ATHENS UNIVERSITY OF ECONOMICS & BUSINESS

M.SC. PROGRAM IN BUSINESS ANALYTICS

SOCIAL NETWORK ANALYSIS

FROM RAW DATA TO TEMPORAL GRAPH STRUCTURE EXPLORATION

INSTRUCTOR | KATIA PAPAKONSTANTINOPOULOU

STUDENT | DESPOTIS SPYRIDON P2822111

- "AUTHORS" DATASET
- IN R LANGUAGE

2022

Contents

Introduction

In this assignment we downloaded a dataset named "authors.csv.gz"[1] from the University of Athens (UOA) database. The dataset contains 2.793.928 rows and 4 columns. In each line above, the first column indicates the year the paper was published, the second column is the title of the paper, and the third column is the conference where the paper was presented. Finally, the fourth column of the line is a comma separated list of the paper's authors.

The main purpose of this assignment is to manipulate the downloaded raw data and create a weighted undirected graph. Then to draw useful insights for each the 5-year evolution with metrics for the graph (e.g. Number of vertices) and design useful plots. After to create dataframes for the 5-year evolution of the top-10 authors with regard to degree and pagerank. Last but not least, to perform community detection and pick the best method.

For the first part we used unix programming language followed by python and for the main analysis R programming language with igraph library. The environment we implemented R was R Studio.

Year	Title	Conference	Authors
2016	Separating-Plane Factorization Models: Scalable Recommendation from One-Class Implicit Feedback.	CIKM	Haolan Chen,Di Niu,Kunfeng Lai,Yu Xu,Masoud Ardakani
2016	Joint Collaborative Ranking with Social Relationships in Top-N Recommendation.	CIKM	Dimitrios Rafailidis,Fabio Crestani
2016	Probabilistic Approaches to Controversy Detection.	CIKM	Myungha Jang,John Foley,Shiri Dori-Hacohen,James Allan
2016	Online Food Recipe Title Semantics: Combining Nutrient Facts and Topics.	CIKM	Tomasz Kusmierczyk,Kjetil NÃ¸rvÃ¥g
2016	Bus Routes Design and Optimization via Taxi Data Analytics.	CIKM	Seong-Ping Chuah,Huayu Wu 0001,Yu Lu 0003,Liang Yu,StÃ©phane Bressan
2016	Distributed Deep Learning for Question Answering.	CIKM	Minwei Feng,Bing Xiang,Bowen Zhou

Table.0.1 Example of Columns in the Primarily Dataset

[1] https://hive.di.uoa.gr/network-analysis/files/authors.csv.gz

1. DBLP Co-authorship Graph

Firstly, we downloaded the zip file named “authors.csv.gz”. Then we decided to use UNIX environment (software: ubuntu for windows), due to the fast performance which provides, in order to filter out the required data for our analysis. So, by using the appropriate functions (e.g. “cat”, “grep”) we filtered out the records that were not related to the required conferences ('WWW', 'KDD', 'ICWSM', 'IEEE BigData', 'CIKM') and we splitted and exported the files by year. As a delimiter we used comma to separate the columns of the primarily dataset.

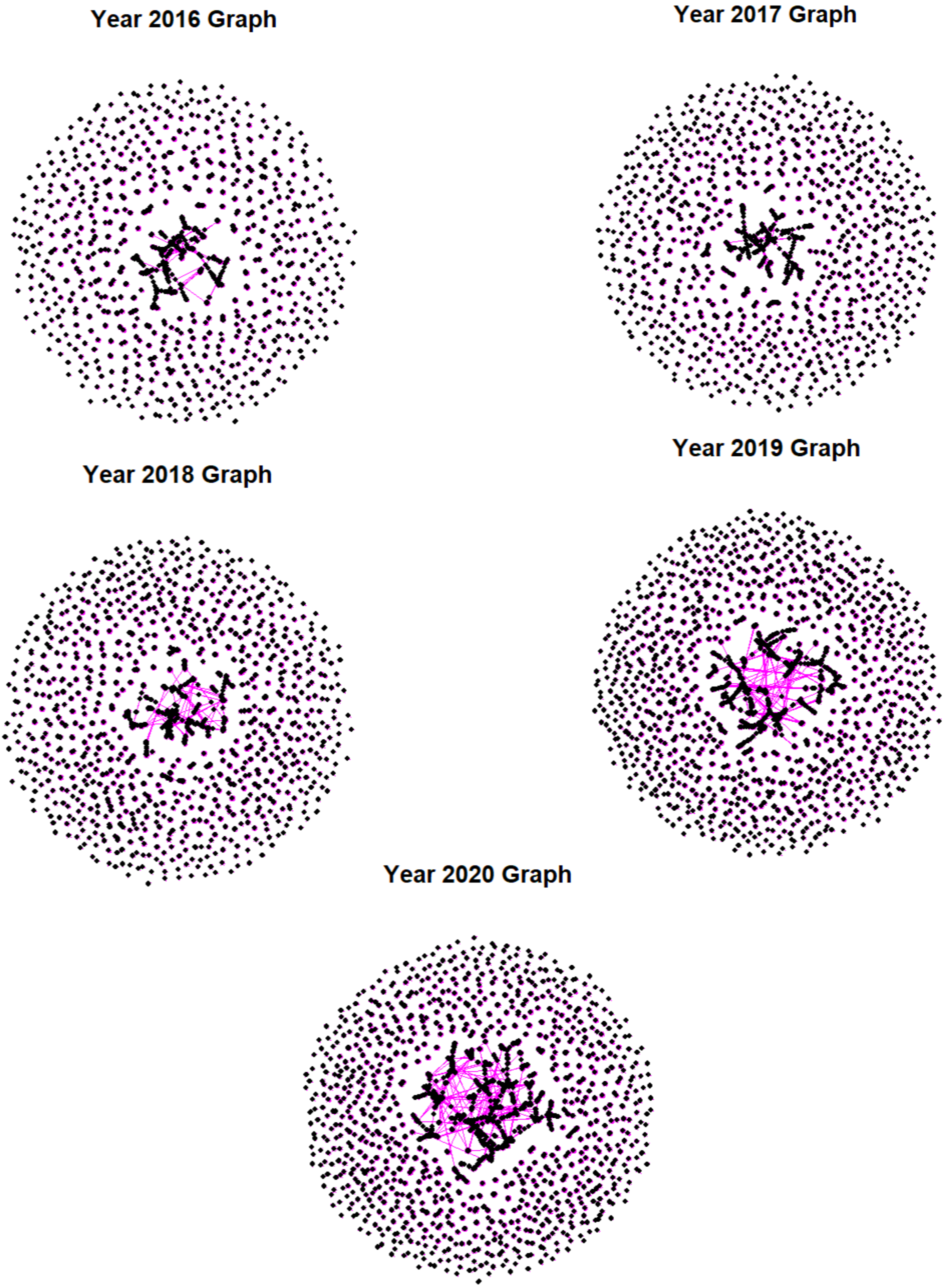
A useful insight was that the year 2021, had only excluded records with conferences that were not required in this analysis (e.g. “VMBO”, “BTW”) and therefore we focused only on the years 2016 to 2020. Probably the conferences took place later and in the future will be worth including 2021 updated data also in the analysis.

Then we proceeded by using python language in Jupiter Notebook environment to do more manipulations. We imported the exported files from Unix in order to achieve the final structure. Specifically, we created a function named “make_graph” that creates links between each author and checks if already that pair already exists or not and adds the appropriate weight. Then we exported the final finals in csv data types in order to import them in R environment later.

The final structure for each dataset of the five years is presented in the above table. The table contains three columns “source”, “target” and “weight”. The first two columns refer to the co-authors and the last column to the number of the papers the co-authors published together. For example, “Haolan Chen” author has co-authored one paper with “Di Niu” author.

Source	Target	Weight
Haolan Chen	Di Niu	1
Haolan Chen	Kunfeng Lai	1
Haolan Chen	Yu Xu	1
Haolan Chen	Masoud Ardakani	1
Di Niu	Kunfeng Lai	1
Di Niu	Yu Xu	1

Table.1.1 Example of the Final Structure of the Dataset



Year 2016 Graph

Year 2017 Graph

Year 2018 Graph

Year 2019 Graph

Year 2020 Graph

Table 1.2 Plot of Each Graph per Year

2. Average Degree Over Time

In this section we create plots that visualize the 5 year evolution of different metrics for the graph. That metrics are important in order to have a straight forward view to our graph.

2.1 Number of Vertices

The names of the graph are the vertices of the graph. (If you're talking about just one of the vertices, it's a vertex.) The order of a graph is its number of vertices (V). In the below plot we can see the how the number of vertices are increasing by the years. As we can see the year with the lowest vertices is 2016 and the year with the highest vertices is 2020. That pattern, maybe indicates that the number of authors who publish papers are increasing. Between the years 2019 and 2020 there is a small increased compared to the other years.

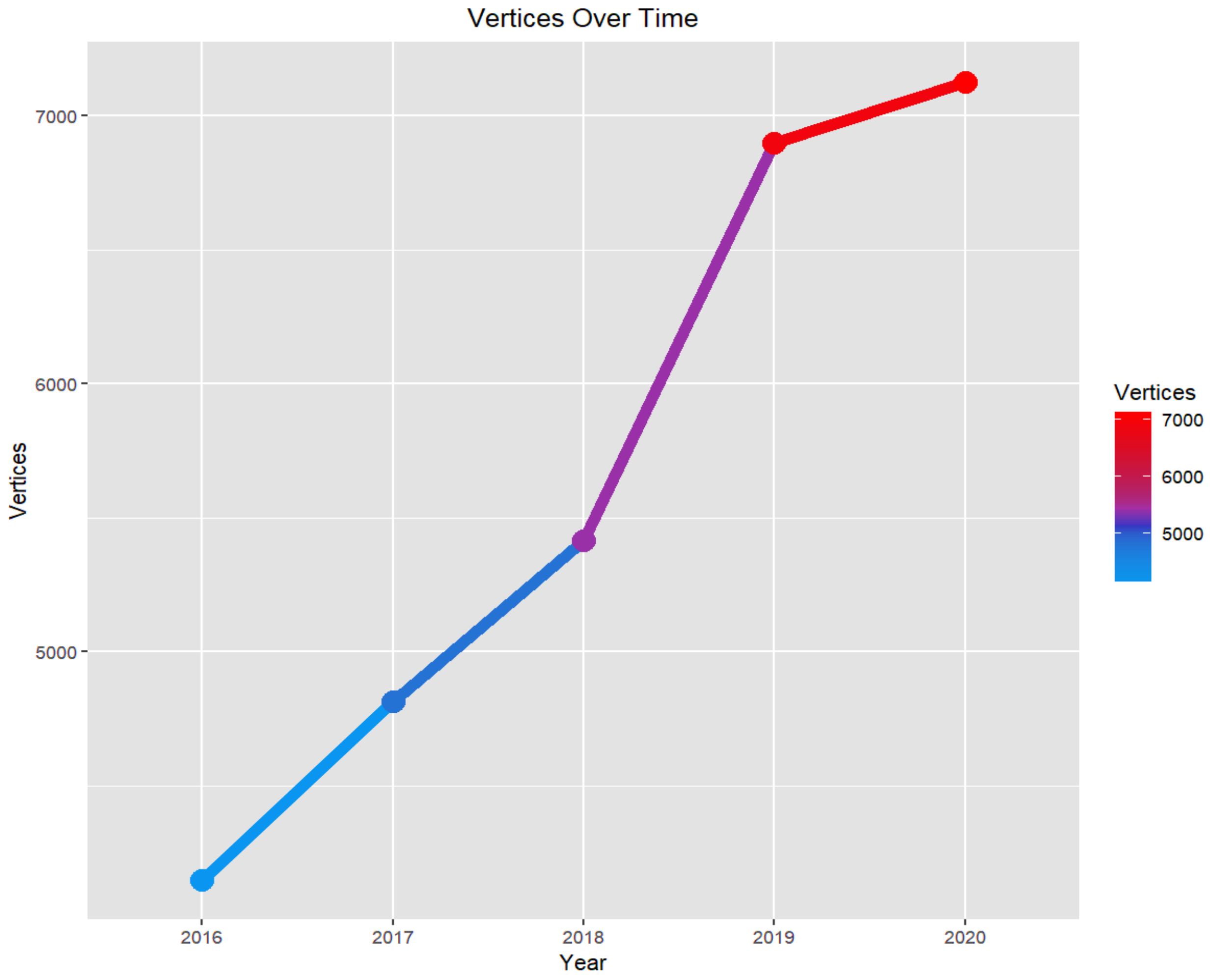


Figure 2-1 Number of Vertices Over Time

2.2 Number of Edges

Edges represent the presence of a connection or relationship between two nodes. In social network analysis these are usually some type of social tie. In the below plot we can observe how the edges are increasing over the years. The uptrend pattern is the same as the vertices. The maximum number of edges is located in 2020 and the lowest in 2016. That indicates that the relationships between the authors is increasing over the years and therefore more and more authors cooperate with each other in order to publish papers.

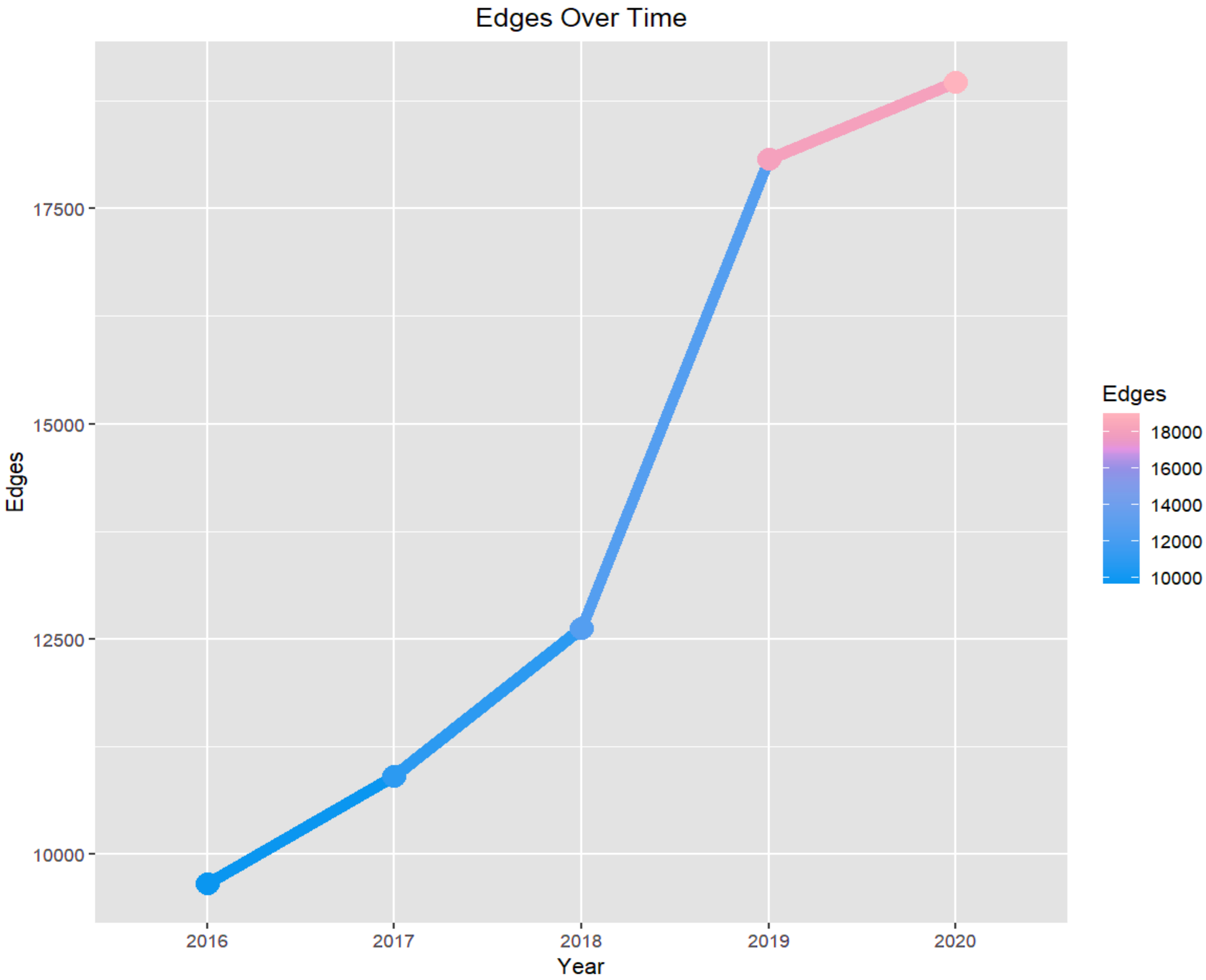


Figure 2-2 Number of Edges Over Time

2.3 Diameter of the Graph

The diameter of a graph is the length of the shortest path between the most distanced nodes. The bellow graph describes how the diameter of the graph is changing during the years. As we can see there is not a continuous uptrend or drop pattern in the diameter during the years and there is high fluctuation. The peak point is in the year 2018 with papers written by 28 authors, while the minimum point was the year 2017. Therefore, we can conclude that in year 2017 there were not significant cooperation between authors to write a paper compared to the other years and in 2018 we had the maximum cooperation between the authors to write a paper.

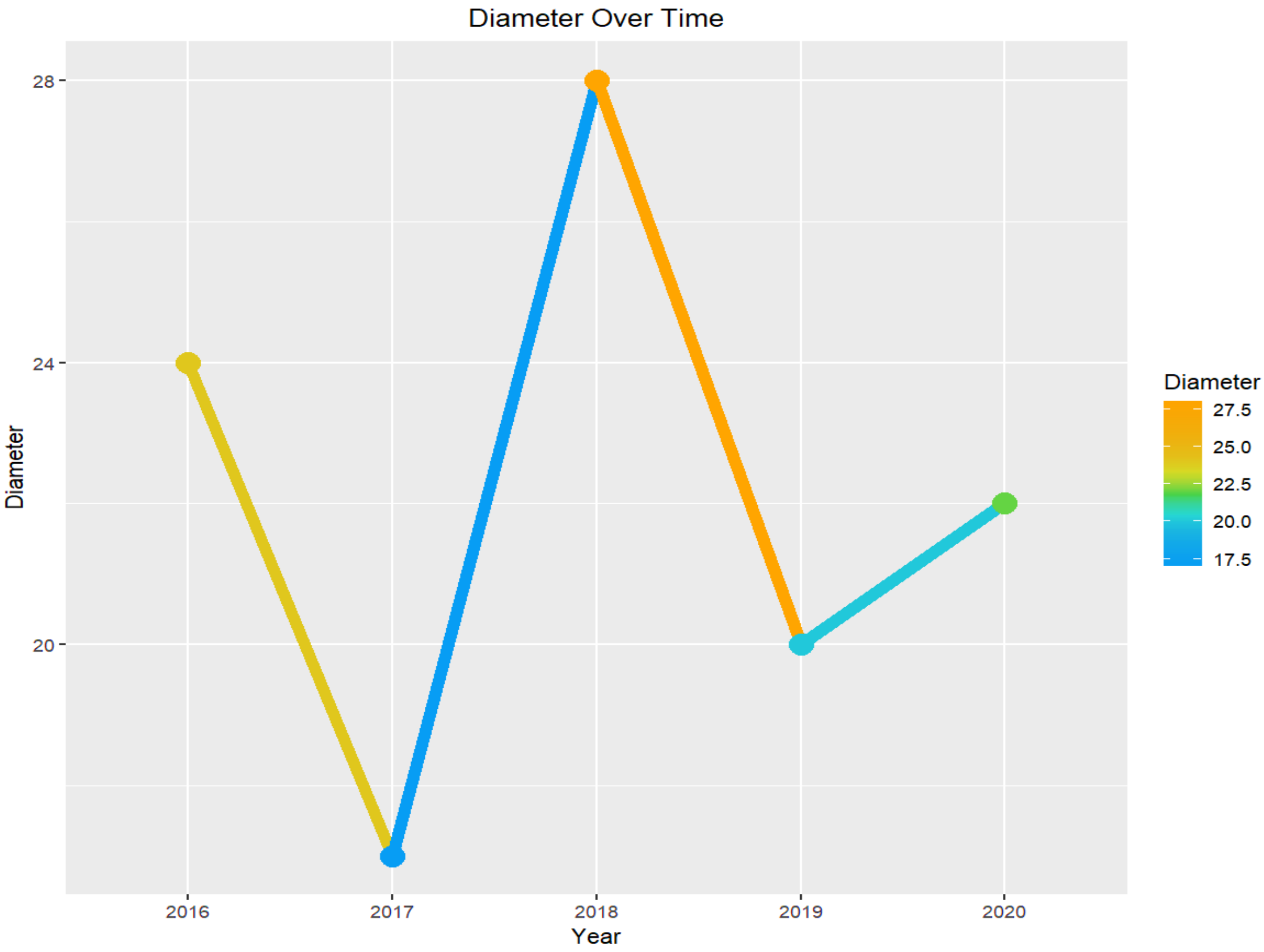


Figure 2-3 Diameter of the Graph over Time

2.4 Average Degree (Simple, not weighted)

Average degree is simply the average number of edges per node in the graph. The below graph describes the changes in average degree over the years. As we can see there is a small fluctuation between the years 2016 and 2017, with the author connections decreasing from one year to the other. In other words, we had fewer cooperations in the papers published this year. Overall, the peak point is the year 2020 with the highest average degree and the lowest point in the year 2017.

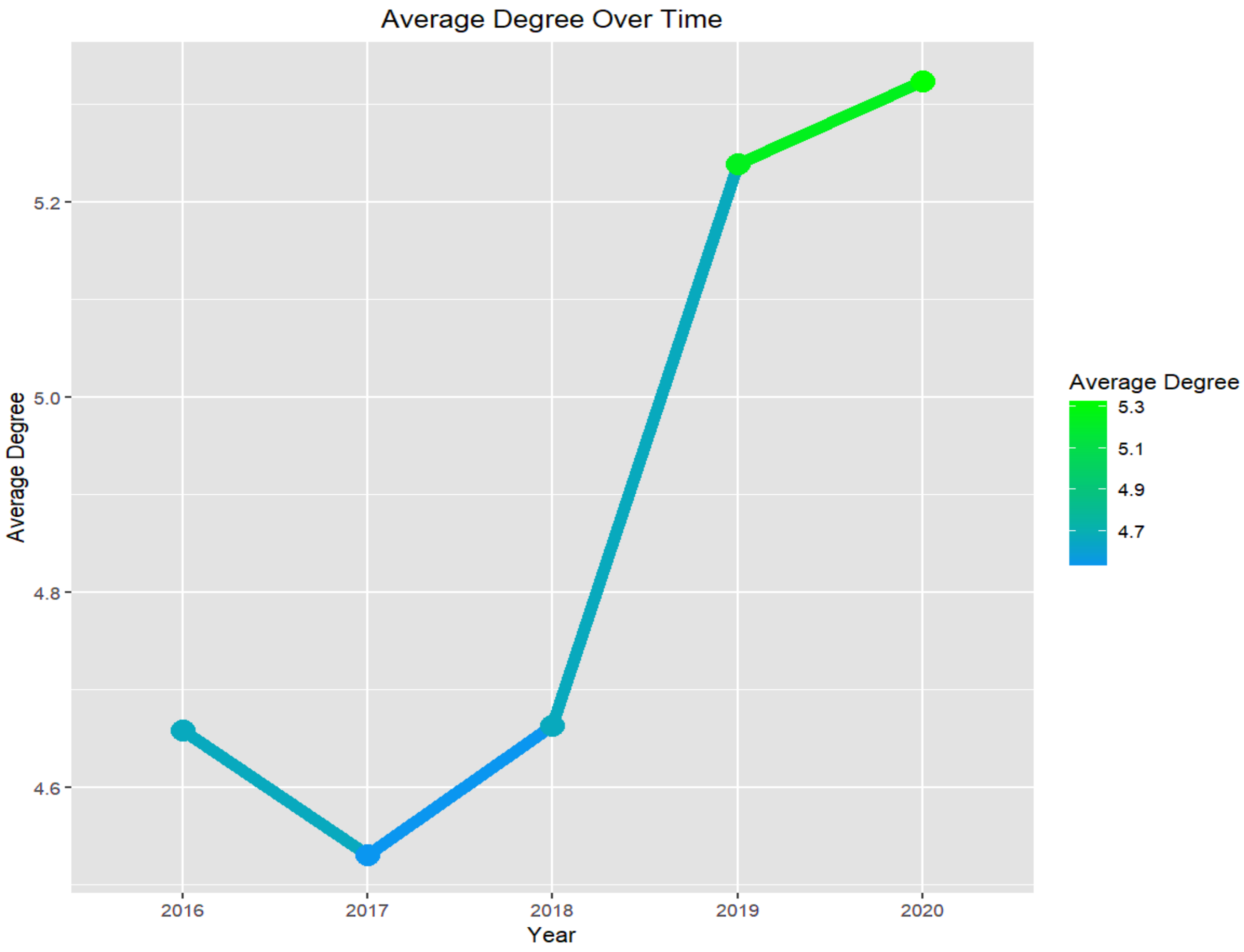


Figure 2-4 Average Degree Over Time

2.5 Insights

1. The number of Vertices and Edges over the years follow an uptrend
2. The Diameter has high fluctuations and the Average Degree small fluctuations
3. The latest years 2019 and 2020 we have significant more co-authors who cooperated on publishing papers. Specifically, on average that years cooperated more than 5 authors. That result combined with Edge and Vertices plot in which we saw an uptrend in authors who publish and cooperation, makes sense.
4. Overall, we have small cooperation between authors, and not big teams above 6 members.
5. Last but not least, although in the year 2018 we found the highest diameter the average degree was low, meaning the maximum distance is high.

3. Important Nodes

In this section we are searching for the top-10 authors regarding to their Degree (simple, not weighted) and PageRank.

3.1 Top 10 Authors based on their Degree

Year: 2016	
Author	Degree
1. Philip S. Yu	46
2. Jiawei Han 0001	41
3. Hui Xiong 0001	39
4. Naren Ramakrishnan	32
5. Jieping Ye	32
6. Yi Chang 0001	31
7. Jiebo Luo	29
8. Rayid Ghani	28
9. Chang-Tien Lu	25
10. Yannis Kotidis	25

Year: 2017	
Author	Degree
1. Philip S. Yu	44
2. Jiawei Han 0001	42
3. Hui Xiong 0001	38
4. Yi Chang 0001	32
5. Claudio Rossi 0003	32
6. Heng-Tze Cheng	31
7. Zakaria Haque	31
8. Mustafa Ispir	31
9. Clemens Mewald	31
10. Martin Wicke	31

Year: 2018	
Author	Degree
1. Philip S. Yu	70
2. Jiawei Han 0001	37
3. Kun Gai	35
4. Wenwu Zhu 0001	28
5. Jing Gao 0004	27
6. Chao Zhang 0014	27
7. Jure Leskovec	27
8. Xing Xie 0001	26
9. Qi Liu 0003	25
10. Enhong Chen	25

Year: 2019	
Author	Degree
1. Philip S. Yu	69
2. Weinan Zhang 0001	59
3. Hui Xiong 0001	49
4. Jieping Ye	41
5. Jie Tang 0001	39
6. Jiawei Han 0001	37
7. Yong Li 0008	36
8. Enhong Chen	36
9. Jingren Zhou	35
10. Jian Pei	35

Year: 2020	
Author	Degree
1. Jiawei Han 0001	69
2. Hongxia Yang	43
3. Hui Xiong 0001	42
4. Xiuqiang He	41
5. Ji Zhang	40
6. Peng Cui 0001	39
7. Christos Faloutsos	38
8. Wei Wang 0010	38
9. Jieping Ye	37
10. Ruiming Tang	35

Table 3.1 Tables with Authors in Descending Order based on their Degree

Network16	Network17	Network18	Network19	Network20
Philip S. Yu	Philip S. Yu	Philip S. Yu	Philip S. Yu	Jiawei Han 0001
Jiawei Han 0001	Jiawei Han 0001	Jiawei Han 0001	Weinan Zhang 0001	Hongxia Yang
Hui Xiong 0001	Hui Xiong 0001	Kun Gai	Hui Xiong 0001	Hui Xiong 0001
Naren Ramakrishnan	Yi Chang 0001	Wenwu Zhu 0001	Jieping Ye	Xiuqiang He
Jieping Ye	Claudio Rossi 0003	Jing Gao 0004	Jie Tang 0001	Ji Zhang
Yi Chang 0001	Heng-Tze Cheng	Chao Zhang 0014	Jiawei Han 0001	Peng Cui 0001
Jiebo Luo	Zakaria Haque	Jure Leskovec	Yong Li 0008	Christos Faloutsos
Rayid Ghani	Mustafa Ispir	Xing Xie 0001	Enhong Chen	Wei Wang 0010
Chang-Tien Lu	Clemens Mewald	Qi Liu 0003	Jingren Zhou	Jieping Ye
Yannis Kotidis	Martin Wicke	Enhong Chen	Jian Pei	Ruiming Tang

Table 3.2 Table with all Author Names ranked by their Degree

3.2 Insights for Degree Metric

In the majority of the years “Philip S. Yu” was the top author, except the year 2020 in which “Jiawei Han 0001” was the top author. But still the degree of 69 is lower than the 70 which is the highest from all the years. That lead us to the conclusion that if a group of authors is cooperating one year, there are many chances to continue cooperating the next years. The minimum degree is 25 in the years 2016 and 2018, and the year with the lowest variance in the number of degree between the authors is 2017. In general we can see that some authors are cooperating one year with others and the next years they decrease the number of publications or they stop publishing. Therefore we have variation between the years, because we have not the same authors every year in the top 10.

3.3 Top 10 Authors per Year based on their PageRank Rate

Year: 2016	
Author	**PageRank**
1. Philip S. Yu	0.0017288334
2. Hui Xiong 0001	0.0014581015
3. Jiawei Han 0001	0.0014119510
4. Jiebo Luo	0.0013099364
5. Jieping Ye	0.0010027077
6. Yi Chang 0001	0.0009601005
7. Hanghang Tong	0.0009272920
8. Christos Faloutsos	0.0009216757
9. Maarten de Rijke	0.0009158533
10. Jiliang Tang	0.0009155034

Year: 2017	
Author	**PageRank**
1. Philip S. Yu	0.0014558956
2. Jiawei Han 0001	0.0013585699
3. Hui Xiong 0001	0.0010997688
4. Jure Leskovec	0.0010681579
5. Jiebo Luo	0.0009454158
6. Hanghang Tong	0.0009285808
7. Jiliang Tang	0.0007750644
8. Yi Chang 0001	0.0007711858
9. Chao Zhang 0014	0.0007510406
10. Ingmar Weber	0.0007208090

Year: 2018	
Author	**PageRank**
1. Philip S. Yu	0.0019809631
2. Jiawei Han 0001	0.0009301987
3. Jure Leskovec	0.0008753490
4. Wenwu Zhu 0001	0.0007842984
5. Chao Zhang 0014	0.0006775310
6. Xing Xie 0001	0.0006263373
7. Jing Gao 0004	0.0006259877
8. Martin Ester	0.0006201636
9. Yiqun Liu 0001	0.0006143691
10. Kun Gai	0.0006129884

Year: 2019	
Author	**PageRank**
1. Philip S. Yu	0.0015871036
2. Hui Xiong 0001	0.0009633261
3. Weinan Zhang 0001	0.0008767308
4. Jieping Ye	0.0007255196
5. Hanghang Tong	0.0007021244
6. Jiawei Han 0001	0.0006855583
7. Peng Cui 0001	0.0006574207
8. Jie Tang 0001	0.0006517701
9. Enhong Chen	0.0006377621
10. Gerhard Weikum	0.0006257373

Year: 2020	
Author	**PageRank**
1. Jiawei Han 0001	0.0010753255
2. Hui Xiong 0001	0.0007594661
3. Hongxia Yang	0.0007284981
4. Elke A. Rundensteiner	0.0006983864
5. Yong Li 0008	0.0006821198
6. Jieping Ye	0.0006800497
7. Peng Cui 0001	0.0006533883
8. Xiuqiang He	0.0006465968
9. Ji-Rong Wen	0.0006450074
10. Jiliang Tang	0.0006423610

Table 3.3 Tables with Author Ranking Based on their PageRank Rate

Network16	Network17	Network18	Network19	Network20
Philip S. Yu	Philip S. Yu	Philip S. Yu	Philip S. Yu	Jiawei Han 0001
Hui Xiong 0001	Jiawei Han 0001	Jiawei Han 0001	Hui Xiong 0001	Hui Xiong 0001
Jiawei Han 0001	Hui Xiong 0001	Jure Leskovec	Weinan Zhang 0001	Hongxia Yang
Jiebo Luo	Jure Leskovec	Wenwu Zhu 0001	Jieping Ye	Elke A. Rundensteiner
Jieping Ye	Jiebo Luo	Chao Zhang 0014	Hanghang Tong	Yong Li 0008
Yi Chang 0001	Hanghang Tong	Xing Xie 0001	Jiawei Han 0001	Jieping Ye
Hanghang Tong	Jiliang Tang	Jing Gao 0004	Peng Cui 0001	Peng Cui 0001
Christos Faloutsos	Yi Chang 0001	Martin Ester	Jie Tang 0001	Xiuqiang He
Maarten de Rijke	Chao Zhang 0014	Yiqun Liu 0001	Enhong Chen	Ji-Rong Wen
Jiliang Tang	Ingmar Weber	Kun Gai	Gerhard Weikum	Jiliang Tang

Table 3.4 Table with all Author Names based on PageRank rate

3.4 Insights for Degree Metric

In terms of PageRank we can also draw some useful insights. We can see the same pattern in the first places like Degree metric (“Philip S. Yu”, “Jiawei Han 0001”) . Therefore, they are important authors with many connections. Also there are many other authors that each year are repeated like “Wenwu Zhu 0001” and “Jieping Ye”. So there is not huge variance between all years.

4. Communities

In the last section, we implemented community detection. The concept of community detection has emerged in network science as a method for finding groups within complex systems through represented on a graph.[2]

4.1 Applying Different Algorithms

The algorithms that we used are fast greedy clustering, infomap clustering, and louvain clustering on the 5 undirected co-authorship graphs. Therefore, we proceeded by applying all algorithms to all networks and compare their performance. In general if a method gives fast result its preferable than a slower method, because the second in a big dataset may not give results or be more slower.

Algorithm	Seconds (all Networks)
Fast Greedy	0.743361
Infomap	14.9883
Louvain	0.2808211

Table 4.1 Performance of Each Algorithm to Complete the Clustering

From the above table we can see that Louvain method has the highest performance followed by Greedy algorithm. Therefore Louvain method is more preferable due to its time efficiency. Also, we got results through all the methods.

Further, we check the modularity of each method. Modularity basically, tells us the strength of division of network into communities.

Year	Fast_Greedy	Infomap	Louvain
2016	0.98	0.961	**0.98**
2017	0.984	0.967	**0.986**
2018	0.982	0.961	**0.984**
2019	0.971	0.94	**0.975**
2020	0.962	0.933	**0.97**

Table 4.2 Modularity Scores of Each Algorithm

[2] https://www.sciencedirect.com/topics/computer-science/community-detection

Again we found Louvain as the best community detection algorithms. Although, there is a small difference in the modularity scores, but the Louvain produced highest modularity in all the years, while, Info Map could produce lowest modularity then other. In addition, we also compared the methods to see how similar communities they are detecting. Overall, we found that Louvain and Fast Greedy methods are detecting most of the communities same. The plot below, confirms our findings.

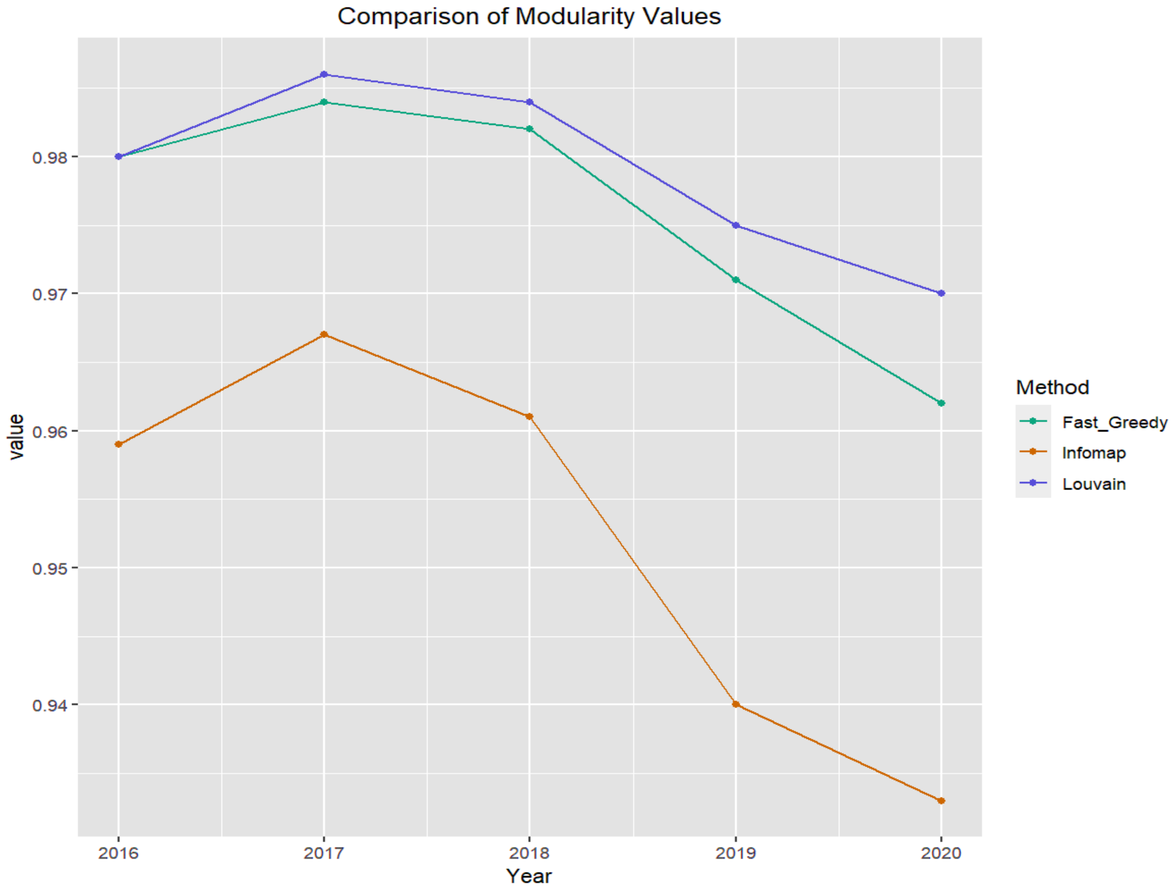


Figure 4-1 Comparison of Modularity Values of Each Method

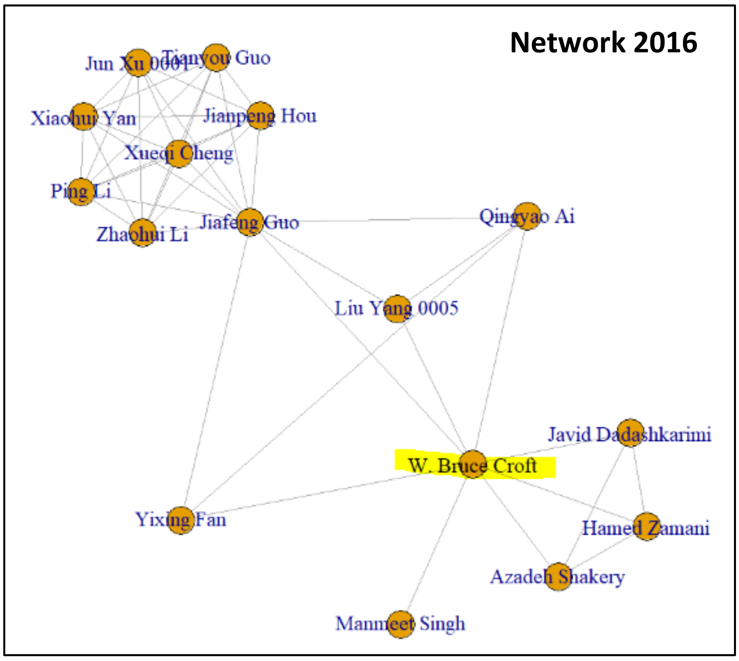
Last but not least, to find the similarity between algorithms, we have used compare() method. Which will tell us the difference between structure of communities detected by the algorithms. Further, we will subtract the difference from 1 to get the similarity i.e., how the similar two algorithms are in community detection. Here similarity means, either the two algorithms are detecting communities in similar way or not. Where, 1 means all the detected communities by both algorithms are same, however, 0 means all the communities are different.

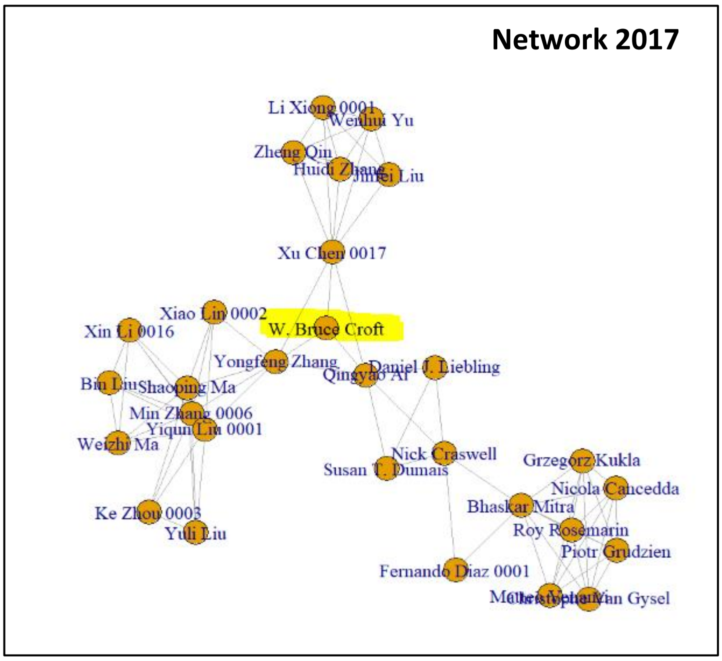
Algorithm 1	Algorithm 2	Year 2016	Year 2017	Year 2018	Year 2019	Year 2020
InfoMap	Clouvain	0.6359537	0.6533883	0.6223782	0.4083424	0.2891448
InfoMap	Fast Greedy	0.6153856	0.6126046	0.5321753	0.2865275	0.2009223
Louvain	Fast Greedy	0.8263803	0.8715111	0.7878936	0.7353485	0.5345587

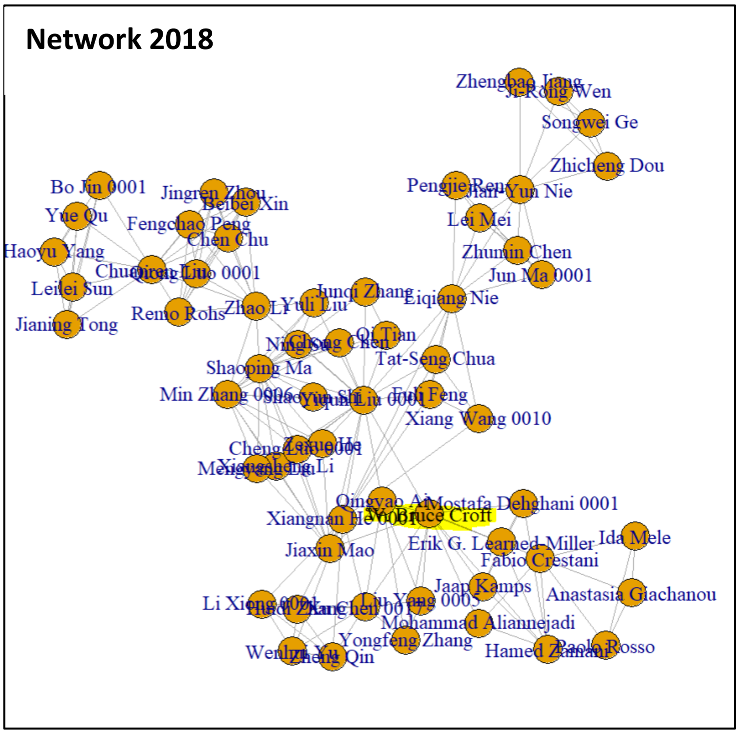
Table 4.3 Results using Compare() function

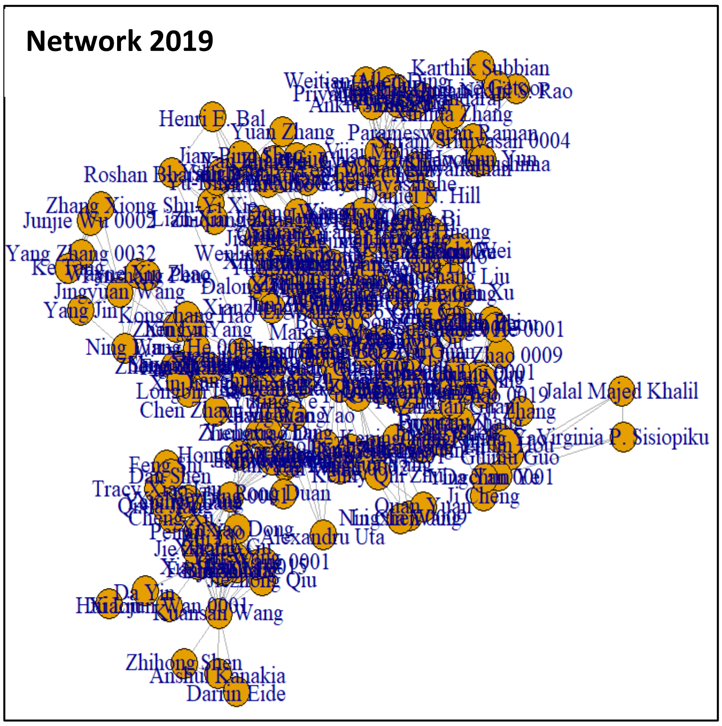
4.2 Picking a Random Author

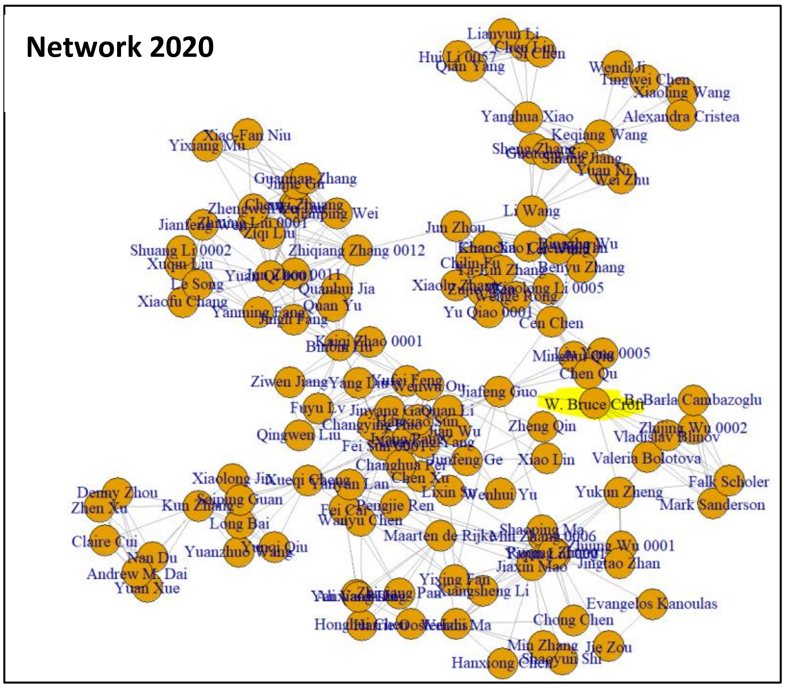
We proceed by picking a random author with sample() function which every time will get a different sample. For our example we take "**W. Bruce Croft**" author that appears in all 5 graphs combined with Louvain method.


Network 2016


Network 2017


Network 2018


Network 2019



As we are picking random numbers, every time, we observe that the node is evolving their circle. Means, in the start, it belongs to small community, then as years goes, the node moves to the large community. Besides, it appears that as years goes the node changes its neighbors too i.e., we found one or 2 neighbors in 2017, but not one stays till 2020. As we can see that there is no common neighbor till 2020, except the random node.

For communities similarity, we have used Jaccard index. The Jaccard index produces the normalize similarity score, where 1 means the communities are 100% similar, while, 0 means the communities have no similarity. The figure shows the similarity between communities by colors, where blue color shows the highest similarity & red color shows the highest dissimilarity.

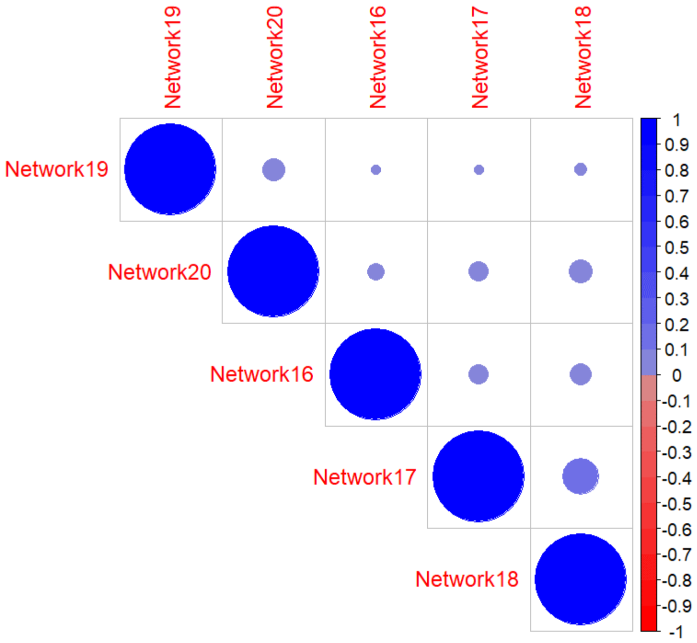
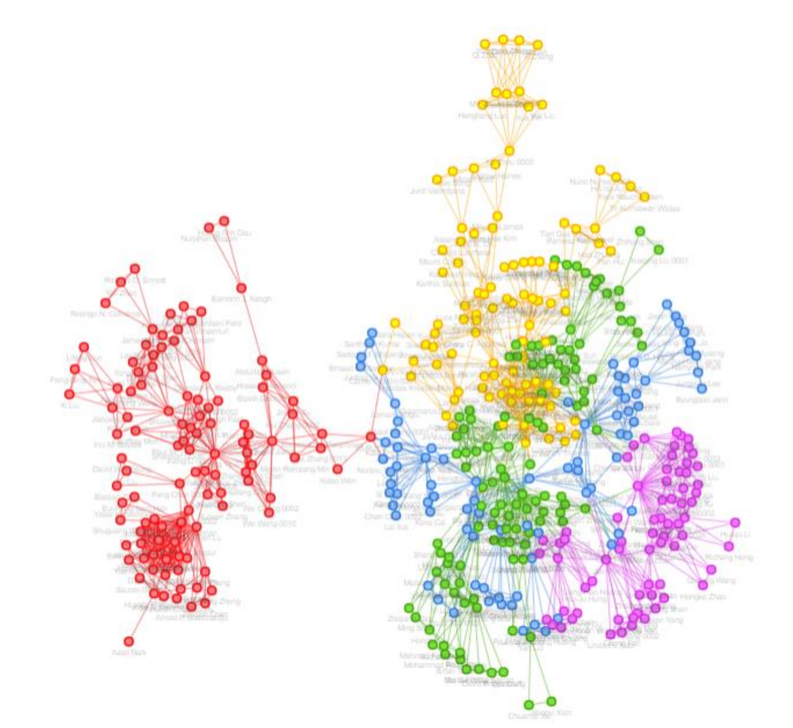


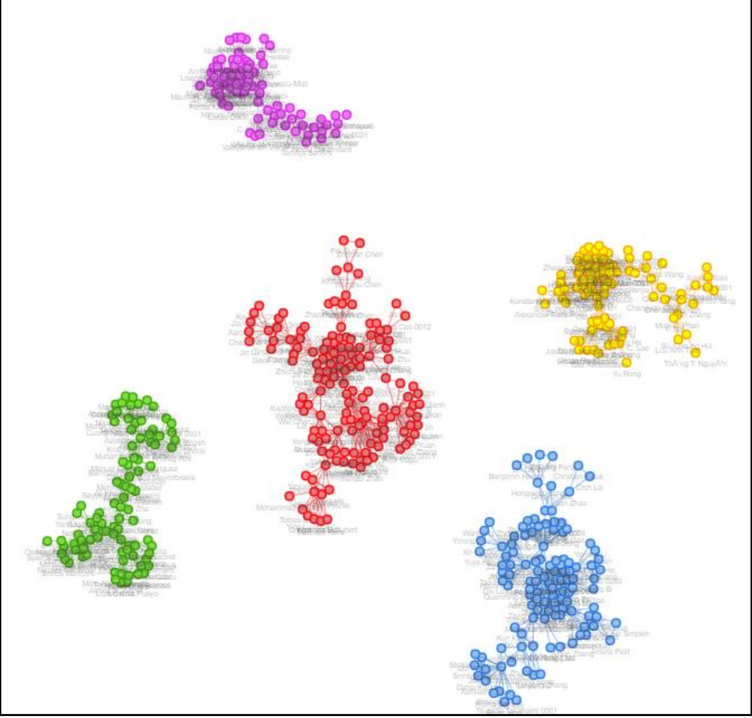
Figure 4-2 Corplot of how many percent user are same in the community.

In the results, it appears that the communities from network 17 and 18 have more similarity then rest of the communities. Which shows that the random author changes its communities as year goes. The results of the communities seem logical since the vast majority of the authors in the large communities have attended at the same conferences ('WWW', 'CIKM'). Also, they were the same authors the same years (2016, 2017, 2018) but this fact is in a reliable metric.

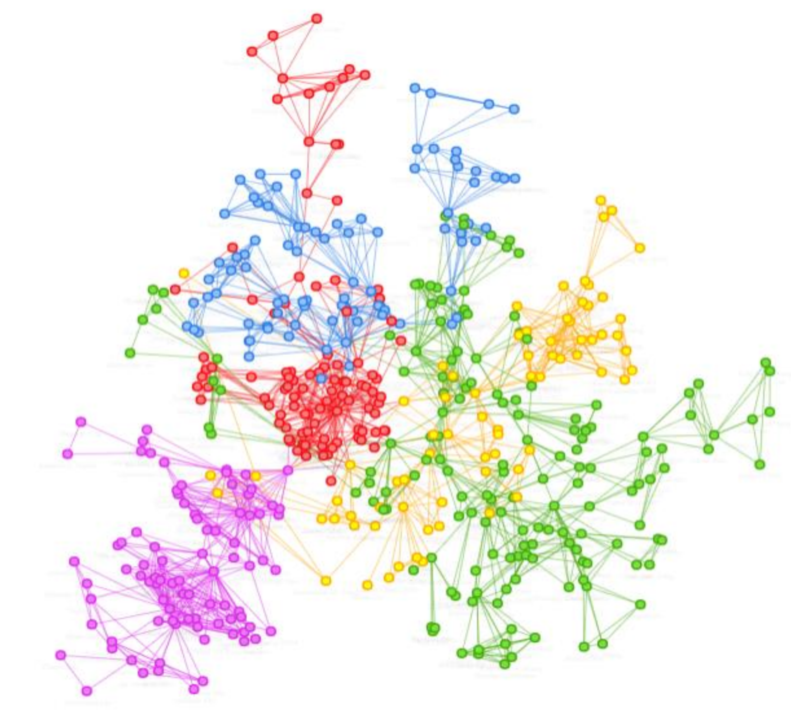
Last but not least, we proceed by visualizing the communities. Although we experimented with different community sizes, we have chosen only to present the top 5 communities in order to achieve an aesthetically pleasing result.



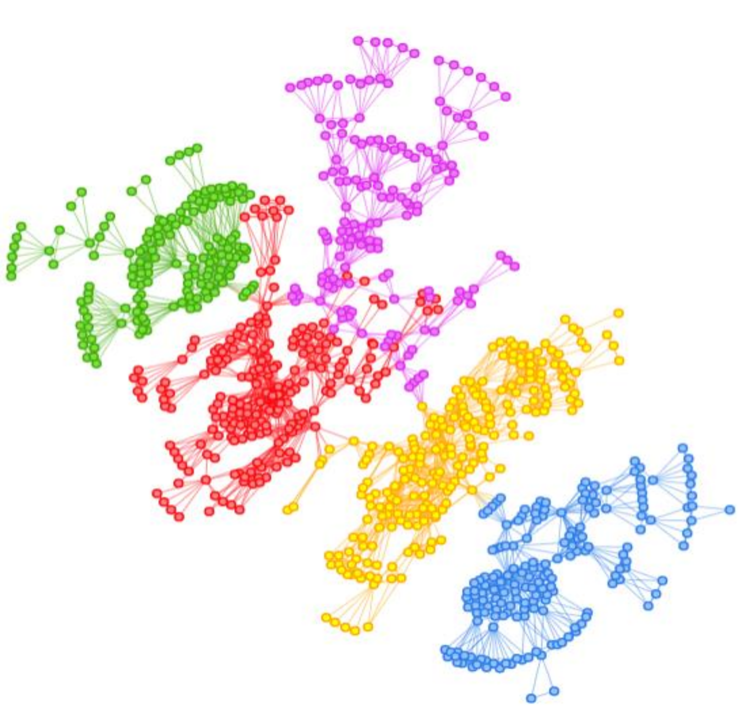
Communities in Network 2016



Communities in Network 2017



Communities in Network 2018



Communities in Network 2019

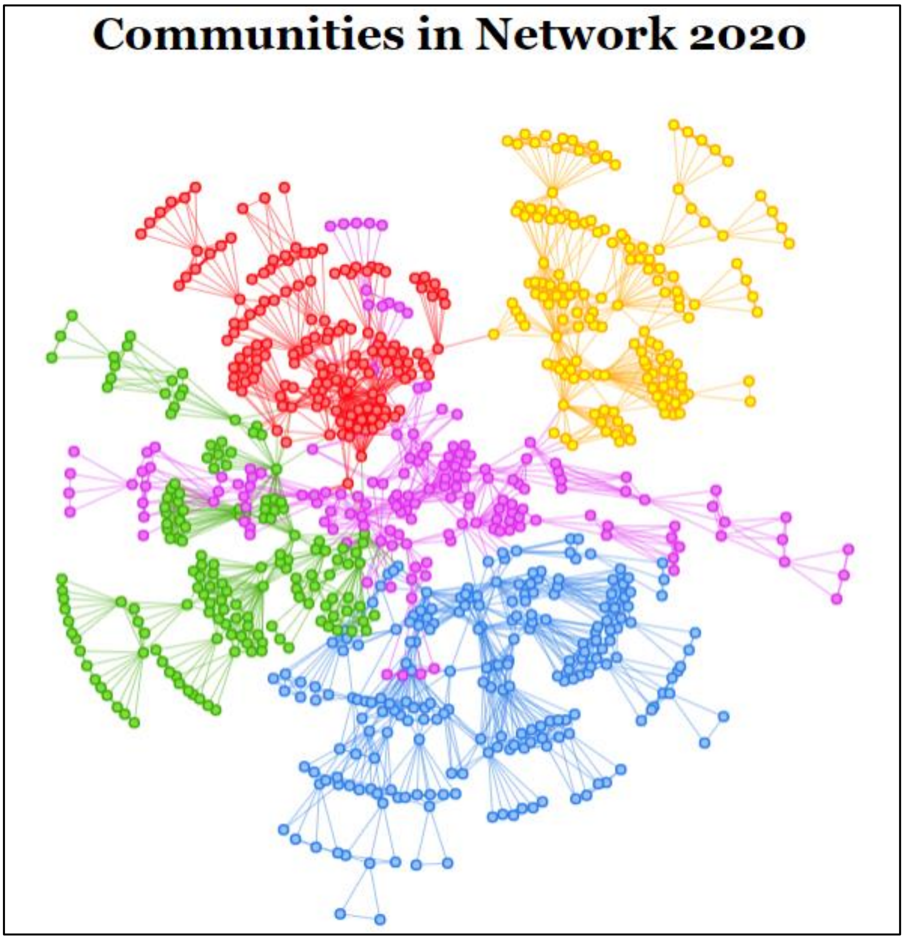


Figure 4-3 Plots for Top 5 Communities per Network

4.3 Conclusions

- We presented the top 5 communities
- We created a different plot for every community
- Every community is presented with different color
- In the majority of the plots the communities are clear to understand where they belong
- The connections between them are also clear

5. Conclusions

Conducting the above analysis with different tests and plots we end up with the final communities. The methodology followed was to first undergo data manipulation and transformation in order to achieve the highest possible accuracy in the results later in our analysis. Second, we conducted exploratory data analysis by checking the number of vertices, edges, diameter of the graph, degree and PageRank. That metrics combined with plots allowed us to have a deeper look into our data relationships and patterns. Third, we performed community detection with Fast Greedy, Infomap, and Louvain clustering followed by performance analysis. Fourth, we created meaningful and aesthetically pleasing visualizations of the communities. Some useful insights conducting the above analysis are:

- The social network analysis requires special attention in data manipulation as it processes raw data and transformations into usable forms for analysis
- A basic analysis with metrics is always needed to find instantly inconsistences in the data and fluctuations between networks
- Algorithm performance plays crucial role in analysis with huge amount of data
- Meaningful and aesthetically pleasing visualizations are important to compare the communities