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

Chapter 1

Introduction

1.1 General Project Description

1.1.1 Framework/Context

This project is a part of a larger one involving researchers in various fields, such as economics, sociology and computer science; it is focused on the evaluation of the impact of funding on scientific research. As expected impact is meant that, for example, given a couple of authors with similar collaboration behaviors, if one of them get a funding, his collaboration rate is expected to grow compared to the others, unfortunately, this is not always true in the analyzed data.

As an example of funding one can indicate LabEx and IdEx, French programs whose scope is to promote collaborations involving different research fields.

The purpose of this project is to analyse the evolution of nodes degree in a collaboration network built upon scientific publications extracted from Scopus Database, that is, **the vertex trajectory** (Section 2.2 - state of the art), where for collaboration network is meant: a set of nodes, the authors, connected by edges, representing collaborations, such that two nodes are connected if there exists at least one collaboration between them.

1.1.2 Motivations

Many systems can be represented as a network, both natural as well as human built, such as the World Wide Web, social networks, collaborations of actors in movies, or even the interaction among molecules. Each of this systems can be viewed as a set of nodes, e.g. routers, computers or people, and a set of edges connecting them, e.g., data-link among computers, social relations among people or, as in our case, **collaborations between scientific researchers**.

Those systems are of practical interest and **attract researchers in different fields of study**, since we reached the computational power to deal with the large amount of data and since we have mathematical models to describe their evolution over time. This can bring to a wide range of tools and applications to analyze the structure of those systems and ways to predict their evolution.

A common property of those networks is that they are **scale-free**, it means that the probability distribution of degrees of nodes over the network follows a **power-law distribution with exponential cut-off**, described in the **state of the art** (Section 2.2) [4]. The insight on it is that the proportion of vertices of a given degree changes following a power-law with exponential cut-off when the degree grows.

Power-law distribution can be seen in many other phenomena such as the frequencies of words in most languages, the sizes of craters on the moon, of Solar flare and so on. This feature is a consequence of the fact that a new node of the network connects preferentially to nodes that already have a huge number of connections. An intuitive example can be that, in social networks, the probability of having new friends is higher for people who already have a lot of them.
All those examples show that "**the development of large networks is governed by robust self-organizing phenomena that go beyond the particulars of the individual systems**"[1].

This project aims at investigating further features of scale-free networks. Particularly, it will concentrate on observing the evolution of degree over time.

1.1.3 Challenges

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

The first challenge would be to **avoid working on misleading data**. There can be authors who have published just once in their carrier or other that had a skyrocketing number of collaborations for a couple of years before disappearing from the network and so on. It is important to identify well the outliers to prevent their big impact on the average behaviour of the theoretical model.

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

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

1.1.4 Goals

The 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.2.5), where a new node appears in the graph, and an **edge event** (Fig.2.5) 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.2.5), 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, already present, to connect with, while in the edge event, two nodes must be selected to place the edge. This selection is carried out by an attachment function as $f(x)$ which defines the probability for any node $v \in V_t$ to be chosen for the attachment.

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

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

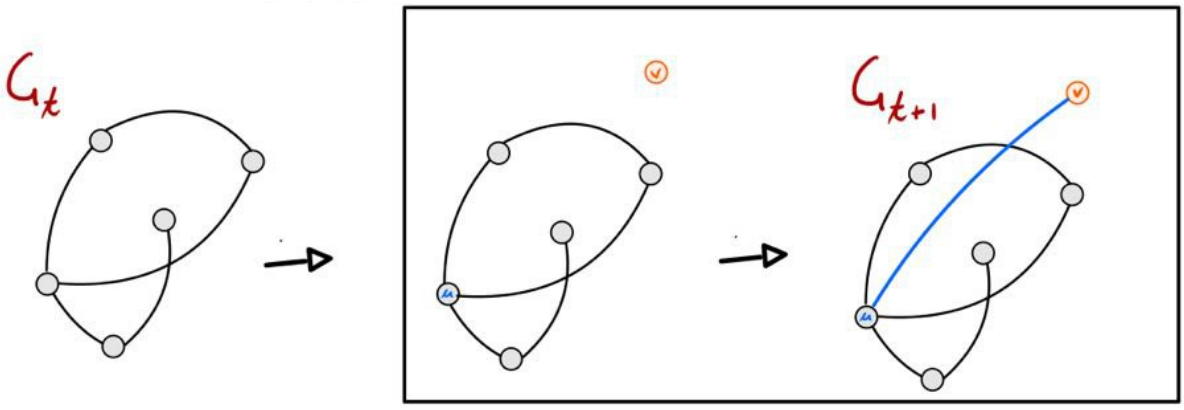


Figure 2.1: *Example of a node event: a new node v join the graph G_t at time t and connect to the node u, after the event the Graph G_{t+1} is obtained.*

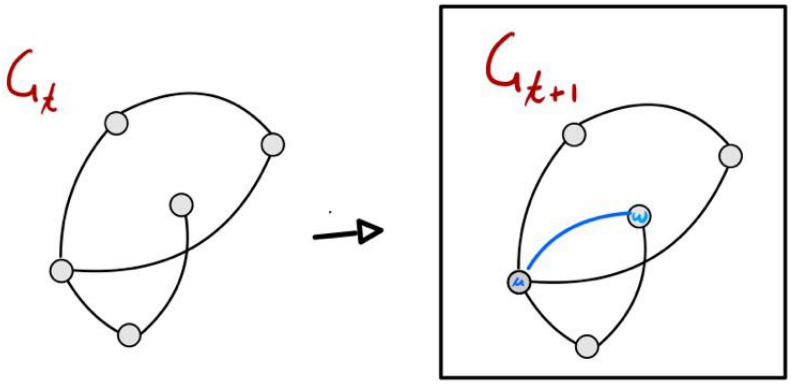


Figure 2.2: *Example of an edge event: a new edge (u, w) appears in the graph G_t at time t, after the event the Graph G_{t+1} is obtained.*

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.

2.1.1 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 is illustrated in Fig.2.1).

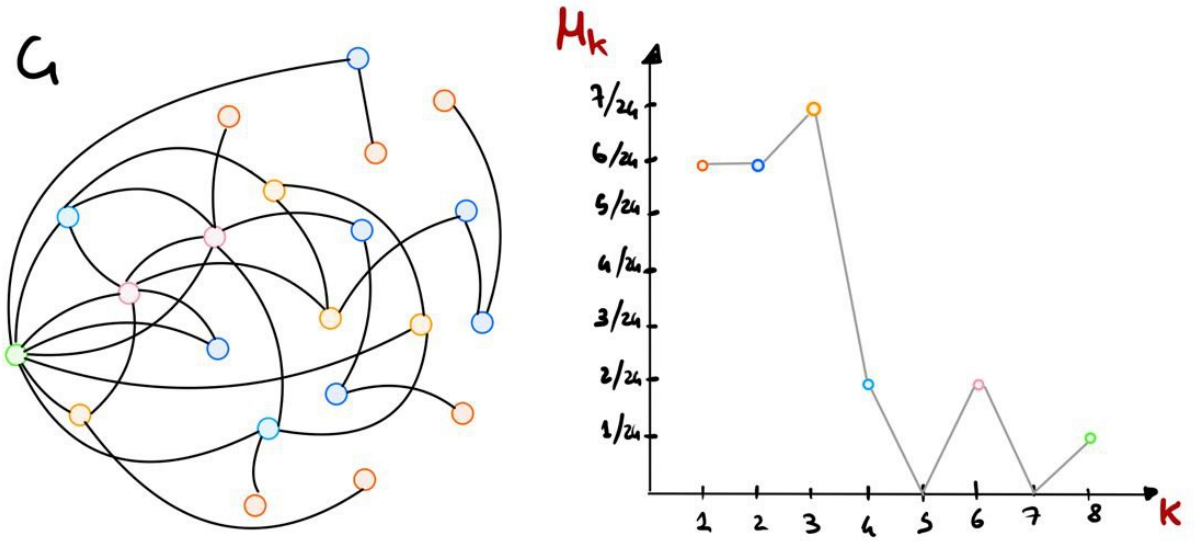


Figure 2.3: *Example of degree distribution: in G nodes of different degree have a different color, the degree distribution of G is plotted on the right.*

2.1.2 Power law distribution and scale free networks

We say that the degree distribution M_k of G follows a power law (Fig.2.2, image on top) 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.

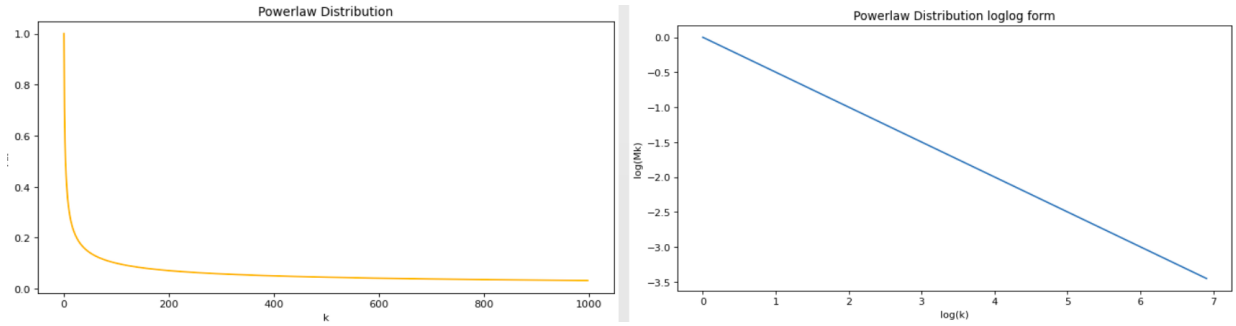


Figure 2.4: *power law distribution and it's log-log form.*

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.2.2, 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.

2.1.3 Power law distribution with exponential cutoff

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

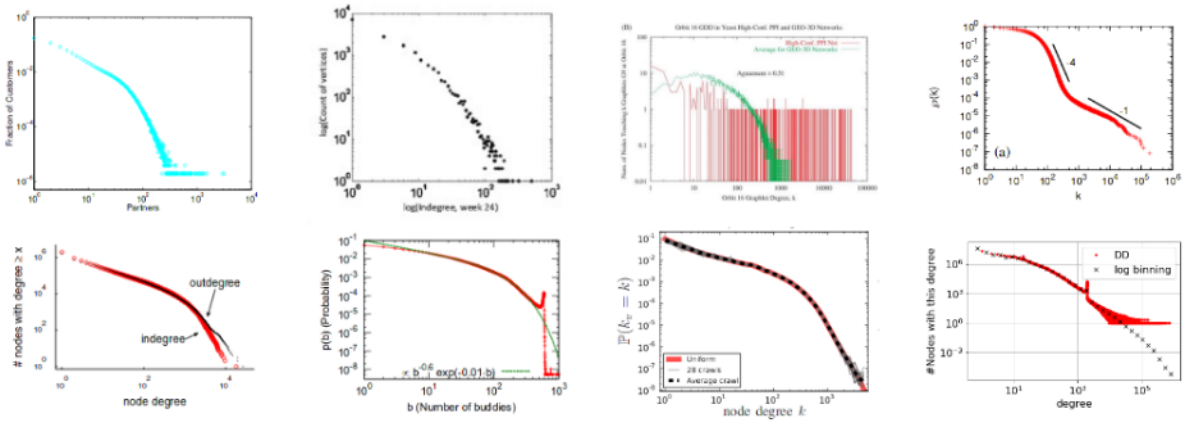


Figure 2.5: *Example of not scale free networks in biology, online market, social networks etc..*

It turns out that many of them (Fig.2.3) follow **power law distribution with an exponential cutoff** (Fig.2.4 - left) of the form:

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

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

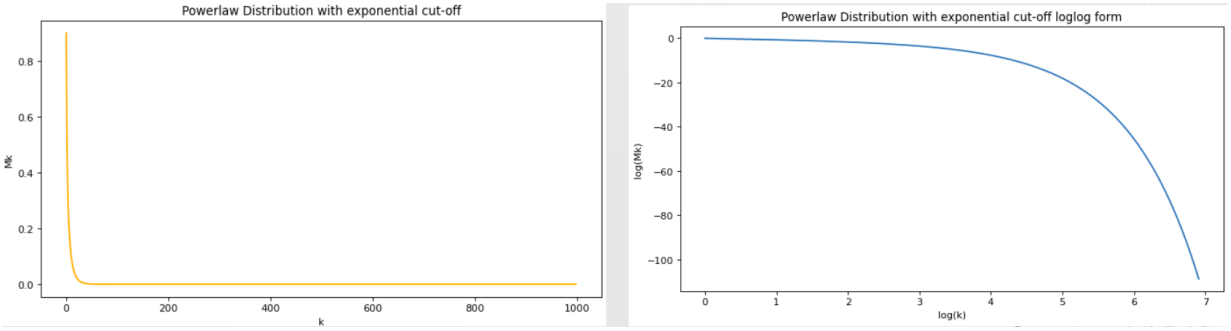


Figure 2.6: *power law distribution with exponential cutoff and it's log-log form.*

2.1.4 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 2.0.4).
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.

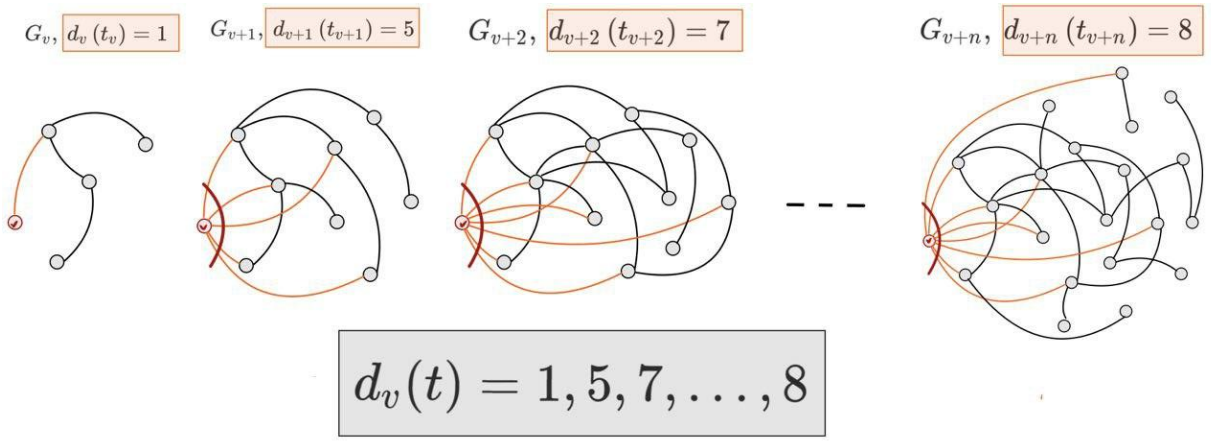


Figure 2.7: *Example of the vertex trajectory of a node v, in red, joining the network at time t_v. The edges of node v, in orange, at each time step they sum up to the degree of v. The sequence of degrees is the vertex trajectory, showed in the bottom.*

2.2 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, is directly related to the probability P of a node event to occur and the way the attachment function $f(x)$ is chosen.

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

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

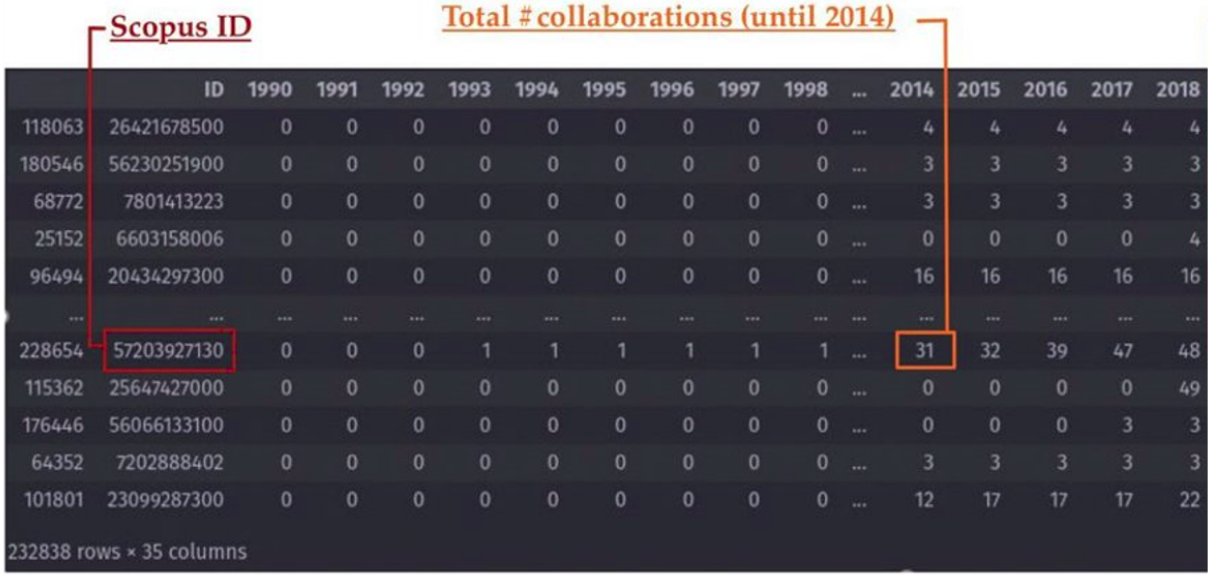
Chapter 3

Data Preparation

3.0.1 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.



	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.*

3.0.2 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.2.1, 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.

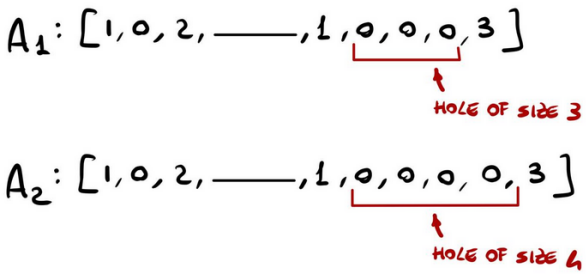


Figure 3.2: *Hole size definition.*

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 vales between 0 and 10, while on the y axis the number of authors associated with such values.

So, in order to build vertex trajectories, and get a better understanding of the data, for each author, is needed the year in which they started to publish as well as the one in which they stopped; this activity period is also a way of making groups for future comparison.

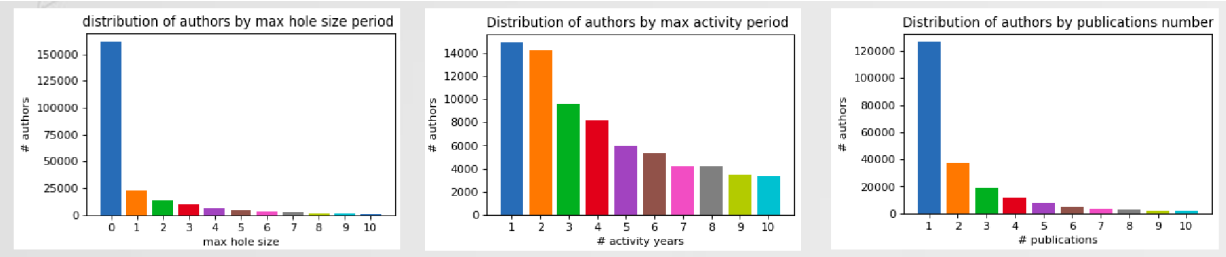


Figure 3.3: *Distribution of authors for hole size (left), activity period (middle) and minimum number of publications. All for values between 1 and 10.*

A new version of the collaboration dataset is built upon the one described in Section 3.0.1 (Fig.3.1), containing new columns with: the starting and ending publication year, the activity period, the maximum hole size and the number of publications for each author (Fig.3.4).

Scopus ID — Total # collaborations (until 2014) — publications years

	ID	1990	1991	1992	1993	1994	1995	1996	1997	1998	...	2014	2015	2016	2017	2018	start_year	end_year
118063	26421678500	0	0	0	0	0	0	0	0	0	...	4	4	4	4	4	2014	2014
180546	56230251900	0	0	0	0	0	0	0	0	0	...	3	3	3	3	3	2013	2014
68772	7801413223	0	0	0	0	0	0	0	0	0	...	3	3	3	3	3	2011	2011
25152	6603158006	0	0	0	0	0	0	0	0	0	...	0	0	0	0	4	2018	2018
96494	20434297300	0	0	0	0	0	0	0	0	0	...	16	16	16	16	16	2013	2013
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
228654	57203927130	0	0	0	1	1	1	1	1	1	...	31	32	39	47	48	1991	2018
115362	25647427000	0	0	0	0	0	0	0	0	0	...	0	0	0	0	49	2018	2018
176446	56066133100	0	0	0	0	0	0	0	0	0	...	0	0	0	3	3	2017	2017
64352	7202888402	0	0	0	0	0	0	0	0	0	...	3	3	3	3	3	2006	2006
101801	23099287300	0	0	0	0	0	0	0	0	0	...	12	17	17	17	22	2009	2018

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 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 publish in 2018 may still be active also nowadays.
Those plots shows also that there are more new active authors who start publishing each year, and that the number of authors who stop to publish is also increasing, but with a lower rate.

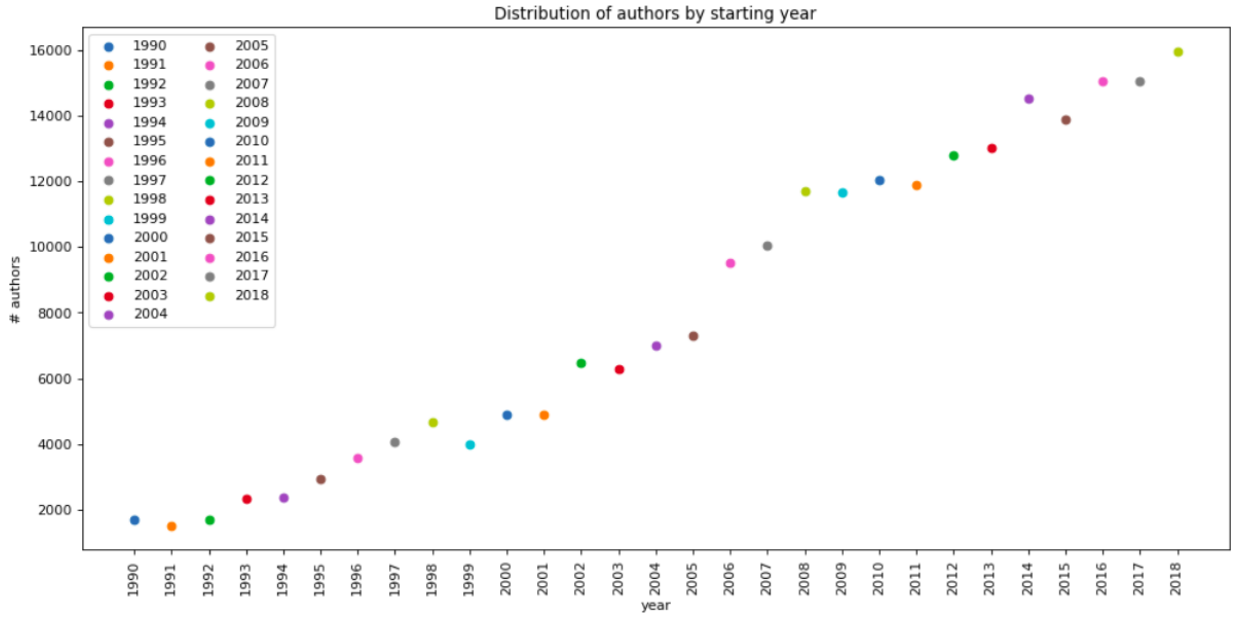


Figure 3.5: *Distribution of authors by starting year*

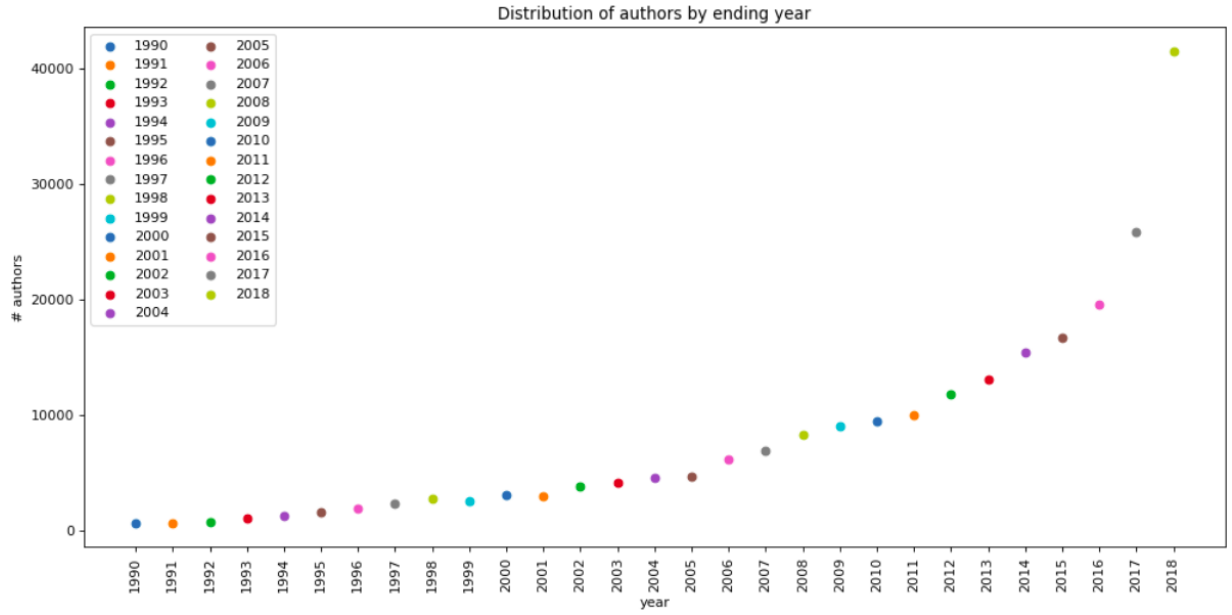


Figure 3.6: *Distribution of authors by ending year*

3.0.3 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. .

	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*

3.0.4 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.

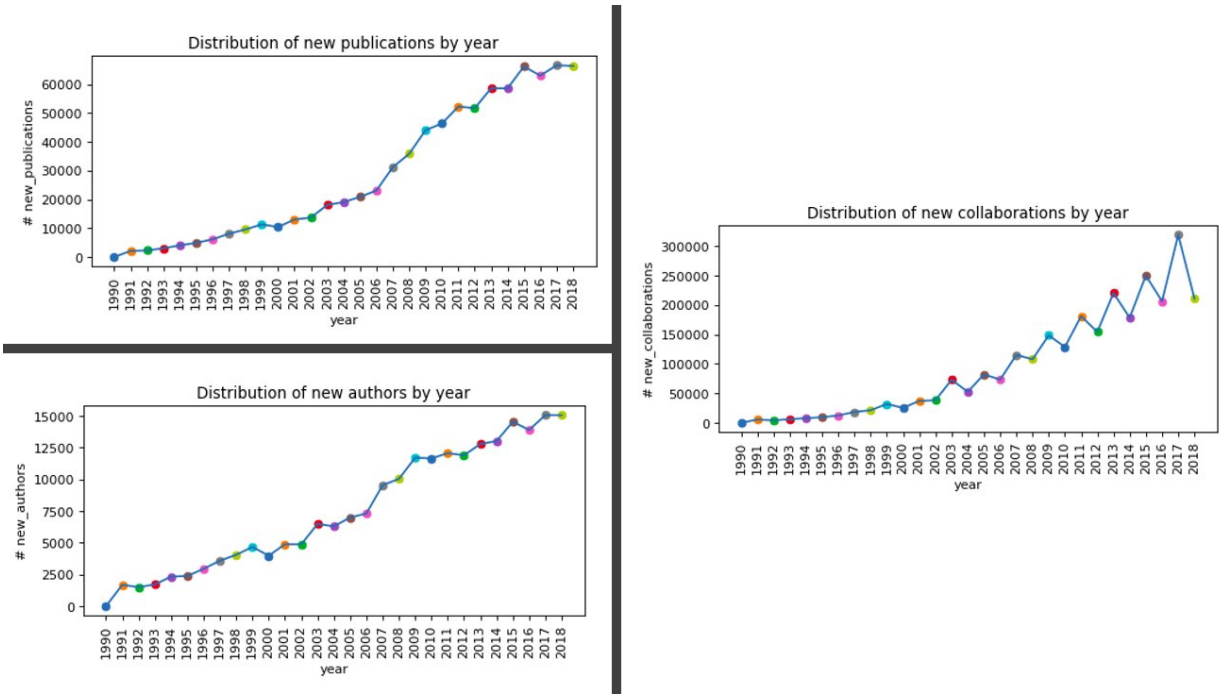


Figure 3.8: *Distribution by year of new collaborations (right), new authors (bottom-left) and new publications (top-left) for all authors in the data.*

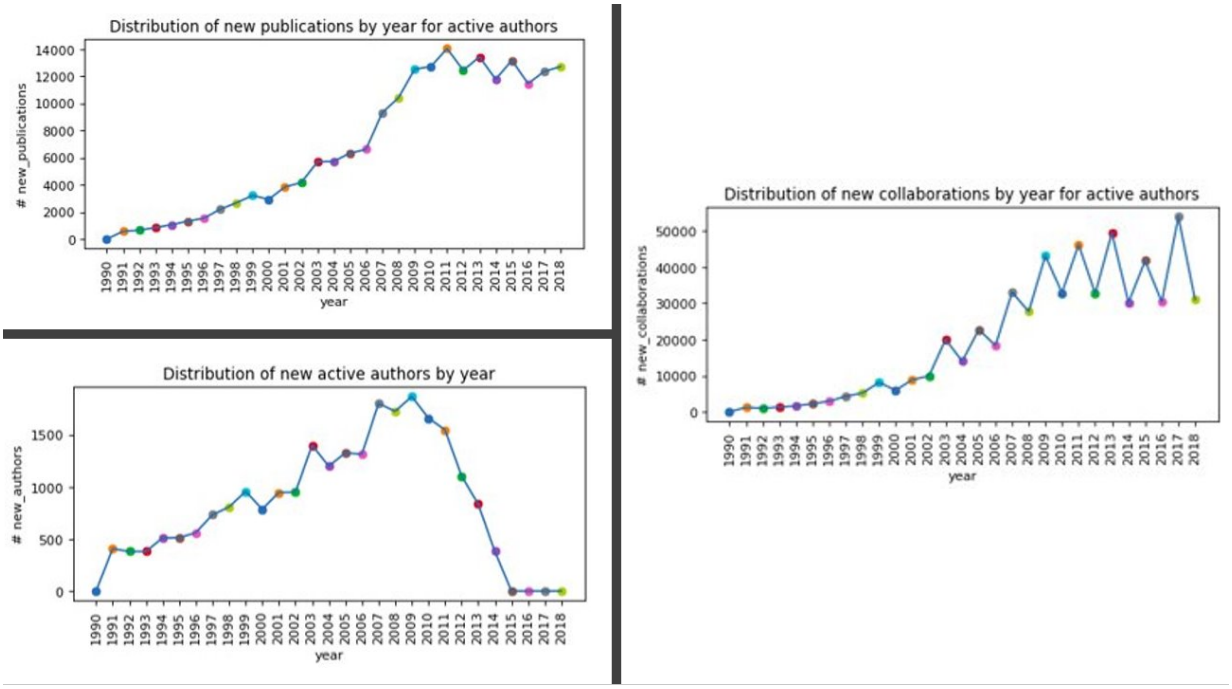


Figure 3.9: *Distribution by year of new collaborations (right), new authors (bottom-left) and new publications (top-left) for the active subset of authors.*

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 2.2).

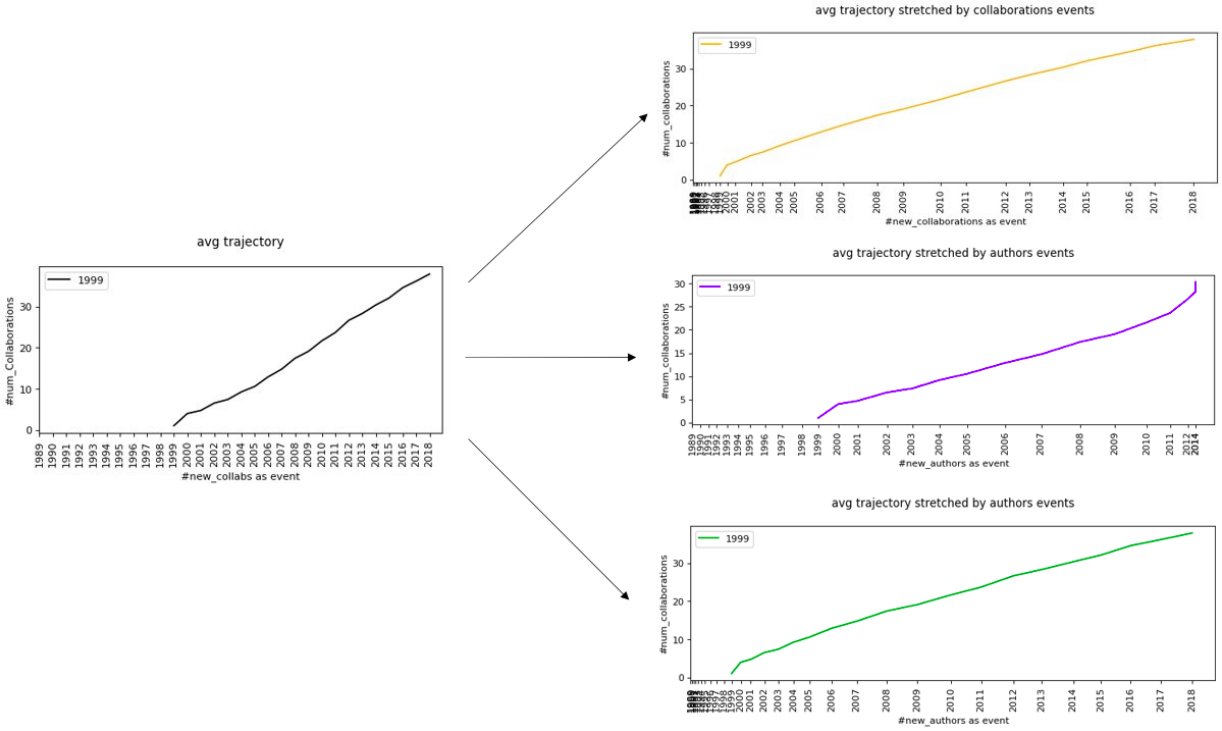


Figure 3.10: *The average vertex trajectory for all authors who started in 1999 (on the left) in stretched using the appearance of new collaborations (top) and of new authors (middle) and of new publications(bottom).*

Chapter 4

Conclusions

T

Bibliography

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

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

Bollobás B. and Riordan O. Handbook of Graphs and Networks: From the Genome to the Internet, 2003. Pages 1–34

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

Newman M. E. J. Coauthorship networks and patterns of scientific collaboration, 2004, doi:10.1073/pnas.03075451

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

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.

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

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

Acknowledgements

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

First, I would like to thanks ....

Nice, 2022

Leonardo Serilli