Instagram graph theory report

Mattia Manna
Paolo Zilviano

March 18, 2025

Contents

1 Introduction

Today we are gonna talk about an analysis of the Instagram network.
To show our work we made a presentation divided in 8 points.
First we are going to talk about our topic of interest and why we chose it, then we briefly describe the research questions, how we collected the data and how we preprocessed them.
Next, with the last 4, the analysis performed will be illustrated.

1.1 Why the social media?

Why we choose to study social media? Nowadays social media is a cross-section of our community; we all use it, to interact with others or using the content offered. The great importance that social media gained in people's lives has meant that over the years they became the subject of studies. To have an idea of their importance let's think of some tweets that influenced the stock market, for instance elon musk's tweets.
As the literature suggest social network are:

- very sparse (few links compared to the maximum that might exist)
- scale free
- there is an ultra small world effect (users on average separated by only 3 profiles)
- there are hubs (few users that holds most of the links)

We wanted to investigate how a network of people is organized, for instance how many people usually a user follows, with how many people a user interact or understand what are the most connected users.
In order to answers to those questions we chose the instagram network.
Instagram has been chosen due to the way it is constructed, for instance allows to gather info on users that aren't friend of yours.

2 Research questions

In this analysis we're not interest in the use and implication of the social media, we have interest just in how they are structured. In this project what we want to do is find an empirical confirmation to what the literature said. In particular we want to find an answers to this questions.

- Does the ultra small world effect exist in the instagram network?
- Is Instagram a scale free network?
- Is it sparse network?
- Are there hubs? Who are they?
- Is the dunbar number respected?

3 Data collection and preprocessing

3.1 Data collection

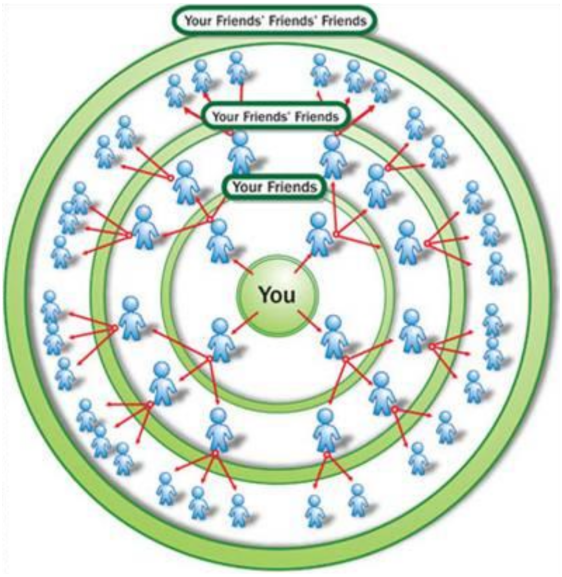
To gather information about the instagram network, we used web scraping.

Initially we tried by building a selenium scraper, although since it was being blocked to much we used some chrome's extensions created to extract info about instagram accounts and use them through a selenium bot to automize the process.
Extracting all the data related to instagram users was certainly not possible. What we did, therefore, was to try to take a portion of this network. Through 3 steps it was possible to build a subgraph that made sense. First we extracted my followings, then...

1. **Tier1:** my followings
2. **Tier2:** following of my followings
3. **Tier3:** following of the followings of my followings

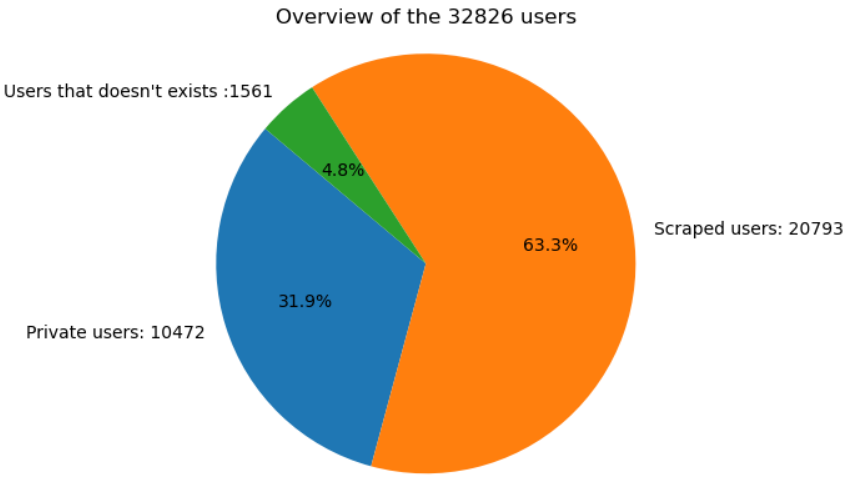
It is important to undestand that from each users we were extracting only the followings, extract the followers was not possible, influencers could reach millions of followers, and to take all of them an excessive amount of time was required.

Extracting other than followings (like posts) wasn't possible , in fact just by taking only following we spend months in the scraping process.
Of course this is a big limitation for the analysis, but still for our research questions was enough.
The subgraph that we extracted has a naive representation like this.



Of course this is a figure of a small world network, we show this just to give an idea of how we structured the tiers for the extraction.

At the end of the extraction process we managed to scrape 32826 users. Among them, we were able to acquire information of about 20k accounts, while 10k were private accounts and 1k were errors (e.g. the user change the name of the account or deleted the profile) so of those we couldn't gather info.



3.2 Data preprocessing and network construction

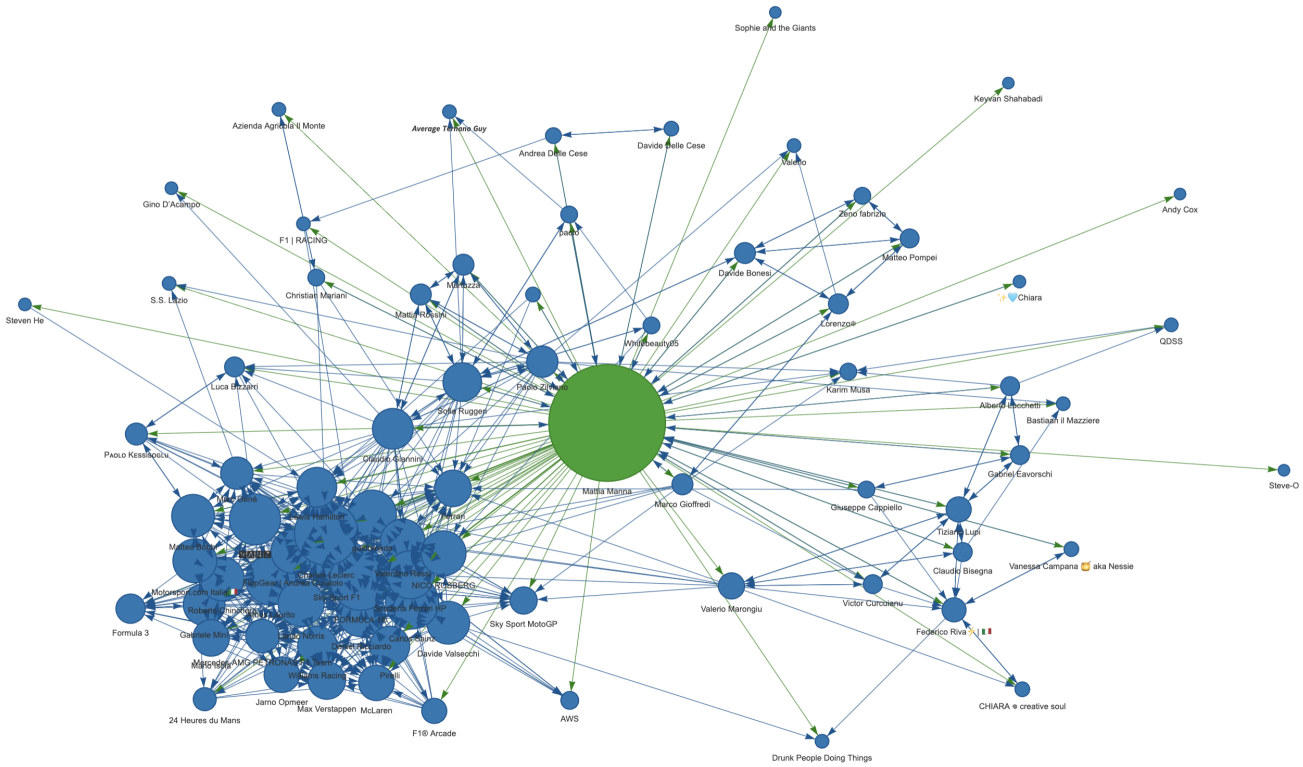
To get a list of the users (nodes) and their following (links) of our network we followed this pipeline:

1. We collected every csv file given by the exporting tools
2. Then through a python script we created 2 datasets, one for the nodes (have list of the users) and one for the directed links (to register who follow who).
3. After we cleaned those two dataset removing useless links and nodes, for instance duplicates and nodes that are outside the 3rd tier.
4. Then importing those two datasets in R we build the subgraph of instagram.

In the end we build a graph that is made of 19992 users (nodes) and 1094224 links.

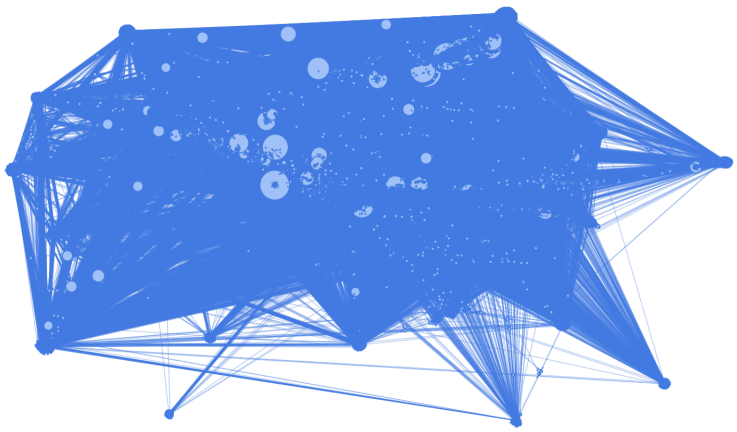
3.3 Partial graph

Microscopically, the graph will have a structure like the following.



3.4 Resulting graph

Looking at the whole graph, it have an almost incomprehensible structure given the large amount of users and links. As we will see this final graph is fully connected and very sparse.



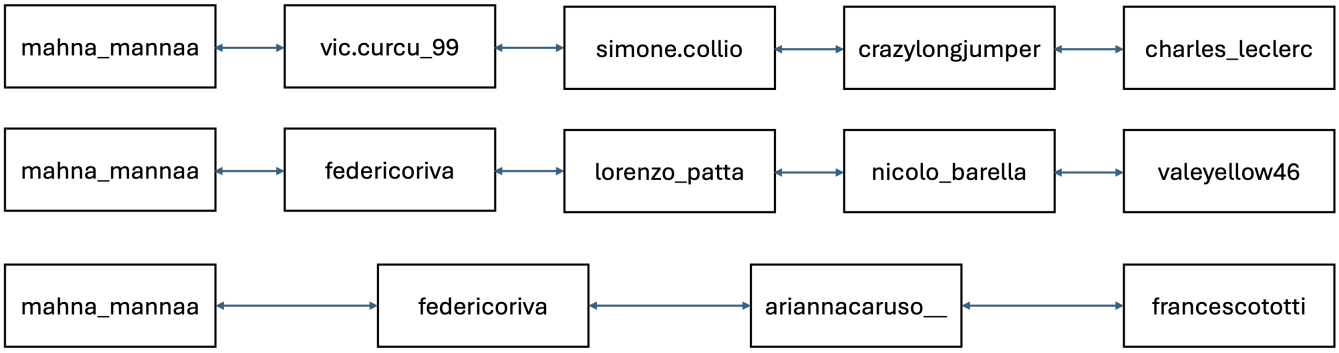
4 Ultra small world effect

It is known as "small world effect" the phenomenon for which two random people in the world are separated by six other people, also known as *six degrees of separation*.
In the social media, scale free networks, there are only 3 degree of separation and we call it ultra small world effect.
In order to find out if there is the ultra small world effect in our network we look at the average path length. The average path length is the average number of links in the shortest path between any two nodes in the graph.

What has emerged in this analysis is that the average path length is equal to 3.75 (that is close to the degrees of separation of a ultra small world effect). So even if is not precisely equal to 3, this is still a good indicator of ultra small world situation considering that a lot of information is missing.

To get a true feedback of the *ultra small world effect* let's see how far my instagram profile is from:

- charles leclerc
- valentino rossi
- francesco totti



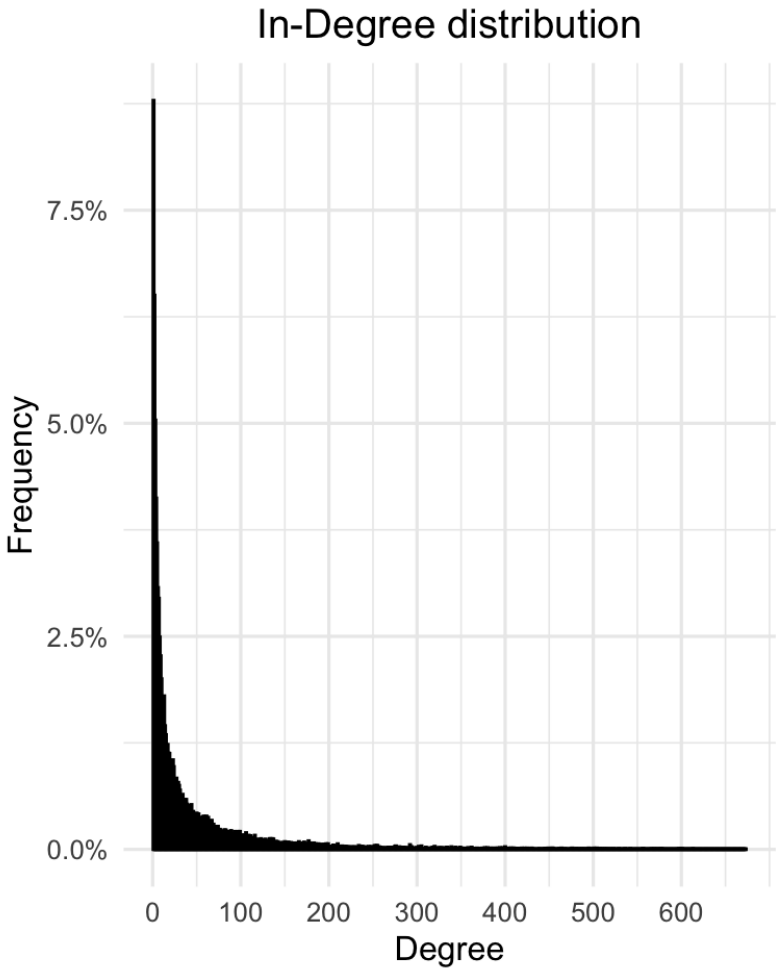
It seem that my instagram profile is separated from charles leclerc and valentino rossi by 3 people and only 2 from francesco totti.

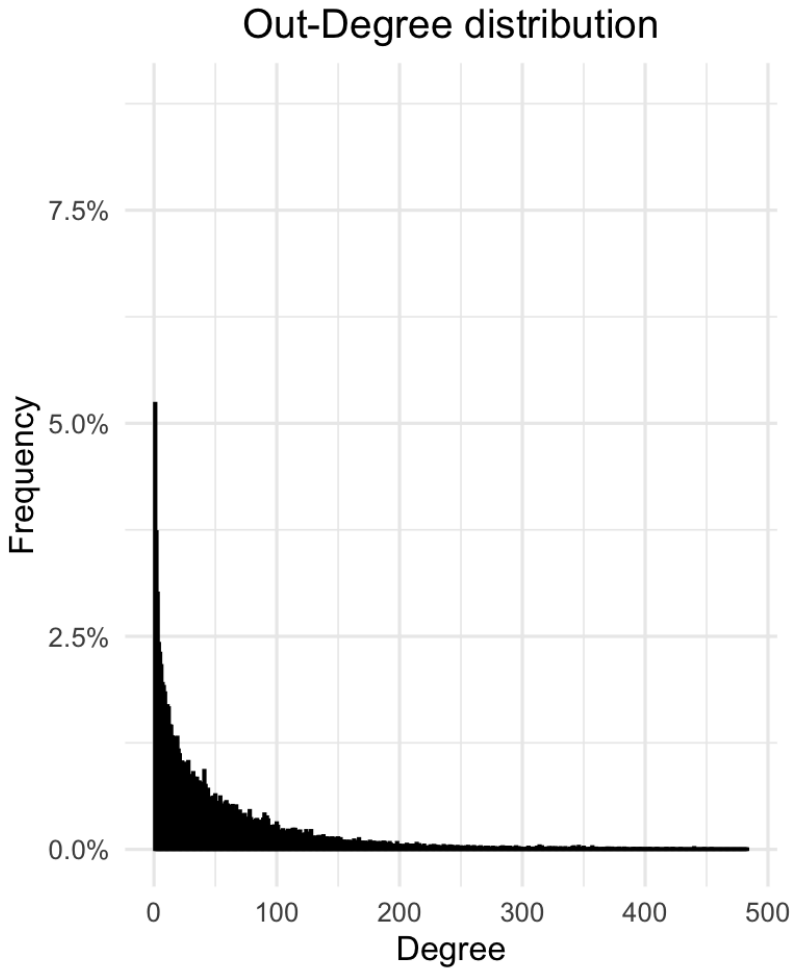
5 Degree distribution

The degree distribution helps to find out which type of network we have.
In order to be more precise we considered the *in degree* distribution and the *out degree* distribution.
In degree distribution take in consideration the links directed towards that node, so in the instagram case only the number of people that follow someone.

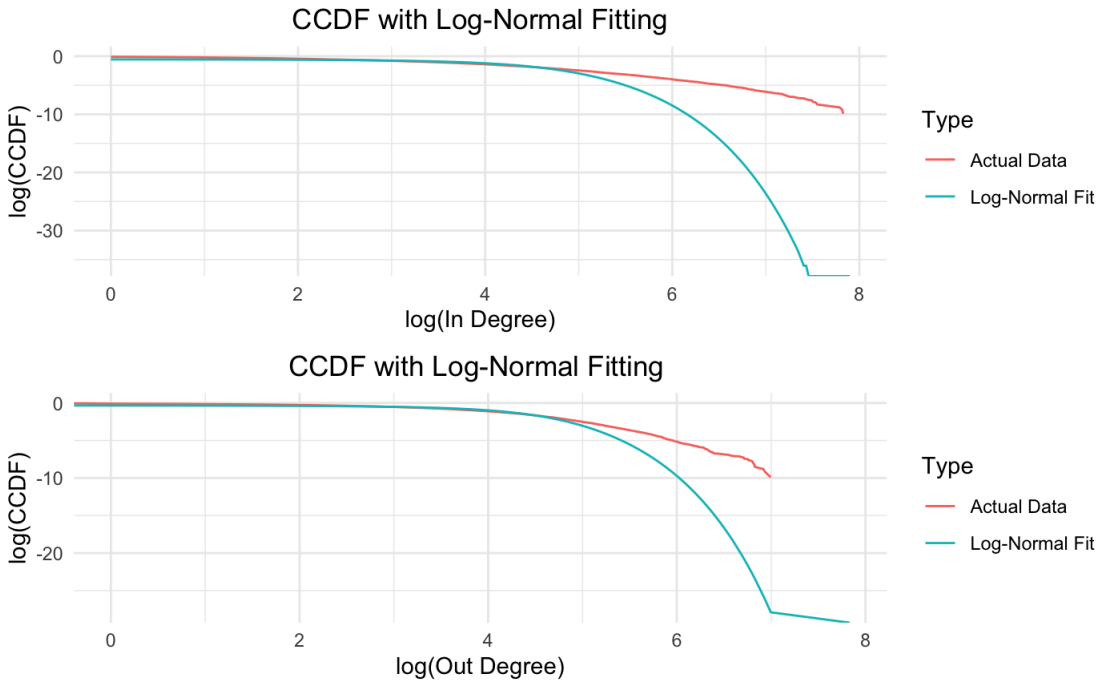
In degree distribution take in consideration the links directed outward from each node rather than inward, so in the instagram case only the number of people that someone is following .

By looking at the resulting histogram they seem to follow a scale free distribution.
Infact we can see that there are a lot of nodes with low degrees and fews highly connected nodes.
That means that a lot of users are followed or follow few people, and few users are followed and follow a lot of people.





However to confirm that the network is a scale free we need to check the ccdf (complementary cumulative distribution function).
By looking at the following charts it is clear that the network doesn't behave like a scale free network, instead it seem more to follow a log normal distribution.



So in the end we can say that NO, the network is not scale free.

6 Hubs

Despite the network is not a scale free hubs can still be present.
To find out the hubs we decided to select the top 5% of users based on the InDegree (incoming links) .
We decided to select the hubs based on the InDegree because an hub in a social media is a profile that is followed by many users, and not a user that follows a lot of others profiles.
By taking the 5% of users in a graph that count nearly 20k of nodes we ended up with a thousand of users identified as hubs.
Those users have the 41 % of ingoing links of the total graph.
The most important hubs are the following (top 30):

Hubs

User	InDegree	User	InDegree	User	InDegree
f1	2689	natgeo	1797	fernandoalo_oficial	1401
charles_leclerc	2508	danielricciardo	1733	kendalljenner	1384
lewishamilton	2482	landonorris	1688	redbullracing	1371
cristiano	2412	nasa	1678	ferrari	1367
chiaraferragni	2102	carlossainz55	1645	433	1349
scuderiaferrari	1900	willsmith	1549	motogp	1346
instagram	1886	dualipa	1531	iamzlatanibrahimovic	1328
valeyellow46	1868	badgalriri	1493	mercedesamgf1	1325
leomessi	1811	champagnepapi	1490	mclaren	1321
maxverstappen1	1800	kingjames	1407	belenrodriguezreal	1312

It's not a surprise that the hubs are profiles belonging to vip or very important corporations.
Infact we can see the official profile of the f1 as the instagram page that as more follows.
Here we can notices a bias due to the extraction process, since it started from my profile and I follow a lot of motorsport personalities the hubs seem to be a lot related to the motorsport environment.

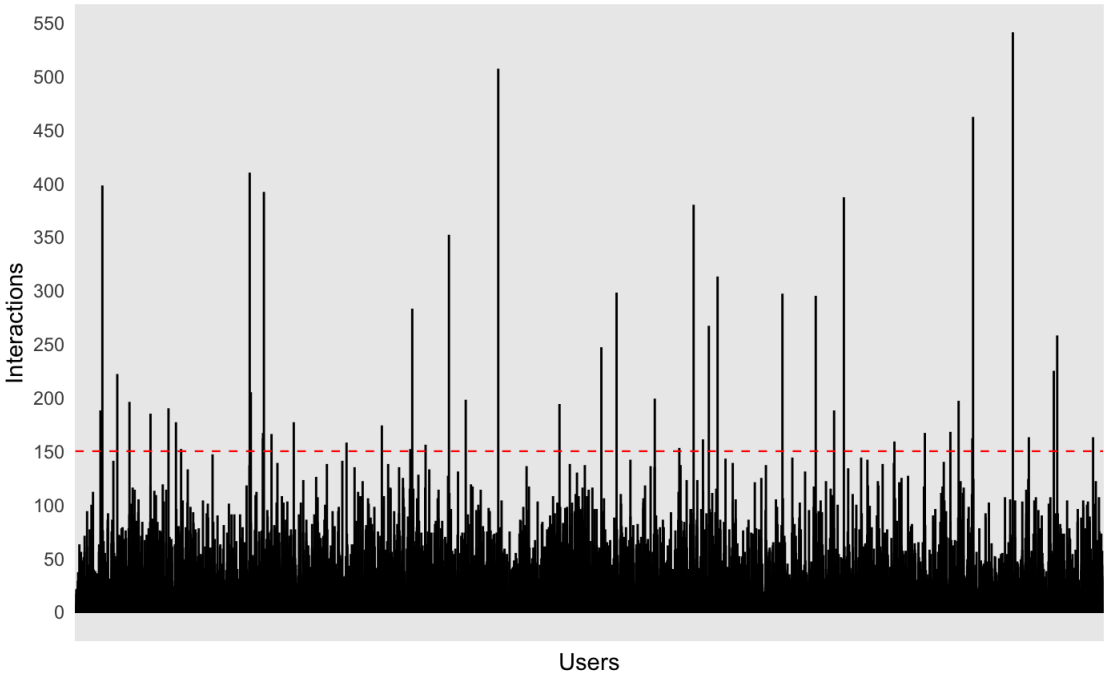
7 Dunbar number

The dunbar number is an upper bound for the human interactions.
It states that people do not interact with more than 150 humans on every day life, it can be interesting to find out if the dunbar number is respected on the social media.

To find the dunbar number we have to look to users interactions. A naive but effective method could have been to monitor users through an observation of their interactions within the posts and find out the level of interaction between each other.

Since it was not possible in our condition, we just consider as interaction when two users follow each other.

In the following chart we can see on the x-axis the users and on the y-axis how many interactions each user have. The red line is the dunbar number.



So is the dunbar number respected?
Yes, it is respected.
In fact only 47 people among the 19992 that are part of the graph (0.23 %) exceed the upper bound.
This number is of course a bias due to the way in which we defined the interactions and also we have to think that outliers exists in the reality (es. manager, journalist, etc...).

8 Conclusions

8.1 Other graph's informations

By looking at the density we can say that the graph is sparse.
Also the graph is connected, that means that all pairs of users in the graph are connected.
Pratically each user is connected by another by a path.
Diameter is equal to 10 mean that the longest path is 10 and so the most distant users are separated by 10 people.

```
[1] "Density: 0.00273788671088464"
[1] "Is the graph connected?: TRUE"
[1] "Average degree <k>: 109.466186474602"
[1] "Average path lenght: 3.75397179553626"
[1] "Diameter: 10"
```

8.2 Final notes

- **Incomplete Data:** The final graph that was used for the analysis is missing a small part of the whole collective, despite we scraped for 4 months. This of course is an issue, but we didn't have the resources to scrape 24/7. Maybe without this problem the analysis would have been more reliable.
  Anyway we thought that this kind of project could have been more interesting than using a Kaggle dataset.
- **Interactions During Collection:** It wasn't possible to build the graph in a freezed timespan. So possible interactions could have happen during the scraping but, again, we didn't have the tool to operate this way. However the instagram network has been tangible for almost a decade, so its structure and the one of its subgraphs shouldn't have changed too much in few months.
- For these reasons, the results of our analysis will no longer be related to the entire instagram network but to its subgraph

8.3 Conclusions on the instagram subgraph

1. Does the ultra small world effect exist? Yes
2. Is Instagram subgraph a scale free network? No
3. Is it a sparse network? Yes
4. Are there hubs? Yes
5. Is the dunbar number respected? Yes