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 Attachment function and network evolution . . . 11
- 2.2 Degree Distribution . . . 12
- 2.3 Power law distribution and scale free networks . . . 13
- 2.4 Power law distribution with exponential cutoff . . . 14
- 2.5 Vertex trajectory . . . 15
- 2.6 Theoretical Model . . . 16

3 Data Preparation 19
- 3.1 Retrieving Data . . . 19
- 3.2 Identifying active authors . . . 20
- 3.3 Filtering active authors . . . 22
- 3.4 Changing definition of event . . . 23

4 Getting to know data characteristic 27
- 4.1 Degree Distribution . . . 27
- 4.2 Average Vertex Trajectories . . . 28

5 Fitting 31
- 5.1 Degree Distribution Fitting . . . 31
- 5.2 Vertex Trajectories Fitting . . . 32
- 5.3 General Fitting function for trajectories . . . 34
  - 5.3.1 Error definition . . . 35
  - 5.3.2 Optimization Results . . . 35
  - 5.3.3 Results implications . . . 36

6 Conclusions **39**

7 Bibliography **41**

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 2.5), 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 2.4) [7]. 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 28 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 goals are, as first, to take confidence with the collected data building the underlying collaboration's network and then study some of it's mathematical properties and their evolution over time, in particular **analyzing vertex trajectories and the degree distribution**, where for vertex trajectory is meant the evolution of node degrees over time, and the degree distribution is the probability distribution over nodes the degrees of nodes.

The final goals of the project are to understand how those metrics behave in the cited networks to implement a mathematical model able to generate similar graphs to firstly understand their evolution pattern and so predict the evolution over time of this and similar networks.

Chapter 2

State of the Art

2.1 Attachment function and network evolution

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** (Fig 2.1), where a new node appears in the graph, and an **edge event** (Fig 2.2) where a new edge appears. Let's define P as the **probability for a node event** to occur and, consequently $(1 - P)$ for an edge event. Notice that in the first case, when a new node appears, is connect to another already present through an edge (Fig 2.1), so is more correct to say that is a **'node+edge event'**, but will be referred for simplicity as 'node event').

In the node event, the new node must select another one, 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** $f(x)$ over which is defined **the probability for any node** $v \in V_t$ **to be chosen for the attachment**.

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

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 $\{G_0, G_1, ...G_n\}$.

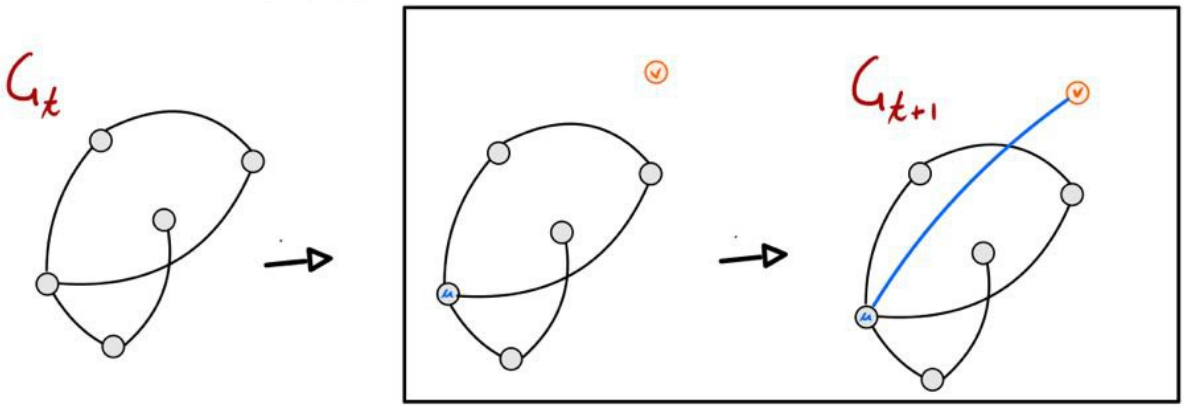

Figure 2.1: *Example of a node event: a new node v join the graph G_t at time t and connect to the node u. After the event the Graph G_{t+1} is obtained.*

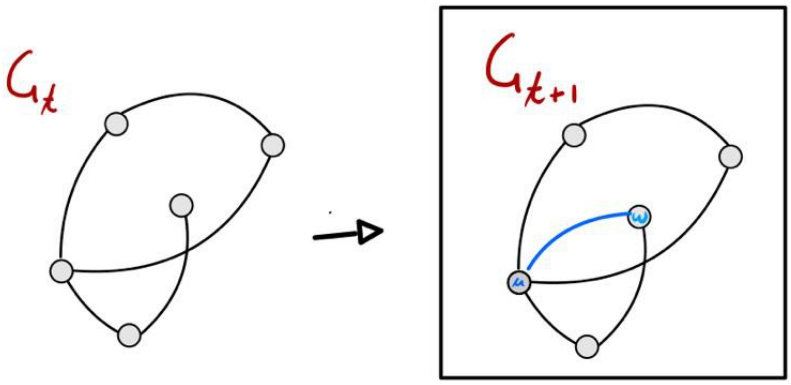

Figure 2.2: *Example of an edge event: a new edge (u, w) appears in the graph G_t at time t. after the event the Graph G_{t+1} is obtained.*

The probability P of occurrence of node events, together with the way the attachment function $f(x)$ is chosen, defines a **model for the evolution of the network**. The interest of this project is focused on **two measures**, directly related to this evolution: The **Degree Distribution** and **The vertex trajectory**, defined in the following subsections.

2.2 Degree Distribution

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. The **Degree distribution** is $M_k = \frac{n_k}{n}$.

A graphical example of degree distribution is illustrated in Fig 2.1. where, on the left, each node o the graph G has been coloured based on it's degree and in the chart, on the right, on the x axis there's the degree k and on he y axis the fraction of node of degree k that is, the degree distribution M_k.

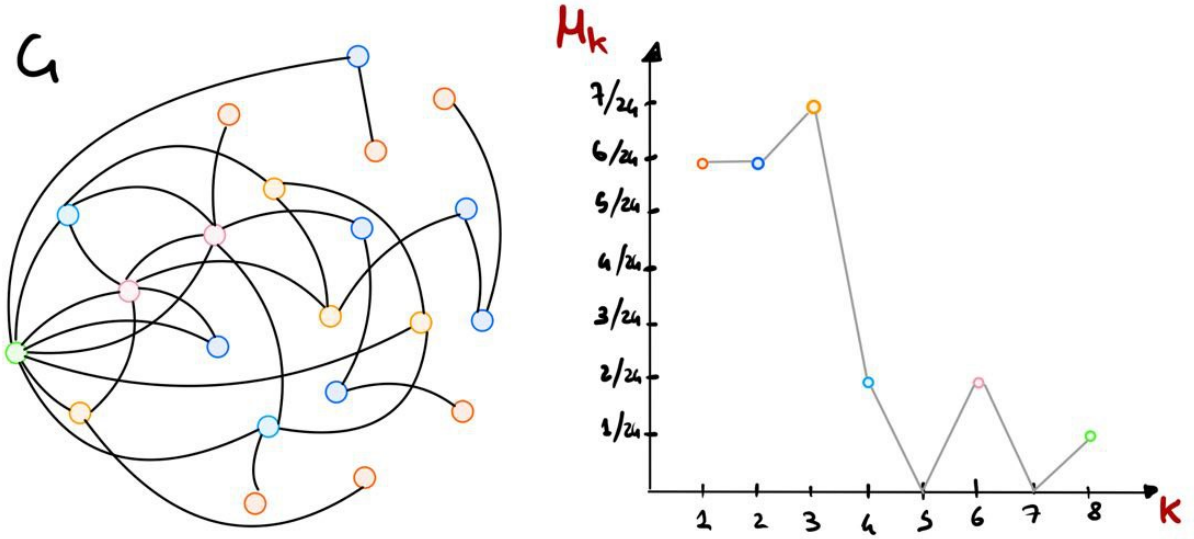

Figure 2.3: *Example of degree distribution: in G nodes of different degree have a different color, the degree distribution of G is plotted on the right.*

2.3 Power law distribution and scale free networks

We say that the degree distribution M_k of G follows a **power law** if:

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

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

In Fig 2.4, on the left, is show the power law distribution with: the degree k on the x axis and the degree distribution M_k on the y axis. On the right there's the same plot but with the logarithm of k on the x axis and the logarithm of M_k on the y axis.

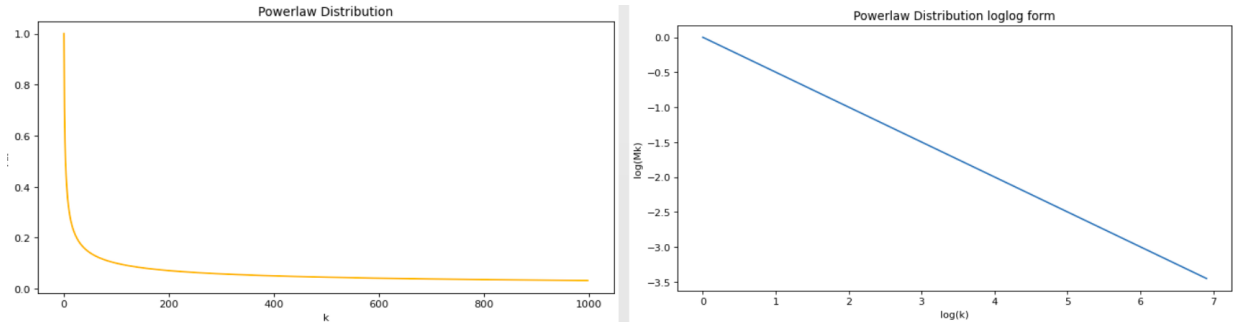

Figure 2.4: *power law distribution and it's log-log form.*

A network whose probability distribution of degrees of nodes respects the power law distribution is said to be **scale-free**. Being scale-free implies that **nodes with an high degree** in the network are **more likely to be chosen for attachment** when a new edge appears in the

network, as an example, on Instagram, a singer with thousand of followers have an higher probability, than anyone with a couple of dozens, of being followed from a new user; this, as showed in Fig 2.4, bring to have **few node of high degree and a lot of small degree**.

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 them and has been often believed that most real-world networks are scale-free [1], [2], [3].

An example as possible applications of these models, can be a tool to **generate random networks** having the same structure as the observed one, as the **Barabàsi–Albert[7] model** does, which is used to generate scale-free networks. In order to build their model, Barabàsi and Albert [7], 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.

2.4 Power law distribution with exponential cutoff

Ubiquitousness of power law has been questioned in the last decades [4] and Investigations shown that **scale-free networks are not so widespread** as thought [5], [6].

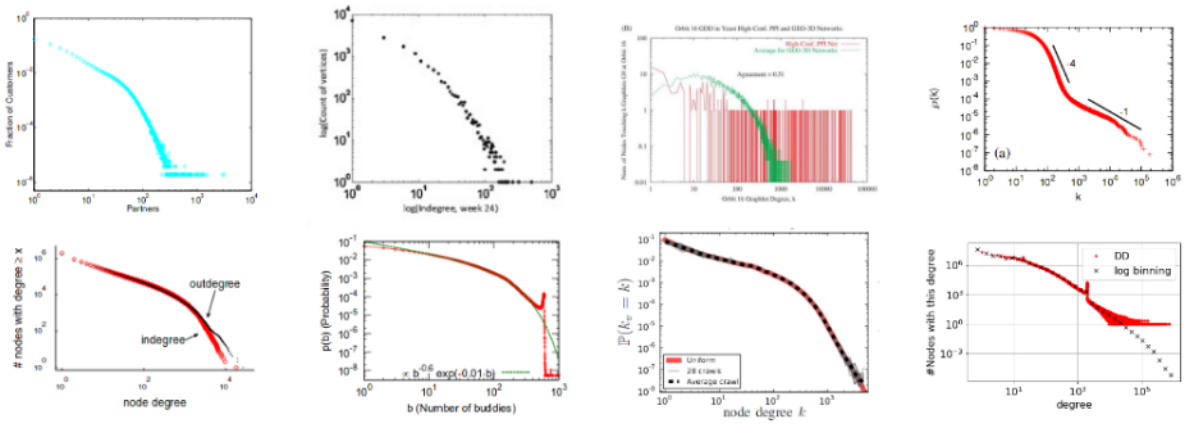

Figure 2.5: *Example of non-scale free networks in biology, online market, social networks etc... . In particular their degree distribution shows the, so called power law with exponential. cutoff*

It turns out that many of them (Fig 2.5) follow **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.

In Fig 2.6, on the left, is show the exponential cutoff power law with: the degree k on the x axis and the degree distribution M_k on the y axis. On the right there's the same plot but with the logarithm of k on the x axis and the logarithm of M_k on the y axis.

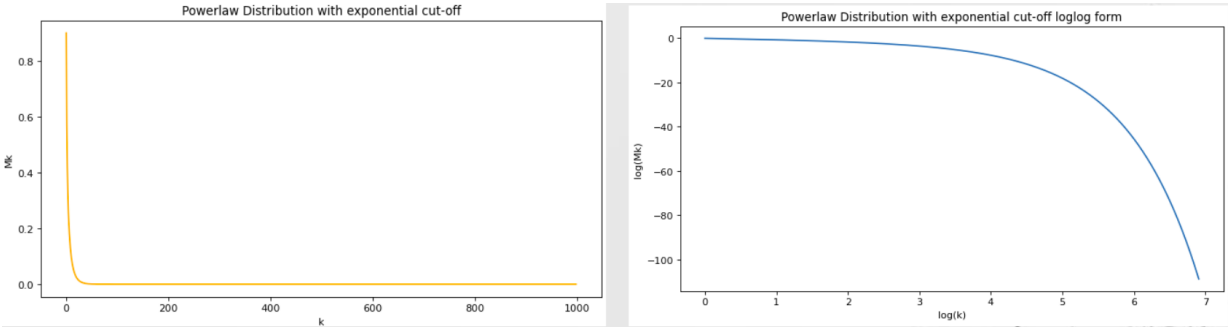

Figure 2.6: *power law distribution with exponential cutoff and it's log-log form.*

2.5 Vertex trajectory

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

Let's define the sequence of graphs $\{G_0, G_1, ...G_n\}$, representing the **evolution over time** of $G_0 = (V_0, E_0)$, and $t = 0, 1, ..., n$, the number of node or edge **event occurred** (section 2.1).
Let also $d_v(t)$ as the degree of vertex v at time t and let t_v 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:

$$\{d_v(t_v), ..., d_v(t_v + 1), ..., d_v(t_v + n)\}$$

.
An example can be found in Fig 2.7. where for each time step $t = 0, 1, ..., n$ the corresponding G_t is shown, in each graph the edges of the node v are coloured in orange, the sum of those edges gives the degree $d_v(t)$ of v at time t. The sequence of degree $\{1, 5, 7, ..., 8\}$ is the vertex trajectory shown on the bottom.

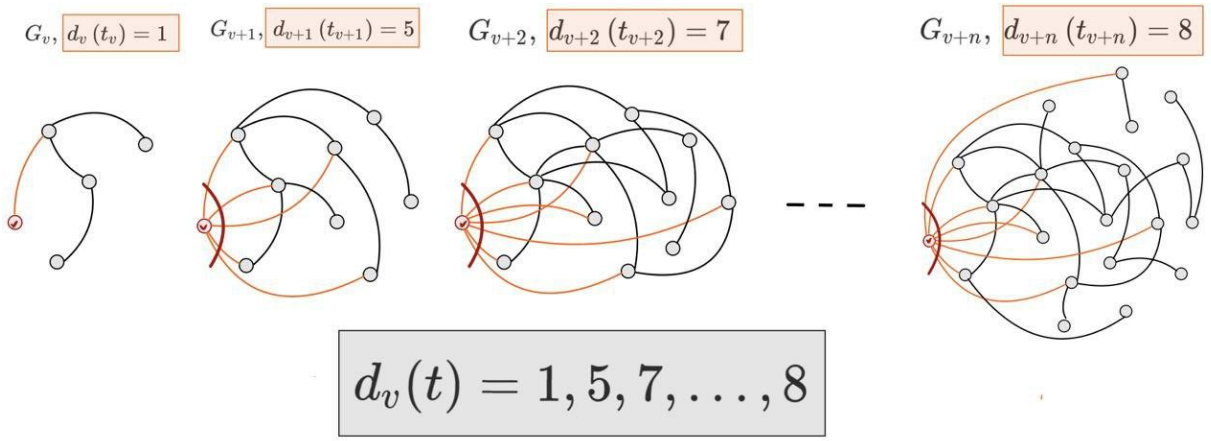

Figure 2.7: *Example of the vertex trajectory of a node v, in red, joining the network at time t_v. The edges of node v, in orange, at each time step they sum up to the degree of v. The sequence of degrees is the vertex trajectory, showed in the bottom.*

2.6 Theoretical Model

The evolution of the two metrics we discussed above, **degree distribution** and **vertex trajectory**, as said at the end of section 2.1, directly depends from: the **probability P of a node event** to occur and the way the **attachment function** $f(x)$ is chosen. From those two parameters depends how the network will evolve, if it will have a **normal power law degree distribution or the one with exponential cutoff** and if the vertex trajectory will be a **square root or a logarithmic function**.

An example of mathematical model which **brings to a power law** degree distribution, is the **Barabàsi and Albert** [7]. In this model $P = 1$, so there arecan **no edge event**, the only way for an edge to appear is trough a new node joining the network and connecting to some other node. The attachment function instead is simply the **degree of the node** v at time t: $f(d_v(t)) = d_v(t)$, so the probability of a node v to be chosen depends only from his current degree $d_v(t)$.

A network evolving under the Barabàsi-Albert's rules brings to a graph with a **power law** degree distribution $M_k \sim Ck^{-\lambda}$ and a **square root** function as vertex trajectory $d_v(t) = (\frac{t}{t_v})^{\frac{1}{2}}$, where t_v is the time step in which the node v appears in the network.

The **theoretical model** we will refer to during this work is the one illustrated in the next table [8]. It is a **generalization of the Barabàsi-Albert** one that is, the **Chung-Lu model** [8],[9].
In this model both **node and edge events** can occur, with probability P and $(1-P)$ respectively; the preferential attachment function $f(x) = x^\gamma$ with $0 \leq \gamma < 1$ is a **sublinear generalization** of the Barabàsi-Albert one (in which $\gamma = 1$).

The degree distribution M_k, in this model, follows a more subtle distribution than the power law with exponential cut off, the so-called **stretched exponential**, while the vertex trajectory

$d_v(t)$ has a **logarithmic shape**.

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 $\sum_{w \in V_t} (\mathrm{d}_w(t))^\gamma \sim \mu t$, where $\mu \in [p, 2]$, where $\mu \in [p, 2]$.

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

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

Chapter 3

Data Preparation

3.1 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.3.1. 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.

Scopus ID

Total # collaborations (until 2014)

	ID	1990	1991	1992	1993	1994	1995	1996	1997	1998	...	2014	2015	2016	2017	2018
118063	26421678500	0	0	0	0	0	0	0	0	0	...	4	4	4	4	4
180546	56230251900	0	0	0	0	0	0	0	0	0	...	3	3	3	3	3
68772	7801413223	0	0	0	0	0	0	0	0	0	...	3	3	3	3	3
25152	6603158006	0	0	0	0	0	0	0	0	0	...	0	0	0	0	4
96494	20434297300	0	0	0	0	0	0	0	0	0	...	16	16	16	16	16
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
228654	57203927130	0	0	0	1	1	1	1	1	1	...	31	32	39	47	48
115362	25647427000	0	0	0	0	0	0	0	0	0	...	0	0	0	0	49
176446	56066133100	0	0	0	0	0	0	0	0	0	...	0	0	0	3	3
64352	7202888402	0	0	0	0	0	0	0	0	0	...	3	3	3	3	3
101801	23099287300	0	0	0	0	0	0	0	0	0	...	12	17	17	17	22

232838 rows × 35 columns

Figure 3.1: *Collaboration data for computer science authors.*

3.2 Identifying 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.3.2 the author A1 is active but A2 is not, and their **maximum hole size** are respectively 3 and 4. 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.

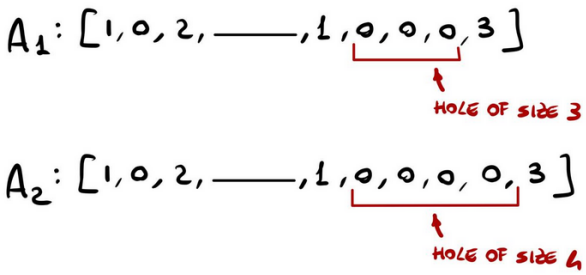

Figure 3.2: *Hole size definition.*

In Fig 3.3 are shown the **distributions of the number of authors** over the yet cited metrics. on the x axis is shown the metric (hole size, activity period or minimum number of publications) for values between 0 and 10, while on the y axis the number of authors associated with such values.

In order to build vertex trajectories, and get a better understanding of the data, for each author, is also needed the year in which they started to publish as well as the one in which they stopped.

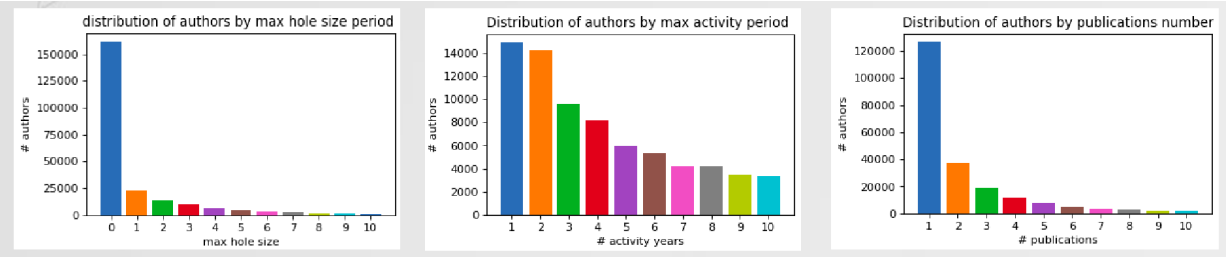

Figure 3.3: *Distribution of authors for hole size (left), activity period (middle) and minimum number of publications. All for values between 1 and 10 for each metric.*

For each author the value of maximum hole size, activity period or minimum number of publications is retrieved in order to be filter them (Section 3.3) in a way of making groups for future comparisons. So a new version of the collaboration dataset is built upon the one described in Section 3.1 (Fig.3.1), containing new columns with: the **starting and ending publication year**, the **activity period**, the **maximum hole size** and the **total number of publications** for each author (Fig.3.4).

	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
34493	6701587952	0	0	0	0	0	0	0	0	0	...	3	3	3	3	3	2005	2005	0	0	2
58480	7102054295	0	0	0	0	0	0	0	0	0	...	11	11	12	12	15	2012	2018	0	6	14
206646	57191595848	0	0	0	0	0	0	0	0	0	...	0	0	0	0	2	2018	2018	0	0	1
85964	14057061000	0	0	0	0	0	0	0	0	0	...	4	4	4	8	8	2012	2017	4	5	2
155327	55351812100	0	0	0	0	0	0	0	0	0	...	1	1	1	1	1	2011	2011	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32798	6701343170	0	0	0	0	0	0	0	0	0	...	8	8	10	10	14	2011	2018	1	7	6
145298	45061005800	0	0	0	0	0	0	0	0	0	...	2	2	2	2	2	2011	2011	0	0	1
123287	35100995100	0	0	0	0	0	0	0	0	0	...	4	4	4	4	4	2009	2009	0	0	1
121565	34467631000	0	0	0	0	0	0	0	0	0	...	5	5	5	5	5	2007	2007	0	0	1
126255	35270434600	0	0	0	0	0	0	0	0	0	...	13	13	13	13	13	2012	2012	0	0	1

232838 rows × 35 columns

Figure 3.4: *Collaboration data with starting, ending year, activity period and total publications number.*

The **distribution of the number of authors by their starting and ending year** has been plotted in Fig.3.5 and Fig.3.6, 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 publishing in 2018 may still be active also nowadays.
Those plots shows also that there are more new authors who start publishing each year, and that the number of authors who stop to publish is also increasing, but with a lower rate.

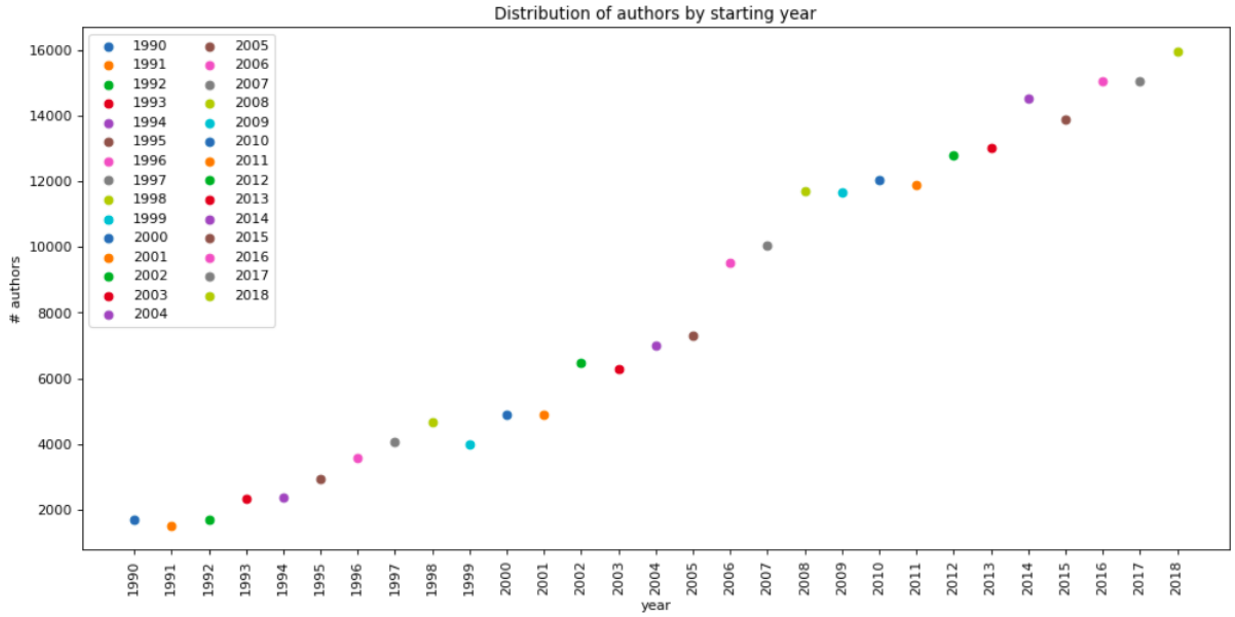

Figure 3.5: *Distribution of authors by starting year*

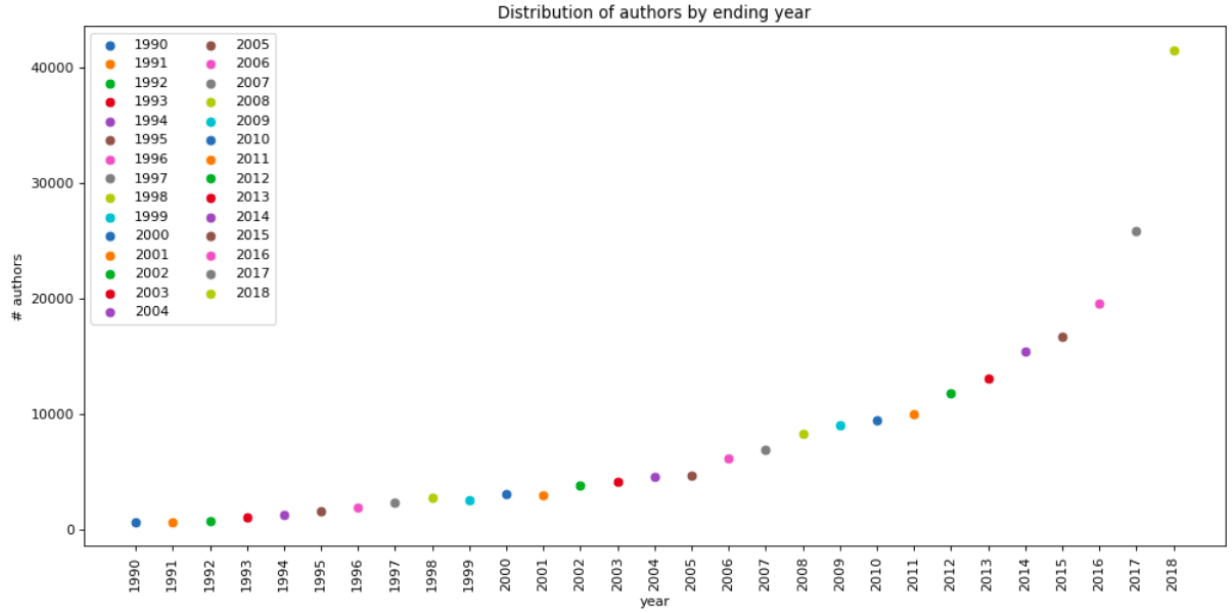

Figure 3.6: *Distribution of authors by ending year*

3.3 Filtering 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**, 36795 authors shown in Fig. 3.7.

	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 3.7: *Identified subset of active author*

3.4 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.
In the possessed data is present the state of the Network for each year, instead our model does not refer to the year, but to the appearance of author or collaboration.
Because **an event can be the appearance of a new node or a new edge**, the set of year, containing only 28 time step, can be too small in order to build meaningful vertex trajectories, that's why from now on, other metrics are used: the **occurrence of a new author**, the **occurrence of a new collaboration** and the **occurrence of a new publication** as time steps.

Their distributions for both all authors and the active subset are showed respectively in Fig.3.8 and Fig. 3.9. where on the x axis are shown the years and on the y axis the number of new publications (top-left), new authors (down-left) and new collaborations (right)

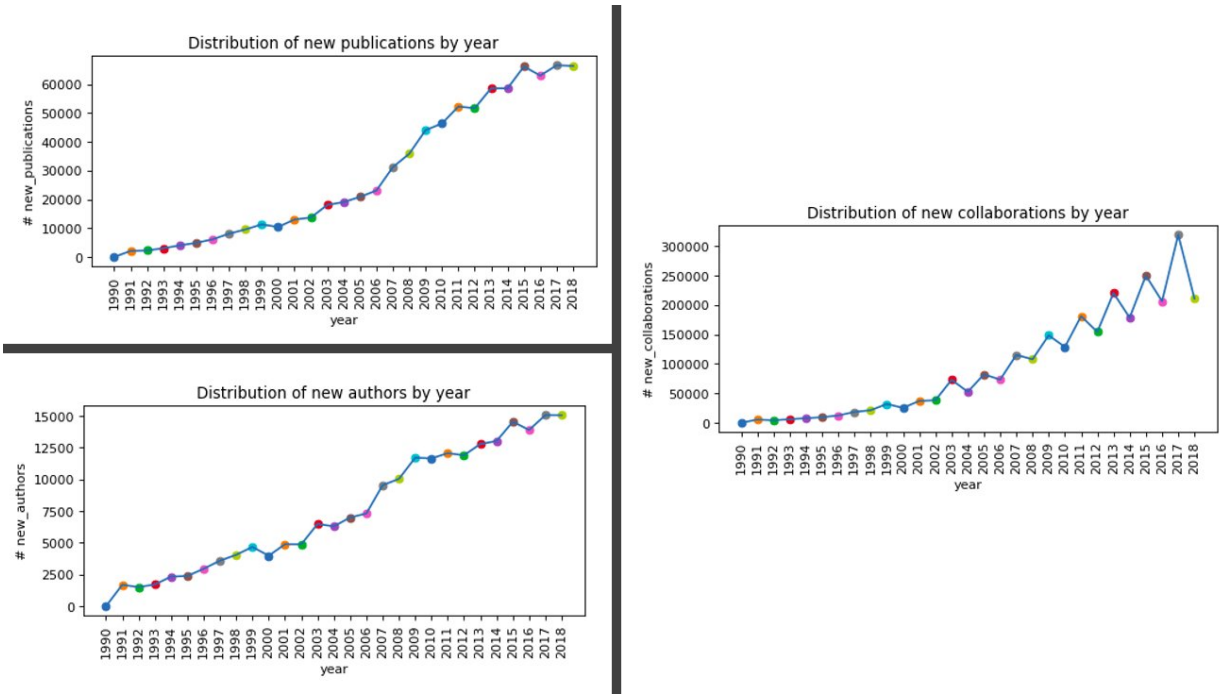

Figure 3.8: *Distribution by year of new collaborations (right), new authors (bottom-left) and new publications (top-left) for all authors in the data.*

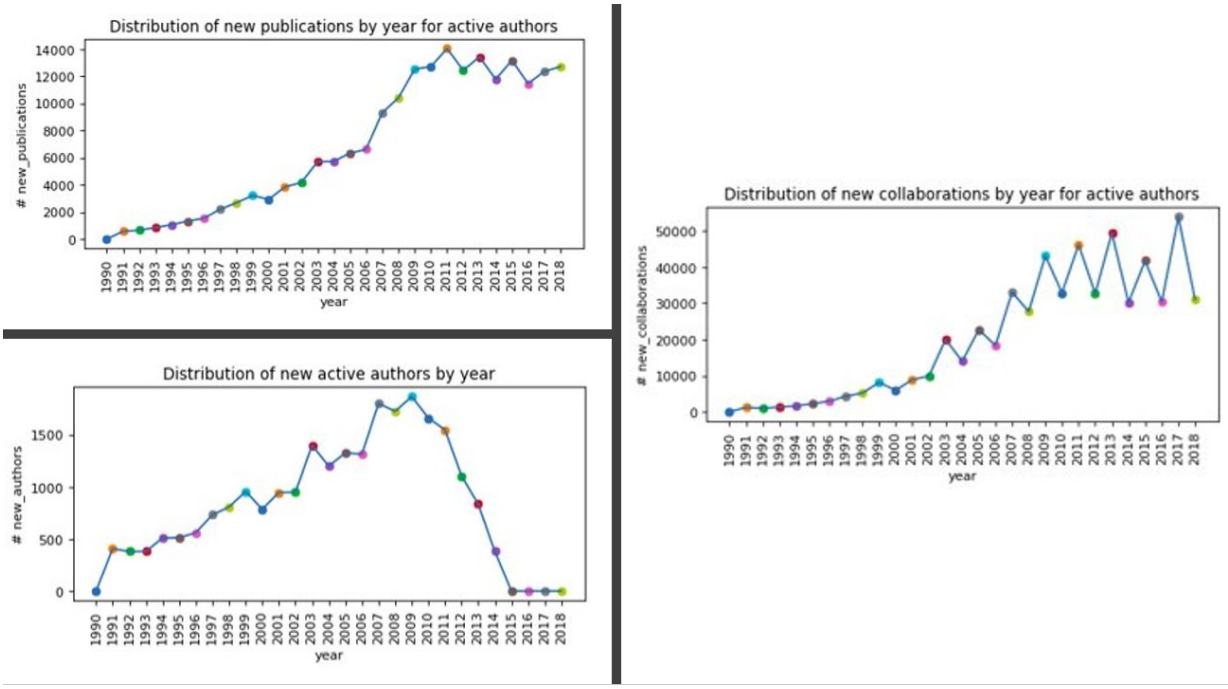

Figure 3.9: *Distribution by year of new collaborations (right), new authors (bottom-left) and new publications (top-left) for the active subset of authors.*

Applying the described metrics results in a **stretching of the x axis** in the plotted trajectories as in Fig.3.10, in this way the trajectory starts to show the **logarithmic shape discussed in the Chang-Lu model** (Section 2.6).

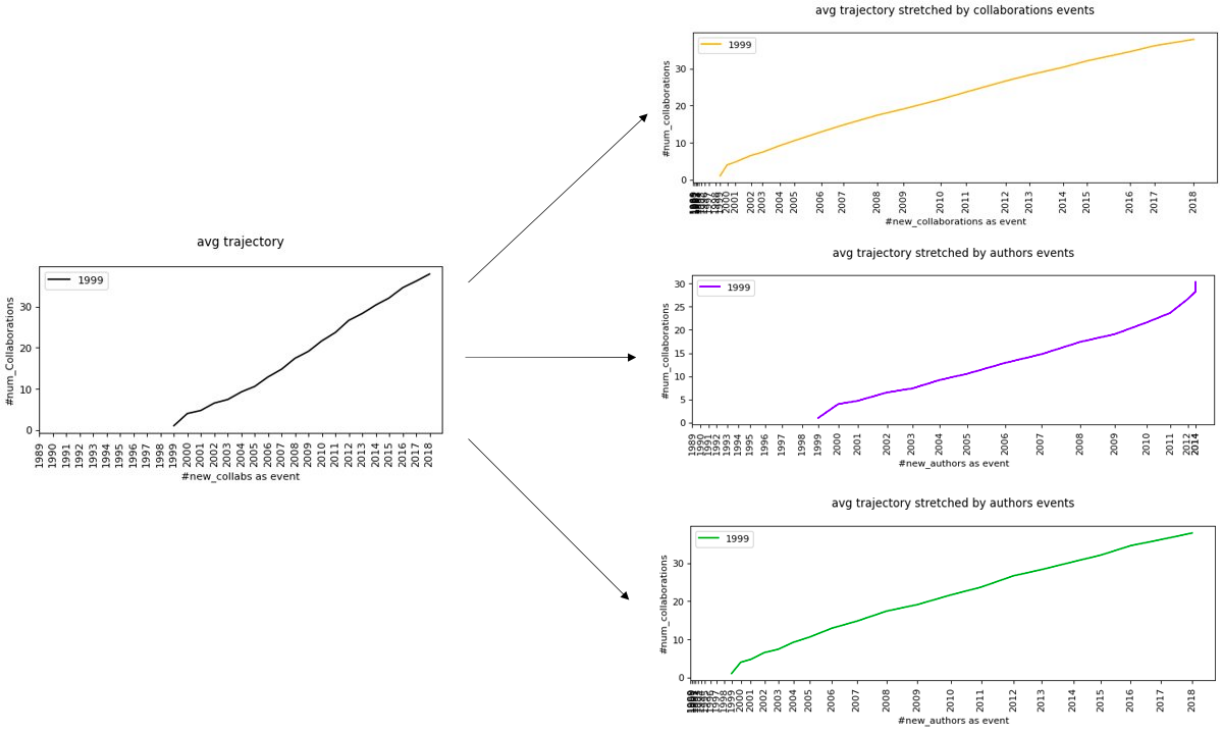

Figure 3.10: *The average vertex trajectory for all authors who started in 1999 (on the left) in stretched using the appearance of new collaborations (top) and of new authors (middle) and of new publications(bottom).*

Chapter 4

Getting to know data characteristic

4.1 Degree Distribution

The degree distribution has been computed to acquire more knowledge from the character of the data.

Fig. 4.1 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. 4.2 contains it's log-log form.

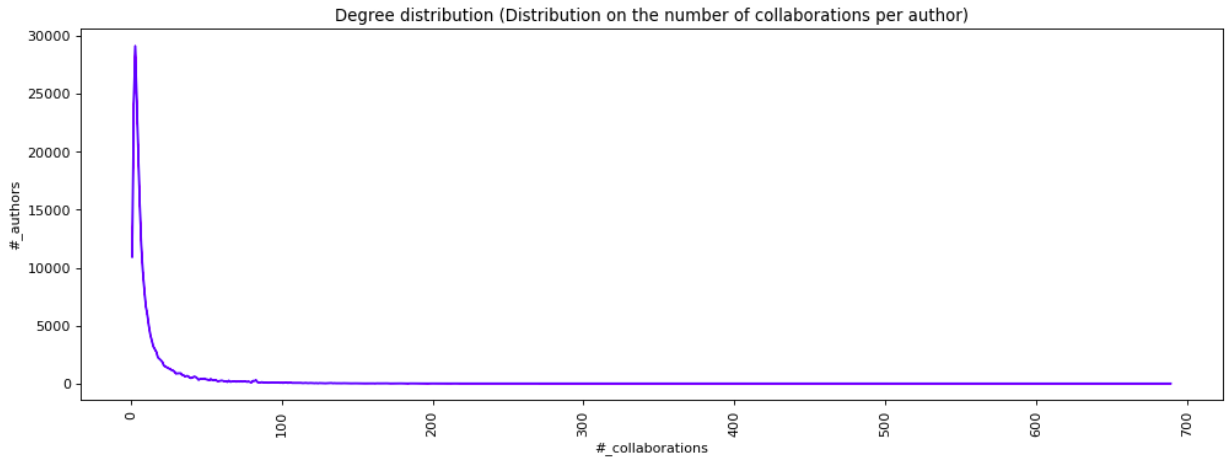

Figure 4.1: *Degree distribution.*

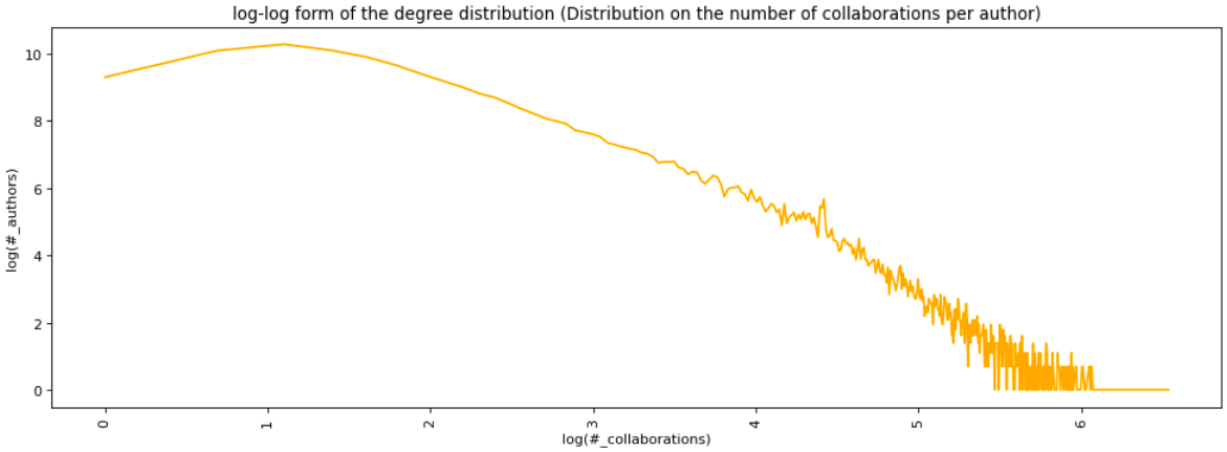

Figure 4.2: *Degree distribution in it's log-log form.*

As said before the scope of this project is to find a matching model better representing the evolution of the network, such model is supposed to be the **Chung-Lu** one, so we expect a **stretched exponential** degree distribution of the form $M_k \sim \alpha \cdot k^{-\gamma} \cdot \exp\left\{-\frac{\alpha}{1-\gamma}k^{1-\gamma}\right\}$. In the next chapters **attempts of fitting** this distribution with the expected one will be made.

4.2 Average Vertex Trajectories

Here **average vertex trajectories** are retrieved, each average curve is associated to **authors with same starting publication year**.

In Fig.4.3. is showed an example of average trajectory for authors who started to publish in 1999, on the x axis are represented the **years stretched by the number of new collaborations** as event and on the y axis the average number of collaborations. In Fig.4.4 are instead plotted all the average trajectories by starting year.

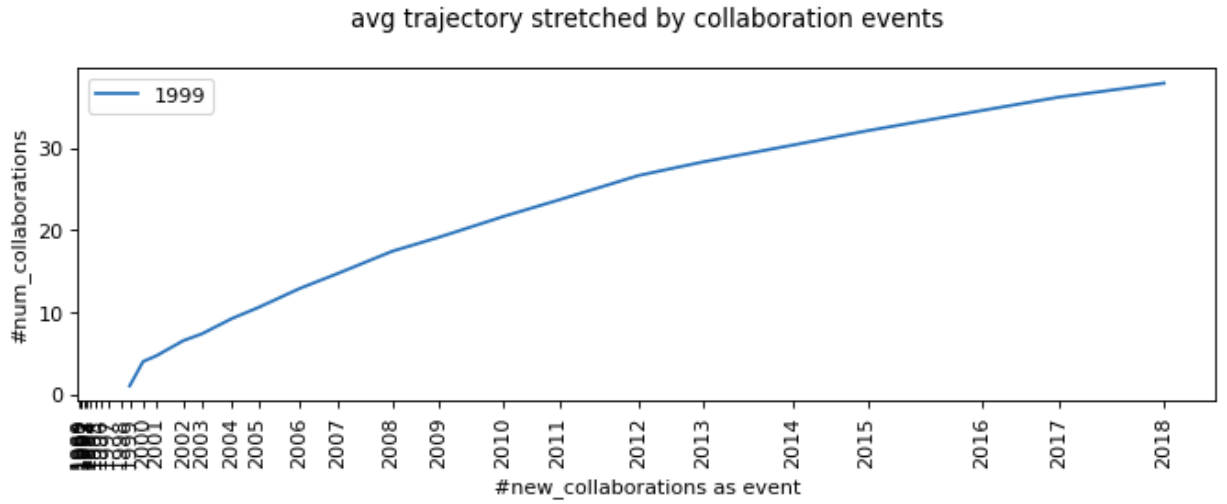

Figure 4.3: *Average trajectory associated with active authors that started their career in 1999 - stretched by the occurrence of new collaborations.*

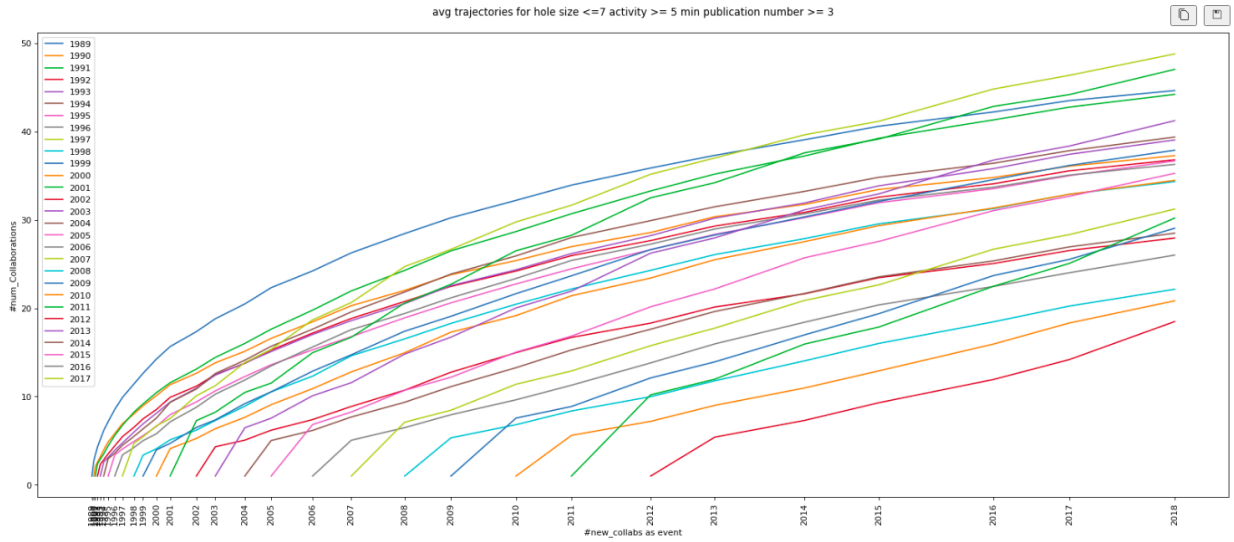

Figure 4.4: *Average trajectories by starting year for all active authors - stretched by the occurrence of new collaborations.*

The investigation of those trajectories is one of the main goals of this project, in fact, in later chapters, **attempts of fitting** those curves to the Chung-Lu function $d_v(t) = \left(\frac{1-\gamma}{\alpha}\ln(t/t_v) + 1\right)^{1/(1-\gamma)}$ described in the state of the art (section 2.2) will be made. Notice that, in Fig.4.4, years after the 2012 are not considered, they contain too few data to represent the logarithmic shape looked for.

Chapter 5

Fitting

5.1 Degree Distribution Fitting

In this section are made attempts to fit the degree distribution of the active subset of authors with both the power law function and the one with exponential cutoff, because, as already mentioned, we expect a **stretched exponential** degree distribution of the form $M_k \sim \alpha \cdot k^{-\gamma} \cdot \exp\left\{-\frac{\alpha}{1-\gamma}k^{1-\gamma}\right\}$, that is, the degree distribution expected in the Chung-Lu model.

The fitting with the **power law** function $M_k \sim Ck^{-\lambda}$ brings to have the constant parameter $\boldsymbol{C = 124534.108}$, and the exponential one $\boldsymbol{\lambda = 1.698}$, the fitted function is the orange curve in the chart shown in Fig. 5.1. Instead fitting our degree distribution with the **exponential cutoff power law** $M_k \sim Ck^{-\lambda}\gamma^k$ results in the couple of parameter $\boldsymbol{C = 22286.576}$, $\boldsymbol{\lambda = 1.04}$ and $\boldsymbol{\gamma = 0.984}$, the fitted curve is the green one in Fig. 5.1.
In Fig. 5.2 is instead shown the log-log form of the degree distribution along with the log-log form of the fitted functions.

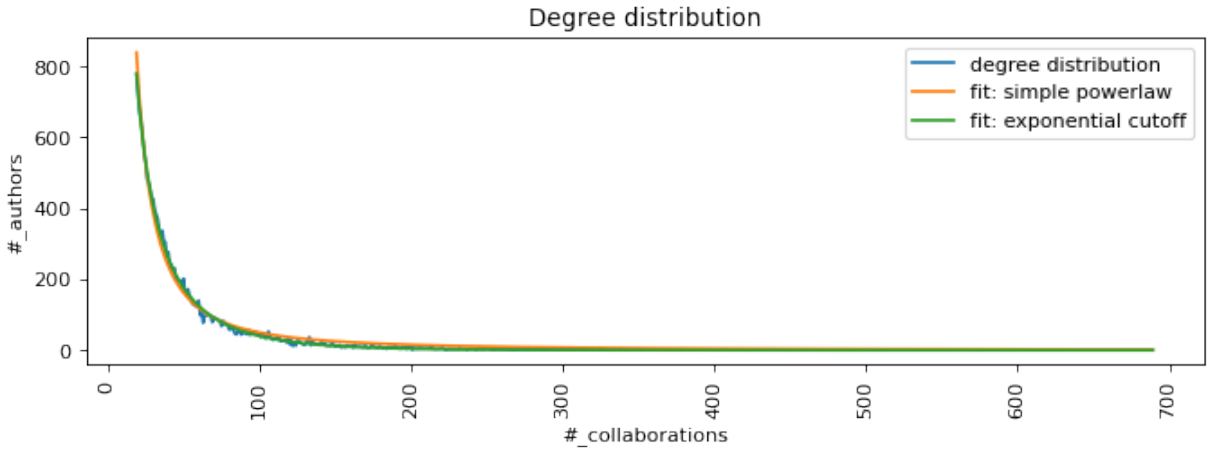

Figure 5.1: *Degree distribution (in blue), fitted power law function (in orange) and fitted power law with exponential cutoff (in green).*

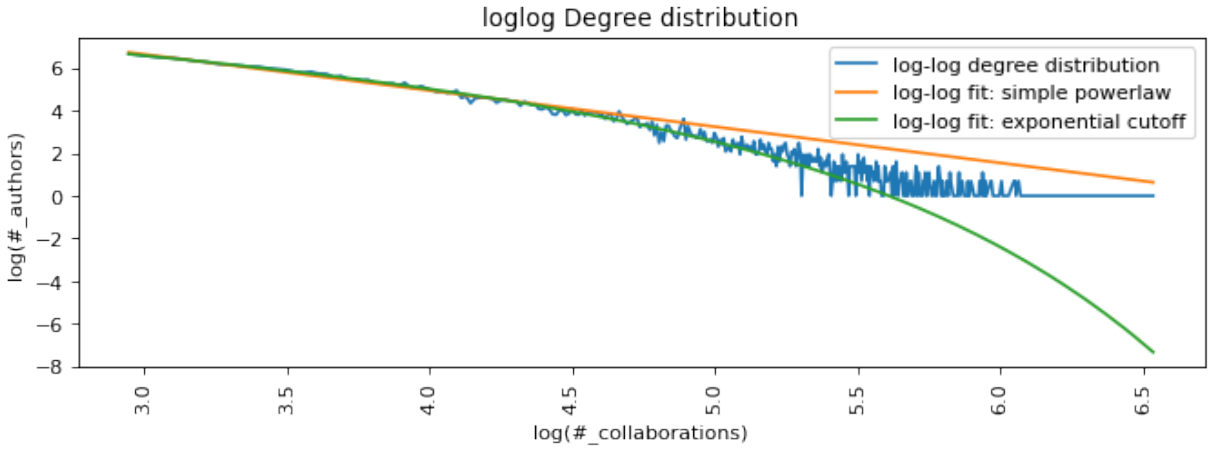

Figure 5.2: *Log-log versions of Degree distribution (in blue), fitted power law function (in orange) and fitted power law with exponential cutoff (in green).*

By this fitting seems that the **exponential cutoff distribution** better fits the given data, so the chances that our network falls in the **Chung-Lu** model are higher.
In the next sections, in order to confirm such assumption, the average vertex trajectories of active authors will be fitted with the logarithmic Chung-Lu trajectory $d_v(t) = \left(\frac{1-\gamma}{\alpha}\ln(t/t_v) + 1\right)^{1/(1-\gamma)}$.

5.2 Vertex Trajectories Fitting

In section 4.2 **average vertex trajectories** are retrieved, where each average curve is associated with **authors with same starting publication year**. The investigation of this curve is needed in order to check if they present the logarithmic shape expected in the **Chung-Lu model**.

The **Chung-Lu trajectory** introduced in section 2.6, $d_v(t) = \left(\frac{1-\gamma}{\alpha}\ln(t/t_v) + 1\right)^{1/(1-\gamma)}$, will be rewritten in the following form:

$$d_v(t) = (a\ln(t/t_v) + 1)^{\beta}$$

where $a = \frac{1-\gamma}{\alpha}$ and $\beta = 1/(1-\gamma)$. The next fitting attempts will refer to this function.

Each average trajectory, retrieved in section 4.2, has been **fitted with the yet cited function**, finding the best **couple of parameters** $\boldsymbol{a}$ and $\boldsymbol{\beta}$ able fit it.

In Fig.5.3, as example, is showed the average trajectory for authors who started to publish in $t_v = 1999$, so the real trajectory $r_{1999}(t)$ (the blue curve) and it's fitting $d_{1999}(t)$ (the dotted red curve). On the x axis are represented the years stretched by the number of new collaborations as event t and on the y axis the average number of collaborations. In Fig.5.4 are instead plotted all the average trajectories by starting year t_v (stretched by number of new collaborations) and their fitting until 2012.

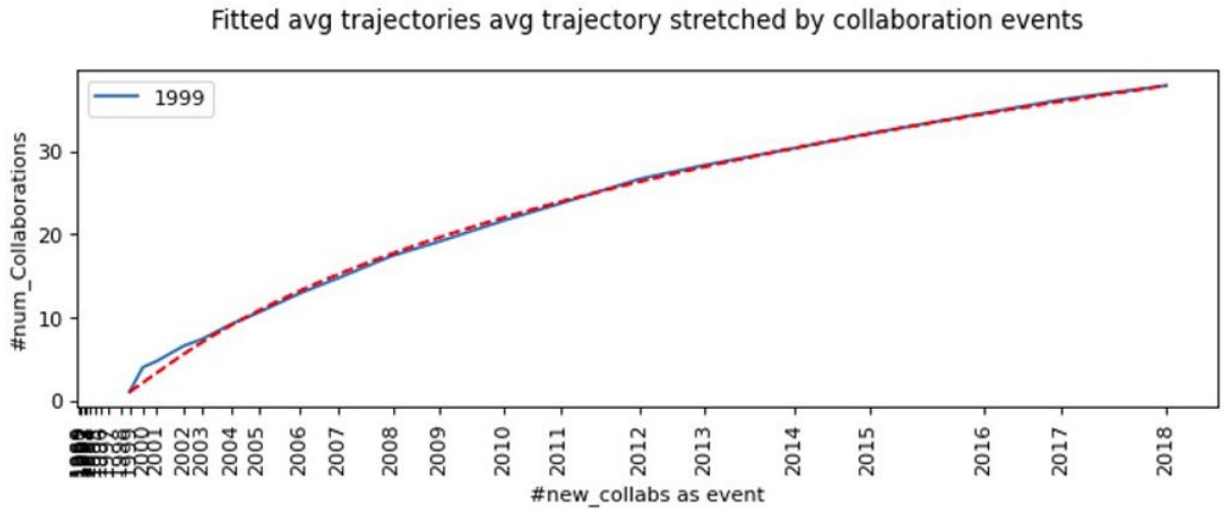

Figure 5.3: *Average trajectory associated with active authors that started their career in 1999 and it's fitting - stretched by the occurrence of new collaborations.*

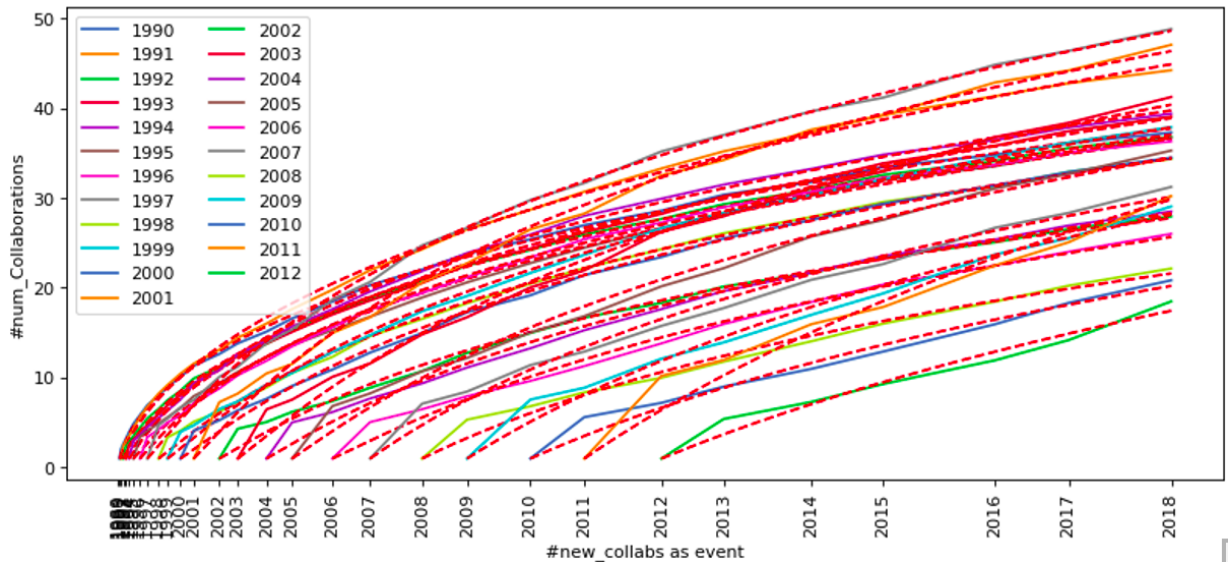

Figure 5.4: *Average trajectories by starting year for all active authors and their fittings - stretched by the occurrence of new collaborations.*

Each of the previous fitting d_{t_v} relative to the average curve r_{t_v}, associated to authors who started in the time step t_v, brings to **a couple $\{a_{t_v}, \beta_{t_v}\}$ of parameters** for each $t_v = 1990, ..., 2012$. The charts in Fig 5.6 and Fig. 5.7, respectively shows a_{t_v} and β_{t_v} (on the y axis) for each starting year t_v (on the x axis).

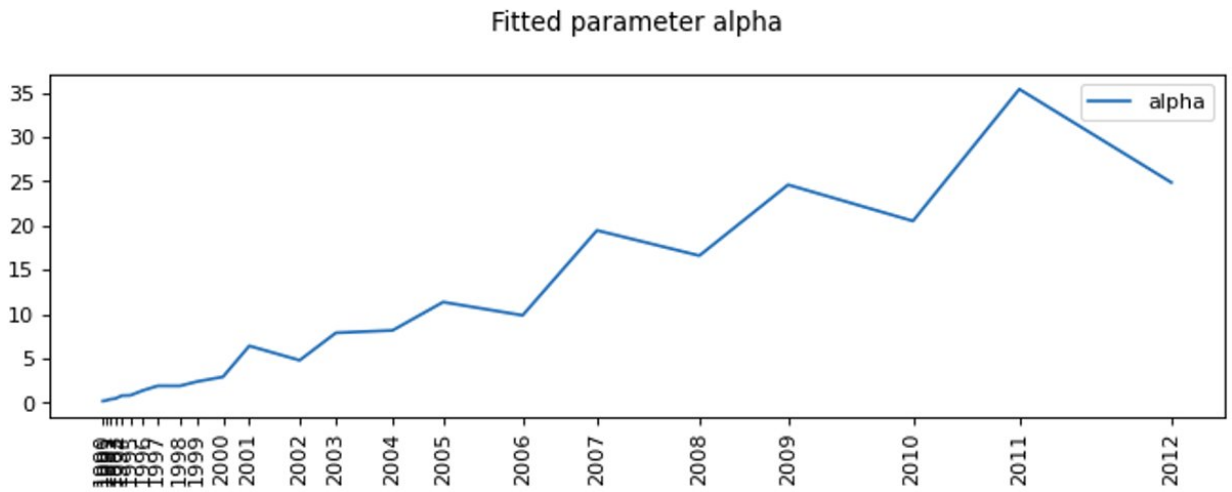

Figure 5.5: *Values of the parameter a_{t_v} (y axis) resulting from the fitting, for each starting year t_v.*

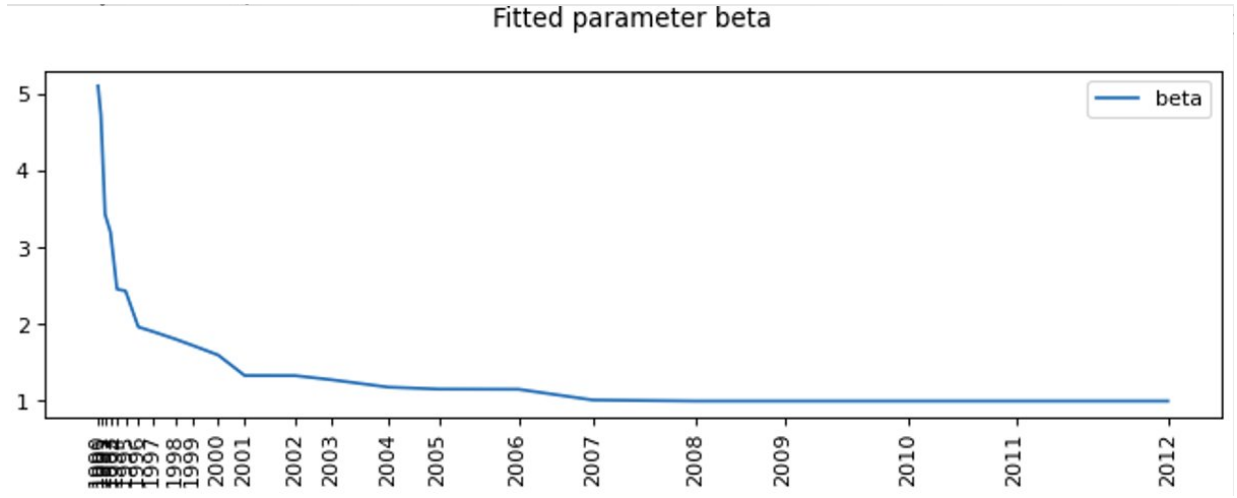

Figure 5.6: *Values of the parameter β_{t_v} (y axis) resulting from the fitting, for each starting year t_v.*

From those chart we can observe that the parameter a **grows linearly** while β seems to **converge to the value 1**.
Can be concluded that the function used is particularly well suited to fit this data, but in this case for each starting year there are different values of a and β. The next step will be to **generalize** our parameters, finding an **optimal couple** $\{a^*, \beta^*\}$, and so an **optimal function** $d^*_{t_v}(t)$ able to fit, with enough accuracy, all the averages $r_{t_v}(t)$

5.3 General Fitting function for trajectories

In the previous section has been retrieved **multiple couples of parameters** $\{a_{t_v}, \beta_{t_v}\}$, each relative to the fitting function $d_{t_v}(t)$ associated with the real average trajectory $r_{t_v}(t)$. In this section will be retrieved a **couple of general parameters** $\{a^*, \beta^*\}$, able to fit each curve $r_{t_v}(t)$.

In order to retrieve those parameters an **error metric** should be defined, the the fitting procedure will then try to find the best couple of parameters in order to **minimize the defined error**.

5.3.1 Error definition

Given a **starting event** t_v, so the time step in which the node v joined the network and t a **generic event**:
- Let $r_{t_v}(t)$ be the function representing the **real average trajectory** for authors who started to publish at the **starting event** t_v;
- Let $d_{t_v}(t)$ be the **fitted function** of $r_{t_v}(t)$;
- Let $d^*_{t_v}(t)$ be the **general fitting function** of which we want to **optimized the parameters**, for authors who started to publish at the event t_v.

The following **four kind of error** are defined:

$$(A)\ min_{\alpha^*,\sigma^*}\left(\sum_{t_v}\sum_{t\geq t_v}|f^*_{t_v}(t)-r_{t_v}(t)|^2\right)$$

$$(B)\ min_{\alpha^*,\sigma^*}\left(\sum_{t_v}\max_{t\geq t_v}|f^*_{t_v}(t)-r_{t_v}(t)|^2\right)$$

$$(C)\ min_{\alpha^*,\sigma^*}\left(\sum_{t_v}\sum_{t\geq t_v}|f^*_{t_v}(t)-f_{t_v}(t)|^2\right)$$

$$(D)\ min_{\alpha^*,\sigma^*}\left(\sum_{t_v}\max_{t\geq t_v}|f^*_{t_v}(t)-f_{t_v}(t)|^2\right)$$

Next those error functions will be optimized in order to find the best couples $\{a^*,\beta^*\}$ able to minimize them.

5.3.2 Optimization Results

In the following table are shown the **optimized values** of a^* and β^* minimizing the relative error function.

	ERROR	a	β
(A)	27755.784282	6.929222	1
(B)	4301.973888	8.973418	1
(C)	444951.754961	7.993638	1
(D)	4714.507436	9.169267	1

It's important to notice that the value of β^* **results always 1**. The implications of $\beta^*=1$ will be discussed later in the chapter.

5.3.3 Results implications

In Fig. 5.7 and Fig. 5.8 are showed an example of the yet retrieved optimal fitting functions for authors who started publishing in 1996 and 2000. The blue curve is the average function $r_{t_v}(t)$, instead the optimal ones $d_{t_v}(t)^*$ relative to errors A, B, C and D are respectively drawn in yellow, green, red and violet. On the x axis are present the years t stretched by the number of new collaborations, while on the y axis there is the average number of collaborations, that is $d_{t_v}(t)$.

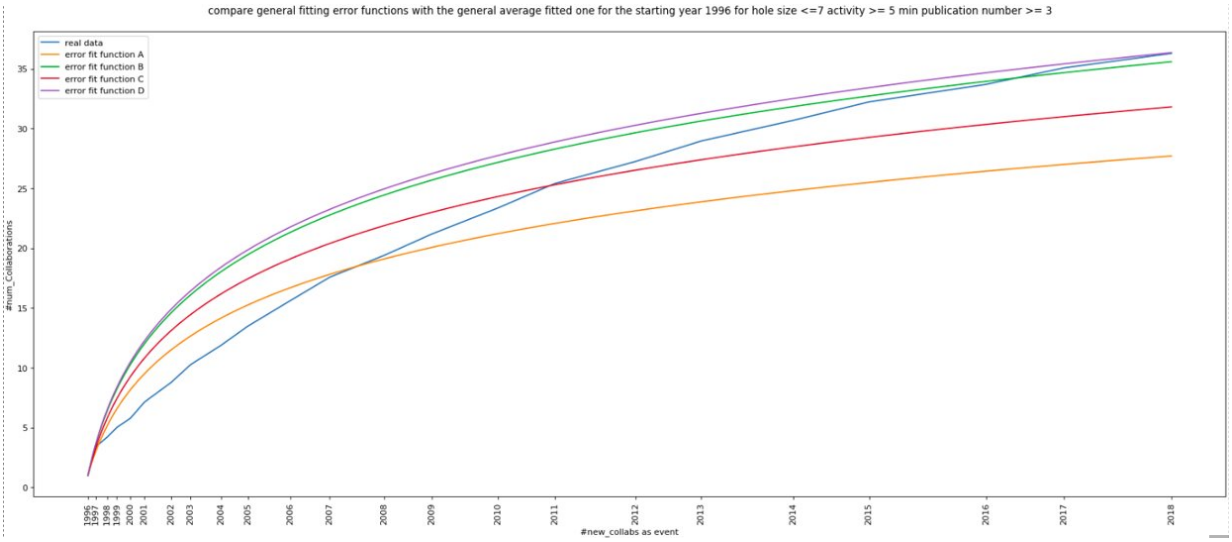

Figure 5.7: *Optimal fitting functions for authors who started publishing in 1996.*

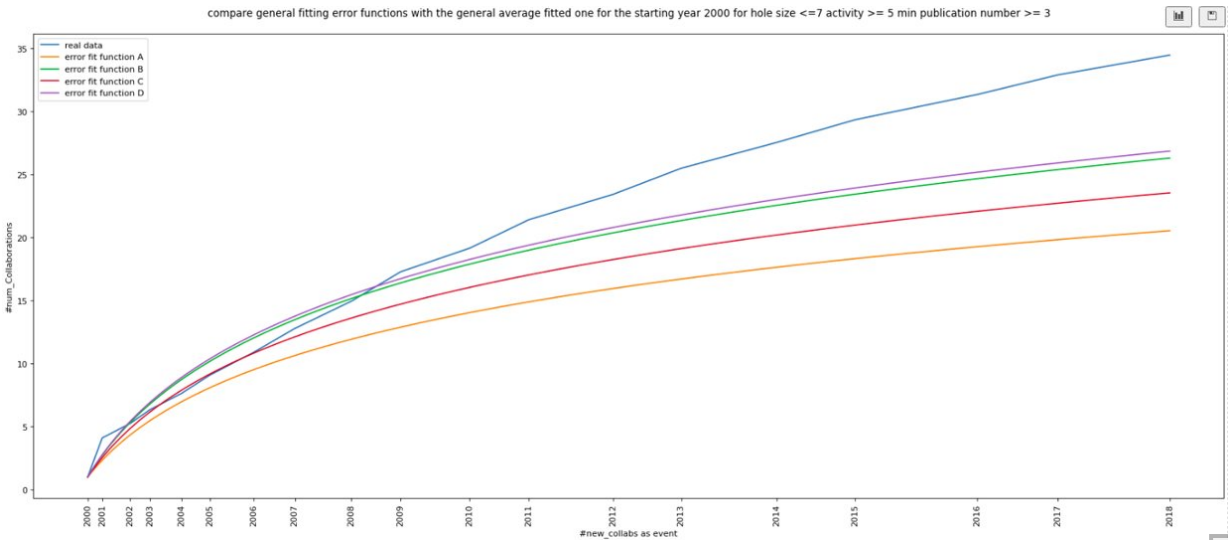

Figure 5.8: *Optimal fitting functions for authors who started publishing in 2000.*

All the optimal fitting functions $d^*_{t_v}(t)$ performs a better fitting for a low starting year t_v, this may be due to the fact that for high values of t_v there's less data to fit.

As said before the optimal value of β^* seems to be **always 1** (section 5.3.2).

Reminding that in the rewritten Chung-Lu function $d_v(t) = (a \ln(t/t_v) + 1)^\beta$ is defined $\boldsymbol{\beta = 1/(1)}$, by assigning a value $\beta = 1$, being $\gamma = 1 - 1/\beta$, we obtain that $\boldsymbol{\gamma = 0}$.

This results **contradict the fact that the data falls in the Chung-Lu model**. In the model the attachment function $f(d_v(t)) = d_v(t)^\gamma$ is directly related to the value of γ.

A value of $\gamma = 0$ implies that the **attachment function is linear** $\boldsymbol{f(d_v(t)) = 1}$, it means that the probability of a node to be chosen for the attachment Pr doesn't depend from the current degree $d_v(t)$ of the node, instead the **probability of being chosen is distributed uniformly** over the set of nodes V_t, that is, each node has the **same probability to be chosen** at each time step t:

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

Chapter 6

Conclusions

The main scope of the project was to find a **matching mathematical model**, able to describe the evolution over time of the graph built upon **collaborations between french computer science researchers from 1990 to 2018**.

The first step has been the **retrieval of the collaboration data** (Section 3.1), through the JSONs obtain through **Scopus** APIs, this data has then been analyzed in order to find a **sufficiently large and meaningful** subset of authors **active** in the research activity. This has been carried on by a filtering on the metrics described in Section 3.2: the maximum **hole size**, the **activity period** and the minimum **total number of publications**. The filtering (Section 3.3) resulted in a set of **36795 active authors**, so the **16%** of the original data.

The expected matching model has been describe in the state of the art (section 2.6), The Chung-Lu one.

A model defines the way the network evolves given the **probability P of node and edge events** t to happen (Section 2.1) and a **preferential attachment function** $f(d_{t_v}(t)) = d_{t_v}(t)^\gamma$ (Section 2.1) which depends from the degree $d_{t_v}(t)$ of a node v at time t. The attachment function defines the **probability for a node to be chosen for the attachment** when a node or an edge event occurs $\Pr[v \text{ is chosen at step t }] = \frac{f(\mathrm{d}_v(t))}{\sum_{w \in V_t} f(\mathrm{d}_w(t))} = \frac{1}{|V_t|}$. (Section 2.1).

The main investigation of this work was relative to two measures: the **degree distribution** and The **Vertex trajectory**, describe respectively in Section 2.2 and Section 2.5 of the state of the Art. The **Chung-Lu model** respect a slight variation of power law degree distribution with exponential cutoff (Section 2.4), the so called **stretched exponential** (table in Section 2.6):

$$M_k \sim \alpha \cdot k^{-\gamma} \cdot \exp\left\{-\frac{\alpha}{1-\gamma} k^{1-\gamma}\right\}.$$

and a logarithmic vertex trajectory (table in Section 2.6)

$$d_v(t) = \left(\frac{1-\gamma}{\alpha}\ln(t/t_v) + 1\right)^{1/(1-\gamma)}$$

In section 4.1 the degree distribution of the network has been retrieved and in section 5.1 it has been fitted with both the classic power law and the one with exponential cutoff, concluding that the **data is better represented by the exponential cutoff one**, so seems to respect the Chung-Lu model.

To define meaningful vertex trajectories $d_{t_v}(t)$, so the degree t, of a vertex v who joined the network at time step t_v, the set of **natural time steps**, so the set of **years from 1990 to 2018**, was not enough representative. In order to face this problem, in section 3.4, have been chosen to use a different sets of events, the number of: **new collaborations**, **new authors** and **new publications**.

The **average vertex trajectories** (Section 4.2), defined over authors with the same starting year, has been fitted, in section 5.2, to the Chung-Lu logarithmic function $d_v(t) = (a\ln(t/t_v) + 1)^{\beta}$.

In order to find a couple of **optimal parameters** $\{a^*, \beta^*\}$ able to fit each average function some **error metrics** has been defined (section 5.3.1). Then by **minimizing the error functions** we obtained four couple of parameters, all showing a value of $\boldsymbol{\beta = 1}$.

The shown value of β implies a value of $\boldsymbol{\gamma = 0}$ in the attachment function (Section 5.3.3). From this result follows that in the model the probability of a node to be chosen for the attachment at time step t is the **same for each node**, so the underlying model is **not the Chung-Lu one**, where the attachment function is sublinear $f(d_{t_v}(t) = d_{t_v}(t)^{\gamma} \mid 0 < \gamma <= 1$.

The result on the vertex trajectory **contradicts** the one on the degree distribution, so or the underlying model is **totally different** from the expected one, or an **error has been committed**. in either case further investigations are required.

A next step on this research may be an analysis of data coming from **research fields different from computer science**, to check weather or not they show the same contradicting conclusions. Also, a network given the supposed underlying theoretical model, so a linear attachment function $f(d_{t_v}(t)) = 1$, can be **generated and the fitted** to our data to double check the validity of those conclusions. In the same way slightly **different model can be generated**, in order to investigates which one is more suited in this case and lastly attempts of fitting can also be made over **other existing models** [11].

Chapter 7

Bibliography

1 Derek J. de Solla P. Networks of scientific papers, 1965, doi:10.1126/science.149.3683.510.

2 Michalis F., Petros F., and Christos F. On power-law relationships of the Internet topology, 1999, doi:10.1145/316188.316229.

3 Bollobás B. and Riordan O. Handbook of Graphs and Networks: From the Genome to the Internet, 2003, ISBN: 978-3-527-60633-7.

4 Broido A. D.and A. Clauset, "Scale-free networks are rare", 2019, doi: 10.1038/s41467-019-08746-5.

5 Newman M. E. J. Coauthorship networks and patterns of scientific collaboration, 2004, doi:10.1073/pnas.0307545100.

6 Newman M. E. J. Clustering and preferential attachment in growing networks, 2001, doi:10.1103/PhysRev

7 A.-L. Barabasi and R. Albert, "Albert, R.: Emergence of Scaling in Random Networks, 1999, doi: 10.1126/science.286.5439.509.

8 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

9 Chung F. and Lu L., Complex Graphs and Networks, 2006.

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

11 Cooper C. , Alan F. A General Model of Web Graphs, 2001