Scientific Modeling Computer Laboratory

Project: Time Evolving Networks

Midterm Report

Ádám Gergely Szabó

28th of March, 2022

1 Introduction

This project is about exploring MTMT's co-partnership network, which evolves in time as more publications get submitted to the site. MTMT's main goal is to host a site that maintains high quality publications, meaning the submitted works are often checked and rated for their quality. The data stored by the site is publicly available, thus data can be gathered from the site without registration.

This project's goal is to explore the co-partnership network of MTMT. In this work, we will look at this network in different given states or its subsets, see how it develops in time, calculate different central indicators, apply different group searching methods and embeddings. The work will be mostly done in Python3 language that will be utilized in the Jupyter Notebook framework.

2 Progress

2.1 Final Dataset

After further exploring how MTMT's ReST API accept queries, it turned out that the 5000 publication limit can be surpassed: the limitation of each response is present, but with the correct ordering, the full data can be obtained in these 5000 publication batches. The final data describes publications published between 2014 and today and narrowed down to one institute: the Eötvös Loránd University.

2.2 Self-loops in the Graph Representations

First things first, it was visible in the previous figures that presented the network that there are self-loops, edges whose start and end points are matching, which was indicated by the Networkx's spiral layout. After further investigation, it turned out that there are wrongly filled fields, namely multiple authors with the same are displayed multiple times in one publication. There were not so many, but they were visible: around 200 edges were removed due to this issue.

2.3 Lowering the Cut-off

Lowering the cut-off was necessary for further progress. Originally, publications with over 200 authors were ignored, now the lowered cut-off is 20, which is more sensible, much faster to process. With this we may look at the new, polished representation:

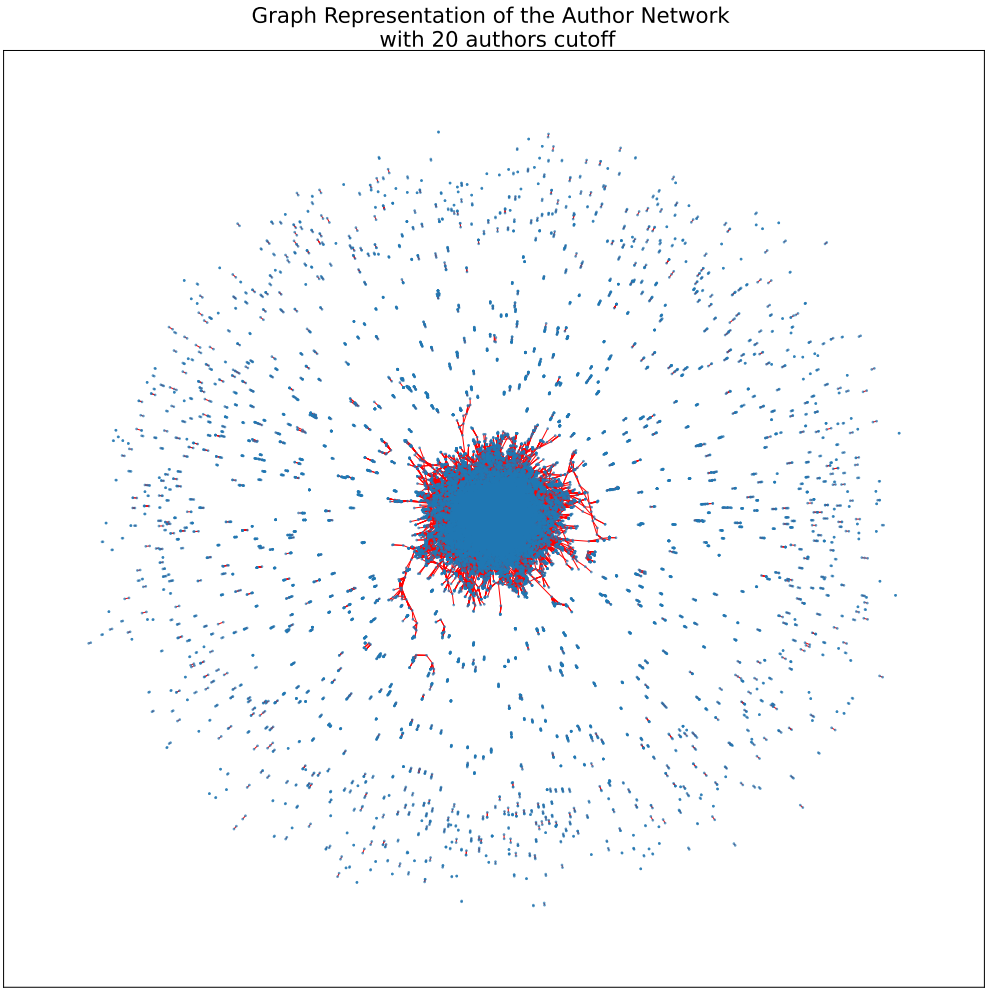

Figure 1: Graph representation of data from final data set. The giant component in the middle is there due to the spring layout smart placement. Smaller components surround the giant one and they are not connected according the data.

With this we may look at the new degree distribution too! The degree distribution may give us some extra information why there are that many nodes at the center of the representation (1). An interesting conflict in the degree distribution is the following: the first one (2) is coming from purely looking at the edge list obtained from the data, thus it is weighted. The second one (3) is the unweighted edge list, given back by Networkx. Both have a meaning: in the first case, two authors working together is counted in, a stronger connection is formed. In the second case, we are only interested

in the establishment of the connection. In the following, I will only use the unweighted results as I am interested in the existence of the connections, not their strength.

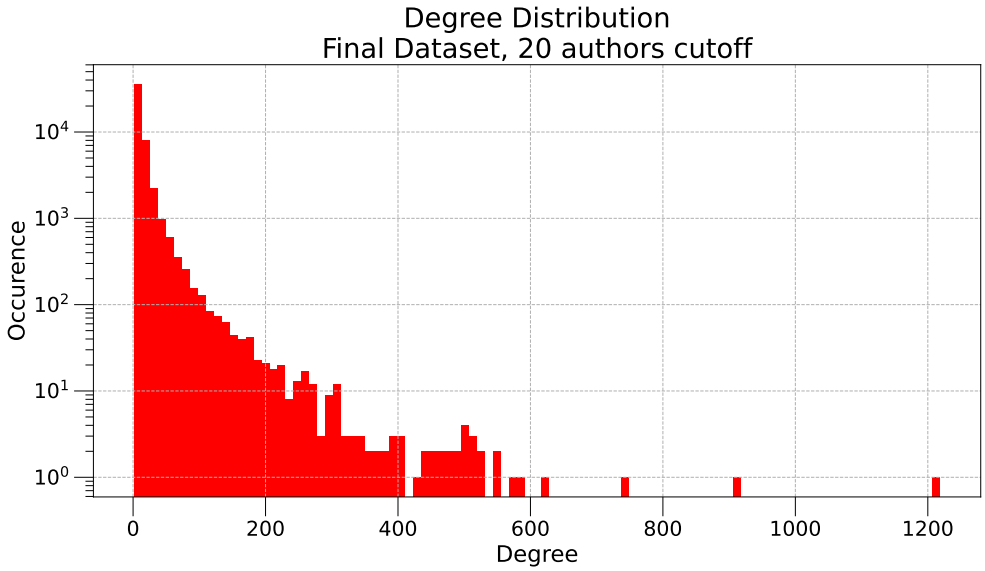

Figure 2: Degree Distribution with allowing more than one connection between nodes

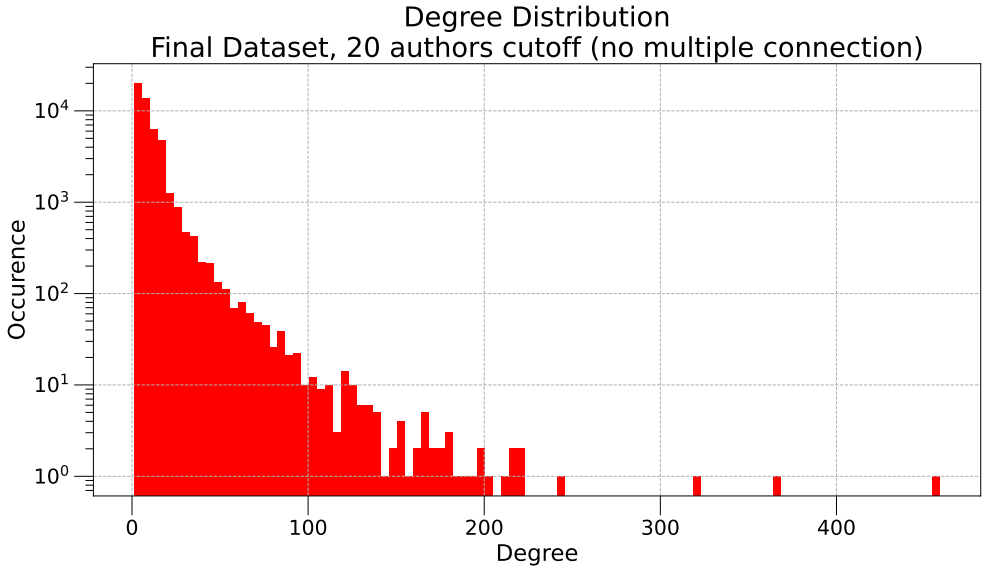

Figure 3: Degree distribution from the unweighted network given by Networkx

Why did the highest degree get lower to its one third? The reason may be that people like to work together who already worked together, so a strengthened bond is presented by the degree distribution that counts multiple connections between nodes.

2.4 Average Shortest Path

It may be interesting to look into the average shortest paths distribution to see how much randomness is present in the network or how much it is a lattice like network.

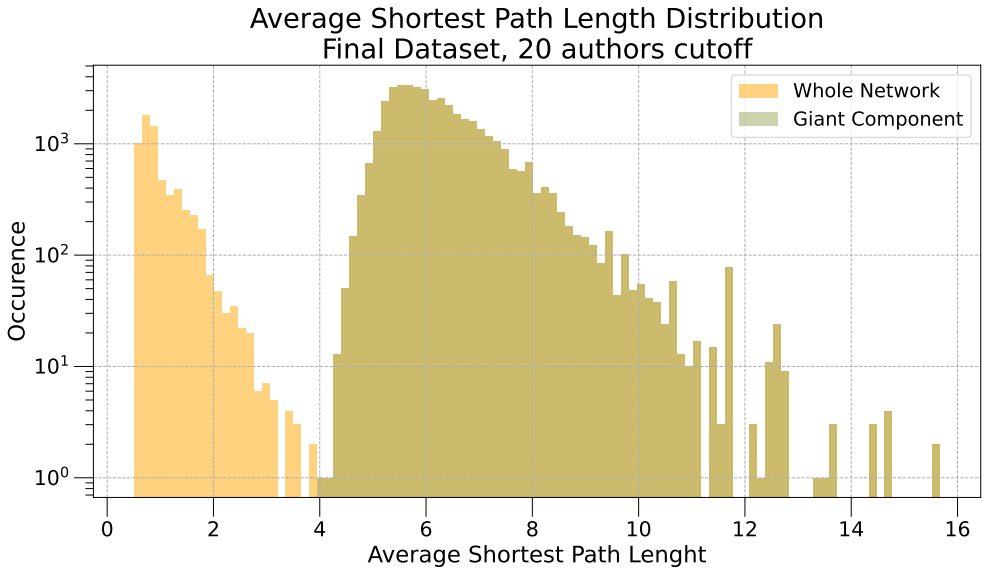

Figure 4: Distribution of the average shortest path lengths. In this case, for the giant component it is being calculated too. The reason is that this component is huge and acts as whole, while it contains ≈43000 nodes, which is a huge portion of the network.

What we can see here is that the average shortest paths lengths almost reach 16 in size. It is somewhat advised to look into the edge cases: the lower values from the giant component are nodes that can reach all the other nodes in relatively low amount of steps or hops, meaning that these nodes are at the center of the giant component. The higher values are nodes that are in the giant component, but at the end of a chain: these have lower degree and or most likely connected to other low degree nodes. In the end, the path connecting the farthest points of the nodes, 26 steps are needed. This could be our first clue that the network is disassortative.

2.5 Components in the Network

While the analysis using only the giant component is good as it represents the major part of the network, we may look into the smaller components and their structure too! (5),(6)

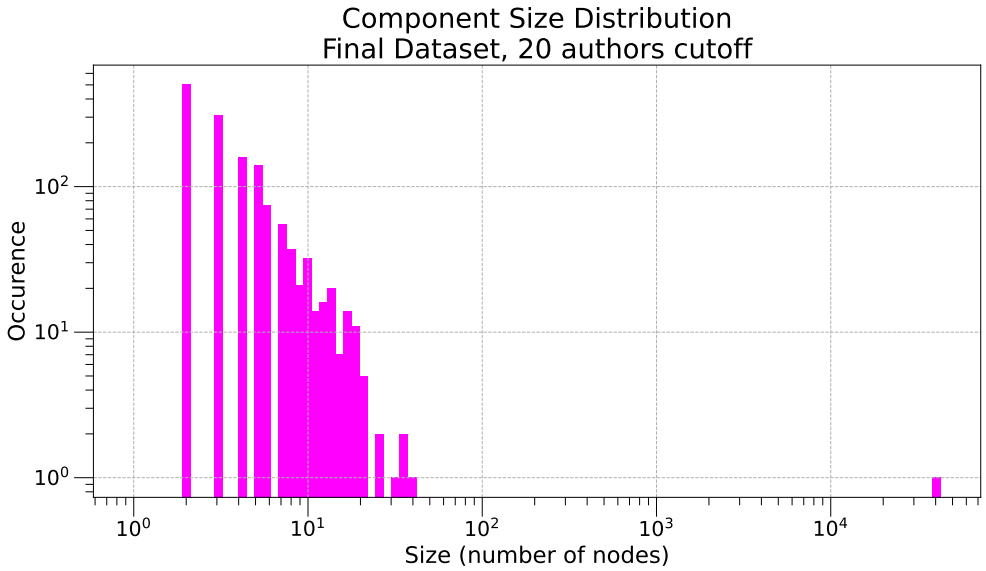

Figure 5: Component size distribution. The network actually consists of 1420 components, most of the have no more than 37 nodes, and at the other end of the spectrum, we have to giant component.

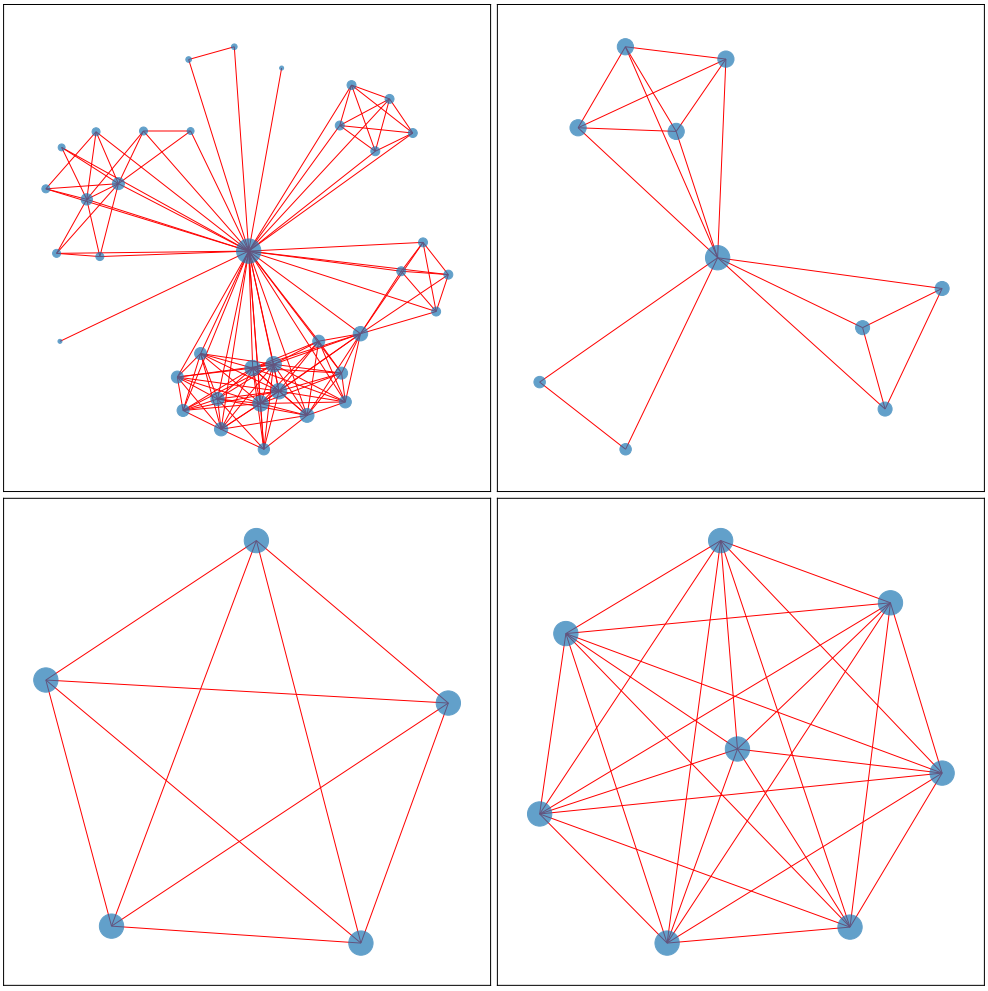

Figure 6: Component graph representation. It is visible that these are different: the top left and right represent more than 2 groups connected together by one node, while the bottom left and right represent groups that are closed and thus coming from one publication.

2.6 Degree Correlations and Assortativity

Previously, the networks assortativity was brought up (4): the co-author network may be disassortative. We can look into this by the means of taking the nodes neighbors degree and averaging them by the choosen nodes degree.

$$k_{nn}(k_i) = \frac{1}{k_i} \sum_{j=1}^{N} A_{ij} k_j \tag{1}$$

where A_{ij} is the adjacency matrix, which describes which node is connected to which and k_j, k_i is the degree of nodes. With this can see that the neutral networks have a constant, disassortative networks have negative, assortative networks have a positive exponent when a power-law function is fitted to the k_{nn} results. (7)

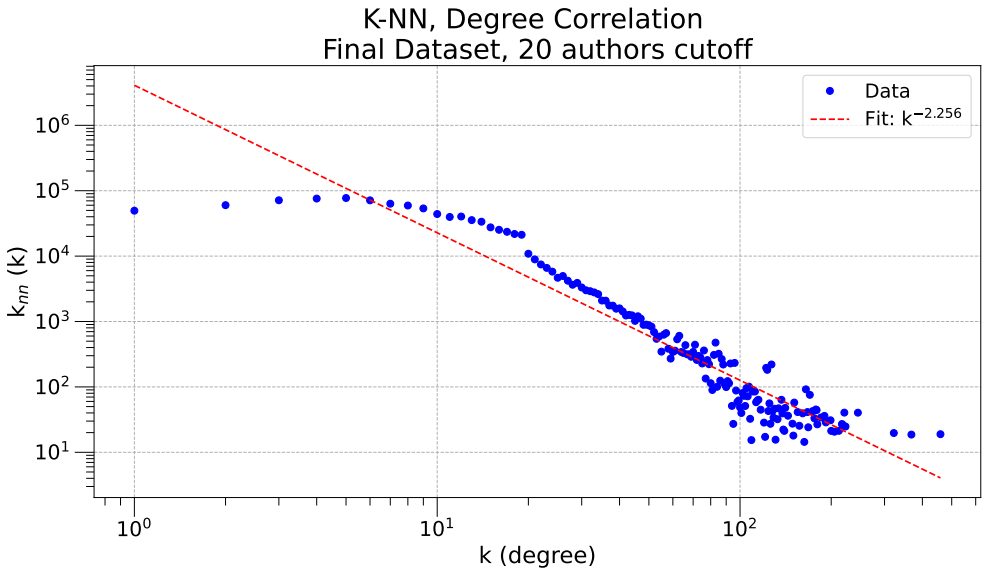

Figure 7: Degree Correlation. We can see that the exponent is

2.7 Time Evolving Network and Animation

While we saw indicators describing the network, we never actually looked into how this network evolves in time. The data structure describing the edges have more information: the time of publication is saved and the number of authors (inverse of the weight of the connection). With this we can see how it evolves! It can be seen here!

References

[1] Albert-László Barabási. Network science. *http://networksciencebook.com*, 2012.

[2] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx, in proceedings of the 7th python in science conference (scipy 2008), 2008.

[3] Xiaoming Liu, Johan Bollen, Michael L. Nelson, and Herbert Van de Sompel. Co-authorship networks in the digital library research community, 2005.