UNIVERSITÀ POLITECNICA DELLE MARCHE

FACOLTÀ DI INGEGNERIA

Corso di Laurea Magistrale in
Ingegneria Informatica e dell'Automazione

Social Network Analysis su una rete di utenti di una piattaforma di scambio Bitcoin

Studenti:
Daniele Pallini 1107326
Matteo Abbruzzetti 1108842

Docenti:
Domenico Ursino
Gianluca Bonifazi
Michele Marchetti

Anno Accademico 2022-2023

Indice

1 Introduzione **3**
1.1 NetworkX . . . 3
1.2 Dataset . . . 4
1.2.1 Preprocessing . . . 4

2 Misure di centralità **5**
2.1 Degree centrality . . . 5
2.2 Closeness centrality . . . 8
2.3 Eigenvector centrality . . . 10
2.4 Betweenness centrality . . . 13

3 Cliques & Ego Networks **15**
3.1 Cliques . . . 15
3.2 Ego Networks . . . 17
3.3 Longest shortest path . . . 18

4 Conclusioni **19**

Elenco delle figure **20**

Capitolo 1

Introduzione

In questo elaborato si effettuerà una campagna di Social Network Analysis (SNA) su una rete di utenti di una piattaforma di scambio Bitcoin. Verrà posta l'attenzione sugli strumenti utilizzati e sulle scelte effettuate al fine di estrarre conoscenza e informazioni utili da una rete sociale, la quale caratteristicamente ha una struttura caotica.

1.1 NetworkX

Figura 1.1: Logo di NetworkX

NetworkX è una libreria open source scritta in *Python* per la creazione, la manipolazione e lo studio delle reti. Si utilizza principalmente per la visualizzazione e l'analisi delle strutture a grafo e fornisce numerose funzionalità e algoritmi standard, oltre che metodi per calcolare rapidamente metriche sui grafi stessi. Si tratta, infatti, di uno strumento standard, utilizzato moltissimo nei task di Social Network Analysis.

1.2 Dataset

Il dataset scelto (https://snap.stanford.edu/data/soc-sign-bitcoin-otc.html) consiste in una rete sociale del tipo "who-trusts-whom" (traducibile con "chi si fida di chi"). I nodi rappresentano persone che effettuano transazioni di Bitcoin sulla piattaforma *Bitcoin OTC*. Trattandosi di Bitcoin, gli utenti sono anonimi e caratterizzati da un ID; è presente un meccanismo di reputazione al fine di prevenire potenziali truffe e transazioni con utenti fraudolenti. Il grafo è costituito da 5881 nodi e 35592 archi pesati. Due nodi sono collegati da un arco diretto che va a esprimere il punteggio (da -10, totale sfiducia, a +10, totale fiducia) assegnato da un utente all'altro. Ad un arco corrisponde, quindi, una transazione e, di conseguenza, una interazione tra due utenti.

1.2.1 Preprocessing

Il grafico costruito su tutti i nodi e archi è risultato essere di difficile visualizzazione, per questo motivo si è deciso di effettuare delle operazioni preliminari volte a ridurre il numero di nodi e archi. Si è scelto di rendere il grafo non orientato utilizzando il metodo di *NetworkX to_undirected()*; in questo modo, si è sicuri di avere al massimo un arco tra due nodi e, di conseguenza, si va a ridurre il numero di archi totale. In questa direzione è andata anche la scelta di non considerare i pesi dei diversi archi, in quanto legati intrinsecamente alla direzione dell'arco stesso. Inoltre, si è deciso di considerare esclusivamente i nodi con grado uguale o superiore a 10, ossia i nodi su cui incidono almeno 10 archi. Ciò è stato implementato con il metodo *subgraph()*.

Con le operazioni di preprocessing si è ottenuto un grafo costituito da 1396 nodi e 13104 archi. La visualizzazione grafica è rappresentata in Figura 1.2.

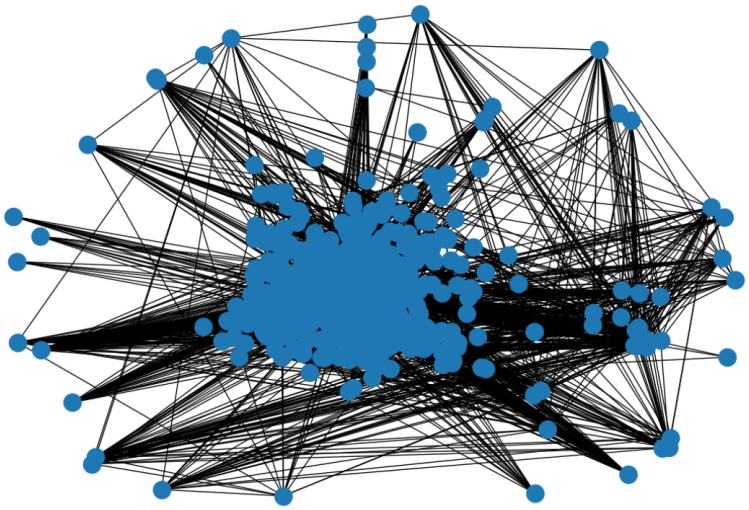

Figura 1.2: Grafo dopo le operazioni di preprocessing

Capitolo 2

Misure di centralità

In questo capitolo verranno presentate le analisi effettuate sul grafo tramite l'utilizzo di diverse metriche di centralità. Queste metriche permettono di comprendere meglio la natura della rete e vanno a individuare i nodi che hanno più importanza e influenza sotto punti di vista differenti.

2.1 Degree centrality

La degree centrality misura il grado di ogni nodo; in questa metrica un nodo è tanto più centrale quanto più alto è il numero di connessioni da cui è caratterizzato. In Figura 2.1 si riporta la distribuzione della degree centrality della rete.

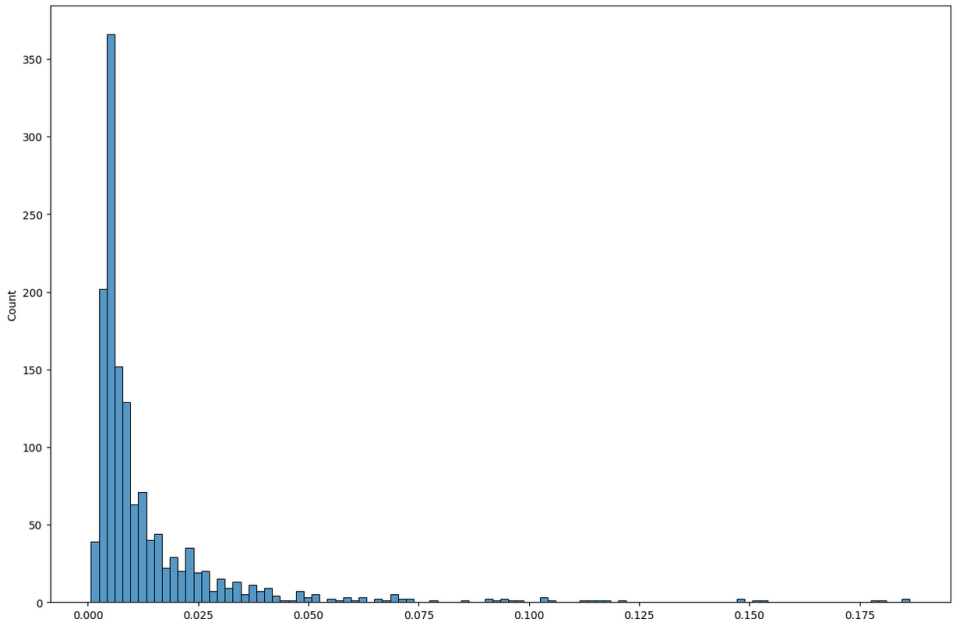

Figura 2.1: Distribuzione Degree centrality della rete

Dalla distribuzione si evince come all'interno della rete sia molto elevato il numero di nodi caratterizzati da degree centrality bassa, tendente allo zero, mentre risultano essere molto pochi i nodi che superano il 10% di connessioni. Questo fattore si riscontra molto frequentemente nelle social networks ed è noto come distribuzione di Pareto. Questa caratteristica si nota bene anche dalla rappresentazione grafica della degree centrality tramite *Spring Layout* di NetworkX, che evidenzia i nodi con colori differenti in base alla misura (Figura 2.2). Risultano, infatti, moltissimi i nodi colorati in blu-violaceo e in rosa-arancione, sintomo di una centrality medio-bassa, mentre sono relativamente pochi i nodi colorati in giallo, ossia quelli con degree centrality più alta. A questo proposito, in Figura 2.3 si riporta la top 10 dei nodi della rete con degree centrality più alta.

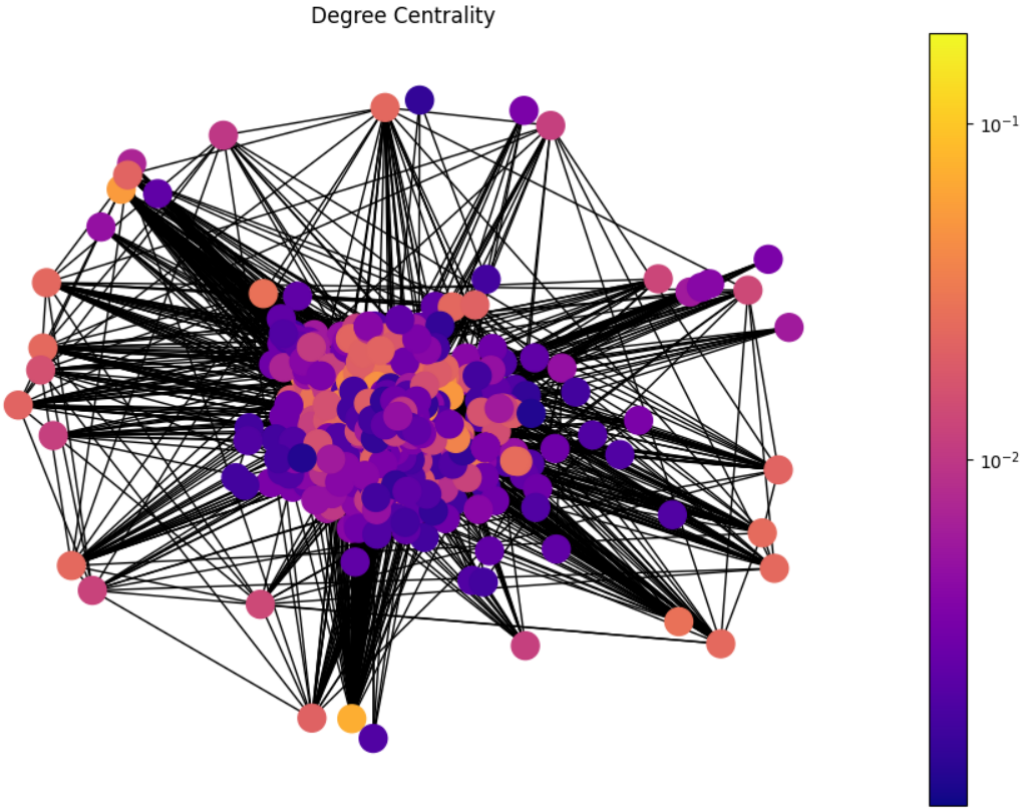

Figura 2.2: Spring Layout Degree Centrality

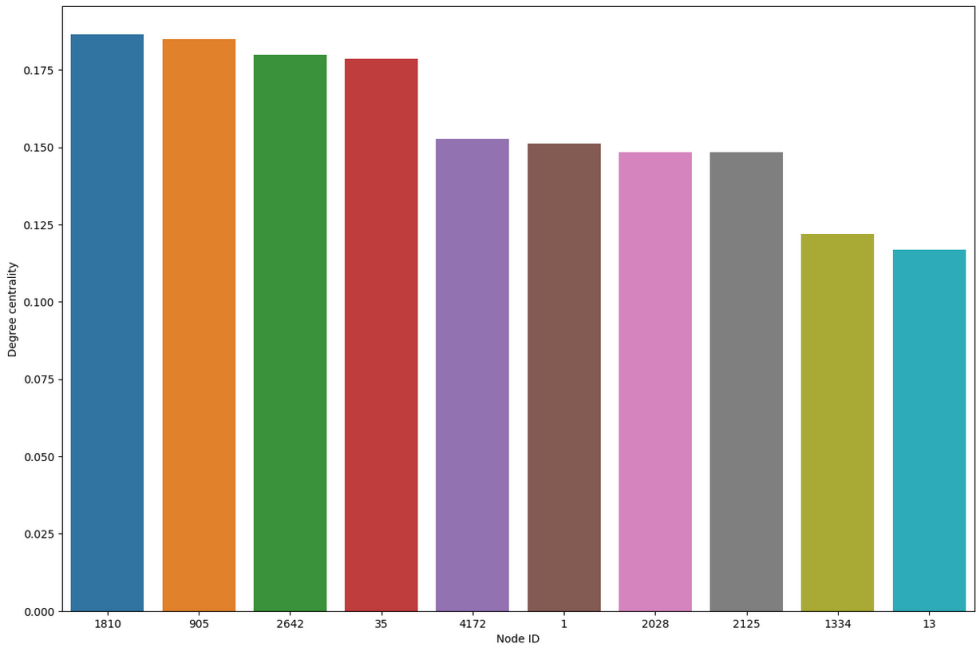

Figura 2.3: Top 10 dei nodi per Degree centrality

2.2 Closeness centrality

La closeness centrality è una metrica che indica quanto un nodo è "vicino" a tutti gli altri nodi della rete. È calcolata come la media della lunghezza dei cammini minimi dal nodo a tutti gli altri nodi della rete ed è utile per identificare i nodi in grado di diffondere informazioni molto efficientemente. In Figura 2.4 si riporta la distribuzione della closeness centrality della rete.

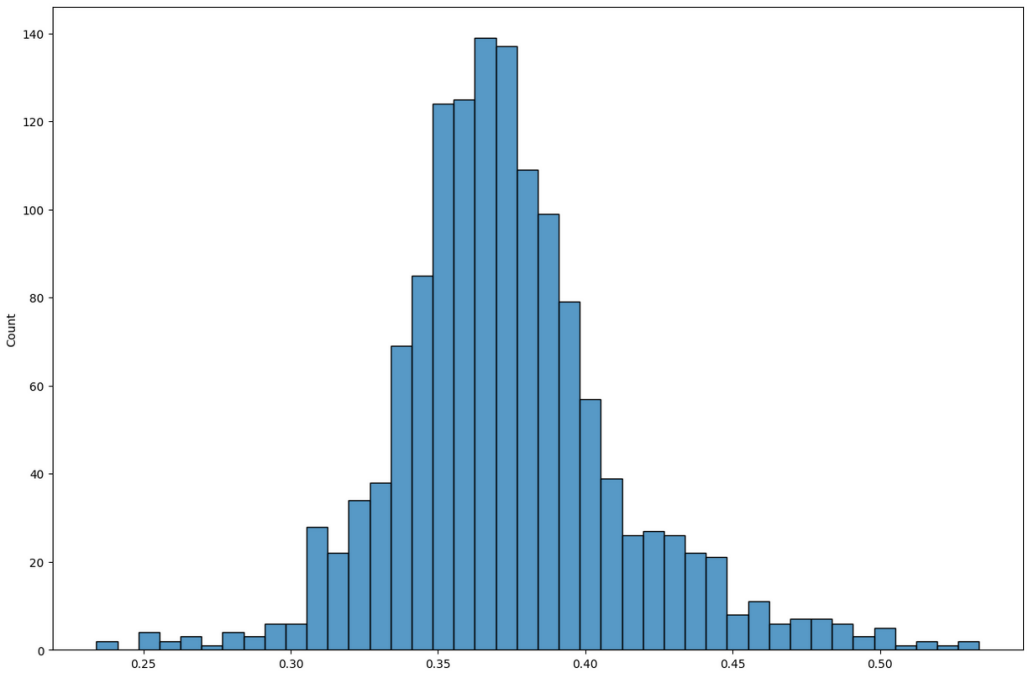

Figura 2.4: Distribuzione Closeness Centrality della rete

Dalla distribuzione a campana si evince come all'interno della rete sia molto elevato il numero di nodi caratterizzati da valori simili di closeness centrality, indicativamente da 0.3 a 0.4. Ciò è indice del fatto che il collegamento tra i vari nodi della rete sia buono e la diffusione di informazioni risulti efficace. Questa caratteristica si può notare bene anche dallo *Spring Layout* (Figura 2.5). Infatti, i colori più diffusi risultano essere il rosa-arancione e l'arancione-giallo, corrispondenti a buoni valori di closeness centrality. Da evidenziare anche la presenza di diversi nodi colorati in giallo, ossia con valori di closeness centrality attorno a 0.5. A questo proposito, in Figura 2.6 si riporta la top 10 dei nodi della rete con closeness centrality più alta. Da notare si ritrovino molti dei nodi con valori più alti di degree centrality; ad esempio il 905, che risulta essere secondo per degree centrality e primo per closeness. I nodi con valori elevati di entrambe le metriche sono dei possibili candidati al ruolo di nodi "più importanti" all'interno della rete.

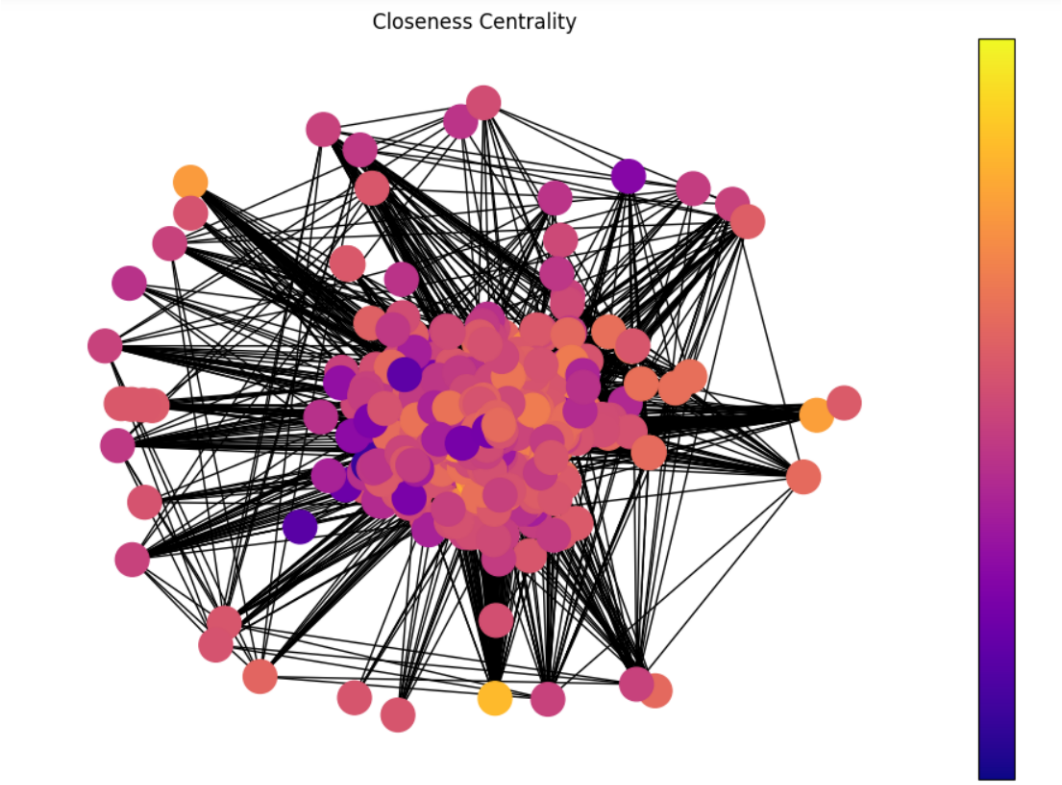

Figura 2.5: Spring Layout Closeness Centrality

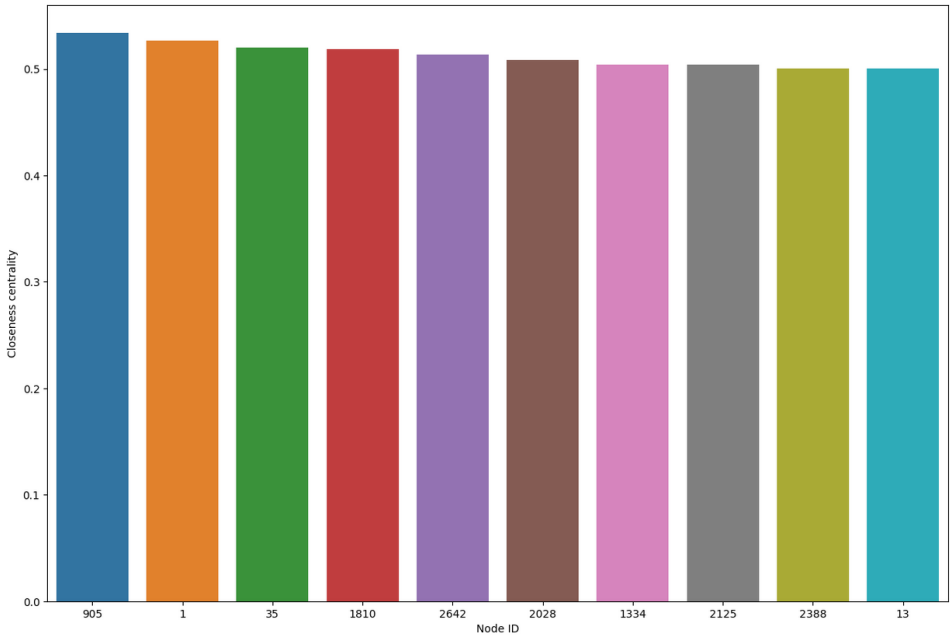

Figura 2.6: Top 10 dei nodi per Closeness Centrality

2.3 Eigenvector centrality

La eigenvector centrality è una misura dell'influenza di un nodo in una rete. Si basa sull'importanza di un nodo rispetto ai suoi vicini; infatti, un nodo è tanto più importante quanto più lo sono i nodi a esso collegati. In Figura 2.7 si riporta la distribuzione della eigenvector centrality della rete.

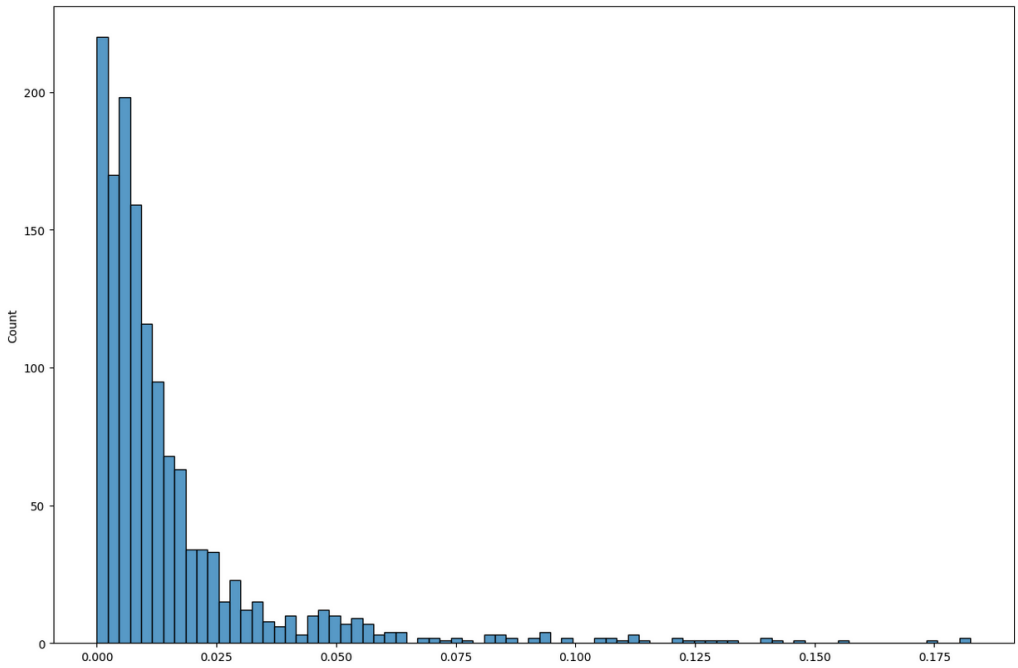

Figura 2.7: Distribuzione Eigenvector Centrality della rete

La distribuzione risulta essere molto simile a quella vista per la degree centrality; ciò non deve sorprendere, in quanto le due metriche sono collegate e danno importanza a fattori simili. Si riscontra, infatti, un alto numero di nodi con coefficienti di eigenvector centrality molto bassi, inferiori a 0.025. Questa caratteristica si può notare bene anche dallo *Spring Layout* (Figura 2.8). Infatti, è netta la predominanza dei nodi colorati in blu e risultano, invece, pochissimi i nodi con colorazioni tendenti al giallo. Andando a vedere in dettaglio la top 10 per eigenvector centrality (Figura 2.9), anche qui si ripresentano molti dei nodi con valori più alti delle due metriche precedenti; in particolare, il 905 risulta essere in top 3 per tutte e 3 le metriche.

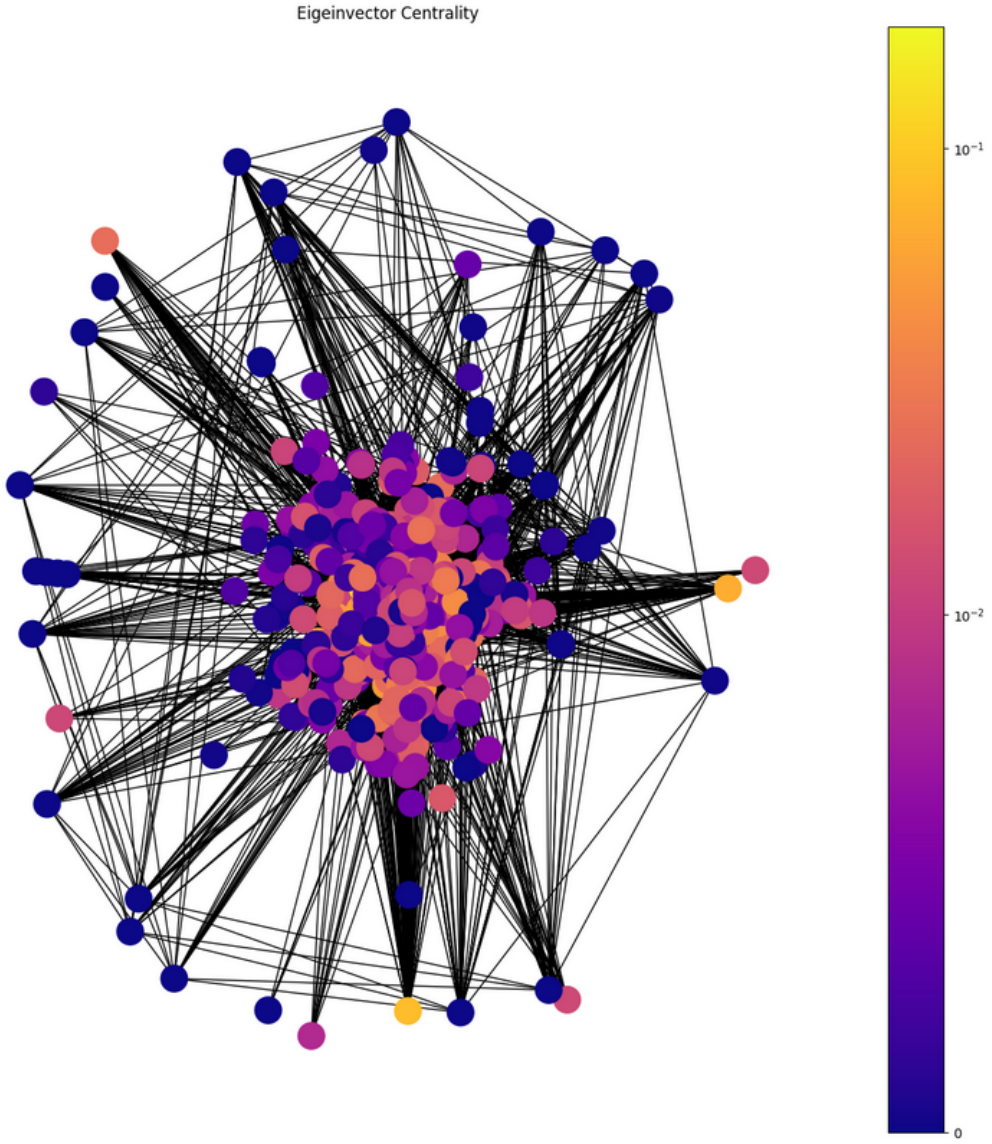

Figura 2.8: Spring Layout Eigenvector Centrality

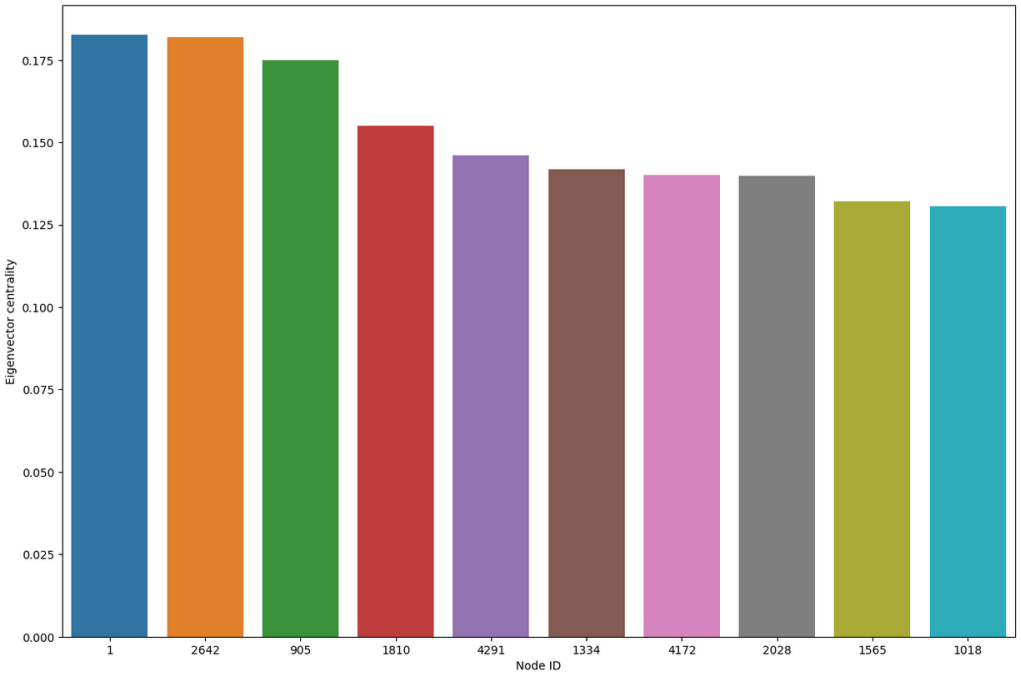

Figura 2.9: Top 10 dei nodi per Eigenvector Centrality

2.4 Betweenness centrality

La betweenness centrality è una metrica utile per misurare l'influenza che ha un nodo sulla propagazione delle informazioni. Viene utilizzata solitamente per individuare i nodi che fungono da "ponte" (o *bridge*) tra le diverse parti della rete. È calcolata in base al numero di volte in cui il nodo specifico è presente all'interno del percorso minimo tra tutte le coppie di nodi. Un nodo con alta betweenness centrality avrà, quindi, un controllo elevato sulla rete, in quanto tante informazioni passeranno per esso.

Nel caso in esame, trattandosi di un sistema basato su blockchain, è logico aspettarsi che il valore di betweenness centrality dei vari nodi risulti in media molto basso. Infatti, le blockchain sono fondate sul concetto di sistema *decentralizzato*, ovvero dall'assenza di enti centrali e regolatori dei flussi di informazione. Non sorprende, di conseguenza, la situazione rappresentata in Figura 2.10, in cui la stragrande maggioranza dei nodi è blu, colore corrispondente a valori molto bassi di betweenness centrality, rasenti lo zero.

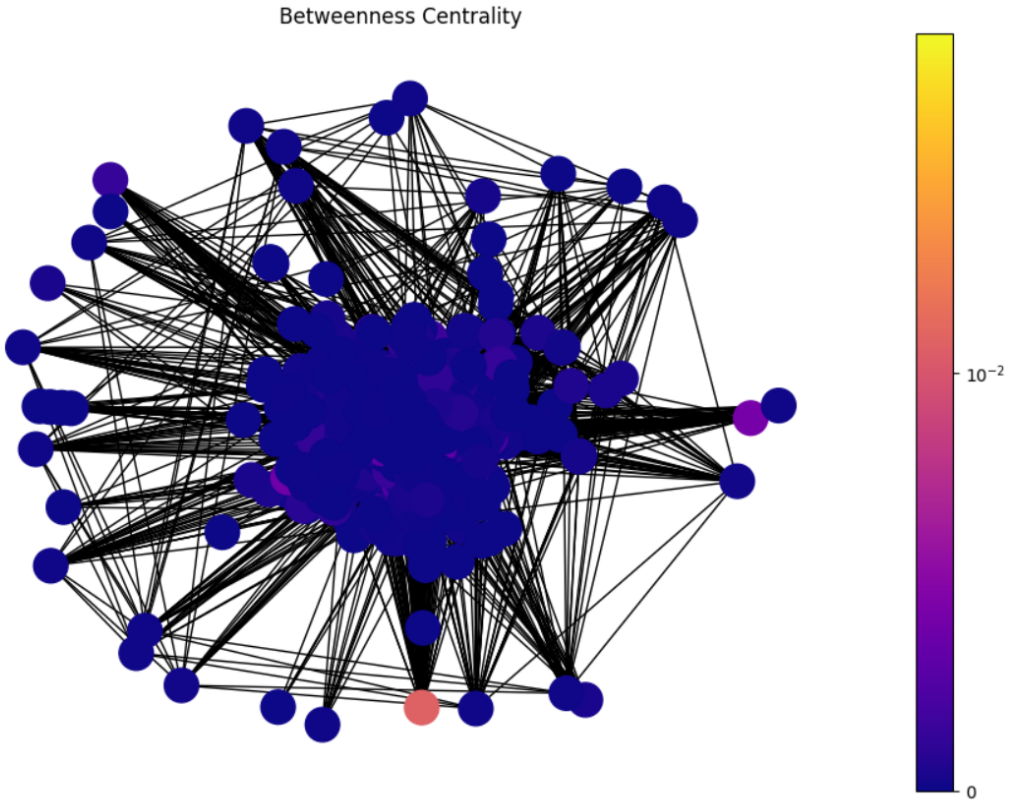

Figura 2.10: Spring Layout Betweenness Centrality

Andando ad analizzare in dettaglio la top 10 dei nodi per betweenness centrality, si evince come anche il valore più alto, relativo al nodo 35, corrisponde

a circa 0.08, una misura indubbiamente bassa (Figura 2.11). Da notare come il nodo 35 compaia nelle prime posizioni anche per degree centrality e per eigenvector centrality, mentre è totalmente assente in quelle riguardanti la closeness centrality. Infine, in seconda posizione è da evidenziare nuovamente la presenza del nodo 905, il quale è, dunque, sul podio in tutte e 4 le metriche analizzate.

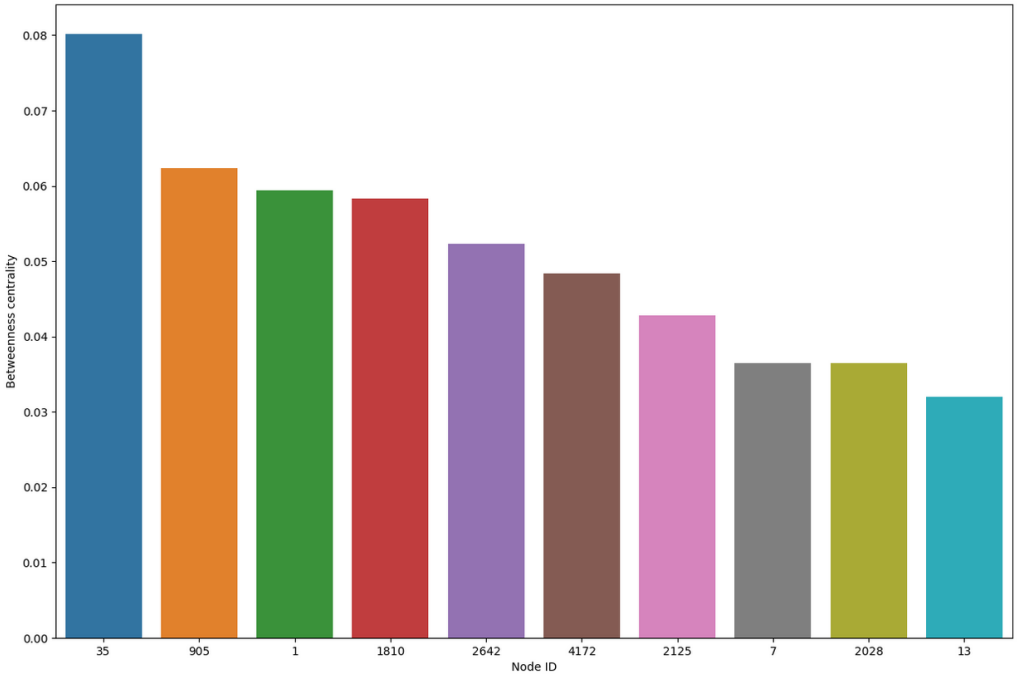

Figura 2.11: Top 10 dei nodi per Betweenness Centrality

Capitolo 3

Cliques & Ego Networks

In questo capitolo verranno individuate e analizzate delle strutture all'interno del grafo note come cliques. Si passerà, poi, all'analisi di singoli nodi e verranno rappresentate le ego networks, ossia l'insieme di tutti i collegamenti del nodo considerato all'interno della rete e di tutti quelli tra i nodi all'interno della ego network.

3.1 Cliques

Una *clique* (o cricca) è un insieme di nodi in un grafo non orientato tale che, per ogni coppia di nodi all'interno dell'insieme, esiste un arco che li collega. In sostanza, una clique è un insieme di nodi totalmente connessi. Lo studio delle clique è utile per individuare la presenza di comunità o di gruppi di nodi fortemente coesi all'interno della rete. Si è partiti individuando tutte le cliques del grafo con il metodo *find_cliques* di *NetworkX*. In Figura 3.1 si riporta la distribuzione delle dimensioni delle cliques trovate.
Dal grafico si può notare come la dimensione più frequente di clique corrisponde al valore 3. La clique più grande, invece, risulta essere costituita da 11 elementi. Un esempio di clique massima è rappresentato in Figura 3.2.
Da evidenziare la presenza del nodo 905, il quale è risultato essere ai vertici delle classifiche per tutte le misure di centralità analizzate nel Capitolo 2. Il fatto di ritrovarlo anche nella clique a dimensione massima non è quindi una sorpresa, al contrario è un'ulteriore dimostrazione del fatto che questo particolare nodo svolge un ruolo cruciale all'interno della rete in esame.

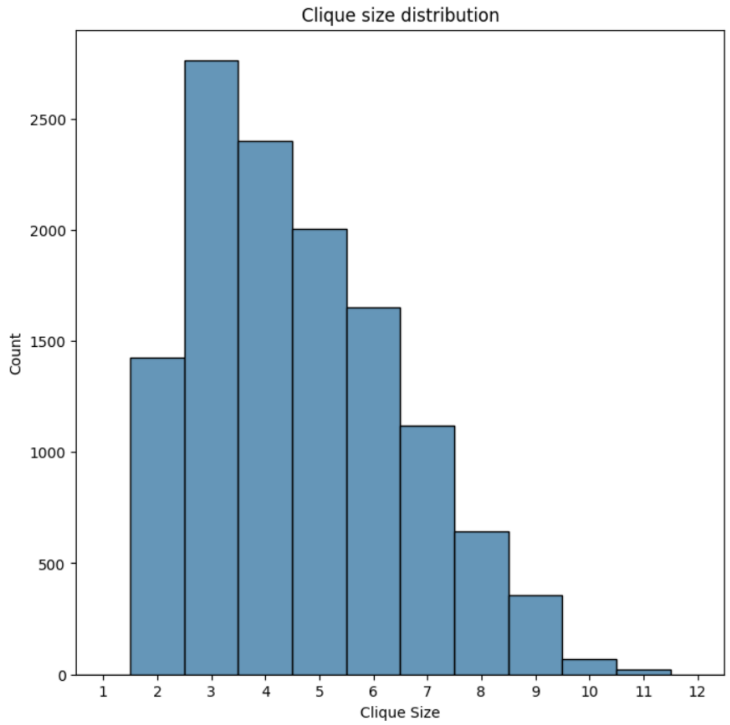

Figura 3.1: Distribuzione dimensione delle cliques della rete

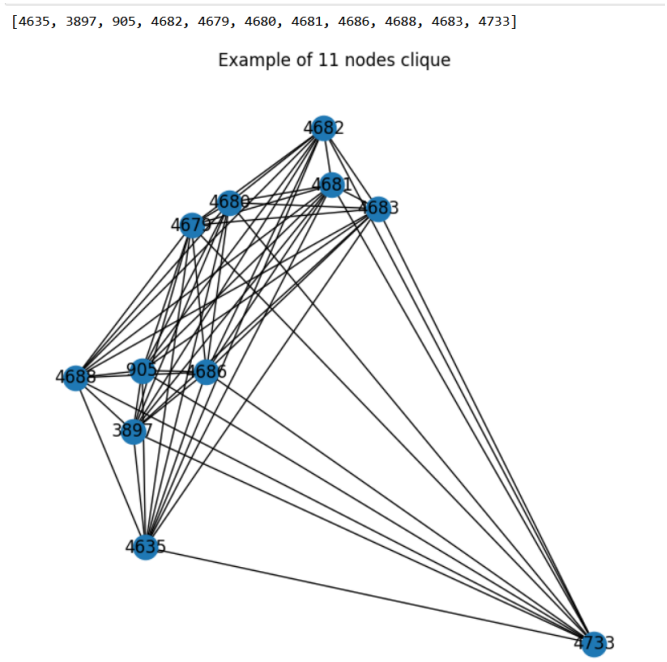

Figura 3.2: Esempio di clique costituita da 11 nodi (massima)

3.2 Ego Networks

Una ego network è una rete centrata su un nodo specifico, definito *ego*. È formata dall'ego e da tutti i nodi collegati direttamente ad esso. Sono inclusi anche gli archi tra tutti i nodi appartenenti alla ego network. Questa tipologia di rete può essere utile per analizzare meglio i nodi con molte connessioni (e dunque importanti) all'interno della rete considerata. Sono state generate due ego networks relative ai nodi **1** e **905**, entrambi rivelatisi in vetta alle classifiche nelle metriche discusse nel Capitolo 2.

L'ego network del nodo **1** (evidenziato in rosso in Figura 3.3) è caratterizzata da 212 nodi e 1895 archi. Ha inoltre una densità di 0.085 e un coefficiente di clustering pari a 0.49. Entrambi i valori non risultano molto alti, a conferma del fatto che non si tratti di una rete molto coesa e densa, come normale che sia in un grafo basato su blockchain.

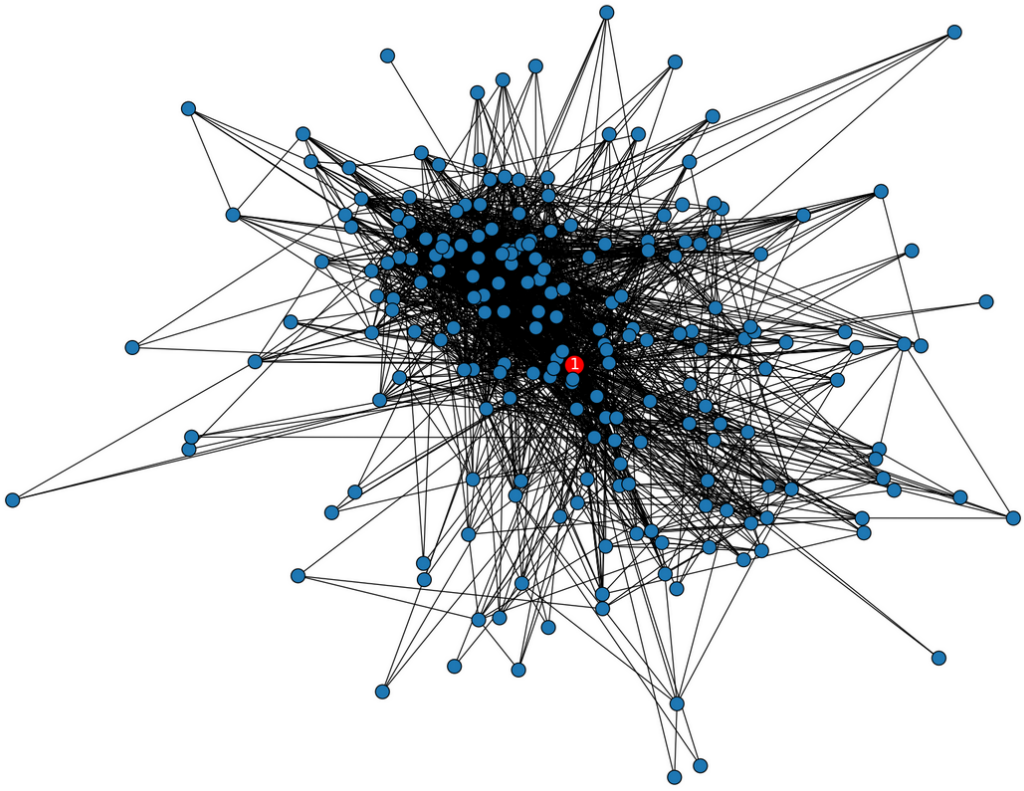

Figura 3.3: Ego network del nodo 1 (evidenziato in rosso)

L'ego network del nodo **905** (evidenziato in rosso in Figura 3.4) è caratterizzata da 259 nodi e 2675 archi. Ha una densità di 0.08 e un coefficiente di clustering pari a 0.52. Rispetto al nodo **1**, la rete è costituita da un numero molto più elevato di archi e nodi, fattore che conferma la fondamentale importanza del nodo all'interno della rete.

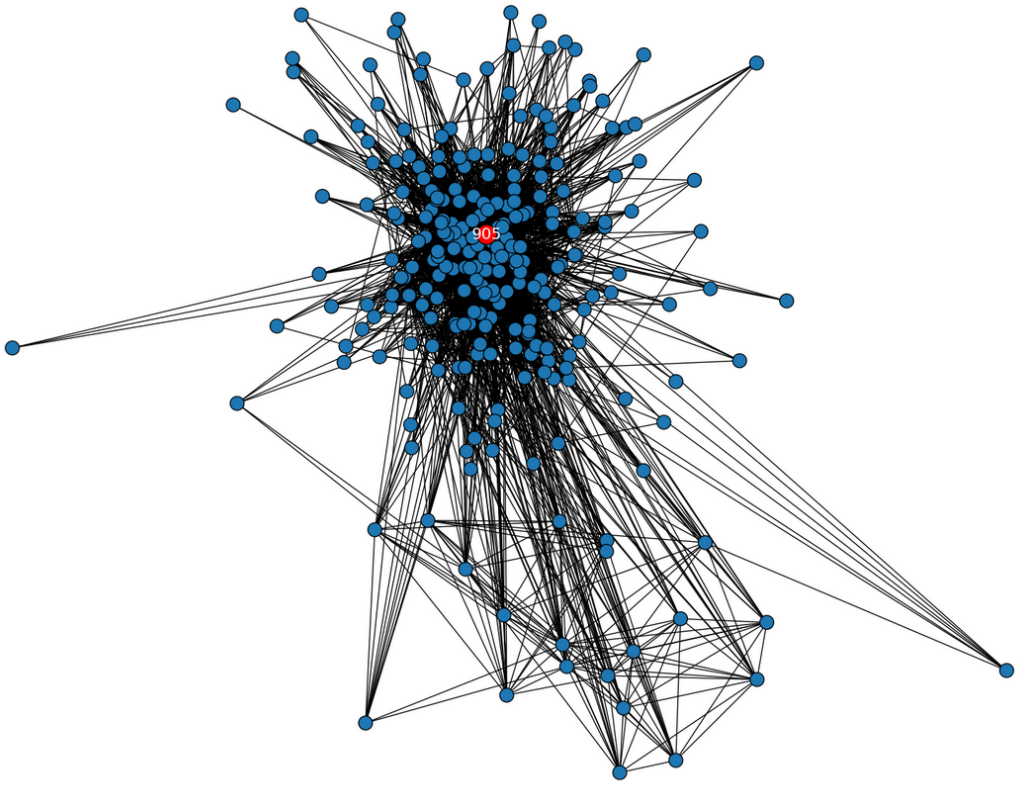

Figura 3.4: Ego network del nodo 905 (evidenziato in rosso)

3.3 Longest shortest path

Come ultimo elemento, si è andati a calcolare il più lungo percorso minimo all'interno del grafo, con il metodo *shortest_path* di *NetworkX*. Globalmente, la lunghezza media del percorso minimo tra due nodi è risultata essere pari a 2.7. Il percorso più lungo, invece, si è rivelato quello che collega il nodo **707** e il **1062** ed è costituito nel seguente modo:

707 -> 570 -> 563 -> 1 -> 36 -> 678 -> 1062

La dimensione del percorso è pari a 6, esito che conferma la teoria dei "6 gradi di separazione", ipotesi secondo la quale ogni persona può essere collegata a qualunque altra persona o cosa attraverso una catena di conoscenze e relazioni con non più di 5 intermediari.

Capitolo 4

Conclusioni

In questo elaborato sono stati utilizzati *Python* e la libreria *NetworkX* per una campagna di Social Network Analysis su una rete sociale basata su una blockchain.
Si è partiti dalla preparazione del dataset, con la rimozione dei nodi con grado basso e degli archi relativi, in modo da poter eseguire una analisi più accurata.
Il passo successivo è stato l'individuazione dei nodi più importanti all'interno della rete in base a 4 diverse misure di centralità: Degree, Closeness, Eigenvector e Betweenness centrality. In questo modo si è avuta una panoramica del grafo sotto punti di vista differenti.
In seguito, si è passati all'analisi dei gruppi di nodi, al fine di identificare le cliques a dimensione massima. Si è posta, poi, l'attenzione sui nodi individuati come più importanti, costruendo le due ego network relative.
Infine, è stato calcolato il percorso minimo più lungo ed è stata verificata la teoria dei 6 gradi di separazione.
In generale, la campagna di Social Network Analysis effettuata ha confermato le caratteristiche tipiche di una rete basata su blockchain, ovvero di un sistema decentralizzato (ad esempio, la bassa betweenness centrality). Nonostante questo, il lavoro effettuato ha allo stesso tempo evidenziato come alcuni nodi siano molto "più centrali" di altri anche in questo tipo di social network.

Elenco delle figure

1.1 Logo di NetworkX 3
1.2 Grafo dopo le operazioni di preprocessing 4

2.1 Distribuzione Degree centrality della rete 5
2.2 Spring Layout Degree Centrality 6
2.3 Top 10 dei nodi per Degree centrality 7
2.4 Distribuzione Closeness Centrality della rete 8
2.5 Spring Layout Closeness Centrality 9
2.6 Top 10 dei nodi per Closeness Centrality 9
2.7 Distribuzione Eigenvector Centrality della rete 10
2.8 Spring Layout Eigenvector Centrality 11
2.9 Top 10 dei nodi per Eigenvector Centrality 12
2.10 Spring Layout Betweenness Centrality 13
2.11 Top 10 dei nodi per Betweenness Centrality 14

3.1 Distribuzione dimensione delle cliques della rete 16
3.2 Esempio di clique costituita da 11 nodi (massima) 16
3.3 Ego network del nodo 1 (evidenziato in rosso) 17
3.4 Ego network del nodo 905 (evidenziato in rosso) 18