TV Show Meta Pages Social Network Analysis

Domenico Izzo, Ciro Maccarone, Adelio Antonini

Dicembre 2023

1 Introduzione

Grazie allo studio delle relazioni che intercorrono tra i diversi profili dei social media, è possibile scoprire informazioni rilevanti sui trend e sui comportamenti sociali degli individui, consentendo di creare dei profili di interesse comuni tra gli utenti e andando a creare degli algoritmi tali da essere in grado di tenere traccia delle azioni e degli interessi mostrati dagli utenti, andando dunque a suggerire profili affini ai propri gusti e di mostrare loro pubblicità interessanti.

Questo genere di analisi non è solo rilevante a scopi commerciali, è possibile infatti osservare come si comporta la popolazione a fronte di determinati eventi, ad esempio a ridosso delle elezioni, è possibile eseguire previsioni sulla base delle preferenze mostrate dalla popolazione andando ad analizzare le loro preferenze ed i loro commenti.

Non solo, è possibile anche osservare anche come si diffonde una informazione, e andare dunque ad osservare come particolari soggetti abbiano capacità di diffusione superiori rispetto agli altri, e quanto questi individui possano influenzare i comportamenti di altri soggetti.

2 Descrizione del dataset

Il dataset del progetto corrente mostra un insieme di relazioni di like di pagine Meta di spettacoli televisivi: in particolare sono un insieme di collegamenti senza direzioni tra le varie pagine a due a due, il cui insieme genera un grafo non orientato. Il dataset si presenta dunque come una lista di coppie a questo modo: (1,2) dove 1 e 2 sono le pagine. Nel dataset per motivi di privacy i nomi delle pagine sono stati sostituiti da numeri.

Il dataset rappresenta dunque un grafo da ben 3892 nodi e da 17262 archi, con i primi rappresentiamo le pagine, e con gli archi le relazioni di like tra le due. Il dataset presenta degli errori, in quanto sono presenti dei cicli. E quindi di pagine che secondo il dataset metterebbero like a sé stesse. Per le meccaniche del social

media sappiamo che questo è impossibile, e data la rarità di questa relazione (23 archi), è riconducibile ad un errore di raccolta dei dati, che dopo le analisi generiche di numerosità del grafo, sono state eliminate in quanto non rappresentano il contesto in esame correttamente, e non sono abbastanza numerose da condurre a rifiutare il dataset.

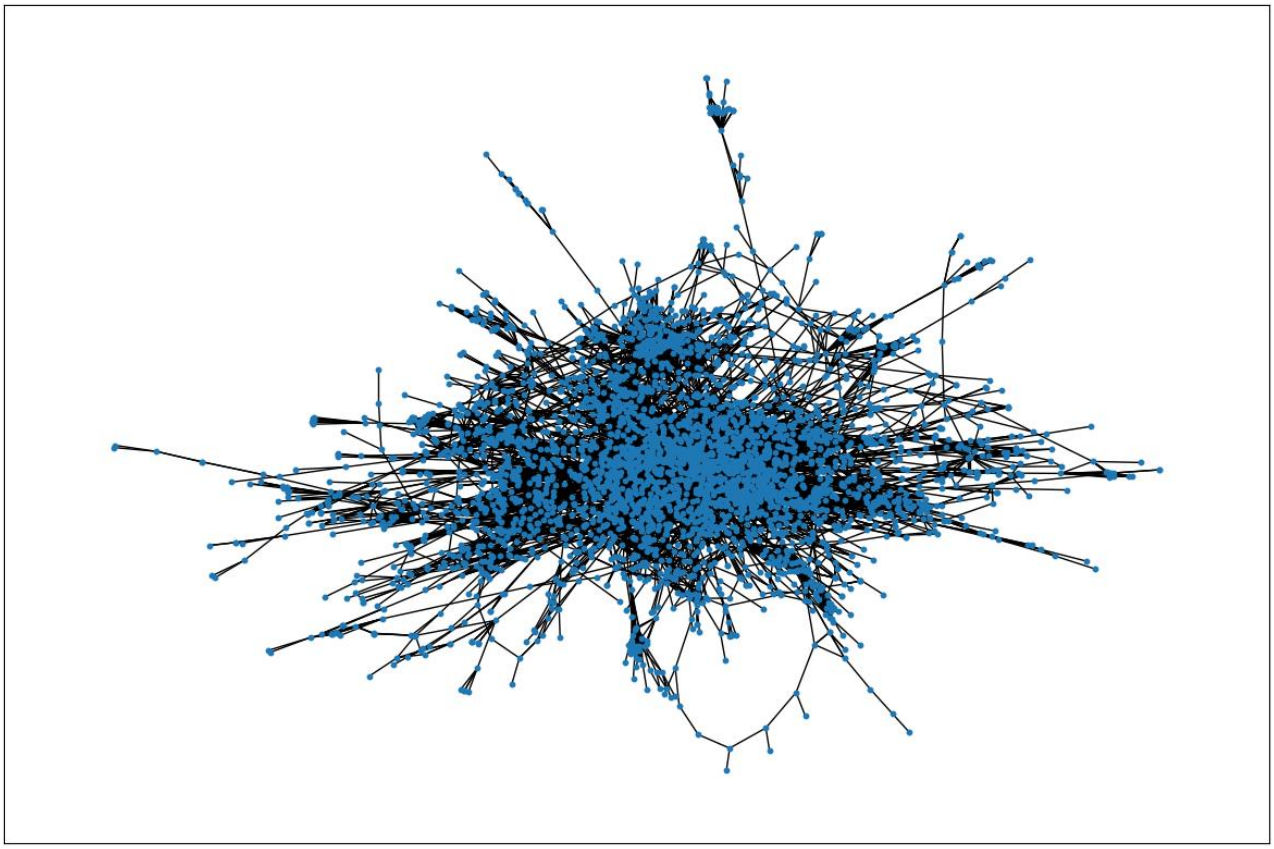
3 Strumenti usati per le analisi

Il software usato per l'analisi è Visual Studio Code, al quale abbiamo usato l'extention di Jupyter Notebook, con il kernel di Python 3.10.2. Le librerie usate sono le seguenti:

- Pandas: libreria standard usata per gestire i dati, ci consente di estrarre e di manipolare i dati prelevati all'interno con facilità. Viene anche usato nel contesto di ML per algoritmi di classificazione e clustering. Nel nostro contesto lo useremo per estrarre i dati e darli in input alla funzione che ci permetterà di costruire il nostro grafo.
- Networkx: la libreria principale del nostro studio, permette di creare e di manipolare il grafo. Ci consente inoltre di ricercare facilmente sotto grafi, di calcolare la numerosità di archi e nodi e di andare a calcolare vari tipi di centralità, nonché di visualizzare i grafi selezionati, garantendo capacità di manipolazione visiva di nodi e archi per mettere in risalto delle caratteristiche salienti.
- Matplotlib: libreria usata in questo contesto per la visualizzazione e per colorare i nodi nelle varie visualizzazioni.
- random: libreria usata per generare numeri casuali, la useremo in alcuni contesti per generare colori casuali da usare nel sistema dei colori RGB.
- seaborn: useremo la libreria per rappresentare la distribuzione dei vari tipi di centralità dei nodi.

4 Analisi del grafo

Dopo aver misurato il numero di nodi e di archi, ed aver eliminato i cicli menzionati nei paragrafi precedenti, visualizziamo il nostro grafo, in modo da avere presente la situazione con la quale ci troviamo ad affrontare:



Il grafo presenta come si nota un cuore ben collegato di pagine e di una serie di diramazioni abbastanza isolate raggiungibili solo con certi colli di bottiglia. Andremo ora a valutare le centralità del nostro grafo.

4.1 Degree Centrality

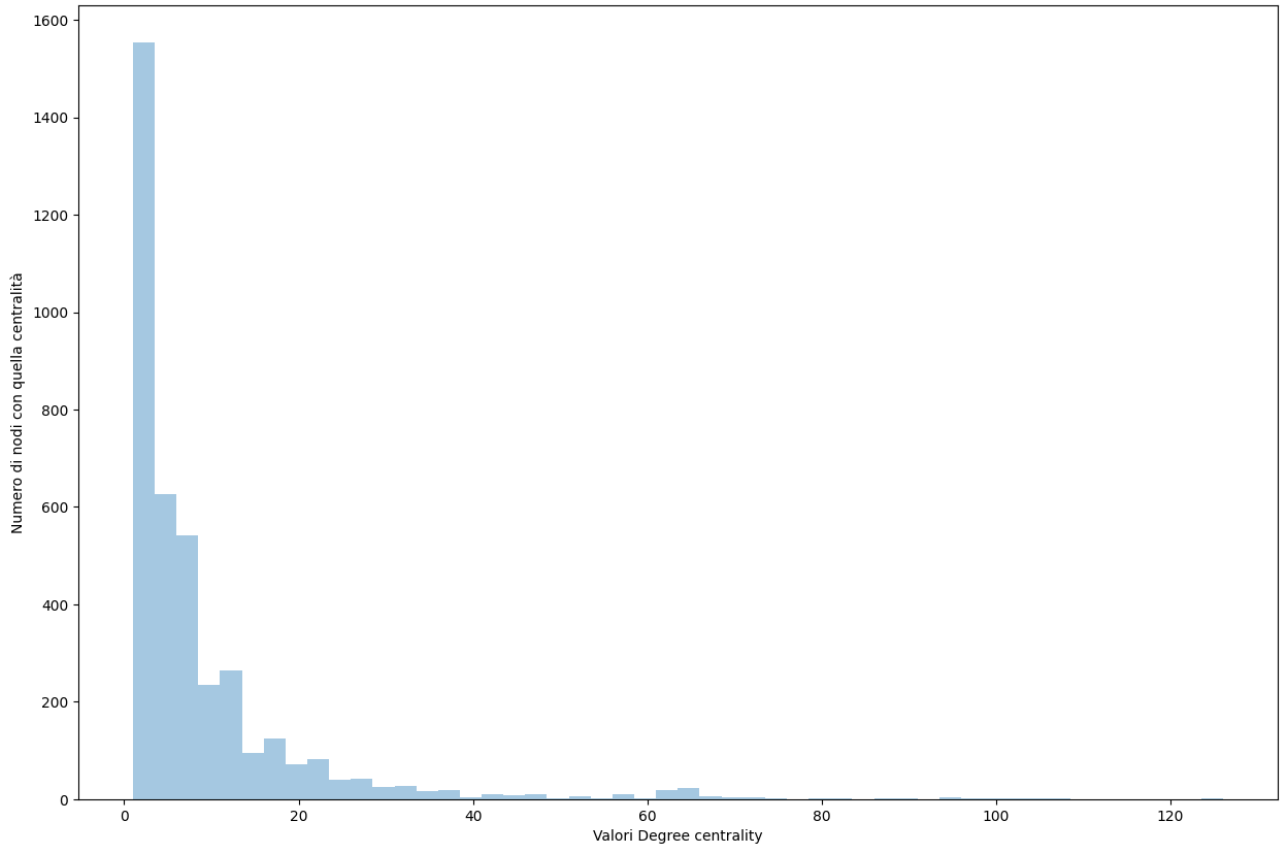
La degree centrality conta per ogni nodo il numero di collegamenti e quindi di archi collegati a sé. Andando quindi a definire nel nostro contesto quali sono le pagine più popolari.

Useremo uno dei metodi dell'oggetto grafo generato dalla libreria Networkx per ottenere un dizionario della nostra Degree Centrality, che contiene al proprio interno una lista di nodi con proprio grado di centralità. Presentiamo una lista dei primi nodi più importanti secondo questa centralità:

Nodi	Degree Centrality
2008	126
3254	126
3525	108
1177	104
1673	102
3156	101
1595	100

3122	100
2659	97
1840	97
2036	95
566	95
1073	94
603	89
3519	86
386	81

Presentiamo ora la rappresentazione della distribuzione delle degree centrality di tutti i nodi, che evidenzia una skewdness alta per valori di centralità bassi:



Presentiamo dunque nuovamente il grafo evidenziando la centralità dei nodi con colori differenti: più i colori sono caldi più la centralità è alta:

L’immagine evidenzia come i nodi più connessi si trovano verso il centro del grafo.

4.2 Closeness centrality

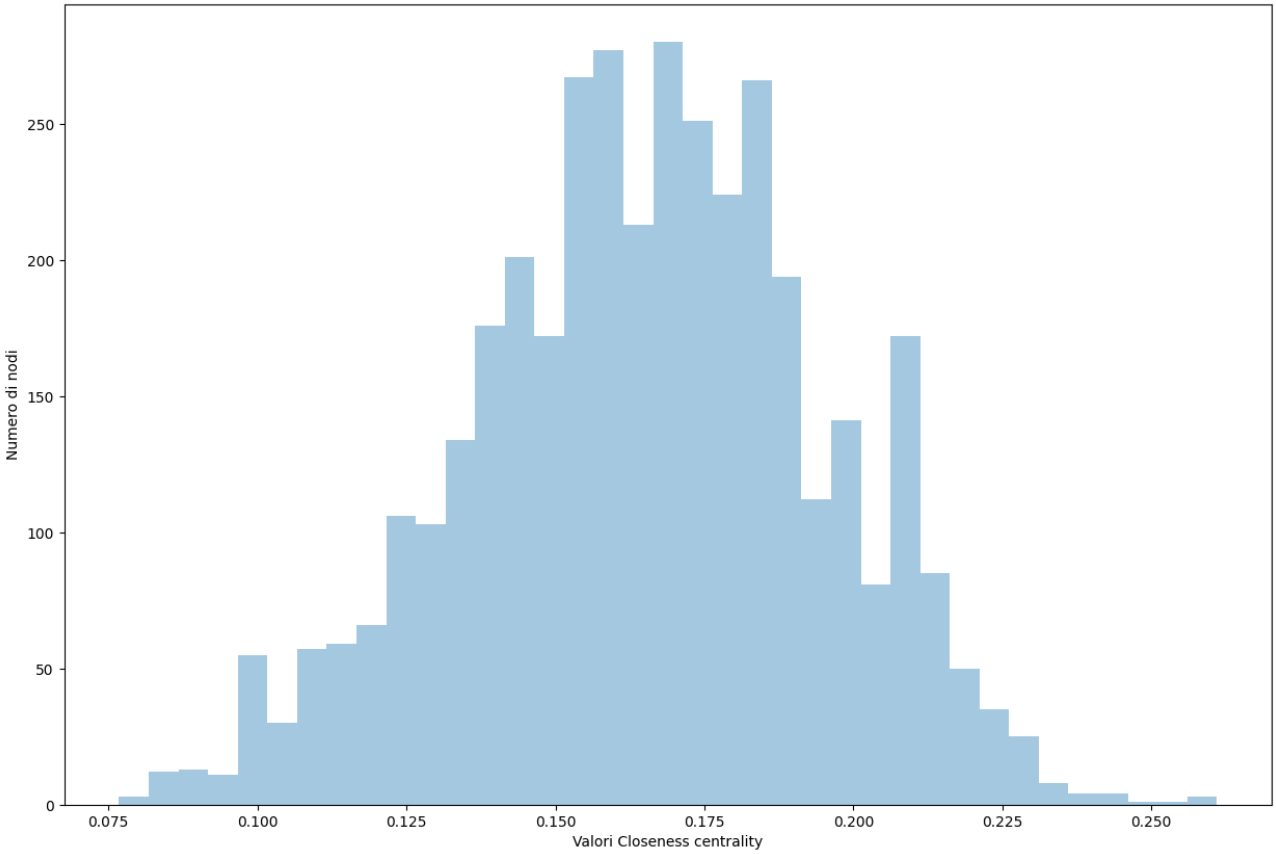
Valutiamo ora la closeness centrality, che misura quanto un nodo sia in media più vicino agli altri nodi. Di fatto se vogliamo decidere quali tragitti scegliere per assicurarci che l’informazione viaggi il più lontano possibile, scegliamo i nodi che abbiano un alto livello di centralità, perché garantisce i percorsi più brevi da un punto ad un altro del grafo.

Ancora una volta la libreria Networkx ci consente di calcolare facilmente le centralità dei nodi, presentiamo una breve lista dei nodi più rilevanti:

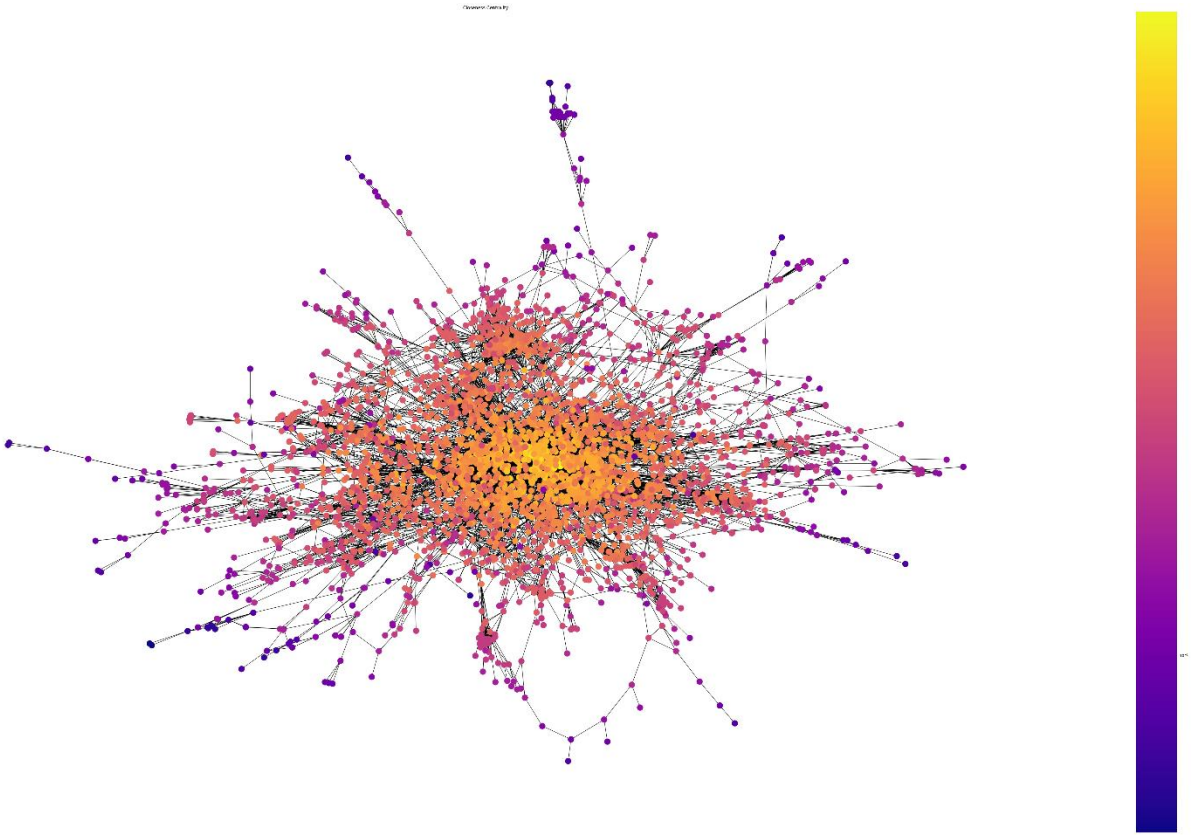
Nodi	Valore di closeness Centrality
3254	0.2609832986786505
2008	0.2593827078194787
2895	0.25618909665525413
819	0.251860961874555
2751	0.2477712684666327
211	0.24425612052730697
160	0.24422545819733868
3837	0.24265668849391955
2885	0.24197761194029851
2035	0.2401407146824662
1214	0.2392253304641869
1053	0.23857992519467777
1206	0.23611869652284725
3122	0.235689623841541
1987	0.23504893077201885
3318	0.23469449303335546
2157	0.23411552346570397

Reincontriamo nuovamente il nodo 3254 come rilevante anche in questa centralità, ma i restanti nodi non sono in genere rilevanti allo stesso modo di quanto visto precedentemente.

Per quanto riguarda la sua distribuzione della centralità in questo caso i punteggi sono distribuiti in maniera più simmetrica, la cui media è comunque relativamente alta sappiamo quindi che le novità pubblicate da una pagina sicuramente saranno visualizzate anche dai follower delle altre pagine, garantendo quindi un amplificatore di diffusione delle notizie molto efficace:



Presentiamo ancora una volta la rappresentazione del nostro grafo, colorando in maniera differente i nodi con differente closeness centrality con il stesso modus operandi del paragrafo precedente:



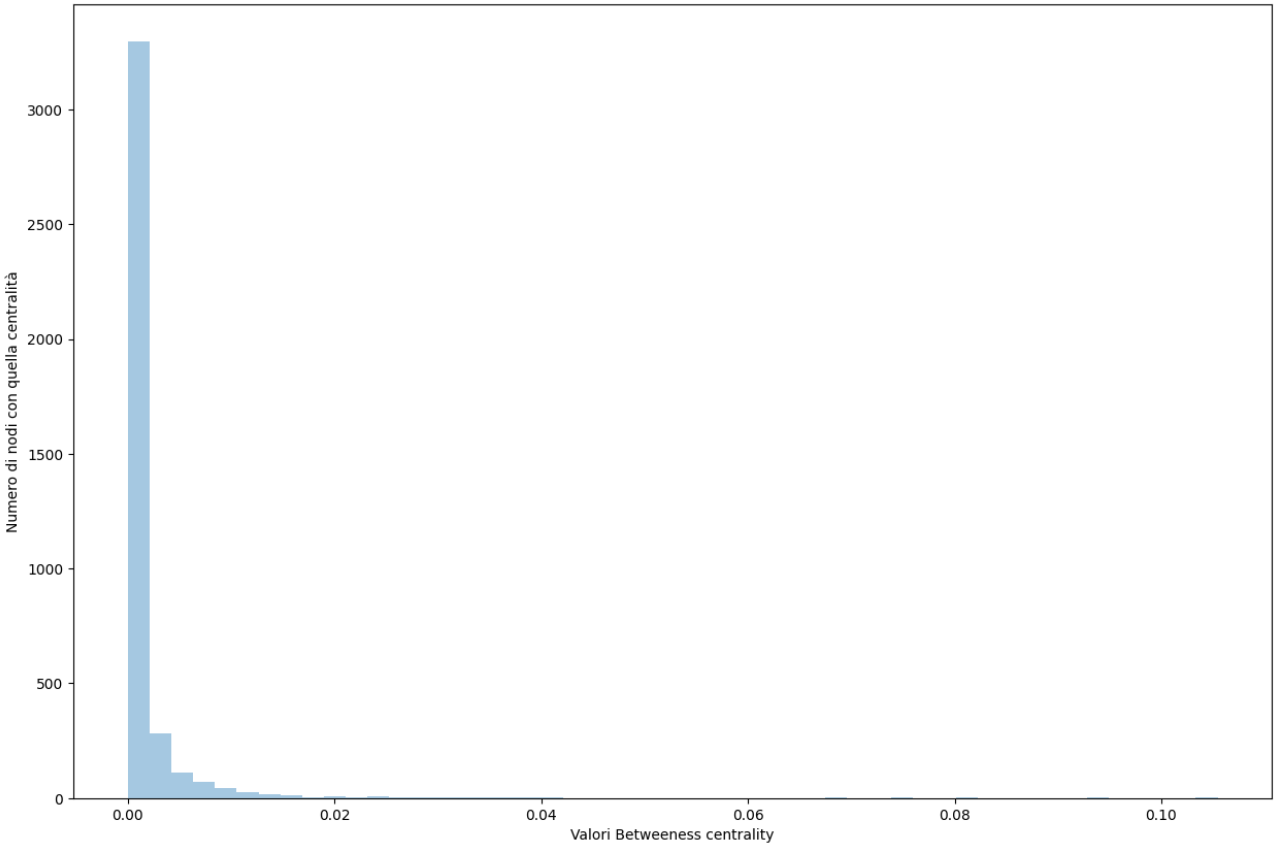
Il grafo conferma nuovamente la conclusione raggiunta: le notizie riescono a passare efficacemente da una parte all'altra del grafo, solo le periferie isolate potrebbero avere difficoltà a raggiungere più utenti, che è una caratteristica normale per pagine poco popolari.

4.3 Betweeness Centrality

Con il seguente paragrafo misuriamo ora la rilevanza posizionale dei nodi per la comunicazione: con questa centralità andiamo ad individuare quelle pagine che sono essenziali affinché le informazioni arrivino a pagine più isolate che sarebbero altrimenti difficilmente o del tutto irraggiungibili. Presentiamo come al solito una lista dei nodi più importanti in questo contesto:

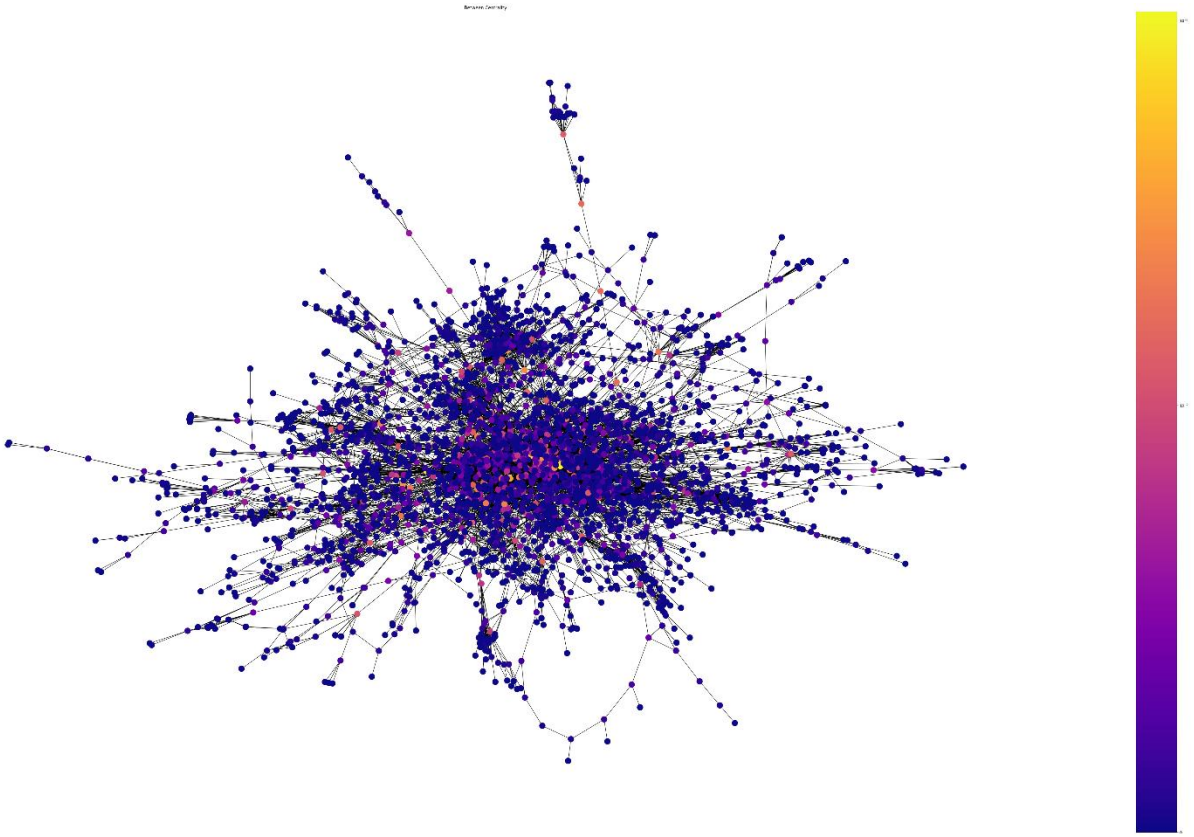
Nodi	Valore di Betweeness Centrality
3254	0.10544488181477074
2008	0.09352541687013526
819	0.0804900367587108
2170	0.07471499425323284
2751	0.07465790776474893
2895	0.06910198604847143
3038	0.040569165908730546
2682	0.03896683285298819
211	0.037912318738099464
2589	0.03424807353742939
160	0.03222334425273599
655	0.031281652512490724
1344	0.030944190757100414
3212	0.028105433487130554

Valutiamo adesso la distribuzione della centralità, vedremo che ha una skewdness elevata verso il basso, essendo poche le periferie da collegare con il centro del grafo.



A differenza della degree centrality, la betweeness è ancora più concentrata nei valori bassi, con una grossa maggioranza di punti a gradazione bassa, **il che conferma ulteriormente che le pagine hanno una buona capacità di diffusione dei propri post.**

Valutiamo adesso nuovamente il grafo mettendo in risalto i nodi a Betweeness Centrality elevata



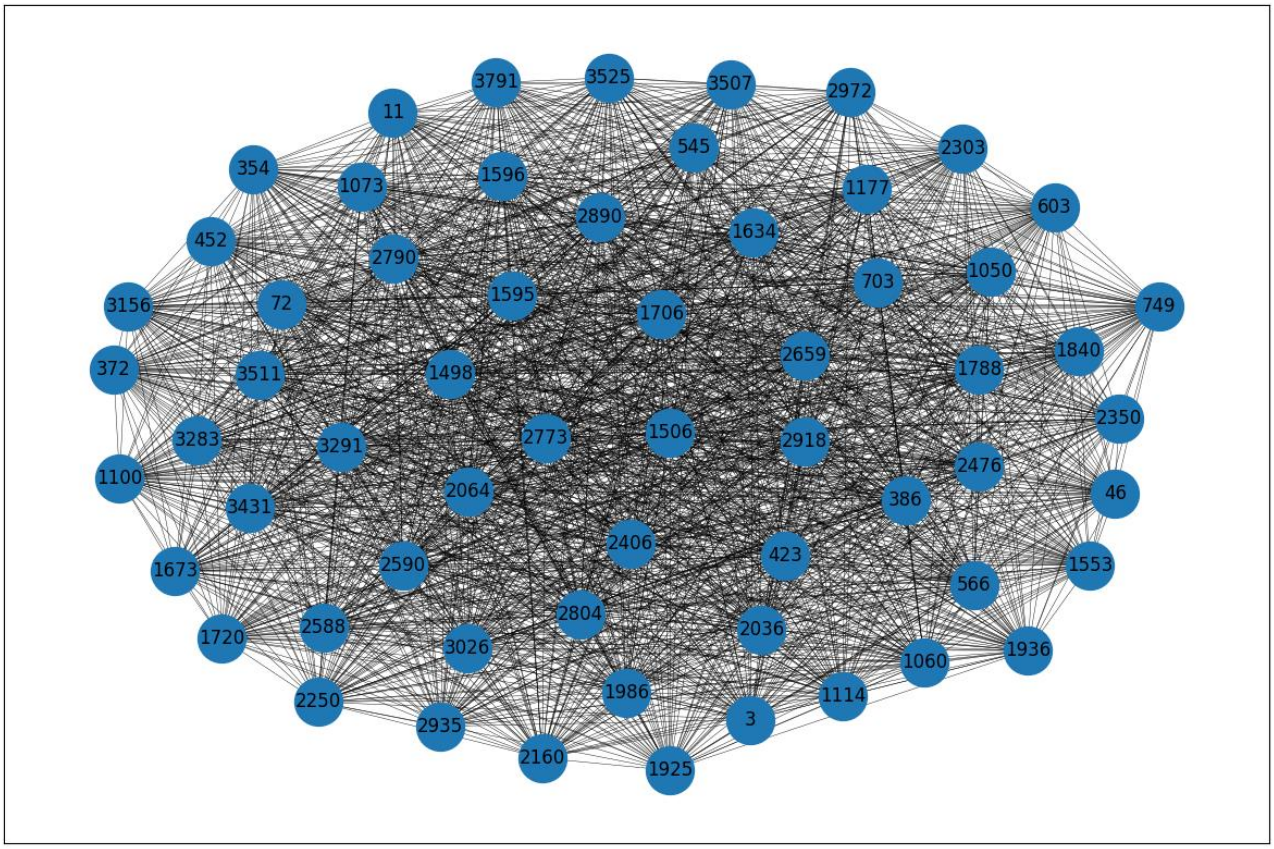
Possiamo osservare come tra i vari colli di bottiglia c'è una gradazione differente, dovuto in base al numero di nodi che collega al grafo, che sarebbero altrimenti isolati.

5 Analisi dei sottografi

Adesso che abbiamo una buona visione d'insieme sulla situazione generale delle pagine televisive, vogliamo analizzare i sottografi, nel particolare vogliamo cercare quelle pagine ben collegate e strettamente connesse l'una con l'altra. Iniziamo dunque ad analizzare il k-core, le cliques e le communities, andando ad individuare quali siano gli algoritmi che nel contesto rappresentano meglio i gruppi.

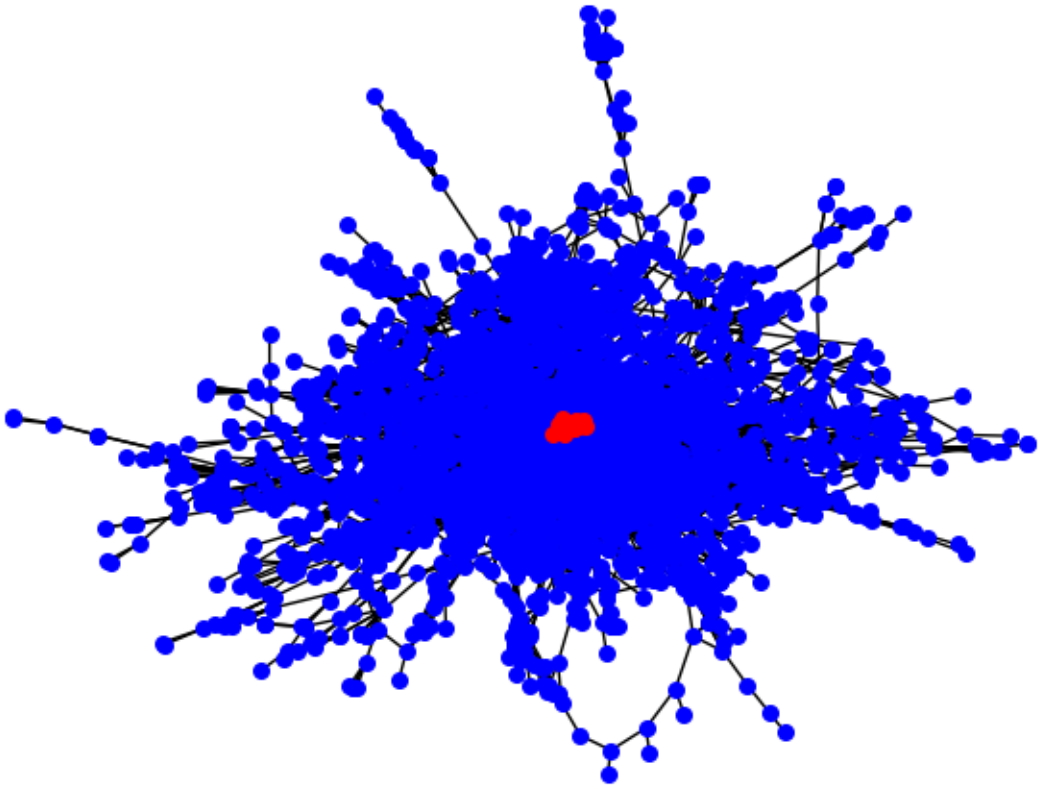
5.1 Analisi del k-core

Usiamo adesso la funzione kcore della libreria network sulla variabile che contiene il nostro grafo per ottenere il k-core massimale, e cioè il sottografo i cui nodi hanno almeno k-connessioni. Mostriamo dunque il risultato in figura:



Riconosciamo alcuni di questi nodi, come il 2036, tra i nodi a degree centrality più alta. Data la natura del k-core tutti i nodi avranno dunque lo stesso valore di centralità. Dunque, non ha senso classificare in ordine i nodi.

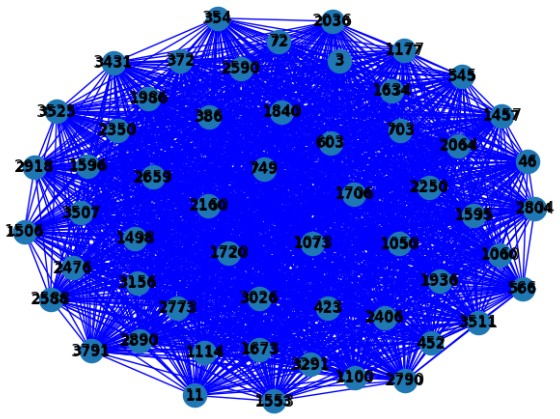
Questo è dunque il “cuore” del nostro grafo, la zona dove i nodi sono a centralità più elevata, posizionato esattamente al centro, come individuato dalla seguente immagine nella quale abbiamo evidenziato il k-core rispetto al resto.



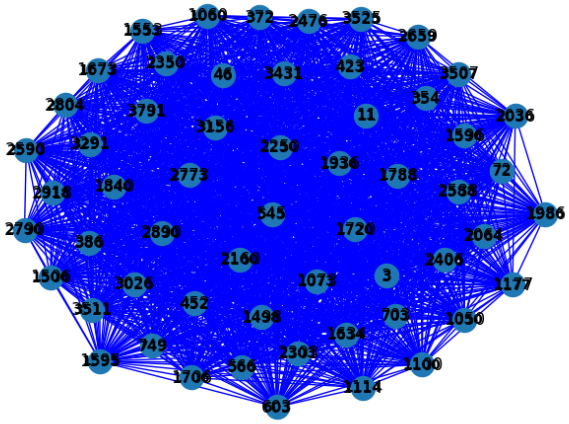
5.2 Analisi delle cliques

Individuiamo adesso le cliques, che sono sotto grafi i cui nodi sono completamente connessi, data la numerosità ci limitiamo a presentare solo alcune tra le più di 500 cliques trovate. L'obiettivo è di cercare i gruppi di pagine strettamente connesse tra loro e osservare il grafo nel loro insieme rispetto a queste.

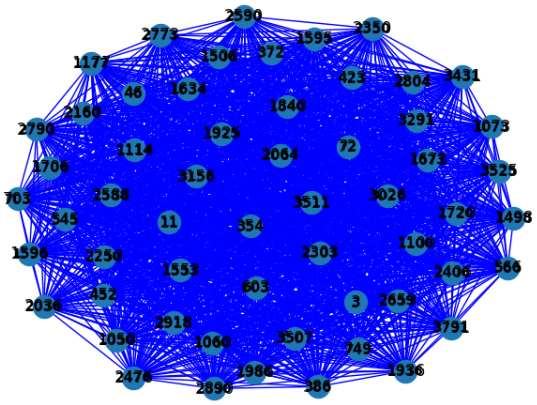
Clique 18



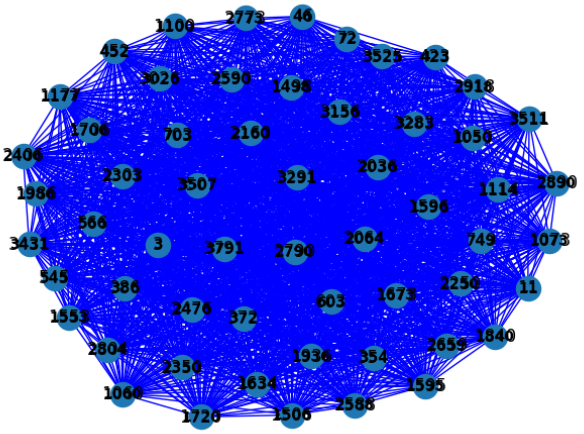
Clique 19



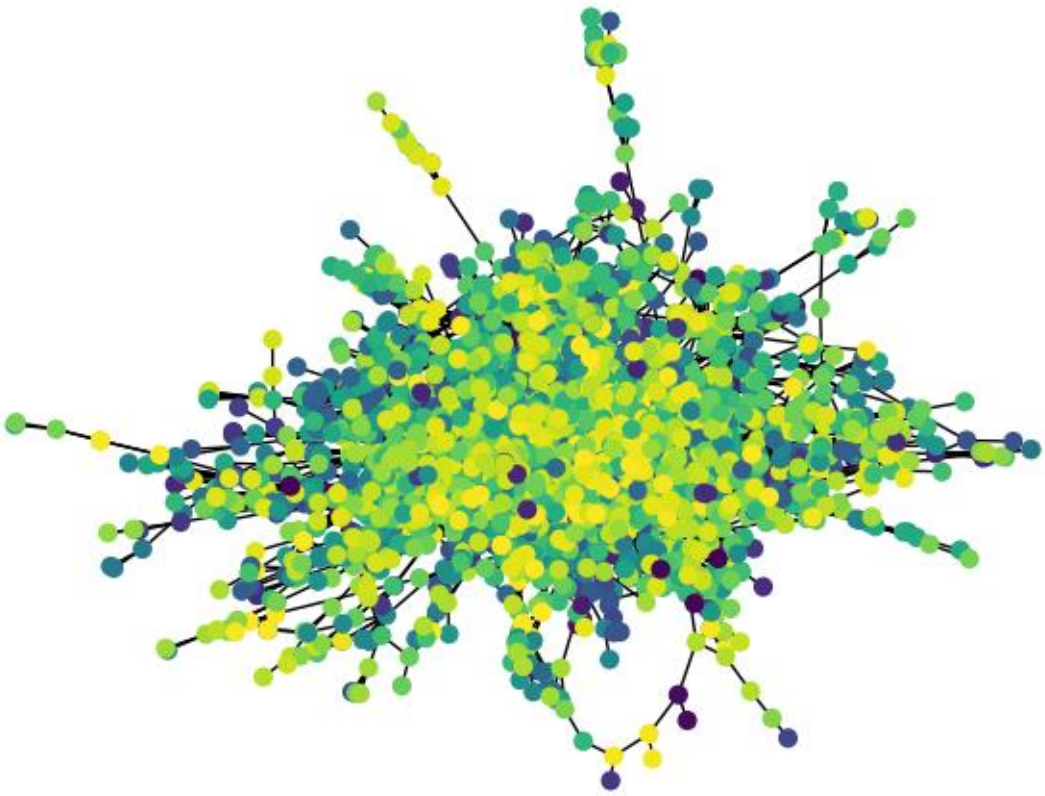
Clique 20



Clique 21



L’individuazione di ogni singola clique non risulta essere particolarmente interessante nel nostro caso, in quanto non vogliamo cercare informazioni di dettaglio sulle pagine, ma valutare le cliques e come si presentano nell’insieme, ma i mezzi a disposizione non ci consentono di apprezzare la suddivisione in cliques del nostro grafo:



5.3 Analisi delle communities di Louvain

Proviamo adesso ad individuare le comunità con l’algoritmo di Louvain, le comunità sono un gruppo di nodi fortemente interconnessi tra di loro rispetto al resto del grafo, rimanendo con l’obiettivo menzionato nel paragrafo precedente, confrontiamo il risultato nell’individuazione delle comunità nel grafo rispetto alla individuazione delle cliques.

L’algoritmo di Louvain è uno tra gli algoritmi di ricerca di communities implementato nella libreria Networkx, per la ricerca delle comunità, inizializza ogni nodo come facente parte di una comunità di cui sono gli unici membri. Gli step successivi sono ripetuti iterativamente, finché la modularità non migliora ulteriormente. La modularità è una misura che valuta la densità delle connessioni della comunità rispetto al resto del grafo. L’algoritmo inizia quindi ad aggregare insieme le comunità per la ricerca delle aggregazioni che aumentino la modularità complessiva del grafo, procede con successive ottimizzazioni e rifiniture finché non trova la combinazione ottimale di aggregazioni, restituendo quindi il grafo separato in gruppi ben connessi tra loro. Il vantaggio rispetto agli altri algoritmi proposti dalla libreria consiste nel lasciare all’algoritmo stesso decidere quante comunità andare a creare, rispetto agli altri algoritmi che richiedono il numero di comunità che si vogliono generare: nella nostra situazione per la quale il dataset è troppo vasto affinché l’utente possa individuarle correttamente, lasciamo che sia l’algoritmo stesso a generare il numero di comunità più appropriato.

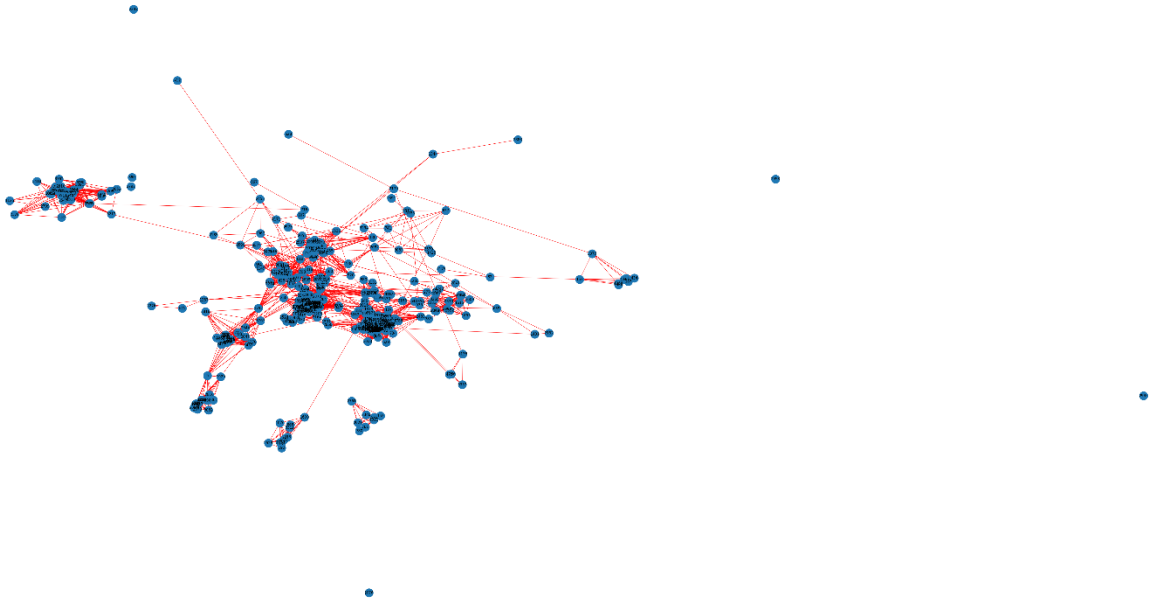
Nel nostro contesto, l'algoritmo ha trovato ben 46 comunità, mostriamo a seguito l'immagine del grafo nel quale mettiamo in risalto le comunità trovate:

Ancora una volta il risultato è confusionario, il centro del grafo è troppo denso per apprezzare le aggregazioni e non ci aiuta a distinguere i gruppi, cioè che si può evincere è che le comunità rimangono comunque ben connesse le una con le altre, data la concentrazione di disparità di colori osservabile.

6 Riduzione del grafo

Nel tentativo di individuare le comunità di pagine rilevanti, abbiamo provato ad individuare il kcore massimale, che ha portato un buon risultato e sembra essere il centro nevralgico delle informazioni, ma dei risultati di poca utilità nel cercare di individuare altri punti importanti per il passaggio delle informazioni, decidiamo dunque di eseguire una ulteriore ricerca di queste aggregazioni, ma lo facciamo su un grafo ridotto: eliminiamo dunque dal grafo tutti i nodi periferici o comunque poco popolari, e lo facciamo andando ad eliminare tutti quei nodi che hanno una degree centrality normalizzata inferiore a 0.005.

Il risultato è visibile nell'immagine sottostante:



Possiamo notare anche ad occhio come questo abbia messo in risalto alcune aggregazioni di nodi, e sebbene alcuni nodi risultino isolati dal resto del grafo (alcuni nodi sono addirittura da soli), è bene ricordare che sono nodi con un grado di centralità superiore rispetto a quelli marginali, e che conservano dunque una buona capacità di connessione con il proprio vicinato. Possiamo dunque pensare che siano i punti per i quali le informazioni passano.

6.1 Valutazione aggregazioni

Valutiamo quale criterio di aggregazione tra le cliques e le communities di Louvain, riesce a definire meglio i gruppi. Confrontiamo le seguenti immagini del nostro grafo ridotto, nel quale in uno abbiamo separato i nodi per cliques, nell'altra per communities.

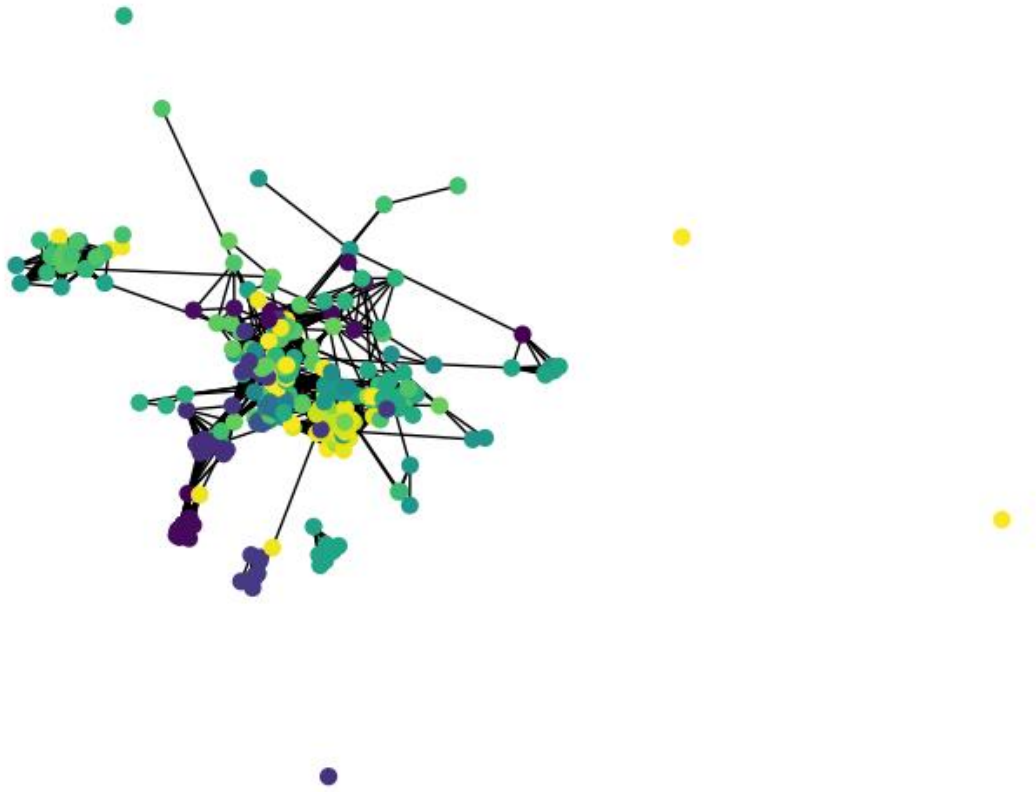


Figura 1Cliques

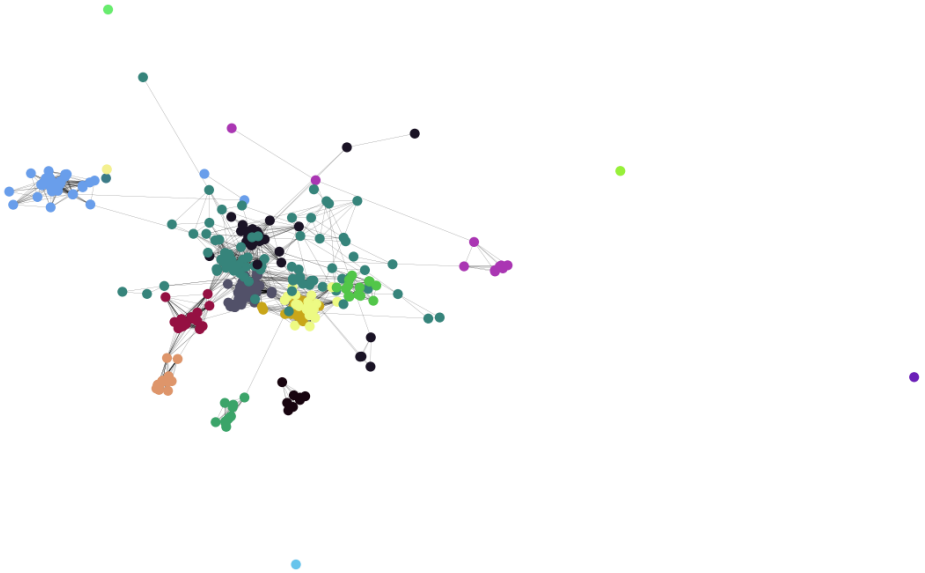
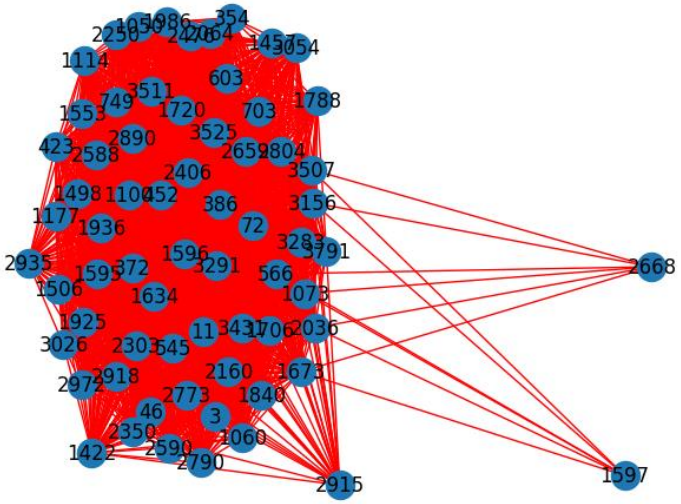
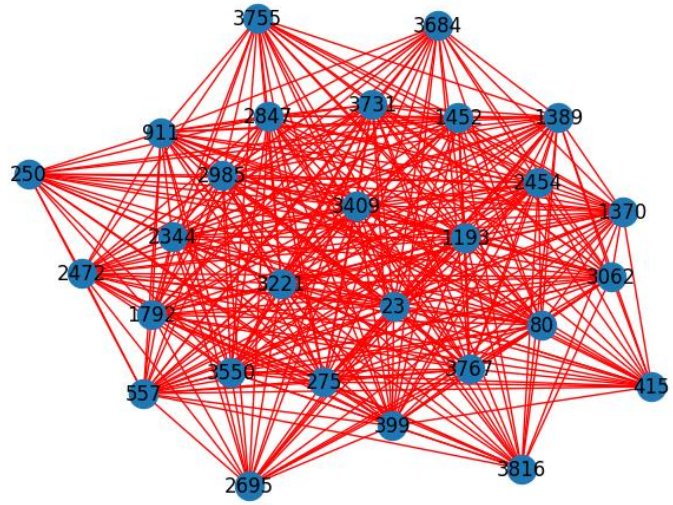
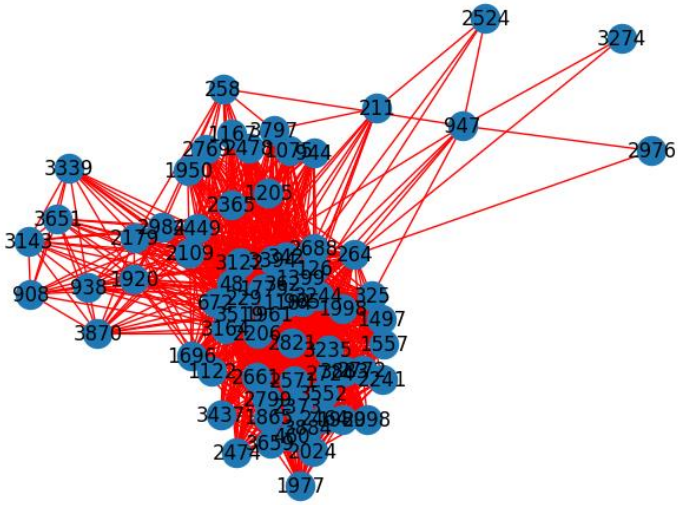


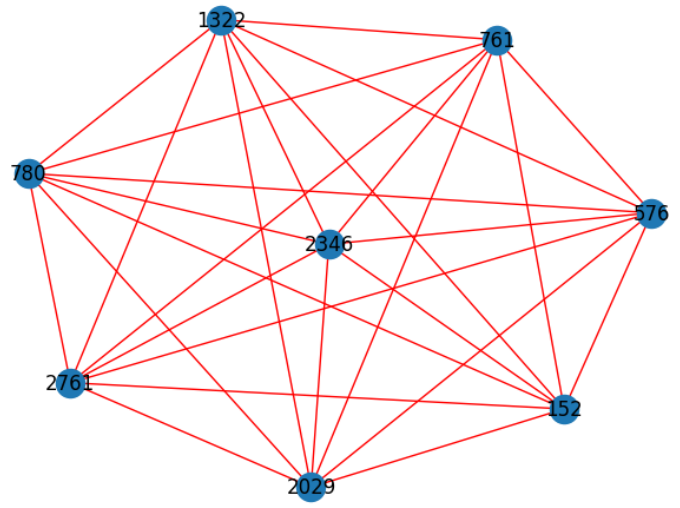
Figura 2Communities

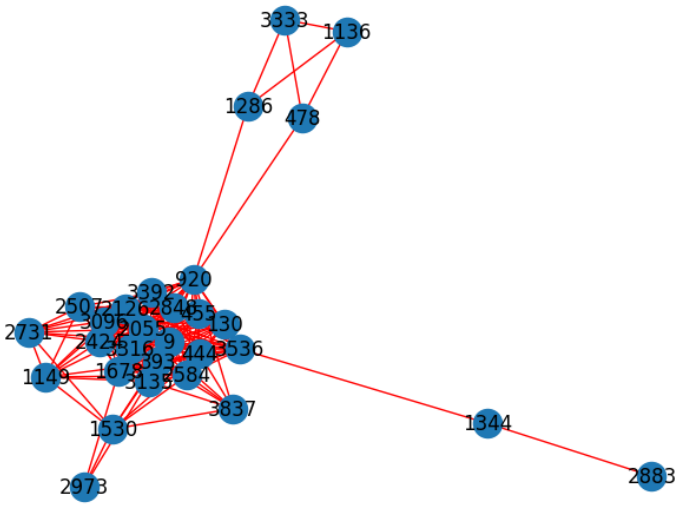
È evidente come le communities riescono ad enfatizzare meglio i gruppi di pagine più correlate tra loro, individuando una serie di centri nevralgici di comunicazione. Pubblichiamo di seguito le communities trovate, per permettere al lettore di apprezzare i nodi individuati all'interno delle varie comunità.

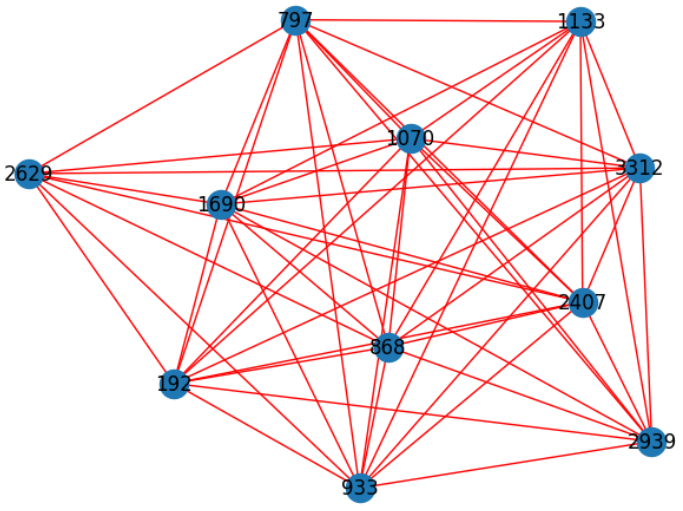


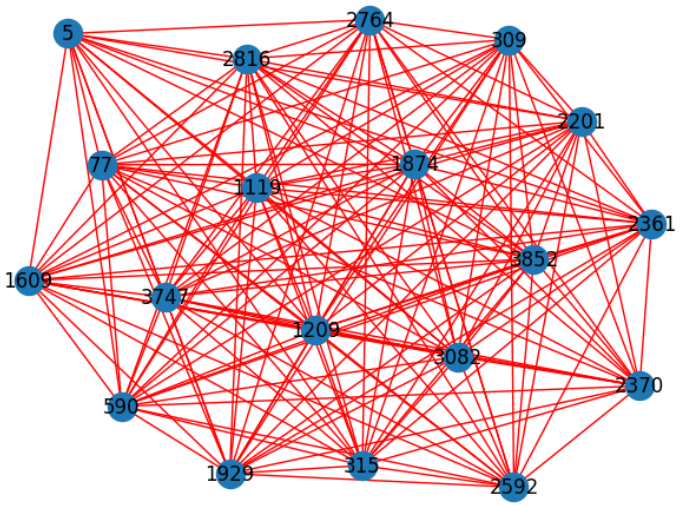


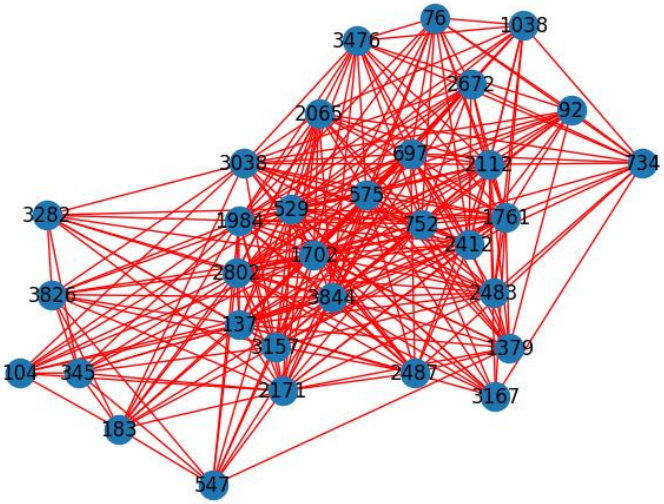


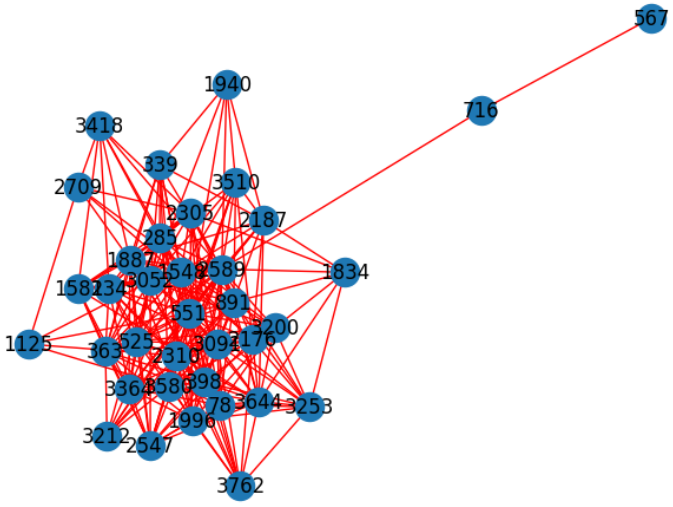

1322
761
780
576
2346
2761
152
2029

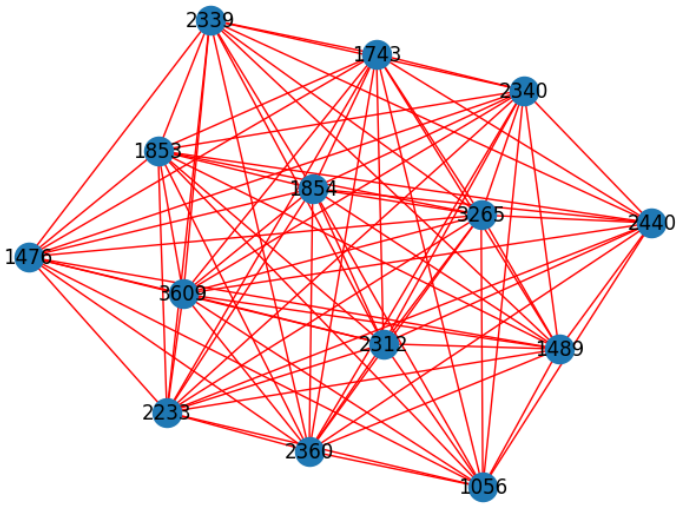

3333
1136
1286
478
920
130
3536
1149
3837
1530
1344
2883
2973

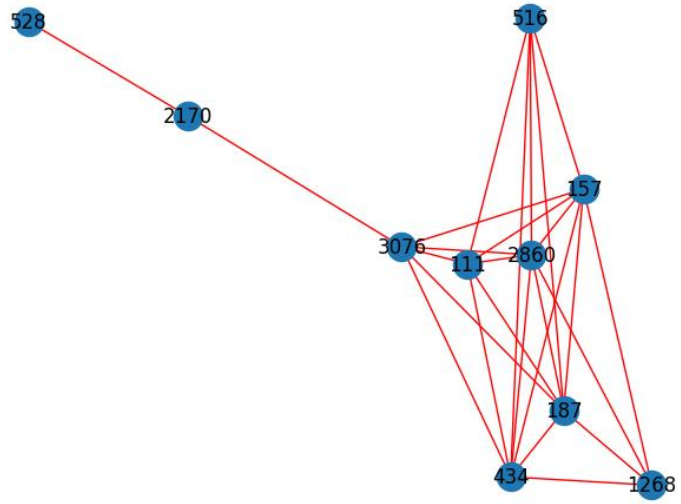

797
1133
1070
2629
3312
1690
2407
868
192
2939
933

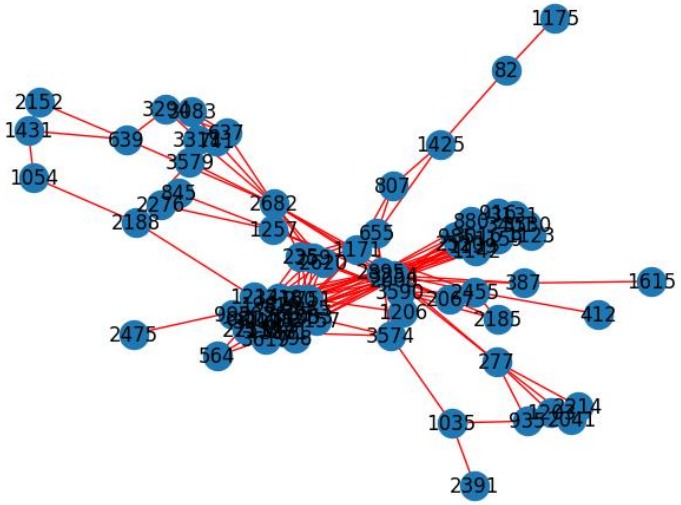

2764
5
309
2816
2201
77
1874
1119
2361
3852
1609
3747
1209
3082
2370
590
1929
315
2592


76
1038
3476
2672
2065
92
697
2112
734
3038
575
529
3282
1984
752
1761
2412
702
2802
3826
3844
2483
137
3157
1379
104
345
2487
2171
3167
183
547


567
1940
716
3418
339
3510
2709
2305
2187
285
1887
1834
1582
1340
3055
2589
891
551
5200
1125
363
525
3094
1176
2310
398
3364
8580
78
3644
3253
1996
3212
2547
3762

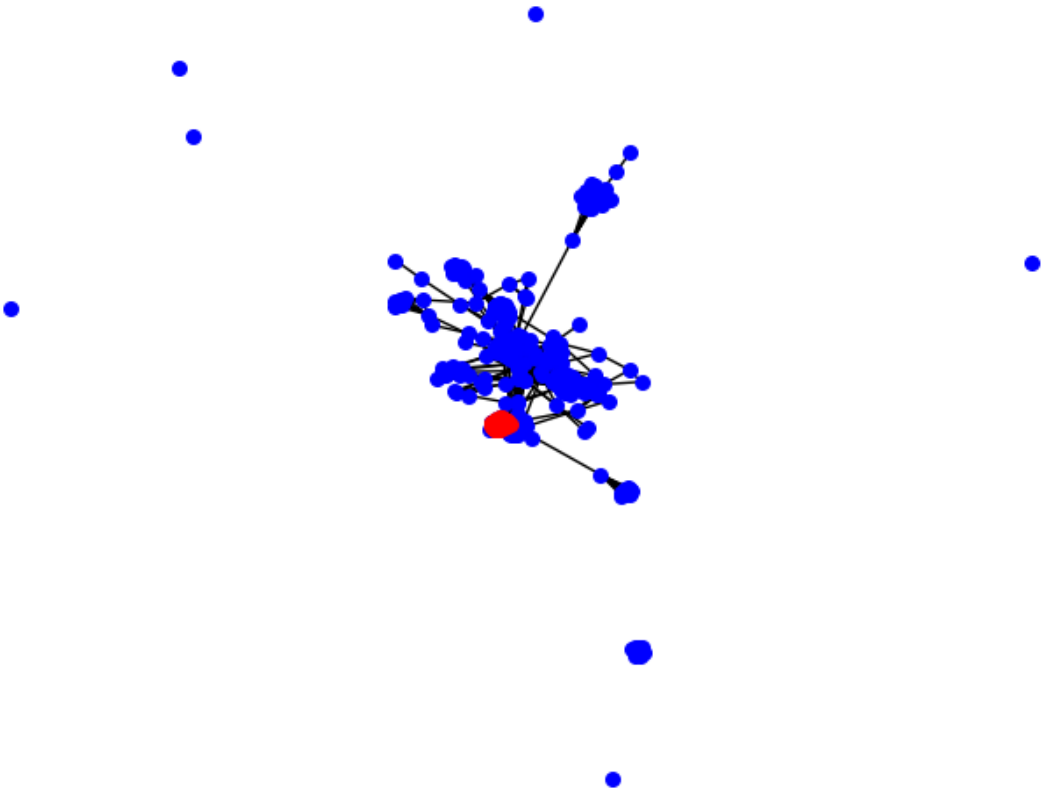

2339
1743
2340
1853
1854
3265
2440
1476
3609
2812
1489
2233
2360
1056

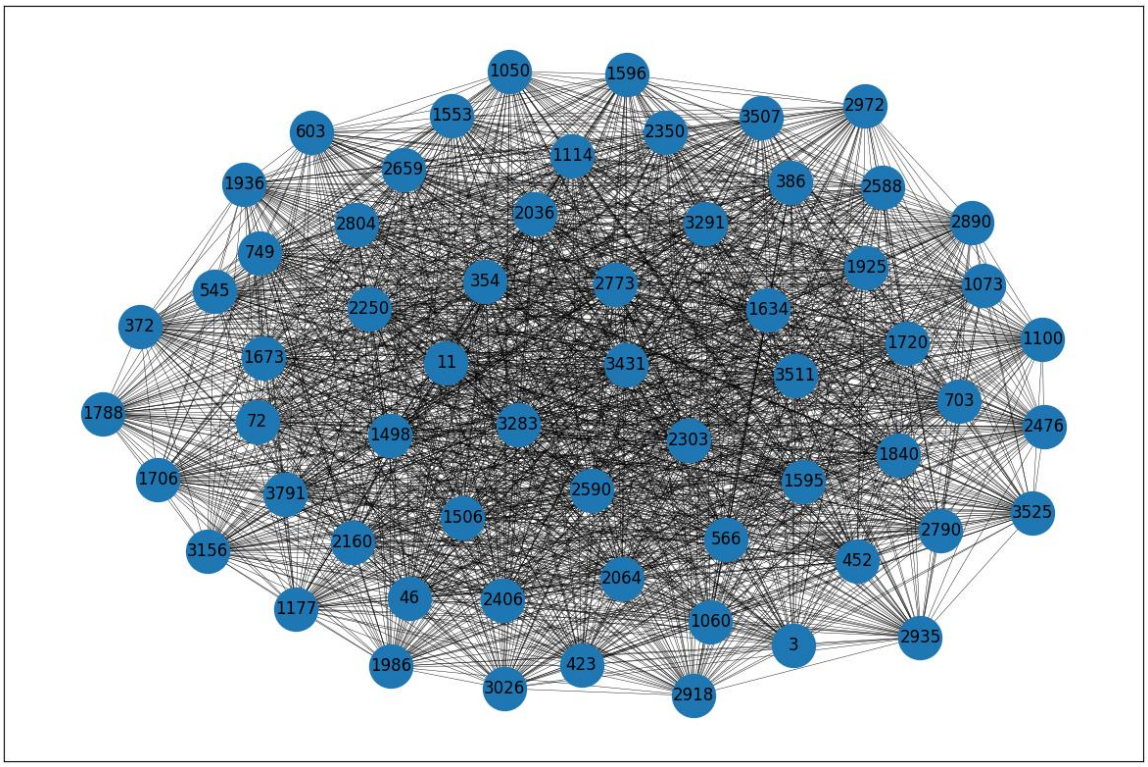

528
516
2170
157
3076
111
2360
187
434
1268


1175
82
2152
1431
639
1425
1054
807
945
2276
2682
2188
1257
655
1615
387
1206
412
2185
2475
3574
564
277
1035
2391

6.1.1 k-core massimale

Ancora una volta rileviamo il k-core massimale del grafo ridotto andando a visualizzarlo singolarmente e all'interno del grafo:





7 Conclusioni

Abbiamo potuto valutare la popolarità delle singole pagine del social Meta e il loro ruolo valutando le loro centralità, appurando come una notizia postata su una singola pagina riesce a raggiungere facilmente le altre pagine e di conseguenza un numero esponenziale di fan che seguono anche solo una di queste, non importa quanto questa pagina sia popolare. Siamo riusciti ad individuare i centri nevralgici di comunicazione o, meglio, i gruppi di pagine centrali vitali per il passaggio di informazioni. E di conseguenza è possibile apprezzare come una singola pagina riesca ad avere un effetto di amplificazione delle proprie notizie, grazie ad una oculata relazione con altre pagine di show televisivi.