

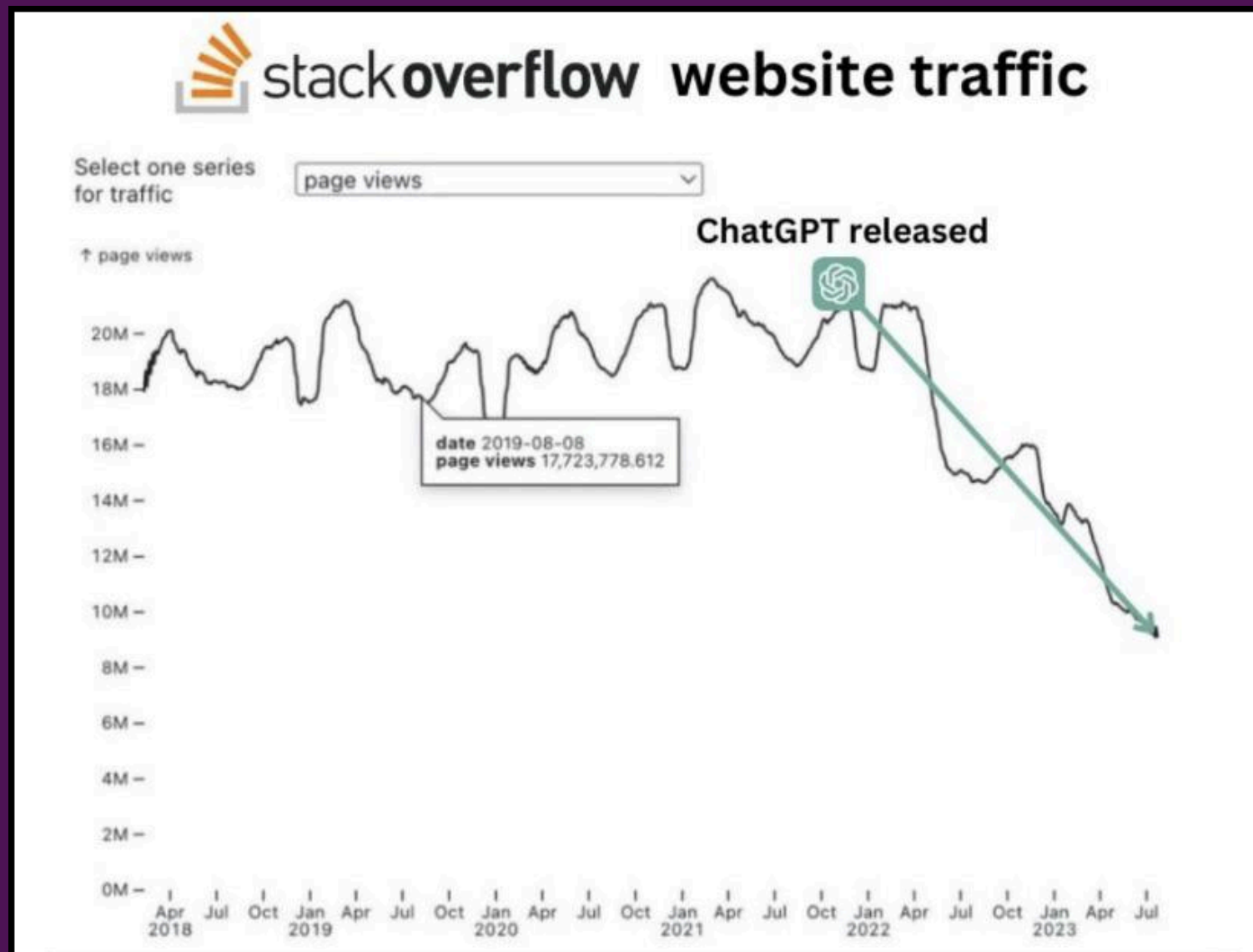
Lorenzo Pannacci – 1948926

LLMs and Stack Overflow

Has the rise of LLMs influenced the behavior of users on Stack Overflow?

Inspiration

Visiting programming and AI-enthusiast communities is easy to find this viral image reposted:

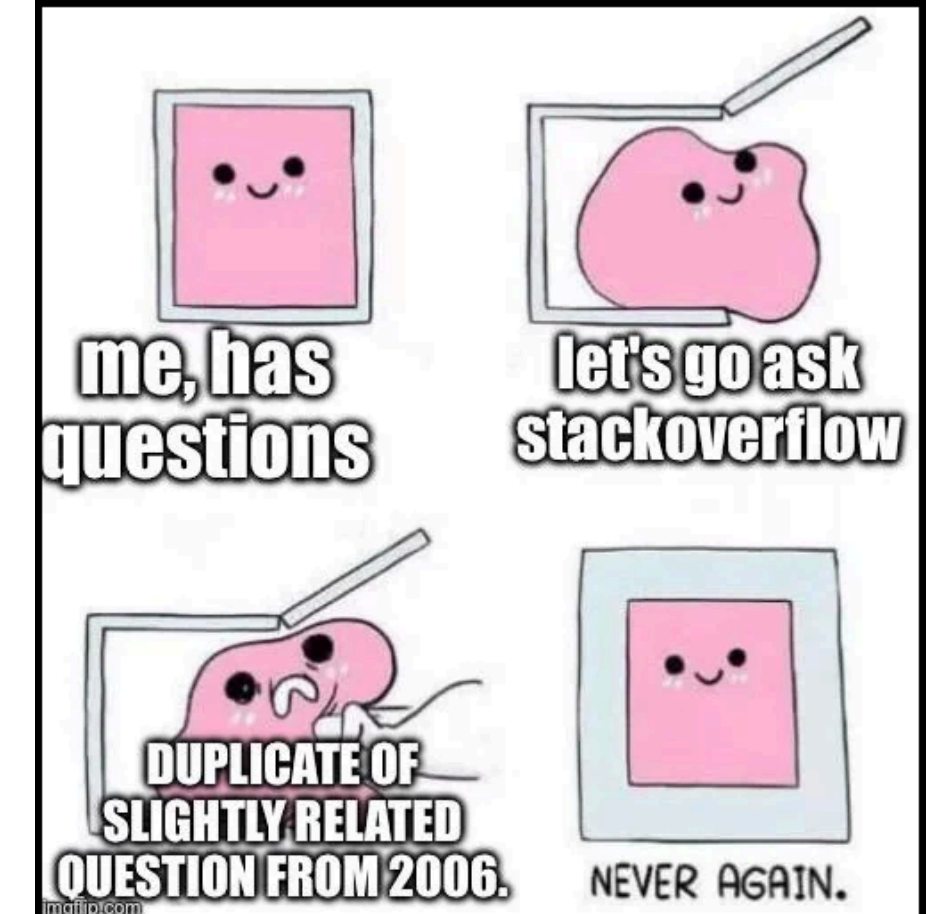


The narrative told by the image is clean, easy and reasonable for our expectations.

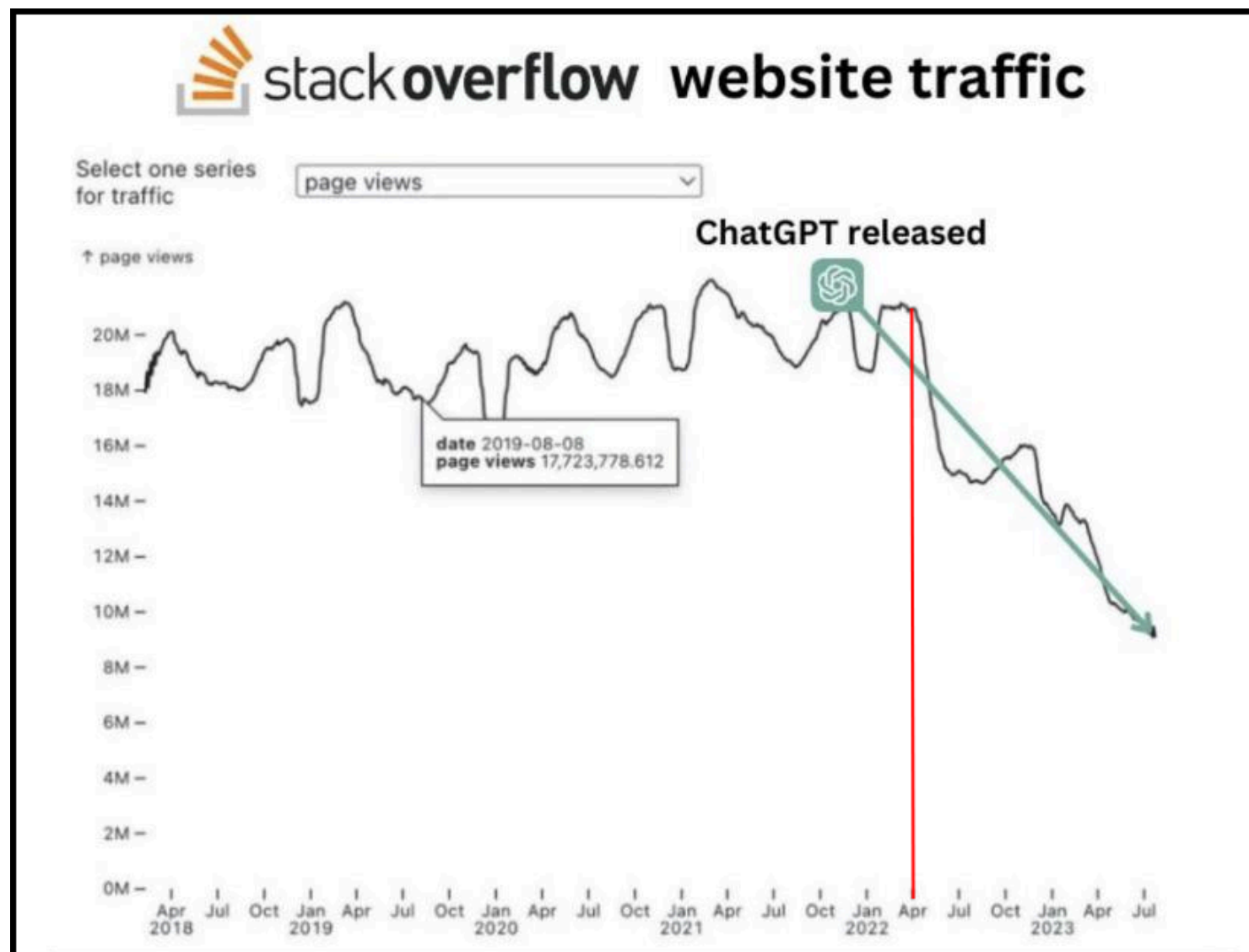
But is it the whole story?

Internet is biased against StackOverflow

Around 2018, in an effort to reduce duplicate questions and increase content quality the moderators of the website started to close low-effort or repeated questions, drawing the antipathies of the general public.



Is the attribution even correct?



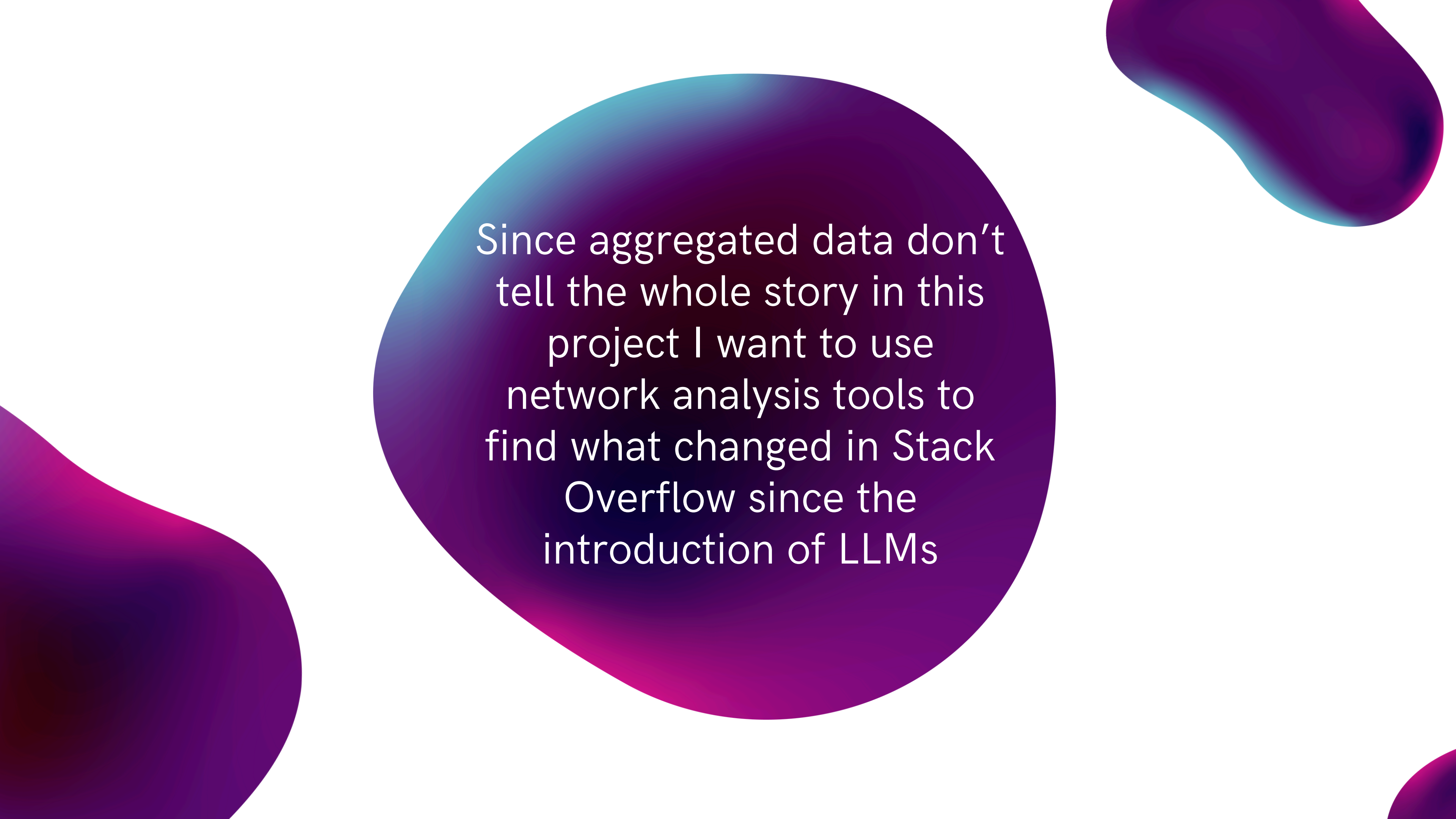
It is not hard to see that the steep decrease in views according to the plot started around the end of **april 2022**, however the first ChatGPT release was in late **november 2022**!

According to the website staff the decrease in traffic was caused to a recategorization of the Google Analytics cookies.

ChatGPT



Developer(s)	OpenAI
Initial release	November 30, 2022 (2 years ago)
Engine	GPT-4o GPT-4o mini GPT-4.5 o3 o4-mini ChatGPT Search
Platform	Cloud computing platforms
Type	Chatbot Large language model Generative pre-trained transformer
License	Proprietary service
Website	chatgpt.com



Since aggregated data don't tell the whole story in this project I want to use network analysis tools to find what changed in Stack Overflow since the introduction of LLMs

The background features three large, organic, fluid shapes in shades of purple and blue. One shape is in the top left, another is a large horizontal shape in the center, and a smaller one is in the bottom left. The word "Data" is written in white on the right side.

Data

What is Stack Overflow?

Stack Exchange is a network of question-and-answer (Q&A) websites on topics in diverse fields.

Stack Overflow is the first and most popular of those sites, dedicated broadly to computer science and programming.

Stack Exchange itself uploads monthly the data dump of every Stack Exchange website on the Internet Archive (<https://archive.org/>).

Structure of a question

How do I undo the most recent local commits in Git?

Title

Asked 15 years, 11 months ago Modified 7 days ago Viewed 16.1m times

27034

Question votes

I accidentally committed the wrong files to [Git](#) but haven't pushed the commit to the server yet.

How do I undo those commits from the *local* repository?

git version-control git-commit undo

Tags

Share Improve this question Follow

edited Oct 21, 2024 at 17:20

community wiki
94 revs, 65 users 11%
Peter Mortensen

838

Comment votes

You know what git needs? `git undo`, that's it. Then the reputation git has for handling mistakes made by us mere mortals disappears. Implement by pushing the current state on a git stack before executing any `git` command. It would affect performance, so it would be best to add a config flag as to whether to enable it. – Yimin Rong Mar 20, 2018 at 1:45

Show 16 more comments

Comment

102 Answers

Sorted by: Highest score (default) ▾

1 2 3 4 Next

29836

Answer votes

Undo a commit & redo

```
$ git commit -m "Something terribly misguided" # (0: Your Accident)
$ git reset HEAD~                               # (1)
# === If you just want to undo the commit, stop here! ===
[ edit files as necessary ]                       # (2)
$ git add .                                       # (3)
$ git commit -c ORIG_HEAD                         # (4)
```

Share Improve this answer Follow

edited Dec 10, 2024 at 12:20

community wiki
54 revs, 44 users 15%
CodeWizard

616

Comment votes

And if the commit was to the wrong branch, you may `git checkout theRightBranch` with all the changes stages. As I just had to do. – Frank Shearar Oct 5, 2010 at 15:44

Show 26 more comments

Comment on answer

Data Structure

The data dump is composed for each website of some .xml files. Those relevant for the study are:

- **Posts.xml**: all questions and answers. Relevant fields are the creation date, the body of the post, the user ID, the tags for questions and the parent post for answers.
- **Comments.xml**: all comments. Relevant fields are creation date, the body, the user ID and the parent post (can be either question or answer).
- **Users.xml**: all users. Relevant fields are creation date, last access date and user ID.

Research Questions

Has the population of the website changed since the arrival of LLMs?
Did the way new and recurrent users approach the website changed?

Has the complexity and the type of questions changed?
Is there a correlation between those changes?

RQ1: Has the population of the website changed since the arrival of LLMs? Has the way new and recurrent users approach the website changed?

To study this question I divided the data into yearly snapshots and built users–questions bipartite networks, with an edge between an user and a question if the user interacted in the question.

I took four snapshots: 2022 as the year in the just before the phenomenon I am studying, 2020 and 2024 as time periods before and after the phenomenon (with time to make the new behavior set on) and 2010 as a year in the past which had similar amounts of traffic and userbase to 2024.

2010

- **Unique users:** 299.840
- **Questions:** 676.010
- **Answers:** 1.430.846
- **Comments:** 2.161.492

2020

- **Unique users:** 833.656
- **Questions:** 1.814.646
- **Answers:** 2.408.406
- **Comments:** 6.675.264

2022

- **Unique users:** 679.743
- **Questions:** 1.322.352
- **Answers:** 1.705.974
- **Comments:** 4.645.471

2024

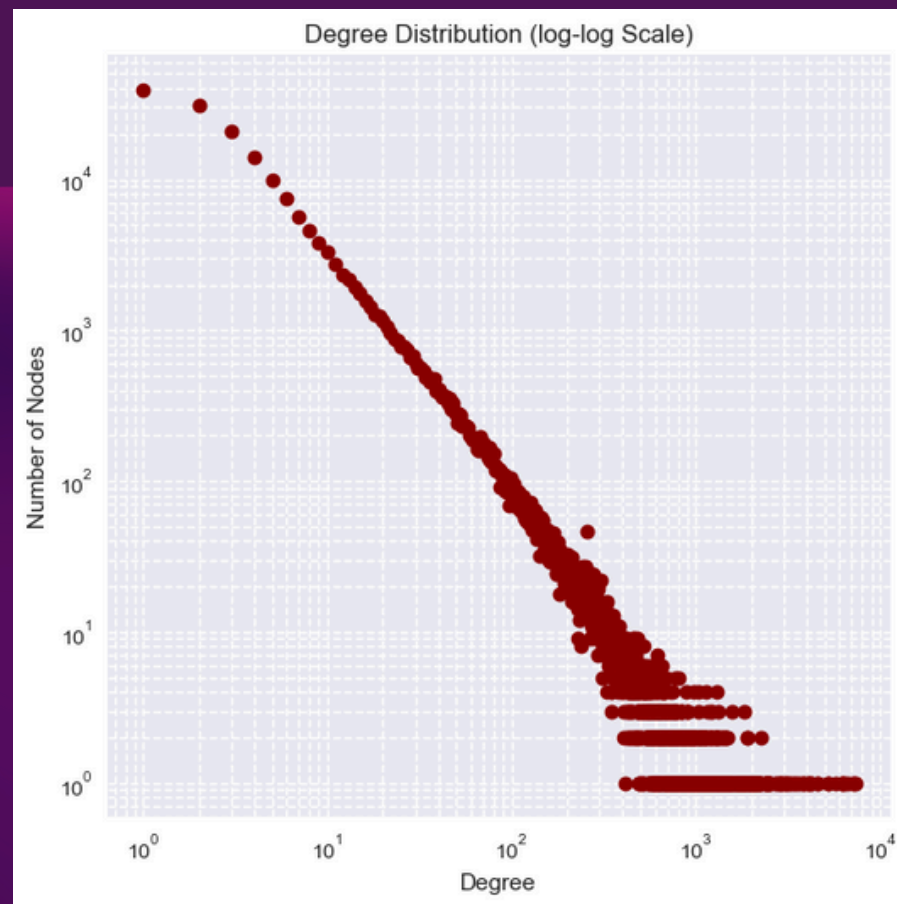
- **Unique users:** 285.945
- **Questions:** 526.916
- **Answers:** 572.331
- **Comments:** 1.786.678

Even from just those metrics we can notice a trend, and in particular a large difference between 2010 and 2024, despite the similar number of users.

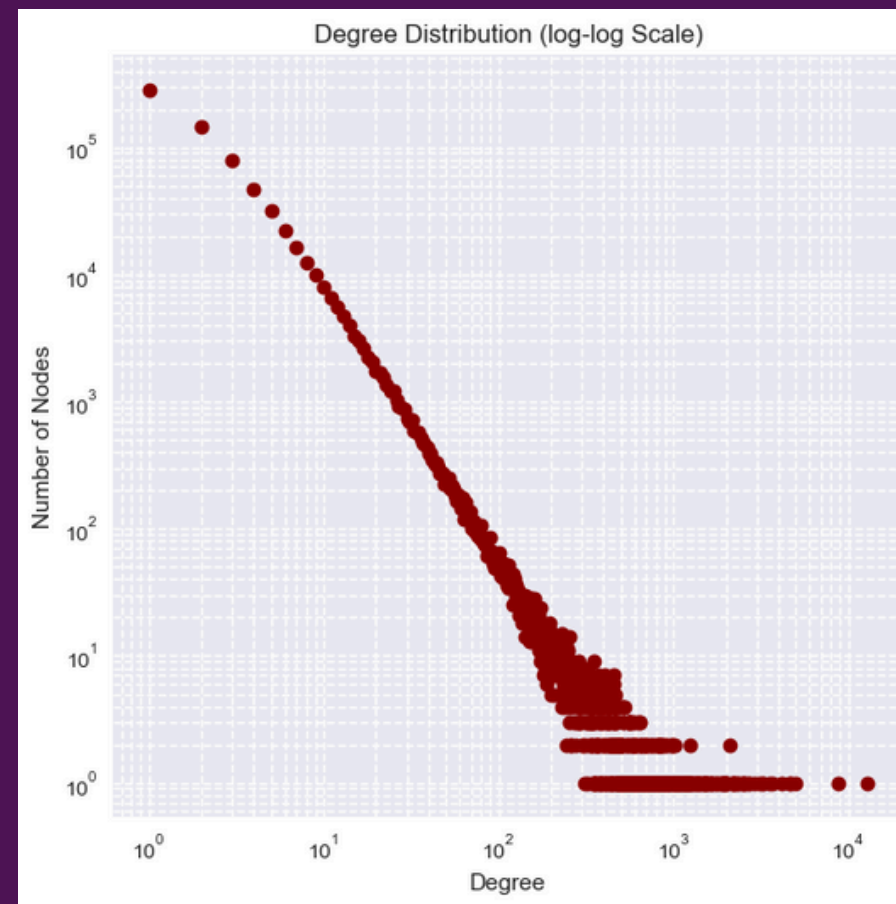
User projection analysis

Let's focus on the user projection of the graphs. We show the degree distribution (log-log scale):

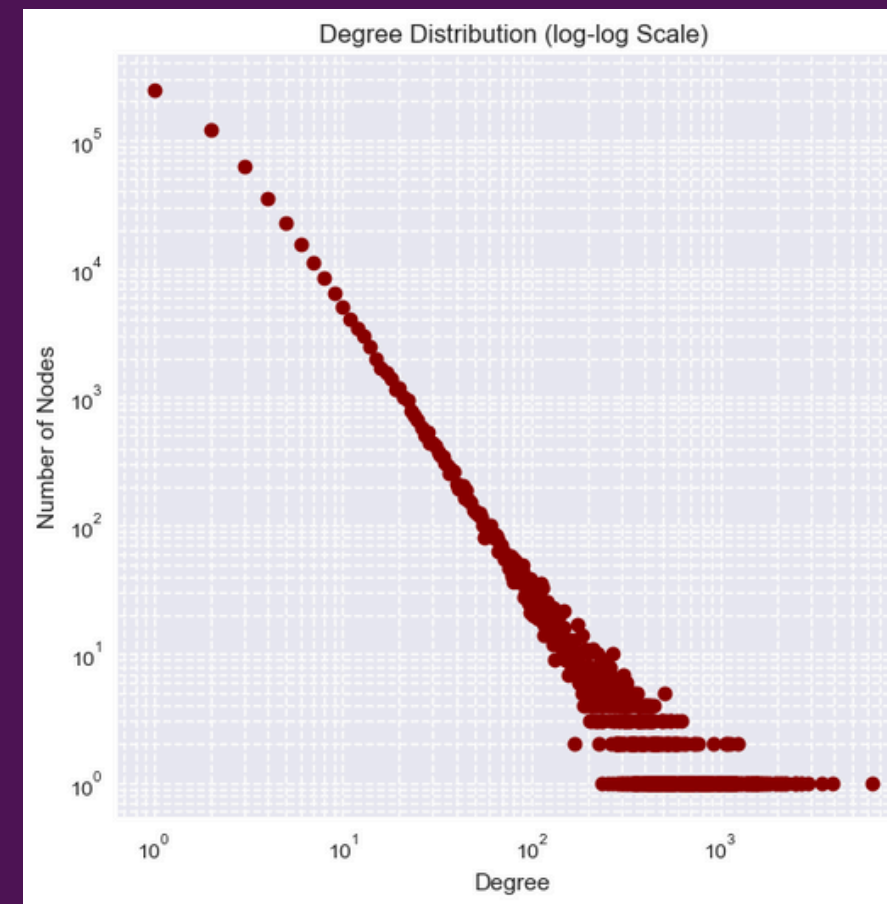
2010



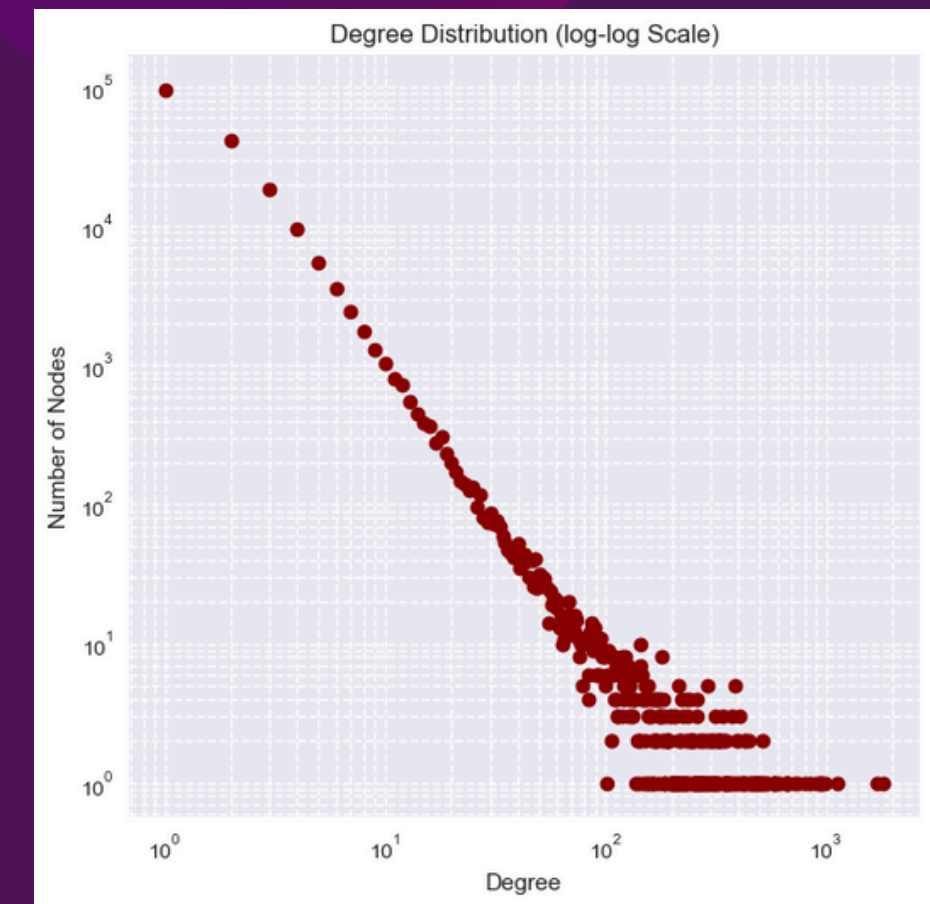
2020



2022



2024

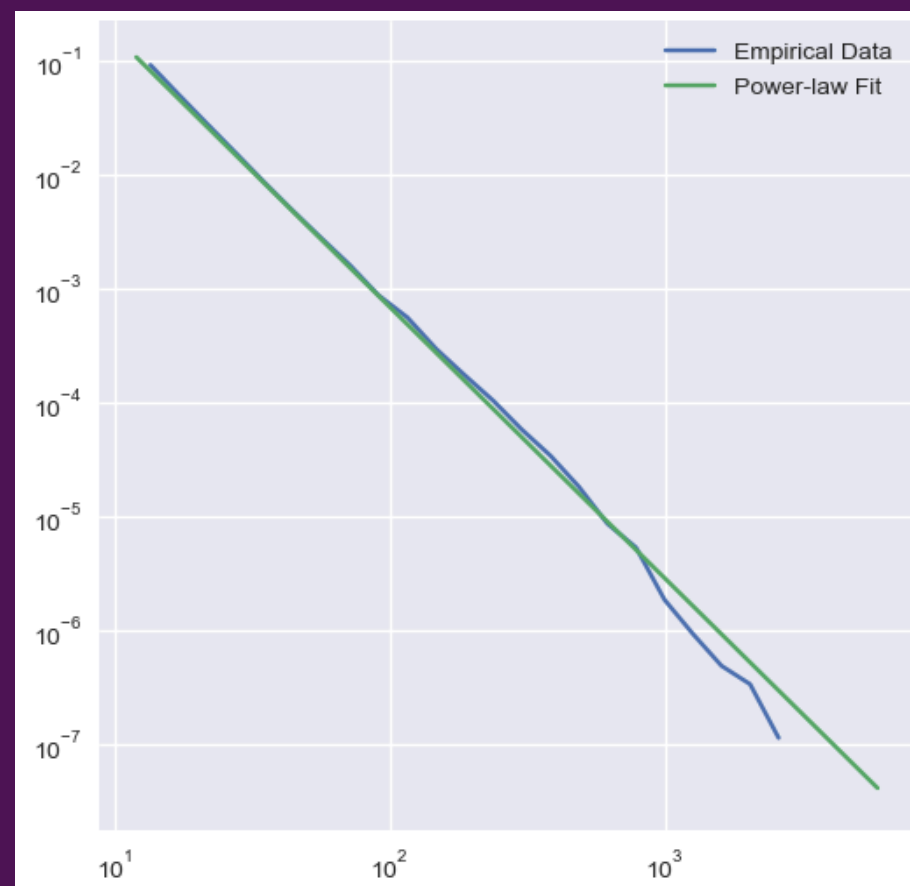


User projection analysis

Since we have a large amount of nodes in every instance of the graph it is reasonable to perform a power-law fit:

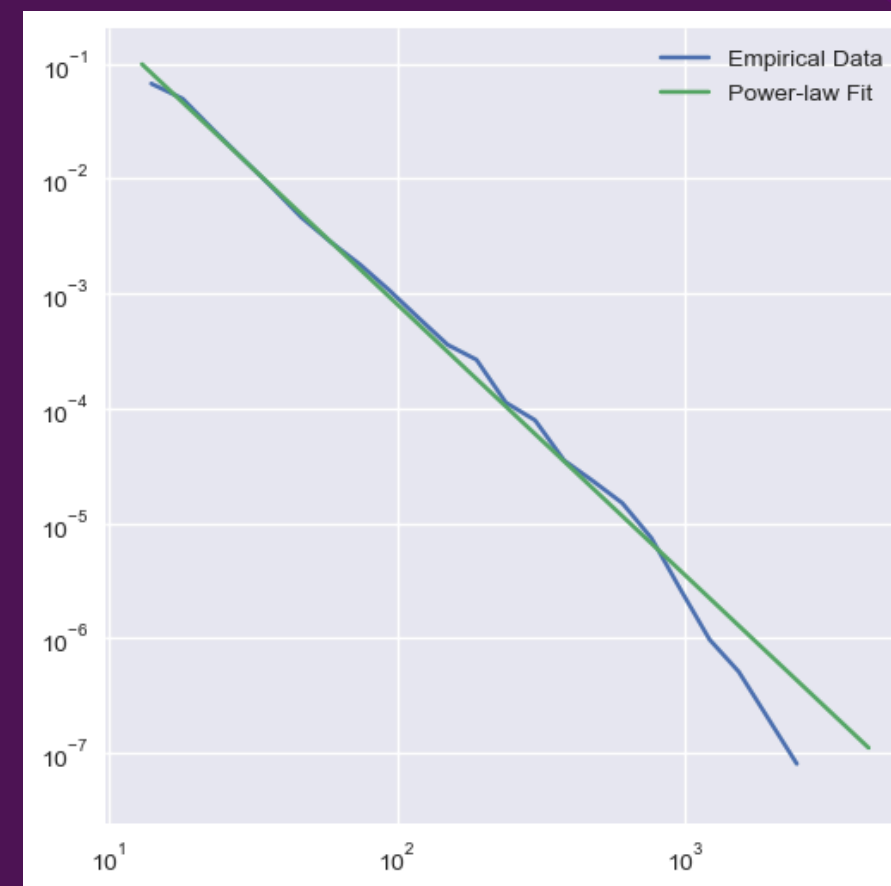
2010

KS statistic: 0.0319
Fitted α : 2.6687
p-value: 7.0824e-6



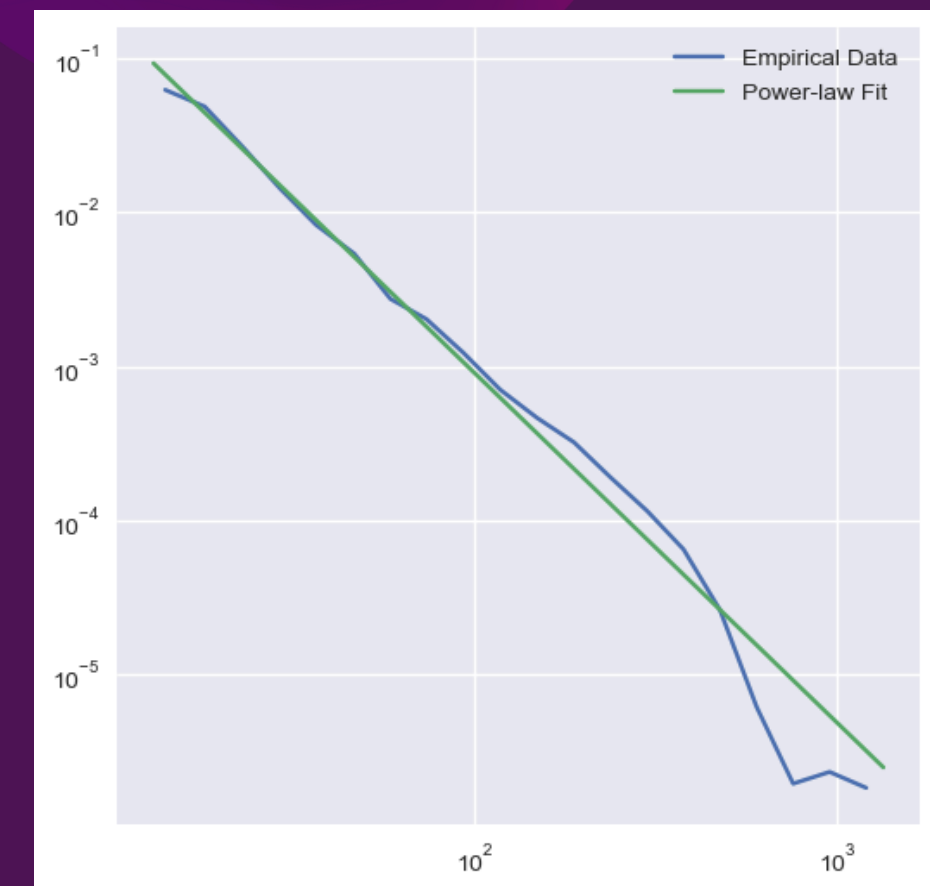
2020

KS statistic: 0.0063
Fitted α : 2.3882
p-value: 0.3010



2022

KS statistic: 0.0168
Fitted α : 2.3574
p-value: 0.2961



2024

KS statistic: 0.0170
Fitted α : 2.2752
p-value: 0.3059

Power-law fit observations

P-value significance

For the 2010 data we refute the null hypothesis that the data follows a power-law distribution as the p-value is off the boundary, even for a very high x-min value.

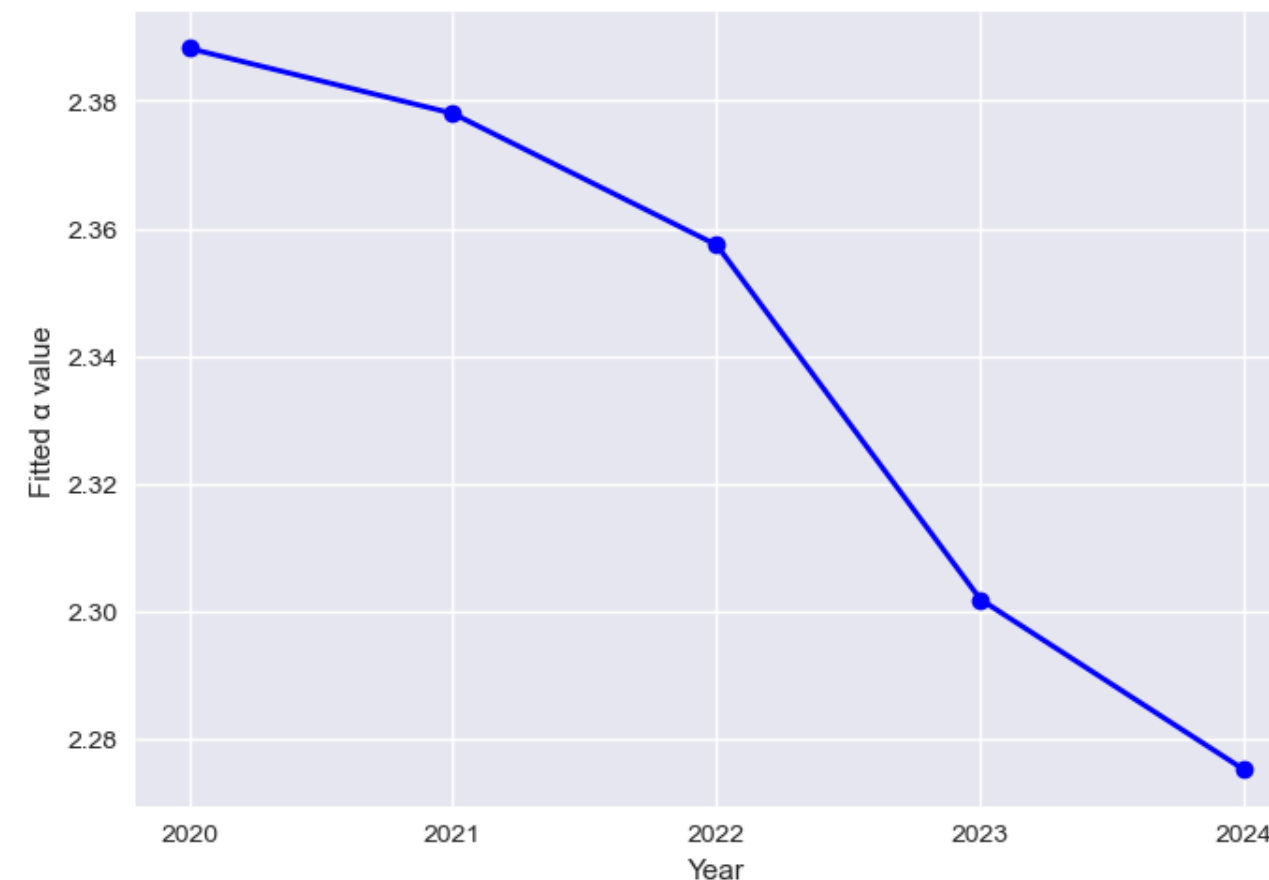
Meanwhile for the other networks taken into exam we cannot refute the null hypothesis.

Decrease of fitted α value

We notice a decreasing trend for the fitted α value over time, with a sharper decrease between the years 2022 and 2023.

This can be interpreted as an increased centralization; the network is becoming more dominated by high-degree nodes.

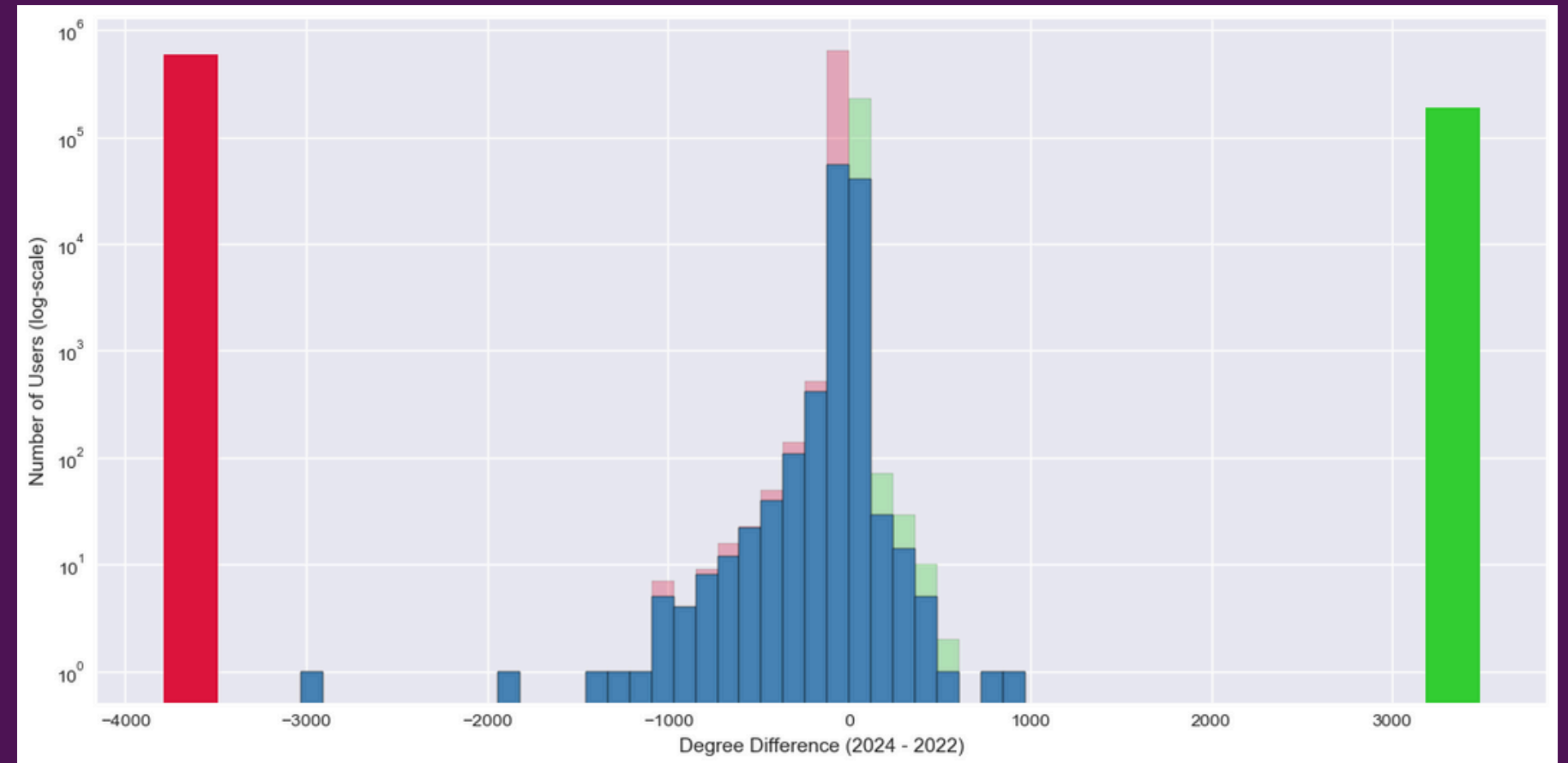
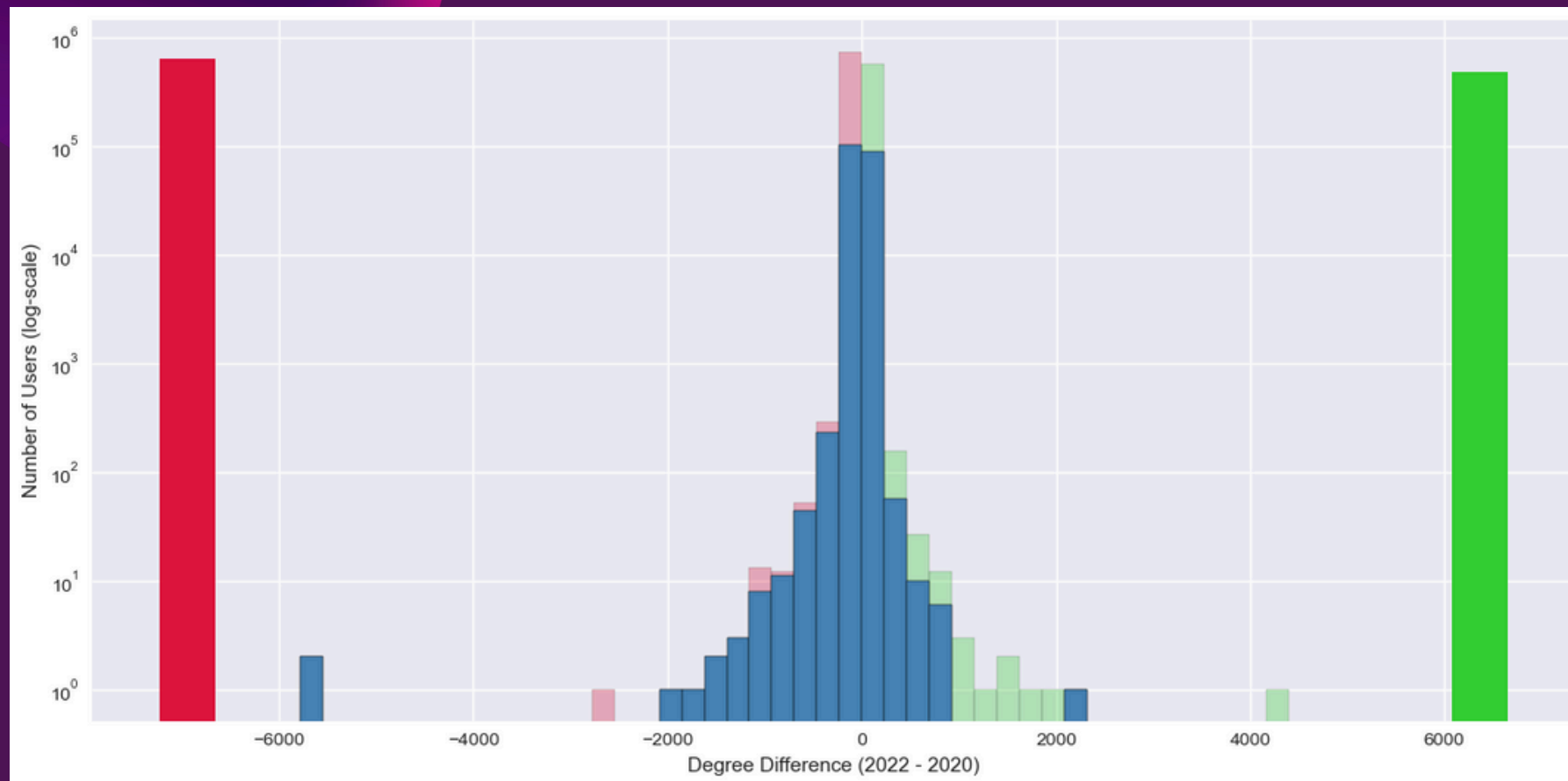
With the network growing more unequal the hub nodes are becoming stonger, even though the specific nodes may differ from year to year.



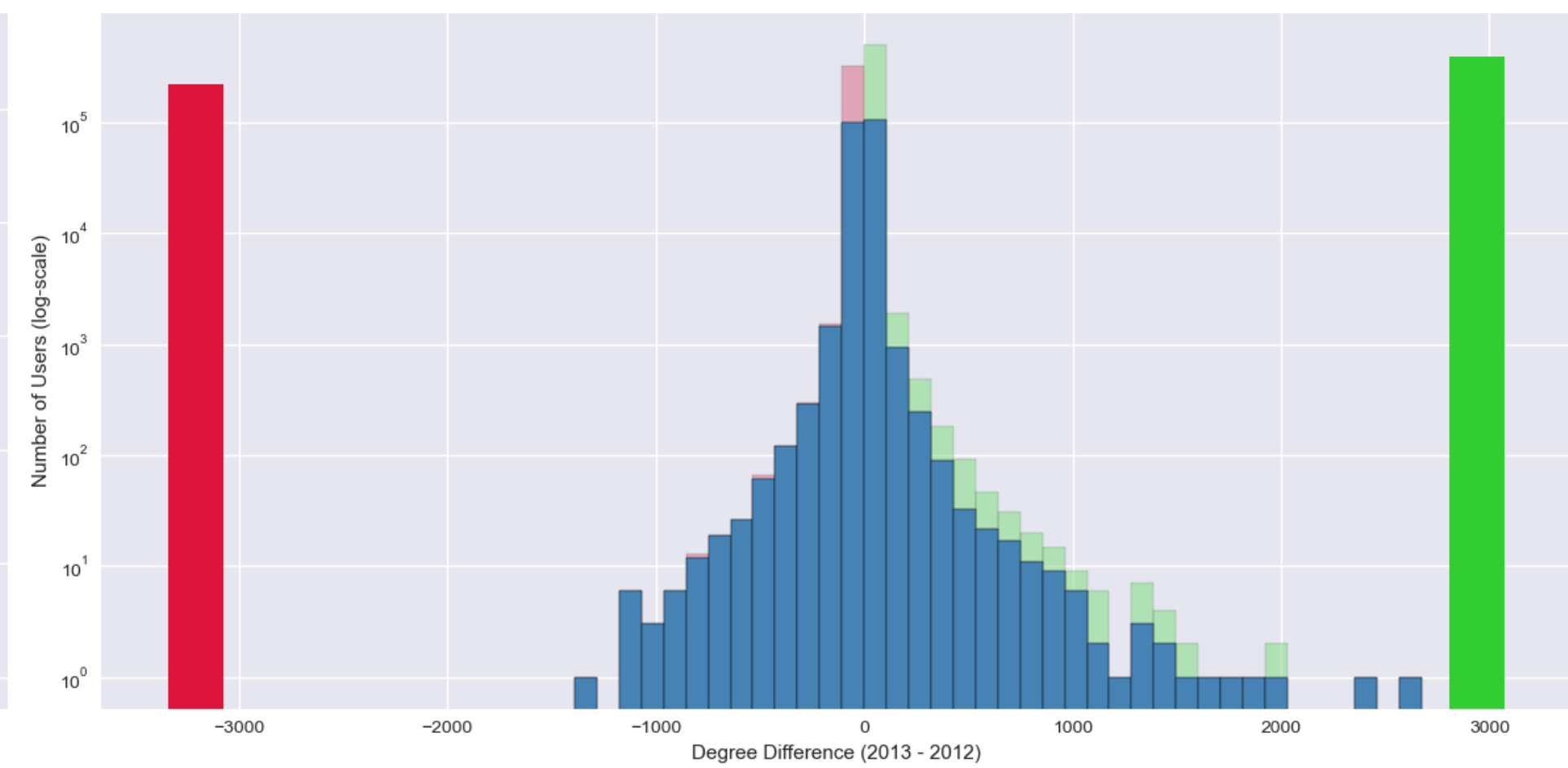
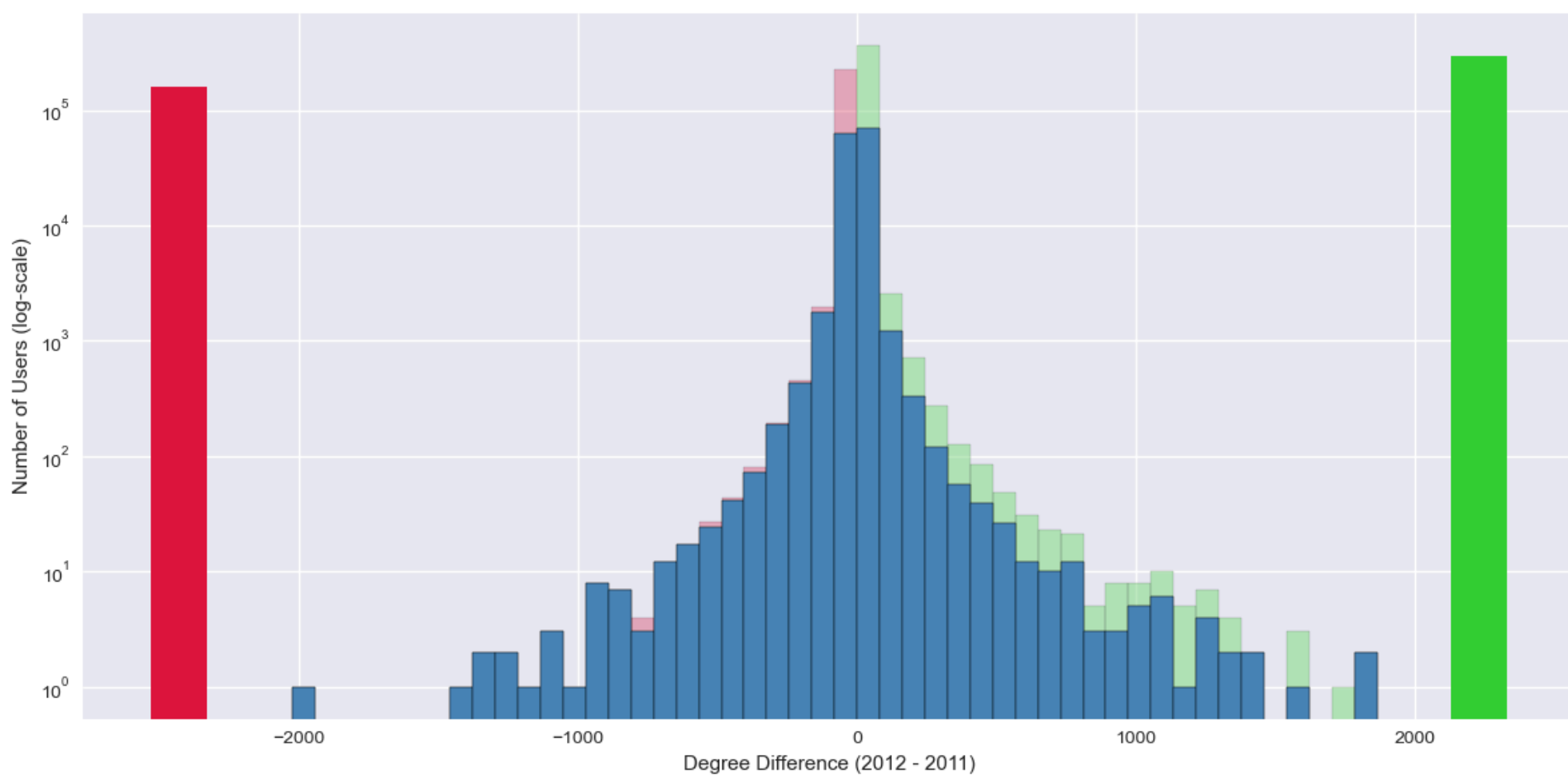
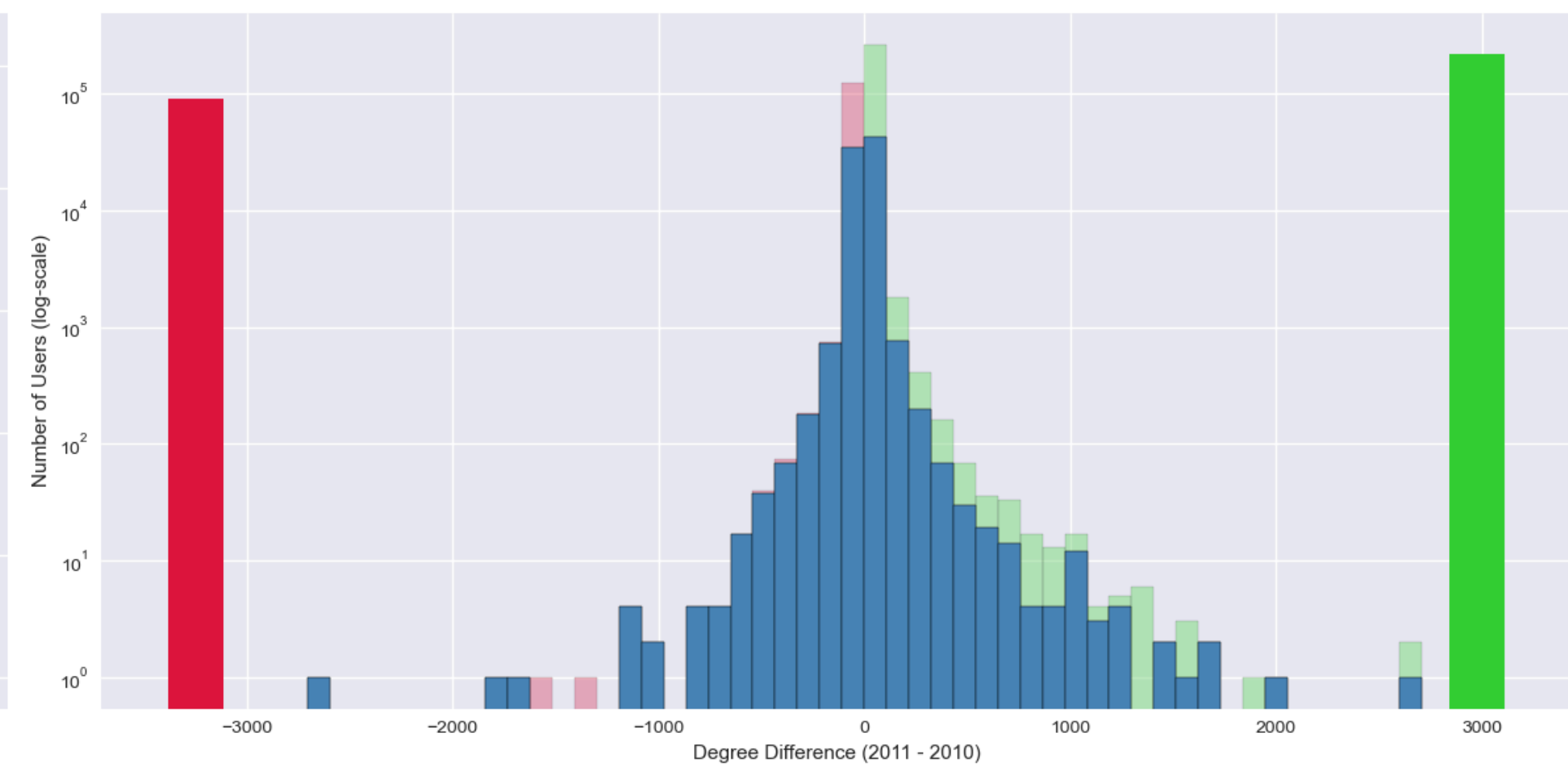
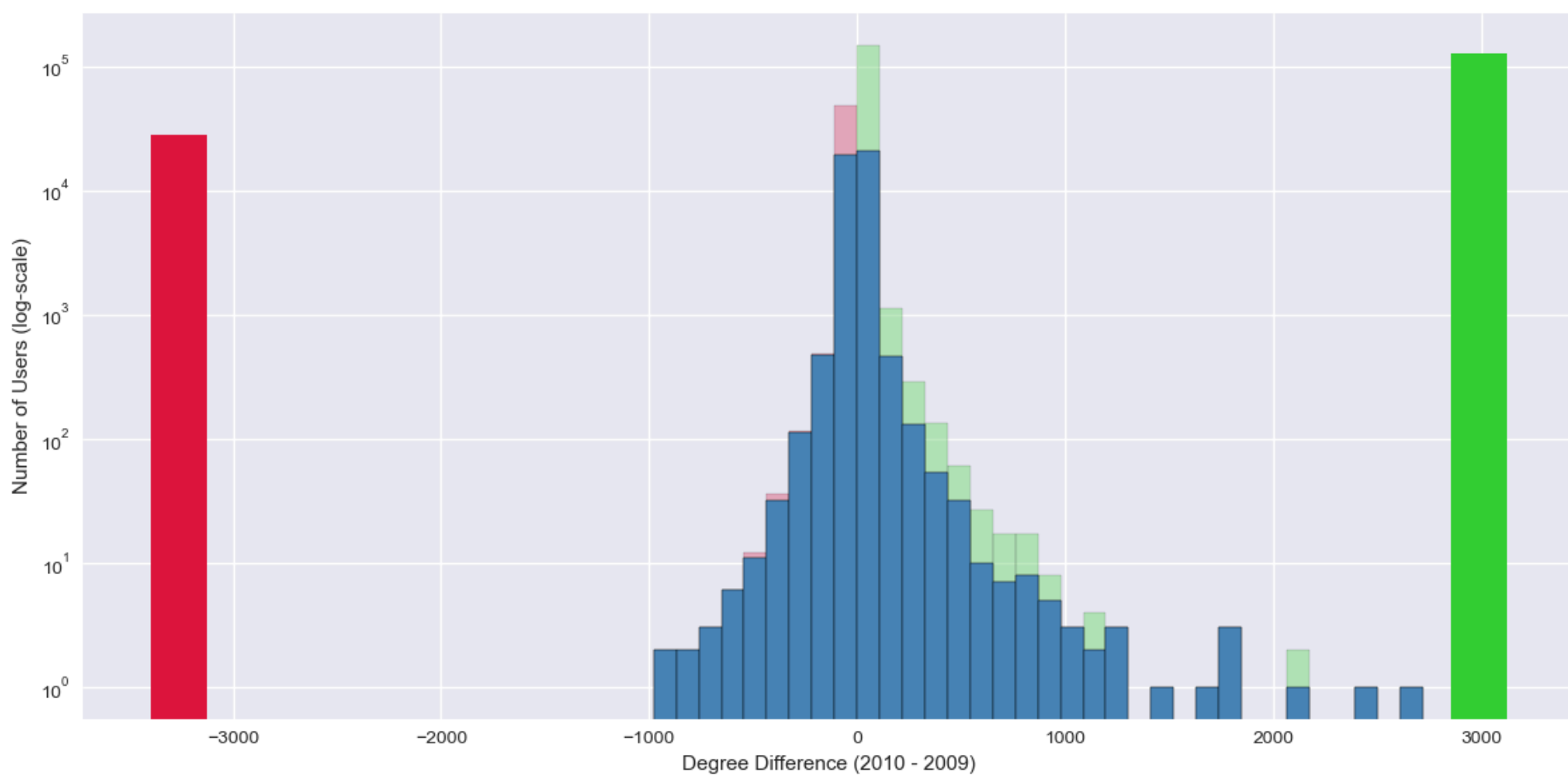
Fitted α value over time

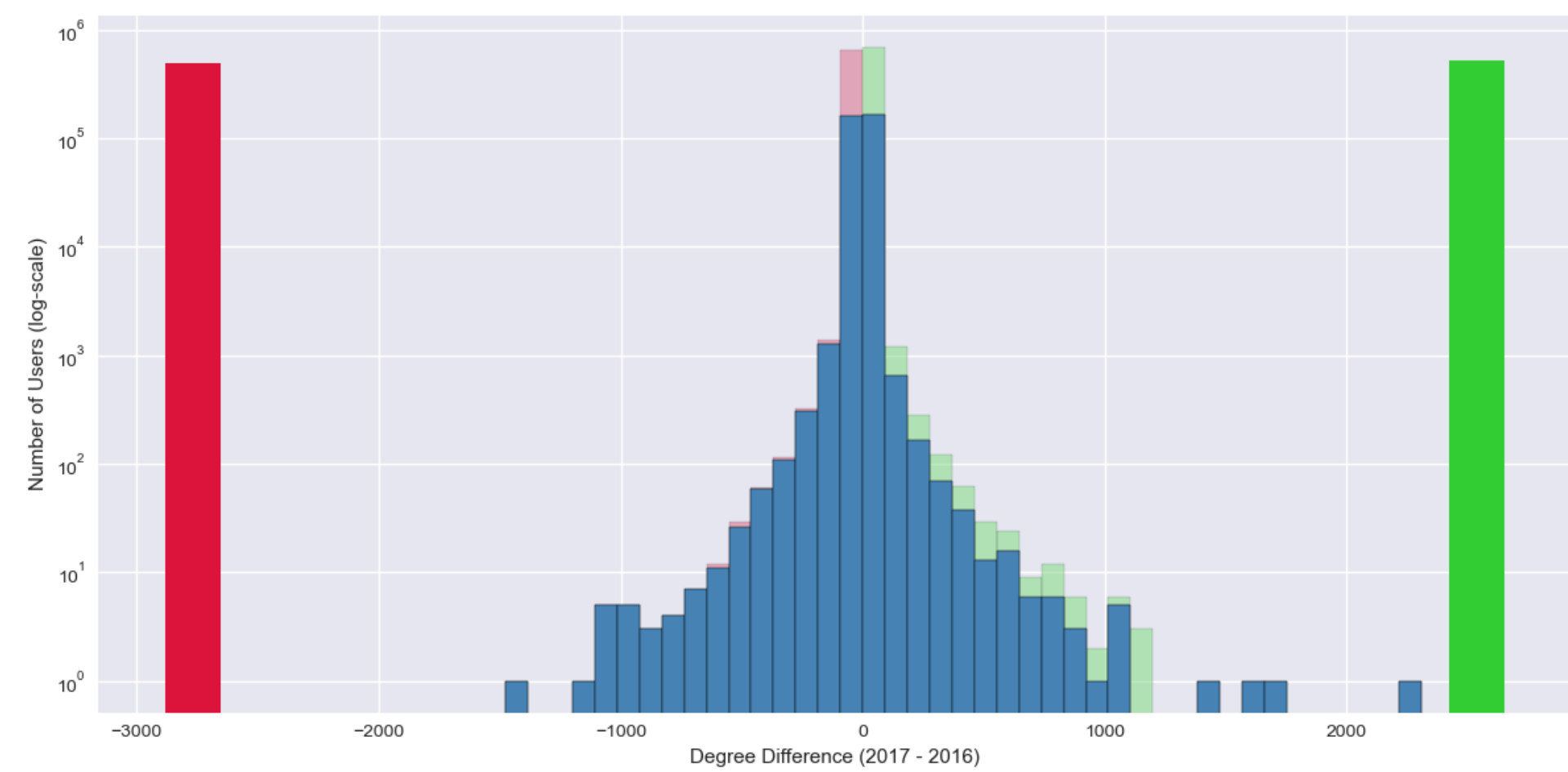
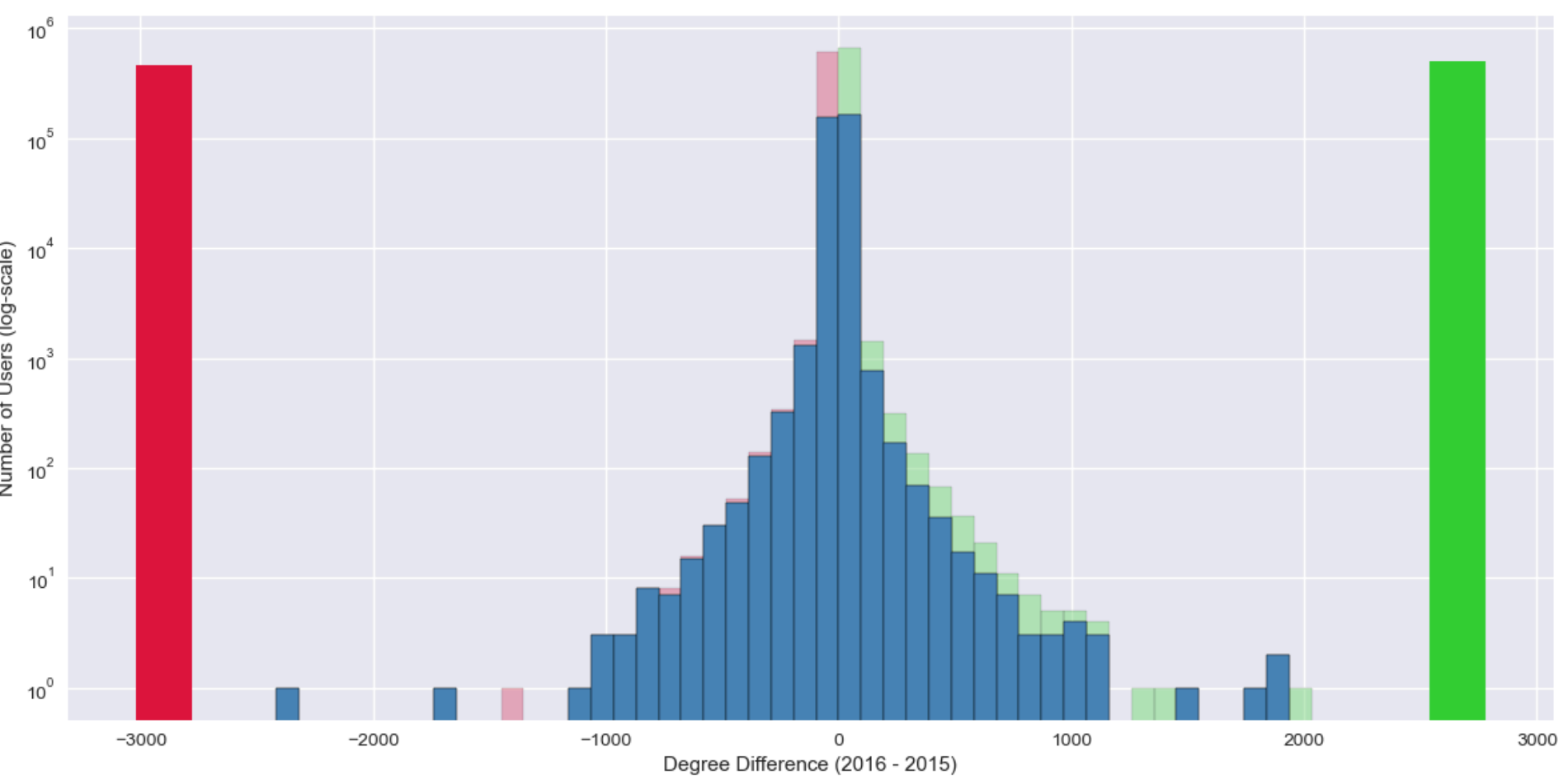
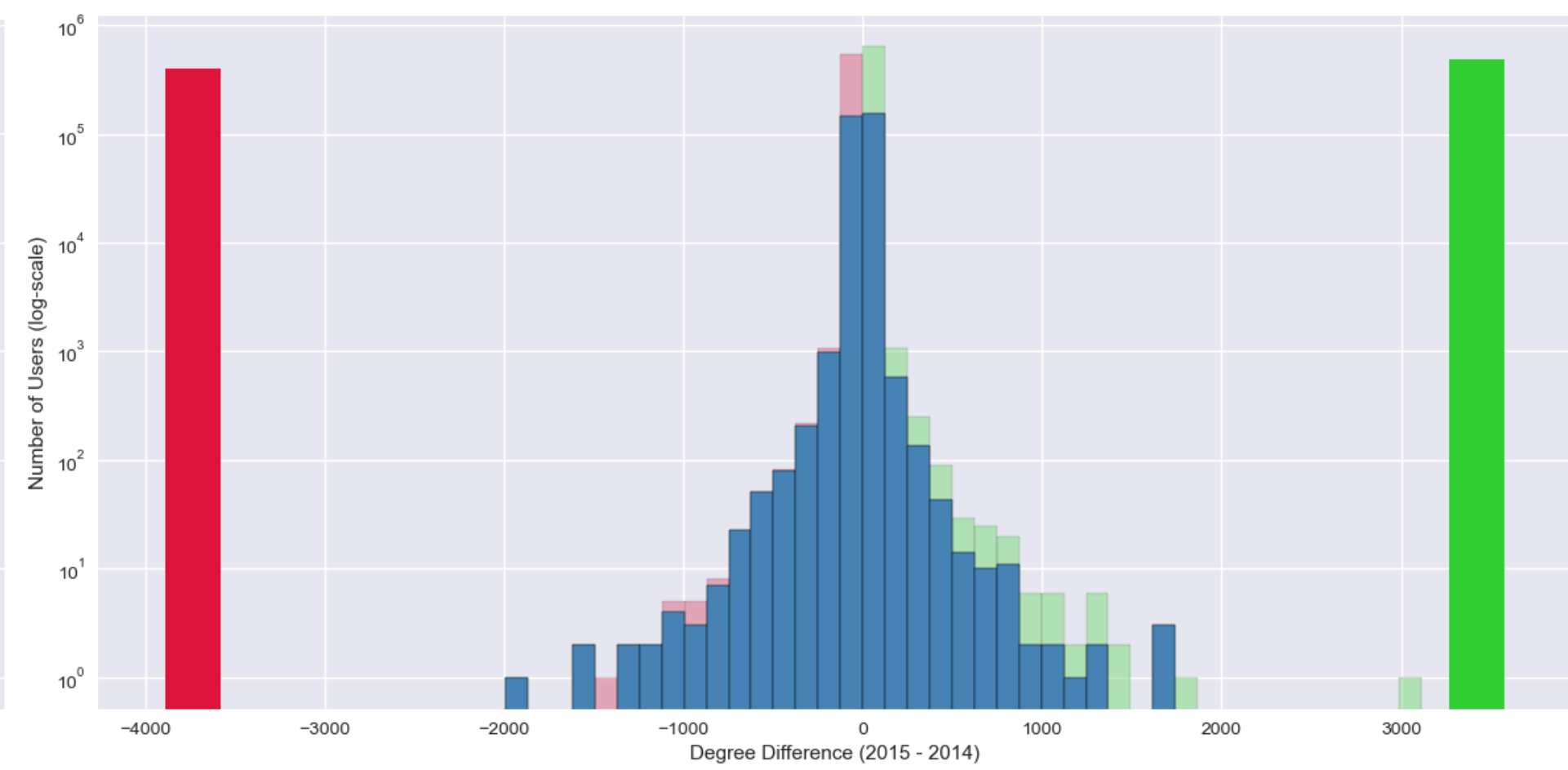
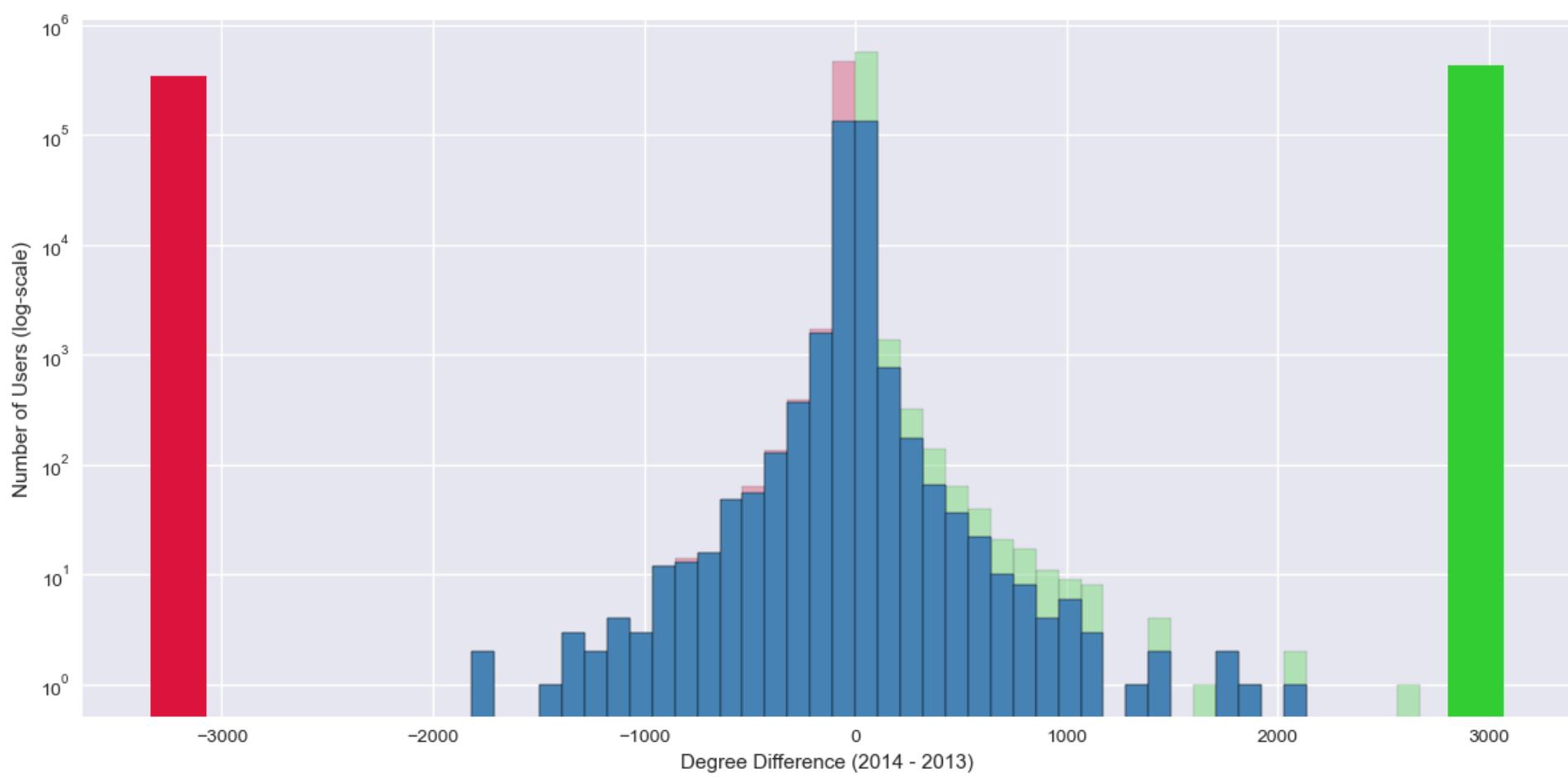
Users evolution over time

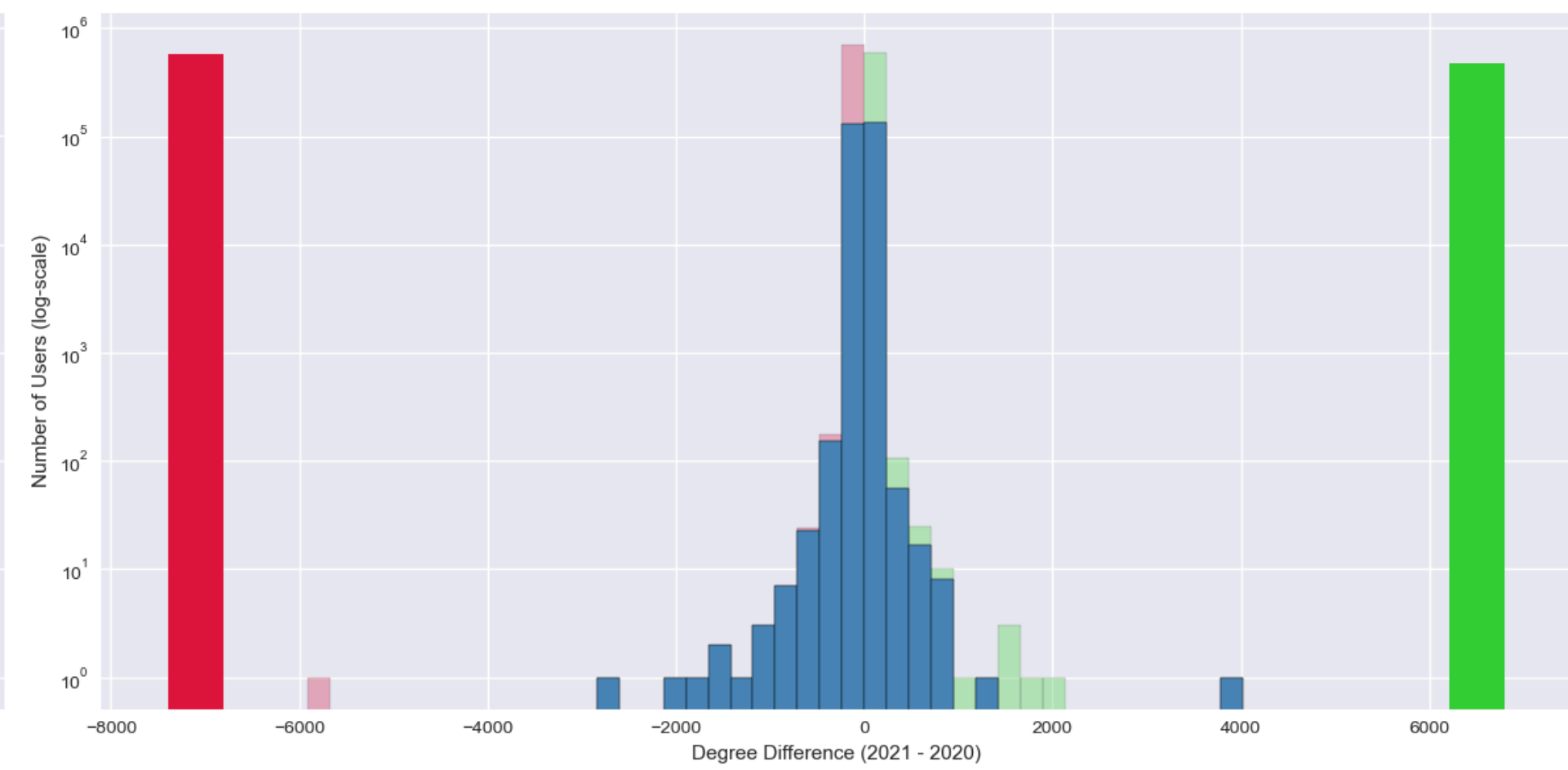
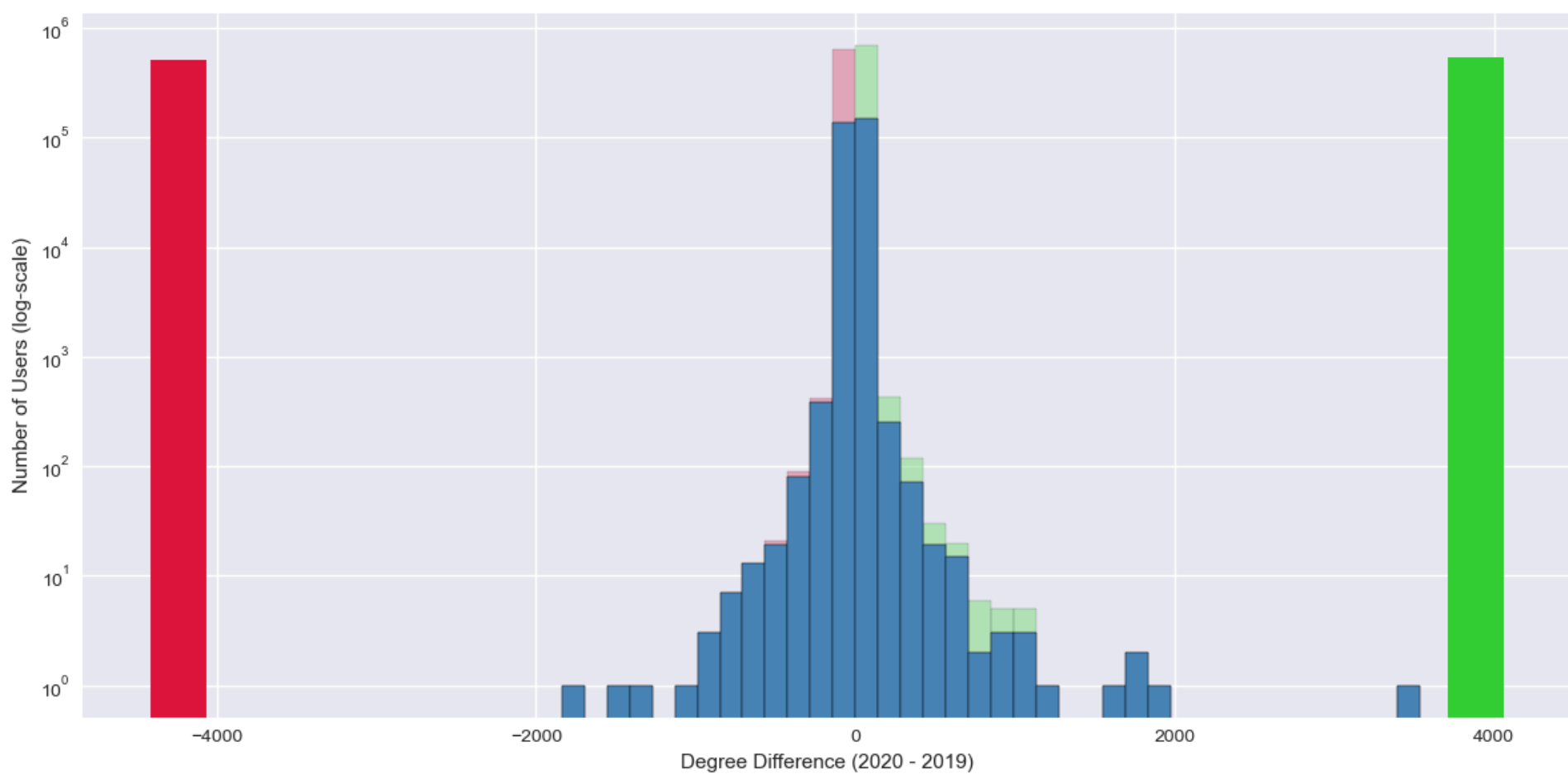
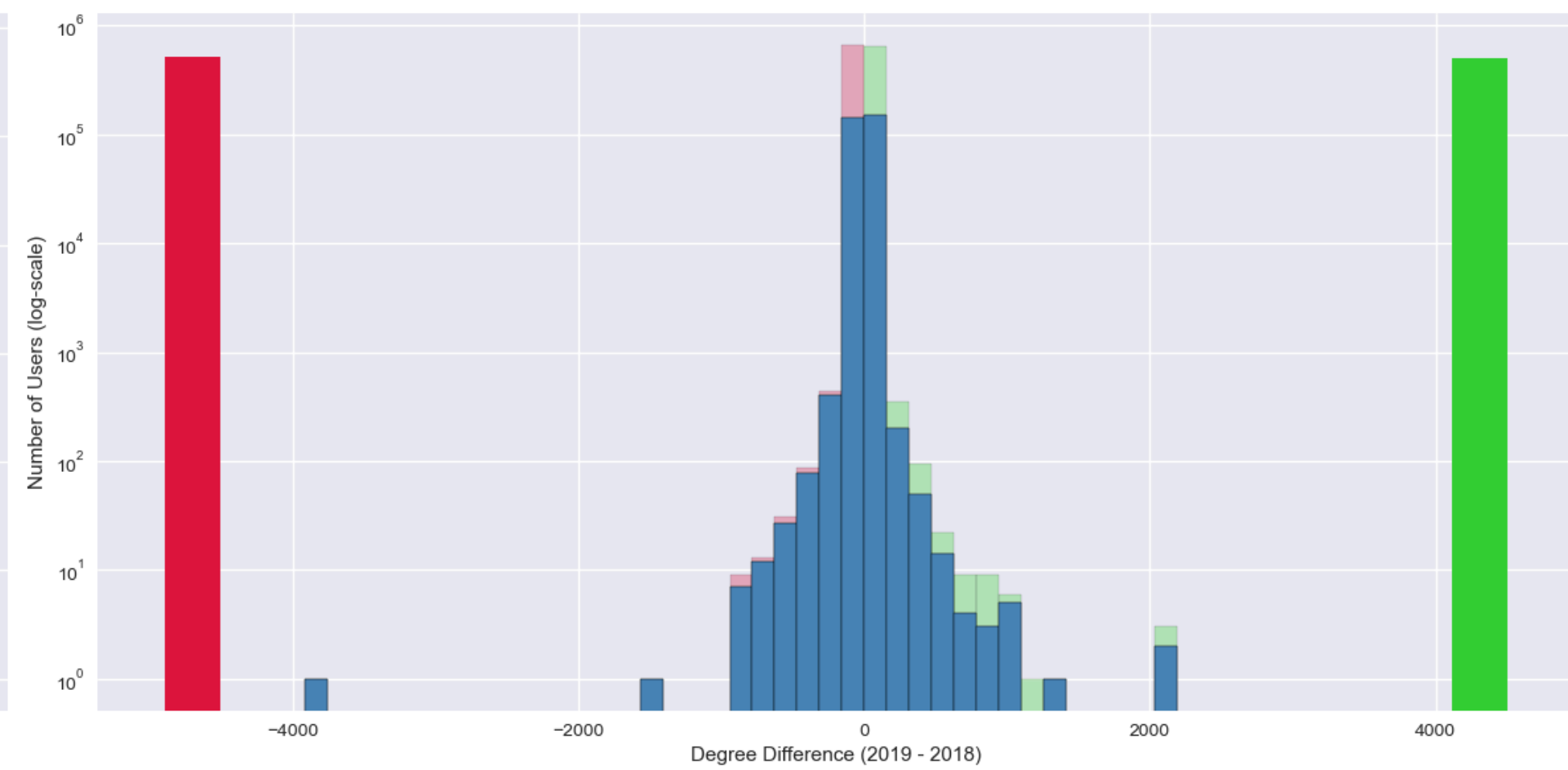
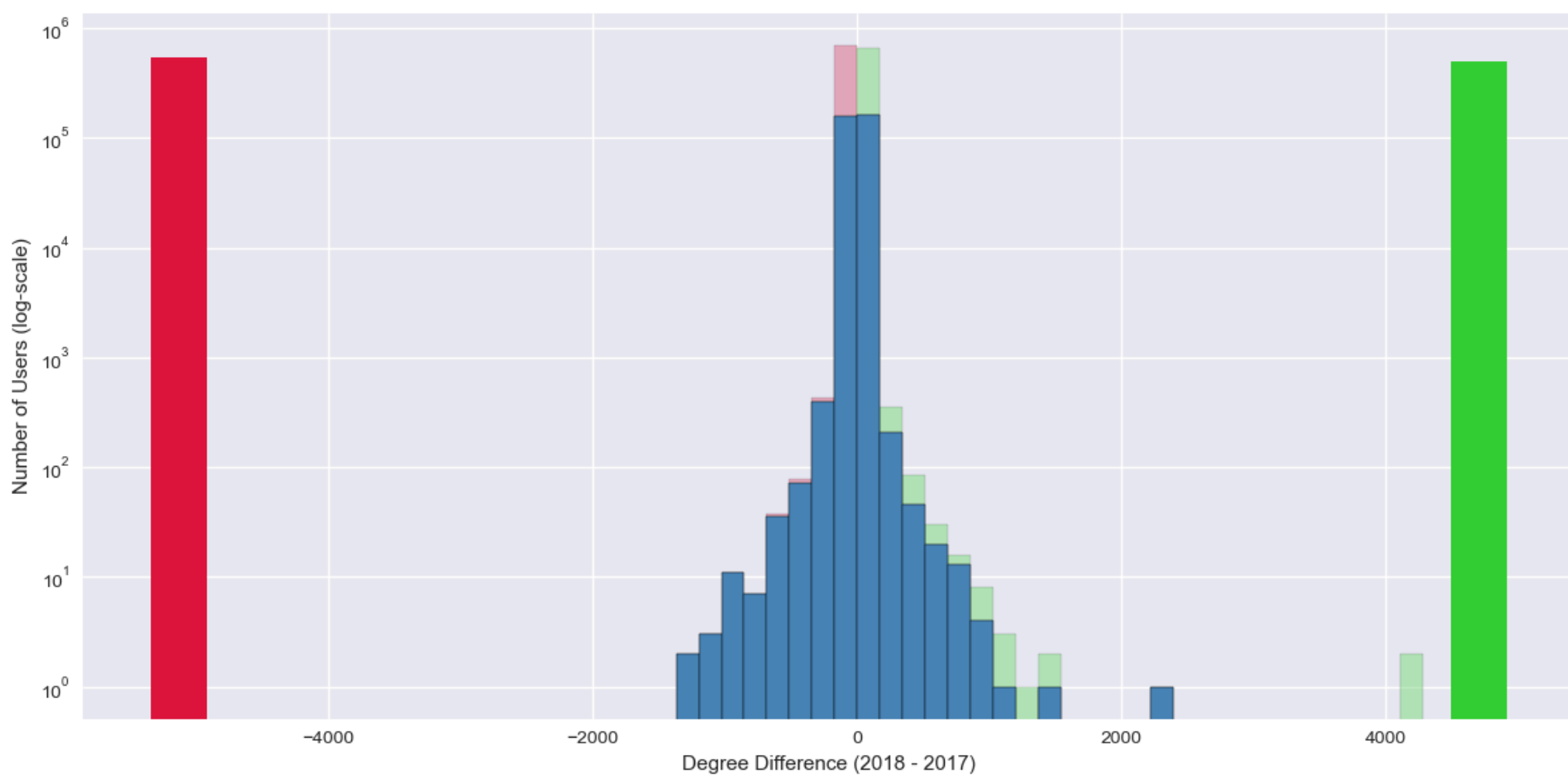
Previous results tell how the activity of the population changed over time, but give us no information on the behavior of the individuals. To study this we track the users over time, with particular attention on users that left and joined the network between two years.

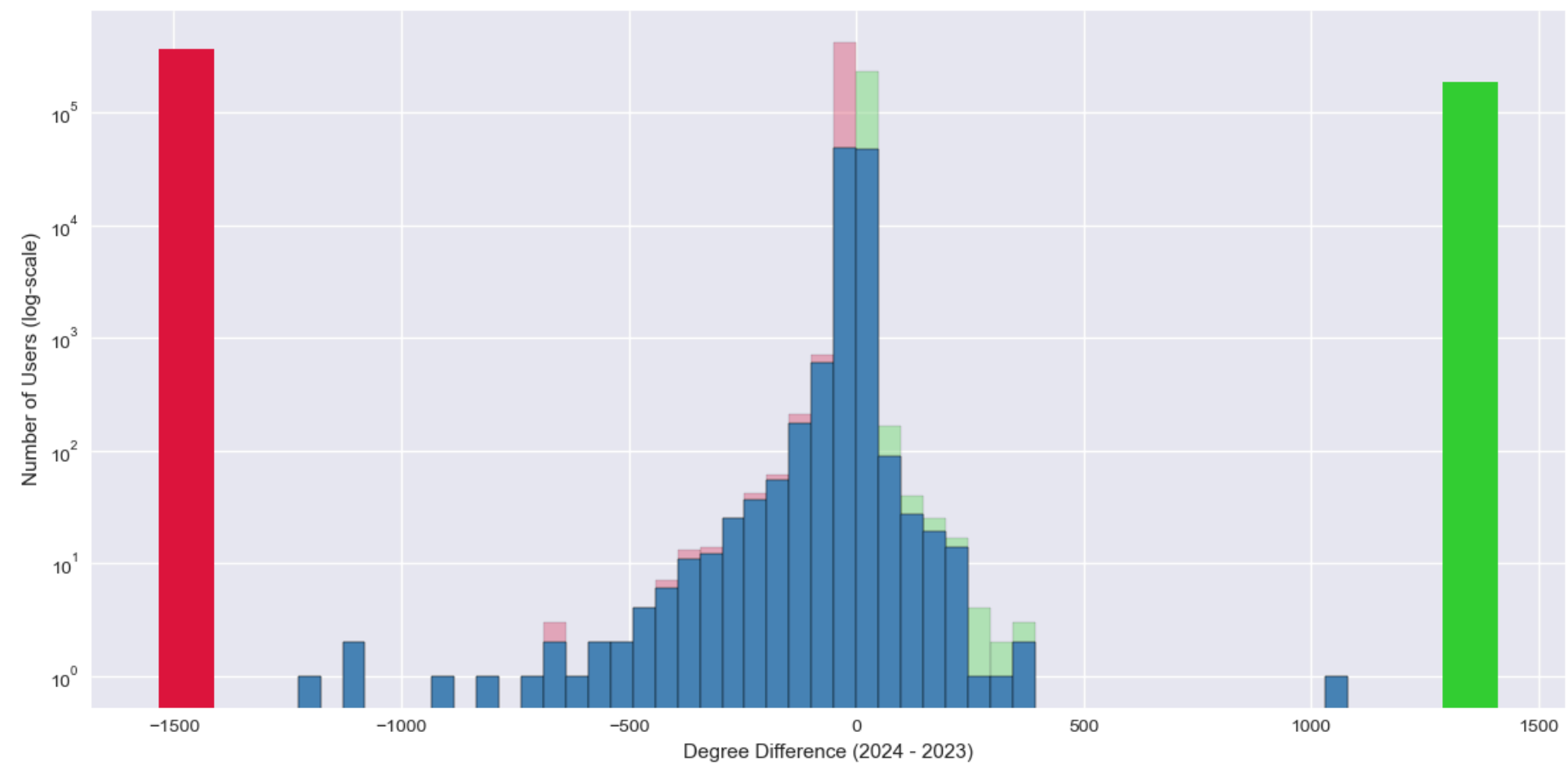
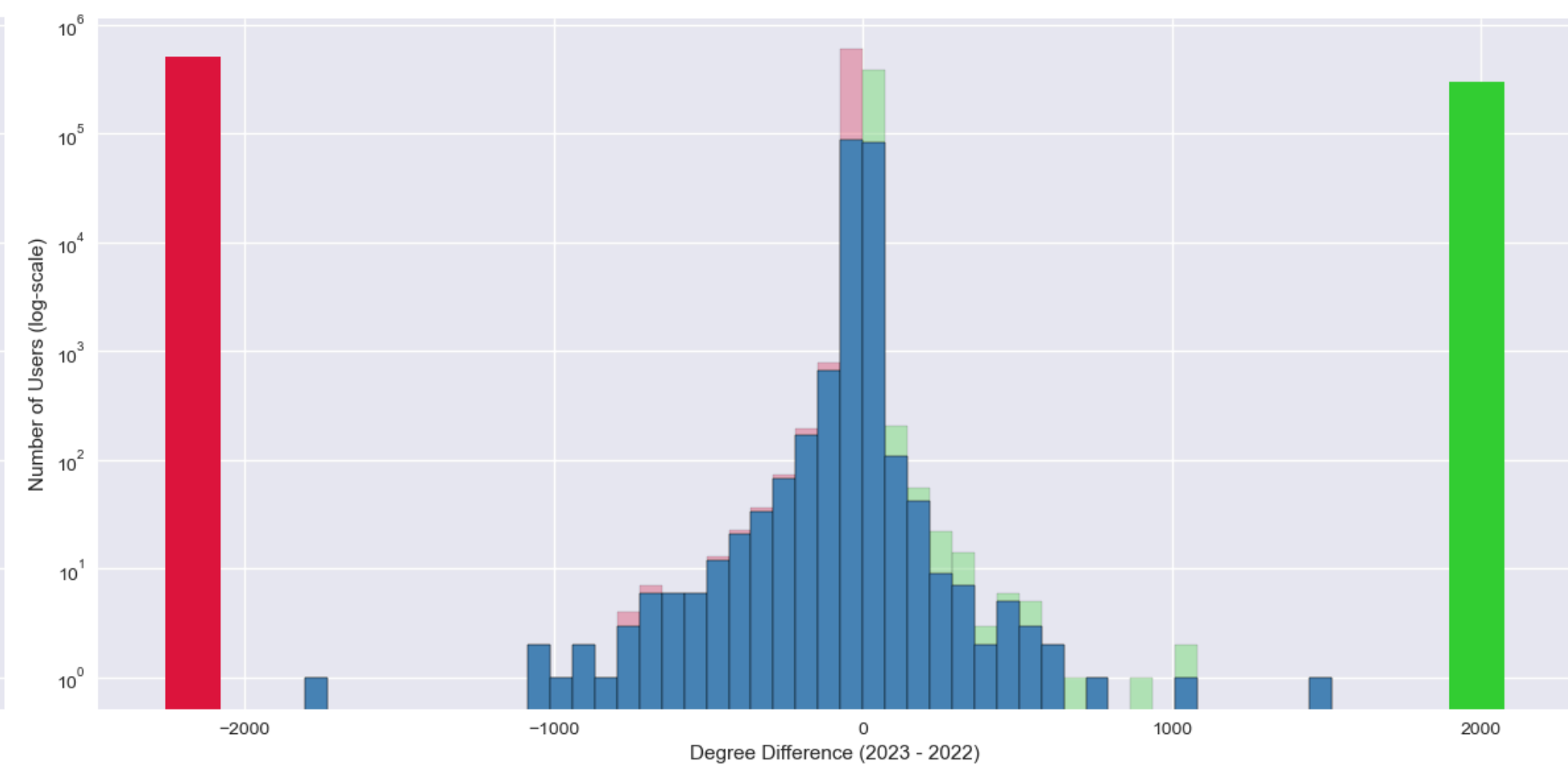
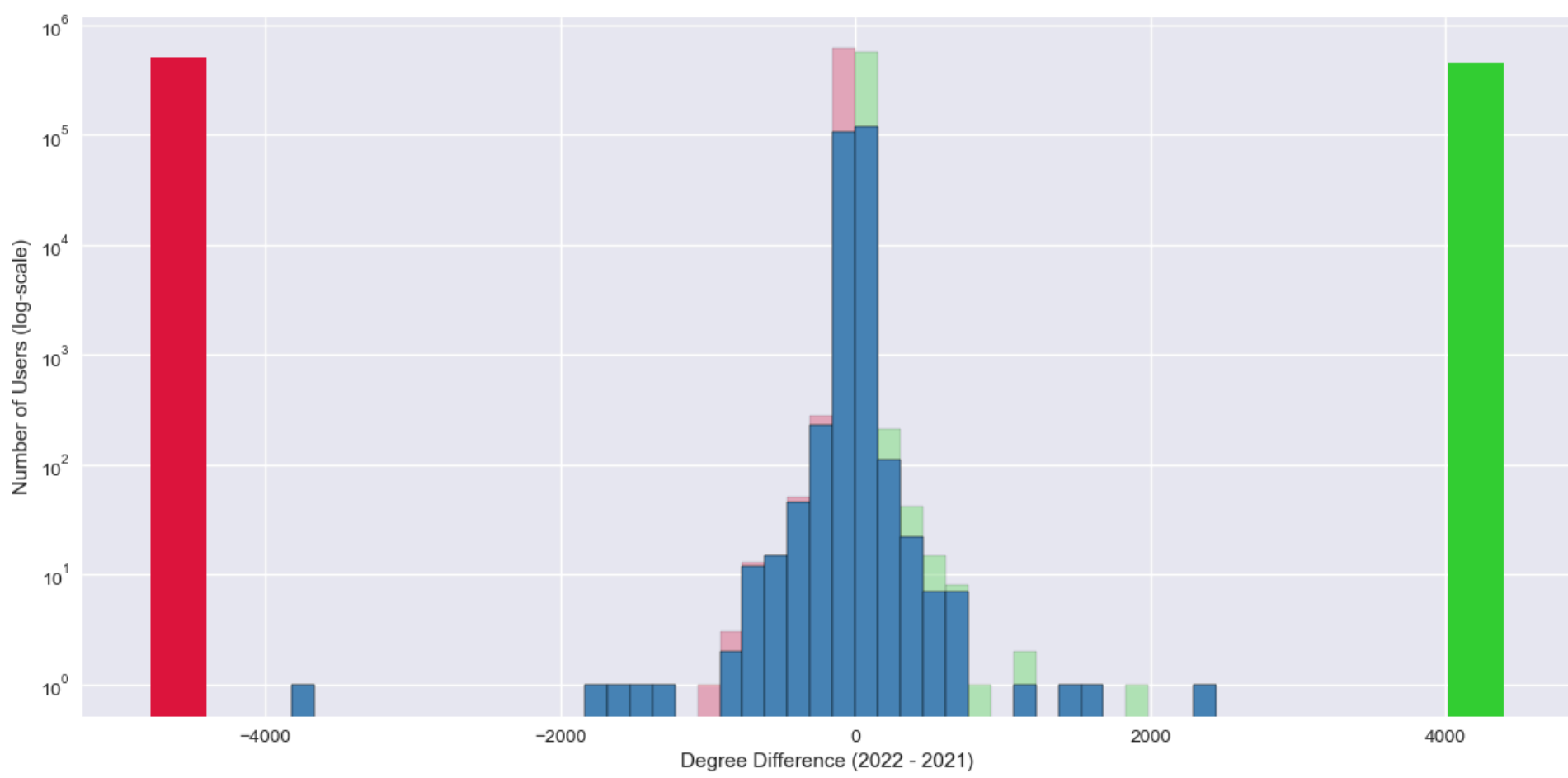


In the plots we have at the extremes the amount of users that joined and left the network, in blue the distribution of the degree difference of users present in both networks and in transparency the distributions counting also users with a missing entry.





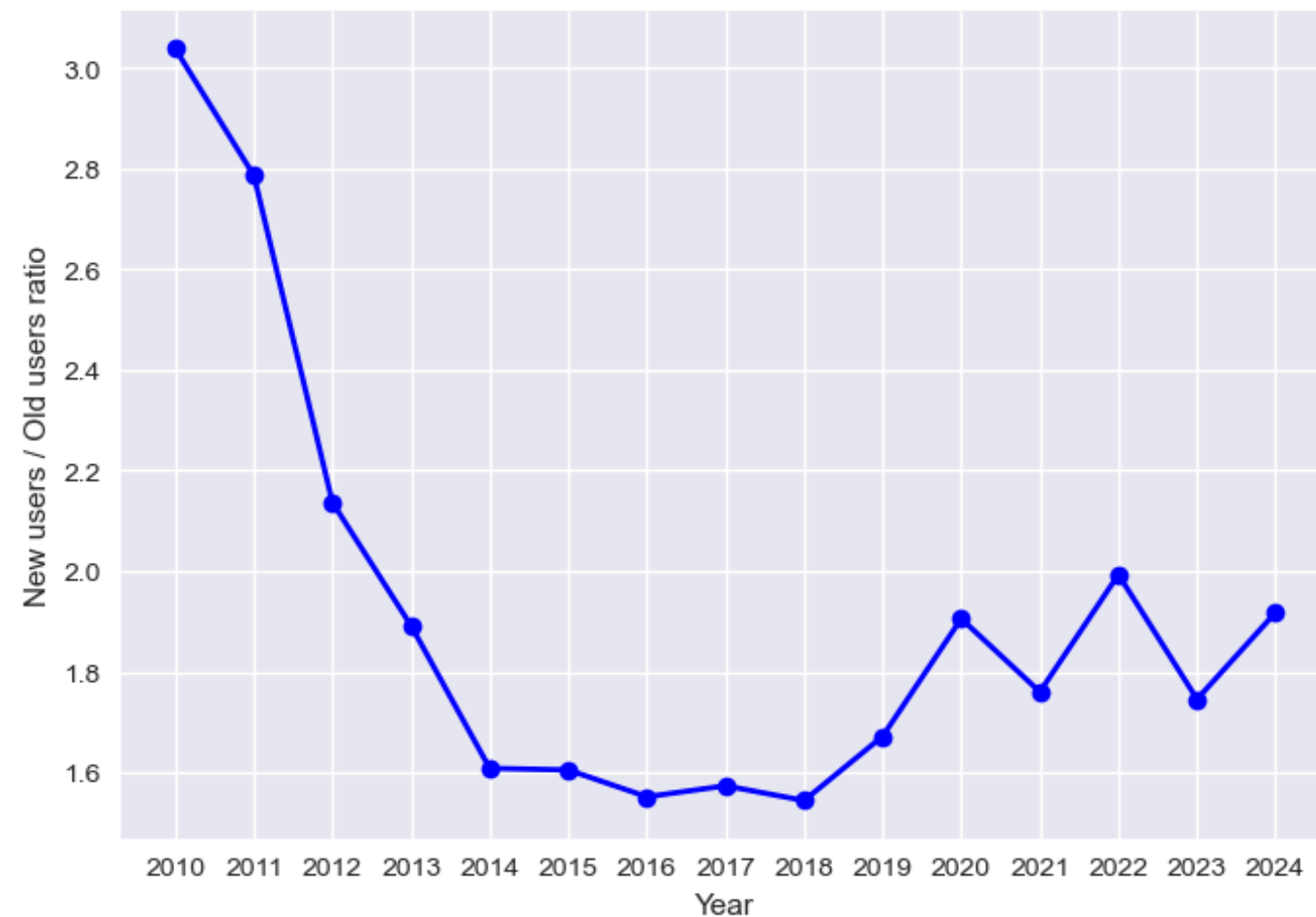




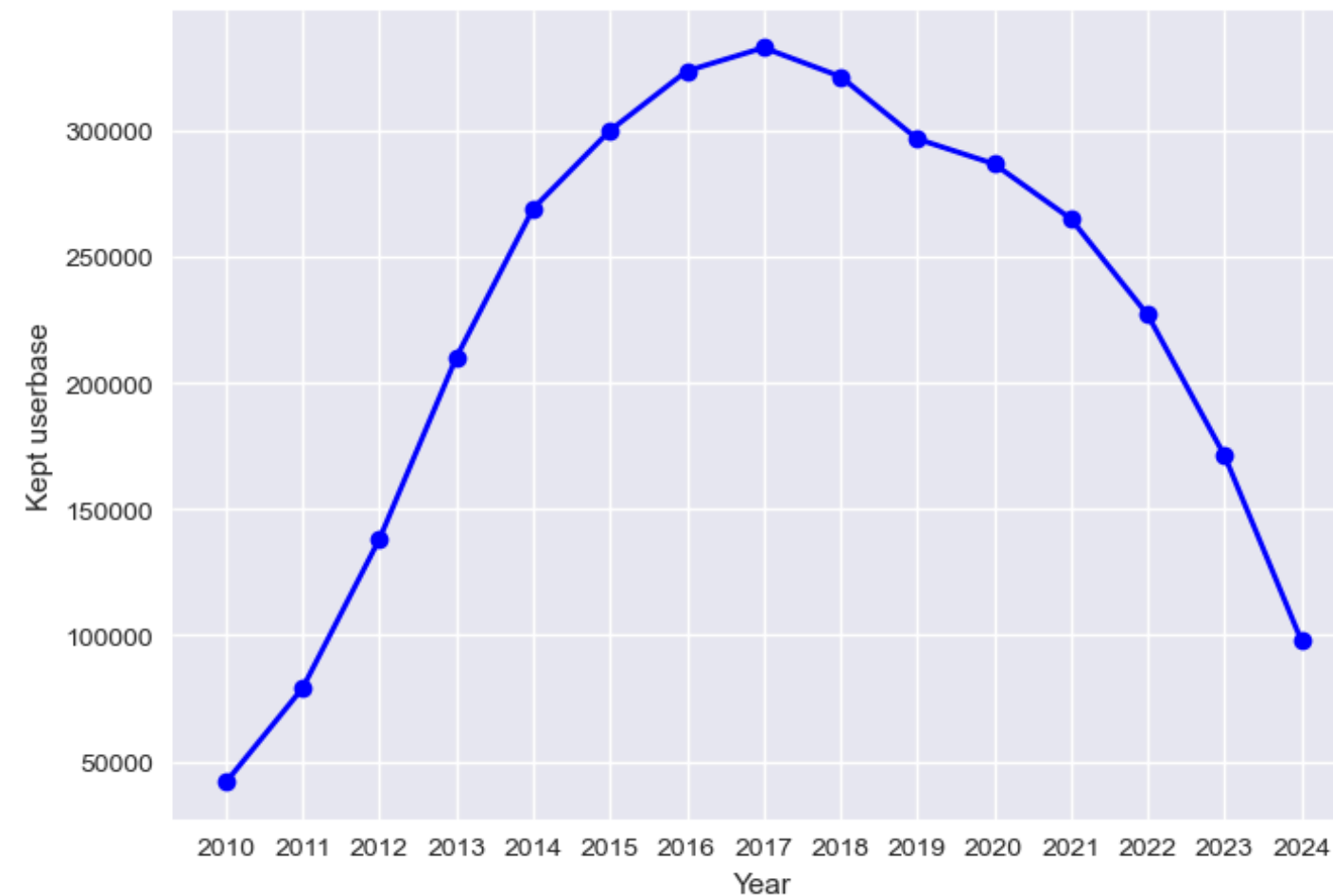
Study of user volatility

The network exhibits a very high volatility: the amounts of users joining and leaving the network every year are far superior to the amount of those that stay, as shown in the plot of the ratio of new users over users already in the network.

This is positive when the network is growing, but when the population is decreasing indicates that retention is low. Other than its instability in the last years we find no noticeable pattern that can be linked to the studied phenomenon.



New users / Old users ratio over time

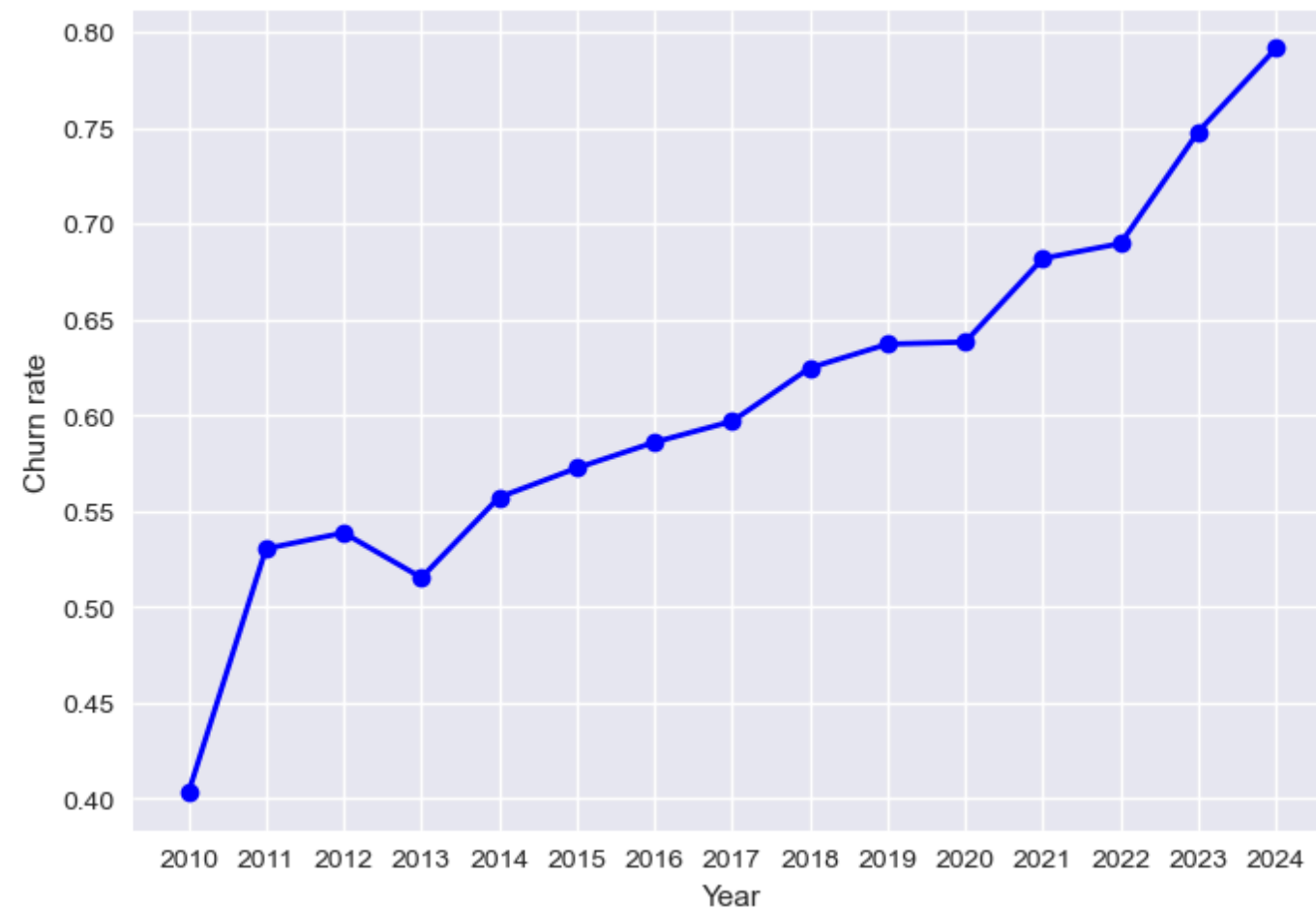


Population kept between years

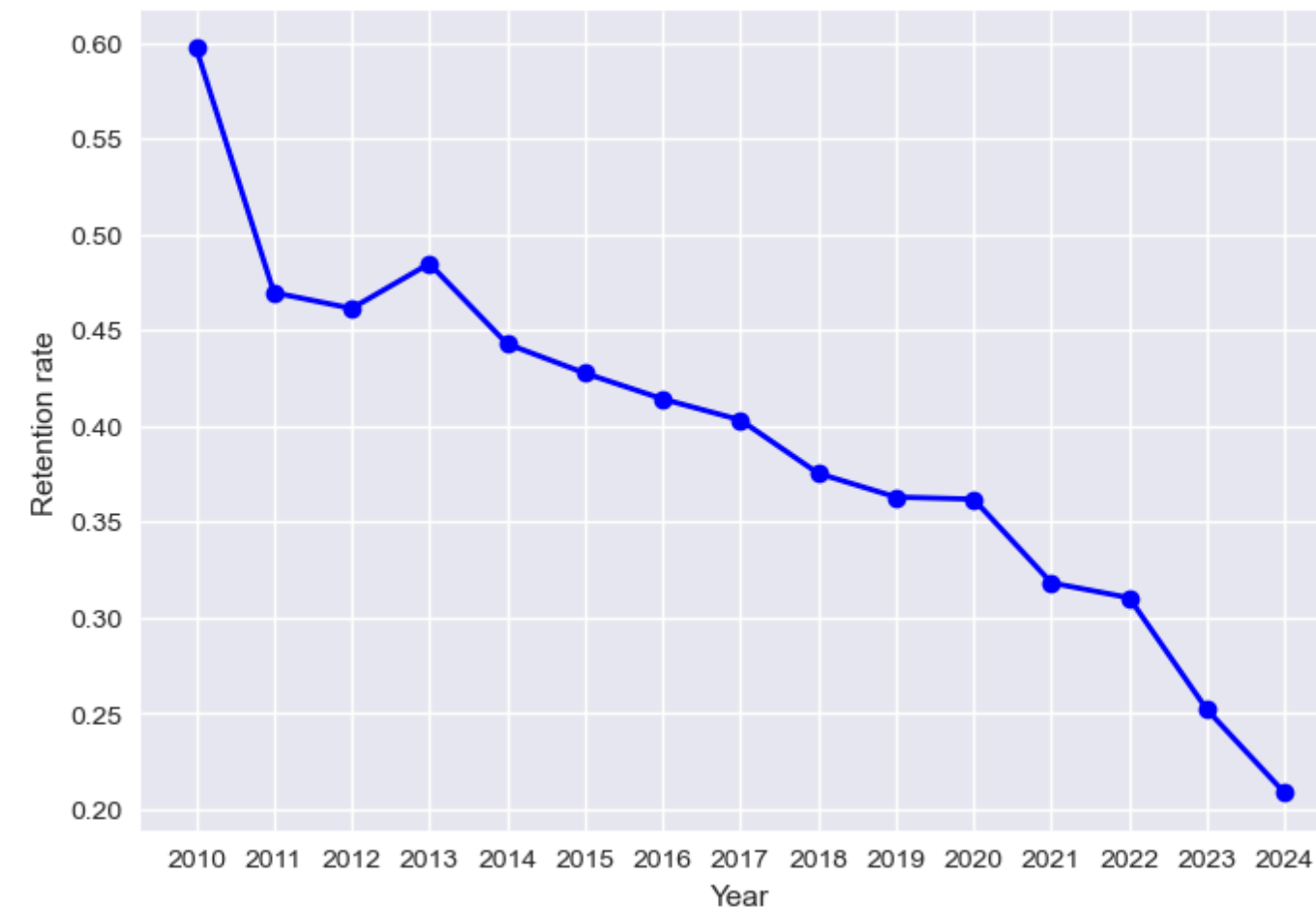
Churn and retention rates

Focussing on this aspect we study the churn rate, defined as the percentage of users who leave the network during a given time period (and the retention rate is just its complementary).

We observe the churn rate already starting high but is continually increasing over time and spiking in the last two years, going from 68% to 80% between years 2022 and 2024.



Churn rate over time



Retention rate over time

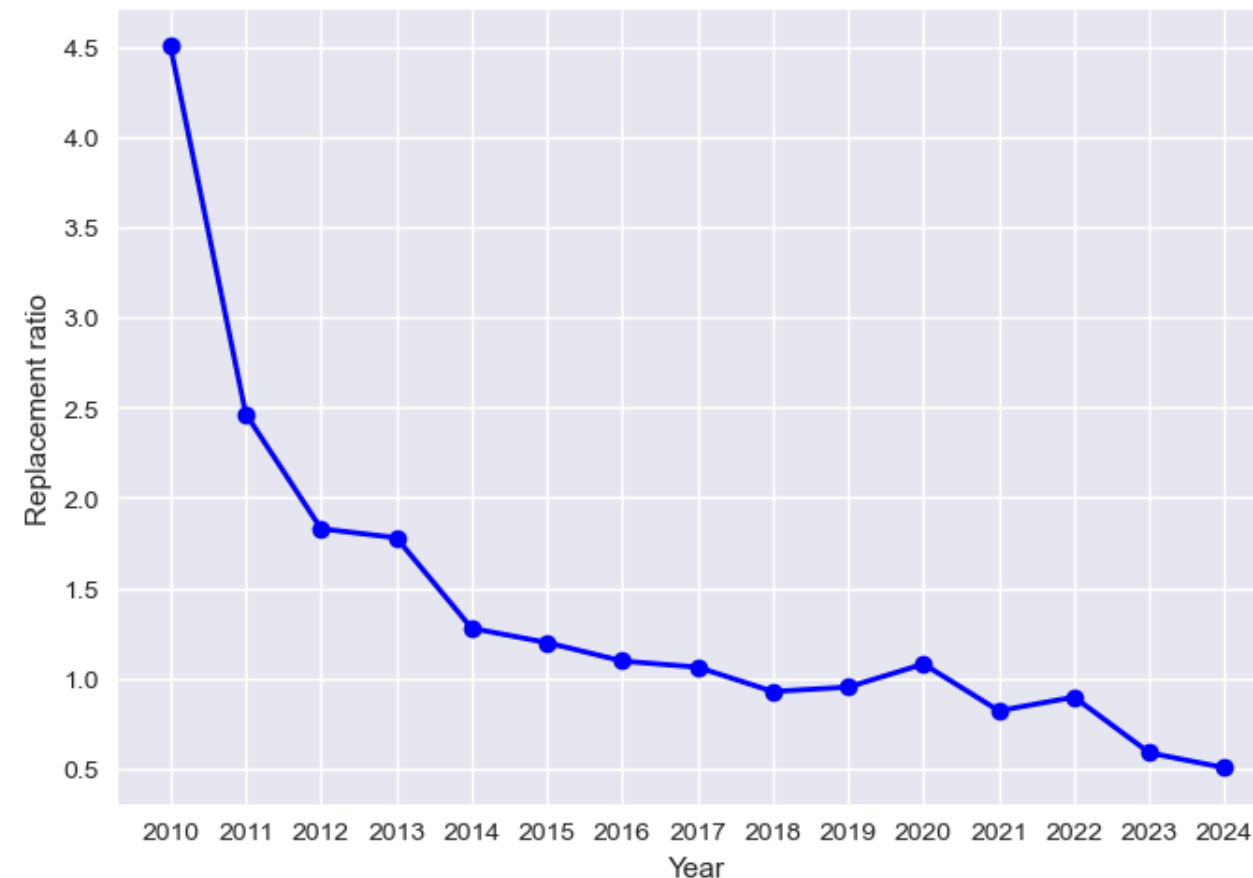
Study of the replacement ratio

Continual decrease

The replacement ratio starts with a value of 4.5 (meaning that for every user that leaves 4.5 users join) and continually decreases over time.

Since 2018 the ratio goes below 1.0, meaning that the active userbase starts to shrink.

The only exception is 2020 where it goes above 1.0 for the last time, probably due to Covid.



Replacement ratio over time

Sharper decrease

We can notice a sharper decrease between the years 2022 and 2023, with the replacement ratio going from 0.90 to 0.58 in just one year.

In conjunction with the churn rate studied before this says that not only the network lost much of its retention in the last years, but it is also failing to gather new users to replace the lost ones.

RQ2: Has the complexity and the type of questions changed? Is there a correlation between those changes?



Questions complexity

We want to study how the questions complexity evolved over time and if there is a correlation with the introduction of LLMs.

There are many ways to define the complexity of a sentence, but given the large amount of data I am handling (24 milion questions) I decided to use a very simple metric, the number of characters in the question body (word count would have conflicted with the code blocks in the questions).

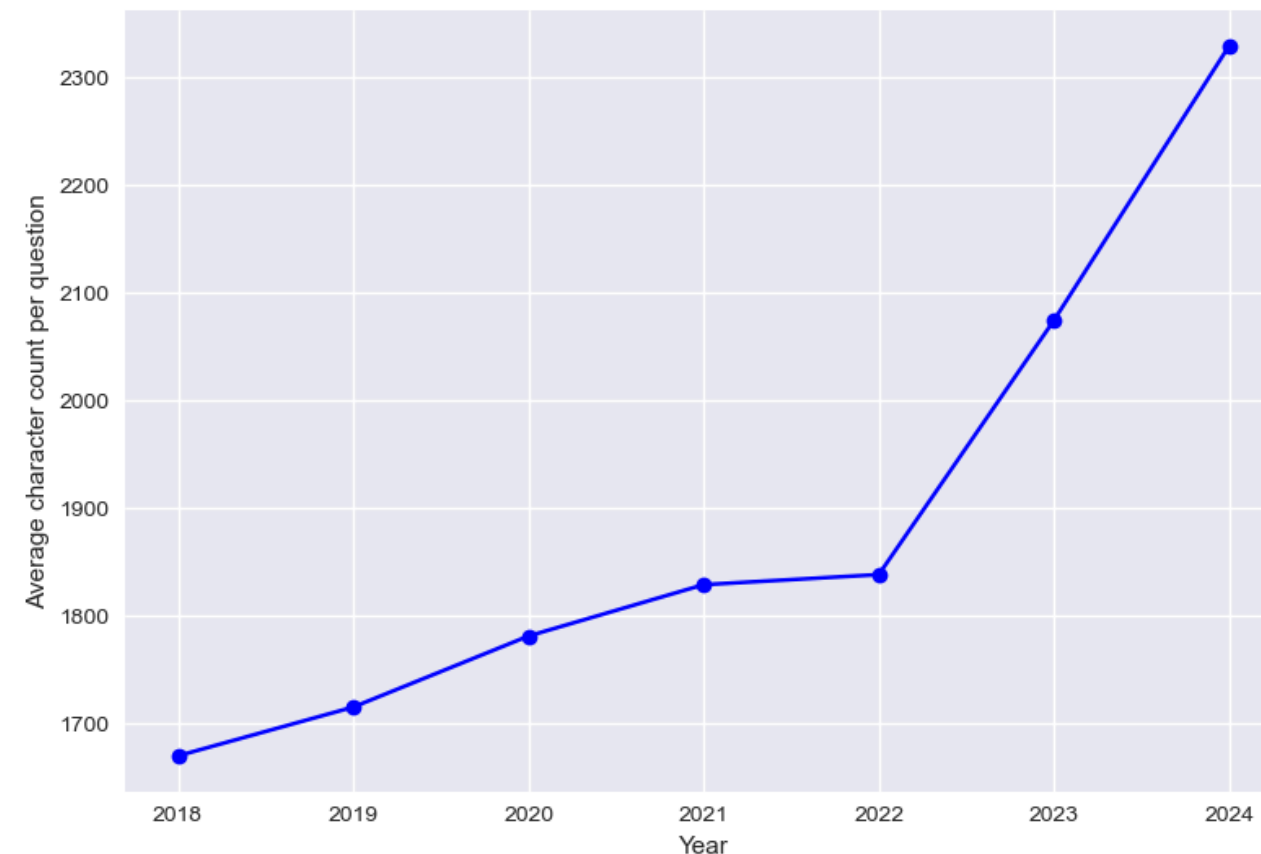
Remembering what we said before about the removal of low-effort questions I decided to consider only data from 2018 onwards.

Questions complexity

An evident effect

The plot shows a continual slow increase in question complexity over time and then an extremely evident increase from 2022 to 2023 and 2024.

In just two years the average question length increased by about 500 characters.



Average question complexity over time

Causes

For the causes of this sharp increase we can think it may be due to the use of LLMs to formulate questions, but as early as december 2022 the website banned all AI generated content.

Another reasonable conclusion could be that easier questions are being answered by LLMs and users may ask the help of other users only for questions LLMs are unable to answer.

Following this we can attribute the subsequent increments of complexity to LLMs becoming able to answer more complex questions over time.

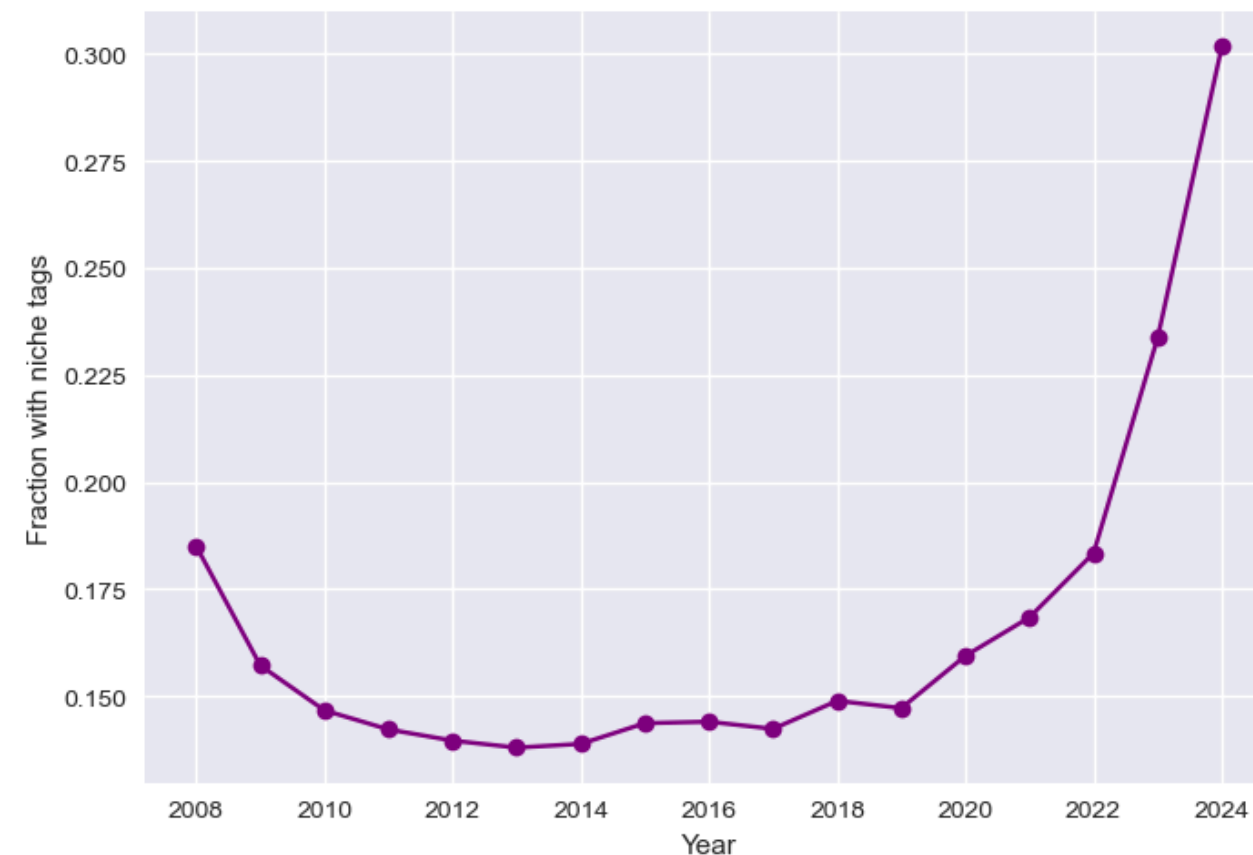
Question tags

Niche tags

Each question comes with a list of tags, which help categorize, filter, and retrieve questions efficiently.

We define a “niche tag” as a tag having less than 500 questions.

There is a total of 65.937 tags in the dataset, with 87% being considered niche.



Proportion of questions containing niche tags over time

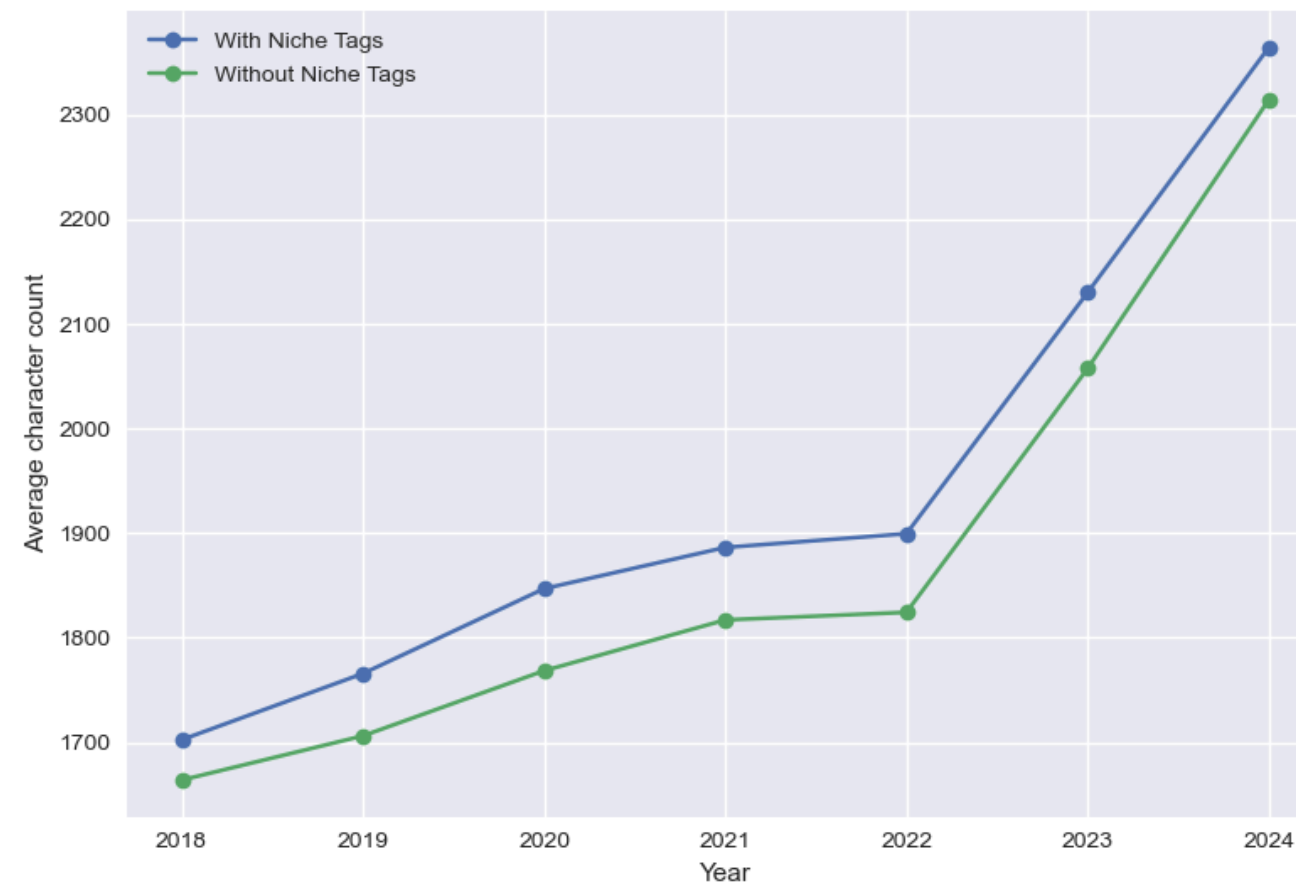
Increase in proportion

The plot shows in the last two years a sharp increase in the proportion of questions containing niche tags, despite a decrease in the amount of tags used per year.

This is reasonable following the previous conclusion as LLMs will have more difficulty answering questions about niche topics.

Niche tags and complexity

It is also interesting to notice that questions with niche tags are consistently slightly more complex than questions without niche tags.



Average character over time: niche vs non-niche questions

Conclusions

While the network decline seems to have been a phenomenon that was slowly occurring for many years (starting around 2018) we noticed in multiple instances a speed-up of this decline starting between 2022 and 2023, in conjunction with the large scale adoption of LLMs by the general public.

Moreover since 2022 the questions greatly increased both in length and in how niche is the topic, making extremely plausible the conclusion that users may first ask their questions to LLMs and if not satisfied with the answer will turn to Stack Overflow for help.

Therefore we can conclude saying that while not being the only factor at play, it is reasonable to think that LLMs might not only have accelerated the decline of Stack Overflow from a point of view of traffic and userbase, but also deeply changed how users approach the website.

The background features four abstract, organic shapes in shades of purple and blue, positioned in the corners of the slide. These shapes have soft, blurred edges and a gradient of colors, ranging from deep purple to a lighter blue. They appear to be floating or emerging from the corners, framing the central text.

Thanks for the attention

References and Data Sources

Stack Exchange Data Dump 2024-12-31 - https://archive.org/details/stackexchange_20241231

The Fall of Stack Overflow - <https://observablehq.com/@ayhanfuat/the-fall-of-stack-overflow>

Stack overflow is almost dead - <https://newsletter.pragmaticengineer.com/p/the-pulse-134>

Are LLMs making StackOverflow irrelevant? - <https://newsletter.pragmaticengineer.com/p/are-llms-making-stackoverflow-irrelevant>

Did ChatGPT Just Kill Stack Overflow? - <https://analyticsindiamag.com/ai-features/did-chatgpt-just-kill-stack-overflow/>

The Consequences of Generative AI for Online Knowledge Communities - <https://www.nature.com/articles/s41598-024-61221-0>