PFE Report

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. 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.

Contents

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 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 - 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.

1.2 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) [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.

1.3 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.

1.4 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.

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

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

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:

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

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.

$$\Pr[v \text{ is chosen}] = \frac{f(\deg_t(v))}{\sum_{w \in V_t} f(\deg_t(w))}.$$

In the equation $deg_v(t)$ 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 P_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.

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

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

3 Data Preparation

3.1 Retrieving Data

The first part of the project concerned the retrieval of data about collaborations and publications 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. and Fig.2. In both figures, each row represent an author with his ID on the Scopus database. In the collaboration data (Fig.1), for each year, each cell shows the cumulative number of his collaborations until each year, while in the publication data (Fig.2), for each year, is represented the number of his publications for the given year, not the cumulated one.

Collaboration 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 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	7003355588	2	2	2	1	4	0	5	5	0	...	7	4	4	15	11	7	11	9	8	6
1	56522848500	3	0	1	0	2	0	6	1	3	...	3	5	6	1	0	0	1	1	1	4
2	7004165433	5	1	1	2	10	5	6	2	6	...	4	3	11	7	6	10	6	3	3	4
3	6603870889	1	0	2	0	1	2	6	4	2	...	8	10	7	20	16	12	9	10	15	16
4	7005944861	10	10	3	7	8	8	4	15	9	...	9	8	12	10	20	19	17	12	7	5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
232833	57200496797	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	2
232834	15137130100	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
232835	57196721826	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
232836	57196401698	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
232837	57195980869	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1

232838 rows × 30 columns

Figure 2: Publication data for computer science authors.

3.2 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 $n \in N$ called **hole size**, A has a **hole of size** n in his publications if he stopped to publish for n consecutive years. 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.3, the author A1 is active but A2 is not, and their **maximum hole size** are respectively 3 and 4.

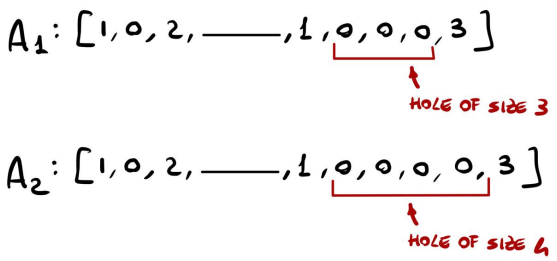

Figure 3: Hole size definition.

For each possible **hole size**, has been built a dataset where all authors considered inactive for the given value, have been filtered out, so a dataset is

built for each possible definition of Inactivity. For each of them the number of active authors is showed in Fig.4, while in Fig.5 is illustrated the difference in the number of active authors between two consecutive **hole size**.

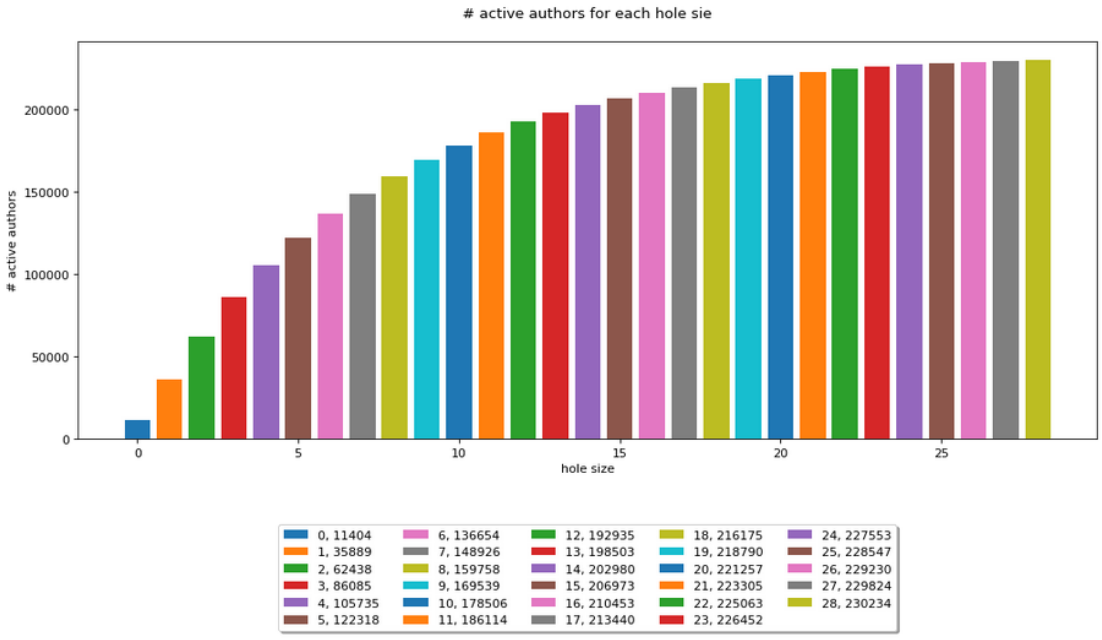

Figure 4: Number of authors kept for each hole size.

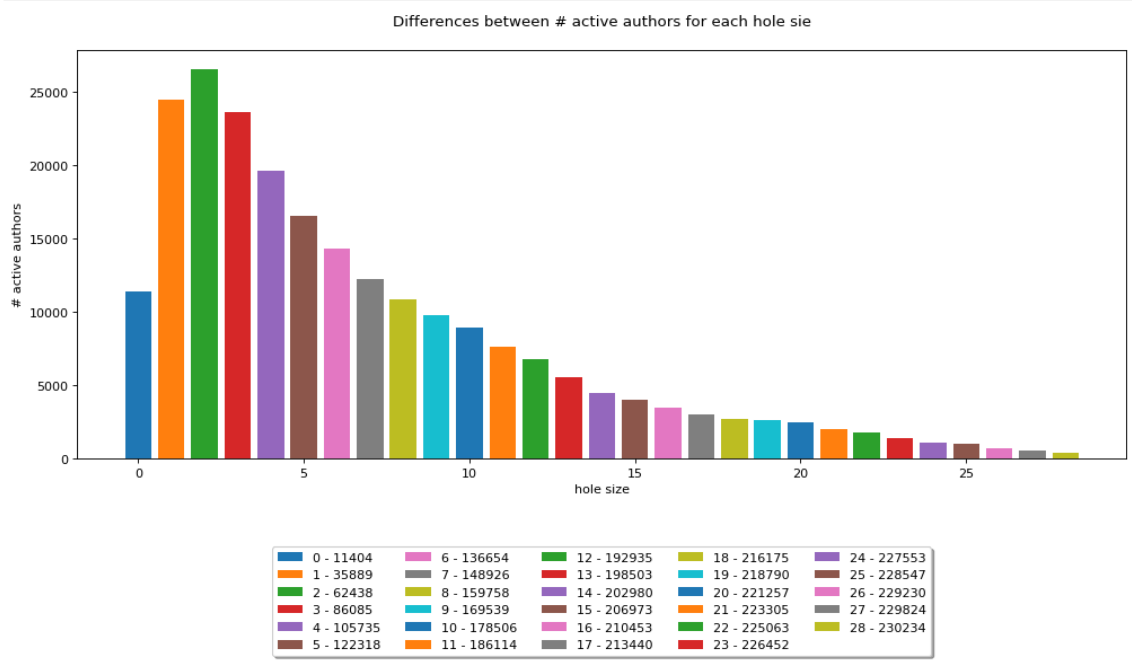

Figure 5: Differences in the number of authors for each hole size.

The same definition of active and inactive authors can be applied also to the collaboration data, such that an author is inactive for a given hole size n if he

has not collaborated for n consecutive years. Applying this definition, for each previously built dataset, we obtain two kind of authors **new collaborators** and **new publishers**: given a year, an author is a new collaborator if he started his first collaboration in the given year, and is a new publisher if he published the first time in the given year.

Next, to get to know better the character of the data, the distribution, for each hole size, has been plotted the number of new authors, both publisher (orange curve in Fig.6) and collaborators (blue curve in Fig.6). As result has been obtained that their distributions doesn't differ too much.
An example of those plots is given in Fig.6, for other hole sizes the difference between the two curves is almost the same.

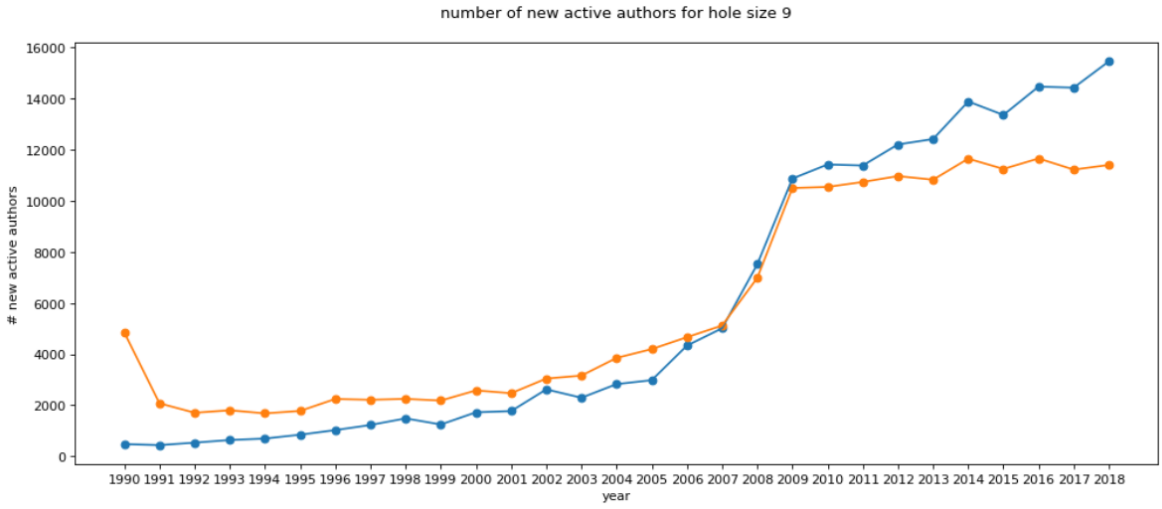

Figure 6: Distribution of new collaborators and new publisher for each year and hole size 9.

3.3 Retrieving starting and ending publication year

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 year they stopped; this is also a way of making groups for future comparison.

Two versions of the collaboration dataset are built upon the one described in Section 3.1.1 (Fig.1), containing respectively a column with the starting publication year (Fig.7) and the ending publication year (Fig.9) for each author.

The distribution of the number of authors by their starting year has been plotted in Fig.8 and Fig.10, respectively. Show an **unnatural increase** in the years 1990 and 2018; this is because 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 pub-

lishing each year, and that the number of authors who stop to publish is also increasing, but with a lower rate.

Collaboration data with starting publication year

	ID	1990	1991	1992	1993	1994	1995	1996	1997	1998	...	2010	2011	2012	2013	2014	2015	2016	2017	2018	start_year
0	8958327900	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	2000
1	6508297663	0	0	0	0	0	0	0	0	0	...	7	7	8	8	8	8	8	8	8	1995
2	7004267341	0	0	0	0	0	0	0	0	0	...	10	10	16	16	16	16	16	16	16	2008
3	8642393600	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	7	7	7	7	2015
4	55873955900	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	8	8	8	8	2014
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
232833	6507630481	0	0	0	0	0	0	0	0	0	...	18	18	18	18	29	29	29	29	29	2002
232834	24577815500	0	0	0	0	0	0	0	0	0	...	4	4	4	6	13	16	16	16	70	2003
232835	57195243976	0	0	0	0	0	0	0	0	0	...	3	3	3	3	3	3	3	8	8	2017
232836	35328962100	0	0	0	0	0	0	0	0	0	...	0	0	0	0	2	2	2	2	3	2010
232837	7403521415	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	29	29	29	2016

232838 rows × 31 columns

Figure 7: Collaboration data with starting year

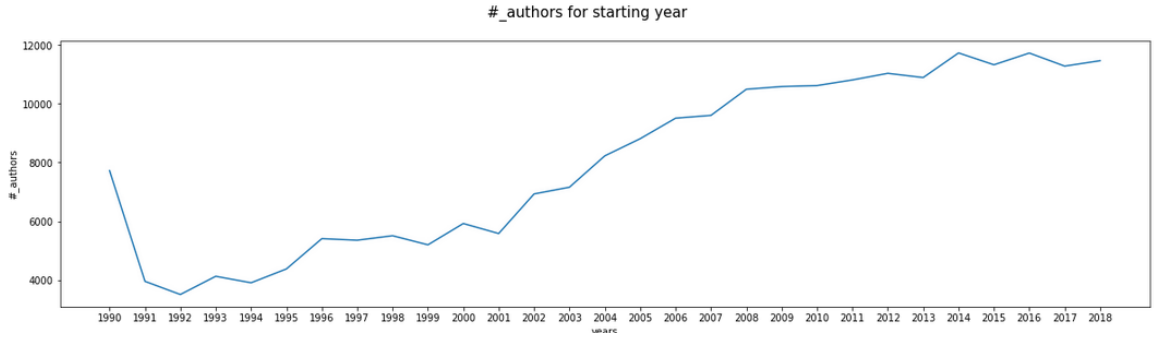

Figure 8: Distribution of authors by starting year

Collaboration data with ending publication year

	ID	1990	1991	1992	1993	1994	1995	1996	1997	1998	...	2010	2011	2012	2013	2014	2015	2016	2017	2018	ending_year
0	8958327900	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	2000
1	6508297663	0	0	0	0	0	0	0	0	0	...	7	7	8	8	8	8	8	8	8	2016
2	7004267341	0	0	0	0	0	0	0	0	0	...	10	10	16	16	16	16	16	16	16	2015
3	8642393600	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	7	7	7	7	2018
4	55873955900	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	8	8	8	8	2015
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
232833	6507630481	0	0	0	0	0	0	0	0	0	...	18	18	18	18	29	29	29	29	29	2015
232834	24577815500	0	0	0	0	0	0	0	0	0	...	4	4	4	6	13	16	16	16	70	2018
232835	57195243976	0	0	0	0	0	0	0	0	0	...	3	3	3	3	3	3	3	8	8	2017
232836	35328962100	0	0	0	0	0	0	0	0	0	...	0	0	0	0	2	2	2	2	3	2018
232837	7403521415	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	29	29	29	2017

232838 rows × 31 columns

Figure 9: Collaboration data with ending year

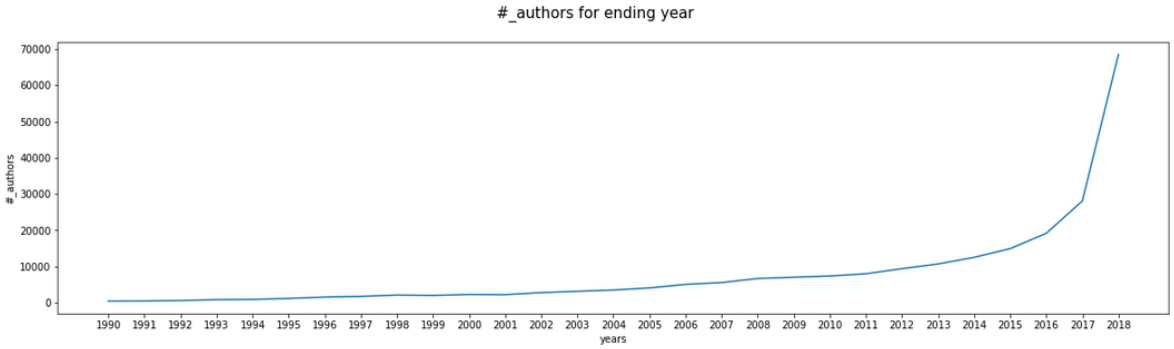

Figure 10: Distribution of authors by ending year

3.4 Splitting data by starting year

Next, to build vertex trajectories, each hole size based dataset is split in 28 subsets based on the starting year of each author.

The set of datasets associated with a hole size has a distribution chart Fig.11-12-13-14 showing the number of authors for each starting year. In those figures can be seen that the distribution stay very similar in each case.

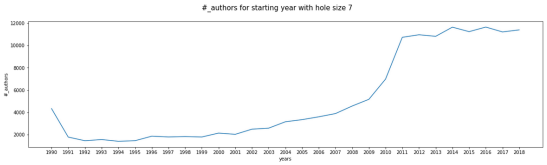

Figure 11: Distribution of authors with hole size 7 by starting year

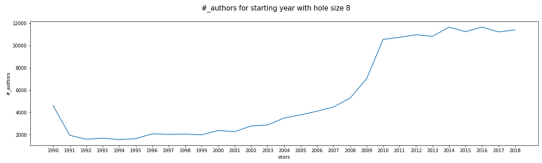

Figure 12: Distribution of authors with hole size 8 by starting year

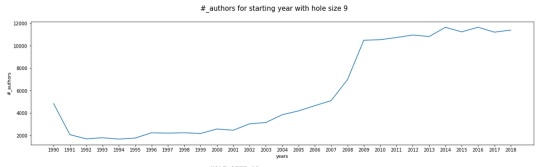

Figure 13: Distribution of authors with hole size 9 by starting year

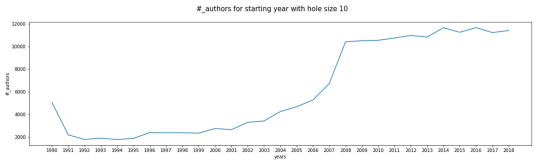

Figure 14: Distribution of authors with hole size 10 by starting year

3.5 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 publication, of a new author or 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.15, Fig.16 and Fig.17.

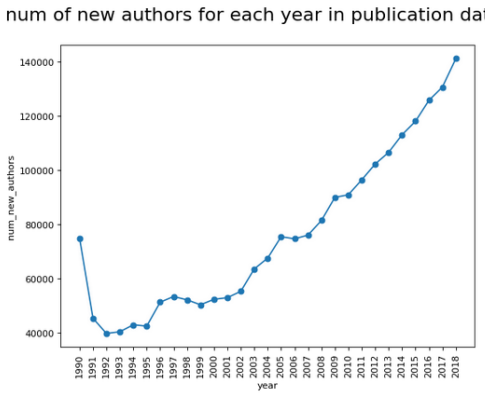

Figure 15: Distribution of new authors by year in the publication dataset

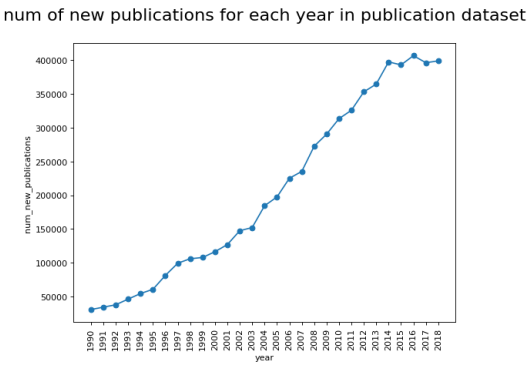

Figure 16: Distribution of new publication by year in the publication dataset

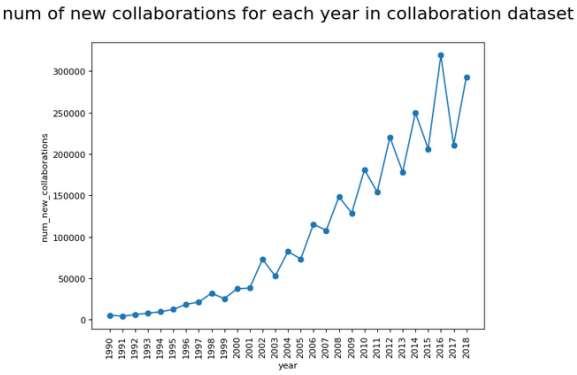

Figure 17: Distribution of new collaborations by year in the publication dataset

3.6 Getting to know data characteristic

3.7 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), have been computed two ratios:

- The ratio between the number of new collaborations and new authors (Fig.18).

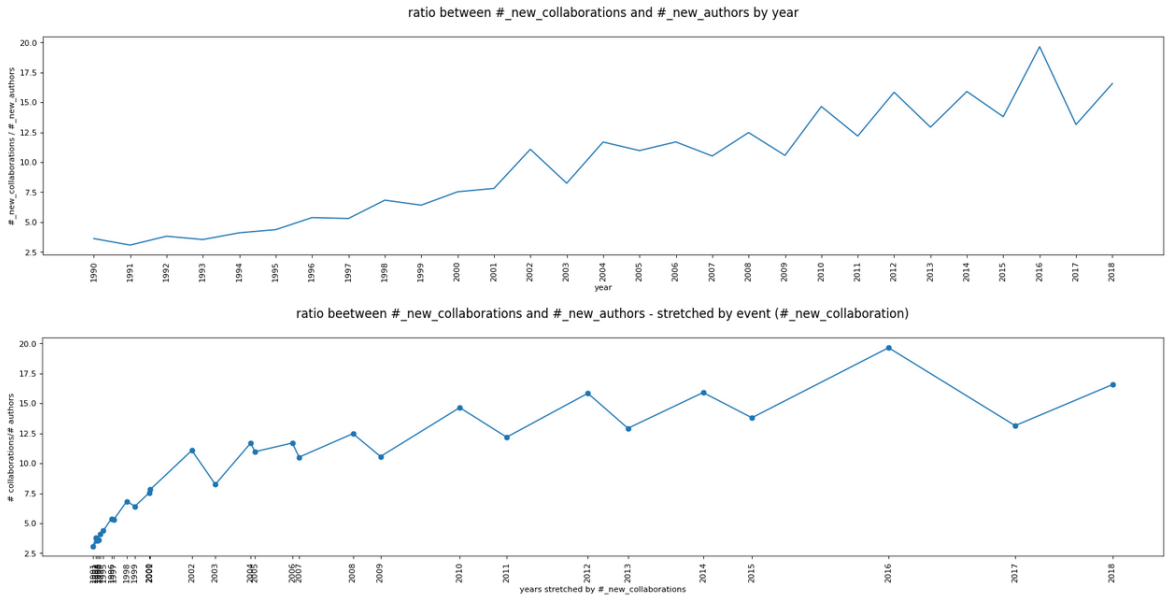

Figure 18: Ratio between the number of new collaborations and new authors.

- The ratio between total number of collaborations and authors (Fig.19) that represents the behavior over time of the average degree of the collaboration network.

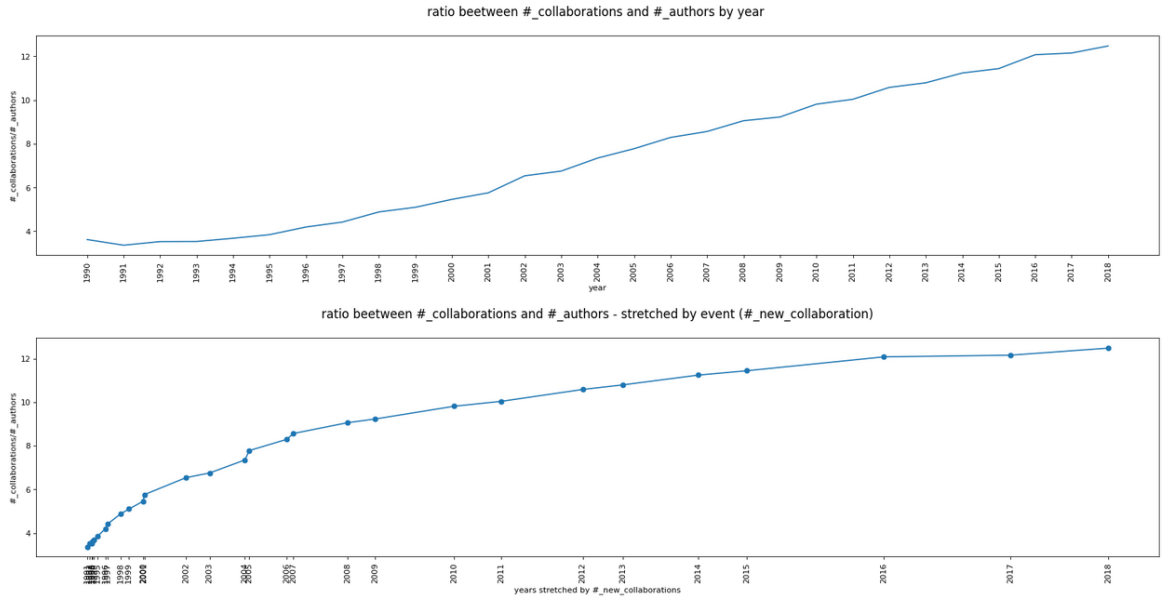

Figure 19: Ratio between total number of collaborations and authors.

By Fig.19 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.

3.8 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.20 shows, in the first plot, 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 the second plot contains it's log-log form.

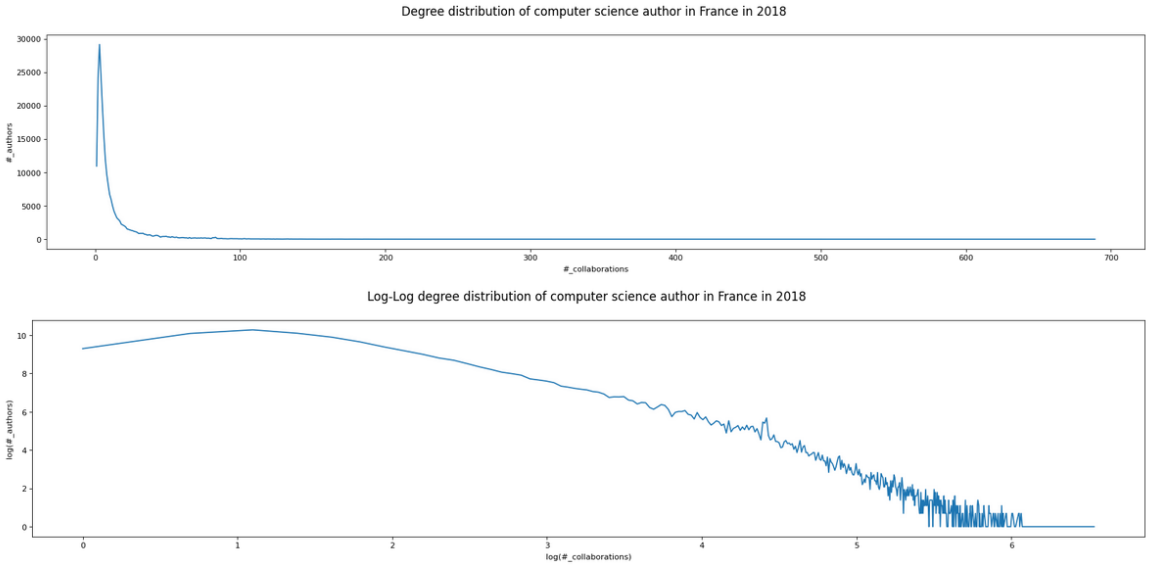

Figure 20: Degree distribution and it's log-log form.

Next has been tried to fit the given distribution with the power-law distribution described in the state of the art (Section 2), both the classic power-law and the one with exponential cut-off.

From Fig.21 can be concluded that a power-law distribution with an exponential cut-off better fits the given data.

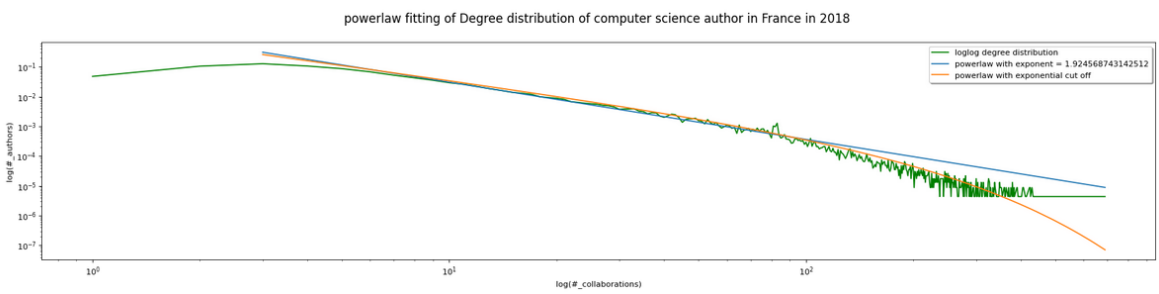

Figure 21: Power-law fitting

Performing the same fitting for data of active authors brings to the same result, that a power-law with exponential cut-off is better for fitting it, not depending on the chosen hole size.

4 Vertex Trajectories

4.1 Plotting vertex trajectories

The main task of this work is the investigation of **vertex trajectories** in the collaboration graph built upon scientific authors and their collaboration. Defined in the state of the art (Section 2), the trajectory of a vertex is the sequence representing the evolution over time of the degree of the vertex.

In Fig.22 a chart containing trajectories of each active author, with a hole size less than 2 (Section 3.1.2), is plotted. It contains the vertex trajectories of the 100 active authors who reached the highest total number of collaborations in the respective dataset.

In the next section all trajectories will be averaged over authors with the same starting year, so that they will refer to the theoretical trajectory (Section 2 - State of the art).

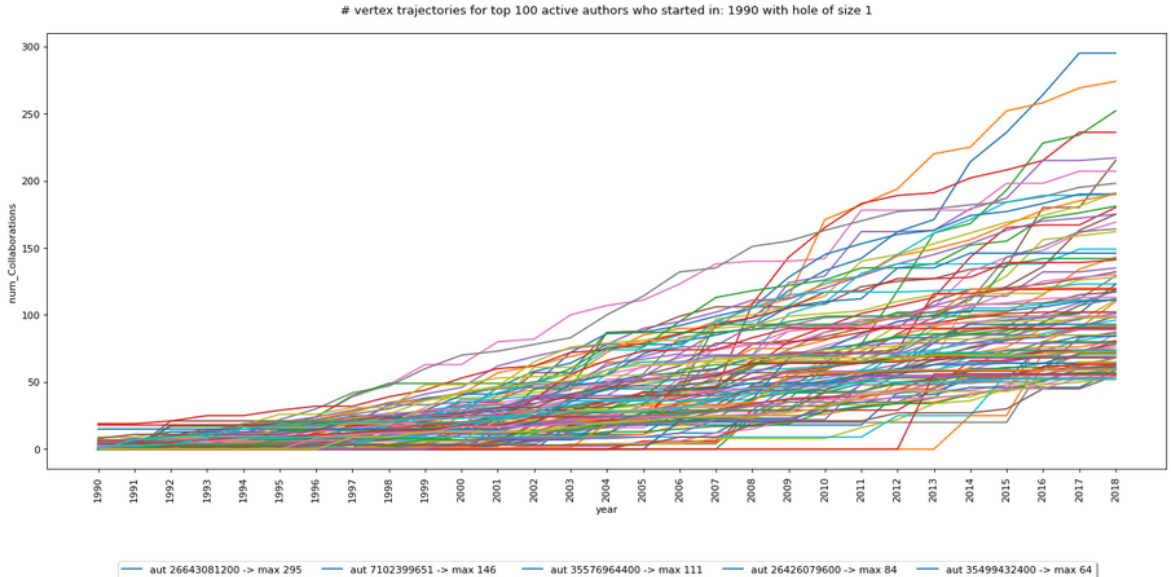

Figure 22: Trajectory for top 100 authors with hole size 1 with starting year 1990

The chart in Fig.22 is stretched by using firstly the occurrence of a new author (Fig.23), then the occurrence of a new collaboration (Fig.24), as an event instead of the years (Section 3.1.5 - Changing definition of event).

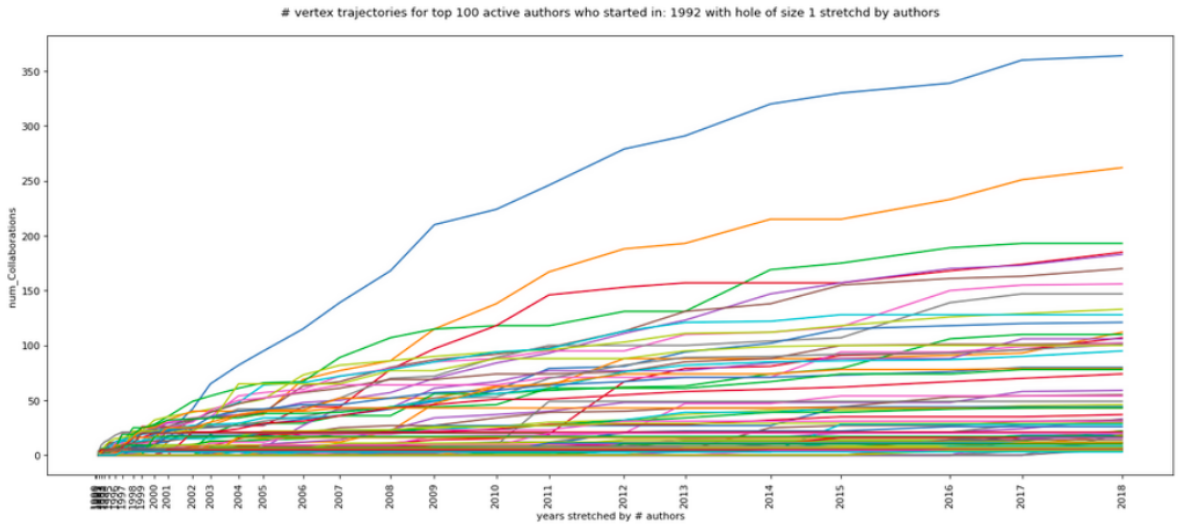

Figure 23: Trajectory for top 100 authors with hole size 1 with starting year 1990 - stretched by #authors

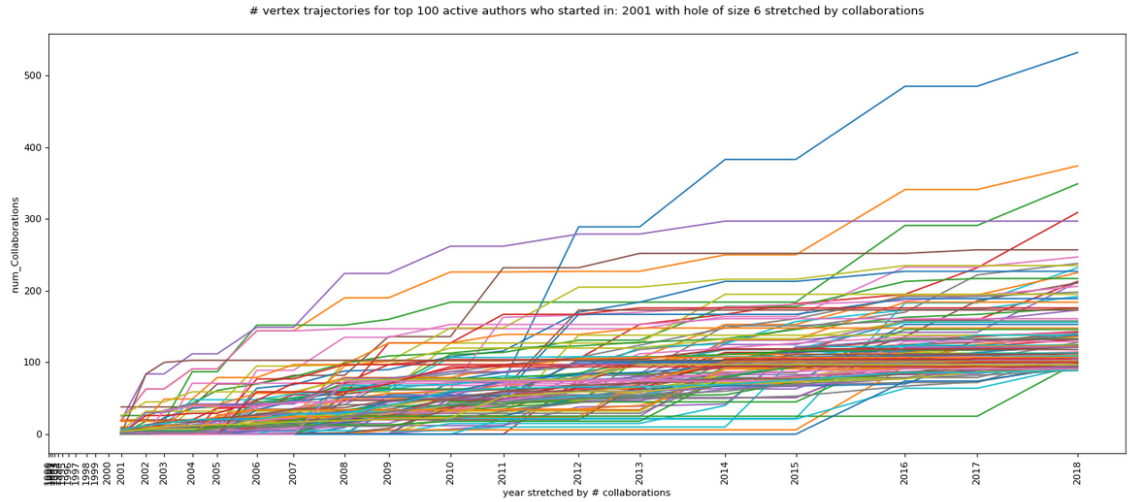

Figure 24: Trajectory for top 100 authors with hole size 1 with starting year 1990 - stretched by #collaborations

4.2 Plotting average vertex trajectories

In the section, in order to refer to the theoretical model described in the state of the art (Section 2), is computed and plotted the average trajectory of active authors by starting year.
In Fig.25 is plotted the set of average trajectories for active authors with a hole size less equal than 8, where for each color is associated a starting year.
In the next sections will be made an attempt of fitting those trajectories with different kind of functions, like those functions previously described in the theoretical model (Section 2 - State of the art).

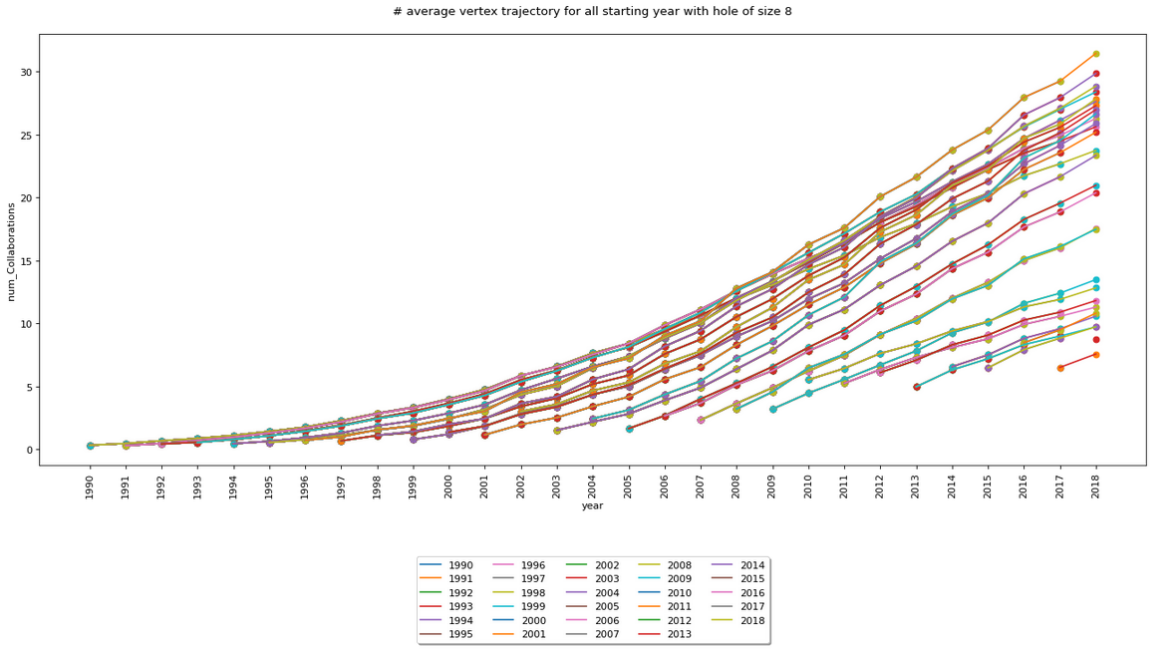

Figure 25: Average trajectories given hole size 8

The above chart is stretched by using firstly the occurrence of a new author (Fig.26), then the occurrence of a new collaboration (Fig.27), as an event instead of the years (Section 3.1.5 - Changing the definition of event).

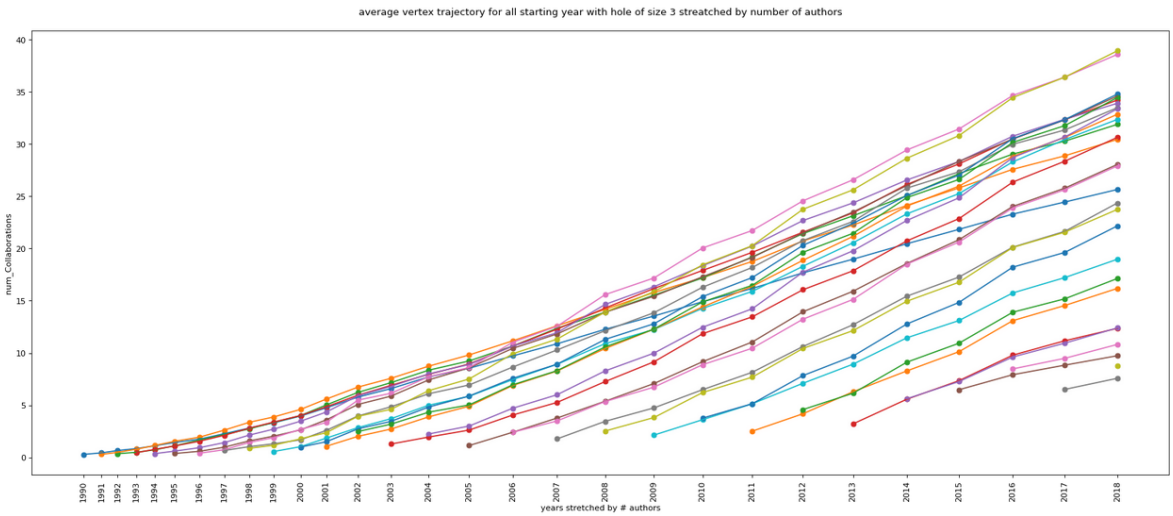

Figure 26: Average trajectories given hole size 8 - stretched by #authors

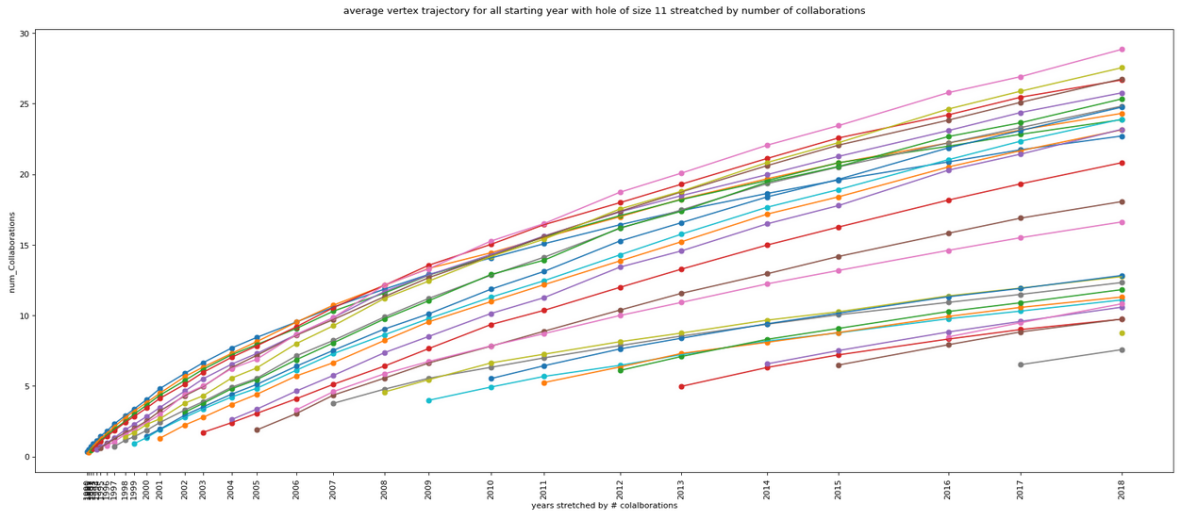

Figure 27: Average trajectories given hole size 8 - stretched by #collaborations

4.3 Fitting average vertex trajectories

For each hole size, each the average trajectory by starting year is fitted, with a logarithmic function of the following form $f(t) = \alpha * ln(t)^{\gamma}$, this introductory fitting is done in order to move, later in the section, on the more detailed functions described in the state of the art (Section 2) that refers to the theoretical model with degree distribution following a power-law with an exponential cut-off.

4.3.1 Logarithmic fitting

In Fig.28 the average trajectory, for active authors with a hole size less equal than one who started publishing in 1993, is fitted using the following function $f(t) = \alpha * ln(t)^{\gamma}$, the same chart is than stretched using the occurrence of a new collaboration as event (section 3.1.5 - Changing the

definition of event).

The results, also found for hole sizes and starting years other then the one in Fig.28, show that logarithmic functions are well suited for fitting our trajectories for each hole size and starting year.

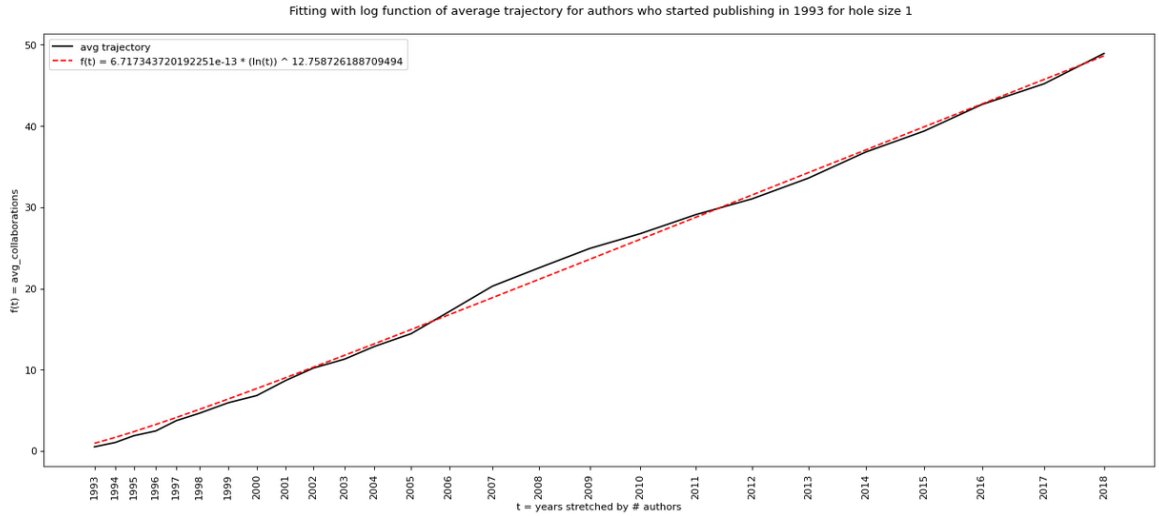

Figure 28: Logarithmic fitting of average trajectory given hole size 1 and starting year 1991 - stretched by #authors

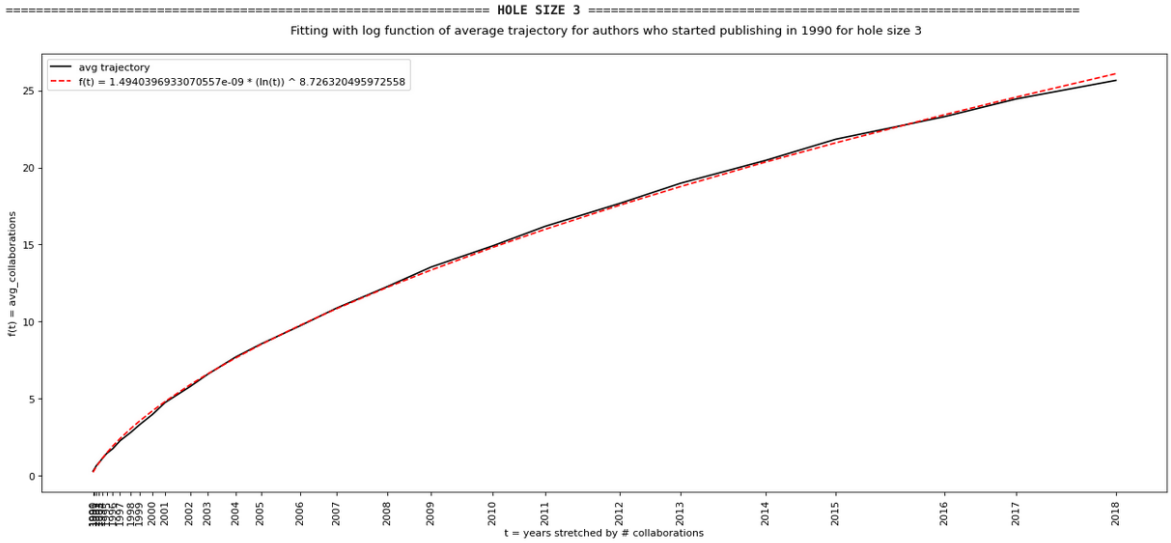

Figure 29: Logarithmic fitting of average trajectory given hole size 1 and starting year 1991 - stretched by #collaborations

4.3.2 Alpha-Sigma logarithmic fitting

In the section the data is fitted using the logarithmic function in Eq.1, that is exactly the one representing the theoretical vertex trajectory described in the state of the art (Section 2).

$$g_v(t) = \left(\alpha * ln\left(\frac{t}{t_v}\right) + 1\right)^{\sigma} \tag{1}$$

In Fig.30 are plotted the results of the fitting for each trajectory. The fitting works better for curves with a low starting year, because, of course, they contain enough data to show the logarithmic behavior we are trying to fit.

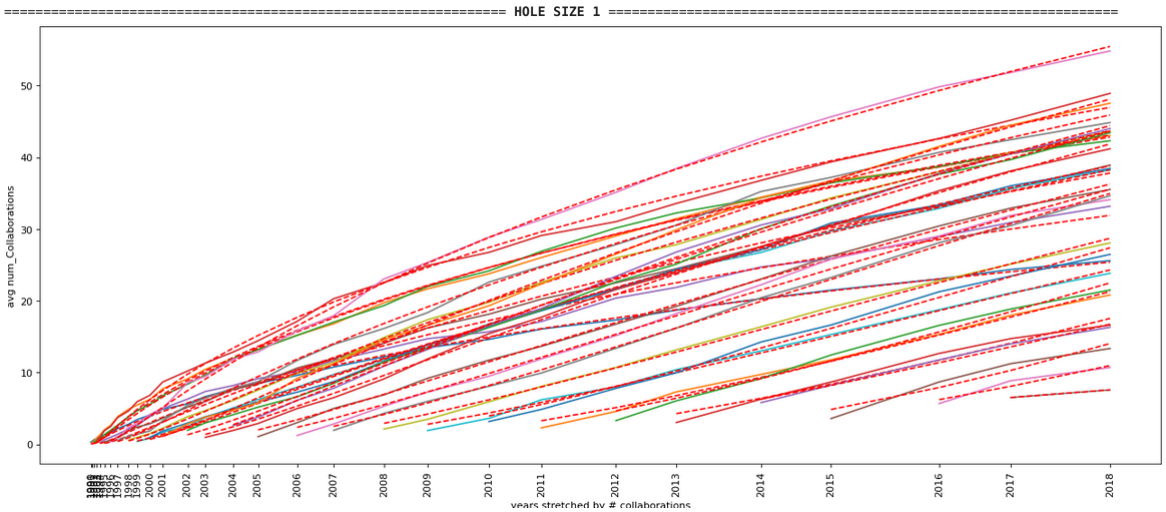

Figure 30: Alpha-sigma logarithmic fitting of average trajectory by starting year for hole size 1 - stretched by #collaborations

To better understand the fitted values for alpha and sigma parameters represented in Fig.31, they have been plotted in Fig.32, where can be noticed the following inconsistency: sigma is rather constant and alpha grows linearly. Later in this section, third parameter is introduced in the fitting function, in order to capture the linear behavior of alpha, so that a better fitting can be achieved (Section 4.3.4 - alpha-beta-sigma logarithmic fitting).

Sarting Year	alpha	sigma
1990	0.0118949	3.85334
1991	0.0374623	3.68901
1992	0.0774668	3.44761
1993	0.203102	3.0804
1994	0.113253	3.30335
1995	0.108768	3.55056
1996	0.446736	3.0516
1997	0.376532	3.16263
1998	0.459725	3.14889
1999	0.454745	3.16802
2000	0.680256	3.01732
2001	1.10439	2.92922
2002	1.31389	2.90598
2003	1.36367	2.96655
2004	2.72389	2.57455
2005	1.99424	2.81892
2006	2.34984	2.85393
2007	2.49052	2.97705
2008	2.907	2.83709
2009	2.77385	2.91749
2010	3.93424	2.93454
2011	3.25826	3.14932
2012	4.65518	3.10196
2013	4.25758	3.21608
2014	6.29613	2.79716
2015	4.83592	3.84465
2016	6.20906	3.37907
2017	6.50722	1.55522

Figure 31: Logarithmic fitting parameters alpha-sigma for each starting year and hole size 1 - stretched by #collaborations

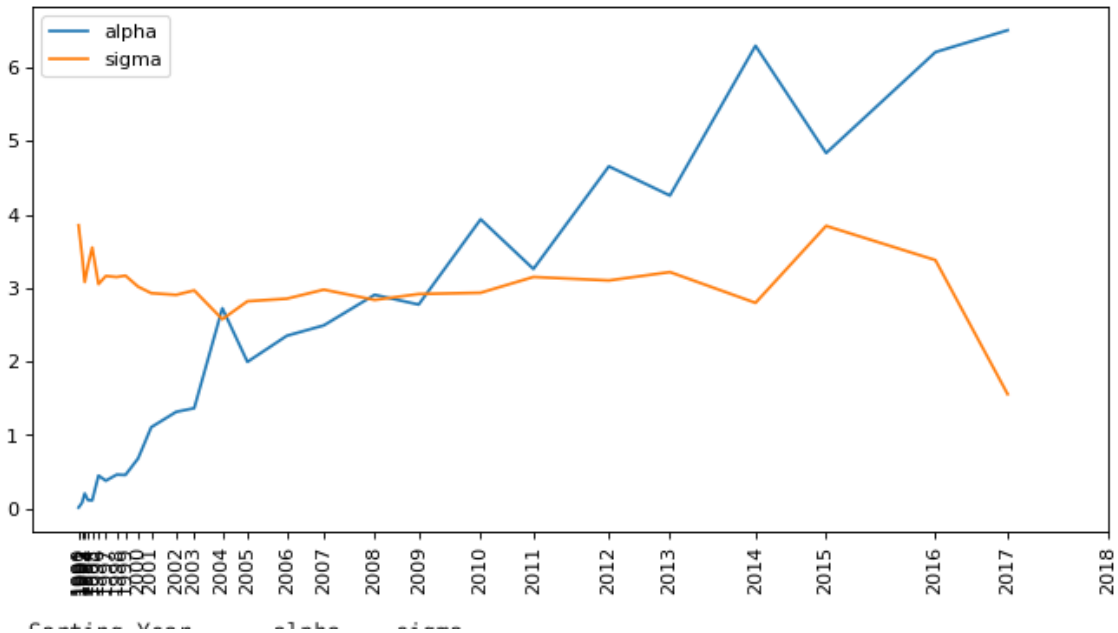

Figure 32: Plot of logarithmic fitting parameters alpha-sigma for each starting year and hole size 1 - stretched by #collaborations

Then has been tried to find the best couple of parameters able to fit all curves minimizing the total error on the fitting. The error is calculated summing up all squared absolute differences between the fitted values and original one.

Given the original function R, the fitted one F and the set of all years Y the error ϵ is $\sum_{y \in Y} |F(y) - R(y)|^2$.

In the performed fitting, all errors, associated with every single functions, are summed up obtaining $\epsilon = 122.818$, in the previously described one-by-one fitting; in the general fitting (Fig.33), whose parameter are $\alpha = 6.713175857951899$ and $\beta = 0.8315715748378055$, the error is $\epsilon = 45890.73$, so the general fitting is not really good.

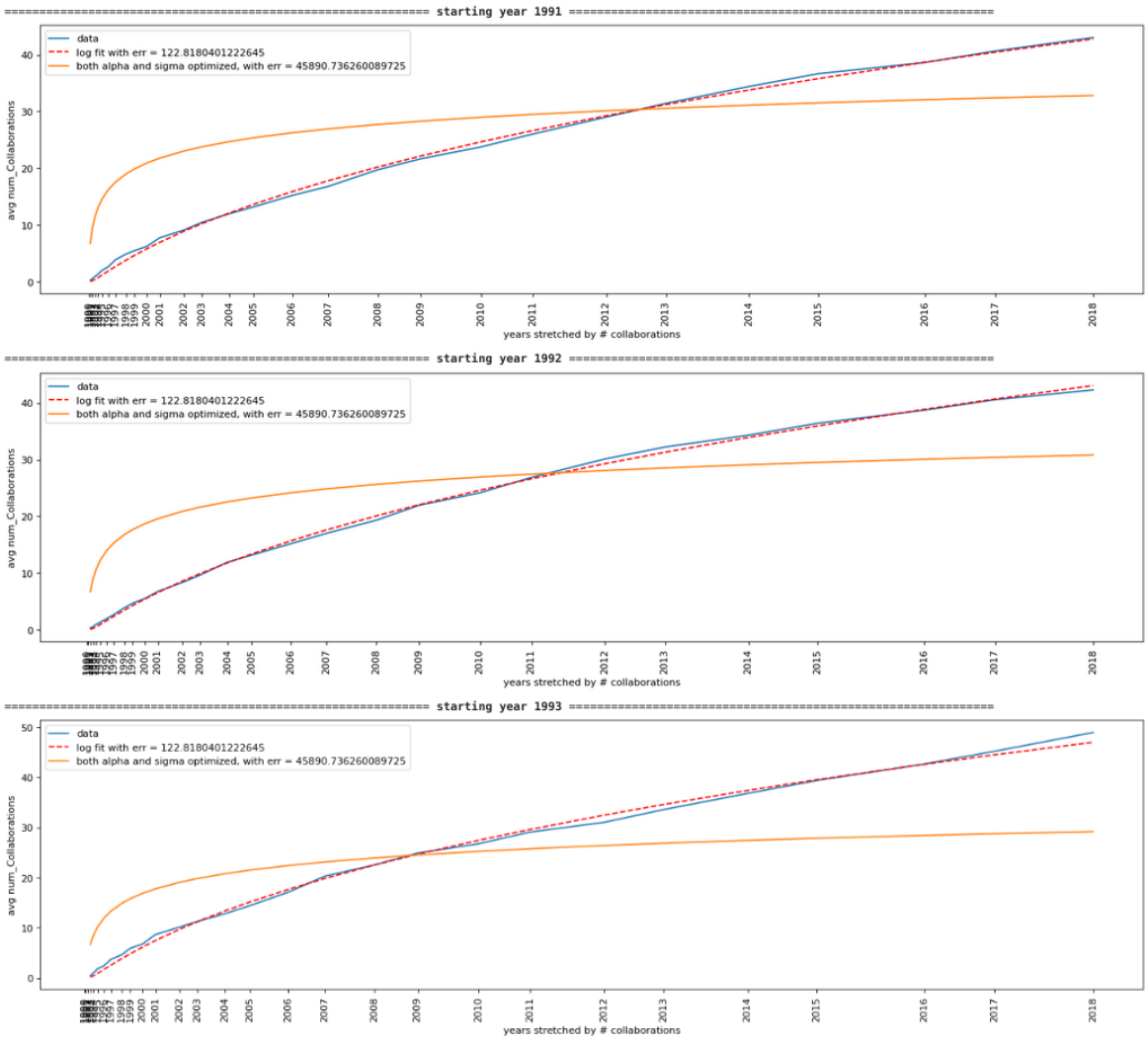

Figure 33: General alpha-sigma fitted function by starting year and hole size 1 - stretched by #collaborations

An explanation behind the obtained big error can be the fact that, as observed in Fig.32, alpha grows linearly while we impose a constant behavior

to it. A possible way to reduce this is introducing a new variable in order to capture the linear behavior of alpha, it's done in the next section (Section 4.3.3 - Alpha-Beta-Sigma logarithmic fitting).

Next the same procedure has been applied for other 2 kinds of events: "new publication" (Fig.34) and "new author" (Fig.35).

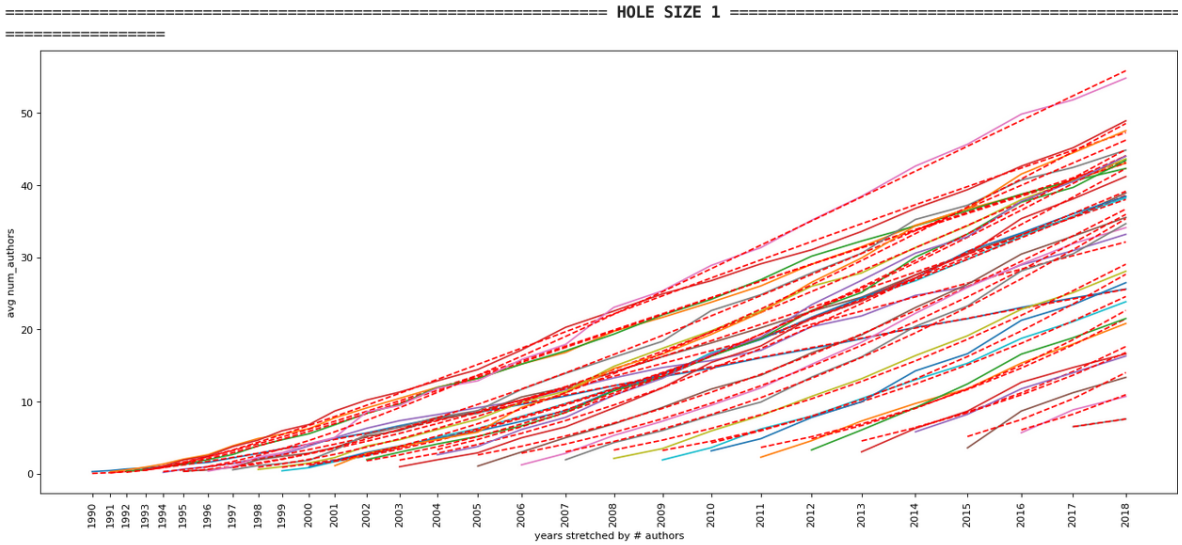

Figure 34: Alpha-Sigma logarithmic fitting of average trajectory by starting year for hole size 1 - stretched by #authors

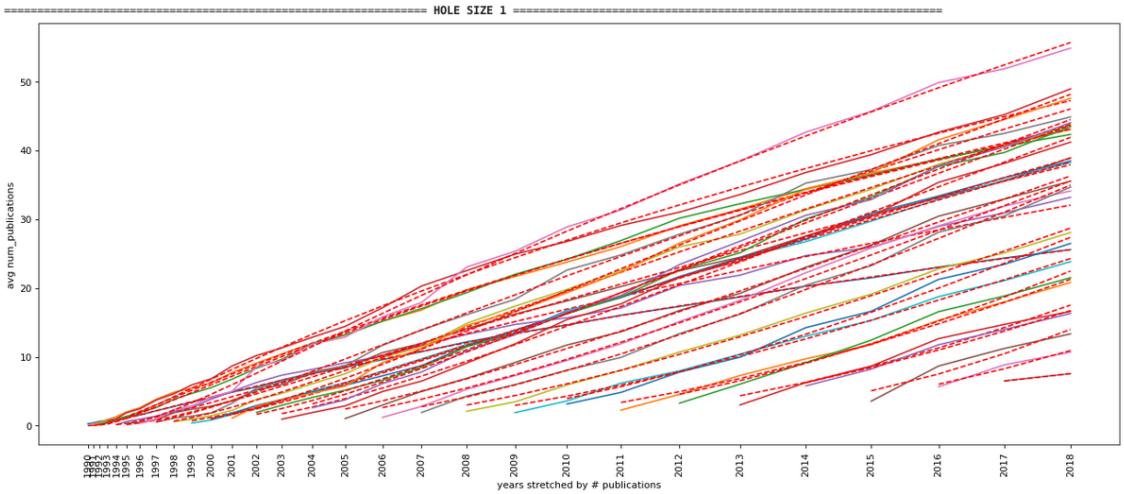

Figure 35: Alpha-Sigma logarithmic fitting of average trajectory by starting year for hole size 1 - stretched by #publications

Their fitted parameters, showed in Fig.38 and Fig.39, have the same behavior as before, alpha is linear and sigma constant (Fig.36 and Fig.37).

The error, for both cases, is a bit bigger than the previous one, both for the one-by-one fitting and the general fitting too; those errors can be

seen respectively in Fig.38 and Fig.39 along with the behavior of the general fitting function in both stretches (Section 3.1.5 - Changing definition of event).

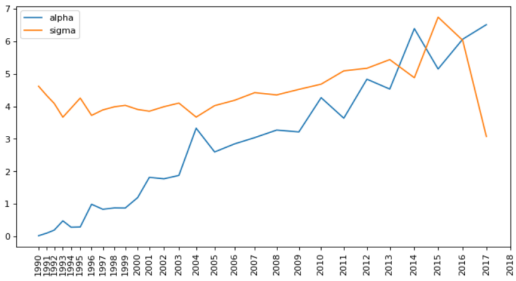

Figure 36: Plot of logarithmic fitting parameters alpha-sigma for each starting year and hole size 1 - stretched by #authors

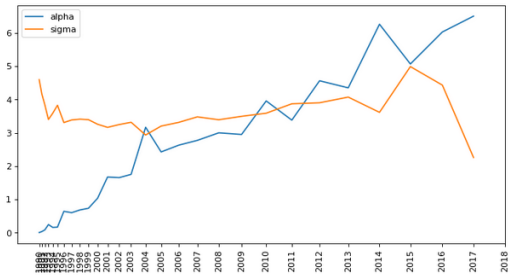

Figure 37: Plot of logarithmic fitting parameters alpha-sigma for each starting year and hole size 1 - stretched by #publications

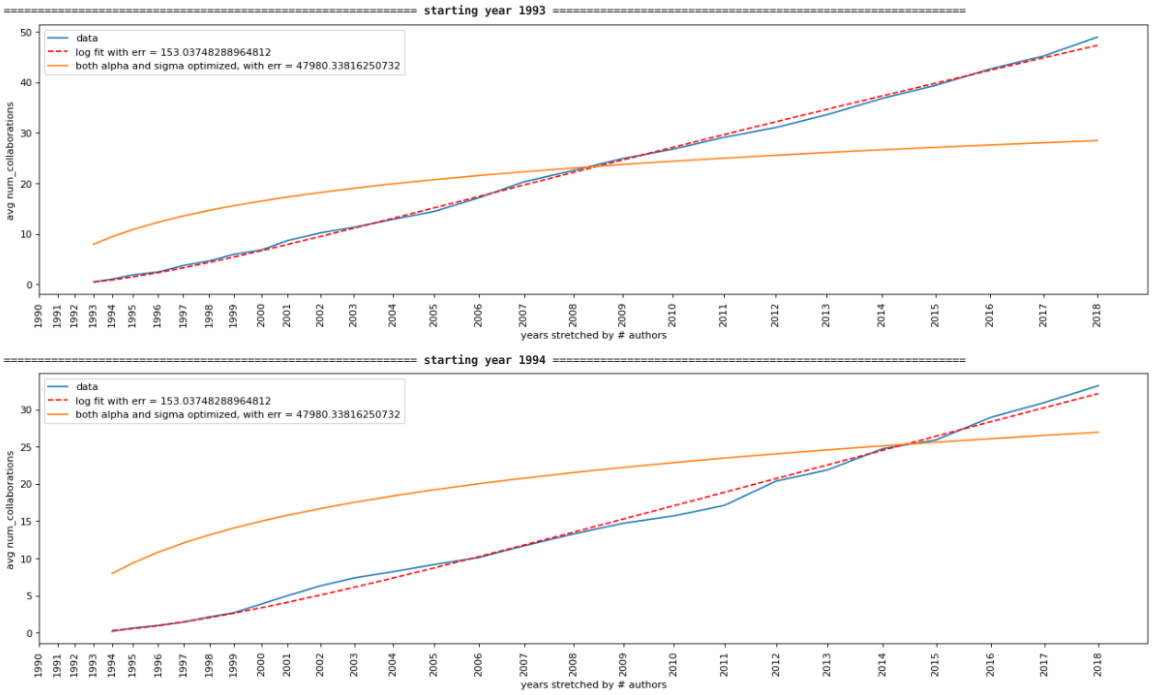

Figure 38: General alpha-sigma fitted function by starting year 1993 and 1994 and hole size 1 - stretched by #authors

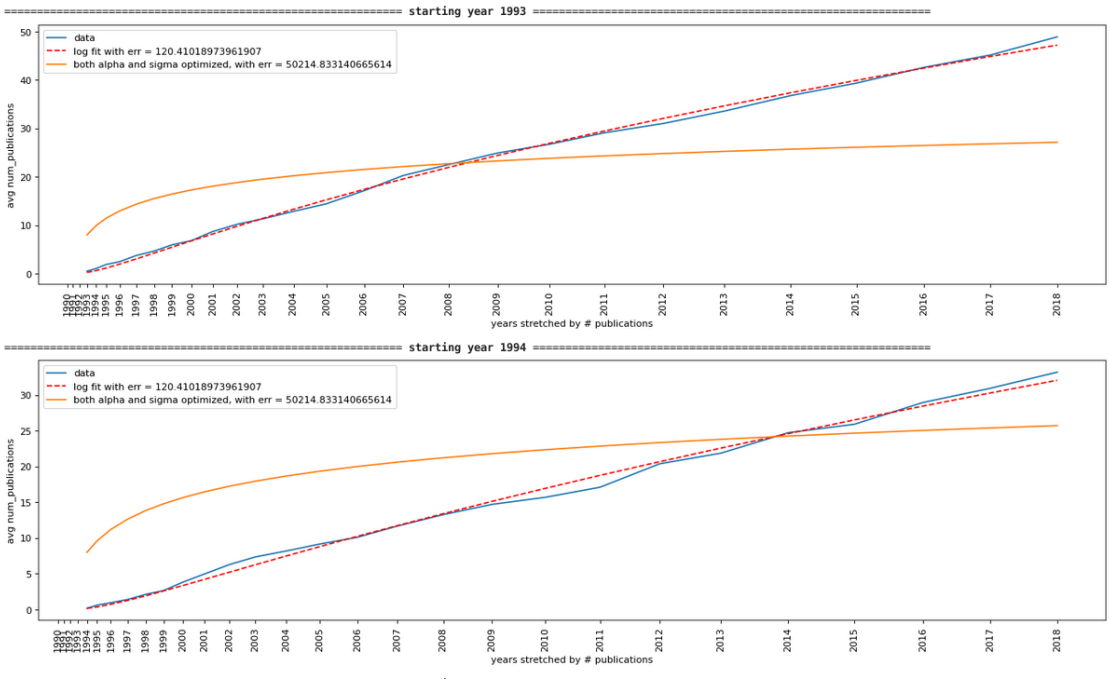

Figure 39: General alpha-sigma fitted function by starting year 1993 and 1994 and hole size 1 - stretched by #publications

4.3.3 Alpha-Beta-Sigma logarithmic fitting

Here the alpha-beta logarithmic function (Eq.1) used in section 4.3.2 has been modified introducing a new parameter: beta. This has been done in order to capture the linear character of alpha in order to reduce the error made by the general fitting function.

The lowest error on the fitting has been obtained using the number of new collaborations as event, so only this kind of stretching is reported here (Fig.40).

The applied procedure remains the same.

$$g_v(t) = \left((\alpha * t + \beta) * ln\left(\frac{t}{t_v}\right) + 1 \right)^{\sigma} \tag{2}$$

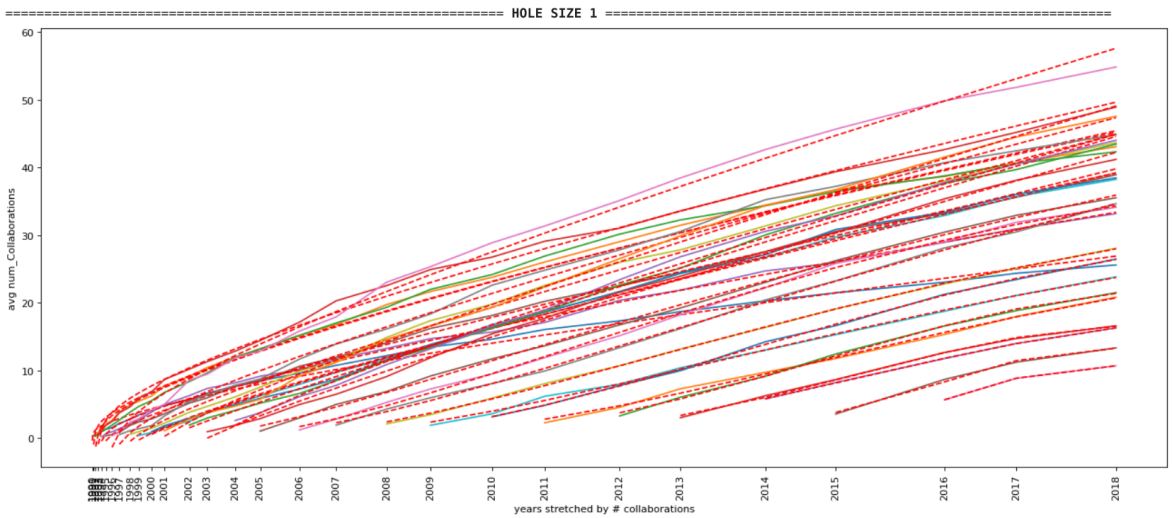

Figure 40: alpha-beta-sigma logarithmic fitting of average trajectory by starting year for hole size 1 - stretched by #collaborations

The fitted values for alpha and sigma parameters are the showed in Fig.41 and their behavior in Fig.42

Sarting Year	alpha	beta	sigma
1990	3.2634	-3.62867	-2.11125
1991	4.91999	-5.86379	-1.84616
1992	5.13772	-6.38	-1.67432
1993	7.68706	-8.57567	-1.68624
1994	3.44468	-3.88234	-1.30748
1995	2.64295	-3.04197	-0.89961
1996	7.29235	-8.64909	-1.11577
1997	5.10166	-6.00725	-0.859194
1998	4.79076	-5.22994	-0.668437
1999	4.17048	-4.43242	-0.543384
2000	4.61766	-4.06242	-0.457184
2001	7.28462	-7.00339	-0.502544
2002	6.03286	-4.49489	-0.218679
2003	8.13492	-8.12544	-0.350182
2004	12.3391	-10.4329	-0.496486
2005	-0.0444474	1.83349	3.04452
2006	-0.148974	1.85116	3.59539
2007	-0.0969829	2.37024	3.23353
2008	-0.293977	2.69651	3.48463
2009	-0.342151	2.70568	3.56568
2010	-0.833991	4.01144	3.9941
2011	-0.790773	3.59249	4.09586
2012	-2.06671	5.81502	4.86355
2013	-2.62224	5.97025	5.59223
2014	-4.5801	10.4223	4.43203
2015	-6.71602	10.5144	8.47151
2016	-17.7267	23.3933	9.6685

Figure 41: Logarithmic fitting parameters alpha-beta-sigma for each starting year and hole size 1 - stretched by #collaborations

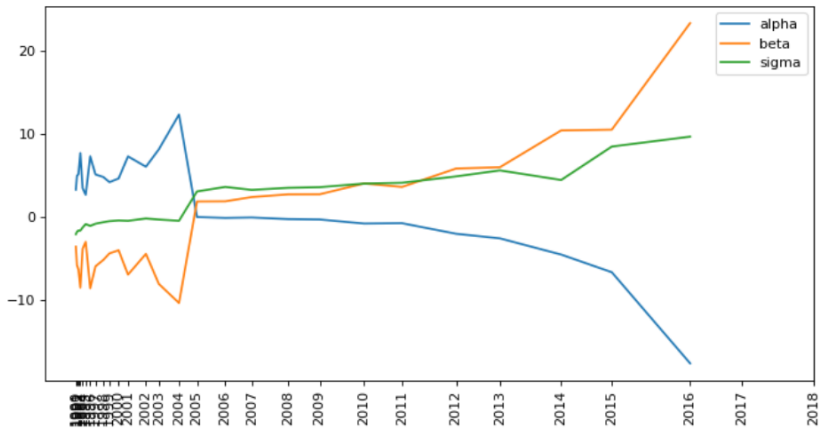

Figure 42: Plot of logarithmic fitting parameters alpha-beta-sigma for each starting year and hole size 1 - stretched by #collaborations

Next has been tried to find the best couple of parameters able to fit all curves and minimizing the total error on the fitting, those parameter for active authors with hole size 1, are the following: $\alpha = 6.297336705250167e-06$, $\sigma = 0.4389174522495641$ and $\beta = 2.1978744505726784$ and they bring to an error on the general fitting, $\epsilon = 11895$, clearly better than the one obtained before.

In Fig.43 can be seen that this function fits much better than the one used before, so probably the theoretical model behind is even more complicated than the model with exponential cutoff.

The fitting probably could be still improved by taking into account the characteristics of alpha and beta observed in Fig.42 as has been done for alpha in this section.

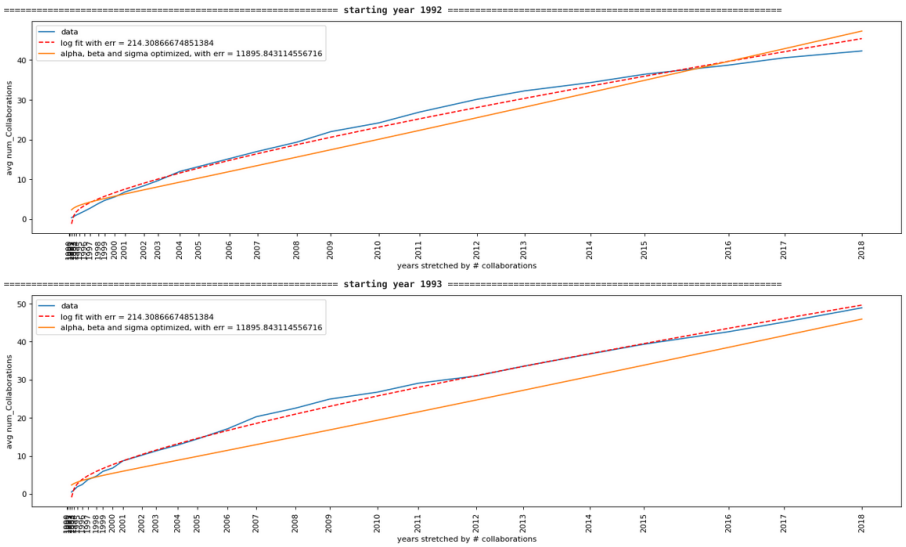

Figure 43: General alpha-beta-sigma fitted function by starting year and hole size 1 - stretched by #collaborations

The last important result is that the error made by the general function decreases for bigger hole size (Fig.44) , maybe because the associated datasets have more authors, as so more data useful for the fitting.

```
Error by hole size

holesize 0 -> 11895.843114556716
holesize 1 -> 6307.941451187858
holesize 2 -> 4855.609080623782
holesize 3 -> 3771.6990334200664
holesize 4 -> 3606.615548113272
holesize 5 -> 3183.8571839528345
holesize 6 -> 2752.313120213234
holesize 7 -> 2544.334087950794
holesize 8 -> 2460.8083212865445
holesize 9 -> 2342.6380478390392
holesize 10 -> 2118.174825468584
holesize 11 -> 1952.465936251338
holesize 12 -> 2134.575761314379
holesize 13 -> 1691.7452633851747
holesize 14 -> 1527.694613837381
holesize 15 -> 1287.864047544826
holesize 16 -> 1162.2438806044945
holesize 17 -> 1069.9276964867058
holesize 18 -> 969.3856720261965
holesize 19 -> 809.644405821902
holesize 20 -> 697.3402291430347
holesize 21 -> 590.5251325372928
holesize 22 -> 564.7740270519718
holesize 23 -> 536.3119940123306
holesize 24 -> 511.96139923090664
holesize 25 -> 548.1252472395685
holesize 26 -> 548.1795549276039
holesize 27 -> 539.2175539919436
```

Figure 44: General alpha-beta-sigma fitted function errors by hole size - stretched by #collaborations

5 Analyzing trajectories for granted and not granted authors

This part of the work has been focused on two subset of authors, one regarding those authors that, at a certain point of their career, received a research grant and those, with a similar collaboration trajectory, that didn't receive the grant.

Collaboration trajectories are similar up to the year of obtaining the grant, we are trying to observe whether, after the year of obtaining the grant, the characteristics of trajectories will start to differ. It is expected, for those who obtained the grant, a higher dynamics of collaboration afterwards.

5.1 Retrieving collaboration data

In Fig.45 and Fig.46 in each row, as for the collaboration dataset, is contained the ID of the author associated with the row, the year in which he started to publish (in the "start_year" column) and the total number of collaborations he has in each year from 1990 to 2018. The "focal" column represent whether the author received (value 1),or not (value 0) a grant, while the "group" column contains an identifier that is equal for each couple of author with similar collaboration trajectory up to the year in which one of them received the grant

(indicated in the "anr_year" column).

	auth.id	group	focal	anr_year	1990	1991	1992	1993	1994	1995	...	2010	2011	2012	2013	2014	2015	2016	2017	2018	start_year
1	10143762200	10	1	2011	0	0	0	0	0	0	...	1	1	1	1	1	1	1	1	1	2003
2	12766889900	100	1	2016	0	0	0	0	0	0	...	20	20	20	20	40	40	40	40	40	2005
4	24766427900	1000	1	2012	0	0	0	0	0	0	...	0	0	0	4	4	4	4	4	4	2003
7	24772993800	1003	1	2011	0	0	0	0	0	0	...	25	25	28	31	31	41	41	46	46	2006
9	24773509500	1004	1	2013	0	0	0	0	0	0	...	6	6	6	6	6	6	6	6	6	1990
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5448	24721362500	987	1	2012	0	0	0	0	0	0	...	3	3	3	3	3	3	3	3	3	1999
5450	12766046000	99	1	2016	0	0	0	0	0	0	...	8	8	8	11	11	11	11	11	19	2006
5452	24725668200	991	1	2014	0	0	0	0	0	0	...	3	6	13	13	17	21	22	26	26	2008
5453	24729414600	992	1	2013	0	0	0	0	0	0	...	0	0	6	6	6	6	12	15	15	2006
5455	24740874700	994	1	2014	0	0	0	0	0	0	...	14	35	40	46	61	79	83	88	111	2008

2861 rows × 34 columns

Figure 45: Granted authors and their collaborations

	auth.id	group	focal	anr_year	1990	1991	1992	1993	1994	1995	...	2010	2011	2012	2013	2014	2015	2016	2017	2018	start_year
0	6602860506	1	0	2010	0	0	0	0	0	0	...	4	11	11	11	11	11	11	11	11	1990
3	23569685800	100	0	2015	0	0	0	0	0	0	...	9	11	12	25	29	33	37	43	46	2007
5	15131147900	1000	0	2005	0	0	0	0	0	0	...	0	7	7	7	7	7	7	7	7	1992
6	16315003800	1001	0	2007	0	0	0	0	0	0	...	3	3	3	3	3	7	7	7	8	1994
8	22135343900	1003	0	2015	0	0	0	0	0	0	...	31	42	59	63	64	72	81	87	103	2006
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5451	7006892189	990	0	2007	0	0	0	0	0	0	...	0	0	0	0	0	0	2	2	2	1991
5454	7005280084	992	0	2014	0	0	0	0	0	0	...	3	3	3	3	3	3	3	3	3	2003
5456	7003626217	994	0	2009	0	0	0	0	0	0	...	6	6	6	6	6	6	6	6	6	1996
5457	6602299637	996	0	2007	0	0	0	0	0	0	...	29	40	42	44	45	45	48	48	52	2001
5458	15128095600	999	0	2015	0	0	0	0	0	0	...	3	3	3	3	3	3	3	3	3	1999

2598 rows × 34 columns

Figure 46: Not granted authors and their collaborations

5.2 Computing weighted average on shifted trajectories

In this section have been computed average trajectories by starting year for both granted (Fig.47) and not granted authors (Fig.48).

Those curves has then been shifted, without prolongation (Fig.49 and Fig.50), and lastly, in Fig.51, their average has been computed, weighted on the number of authors each curve contains.

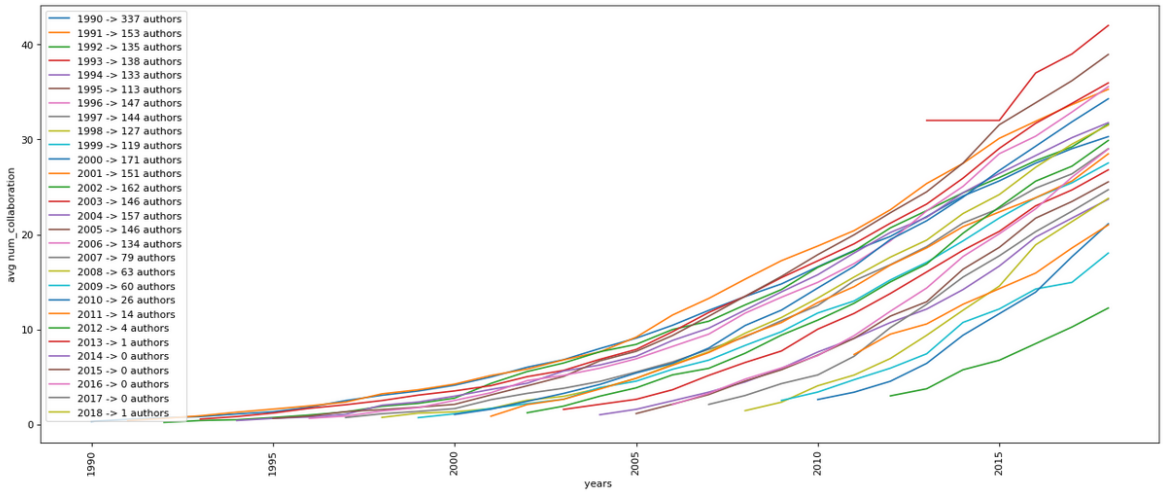

Figure 47: granted average trajectories by starting year

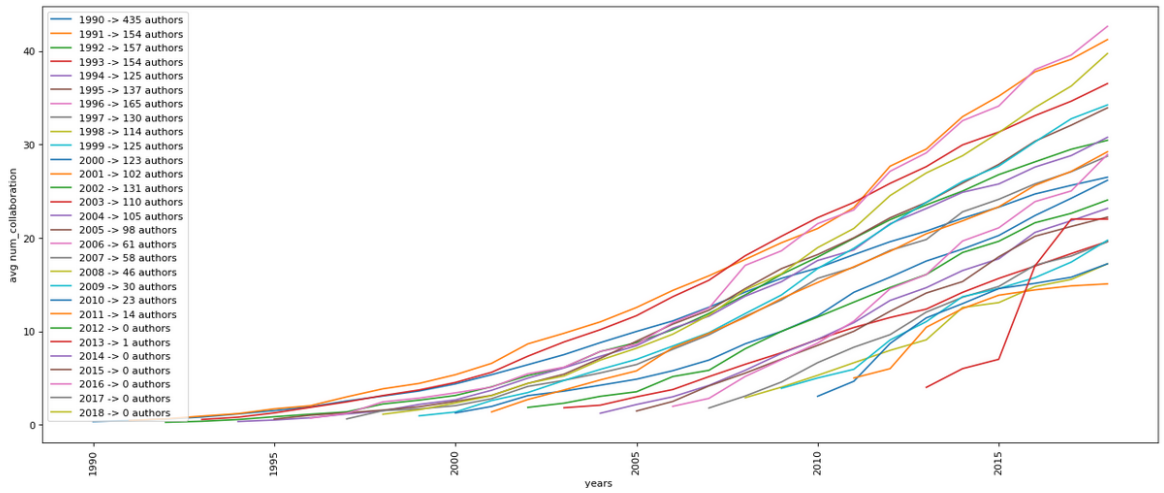

Figure 48: not granted average trajectories by starting year

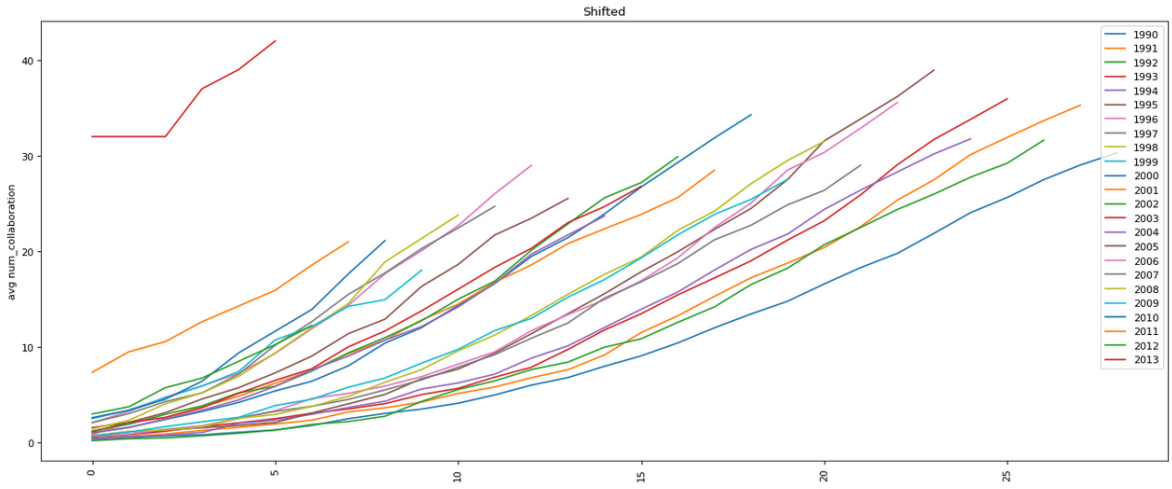

Figure 49: Granted average shifted trajectories

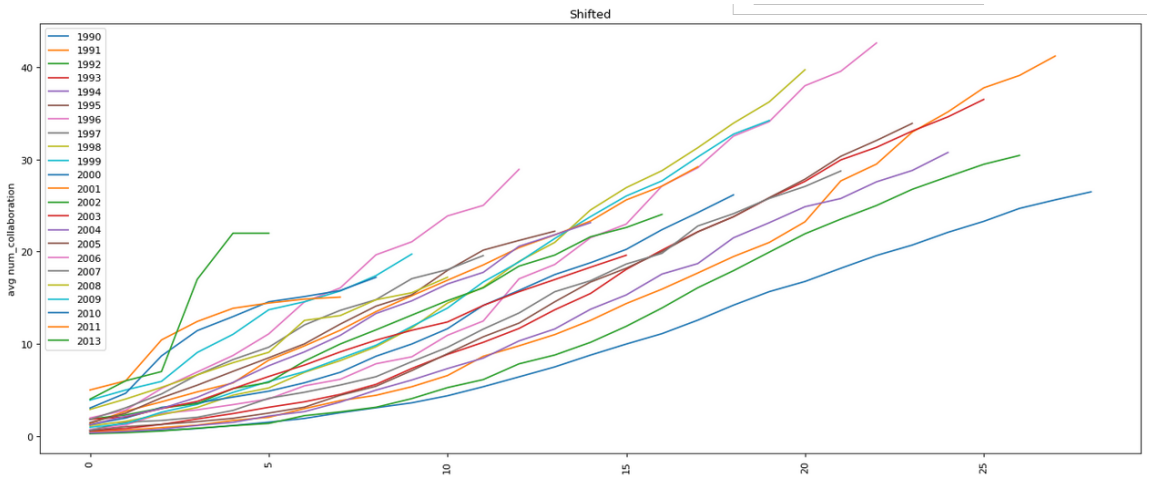

Figure 50: Not granted average shifted trajectories

We expected granted authors to reach a higher number of collaboration, so in Fig.51 the average trajectory associated with them should grow more than the one associated with not granted authors, but as can be seen, this is not the behavior they have in the real data.

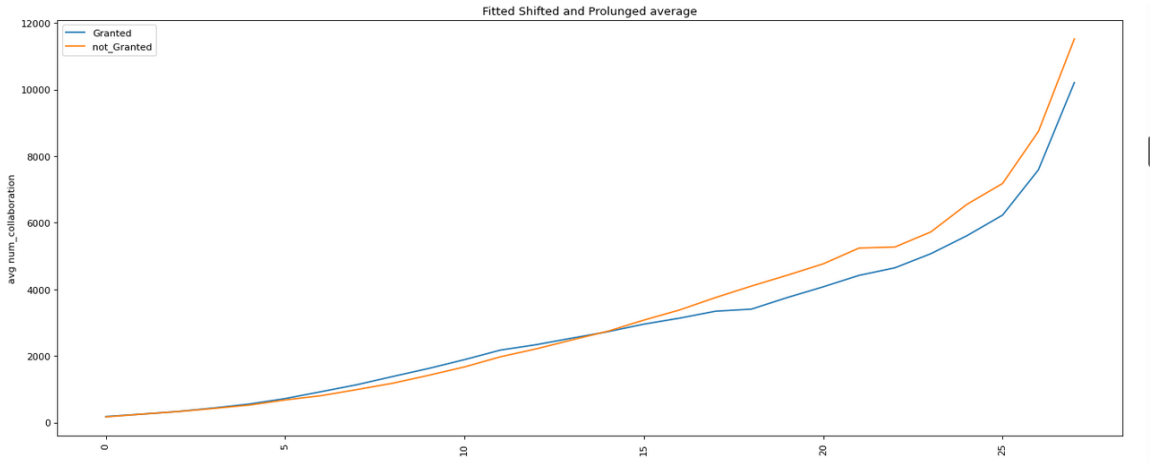

Figure 51: Total weighted average of shifted trajectories

5.3 Computing average on shifted fitting curves on trajectories

Here the average trajectories computed in the previous section, have been fitted with the function in Eq.3, then those fitting curves has been shifted (Fig.52 and Fig.53) and prolongated (Fig.54), lastly their average has been computed (Fig.55) in order to compare again collaboration trajectories for granted and not granted authors (Fig.56).

In order to do the fitting the previously logarithmic alpha-sigma function used before (Section 4.3.2) has been used:

$$g_v(t) = \left(\alpha * ln\left(\frac{t}{t_v}\right) + 1\right)^{\sigma} \tag{3}$$

Notice from Fig.52 and Fig.53 that the trajectory associated with the latest starting years contain few collaboration data, so when prolongated they will have a higher grow, and so a higher impact on the average. We can then say that the smaller the starting year is the more precise is the fitting.

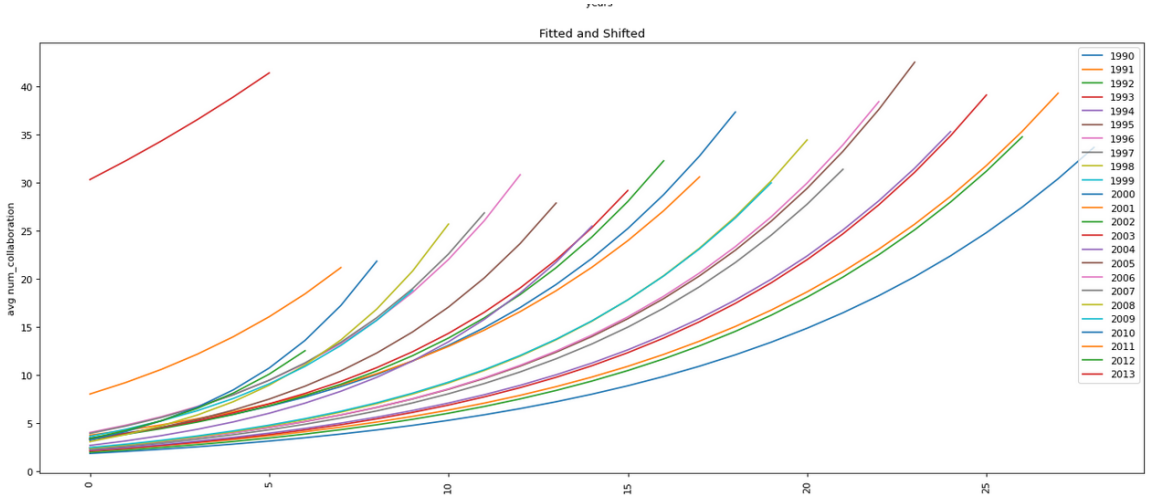

Figure 52: Granted fitted and shifted average trajectories

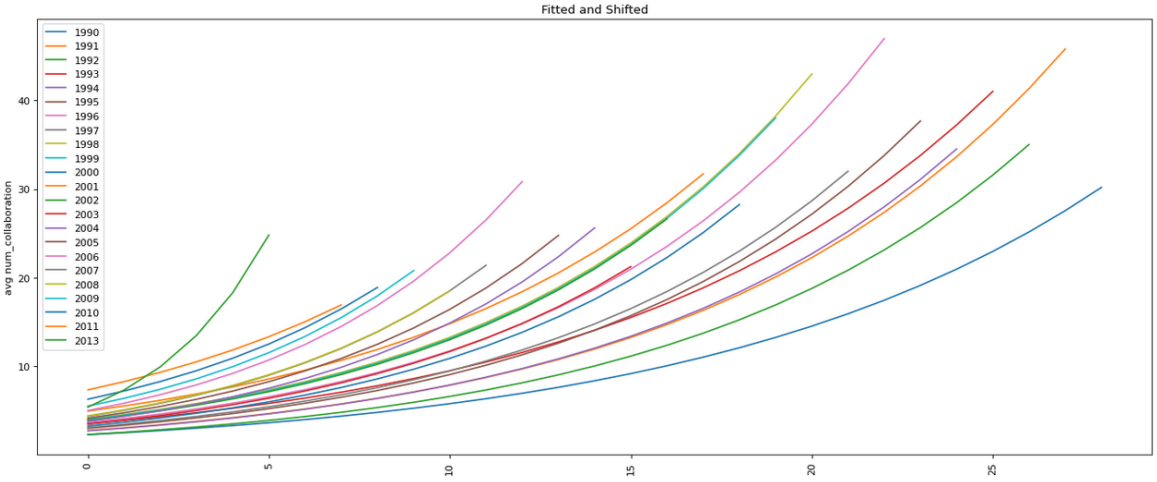

Figure 53: Not granted fitted and shifted average trajectories

As said above, some trajectories, because the lack of data, grows faster than others after the prolongation, this fact is evident in Fig.54 and Fig.55.

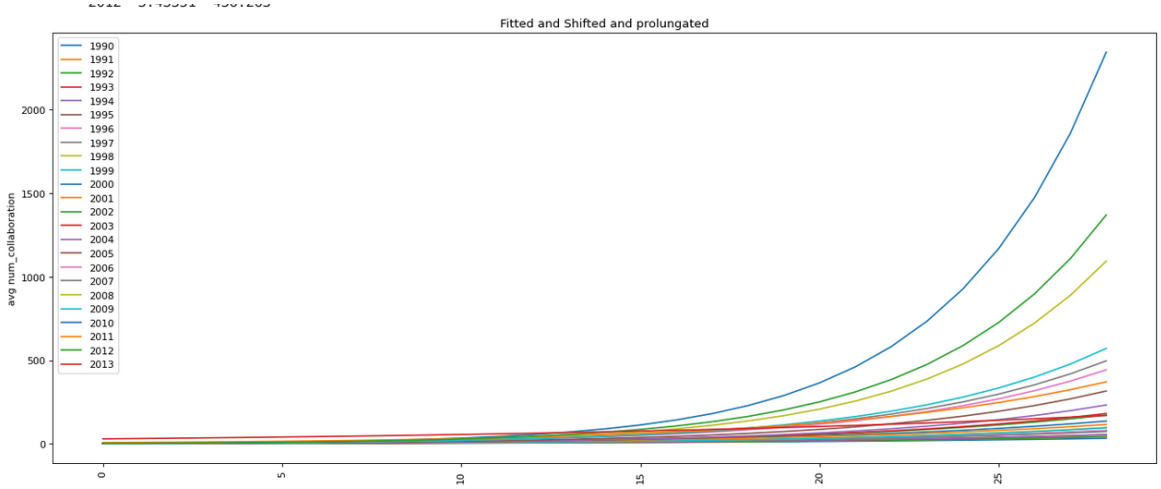

Figure 54: Granted fitted, shifted and prolongated average trajectories

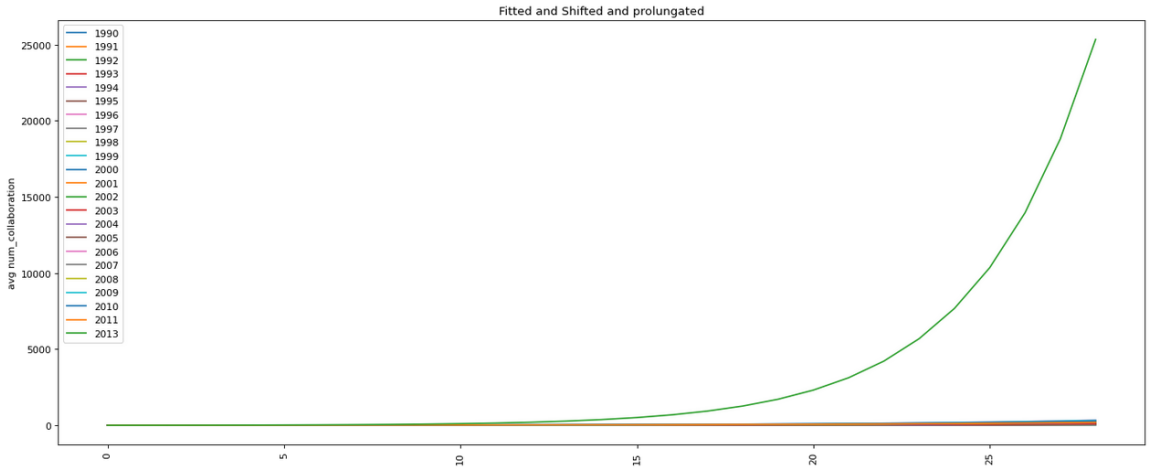

Figure 55: Not granted fitted, shifted and prolongated average trajectories

As expected at the start of this section, trajectories of granted authors reach a higher average number of collaboration (Fig.56).

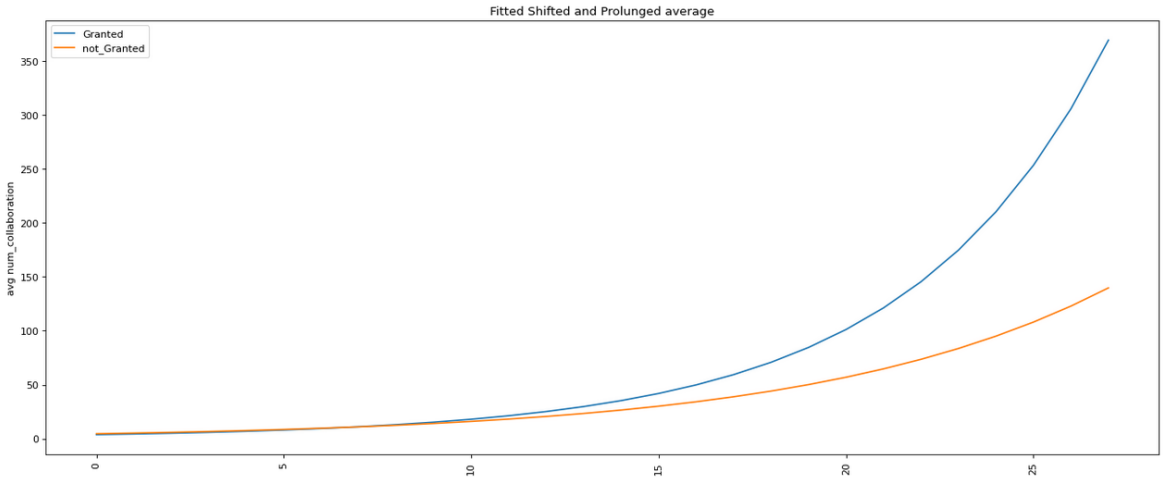

Figure 56: Total average of fitted, shifted and prolongated average trajectories

6 Conclusions

The main scope of the project was the analysis of vertex trajectories in the given collaboration network, under the assumption that the underlying data respects a power-law distribution with exponential cutoff.
By results given in Section 3.1.8 we can say that it does respect the expected distribution, so in Section 4 has been performed a more sophisticated analysis on vertex trajectories, in particular, we have found that the function in Eq.4 is, so far, the most suited to fit our data.

$$g_v(t) = \left((\alpha * t + \beta) * ln\left(\frac{t}{t_v}\right) + 1 \right)^{\sigma} \tag{4}$$

We concluded that this fitting function works better if we consider the number of new collaborations as events (Section 3.1.5) and that, the bigger is the hole size considered the smaller is the error made; this can be a consequence of the fact that for a big hole size, the dataset associated contains more authors, and so, more data is given for the fitting.

We are brought to suppose that the underlying theoretical model is much more complex than expected, and so the fitting function used can be improved to perform a better fit.

Regarding the trajectory analysis for granted and not granted authors (Section 5), has been found out that their number of collaboration, on raw data, doesn't grow as expected: the average number of collaborations achieved by the two groups of authors is almost the same, while we were expecting a higher collaboration rate for those who received a funding. Has been tried so, by the usage of a fitting function, to predict the evolution of those trajectories by shifting and prolongating their fittings, discovering that, after those operations, the average trajectory for granted authors grows more than the one of not granted, as expected.

A next step for this work could be to analyze data coming from other research field than computer science, to see, if they bring to same, or even better, results.

So far we can conclude to be in the right direction of finding a model able to represent, with enough precision, the behavior of collaboration trajectories in networks as the one analyzed in this work.

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.

[4] S. Bornholdt and H. G. Schuster, Handbook of Graphs and Networks: From the Genome to the Internet. John Wiley Sons, 2006.

[5] A. D. Broido and A. Clauset, "Scale-free networks are rare," Nat Commun, vol. 10, no. 1, p. 1017, Mar. 2019, doi: 10.1038/s41467-019-08746-5.

[6] R. van der Hofstad, Random Graphs and Complex Networks: Volume 1, 1st ed. USA: Cambridge University Press, 2016.

[7] F.Giroire, N.Nisse, M.Sulkowska, Study of a degree distribution and a vertex trajectory in the Chung-Lu model with a generalized attachment function, 2022

[8] P. L. Krapivsky, S. Redner, and F. Leyvraz, "Connectivity of Growing Random Networks," Phys. Rev. Lett., vol. 85, no. 21, pp. 4629–4632, Nov. 2000, doi: 10.1103/PhysRevLett.85.4629.

[9] P. L. Krapivsky and S. Redner, "Organization of growing random networks," Phys. Rev. E, vol. 63, no. 6, p. 066123, May 2001, doi: 10.1103/PhysRevE.63.066123.

[10] R. Durrett, Random Graph Dynamics. Cambridge University Press, 2006. doi: 10.1017/CBO9780511546594.