University of Côte d'Azur
Polytech Nice-Sophia
Master II - Ubiquitous Networking (Ubinet)

Evolution over time of the structure of social graphs

Internship report

LEONARDO SERILLI

Supervisors: Małgorzata Sulkowska, Nicolas Nisse, Frédéric Giroire

Nice, 2022

Inria

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. During this study we discovered that the underlying theoretical model is probably more complex than we assumed at the beginning. Nevertheless, our research may be seen as a step forward in constructing functions representing well the evolution of the node degree in the networks.

Short inspired phrase here

Contents

Abstract 1

1 Introduction 7
- 1.1 General Project Description . . . 7
  - 1.1.1 Framework/Context . . . 7
  - 1.1.2 Motivations . . . 7
  - 1.1.3 Challenges . . . 8
  - 1.1.4 Goals . . . 9

2 State of the Art 11
- 2.1 State of the Art . . . 11
- 2.2 Data Preparation . . . 13
  - 2.2.1 Retrieving Data . . . 13
  - 2.2.2 Filtering active authors . . . 13
  - 2.2.3 Identifying active authors . . . 16
  - 2.2.4 Changing definition of event . . . 16
  - 2.2.5 Getting to know data characteristic . . . 18
  - 2.2.6 Plotting Ratios . . . 18
  - 2.2.7 Degree Distribution . . . 20

Conclusions 23

Appendices 25

A General details 27

Acknowledgements 29

Bibliography 31

Abbreviations 33

Chapter 1

Introduction

1.1 General Project Description

1.1.1 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 expected impact is meant that, for example, given a couple of authors with similar collaboration behaviors, if one of them get a funding, his collaboration rate is expected to grow compared to the others, unfortunately, this is not always true in the analyzed data.

As an example of funding one can indicate LabEx and IdEx, French programs whose 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** (Section 2.2 - state of the art), where for collaboration network is meant: a set of nodes, the authors, connected by edges, representing collaborations, such that two nodes are connected if there exists at least one collaboration between them.

1.1.2 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, e.g. routers, computers or people, and a set of edges connecting them, e.g., 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 with exponential cut-off**, described in the **state of the art** (Section 2.2) [4]. The insight on it is that the proportion of vertices of a given degree changes following a power-law with exponential cut-off when the degree grows.

Power-law distribution can be seen in many other phenomena such as the frequencies of words 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 connects 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 examples 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.

1.1.3 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 and publications they had **since 1990 until 2018**, along with the list of their co-authors and publications.

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. It is important to identify well the outliers to prevent their big impact on the average behaviour of the theoretical model.

Another challenge concerns dealing with the **lack of temporal step**. We want to investigate the evolution of vertex degree over time from the collected data but the only time information that can be used from the collected data is the year in which a collaboration appears, and there are present only 29 years of collaborations.

The final challenge is to have a bunch of **good metrics to build vertex trajectory** upon the yet cited network. Where the vertex trajectory is a sequence defined by the total number of collaborations each author has for each year.

1.1.4 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 gets a funding. The expected behavior is that the funded author will have an increase in the number of collaborations. In this work one will check if the real data follows this expected behavior.

The final goal of the project is to obtain methods to **build meaningful vertex trajectories** and to extract useful data from them, like fitting functions able to represent their evolution.

Chapter 2

State of the Art

2.1 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 in G. We say that the degree distribution M_k of G follows a power-law if:

$$M_k = \frac{n_k}{n}$$

and

$$M_k \sim Ck^{-\lambda}$$

where $\lambda > 0$ is an exponential parameter and $C > 0$ a scaling constant, and the sign "$\sim$" stands for almost equal.

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 in order to 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 as possible applications can be a tool to generate random networks having the same structure as the observed one, as the **Barabàsi–Albert[1] model** does, which is used to generate scale-free networks. In order to build their model, Barabàsi and Albert [1], 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 [5], it turns out that many of them follow a **power-law distribution with an exponential cutoff** of the form:

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

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

The experimental study showed that data we are working with also falls into "exponential cutoff" case.

Next, reminding that the main task of this work is their investigation, **vertex trajectories** are defined.

Given a graph $G_t = (V_t, E_t)$, where t is a given time step, let's define two kind of events: a **node event**, where a new node appears in the graph, and an **edge event** where a new edge appears.

In the node event, the new node must select another, already present, to connect with, while in the edge event two nodes must be selected to place the edge. This selection is carried out by an attachment function as $f(x)$ which defines the probability for any node $v \in V_t$ to be chosen for the attachment.

$$\Pr[v \text{ is chosen at step t+1 }] = \frac{f\left(\mathrm{d}_v(t)\right)}{\sum_{w \in V_t} f\left(\mathrm{d}_w(t)\right)}$$

In the equation $d_t(v)$ is the degree of vertex v at time t .

Time steps are defined by the occurrence of an event, after which we obtain a new graph $G_{t+1} = (V_{t+1}, E_{t+1})$. Given n time steps , we can define the evolution of a collaboration graph G_0 by the sequence of graphs $\{G_0, G_1, ...G_n\}$.

Let $d_v(t)$ be the degree of vertex v at time t and let t_0 be the time at which v appears in the graph, then the vertex trajectory of v is the evolution over time of it's degree, so the sequence $\{d_v(t_0), ..., d_v(t), ..., d_v(t_n)\}$

The **theoretical model** we will refer to during this work is the one illustrated in the next table [7], in which the degree distribution M_k follows a more subtle distribution than the power-law with exponential cut-off, the so-called **stretched exponential**, and the vertex trajectory has a logarithmic shape $g_v(t)$. In the table α is a positive constant dependant from the parameters of the model, while t_v is the time step in which the node v joined the network. Lastly we assume that $\sum_{w \in V_t} \left(\mathrm{d}_w(t)\right)^\gamma \sim \mu t,$ where $\mu \in [p, 2]$, where $\mu \in [p, 2]$.

Attachment function	**Degree distribution**	**Vertex trajectory**
$f(x) = x^\gamma$ $0 \leq \gamma < 1$ $\alpha = \frac{\mu}{2-p}$	$M_k = \frac{\alpha}{k^\gamma} \prod_{j=1}^{k} \left(\frac{j^\gamma}{\alpha + j^\gamma}\right)$ $\sim \alpha \cdot k^{-\gamma} \cdot \exp\left\{-\frac{\alpha}{1-\gamma} k^{1-\gamma}\right\}$	$g_v(t) = \left(\frac{1-\gamma}{\alpha} \ln(t/t_v) + 1\right)^{1/(1-\gamma)}$

Later the function $g_v(t)$ will be used for fitting average vertex trajectories extracted from the provided data.

2.2 Data Preparation

2.2.1 Retrieving Data

The first part of the project concerned the retrieval of data about collaborations regarding all computer science authors in France since 1990 to 2018. Where a collaboration between two authors exists if they have published together the same paper.

Exemplary data are given in Fig.1. where each row represent an author with his ID on the Scopus database. In it, for each year, each cell shows the cumulative number of his collaborations until each year.

	ID	1990	1991	1992	1993	1994	1995	1996	1997	1998	...	2015	2016	2017	2018
0	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0
1	8	0	0	0	5	5	5	5	5	5	...	5	5	5	5
2	9523235	0	0	0	0	0	0	0	0	0	...	0	0	14	14
3	6503845773	0	0	0	0	0	0	0	0	0	...	3	3	3	3
4	6503846520	0	0	0	0	0	0	0	0	4	...	4	4	4	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
232833	57207860893	0	0	0	0	0	0	0	0	0	...	0	0	0	3
232834	57207860977	0	0	0	0	0	0	0	0	0	...	0	0	0	9
232835	57207866601	0	0	0	0	0	0	0	0	0	...	0	0	0	6
232836	57207868309	0	0	0	0	0	0	0	0	0	...	0	0	0	6
232837	57207872558	0	0	0	0	0	0	0	0	0	...	0	0	0	13

232838 rows × 36 columns

Figure 2.1: *Collaboration data for computer science authors.*

2.2.2 Filtering active authors

There are authors that can bring to a misleading analysis, for example those who have published just once before disappearing from the network or simply published too few to be considered representative. Because of them, in this section, will be given the definition of what an active author is, along with the concept of hole in publications.

Given an author A and an integer value $n \in \mathbb{N}$ called **hole size**, A has a **hole of**

size n in his publications if he stopped to publish for n consecutive years in his activity period, where the activity period are the set of years between his first and last publication. Follow that the **maximum hole size** of A is the maximum number of years he has passed without publishing.

An author is considered **inactive** for a given **hole size** if he has a **hole**, in his publication data, greater than the given **hole sizes**.
For example, given a **hole size** = **3**, in Fig.2.1, the author A1 is active but A2 is not, and their **maximum hole size** are respectively 3 and 4.

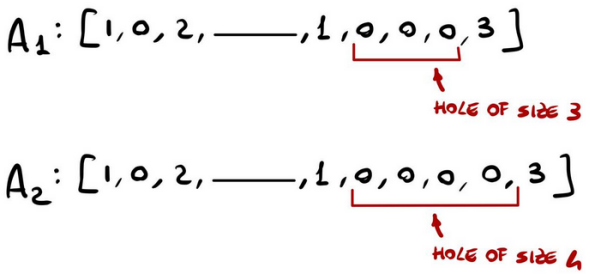

Figure 2.2: *Hole size definition.*

The hole size is not a sufficient metric to define active authors, there can be authors with hole size 0 that have published just once, so also the activity period must be used, lastly we can use a threshold on the minimum number of publications required to be considered active. So, in order to build vertex trajectories, and get a better understanding of the data, for each author, is needed the year in which they started to publish as well as the one in which they stopped; this activity period is also a way of making groups for future comparison.

A new version of the collaboration dataset is built upon the one described in Section 2.2.1 (Fig.2.1), containing new columns with: the starting and ending publication year, the activity period, the maximum hole size and the number of publications for each author (Fig.2.3).

	ID	1990	1991	1992	1993	1994	1995	1996	1997	1998	...	2014	2015	2016	2017	2018	start_year	end_year	max_hole_size	activity	tot_pubs
0	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	2000	2000	0	0	1
1	8	0	0	0	5	5	5	5	5	5	...	5	5	5	5	5	1993	1993	0	0	1
2	9523235	0	0	0	0	0	0	0	0	0	...	0	0	0	14	14	2017	2017	0	0	2
3	6503845773	0	0	0	0	0	0	0	0	0	...	3	3	3	3	3	2001	2001	0	0	1
4	6503846520	0	0	0	0	0	0	0	0	4	...	4	4	4	4	4	1998	1998	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
232833	57207860893	0	0	0	0	0	0	0	0	0	...	0	0	0	0	3	2018	2018	0	0	1
232834	57207860977	0	0	0	0	0	0	0	0	0	...	0	0	0	0	9	2018	2018	0	0	1
232835	57207866601	0	0	0	0	0	0	0	0	0	...	0	0	0	0	6	2018	2018	0	0	1
232836	57207868309	0	0	0	0	0	0	0	0	0	...	0	0	0	0	6	2018	2018	0	0	1
232837	57207872558	0	0	0	0	0	0	0	0	0	...	0	0	0	0	13	2018	2018	0	0	1

232838 rows × 35 columns

Figure 2.3: *Collaboration data with starting, ending year, activity period and total publications number.*

The distribution of the number of authors by their starting year has been plotted in Fig.2.4 and Fig.2.5, respectively. Notice that the data doesn't contain information about the years before 1990 and after 2018. So, those authors who started publishing in 1990 in the given data, have probably started before, as well as those who stopped publish in 2018 may still be active also nowadays.
Those plots shows also that there are more new active authors who start publishing each year, and that the number of authors who stop to publish is also increasing, but with a lower rate.

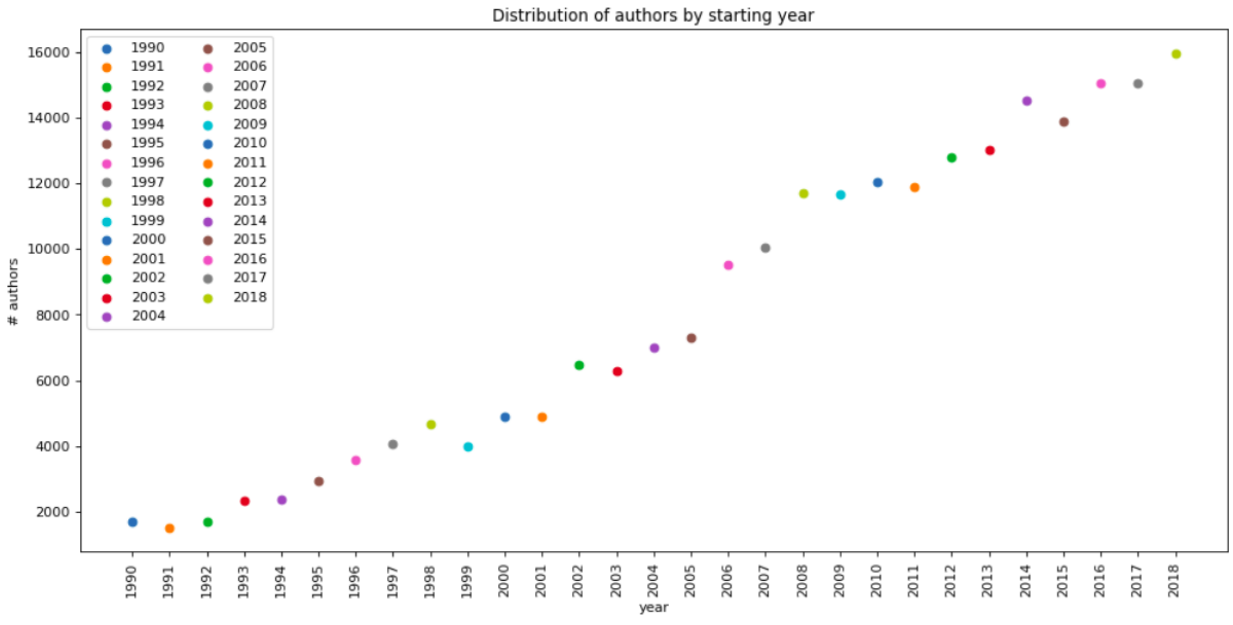

Figure 2.4: *Distribution of authors by starting year*

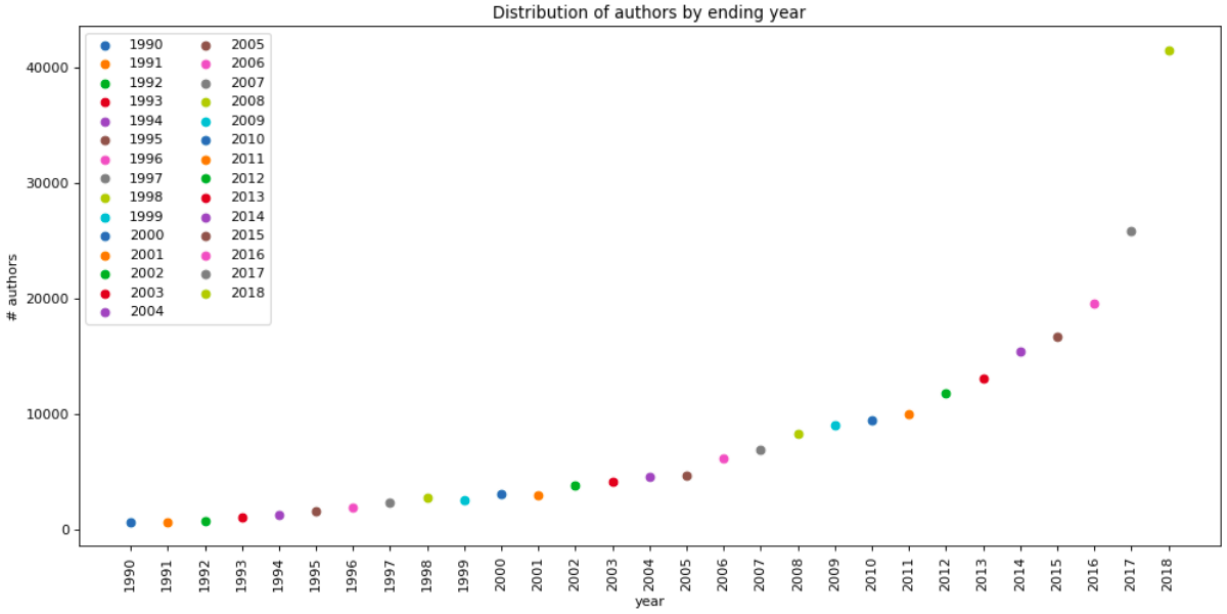

Figure 2.5: *Distribution of authors by ending year*

2.2.3 Identifying active authors

Next, values have been chosen for the previously described metrics: maximum hole size, minimum activity period and minimum publications number. In order to identify a sufficiently large and meaningful subset of active authors have been chosen an activity period of at least five years, to include all students that didn't stop the research activity after their Phd, an hole size of at most 7 years, to include those researchers who take a sabbatical year to do research every seven, years in which they teach, and lastly at least three publications are required to be active, this value has been chosen to be able to obtain a subset containing the 16% of the data. .

	ID	1990	1991	1992	1993	1994	1995	1996	1997	1998	...	2014	2015	2016	2017	2018	start_year	end_year	max_hole_size	activity	tot_pubs
6	6503847168	0	0	0	0	0	0	0	0	0	...	12	12	15	16	21	2004	2018	4	14	14
8	6503849838	0	0	0	0	0	0	0	0	0	...	13	13	13	32	32	2006	2017	7	11	4
20	6503858724	0	0	0	0	0	0	0	0	0	...	16	16	16	16	16	1999	2013	7	14	5
31	6503866265	0	0	0	0	0	0	0	0	0	...	20	20	20	20	20	2002	2012	6	10	3
70	6503889335	0	0	0	0	0	0	0	0	0	...	16	20	20	25	25	2004	2018	3	14	21
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
232590	57207536959	0	0	0	0	0	0	0	0	0	...	30	30	46	46	47	2009	2018	4	9	16
232623	57207585229	0	0	0	0	0	0	0	0	0	...	3	3	3	3	3	2009	2016	6	7	4
232638	57207598135	0	0	0	0	0	0	0	0	0	...	11	11	25	25	27	2009	2018	6	9	6
232647	57207604191	0	0	0	0	0	0	0	0	0	...	25	25	25	25	25	2009	2016	4	7	11
232654	57207607528	0	0	0	0	0	0	0	0	0	...	15	15	25	25	30	2009	2018	4	9	18

36795 rows × 35 columns

Figure 2.6: *Identified subset of active author*

2.2.4 Changing definition of event

Until now as time steps for the evolution of the collaboration graph have been considered 28 years, since 1990 to 2018.

The theoretical model, explained in the state of the art (Section 2) of this report, as a time step consider an event, where an event can be the appearance of a new node or a new edge, so the set of year, containing only 28 time step, can be too small in order to build meaningful vertex trajectories, that's why in next sections, other metrics are used: the occurrence of a new author and a new collaboration as time steps.

Applying the described metrics results in a stretching of the x axis in the plotted data.

Their distributions are showed in Fig.2.7 and Fig.2.8.

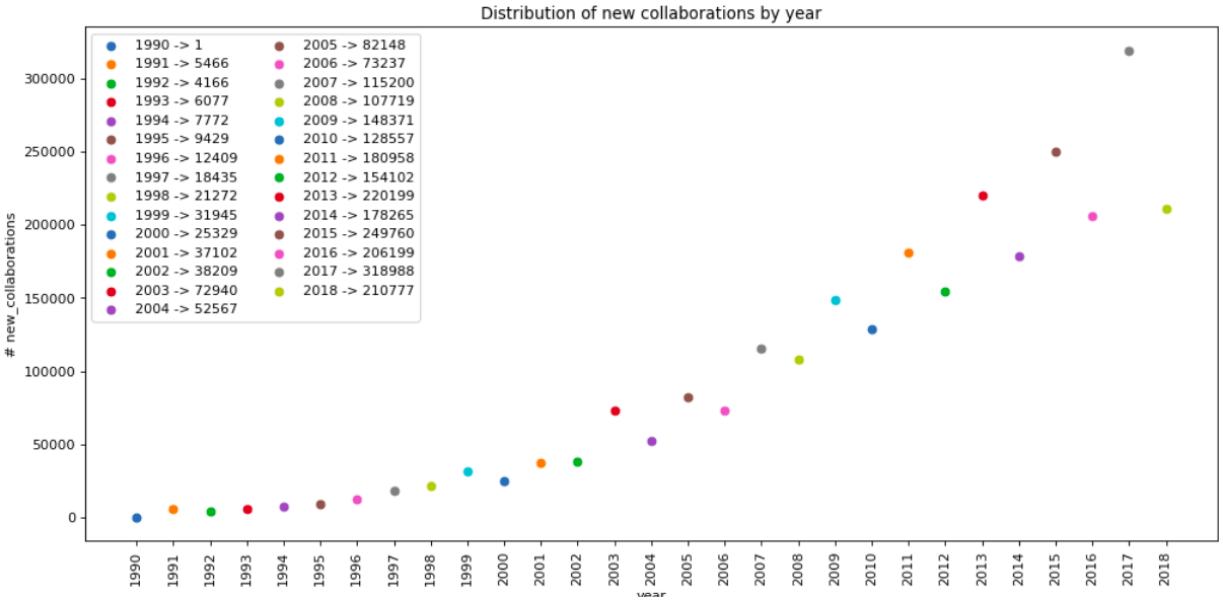

Figure 2.7: *Distribution of new collaborations by year*

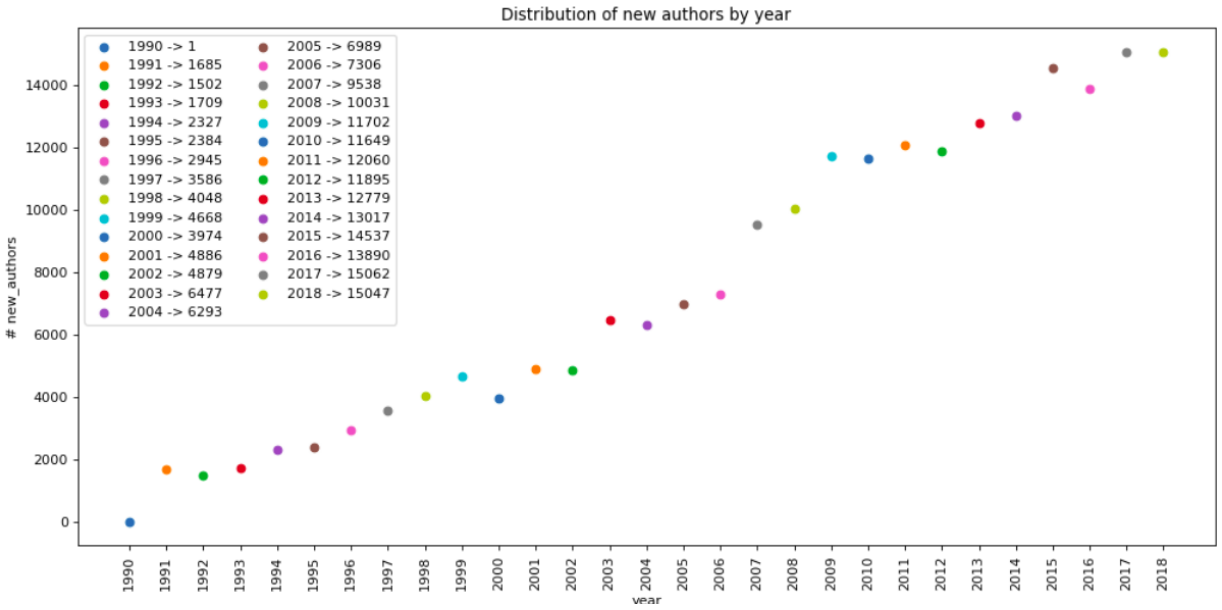

Figure 2.8: *Distribution of new authors by year*

2.2.5 Getting to know data characteristic

2.2.6 Plotting Ratios

Here, for a better understanding the character of the data, remembering that stretching permits comparisons with the theoretical model described in the state of the art (Section 2.1), have been computed two ratios: the ratio between the number of new collaborations and new authors (Fig.2.9, 2.10, 2.11) and the ratio between total number of collaborations and authors (Fig.2.12, 2.13, 2.14) that represents the behavior over time of the average degree of the collaboration network.

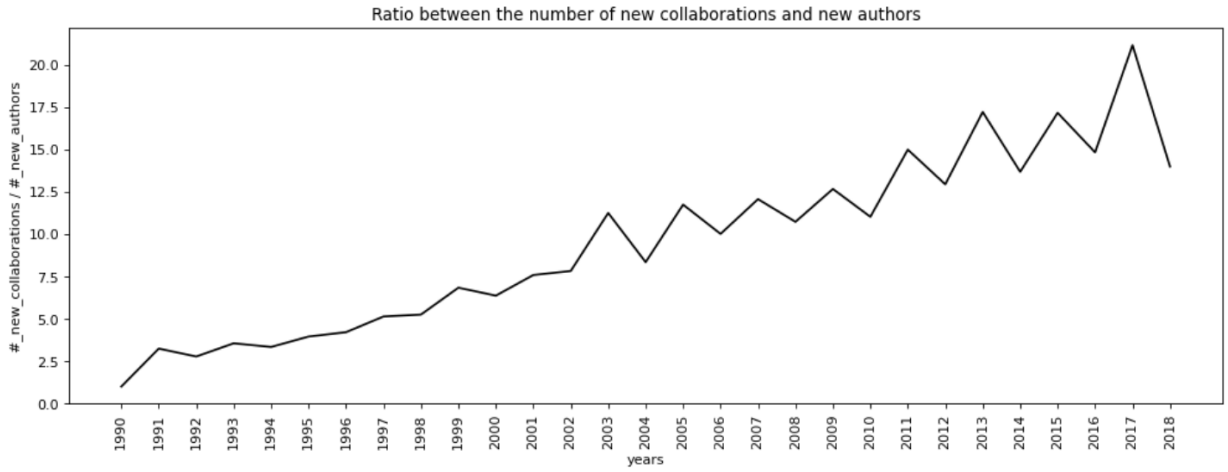

Figure 2.9: *Ratio between the number of new collaborations and new authors.*

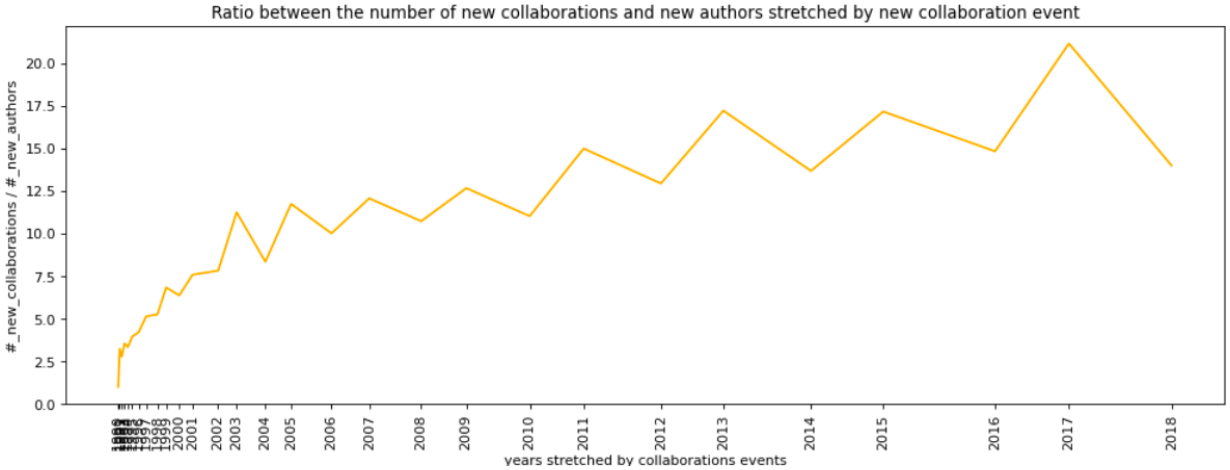

Figure 2.10: *Ratio between the number of new collaborations and new authors - stretched by collaborations.*

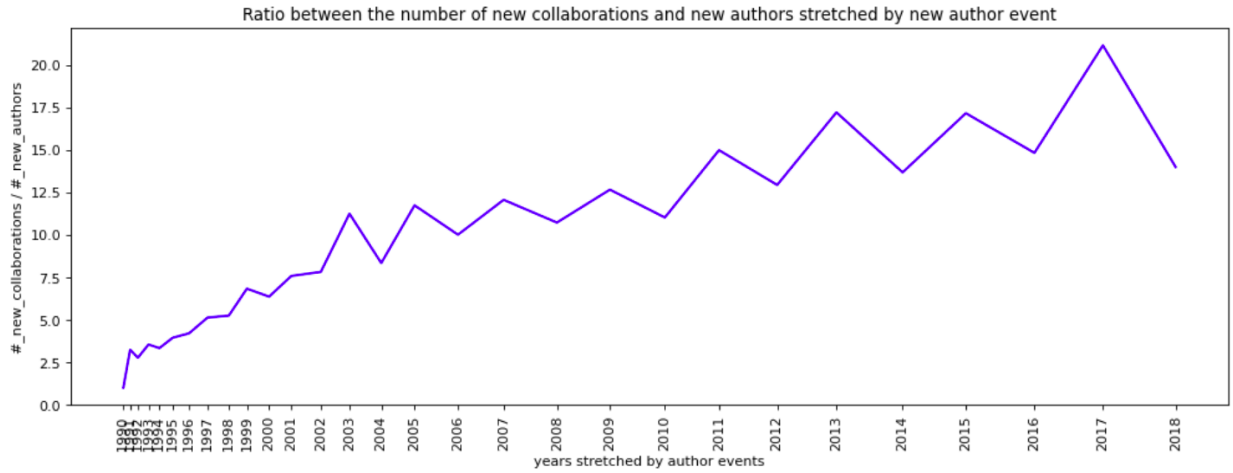

Figure 2.11: *Ratio between the number of new collaborations and new authors stretched by authors.*

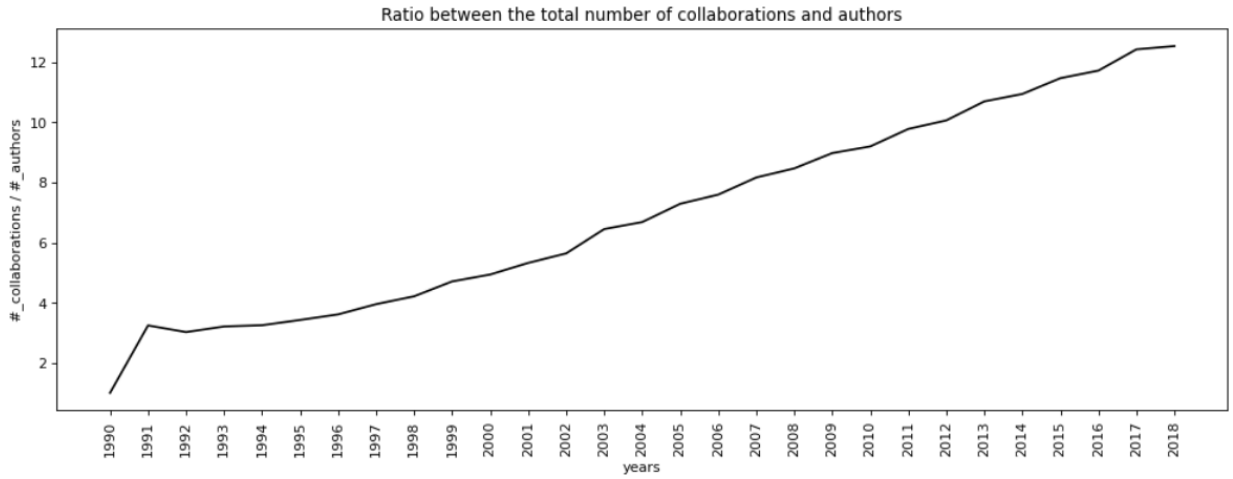

Figure 2.12: *Ratio between the total number of collaborations and authors.*

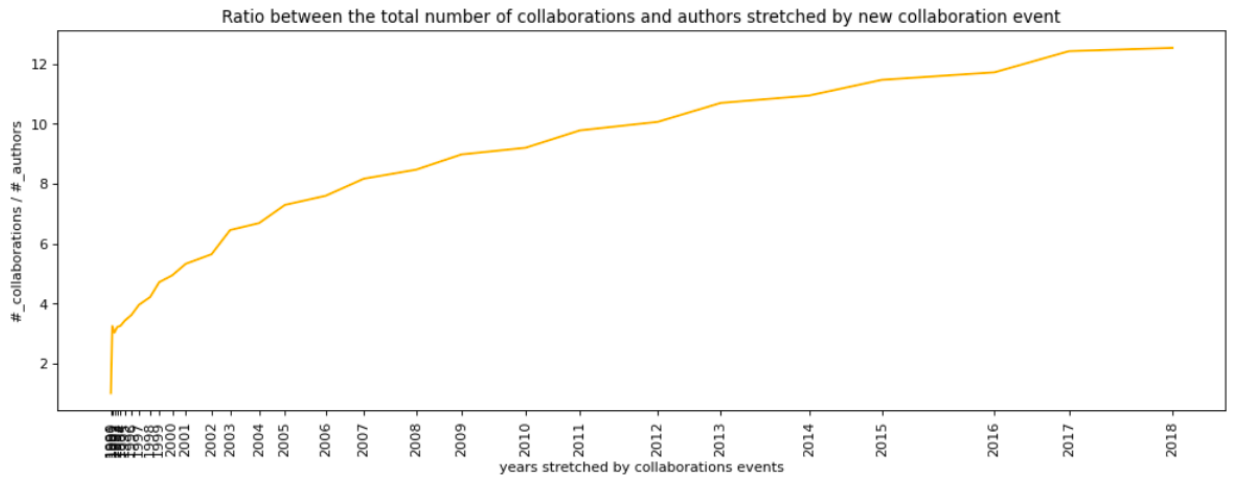

Figure 2.13: *Ratio between the total number of collaborations and authors - stretched by collaborations.*

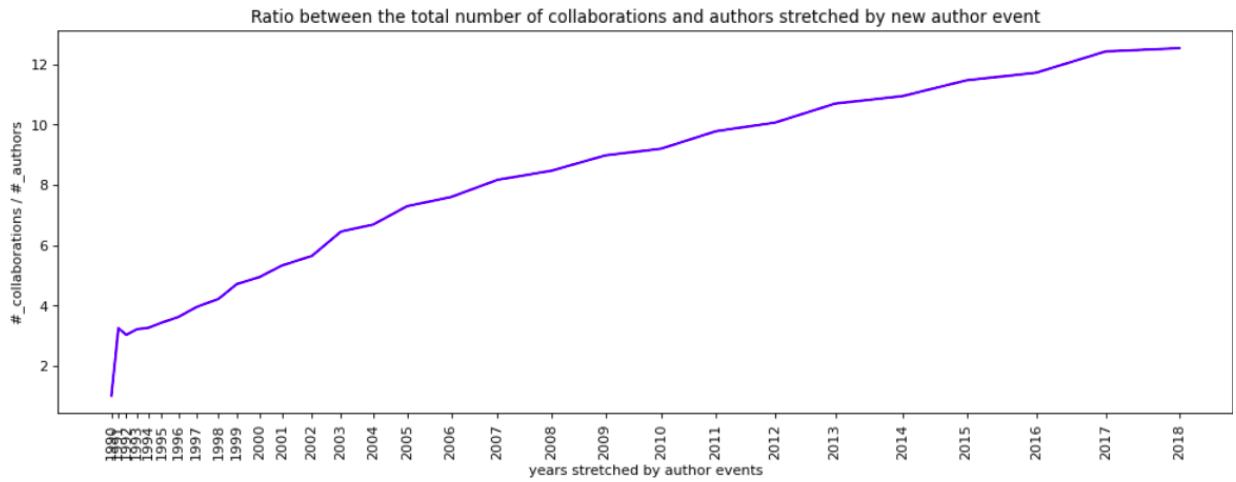

Figure 2.14: *Ratio between the total number of collaborations and authors - stretched by authors.*

By Fig.2.12, 2.13, 2.14 can be noticed that the average degree grows, while most of theoretical models assume that it is constant over time, e.g., Barabasi-Albert [1]; we add a new edge per step, thus at time t we have sum of degrees $2t$, thus the average degree is $\frac{2t}{t} = 2$, which is constant.

2.2.7 Degree Distribution

The degree distribution is found by taking the number of authors that have a given amount of publication, and it's computed to acquire more knowledge from the character of the data.

Fig.2.15 shows on the y axis the number of authors with a total number of collaboration equal to the one indicated in the x axis, so the degree distribution; while Fig.2.15 contains it's log-log form.

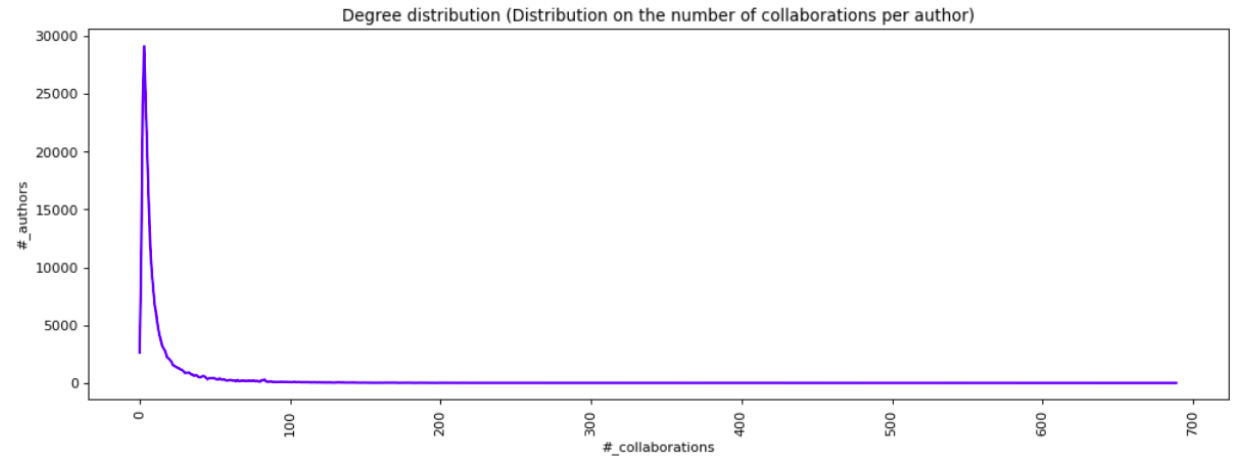

Figure 2.15: *Degree distribution.*

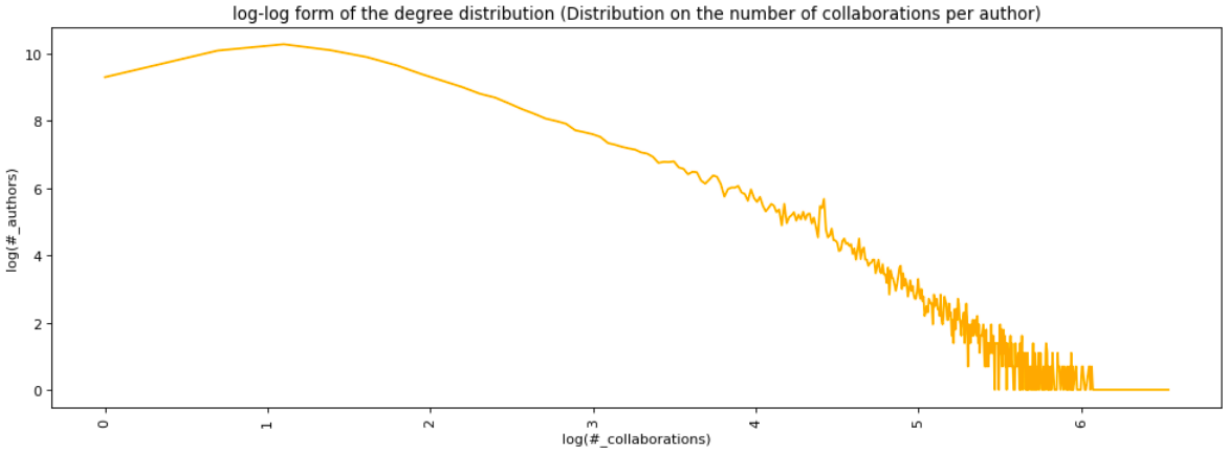

Figure 2.16: *Degree distribution in it's log-log form.*

Conclusions

We have have solve all the problems and created a new ones.

Appendices

Appendix A

General details

It is known that ...

Acknowledgements

I would like to thanks all the people that in some way have helped me with the elaboration of this thesis.

First, I would like to thanks ....

Nice, 2022

Leonardo Serilli

Bibliography

Abbreviations

BAO	Baryon acoustic oscillation
CDM	Cold dark matter
CMB	Cosmic microwave background
CTA	Cherenkov telescope array
DM	Dark matter
EW	Electroweak
EWPO	Electroweak precision observables
FCNC	Flavor changing neutral currents
GUT	Gran unified theory
GDE	Gamma diffuse emission
GC	Galactic center
GCE	Galactic center excess
ICS	Inverse Compton scattering
ILC	International linear collider
IH	Inverse hierarchy
LHC	Large hadron collider
LUX	Large underground Xenon experiment
LEP	Large Electron-Positron Collider
LFV	Lepton flavor violation
LOP	Lightest odd particle
LAT	Large area telescope
LZ	LUX-Zeplin experiment
MSM	Millisecond pulsar
MSSM	Minimal supersymmetry standard model
NLO	Next to leading-order
NH	Normal hierarchy
NFW	Navarro-Frenk-White
SDFDM	Singlet-doublet fermion dark matter
SSDM	Singlet scalar dark matter
SI	Spin independent
SD	Spin independent
SM	Standard model
WIMP	Weakly interacting massive particle