Evolution Over Time of the Structure of Social Graphs

Description of Work

Kostiantyn Ohulchanskyi
Sofiia Shelest

Supervised by

Frédéric Giroire Nicolas Nisse
Thibaud Trolliet Małgorzata Sulkowska

November 2020

Abstract. The goal of this personal project is to continue previous research conducted by Abdelkrim EL Merss on how funding impacts future scientific collaborations between research teams by studying experiences of the IDEX and LABEX funding programs. However, in this research only collaborations between research teams in the field of IT and telecommunications where at least one of collaborators is from Sophia Antipolis were considered. In this PFE we will extend the scale of the investigation by considering a bigger collaboration network with a broader range of researchers. We will also check whether the network is scale-free, and study a few properties of communities in the network.

1 General Project Description

1.1 Motivation

The purpose of multiple funding programs, such as IDEX and LABEX, is to facilitate scientific collaborations between research teams. For example, the idea of the LABEX program is to fund PhD students that are supervised by different research teams in order to increase the number of collaborations between those teams in the future.

This PFE is a part of the SNIF project which aims to measure the success of funding of the aforementioned programs by studying their influence on multiple metrics, such as the number of collaborations between researchers from different scientific disciplines, evolution of clustering of the network, et cetera. These metrics allow analysing how collaborations between different research teams evolve over time.

1.2 Framework/Context

This PFE is a continuation of the research conducted by Abdelkrim EL Merss during an internship from March to August 2020 which mainly concentrated on two topics: analysis of a real collaboration network and construction of a theoretical model.

1.2.1 Real Data Analysis

In this part of the research, real-world collaboration network was investigated.

Real data was extracted from the Scopus database, Elsevier's abstract and citation database that covers millions of scientific papers on different disciplines. Then, the data was processed and filtered to consider research teams in the field of IT and telecommunications where at least one of collaborators is from Sophia Antipolis.

The collected data was used to construct social graphs modelling the collaboration network at different periods of time (with a step of one year). In these graphs, nodes represent authors, and there is an edge between two nodes if corresponding authors collaborated on the same publication. The modelled graphs allowed calculating multiple metrics and properties of the collaboration network, such as clustering (communities of researchers), to measure the impact of funding on the evolution of scientific collaborations.

In the final report of the research, it has been concluded that funding has a significant impact on the development of scientific collaborations between different research teams.

1.2.2 Theoretical Model

This part of the research aimed to propose a theoretical model to build and study random collaboration networks that have the same properties as the real ones.

In order to devise a model, analysis of a real collaboration network was conducted (see section 1.2.1), which determined properties that such networks have. This allowed introducing parameters to the model to generate random graphs that have the same properties.

The model was implemented in the Python programming language and the NetworkX library was used to represent and process graphs.

1.3 Goals

1.3.1 Communities

In order to better understand how researchers form communities based on collaborations, we aim to investigate the following.

1. Study the distribution of publications with different numbers of communities that collaborate on a single publication.
2. Study the ratio of co-authors from different communities.

   For example, if there is a publication with 4 authors from 2 communities, there may be different cases:

   - there are 2 authors from one community and 2 authors from the other one, which means that a paper is co-authored evenly by researchers from different communities;
   - there are 3 authors from one community and 1 author from the other one, which means that a paper is co-authored predominantly by researchers from a single community.

1.3.2 Scale-Free Network

A scale-free network is a network where the probability distribution of degrees of nodes over the network follows a power-law distribution.

Multiple studies of real-world collaboration networks (scientific collaborations, collaborations of actors in movies, etc.) have shown that these networks are scale-free. In the former research, it was observed that the degree distribution of such a network appears to follow a power-law distribution.

One of our goals is to construct a social network based on the data on collaborations between researchers and to check whether this network is indeed scale-free in order to better understand the structure of such network and its properties.

Studying whether the collaboration network is scale-free is crucial for improving the theoretical model since this property could show us how to generate a random network that has the same structure as the real one. For example, the Barabási–Albert model can be used to generate random scale-free networks.

1.3.3 Larger Scale

In the former research, only research teams in the field of IT and telecommunications where at least one of collaborators is from Sophia Antipolis area were considered.

One of the main goals is to investigate the influence of funding on the collaborations of research teams from a broader range of scientific disciplines, universities, and cities.

An extension of the scale of investigation may help us discover new properties of the network that are too subtle on a smaller scale.

In order to extend the scale of the investigation, we need to prepare a large amount of data and then use it to build a bigger collaboration network. When the network is built, we have to analyse it by utilising algorithms implemented for the former research. However, implementations of those algorithms should be optimised to perform better on larger graphs since, according to Abdelkrim EL Merss's final report, they may take days to process.

One of our current goals is to improve the implementation of the Louvain method for community detection.

1.4 Challenges

1.4.1 Large Amount of Data

The first challenge is in storing and processing a large amount of data. In the former research, the final size of files that stored processed and filtered data was 16 gigabytes. By increasing the scale of investigation, we also increase the amount of data we need to process. This complicates the analysis of a network due to limitations of the RAM size of a machine.

1.4.2 Processing Complexity

The second challenge is in processing a graph with a significant number of nodes and edges. An increase in the size of a graph will impact the computational time of analysis of a network. In order to handle this, code optimisations may be required.

2 Workplan, Tasks and Milestones

2.1 Tasks

2.1.1 Communities

This task is assigned to Kostiantyn.

1. The fist step is to construct a social graph from the extracted data.

   We will use the Python programming language due to its rich scientific environment. To represent and manipulate graphs, we will use the NetworkX library.

2. Then, we need to detect communities using the Louvain method. This method has been implemented as a part of the former research to analyse the evolution of the clustering of the collaboration network.

3. Finally, calculate the corresponding metrics.

If there is enough time, other metrics (besides those that have been described in section 1.3.1) may be calculated and analysed.

2.1.2 Scale-Free Network

This task is assigned to Kostiantyn.

In order to check whether a real collaboration network is scale-free, the following steps should be taken.

1. Like in the previous task, the fist step is to construct a social graph from the extracted data using the Python programming language and the NetworkX library.
2. Then, the degree distribution of the constructed graph should be calculated. The degree distribution $P(k)$ of a graph is defined as a fraction of nodes with degree k,

$$P(k) = \frac{n_k}{n},$$

   where n is the total number of nodes, and n_k is the number of nodes with degree k. This can be easily calculated with an extensive API that the NetworkX library provides.
3. Finally, check whether the distribution of degrees of nodes follows a power-law distribution using the Kolmogorov—Smirnov test.

   This can be done with the SciPy package and its module `scipy.stats`. This module contains numerous statistical functions, including `scipy.stats.kstest` for performing the (one sample or two samples) Kolmogorov—Smirnov test.

2.1.3 Larger Scale

This task is assigned to Sofiia.

1. Optimise the implementation of the Louvain method to make it perform faster on larger graphs.
2. Run analysis on a larger social graph that includes researchers from different scientific fields.
3. In case a if the graph uses too much memory, suggest another representation of the graph in memory.

   In the former research, the total size of the extracted data was 16 gigabytes, and it can be held in RAM only by a small portion of the modern personal computers. Thus, an optimisation of the graph representation may be required.

2.2 Milestones and Workplan

1. Learn the existing implementation of the Louvain method and apply it to the collaboration network to calculate the determined metrics. (Kostiantyn)
2. Survey the literature on power-law distributions in social networks. (Kostiantyn)
3. Test the hypothesis that the degree distribution of the collaboration network follows a power-law distribution. (Kostiantyn)
4. Optimise the implementation of the Louvain method. (Sofiia)
5. Analyse a larger social graph using the optimised Louvain method. (Sofiia)
6. Compare performance of optimised Louvain method with previous implementation. (Sofiia)
7. Suggest another representation of the graph if it uses too much RAM. (Sofiia)
8. Write a final report about the research that has been conducted. (Kostiantyn, Sofiia)

3 Bibliography

[1] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, pp. 661–703, 2009.

[2] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review*, vol. 70, no. 066111, 2004.

[3] A. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

[5] C. Avin, Z. Lotker, and D. Peleg, "Random preferential attachment hypergraphs," *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 398–405, 2019.