Evolution over time of the structure of social graphs

Student Leonardo Serilli
Supervisors Nicolas Nisse, Malgorzata Sulkowska, Frédéric Giroire
Year 2021/2022

TABLE OF CONTENTS

01

DATA

COLLABORATION DATA

The following data is about the total number of **collaborations for computer science authors in France since 1990 to 2018**.

	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

PUBLICATION DATA

The following data is about the number of publications in each year, **for computer science authors in France since 1990 to 2018**.

	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

02

STATE OF THE ART

Scale-free networks
Vertex trajectories
Theoretical model

SCALE-FREE NETWORKS

Let:

- **G = (V, E)** be a graph of **n** nodes.
- **n_k** be the #nodes of **degree k** in G,
- **λ** > 0 an exponential parameter,
- **C** > 0 a scaling constant.

The **degree distribution P_k** of G follows a **power-law** if:

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

and

$$P_k \sim C k^{-\lambda}$$

A network whose probability distribution of degrees of nodes respects the power law distribution is said to be **scale-free**.

SCALE-FREE NETWORKS

Scale-free networks are not so widespread as thought, it turns out that many of them follow a **power-law distribution with an exponential cut-off**:

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

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

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

VERTEX TRAJECTORIES

Given a graph **G_t = (V_t , E_t)**, where **t** is a given time step are defined:

- **Node Event**: a new node appears in the graph,
- **Edge event**: a new edge appears.

New nodes and edges must select already present nodes to attach. The probability for any node **v** to be chosen is defined from an **attachment function f(x):**

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

Where **deg_t(v)** is the degree of vertex **v** at timestep **t** .

VERTEX TRAJECTORIES

Given **n** time steps, defined by the occurrence of an event, the evolution of a graph **G_0** is the sequence:

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

Let:

- **d_v(t)** degree of vertex **v** at timestep **t,**
- **t_0** timestep in which **v** appears.

=> The **vertex trajectory of v** is the evolution over time of it's degree, so the sequence:

$$\{d_v(t_0), ..., d_v(t), ..., d_v(t_n)\}$$

THEORETICAL MODEL

Given:

- **α > 0**,
- **0 ≤ σ < 1**,
- **t_v** be the time step in which **v** join the network ,

f(x) = x^σ is the attachment function; this time is **sublinear**, which gives **power-law with exponential cutoff** (so far we used linear and constant functions, but they give normal power-law). This produces logarithmic vertex trajectory. We will check whether they fit our real data.

The **theoretical model,** referred from now on, is the one illustrated next, 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:**

$$P_k \sim \beta \cdot k^{-\sigma} \cdot \exp\left\{-\frac{1}{\alpha} k^{1-\sigma}\right\}$$

and the **vertex trajectory** has logarithmic shape:

$$g_v(t) = (\alpha * \ln(t/t_v) + 1)^{\frac{1}{1-\sigma}}$$

03

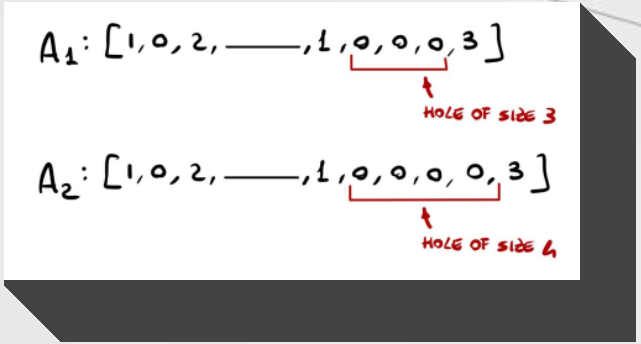
FACED PROBLEMS

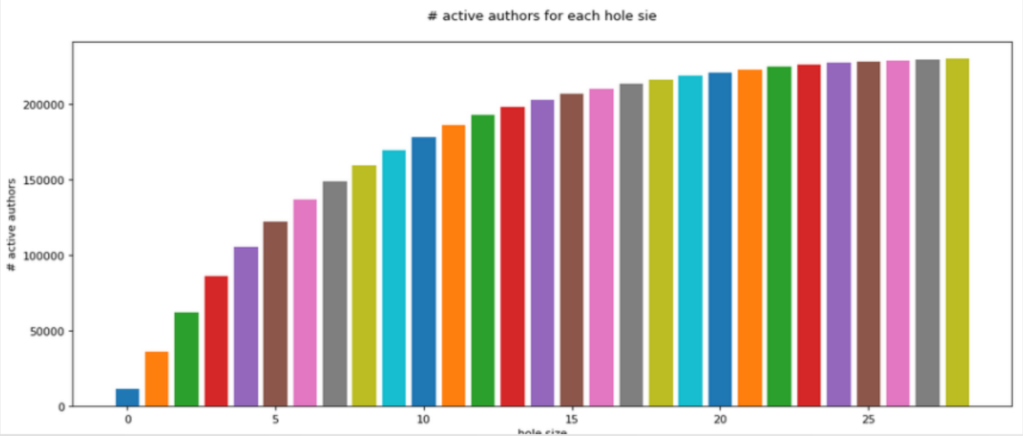
Which researchers should be taken into account?
How to interpret a time step?

Which researchers should be taken into account?

An **author** has a **hole of size n ∈ N** , in his publication history, if he stopped to publish for n consecutive years.

Follow that the **maximum hole size** is the maximum number of years he has passed without publishing.





An author is **inactive** for a given hole size if he has a hole, in his publication data, **greater than the given hole sizes.**

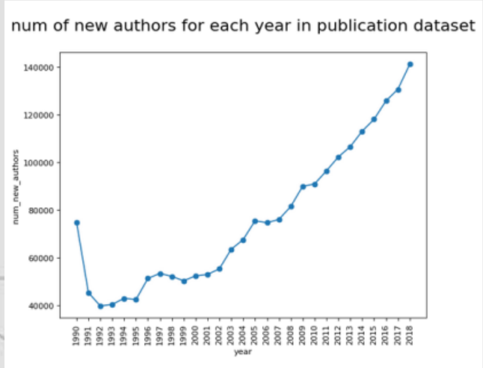
For each possible hole size, has been built a dataset where all authors considered inactive have been filtered out.

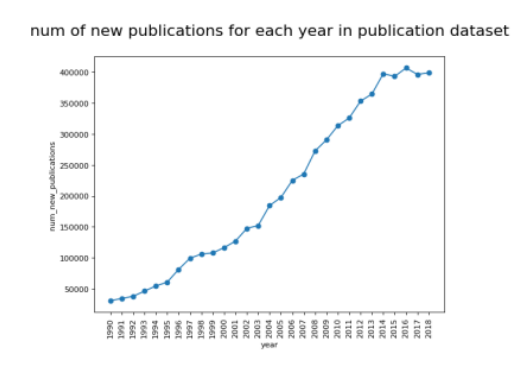
How to interpret a time step?

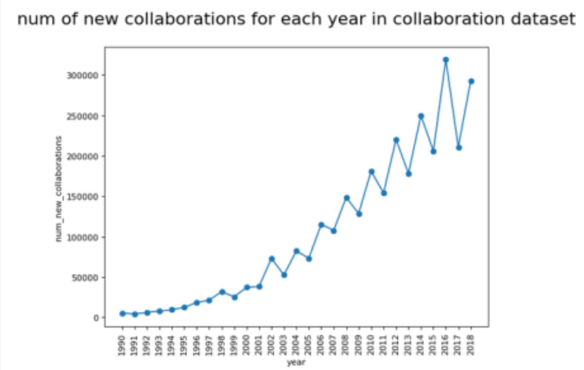
The **theoretical model**, as a time step, consider events, so the set of year, containing only **28 time step,** can be **too small in order to build meaningful vertex trajectories.**

Other metrics are so considered as event : the occurrence of a **new publication**, of a **new author** or a **new collaboration**.

This results in a stretching of the vertex trajectories, which shows the logarithmic shape described in the theoretical model







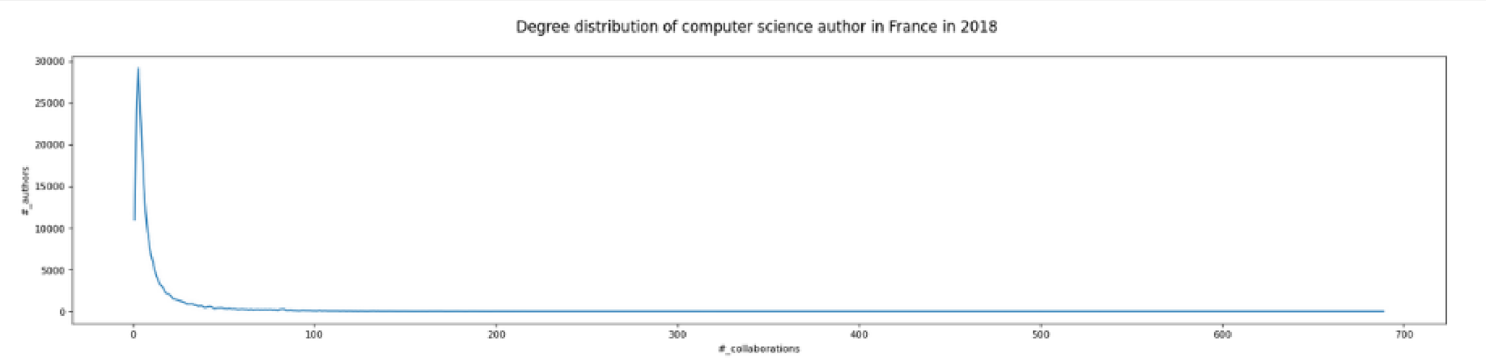
04

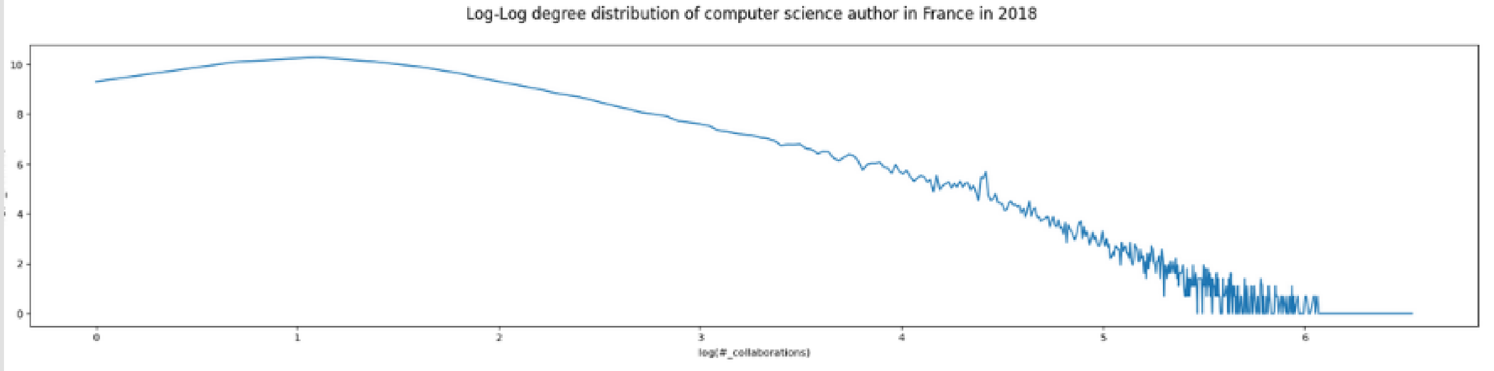
DEGREE DISTRIBUTION

Degree Distribution Retrieval
Degree Distribution Fitting

DEGREE DISTRIBUTION RETRIEVAL

To acquire more knowledge from the character of the data the **degree distribution** is found, taking the **#authors** with a given **#collaborations.**

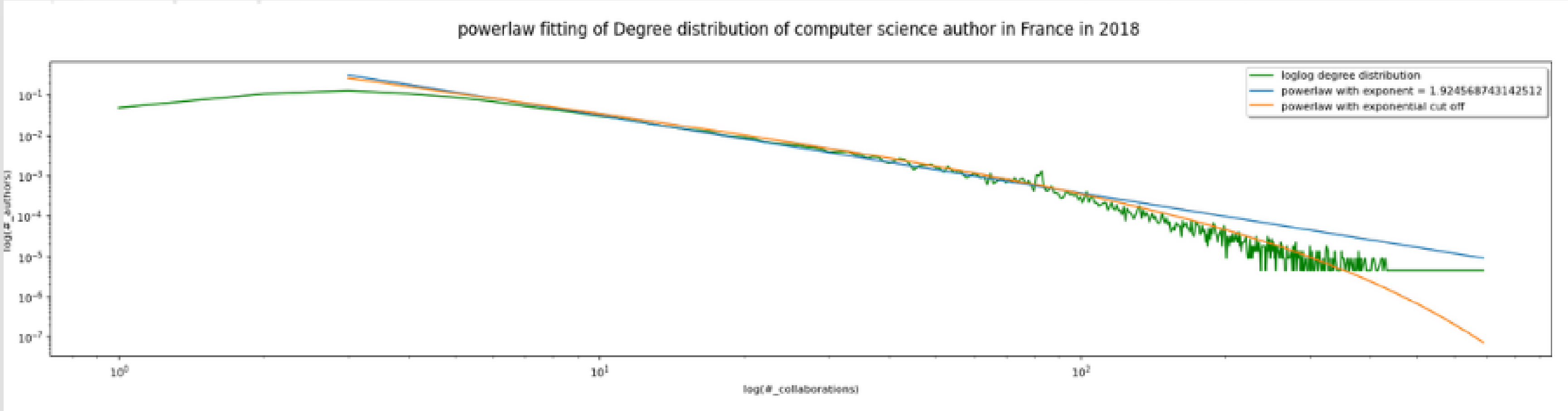




DEGREE DISTRIBUTION FITTING

Fitting the given distribution with the power-law, both the classic and the one with exponential cut-off.

Results that a **power-law with exponential cut-off is better for it's fitting**.



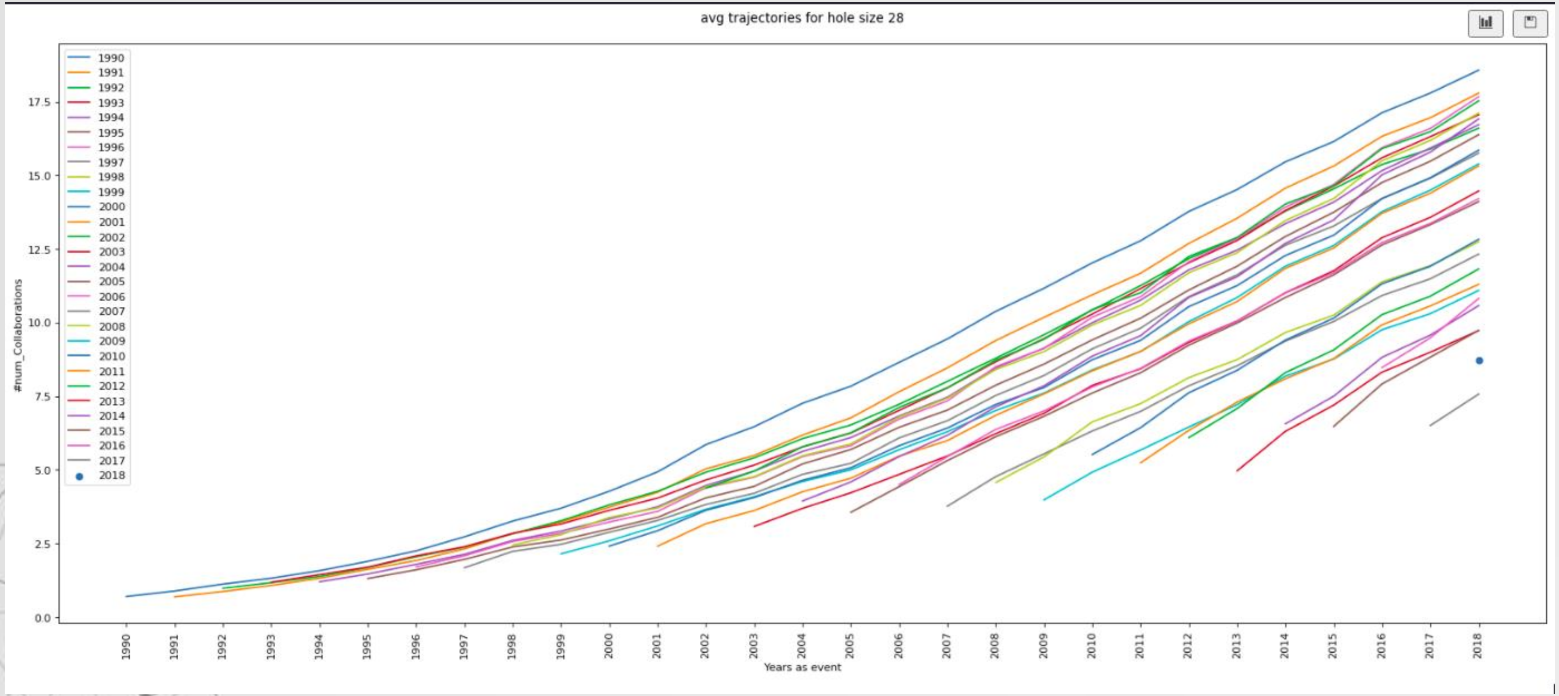
05

VERTEX TRAJECTORIES

Average Trajectories
Stretching
Fitting Trajectories
General Fitting

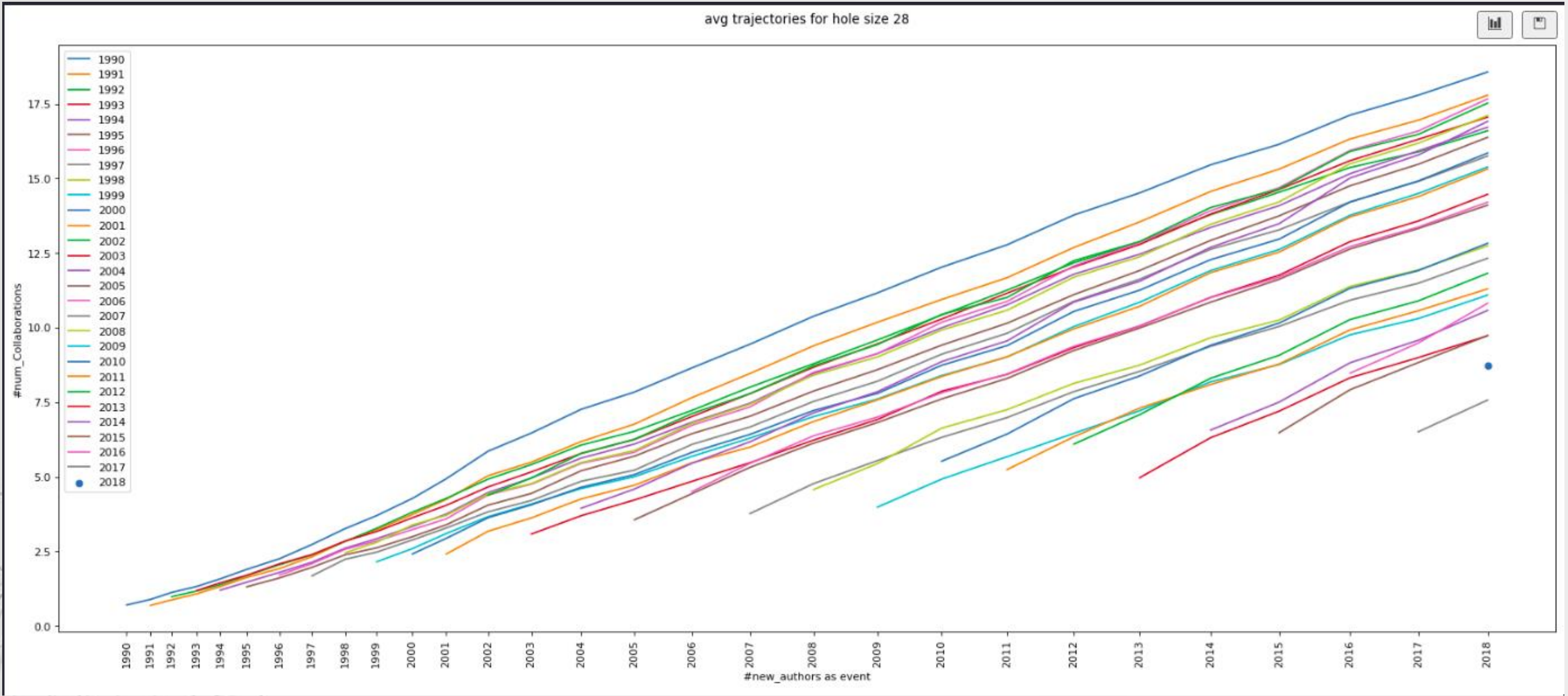
AVERAGE TRAJECTORIES

In order to refer to the theoretical model described in the state of the art, is computed and plotted the **average trajectory of authors** by starting year.



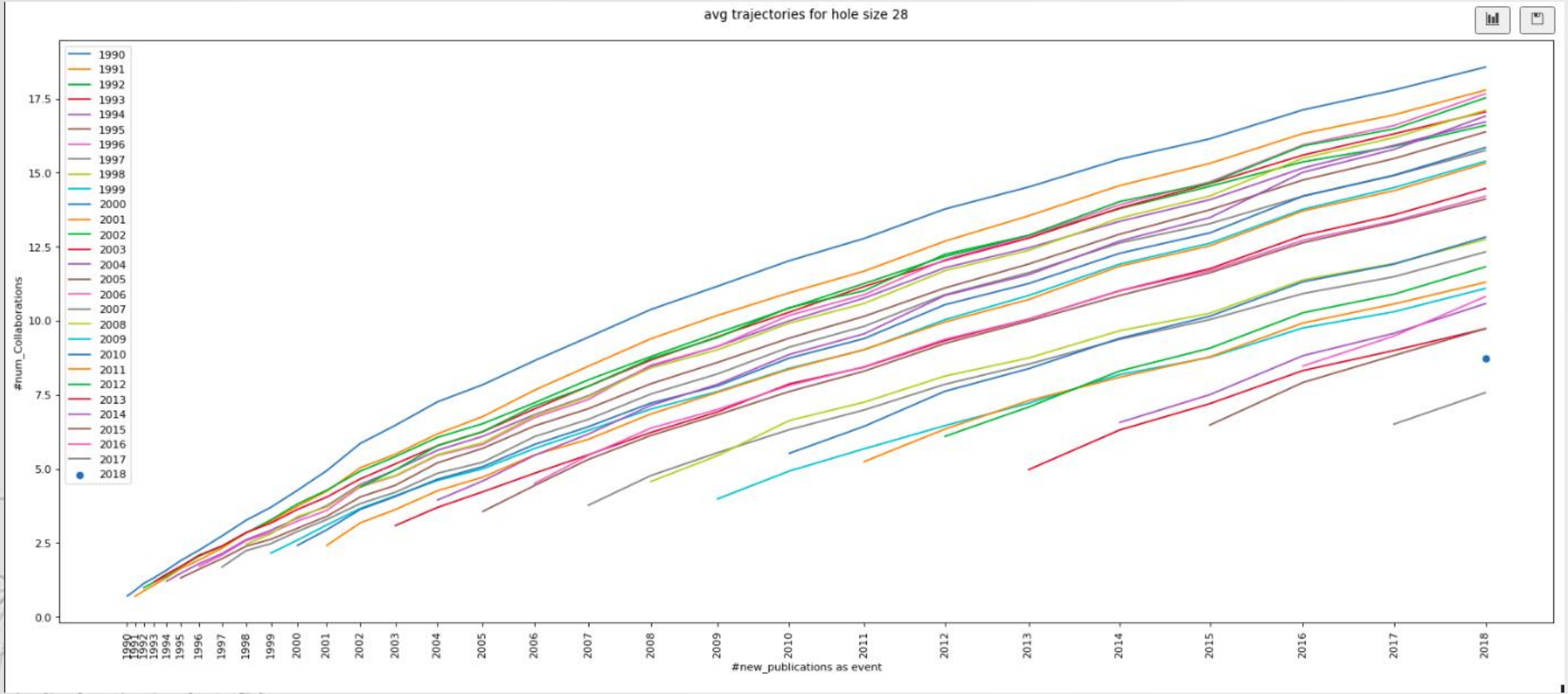
STRETCHING

Trajectories are stretched considering the **number of new authors** as event.



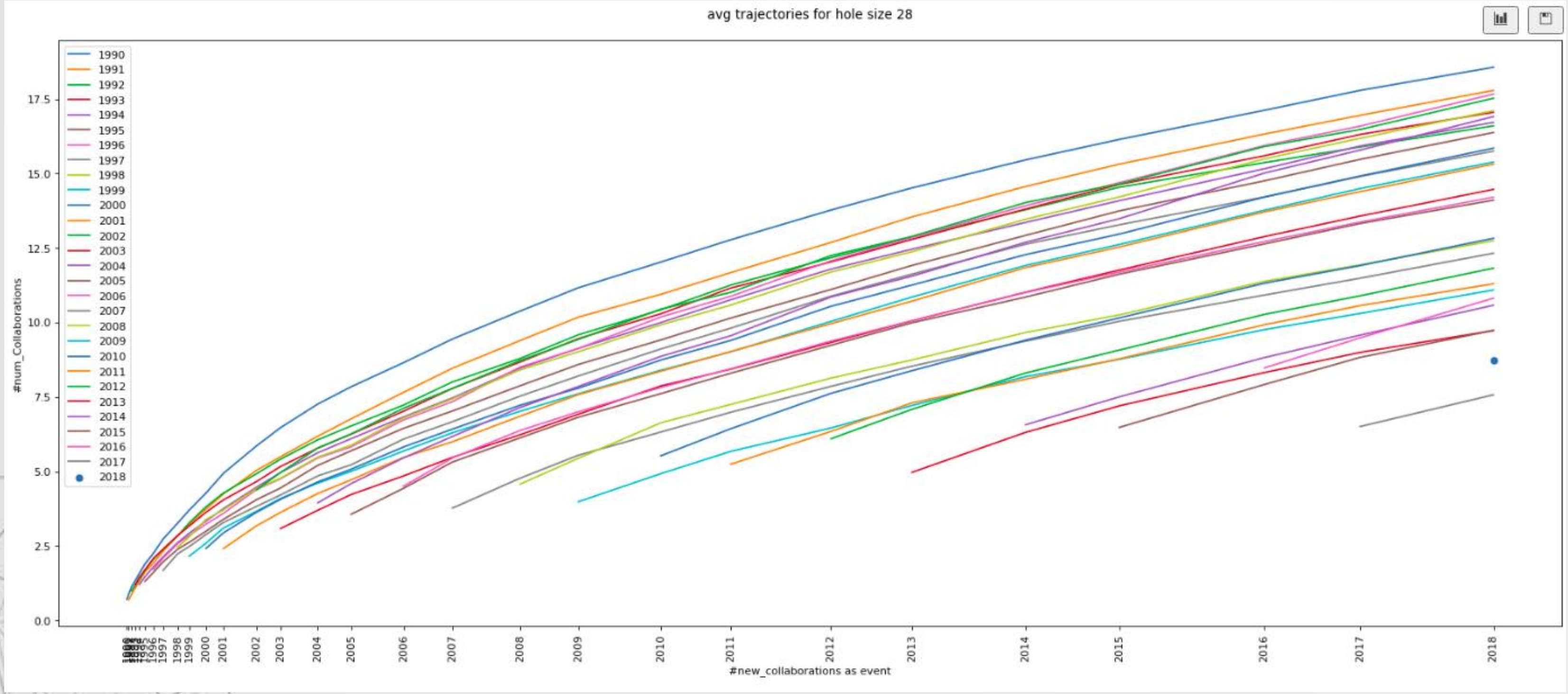
STRETCHING

Trajectories are stretched considering the **number of new publications** as event.



STRETCHING

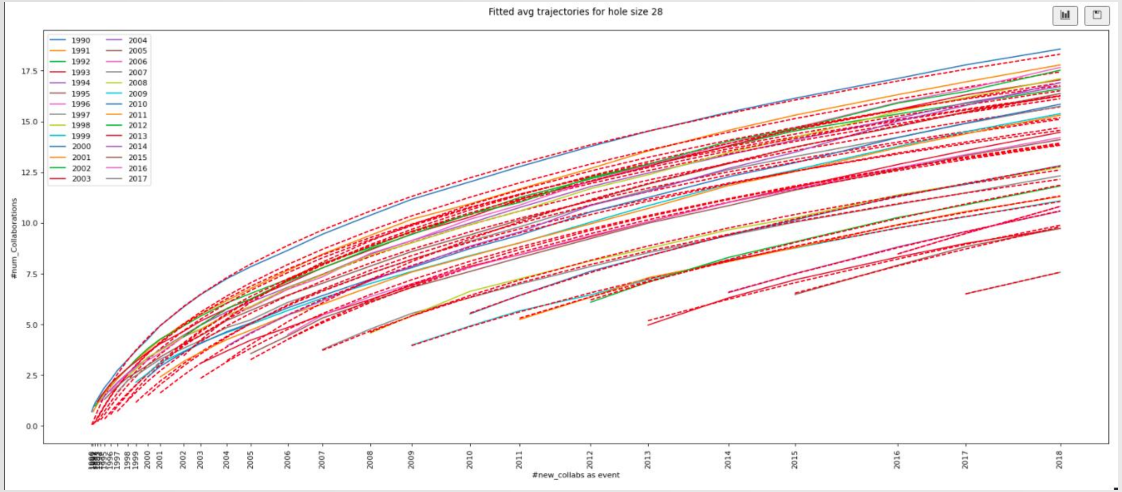
For the fitting presented in the next section, is considered the **number of new collaborations** as event, representing better the logarithmic shape of the theoretical model.



FITTING TRAJECTORIES

Average trajectories are fitted one by one using the **logarithmic function** representing the theoretical vertex trajectory.

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



The fitting works better for trajectories with a low starting year, they contain enough data to show the logarithmic behavior we are trying to fit.

GENERAL FITTING

Then has been tried to find the best couple of parameters α and σ able to fit all curves minimizing the total error on the fitting, for four different kind of error, and obtaining so four couples of **optimal parameters.**

Given a **starting event i** and a **generic event e:**

- **r_i(e)** is the **real average trajectory** for authors who started to publish at the event **i;**

- **f_i(e)** is the **fitting function** associated with **r_i(e);**

- **f_i^*(e)** is the **general fitting function** of which we want to optimize the parameters, for authors who started to publish at the event **i.**

A) $min_{\alpha^*,\sigma^*}\left(\sum_i \sum_{e\geq i} |f_i^*(e) - r_i(e)|^2\right)$

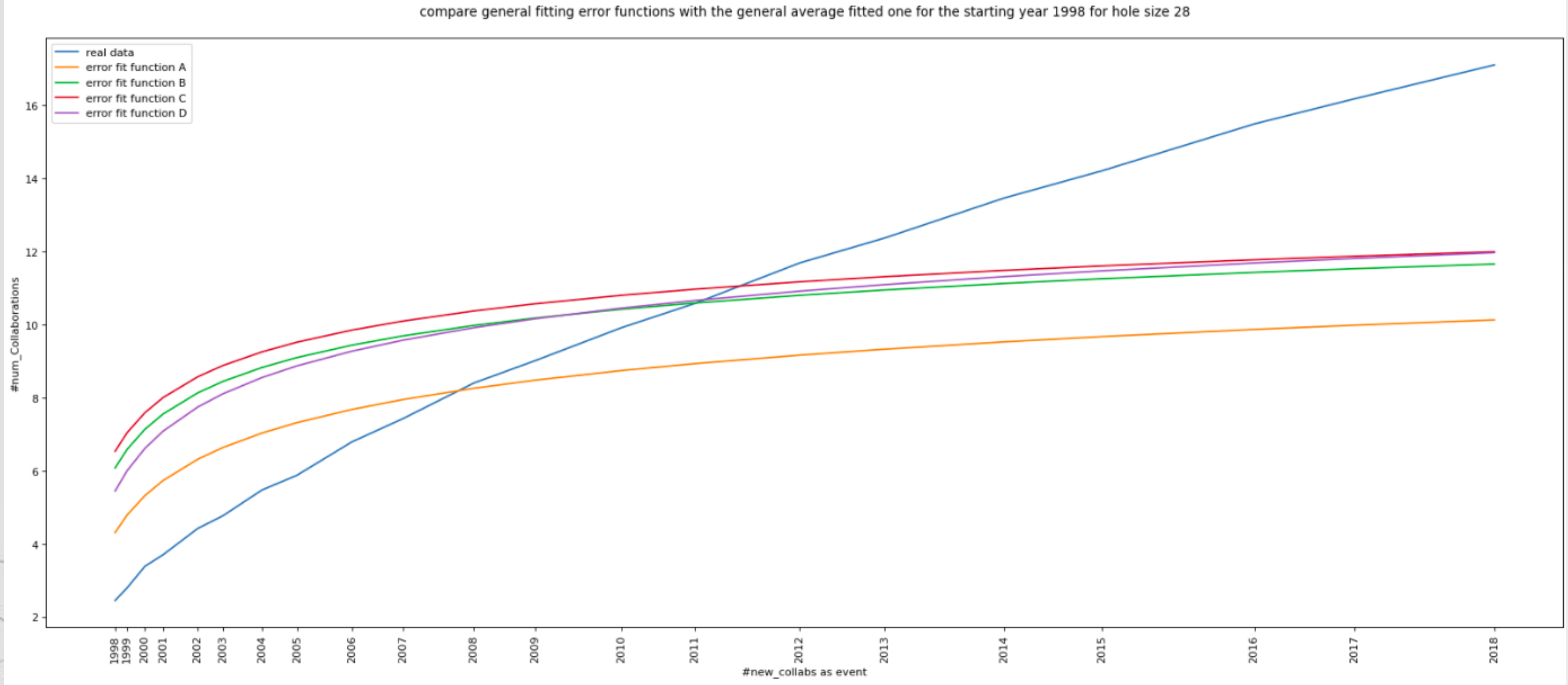
B) $min_{\alpha^*,\sigma^*}\left(\sum_i \max_{e\geq i} |f_i^*(e) - r_i(e)|^2\right)$

C) $min_{\alpha^*,\sigma^*}\left(\sum_i \sum_{e\geq i} |f_i^*(e) - f_i(e)|^2\right)$

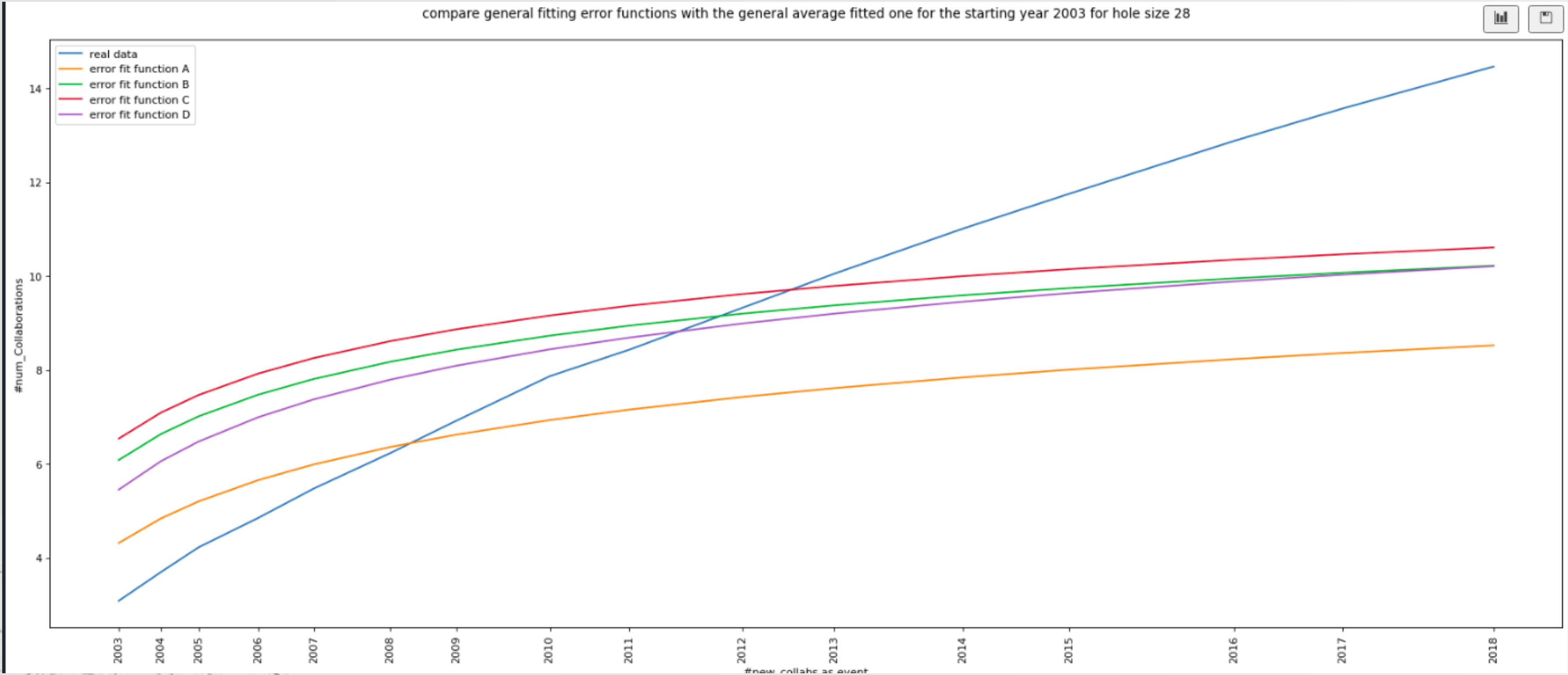
D) $min_{\alpha^*,\sigma^*}\left(\sum_i \max_{e\geq i} |f_i^*(e) - f_i(e)|^2\right)$

	ERROR	alpha	sigma
err_f_A	4529.074755	4.314193	0.590249
err_f_B	620.851259	6.080949	0.450102
err_f_C	32466.473974	6.537745	0.419692
err_f_D	747.430665	5.449991	0.544075

GENERAL FITTING



GENERAL FITTING



06

NEXT STEPS

NEXT STEPS

The underlying theoretical model is probably more complex than we assumed at the beginning

Next step: the analysis of data from other research field.

Next step: two subset of authors ,granted and non-granted, with similar trajectories up to the year of the grant; comparing trajectories of granted who started at the same time with their controls


The End