D1.1 - Description of Work (DOW)

Evolution over time of the structure of social graphs

Student Leonardo Serilli
Supervisors Małgorzata Sulkowska, Nicolas Nisse, Frédéric Giroire

Year 2021/2022

Abstract

Many natural and human made systems can be represented by networks, that is, graphs, sets of nodes and edges. For example the World Wide Web is just a set of hosts interconnected by data links and social networks are made of people and their relationships. Those structures **respect similar mathematical properties** like the power-law distribution, which intuitively says that majority of nodes have just a few connections while there are several nodes with a large number of connections. It is just the way nature want those kinds of networks to be: **self-organizing structures**.
Finding this peculiar characteristic on data we can collect and analyze can bring us to the development of tools to study them, and even to predict, with high accuracy, their future evolution.
The scope of this project is to build a **network of scientific authors and their collaborations**, collected from the **Scopus database**, and to analyze the distribution of their collaborations over time.

Contents

1 General Project Description **1**
1.1 Framework/Context 1
1.2 Motivations 2
1.3 Challenges 2
1.4 Goals 3

2 State of the Art **3**

3 Tasks and Milestones **4**
3.1 Retrieving Data 4
3.2 Filtering active authors 5
3.3 Workplan and Milestones 6

1 General Project Description

1.1 Framework/Context

This project is a part of a larger one involving researchers in various fields, such as economics, sociology and computer science; it is focused on the evaluation of the impact of funding on scientific research.
As an example of funding one can indicate LabEx and IdEx, French programs which

scope is to promote collaborations involving different research fields.

The purpose of this project is to analyse the evolution of nodes degree in a collaboration network built upon scientific publications extracted from Scopus Database, that is, **the vertex trajectory**.

1.2 Motivations

Many systems can be represented as a network, both natural as well as human built, such as the World Wide Web, social networks, collaborations of actors in movies, or even the interaction among molecules. Each of this systems can be viewed as a set of nodes, routers, computers or people, and a set of edges connecting them, data-link among computers, social relations among people or, as in our case, **collaborations between scientific researchers**.
Those systems are of practical interest and **attract researchers in different fields of study**, since we reached the computational power to deal with the large amount of data and since we have mathematical models to describe their evolution over time. This can bring to a wide range of tools and applications to analyze the structure of those systems and ways to predict their evolution.

A common property of those networks is that they are **scale-free**, it means that the probability distribution of degrees of nodes over the network follows a **power-law distribution**. The insight on it is that the proportion of vertices of a given degree changes following a power law when the degree grows. Power-law distribution can be seen in many other phenomena such as the frequencies of word in most languages, the sizes of craters on the moon, of Solar flare and so on. This feature is a consequence of the fact that a new node of the network sites preferentially to nodes that already have a huge number of connections. An intuitive example can be that, in social networks, the probability of having new friends is higher for people who already have a lot of them.

All those example show that "**the development of large networks is governed by robust self-organizing phenomena that go beyond the particulars of the individual systems**"[1].
This project aims at investigating further features of scale-free networks. Particularly, it will concentrate on observing the evolution of degree over time.

1.3 Challenges

In this work a collaboration network built upon data collected from **Scopus database** is analyzed. A node in the network represents an author and there is an edge between two authors if they have collaborated at least once on a scientific paper.
The data is composed of **258145 French computer science authors** and the amount of collaborations they had **since 1990 until 2018**.

The first challenge would be to **avoid working on misleading data**, there can be authors who have published just once in their carrier or other that had a skyrocketing number of collaborations for a couple of years before disappearing from the network and so on.
All of this data can bring to a model diverging from the expected one. Another chal-

lenge concerns dealing with the **lack of temporal step**, from the collected data the only information that can be used for this purpose is the year in which a collaboration appears, and there are present only 29 years of collaborations.

The final purpose would be to have a bunch of **good metrics to build vertex trajectory** upon the yet cited network.

1.4 Goals

The first goal is to take confidence with the collected data, building a **meaningful dataset** and **analyzing vertex trajectories**, that is, the evolution of node degrees over time, by using simple metrics like the average of the amount of collaborations for authors who started to publish in the same year.

Another goal is about understanding how the structure of the collaboration network varies for authors with a similar vertex trajectory when a subset of them get a funding. The expected behavior is that the funded author will have an increase in the number of collaborations. In this work will be check if the real data follow this expected behavior.

The final goal of the project is to have methods to **build meaningful vertex trajectories** and to extract useful data from them.

2 State of the Art

Let $G = (V, E)$ be a graph of $|V| = n$ nodes, and let n_k be the number of nodes of degree k then power-law distribution P_k is defined as follows:

$$P_k = \frac{n_k}{n}$$
$$P_k \sim Ck^{-\lambda}$$

where $\lambda > 0$ is an exponential parameter and $C > 0$ a scaling constant.
A network whose probability distribution of degrees of nodes respects the power law distribution is said to be **scale-free**.
In past decades researchers from different fields have worked for establish the scale free properties of networks. Observing this property of real-world network makes possible to develop theoretical models to study those networks.

An example of possible applications can be a tool to generate random networks having the same structure of the observed one, as the **Barabàsi–Albert[1] model** does, which is used to generate scale-free networks. In order to build his model Barabàsi mapped the topology of a portion of the Web observing that some nodes, called hubs, has a higher number of connections, and with it, a higher probability to develop connections with new nodes in future.

Recent studies show that scale-free networks are not so widespread as thought, the majority follows a **power-law distribution with an exponential cutoff** of the form:

$$P_k \sim Ck^{-\lambda}\gamma^k$$

where $0 \leq \gamma < 1$ is a constant parameter of the distribution.

3 Tasks and Milestones

3.1 Retrieving Data

The first part of the project concerned the retrival of data about collaborations and publications regarding all computer science authors in France since 1990 to 2018.

Collaboration data

	ID	1990	1991	1992	1993	1994	1995	1996	1997	1998	...	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
0	6508297663	0	0	0	0	0	0	0	0	0	...	4	7	7	8	8	8	8	8	8	8
1	34571759000	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	5	5	5	5
2	7004267341	0	0	0	0	0	0	0	0	0	...	10	10	10	16	16	16	16	16	16	16
3	8642393600	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	7	7	7	7
4	55873955900	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	8	8	8	8
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
255095	6507630481	0	0	0	0	0	0	0	0	0	...	18	18	18	18	18	29	29	29	29	29
255096	24577815500	0	0	0	0	0	0	0	0	0	...	4	4	4	4	6	13	16	16	16	70
255097	57195243976	0	0	0	0	0	0	0	0	0	...	0	3	3	3	3	3	3	3	8	8
255098	35328962100	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	2	2	2	2	3
255099	7403521415	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	29	29	29

255100 rows × 30 columns

Figure 1: Collaboration data for computer science authors.

Publication data

	ID	1990	1991	1992	1993	1994	1995	1996	1997	1998	...	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
0	8958327900	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	6508297663	0	0	0	0	0	0	0	0	0	...	4	7	7	8	8	8	8	8	8	8
2	7004267341	0	0	0	0	0	0	0	0	0	...	10	10	10	16	16	16	16	16	16	16
3	8642393600	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	7	7	7	7
4	55873955900	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	8	8	8	8
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
232833	6507630481	0	0	0	0	0	0	0	0	0	...	18	18	18	18	18	29	29	29	29	29
232834	24577815500	0	0	0	0	0	0	0	0	0	...	4	4	4	4	6	13	16	16	16	70
232835	57195243976	0	0	0	0	0	0	0	0	0	...	0	3	3	3	3	3	3	3	8	8
232836	35328962100	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	2	2	2	2	3
232837	7403521415	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	29	29	29

232838 rows × 30 columns

Figure 2: Publication data for computer science authors.

3.2 Filtering active authors

All inactive authors are filtered out, so a dataset is built for each possible definition of Inactivity.

An author is inactive if it has a hole in publications greater than a given value. An author has a hole in publication if he haven't published for n consecutive years. For example, given an hole size = 3, the author A1 is active but A2 is not.

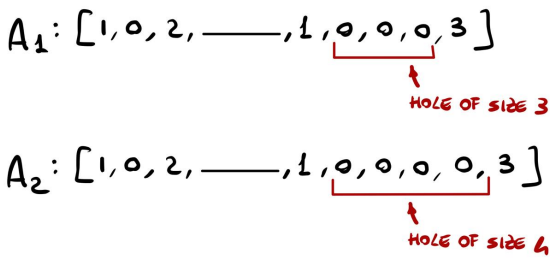


Figure 3: Hole size definition.

The number of authors kept for each hole size is showed in the following chart and also their differences in the number of authors:

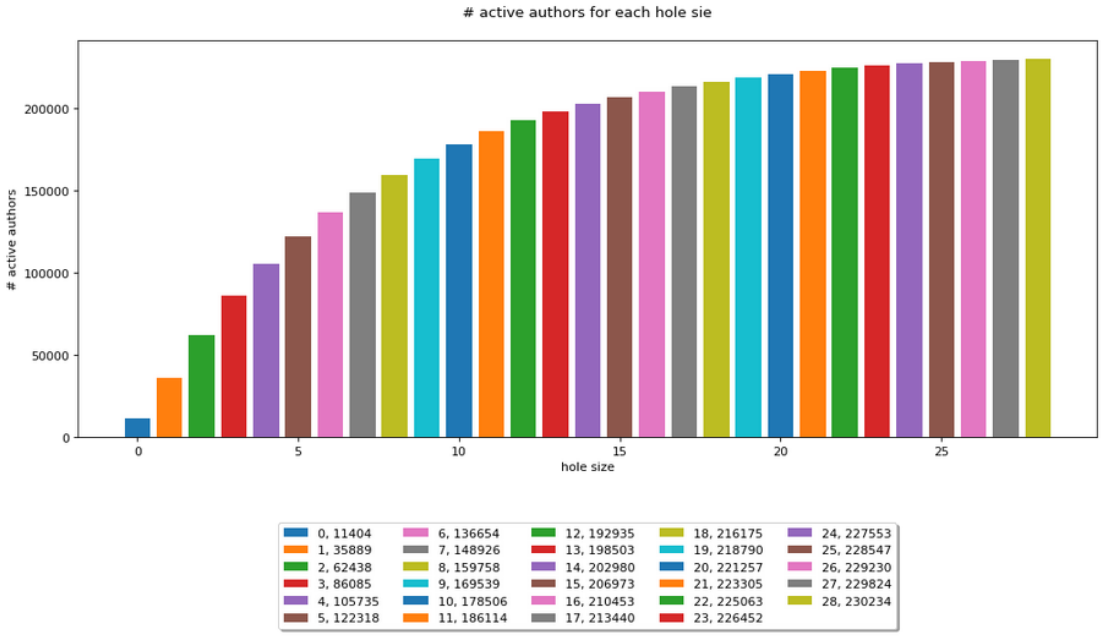


Figure 4: Number of authors kept for each hole size.

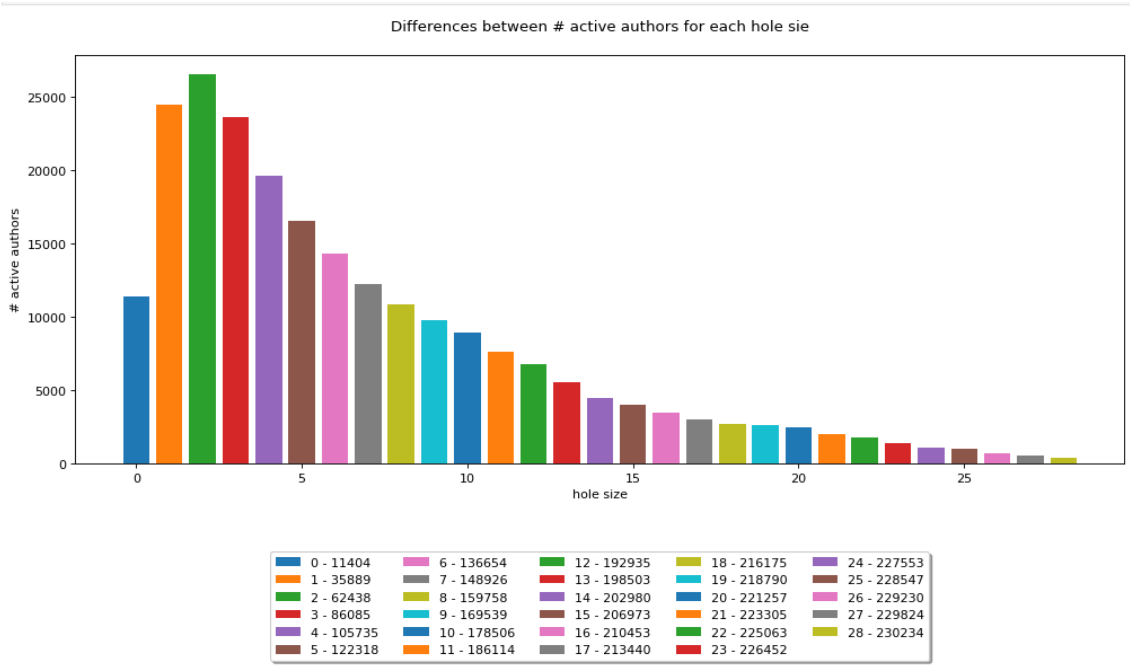
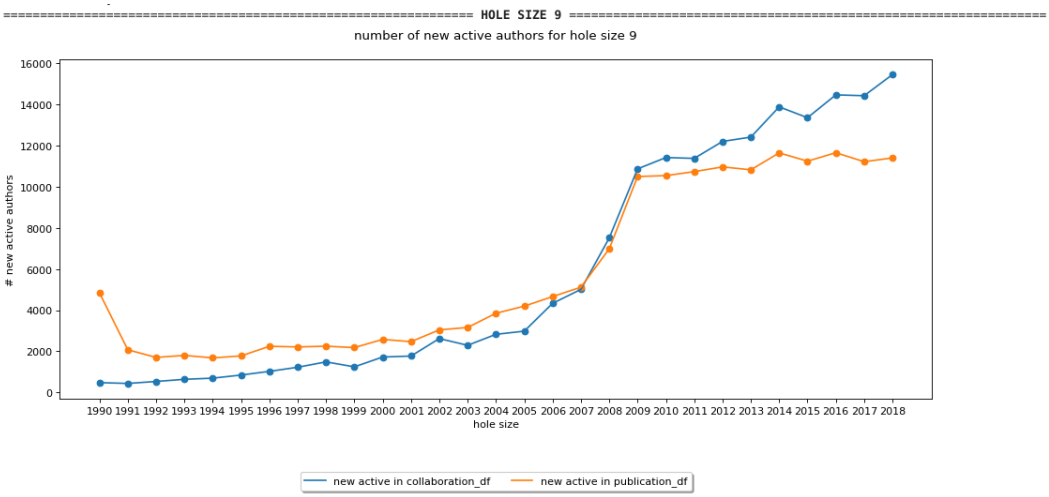


Figure 5: Differences in the number of authors for each hole size.

Also the distribution of new collaborators and new publicators for each year and hole size as been found, for example for an hole size of 9 is the following:



collaboration_df {'1990': '475', '1991': '435', '1992': '531', '1993': '636', '1994': '692', '1995': '843', '1996': '1023', '1997': '1223', '1998': '1483', '1999': '1242', '2000': '1720', '2001': '1767', '2002': '2617', '2003': '2290', '2004': '2824', '2005': '2978', '2006': '4331', '2007': '5010', '2008': '7514', '2009': '10871', '2010': '11422', '2011': '11382', '2012': '12206', '2013': '12417', '2014': '13888', '2015': '13360', '2016': '14470', '2017': '14430', '2018': '15459'}

publication_df {'1990': '4844', '1991': '2072', '1992': '1702', '1993': '1799', '1994': '1678', '1995': '1774', '1996': '2242', '1997': '2211', '1998': '2246', '1999': '2179', '2000': '2577', '2001': '2467', '2002': '3040', '2003': '3157', '2004': '3849', '2005': '4195', '2006': '4660', '2007': '5116', '2008': '6982', '2009': '10500', '2010': '10543', '2011': '10737', '2012': '10967', '2013': '10826', '2014': '11649', '2015': '11246', '2016': '11657', '2017': '11220', '2018': '11404'}

Figure 6: distribution of new collaborators and new publicators for each year and hole size 9.

3.3 Workplan and Milestones

- Collaboration dataset
- Refined Collaboration datasets

- Dataset splitted by starting publication year
- Average vertex trajectories for authors who started collaborating in the same year
- Plots about obtained data

References

[1] A.-L. Barabasi and R. Albert, "Albert, R.: Emergence of Scaling in Random Networks. Science 286, 509-512," Science (New York, N.Y.), vol. 286, pp. 509–12, Nov. 1999, doi: 10.1126/science.286.5439.509.

[2] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," Phys. Rev. E, vol. 70, no. 6, p. 066111, Dec. 2004, doi: 10.1103/PhysRevE.70.066111.

[3] "Scopus": https://www.scopus.com/home.uri.