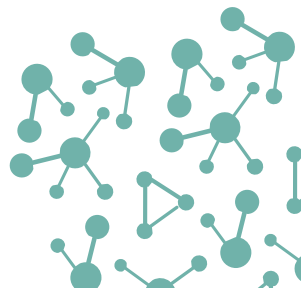


# LICHESS.ORG

Uno studio di social media mining



# Cos'è Lichess?

Lichess è un server di scacchi open-source e gratuito che offre la possibilità di giocare online contro persone di tutto il mondo a diverse varianti di scacchi.

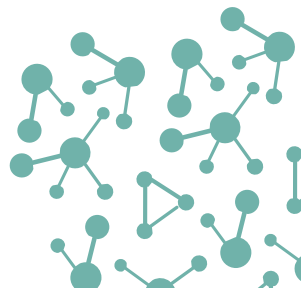
- ✓ Accesso facile ai suoi dati tramite API
- ✓ Dati sugli utenti e sulle partite molto ricchi
- ✓ Presenza di gruppi e possibilità di creare relazioni tra utenti



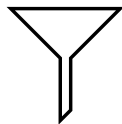
# Qual è la rete che si vuole studiare?

Con questo studio si vuole analizzare la rete di utenti iscritti al gruppo 'ChessNetwork', una delle community più popolari sulla piattaforma, composta da 10629 utenti. Ogni utente viene modellato come un nodo, ogni relazione di following viene modellata come un link orientato mentre le partite disputate determinano i pesi dei link.

*Nota: sono stati esclusi dallo studio i nodi singoletto, ovvero quelli che non avevano alcun link entrante/uscente nel grafo, in totale i nodi rimanenti sono 6775.*



# Data gathering

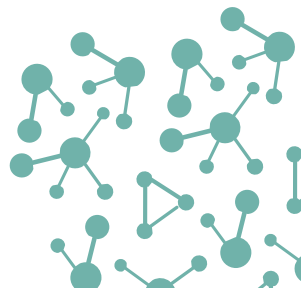


La fase di data gathering è avvenuta utilizzando le API messe a disposizione da lichess.org e la libreria python berserk. Per velocizzare la raccolta dei dati sono stati creati più bearer token.

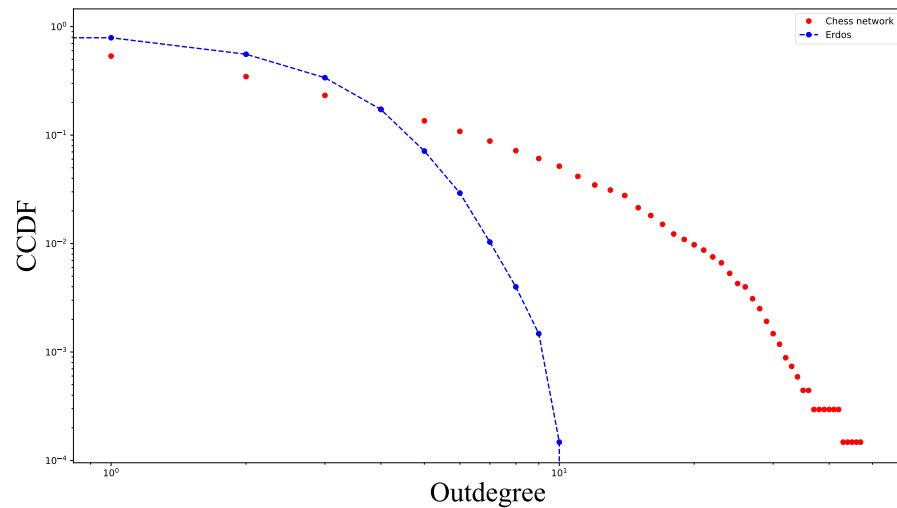
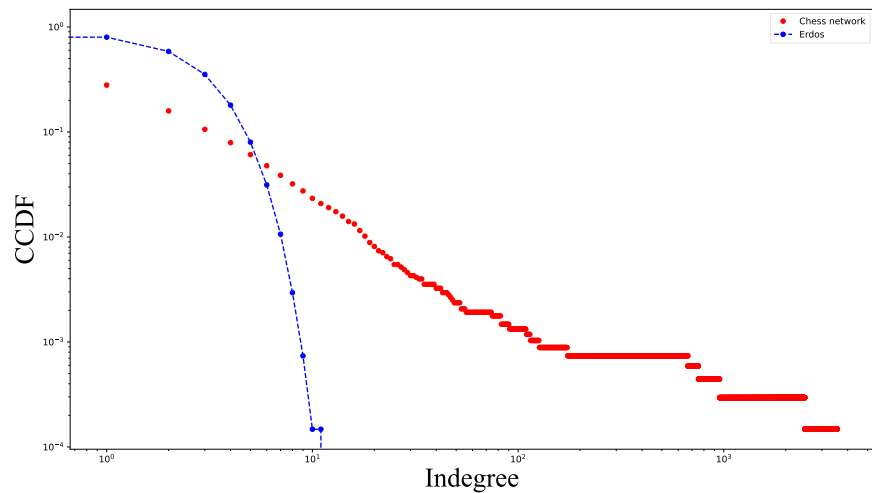
Il processo è stato particolarmente rallentato dalla raccolta dei dati sulle partite. Non avendo a disposizione un endpoint che restituisce le partite disputate tra due utenti, è stato necessario raccogliere tutte le partite di tutti gli utenti per poi fare l'intersezione tra gli insiemi di partite di ogni coppia.

<https://berserk.readthedocs.io/en/master/>

<https://lichess.org/api>



# Degree distributions



$$\langle k_{in} \rangle = \langle k_{out} \rangle = 2.966$$

$$\Delta = 0.0004378$$

# È una rete scale free?

Vediamolo attraverso delle statistiche puntuali sulle distribuzioni del grado.

Indegree:

$$\sigma_{in}^2 = 3075.14$$

$$Moda = 0$$

$$Mediana = 1$$

$$k_{min} = 0$$

$$k_{max} = 3556$$

$$quantile_{95} = 6$$

Outdegree:

$$\sigma_{out}^2 = 15.56$$

$$Moda = 1$$

$$Mediana = 2$$

$$k_{min} = 0$$

$$k_{max} = 48$$

$$quantile_{95} = 11$$

Tipicamente nelle reti scale free si osserva che:  $Mediana \ll Media \ll quantile_{95}$

Nel nostro caso questo non si verifica in nessuna delle due distribuzioni.

Tuttavia la varianza dell'in-degree è molto alta, questo ci fa capire che esistono nodi influencer.

# I nodi hub hanno un titolo FIDE?



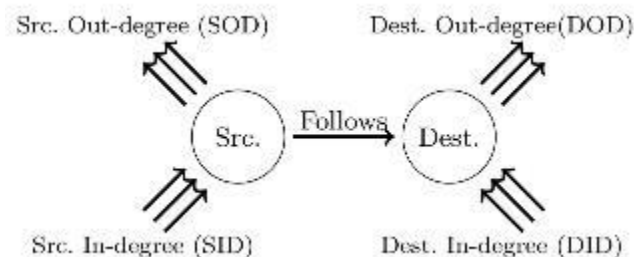
Dato che il 95° percentile del campione in-degree è 6, verrà considerato un percentile più alto per individuare gli hub (99.9 ). I nodi HUB rilevati sono 7 e come si può vedere dal grafico più della metà ha un titolo FIDE. Questo ci fa dedurre che i nodi influencer hanno in-degree alto grazie alla loro rilevanza nell'ambiente scacchistico.



# Degree assortativity and hub's influence

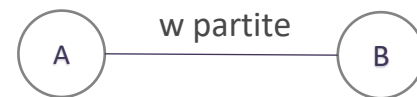
Coefficiente di Pearson	Come si interpreta?
SOD vs DOD: 0.043	non c'è una correlazione significativa
SID vs DOD: -0.003	non c'è una correlazione significativa
SOD vs DID: -0.360	tante più persone la sorgente segue tanto più queste sono meno popolari
SID vs DID: -0.0344	non c'è una correlazione significativa, i nodi hub seguono altri nodi hub ma non seguono utenti con basso in-degree

Coefficiente di Spearman	Come si interpreta?
In-degree vs blitz rating: 0.642	Il grado dei nodi è correlato positivamente con il punteggio dei giocatori nelle partite blitz
In-degree vs rapid rating: 0.285	Il grado dei nodi è correlato positivamente con il punteggio dei giocatori nelle partite rapid
In-degree vs classical rating: 0.035	Non ci si aspettava una correlazione in quanto le partite classiche sono le meno diffuse su Lichess





# Ma...



*‘Dato che la rete in analisi è orientata al gioco, gli individui interagiscono tra loro anche se non stringono un legame esplicito?’*

La risposta è sì! La rete che otteniamo è molto più densa di link e ha un andamento scale free. Per le successive analisi verrà usata questa rete in quanto si ritiene più interessante da analizzare.

$$\langle k \rangle = 72,88$$

$$N = 9852$$

$$\Delta = 0,0074$$

$$L = 359022$$

$$\sigma_{in}^2 = 12228,08$$

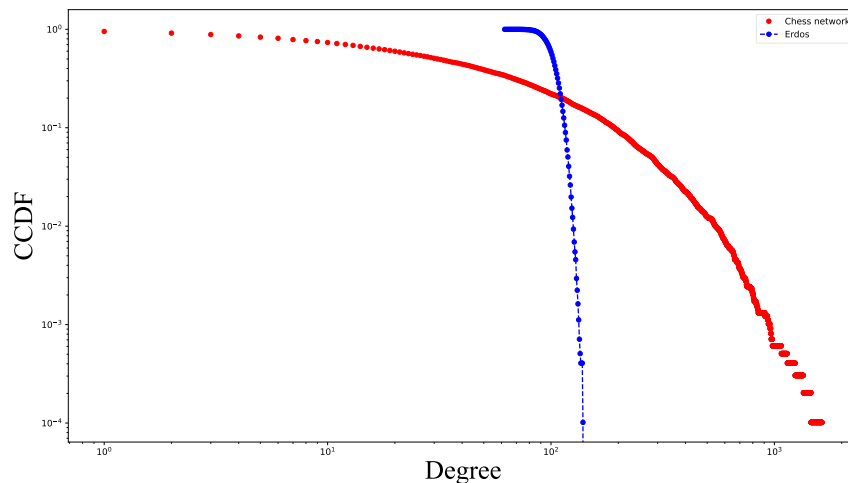
$$Moda = 1$$

$$Mediana = 32$$

$$k_{min} = 1$$

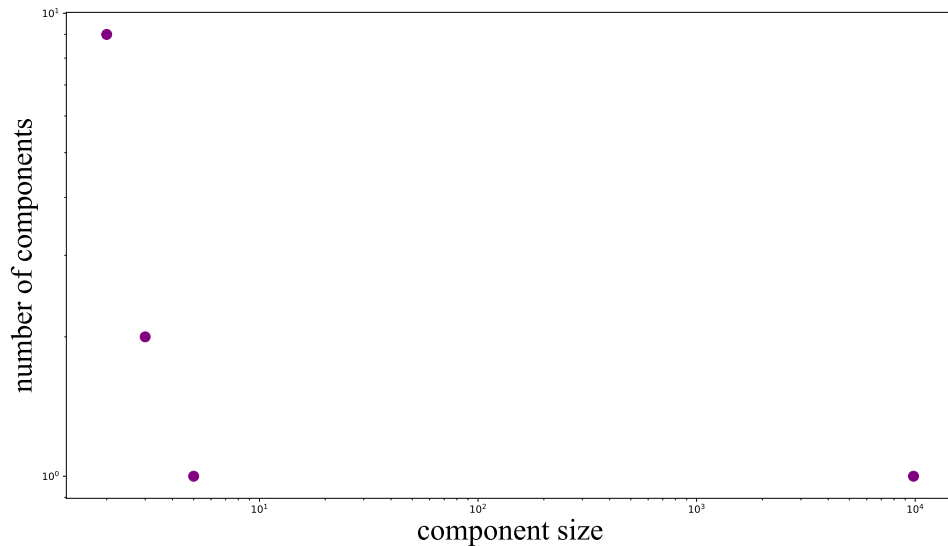
$$k_{max} = 1633$$

$$quantile_{95} = 284$$



# Giant Component?

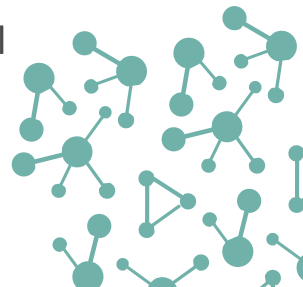
La presenza dell'outlier ci fa capire che esiste la giant component, inoltre  $\langle k \rangle > \ln N$  quindi siamo nel regime 'grafo connesso'



$$N_G = 9823$$

$$S_G = 0.9970$$

Le componenti  
connesse rimanenti  
sono piccole  
componenti composte  
da minimo 2 nodi e al  
più 5.



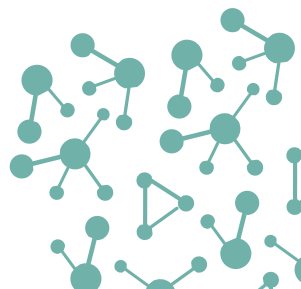
# La rete è assortativa per rating e lingua?

Il valore di assortatività per rating trovato indica che i giocatori con punteggio simile tendono a giocare di più tra loro. Mentre per quanto riguarda la lingua la rete non risulta essere né assortativa né disassortativa, dimostrando che non è necessario parlare la stessa lingua per giocare.

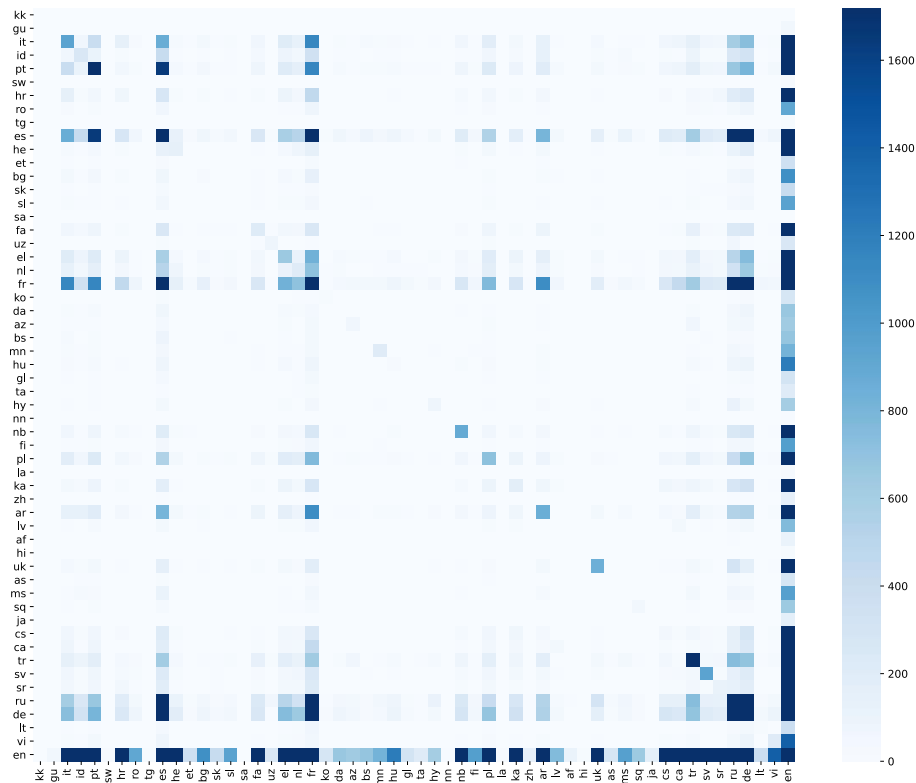
$$\rho(\textit{rating left}, \textit{rating right}) = 0.35$$

$$Q(\textit{language}) = 0.035$$

Di seguito è riportata una heat-map per studiare le partite giocate tra le nazioni.

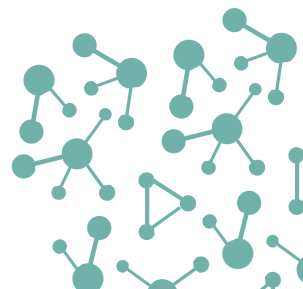


# La rete è assortativa per rating e lingua?

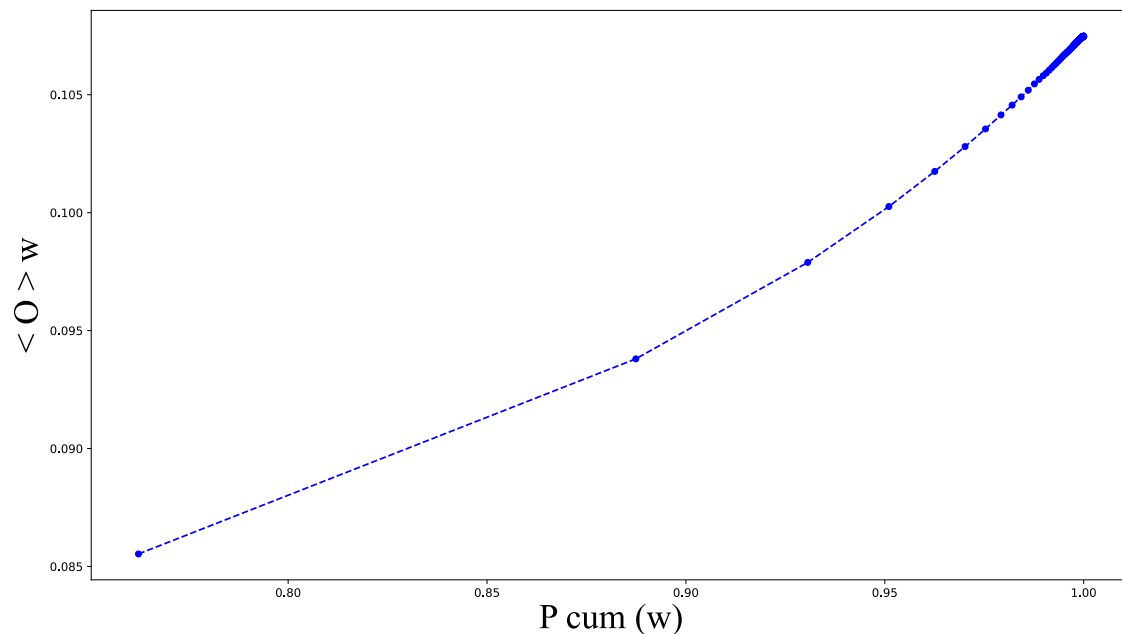


Come visto nella slide precedente la rete non è assortativa per lingua, altrimenti ci sarebbero più link sulla diagonale.

Osservazione interessante:  
il 67 % dei giocatori della rete utilizza la lingua inglese.



# C'è una sovrapposizione del vicinato?



$$O(A, B) = \frac{|N(A) \cap N(B)|}{|N(A) \cup N(B)|}$$

Nel grafico viene mostrato l'andamento dell'overlapping medio dei nodi all'aumentare della forza dei legami.

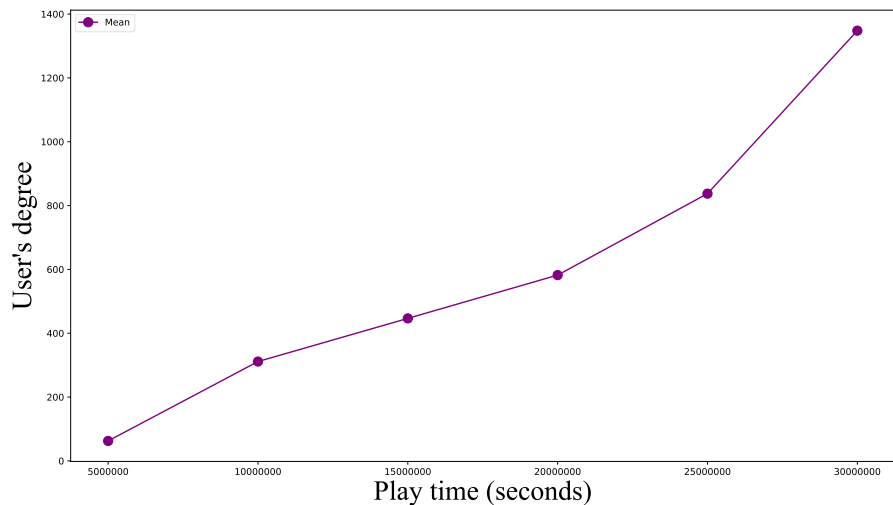
$P_{cum}(w)$  rappresenta la frazione di link con forza dei legami minore o uguale di  $w$ .  $\langle O \rangle_w$  rappresenta l'overlapping medio tra gli endpoint il cui link ha forza minore di  $w$ .



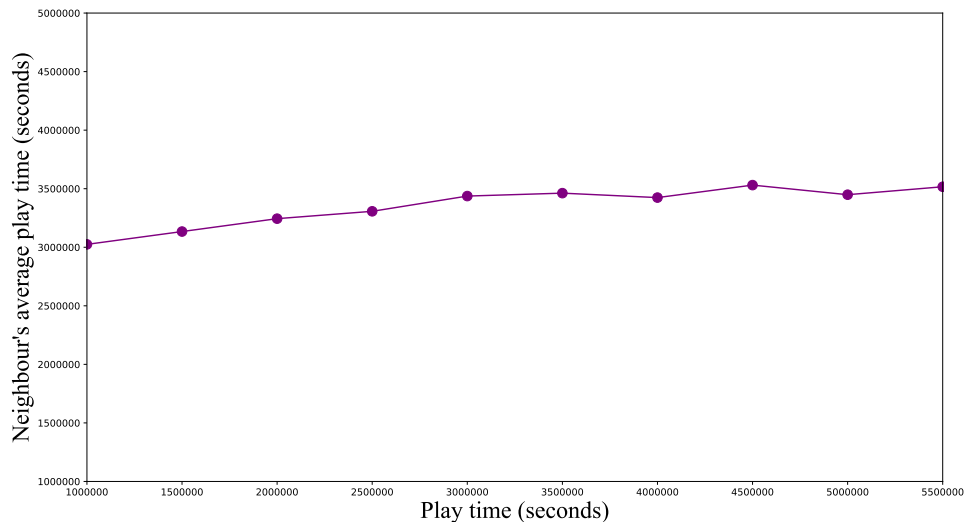
# Engagement correlations

Per misurare l'engagement è stato utilizzato il *playTime*, informazione che rappresenta il tempo trascorso in partita dell'utente.

All'aumentare del playTime aumenta il grado medio degli utenti

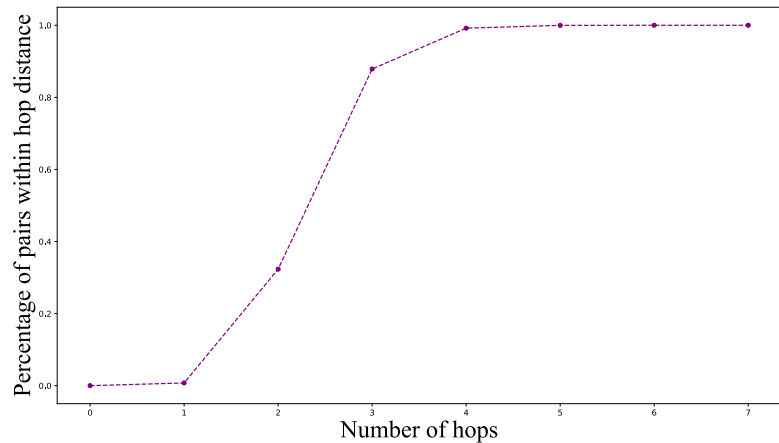


Finché ci sono utenti con un playTime di 35,000,000 in media osservo che i loro amici hanno un playTime maggiore.



# Small world?

Le statistiche riportate ci fanno subito comprendere la natura small world della rete. Oltre al diametro è stata plottata la distribuzione della funzione  $N()$  che dato un numero di hop restituisce la percentuale di coppie con al massimo tale numero di hop.








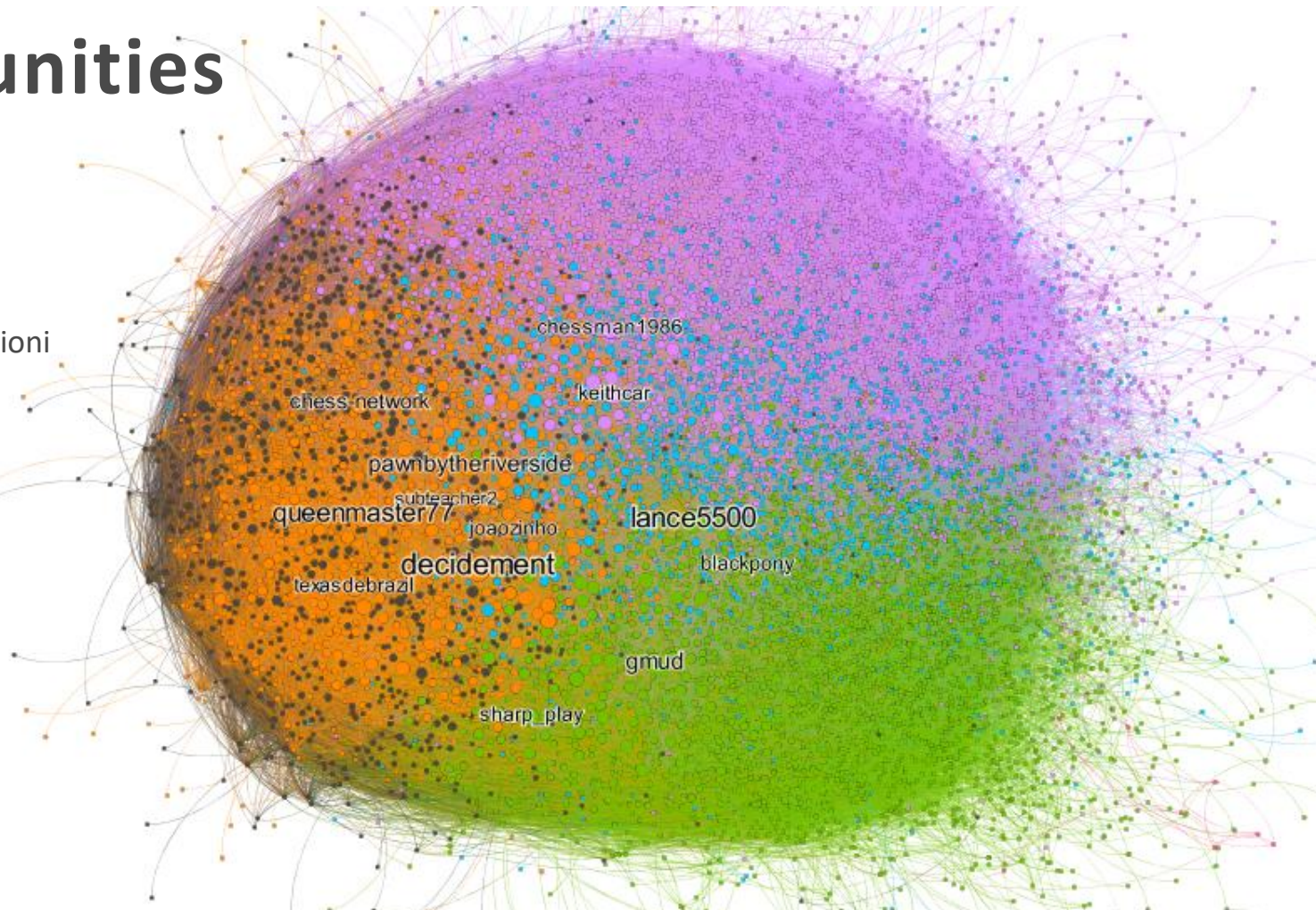
$N, L$	Clustering coefficient	$\langle d \rangle$	$\frac{\log(N)}{\log \langle k \rangle}$	$d_{max}$
9852, 359022	0,13	2,79	2,14	7

# Communities

$$Q = 0,273$$

Mediana distribuzioni  
Rating blitz:

1. 1507 
2. 1839.5 
3. 1624.0 
4. 1946.0 
5. 2086.0 





# Link prediction



Non avendo a disposizione dati temporali sul momento in cui sono state create le connessioni, è stata introdotto in modo artificiale una 'temporalità'. Per far ciò l'insieme di tutti gli archi è stato suddiviso in due parti  $G_0$  (60% del totale),  $G_1$  (40% del totale).

Per creare i data point positivi sono state prese le coppie connesse in  $G_1$  i cui estremi 'erano' presenti in  $G_0$ , mentre per creare i data-point negativi sono state prese le coppie non connesse in  $G_0$  che si sono connesse neanche in  $G_1$ . I data-point sono stati quindi uniti in un unico Data Frame per poi effettuare uno split tra training e testing.

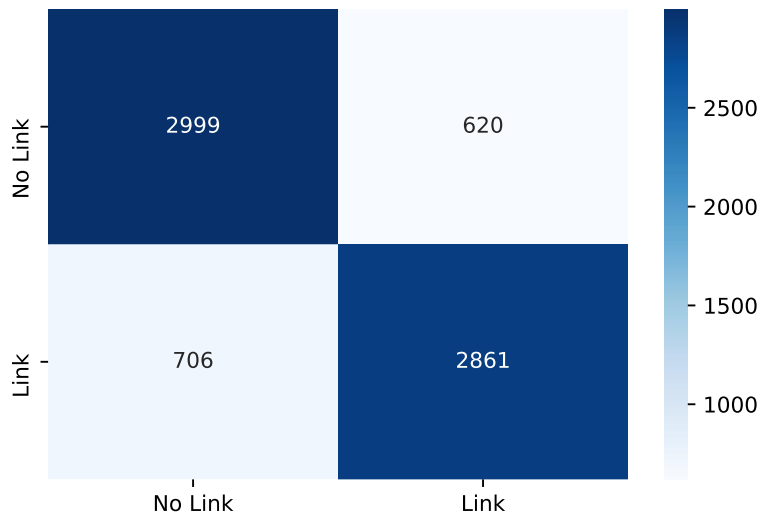
Ricordiamo che ci troviamo nel campo dell'apprendimento supervisionato, quindi scegliamo RandomForest come classificatore, e facciamone il tuning degli iperparametri usando *RandomizedSearchCV*. I migliori iper-parametri rilevati sono:

n\_estimators: 93,  
min\_samples\_split: 2,  
min\_samples\_leaf: 5,

max\_leaf\_nodes: 52,  
max\_features: auto,  
max\_depth: 8, 'bootstrap': True



# Link prediction

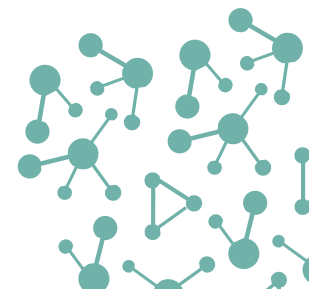


*Accuratezza: 0.815*

*F1-score: 0.812*

*Sensibilità: 0.802*

*Specificità: 0.829*



# Grazie!

Domande?



<https://github.com/vincenzoconv99>