Università
Ca'Foscari
Venezia

Twitch Network Analysis

Autore: Nicola Aggio

Matricola: 880008

E-mail: 880008@stud.unive.it

Materia: Social Network Analysis (A.A. 2021-2022)

Università
Ca'Foscari
Venezia

INDICE

1. OVERVIEW DELL'ANALISI

Lo scopo di questo progetto è quello di sviluppare un'analisi originale di una rete sociale mettendo in pratica metodi, strumenti e tecniche di analisi e visualizzazione affrontati durante il corso "Social Network Analysis".
Nel caso specifico di questo progetto, la **ricerca** è stata effettuata su una rete sociale composta da utenti portoghesi della piattaforma di streaming Twitch, ed i **dati**, risalenti al maggio 2018 e consistenti in due file .csv, sono stati scaricati dalla sezione "Stanford Network Analysis Project" (SNAP) dell'università di Stanford. All'interno della rete:

- i nodi rappresentano streamer portoghesi della piattaforma Twitch, e gli attributi associati riguardano le visualizzazioni totali ottenute, il periodo di attività (espresso in giorni), informazioni sull'età dello streamer (maggiorenne o meno) e sulle condizioni contrattuali con la piattaforma (partner o meno);
- gli archi rappresentano le amicizie tra gli streamer.

Il **software** che si è deciso di utilizzare per condurre l'analisi della rete è stato R, per la sua grande versatilità nell'analisi dei dati e, in particolare, per la presenza della libreria "igraph", che permette analizzare e visualizzare grafi in maniera dinamica e non troppo complessa.

2. ANALISI DELLA RETE

In questo capitolo è stata analizzata la rete seguendo questi step:

- visualizzazione della rete ed estrazione di informazioni utili;
- calcolo della centralità della rete attraverso diverse metriche;
- identificazione ed analisi di clusters ottenuti attraverso diversi algoritmi di community detection;
- misurazione di un eventuale fenomeno omofilia all'interno della rete rispetto agli attributi dei nodi.

2.1. VISUALIZZAZIONE DELLA RETE ED INFORMAZIONI UTILI

Il primo passo dell'analisi è stato scaricare i dati dalla piattaforma e memorizzare in R nodi ed archi in due liste distinte, attraverso le quali è

stata poi creata la network. Successivamente, quest'ultima è stata semplificata, eliminando gli archi duplicati ed i self-loops, poco significativi per la nostra ricerca, e si è ottenuta la rete rappresentata nella Figura 1. Nella rappresentazione, la colorazione dei nodi varia in base al numero di visualizzazioni come mostrato in legenda.

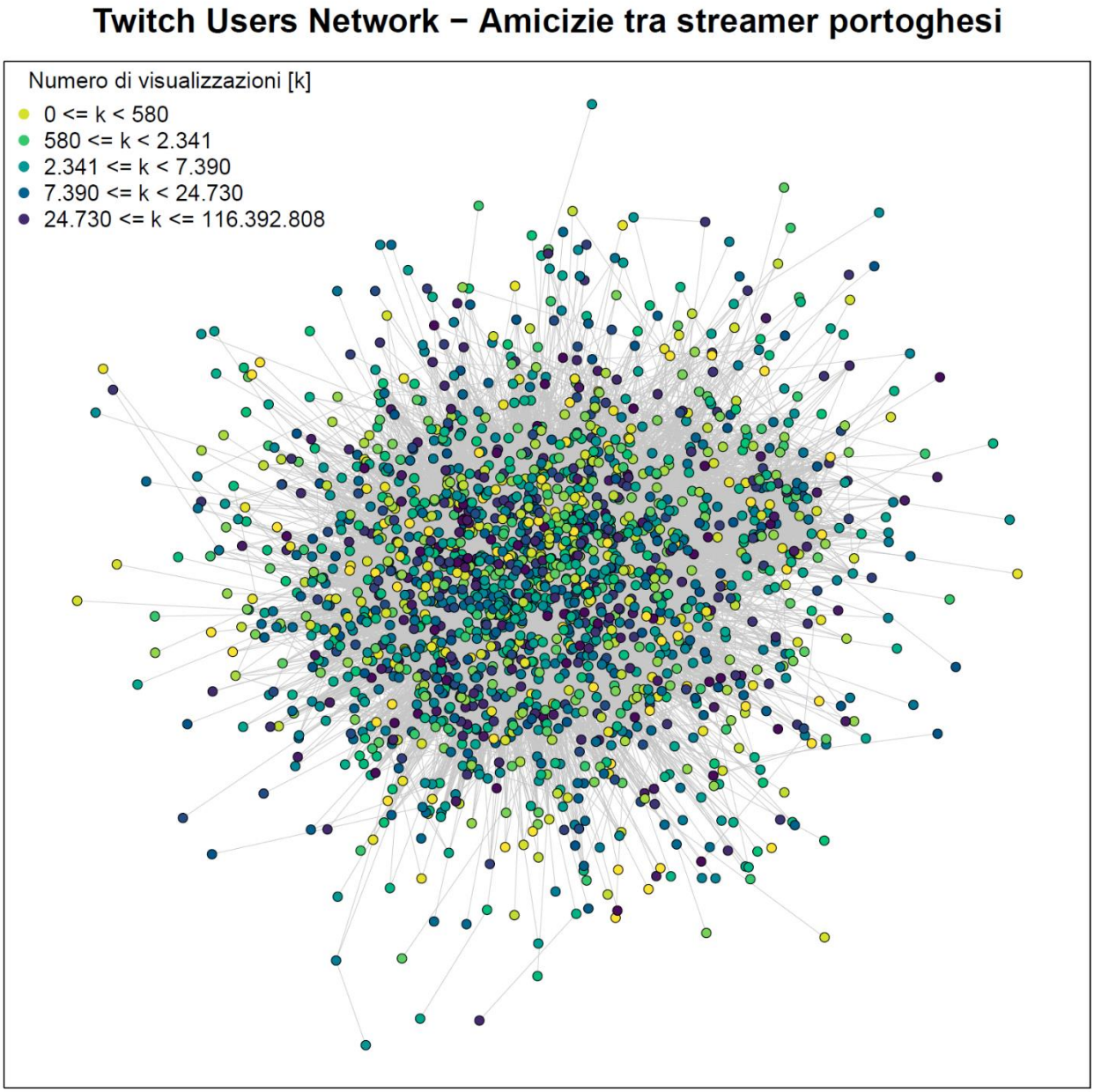

Figura 1 *– Visualizzazione della rete*

Nella Tabella 1 vengono riportate le informazioni più rilevanti della rete ottenuta.

Numero di nodi	1.912
Numero di archi	31.299
Densità	0,017
Average clustering coefficient	0,131

Tabella 1 *- Informazioni sulla rete*

Come si può notare dal valore di densità molto basso il grafo è sparso, mentre l'average clustering coefficient mostra che la probabilità media che i vicini di un qualunque streamer siano collegati è del 13,1%.
All'interno della Tabella 1 non è stato considerato l'average degree come informazione rilevante in quanto, come mostrato nella Figura 2, la degree distribution del grafo segue una Power-Law distribution; pertanto, la media non risulta una metrica significativa.

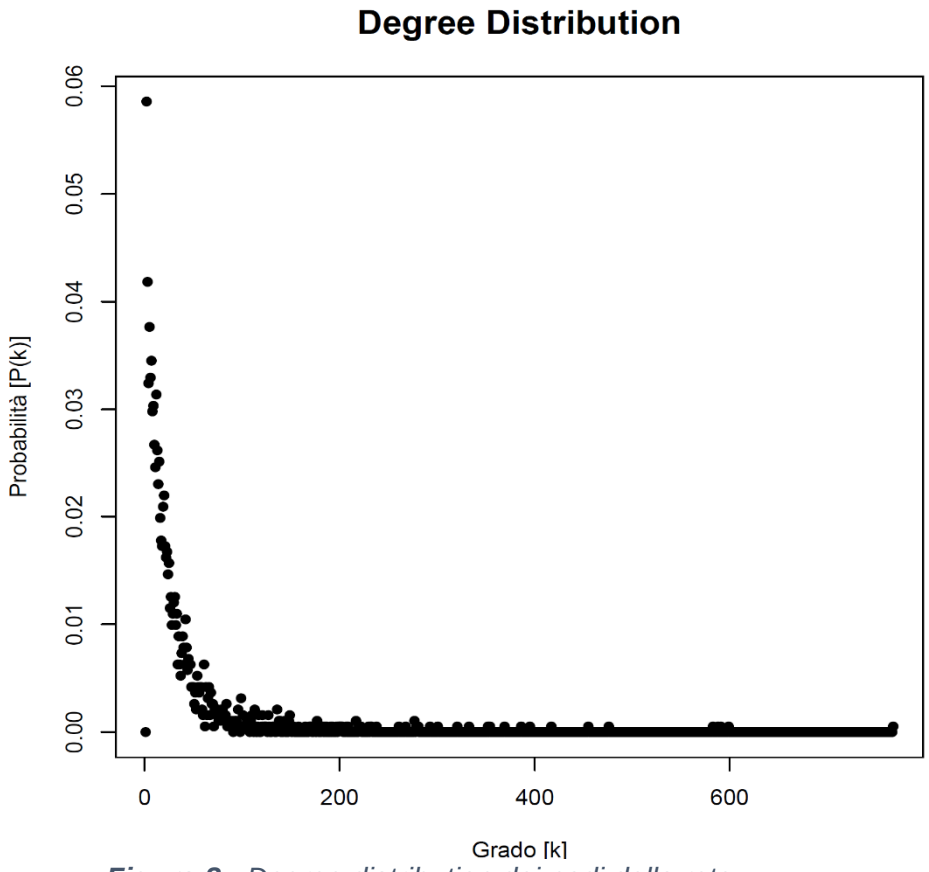

Figura 2 *- Degree distribution dei nodi della rete*

2.2. MISURE DI CENTRALITÀ

Nella seconda parte dell'analisi sono stati applicate diverse misure per calcolare la centralità della rete, in modo tale da individuarne il/i nodo/i più rilevanti e visualizzarli graficamente. Di seguito vengono riportate le metriche utilizzate accompagnate da una breve descrizione:

- Degree centrality: misura di centralità basata sul grado del nodo;
- Eigenvector centrality: misura di centralità basata sull'influenza del nodo all'interno della rete. In generale, nodi con un valore alto di eigenvector centrality sono nodi connessi ad altri nodi che a loro volta possiedono valori alti di centralità. In questo senso, i valori più alti si registrano per nodi all'interno di clique o sottografi molto densi;
- Betweenness centrality: misura di centralità basata sul numero di cammini minimi che attraversano il nodo. Questa particolare misura assegna valori di centralità più alti a nodi che collegano parti disgiunte della rete, e si deriva calcolando, per ogni nodo, il numero di cammini minimi che lo attraversano;
- Closeness centrality: misura di centralità basata sulla lunghezza media dei cammini minimi tra il nodo e tutti gli altri.

  Per ogni metrica utilizzata, vengono classificati nodi in base al loro grado di centralità, e viene visualizzato il risultato direttamente nella rete, in modo tale da poter confrontare anche visivamente le misure. Le Figure 3, 4, 5 e 6 mostrano rispettivamente i risultati di degree centrality, eigenvector centrality, betweenness centrality e closeness centrality.

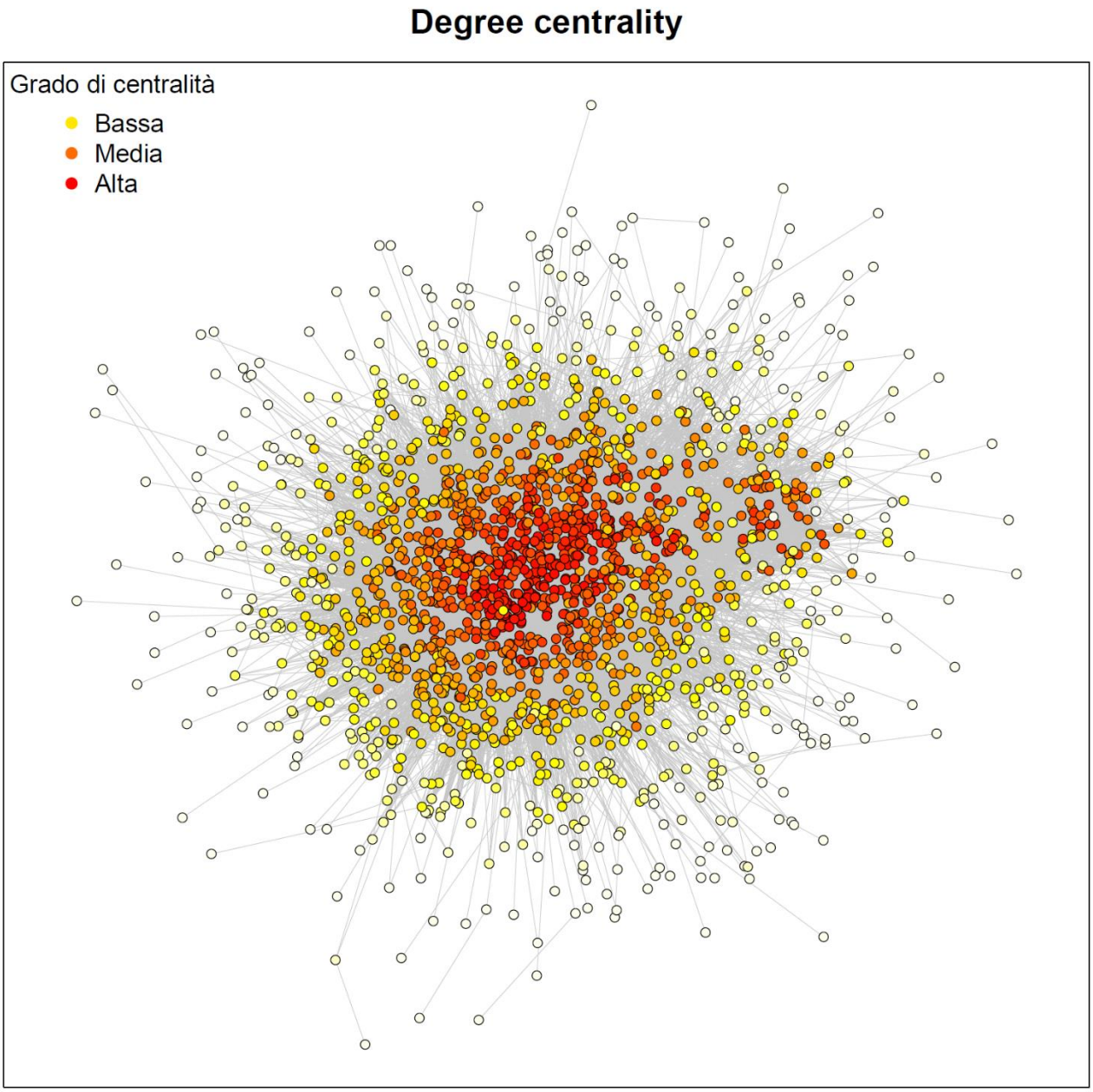

Figura 3 *- Degree centrality*

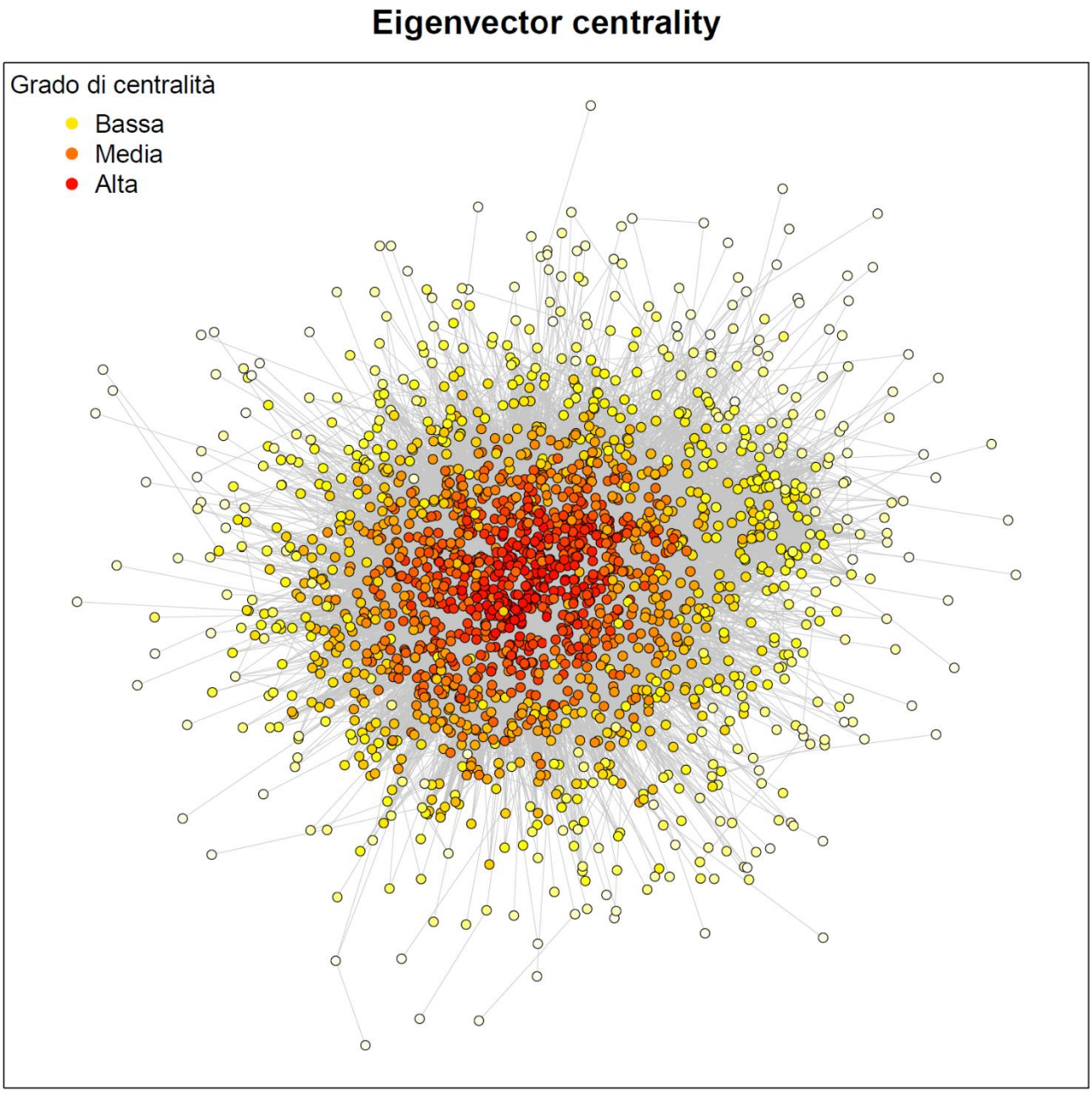

Figura 4 *- Eigenvector centrality*

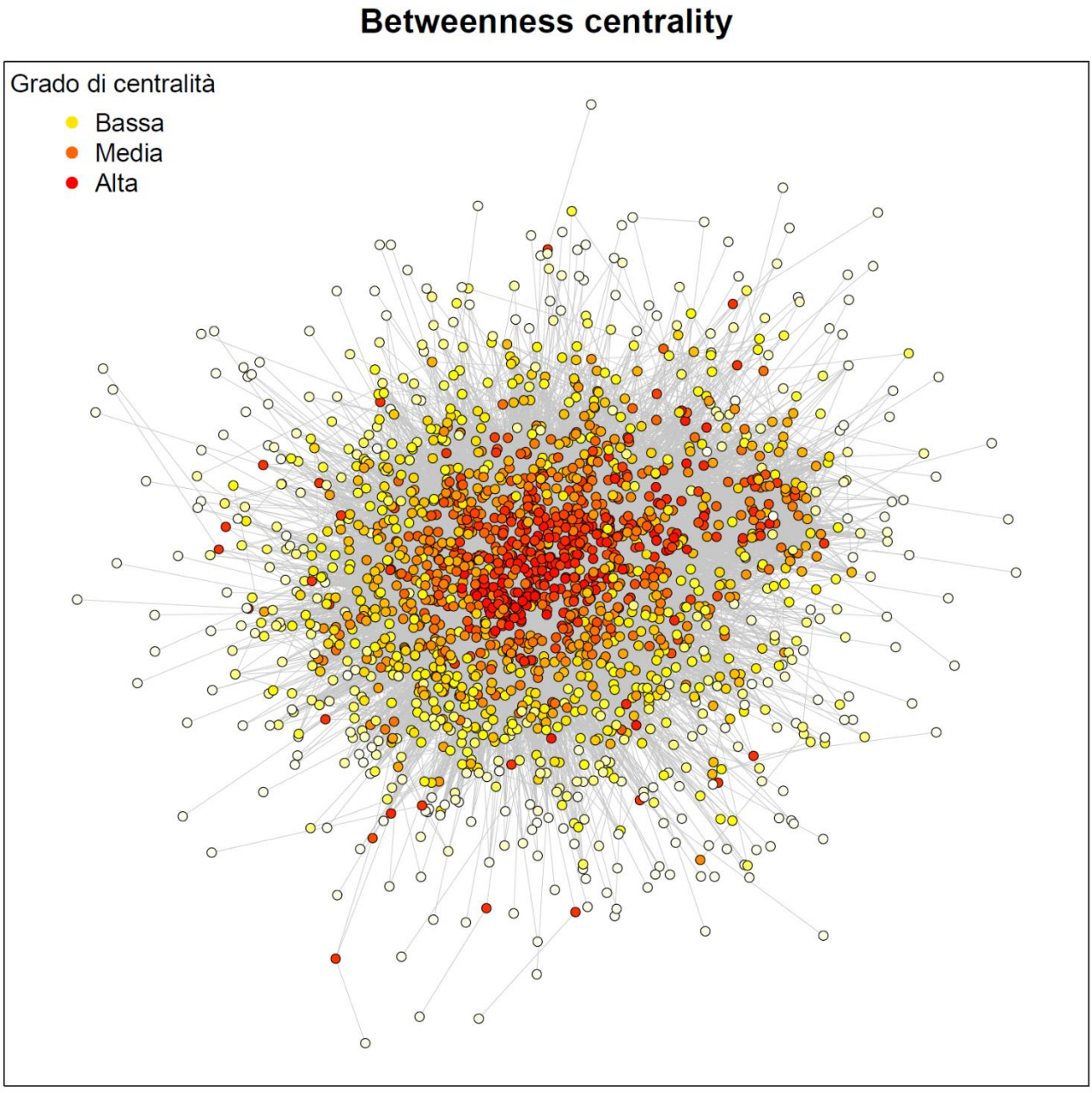

Figura 5 *- Betweenness centrality*

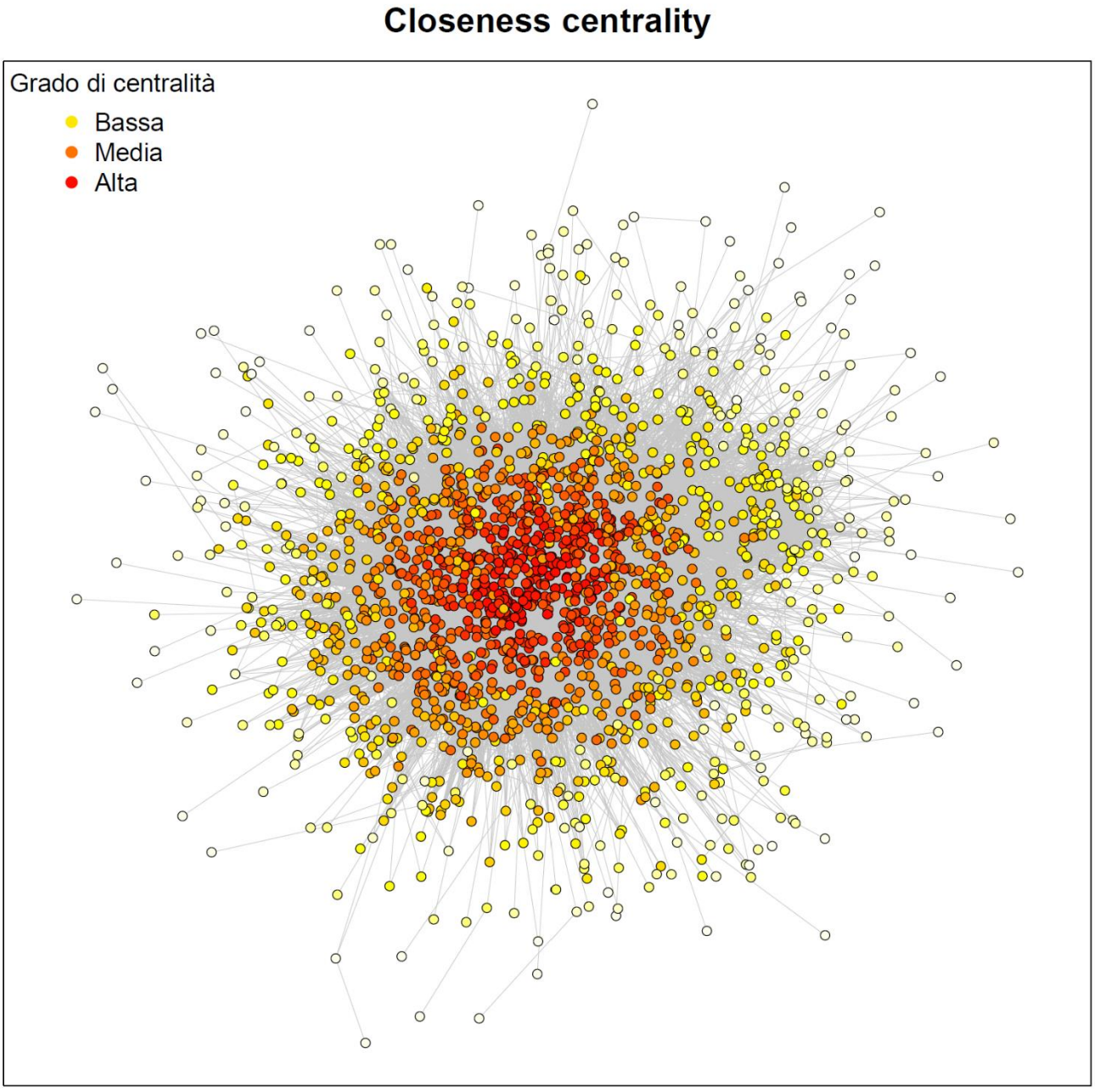

Figura 6 *- Closeness centrality*

A questo punto, si può utilizzare la correlazione per confrontare le diverse misure di centralità, come mostrato nella Tabella 2:

	Degree	Eigenvector	Betweennes	Closeness
Degree	1	**0,93**	**0,86**	**0,61**
Eigenvector	0,93	1	**0,70**	**0,77**
Betweenness	0,86	0,70	1	**0,36**
Closeness	0,61	0,77	0,36	1

***Tabella 2** - Correlazione tra le misure di centralità*

Come si nota, la degree centrality è fortemente correlata sia con eigenvector centrality che con betweenness centrality; perciò, possiamo affermare che streamer con un maggior numero di amicizie hanno la caratteristica di relazionarsi con altri streamer "centrali" e permettono di connettere parti disgiunte della rete.

Al contrario, si nota che la misura di closeness centrality risulta essere abbastanza scorrelata con la betweennes, a testimoniare il fatto che nodi più vicini "geograficamente" a tutti gli altri non necessariamente sono anche centrali dal punto di vista strutturale.

In ogni caso, il nodo con centralità più alta ritornato da ciascuna metrica è lo stesso (id = 33195858), e nella Tabella 3 confrontiamo i suoi attributi rispetto alla media della popolazione

Attributo	Valore	Valore medio della popolazione
Days	2.251	1.327
Mature	True	False
Views	79.248	408.715
Partner	False	False

***Tabella 3** - Valori del nodo centrale e confronto*

Dai dati mostrati si nota che lo streamer centrale non possiede un numero alto di visualizzazioni al canale (valore molto minore rispetto alla media), tuttavia è uno degli utenti più "anziani" dal punto di vista dell'utilizzo della piattaforma.

2.3. COMMUNITY DETECTION

Il terzo step dell'analisi prevede l'identificazione di clusters all'interno della rete attraverso algoritmi di community detection e la visualizzazione grafica dei risultati ottenuti. Gli algoritmi utilizzati sono:

- Louvain community detection algorithm: è un algoritmo basato sulla massimizzazione della modularità, un indice che misura la densità relativa degli archi all'interno della comunità rispetto a quella degli archi esterni alla comunità;
- Fast greedy algorithm: anche questo algoritmo si basa sulla modularità, utilizzando un approccio greedy;
- Walktrap algorithm: questo algoritmo identifica le comunità attraverso i "random walks" ed un approccio bottom-up.

Anche in questo caso, per ogni algoritmo visualizziamo i risultati ottenuti direttamente sulla rete, in modo tale da confrontare anche visivamente i metodi. Per motivi di leggibilità dei grafi e per distinguere meglio i cluster è stato leggermente modificato il layout originale.

Nota: per ciascun algoritmo è stato deciso di considerare solamente le 5 community più popolose, in modo tale da non includere all'interno della ricerca i clusters poco significativi.

Le figure 7, 8 e 9 mostrano rispettivamente i risultati degli algoritmi Louvain, fast greedy e walktrap.

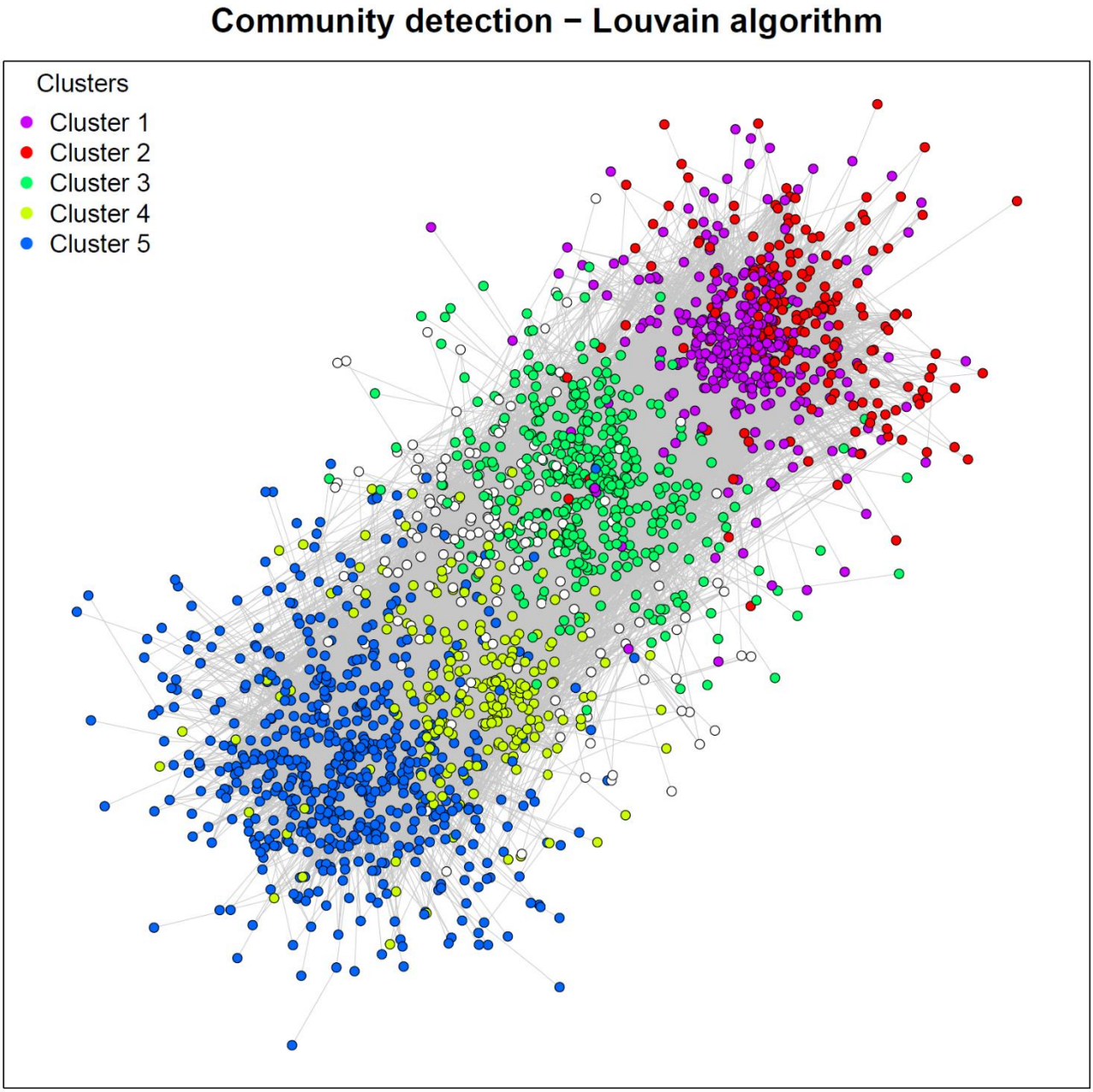

Figura 7 *- Louvain algorithm*

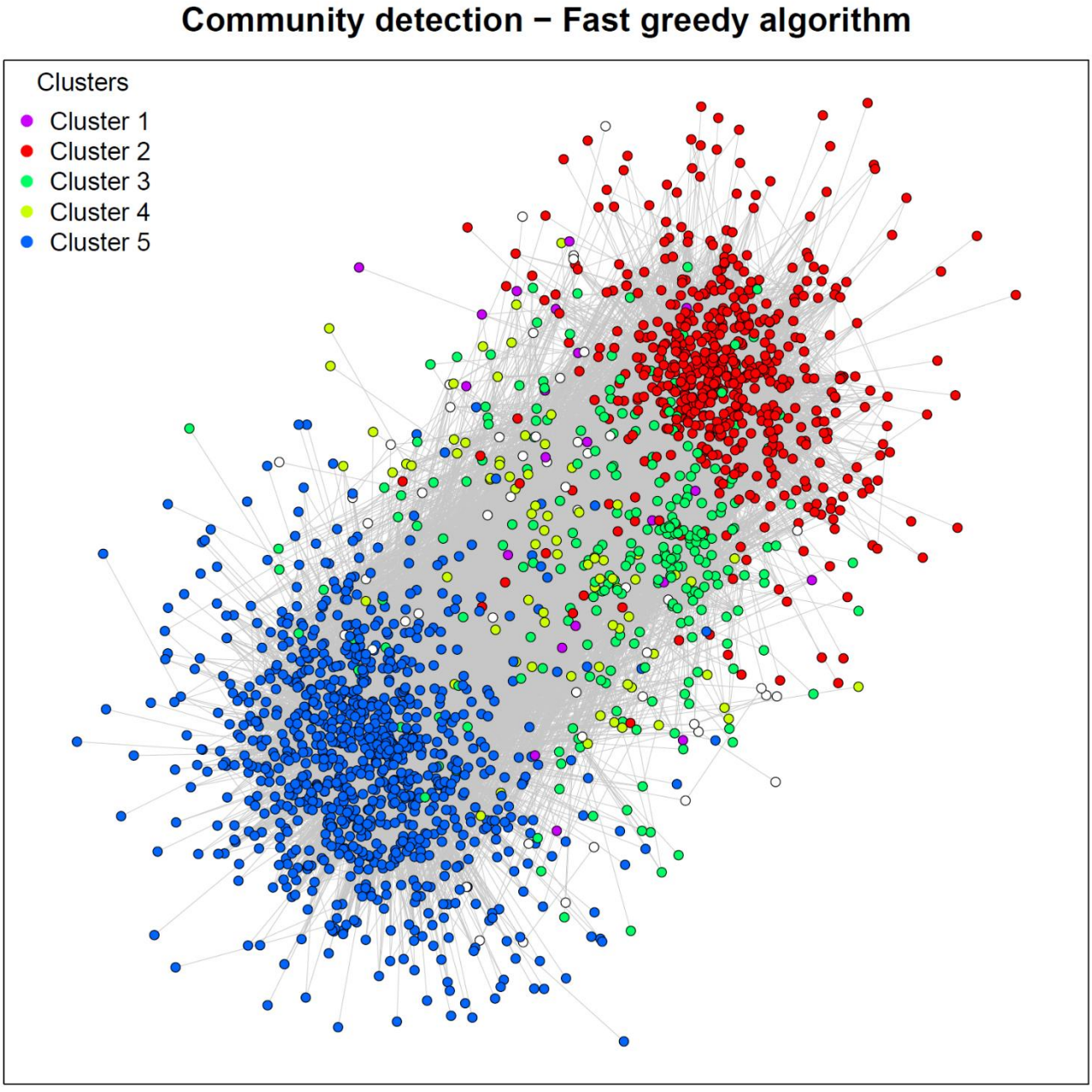

Figura 8 *– Fast greedy algorithm*

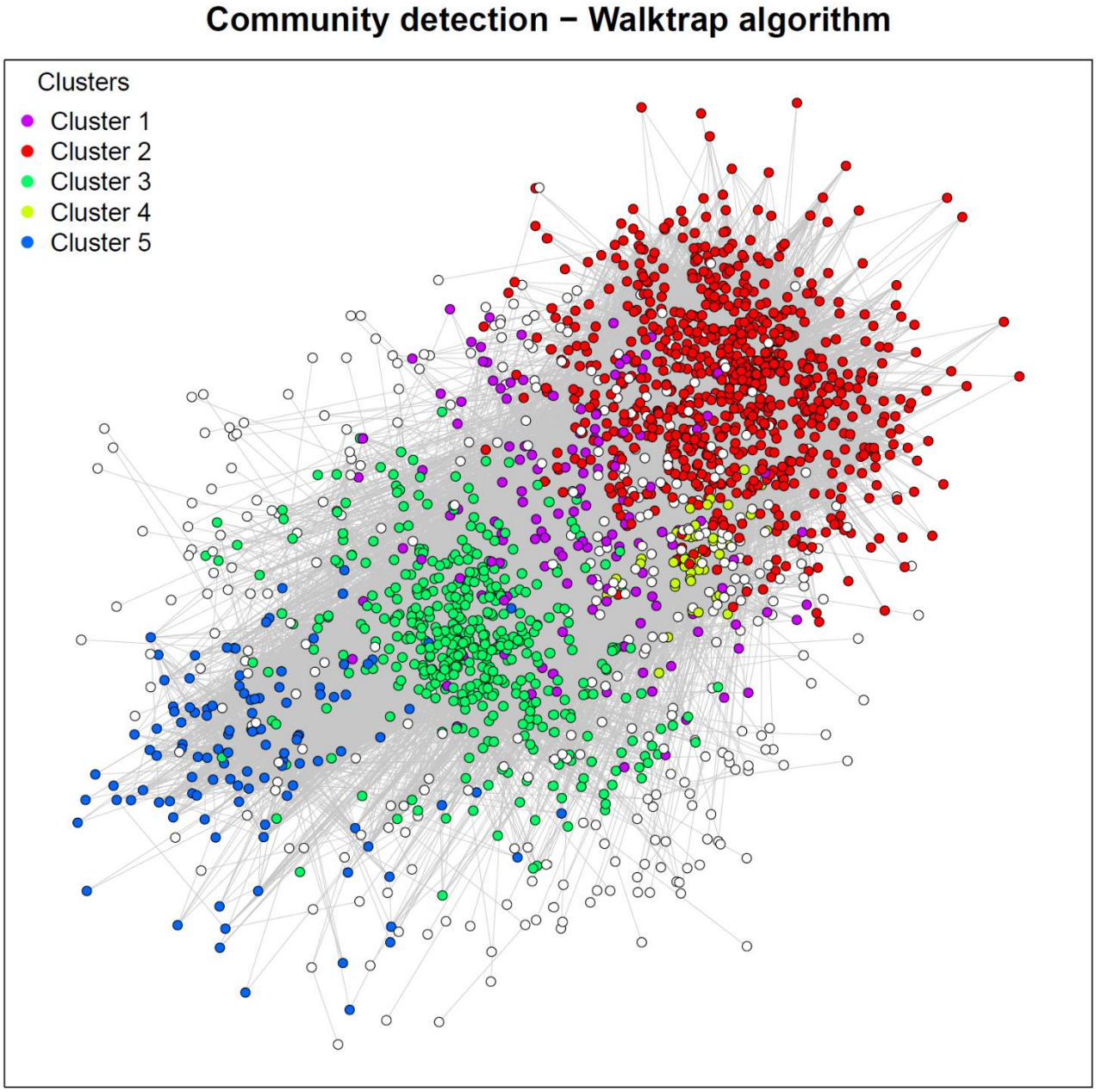

Figura 9 *- Walktrap algorithm*

Per confrontare i quattro algoritmi può essere utile comparare le community che sono state identificate: nella Tabella 4 vengono riportati, per ogni algoritmo, il numero di elementi che formano le community in ordine crescente (dal cluster meno popolato a quello più popolato).

Louvain	202, 224, 321, 405, 584
Fast greedy	21, 77, 253, 577, 933
Walktrap	42, 117, 141, 452, 774

***Tabella 4** - Numero di nodi per ciascuna community*

Da queste informazioni si può facilmente notare che l'algoritmo Fast greedy è quello che clusterizza il maggior numero di nodi (1.861 su 1.912 totali) e che contiene nel suo cluster più popoloso molti più nodi rispetto altri due algoritmi (933 contro 774 e 584). Allo stesso tempo si osserva che l'algoritmo Louvain produce cluster con deviazione standard molto minore rispetto agli altri (155 contro 381 di fast greedy e 305 di walktrap), riscontrabile dalla poca differenza nelle dimensioni dei clusters individuati.

Un'ulteriore metrica utile con confrontare i clusters ottenuti è l'indice Rand. Questo indice, il cui valore è compreso tra 0 ed 1, misura la somiglianza tra due metodi di raggruppamento contando il numero di elementi raggruppati nello stesso modo rispetto al totale. Nella Tabella 5 vengono mostrati gli indici Rand calcolati per le coppie di algoritmi utilizzati.

	Louvain	Fast greedy	Walktrap
Louvain	100%	**72%**	**75%**
Fast Greedy	72%	100%	**72%**
Walktrap	75%	72%	100%

***Tabella 5** - Indice Rand degli algoritmi*

Le alte percentuali degli indici Rand ci permettono di affermare che le community individuate non dipendono dall'algoritmo utilizzato, bensì sembrano essere una caratteristica "intrinseca" della rete. Ciò significa che utilizzando altri algoritmi, le communities individuate saranno simili a quelle mostrate in queste pagine.

Il prossimo passo consiste nell'analizzare più in profondità i cluster ottenuti: per semplicità e visto l'alto indice di somiglianza tra gli

algoritmi utilizzati, verranno analizzati solamente i cluster ottenuti dall'algoritmo Louvain. È stato scelto questo algoritmo poiché, come detto in precedenza, è quello con un numero di nodi per cluster più omogeneo e con deviazione standard minore, caratteristica che ci permette di condurre un'analisi caratterizzata anch'essa da una maggior omogeneità.

Nei boxplot riportati nelle Figure 10 e 11 vengono confrontati i cluster rispettivamente sul numero di visualizzazioni ed i giorni di attività degli utenti che li compongono.

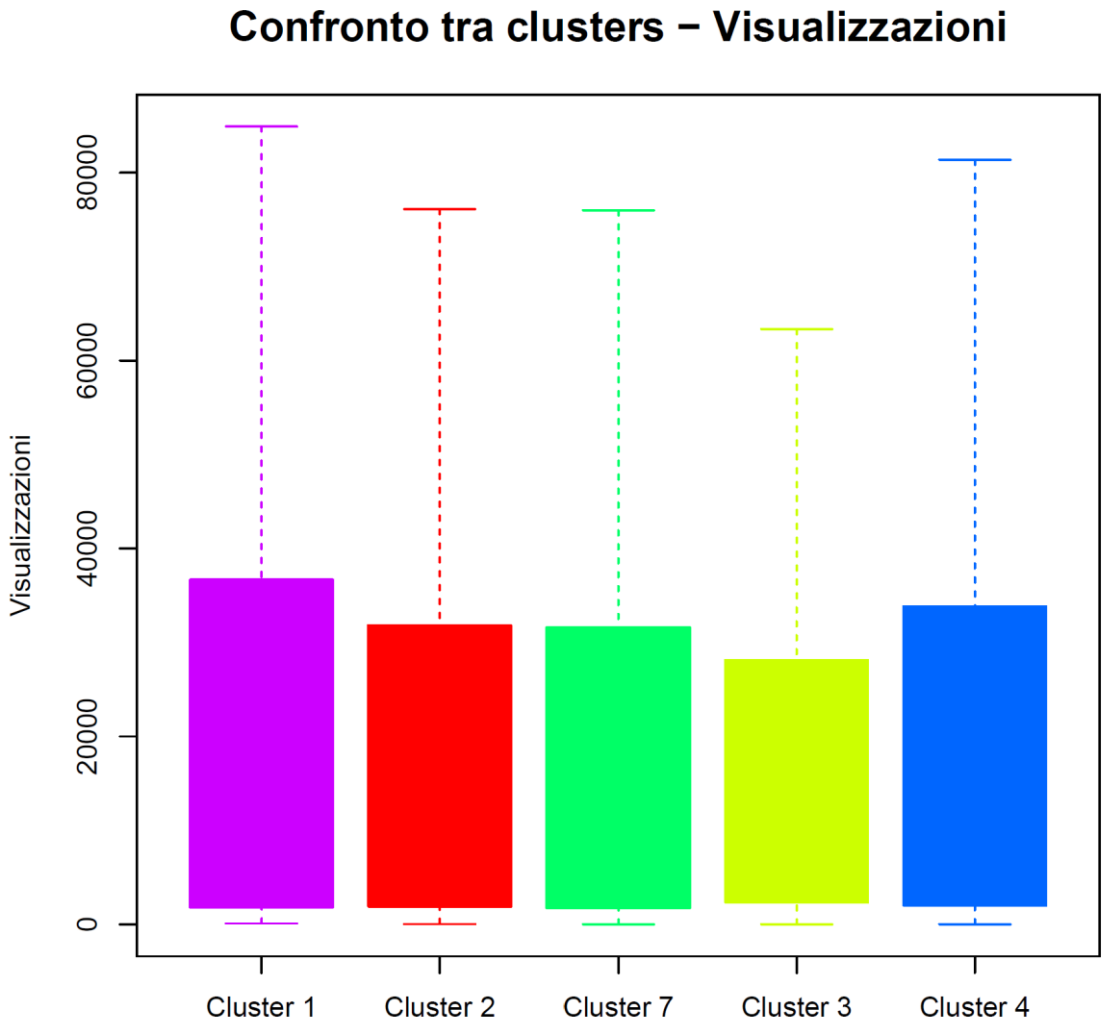

Figura 10 *- Boxplot del numero di visualizzazioni nei clusters*

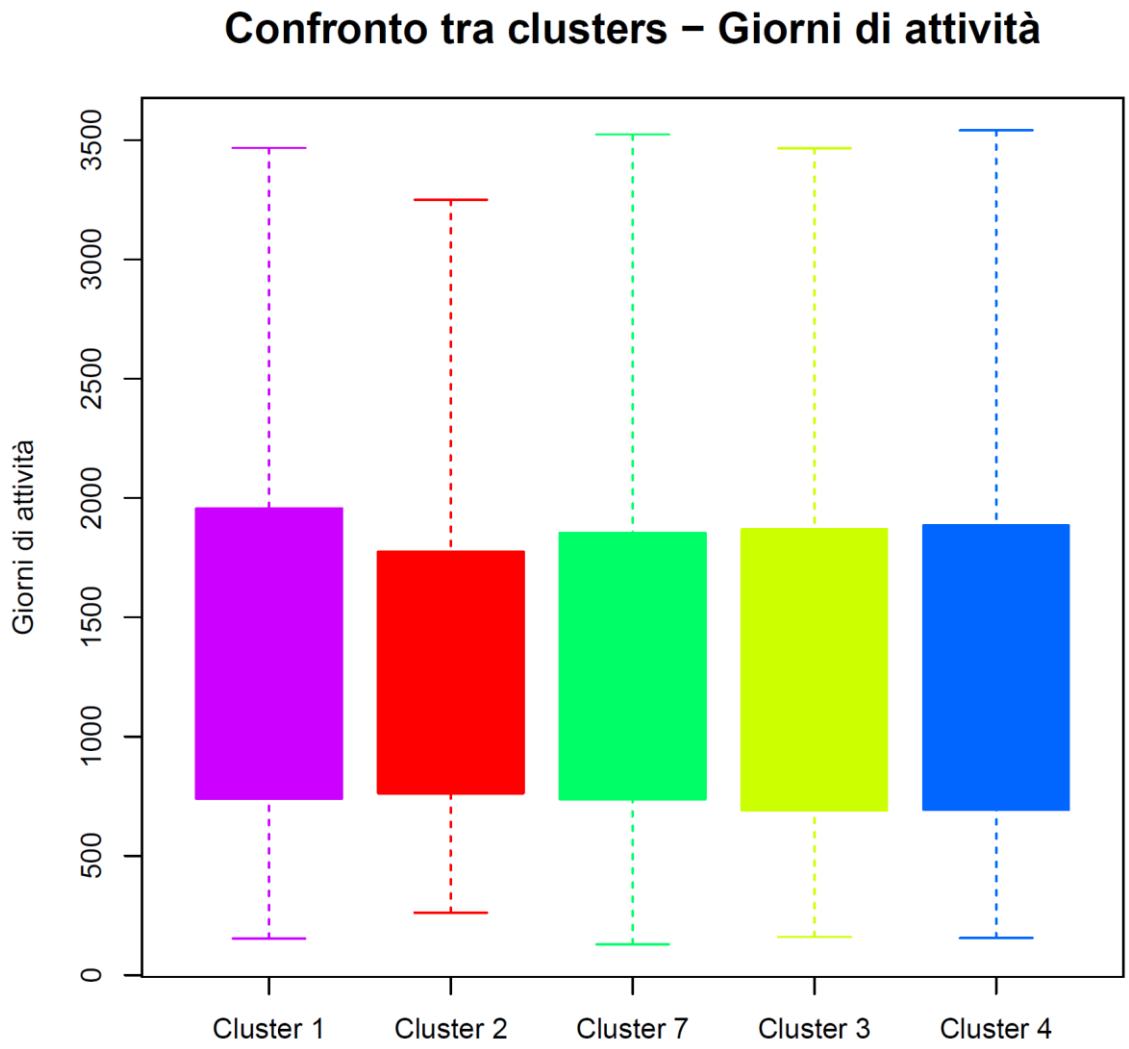

Figura 11 *- Boxplot del numero di giorni di attività nei clusters*

Da questi boxplot si nota come streamer appartenenti a cluster diversi siano in realtà simili nei loro due attributi principali; perciò, si può concludere che le communities non abbiano particolari relazioni con gli streamer che le compongono.

Al contrario, il barplot della Figura 12 ci mostra come il cluster 4 abbia un numero di streamer maggiorenni molto maggiore rispetto agli altri, e ciò ci traduce anche in numero maggiore di streamer con la partner (verosimilmente, vi sono molti più utenti partner maggiorenni che non).

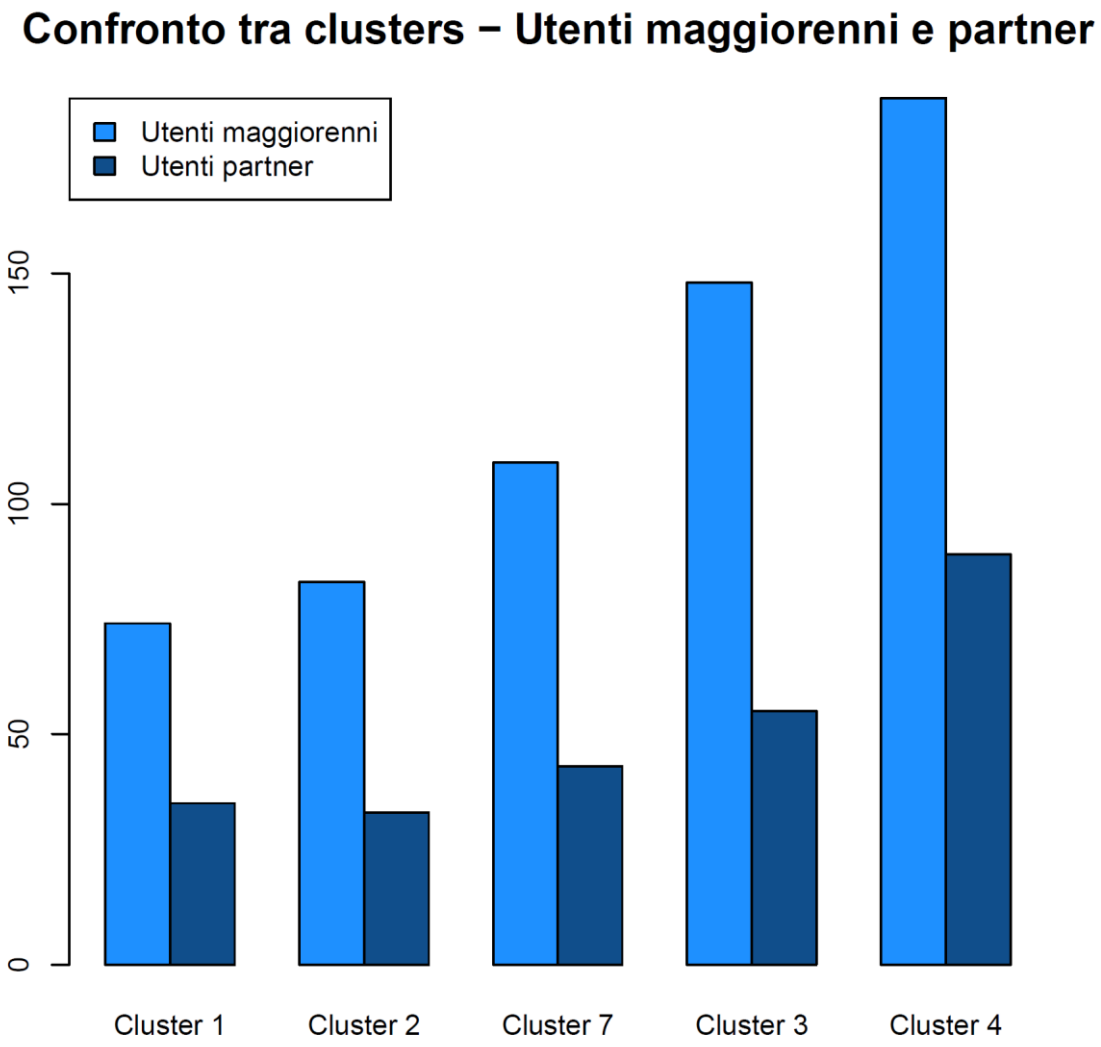

Figura 12 *- Barplot del numero di utenti partner tra quelli maggiorenni*

2.4. MISURAZIONE DELL'OMOFILIA

L'ultima parte dell'analisi si concentra sulla misurazione di un eventuale fenomeno di omofilia all'interno della rete rispetto agli attributi degli utenti, concentrandoci sul numero di visualizzazioni e il periodo di attività. In questo senso, l'obiettivo è verificare se gli streamer tendono a legarsi maggiormente a streamer con un numero simile di visualizzazioni oppure di giorni di attività.
Per misurare l'omofilia rispetto agli attributi citati è stata usata la funzione "assortativity" del pacchetto R "igraph", e nella Tabella 6 vengono mostrati i risultati ottenuti.

Attributo	Indice di omofilia
Views	-0,004
Days	-0,020
Partner	0,006
Mature	0,006

***Tabella 6** - Indici di omofilia*

I valori molto bassi di omofilia per tutti e quattro gli attributi ci suggeriscono che le amicizie che si stringono tra gli utenti non sono influenzate da alcun tipo di caratteristica, tra quelle prese in considerazione. Al contrario, i valori negativi (ma comunque vicinissimi a 0) per views e days sono sintomo di una leggerissima omofilia inversa.

Per sostenere i risultati ottenuti misurando l'omofilia, nelle Figure 13 e 14 è stata rappresentata la rete evidenziando rispettivamente, i top 5 streamer per numero di visualizzazioni e di giorni di attività. Come si può notare, gli utenti con maggiori visualizzazioni sono distribuiti in 3 cluster diversi, mentre quelli più "anziani" addirittura in 5: questo risultato sostiene le misure ottenute nella Tabella 6, in quanto in nessuno dei due casi si può affermare con certezza che le relazioni tra gli utenti siano influenzate da questi attributi.

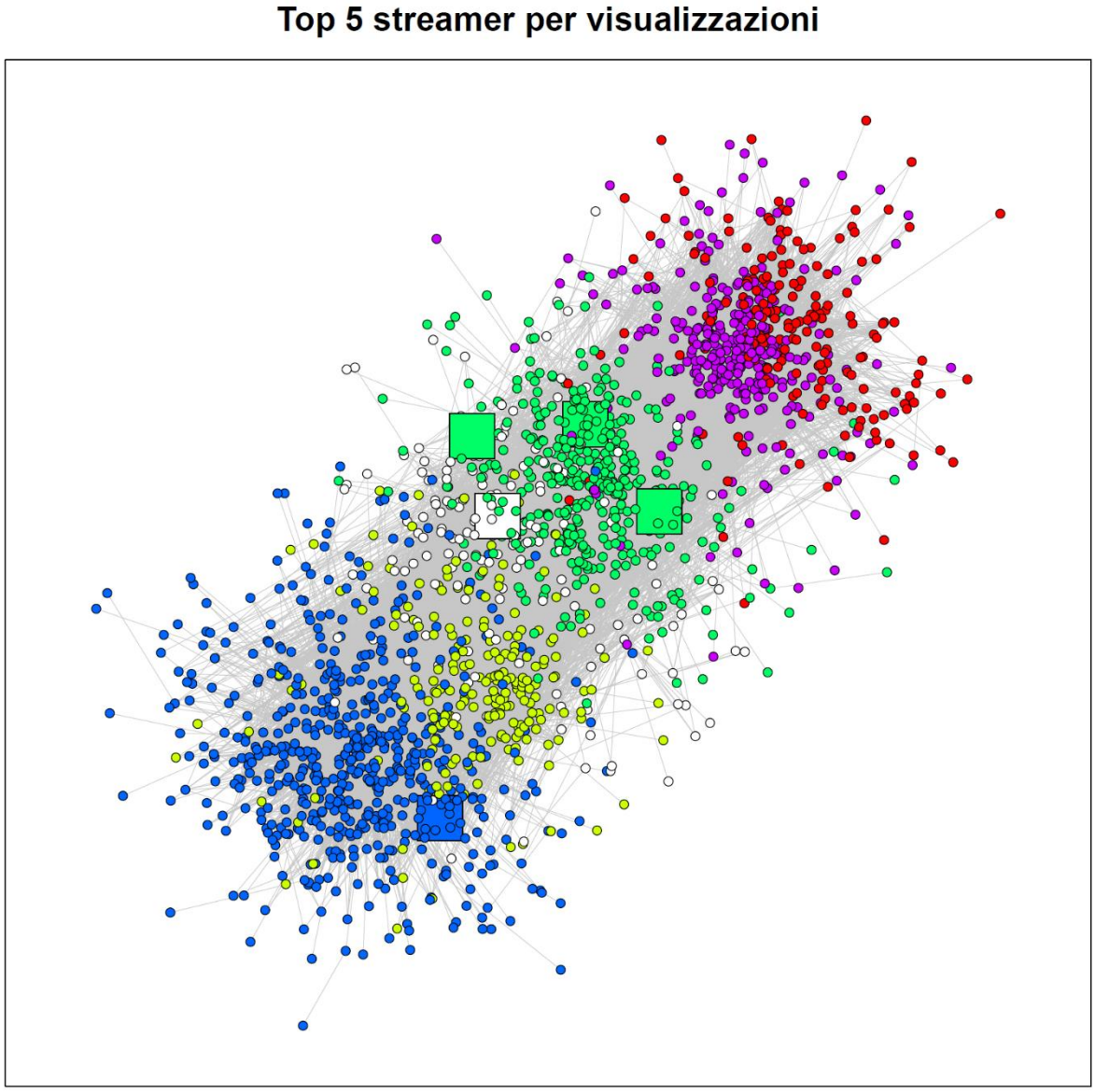

Figura 13 *- Top 5 streamer (numero di visualizzazioni)*

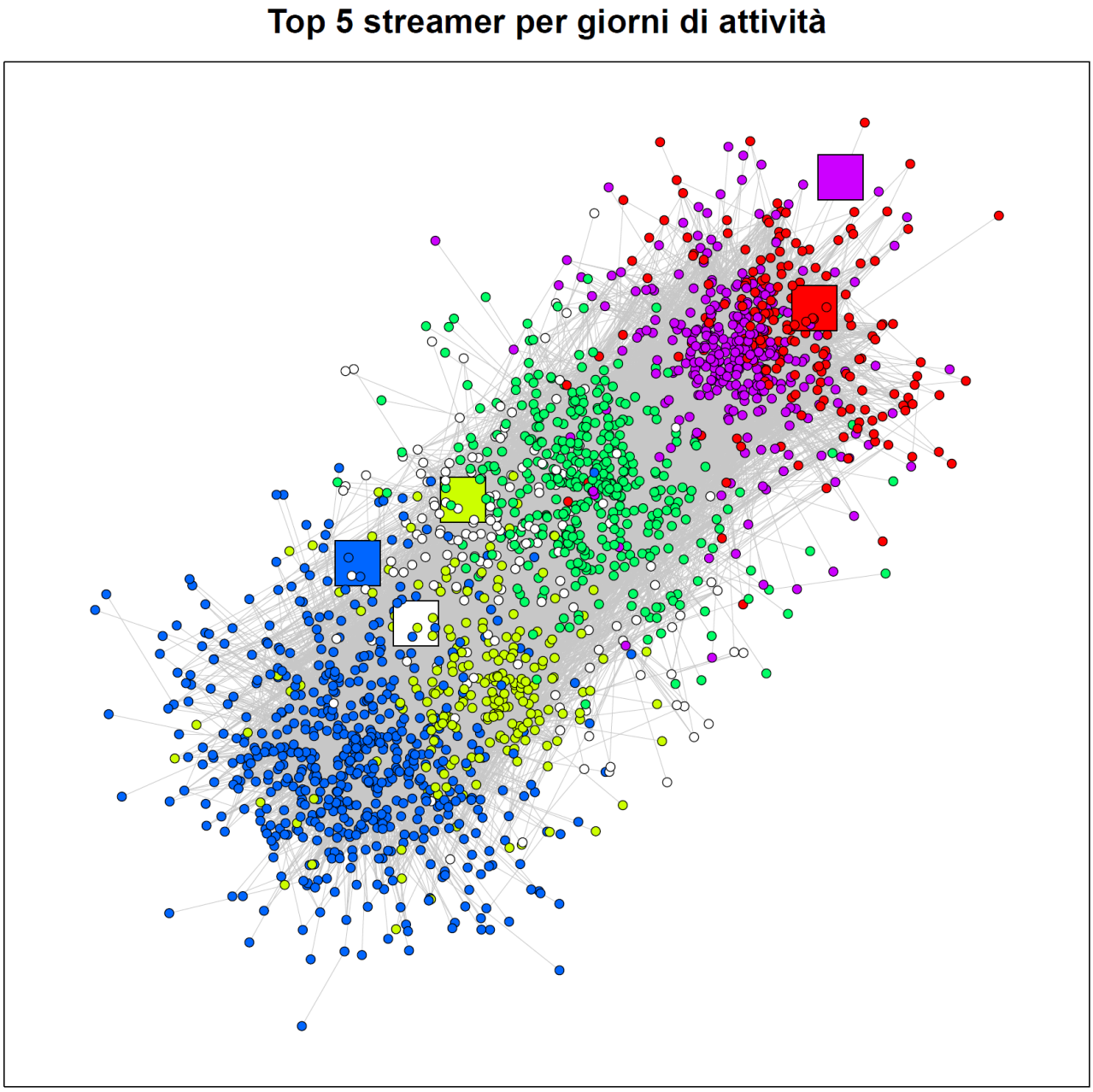

Figura 14 *- Top 5 streamer (giorni di attività)*

3. LIMITI E BIAS DELL'ANALISI

Come in ogni genere di analisi e di ricerca, anche questo progetto è caratterizzato da alcuni limiti e bias.

Innanzitutto, il limite più importate che è stato riscontrato è quello legato agli strumenti ed alla componente software utilizzata: viste le dimensioni della rete analizzata (più di 30.000 archi), non è stato possibile eseguire alcuni algoritmi di community detection (es: edge betweenness oppure spin glass algorithm) in quanto troppo lenti nell'esecuzione; per questo motivo si è innanzitutto semplificata la rete, come spiegato all'inizio dell'analisi, e si è sempre cercato di ottimizzare il più possibile le operazioni da eseguire, in modo tale da poter coniugare la potenza computazionale del software con l'obiettivo di ricerca.

Un ulteriore bias si riscontra nella sezione dedicata allo studio dell'omofilia: in questo caso abbiamo concluso che, rispetto agli attributi disponibili per questo insieme di nodi, non è possibile affermare con certezza che le amicizie formate tra gli streamer siano caratterizzate da omofilia. Questa conclusione è corretta, ma allo stesso tempo non preclude che vi siano altri attributi, non presenti in questo dataset e quindi non presi in considerazione, per i quali possa effettivamente essere osservato il fenomeno dell'omofilia. In questo senso, è corretto affermare che, per i dati a nostra disposizione, non vi sembra essere un fenomeno di omofilia, ma nulla esclude che possa esserci per altri attributi non esaminati.

Infine, un altro limite riscontrabile in questo lavoro è legato alle caratteristiche dei dati analizzati: il dataset, infatti, contiene esclusivamente dati di streamer portoghesi e risale al 2018. In questo senso, da una parte i risultati ottenuti sono limitati ad un campione di una popolazione molto più grande, dall'altra, dal momento che le osservazioni raccolte risalgono a quattro anni fa, è possibile che le conclusioni a cui si è giunti possano cambiare se viene analizzato un dataset più recente.

In ogni caso, l'intera analisi è stata condotta cercando di limitare il più possibile le problematiche relative ai bias citati e cercando di giungere a conclusioni il più oggettive possibili.

4. BIBLIOGRAFIA

- Appunti delle lezioni di Social Network Analysis A.A. 2021-2022 tenute dalla professoressa Fabiana Zollo;
- Sezione "Social Network Analysis Project" (SNAP) dell'Università di Stanford: http://snap.stanford.edu/data/twitch-social-networks.html
- Documentazione di R: https://www.rdocumentation.org
- Documentazione della libreria "igraph": https://igraph.org/r/html/latest