Scientific Modeling Computer Laboratory

Project: MTMT's Co-author Network

First Bi-weekly Report

Ádám Gergely Szabó

21st of February, 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 Gathering and Understanding the Data

In this weeks report, the advanced in understanding the data structure will be presented. The data from the site is easy to gather as the site uses a ReST API that accepts queries. There are two datastructure options: json or xml. In my case, I will use the json format and save the result. The returned json file is hard to understand by simple opening it in a text editor: the objects presented in the file has many fields that needs exploration and explanation. After this, it becomes obvious: the returned json stores the publications information in its 'content' field, and each publication has an 'authorships' field to describe the different authors for a given publications.

2.2 Sample: 5000 Freshest Publications

I gathered a smaller test sample consisting of the 5000 freshly created publications according the MTMT's records. This sample is large enough to be representative of the publications available under the topic 'Science'. To create the network, we should know how to create the edge list for the given network: people that work together (listed as authors for one publications) are connected. Problems that were faced during the creation of the node list and edge list of the network:

- Some people don't give their family name,
- Some publications are done with collaboration organization that is listed as an author,
- Some publications have huge amount of authors.

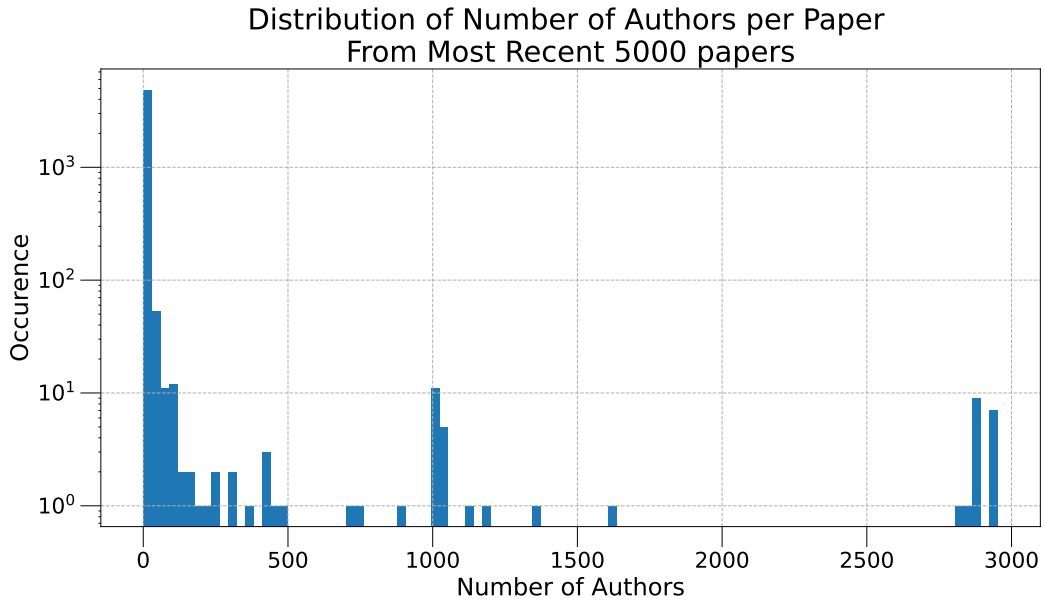

Figure 1: Distribution of the number of authors for publications.
The highest amount of authors for a given paper reached 2800 entity.

This in the end resulted in a network having 33000 nodes and 170 million edges, which is heavily affected by publications having large number of authors and as for such publications, all authors needs to be connected and presented in the network. Even though the network has many edges, compared to the full graph created from connecting all nodes with each other, this network could be considered sparse.

3 Problems

Problems that were faced till this report was creating are the following: the edge list is huge and it takes up huge amount of space in memory or in a file, if not handled appropriately, meaning that the nodes cannot be kept with the authors name, causing the need of some kind of encoding. Even though this is just a sample of the data, the network cannot be presented by the tools available by networkx or igraph Python packages, meaning that further tools needs to be explored.

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.