Andrea Paciolla (matr. 874512)

Metodi & applicazioni per social network

Presentazione progetto d‘esame a.a. 2017-2018


500px

500px

In breve

Community & marketplace allo stesso tempo, 500px offre visibilità a fotografi neofiti e professionisti, permette di trovare la giusta porta verso l‘ispirazione e mette in contatto fotografi da ogni parte del mondo.

Un „luogo“ importante

per appassionati e professionisti.

Nato in Canada

OGGI PRESENTE
IN TUTTO IL MONDO

Fondatori

Oleg Gutsol

Imprenditore canadese, di origini ucraine, nato il 3 Febbraio 1982, è oggi un co-founder di 500px.

Evgeny Tchebotarev

Originario di Mosca, creò la prima community di 500px ed ora ne è CPO e CFO.

Gli albori

500 pixel era la dimensione considerata ottimale per essere mostrata sul web e quindi impostato come limite sulle foto postate dalla community per le review.

2003

500px: un hobby durante gli studi

2003

Evgeny gettò le basi di 500px sulla piattaforma di social blogging LiveJournal per hobby.
Le foto erano moderate: da sempre l'obiettivo è stato quello di promuovere gli scatti migliori.

500px

Il lancio del primo 500px

2009

Tchebotarev si unisce a Gutsol

Primi mesi del 2009 – Novembre 2012

Si inizia a lavorare su una versione automatizzata di 500px.
Il sito ufficiale di 500px viene lanciato nell'Ottobre 2009.

Mille utenti nel 2009, divennero più di 1.500.000 nel Novembre 2012

500px

Visual China Group

Entra come partner strategico e offre un finanziamento pari a 13 milioni di dollari

2015

L‘ascesa di 500px

Gennaio 2015 – Novembre 2016

13 Milioni di dollari come finanziamento permisero non solo di entrare sul mercato con la nuova app iOS ma anche di far entrare come partner strategico Visual China Group.

A Settembre, Google annuncia 500px come partner fotografico principale esterno per Chromecast: più di 20 milioni di utenti possono avere tutti gli scatti su 500px.com sui loro tv.

In Aprile 2016 vengono lanciati i “verified accounts”: Red Bull primo a partecipare.

Novembre 2016: viene lanciato 500px Studio. Scatti custom e a richiesta, a disposizione per i brand.

2018

L‘Acquisizione

Gennaio 2017 – Presente

500px

Acquisizione Visual China Group

Il 26 Febbraio 2018, Visual China Group acquisisce 500px dopo che, nell’Agosto 2017, 500px annuncia il support alle immagini ad altissima risoluzione.

500px

Oltre le “semplici” foto

Non solo un hosting di foto.
500px facilita il contatto tra fotografi e clienti,
permettendo ai clienti stessi l’acquisto di interi photobook.

PULSE

L’algoritmo di 500px prende in considerazione view, likes e commenti e genera un rating chiamato **pulse**.

Più alto è questo parametro, più probabilità si hanno di entrare nella **popular page**, ottenendo milioni di visualizzazioni.

L’algoritmo permette a tutti di ottenere visibilità, anche agli utenti con ancora poca partecipazione.

AFFECTION

Ogni utente ha un parametro complessivo di rate chiamato **affection**.

Il parametro tiene conto dei **like** e salvataggio nei preferiti che tutte le foto hanno ricevuto. E’ un indicatore di quanto è popolare un membro all’interno della community. Potrebbe essere usato per individuare gli **hub**.

Obiettivi

Perchè studiarlo e quali dati gestire

Qualità

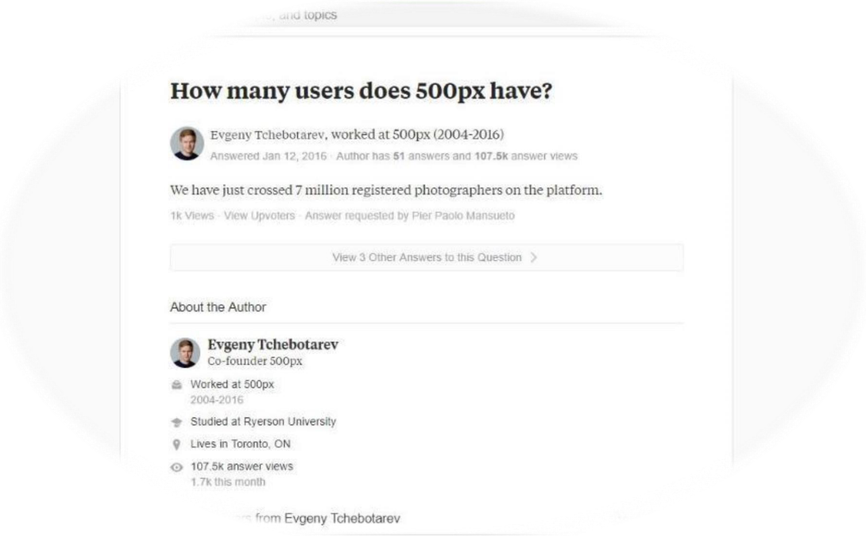
500px, al contrario di altre community fotografiche ha una qualità fuori dal comune che ha attirato, a Gennaio 2018, **13 milioni di utenti.**

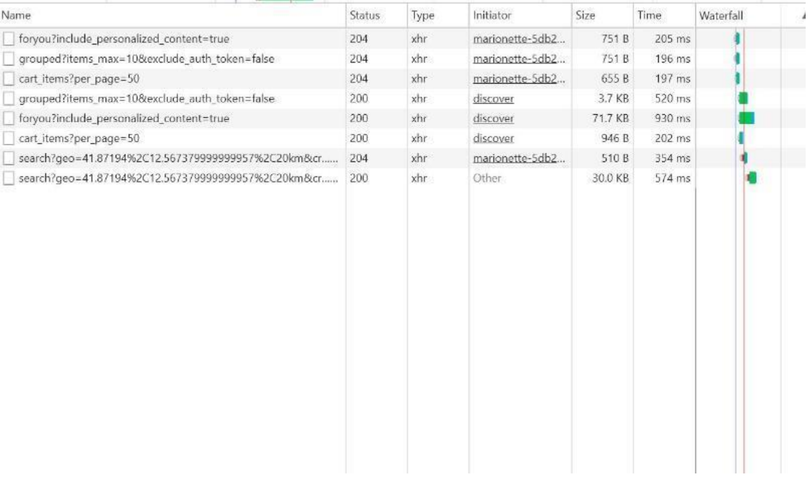
Domande di ricerca

1. 500px è come dicono tutti un social network?
2. Rispetta il modello scale free?
3. Chi sono gli utenti più importanti?

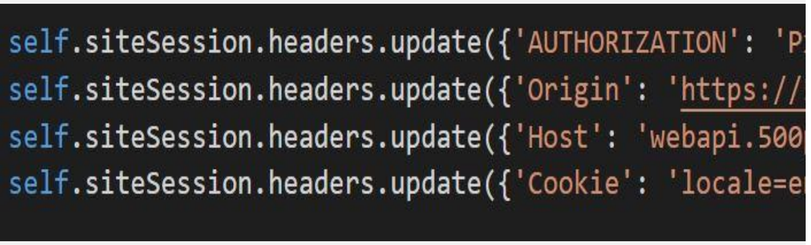
Dati

Riesco ad estrarre dati relativi agli utenti attraverso chiamate ad API scoperte osservando la network console del browser durante la navigazione sul sito.

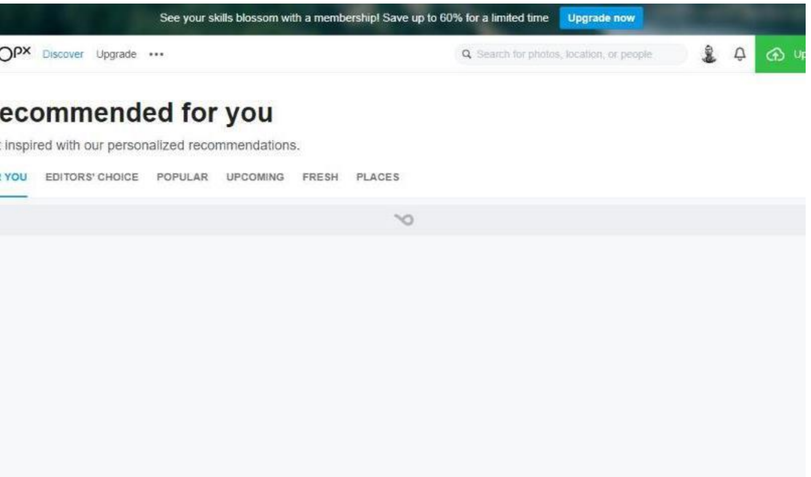




Name	Status	Type	Initiator	Size	Time	Waterfall
foryou?include_personalized_content=true	204	xhr	marionette-5db2...	751 B	205 ms	
grouped?items_max=10&exclude_auth_token=false	204	xhr	marionette-5db2...	751 B	196 ms	
cart_items?per_page=50	204	xhr	marionette-5db2...	655 B	197 ms	
grouped?items_max=10&exclude_auth_token=false	200	xhr	discover	3.7 KB	520 ms	
foryou?include_personalized_content=true	200	xhr	discover	71.7 KB	930 ms	
cart_items?per_page=50	200	xhr	discover	946 B	202 ms	
search?geo=41.87194%2C12.567379999999957%2C20km&cr.....	204	xhr	marionette-5db2...	510 B	354 ms	
search?geo=41.87194%2C12.567379999999957%2C20km&cr.....	200	xhr	Other	30.0 KB	574 ms	



```
self.siteSession.headers.update({'AUTHORIZATION': 'P
self.siteSession.headers.update({'Origin': 'https://
self.siteSession.headers.update({'Host': 'webapi.500
self.siteSession.headers.update({'Cookie': 'locale=e
```



Modalità di estrazione delle informazioni

Punti critici

Nell‘analisi di questo social media è stato necessario dover tenere conto di alcune particolarità.

ASSENTE documentazione API

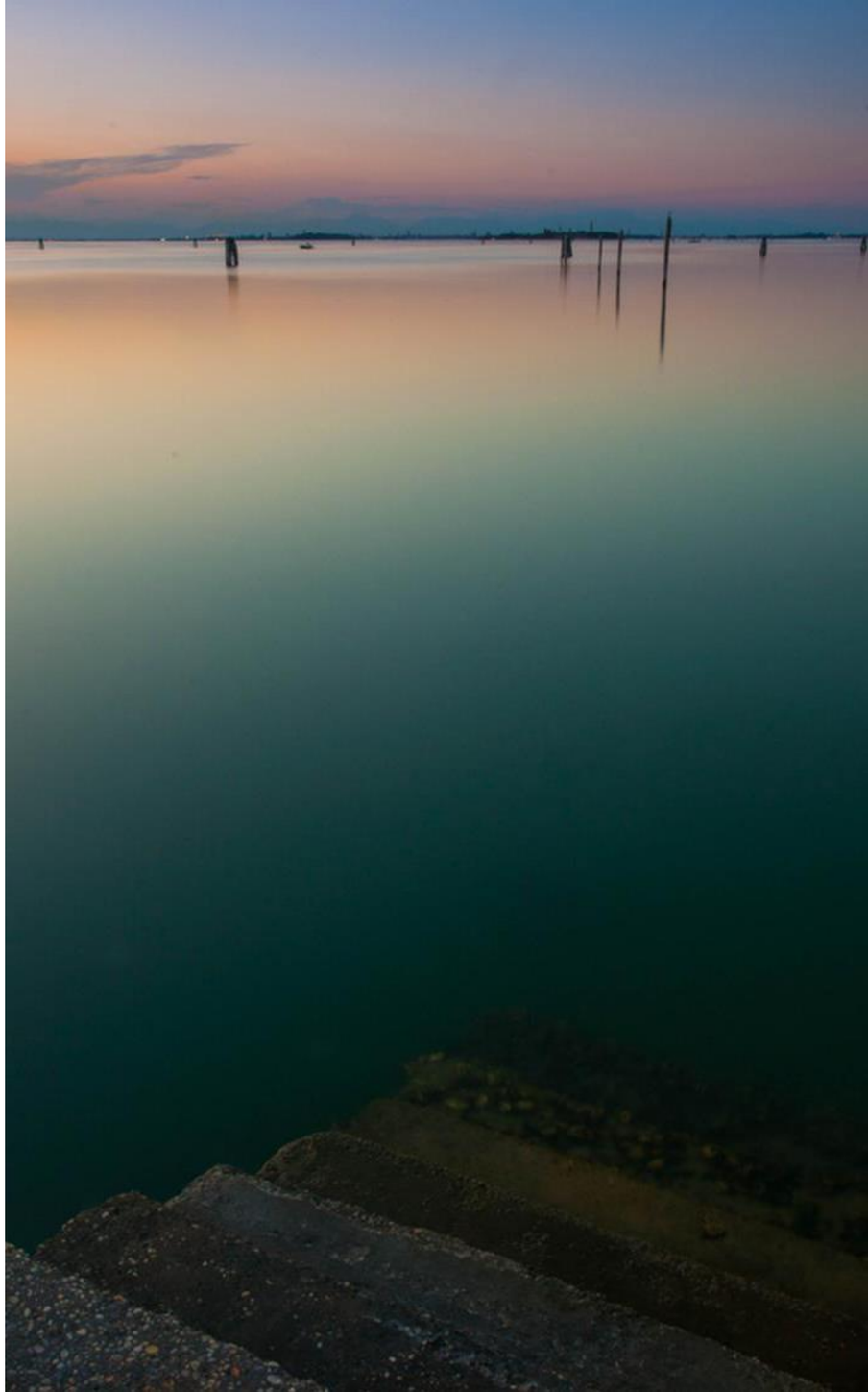
Trovati “manualmente”, attraverso la network console di Chrome, gli endpoint che potevano tornare utili per lo scopo del progetto.

PROTEZIONE API

Ad ogni chiamata tutti gli HTTP request header sono riscritti in quanto controllati e validati dai server di 500px

ASYNC content

Il contenuto viene renderizzato da script client Javascript: impossibilità di adottare tecniche di scraping del sorgente.
Il data set è costruito chiamando iterativamente le API di follower e users.



Come avviene la fase di

Auth

L’autenticazione avviene chiamando l’endpoint di login con I dati di autenticazione del mio account personale. A questi aggiungo una property “authentication_token”, richiesta dall’interfaccia del servizio chiamato.

Una volta autenticato richiamo le API users e followers per la costruzione del dataset.

```
29      def doLogin(self):
30          try:
31              print('-> Attempting login...')
32
33              self.siteSession.headers.update({'Origin': 'https://500px.com'})
34              self.siteSession.headers.update({'Host': 'api.500px.com'})
35              self.siteSession.headers.update({'X-CSRF-Token': 'siu+B2gEwPxM4ZOwmqY9iSmDIWu/aHFrnrmFQM8qvJ3IE/C1/
36              self.siteSession.headers.update({'Cookie': 'device_uuid=1ecb2213-61aa-4521-aa8a-5e4bdeb6f2e4; local
37
38              response_login_request = self.siteSession.post(
39                  self['pages']['login'],
40                  data=self.loginPayload,
41                  allow_redirects=False
42              )
43              # Print the cookies just got
44              print('-> Cookies got {}'.format(self.siteSession.cookies))
45
46              # Prepare another request to get the homepage and check if we've logged in correctly
47              print('-> Retrieving /profile page... ')
48
49
50              self.siteSession.headers.update({'Origin': 'https://500px.com'})
51              self.siteSession.headers.update({'Host': '500px.com'})
52              responseObjHomeURL = self.siteSession.get(self['pages']['profile'])
53
54              if responseObjHomeURL.text.find(self.loginSuccessAttr) > 0:
55                  return True
56              return False
57
58          except requests.exceptions.RequestException as e:
59              print(e)
60
```

Analisi 500px.com

AGENDA

Analisi del network

Statistiche di base

Sono stati raccolti dati per un totale di circa 30.000 utenti.

L‘analisi delle statistiche di base mette in evidenza una bassa moda e mediana, un basso average degree (*limitato inferiormente dal valore 0 e superiormente dal valore 502*) ed una varianza molto alta.

Questi parametri ci danno modo di pensare ad una degree distribution che segue una power law.

47.000
Archi

27.495
Nodi

850.210
Varianza

0
Grado minimo

502
Grado massimo

3.42
Grado medio

1.0
Mediana

1.0
Moda

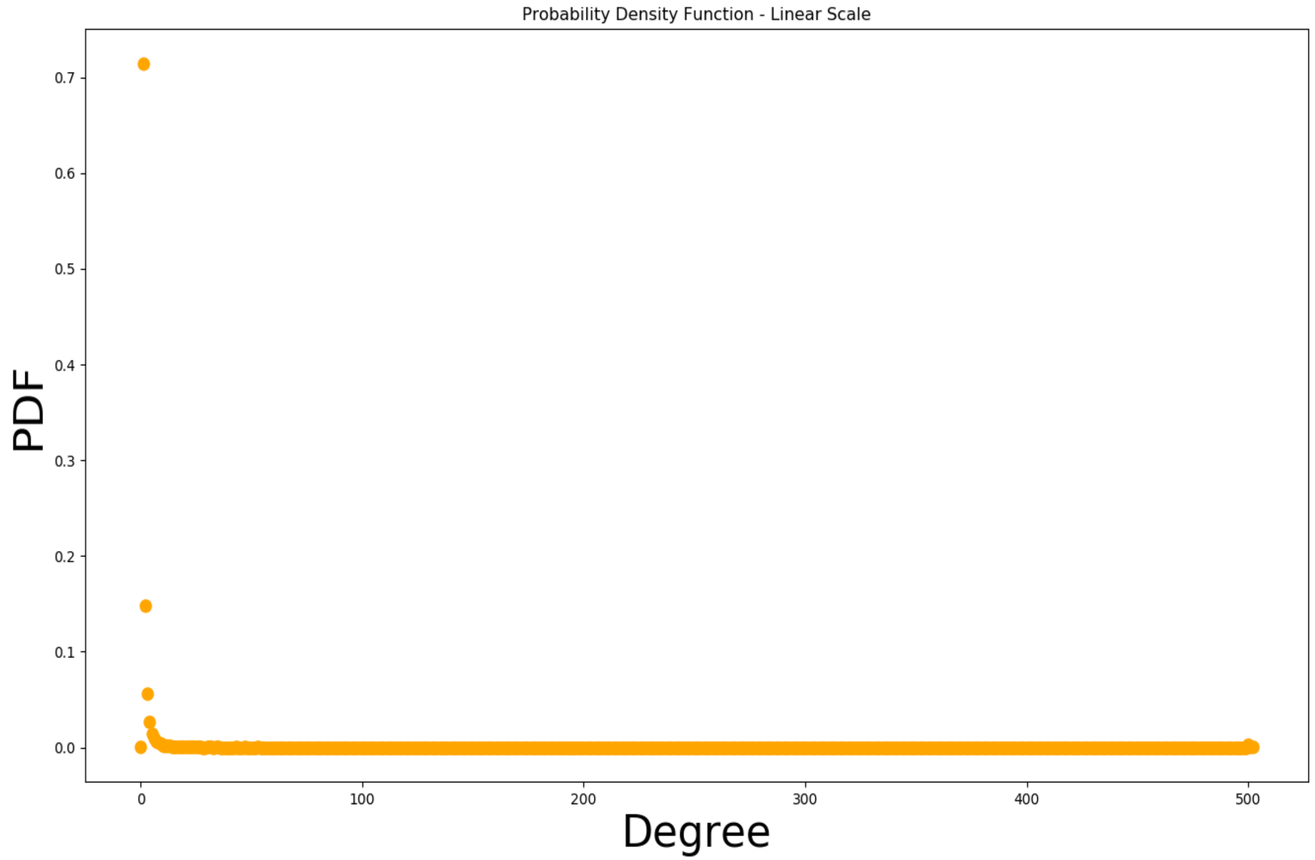
Distribuzione del grado

PDF Lineare

Accanto è possibile notare come la Probabilità p_k si distribuisca seguendo la curva caratteristica delle real network.

Il grafico di fianco non distingue tra in-degree ed out-degree.

Il grado Massimo osservato è 502, coerente con quanto si osserva sull'asse x.



Distribuzione del grado

PDF LogLog

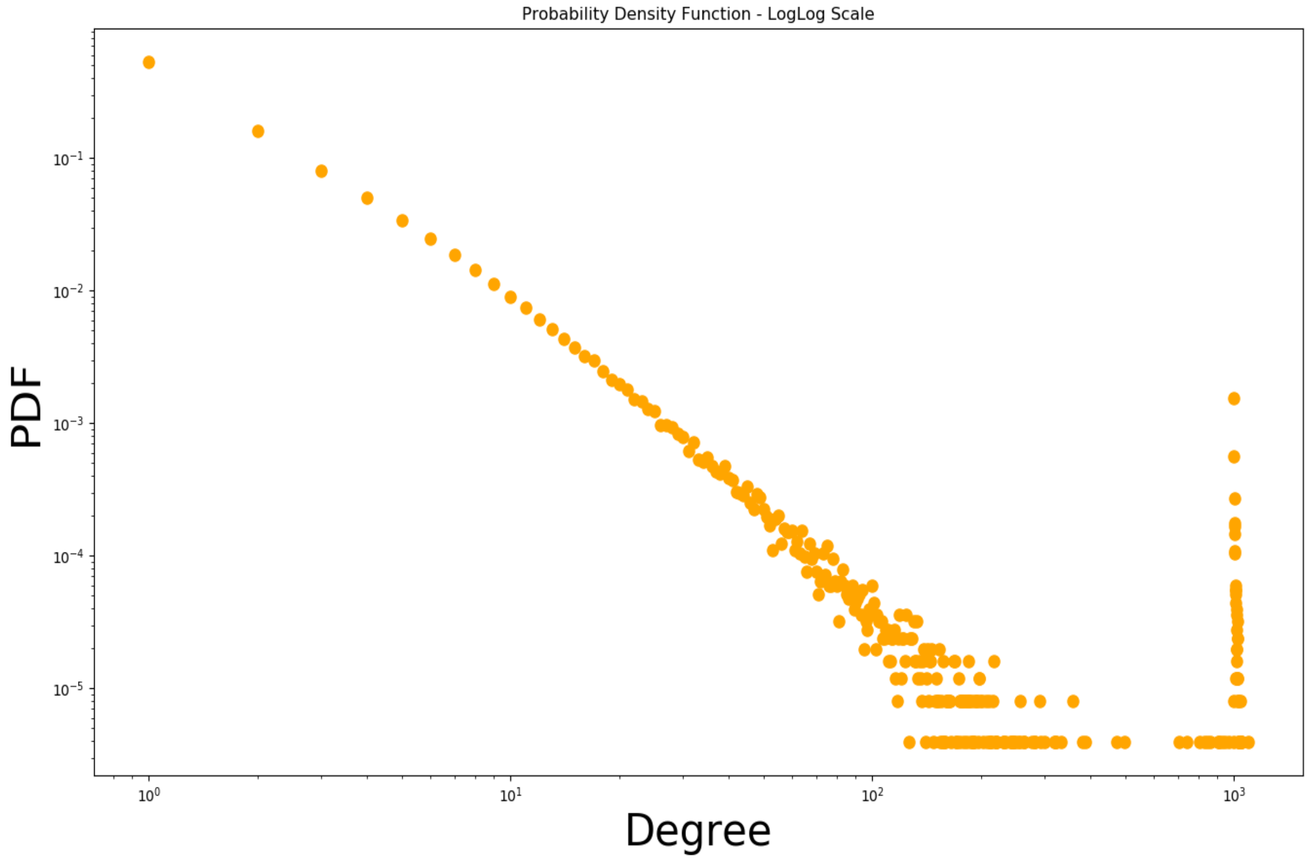
Qui troviamo una PDF su scala logaritmica loglog.

Abbiamo modo di notare che la curva caratteristica si avvicina ad una retta, dandoci modo di credere di essere in presenza di una degree distribution che segue la power law.

Notiamo inoltre una concentrazione di punti intorno a k = 4000, con k = degree.

Ciò dipende dal fatto che gli out-degree, per ciascun utente, sono limitati da 500px 4000 fotografi massimo.

(rif. https://support.500px.com/hc/en-us/articles/203999518-How-do-I-manage-my friends-following-and-followers-)



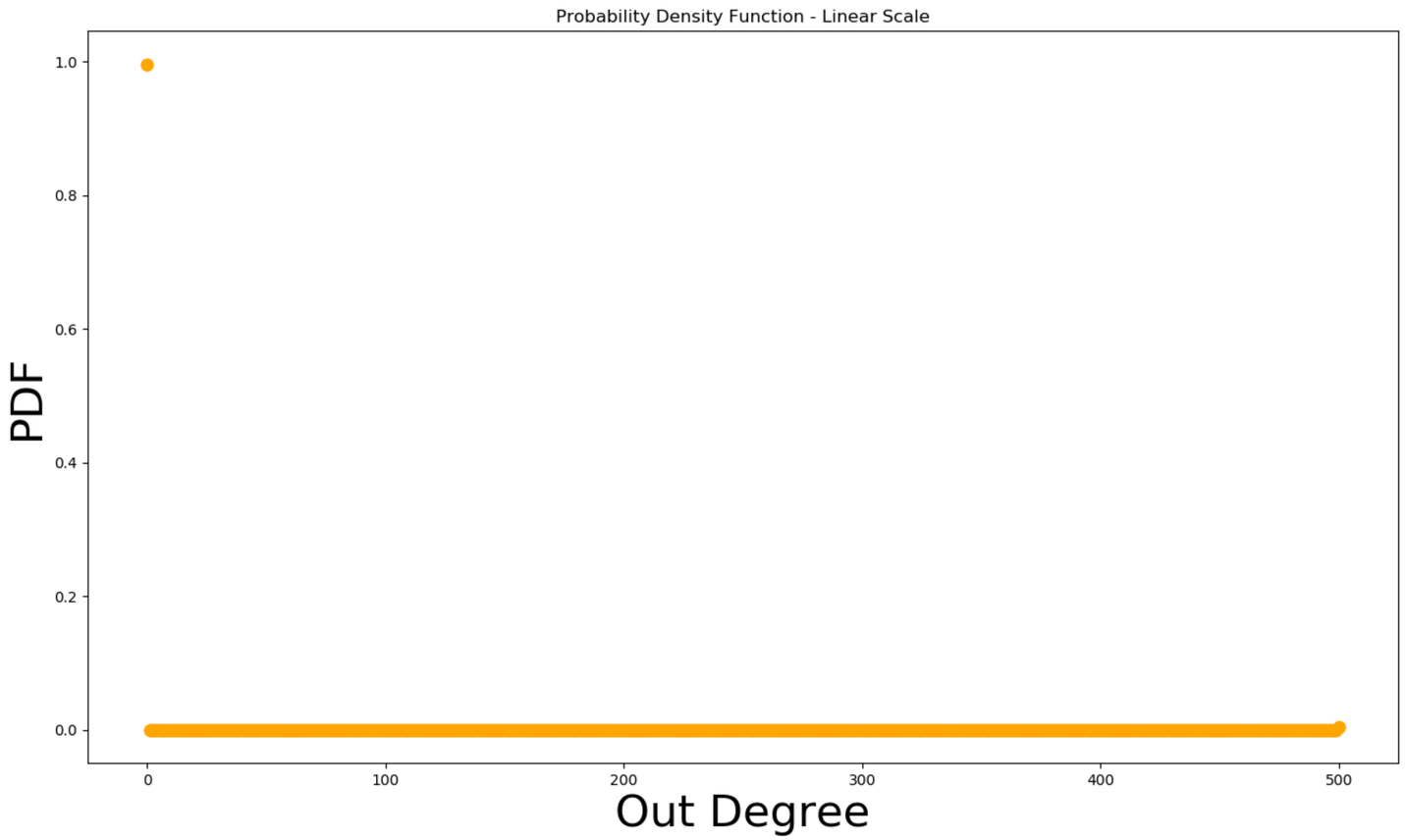
Distribuzione del grado

PDF Linear

Con gephi vado a controllare la distribuzione dell'out-degree.

La presenza di **utenti con 500 e più utenti seguiti** (*considerando il cap imposto durante la fase di scraping potrebbero anche essere migliaia*…) è **sintomo di un comportamento non social**.

E' stato infatti ben stabilito che un individuo può mantenere solo 150 relazioni alla volta (*vd. Numero di Dunbar*)

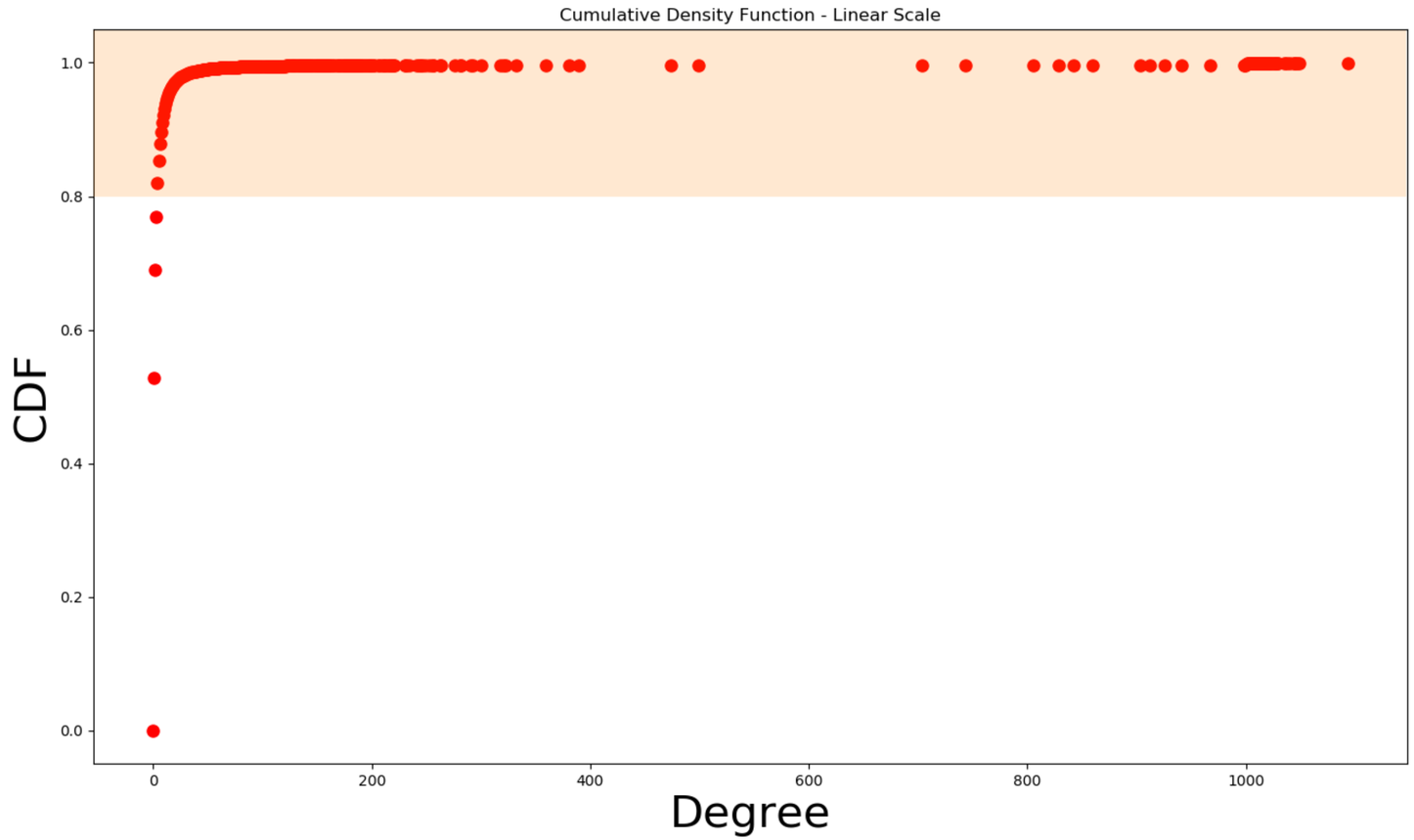


Distribuzione del grado

CDF

Accanto la Cumulative Distribution Function per la rete oggetto di studio.

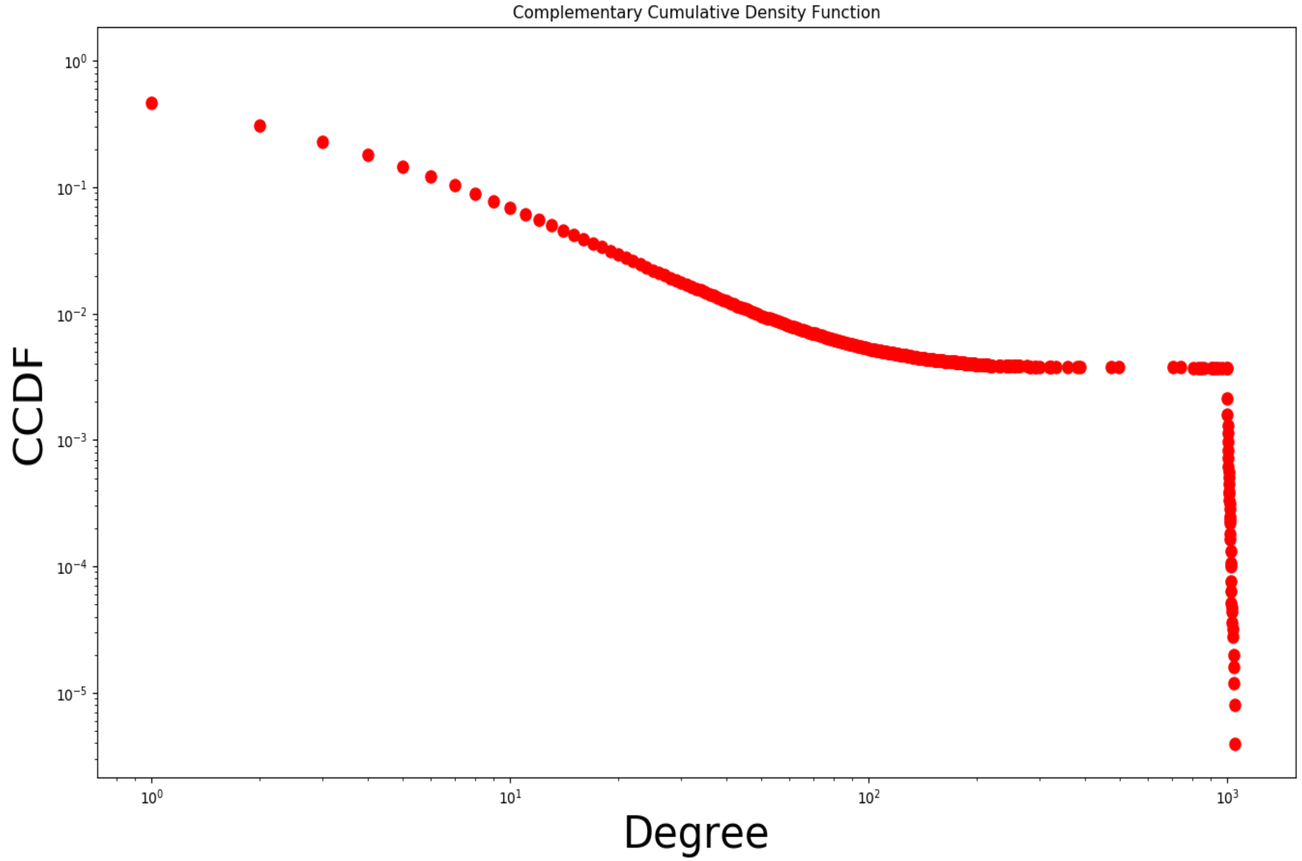
Qualitativamente possiamo inoltre dire che il 20% della popolazione (*evidenziato in arancione*) risulta avere la maggior parte del degree, rispettando la regola di Pareto



Distribuzione del grado

CCDF

La Complementary Cumulative Distribution Function conferma le analisi sin’ora fatte, mettendo ancora una volta in evidenza l’accumulo di degree intorno a k = 4000



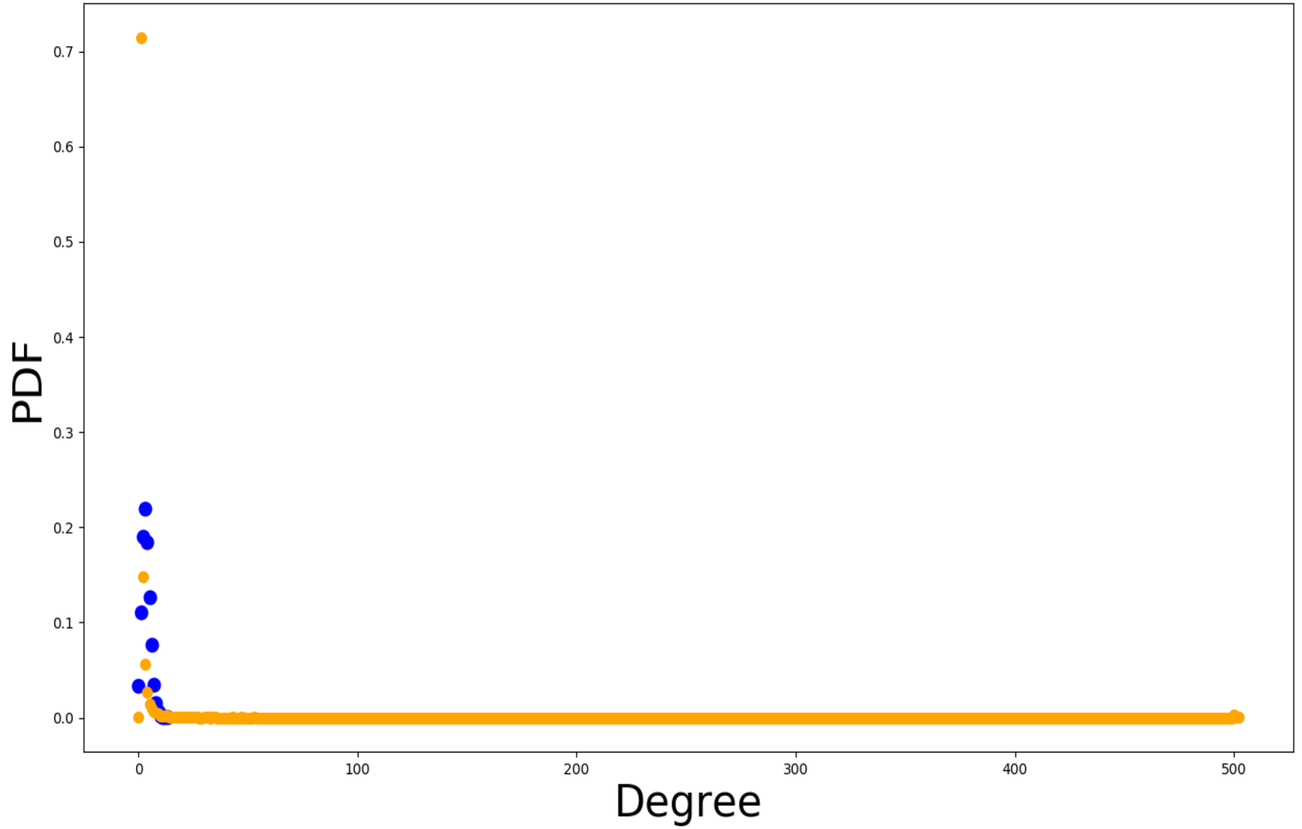
Confronto con modello Random Network

Erdos-Renyi

Qui abbiamo un confronto col modello delle random network (*in blu*), generato usando lo stesso campione acquisito da 500px.

Sfruttando le previsioni a partire dal modello random, sappiamo che:

- La rete di 500px è in regime supercritico, <k> > 1 e quindi ci aspettiamo di trovare un giant component (*vedremo che è effettivamente così*)
- Non siamo in regime di connessione in quanto <k> non è maggiore di logN = 4,4392 (*ricordiamo che abbiamo un <k> pari a 3,42 e che tale prevesione è pienamente attendibile solo su dataset generate con modello G = {N,P}*

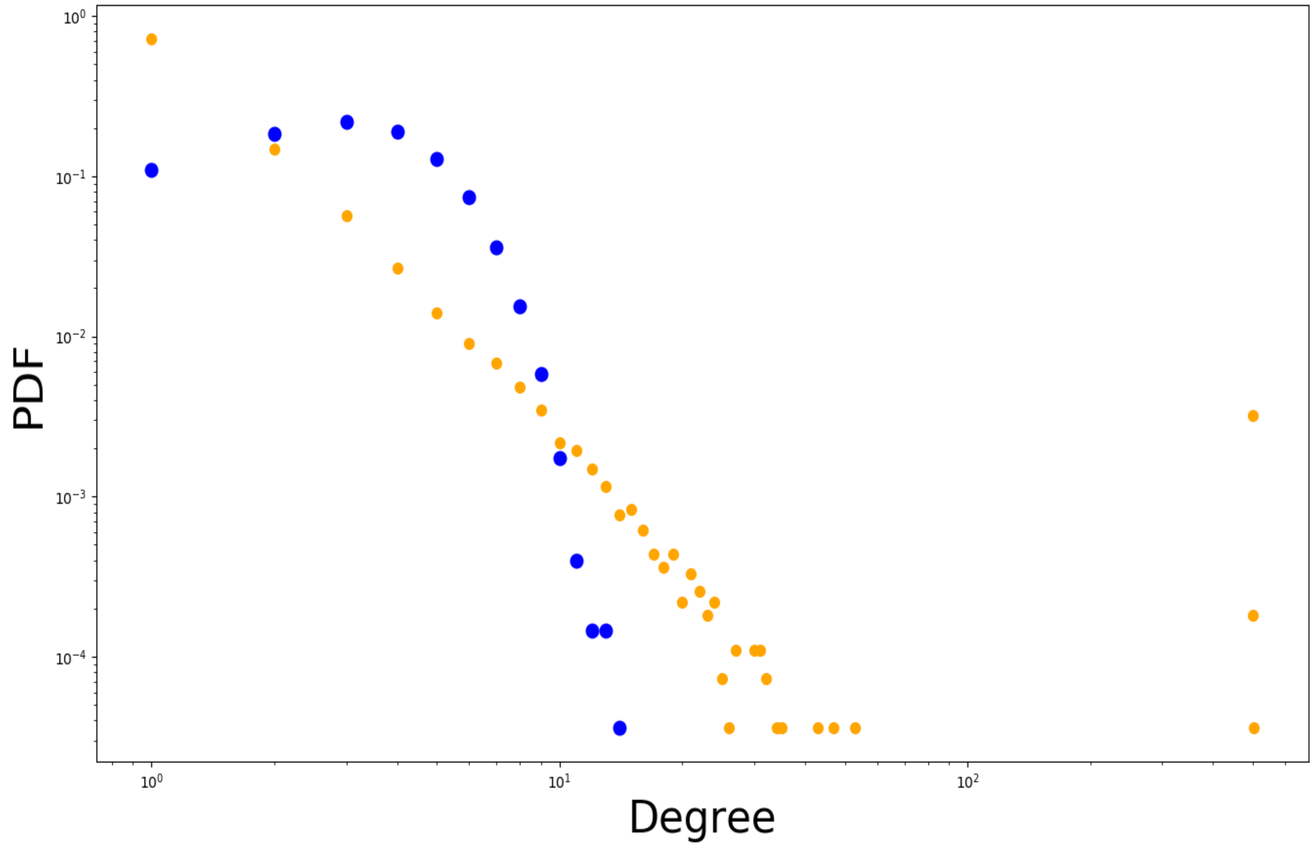


Confronto con modello Random Network

Erdos-Renyi

Qui abbiamo un confronto col modello delle random network (*in blu*), generato usando lo stesso campione acquisito da 500px, in scala loglog.

L'approssimazione del modello random si presenta come da conoscenze teoriche.

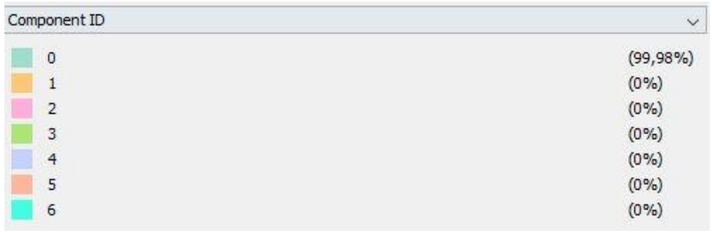


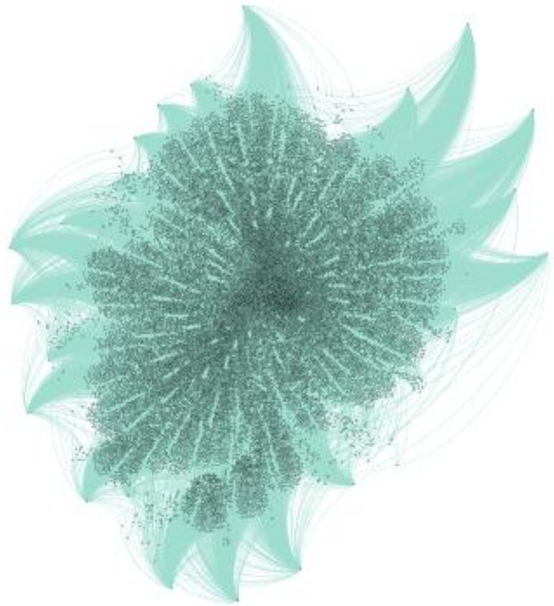
Erdos-Renyi

Essendo la rete in regime supercritico, ci aspettiamo di trovare un giant component.

Esiste un componente che comprende la grande maggioranza (il 99,98%) di utenti di 500px, un po’ come accade in Facebook con il 99,91% del network.

Quindi non solo abbiamo dei path corti (lo vedremo più Avanti) tra le coppie di nodi, ma queste connessioni esistono quasi tra tutti in 500px.





La network che stiamo studiando è una

Scale free?

Dai risultati in nostro possesso sappiamo che per la rete di 500px vale:

- Varianza: 850
- Dev. Standard: 29,16

Sappiamo che nelle random network la «scala» viene definita come:
average degree ± dev. Std.

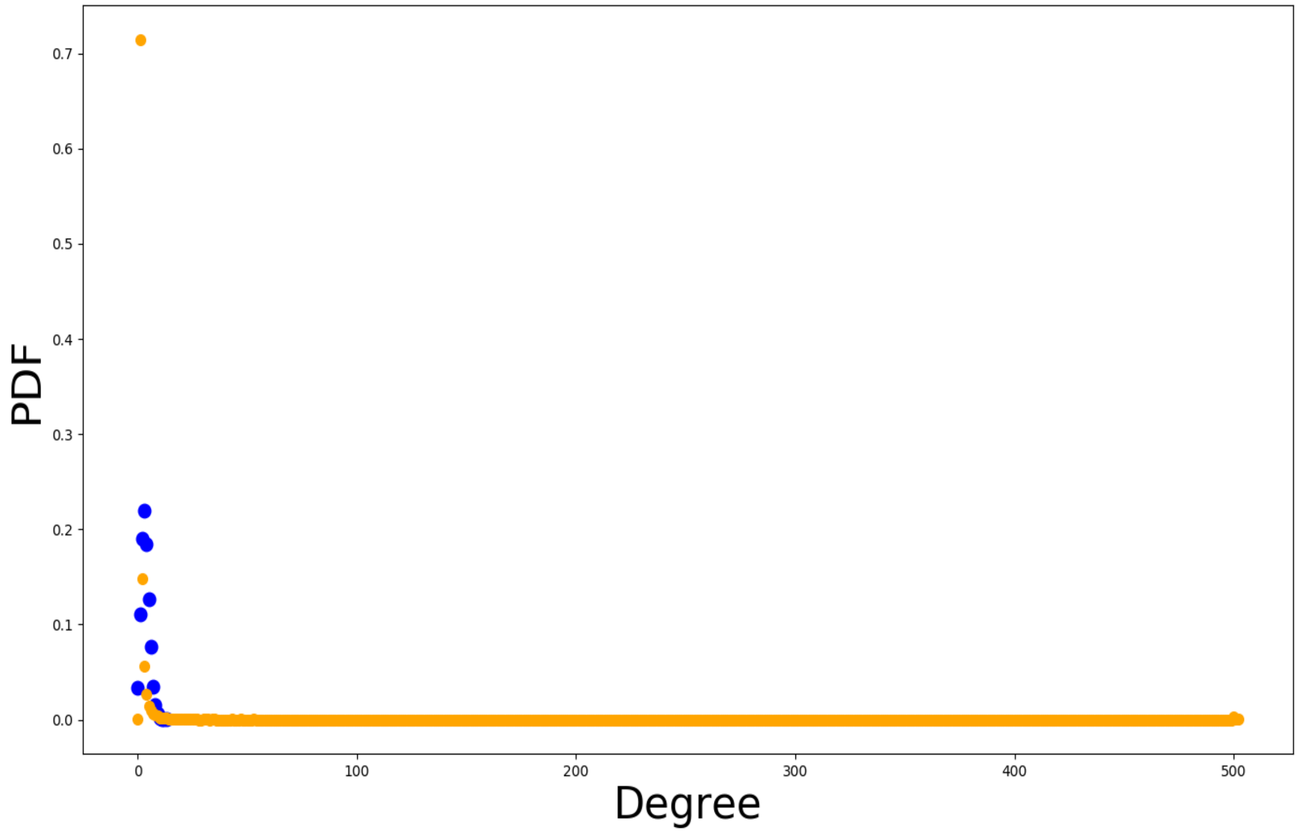
Proviamo a vedere se il nostro range di variazione del degree ricade in questa scala.
Abbiamo:

- Scala = 3,42 ± 29,16

Ovvero [-25,74 ; 32,58]

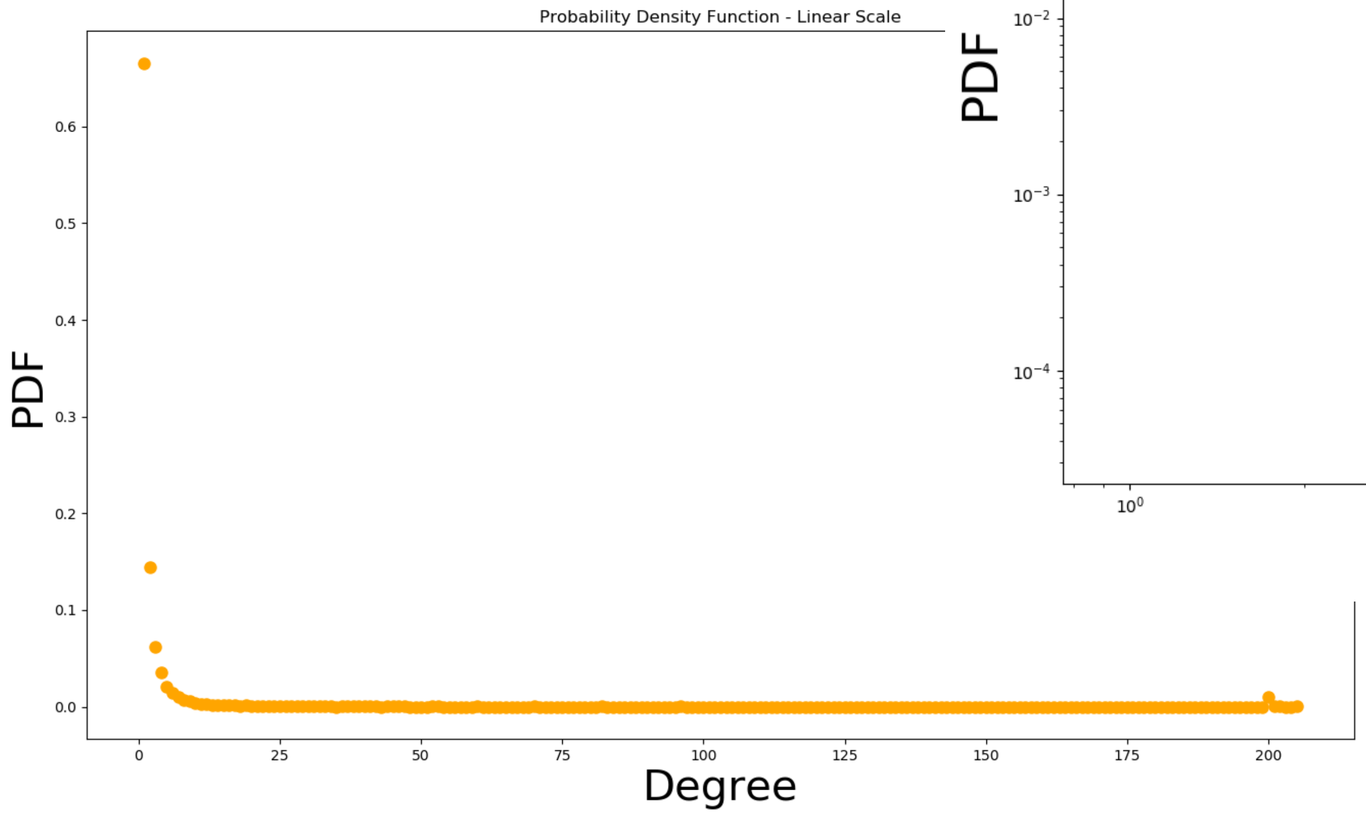
Decisamente l'intervallo che definirebbe la nostra scala è troppo limitato per i nostri valori.

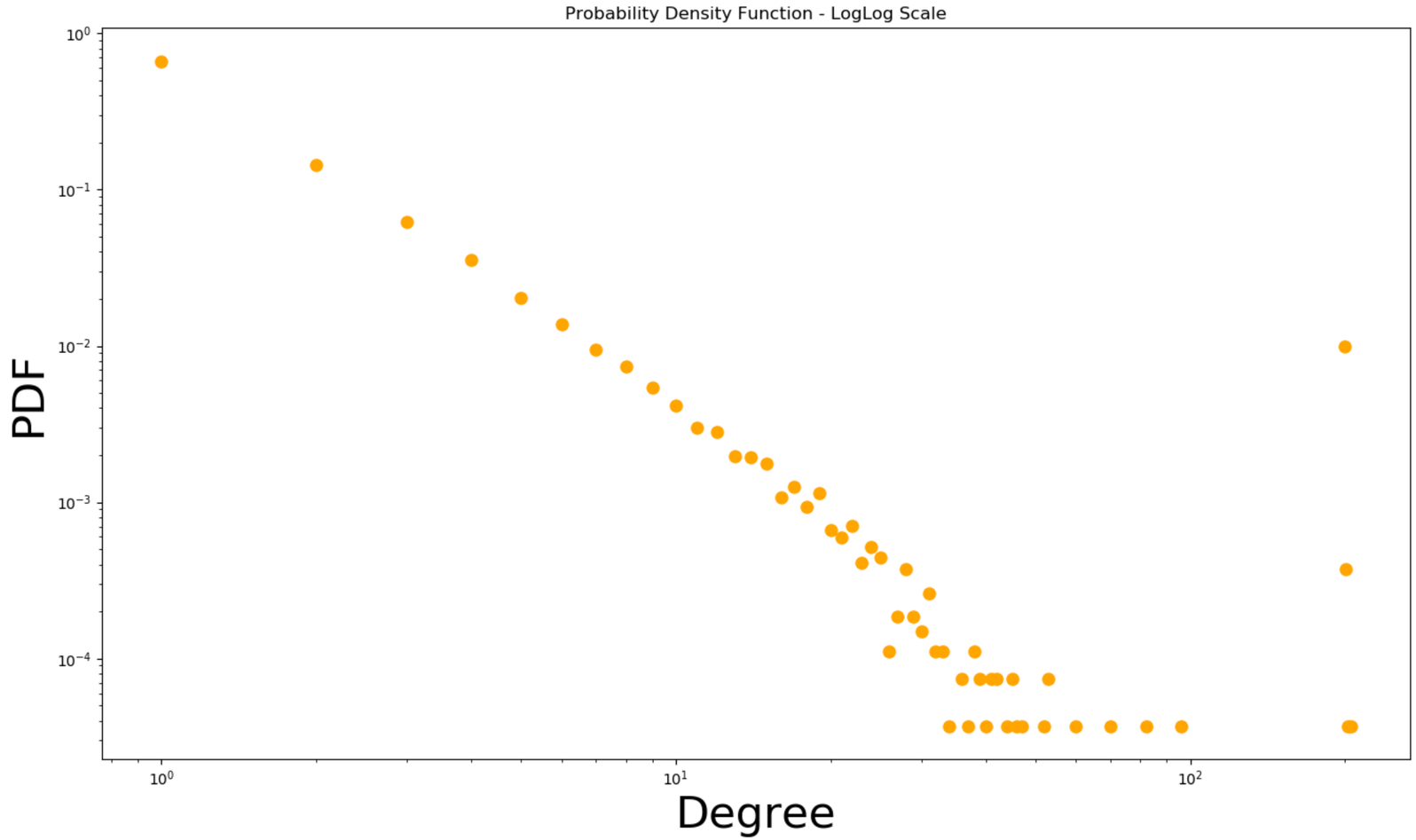
Siamo quindi in presenza di una scala free network.



Italy Degree

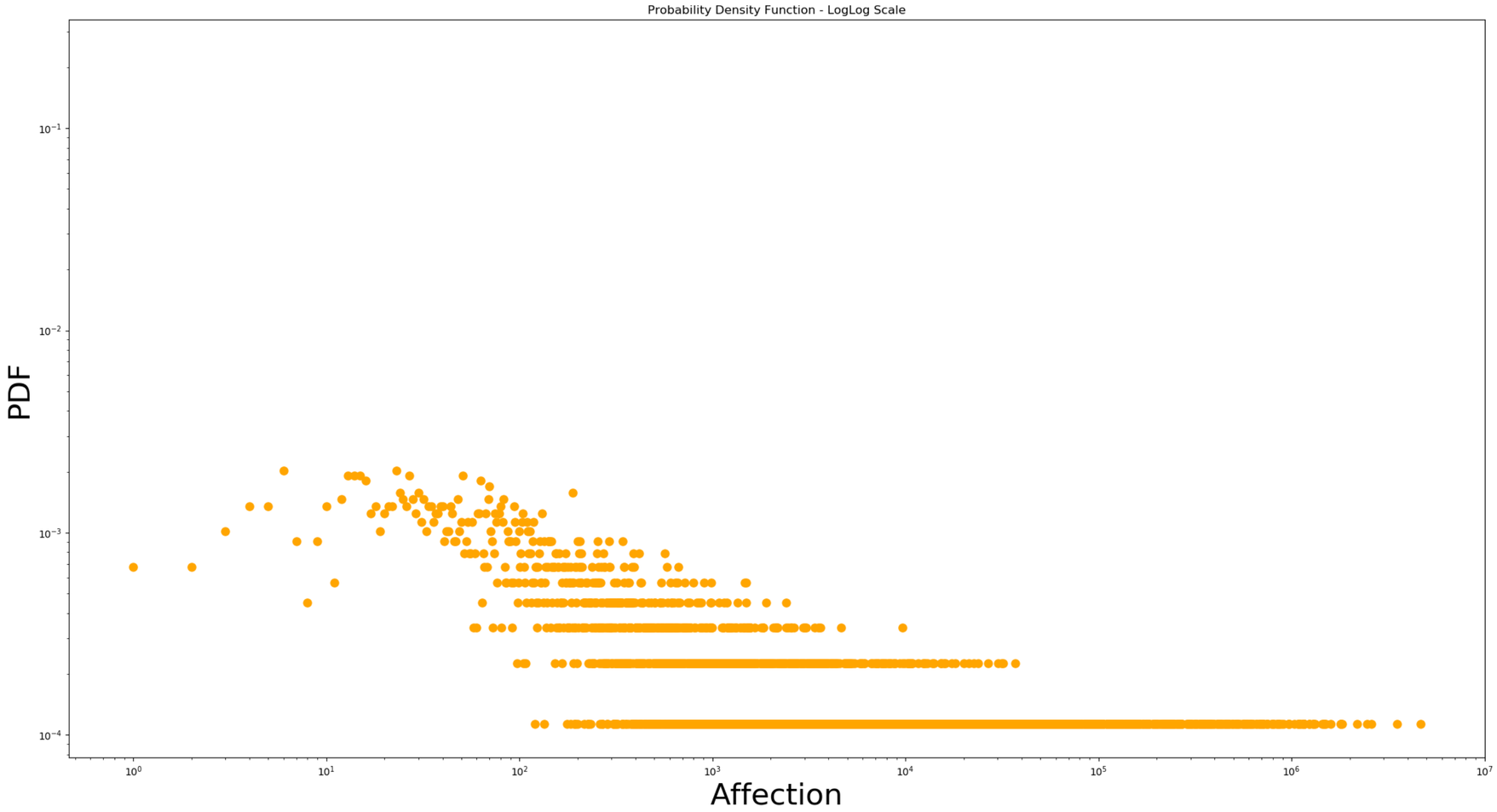
Come si presenta la degree distribution per i soli utenti italiani?

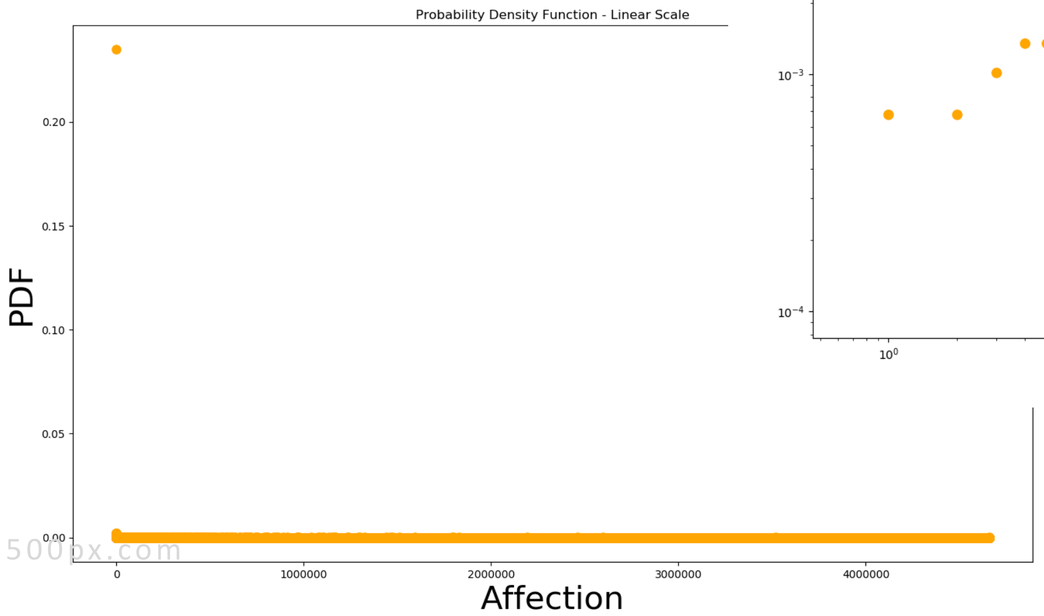




Affection distribution

Come si presenta la distribution dell’affection?





Scale-free network & hubs

hubs

Ipotizzando di trovarci in una scala free network (tesi avvalorata dalle indagini che stiamo facendo), è lecito individuare i tre hub più grandi.

Di fianco il codice usato per l'individuazione.

```
#
# Get the hubs
#
def get_hubs(self):
    network_degree = list(dict(self.network_graph.degree).values())
    # Define a thresold to identify what is "high degree"
    # I choose to check percentile with p = 0.98
    # I know that 98% of total nodes have got a degree which is lower
    quantile_98 = numpy.percentile(network_degree, 98)
    # Now get the nodes
    hub_nodes = [k for k,v in dict(self.network_graph.degree).items()
    # Extend the hub_nodes array list with information about users
    # Getting Array<User>
    hubs = []
    for hubId in hub_nodes:
        hubs.append( self.network_graph.node[hubId].get('info') )
    return hubs
```

```
#
# Get the hubs
#
def get_hubs(self):
    network_degree = list(dict(self.network_graph.degree).values())
    # Define a thresold to identify what is "high degree"
    # I choose to check percentile with p = 0.98
    # I know that 98% of total nodes have got a degree which is lower than quantile_98 variable
    quantile_98 = numpy.percentile(network_degree, 98)
    # Now get the nodes
    hub_nodes = [k for k,v in dict(self.network_graph.degree).items() if v >= quantile_98]
    # Extend the hub_nodes array list with information about users
    # Getting Array<User>
    hubs = []
    for hubId in hub_nodes:
        hubs.append( self.network_graph.node[hubId].get('info') )
    return hubs
```

hubs

01 **Sean Archer**
500px id (777395)

02 **Mark Bridger**
500px id (99604)

03 **Paul Zizka**
500px id (75551)

Hub #1 (id 777395)

Sean Archer

Affection **4.598.598** (*più alto valore nel campione*)

Followers **165.072**

Degree Centrality **0.01818578598967O473**

Nota: la degree centrality non corrisponde all'affection. Questa viene infatti misurata con altri parametri.


777395

Hub #2 (id 99604)

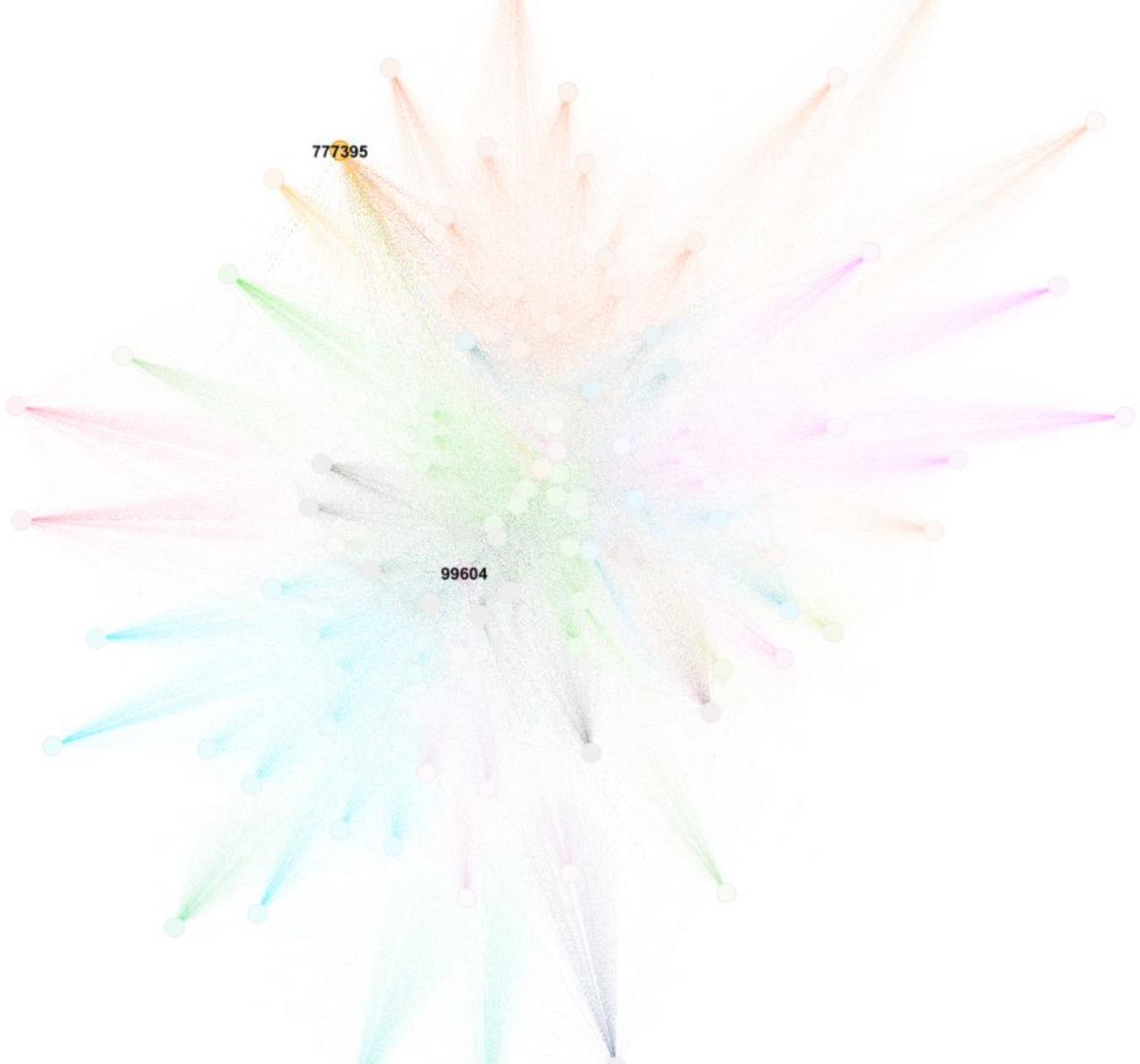
Mark Bridger

Affection **856.454**

Followers **131.878**

Degree Centrality **0.01818578598967O473**

Nota: abbiamo avuto, rispetto a Sean Archer, una drastica diminuzione dell'affection e dei followers, ma la degree centrality è la medesima. Ciò perché la fase di creazione del dataset ha raccolto i follower con un cap impostato a 500 circa, per ogni utente.


777395
99604

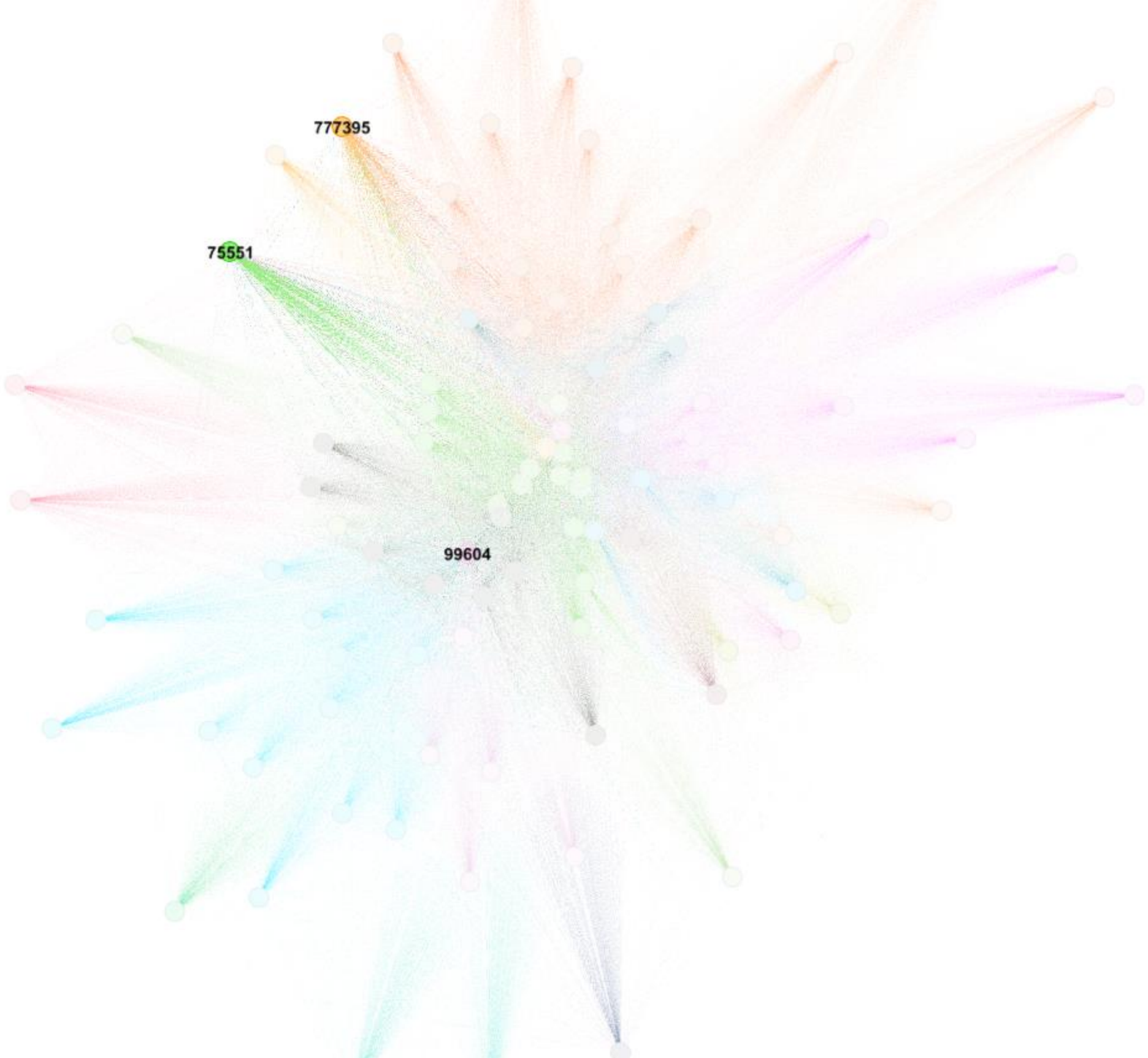
Hub #3 (id 75551)

Paul Zizka

Affection **540.943**

Followers **131.835**

Degree Centrality **0.01818578598967O473**


777395
75551
99604

Analisi del

Natural cutoff

Il natural cutoff è definito come il maximum expected degree all'interno della network.

Sappiamo che:

- **Grafo completo**
  $k_{max} = n - 1$
- **Modello di Erdos-Renyi**
  $k_{max} = \ln(N)$
- **modello power law**
  $k_{max} = k_{min} \times N^{\wedge}(1/\gamma-1)$

Tenendo a mente il grafico a destra, dovrebbe sussistere la relazione:

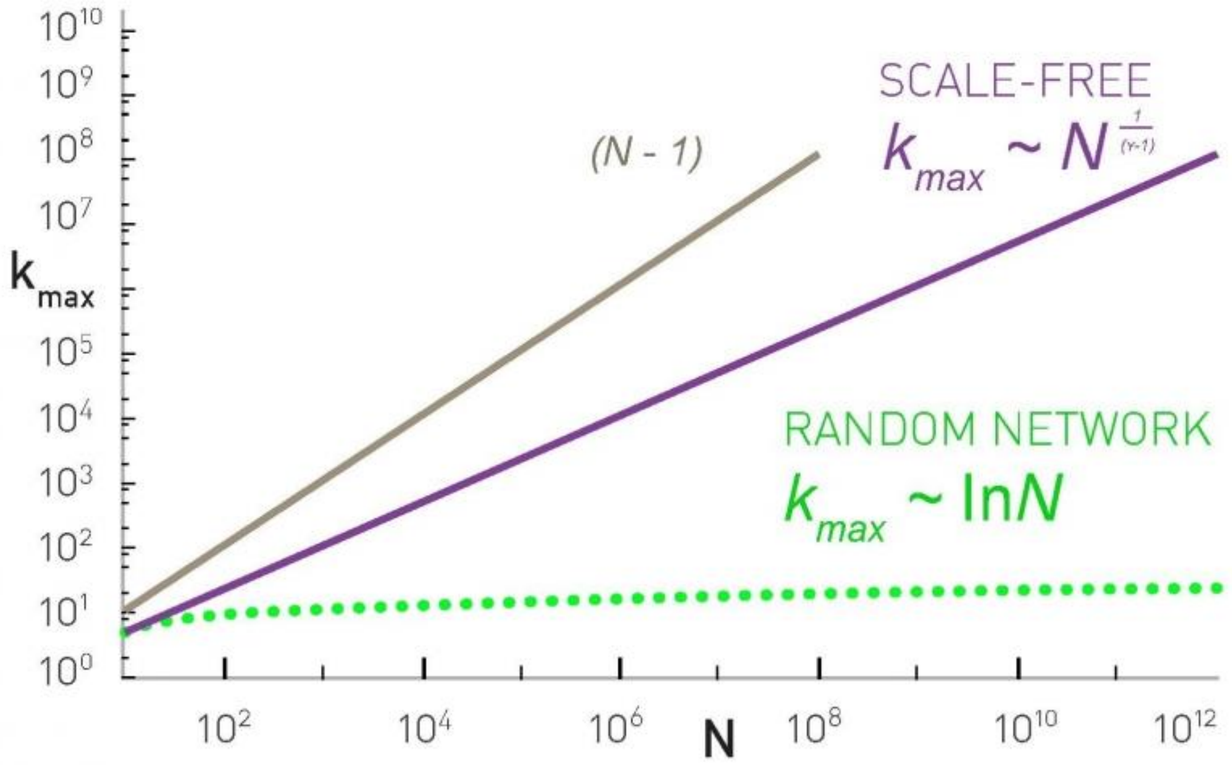
$$\ln(N) < k_{max} < N - 1$$

Cioè

$$10{,}2217 < k_{max} < 27.494$$

E rapportandola ai dati della popolazione

$$16{,}3804 < k_{max} < 13.000.000$$



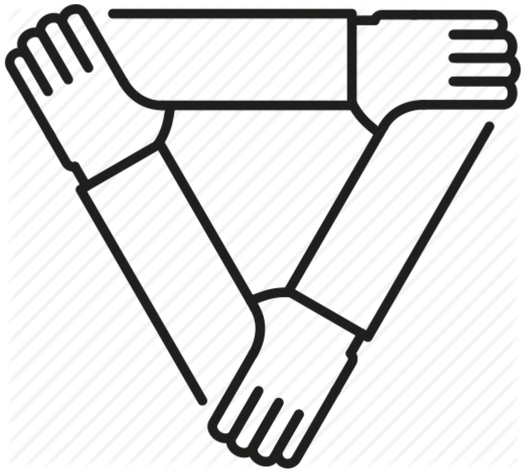
Nel dataset in analisi troviamo 502 come grado massimo, valore che rispetta la previsione.
Tuttavia, ricordando che l'hub #1 ha k=165.072 e rapportandolo ad una popolazione di 13 mln di utenti, abbiamo un ulteriore conferma del fatto che siamo in presenza di una network che segue la power law.

“L’amico del mio amico è mio amico”

Transitivity

Il Global Clustering Coefficient, usato per il calcolo della transitività, che esprime il concetto di „*un amico del mio amico è mio amico*“ è notevolmente basso: 0,00000511.

Il valore così basso ci consente di pensare che la network sia in uno stato embrionale se considerata da un punto di vista **social network**.
E’ doveroso ricordare però che si è considerato un campione rappresentativo del solo 0,20% dell’intera popolazione (*27.000 nodi contro I quasi 14 milioni di utenti totali*).



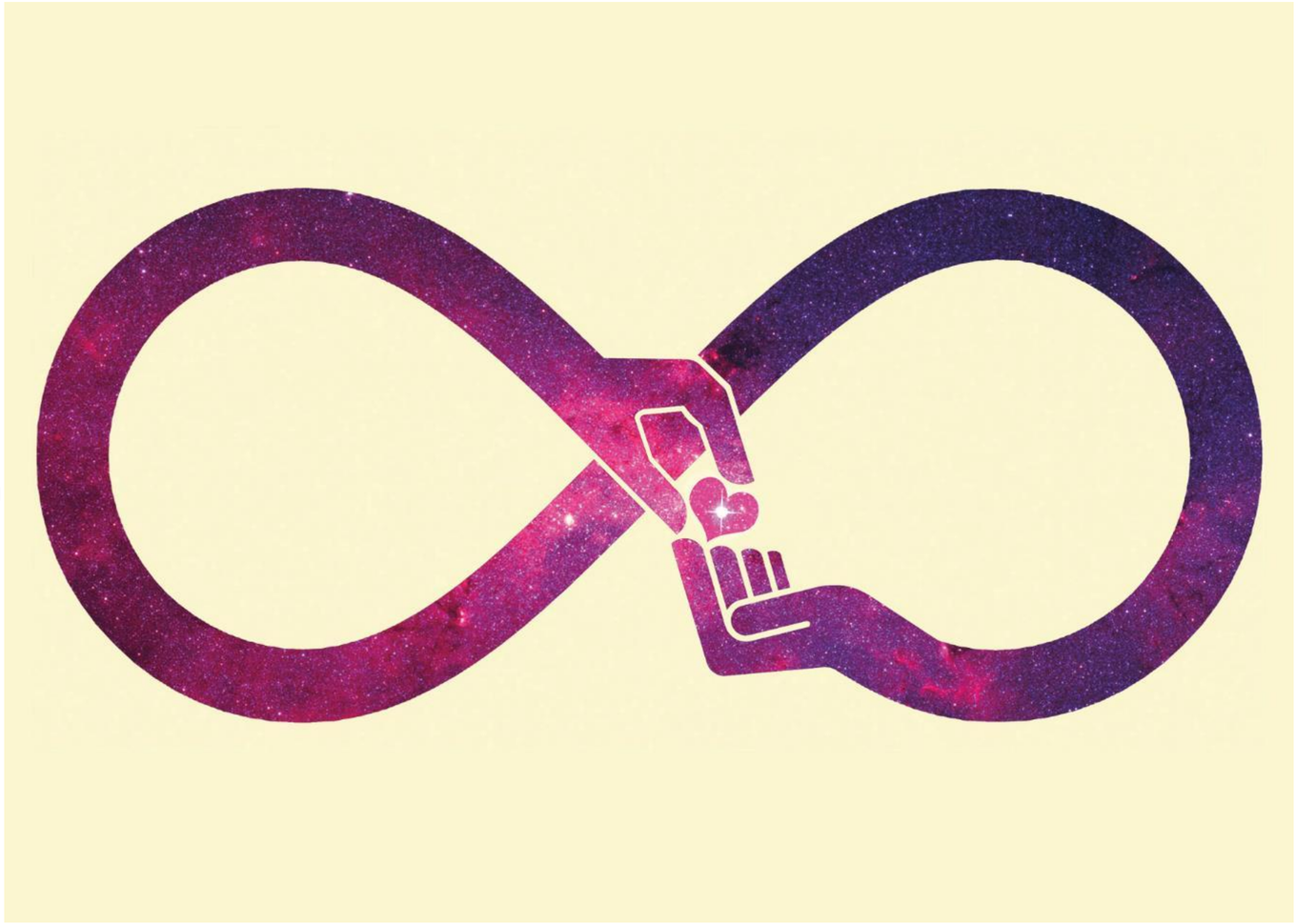
Analisi della

Reciprocity

La misura della probabilità di «essere seguiti, seguendo» cioè è più probabile che un utente mi segua se io lo seguo,

è nel nostro caso 0 ma potrebbe dipendere da un dataset rappresentativo del solo 0,20% della totalità della popolazione.

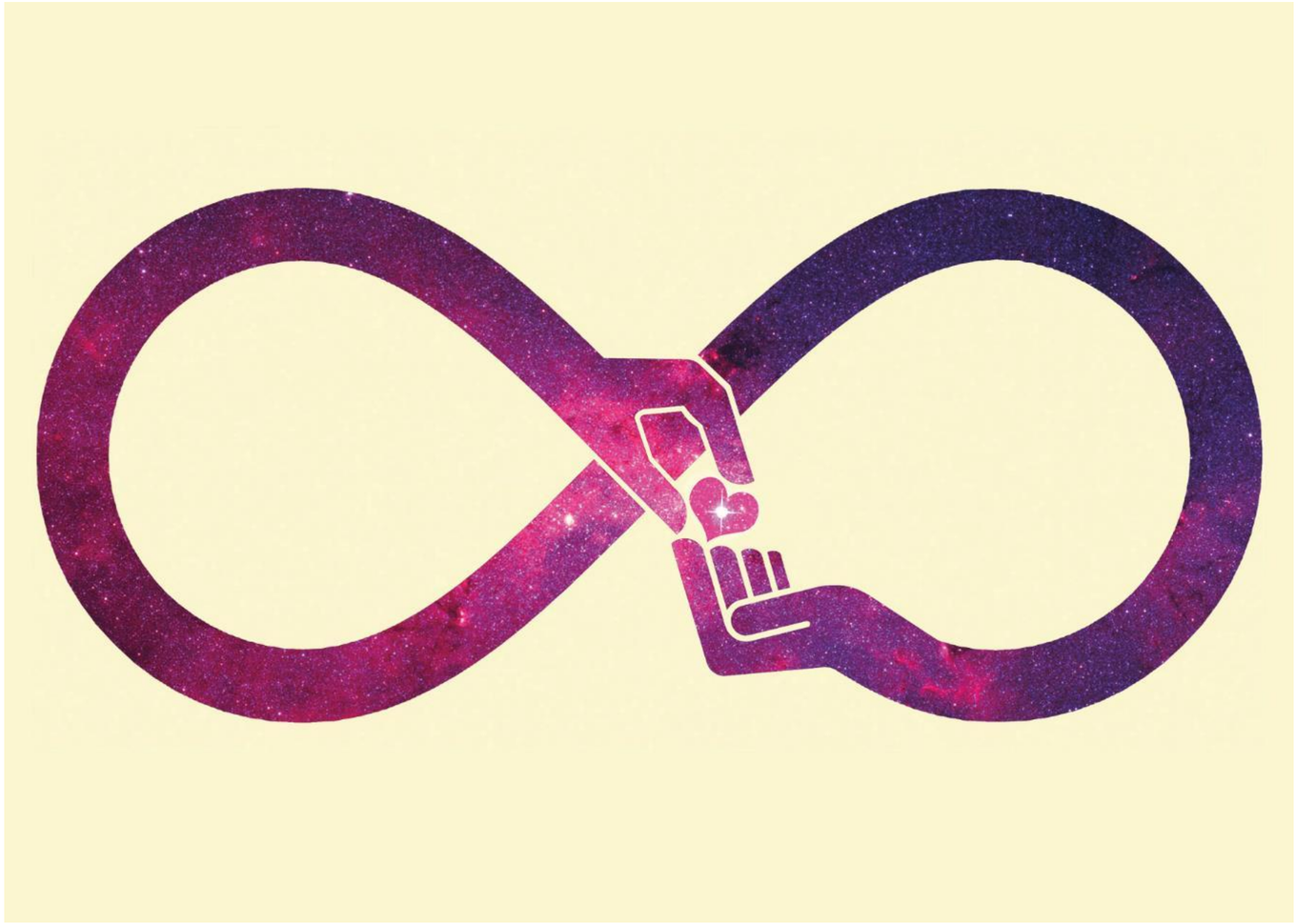
Potrebbero infatti essere stati esclusi dalla fase di scraping dei cicli di lunghezza due falsificandone il risultato finale.



Social o information?

I valori di transitivity e reciprocity ci danno l’idea di essere in presenza di una information network.

Tuttavia, considerando la bassa rappresentatività del campione a causa della dimensione, potremmo supporre che si tratti di un social network in fase embrionale.



Analisi della

Community

Modularity Q : 0,612
Number of Communities: 25

L’algoritmo applicato in Gephi per la ricerca delle community permette di identificare I gruppi più densamente connessi rispetto a tutta la network.

Modularity Class	
8	(14,51%)
0	(10,04%)
5	(9,26%)
12	(8,08%)
14	(5,93%)
3	(5,85%)
15	(5,36%)
1	(5,02%)
9	(4,31%)
4	(3,69%)
6	(3,46%)
17	(3,07%)
2	(2,88%)
16	(2,43%)
21	(2,42%)
20	(2,27%)
7	(2,24%)
11	(2,08%)
13	(1,97%)
10	(1,82%)
19	(1,66%)
18	(1,64%)

Analisi della

Small world?

Le analisi ci mostrano che il network non tende a comportarsi come un modello small world, avendo delle lunghezze dei path, in media, molto basse (*facebook è intorno a 4,7*) ma un average clustering coefficient altrettanto basso.

	GRAFO ORIENTATO	GRAFO NON ORIENTATO
DIAMETRO	2	6
AVERAGE PATH LENGTH	1,067	3,993
AVERAGE CLUSTERING COEFF.		0,002
TRIANGLES		120

Analisi della

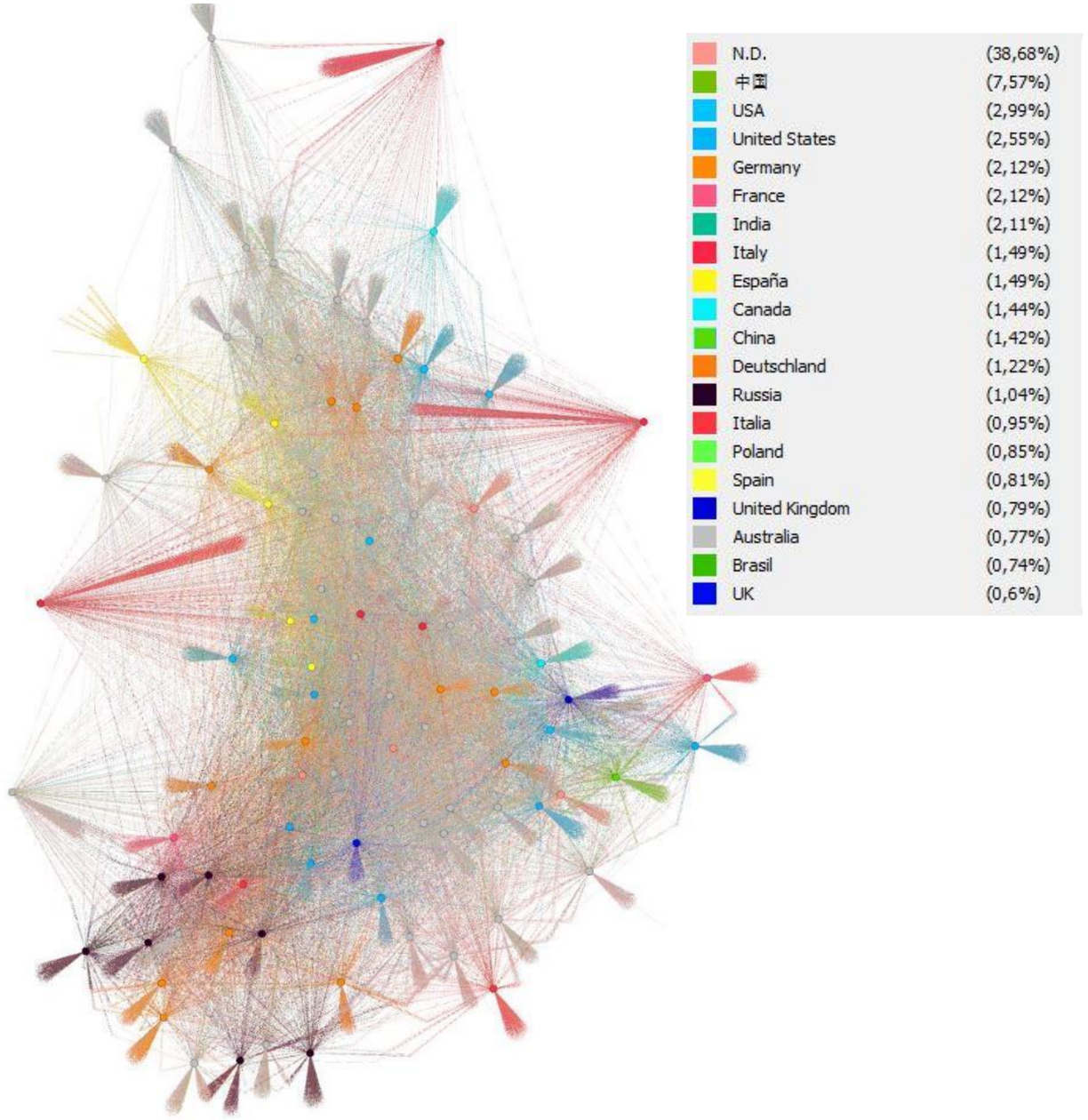
Assortativity

Valutiamo cioè quanto i nodi tendano a seguire utenti simili per provenienza geografica.

L’attributo considerato, estratto dal dataset, è country.

L’assortativity calcolata è pari a
-0.03276

A significare che ci sono meno archi di quelli attesi tra utenti dalla stessa Nazione.
Ciò può essere dovuto a stesse Nazioni indicate con nomi differenti oppure al 39% di popolazione che non ha indicato la provenienza geografica.



Analisi della

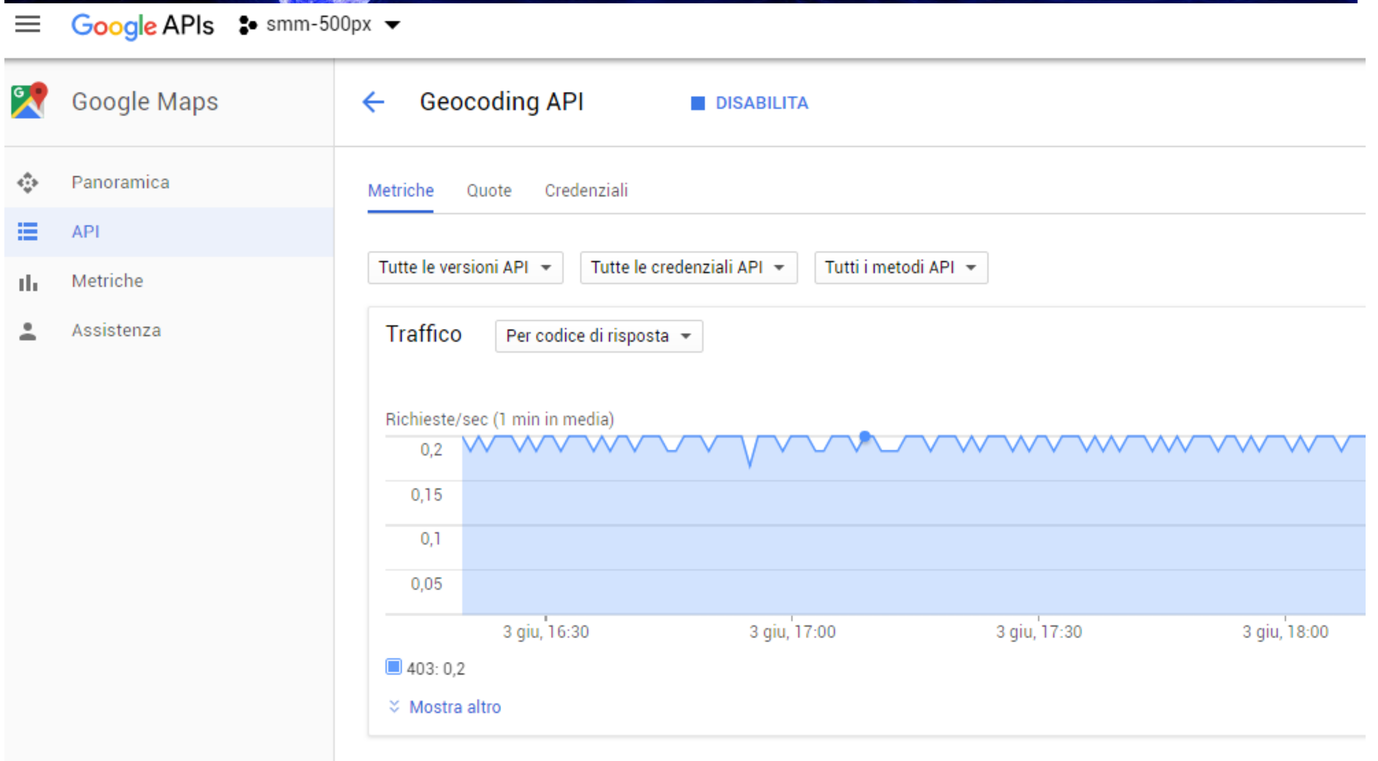
Assortativity

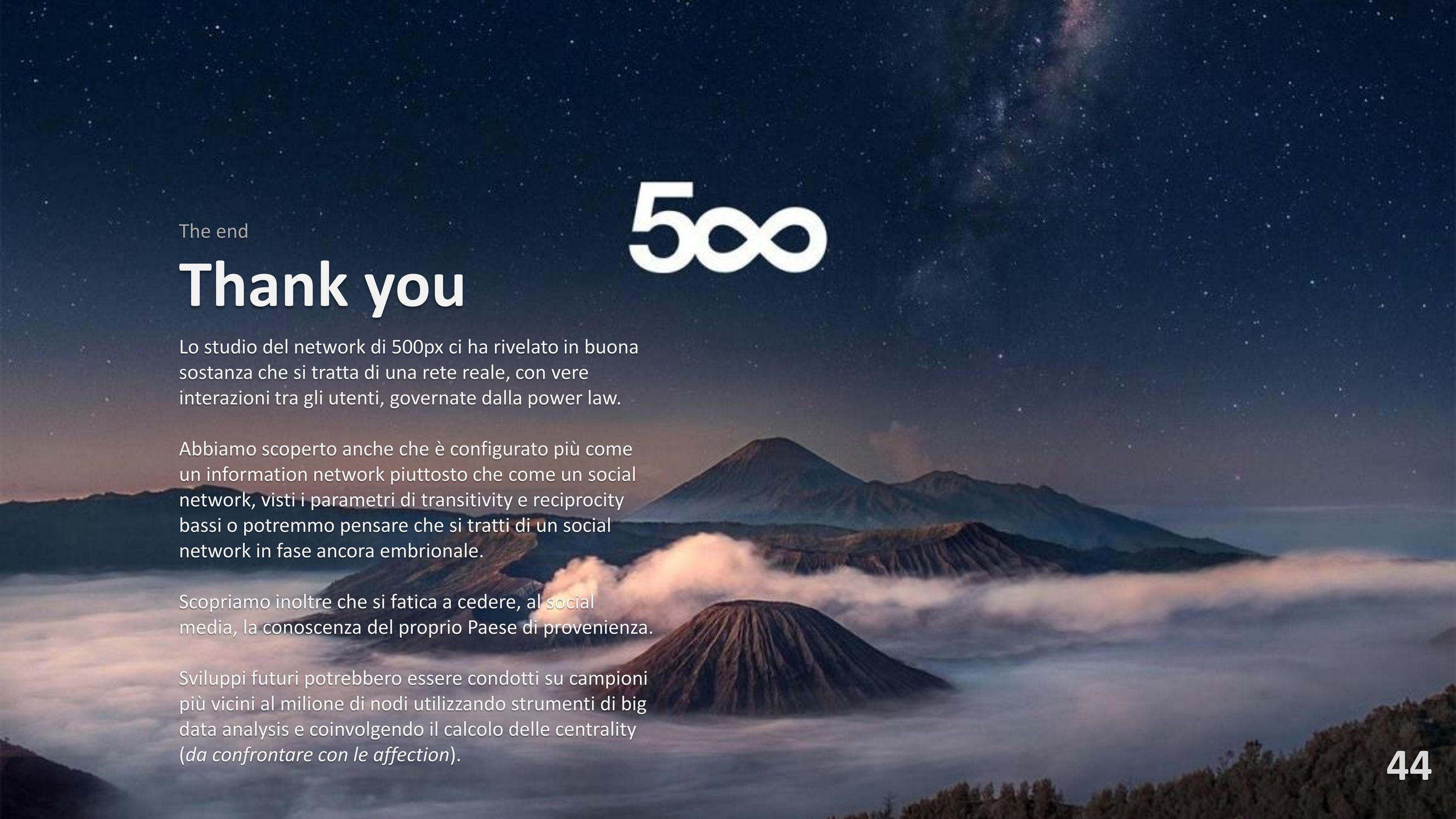
Valutiamo cioè quanto i nodi tendano a seguire utenti simili per provenienza geografica.

Gli attributi considerati, estratti dal dataset, sono country e city.

Grazie al **geolayout** (*plugin esterno di Gephi*) e grazie alla libreria **geocoder** di python (*progetto free github*) e grazie all'integrazione con le **Google geocoding API**, posso utilizzare le coordinate delle città/country di provenienza degli utenti per dar modo agli *edge* di creare un planisfero.

N.B. per il dataset estratto la disponibilità di dati è stata talmente minima da non poter consentire una visualizzazione significativa.





The end

Thank you

Lo studio del network di 500px ci ha rivelato in buona sostanza che si tratta di una rete reale, con vere interazioni tra gli utenti, governate dalla power law.

Abbiamo scoperto anche che è configurato più come un information network piuttosto che come un social network, visti i parametri di transitivity e reciprocity bassi o potremmo pensare che si tratti di un social network in fase ancora embrionale.

Scopriamo inoltre che si fatica a cedere, al social media, la conoscenza del proprio Paese di provenienza.

Sviluppi futuri potrebbero essere condotti su campioni più vicini al milione di nodi utilizzando strumenti di big data analysis e coinvolgendo il calcolo delle centrality (*da confrontare con le affection*).

500px

Sviluppi futuri

L‘indagine condotta ha analizzato la maggior parte delle metriche principali di una network, ma quale **strada** seguire **per futuri sviluppi?**