



# komoot

## NETWORK ANALYSIS

Riccardo Carissimi, 962766. Progetto per il  
corso di Social Media Mining A.A. 2022-23



[Link repo GitHub](#)

# Il Contesto

Cos'è *Komoot* e per cosa viene usato, quali sono le domande di ricerca e perché è interessante

# Cos'è Komoot

*Komoot* è una piattaforma per pianificare e tenere traccia degli **sport all'aperto**. Nello specifico è un *route planner* per il **ciclismo**, l'hiking e sport affini. Negli anni si è evoluto: tra le funzionalità interessanti è possibile usarlo come navigatore, ma anche come **social network** per condividere i propri itinerari e progressi col mondo.

Nonostante sia un'azienda senza una sede di lavoro fisica, è stata fondata in **Germania**. Questo dettaglio ritornerà nella nostra analisi.

È il principale competitor di Strava, piattaforma statunitense che fornisce le stesse funzionalità base.





# La parte social

Nonostante la user base molto settoriale, i competitor e il focus dell'app sia un altro la parte social è abbastanza **sviluppata**, caratterizzata da follower e followings.

Sulla piattaforma sono presenti circa **30 milioni** di utenti<sup>1</sup> tra cui alcuni veri e propri **influencer** e account istituzionali.

La piattaforma viene molto usata per condividere gli itinerari con la community e per mostrare agli altri utenti i propri successi.

1. <https://www.komoot.com/it-it/jobs>





# Domande di Ricerca

1. Gli utenti della stessa nazione tendono a creare più spesso link tra di loro? E quelli con lo stesso sport preferito? Considerando invece i chilometri percorsi?
2. Il numero di chilometri percorsi influisce sulla centralità? Ci stiamo chiedendo se gli utenti che **pubblicano** di più sono anche più centrali.
3. C'è una correlazione tra la centralità e la nazionalità degli utenti?
4. Quali sono gli sport più diffusi in ogni paese?
5. Quali sono i paesi da cui provengono gli utenti? Il fatto che sia una piattaforma tedesca influisce su questo dato?
6. Proveremo anche a predire la formazione di nuovi link → **link prediction**

# Data Gathering

Come sono stati raccolti tutti i dati necessari per la costruzione della rete e la sua analisi

# Un (*lento*) approccio iniziale

Ho cominciato analizzando la piattaforma alla ricerca di eventuali API. Ho trovato delle API **senza documentazione** (o quasi). Queste API non erano in grado di fornire i dati di cui avevo bisogno riguardo i follower e followings di ogni utente. Tuttavia queste informazioni sono accessibili pubblicamente.

**PROBLEMA!** Le informazioni vengono caricate dinamicamente e facendo una semplice richiesta HTTP non ottenevo tutti i following.

Mi ero rassegnato a usare **Selenium**, una famosa libreria di *browser automation*. Purtroppo riuscivo ad analizzare meno di 10 utenti al minuto.

Grazie al consiglio di un'amica e con un po' di intuito sono riuscito a sfruttare le chiamate API della pagina web. Sono passato da meno di 10 a circa 200 utenti analizzati al minuto.



# Gli attributi degli utenti

Ho pensato di aggiungere delle informazioni per ogni nodo:



- **nazionalità** → purtroppo il paese di provenienza non è un'informazione fornita né dal sito né dalle API. Tuttavia possiamo ottenere i *tour* pubblici degli utenti, cioè i percorsi che hanno seguito. Ho estratto la nazionalità di un utente considerando la **moda del paese degli ultimi 5 tour**.
- **sport preferito** → con lo stesso approccio ho estrapolato lo sport preferito considerando la moda degli sport degli ultimi 5 tour. Ad ogni tour, infatti, è associato lo sport per cui è stato pensato.
- **km percorsi** → questa informazione era disponibile nella versione web della piattaforma ma non nelle API. Ho sfruttato le librerie *Requests* e *BeautifulSoup* per ottenere questo dato.



**Riccardo**

2

3

Follower Seguiti

Segui



Profilo

Tour

6



## Informazioni su Riccardo

Distanza percorsa

216 km

Tempo in movimento

11:20 h

## Attività recenti



# L'implementazione iniziale

Purtroppo ottenere queste informazioni era troppo lento. Avessi voluto ottenere gli attributi di tutti i nodi avrei dovuto aspettare più di 50 ore: le risposte alle richieste HTTP erano troppo lente. **Decisamente troppo!**

```
1  resp = session.get(url=url, params=params)
2  data = resp.json()
3
4  for tour in data['_embedded']['tours']:
5      countries.append(get_country(tour['start_point']['lat'], tour['start_point']['lng']))
6      sport_type.append(tour['sport'])
7
8  return max(set(countries), key=countries.count), max(set(sport_type), key=sport_type.count)
```



# Facciamolo multithread

```
1  from threading import Thread, Lock
2  from time import sleep
3
4  procs = []
5  lock = Lock()
6  threads = 40
7  nodes = g.nodes()
8
9  step = int(len(nodes)/threads)
10 for i in range(threads-1):
11     print(f"Thread Started - Range [{i*step}-{(i+1)*step}]")
12     p = Thread(target=bounded_infos, args=(lock, g, i*step, (i+1)*step, nodes))
13     procs.append(p)
14     p.start()
```



40 Python threads



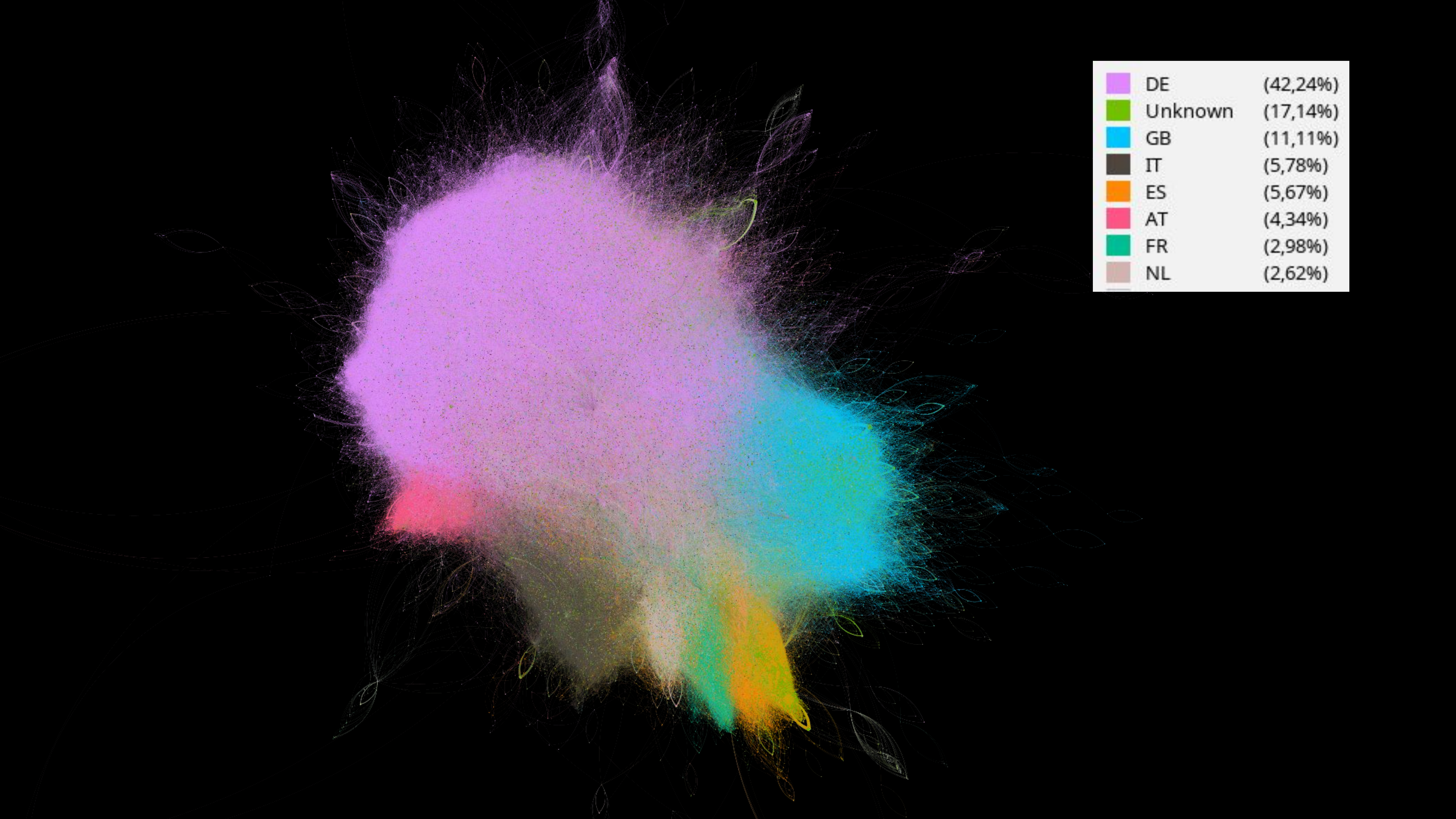
My laptop



# Analisi della rete

Vediamo le proprietà fondamentali della rete, la centralità e rispondiamo alle domande che ci siamo posti







# Proprietà fondamentali della rete

Order:	74139	Weakly connected components: 1
Size:	1113417	Strongly connected components: 283
Density:	$0.2025 \times 10^{-3}$	Largest strong connected component: 73571
Reciprocity :	0.4497	Average clustering : 0.0272
		Average local clustering: 0.2665

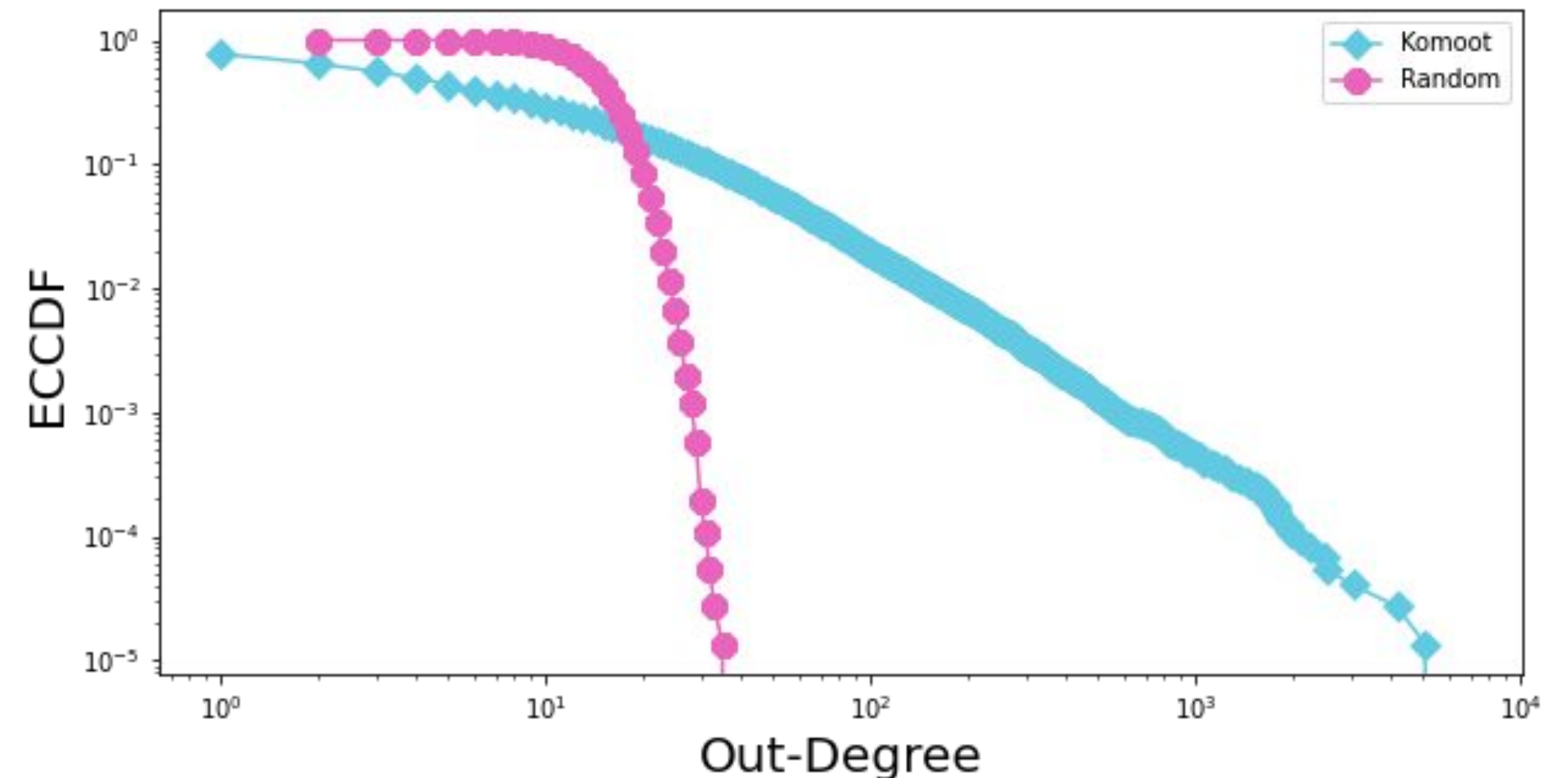
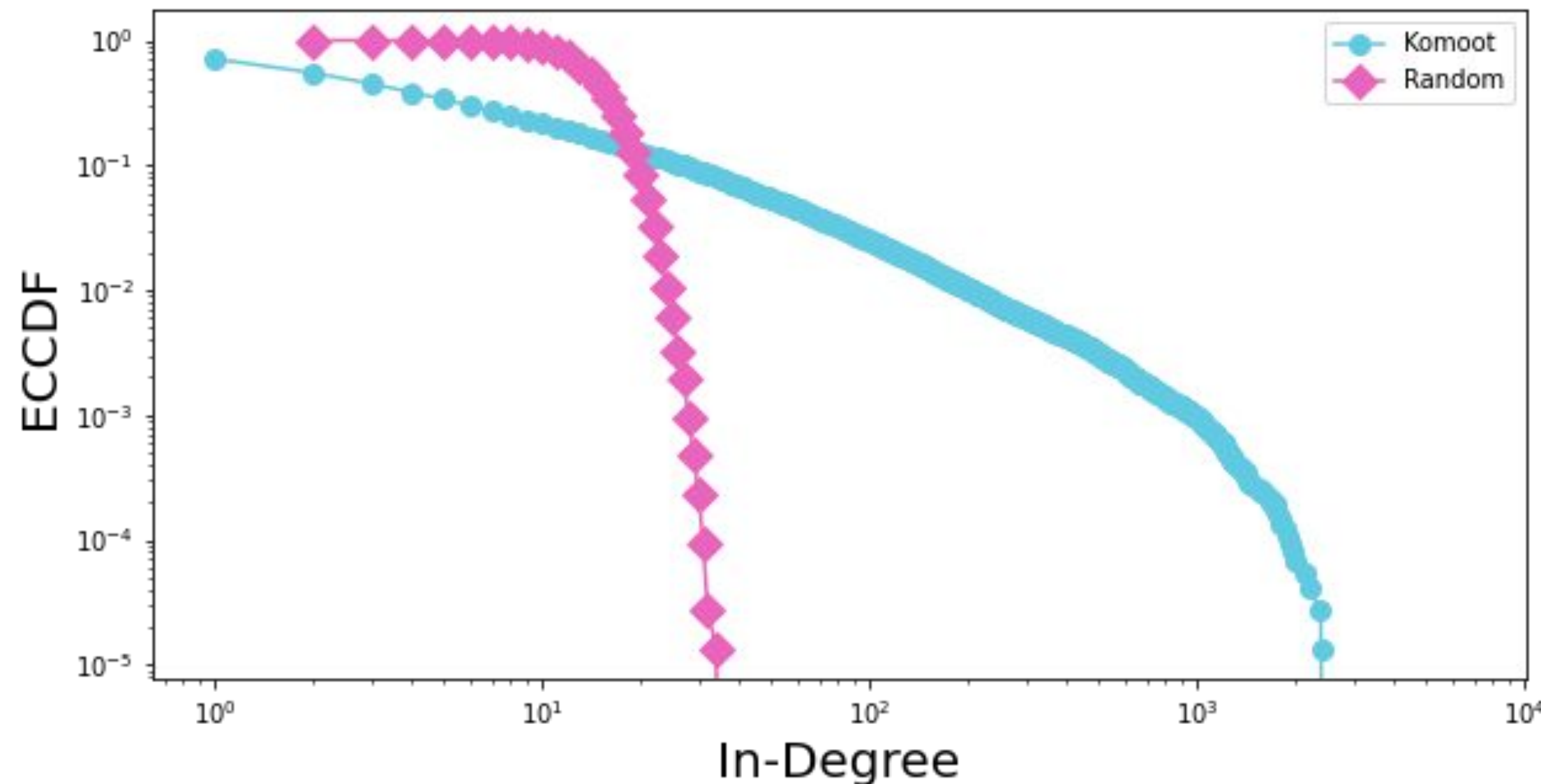
# Ulteriori informazioni sul grado

## In-degree:

Grado medio: 15.017966252579614  
Standard deviation: 67.15731144375731  
Median: 3.0  
Min: 1  
Max: 6512

## Out-degree:

Grado medio: 15.017966252579614  
Standard deviation: 60.01526038966208  
Median: 4.0  
Min: 1  
Max: 6531

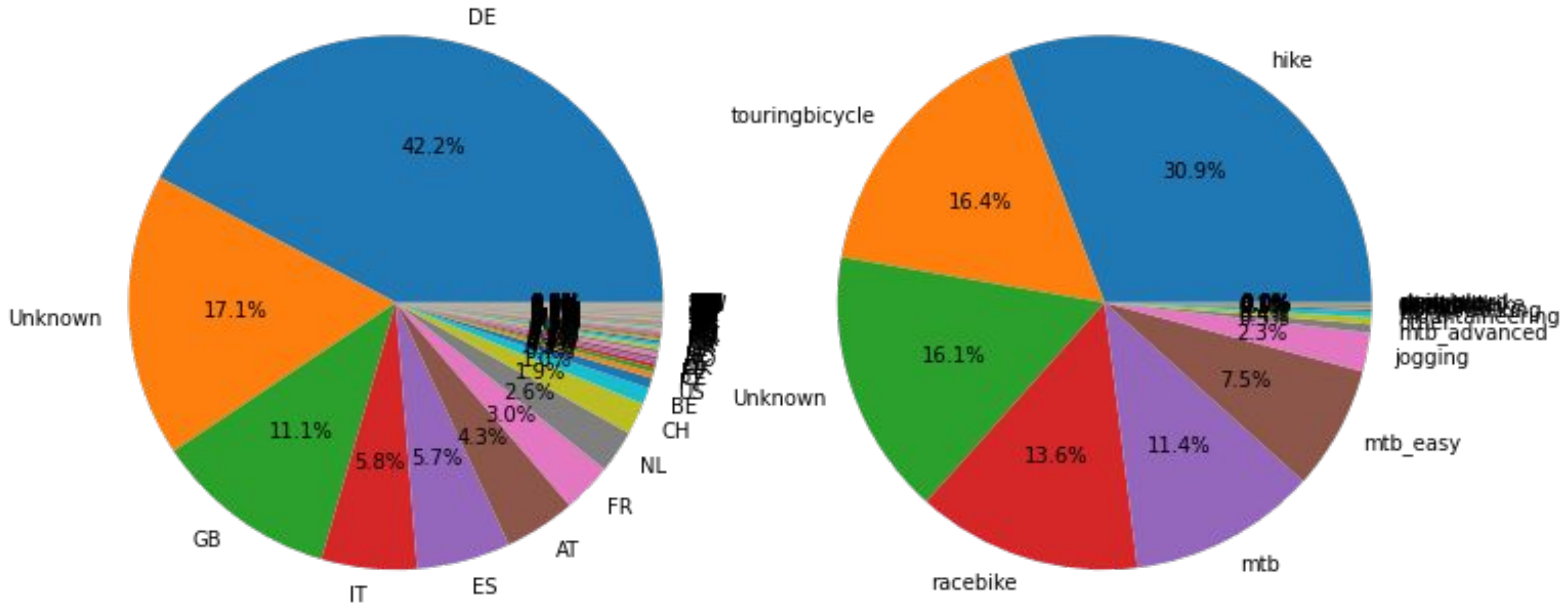




# Utenti più centrali

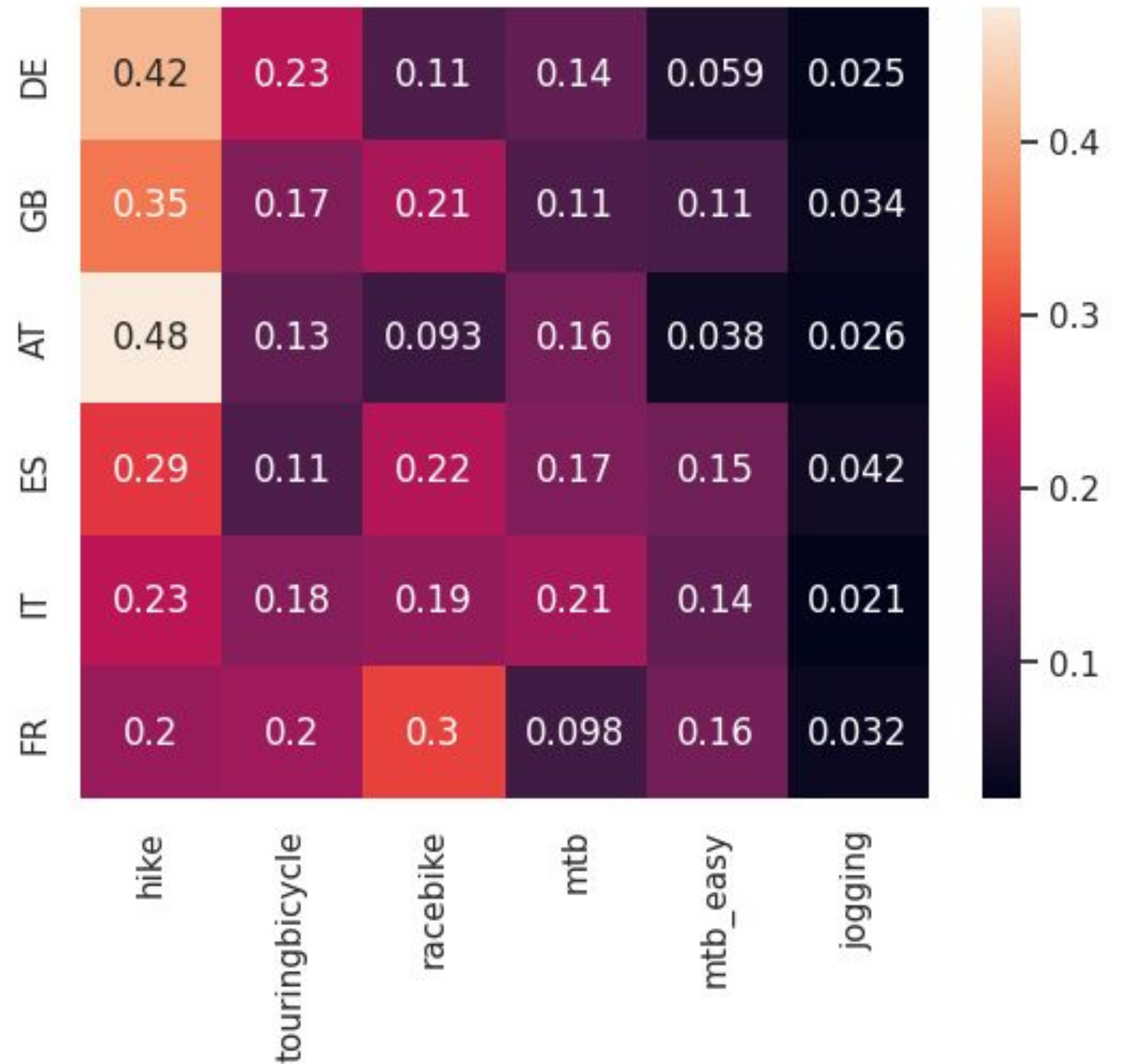
In-degree centrality	Out-degree centrality	Eigenvector centrality
Komoot	Pasquale Albano	Komoot
Orbit360	Thomas	Bea
Adventurer Nic	Steffen	Carsten
Xavier Farràs	Michael Hofmann	Söhni
Katherine Moore	Adventurer Nic	Ewa und Christof

# Utenti per nazionalità e sport















# Correlazione tra nazionalità e sport



# Correlazione tra centralità e nazione

Nazione	Correlazione correlazione punto biseriale
Germania 	0.0561
Montenegro 	0.0325
Kenya 	0.0321
Costa d'Avorio 	0.0177
Austria 	0.0173

Nazione	Correlazione correlazione punto biseriale
Belgio 	-0.0037
Norvegia 	-0.0038
Polonia 	-0.0058
Regno Unito 	-0.0224
Sconosciuta 	-0.0833



# Altre risposte

**Correlazione tra centralità e km percorsi:**

0.1834

La correlazione c'è ma non è così importante quanto è ragionevole pensare. Per essere *popolari* sulla piattaforma non basta pubblicare tanti itinerari.

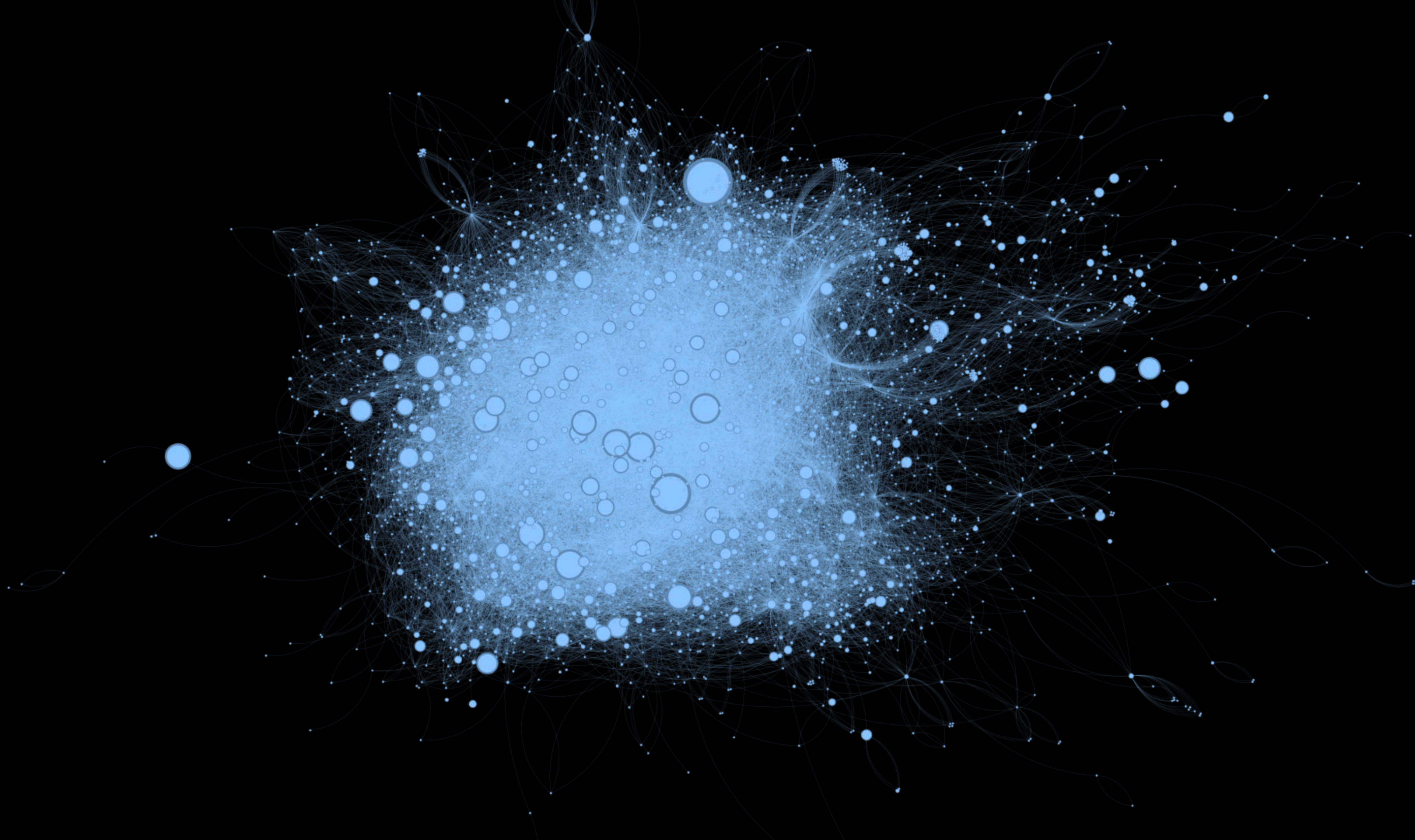
## Assortatività

- sulla nazionalità 0.3928
- sullo sport preferito 0.1953
- sui km percorsi 0.0178
- sul grado del nodo -0.0985

# Analisi della rete italiana

Analizziamo brevemente alcune informazioni  
interessanti sulla sottorete composta dagli utenti  
italiani







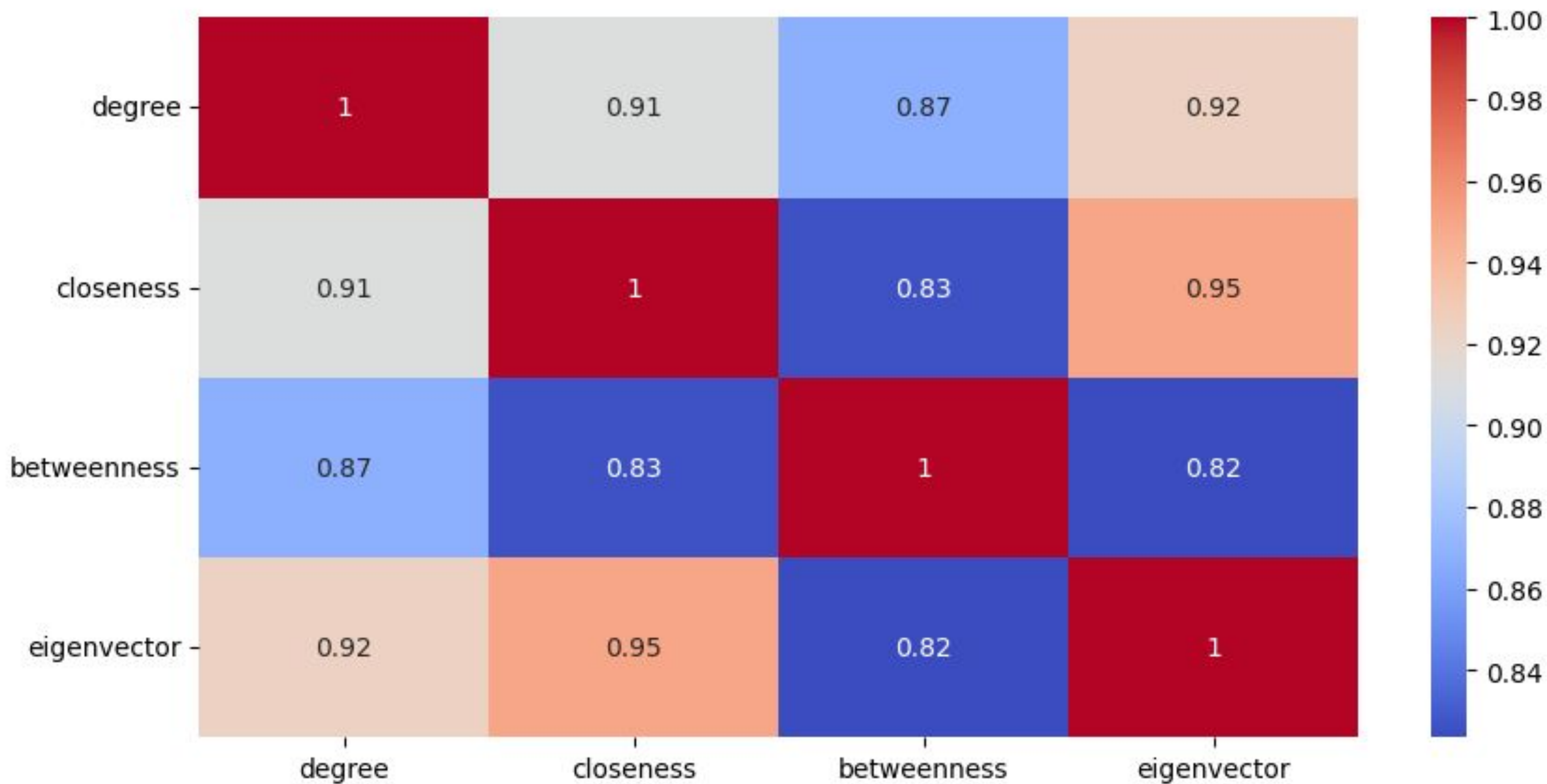
# Informazioni sulla sottorete

Order:	4286
Size:	35676
Density:	$0.1942 \times 10^{-2}$
Diametro:	8
Grado medio:	8.3238
Standard deviation:	29.7475
Median:	2.0
Min:	0
Max:	543

In-degree centrality
Omar Di Felice
Gravel Club
Elena Martinello
Niccolò Varanini
Cento Canesio



## Pearson correlation



# Link prediction

Sfruttiamo le conoscenze di machine learning per provare a predire i link tra nodi



# Costruzione del dataset

Ho cominciato creando il dataset. L'ho poi diviso nella **feature matrix** e nel **vettore delle label**.

Ho poi diviso il dataset in training e testing (adottando una politica del 30%) ed effettuato lo scaling delle feature.

Ho addestrato due modelli: uno basato su **Logistic Regression** e uno su **Random Forest**.

	jacca rd	rai	aai	pref	same_co untry	same_ sport	km_ diff
0	0.08 1478	1.47 7435	37.08 6708	148 914 6	0	0	180 84
1	0.00 2232	0.00 4735	0.313 497	184 17	1	0	180 84
...	...	...	...	...	...	...	...
196 88	0.00 0000	0.00 0000	0.000 000	5	0	0	0
196 89	0.00 0000	0.00 0000	0.000 000	24	0	0	821 2

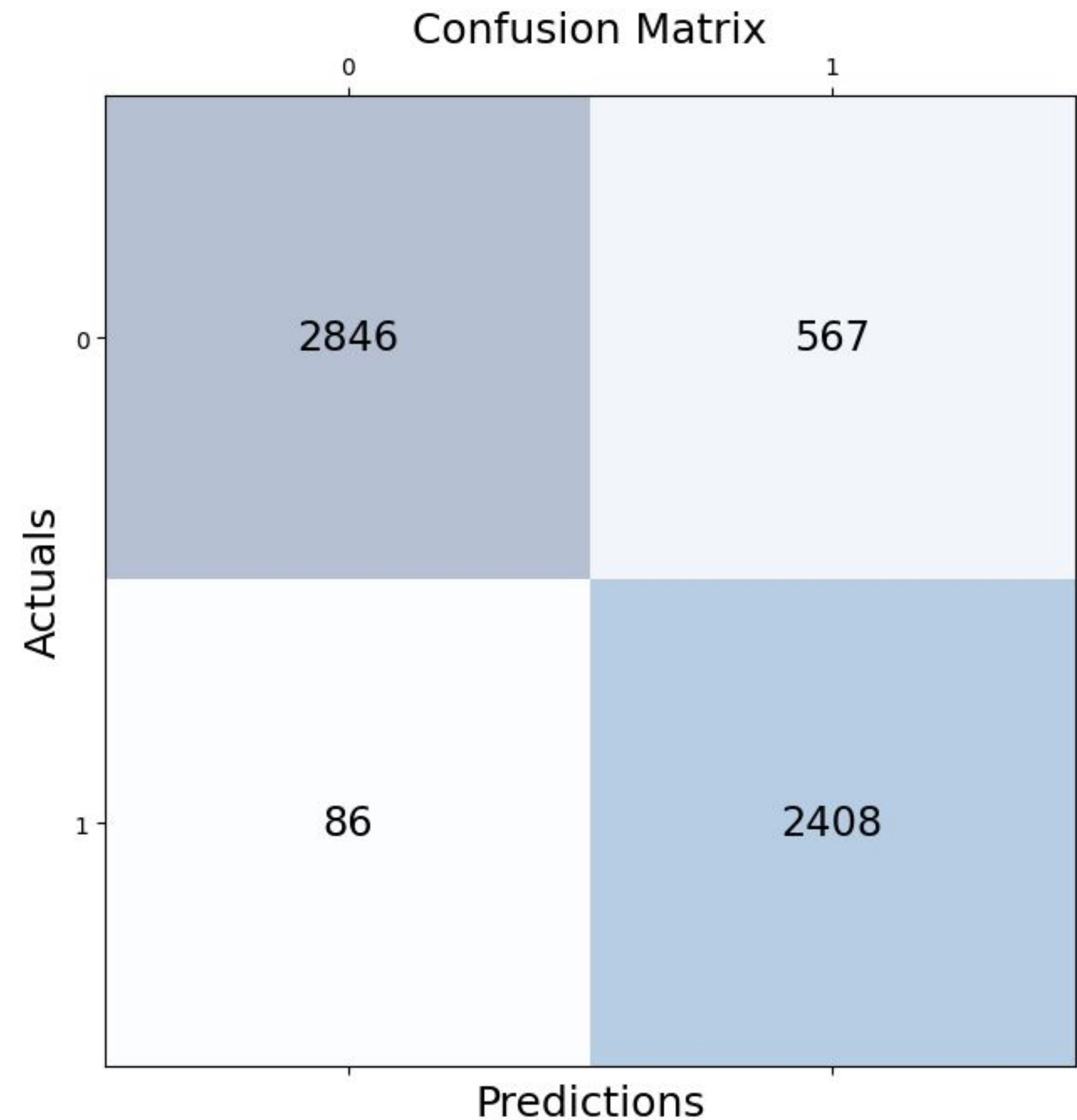
# Logistic regression

**Accuracy:** 0.8895

**Recall:** 0.8094

**Precision:** 0.9655

**F1 Score:** 0.8806





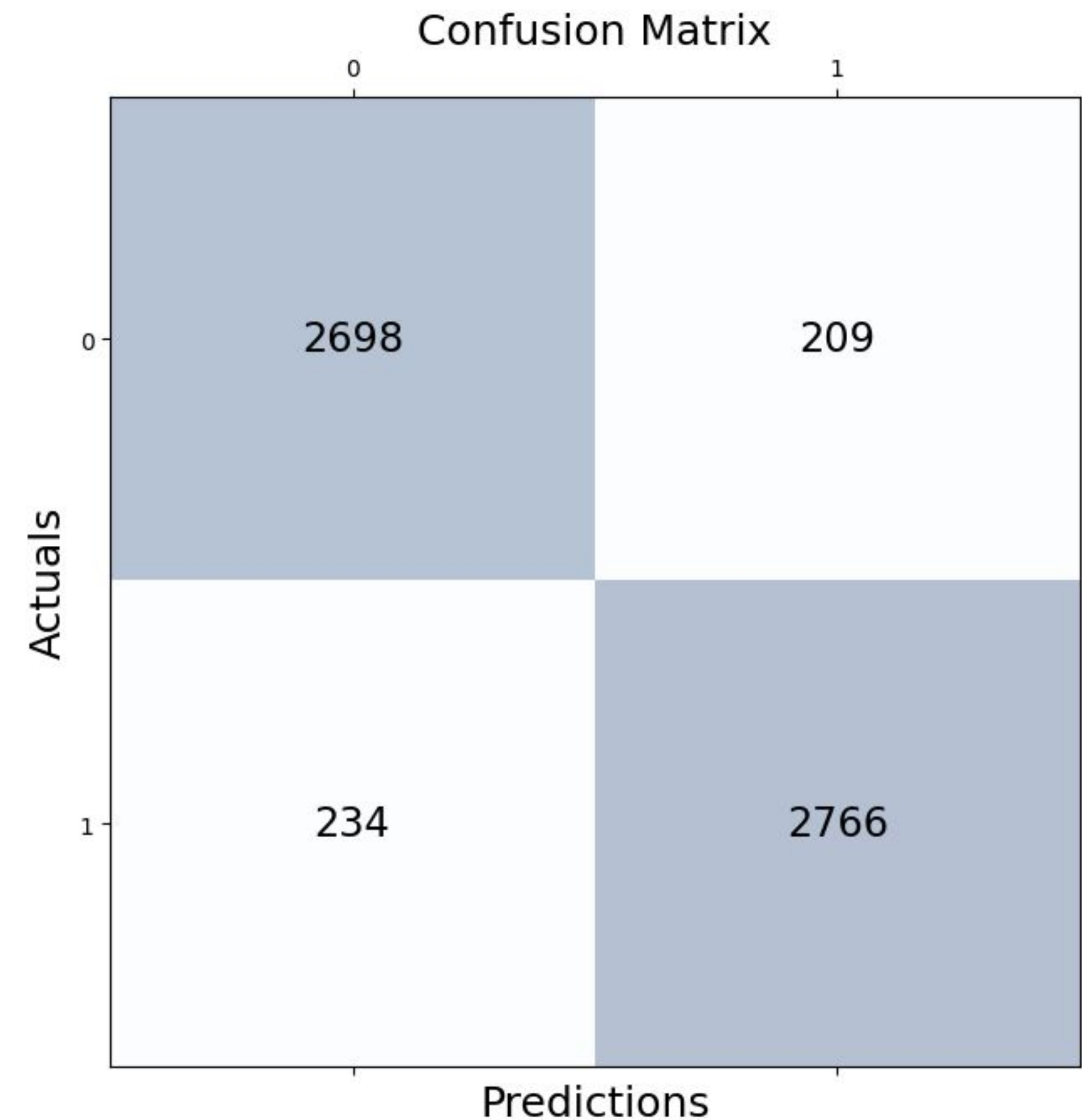
# Random forest

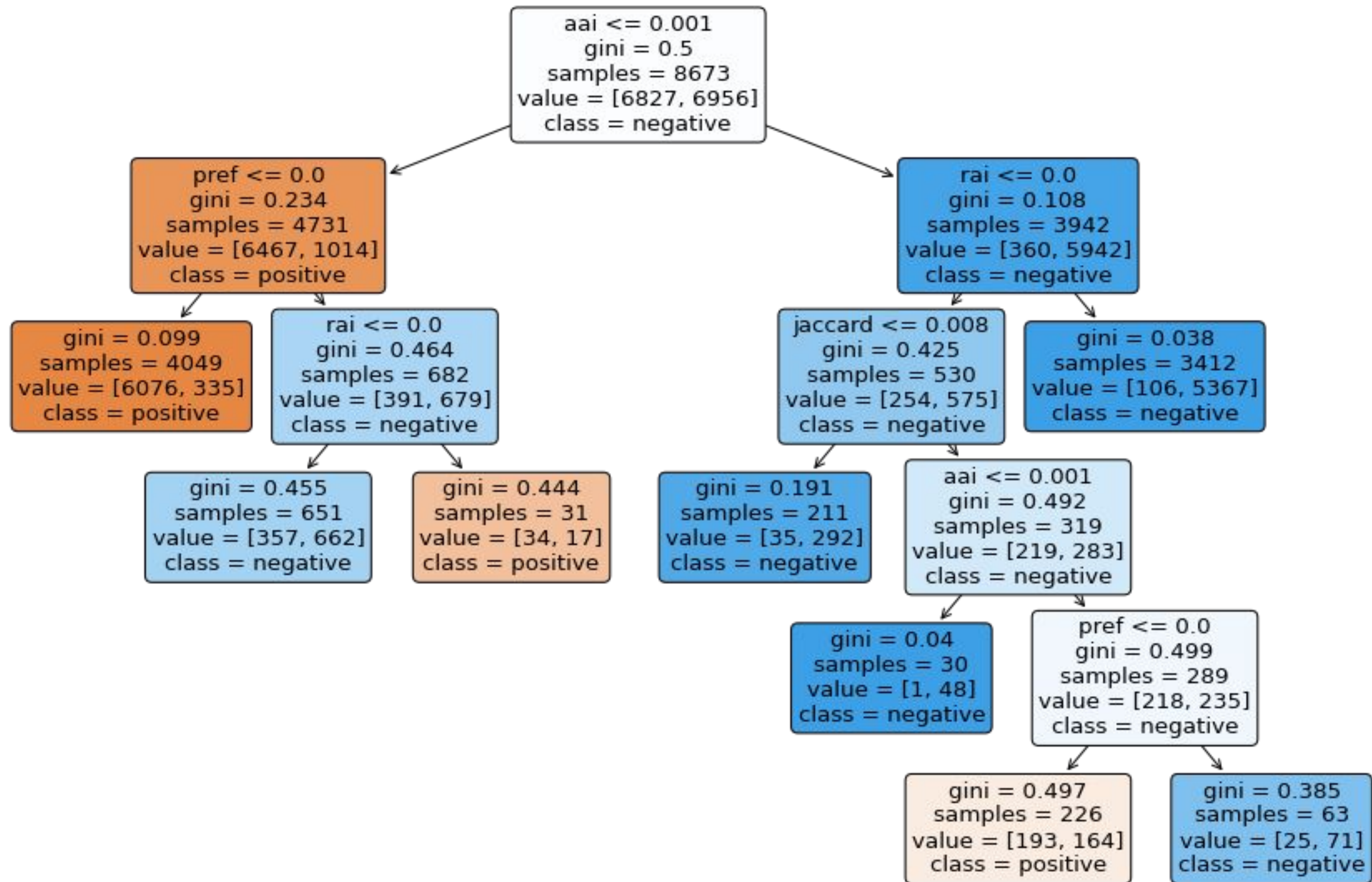
**Accuracy:** 0.9250

**Recall:** 0.9297

**Precision:** 0.9220

**F1 Score:** 0.9259







# Sviluppi Futuri

Come può ancora evolvere questo progetto?

# Sviluppi futuri

1. Si potrebbero analizzare altri dati riguardanti il tour, come il tipo di superficie, o il dislivello. Queste informazioni potrebbero portare ad analisi più puntuali
2. Si può entrare più in dettaglio sulla zona geografica dell'utente, fino ad arrivare al livello di regioni o singoli comuni. Questo potrebbe portare a classificatori molto più accurati per la link prediction
3. Si può analizzare il numero di tour a cui due amici hanno preso parte insieme
4. Potrebbe essere interessante analizzare la correlazione tra la centralità e la distanza geografica dei propri tour: un utente che viaggia molto e crea itinerari in giro per il mondo è più seguito?



# Grazie!

Riccardo Carissimi, 962766. Progetto per il corso di Social Media Mining A.A. 2022-23



[Link repo GitHub](#)

Tema delle slide ispirato al progetto analogo di Margherita Pindaro. [Link al progetto.](#)