SCHOOL OF BUSINESS

DEPARTMENT OF MANAGEMENT SCIENCE & TECHNOLOGY

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ACADEMIC YEAR OF 2021 – 2022

SOCIAL NETWORK ANALYSIS – PROJECT II

NINAS KONSTANTINOS

f2822108

SUPERVISING INSTRUCTOR

KATIA PAPAKONSTANTINOPOULOU

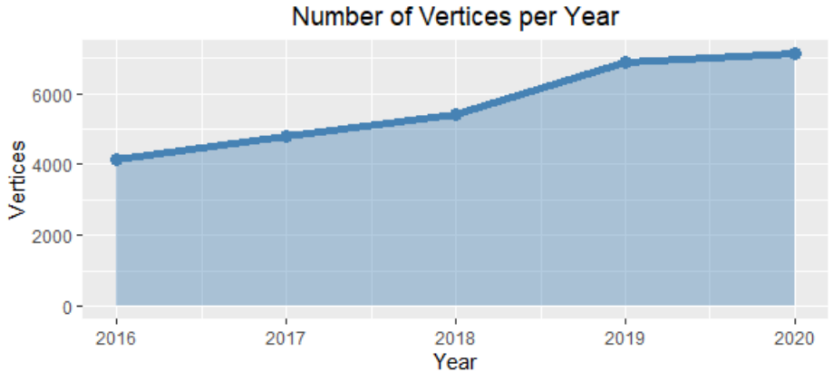
Table of Contents

1 – Co-authorship Graph Creation

The first task was the creation of a weighted undirected graph given the co-authorship data from dblp for the conferences CIKM, KDD, ICWSM, WWW and IEEE BigData in the last 5 years. Prior to that, the data should be converted to an 'edge list' format. The first step to that transformation was to filter out all the observations that were not associated with the selected conferences or were older than 5 years. Unix commands were used to resolve that issue. Specifically, the initial file was divided into 5 new (one for each conference), which then were united into one. The new file, which only contained the selected conferences, was again divided into 5 new, one for each year. Next, it was observed that some observations were not separated appropriately to their corresponding columns, and as a result, some observations (approximately 5) had more columns than the rest. This issue was solved with the assistance of Microsoft Excel, with which those observations were detected and fixed. Following, the divided files were transformed into edge list format with the assistance of Python scripts. In detail, all the combinations of authors were found along with their frequencies, which would be used as weights in the creation of the graph. This process was applied to all 5 files previously created. The files, at that point, had taken the form of an edge list, and were exported to csv format. The edge list files were imported in R, in order to create the graphs. Before the creation of the graphs, the edge lists were checked for the existence of multiple edges. Some were found, and removed from each edge list. Finally, the graphs were created.

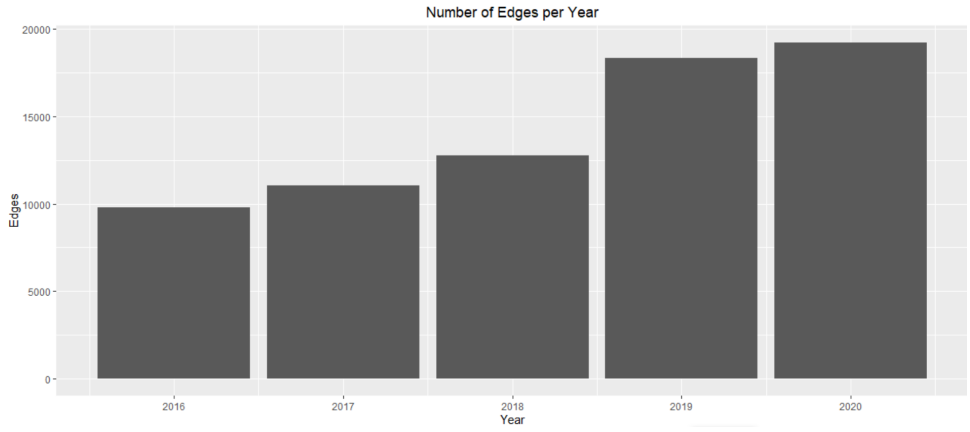
2 – Metrics Over Time

In the next task, we were required to identify the evolution of certain graph metrics over the years and comment on our findings. The first metric we were tasked to identify was the number of vertices in the graphs throughout the years (Plot 1). It is identified that the number of vertices is increasing each year. It is worth mentioning that the greatest increase was noted from 2018 to 2019. It can be inferred that the authors' community is growing over the years.



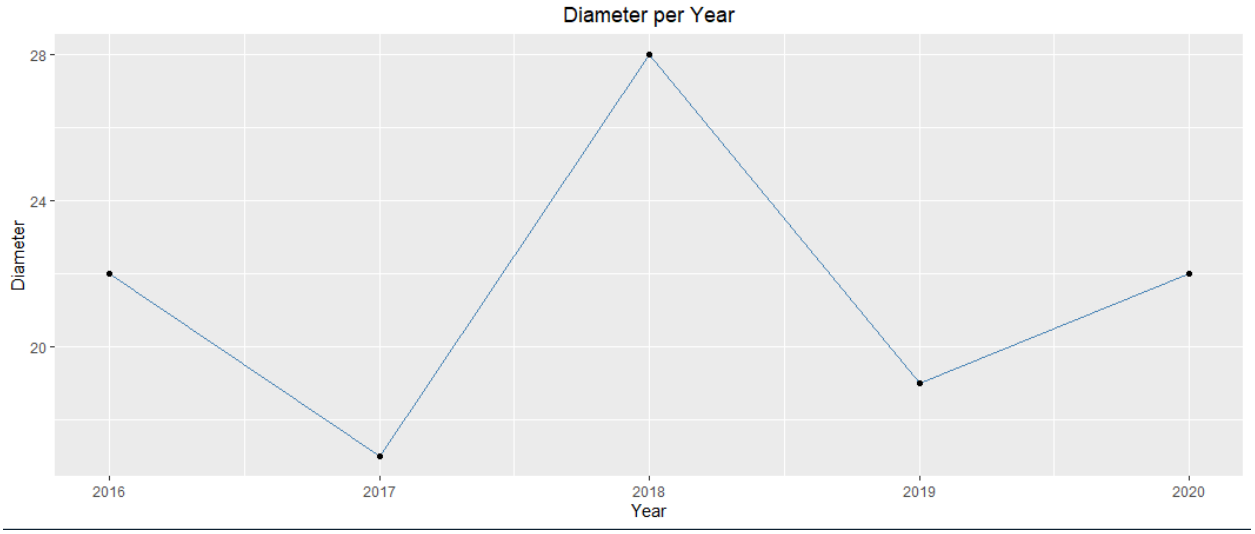
Plot 1. Number of vertices per year

The second metric we were tasked to identify was the number of edges per year (Plot 2). Likewise, the number of edges is increasing each year, with the greatest increase from 2018 to 2019. Given that metric alone, it can be deduced that authors are connecting more with each other throughout the years. It is also a direct result from the fact that the graph grows since more and more authors become a part of it every year.



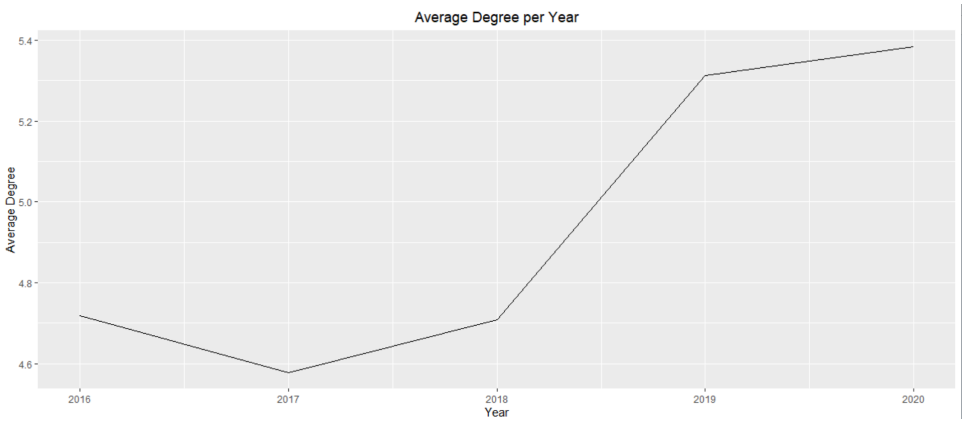
Plot 2. Number of edges per year

The third metric we were tasked to identify was the diameter of the graphs throughout the years (Plot 3). It is noticed that in most years there are fluctuations in the diameter of the graphs (with the exception of the years from 2019 to 2020). Which means that the distance between the two most distant authors was not stabilized throughout the years.



Plot 3. Diameter per year

The final metric we were tasked to identify was the average degree (number of connections) between the authors in the passing of years (Plot 4). It is noticed that the average degree does not change significantly over the years. In other words, the authors are actually not getting significantly more connected with each other, even though more and more of them are present over the years.



Plot 4. Average Degree per Year

3 – Identification of Important Authors

Following, we were tasked to detect the 10 most important authors over the years according to their degree (number of connections) (Plot 5) and their pagerank scores (Plot 6). For the 10 most important authors according to both their degree and their pagerank scores, it can be observed that there is great variation over the years. In detail, the 3 most important authors of 2016 can be observed in most of the rest of the succeeding years. The opposite can be inferred for the rest of the authors in any year, most of which do not appear more than once throughout the years. Additionally, it is worth mentioning that the 4 most important authors according to their degree score are quite similar with those that occur when considering their pagerank scores each year. Which means that, in some sense, the authors that have the most co-authorships (greatest degree) each year are also the most important ones according to the pagerank metric.

	top_2016	top_2017	top_2018	top_2019	top_2020
1	Philip S. Yu	Philip S. Yu	Philip S. Yu	Philip S. Yu	Jiawei Han 0001
2	Jiawei Han 0001	Jiawei Han 0001	Jiawei Han 0001	Weinan Zhang 0001	Hongxia Yang
3	Hui Xiong 0001	Hui Xiong 0001	Kun Gai	Hui Xiong 0001	Hui Xiong 0001
4	Jieping Ye	Claudio Rossi 0003	Wenwu Zhu 0001	Jieping Ye	Xiuqiang He
5	Naren Ramakrishnan	Yi Chang 0001	Chao Zhang 0014	Jie Tang 0001	Ji Zhang
6	Yi Chang 0001	Clemens Mewald	Jing Gao 0004	Jiawei Han 0001	Peng Cui 0001
7	Jiebo Luo	Heng-Tze Cheng	Jure Leskovec	Enhong Chen	Christos Faloutsos
8	Rayid Ghani	Martin Wicke	Xing Xie 0001	Yong Li 0008	Wei Wang 0010
9	Chang-Tien Lu	Mustafa Ispir	Enhong Chen	Jian Pei	Jieping Ye
10	Yannis Kotidis	Zakaria Haque	Haifeng Chen	Jingren Zhou	Jiliang Tang

Plot 5. Top 10 authors per year based on degree

	top_2016	top_2017	top_2018	top_2019	top_2020
1	Philip S. Yu	Philip S. Yu	Philip S. Yu	Philip S. Yu	Jiawei Han 0001
2	Jiawei Han 0001	Jiawei Han 0001	Jure Leskovec	Hui Xiong 0001	Hongxia Yang
3	Hui Xiong 0001	Jure Leskovec	Jiawei Han 0001	Weinan Zhang 0001	Hui Xiong 0001
4	Jiebo Luo	Hui Xiong 0001	Wenwu Zhu 0001	Jiawei Han 0001	Jieping Ye
5	Jieping Ye	Jiebo Luo	Jieping Ye	Hanghang Tong	Elke A. Rundensteiner
6	Christos Faloutsos	Hanghang Tong	Chao Zhang 0014	Gerhard Weikum	Ji Zhang
7	Hanghang Tong	Yi Chang 0001	Xing Xie 0001	Jie Tang 0001	Peng Cui 0001
8	Maarten de Rijke	Ingmar Weber	Kun Gai	Jieping Ye	Yong Li 0008
9	Yi Chang 0001	Jiliang Tang	Robert West 0001	Jian Pei	Jiliang Tang
10	Rayid Ghani	Yizhou Sun	Meng Jiang 0001	Peng Cui 0001	Xiuqiang He

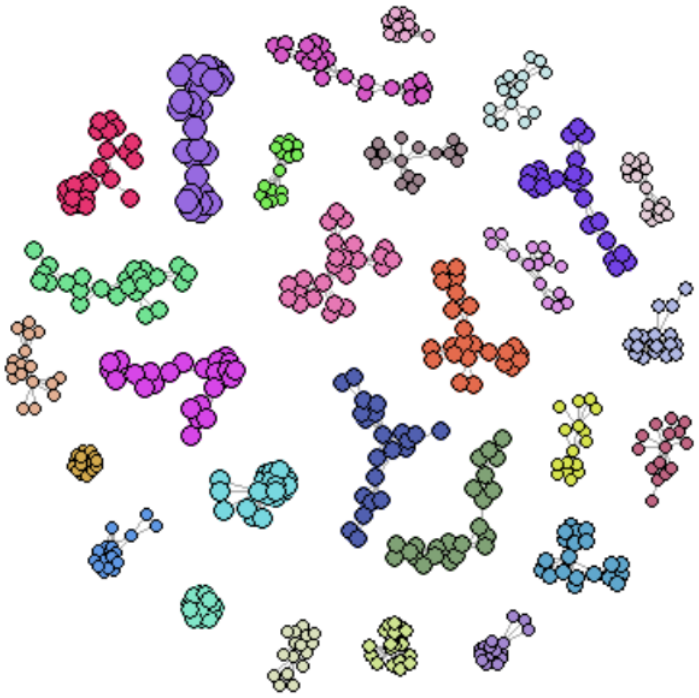
Plot 6. Top 10 authors per year base on Pagerank

4 – Communities of Authors

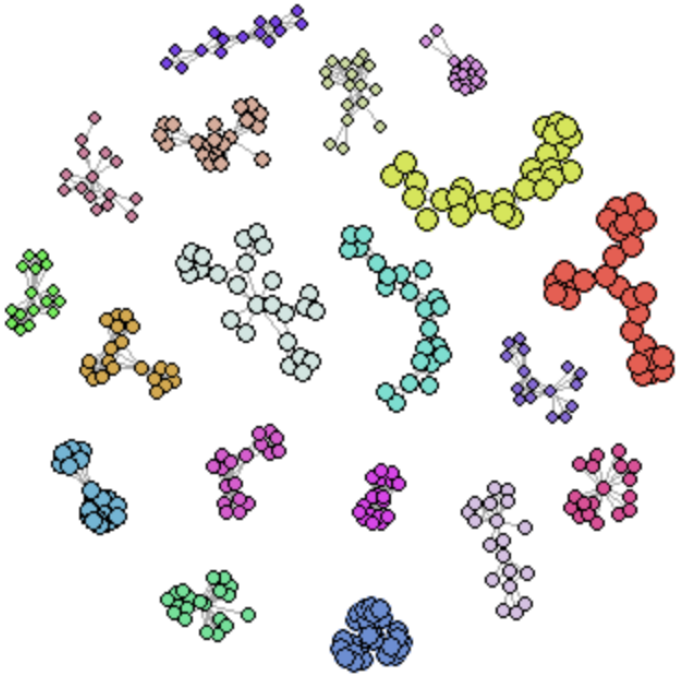
The final task was to create communities in each graph with multiple community detection methods, to identify their performance and to conduct further analysis on the communities created. The first method was the fast greedy clustering method, followed by the infomap clustering and the louvain clustering methods. Each one of those methods was able to detect communities in all graphs. Regarding their performance, the infomap clustering method was the slowest in detecting communities and detected way more communities in each graph than any of the other community detection methods. On the other hand, fast greedy and louvain clustering methods were both significantly faster than the infomap method in their community detection and identified almost the same number of communities in each graph.

Following, we were asked to identify a random author that exists in all 5 graphs and identify the evolution of the communities she belongs to over the years. Specifically, we examined the author "Ruizhe Ma". The first thing that was perceived was that the size of the communities she belongs to is similar throughout the years (approximately equal to 10), with the exception of the year 2018, in which her community shrunk significantly. Additionally, it is worth mentioning that most of the authors in her communities appear more than once, which means that she is collaborating quite often with them.

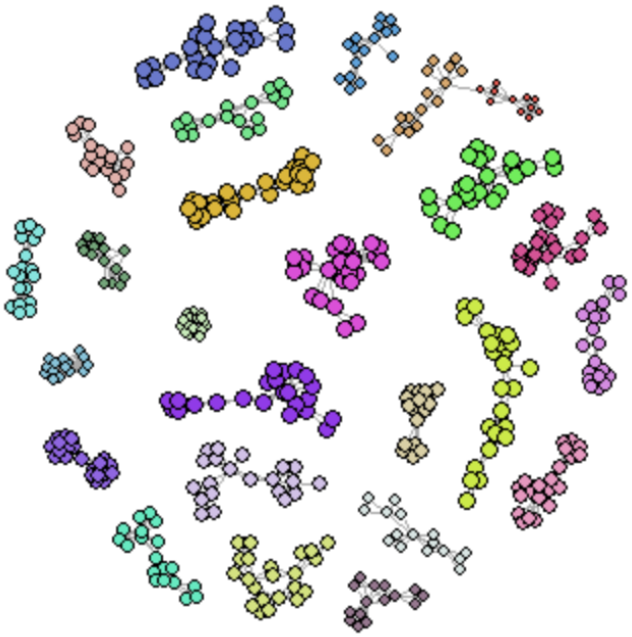
Concluding, we were asked to plot all the communities of each graph. To create more meaningful plots of the communities, we disregarded too small (with less than 15 authors) and too big communities (with more than 30 authors) in order to produce more meaningful and aesthetically pleasing representations of the graphs. In addition, we assigned a color to each author according to their community to make them more distinct and a size equivalent to the size of the community they belong to, in order to make bigger communities more obvious than the rest. The results of the communities for each year can be observed in plots 7, 8, 9, 10 and 11.



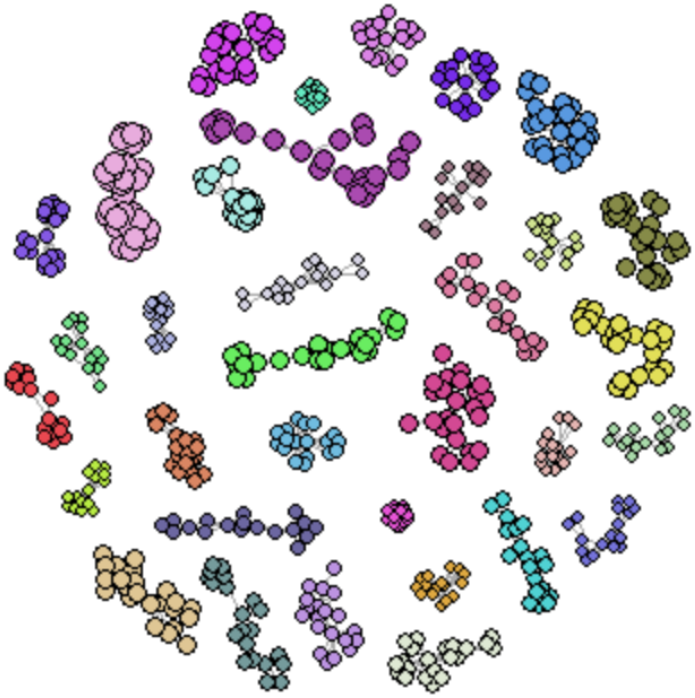
Plot 7. Communities in 2016



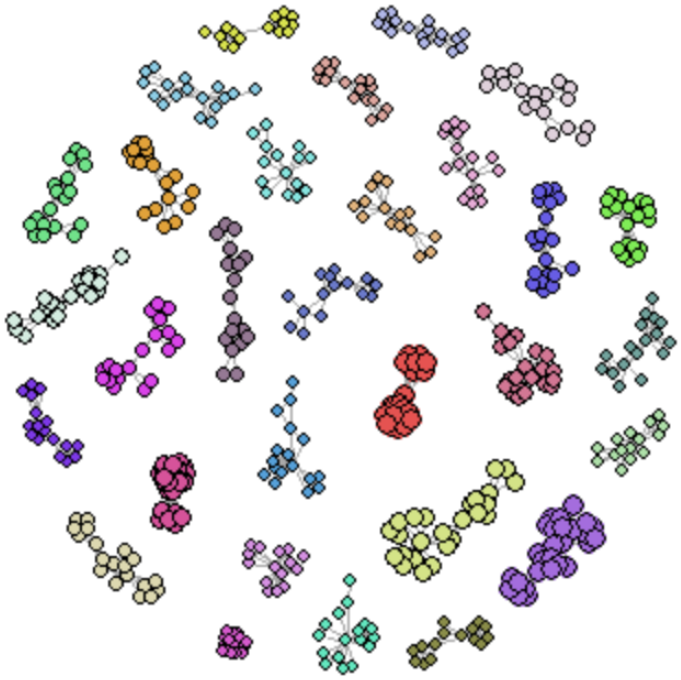
Plot 8. Communities 2017



Plot 9. Communities 2018



Plot 10. Communities 2019



Plot 11. Communities 2020