Scientific Modeling Computer Laboratory

Project: Time Evolving Networks

Second Bi-weekly Report

Ádám Gergely Szabó

7th 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 searcing 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 Storing The Tables

It was problematic to handle the data in the means of strings, thus some kind of encoding had to be done. Beforehand, uuid5s were used in strings and is it turns out, strings take up a lot of space in the memory. I have transitioned to 32 bit unsigned integers, which is like by numpy as these use much less space. The tables will be prepared for evolving networks and will have the following structure:

Author1	Author2	Publication Year	Number of Authors
0	1	2021	2870
...	...	...	...

Table 1: Table structure for storing information of the tie evolving network. In the following, the publication year won't be used.

2.2 Drawing Networks

The main problem with drawing networks is that the co-authorship network has many connection: around 90 million. This is too many connection, even for specialized soft-

wares / libraries and it is quite unnecessary to draw these many edges for a given network. Most of these edges are coming for multi-author publications, in which all of the authors needs to be connected and there are hundred to thousand authors contributing to this publication. (This is often done through a collaboration organizations). In this case a cutoff needs to be placed. The cutoff was placed in a way that publications with over 200 authors are neglected, resulting into a network that has $\sim$ 350000 edge. (In this case there are less authors presented too.) But it is necessary to look into the effect of the cutoff.

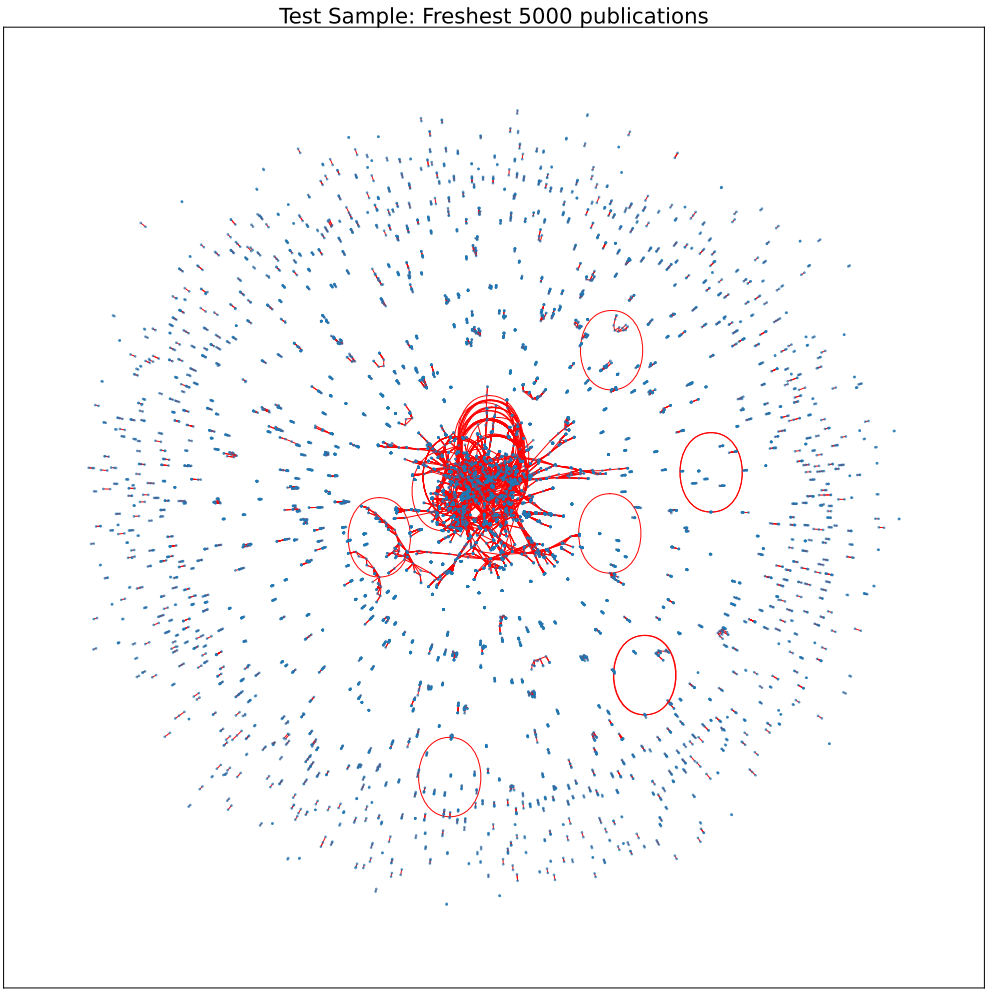


Figure 1: Co-author network presented with cutoff in place at. Looking at the network, it is visible that a lot of small groups are present, with a large component in the middle. Using networkx's spring layout.

It is visible that there are a lot of small groups present in the network and there is huge component in the middle. This is a result of the cutoff being too high even now: 200 authors for a publication still heavily impacts the formation of the graph and there are a lot of publication with lot of authors. Choosing 200 for the cutoff was arbitrary:

later on with different group searching methods we could obtain a better number for the cutoff but for now it is sufficient.

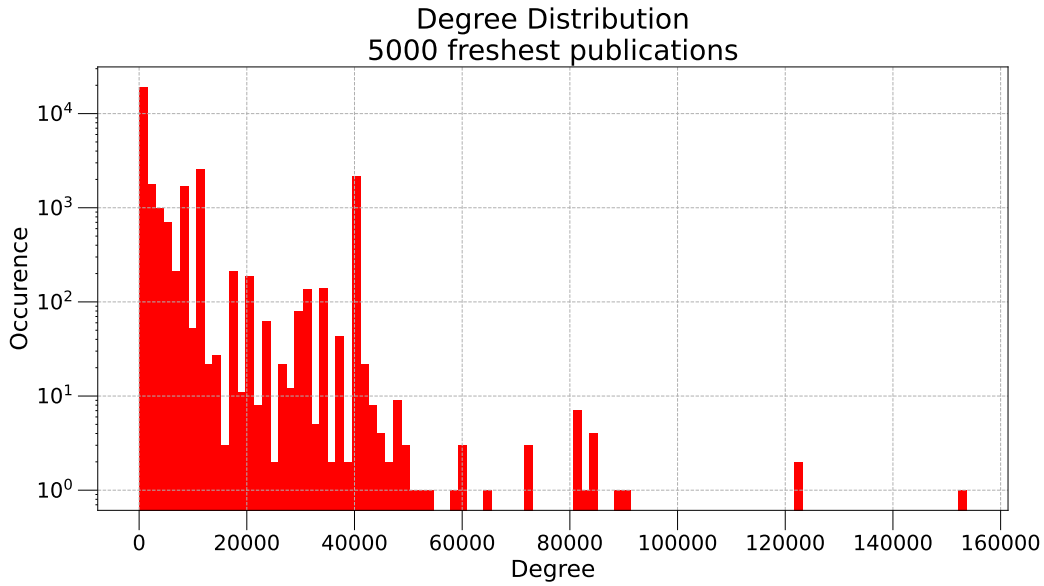


Figure 2: Degree distribution without the cutoff.

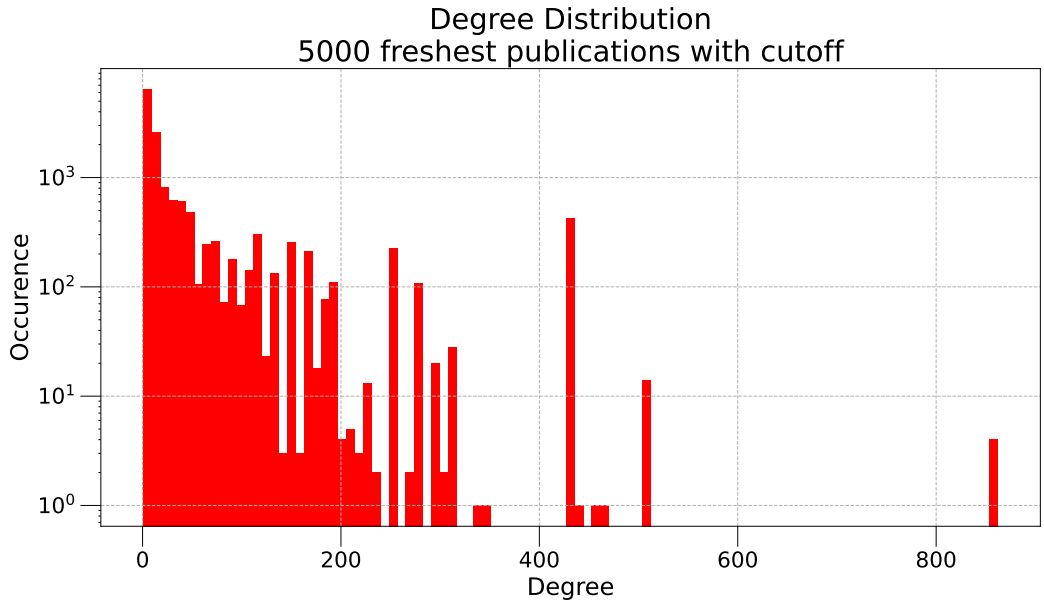


Figure 3: Degree distribution with the cutoff.

The degree distribution in the two case behave similar while it latter case the highest degree is smaller which is the result of the cutoff. This means that the cutoff was well chosen, but more indicators needs to be calculated to see further impacts of the cutoff.

2.3 Further Networks

Authors are connected if there are present in the list of authors for a given publication resulting into the co-author network. Is this the only network that could be formed?

Further exploration of the publication fields could bring new networks to life. Most of the fields are there for administrative reasons, while others carry other kind of information: subjects and keywords. It is possible to connect an author with a subject which brings light to what subjects are liked more. With keywords, we can look into different specialisations. It is advised to see how persistent these fields are: subjects are always present, but keywords are rarely missing. In the following, the subject-author network is presented:

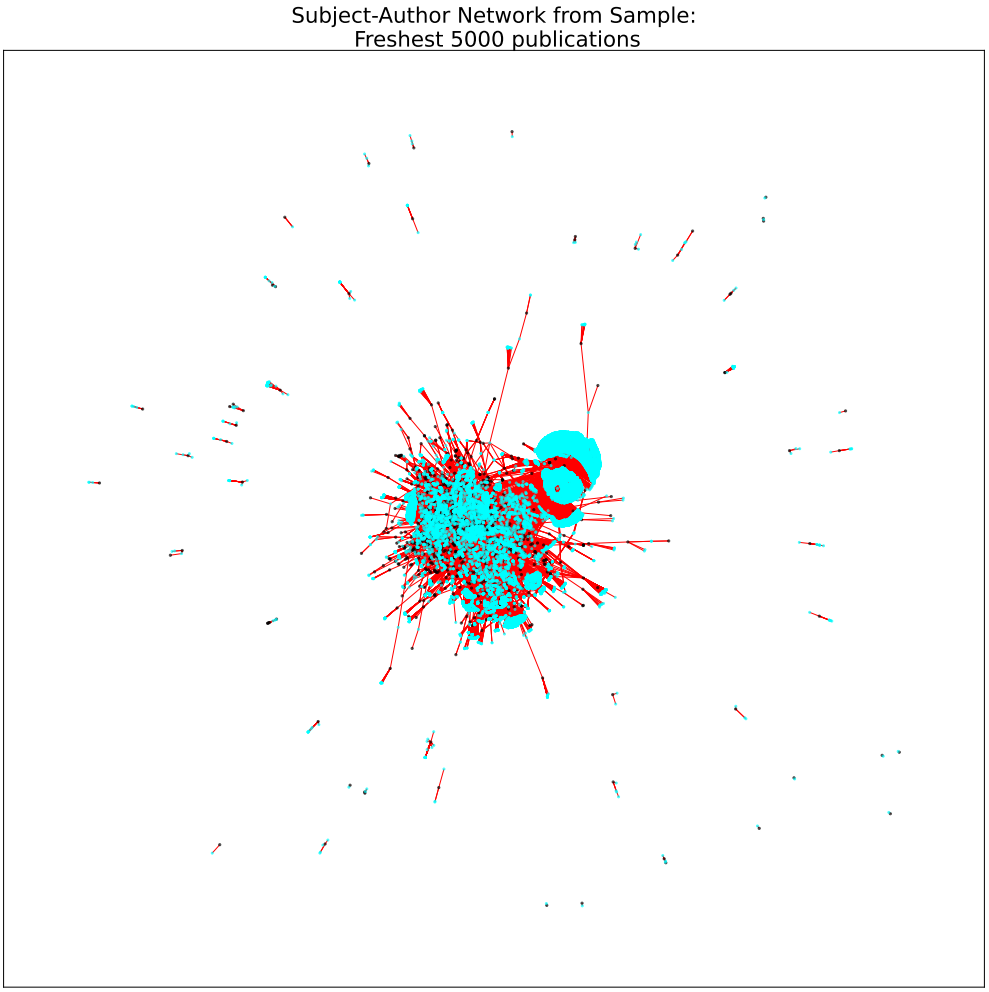


Figure 4: The subject-author network. Black dots are subjects, cyan colored dots are authors. It is a bipartite network and as such, networkx's spring layout has hard time to unfold it.

It is visible that there are subjects that are like and a lot of authors work with while in some cases the subject is specialized and only a few authors write in such subject. For example, 'Physics' have a lot of authors attached to it.

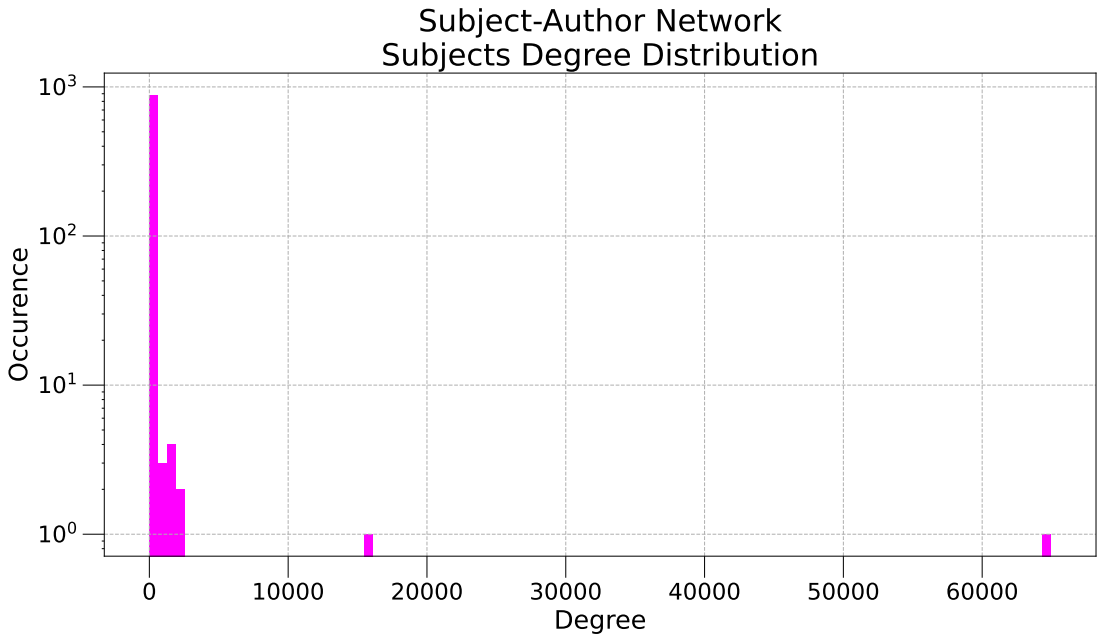
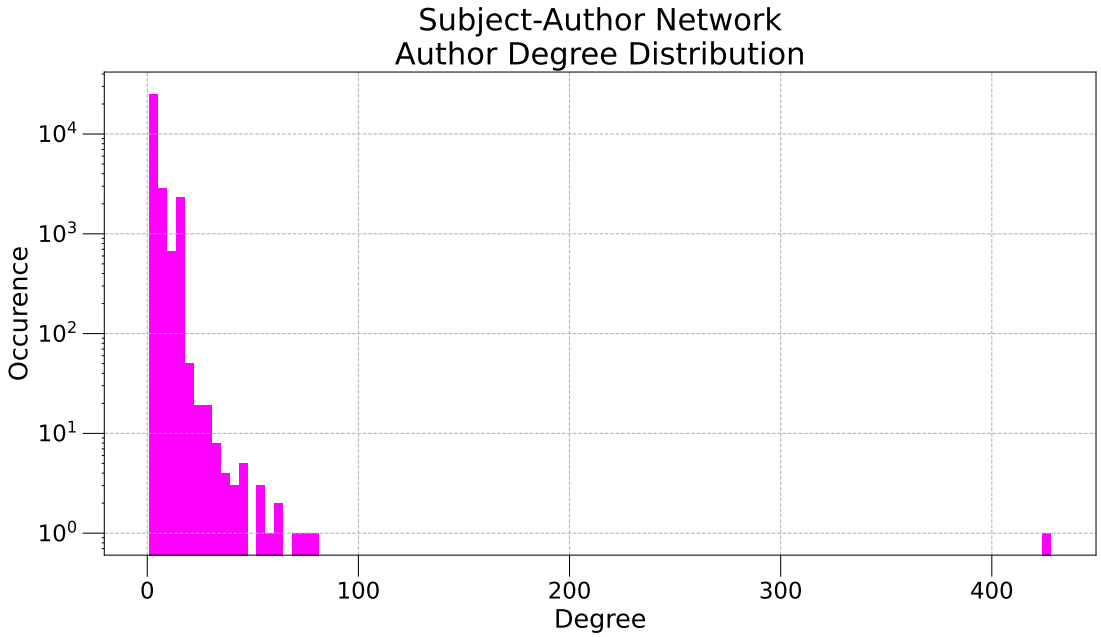


Figure 5



Unsurprisingly, an author is only familiar with a few topics, but there is one author with over 400 subjects, while there is only 890 subjects present in the sample. This person maybe connected to a lot of publications which have a lot of different topics attached to them.

3 Problems

Now the only problem remaining is surpassing the MTMT's site's limitation of 5000 publications. This is the last problem that is an obstacle in further progress.

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.