

00 - build_collaboration_df.ipynb

Retrieved the dataset containing the **cumulative number of collaboration** for computer science authors in French between **1990 and 2018**.

| | ID | 1990 | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | ... | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 |
|--------|-------------|------|------|------|------|------|------|------|------|------|-----|------|------|------|------|------|------|------|------|------|------|
| 0 | 8958327900 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 6508297663 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 4 | 7 | 7 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| 2 | 7004267341 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 10 | 10 | 10 | 16 | 16 | 16 | 16 | 16 | 16 | 16 |
| 3 | 8642393600 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 7 | 7 | 7 |
| 4 | 55873955900 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 8 | 8 | 8 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 232833 | 6507630481 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 18 | 18 | 18 | 18 | 18 | 29 | 29 | 29 | 29 | 29 |
| 232834 | 24577815500 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 4 | 4 | 4 | 4 | 6 | 13 | 16 | 16 | 16 | 70 |
| 232835 | 57195243976 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 8 | 8 |
| 232836 | 35328962100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 2 | 2 | 3 |
| 232837 | 7403521415 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 29 | 29 | 29 |

232838 rows × 30 columns

The relative csv is located at `myDATA/00-collaboration_df.csv`.

02 - build_publication_df.ipynb

Retrieved the dataset containing the **number of publications** for computer science authors in French between **1990 and 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 |

232838 rows × 30 columns

The relative csv is located at `02-publication_df.csv`

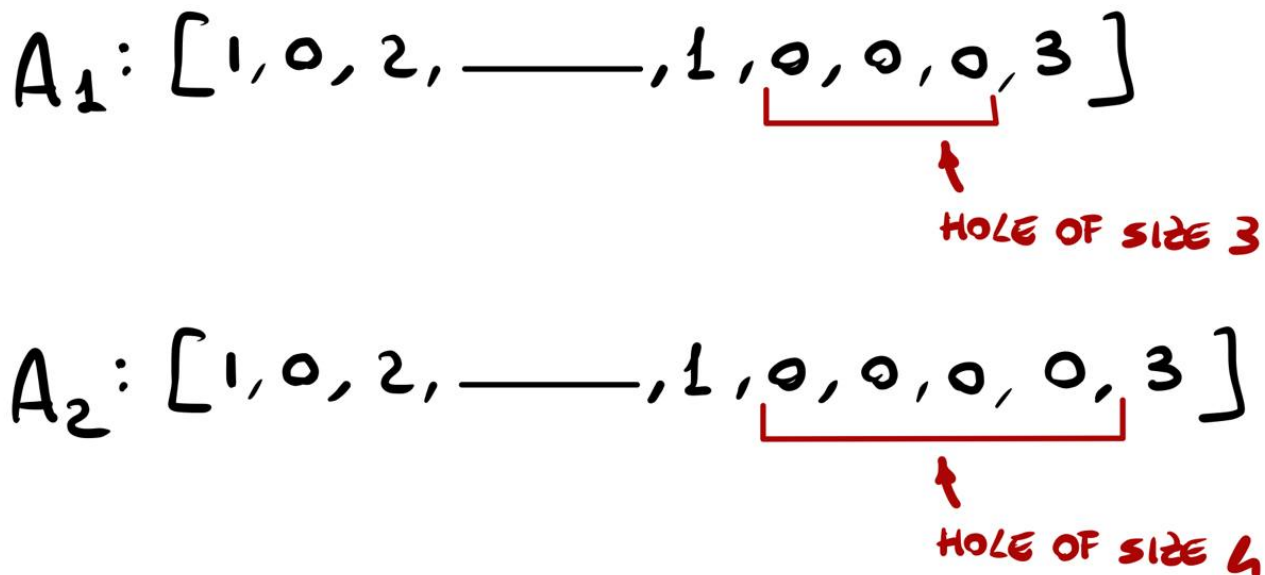
05 - filtering_active_authors.ipynb

In the associated notebook are filtered out all inactive authors, so a dataset is built for each possible definition of **Inactivity**.

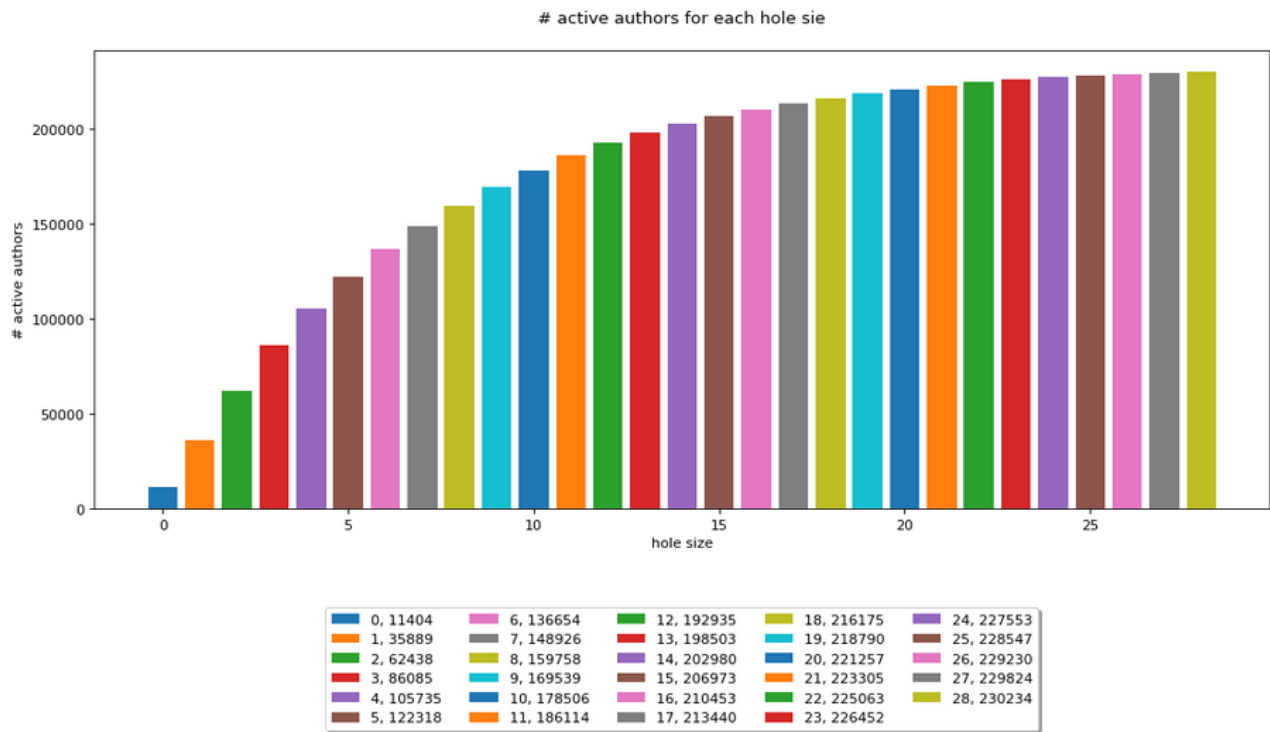
An author is inactive if it has a hole in publications **greater** than a given value.

An author has a hole in publication if he haven't published for **n consecutive years**.

For example, given an **hole size = 3**, the author A1 is active but A2 is not.

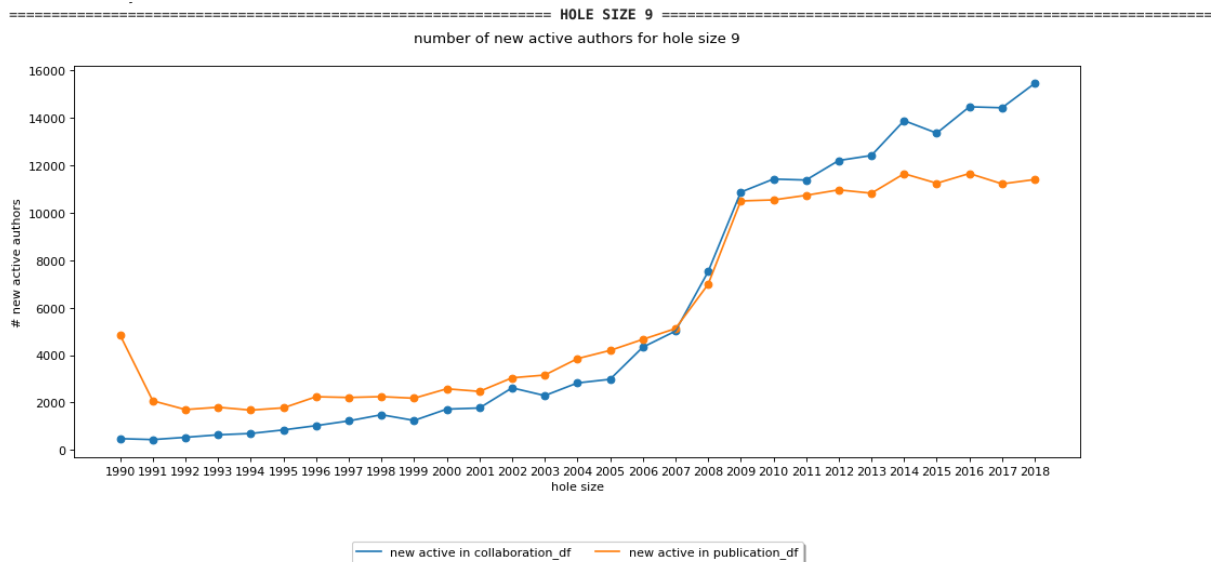


The number of authors kept for each hole size is showed in the following chart:



The **output of the code** are **28 csv** located in the directory `/myata/05-filtered_by_hole_size/`

The distribution of new collaborators and new publicators for each year and hole size is located at `myDATA/05-filtered_by_hole_size`, it's shape is like:



```
collaboration_df {'1990': '475', '1991': '435', '1992': '531', '1993': '636', '1994': '692', '1995': '843', '1996': '1023', '1997': '1223', '1998': '1483', '1999': '1242', '2000': '1720', '2001': '1767', '2002': '2617', '2003': '2290', '2004': '2824', '2005': '2978', '2006': '4331', '2007': '5010', '2008': '7514', '2009': '10871', '2010': '11422', '2011': '11382', '2012': '12206', '2013': '12417', '2014': '13888', '2015': '13360', '2016': '14470', '2017': '14430', '2018': '15459'}
```

```
publication_df {'1990': '4844', '1991': '2072', '1992': '1702', '1993': '1799', '1994': '1678', '1995': '1774', '1996': '2242', '1997': '2211', '1998': '2246', '1999': '2179', '2000': '2577', '2001': '2467', '2002': '3040', '2003': '3157', '2004': '3849', '2005': '4195', '2006': '4660', '2007': '5116', '2008': '6982', '2009': '10500', '2010': '10543', '2011': '10737', '2012': '10967', '2013': '10826', '2014': '11649', '2015': '11246', '2016': '11657', '2017': '11220', '2018': '11404'}
```

08 - starting_publication_year.ipynb / 09 - ending_publication_year.ipynb

There are built two version of the **collabaration dataset** containing respectively a column with the **starting publication year** and the **ending publication year** for each author.

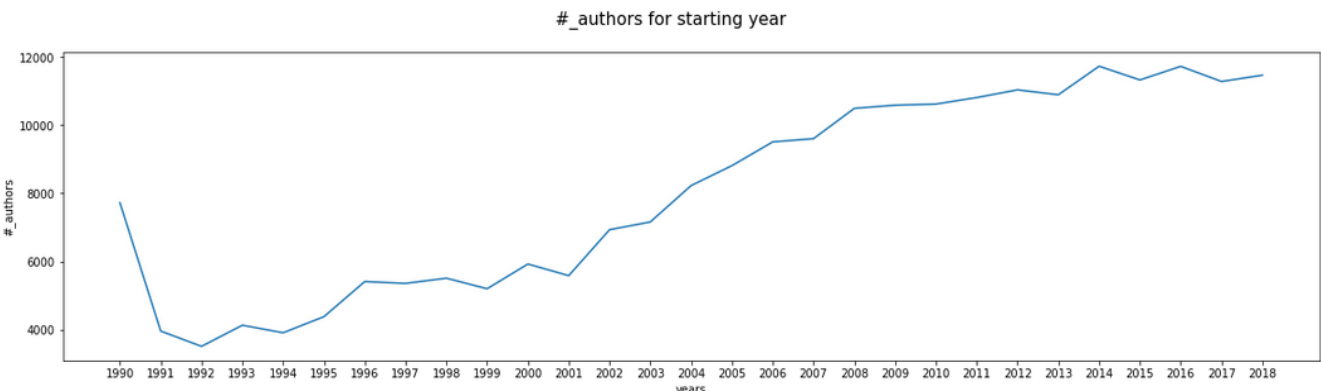
starting year

Located at → `myDATA/00-collaboration_df_with_starting_years.csv`.

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

232838 rows × 31 columns

distribution of the number of authors by starting year



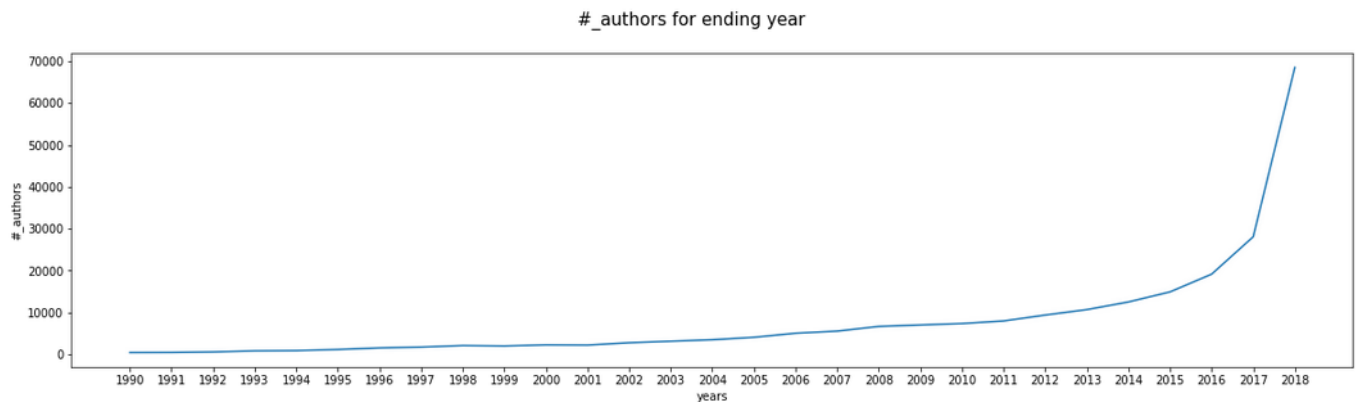
ending year

Located at → `myDATA/00-collaboration_df_with_ending_years.csv`.

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

232838 rows × 31 columns

distribution of the number of authors by ending year

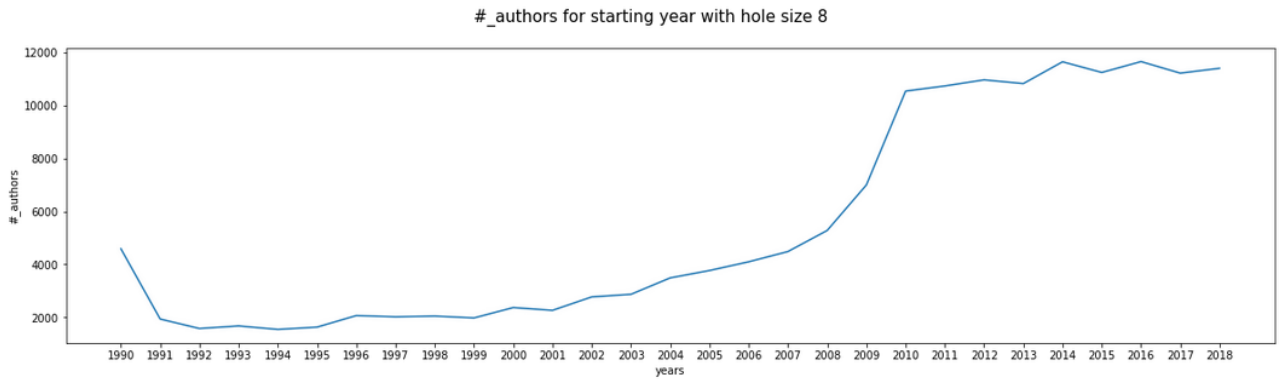


10 - splitting_collaborations.ipynb

Each hole lenght based dataset is **split in 28 subsets based on the starting year** of each author.

- They are located at `myDATA/10-splitting_by_year/`
- Each hole size based dataset has an associate directory at `myDATA/10-splitting_by_year/<HOLE_SIZE>_hole_size_splitting/`
- The set of datasets associated with an hole size has a distribution chart showing the number of authors for each starting year, and it's located at `myDATA/10-splitting_by_year/<HOLE_SIZE>_hole_size_splitting/distribution_chart.png`.

It's shape is the following:

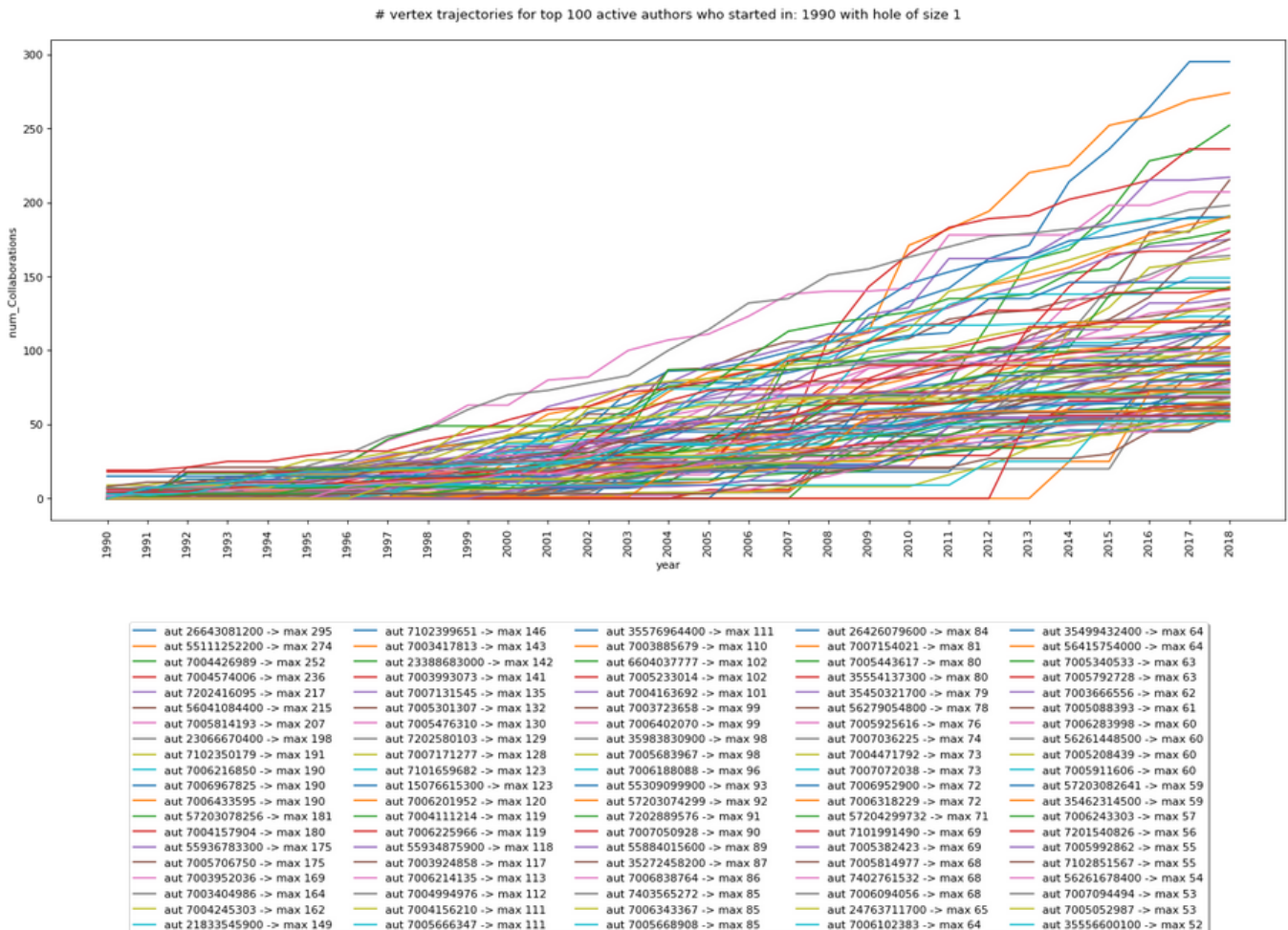


15 - plotting_splitted_data_by_year.ipynb

A chart containing the number of collaborations by year for each dataset built in the above described notebook is plotted here. Each one contains only the top 100 collaborative authors.

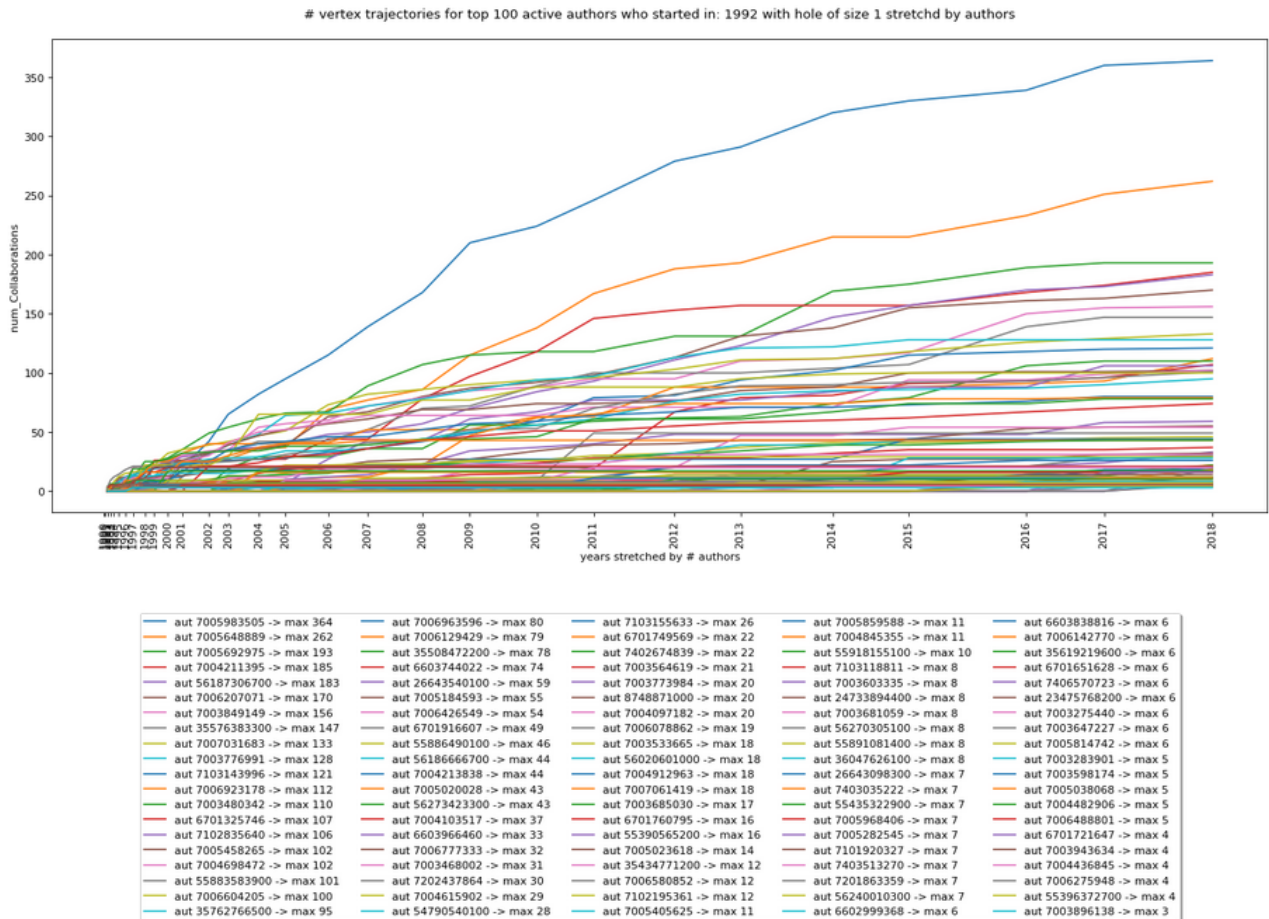
Each of them is located in the directory `myDATA/10-`

`splitted_by_year/<HOLE_SIZE>_hole_size_splitted/trajectories_plt/` and they looks as follow:

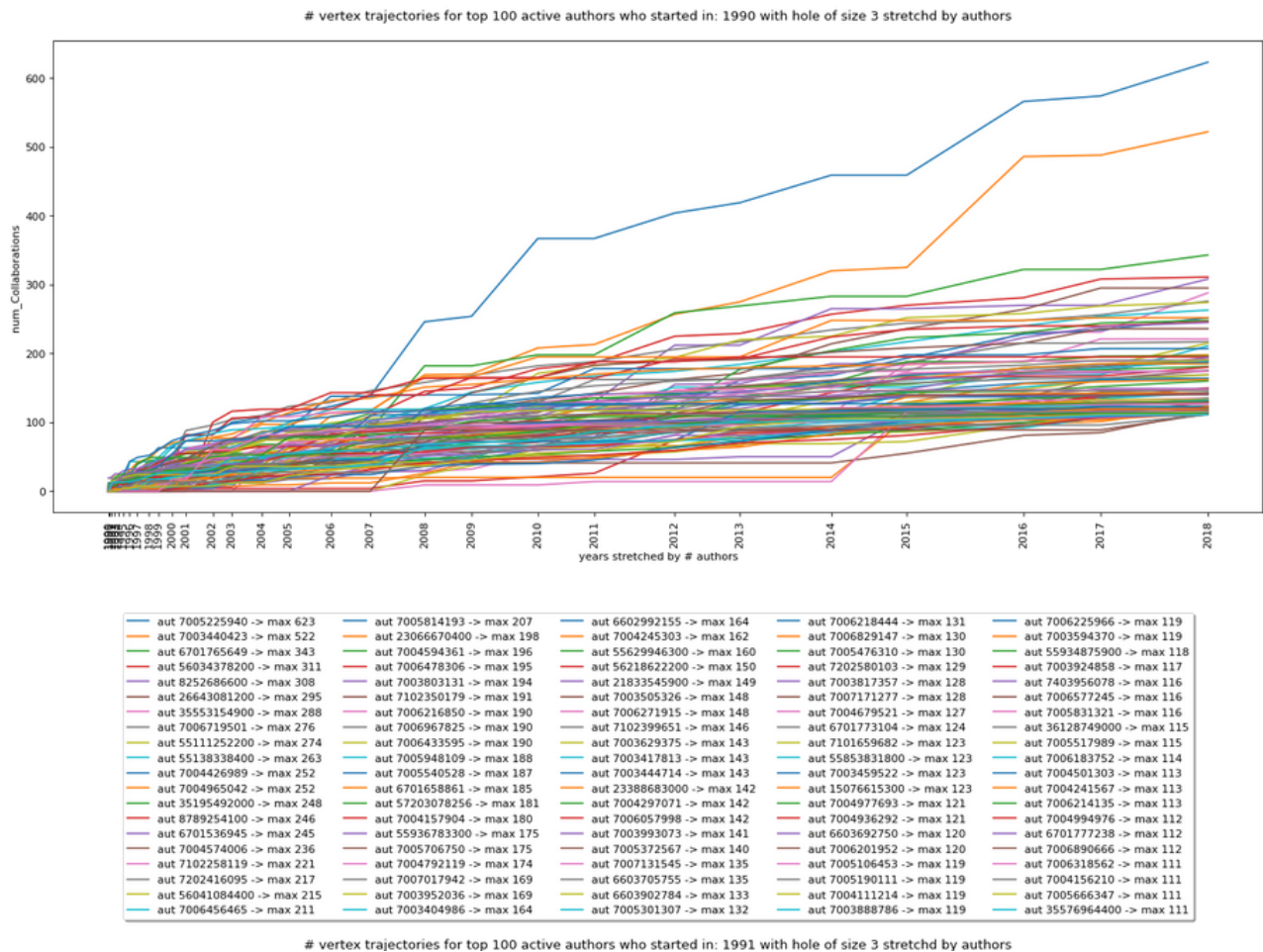


16 - plotting_splitted_data_by_event.ipynb

- Firstly the same chart of the previous notebook is stretched by using an **occurrence of a new author as an event** instead of the years.



- located at `myDATA/10-splittd_by_year//<HOLE_SIZE>_hole_size_splittd/trajectories_plt_by_events`
- Then the same chart of the previous notebook is stretched by using an **occurrence of a new author as an event** instead of the years.



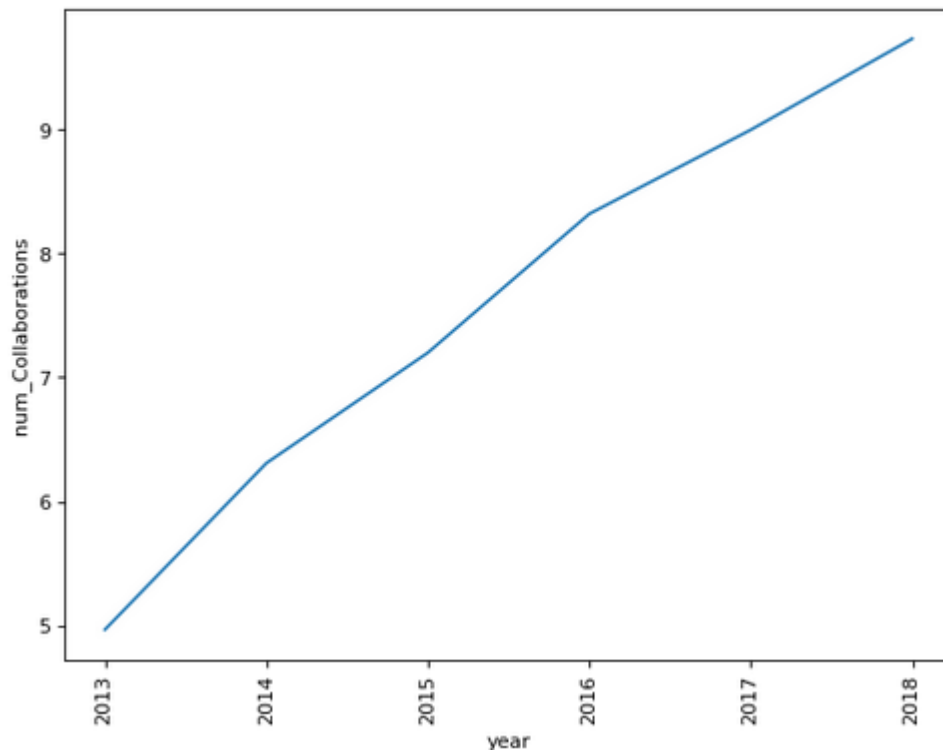
- located at `myDATA/10-`
- `splitted_by_year//<HOLE_SIZE>_hole_size_splitted/trajectories_plt_by_events`

20 - plotting_avg.ipynb

Same as the previous notebbok but, instead of the top 100 collaborative authors, here is plotted the average number of publications for each year.

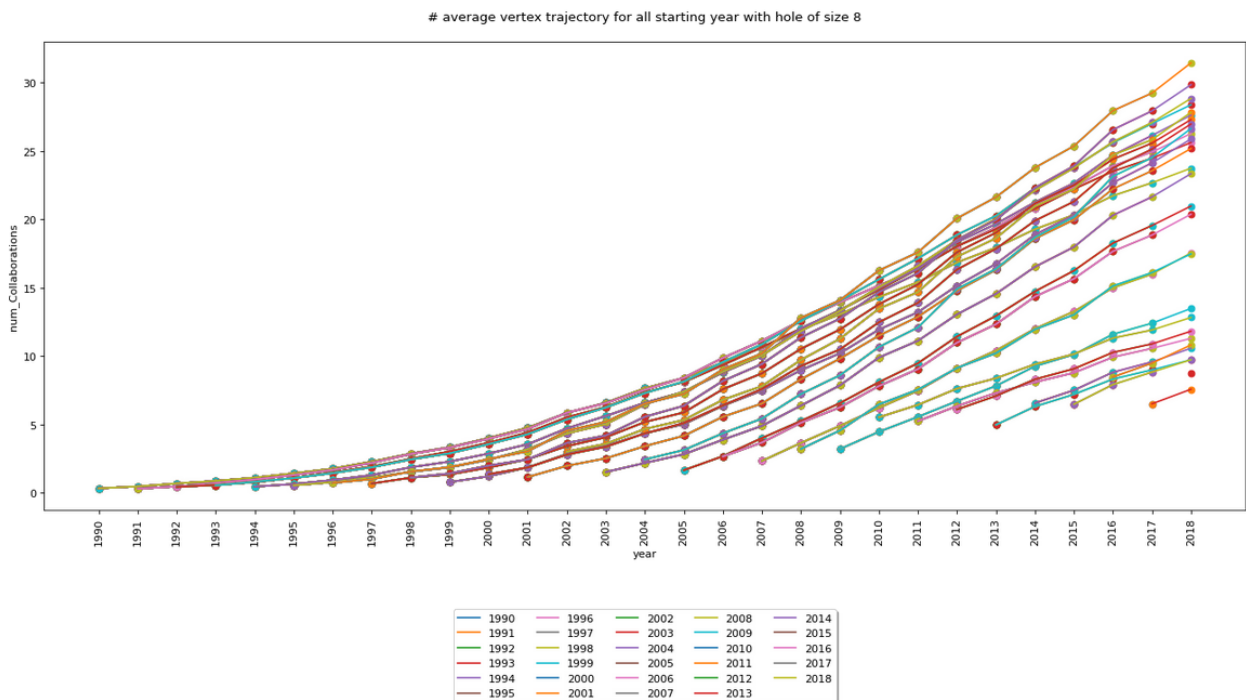
Each of them is located into the directory `trajectories_avg_plt` and look as follow:

average vertex trajectory for all active authors who started in 2013 with hole of size 13



21 - plotting_avg_by_event.ipynb

- Firstly are plotted the same averages of the previous notebook but all years associated with the same hole size, are collected in the same chart:

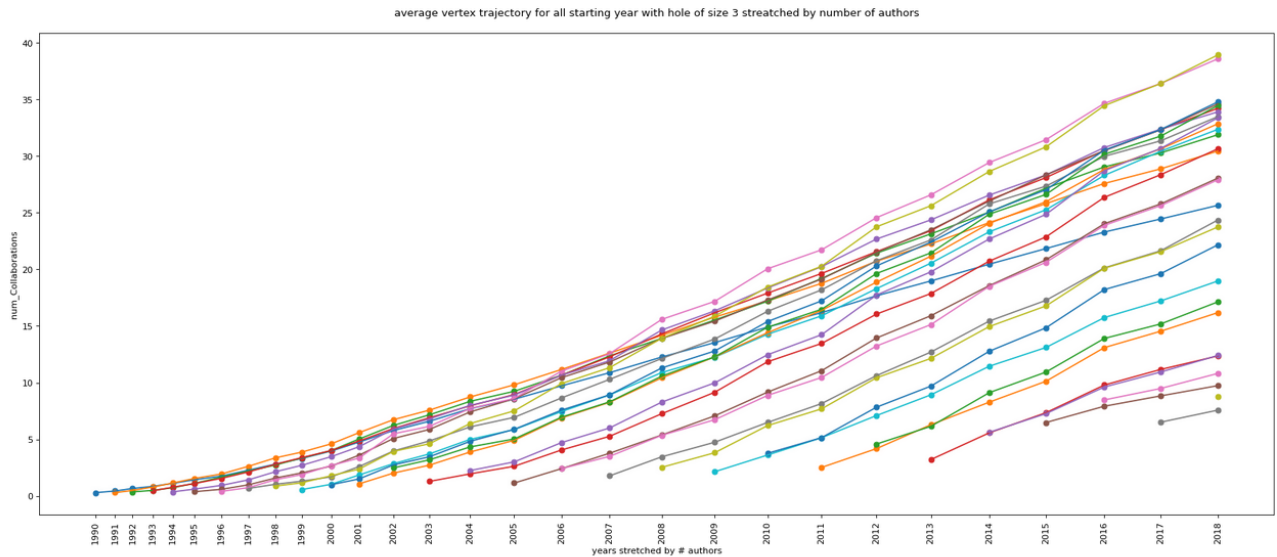


- They are located at `myDATA/10-splitting_by_year/<HOLE_SIZE>_hole_size_splitting/trajectories_avg_plt/aver`

ages_by_year.png

new author as event

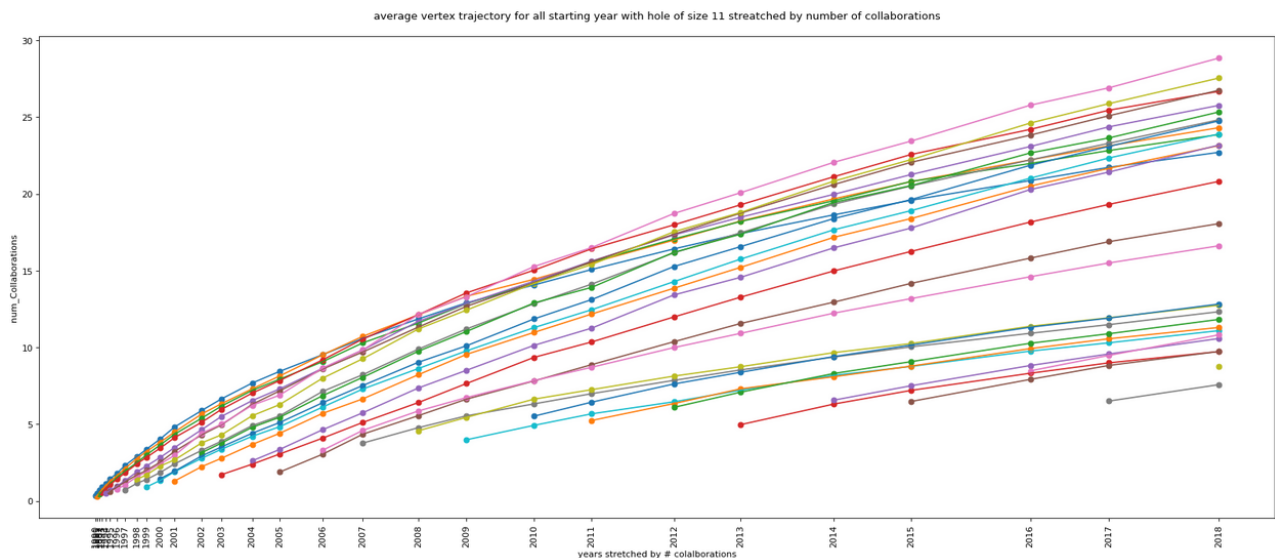
- Then the same chart is stretched by using an **occurrence of a new author as an event** instead of the years.



- They are located at `myDATA/10-splitted_by_year/<HOLE_SIZE>_hole_size_splitted/trajectories_avg_plt/average_by_num_authors.png`

new collaboration as event

- lastly the same chart is stretched by using an **occurrence of a new collaboration as an event** instead of the years.

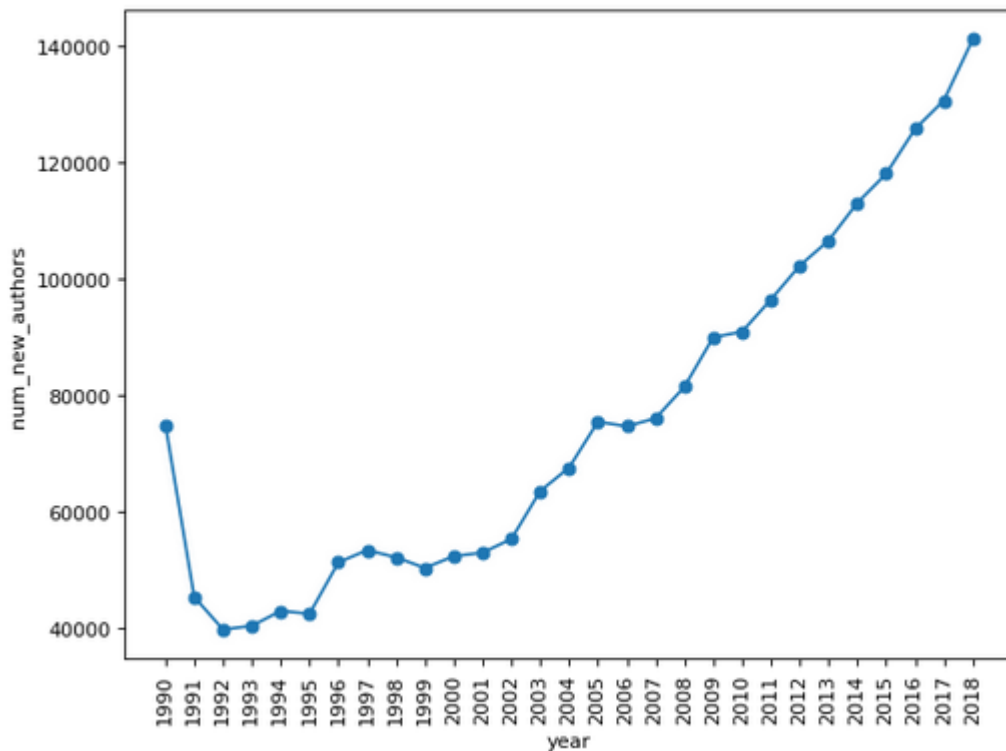


- They are located at `myDATA/10-splitted_by_year/<HOLE_SIZE>_hole_size_splitted/trajectories_avg_plt/averages_by_num_collaborations.png`

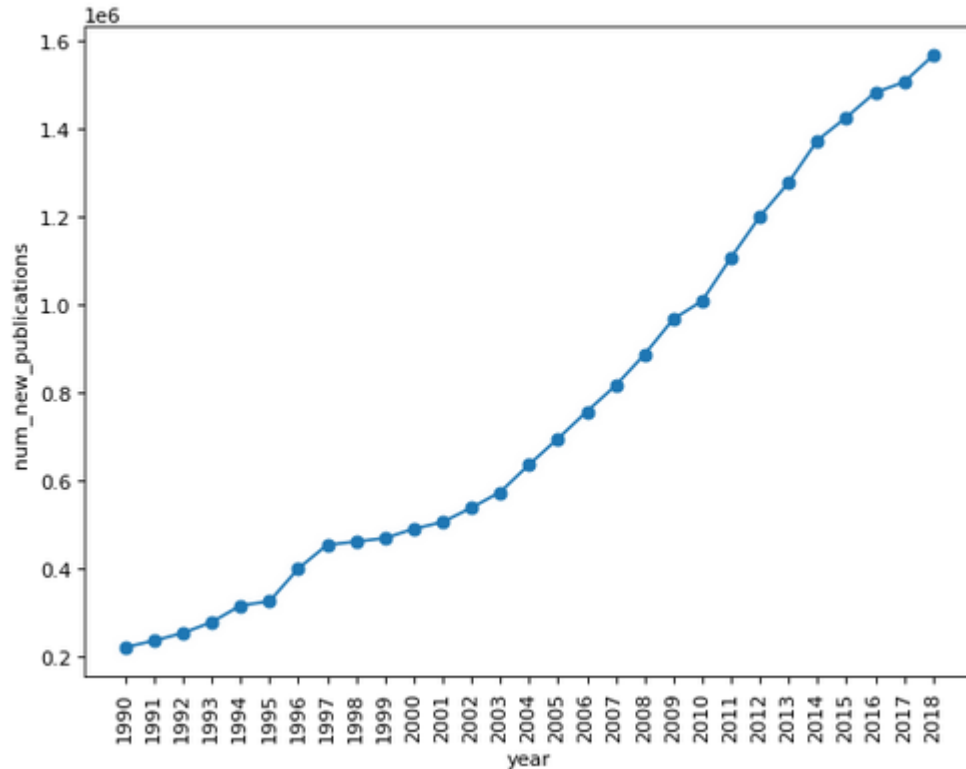
25 - plotting_functions._of_publications_df.ipynb

Here are plotted:

- The **number of new authors** by year in the publication dataset
num of new authors for each year in publication dataset



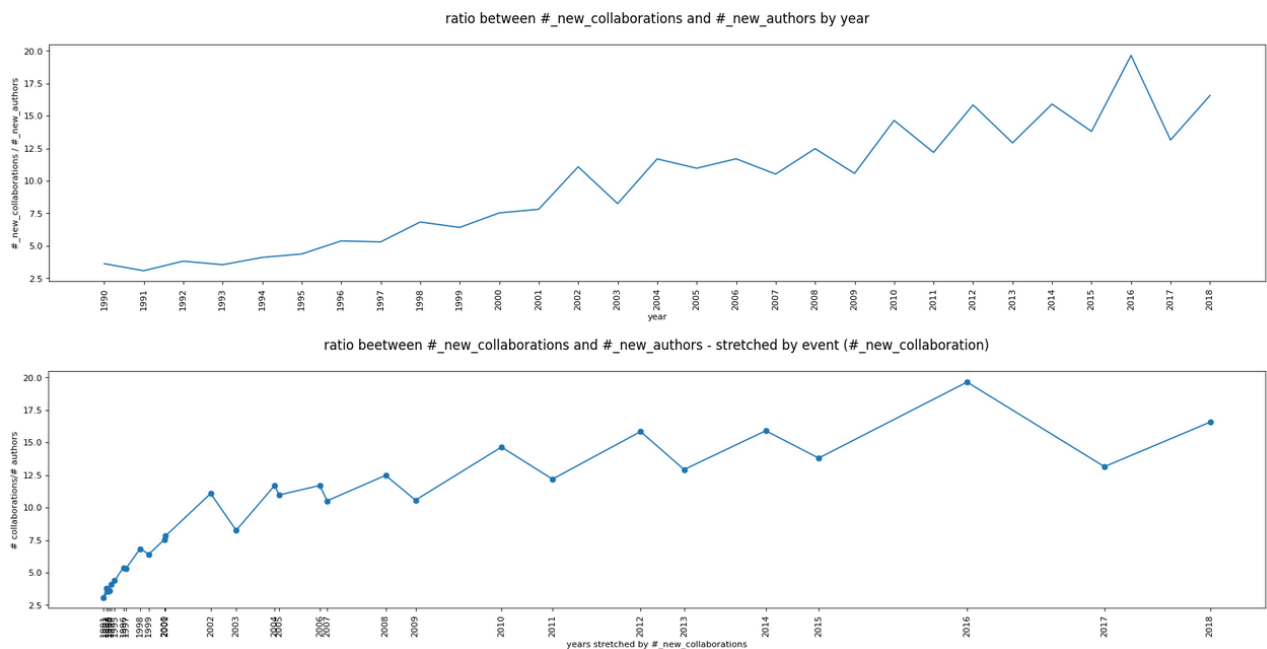
- The **number of new publications** by year in the publication dataset
num of new publications for each year in publication dataset



35 - Ratios.ipynb

Here, both by year and using the occurrence of a new collaboration as event, are plotted:

- The **Ratio** between the number of **new collaboration** and **new authors**.



- The **Ratio** between **total** number of **collaboration** and **authors**.

