum	6e–04

2020

	PageRank
Jiawei Han 0001	0.0011
Hui Xiong 0001	8e–04
Hongxia Yang	7e–04
Elke A. Rundensteiner	7e–04
Yong Li 0008	7e–04
Jieping Ye	7e–04
Peng Cui 0001	7e–04
Xiuqiang He	6e–04
Ji–Rong Wen	6e–04
Jiliang Tang	6e–04

Table 2: *Tables for top 10 authors based on their PageRank from 2016 to 2020*

From the above tables (Table 2), we can observe that there are again variations in the top 10 authors - according to their PageRank - through the years. Again, we can notice that some authors are keep reappearing through some of the years, such as “Philip S. Yu”, “Jiawei Han 0001”, “Hui Xiong 0001” and “Hanghang Tong”.

TASK 4

In the final task of this assignment, we have to apply some clustering methods in order to detect the communities on each of the 5 co-authorship graphs. We performed the following three methods: fast greedy clustering, infomap clustering, and louvain clustering. We did not face any issue applying them as no one gave an error message. We observed that fast greedy clustering and Louvain clustering were detecting approximately the same number of clusters, while the infomap clustering method was detecting a larger number of cluster than them. Furthermore, we observed that infomap method was the slowest one in terms of execution time, while Louvain was the fastest one.

Then, we were asked to pick a random author that appears in all 5 graphs. We were also asked to select one of the three clustering methods mentioned above and according to it, we had to detect the evolution of the communities (through 2016-2020) the author belongs to. We selected the method of "Louvain" clustering method for performance reasons (fast execution). After writing some code we noticed that communities have many similarities in terms of their members. Communities (Table 3) may not be identical through the years, but we can see many members remaining in the same community with our author ("Shusaku Tsumoto"). We observed that from 2016 to 2017 the community lost half of its members, while it remained completely unchanged for 3 consecutive years (2017-2019). During the final year (2020) community lost only one member with the rest remaining the same.

2016	2017	2018	2019	2020
Shusaku Tsumoto	Shusaku Tsumoto	Shusaku Tsumoto	Shusaku Tsumoto	Shusaku Tsumoto
Tomohiro Kimura	Tomohiro Kimura	Tomohiro Kimura	Tomohiro Kimura	Tomohiro Kimura
Shoji Hirano	Shoji Hirano	Shoji Hirano	Shoji Hirano	Shoji Hirano
Haruko Iwata	Haruko Iwata	Haruko Iwata	Haruko Iwata	X
Norio Yoshimoto	X	X	X	X
Chenxi Liu	X	X	X	X
Hiroshi Sakai	X	X	X	X
Michinori Nakata	X	X	X	X

Table 3: *Evolution of "Shusaku Tsumoto" community.*

Continuing, we created a visualization of the communities for every year. We have to mention that we filtered out communities that consist of less than 50 members (small communities) as well as communities that have more than 90 members (large communities). We should also add that we have used different colors for the representation of each community. The final form of the communities for each year are demonstrated on the graphs below (Figure 5, Figure 6, Figure 7, Figure 8, Figure 9).

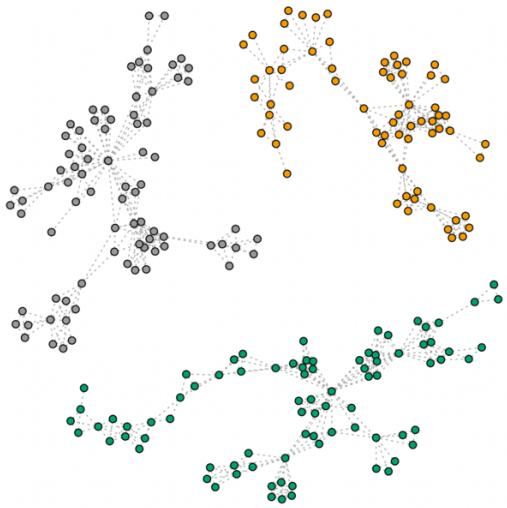

Figure 5: *Graph of Communities for 2016.*

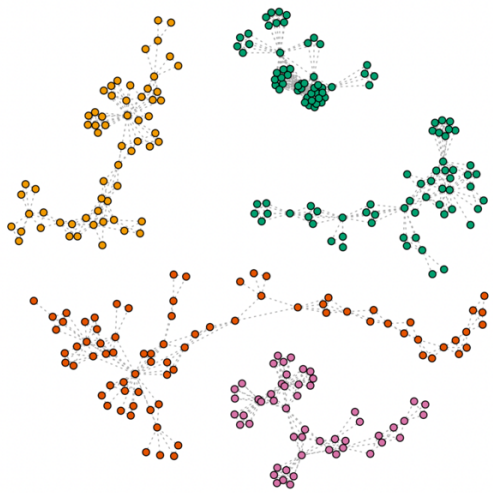

Figure 6: *Graph of Communities for 2017.*

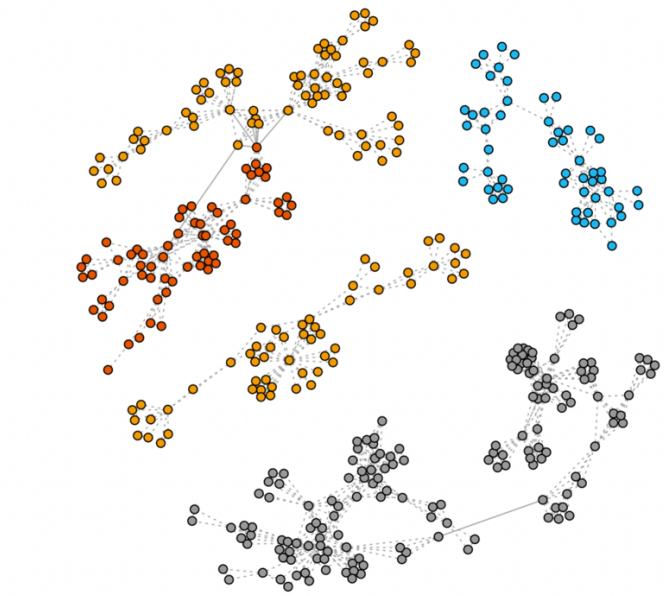

Figure 7: *Graph of Communities for 2018.*

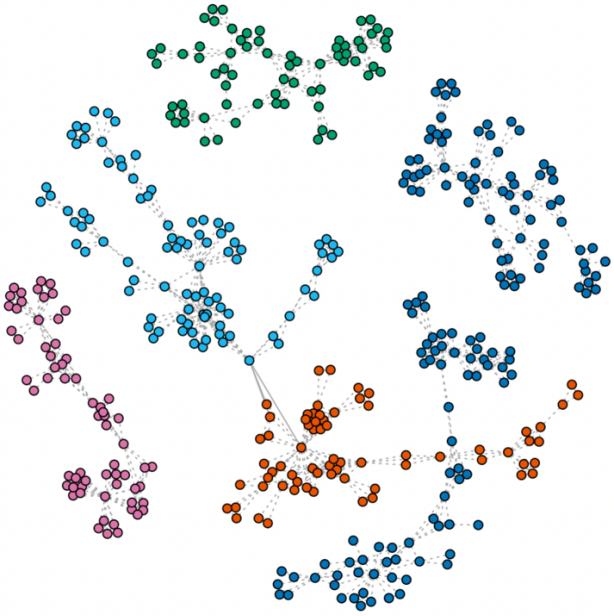

Figure 8: *Graph of Communities for 2019.*

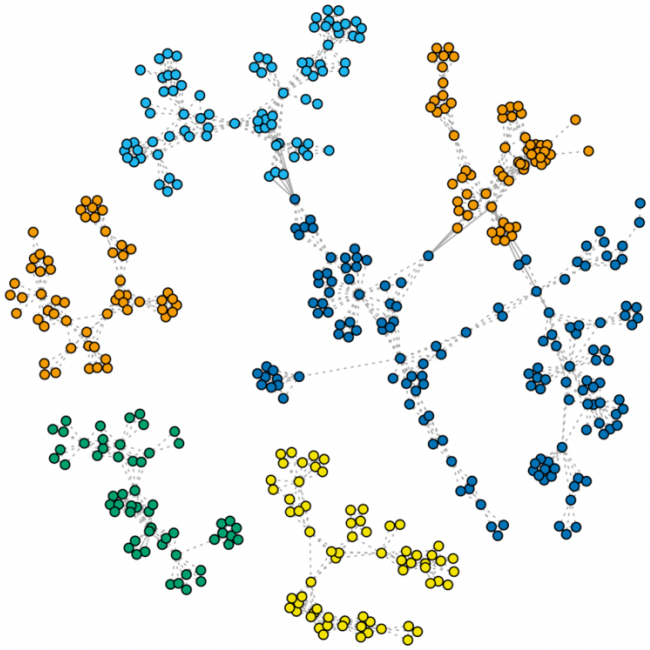

Figure 9: *Graph of Communities for 2020.*