UNIVERSITÀ DI PISA

Dipartimento di Informatica

Corso di Laurea in Informatica

Analisi di centralità degli utenti della social network Steemit

Relatori:

Prof.ssa Laura Ricci

Dott. Andrea Michienzi

Presentata da:

Simone Lissandrello

Sessione primaverile

Anno Accademico 2020/2021

Indice

Capitolo 1

Introduzione

La Social Network Analysis (SNA) è un processo di investigazione della struttura sociale attraverso l'utilizzo di reti e strumenti appartenenti alla teoria dei grafi. La metodologia attuata dalla social network analysis ha scaturito interesse e curiosità tra le scienze comportamentali e sociali. L'interesse è dato dall'utilizzo di strumenti empirici per analizzare le relazioni che intercorrono tra entità sociali, e inoltre dalla formulazione di modelli e implicazioni dovute a queste relazioni. Molti ricercatori si sono resi conto che la prospettiva della rete consente una nuova leva per rispondere alle domande standard della ricerca nelle scienze sociali e comportamentali fornendo una definizione formale e precisa sugli aspetti economici, politici e sociali. In una concezione mai vista nella maggior parte delle altre discipline delle scienze sociali, i metodi della SNA si sono sviluppati negli ultimi cinquant'anni come parte integrante dei progressi nella teoria sociale, nella ricerca empirica e nella matematica e statistica formale. In contrapposizione alle altre prospettive di ricerca all'interno delle scienze sociali e comportamentali nella quale si fa riferimento in maggior parte agli attributi personali dei membri del sistema, nella social network analysis si considerano una componente fondamentale le possibili interazioni tra le unità interagenti. Tramite il crescente interesse e l'aumento dell'uso di questa nuova scienza si sono instaurati dei principi centrali a cui si fa riferimento [25]. Oltre all'uso di concetti relazionali, notiamo quanto segue come importante:

- Gli attori sociali e le loro azioni sono visti come unità interdipendenti piuttosto che come unità autonome e indipendenti;

- I legami relazionali (collegamenti) tra gli attori sociali sono canali per il trasferimento o ”flusso” di risorse;
- I modelli di rete incentrati sugli individui vedono l'ambiente strutturale della rete come una fonte di opportunità o vincoli all'azione individuale;
- I modelli di rete concettualizzano la struttura (sociale, economica, politica e così via) come modelli duraturi di relazioni tra attori sociali.

Tra i concetti fondamentali instaurati dalla SNA, troviamo la definizione di ”struttura”. Senza negare l'utilità dello studio delle funzioni svolte dai vari componenti di un sistema di interdipendenza, la struttura offre la possibilità di poter studiare separatamente le diverse forme delle relazioni in quanto tali (Chiesi, 1980, p.297) [5]: infatti, ”analisi strutturale” è assunto come sinonimo di analisi di rete [9]. L'analisi strutturale lavora con un concetto di struttura che enfatizza gli elementi posizionali, cancellando completamente gli elementi simbolici e culturali. Il noto sociologo Blau afferma che: “L'analisi strutturale focalizza la sua attenzione precisamente sulla struttura sociale e assume che la vita sociale, incluse le sue manifestazioni culturali, è radicata nella struttura delle posizioni e delle relazioni sociale e che la vita sociale può essere spiegata analizzando questi modelli o le distribuzioni di posizioni [4]. Come già esposto, la relazione è l'unità di base della struttura sociale. E dunque, come affermano Wellman e Berkowitz, che hanno elaborato il paradigma più completo dell'analisi strutturale: “Le strutture sociali possono essere rappresentate come networks, cioè come insiemi di nodi (o membri del sistema sociale) e come insiemi di legami che indicano le loro interconnessioni”. I nodi possono rappresentare persone, ma anche ”gruppi, grandi società, stati-nazione o altre collettività”. I ”legami” sono usati per rappresentare flussi di risorse, relazioni di amicizia simmetrica, trasferimenti, o relazioni strutturate tra nodi” [27]. Bavelas fu il primo a realizzare che una posizione centrale nella struttura del network corrisponde a potere in termini di indipendenza, da, influenza e controllo sugli altri. I lavori successivi sulla centralità effettuati da Freeman rivisitarono e coordinarono un corpo di ricerche prodotte fin dall'inizio degli anni Cinquanta e definì un insieme di indici di centralità [16]. Oltre a tali misurazioni basate sugli elementi centrali, l'analisi strutturale comprende ulteriori metriche come: la densità, la coesione, individuazioni di cluster, ecc.

L'applicazione della metodologia di analisi, in questo contesto, viene attuata a un social media basato sulla tecnologia delle blockchain. Tale sistema sociale, oggetto di studio per questa ricerca, il quale vanta un milione e più utenti registrati dal marzo del 2016 è Steemit. La tesi percorre e descrive il cammino svolto durante il tirocinio, dividendo in capitoli i diversi argomenti studiati in modo tale da dividere logicamente le diverse materie che entrano in gioco nella ricerca. In linea generale, tralasciando per un attimo la suddivisione in capitoli, possiamo organizzare la struttura della tesi in due parti: la prima utile a spiegare l'ambito del tirocinio e i concetti appresi per svolgere correttamente il lavoro (dal capitolo 2 al 5, compresi); la seconda descrive l'attuazione nella pratica dei concetti studiati in teoria. In ordine di stesura dei capitoli, tralasciando il capitolo dell'introduzione, presentiamo il preambolo del tirocinio come tutto ciò che riguarda le fondamenta della ricerca: cioè una descrizione sommaria sul sistema su cui attuiamo le analisi. Il capitolo successivo abbraccia una descrizione teorica delle nozioni statistiche che entrano in gioco nella fase di analisi della ricerca sperimentale, descrivendo i metodi utilizzati durante la fase di statistica descrittiva e i test statistici nella statistica inferenziale, con un discreto livello di astrazione senza dilungarci troppo nella vasta complessità della materia, così da garantire una comprensione più immediata. Una volta presa coscienza dell'ambito su cui si svolge la ricerca e delle nozioni teoriche di base, descriviamo quali tecnologie vengono effettivamente impiegate durante il tirocinio, includendo una descrizione dei linguaggi di programmazione e delle librerie implementate. Terminato il quarto capitolo, entriamo nel fulcro del tirocinio descrivendo le proprietà che caratterizzano i sistemi sociali dal punto di vista matematico e le varie metriche analizzate e studiate sui nodi della rete Steemit con alcuni accenni agli algoritmi implementati nelle librerie Networkit. Una volta introdotti alle teorie e le fondamenta della ricerca si è in grado di comprendere all'atto pratico il lavoro e le motivazioni per cui si sono scelti certi strumenti; infatti, il sesto capitolo può essere considerato il cuore del tirocinio, descrivendo minuziosamente tutte le fasi del lavoro con l'attuazione pratica di tutti i concetti descritti nei capitoli introduttivi. L'ultimo capitolo offre le conclusioni raccolte al termine del tirocinio con gli obiettivi raggiunti e i possibili lavori futuri. Gli appendici conclusivi illustrano, rispettivamente: i codici Python e gli script in R nell'appendice A, i grafici delle distribuzioni nell'appendice B ed eventuali grafici integrativi al lavoro.

Capitolo 2

I Blockchain Online Social Media

Il capitolo è suddiviso in più sezioni, utili per illustrare il funzionamento, i vantaggi e svantaggi dei sistemi presi in esame in questo tirocinio. Le prime quattro sezioni espongono una descrizione sommaria della tecnologia (Blockchain) sul quale si fonda il social media (Steemit), con una descrizione sulla sua architettura, le sue dinamiche e la monetizzazione dei contenuti.

2.1 Blockchain

La blockchain (letteralmente “catena di blocchi”) è una struttura dati condivisa e “immutabile”. Essa è pensata e implementata come un vero e proprio registro digitale, con la caratteristica intrinseca di crescere nel tempo, in grado di contenere informazioni condivise con tutta la rete. Le interazioni all’interno del sistema vengono raggruppate in “blocchi”, che vengono concatenati in ordine cronologico, e la cui integrità è garantita dall’uso della crittografia. Tale struttura dati essendo immutabile specifica che, di norma, il suo contenuto una volta scritto non è più né modificabile né eliminabile, a meno di non invalidare l’intera struttura. Ogni blocco è suddiviso in due componenti: una prima parte utile ad identificare il blocco chiamato header contenente il timestamp del blocco (quindi la data di creazione) e un insieme di informazioni utili a identificarlo; la seconda parte contiene l’informazione vera e propria, quindi un record di transazioni registrate in quel momento nel sistema.

In principio le blockchain sono nate come implementazione di ”distributed ledger”, cioè registri distribuiti tra tutte le entità del sistema, per il semplice scambio di informazioni. Negli

ultimi anni, si è aggiunta un'ulteriore funzionalità alla tecnologia delle blockchain riguardante l'implementazione e la gestione in totale sicurezza di criptovalute. Soffermandoci sul fattore di immutabilità e di condivisione, le blockchain sono particolarmente utili in situazioni in cui due soggetti vorrebbero effettuare una transizione, ma l'uno è diffidente nei confronti dell'altro; nella società civile, sono state pensate e sviluppate istituzioni come gli studi legali e le banche, per assicurare lo scambio di denaro o di proprietà. Immaginandoci il registro digitale come un'istituzione, esso consente lo sviluppo di un rapporto sociale dove, grazie alla partecipazione di tutti, si è in grado di verificare le transazioni, in totale trasparenza, grazie al fatto che le transazioni vengono registrate su archivi inalterabili. Questo è possibile grazie all'uso di un sistema decentralizzato e distribuito, senza la presenza di un'entità centrale che decide e garantisce il corretto funzionamento della piattaforma, e soprattutto grazie alla distribuzione delle copie del registro visibile a tutti gli utenti del sistema, cosicché sia facile risalire alla veridicità delle transazioni eseguite. I vantaggi dall'utilizzo, in generale, delle architetture distribuite porta all'inattaccabilità da parte di utenti malevoli al registro distribuito che forma la rete. Un utente malintenzionato per poter compromettere un sistema così costruito dovrebbe poter accedere all'archivio di tutti i possessori di una copia del registro ed effettuare delle modifiche che siano coerenti tra loro, totalmente differente da un'architettura client-server dove tutti i dati dei client sono contenuti in un'unica macchina la quale offre, con semplicità, un bersaglio da violare. Le entità che creano la rete vengono chiamati nodi, tali nodi partecipano e contribuiscono alla crescita della catena di blocchi in totale democrazia, attraverso l'uso di un algoritmo di consenso, protocollo fondamentale per il corretto funzionamento del sistema. Tale protocollo è in grado di garantire il rispetto delle regole della blockchain da parte dei nodi, in modo che tutte le transazioni avvengano correttamente e oltre a tutto ciò è uno dei motivi principali che rende difficile modificare una blockchain. All'atto pratico, l'algoritmo di consenso, viene utilizzato per decide qual è il prossimo blocco da aggiungere alla catena di blocchi. Presentiamo tre tipi di protocolli implementati in diversi sistemi reali: Proof of Work (PoW), Proof of Stake (PoS) e Delegated Proof of Stake (DPos). In ordine di citazione descriviamo:

- La PoW è il primo algoritmo di consenso progettato in ambito blockchain, creato per validare le transazioni e raggiungere il consenso all'interno della rete. La caratteristica

fondamentale del protocollo consiste nella competizione creata tra gli utenti per la pubblicazione dei nuovi blocchi. Tale funzione di miner è permessa a tutti i nodi della rete a patto di risolvere un problema matematico con difficoltà crescente che solo pochi utenti possono permettersi di risolvere, a causa dello spreco di corrente e delle caratteristiche hardware necessarie. Il nodo che, prima di tutti gli altri, risolve il problema in questione, riceve una ricompensa in denaro dopo aver effettivamente creato il blocco;

- La PoS è una possibile alternativa al PoW. Il protocollo in questione, oltre a validare le transazioni e raggiungere il consenso all'interno della rete, introduce all'interno dell'algoritmo un processo di elezione pseudo-casuale del validatore del nuovo blocco. Il validatore di ciascun blocco viene determinato da un investimento della stessa criptovaluta e non dalla potenza di calcolo utilizzata, rispetto la PoW. Nella scelta del nodo miner vengono valutati la quantità di criptovaluta "bloccata" e messa a disposizione per la convalida di un nuovo blocco (questo fondo, messo a disposizione, è chiamato "stake"), e la ricompensa sarà esclusivamente legata alle tasse generate dalle transizioni effettuate sulla rete. Non vi è alcuna nuova creazione di criptomoneta, in quanto è già stata tutta distribuita durante la fase di lancio del progetto e della criptovaluta.

- Il DPoS può essere considerato un'evoluzione del PoS. Il Delegated Proof of Stake venne creato per cercare di velocizzare il processo di creazione dei blocchi, rallentato a causa di alcune problematiche presenti nei protocolli illustrati precedentemente. Il DPoS è costituito da due fasi: l'elezione di un gruppo di produttori (i miner) e lo scheduling della produzione. In principio un numero limitato di nodi viene candidato dal protocollo per aggiungere un nuovo blocco alla blockchain in base allo stake e al comportamento tenuto nel corso del tempo. La votazione viene eseguita dagli utenti del sistema che scelgono attraverso un voto, nella quale il voto viene ponderato in base allo stake posseduto dall'utente votante (maggiore è l'ammontare di denaro posseduto dall'utente votante, maggiore sarà il peso del suo voto nella scelta del miner).

Elencando brevemente gli svantaggi e vantaggi dei tre procolli, vale che il grande difetto della PoS è la libertà da parte degli utenti di poter validare, in caso di fork, più blocchi in due o più catene. La fork è un cambiamento effettuato nel protocollo, che può sfociare nel caso di un

cambiamento non retro-compatibile al protocollo precedente nella creazione di due diverse catene. I validatori di blocchi per aumentare il loro compenso guadagnato dal loro ruolo di miner, tentano di convalidare tutti i blocchi possibili poiché questo non comporta alcuna perdita nelle loro finanze. DPoS tenta di risolvere questo problema, perché lo stake è in parte spostato sul voto dato ai produttori, in quanto gli utenti del sistema possono investire un quantitativo di coin a loro scelta per assegnare il compito di validatore. Inoltre, un comportamento malevolo da parte di un miner che tenta di alterare le informazioni, porterà ad essere sfiduciato nelle votazioni successive. Un ulteriore vantaggio scaturito dall'utilizzo di una DPos rispetto a una PoW o una PoS è la democrazia creata attraverso la fase di votazione, permettendo a chiunque di ricoprire il ruolo di miner, eliminando il vincolo di dover possedere un grosso ammontare di "stake" oppure una macchina dalle prestazioni hardware notevoli. In un generico algoritmo di consenso, il blocco una volta creato nella fase finale del protocollo e aggiunto alla struttura dati non può essere più alterato in maniera retroattiva, senza alterare tutti i blocchi successivi. Questo è possibile tramite l'adozione di una funzione hash, funzione che mappa dati di dimensione arbitraria in dati di dimensione fissata. Quando un nodo esegue una nuova transazione, deve firmarla attraverso un protocollo di firma digitale e segnalare a tutti gli altri nodi della piattaforma che una nuova transazione deve essere aggiunta a un blocco. Ogni blocco, quindi, è processato in modo che la funzione generi un hash value, una stringa alfanumerica di lunghezza fissata. La funzione hash solitamente utilizzata all'interno delle applicazioni di alta sicurezza è la SHA256, che genera un'immagine di 256 bit. Quando viene creato un nuovo blocco i dati appartenenti al blocco stesso vengono elaborati attraverso SHA256, incorporando l'hash dei dati del blocco precedente. Così facendo, ogni blocco è collegato al blocco precedente e, se dei dati contenuti nella blockchain venissero manipolati o cancellati, i blocchi successivi al blocco corrotto rifiuterebbero i tentativi di modifica [22].

2.2 Blockchain Online Social Media

I social network distribuiti nascono per la possibilità di proteggere eventuali dati sensibili appartenenti agli utenti iscritti nella rete, così da eliminare la controindicazione per cui il proprietario del sistema legge e manipola i dati degli utenti senza alcuna limitazione. Con l'utilizzo della tecnologia delle blockchain, sono stati formulati nuovi casi d'uso nell'ambito dei social network distribuiti. Uno di essi è la possibilità di poter mettere in risalto solamente le informazioni veritiere all'interno del sistema tramite un meccanismo automatico, realizzato attraverso i comportamenti degli utenti, riguardante il poter votare in positivo o in negativo i post degli utenti. Inoltre, per spronare gli utenti a produrre materiale di qualità, vengono incentivati tramite delle ricompense in criptomoneta, che variano in base alla qualità del post creato (determinato per mezzo degli up-vote di altri utenti). Un sistema economico così costruito è basato sull'impiego dei principi della token economy [12] e della attention economy [8]. L'attention economy è un tipo di economia che fonda il suo sviluppo sull'attenzione degli utenti, in cambio di servizi offerti ai consumatori; mentre la token economy è una tecnica psicologica secondo cui, sfruttando determinati incentivi come i token, il soggetto è portato a comportarsi come desiderato. Oltre ai vantaggi, rispetto ai tradizionali social network centralizzati, esistono anche degli svantaggi non banali. Uno di questi è la non possibilità di riconoscere l'identità di un determinato utente, sia perché è possibile avere più di un account associato alla stessa persona sia perché quell'account potrebbe essere un bot programmato per eseguire determinate azioni all'interno della società. Un bot è un particolare tipo di utente iscritto al sistema online, con la sostanziale differenza di essere un programma in esecuzione a fronte di svolgere particolari compiti automaticamente confondendosi all'interno della rete sociale con gli altri umani. Un ulteriore svantaggio è determinato dalla struttura delle blockchain, che per come sono costruite permettono di registrare una decina di operazioni al secondo. La scalabilità è un problema che affligge anche i sistemi centralizzati, ma può essere attenuato dall'impiego di numerosi server sparsi nel mondo che permettono di effettuare parallelamente molte più operazioni. Tra gli obiettivi comuni degli sviluppatori di social media blockchain-based, oltre al rinnovare la tecnologia in modo da aumentare la scalabilità tentano di rinnovare la gestione dei dati, in modo da renderla ancora più trasparente e dando la possibilità di verificare l'autenticità di

ogni contenuto. Descrivendo le caratteristiche delle BOSM (Blockchain Online Social Media) risultano dei particolari da tenere in considerazione nell'utilizzo pratico. Infatti, affinché gli utenti siano premiati legittimamente, i contenuti pubblicati devono essere visibili a tutta la community e non c'è la possibilità di limitare la visione a un gruppo di utenti. Questo non garantisce una totale libertà di utilizzo da parte dei membri.

Attualmente le applicazioni sociali basate su blockchain sono più improntate a essere social media rispetto che social networks. Una prima differenza sostanziale è che nei social media l'attenzione è dedicata maggiormente sul materiale pubblicato, mentre nei social network l'attenzione si sposta sulle relazioni interpersonali che si vengono a creare. La seconda differenza è che nei social network è possibile decidere quali informazioni condividere e con che livello di granularità sociale, questa scelta nei social media non è permessa [22].

2.3 Steemit

Steemit è un social media che si basa sulla blockchain di nome Steem. Dal punto di vista implementativo, Steemit ha molte caratteristiche comuni alla piattaforma di social news Reddit [1], molto conosciuta negli Stati Uniti. Su Steemit è possibile registrarsi in maniera gratuita o a pagamento e una volta iscritto ogni utente può "postare" foto, video oppure contenuti testuali. Ogni post pubblicato dagli utenti può essere commentato oppure può essere votato in positivo (in inglese, upvote) o in negativo (in inglese, downvote). Anche i commenti possono essere votati dall'utente. Infine, esattamente come su Reddit, è possibile seguire un utente (equivale ad aggiungere alla lista amici). Steemit utilizza tre tipi di valute:

- STEEM – Rappresentano la vera e propria cripto-moneta. È possibile acquistarli, scambiarli con altre criptovalute o convertirli in Steem Dollars e/o Steem Power;
- STEEM DOLLARS (SBD) – Rappresentano la liquidità all'interno di Steemit e sono costruiti in modo tale che 1 SBD corrisponderà a 1 USD. Tale valuta può essere guadagnata dalla creazione di un post o dal proprio voto, possono essere scambiati con altri utenti

[1] https://www.redditinc.com

oppure possono essere convertiti in STEEM in un processo che varia dai tre ai cinque giorni.

- STEEM POWER (SP) – Rappresentano la propria reputazione all'interno della piattaforma. La quantità di SP influisce in maniera proporzionale al voto, più SP si posseggono, più il voto sarà influente. Tramite un processo chiamato power up è possibile convertire gli STEEM in SP, viceversa è possibile convertire gli SP in STEEM tramite un processo chiamato power down che richiede, in confronto al power up, molto tempo per essere portato a termine. Gli SP, in termini finanziari, vengono considerati come una sorta di investimento a lungo termine; i possessori di SP, infatti, vengono ricompensati con il 15% dell'inflazione annuale, secondo una formula che assegna un numero di SP direttamente proporzionale all'ammontare di SP posseduti rispetto al totale posseduto da tutti gli utenti.

Chi crea e chi vota contenuti vengono ricompensati con una certa somma di SBD e SP, opportunamente suddivise tra tutti. Ogni post, in base al numero di voti e quanta influenza hanno gli utenti che li hanno espressi, accumula una certa quantità di denaro che viene divisa dopo 7 giorni dalla pubblicazione del post. Gli utenti possono essere categorizzati in due gruppi:

- Autore, ovvero colui che pubblica contenuti;
- Curatore, ovvero colui che premia i contenuti attraverso voti e commenti.

L'*upvote* di un post o di un commento è un segno di apprezzamento e indica la volontà di voler premiare l'autore con una ricompensa. Gli utenti possono valutare un numero limitato di post al giorno, dovuto all'introduzione di un parametro importante chiamato *voting mana*. Quest'ultimo è legato a un ulteriore parametro chiamato *voting weight* utilizzato per assegnare, in base agli SP posseduti, all'atto di votazione un "peso" al proprio voto nel post. Maggiore è il voting weight assegnato a un voto, maggiore sarà il consumo di voting mana e conseguentemente minore la possibilità di assegnare ulteriori voti ad altri post o commenti. Il bacino delle ricompense accumulato allo scadere dei 7 giorni viene suddiviso con il 75% del totale all'autore del post e il 25% ai curatori del post, divisi in valuta SBD e SP [22].

2.4 Steem e i principi di funzionamento della blockchain

Steem è una blockchain che supporta la costruzione di una comunità e l'interazione sociale attraverso ricompense in criptomoneta. Secondo quanto dichiarato nel paper [20], si tratta della prima criptomoneta che tenta, in maniera accurata e chiara, di ricompensare il contributo soggettivo dato alla community da un certo gruppo di individui. Gli incentivi economici permessi da tale token possono contemporaneamente facilitare la crescita di una nuova piattaforma social media e nel frattempo avviare una nuova economia di successo. Gli sviluppatori della blockchain sono alla ricerca di un algoritmo in grado di dare un punteggio ai contributi degli utenti, che la maggioranza dei membri della community consideri equa.

Steem è una blockchain basata su Graphene, la stessa tecnologia che alimenta BitShares. Essa produce nuovi token ogni volta che un blocco è stato prodotto. L'algoritmo di consenso utilizzato da Steem è il DPoS e a conseguenza di ciò nella blockchain Steem il compito di miner viene delegato a un numero fissato di utenti fidati e conoscitori della piattaforma, chiamati Witness. Tali soggetti sono pagati dal sistema per creare in totale sicurezza i blocchi che compongono la blockchain. Essenzialmente, un Witness rappresenta un nodo della rete Steem affidabile, in quanto eletto attraverso i voti della Steemit Community, poiché ritenuto in grado di eseguire diversi compiti come:

- **Price feed** – Steem si basa sul concetto di libero mercato, di conseguenza il valore degli STEEM viene deciso dai Witness, che hanno anche il compito di produrre i price feed. I price feed rappresentano quanto vale uno STEEM in Dollari Statunitensi (USD). Ognuno dei Witness pubblica il proprio price feed e il price feed effettivo è calcolato come la mediana dei top 20 Witness che sono stati scelti in un periodo tra i tre e i cinque giorni;
- **Consenso** - Essendo un gruppo di persone elette e fidate, i Witness attivi hanno il compito di esprimere il proprio consenso sull'applicazione degli effetti delle hardfork. Un hardfork è un cambio nella logica dell'attività che alimenta la rete Steem. Gli sviluppatori del sistema permettono agli utenti di poter scegliere se sia il caso di aggiornare il codice all'ultima versione e quindi seguire un protocollo aggiornato oppure rimanere nella versione precedente con una scissione della catena di blocchi in due catene totalmente distinte;

- **Produzione di blocchi ogni 3 secondi** – Un altro importante task che rende le blockchain stabile e aggiornata agli occhi dei mercati esterni. Dato che il sistema conosce in anticipo i Witness attivi in quel round, esso è in grado di schedulare i Witness per produrre blocchi ogni tre secondi. Al tempo stesso, questi ultimi sincronizzano la loro produzione di blocchi grazie al protocollo NTP (Network Time Protocol). Ovviamente, nel fare questo lavoro i Witness vengono ricompensati con uno STEEM per ogni blocco prodotto.

Gli utenti della piattaforma Steemit eleggono i Witness attraverso una votazione. Ognuno ha 30 voti giornalieri e i voti sono pesati sullo stake in base al quantitativo di SP investito dall'utente. In questo modo, i casi di Witness maliziosi sono poco probabili, per via dell'algoritmo di consenso DPoS illustrato precedentemente, che, a lungo andare, fa sì che gli utenti siano a conoscenza di quali Witness siano affidabili e onesti. Inoltre, ciascun Witness è pagato molto bene per svolgere il suo compito e comportarsi in maniera maliziosa farebbe perdere l'incarico nelle votazioni future [22].

La blockchain Steem, come detto in precedenza, contiene nei suoi blocchi una lista di interazioni avvenute nella piattaforma Steemit. Un'interazione può essere vista come un'azione effettuata da un certo utente nei confronti di sé medesimo o di un secondo utente. In totale su Steemit sono presenti 38 transazioni differenti, raggruppate secondo queste categorie:

- Transazioni per la creazione di un account;
- Transazioni di gestione dell'account;
- Transazioni riguardanti le criptovalute di Steem;
- Transazioni di tipo sociale;
- Transazioni riguardanti i Witness.

In Steem è possibile creare un account, a pagamento o meno, attraverso due transazioni facente parti della prima categoria: tramite account_create si ottiene un nuovo account gratuito oppure pagando; invece, tramite account_create_with_delegation si ottiene un nuovo account pagando una somma che verrà risarcita in SP.

L'operazione account_create_with_delegation è stata successivamente sostituita dall'implementazione di due nuove transazioni: claim_account, attraverso cui un utente converte i propri Resource Credits (crediti non trasferibili introdotti con l'implementazione di queste nuove transazioni) in un token, chiamato ATC, utile specificatamente per la creazione di un nuovo account; create_claimed_account, con cui l'utente, una volta ottenuto l'ATC, crea effettivamente l'account. Alla seconda categoria appartengono le transazioni relative alla gestione delle funzionalità del proprio account. Questa categoria viene divisa in base al tipo di funzionalità trattata, in particolare:

- La gestione dei propri SBD e STEEM avviene attraverso le transazione transfer_to_savings, transfer_from_savings e cancel_transfer_from_savings con cui vengono gestiti o annullati i trasferimenti da o per il saving balance;
- Nel caso in cui un utente non riesca ad accedere al proprio account, il recupero deve essere effettuato da un Witness, tramite le transazioni: request_account_recovery con cui inoltra la richiesta di recupero e recover_account con cui avviene l'effettivo recupero. Il Witness, da parte sua, recupera l'account richiesto tramite la transazione change_recovery_account;
- L'utente iscritto può attuare delle modifiche nel proprio account, attuabile con la transazione account_update;
- Un utente può rinunciare al proprio diritto di voto tramite la transazione decline_voting_rights.

La gestione delle funzionalità appartenenti all'ambito economico, terza categoria, in Steem avvengono attraverso molteplici transazioni. Nello specifico:

- È possibile impostare una soglia di prezzo degli SBD oltre la quale gli utenti decidono di venderli in cambio di USD. Le transazioni che gestiscono questa feature sono: limit_order_create, limit_order_create2 che impostano il lower bound (soglia minima) oltre il quale vendere gli SBD e limit_order_cancel con cui viene cancellata una soglia impostata precedentemente;

- È possibile gestire delle operazioni di escrow, usate per sancire un accordo in STEEM tra due utenti, con il presidio di un intermediario. Le transazioni che permettono questa funzionalità sono: escrow_transfer che permette di iniziare un accordo, escrow_approve che sancisce l'avvenuto accordo tra le due parti da parte dell'intermediario, escrow_dispute utilizzato per rifiutare l'accordo e affidare la decisione all'intermediario ed escrow_release con cui annullare definitivamente l'accordo.

- La gestione delle operazioni di conversione delle criptovalute di Steem è offerta dalle transazioni: convert con cui convertire STEEM in SBD, tansfer_to_vesting con cui effettuare un power up, withdraw_vesting con cui effettuare un power down;

- È possibile delegare una somma di SP posseduti a un altro account tramite la transazione delegate_vesting_shares;

- Tramite la transazione transfer possono essere trasferiti STEEM ad un account beneficiario;

- In ultimo, è possibile effettuare una conversione power down, specificando l'account a cui trasferire gli STEEM prodotti, con la transazione set_withdraw_vesting_route.

La quarta categoria comprende tutte le transazioni di tipo sociale permesse dal sistema: azione di vote, comment, custom_json, comment_options, claim_reward_balance, delete_comment. Tralasciando le ovvie, l'azione custom_json permette di seguire un utente, effettuare la condivisione di un post, oppure richiedere la modifica di un post a un Witness; l'interazione comment_options permette di modificare alcune impostazioni di un post e in ultimo il claim_reward_balance permette di aggiungere effettivamente le ricompense ricevute al bilancio dell'utente.

L'ultima categoria comprende le azioni legate al contesto dei Witness. Elenchiamo, innanzitutto, le azioni che possono essere eseguite da qualsiasi utente, cioè: account_witness_vote con cui esprimere il voto a favore di un certo Witness; witness_set_properties e witness_update con cui effettuare la richiesta di diventare un Witness; account_witness_proxy con cui votare automaticamente gli stessi Witness. E infine, citiamo l'azione eseguibile esclusivamente da chi ha il ruolo di Witness, ovvero il feed_publish con cui quest'ultimi pubblicano i price feed.

Capitolo 3

Elementi di statistica

In questo capitolo presentiamo i metodi ricorrenti nella ricerca statistica e sfruttati nel nostro lavoro.

La scienza statistica è nata per poter studiare quantitativamente e qualitativamente attraverso metodi empirici le proprietà di un particolare fenomeno e le sue correlazioni. Questa tipo di scienza è considerata, appunto, una scienza empirica che si distingue rispetto a una scienza deduttiva (o non empirica) come la logica e la matematica. La sostanziale differenza sta dall'utilizzo di asserzioni iniziali (detti assiomi) ritenuti veritieri il quale attraverso questi vengono dedotti teoremi, lemmi, corollari ecc. Viceversa, le scienze empiriche impiegano come metodo di indagine non solo la deduzione ma anche l'induzione [28]. Tale scienza è comunemente suddivisa in due categorie:

- **Statistica descrittiva**: studia i criteri di rilevazione, classificazione, sintesi e rappresentazione dei dati appresi dallo studio di una popolazione o di un campione (parte di una popolazione);
- **Statistica inferenziale**: estende i risultati ottenuti su un campione, all'intera popolazione.

L'impiego di metodi grafici per la descrizione delle distribuzioni e i confronti tra queste, risultano degli strumenti fondamentali nella ricerca scientifica per la caratterizzazione dei fenomeni e delle loro correlazioni. Le due sottosezioni seguenti sono divise in base a queste due branche della statistica.

3.1 Statistica descrittiva

Nella statistica descrittiva sono definiti gli strumenti utili a sintetizzare e classificare le diverse distribuzioni relative a particolari fenomeni. I dati di un'indagine sperimentale possono essere visualizzati attraverso molteplici rappresentazioni grafiche e risultati analitici. Le sezioni successive comprendono una descrizione di strumenti e metodi che abbiamo impiegato nel corso della ricerca. Tra questi vi è la presenza di concetti appartenenti alla teoria delle probabilità, questa branca della statistica in alcuni testi è chiamata **statistica matematica** cioè quella parte della statistica che presenta le distribuzioni teoriche allo scopo di illustrarne le caratteristiche fondamentali, le relazioni che esistono tra esse e gli usi possibili. Per semplicità, introduciamo queste nozioni nel capitolo di statistica descrittiva.

3.1.1 Istogramma e Stem-and-Leaf

Per dati quantitativi, riferiti a variabili continue misurate su scale a intervalli, di norma si ricorre a istogrammi oppure tramite un differente tipo di rappresentazione chiamato stem-and-leaf. Gli istogrammi sono grafici a barre verticali nei quali:

- Le misure della variabile casuale sono riportato lungo l'asse orizzontale;
- Mentre l'asse verticale rappresenta il numero assoluto, oppure la frequenza relativa o quella percentuale, con cui compaiono i valori di ogni classe.

I lati dei rettangoli sono costruiti in corrispondenza degli estremi di ciascuna classe. Un istogramma deve essere inteso come una rappresentazione areale: sono, infatti, le superfici dei vari rettangoli che devono essere proporzionali alle frequenze corrispondenti. Quando le classi hanno la stessa ampiezza, le basi dei rettangoli sono uguali e di conseguenza, le loro altezze risultano proporzionali alle frequenze che rappresentano. Nel caso in cui le basi sono uguali, è indifferente ragionare in termini di altezze o di aree di ogni rettangolo. Nel particolare caso, meno comune, in cui le ampiezze delle classi sono diverse, bisogna ricordare il concetto generale che le frequenze sono rappresentate dalle superfici e quindi è necessario rendere l'altezza proporzionale.

Descrivendo, invece, il diagramma a ramo e foglia (stem-and-leaf plot), esso è un tipo di tecnica semi-grafica che può essere descritta come un incrocio tra un istogramma e una tabella di frequenza. Il metodo è utile per una prima descrizione di una distribuzione di dati. I principi di costruzione sono semplici:

- Ogni numero è diviso in due parti, il ramo (stem) e la foglia (leaf);
- Il ramo è il numero, collocato a sinistra, che include tutte le cifre, o almeno le cifre significative alla corretta comprensione, eccetto l'ultima;
- La foglia, collocata a destra, è sempre un numero con una cifra sola (single digit), che può essere esclusivamente l'ultima di tutto il numero.

Anche questo grafico insieme all'istogramma ha lo scopo di mostrare le caratteristiche fondamentali di una distribuzione di dati:

- Valore minimo e massimo e quindi l'intervallo di variazione;
- I valori più frequenti;
- La presenza di uno o più picchi;
- La forma della distribuzione, in relazione soprattutto alla simmetria;
- La presenza di outlier o valori anomali cioè quelli troppo distanti dal gruppo principale di valori.

3.1.2 Distribuzione di probabilità

In ordine per poter illustrare nello specifico le distribuzioni che interessano il fenomeno da noi analizzato, partiamo nel definire uno strumento matematico di base chiamata **variabile aleatoria**. In matematica, e in particolare nella teoria della probabilità, una variabile casuale (detta anche variabile aleatoria o variabile stocastica) è una variabile che può assumere valori diversi in dipendenza da qualche fenomeno aleatorio. I possibili esiti di un esperimento e le loro probabilità possono essere sintetizzati in distribuzioni di probabilità. Dal momento che una variabile casuale si riferisce all'esito di un fenomeno casuale, a ogni possibile esito resta

associata una probabilità di verificarsi. Le distribuzioni di probabilità possono essere continue o discrete, in base al tipo di valore che la variabile aleatoria può assumere sull'insieme degli eventi. Poiché a noi interessano le variabili casuali continue, le definiamo come: una variabile casuale X è detta continua se può assumere qualunque valore reale in un determinato intervallo. Una distribuzione di probabilità per variabili continue può essere graficamente rappresentata con un istogramma. Una variabile casuale continua assume valori che costituiscono un intervallo. La sua distribuzione di probabilità è rappresentata da una curva che consente di determinare la probabilità che la variabile casuale assuma valori in un prestabilito intervallo. Ciascun intervallo ha probabilità di verificarsi comprese tra 0 e 1. L'intervallo che contiene tutti i possibili valori della variabile casuale ha probabilità pari a 1, poiché la variabile deve assumere almeno un valore tra quelli definiti nello spazio degli eventi.

Per una variabile continua non ha senso calcolare la probabilità per una specifica realizzazione della variabile, questa probabilità sarà necessariamente 0; invece è possibile calcolare le probabilità di intervalli. Nello specifico della materia introduciamo una funzione che definisce la probabilità che una variabile X non superi un certo valore x e prende il nome di funzione di ripartizione e rappresenta le probabilità cumulate, chiamata anche Cumulative Distribution Function (CDF):

$$F(x) = P(X \leq x).$$

Esistono diverse distribuzioni di probabilità generate da fenomeni diversi in natura che hanno come comun denominatore l'andamento della loro distribuzione dei dati. Nel nostro caso descriviamo le distribuzioni correlati a fenomeni che seguono la legge di potenza (power law), la distribuzione log-normale, e in modo da comprendere quest'ultima, la distribuzione normale.

Distribuzione normale

La distribuzione normale (o distribuzione di Gauss dal nome del matematico tedesco Carl Friedrich Gauss), nella teoria della probabilità, è una distribuzione di probabilità continua che è spesso usata come prima approssimazione per descrivere variabili casuali a valori reali che tendono a concentrarsi attorno a un singolo valore medio. Tenendo conto di questa sua particolare caratteristica, vale che il grafico della funzione di densità di probabilità associata è

simmetrico e ha una forma a campana, nota come campana di Gauss. La funzione matematica della curva normale è definita in maniera univoca da due soli parametri: il valore medio e lo scarto quadratico medio della distribuzione stessa. La funzione di densità è calcolata con la relazione seguente:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Come si può notare dalla formula scritta sopra, la funzione $f(x)$ descrive, al variare dei valori assunti dei due parametri μ e σ, una famiglia di curve normali:

- se si varia μ: si sposta orizzontalmente l'asse di simmetria della curva;
- se si varia σ: la curva si allarga e appiattisce al crescere del valore di σ

Per indicare che la variabile X si distribuisce come una normale, è di usanza scrivere $X \sim N(\mu, \sigma)$.

Molte distribuzioni che si incontrano nel mondo reale seguono effettivamente la distribuzione normale, ovvero sono ben approssimate dalla funzione vista prima. Elenchiamo dei possibili esempi di fenomeni naturali descrivibili tramite una normale, in modo da avere un'idea più intuitiva:

- La funzione di densità riguardante la distribuzione dell'altezza in una popolazione di umani si mostra come una normale;
- La funzione di densità riguardante la distribuzione del peso in una popolazione di umani si comporta come una normale;
- i valori proodotti da un processo di misura i quali producono lo stesso valore a meno di un errore di misura è stato dimostrato che si comporta secondo una normale.

Molte distribuzioni, di per sé anche se non presentano una forma normale magari anche totalmente asimmetriche (purché unimodali), possono essere normalizzate mediante una trasformazione di variabile utilizzando una funzione matematica invertibile (es. $y = log(x)$).

In ultimo mostriamo un esempio di grafico avente una distribuzione normale (figura 3.1) e la sua funzione di ripartizione per diversi valori di μ e σ (figura 3.2) .

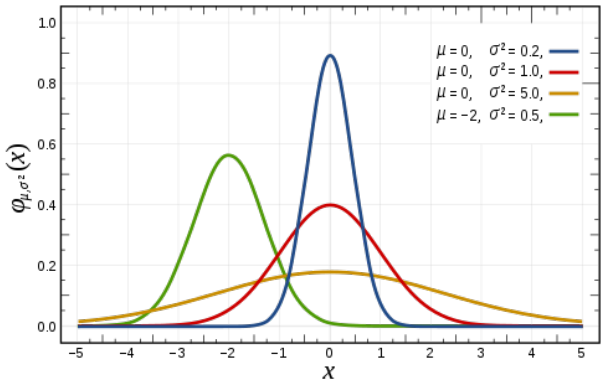


Figura 3.1: Distribuzione simmetrica (o normale)

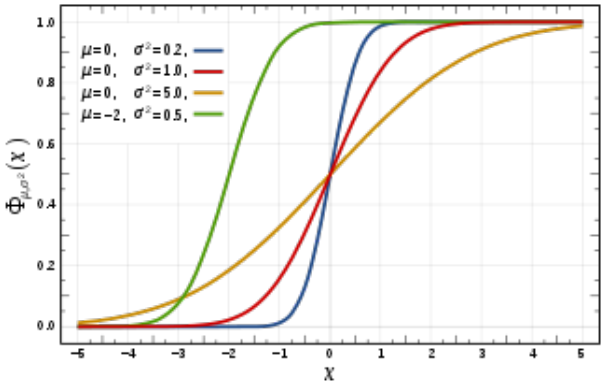


Figura 3.2: Funzione di ripartizione per una distribuzione normale con diversi valori della media e dello scarto quadratico medio.

Distribuzione log-normale

Nel particolare caso in cui i dati raccolti sul fenomeno presentano una distribuzione differente dalla normale, spesso una semplice trasformazione dei valori utilizzando una funzione matematica di base, invertibile, conduce a una distribuzione normale. È il caso della trasformazione con i logaritmi. Nelle circostanze in cui la variabile X tende ad avere una forma circa asimmetrica a destra, con la trasformazione logaritmica la variabile $log(X)$ assume, appunto, una forma molto simile alla distribuzione di Gauss. Questa distribuzione può approssimare il prodotto di molte variabili aleatorie positive indipendenti. La variabile aleatoria $X = e^N$ segue la distribuzione log-normale $logX(\mu, \sigma^2)$ se e solo se $N = log(X)$ segue la distribuzione normale.

Per completezza la sua funzione di densità di probabilità è

$$f(x) = \frac{e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}}{x\sqrt{2\pi}\sigma} \text{ per } x > 0.$$

Semplificando la spiegazione della distribuzione, dal calcolo di tutti i suoi parametri, possiamo mostrare immediatamente la funzione di densità di probabilità e la sua funzione di ripartizione (rispettivamente, figure 3.3 e 3.4).

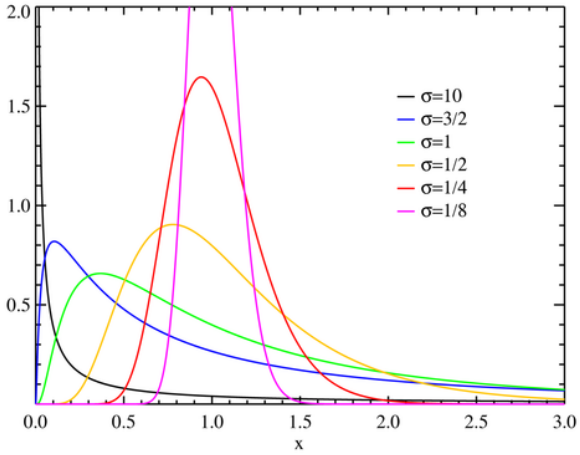


Figura 3.3: Funzione di densità della log-normale al variare del parametro σ.

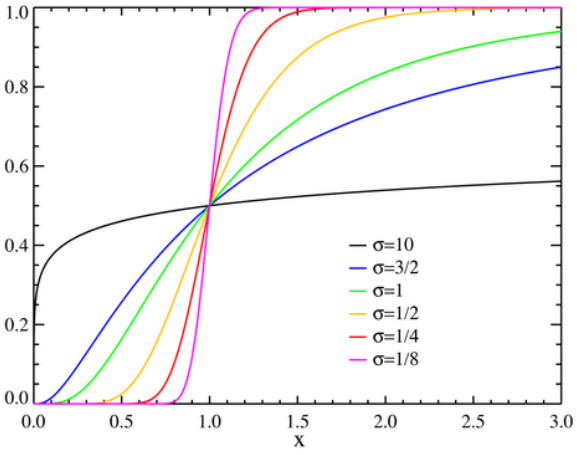


Figura 3.4: Funzione di ripartizione della log-normale.

Distribuzione Power law

Una distribuzione viene detta legge di potenza (power-law) se la probabilità che una certa variabile casuale assuma un certo valore x a una potenza $-\alpha$ di x con α valore costante (in generale α compreso tra 2 e 3).

Di fondamentale importanza introdurre la storia che sta dietro alla legge di potenza e quali esperimenti nell'ambito reale hanno portato all'affermarsi della legge di probabilità come la conosciamo oggi. L'economista Vilfredo Pareto battezzò la nota legge di potenza come distribuzione di Pareto. L'economista analizzando la distribuzione del reddito della popolazione del Canton Ticino, osservò che tale distribuzione di probabilità era caratterizzata da un decadimento polinomiale [21]. Il decadimento polinomiale, e quindi la legge di potenza, può essere riassunto tramite la seguente formula matematica:

$$P(x) = Cx^{-\alpha} \text{ con C costante e } \alpha > 1.$$

Una power law viene rappresentata come una retta se si utilizza una scala logaritmica per entrambi gli assi (log-log graph); conseguenza derivata dal fatto che se modelliamo l'espressione della power law:

$$y = k \cdot x^{\alpha}$$

se si passa ai logaritmi si ottiene

$$log(y) = \alpha log(x) + log(k)$$

applicando la trasformazione $X = log(x), Y = log(y)$ per $m = \alpha$ e $q = log(k)$ possiamo ricondurci alla retta

$$Y = mX + q$$

dove m è il coefficiente angolare della retta e q è l'interdetta.

Le proprietà fondamentali che caratterizzano le distribuzioni power law sono:

- grafico con lunga coda a destra (right skew, heavy tailed): la maggior parte delle persone ha poca ricchezza, ma esistono anche persone con molta ricchezza che in proporzione sono poche e tutte con una ricchezza differente;
- alto rapporto nella distribuzione tra il valore massimo e il valore minimo.

Un modo di dire generato da questa legge è la famosa regola 20-80 valida per tutti i sistemi che seguono una distribuzione power law:

- 20% dei siti Web ricevono l'80% delle visite;
- 20% dei routers Internet gestisce l'80% del traffico;
- 20% delle industrie mondiali possiede l'80% dei ricavi;
- 20% della popolazione mondiale consuma l'80% delle risorse.

Per rappresentare e studiare una power law, metodo utile, consiste nel rappresentarla mediante la distribuzione di probabilità cumulativa. Se indichiamo con $P(x)$ la probabilità che X (x maiuscolo) abbia un valore maggiore o uguale a x (x minuscolo)

$$P(x) = \int_x^\infty p(x')dx'$$

poiché la $p(x)$ è una power law di tipo

$$p(x) = Cx^{-\alpha}$$

possiamo scrivere

$$P(x) = C\int_x^\infty x'^{-\alpha}dx' = \frac{C}{\alpha - 1}x^{-(\alpha-1)}$$

quindi la CDF $P(x)$ di una power law è a sua volta, una power law ma con esponente $\alpha - 1$ e come costante $\frac{C}{\alpha - 1}$. Quindi, riassumendo i concetti, una distribuzione che segue una legge di potenza può essere analizzata tramite la CDF e nel caso in cui il grafico risultasse poco chiaro, è possibile utilizzare una scala log-log; in questo modo, in ambo i casi, la distribuzione seguirà un andamento rettilineo più comprensibile, ma con una pendenza diversa. Per spiegare il concetto della scala log-log consideriamo l'asse cartesiano x e una serie di valori $x_1, x_2, .., x_n$ dove ogni valore x_i è rappresentato sull'asse a una distanza dall'origine pari a $\log_{10}(x)$. Risulta che per valori crescenti dell'argomento la curva logaritmica cresce sempre più lentamente e sua volta la distanza di un valore dall'origine aumenta sempre più lentamente (la distanza dei valori intermedi tra una potenza di 10 e la successiva è sempre più smorzata).

Anche in questo caso mostriamo le distribuzioni tramite un istogramma e la distribuzione cumulata in scala log-log (immagini 3.5 e 3.6, rispettivamente).

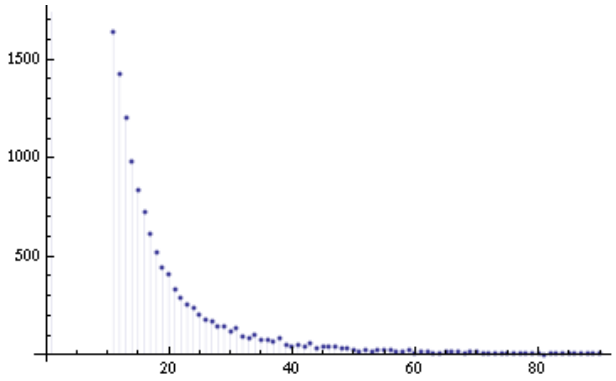


Figura 3.5: Istogramma di densità di una distribuzione power-law.

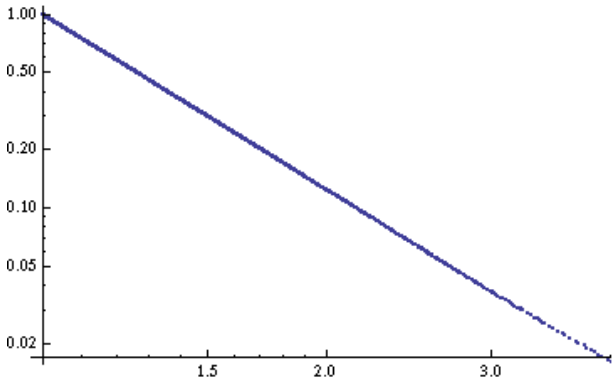


Figura 3.6: Funzione di ripartizione di una power law in scala log-log.

3.1.3 Heatmap

La mappa di calore (detta in inglese heatmap) è una rappresentazione grafica dei dati dove i singoli valori contenuti in una matrice sono rappresentati da colori. Essendo la mappa una matrice avremo dei valori nelle colonne e dei valori nelle righe nella quale: ogni cella rappresenta il valore che assumono queste variabili in base al tipo di correlazione studiato tra le due variabili.

Fondamentalmente ci sono due tipi di mappe di calore: il cluster heatmap e l'heatmap spaziale. La mappa rappresenta esattamente una matrice di dimensioni di cella fisse le cui righe e colonne sono fenomeni distinti, chiaramente le dimensioni delle celle sono arbitrarie ma abbastanza grandi da essere visibili. Al contrario, la posizione in una mappa del secondo tipo è forzata dalla grandezza del fenomeno analizzato e non esiste la nozione di cella; in questi particolari casi si ritiene che il fenomeno vari continuamente.

3.1.4 Famiglie di indici

Viene fornito una lista di indici comunemente calcolati in un'indagine statistica utili ad avere una visione numerica dei dati raccolti. Principalmente citiamo le tre caratteristiche fondamentali di un insieme di dati statistico:

- Media aritmetica - Indice matematico calcolato attraverso il rapporto tra la somma dei dati numeri e la cardinalità dei dati della popolazione o del campione;
- Moda - Indice statistico riguardante il valore che si presenta con maggior frequenza nei dati raccolti;
- Mediana - Appartenente alla famiglia dei quantili (indici che dividono in due parti, in base al quantile, il campione) indica il valore centrale tra i dati numerici in una distribuzione ordinata.

Oltre agli indici di "base" precedenti, introduciamo ulteriori indici di posizione che rispecchiano altre caratteristiche di un campione o popolazione in esame. Gli indici di cui vale la pena prendere nota sono:

- Deviazione standard (σ) – Lo scarto quadratico medio, o appunto deviazione standard, è un indice di dispersione statistico, vale a dire una stima della variabilità di una popolazione di dati. È uno dei modi per esprimere la dispersione dei dati intorno a un indice di posizione, come la media aritmetica o una sua stima. Dato che è una misura effettuata con l'utilizzo dei dati espressi in una certa misura, tale indice avrà la medesima misura. L'espressione per il calcolo della deviazione standard è:

$$\sigma_X = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \overline{x})^2}{N}}$$

dove $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ è la media artimetica della variabile casuale X

N è la cardinalità della popolazione X.

- Indice di asimmetria – Nella famiglia degli indici di simmetria sono presenti diverse formule matematiche per il calcolo dell'asimmetria di una distribuzione. Un indice di

asimmetria di una distribuzione è un valore che tenta di fornire numericamente una misura della sua mancanza di simmetria. Una possibile asimmetria nella distribuzione dei dati indica una densità maggiore nei valori più bassi o nei valori più alti rispetto a una misura statistica nota, come potrebbe essere la mediana o la media. L'aspetto comune a tutti gli indici appartenenti a questa famiglia è che il valore 0 ritornato fornisce una condizione necessaria, ma non sufficiente, affinché una distribuzione sia simmetrica. Gli indici di asimmetria comunemente si basano su alcune proprietà delle distribuzioni simmetriche, o in particolare, della distribuzione normale. Definiamo l'indice skewness (indicato con γ_3):

$$\gamma_3 = \frac{1}{N}\sum_{i=1}^{N}(\frac{x_i - \overline{x}}{\sigma_X})^3$$

e l'indice di curtosi (indicato con γ_4):

$$\gamma_4 = \frac{1}{N}\sum_{i=1}^{N}(\frac{x_i - \overline{x}}{\sigma_X})^4$$

nel quale l'ultimo ha come obiettivo la visualizzazione della forma di una distribuzione, che costituisce una misura dello "spessore" delle code di una funzione di densità, ovvero il grado di "appiattimento" di una distribuzione, in confronto a una distribuzione normale;

- Coefficiente di variazione – Il coefficiente di variazione è un indicatore statistico di dispersione relativo indipendentemente dall'unità di misura, calcolato come rapporto tra la deviazione standard e la media della distribuzione:

$$CV_X = \frac{\sigma_x}{|\overline{x}|}$$

La dispersione relativa serve a indicare la variabilità di un fenomeno in termini percentuali. È usata per confrontare la variabilità dei fenomeni, senza prendere in considerazione l'unità di misura.

3.2 Statistica inferenziale

Nella statistica inferenziale vengono descritti i test utilizzati per affermare una determinata ipotesi. In statistica il test è un processo logico-matematico che porta alla conclusione di non poter respingere oppure di poter respingere l'ipotesi di casualità, mediante il calcolo di probabilità specifiche di commettere un errore con queste affermazioni. L'ipotesi che il risultato ottenuto con i dati sperimentali raccolti sia dovuto solo al caso è chiamato ipotesi nulla ed è indicata con H_0. Di norma, con essa si afferma che le differenze tra due o più gruppi, quelle tra un gruppo e il valore atteso oppure le tendenze riscontrate siano imputabili essenzialmente al caso; per giungere a queste conclusioni si deve ricorrere all'inferenza. La domanda inferenziale di verifica dell'ipotesi nulla a cui si deve rispondere tramite questi strumenti è:

> "Nell'ipotesi che le differenze fra gruppi di osservazioni empiriche siano dovute a fattori esclusivamente casuali, qual è la probabilità che fra tutte le alternative possibili si presenti proprio la situazione descritta dei dati raccolti o una ancora più estrema?"

Ogni test è associato a 4 parametri interdipendenti che esprimono il rischio che si corre, ovvero della sicurezza che si ha, nel formulare una conclusione:

- **Errore di primo tipo** (rischio α o livello di significatività a cui corrisponde il p-value): probabilità che esprime il rischio di rifiutare H_0 quando è vera (falso negativo);
- **Errore di secondo tipo** (rischio β): probabilità del rischio di accettare H_0 quando è falsa (falso positivo);
- **Protezione del test**$(1 - \alpha)$: probabilità di accettare H_0 quando è vera;
- **Protezione del test**$(1 - \beta)$: probabilità di rifiutare H_0 quando è falsa

Poiché i due parametri sono interdipendenti, nel caso in cui aumentasse α diminuirebbe β o viceversa, non è possibile ridurre entrambi i tipi di errore contemporaneamente. Il metodo che utilizziamo per verificare se l'ipotesi nulla va rifiutata o meno è quella di calcolare il p-value. Il valore p è la probabilità che il possibile rifiuto dell'ipotesi nulla sia solo dovuta al caso.

Comunemente, il livello di significatività (α) è fissato al valore 0.05 e alla fine del test si può incorrere in due casistiche:

- Caso p-value $> \alpha$ l'evidenza empirica non è sufficientemente contraria all'ipotesi H_0 che quindi NON può essere rifiutata;
- Caso p-value $< \alpha$ l'evidenza empirica è fortemente contraria all'ipotesi H_0 che quindi va rifiutata, accetto H_1.

I test a cui faremo riferimento vengono detti test non parametrici, un test non parametrico non presuppone alcun tipo di distribuzione, quindi non utilizzano i valori effettivi dei dati. In contrapposizione, un test statistico è parametrico quando nel controllo delle ipotesi vi è la presenza di un parametro come la media, la varianza o la deviazione standard calcolati tramite il valore effettivo delle misurazioni, quindi per poter prendere in considerazione l'utilizzo di questi ultimi bisogna rispettare il ristrettissimo vincolo di poter essere applicato solamente a una distribuzione gaussiana (normale) o comunque che la distribuzione si avvicini a una distribuzione simmetrica, con il vantaggio che i risultati riscontrati con questi test risultano attendibili rispetto a quelli non parametrici poiché si basano su grandezze matematiche calcolate sui reali valori dei dati. Descrivendo sommariamente i test non parametrici, molti hanno la caratteristica identica di trasformare i valori della distribuzione in ranghi. Il punto cardine del metodo non parametrico sta nel fatto che la distribuzione perde una quantità di informazione, passando da una scala di rapporti che trova il vantaggio di avere un'origine reale e la proprietà di poter rapportare una coppia di valori in una scala solamente ordinale (ranghi dei dati), con le sole proprietà di equivalenza e di relazioni di maggiore e minore tra i ranghi.

3.2.1 Test di correlazione per ranghi (Spearman e Kendall)

La correlazione è uno dei metodi statistici più antichi, diffuso già all'inizio del 900. La metodologia non parametrica proposta da Charles Spearman è una correlazione basata sui ranghi, che ricorre agli stessi concetti della correlazione parametrica r di Pearson [18]. La correlazione di Pearson è uno degli stimatori statistici più utilizzati. Ma il calcolo può essere influenzato da anche un unico valore anomalo e in aggiunta le distribuzioni richieste devono attenersi secondo una distribuzione normale in modo da avere un'alta probabilità di sicurezza nel considerare i possibili risultati prodotti dal test. In riferimento ai test di correlazione non parametrica, in letteratura sono ricorrenti i termini di correlazione tra ranghi e cograduazione utilizzati in modo appropriato solo con scale almeno di tipo ordinale, mentre associazione e concordanza con scale qualitative o categoriali. I test di correlazioni conosciuti oltre al test di Spearman è il test di Kendall, proposto in tempi successivi da Maurice Kendall. Si può dimostrare che i test appena citati mantengono un'efficienza statistica sopra al 70%, resistendo alla possibile presenza di valori anomali, fornendo un buon compromesso tra robustezza ed efficienza [7]. Soffermandoci sul concetto di correlazione tra ranghi, descriviamo due metodi concettualmente differenti ma che rispondono allo stesso tipo di ipotesi. Il primo metodo calcola il coefficiente di correlazione di Spearman, indicato con il simbolo greco ρ (rho). Questa quantità può variare tra +1 e -1 quando la correlazione è massima, con valore positivo oppure negativo, ed è vicino a zero quando non esiste correlazione. Il metodo richiede che entrambe le variabili siamo misurate su una scala almeno ordinale. Nel corso di questo paragrafo indicheremo il primo set di dati con X e il secondo set di dati con Y per semplicità.

Il coefficiente di correlazione per ranghi di Spearman serve per verificare l'ipotesi nulla (H_0) dell'indipendenza tra due variabili, nel senso che gli N valori della variabile Y hanno le stesse probabilità di associarsi con ognuno degli N valori di X. L'ipotesi alternativa di esistenza di una associazione può prevedere un risultato positivo oppure negativo, nel primo caso è detta associazione diretta: le coppie di valori sono contemporaneamente alti o bassi sia per X che per Y; nel secondo caso è chiamata associazione indiretta e a valori alti di X corrispondono valori bassi di Y o viceversa.

La metodologia di calcolo è la seguente:

1. Indichiamo l'ipotesi nulla H_0, in base a ciò che vogliamo scoprire. L'ipotesi può essere bilaterale o unilaterale:

   - Bilaterale, quando vogliamo che un valore sia diverso o uguale a una certa quantità; un esempio $H_0 : \rho = 0$;
   - Unilaterale in una direzione, quando vogliamo osservare se ρ sia maggiore o minore di un determinato valore; un esempio $H_0 : \rho >= 0$.

2. Successivamente, occorre ordinare i ranghi della variabile *X*, assegnando 1 al valore più piccolo e progressivamente valori interi maggiori, fino ad N per il valore più alto. Nel caso in cui i dati della variabile hanno due o più valori uguali, è necessario assegnare a ognuno di essi come rango la media delle loro posizioni;

3. Sostituire anche gli N valori di Y con i ranghi rispettivi;

4. A questo punto l'algoritmo si divide in due parti a seconda dei risultati:

   - Nel caso in cui le distribuzioni siano correlate in modo positivo $(\rho = +1)$, i valori della variabile X e della Y relativi allo stesso soggetto saranno uguali;
   - Nel caso in cui le distribuzioni siano correlate in modo negativo $(\rho = -1)$, a valori alti di X saranno associati valori bassi di Y o viceversa;
   - Nel caso in cui le distribuzioni presentano correlazione nulla $(\rho = 0)$, i valori di *X* e di Y relativi agli stessi soggetti saranno associati in modo casuale.

   Per quantificare questo grado di correlazione o concordanza, Spearman ha proposto la distanza tra le coppie dei ranghi (d_{R_i}). Questo valore è interpretato come la quantità utile per esprimere l'indicatore di correlazione, esso è calcolato come la somma dei quadrati delle differenze dei ranghi di Y con i ranghi di *X*, in formula:

   $$\sum d_{R_i}^2 \text{ dove } d_{R_i} = rank(y_i) - rank(x_i)$$

5. Il valore dell'indice $\rho = +1$ indica che le coppie di osservazioni di X e Y hanno lo stesso rango e pertanto questa sommatoria risulta uguale a 0. Se il valore dell'indice di Spearman $\rho = -1$ vale che se X è ordinato in modo crescente allora Y è ordinato in

modo decrescente: di conseguenza, le differenze sono massime e la sommatoria raggiunge un valore massimo determinato dal numero di coppie di osservazioni *N*. Ultimo caso da considerare è l'indice $\rho = 0$, mentre i ranghi di X sono ordinati in modo crescente quelli di Y hanno una distribuzione casuale: la sommatoria delle $d^2_{R_i}$ tende a un valore medio, determinato dal numero di coppie di osservazioni *N*.

Il secondo metodo nato 30 anni dopo la ρ di Spearman, è stato scoperto da Kendall proponendo il suo test τ [18]. Questo metodo ha le stesse assunzioni di Spearman, pertanto, può essere utilizzato nelle medesime condizioni e sui medesimi dati del test ρ. I risultati tra i due test sono molto simili, anche se matematicamente non equivalenti. La metodologia utilizzata per calcolare la τ di Kendall segue sei passaggi:

1. Occorre ordinare per ranghi la variabile *X*, assegnando il rango 1 al valore più piccolo e progressivamente un rango maggiore, fino ad N assegnato al valore più grande. In comune al test di Spearman nel caso in cui siano presenti due o più valori uguali nella variabile X, bisogna assegnare a ognuno come rango la media delle loro posizioni;

2. Sostituire gli N valori di Y anch'essi con i ranghi rispettivi;

3. Nel caso in cui le distribuzioni siano correlate in modo positivo $(\tau = +1)$, anche i ranghi della variabile Y sono ordinati in modo crescente, concordanti con l'ordine naturale. Riguardo al caso in cui sia presente una correlazione negativa con $(\tau = -1)$, vale che i valori di Y risulteranno ordinati in modo decrescente e saranno discordanti dall'ordine naturale. In ultimo, se tra le due variabili non esiste correlazione $(\tau = 0)$, l'ordine della variabile Y risulterà casuale e il numero di ranghi concordanti e di quelli discordanti dall'ordine naturale tenderà ad essere uguale, con somma 0;

4. Per quantificare il grado di correlazione o concordanza, Kendall ha proposto di contare la sola variabile *Y*. Contando nella successione, creata tramite la trasformazione in ranghi, quante sono le coppie di ranghi adiacenti che sono concordanti dall'ordine naturale e quante quelle discordanti (l'ordine naturale è avere una scala crescente di valori nella successione), vale che le coppie discordanti contribuiranno con un valore di -1 alla correlazione, viceversa nel caso in cui nell'ordine della variabile Y due valori contigui sono

nell'ordine naturale questo concorrerà con un valore $+1$. La misura della concordanza complessiva con la variabile X è data dalla somma algebrica di tutte le concordanze e discordanze;

5. Per ricondurre il valore calcolato a un campo di variazione compreso tra $+1$ e -1, il numero totale di concordanze e discordanze di una serie di valori deve essere rapportato al massimo totale possibile, quindi secondo il metodo proposto da Kendall, il grado di relazione o concordanza (τ) tra la variabile X e Y può essere quantifica dal rapporto:

$$\tau = \frac{totale(concordanze - discordanze)}{massimototalepossibile}$$

3.2.2 Test di Friedman per k campioni dipendenti

Il test proposto da Milton Friedman nel 1937 è utile quando si vuole osservare se le diverse mediane di più campioni dipendenti coincidono o meno tra loro. Tale tipo di test è utilizzabile quando si dispone di dati categorizzati su una scala quantitativa continua. Il test verifica l'ipotesi nulla sulla tendenza centrale:

$$H_0 : me_1 = me_2 = ... = me_k$$

con l'ipotesi alternativa:

$$H_1 : \text{non tutte le k mediane sono uguali}$$

dove me_i rappresenta la mediana della variabile *i*.

Il test di Friedman in alcuni testi di statistica introduttiva è considerata l'alternativa non parametrica all'ANOVA [1] a due criteri di classificazione o a blocchi randomizzati, quando non sono rispettate le condizioni di validità richieste dai test parametrici. Tuttavia, è noto nella statistica teorica che il test di Friedman è una generalizzazione dei test dei segni [29]. Esso è uno dei test non parametrici più potenti e di uso più generale.

Il procedimento è semplice e richiede pochi passaggi:

1. Creare una tabella a doppia entrata dove nelle righe vengono inseriti gli individui da analizzare e nelle colonne le diverse metriche da confrontare;

2. Trasformare i punteggi o le misure in ranghi entro la stessa riga, assegnando 1 al punteggio minore e progressivamente valori maggiori fino a k ;

3. Successivamente bisogna sommare per colonna i valori dei ranghi;

4. Se l'ipotesi nulla H_0 è vera, nelle colonne a confronto i ranghi minori e quelli maggiori dovrebbero essere distribuiti casualmente, pertanto le somme dei ranghi nelle k colonne dovrebbero essere tra loro tutte equivalente ed essere uguali a un valore atteso (T_i) che dipende dal numero di osservazioni. Nel caso in cui l'ipotesi è falsa, in almeno una

[1] https://it.wikipedia.org/wiki/Analisi_della_varianza

colonna si concentrano i ranghi minori o maggiori e di conseguenza tale somma tende ad essere significativamente differente dal valore T_i atteso, dove:

$$T_i(attesi) = \frac{N \cdot (k+1)}{2}$$

5. Per decidere se queste somme sono significativamente differenti dall'atteso, si calcola la statistica *Fr* che rappresenta la sommatoria dei quadrati degli scarti tra i k totali osservati e i corrispondenti attesi. Il valore di *Fr* tenderà a 0 nel caso di accordo tra totali osservati e totali attesi (H_0 vera), mentre tenderà a un valore alto al crescere dello scarto tra essi (H_0 falsa):

$$Fr = \sum (T_i - \frac{N \cdot (k+1)}{2})^2$$

Nel caso di grandi campioni (caso in cui $k \cdot N > 40$), per poter rifiutare o accettare l'ipotesi nulla, il test di Friedman ricorre a un indice χ^2_F che si distribuisce approssimativamente come il $\chi^2_{(k-1)}$ con gradi di libertà *k-1*. Può essere stimato mediante la formula:

$$\chi^2_F = \frac{12}{N \cdot k \cdot (k+1)} \sum_{i=1}^{k} (T_i - \frac{N \cdot (k+1)}{2})^2$$

in cui

- la seconda parte è data dagli scarti al quadrato tra somma osservata e attesa,
- mentre la prima dipende dall'errore standard.

Tralasciando i particolari sulla statistica χ^2_F e sul concetto di gradi di libertà che consideriamo futili ai fini della ricerca, vale che se il valore di quest'ultima è minore della statistica *Fr* si può rifiutare l'ipotesi nulla, con probabilità *p* di commettere un errore di primo tipo [18].

3.2.3 Test di Wilcoxon per due campioni dipendenti

Il test dei segni per ranghi di Wilcoxon è un test non parametrico che si applica nel caso di un singolo campione con due misure appaiate. Esso normalmente viene utilizzato per verificare se esistono differenze significative nei livelli mediani di due quantità dello stesso campione. Per semplicità potremmo dire che piuttosto di verificare l'ipotesi su k valori dello stesso campione, come nel caso del test di Friedman, esso verifica l'ipotesi nulla su una coppia di valori.

La metodologia risponde al confronto delle due mediane come ipotesi nulla H_0 che può essere unilaterale o bilaterale.

I passaggi per calcolare la statistica W del test di Wilcoxon è la seguente:

1. Inserire come righe di una tabella il campione a cui sono associate quattro colonne. Nella prima colonna viene inserito il valore della prima misura, nella seconda il valore della seconda misura, nella terza la differenza tra le precedenti misure e nell'ultima il rango corrispondente alla differenza delle misure;

2. Dopodiché vengono eliminate dall'analisi le differenze nulle, quindi risulterà che la numerosità del campione sarà proporzionalmente ridotta;

3. Trasformare le differenze considerate in valore assoluto nel loro rango. Nel caso di due o più dati uguali, assegnare lo stesso valore, calcolato come media dei ranghi;

4. Attribuire ad ogni rango il segno della differenza corrispondente;

5. Sommare i ranghi con lo stesso segno;

6. Scegliere il totale minore, esso sarà il valore di W;

7. Secondo l'ipotesi nulla H_0, la differenza tra le due serie di osservazioni appaiate dovrebbe essere uguale a 0. Di conseguenza, nella colonna delle differenze la somma dei ranghi con segno positivo e la somma dei ranghi con segno negativo dovrebbero essere uguali. Perciò, il totale minore dovrebbe tendere a un valore medio atteso μ_W determinato da N, secondo la relazione:

$$\mu_W = \frac{N \cdot (N+1)}{4}$$

8. La significatività della differenza tra le due serie di dati appaiati è tradotta nella significatività della differenza tra W e μ_W. Dato che facciamo riferimento a campioni formati da molti elementi ($N > 25$), il valore della somma dei ranghi (W) è distribuito in modo approssimativamente normale: la significatività della differenza può essere saggiata con la distribuzione Z, usando la relazione

$$Z = \frac{W - \mu_W}{\sigma_W}$$

dove la mediana attesa μ_W è data da

$$\mu_W = \frac{N \cdot (N+1)}{4}$$

e la deviazione standard σ_W è data da

$$\sigma_W = \sqrt{\frac{N \cdot (N+1) \cdot (2N+1)}{24}}$$

Tralasciando anche qui i particolari sulla statistica Z, la quale verrà introdotta parzialmente durante la presentazione dello svolgimento del tirocinio, nei prossimi capitoli, poiché utile nell'ambito della *Standardizzazione*. Possiamo concludere che nel caso in cui il valore della statistica Z è minore o uguale al valore della statistica Z_α (per ipotesi alternative unilaterali) si rifiuta l'ipotesi nulla e si accetta l'ipotesi alternativa unilaterale con p probabilità di commettere un errore di primo tipo [18].

Capitolo 4

Tecnologie e linguaggi adoperati

Nel seguente capitolo vengono trattati i concetti che caratterizzano l'ambiente in cui è stato svolto il tirocinio. La descrizione è stilata in modo tale da descrivere gli strumenti utilizzati durante tutte le fasi della ricerca sperimentale.

Le sezioni successive comprendono una descrizione dei protocolli di comunicazione e dei comandi fondamentali impiegati per la comunicazione remota e per l'esecuzione degli script. Gli script li abbiamo scritti utilizzando i linguaggi di programmazione Python e l'ambiente statistico R, di cui verrà fornita una descrizione nelle sezioni adatte. Infine, data un'introduzione dei linguaggi descriviamo le librerie cardine contenente le funzioni e gli algoritmi impiegati sui dati.

4.1 Protocolli di comunicazione e comandi

Il requisito fondamentale per cui è stato necessario l'utilizzo della macchina dipartimentale per l'esecuzione degli script è la quantità di memoria RAM installata all'interno, conseguenza del fatto che trattiamo algoritmi i quali ricevono in input file dalle dimensioni considerevoli. Tale vincolo impone l'impossibilità di utilizzare la macchina personale per il calcolo, in quanto l'esecuzione incorre in problemi di memoria. Trattando la modalità di collegamento in remoto, la soluzione più rapida e ovvia è l'utilizzo del servizio per la comunicazione tra macchine remote, chiamato SSH (Secure SHell). In informatica e telecomunicazioni il protocollo SSH permette di stabilire una sessione remota cifrata (protetta da utenti malintenzionati durante

il flusso di dati nella rete) tramite interfaccia a riga di comando con un altro host di una rete informatica. Per essere in grado di utilizzare tale funzionalità, in ambiente GNU/Linux, è sufficiente installare il pacchetto *openssh-server*, tramite il comando

```
1 apt-get install openssh-server
```

Installato il pacchetto, bisogna tenere conto che il protocollo SSH utilizza la porta TCP 22 il che non deve essere utilizzata da nessun'altro servizio della macchina locale (e, in modo ovvio, della macchina remota). Essendo un protocollo client-server, abbiamo bisogno di una struttura client che si interfacci con la struttura server. Nel caso di un sistema operativo Unix/Linux, il client per la shell remota è incluso nel pacchetto OpenSSH. La connessione a un server SSH, nei sistemi Unix-like, avviene tramite il comando con sintassi generale:

```
1 ssh [opzioni] nomeutente@host [comando]
```

Conclusa la descrizione riguardante la modalità di connessione a un host remoto, trattiamo un ulteriore comando utile in determinate situazioni. Un esempio di dinamica sconveniente durante l'utilizzo del protocollo SSH riguarda l'esecuzione di programmi avviati dal terminale locale sulla macchina remota. Per il corretto funzionamento del protocollo, l'host locale e l'host remoto devono essere in possesso di una connessione ADSL attiva e del funzionamento costante delle due macchine durante il periodo di connessione SSH. Nel caso di un malfunzionamento da parte della linea ADSL o di un host tra i due collegati, si rischia di perdere tutto il lavoro svolto dall'esecuzione di qualche programma avviato in remoto. Per ovviare a questo problema viene utilizzato un comando presente nei sistemi Unix e Unix-like, più in generale di sistemi POSIX, in grado di far ignorare al sistema operativo remoto il segnale di SIGHUP. Solitamente, un programma avviato tramite host remoto riceve tale segnale nel caso di interruzione di connessione o interruzione di corrente, interrompendo la sua esecuzione. Il comando *nohup* risolve tali tipi di problematiche, evitando l'interruzione dei comandi eseguiti tramite tale strumento e quindi l'eventuale segnale di SIGHUP verrà ignorato. La sintassi per il comando nohup è la seguente:

```
1 nohup comando [arg1 [arg2, ] ]
```

Nohup eseguirà il comando specificato nell'argomento in foreground e reindirizzerà l'output al file di log *nohup.out* conseguentemente creato nella directory di lavoro corrente. Inoltre, la sintassi del simbolo '&' per eseguire i comandi in background è utilizzabile in parallelo con il comando nohup.

Ultima utility il cui vale la pena considerare per trasferire file in modo sicuro da e per sistemi Linux è il comando SCP. Il protocollo SCP (secure copy) è in grado, sfruttando una connessione SSH, di copiare file e directory in sicurezza tra due host remoti. Con l'utilizzo di tale protocollo, dato che si tratta di trasferimento sicuro, sia i file che le password vengono correttamente crittografate.

La sintassi del comando generale è la seguente:

```
1 scp [opzioni] nomeutente@sorg:"file1" nomeutente@dest:"directory"
```

In cui i vari componenti del comando specificano:

- opzioni – Eventuali opzioni come cifratura, configurazione ssh, porta ssh, limite, copia ricorsiva, ecc;
- nomeutente@sorg:"file1" - File di origine;
- nomeutente@dest:"directory" – directory dove inserire il file di origine nella macchina remota.

A seconda della sintassi utilizzata nel comando si specifica di conseguenza la semantica dell'operazione. SCP fornisce, come visto, una serie di opzioni che controllano ogni aspetto del suo comportamento.

4.2 Linguaggi di programmazione impiegati

Nel corso della ricerca abbiamo avuto modo di prendere familiarità con due diversi linguaggi di programmazione mai utilizzati nel corso della carriera universitaria, il linguaggio Python e il linguaggio R. Il primo linguaggio di programmazione lo abbiamo scelto, in particolare, per la libreria Networkit [1] di cui forniremo una descrizione nelle sezioni successive. Il secondo

[1] https://networkit.github.io/

linguaggio lo abbiamo scelto per la sua semplicità nel calcolo e visualizzazione di parametri statistici e grafici, e dalla rapidità nell'integrazione di librerie presenti nel suo database CRAN (The Comprehensive R Archive Network) [2]

4.2.1 Python 3

Il linguaggio di programmazione Python3 lo abbiamo scelto come conseguenza del fatto che: è possibile integrare la libreria Networkit, pacchetto software open-source, in grado di effettuare analisi ad alte prestazioni su reti complesse [19]. Il requisito da rispettare è l'impiego della versione Python 3 [3] in modo da potersi interfacciare, correttamente, alle funzioni implementate all'interno del pacchetto sopra citato. Inoltre, abbiamo usufruito di un ulteriore libreria open-source definita in egual modo in Python, di nome SciPy, conosciuto per la modellazione e la risoluzione di problemi scientifici [24].

Python è un linguaggio di programmazione ad alto livello, orientato a oggetti, adatto, tra i diversi usi, a sviluppare applicazioni distribuite, scripting, computazione numerica e system testing. Esso è un linguaggio multi-paradigma che ha tra i principali obiettivi: dinamicità, semplicità e flessibilità. Supporta, come già detto, il paradigma *Object-oriented*, la programmazione strutturata e molte caratteristiche di programmazione funzionale e riflessione.

4.2.2 R (Software)

R [4] è un linguaggio di programmazione e un ambiente di sviluppo specifico per l'analisi statistica dei dati. Esso è un software libero in quanto viene distribuito con la licenza GNU GPL (General Public License), ed è disponibile per diversi sistemi operativi. Il suo linguaggio orientato agli oggetti deriva direttamente dal pacchetto S distribuito con una licenza non open source. La sua popolarità e semplicità è dovuta anche all'ampia disponibilità di moduli distribuiti con la licenza GPL e organizzati in un apposito sito chiamato CRAN. Tramite questi moduli è possibile estendere di molto le capacità del programma, aggiungendo anche di propria iniziativa moduli utili agli utilizzatori del sistema. Anche se il linguaggio è fornito con un'interfaccia a riga di

[2] https://cran.r-project.org/

[3] https://www.python.org/doc/

[4] https://www.rdocumentation.org

comando, sono disponibili diverse interfacce grafiche che consentono di integrare R con diversi pacchetti come Emacs (Editor di testo libero).

Abbiamo scelto di utilizzare il linguaggio R per la sua semplicità nell'organizzazione dei pacchetti e nel calcolo di determinati risultati, che attraverso l'utilizzo di altre "routine" di altri linguaggi avrebbe portato a una maggior complessità di svolgimento. Tra gli innumerevoli pacchetti presenti nel sito CRAN, abbiamo selezionato le librerie "labstatR [5]" per il calcolo di parametri statistici e "poweRlaw [6]" per delle rappresentazioni grafiche di distribuzioni.

4.3 Libreria Networkit

Questa sezione comprende una descrizione della libreria Networkit, utilizzata nel corso della ricerca sperimentale. Come già citato nel capitolo introduttivo, la libreria Networkit è un pacchetto software open-source per l'analisi ad alte prestazioni di reti complesse. Come reti complesse si intende tutte quelle strutture, anche eterogenee, che presentano una struttura comune. Degli esempi comuni di fenomeni avente determinate caratteristiche sono: le reti sociali, le reti biologiche, le reti informatiche il quale presentano la proprietà di poter essere manipolati tramite l'utilizzo di strumenti definiti nella teoria dei grafi. Conseguentemente a questa proprietà, l'utilizzo di tali metodi per la manipolazione di determinate strutture avviene nel campo della moderna ricerca web per visualizzare i nodi (le pagine web) con maggior prestigio, vengono utilizzati nell'ambito social media per i metodi di "community detection", ecc. Poiché nelle casistiche citate sopra consideriamo reti caratterizzate da un numero di archi considerevole (da 10^6 fino ad arrivare a 10^10 o superiore) occorrono algoritmi in grado di manipolare in tempi e spazi ottimali tali reti complesse. Attraverso Networkit l'analisi di così ampie reti complesse è possibile. Esso fornisce tutti gli strumenti utili per l'analisi di grafi riguardante: centralità, community detection, visite, calcolo del diametro e altre proprietà; il tutto già implementato in un pacchetto utilizzabile anche da utenti senza competenze di programmazione approfondite.

[5] https://cran.r-project.org/web/packages/labstatR/labstatR.pdf

[6] https://cran.r-project.org/web/packages/poweRlaw/poweRlaw.pdf

Una delle motivazioni che ha portato a preferire l'uso di Networkit è sicuramente l'ottimo compromesso tra buone prestazioni e perdita delle informazioni. Dalla documentazione [19] fornita dai produttori del sistema, sorge un importante lavoro di software engineering studiato per ottenere alte prestazioni e parallelismo, con un codice ibrido in modo da combinare le prestazioni del kernel facendo uso del linguaggio C++ e dalla scrittura di un front-end Python. È utile fornire un'introduzione di alcuni moduli implementati nella libreria: *networkit.centrality* e *networkit.base.* Nella composizione dei suddetti moduli è utilizzato il meccanismo di ereditarietà (concetto proprio della programmazione object-oriented per la scomposizione di un modello dati in una gerarchia di classi). La gerarchia definita tra le classi nei moduli è così composta (vedi anche figura 4.1):

- La classe astratta di base Oggetto;
- La classe astratta networkit.base.Algorithm, specializzazione della classe Oggetto;
- La classe astratta networkit.centrality.Centrality, specializzione della classe Algorithm;
- Le specializzazioni della classe Centrality, data dalle differenti classi:
  - networkit.centrality.ApproxBetweenness;
  - networkit.centrality.Closeness;
  - networkit.centrality.Eigenvector;
  - networkit.centrality.DegreeCentrality;
  - networkit.centrality.KatzCentrality;
  - networkit.centrality.PageRank

Di seguito una descrizione sommaria delle diverse classi:

- **networkit.base.Algorithm** – Classe container astratta per la definizione di algoritmi, tra i metodi implementati vi è il metodo run() che esegue l'algoritmo rappresentato dall'oggetto in tempo di esecuzione;
- **networkit.centrality.Centrality** - Classe astratta per la definizione delle misure di centralità, i metodi implementati tra gli altri sono ranking() che restituisce una lista i

cui elementi rappresentano il numero del nodo e il suo valore di centralità in ordine decrescente, e il metodo scores() che restituisce i valori di centralità ordinati per numero di nodo;

- **networkit.centrality.Valore** – Classe specializzata per il calcolo di una ben precisa centralità tra quelle elencate, la descrizione è fornita nel prossimo capitolo.

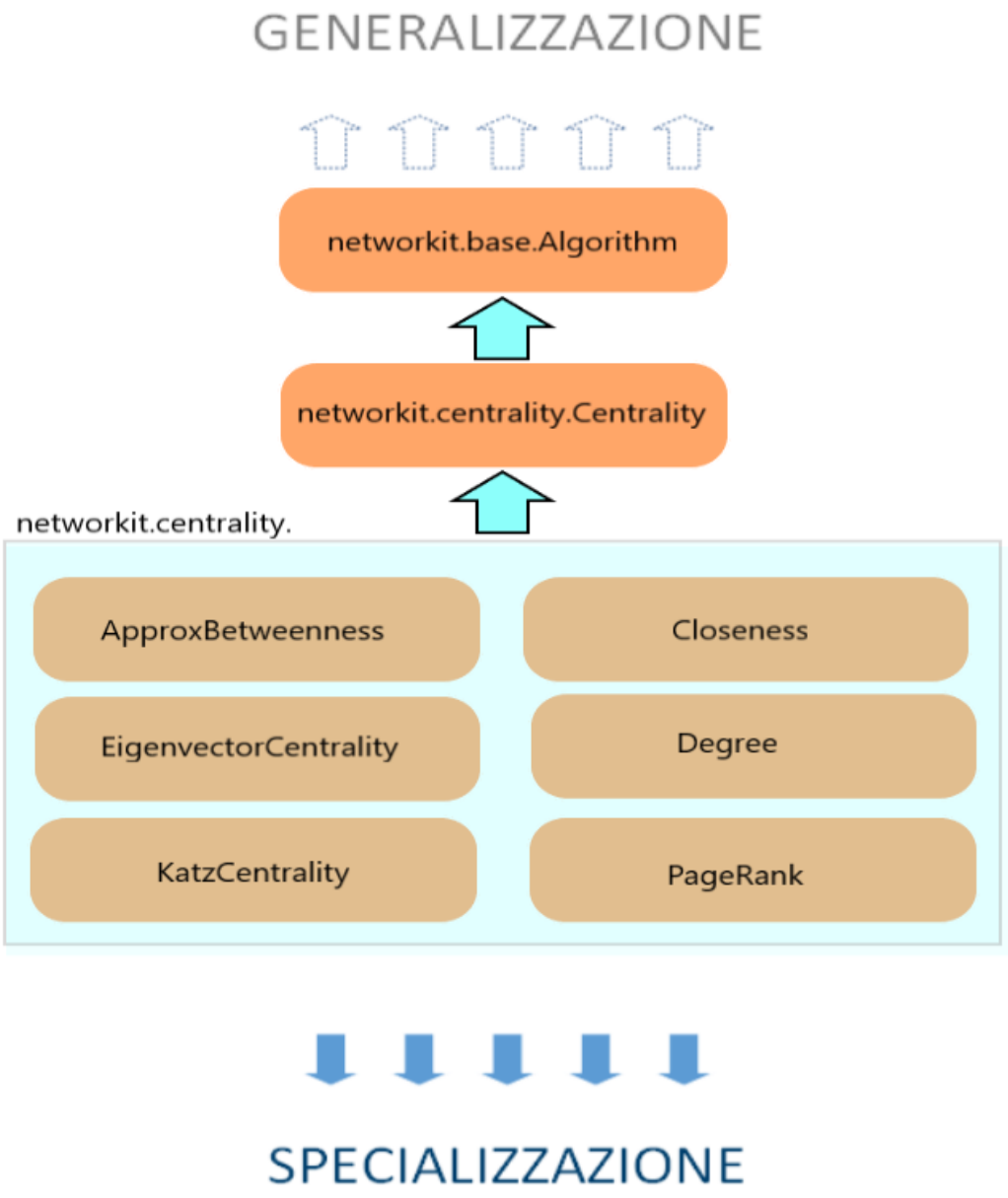


Figura 4.1: Ereditarietà dei moduli di Networkit.

4.4 Libreria poweRlaw e labstatR

Questa sezione comprende una descrizione sommaria tratte dalle librerie principali utilizzate durante la ricerca nell'ambiente R.

La libreria poweRlaw è utilizzata per l'analisi statistica di particolari tipi di distribuzioni di dati, più in particolare le distribuzioni di probabilità che hanno come caratteristica code lunghe (heavy tailed). La libreria è particolarmente utile per verificare se una distribuzione segue una legge power law, e nel caso affermativo analizzarla. Il pacchetto poweRlaw, dalla documentazione nell'archivio CRAN, implementa degli stimatori di massima verosimiglianza discreti e continui per adattare la distribuzione della legge di potenza ai dati utilizzando i metodi descritti in Clauset et al, 2009 [6]. Il pacchetto fornisce anche funzioni per adattare le distribuzioni log-normale ed esponenziali. Inoltre, è possibile stimare un possibile fitting, calcolando il valore x_{min} più adatto, per le leggi di potenza, log-normale ed esponenziale. La libreria labstatR comprende un insieme di funzioni per il calcolo di quasi tutti i parametri statistici più importanti come il calcolo dei quantili, medie, deviazioni standard, coefficienti, ecc.

Capitolo 5

Scale free network e indicatori di centralità

In questo capitolo andremo a trattare alcune proprietà molto importanti dei grafi quali centralità e una proprietà che caratterizza alcune reti reali. Uno tra i primi a impostare delle basi sulle metriche di centralità, sfruttando i lavori effettuati da Bavelas, fu Freeman che definì un insieme di indici da allora considerati classici nella letteratura scientifica: centralità di grado, centralità betweeness e centralità closeness [16]. A distanza di anni, a questi gruppi di indici strutturali sono stati aggiunti ulteriori indici di classificazione che descriveremo.

Al fine di rendere chiare le caratteristiche che queste centralità esprimono, possiamo raggrupparle in due macro-gruppi:

1. Il primo gruppo riguarda le misure che quantizzano tramite i propri collegamenti la fama di un utente all'interno di un contesto sociale. Tra queste metriche analizzeremo la centralità betweeness, centralità closeness e centralità in base al grado di uscita o di entrata (chiamante comunemente, rispettivamente, centralità out-degree e centralità in-degree);

2. Il secondo gruppo comprende tutte quelle metriche che come metro di misura utilizzando la "fama" (prestigio) degli utenti a cui sono direttamente collegati, tra questi abbiamo la centralità eigenvector, centralità secondo Katz e la centralità Page Rank.

5.1 Rete a invarianza di scala

Introduciamo le reti a invarianza di scala (scale free network) come quei particolari grafi che godono della proprietà per cui pochi vertici hanno il grado di connessioni alto e un altissimo numero di vertici hanno il grado di connessioni basso ed esso può essere descritto tramite una distribuzione power law. Le scale free networks sono state scoperte attraverso un esperimento condotto da un fisico, informatico e ricercatore ungherese chiamato Albert-László Barabási, il quale ha esaminato il grafo del web ponendo come ipotesi iniziale il fatto che il web potesse essere descritto mediante un grafo random (cioè, ogni nodo sceglie casualmente a chi collegarsi tra gli altri nodi della rete). L'esperimento contraddice l'ipotesi iniziale in quanto più dell'80% delle pagine ha meno di quattro links, ma lo $0,01\%$ delle pagine include più di mille links, quindi esistono alcuni hubs che "dominano" la rete [2]. Un altro modello nato provando a simulare le reti sociali e cercando di trovare proprietà generali comuni a tutti le reti reali è il modello del mondo piccolo (small world). Watts e Strogatz modellano in modo appropriato una teoria matematica e sociologica il quale sostiene che tutte le reti complesse presenti in natura sono tali che una qualunque coppia di nodi possono essere collegati da un cammino costituito da un numero relativamente piccolo di archi. Talvolta, si fa riferimento ad essa come teoria dei sei gradi di separazione [26]. Tuttavia, quest'ultimo modello non cattura bene altre proprietà importanti di alcune reti reali. Infatti, come già anticipato in molte reti esistono pochi nodi con grado alto e molti con grado basso. Da qui nasce il concetto di funzione detta scale free. Una funzione f è scale free se e solo se:

$$f(bx) = C(b)f(x)$$

dove $C(b)$ è una costante dipendente solo da b, quindi la "forma" della funzione f non cambia quando si considerano valori di *x* che sono moltiplicati per un fattore *b*, la power law soddisfa queste proprietà, infatti:

$$f(bx) = (bx)^{-\alpha} = b^{-\alpha}x^{-\alpha} = b^{-\alpha}f(x)$$

In conclusione, mostriamo graficamente le differenze che intercorrono tra una rete costruita casualmente (random graph), una rete a invarianza di scala (scale-free network) e una rete

costruita mediante l'attuazione della teoria del mondo piccolo (small-world network) (figura 5.1).

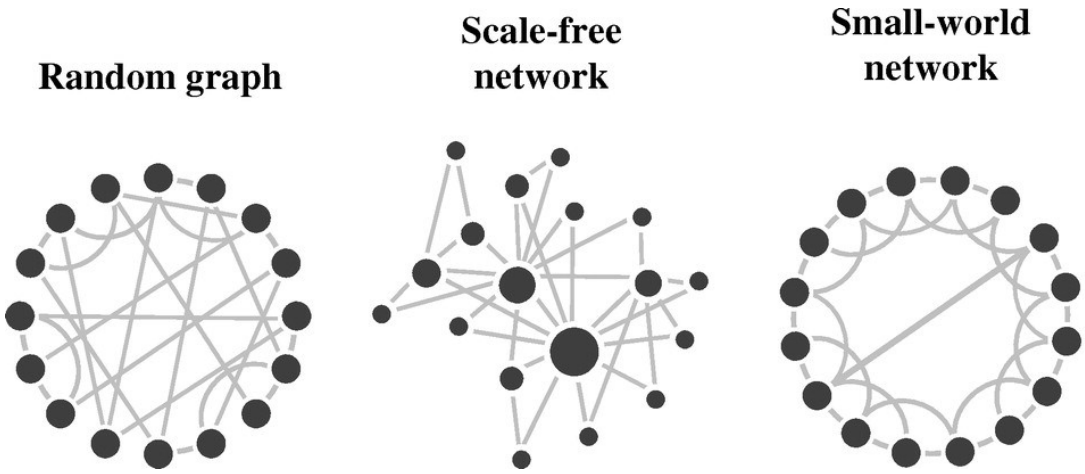


Figura 5.1: Differenze grafiche dei diversi tipi di rete.

5.2 Centralità In-degree e Out-degree

toricamente la prima misura di centralità considerata per la sua intuitività è la misura del grado. Come grado di un nodo si intende la cardinalità degli archi che puntano e vengono puntati per e dal nodo in questione. La definizione di centralità di grado può essere divisa in centralità di grado in entrata (in-degree) e centralità di grado in uscita (out-degree). Il grado di collegamento in entrata è la caratterizzazione del rischio immediato che quel nodo riceva ciò che sta fluendo all'interno della rete. Nel caso specifico di una rete sociale, gli archi associati sono aspetti positivi (cioè scambio di informazioni) e quindi la centralità in-degree può essere considerata come una forma di popolarità nella rete in termini di "quante persone prendono in considerazione il nodo per far fluire l'informazione". Il grado di centralità in-degree di un vertice *v*, per un grafo $G := (V, E)$ con $|V|$ (cardinalità di *V*) vertici e $|E|$ archi, è definita come:

Indegree(v) = #(nodi che puntano a v)

Equivalentemente alla misura della centralità in-degree, la misura out-degree utilizza il grado in uscita dei nodi, cioè il numero di archi originati dal nodo in questione che si vanno a collegare ad altri nodi della rete. Quindi, equivalentemente, alla misura in-degree rappresenta

una forma di popolarità all'interno della rete in termini di "quanti contatti possiede il nodo per far fluire l'informazione". Il grado di centralità di un vertice *v*, per un grafo $G := (V, E)$ con $|V|$ vertici e $|E|$ archi, è definita come:

Outdegree(v)= #(nodi che vengono puntati da v)

La segnatura del metodo costruttore per il calcolo di entrambe le centralità, all'interno del modulo Networkit, è la seguente: DegreeCentrality(G, normalized=False, outDeg, ignoreSelfLoops=True) Il metodo si aspetta come parametri il grafo *G*, la variabile booleana per la normalizzazione dei risultati e un'opzione *ignoreSelfLoops* utilizzato, nel caso fosse impostato a *True*, per ignorare gli archi con coda e testa nello stesso nodo. Il parametro *outDeg* viene impostato per calcolare la centralità in-degree nel caso assumesse il valore *False* oppure la centralità out-degree nel caso assumesse il valore *True*. Nell'algoritmo utilizzato su Networkit, il metodo run() per il calcolo del grado in entrata ha complessità in tempo $O(m)$ dove *m* è il numero di archi che contiene il grafo. Riguardante il metodo run() per il calcolo del grado in uscita, esso ha complessità $O(1)$.

5.3 Centralità Betweeness

Nella teoria dei grafi, la quantità betweeness è una misura di centralità nei grafi basata sul concetto di "cammino minimo". Come cammino minimo in un grafo non pesato viene inteso rispetto la lunghezza di un cammino, cioè il numero di collegamenti attraversati per andare da un nodo all'altro (in contrapposizione alla lunghezza di un cammino in un grafo pesato, dove ogni collegamento ha un peso che lo caratterizza secondo un'unità di misura, ad esempio, fisica).

La centralità betweeness esprime quanto un nodo sia da mediatore nei flussi d'informazione tra gli altri abitanti della rete, cioè un nodo risulta centrale secondo la betweeness se si trova nel maggior numero di distanze geodetiche (percorsi più brevi) che collegano ogni coppia di attori non adiacenti. Il calcolo della misura per il nodo v avviene tramite il calcolo dell'espressione:

$$\text{Betweeness(v)} = \sum_{s \neq v \neq t} \frac{\sigma_{st}(\text{v})}{\sigma_{st}}$$

nella quale σ_{st} è il numero totale di cammini minimi dal nodo s al nodo t e $\sigma_{st}(v)$ è la cardinalità del sottoinsieme formato dai cammini minimi che passano per il nodo v. La centralità betweeness è estremamente costosa da calcolare in termini di tempo, dato che essa è un problema di minimo e quindi incorre nella problematica dell'enumerazione di tutti i possibili cammini (chiamati path) tra tutti i nodi, scegliendo tra essi quelli di minor lunghezza. Per ridurre il problema della complessità in tempo, incorrendo però nella precisione dei risultati ottenuti, facciamo uso di un algoritmo approssimato rappresentato dalla classe networkit.centrality.ApproxBetweenness [17]. Il costruttore della classe networkit.centrality.ApproxBetweenness prende in ingresso quattro parametri: il *grafo*, *epsilon*, *delta* e *universalConstant*.

La segnatura del metodo costruttore nel modulo corrispondente di Networkit è la seguente:

ApproxBetweenness(G, epsilon=0.01, delta=0.1, universalConstant=1.0)

Dalla documentazione, l'algoritmo si basa sul campionamento casuale dei cammini minimi e offre garanzie probabilistiche sulla qualità dell'osservazione, tutti i valori calcolati sono all'interno di un fattore additivo ϵ dai valori reali, con probabilità almeno $1 - \delta$. I valori vengono scalati di default. Il metodo run() ha un costo in tempo di $O(m)$ per campione, dove m è il numero di archi.

5.4 Centralità Closeness

In un grafo connesso, la centralità closeness di un nodo è calcolata come il reciproco della somma delle lunghezze dei cammini minimi tra il nodo da esaminare e tutti i nodi all'interno del grafo. Quindi, più la quantità di closeness di un nodo è alta più il nodo è "vicino", in media, a tutti gli altri nodi della rete, quindi interagisce velocemente con tutti gli altri attori. La misura è stata definita da Bavelas (1950) come il reciproco della farness [3] (misura che calcola la somma dei cammini):

$$\text{Closeness(x)} = \frac{1}{\sum_y d(y, x)}$$

dove d(y, x) è la distanza tra i vertici x e y.

La classe definita in networkit crea un algoritmo che calcola l'esatta quantità di closeness di tutti i nodi del grafo.

Il costruttore in Networkit presenta la seguente segnatura:

Closeness(G, normalized=FALSE, variant=networkit.centrality.ClosenessVariant.Standard)

L'algoritmo si aspetta due parametri oltre al grafo da analizzare, di cui una variabile booleana per la normalizzazione dei risultati e una variabile *variant* per decidere quale definizione di closeness scegliere per il calcolo. Nell'implementazione le possibilità per quest'ultima variabile variant sono due: la versione standard (utilizzata nel nostro caso) per il calcolo su grafi connessi, e una versione generalizzata per il calcolo su grafi non connessi. L'algoritmo calcola i valori non normalizzati, dopo il settaggio del parametro *Normalized* a *False*. Il metodo run() ha un costo in tempo di $O(n \cdot m)$, dove n è il numero di nodi e m il numero di archi del grafo.

5.5 Centralità Eigenvector

Nella teoria dei grafi, la centralità Eigenvector (anche chiamata "eigencentrality" o "punteggio prestigio") è una misura dell'influenza di un nodo all'interno di una rete. I relativi punteggi vengono assegnati a tutti i nodi della rete in base al concetto che le connessioni ai nodi con punteggio alto contribuiscono maggiormente al punteggio del nodo in questione rispetto alle connessioni ai nodi con punteggio basso. Un punteggio di eigenvector alto significa che un nodo è connesso a molti nodi che hanno punteggi alti [13]. Dato un grafo $G := (V, E)$ con $|V|$ vertici e $|E|$ archi, consideriamo la matrice di adiacenza $A = (a_{v,t})$, nella quale $a_{v,t} = 1$ se il vertice v è collegato al vertice t, e $a_{v,t} = 0$ altrimenti. La centralità eigenvector del nodo v può essa può essere definita come:

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t$$

dove $M(v)$ è l'insieme dei nodi vicini a v e λ è una costante. Con un piccolo riarrangiamento questo può essere riscritto in notazione vettoriale come equazione per il calcolo degli autovettori relativi all'autovalore λ:

$$Ax = \lambda x$$

In generale, ci sono diversi autovalori λ corrispondenti ad autovettori diversi da zero. Tuttavia, la matrice di adiacenza ha tutte le sue componenti non negative e ciò implica che, per il teorema di Perron-Frobenius [14], l'autovalore di modulo massimo λ di A è reale positivo e l'autovettore corrispondente ha tutte le componenti positive. La componente *v-esima* del relativo autovettore fornisce quindi il punteggio di centralità relativa al vertice v nella rete.

Il costruttore della classe networkit.centrality.Eigenvector inizializza, tramite i suoi parametri, il grafo e la tolleranza per il calcolo dell'autovettore. Il parametro *tol* (tolleranza) è utilizzato per questioni di ottimizzazione numerica durante il calcolo.

La dicitura del costruttore della classe, nel modulo su Networkit, è la seguente:

EigenvectorCentrality(G,tol=1e-9)

5.6 Centralità secondo Katz

La centralità secondo Katz è una generalizzazione della centralità di grado, introdotto da Leo Katz nel 1953 [11]. Come abbiamo visto, la centralità di grado tiene conto del numero di nodi direttamente collegati al nodo a cui si sta applicando la misurazione. La centralità secondo Katz, invece, misura il numero di tutti i nodi la quale sono connessi attraverso un cammino e allo stesso tempo quantizzando tali "pesi" secondo una classificazione rispetto alla distanza geodetica di tali nodi [10].

La centralità per un nodo v è matematicamente calcolata come:

$$\text{Katz(v)} = \sum_{k=1}^{+\infty} \sum_{j=1}^{N} \alpha^k (A^k)_{jv}$$

dove α è il fattore di attenuazione compreso tra (0,1), estremi esclusi.

La centralità secondo Katz può essere vista come variante della centralità per autovettore (eigenvector) [23]. Un ulteriore modo di calcolare la centralità per Katz è:

$$\text{Katz(v)} = \sum_{j=1}^{N} a_{vj}(x_j + 1)$$

Messa a confronto con l'espressione della centralità per autovettore, si nota che al posto di $(x_j + 1)$ abbiamo solamente il termine x_j. Per quanto riguarda l'algoritmo contenuto nella

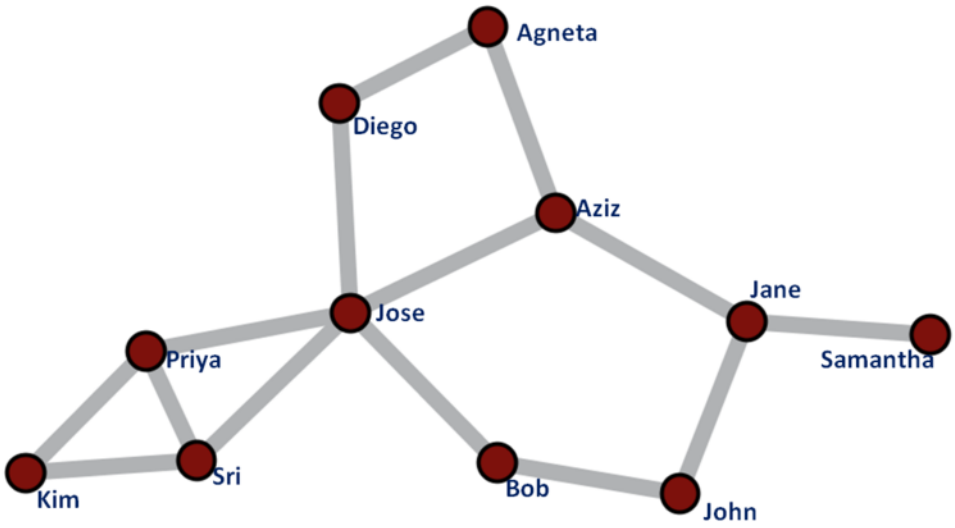


Figura 5.2: Esempio di grafo in cui calcolare la centralità secondo Katz.

libreria Networkit, esso richiede tempo $O(m)$ dove m è il numero di archi del grafo. Il numero di iterazioni dipende da quanto tempo è necessario per raggiungere la convergenza (e quindi sulla tolleranza desiderata, specificata come parametro della classe).

La segnatura del costruttore è:

KatzCentrality(G,alpha=0.05,beta=0.1,tol=1e-8)

Di seguito mostriamo un esempio per il calcolo della centralità secondo Katz, prendendo come riferimento il grafo della figura sopra (figura 5.2). Assumiamo di voler calcolare la centralità di John con $\alpha = 0.5$. Il peso assegnato ad ogni collegamento con i nodi direttamente collegati a John avranno un peso nella sommatoria di $(0.5)^1 = 0.5$. Jose è connesso indirettamente a John, attraverso la connessione con Bob, dato che il cammino è composto da due collegamenti avremo che il peso nella sommatoria sarà $(0.5)^2 = 0.25$. Similmente i collegamenti con Agneta attraverso Aziz e Jane sarà $(0.5)^3 = 0.125$ e così via. Una volta calcolati tutti i fattori, possiamo calcolare la centralità, tramite una sommatoria di tutti questi termini.

5.7 Centralità PageRank

La centralità che si affianca alle centralità eigenvector e alla centralità secondo Katz, è la centralità PageRank [1] [15]. Considerando le centralità già viste, notiamo che la centralità secondo Katz ha il problema seguente: se il nodo con alta centralità è collegato a molti altri nodi, allora anche questi ultimi automaticamente avranno un'alta centralità. In molti casi, tuttavia, significa meno se un nodo è solo uno tra i tanti da collegare, cioè la centralità acquisita da un nodo dovrebbe essere maggiore nel caso in cui questo nodo abbia un basso grado di connessioni. La centralità PageRank è considerata una correzione della centralità secondo Katz poiché tiene in considerazione questo problema.

Ci sono tre distinti fattori che determinano tale centralità per un nodo:

1. Il numero di archi collegati allo stesso;
2. La propensione al collegamento dei suoi vicini;
3. La centralità dei suoi vicini.

Il primo fattore è ovvio, infatti più un nodo è "linkato" (riferito) dagli altri nodi più è centrale. Ragionevolmente, il valore dell'appoggio si deprezza proporzionalmente al numero di collegamenti forniti dal nodo: i link provenienti da nodi parsimoniosi sono più degni di quelli emanati da nodi con un eccesso di link (ad essere considerati, detto brutalmente, nodi "spam"). Infine, non tutti i nodi anche con egual grado di connessione vengono considerati allo stesso modo: i collegamenti offerti da vertici importanti sono più prestigiosi di collegamenti offerti da vertici sconosciuti.

Questa classificazione è stata utilizzata da Google per assegnare un valore di "prestigio" alle pagine Web. Si tratta di una distribuzione probabilistica che misura la probabilità che un utente generico, navigando in maniera random attraverso i link delle pagine visitate, arrivi a una pagina target. L'algoritmo Page Rank è composto da una fase di inizializzazione e una di aggiornamento:

$$PR(p_i) = \frac{1}{n} \qquad \text{Inizializzazione}$$

$$PR(p_i) = \frac{1-d}{n} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \qquad \text{Aggiornamento}$$

- $PR(p_i)$ è il PageRank di un nodo p_i;
- $M(P_i)$ è l'insieme di tutti i nodi che puntano p_i;
- $L(p_j)$ è l'out-degree del nodo p_j.

d è il damping factor. La differenza 1-d è la probabilità che l'utente che naviga in maniera random seguendo i vari collegamenti web, possa decidere di smettere di seguire i collegamenti e di scegliere di aprire una pagina a caso.

Nel caso in cui la centralità PageRank venga riferita a una rete sociale, semanticamente parlando, un nodo avrà un alto valore di PageRank quando gli utenti saranno interessati a interagire con quel particolare utente. Un esempio di valore acquisito dal dumping factor, nell'algoritmo di ranking di google come indicato dal paper di Brin e Page [15], è di $0,85$.

La segnatura del costruttore, definita nel modulo Networkit, è la seguente:

PageRank(G, damp=0,85, tol=1e-9)

Capitolo 6

Svolgimento del tirocinio

In tutta la ricerca scientifica sperimentale è indispensabile una corretta impostazione dell'indagine statistica composta da una chiara presentazione dei dati e della loro elaborazione che seguano criteri validi universalmente. Al fine di facilitare la corretta comprensione dei risultati si ritiene necessario e utile seguire uno schema preciso che è fondato sullo sviluppo di tre fasi, oltre a un'introduzione già fornita nei capitoli precedenti:

- La descrizione del **disegno sperimentale** nella quale sono definiti le modalità di raccolta dati, il tipo di scala utilizzato e ulteriori modellazione dei dati;
- La **descrizione statistica** che comprende una descrizione delle distribuzioni dei dati tramite metodi grafici, semi-grafici e analitici;
- La **descrizione inferenziale** che comprende le citazioni ai test d'inferenza utilizzati chiarendo le ipotesi che si intendono verificare e dei risultati generati effettivamente dai test.

6.1 Disegno sperimentale: Progettazione dell'analisi sperimentale

L'obiettivo della ricerca consiste nell'analizzare le proprietà strutturali della rete Steemit, in termini di centralità. Trattandosi di un'analisi statistica sperimentale, il primo passo fondamentale è la raccolta dei dati a cui vogliamo fare riferimento per le successive fasi di indagine. Dato che il sistema Steemit è caratterizzato da una struttura blockchain, il metodo più rapido per l'acquisizione dei dati, è tramite l'utilizzo di librerie fornite dagli sviluppatori di Steemit, di nome steem-python [1]. La libreria steem-python permette di effettuare il download dei blocchi della blockchain in modo tale da poter estrarre le informazioni al loro interno. I blocchi selezionati per la modellazione partono dal giorno 8 Marzo 2017 fino al blocco con data 1° gennaio 2019 formando un dataset, contenente tutte le interazioni dei blocchi estratti, di dimensione complessiva maggiore di 430 GB. Per interazioni intendiamo una qualsiasi azione permessa dal sistema che eseguita da un utente influenzerà un altro utente diverso dal primo. All'interno di Steemit abbiamo classificato gli utenti come: utenti bot e utenti non bot. Inoltre, possiamo classificare le interazioni in due diversi gruppi: interazioni sociali e interazioni economiche/management. Una volta che abbiamo ottenuto tutte le informazioni utili per rappresentare la composizione strutturale della rete Steemit, abbiamo usufruito dell'unico strumento matematico in grado di sintetizzare le informazioni della rete: un grafo. A partire dal dataset abbiamo costruito, per mezzo delle funzioni fornite da Networkit, un grafo di tutte le interazioni (vedi figura 6.1).

Il grafo di Steemit rappresenta le interazioni tra gli utenti della piattaforma. Si tratta di un grafo orientato, in cui gli utenti di Steemit sono rappresentati dai nodi e le azioni che intraprendono sono rappresentate da archi. Dall'analisi delle proprietà base del grafo risulta un grafo diretto e non pesato che rappresenta 1.24 milioni di nodi con 191.2 milioni di archi. Considerando la classificazione degli utenti e delle interazioni, abbiamo derivato tre ulteriori grafi dal grafo principale. Le tre derivazioni del grafo globale sono caratterizzate come:

- **Grafo delle interazioni sociali** – Il grafo delle interazioni sociali è stato pensato e costruito considerando solamente le transazioni col quale si interagisce nell'ambito del social network. Il grafo considerato ha proprietà con valori piuttosto simili al grafo

[1] https://steem.readthedocs.io/en/latest/

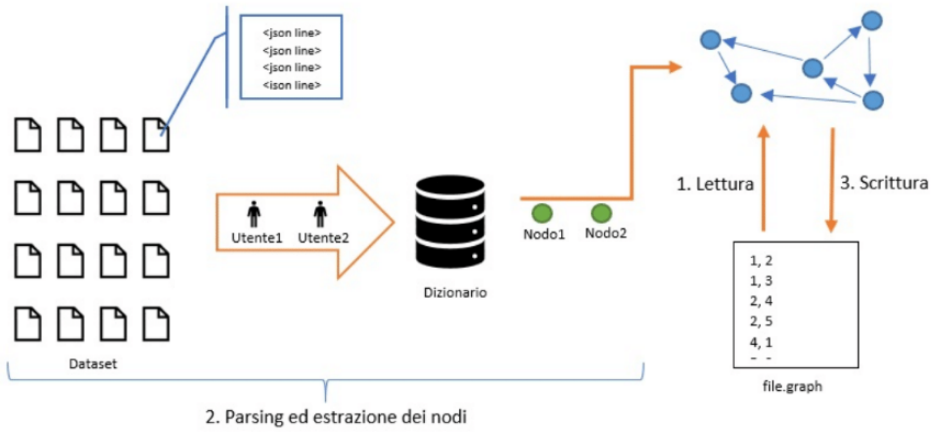


Figura 6.1: Creazione del grafo globale.

globale. Il numero dei nodi è 1.23 milioni. Di conseguenza, il numero di utenti che non ha mai effettuato alcuna operazione di tipo sociale è di circa 14 mila, cioè l'1% degli utenti totali del grafo di tutte le interazioni;

- **Grafo sociale senza bot** – Alla classifica degli utenti centrali per out-degree e in-degree nel grafo social, emerge la forte presenza di account di proprietà di Steem e di account bot. I corrispondenti nodi giocano sicuramente un ruolo fondamentale nel mantenere la macrostruttura del grafo coesa, per questo motivo abbiamo eliminato 200 bot tra quelli più centrali della rete.

- **Grafo delle interazioni monetarie** – Il sottografo rappresenta solamente interazioni di tipo finanziario. Il grafo delle interazioni monetarie risulta piuttosto piccolo rispetto ai grafi precedenti. Questo va ad evidenziare quella che è la natura di Steemit: la parte monetaria è ridotta e gli utenti ricevono valuta grazie al rewarding system. Il grafo contiene circa 1 milione di nodi e 4 milioni di archi.

Sfruttando i concetti della teoria dei grafi e della teoria delle reti sociali, valutiamo i grafi a livello macroscopico, studiandone la struttura e la connettività, e a livello microscopico, analizzando il comportamento che tendono ad avere i nodi. Per valutare il grafo a livello macroscopico abbiamo utilizzato il concetto di centralità dei nodi. Esistono diverse definizioni di centralità, e ognuna di essi quantifica e qualifica tutti i nodi della rete sociale per mezzo

dei loro collegamenti. Ogni algoritmo di centralità considerato sfrutta in maniera differente i collegamenti di un nodo. Nello specifico, abbiamo deciso di calcolare sette tipi di centralità sui quattro grafi utilizzando gli algoritmi predefiniti nella libreria Networkit. Quindi, abbiamo scritto un programma che includesse le librerie Networkit, e che ricevuto in input un grafo sotto forma di file restituisse in output sette differenti file contenente i risultati degli algoritmi per ogni nodo (uno per riga). In totale, contiamo quattro programmi in esecuzione poiché quattro sono i grafi considerati. Il programma riguardante il grafo monetario ha terminato la sua esecuzione nell'arco di circa 72 ore, fornendoci di tutte le misure per tutte le centralità prestabilite. I programmi riguardante i tre grafi di dimensioni maggiori hanno mostrato due inconvenienti non banali:

- La centralità secondo Katz ha prodotto un file contenente tutte le misure dei nodi di valore uguale a 0;
- La centralità closeness, non essendo un'approssimazione, per la complessità in tempo dell'algoritmo e per il numero troppo alto di nodi e archi richiedeva un tempo di esecuzione troppo alto.

Tenuto conto di queste problematiche, abbiamo deciso di eliminare dall'analisi statistica descrittiva (ad eccezione del grafo monetario nell'appendice C) e inferenziale l'uso di queste due centralità, tenendo le cinque centralità invece calcolate con successo. Tali misure considerate in questa ricerca sono valori continui (appartenenti alla retta reale positiva, $\mathbb{R}^+$) che possono essere raggruppati sotto un unico tipo di scala indicata come **scala di rapporti** la quale ha il vantaggio di avere un'origine reale (caratteristica fondamentale di queste misure è la definizione del valore 0 (zero) che indica quantità nulla). Un nodo può avere tranquillamente zero di centralità il che indica assenza di collegamenti con il resto del grafo. L'individuazione del tipo di scala è fondamentale per l'utilizzo dei corretti metodi statistici descrittivi e inferenziali tra i tanti conosciuti.

Una volta posti i concetti di base utili a una corretta indagine statistica, introduciamo le ipotesi che intendiamo verificare con i lavori successivi all'acquisizione dei dati. In principio è utile chiarire quali sono le proprietà a cui gli utenti del social media si affidano, anche involontariamente, per l'interazione con il resto della società. Ulteriori studi effettuati su altri

tipi di reti reali, condotti da sociologi e informatici nel passato, hanno portato ad intravedere una predisposizione dei nodi a interagire o collegarsi con i nodi più influenti della rete, ad esempio nel grafo web i nodi tendevano ad interagire con i nodi più ricchi in termini di collegamenti. Tutto ciò in contrapposizione con l'ipotesi iniziale in cui si riteneva che i collegamenti effettuati dei nodi avessero un andamento "normale": pochi nodi con pochi collegamenti, molti nodi con molti collegamenti e un valore medio in cui si concentra la maggior parte della popolazione. Questo comportamento scaturisce una legge fondamentale, nota come legge di potenza, che rende i nodi ricchi (nel caso del grafo web, ricchi in collegamenti) ancora più ricchi con lo scorrere del tempo. Detto ciò, parte della ricerca consiste nel visualizzare le distribuzioni di tutte le centralità per tutti i grafi estrapolando una possibile somiglianza con distribuzioni già note.

In secondo luogo, esponiamo le ipotesi che intendiamo verificare attraverso i test d'inferenza. L'analisi può essere divisa logicamente in due parti: una parte effettuata al livello dei singoli grafi e quindi sulle metriche di centralità; una parte comprende una visione un po' più ampia, studiando, attraverso i risultati acquisiti, le differenze che intercorrono tra i diversi grafi. Una prima ipotesi che vogliamo verificare è la possibile esistenza di una qualsiasi relazione tra le diverse misure di centralità, e nel caso in cui la risposta sia affermativa verificare per quali fattori intrinsechi esse siano correlate. Una seconda ipotesi a cui facciamo riferimento è notare se le mediane delle diverse centralità calcolati sui singoli grafi siano considerevolmente diverse tra loro.

Chiarito se smentire o accettare tali ipotesi, possiamo discutere sul significato sociale dei risultati con una successiva comparazione tra i diversi grafi, estrapolando quali gruppi di interazioni siano rilevanti all'interno del sistema e quali utenti (bot e non) influiscano maggiormente nelle operazioni svolte sul social media.

6.2 Descrizione statistica: Analisi delle frequenze e distribuzioni

La descrizione statistica viene mostrata suddivisa in cinque sezioni, distinte in termini di centralità. Ogni sezione comprende una descrizione dei grafici stem-and-leaf, una descrizione tabellare contenente i parametri statistici e i grafici delle funzioni di ripartizione con un eventuale "fitting" rispetto a delle leggi di probabilità già note. Inoltre, viene esposta una descrizione delle eventuali differenze tra i grafi nella visione delle distribuzioni. La scelta di un grafico di tipo stem-and-leaf è stata ponderata per la sua capacità nell'astrarre le informazioni rispetto, ad esempio, l'utilizzo di un istogramma. Infatti, tramite un istogramma abbiamo notato la rappresentazione di un'unica classe contenente tutti i valori minori, tralasciando le informazioni essenziali riguardo le frequenze minori. Poiché trattiamo misure discrete e misure continue, è utile spendere un commento sulla semantica di un grafico stem-and-leaf nei due casi:

- Caso discreto – Nel ramo (cifra a sinistra della pipe) è inserita la prima cifra o le prime cifre, in base alla scala utilizzata (visibile nel commento del grafico), del valore discreto preso in considerazione. Nella foglia è considerata la singola cifra successiva alla cifra utilizzata nel ramo.

- Caso continuo – Nel ramo (cifra a sinistra della pipe) è inserita la prima cifra o le prime cifre significative, in base alla scala utilizzata (visibile nel commento del grafico), della parte decimale. Nella foglia è considerata la singola cifra successiva alla cifra significativa inserita nel ramo.

Per la rappresentazione dei grafici e il calcolo dei parametri statistici abbiamo utilizzato il software R, attraverso l'utilizzo del pacchetto base per i grafici stem-and-leaf, la libreria "labstatR" per il calcolo dei parametri e il pacchetto "poweRlaw" per il calcolo delle funzioni di ripartizione, presentati in Sezione 4.4. Riguardo la funzione cumulativa abbiamo dovuto imporre il vincolo secondo cui debbano essere considerati solamente valori maggiori strettamente di zero. Riguardante la centralità eigenvector abbiamo dovuto imporre un vincolo ancor più ristretto, scegliendo come soglia il valore $1e-05$ in modo da eliminare quei pochi nodi con centralità

troppo basse per essere presi in considerazione. La scelta di eliminare i valori uguali a 0 deriva dal dominio definito nella relazione della legge di potenza, infatti si ha $f(x) = \frac{1}{x^{-\alpha}}$ e risulta che la funzione è definita per valori strettamente maggiori di zero.

6.2.1 Comparazioni: In-degree

Analizziamo in questa sezione una centralità intuitiva di fondamentale importanza per determinare quantitativamente il numero di interazioni accumulate da un utente rispetto gli altri utenti del social media. Presentiamo la centralità in-degree rispettando i passi guida presentati all'inizio del paragrafo. Rappresentiamo le distribuzioni di frequenza della centralità tramite il grafico stem-and-leaf per i quattro grafi, tenendo in considerazione la corretta semantica trattandosi di valori discreti.

Per motivi di rapida comparazione e visualizzazione, i diagrammi vengono posti in una singola pagina così da rendere chiare le differenze tra i quattro grafi presi in esame secondo tale centralità

```
The decimal point is 4 digit(s) to the right of the |

 0 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+1243850
 1 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+666
 2 | 00000000000000000000000000011111111111111111111222222222222222223333333+61
 3 | 000000000111111233333444444556666778888
 4 | 002234555566678
 5 | 025559
 6 | 12368
 7 | 045
 8 | 14
 9 |
10 | 6
11 | 6
```

(a) Grafo steemit

```
The decimal point is 4 digit(s) to the right of the |

 0 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+1229560
 1 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+653
 2 | 00000000000000000000001111111111111111122222222222222233333333333344+51
 3 | 000000001111122334444455556778889
 4 | 022344555666678
 5 | 234455
 6 | 168
 7 | 455
 8 | 4
 9 |
10 | 5
11 | 6
```

(b) Grafo social

```
The decimal point is 4 digit(s) to the right of the |

 0 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+1229575
 1 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+643
 2 | 00000000000000000000001111111111111111122222222222233333333333444+48
 3 | 000000001111122334444444556777889
 4 | 022344555666678
 5 | 233455
 6 | 168
 7 | 45
 8 | 4
 9 |
10 | 5
11 | 6
```

(c) Grafo subsocial

```
The decimal point is 3 digit(s) to the right of the |

 0 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+1074547
 2 | 011112222223444444556666778889901111122333667778
 4 | 001233460246778
 6 | 3456345588
 8 | 131569
10 | 12245581
12 | 2
14 | 36
16 |
18 | 3
20 |
22 |
24 | 14
26 |
28 |
30 |
32 |
34 |
36 |
38 | 3
40 | 5
```

(d) Grafo monetario

Le distribuzioni di frequenze presentano le seguenti caratteristiche, divise per punti:

- Caratteristica comune ai quattro grafi è la presenza di valori più frequenti nelle classi di valore minore. Dato questo risultato, notiamo la presenza di un solo picco nelle distribuzioni di frequenza denotata dalle classi di valore 0;
- La forma della distribuzione risulta comune ai quattro grafi e si presenta secondo un'asimmetrica positiva. La curva della distribuzione per i quattro grafi è leptocurtica, con l'eccezione per il grafo monetario di avere una curva più "appuntita";
- I grafici contengono dei valori anomali rappresentati dai valori maggiori delle distribuzioni;
- I valori massimi sono approssimativamente: 116.000 per il grafo steemit, 116.000 per il grafo social, 116.000 per il grafo subsocial, 40.500 per il grafo monetario. Quindi, il massimo tra i valori massimi dei quattro grafi risultano essere i valori del grafo steemit, social e subsocial.

Integriamo l'analisi delle distribuzioni di frequenze calcolando e commentando i parametri statistici più importanti delle distribuzioni di centralità in-degree (Tabella 6.1).

(a) Grafo steemit

Parametro	**Valore**
Minimo	0
Q_1	4
Mediana	12
Media	153, 6
Q_3	85
Massimo	115.867
Asimmetria	30, 98
Curtosi	2.309
Dev. Standard	727, 33
Coeff. Variazione	4, 73

(b) Grafo social

Parametro	**Valore**
Minimo	0
Q_1	3
Mediana	11
Media	152, 7
Q_3	86
Massimo	115.825
Asimmetria	30, 12
Curtosi	2.256
Dev. Standard	714, 57
Coeff. Variazione	4, 68

(c) Grafo subsocial

Parametro	**Valore**
Minimo	0
Q_1	3
Mediana	11
Media	151, 8
Q_3	86
Massimo	115.751
Asimmetria	29, 78
Curtosi	2.233
Dev. Standard	706, 98
Coeff. Variazione	4, 66

(d) Grafo monetario

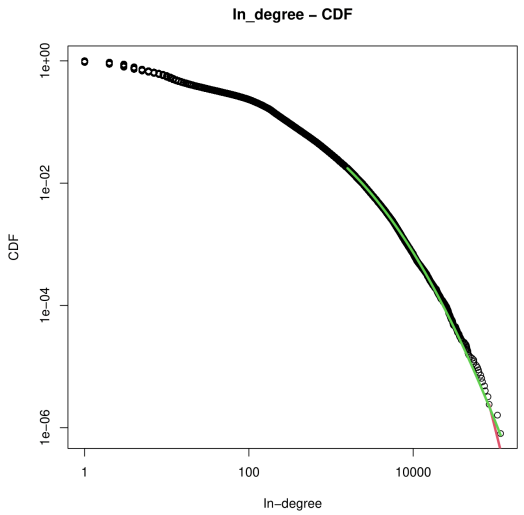
Parametro	**Valore**
Minimo	0
Q_1	1
Mediana	1
Media	3, 91
Q_3	2
Massimo	41.542
Asimmetria	268, 13
Curtosi	100.271
Dev. Standard	88, 05
Coeff. Variazione	22, 51

Tabella 6.1: Parametri statistici delle centralità di grado in entrata.

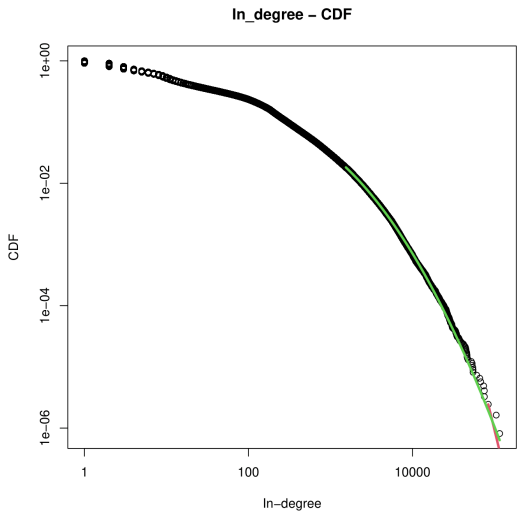
Commentiamo le quattro tabelle statistiche, confermando i parametri visualizzati tramite i grafici stem-and-leaf attraverso un valore numerico calcolato analiticamente:

- Il minimo e il massimo nei quattro grafi coincidono, in modo ovvio, con i valori approssimati nei grafici stem-and-leaf;
- Le mediane dei grafi steemit, social e subsocial assumono circa lo stesso valore. Il terzo quartile per questi ultimi risulta essere approssimativamente uguale e si distacca notevolmente dalla corrispondente mediana. Tali valori mostrano quanto sia alto il *gap* tra i pochi nodi ricchi di archi entranti e la restante popolazione rappresentata dal terzo quartile con basso valore in proporzione alla grandezza del grafo. La mediana del grafo monetario è uguale a 1 quindi il 50% dei nodi ha centralità uguale o minore a 1. Il terzo quartile si distacca di un'unità rispetto al valore della mediana, conseguenza del fatto che è raro avere nodi con alto numero di archi entranti.
- Gli indici di asimmetria e curtosi confermano i risultati analizzati nei grafici stem-and-leaf, mostrandosi maggiori nel grafo monetario rispetto ai grafi steemit, social e subsocial;
- Il coefficiente di variazione contiene un valore simile per i grafi steemit, social e subsocial dovuti al rapporto tra un valore molto alto di deviazione standard e la media. Il coefficiente di variazione per il grafo monetario si distacca dal coefficiente calcolato per i restanti grafi presentando un valore cinque volte maggiore. Conseguenza della differenza tra i coefficienti di variazione dovuta al fatto che tutti i nodi del grafo hanno una centralità in-degree molto bassa a differenza di pochi con centralità, a confronto, molto alta.

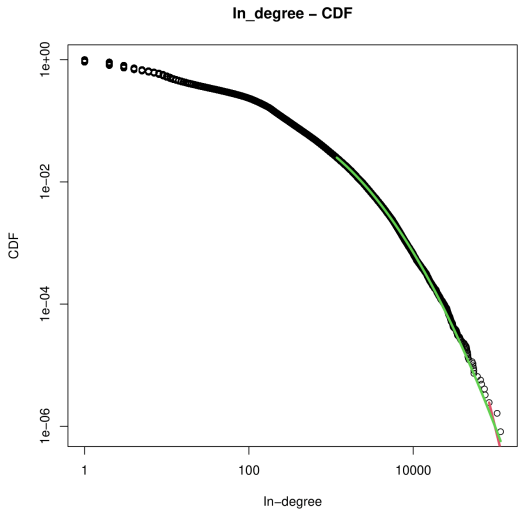
In ultimo, mostriamo le distribuzioni per la centralità in-degree ponendo in risalto la lunga coda (heavy-tailed) attraverso l'utilizzo della funzione di distribuzione cumulativa (Figura 6.2). Utilizzando il software statistico, si è eseguito il fitting secondo le leggi già note: log-normale e power-law.



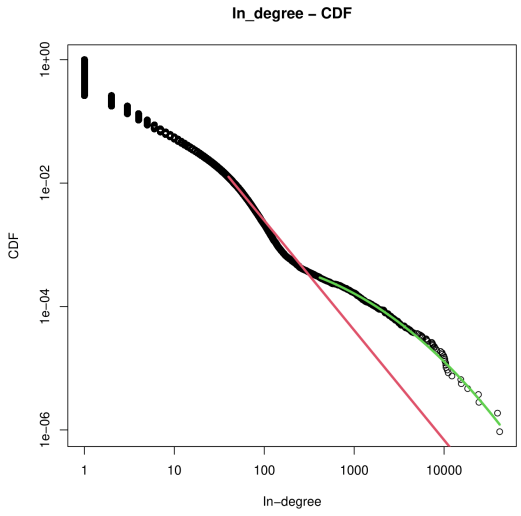
(e) Grafo steemit



(f) Grafo social



(g) Grafo subsocial



(h) Grafo monetario

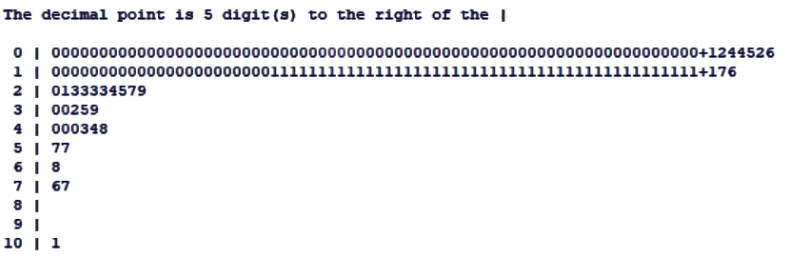
Figura 6.2: CDF delle centralità di grado in entrata.

Studiando le quattro funzioni cumulative, notiamo la netta differenza tra le distribuzioni dei grafi steemit, social e subsocial rispetto alla distribuzione del grafo monetario. Le funzioni dei primi tre grafi si mostrano chiaramente come una curva; il fitting effettuato secondo la legge log-normale (curva verde) sulle tre distribuzioni in questione si adatta perfettamente all'andamento dei dati. Da escludere l'andamento secondo una power-law a seguito del fitting effettuato (retta viola) il quale ha stimato un valore x_{min} troppo alto. La funzione del grafo

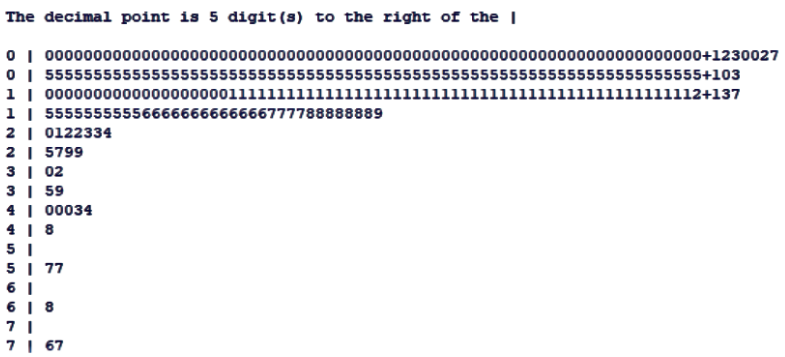
monetario, invece, si comporta diversamente rispetto ai restanti tre grafi. Notiamo non un andamento curvo, bensì la distribuzione tende ad essere più rettilinea. I fitting eseguiti mostrano un andamento non chiaramente definibile, infatti, il fitting secondo una legge di potenza non riesce a riassumere i dati. Il fitting secondo una distribuzione log-normale riesce, a partire da un valore compreso tra $(500, 1000)$, a riassumere in parte la coda della distribuzione.

6.2.2 Comparazioni: Out-degree

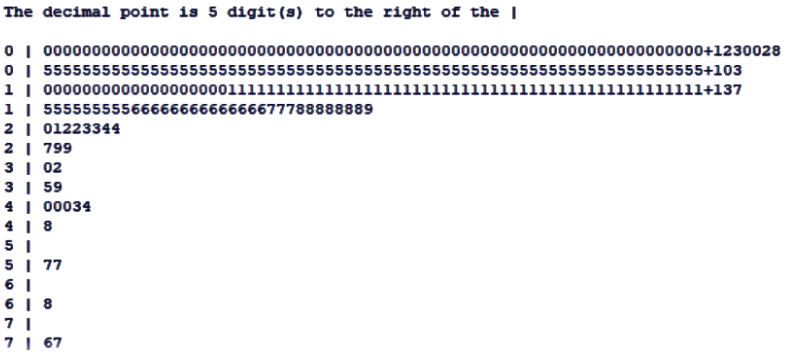
In questa sezione analizziamo una centralità intuitiva nella teoria dei grafi utile per analizzare la ricchezza dei nodi in termini di collegamenti in uscita. Questa centralità, in termini di social network, si riferisce a quante interazioni rivolte ad altri utenti, quantitativamente, ha accumulato l'utente analizzato nel corso del tempo all'interno del social media. Coerentemente alla centralità in-degree, la semantica dei diagrammi stem-and-leaf è riferita a valori discreti.



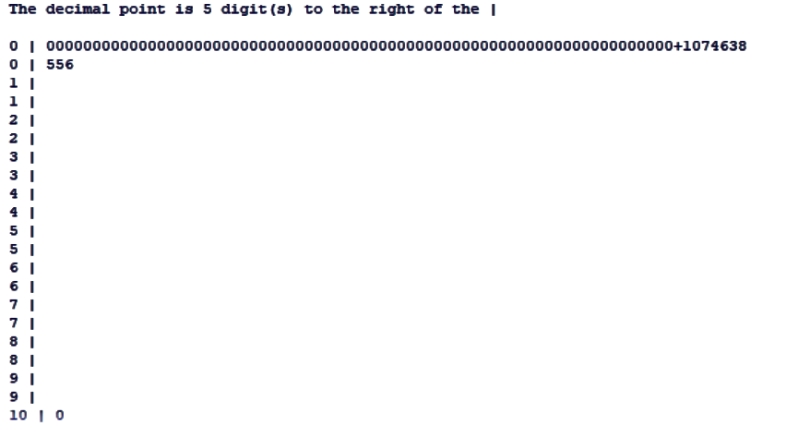
(a) Grafo steemit



(b) Grafo social



(c) Grafo subsocial



(d) Grafo monetario

Prima di analizzare i quattro grafici stem-and-leaf, commentiamo lo stem-and-leaf del grafo monetario il quale è stato semplificato in modo da poter mostrare chiaramente i parametri più importanti della distribuzione. Non teniamo conto del fatto che tutti i valori hanno 0 di centralità ma consideriamo solo il distacco che vi è tra il valore massimo e il resto dei valori. Studiamo i quattro grafici stem-and-leaf:

- I valori più frequenti all'interno dei quattro grafi si concentrano nelle classi più basse. Comunemente nei quattro grafi, la classe con maggiore frequenza è la classe di valore 0. Notiamo, inoltre, che le quattro distribuzioni sono caratterizzate da un unico picco di frequenze;

- Le distribuzioni mostrano una forma asimmetrica a destra per tutti e quattro i grafi. La forma della curva risulta più "appiattita" nelle distribuzioni dei grafi steemit, social e subsocial mentre una forma più "appuntita" nella distribuzione del grafo monetario;

- Il grafo steemit, social e subsocial presentano approssimativamente un numero uguale di outliers. Il grafo monetario contiene un unico valore anomalo che si distingue particolarmente rappresentato dal massimo della distribuzione;

- I massimi del grafo steemit, social e subsocial hanno un valore approssimato, rispettivamente: $1.010.000$, 770.000, 770.000. Il massimo del grafo monetario ha un valore approssimato di $1.000.000$. Quindi, in conclusione, notiamo che il massimo del grafo monetario si avvicina al massimo del grafo steemit a differenza dei rimanenti due grafi che presentano un valore di massimo minore.

.

L'analisi grafica delle frequenze la integriamo con una tabella riassuntiva dei parametri statistici più importanti (Tabella 6.2).

(a) Grafo steemit

Parametro	**Valore**
Minimo	0
Q_1	0
Mediana	2
Media	153, 67
Q_3	15
Massimo	1.005.838
Asimmetria	122, 70
Curtosi	26.111
Dev. Standard	2.958
Coeff. Variazione	19, 26

(b) Grafo social

Parametro	**Valore**
Minimo	0
Q_1	0
Mediana	1
Media	152, 7
Q_3	15
Massimo	765.879
Asimmetria	105, 60
Curtosi	18.717
Dev. Standard	2.823
Coeff. Variazione	18, 49

(c) Grafo subsocial

Parametro	**Valore**
Minimo	0
Q_1	0
Mediana	1
Media	151, 8
Q_3	15
Massimo	765.685
Asimmetria	105, 95
Curtosi	18.829
Dev. Standard	2.818
Coeff. Variazione	18, 56

(d) Grafo monetario

Parametro	**Valore**
Minimo	0
Q_1	0
Mediana	0
Media	3, 9
Q_3	0
Massimo	999.576
Asimmetria	993, 84
Curtosi	1.014.722
Dev. Standard	978, 15
Coeff. Variazione	250, 03

Tabella 6.2: Parametri statistici della centralità di grado in uscita.

Studiamo le quattro tabelle statistiche, confermando ciò che abbiamo visto nei grafici stem-and-leaf:

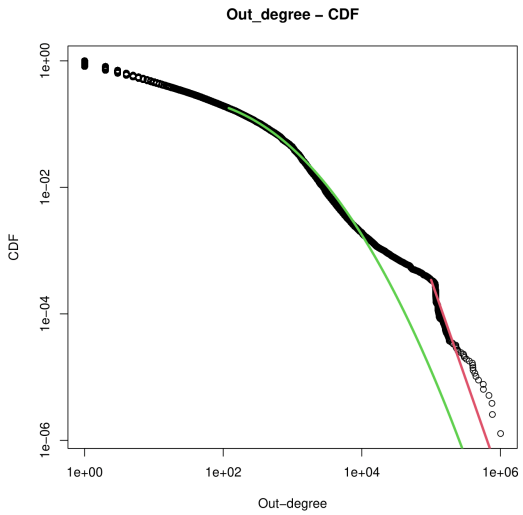
- Come notato dal grafico stem-and-leaf, il minimo per i quattro grafi è 0 e i massimi coincidono con quelli già approssimati. In particolare, notiamo che il massimo del grafo steemit contiene esattamente 6.262 archi (in uscita) in più rispetto al massimo del grafo monetario. Quindi, l'utente con più alta centralità su steemit riceve per la maggior parte interazioni monetarie e le restanti 6.262 provengono dal gruppo delle interazioni sociali;

- Le mediane in tutti e quattro i grafi risultano particolarmente basse. Il terzo quartile per i grafi steemit, social e subsocial assume valore 15, quindi il 75% dei nodi ha al massimo 15 archi in uscita. Il grafo monetario, invece, è caratterizzato dal 75% dei nodi con centralità uguale a 0.

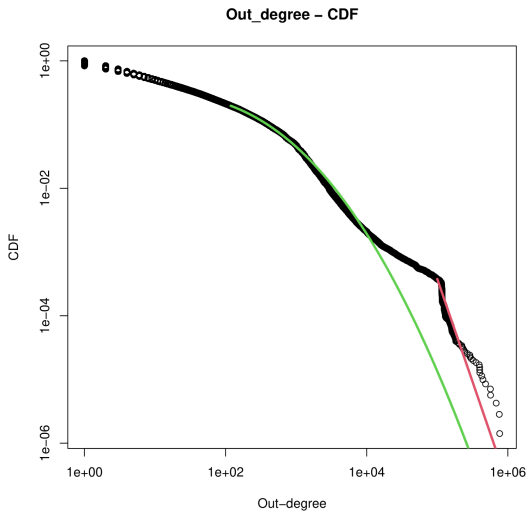
- Gli indici di asimmetria e curtosi confermando l'andamento dei grafici stem-and-leaf, con un curtosi piuttosto elevato per il grafo monetario.

- Il coefficiente di variazione per il grafo monetario risulta particolarmente alto rispetto ai valori dei grafi steemit, social e subsocial. Tale valore del grafo monetario è dovuto al basso valore di deviazione standard, rispetto all'alto valore assunto dai restanti tre grafi.

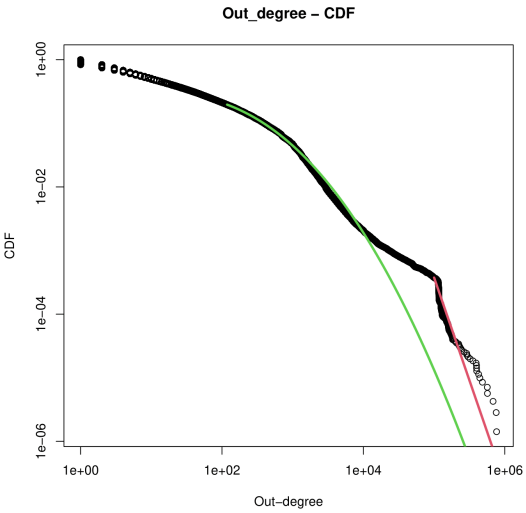
Una volta che abbiamo sintetizzato i dati per i quattro grafi nella centralità in-degree, ci interessiamo alla visione del tipo di distribuzione analizzata attraverso l'utilizzo della funzione di distribuzione cumulativa. Su tali distribuzioni verranno calcolato eventuali fitting rispetto alle distribuzioni log-normale e power law.



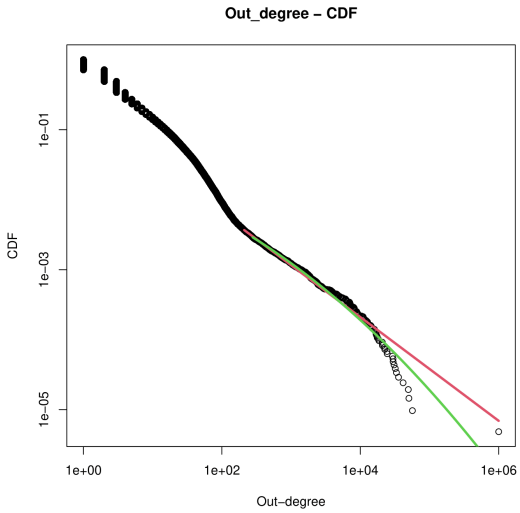
(e) Grafo steemit



(f) Grafo social



(g) Grafo subsocial



(h) Grafo monetario

ILe distribuzioni si presentano simili per i grafi steemit, social e subsocial rispetto al grafo monetario. La funzione calcolata sui primi tre grafi risulta parecchio frastagliata, comprese di curve anomale nei valori dell'intervallo $(1e-04, 1e-06)$. La funzione calcolata sul grafo monetario, invece, si presenta con un andamento più lineare.

Studiando il fitting rispetto la distribuzione log-normale (curva verde) e la distribuzione power law (retta viola) notiamo che entrambe non riescono a riassumere le quattro distribuzioni analizzate. Nei grafi steemit, social e subsocial abbiamo notato un andamento rispetto la

distribuzione log-normale nei valori intermedi. Nel grafo monetario, il fitting secondo la power law riesce in parte a seguire la distribuzione dei dati, anche se non in maniera accurata.

6.2.3 Comparazioni: Betweeness

In questa sezione presentiamo l'analisi che abbiamo effettuato sui quattro grafi rispetto la centralità betweeness. Proseguendo per gradi, mostriamo innanzitutto i quattro grafici stem-and-leaf che mostrano le distribuzioni di frequenze secondo tale centralità.

```
The decimal point is 2 digit(s) to the left of the |

 0 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+1244798
 1 | 001112
 2 | 45
 3 | 8
 4 |
 5 |
 6 | 8
 7 |
 8 |
 9 |
10 |
11 |
12 | 2
```

(i) Grafo steemit

```
The decimal point is 2 digit(s) to the left of the |

 0 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+1230479
 1 | 00126
 2 | 5
 3 | 9
 4 |
 5 |
 6 |
 7 | 2
 8 |
 9 |
10 |
11 |
12 |
13 | 1
```

(j) Grafo social

```
The decimal point is 2 digit(s) to the left of the |

 0 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+1230479
 1 | 01126
 2 | 6
 3 | 9
 4 |
 5 |
 6 |
 7 | 2
 8 |
 9 |
10 |
11 |
12 |
13 | 0
```

(k) Grafo subsocial

```
The decimal point is 2 digit(s) to the left of the |

 0 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+1074637
 1 | 4
 2 |
 3 | 67
 4 | 1
 5 |
 6 |
 7 |
 8 |
 9 |
10 |
11 |
12 |
13 | 1
```

(l) Grafo monetario

Innanzitutto, dato che le centralità betweeness presentano valori continui dobbiamo attenerci alla giusta semantica dei grafici stem-and-leaf. Detto ciò, commentiamo i diversi grafici che risultano in generale piuttosto semplici da decifrare. Presentiamo i seguenti punti:

- Notiamo che i valori più frequenti nei quattro stem-and-leaf coincidono, mostrando un unico picco nelle distribuzioni;
- Le distribuzioni presentano pressocché la stessa forma asimmetrica positiva con una curva leptocurtica piuttosto pronunciata;
- Contiamo una minima presenza di valori anomali nelle distribuzioni di frequenza nei quattro grafi. Osserviamo che tra gli outliers, questi si presentano piuttosto isolati tra loro; in modo pronunciato questo comportamento è assunto nel grafo monetario in cui il massimo del grafo si distacca dal resto degli outliers e di conseguenza dal picco di valori più frequente.
- Approssimativamente il massimo del grafo steemit è di valore $0,122$, il massimo del grafo social è $0,131$, il massimo del grafo subsocial è $0,130$ e il massimo del grafo monetario è $0,131$. Confrontando tali valori notiamo che i massimi tendono ad avere lo stesso valore. I valori maggiori risultano i massimi del grafo social e del grafo monetario. Il massimo del grafo steemit tende ad avere minore centralità betweeness rispetto al resto dei grafi, conseguenza del fatto che essendo un grafo molto ampio i nodi fanno più fatica a fungere da "ponte" per gli altri nodi all'interno del grafo.

Oltre a un'analisi grafica, proponiamo un'analisi tabellare (Tabella 6.3) sui parametri statistici più importanti delle quattro distribuzioni.

(a) Grafo steemit

Parametro	**Valore**
Minimo	0
Q_1	0
Mediana	0
Media	$9,00e-07$
Q_3	0
Massimo	$0,12$
Asimmetria	$672,98$
Curtosi	534.764
Dev. Standard	$0,00014$
Coeff. Variazione	$153,68$

(b) Grafo social

Parametro	**Valore**
Minimo	0
Q_1	0
Mediana	0
Media	$8,8e-07$
Q_3	0
Massimo	$0,13$
Asimmetria	$716,63$
Curtosi	590.598
Dev. Standard	$0,00014$
Coeff. Variazione	$165,96$

(c) Grafo subsocial

Parametro	**Valore**
Minimo	0
Q_1	0
Mediana	0
Media	$8,7e-07$
Q_3	0
Massimo	$0,13$
Asimmetria	$713,58$
Curtosi	585.582
Dev. Standard	$0,00014$
Coeff. Variazione	$167,01$

(d) Grafo monetario

Parametro	**Valore**
Minimo	0
Q_1	0
Mediana	0
Media	$3,8e-07$
Q_3	0
Massimo	$0,13$
Asimmetria	$774,60$
Curtosi	676.020
Dev. Standard	$0,00014$
Coeff. Variazione	$370,741$

Tabella 6.3: Parametri statistici per la centralità betweeness.

Commentiamo le misure statistiche calcolate sulle quattro distribuzioni:

- Il valore minimo per le quattro distribuzioni coincide ed è 0. I massimi approssimati nei grafici stem-and-leaf risultano all'incirca identici anche nelle tabelle statistiche. In aggiunta, anche in questo caso, si nota il nodo con maggiore centralità betweeness per il grafo steemit minore dei nodi con maggiore centralità betweeness per i grafi social, subsocial e monetario;

- La mediana e il terzo quartile per i quattro grafi risultano tutti di valore 0. Quindi, almeno il 75% dei nodi ha centralità betweeness uguale a 0;

- Gli indici di asimmetria e curtosi risultano approssimativamente dello stesso valore nei quattro grafi, confermando l'andamento osservato nei grafici stem-and-leaf;

- Il coefficiente di variazione nei grafi steemit, social e subsocial all'incirca presenta lo stesso valore. Nel grafo monetario, tale indice, risulta quasi il doppio rispetto ai precedenti tre grafi. Tale differenza è dovuta alla differenza delle medie poiché la deviazione standard risulta uguale per tutti e quattro i grafi.

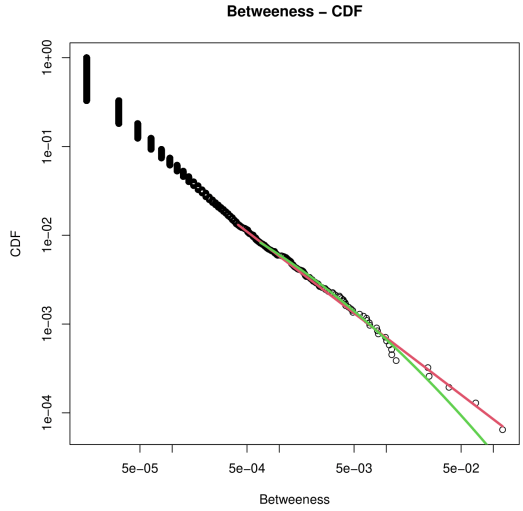
Completiamo la descrizione statistica della centralità betweeness con la visualizzazione delle distribuzioni attraverso la funzione di ripartizione.

L'ipotesi iniziale per cui consideriamo solo i valori maggiori strettamente di 0, impone che la maggior parte dei nodi della distribuzione vengano eliminati dall'analisi del grafico. Coerentemente con la descrizione statistica effettuata fino a questo momento sulla centralità betweeness, anche la funzione di ripartizione si presenta piuttosto semplice da commentare. Le funzioni di ripartizione calcolate sui grafi steemit, social e subsocial si presentano piuttosto simili, con un andamento rettilineo. Effettuando il fitting sulle distribuzioni, abbiamo notato che la legge di potenza (retta viola) riassume in maniera quasi impeccabile la distribuzione dei dati in contrapposizione al fitting della log-normale (curva verde).

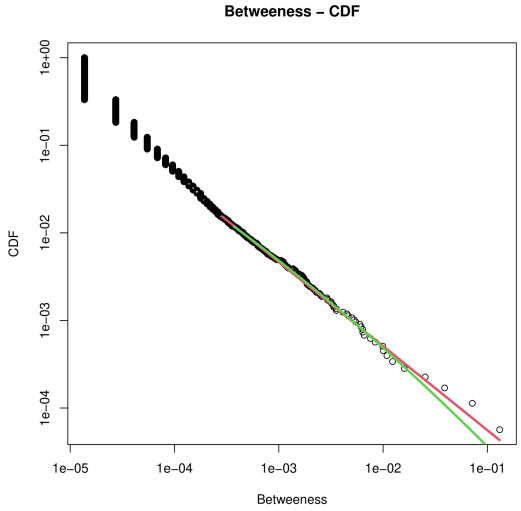
La funzione di ripartizione per il grafo monetario mostra comunque un andamento rettilineo, ma con una pendenza diversa conseguenza data dal diverso valore del coefficiente angolare della retta considerata. Il fitting della legge di potenza riassume approssimativamente l'andamento dei dati.

6.2.4 Comparazioni: Eigenvector

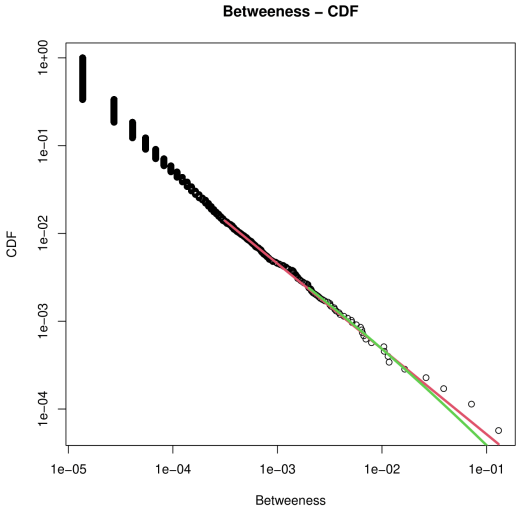
Introduciamo la descrizione statistica della centralità eigenvector sui quattro grafi. I metodi utilizzati per la corretta visione dei dati seguono le solite linee guida utilizzate fino ad adesso per le precedenti centralità. Analizziamo, in principio, i grafici stem-and-leaf che abbiamo calcolato.



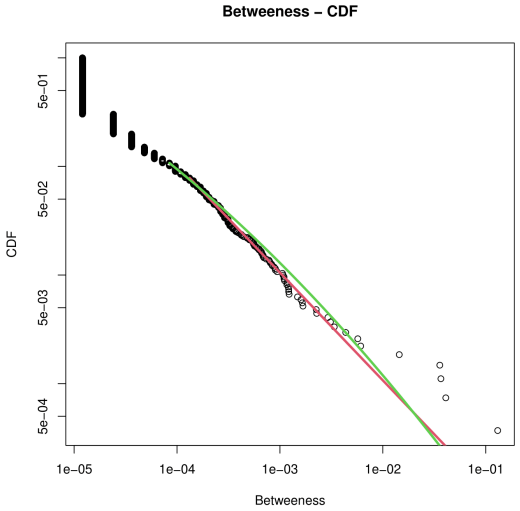
(m) Grafo steemit



(n) Grafo social



(o) Grafo subsocial



(p) Grafo monetario

```
The decimal point is 2 digit(s) to the left of the |

0 | 00000000000000000000000000000000000000000000000000000000000000000000000000+1236808
0 | 55555555555555555555555555555555555555555555555555555555555555555555555+6350
1 | 00000000000000000000000000000000000000000000000000000000000000000000000000+1005
1 | 55555555555555555555555555555555555555555555555555555555555555555555555+188
2 | 00000000000000000000000000000000011111111111111111111111122222222222+44
2 | 55555555555666666777777788888888999999
3 | 000000111111122223333444444
3 | 56666777888
4 | 01114
4 | 5556679
5 | 3
5 | 68
6 | 3
```

(q) Grafo steemit

```
The decimal point is 2 digit(s) to the left of the |

0 | 00000000000000000000000000000000000000000000000000000000000000000000000000+1222461
0 | 55555555555555555555555555555555555555555555555555555555555555555555555+6366
1 | 00000000000000000000000000000000000000000000000000000000000000000000000000+1017
1 | 55555555555555555555555555555555555555555555555555555555555555555555555+197
2 | 00000000000000000000000000011111111111111111111111111122222222222222+36
2 | 55555566666666666777778888889999999
3 | 000000000011122222233334444
3 | 5556667788889
4 | 002234
4 | 5568
5 | 04
5 | 9
6 | 4
```

(r) Grafo social

```
The decimal point is 2 digit(s) to the left of the |

0 | 00000000000000000000000000000000000000000000000000000000000000000000000000+1222463
0 | 55555555555555555555555555555555555555555555555555555555555555555555555+6361
1 | 00000000000000000000000000000000000000000000000000000000000000000000000000+1025
1 | 55555555555555555555555555555555555555555555555555555555555555555555555+195
2 | 00000000000000000000000000001111111111111111111111111122222222222222+33
2 | 55555556666666666677777888899999999
3 | 000000000111122222333334444
3 | 55556677788999
4 | 012234
4 | 5669
5 | 0
5 |
6 | 04
```

(s) Grafo subsocial

```
The decimal point is 2 digit(s) to the left of the |

 0 | 00000000000000000000000000000000000000000000000000000000000000000000000000+1074526
 2 | 000001122233333445555666678888889999000011223344555667777788
 4 | 000111222446678223334777
 6 | 0224891123
 8 | 045588368
10 | 011369
12 | 1259
14 | 22
16 |
18 |
20 |
22 | 0
24 |
26 | 7
```

(t) Grafo monetario

Descriviamo i grafici stem-and-leaf calcolati, notando che:

- Nelle distribuzioni si nota il valore 0 come valore più frequente, comune a tutti e quattro i grafi. Conseguenza di ciò, confermiamo la presenza di un unico picco rappresentato dalla classe di valore 0;
- Le distribuzioni seguono una forma asimmetrica positiva, con una curva leptocurtica ma meno pronunciata rispetto alle precedenti centralità visualizzate;
- Nei grafi steemit, social e subsocial i valori anomali non si distaccano troppo dai valori più frequenti. Nel grafo monetario, invece, si notano due outliers che si distaccano dai valori dei restanti nodi;
- Il grafo steemit ha come massimo il valore $0,063$, il grafo social ha come massimo il valore $0,064$, il grafo subsocial ha come massimo il valore $0,064$ e il grafo monetario ha come massimo il valore $0,267$. Da tali valori osserviamo che il massimo della centralità eigenvector nel grafo monetario risulta maggiore rispetto al massimo della centralità eigenvector nei rimanenti tre grafi. Tale risultato, proviene dalla conseguenza per cui pochi nodi posseggono elevata centralità e tutti i rimanenti nodi tentano di collegarsi a questi pochi nodi. Poiché tale gruppo di nodi centrali sono pochi questi ricevono più centralità eigenvector quando vengono riferiti; mentre dalla parte dei nodi meno centrali, l'unico modo per acquisire centralità è collegandosi a questi pochi nodi che risultano più visibili.

Successivamente alla visione grafica della distribuzione, elenchiamo i principali parametri statistici i quali riassumono analiticamente le distribuzioni in alcuni suoi aspetti, fornendo informazioni più precise (Tabella 6.4.

(a) Grafo steemit

Parametro	**Valore**
Minimo	0
Q_1	$2,33e-05$
Mediana	$5,56e-05$
Media	$2,87e-04$
Q_3	$2,36e-04$
Massimo	$6,25e-02$
Asimmetria	$13,09$
Curtosi	$360,21$
Dev. Standard	$0,00085$
Coeff. Variazione	$2,96$

(b) Grafo social

Parametro	**Valore**
Minimo	0
Q_1	$2,34e-05$
Mediana	$5,67e-05$
Media	$2,92e-04$
Q_3	$2,43e-04$
Massimo	$6,35e-02$
Asimmetria	$12,78$
Curtosi	$341,71$
Dev. Standard	$0,00085$
Coeff. Variazione	$2,92$

(c) Grafo subsocial

Parametro	**Valore**
Minimo	0
Q_1	$2,37e-05$
Mediana	$5,74e-05$
Media	$2,93e-04$
Q_3	$2,45e-04$
Massimo	$6,40e-02$
Asimmetria	$12,68$
Curtosi	$335,17$
Dev. Standard	$0,00085$
Coeff. Variazione	$2,91$

(d) Grafo monetario

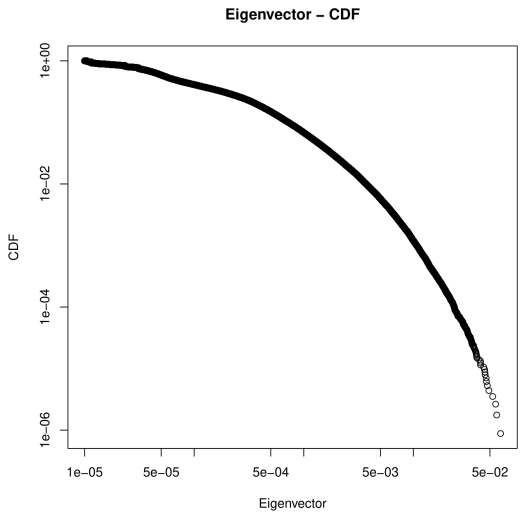
Parametro	**Valore**
Minimo	0
Q_1	$1,08e-05$
Mediana	$1,08e-05$
Media	$1,79e-04$
Q_3	$1,69e-05$
Massimo	$2,77e-01$
Asimmetria	$76,84$
Curtosi	13.961
Dev. Standard	$0,00095$
Coeff. Variazione	$5,28$

Tabella 6.4: Parametri statistici per la centralità eigenvector.

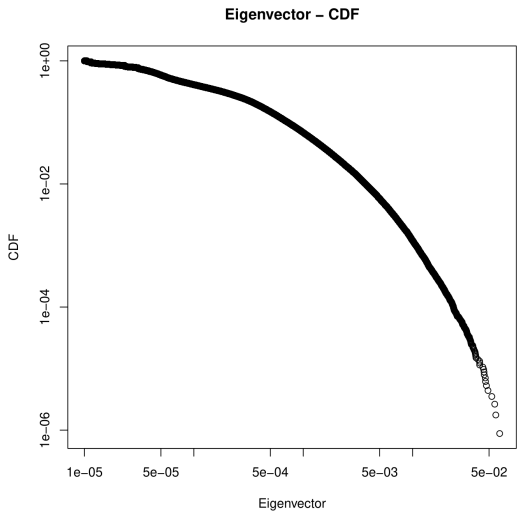
Dall'analisi tabellare è possibile confermare i seguenti concetti:

- Il minimo in tutti i grafi ha valore 0. Il massimo mostrato nell'analisi tabellare coincide con il massimo mostrato nel grafico stem-and-leaf. Di conseguenza, confermiamo la sostanziale differenza tra i grafi steemit, social e subsocial rispetto al grafo monetario;
- La mediana per i grafi steemit, social e subsocial assumono valori quasi identici. Stessa situazione per il terzo quartile. La mediana e il terzo quartile del grafo monetario risultano più bassi rispetto ai restanti alla mediana e terzo quantile dei rimanenti tre grafi. Tale risultato porta alla conclusione che, in linea generale, i nodi nei grafi steemit, social e subsocial tendono ad avere una più alta centralità rispetto al grafo monetario. La media conferma tale risultato;
- L'indice di simmetria e curtosi riflettono il risultato mostrato dai grafici stem-and-leaf, con un maggior indice di curtosi per il grafo monetario rispetto ai grafi steemit, social e subsocial;
- Il coefficiente di variazione coincide approssimativamente per i grafi steemit, social e subsocial. Per il grafo monetario, tale coefficiente risulta più alto dovuto ai valori di deviazione standard e di media più bassi.

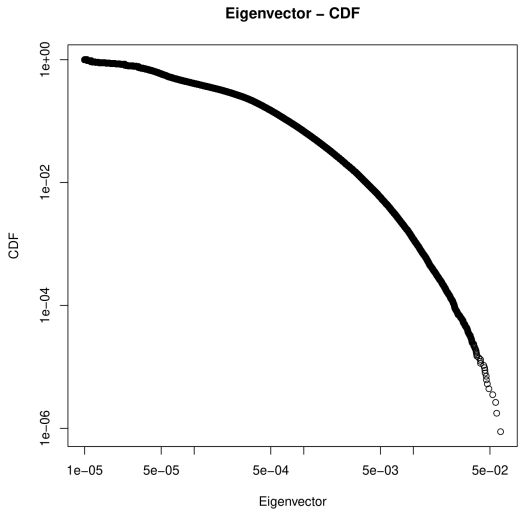
Terminata l'analisi tabellare, introduciamo l'analisi della funzione di distribuzione cumulativa. Come già detto, l'ipotesi delle centralità strettamente maggiore di 0 viene ulteriormente ristretta eliminando alcuni valori troppo bassi per visualizzare in maniera corretta la CDF (Figura 6.3). Con l'imposizione di tale vincolo, abbiamo eliminato dal calcolo del grafico all'incirca 200 nodi che avrebbero portato a un aumento dei valori considerati nel grafico, creando un'inutile espansione dell'intervallo considerato (un possibile esempio di centralità bassa è $1e-140$).



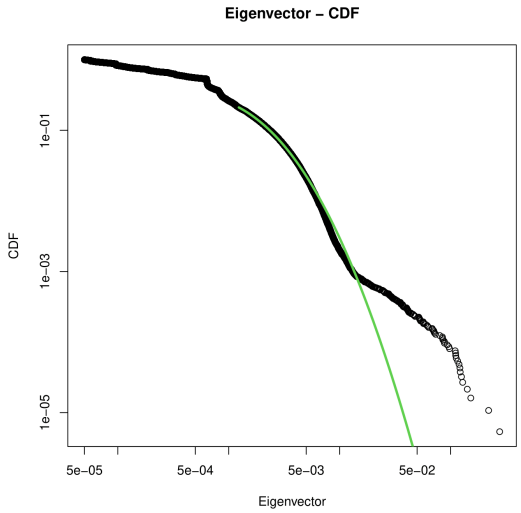
(u) Grafo steemit



(v) Grafo social



(w) Grafo subsocial



(x) Grafo monetario

Figura 6.3: CDF delle centralità eigenvector.

La CDF per le centralità eigenvector nei grafi steemit, social e subsocial si presentano con un'andatura curva, piuttosto simile tra loro. Data la complessità in tempo dovuta al calcolo di un corretto fitting secondo le distribuzioni log-normale e power law, mostriamo i grafici senza alcuna retta o curva riassuntiva. Notando l'andamento, concludiamo che nella parte finale della coda è presente una possibile legge di potenza; ma, in generale, non possiamo essere certi che la centralità segua una legge nota. Nel grafo monetario la situazione è parecchio diversa e

complicata. La curva presenta parecchie anomalie nel suo andamento. Effettuando il fitting secondo la distribuzione log-normale (curva verde), notiamo che nei valori intermedi i dati seguono tale distribuzione.

6.2.5 Comparazioni: PageRank

Esaminiamo una centralità derivata dal calcolo dell'algoritmo pagerank. Tale algoritmo lo si associa per le sue caratteristiche alla centralità eigenvector, poiché entrambe puntano ad aumentare il valore della metrica sul nodo considerando a sua volta la centralità dei suoi collegamenti diretti o indiretti all'interno del grafo. Nelle distribuzioni di frequenza, data la natura della centralità pagerank, dovremmo notare una classificazione più ristretta dei nodi rispetto alla classificazione attuata con l'algoritmo eigenvector.

Illustriamo nella pagina seguente i quattro grafici stem-and-leaf riferiti alle distribuzioni di frequenze della centralità pagerank. I stem-and-leaf vengono costruiti con scale diverse per una corretta visualizzazione dei parametri più importanti.

```
The decimal point is 4 digit(s) to the left of the |

 0 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+1244419
 1 | 00000000000000000000000000000000000000000000000000000000000000000000000011111111+174
 2 | 00000000001111111111111122222233333444444455555555666777777888888899
 3 | 00000000111111222233333344456678889
 4 | 0333455588
 5 | 01478
 6 | 0388
 7 | 79
 8 | 348
 9 | 39
10 | 9
11 |
12 |
13 | 6
14 | 35
15 | 9
16 |
17 |
18 |
19 |
20 |
21 |
22 | 6
```

(a) Grafo steemit

```
The decimal point is 4 digit(s) to the left of the |

 0 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+1230356
 2 | 00000000111111112222222233333333333444444455556666777888888889999900+19
 4 | 00111224477799902236
 6 | 4455
 8 | 24
10 | 13
12 | 2
14 | 7
16 | 15
18 |
20 |
22 |
24 | 5
```

(b) Grafo social

```
The decimal point is 4 digit(s) to the left of the |

 0 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+1230357
 2 | 00000001111111222222222233333334444444455555666667778888889999000l+18
 4 | 001112334588890012347
 6 | 4567
 8 | 35
10 | 4
12 | 2
14 | 9
16 | 37
18 |
20 |
22 |
24 | 7
```

(c) Grafo subsocial

```
The decimal point is 3 digit(s) to the left of the |

 0 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+1074608
 1 | 012334445555666788889
 2 | 177
 3 | 35
 4 | 2449
 5 |
 6 |
 7 | 1
 8 |
 9 |
10 | 0
11 | 2
12 | 8
```

(d) Grafo monetario

Dal momento che trattiamo valori di centralità continui, consideriamo la giusta semantica da utilizzare per interpretare i grafici stem-and-leaf. Commentiamo i grafici, considerando tutti i parametri visibili ed eventuali confronti tra loro:

- I valori più frequenti nei quattro grafici risultano essere i valori uguali a 0. Congruentemente al resto delle centralità, notiamo che l'unico picco notato nei quattro grafici è la classe con valore 0 che contiene la maggior parte delle osservazioni;

- Le distribuzioni hanno una forma asimmetrica positiva, caratterizzate da una curva leptocurtica rispetto alla normale. Inoltre, abbiamo notato una possibile relazione tra i 4 grafici: dal grafo col numero di nodi e archi maggiore fino ad arrivare al grafo col numero di nodi e archi minori, quindi in questo ordine: grafo steemit, social, subsocial e monetario si intravede uno scemarsi della curva delle distribuzioni. Infatti, si intravede una curva molto più piatta per il grafo steemit, una curva meno piatta per il grafo social, una curva più appuntita per il grafo subsocial e in ultimo una curva molto appuntita per il grafo monetario;

- Nei diversi grafici si nota una discreta presenza di valori anomali nelle classi di valore maggiore rispetto le classi con maggiore frequenza. Osserviamo che i grafici riguardanti i grafi social e subsocial mantengono una coda destra parzialmente continua nei suoi valori, con l'eccezione per il valore massimo che si trova distaccato rispetto agli outliers di valore minore. I grafici riguardanti i grafi steemit e monetario sono caratterizzati da una coda destra più discontinua, avendo dei valori anomali molto distaccati tra loro;

- Il massimo del grafo steemit si distacca dai restanti nodi e assume valore $0,00226$, il massimo del grafo social assume valore $0,00245$, il massimo del grafo subsocial assume valore $0,00247$ e infine, il massimo del grafo monetario assume valore $0,0128$. Esattamente come per la centralità eigenvector, notiamo che il massimo tra i quattro grafi risulta essere il massimo del grafo monetario. Tale risultato deriva dalla caratteristica intrinseca della centralità pagerank per il quale essendo pochi i nodi centrali, i rimanenti nodi tentano di collegarsi ai nodi più "prestigiosi" aumentando ancor di più la centralità di quest'ultimi.

Introduciamo alla visione grafica delle distribuzioni, una visione numerica sintetica dei diversi grafici tramite i parametri statistici più importanti. (Tabella 6.5).

(a) Grafo steemit

Parametro	**Valore**
Minimo	$1,84e-07$
Q_1	$1,87e-07$
Mediana	$2,22e-07$
Media	$8,03e-07$
Q_3	$4,15e-07$
Massimo	$2,26e-03$
Asimmetria	$128,54$
Curtosi	30.808
Dev. Standard	$6,07e-06$
Coeff. Variazione	$7,56$

(b) Grafo social

Parametro	**Valore**
Minimo	$2e-07$
Q_1	$2,02e-07$
Mediana	$2,09e-07$
Media	$8,13e-07$
Q_3	$3,76e-07$
Massimo	$2,55e-03$
Asimmetria	$140,86$
Curtosi	39.038
Dev. Standard	$6,24e-06$
Coeff. Variazione	$7,68$

(c) Grafo subsocial

Parametro	**Valore**
Minimo	$2e-07$
Q_1	$2,03e-07$
Mediana	$2,09e-07$
Media	$8,13e-07$
Q_3	$3,77e-07$
Massimo	$2,57e-03$
Asimmetria	$143,20$
Curtosi	40.713
Dev. Standard	$6,21e-06$
Coeff. Variazione	$7,64$

(d) Grafo monetario

Parametro	**Valore**
Minimo	$4,84e-07$
Q_1	$4,84e-07$
Mediana	$4,84e-07$
Media	$9,3e-07$
Q_3	$4,86e-07$
Massimo	$1,28e-02$
Asimmetria	$347,81$
Curtosi	149.906
Dev. Standard	$2,44e-05$
Coeff. Variazione	$26,24$

Tabella 6.5: Parametri statistici della centralità page rank.

Dai valori statistici possiamo trarre le seguenti conclusioni:

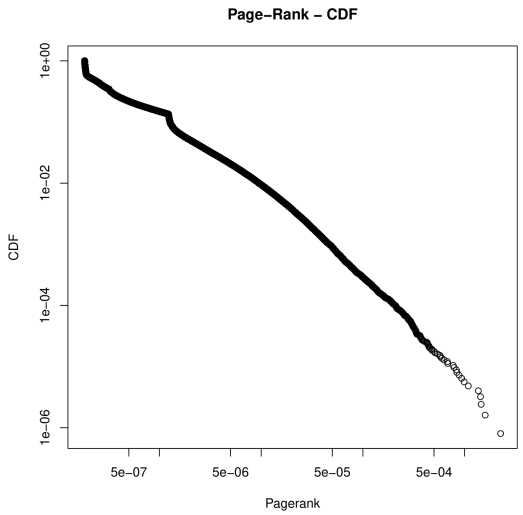
- I minimi delle centralità pagerank risultano maggiori di 0, differentemente dalla centralità eigenvector. Il minimo dei grafi steemit, social e subsocial approssimativamente coincide. Il minimo del grafo monetario risulta maggiore rispetto ai rimanenti grafi. Identico ragionamento può essere attuato al massimo della centralità: il massimo dei grafi steemit, social e subsocial risulta approssimativamente uguale. Il massimo del grafo monetario assume un valore maggiore rispetto ai restanti tre grafi.

- La mediana e il terzo quartile dei grafi steemit, social e subsocial hanno approssimativamente lo stesso valore. La mediana e il terzo quartile del grafo monetario risultano avere un valore maggiore rispetto ai rimanenti tre grafi;

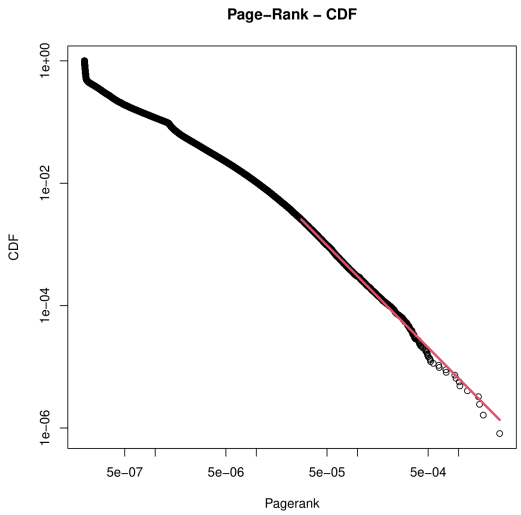
- Gli indici di asimmetria e gli indici di curtosi confermano il risultato analizzato per cui la forma della distribuzione, coerentemente con la grandezza del grafo, va a scemarsi e di conseguenza i valori dei due indici ad aumentare;

- Il coefficiente di variazione per il grafo steemit, social e subsocial contengono un valore basso rispetto al coefficiente di variazione per il grafo monetario. Tale risultato è dovuto dalla deviazione standard e dalla media che nel grafo monetario risultano maggiori rispetto ai rimanenti tre grafi.

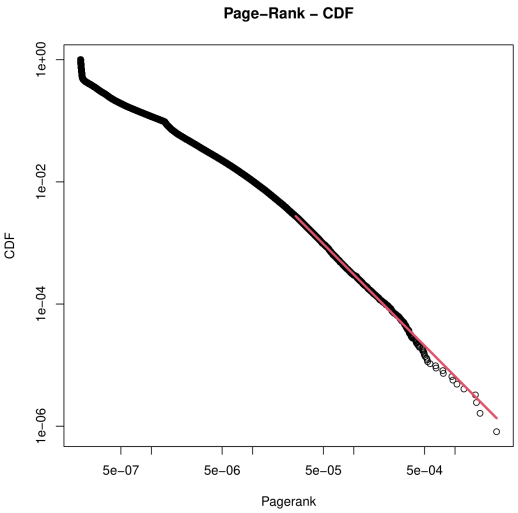
Caratterizzati i grafi rispetto la centralità pagerank secondo una visione grafica e una analitica, tentiamo di classificare la distribuzione secondo una qualche legge di distribuzione di probabilità già nota. Rappresentiamo le quattro distribuzioni secondo la funzione di ripartizione associata.



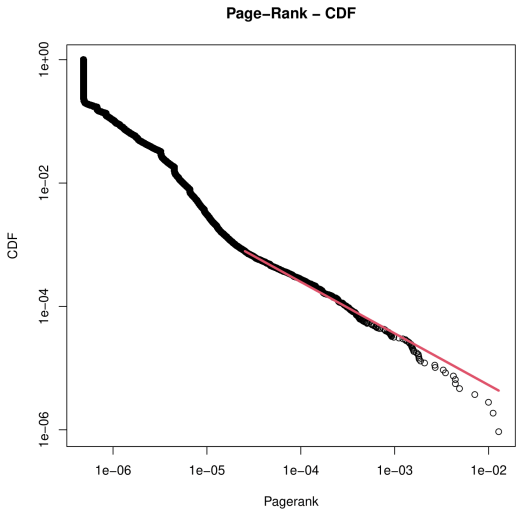
(e) Grafo steemit



(f) Grafo social



(g) Grafo subsocial



(h) Grafo monetario

I grafici della funzione di ripartizione per i grafi steemit, social e subsocial risultano indicativamente identici. Il loro andamento si osserva essere lineare con la stessa pendenza. Effettuando il fitting rispetto la power law (retta viola) notiamo che a partire da un valore x_{min} intermedio, la distribuzione segue nella sua coda tale legge. Per il grafo monetario la situazione è differente: l'andamento risulta più frastagliato anche se nella coda destra tende ad assumere un andamento lineare. Anche in questo caso, la power law tenta di riassumere la coda destra della distribuzione, non riuscendo nella parte finale.

6.3 Descrizione inferenziale: Analisi delle ipotesi e risultati

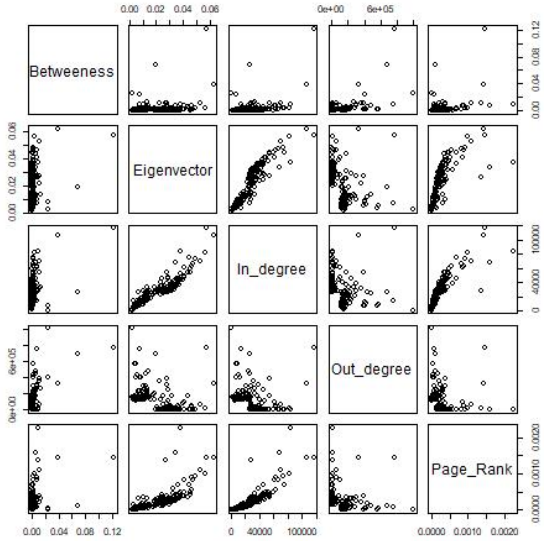
Il capitolo seguente descrive la fase finale dell'analisi oltre al trarre delle conclusioni su tutta la ricerca. Proseguendo per gradi, descriviamo l'attuazione dei test d'inferenza attenendoci alla scala delle ipotesi, presentate nella sezione del disegno sperimentale.

La prima ipotesi a cui dobbiamo trovare verifica consiste nel capire se siano presenti o meno determinate correlazioni rispetto le diverse misure di centralità. Per correlazioni intendiamo che: all'aumentare di una misura X anche la misura Y aumenta o diminuisce a sua volta per un'osservazione riguardante lo stesso campione. Nel caso particolare in cui questa situazione dovesse presentarsi per un numero non banale di osservazioni nel campione, allora si sospetta una possibile correlazione positiva o negativa in base alla situazione. In modo da ottenere un risultato di questo genere, utilizziamo i test non parametrici di Spearman e Kendall sulle cinque centralità prese in considerazione considerate a coppie, presentate in Sezione 3.2.1.

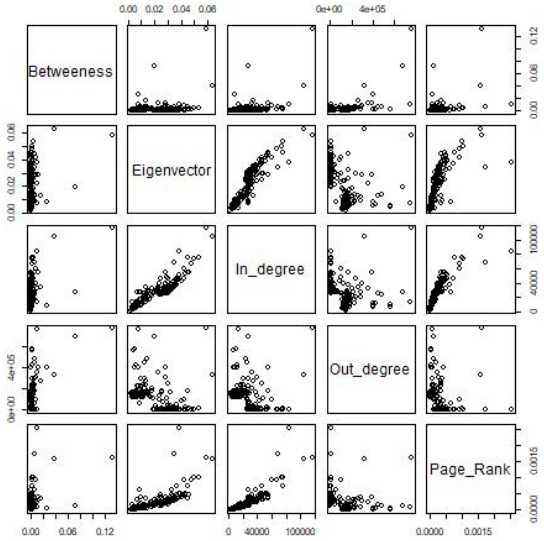
Innanzitutto, per poter considerare valido il risultato di un test non parametro di correlazione, abbiamo visualizzato i diagrammi di dispersione calcolati tramite il pacchetto di base del software R su tutte le centralità considerate. Quindi, in modo tale da poter confermare le correlazioni in modo più certo, abbiamo calcolato i test e costruito i diagrammi di dispersione per tutti e quattro i grafi. Il concetto da considerare fondamentale quando si calcolato i coefficienti di Spearman e Kendall, ma anche per Pearson, è il controllo della correlazione grafica in modo da visualizzare se davvero è presente una qualsiasi funzione monotòna crescente/decrescente o addirittura lineare. Nel caso in cui notiamo un diagramma di dispersione abbastanza sparso, allora è inutile controllare l'indice di correlazione associato.

Durante la sperimentazione, abbiamo costruito tutti i possibili diagrammi di dispersione per tutti i grafi e tutti i corrispondenti indici di correlazione in modo da comunque verificare tutti i possibili risultati. L'ipotesi che abbiamo deciso di utilizzare per la corretta costruzione dei diagrammi di dispersione e del calcolo degli indici di correlazione è l'eliminazione di quasi tutti i nodi meno centrali, selezionando in modo accurato 200 nodi per ogni grafo considerato. Per effettuate tale selezione, abbiamo pensato di scrivere un programma che selezionasse il numero dei 100 nodi più alti secondo la centralità per grado in entrata e 100 nodi più alti secondo

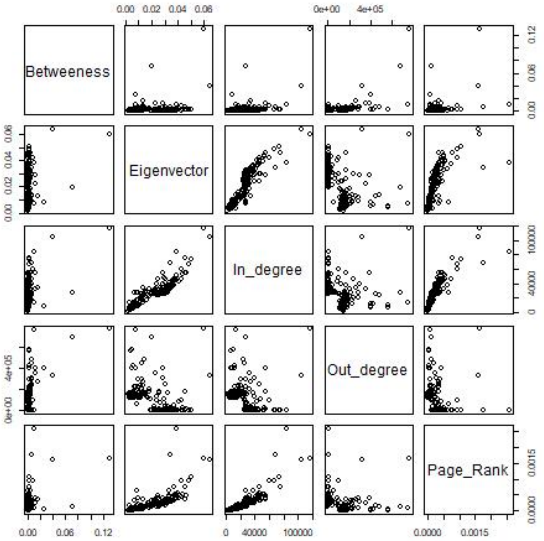
la centralità per grado in uscita. Creata la lista dei 200 nodi, per ogni grafo, creiamo le liste contenenti i valori di centralità riferiti a tale gruppo costruendo un campione appaiato più ridotto rispetto al campione iniziale.



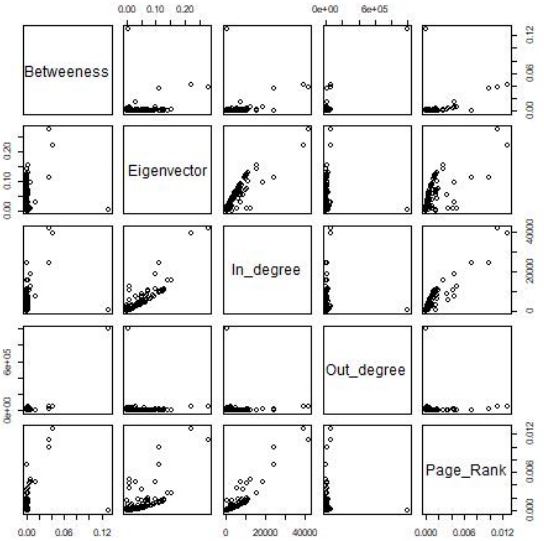
(i) Grafo steemit



(j) Grafo social



(k) Grafo subsocial



(l) Grafo monetario

(a) Grafo steemit

	Betweeness	**Eigenvector**	**Indegree**	**Outdegree**	**Pagerank**
Betweeness	1	$0,47$	$0,53$	$0,29$	$0,58$
Eigenvector	$0,47$	1	$0,94$	$-0,53$	$0,92$
Indegree	$0,53$	$0,94$	1	$-0,47$	$0,96$
Outdegree	$0,29$	$-0,53$	$-0,47$	1	$-0,49$
Pagerank	$0,58$	$0,92$	$0,96$	$-0,49$	1

(b) Grafo social

	Betweeness	**Eigenvector**	**Indegree**	**Outdegree**	**Pagerank**
Betweeness	1	$0,52$	$0,58$	$0,27$	$0,62$
Eigenvector	$0,52$	1	$0,94$	$-0,50$	$0,91$
Indegree	$0,58$	$0,94$	1	$-0,45$	$0,96$
Outdegree	$0,27$	$-0,50$	$-0,45$	1	$-0,46$
Pagerank	$0,62$	$0,91$	$0,96$	$-0,46$	1

(c) Grafo sub-social

	Betweeness	**Eigenvector**	**Indegree**	**Outdegree**	**Pagerank**
Betweeness	1	$0,51$	$0,57$	$0,29$	$0,61$
Eigenvector	$0,51$	1	$0,94$	$-0,50$	$0,91$
Indegree	$0,57$	$0,94$	1	$-0,43$	$0,95$
Outdegree	$0,29$	$-0,50$	$-0,43$	1	$-0,46$
Pagerank	$0,61$	$0,91$	$0,95$	$-0,46$	1

(d) Grafo monetario

	Betweeness	**Eigenvector**	**Indegree**	**Outdegree**	**Pagerank**
Betweeness	1	$0,32$	$0,41$	$0,30$	$0,50$
Eigenvector	$0,32$	1	$0,92$	$-0,31$	$0,84$
Indegree	$0,41$	$0,92$	1	$-0,35$	$0,95$
Outdegree	$0,30$	$-0,31$	$0,35$	1	$-0,34$
Pagerank	$0,50$	$0,84$	$0,95$	$-0,34$	1

Tabella 6.6: Tabelle per l'indice di correlazione di Spearman.

Nella tabella 6.6 è possibile notare una tabella a doppia entrata per ogni grafo, a ogni riga e colonna sono assegnate le stesse centralità formando una matrice in cui nelle celle sono contenuti i valori risultanti dalla correlazione di Spearman. I valori sono compresi tra -1 e $+1$, nella diagonale della tabella sono presenti valori uguali a $+1$ poiché si tratta della correlazione sulla stessa centralità.

Un valore nella cella può essere considerato consistente nel caso in cui il diagramma di dispersione realizzato per mezzo delle centralità produce una funzione monotòna o lineare crescente nel caso di un indice positivo e monotòna o lineare decrescente nel caso di un indice negativo. Visualizzando i valori nelle quattro tabelle, notiamo una certa somiglianza tra di esse, l'unica differenza sostanziale che si intravede tra le celle è rilevata nel grafo monetario rispetto ai grafi steemit, social e subsocial. Tralasciando le motivazioni per cui nelle correlazioni del grafo monetario i valori risultino diversi rispetto gli altri grafi, prenderemo in considerazione la tabella di correlazione del grafo steemit. In tale tabella notiamo che le uniche correlazioni certe (maggiore di $0,90$) sono:

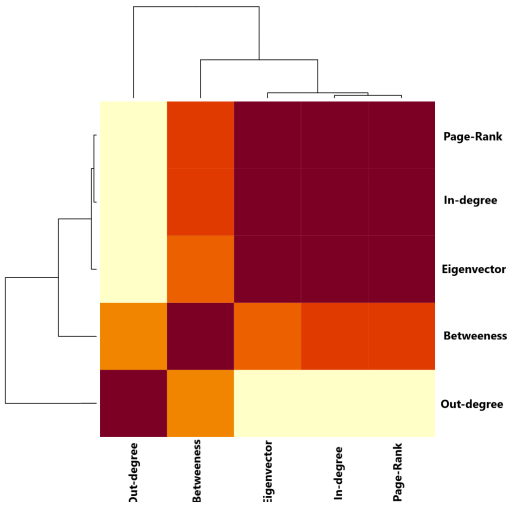
- Centralità eigenvector correlata positivamente con la centralità in-degree;
- Centralità eigenvector correlata positivamente con la centralità pagerank; item Centralità in-degree correlata positivamente con la centralità pagerank.

Inoltre, se teniamo conto dei loro diagrammi di dispersione, notiamo che:

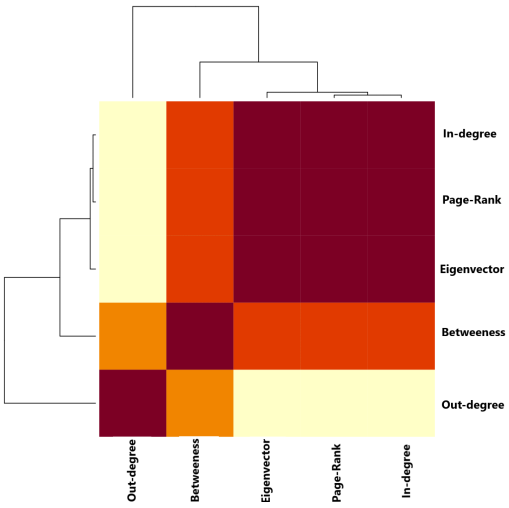
- Diagramma riferito alla correlazione eigenvector/in-degree risulta approssimativamente una funzione lineare crescente;
- Diagramma riferito alla correlazione eigenvector/pagerank risulta approssimativamente una funzione monotòna crescente;
- Diagramma riferito alla correlazione in-degree/pagerank risulta approssimativamente una funzione monotòna crescente.

Allo stesso modo per cui le tabelle dei diversi grafi contenente gli indici di Spearman si somigliano tra loro, anche i relativi diagrammi di dispersione dei diversi grafi presentano la stessa caratteristica. Per poter visualizzare gli indici di Spearman rapidamente per tutte le centralità,

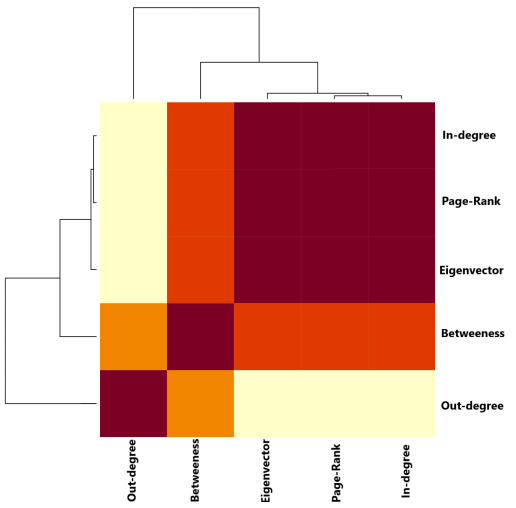
invece di affidarci a rappresentazioni numeriche possiamo affidarci a delle mappe di calore. Nelle mappe di calori abbiamo utilizzato colori più chiari per indici di correlazione negativi e bassi e colori più scuri per indici di correlazioni alti e positivi. Mostriamo, di seguito, le quattro heatmap riferite alle quattro tabelle di Spearman in modo da essere coerenti con il resto dell'esposizione. In generale, tali mappe risulteranno identiche tra loro rispetto il grafo steemit, social e subsocial; si presenterà differente nel caso del grafo monetario (Tabella 6.4).



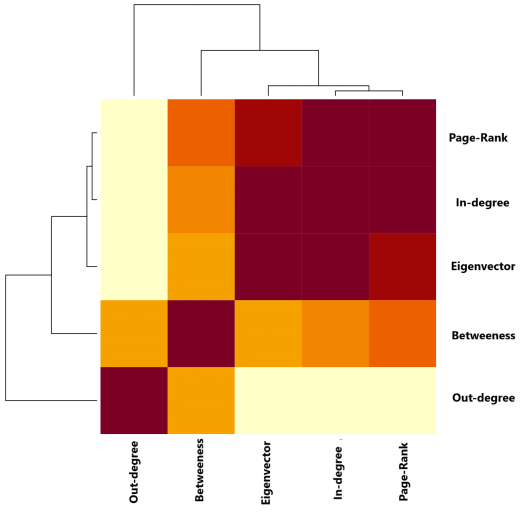
(m) Grafo steemit



(n) Grafo social



(o) Grafo subsocial



(p) Grafo monetario

Figura 6.4: Heatmap degli indici di spearman.

Un ulteriore correlazione già nominata in precedenza, e che utilizza la stessa metrica di misura della correlazione di Spearman è la correlazione di Kendall. L'algoritmo per il calcolo di quest'ultimo è differente dalla precedente correlazione, quindi risulta utile per confrontare effettivamente i valori e notare un riscontro tra le diverse mappe di calore. Il procedimento attuato per Spearman viene riportato interamente anche per l'indice di Kendall. Per non essere ripetitivi non verranno illustrati i diagrammi di dispersione poiché identici a quelli visti per l'indice di Spearman. Le tabelle a doppia entrata, invece, saranno differenti (Tabella 6.7).

(a) Grafo steemit

	Betweeness	**Eigenvector**	**Indegree**	**Outdegree**	**Pagerank**
Betweeness	1	$0,34$	$0,38$	$0,21$	$0,42$
Eigenvector	$0,34$	1	$0,81$	$-0,31$	$0,77$
Indegree	$0,38$	$0,81$	1	$-0,28$	$0,85$
Outdegree	$0,21$	$-0,31$	$-0,28$	1	$-0,29$
Pagerank	$0,42$	$0,77$	$0,85$	$-0,29$	1

(b) Grafo social

	Betweeness	**Eigenvector**	**Indegree**	**Outdegree**	**Pagerank**
Betweeness	1	$0,37$	$0,42$	$0,19$	$0,45$
Eigenvector	$0,37$	1	$0,80$	$-0,29$	$0,76$
Indegree	$0,41$	$0,80$	1	$-0,26$	$0,84$
Outdegree	$0,19$	$-0,29$	$-0,26$	1	$-0,27$
Pagerank	$0,45$	$0,76$	$0,84$	$-0,27$	1

(c) Grafo sub-social

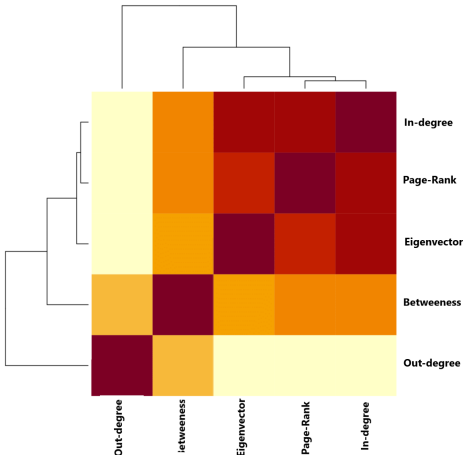
	Betweeness	**Eigenvector**	**Indegree**	**Outdegree**	**Pagerank**
Betweeness	1	$0,36$	$0,40$	$0,20$	$0,44$
Eigenvector	$0,36$	1	$0,80$	$-0,29$	$0,76$
Indegree	$0,40$	$0,80$	1	$-0,24$	$0,83$
Outdegree	$0,20$	$-0,29$	$-0,24$	1	$-0,26$
Pagerank	$0,44$	$0,76$	$0,83$	$-0,26$	1

(d) Grafo monetario

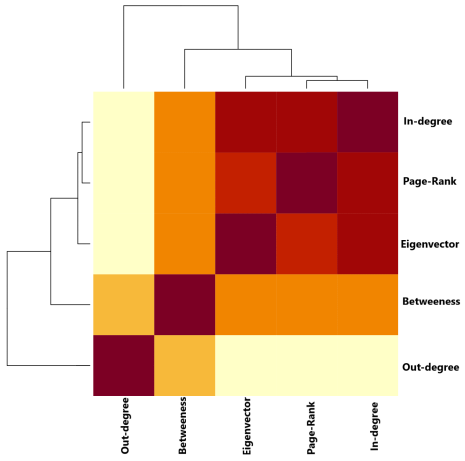
	Betweeness	**Eigenvector**	**Indegree**	**Outdegree**	**Pagerank**
Betweeness	1	$0,23$	$0,29$	$0,22$	$0,37$
Eigenvector	$0,23$	1	$0,82$	$-0,18$	$0,73$
Indegree	$0,29$	$0,82$	1	$-0,20$	$0,88$
Outdegree	$0,22$	$-0,18$	$-0,20$	1	$-0,20$
Pagerank	$0,37$	$0,73$	$0,88$	$-0,20$	1

Tabella 6.7: Tabelle per l'indice di correlazione di Kendall.

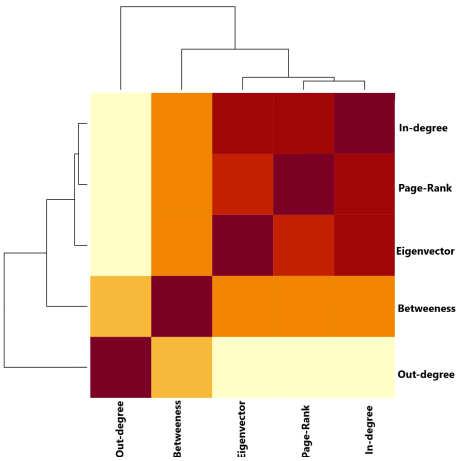
Da come notiamo dall'ultima tabella contenente le correlazioni di Kendall, tutti gli indici perdono un fattore di circa $0,15$ rispetto agli indici calcolati con la correlazione di Spearman. Quindi, vale che la correlazione di Kendall è più restringente.



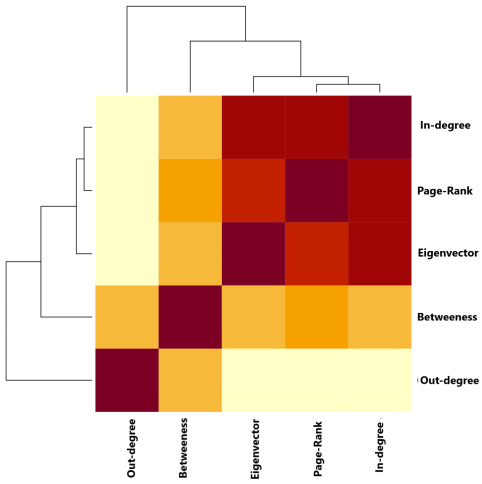
(a) Grafo steemit



(b) Grafo social



(c) Grafo subsocial



(d) Grafo monetario

Le correlazioni di Kendall tramite le Heatmap si mostrano coincidenti, in proporzione, alle correlazioni di Spearman nelle differenze tra i grafi steemit, social e subsocial rispetto al grafo monetario. Quindi, abbiamo confermato dai diagrammi di dispersione e dai risultati delle statistiche test le seguenti correlazioni:

- Centralità eigenvector correlata positivamente con la centralità in-degree;
- Centralità eigenvector correlata positivamente con la centralità pagerank;

- Centralità in-degree correlata positivamente con la centralità pagerank.

Con un $mboxp - value < 0,05 = \alpha$.

Interpretando i risultati possiamo dedurre che la correlazione per la centralità eigenvector e la centralità in-degree è spiegata dall'implicazione ovvia che più utenti riferiscono un particolare utente e più vi è la probabilità che siano presenti, tra gli utenti che riferiscono, utenti con alto prestigio (importanza) e quindi più centrali a sua volta all'interno della rete. Stessa interpretazione può essere attuata per la centralità in-degree e la centralità pagerank. Infine, la correlazione tra la centralità eigenvector e la centralità pagerank è spiegata dalla loro natura, ripresa più volte all'interno di questa tesi, che entrambe le misure caratterizzano gli utenti con più alto prestigio nella rete.

L'analisi inferenziale prosegue con un'ulteriore ipotesi da verificare riguardante la comparazione tra le diverse centralità. L'ipotesi da verificare è la seguente: nell'ipotesi di avere lo stesso campione per tutte le centralità analizzare, qual è all'interno della popolazione (popolazione e non campione, differenza sostanziale) la mediana più alta tra le diverse centralità. Quindi, quale centralità risulta essere più alta comunemente tra gli utenti all'interno del social media. Per poter verificare tale ipotesi, utilizziamo il test di Friedman su cinque campioni e a seguire il test di Wilcoxon per dati appaiati su due centralità.

Le statistiche test appena citate, solitamente, vengono utilizzate su campioni che posseggono lo stesso tipo di misurazione. Tale caratteristica è dovuta dal modo in cui vengono calcolati i risultati del test poiché sfruttano le relazioni tra misure diverse (solitamente, viene utilizzato per i test del tipo: misurazione giorno prima comparato a misurazione giorno dopo e notare le eventuali differenze). Noi abbiamo modificato tale comportamento standard. Le misure di centralità hanno la maggior parte unità di misura differenti, quindi per poter attuare i test di Friedman e di Wilcoxon portiamo tutte le metriche a una stessa unità di misura. Per effettuare questo cambiamento, abbiamo utilizzato il processo di standardizzazione su tutte le centralità. Attraverso la standardizzazione delle variabili è possibile rendere confrontabili variabili identiche appartenenti a distribuzioni diverse e variabili diverse, oppure variabili espresse in unità di misura differenti. La standardizzazione consiste in una doppia normalizzazione, dove ogni dato è trasformato nel suo scarto dalla media della distribuzione, dopodiché il dato viene trasformato dall'unità di misura di quella variabile in unità del suo scarto tipo (deviazione standard), ovvero

viene normalizzato per il miglior coefficiente sintetico. Tale operazione consiste nel ricondurre una variabile aleatoria con media μ e varianza σ a una variabile con distribuzione "standard" Z, ossia di media 0 e varianza 1.

Abbiamo creato un programma in Python che eseguisse tale trasformazione sulle centralità dei 200 nodi selezionati in precedenza, lo stesso campione utilizzato per i test di correlazione. Una volta costruito cinque liste, per ogni grafo, con i valori di centralità standardizzati possiamo attuare il test di Friedman.

Nel grafo Steemit, il risultato del test è dato da due valori fondamentali:

$$Fr = 8 \text{ con } p - value = 0,09$$

Dalla documentazione, sappiamo che a causa del presupposto che la statistica test abbia una distribuzione chi-quadro, il p-value è affidabile solo per $n > 10$ e per più di 6 misurazioni ripetute [2]. Nei nostri quattro grafi, le centralità a cui stiamo attuando la statistica test sono cinque il che non rientrano tra le ipotesi per poter considerare il p-value. Quindi, consideriamo solamente il valore della statistica di Friedman.

Dato che i valori critici per la statistica di Friedman per l'indice di significatività uguale a $0,05$ e 4 gradi di libertà (calcolato dall'espressione $gdl = k - 1$), è uguale a $\chi^2_F = 9,5$. La statistica chi-quadro è maggiore della statistica test, allora in teoria, accettiamo l'ipotesi nulla per il quale tutte le mediane delle cinque centralità sono uguali.

Dal risultato ottenuto dal test di Friedman, per cui le cinque mediane risultano approssimativamente uguali, ci inducono a non utilizzare il test di Wilcoxon sulle centralità. Per completezza, nel caso di qualche errore commesso dal test di Friedman, impieghiamo comunque il test di Wilcoxon nelle centralità a coppie. Il risultato che contraddice l'ipotesi nulla si nota per la centralità betweeness e la centralità in-degree con i seguenti valori del test:

$$W = 11.434 \text{ con } p - value = 0,04$$

Poiché notiamo un valore di W abbastanza alto e un p-value minore di $0,05$ accettiamo l'ipotesi alternativa per il quale la mediana della betweeness risulta maggiore della mediana per la in-degree nella popolazione.

[2] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.friedmanchisquare.html

Continuiamo lo studio sugli altri grafi proseguendo in ordine di grandezza sul grafo social. Calcolando la statistica Friedman sulle centralità standardizzate del grafo social, notiamo i seguenti valori:

$$Fr = 12,60 \text{ con } p - value = 0,01$$

Il valore *Fr* calcolato dal test è maggiore rispetto alla statistica test della variabile aleatoria chi-quadro riferito a un livello di significatività di $0,05$ e 4 gradi di libertà, di valore uguale a $9,5$. Quindi, avendo $12,60 > 9,5$ implica che, in teoria, accettiamo l'ipotesi alternativa che alcune mediane delle centralità siano differenti tra loro nella popolazione. A differenza dei risultati ottenuti sul grafo steemit, per il grafo social le statistiche di Wilcoxon hanno mostrato uguaglianza nelle mediane sulle diverse centralità.

Proseguiamo il calcolo dei test per il grafo sub-sociale. I valori calcolati dal test di Friedman per le cinque centralità standardizzate sono:

$$Fr = 12,21 \text{ con } p - value = 0,01$$

Osserviamo che il valore risultato del test di Friedman è minore della statistica chi-quadro di valore $9,5$. Anche in questo grafo, accettiamo l'ipotesi alternativa che le mediane per le centralità sono differenti tra loro all'interno della popolazione. Anche in tale grafo, come per il grafo social, non abbiamo trovato alcuna sostanziale differenza tra le mediane delle centralità utilizzando le statistiche test di Wilcoxon.

La ricerca termina con l'impiego delle due statistiche test per il grafo monetario. La situazione in questo caso cambia notevolmente. La statistica Friedman sulle cinque centralità per il grafo monetario produce i seguenti risultati:

$$Fr = 46,25 \text{ con } p - value = 2,18e - 09$$

La distribuzione chi-quadro per il livello di significatività uguale a $0,05$ e 4 gradi di libertà produce il valore $9,5$. Poiché il valore della statistica di Friedman è parecchio maggiore del valore χ^2_F accettiamo l'ipotesi alternativa che le mediane tra le centralità non sono uguali, e notando il valore molto basso del p-value siamo certi di questo risultato.

Approfondiamo questo risultato interessante, tramite la statistica di Wilcoxon in modo da mettere in risalto quali centralità, effettivamente, in termini di mediane siano inferiori rispetto

le altre. In principio calcoliamo la mediana attesa (μ_W) e la deviazione standard attesa (σ_W) utile per calcolare la variabile "standardizzata" Z per effettuare i confronti con il risultato della statistica test W. Notiamo che per $N = 200$(numero di esempi nel campione), i parametri attesi assumono i seguenti valori:

$$\mu_W = \frac{N(N+1)}{4} = 10.050$$

$$\sigma_W = \sqrt{\frac{(N(N+1)(2N+1)}{24}} = 819,56$$

Una volta aver effettuato i soliti $\binom{5}{2}$ confronti, arriviamo a sei risultati importanti da tenere in conto. Il primo confronto da sottolineare è tra la centralità betweeness e la centralità in-degree, i valori prodotti dalla statistica sono:

$$W = 11.876 \text{ con } p - value = 0,01$$

poiché il valore della statistica W standardizzata è:

$$Z = \frac{W - \mu_W}{\sigma_W} = 2,23 \text{ e di conseguenza } Z_\alpha = 0,99 \text{ quindi } Z_\alpha < Z$$

con un p-value minore del livello di significatività ($\alpha = 0,05$), accettiamo l'ipotesi alternativa che la mediana per la betweeness è maggiore della mediana per il grado in entrata in tutta la popolazione.

I calcoli della statistica standardizzata e dei confronti, d'ora in avanti, li prenderemo per buoni poiché effettuati implicitamente dalla funzione implementata nel software di R. Proseguendo i confronti, arriviamo al confronto tra la centralità betweeness e la centralità di grado in uscita che producono i seguenti valori:

$$W = 5.746 \text{con} p - value = 4,26e - 08$$

Anche qui notiamo un valore abbastanza alto con un p-value molto basso, accettiamo l'ipotesi alternativa che la mediana della betweeness è minore della mediana riferita alla centralità di grado in uscita.

Il terzo confronto è dato dalla centralità betweeness con la centralità pagerank, che producono i seguenti valori:

$$W = 13.065 \text{ con } p - value = 0,0001$$

il valore alto di W e il valore basso del p-value confermano l'ipotesi alternativa per cui la mediana della betweeness è maggiore della mediana per la centralità pagerank.

Il quarto confronto è scaturito dalla centralità eigenvector e la centralità di grado in uscita, con $Me(Eig) < Me(Out)$ come ipotesi alternativa. I valori prodotti dalla statistica test di Wilcoxon sono:

$$W = 8.643 \text{ con } p - value = 0,04$$

Dato che abbiamo messo in risalto questo risultato, accettiamo l'ipotesi alternativa che la mediana dell'eigenvector è minore della mediana del grado in uscita.

Il penultimo confronto è dato dalla centralità di grado in entrata e la centralità di grado in uscita. I valori W e p-value sono:

$$W = 8.233 \text{ con } p - value = 0,01$$

Il livello del p-value minore di $0,05$ conferma il risultato di *W* il quale accetta l'ipotesi alternativa che la mediana della centralità di grado in entrata è minore della mediana di grado in uscita.

L'ultimo confronto, per cui vale la pena tenere nota, è la statistica test applicata alla centralità di grado in uscita e la centralità pagerank. Impostiamo come ipotesi nulla il fatto che le due mediane siano uguali e come ipotesi alternativa il fatto che la mediana per la centralità in grado di uscita sia maggiore della mediana per la centralità pagerank. I valori prodotti sono:

$$W = 13.379 \text{ con } p - value = 2,03e - 05$$

Tenendo conto dei risultati, accettiamo l'ipotesi alternativa per il quale tutta la popolazione avrà una mediana della centralità di grado in uscita maggiore della mediana della centralità pagerank.

Riassumendo le statistiche test effettuate su grafi diversi notiamo la contrapposizione tra la statistica test applicata alla centralità betweeness e la centralità di grado in uscita sul grafo globale e il grafo monetario. Infatti, si nota che la statistica applicata al grafo globale produce il risultato per cui la mediana della centralità betweeness è maggiore della mediana per la centralità di grado in uscita; mentre per il grafo monetario, la statistica test applicata alle stesse centralità offre il risultato per cui la mediana della centralità betweeness è minore della mediana per la centralità di grado in uscita.

Tenendo in considerazione solo il grafo monetario, possiamo costruire una gerarchia tra le mediane delle diverse centralità ottenendo la seguente scala decrescente, riferita alle mediane:

1. Me(Outdegree);
2. Me(Eigenvector);
3. Me(Betweeness);
4. Me(Page rank);
5. Me(Indegree).

In modo ovvio, non è possibile garantire questa gerarchia poiché non sono stati considerati tutti i possibili confronti. L'unica centralità per cui possiamo essere certi che la mediana è maggiore rispetto alle altre, è la centralità di grado in uscita. Quindi, la popolazione nell'ambito monetario tende ad effettuare più interazioni in uscita. Un altro fatto confermabile con certezza, è che le ultime tre centralità della scala per la proprietà transitiva ricoprono esattamente quelle posizioni tra loro, escludendo altre centralità (nel senso che non sappiamo effettivamente la centralità eigenvector che posizione assumi). Quindi, chiaro questo risultato, i nodi all'interno della rete tenderanno ad essere più da ponte verso altri nodi (betweeness) che tendere ad avere alto prestigio (pagerank) sfruttando i loro collegamenti e questo si riversa sull'avere poche interazioni in uscita (in degree).

Capitolo 7

Conclusioni

La ricerca si conclude effettuando il punto sui risultati ottenuti tramite la statistica descrittiva e la statistica inferenziale.

Tramite l'analisi dei risultati sulle varie distribuzioni notiamo che tutte le centralità appartengono alla categoria delle distribuzioni "heavy-tailed". In particolare, dividiamo i risultati per le diverse centralità analizzate:

- Centralità in-degree – I grafi, in generale, presentano una notevole differenza tra quei pochi nodi con alta centralità e molti nodi con bassissima centralità. Tale differenza si nota ancor di più nel grafo monetario il quale risulta avere tutti i nodi con centralità in-degree quasi nulla mentre un numero limitato di nodi con altissima centralità. Le distribuzioni seguono un andamento log-normale.

- Centralità out-degree – I grafi, in generale, presentano una notevole differenza tra il nodo con centralità massima e i rimanenti nodi. Inoltre, notiamo una somiglianza tra il nodo con massima centralità per il grafo steemit e il nodo con massima centralità per il grafo monetario che entrambi si distaccano dai massimi del grafo social e subsocial. Le distribuzioni per i grafi steemit, social e subsocial tendono ad avere un andamento log-normale nei valori intermedi. La distribuzione del grafo monetario risulta somigliare più a una power-law.

- Centralità betweeness – I valori massimi nei quattro grafi risultano circa coincidenti, con una centralità minore nel grafo globale derivata dalla relativa grandezza del grafo. La

distribuzione betweeness segue un andamento quasi uguale all'andamento di una power law per tutti e quattro i grafi.

- Centralità eigenvector – Il nodo con massima centralità del grafo monetario risulta maggiore rispetto ai nodi massimi dei grafi steemit, social e subsocial. Tale risultato derivato dalla grandezza del grafo monetario che risulta minore rispetto ai tre rimanenti, il quale valorizza quei pochi nodi centrali e visibili all'interno del grafo. Le distribuzioni seguono un andamento log-normale per i quattro grafi.

- Centralità pagerank – Esattamente come per la centralità eigenvector, il massimo del grafo monetario risulta essere maggiore del massimo dei grafi steemit, social e subsocial. Le distribuzioni tendono a somigliare a una legge power law.

Per mezzo dei risultati prodotti dai test inferenziali, li riassumiamo descrivendoli secondo un'interpretazione sociologica. Effettuando i test di correlazione notiamo i seguenti risultati:

- In-degree correlato positivamente alla pagerank - conseguente al fattore:"maggiori sono le interazioni rivolte al nodo in questione più vi è la probabilità di trasmettere "prestigio", attraverso il collegamento del nodo, calcolato dall'algoritmo page rank". Ad esempio, dati due nodi, con il primo nodo avente un numero di interazioni maggiore rispetto al secondo, vi è più probabilità che siano presenti nodi prestigiosi nel gruppo del primo nodo rispetto al gruppo di interazioni del secondo nodo.

- In-degree correlato positivamente alla eigenvector - consenguente allo stesso fattore causato dalla correlazione tra centralità di grado in entrata e centralità eigenvector;

- Eigenvector correlato positivamente alla pagerank - conseguente al fattore che entrambe le centralità cercano di riassumere i valori dei nodi tenendo in considerazione il "prestigio" dei nodi collegati.

Per mezzo dei risultati prodotti dai test inferenziali, li riassumiamo descrivendoli secondo un'interpretazione sociologica. Tramite i test attuati sulle mediane dei campioni siamo riusciti a intravedere una classifica sulle mediane delle diverse centralità analizzate. La classificazione può essere riassunta e spiegata in tale maniera, in ordine decrescente di mediane:

- Mediana(out-degree) - ragion per cui sospettiamo che le interazioni in uscita all'interno dell'ambito finanziario siano più comuni rispetto le altre centralità;
- Mediana(betweeness) – gli utenti tendono a far da "ponte" all'interno dell'ambito finanziario, sfruttando i loro collegamenti;
- Mediana(pagerank) – gli utenti tendono ad avere poco "prestigio" all'interno dell'ambito finanziario, rispetto alle caratteristiche precedenti;
- Mediana(in-degree) – gli utenti tendono a non effettuare interazioni monetarie nei confronti di altri utenti.

A concludere tali risultati, diciamo che all'interno dell'ambito monetario la maggior parte degli utenti non riesce ad avere alta centralità per grado in entrata e quindi tendono ad essere pochi gli utenti a cui sono rivolte tutte le interazioni monetarie. Dal fatto che non siamo riusciti ad avere alcun risultato per i tre grafi maggiori riguardo le mediane potrebbe essere spiegato dall'equilibrio che tendono ad avere i nodi all'interno della rete dal lato di tutte le interazioni e dal lato delle interazioni sociali. Al contrario, siamo riusciti a concludere molte informazioni sul grafo monetario per il semplice fatto che quest'ultimo è molto squilibrato a livello di struttura in quanto sono pochi i nodi che controllano questo aspetto e molti che ne dipendono.

In conclusione, il tirocinio è stato essenzialmente utile per conoscere molti aspetti della SNA e più in generale dei metodi e degli strumenti nella teoria delle probabilità e più in generale nella statistica. Per poter applicare i metodi e gli strumenti è stato necessario applicarsi nell'acquisizione di due nuovi linguaggi di programmazione, entrambi con caratteristiche prettamente differenti, molto utili per il futuro immediato accademico e prossimo lavorativo.

7.1 Lavori futuri

I risultati ottenuti con questa ricerca possono essere impiegate per un possibile confronto con un ulteriore analisi applicando una diversa scelta dei nodi da considerare nel campione nelle statistiche di Friedman e Wilcoxon. Inoltre, potrebbe essere utile analizzare i blocchi più recenti della blockchain inserirli nel campione e rianalizzare le statistiche in modo da confermare o sminuire tutto ciò che è stato analizzato in questo tirocinio.

Appendice A

Codice python

Script python utilizzato per la costruzione del grafo Subsocial aggiornato

Codice A.1: Creazione grafo

```
import networkit as nk
import networkit.graph as graph
import networkit.community as cm
import pickle
import gzip

def getlistnodes(p):
    l = []
    for person in p:
      print(person)
      l.append(users[person])
    return l
filebot = open("/../path/bots.txt",'r')
persons = []
for line in filebot:
  print(line); print(str(line)[:-1])
  persons.append(str(line)[:-1])
f = gzip.open("userssocial.gz", 'rb')
users = pickle.load(f)
```

```
g = nk.graphio.EdgeListReader(separator=' ',
firstNode=0, directed=True).read(
"/../path/socialgraph.graph")
l1 = getlistnodes(persons)
l2 = g.nodes()
difference = list(set(l2) - set(l1))
print(f'list of nodes: -l1 ')
g2 = g.subgraphFromNodes(difference) #creo il nuovo grafo
print(f'#nodes: -g2.numberOfNodes() ,
#edges: -g2.numberOfEdges() ')
#Scrittura del grafo
nodes = g2.nodes()
nodes.sort(reverse=False)
nk.graphio.EdgeListWriter(separator=' ',
firstNode=nodes[0], bothDirections=True).write(g2,
"/../path/data/subsubsocialgraph.graph")
```

Script python per il calcolo delle misure di centralità e successiva scrittura su file:

Codice A.2: Calcolo centralità sui grafi

```
import networkit as nk
import numpy as np
import datetime
import sys

namefolder=''

def calccentrality(path,namefolder):
 #apro il file in lettura contenente
 #il grafo in formato testuale
```

```
11   f= open(path,'r')
12  #variabile globale che mi indica il path dove
13  #salvare le varie distribuzioni
14   global namefolder
15  #assegno alla variabile globale il folder passato
16  #come parametro dalla riga di comando
17   namefolder=namefolder
18  #carico i grafi in memoria
19   reader= nk.graphio.EdgeListReader(separator=' ',
20   firstNode=0, directed=True)
21   G= reader.read(path)
22
23
24  #Out-degree centrality
25   print("Inizio Out-degree,
26   Current time:-", datetime.datetime.now())
27   centrality = nk.centrality.DegreeCentrality(G,
28   False, True, True).run()
29   writeonfile(centrality.scores(),
30   'DegreeCentrality(Out-degree)')
31   writeonfile(centrality.ranking(),
32   'DegreeCentrality(Out-degree)RANK')
33
34  #In-degree centrality
35   print("Inizio In-degree, Current time:-",
36   datetime.datetime.now())
37   centrality = nk.centrality.DegreeCentrality(G,
38   False, False, True).run()
39   writeonfile(centrality.scores(),
40   'DegreeCentrality(In-degree)')
41   writeonfile(centrality.ranking(),
42   'DegreeCentrality(In-degree)RANK')
```

```
43
44  #Eigenvector centrality
45   print(”Inizio Eigenvector,
46   Current time:-”, datetime.datetime.now())
47   centrality = nk.centrality.EigenvectorCentrality(G,
48   tol=1e-9).run()
49   writeonfile(centrality.scores(),
50   ’EigenvectorCentrality’)
51   writeonfile(centrality.ranking(),
52   ’EigenvectorCentralityRANK’)
53
54  #PageRank centrality
55   print(”Inizio PageRank,
56   Current time:-”, datetime.datetime.now())
57   centrality = nk.centrality.PageRank(G,
58   damp=0.85, tol=1e-9).run()
59   writeonfile(centrality.scores(),’PageRank’)
60   writeonfile(centrality.ranking(),’PageRankRANK’)
61
62  #Katz centrality
63   print(”Inizio Katz, Current time:-”,
64   datetime.datetime.now())
65   centrality = nk.centrality.KatzCentrality(G,
66   alpha=0.0005,beta=0.1,tol=1e-08).run()
67   writeonfile(centrality.scores(),’KatzCentrality’)
68   writeonfile(centrality.ranking(),’KatzCentralityRANK’)
69
70  #ApproxBetweenness centrality
71   print(”Inizio Betweenness,
72   Current time:-”, datetime.datetime.now())
73   centrality =nk.centrality.ApproxBetweenness(G,
74   epsilon=0.01, delta=0.1, universalConstant=1.0).run()
```

```
75    writeonfile(centrality.scores(),'ApproxBetweenness')
76    writeonfile(centrality.ranking(),'ApproxBetweennessRANK')
77
78   #Closeness centrality
79    print("Inizio Closeness, Current time:-",
80    datetime.datetime.now())
81    centrality = nk.centrality.Closeness(G,
82    False, nk.centrality.ClosenessVariant.Generalized).run()
83    writeonfile(centrality.scores(),"ClosenessCentrality")
84    writeonfile(centrality.ranking(),"ClosenessCentralityRANK")
85     #chiudo il file dei grafi aperto precedentemente
86    print("Fine Closeness,
87    Current time:-", datetime.datetime.now())
88    f.close()
89
90   #procedura per salvare i risultati di centralita' in un file
91  def writeonfile(listcentrality, name):
92    global namefolder
93    #diamo il nome al file
94    #dove salvare i risultati uno per riga
95    namefile = "/../path/data/"+namefolder+'/'+name+'.txt'
96    #apriamo il file (lo creiamo se non e presente)
97    f = open(namefile, 'w')
98    #convertiamo la lista di float
99    #in str per poterle scrivere nel file
100   newlist = []
101   for item in listcentrality:
102      newlist.append(str(item)+'"n')
103   f.writelines(newlist)
104   #chiudiamo il file
105   f.close()
106
```

```
107 if( name ==' main '):
108     calccentrality(sys.argv[1],sys.argv[2])
```

Codice python per la selezione mirata dei 100 nodi con più alta incoming degree e 100 nodi con la più alta outcoming degree, arrivando a un totale di 200 nodi, codice riutilizzato con la definizione di altre due funzioni utili alla selezione dei valori dei 200 nodi scelti sulle altre centralità e per la standardizzazione dei valori:

Codice A.3: Ranking

```
1  import sys
2  import statistics
3
4  def rank(path, namefile, n):
5      f = open(path, 'r')
6      out = open("./"+namefile, 'w')
7      cont = 0
8      for riga in f:
9        tmp = riga[1:riga.find(',')]
10       out.writelines(tmp+'"n')
11       cont = cont+1
12       if(cont == n):
13         break
14     f.close()
15     out.close()
16 def misure(path,pathresult):
17   result=open("./Lista200.txt",'r')
18   misurain=open(path,'r')
19   resultout=open(pathresult,'w')
20   lresult =[]; lmisure=[]; lout=[]
21   for riga in result:
22     lresult.append(riga)
23   for riga in misurain:
24     lmisure.append(riga)
```

```
25    for elem in lresult:
26      for elem2 in lmisure:
27        x=elem2[1:elem2.find(”,”)]
28        y1=int(x)
29        y2=int(elem)
30        if(int(elem)==int(x)):
31          print(”trovato”)
32          lout.append(elem2)
33          break
34    for e in lout:
35      resultout.writelines(e)
36    result.close(); misurain.close() ; resultout.close()
37
38  def valori(path,namepath):
39    f=open(path,’r’)
40    f2=open(namepath,’w’)
41    l = []; l1 = []
42    for riga in f:
43      tmp=riga[riga.find(’,’)+2:-2]
44      l.append(float(tmp))
45    for i in range(200):
46      y=l[i]
47      f2.writelines(str(y)+’“n’)
48    f.close(); f2.close()
49
50  def standardization(path, namepath):
51    f=open(path,’r’)
52    f2=open(namepath,’w’)
53    l = [] ; l1 = []
54    for riga in f:
55      l.append(float(riga))
56
```

```
  somma = 0
  for elem in l:
    tmp = somma + elem
    somma = tmp

  media = somma / 200
  sd = statistics.stdev(l)
  for elem in l:
    z = (elem - media) / sd
    l1.append(z)

  for i in range(200):
    f2.writelines(str(l1[i])+'“n')

  f.close(); f2.close()

rank("/../path/DegreeCentrality(In-degree)RANK.txt",
"Lista200.txt",100)
resultin=open("./Lista200.txt",'r')
rank("/../path/DegreeCentrality(Out-degree)RANK.txt",
"Lista1000outdegree.txt",1000)

res=open("./Lista200.txt",'a')
outdegree=open("./Lista1000outdegree.txt",'r')
count=0
trovato=0;
l=[]
l2=[]
for riga in outdegree:
  l.append(riga)
for riga in resultin:
  l2.append(riga)
```

```
for elem in l:
  for elem2 in l2:
    if(elem==elem2):
      trovato=1
      break
  if(trovato==0):
   res.writelines(elem)
   count=count+1
  else:
    trovato=0
  if(count==100):
    break
outdegree.close(); res.close()
#Ad ogni esecuzione del programma
#viene chiamata una e una sola funzione in base
#al valore del primo argomento

if(sys.argv[1]==”m”):
  misure(sys.argv[2],sys.argv[3])
if(sys.argv[1]==”v”):
  valori(sys.argv[2],sys.argv[3])
if(sys.argv[1]==”s”):
  standardization(sys.argv[2], sys.argv[3])
```

Script python utilizzato per l’esecuzione dell’algoritmo di Friedman per i confronti su k campioni dipendenti

Codice A.4: Statistica di Friedman

```
from scipy import stats
import sys

#procedura per calcolare la statistica di friedman
#prendendo in ingresso il nome del file
```

```
#dove salvare i risultati calcolati
def  confrontiF(name):
    #apro in lettura i file delle misure
    path="/media/sdc/lissandrello/code/"
    f1 = open(path+"StandBetweeness.txt",'r')
    f3 = open(path+"StandEigenvector.txt",'r')
    f4 = open(path+"StandIndegree.txt",'r')
    f5 = open(path+"StandOutdegree.txt",'r')
    f7 = open(path+"StandPagerank.txt",'r')
    #creo sette liste vuote per inserire i valori
    l1 = []; l2 = []; l3 = []
    l4 = []; l5 = []
    #sette cicli for per caricare
    #all'interno delle liste le varie misure
    for line in f1:
      l1.append(float(line))
    for line in f2:
      l2.append(float(line))
    for line in f3:
      l3.append(float(line))
    for line in f4:
      l4.append(float(line))
    for line in f5:
      l5.append(float(line))
    for line in f6:
    #calcolo le statistiche tramite
    #il modulo friedmanchisquare
    (statistic, pvalue) = stats.friedmanchisquare(l1,l2,l3,
    l4,l5)

        #scrivo i risultati in un file
    writeonfile(name, statistic, pvalue)
```

```
38
39 #procedura per scrivere all'interno di un file
40 #due misure (prima riga separati da uno spazio)
41 #name: contiene il nome del file dove scrivere le misure
42 #parameter1:prima misura da salvare
43 #parameter2:seconda misura da salvare
44 def writeonfile(name, parameter1, parameter2):
45  #apro/creo il file in cui inserire le due misure
46     fresult= open("./"+name,'w')
47     #scrivo la prima misura, separo
48     #da uno spazio e scrivo la seconda
49     fresult.write(str(parameter1)+' ')
50     fresult.write(str(parameter2)+'"n')
51     #chiudo il file aperto/creato precedentemente
52     fresult.close()
53
54 if( name ==' main '):
55    confrontiF(sys.argv[1])
```

Appendice B

Codice R

Script implementato in R per il caricamento in memoria del dataset contenente tutte le misura di centralità; il programma legge dai file con estensione.txt le misure per ogni valore di tutti i nodi, infine vengono calcolate e stampate le distribuzioni con successivo fitting per legge di potenza e distribuzione log-normale.

Codice B.1: Calcolo distribuzioni

```
1  bet = read.table(file="ApproxBetweenness.txt",
2  header=FALSE,sep="/n",dec=".")
3  eig = read.table(file="EigenvectorCentrality.txt",
4  header=FALSE,sep="/n",dec=".")
5  indegree = read.table(file="Indegree.txt",
6  header=FALSE,sep="/n"",dec=".")
7  outdegree = read.table(file="Outdegree.txt",
8  header=FALSE,sep="/n",dec=".")
9  page = read.table(file="PageRank.txt",
10 header=FALSE,sep="/n",dec=".")
11 dataset = data.frame(Betweeness=bet$V1,
12 Eigenvector=eig$V1,Indegree=indegree$V1,
13 Outdegree=outdegree$V1, PageRank=page$V1)
14
15 b = bet[bet¿0]
```

```
c = close[close¿0]
e = eig[eig¿1e-05]
i = indegree[indegree¿0]
o = outdegree[outdegree¿0]
k = katz[katz¿0]
p = page[page¿0]
library("poweRlaw")

mpl = conpl$new(b)
mln = conlnorm$new(b)
est = estimatexmin(mpl)
mpl$setXmin(est)
est = estimatexmin(mln)
mln$setXmin(est)
pdf("PowerlawBet.pdf")
plot(mpl,xlab="Betweeness",ylab="CDF",
main="Betweeness - CDF")
lines(mpl, col = 2, lwd=3)
lines(mln, col = 3, lwd=3)
dev.off()

mpl = conpl$new(c)
mln = conlnorm$new(c)
est = estimatexmin(mpl)
mpl$setXmin(est)
est = estimatexmin(mln)
mln$setXmin(est)
pdf("PowerlawClose.pdf")
plot(mpl,xlab="Closeness",ylab="CDF",
main="Closeness - CDF")
lines(mpl, col = 2, lwd=3)
lines(mln, col = 3, lwd=3)
```

```
48 dev.off()
49
50 mpl = conpl$new(e)
51 mln = conlnorm$new(e)
52 est = estimatexmin(mpl)
53 mpl$setXmin(est)
54 est = estimatexmin(mln)
55 mln$setXmin(est)
56 pdf("PowerlawEig.pdf")
57 plot(mpl,xlab="Eigenvector",ylab="CDF",
58 main="Eigenvector - CDF")
59 lines(mpl, col = 2, lwd=3)
60 lines(mln, col = 3, lwd=3)
61 dev.off()
62
63 mpl = conpl$new(i)
64 mln = conlnorm$new(i)
65 est = estimatexmin(mpl)
66 mpl$setXmin(est)
67 est = estimatexmin(mln)
68 mln$setXmin(est)
69 pdf("PowerlawIndegree.pdf")
70 plot(mpl,xlab="In-degree",ylab="CDF",
71 main="Indegree - CDF")
72 lines(mpl, col = 2, lwd=3)
73 lines(mln, col = 3, lwd=3)
74 dev.off()
75
76 mpl = conpl$new(o)
77 mln = conlnorm$new(o)
78 est = estimatexmin(mpl)
79 mpl$setXmin(est)
```

```
est = estimatexmin(mln)
mln$setXmin(est)
pdf("PowerlawOutdegree.pdf")
plot(mpl,xlab="Out-degree",ylab="CDF",
main="Outdegree - CDF")
lines(mpl, col = 2, lwd=3)
lines(mln, col = 3, lwd=3)
dev.off()

mpl = conpl$new(k)
mln = conlnorm$new(k)
est = estimatexmin(mpl)
mpl$setXmin(est)
est = estimatexmin(mln)
mln$setXmin(est)
pdf("PowerlawKatz.pdf")
plot(mpl,xlab="Katz",ylab="CDF",
main="Katz - CDF")
lines(mpl, col = 2, lwd=3)
lines(mln, col = 3, lwd=3)
dev.off()

mpl = conpl$new(p)
mln = conlnorm$new(p)
est = estimatexmin(mpl)
mpl$setXmin(est)
est = estimatexmin(mln)
mln$setXmin(est)
pdf("PowerlawPagerank.pdf")
plot(mpl,xlab="Pagerank",ylab="CDF",
main="Page-Rank - CDF")
lines(mpl, col = 2, lwd=3)
```

```
lines(mln, col = 3, lwd=3)
dev.off()
```

Script in R per il calcolo degli stem and leaf, in più calcoliamo gli indici statistici.

Codice B.2: Calcolo grafico e indici

```
stem = stem(dataset$Centrality)
summary(dataset$Centrality)#calcolo minimo, massimo, media,
                           #mediana, primo e terzo quartile
sd(dataset$Centrality) #calcolo della deviazione standard
library(labstatR)
skew(dataset$Centrality) #coefficiente di skewness
kurt(dataset$Centrality) #coefficiente di curtosi
cv(dataset$Centralitys) #coefficiente di variazione
```

Script in R per la stampa dei diagrammi di dispersione tra tutte le centralità in modo da vedere graficamente le correlazioni e successivamente il calcolo analitico dei test di Spearman e Kendall con la stampa di una heatmap per ogni matrice di correlazioni prodotta dalla statistica test.

Codice B.3: Calcolo delle correlazioni

```
pdf("Diagrammididispersione.pdf")
pairs(dataset)# per visualizzare le
              #correlazioni tramite i dataplot
dev.off()
mapsp = cor.test(dataset,method="spearman")
mapken = cor.test(dataset,method="kendall")
pdf("HeatmapSperaman.pdf")
heatmap(mapsp,scale="none")
dev.off()
pdf("HeatmapKendall.pdf")
heatmap(mapken,scale="none")
dev.off()
```

Script in R per l'esecuzione del test di Wilcoxon su due campioni dipendenti appaiati, con le tre diverse ipotesi alternative; ovviamente da applicare a due a due per tutte le centralità, quindi con un rapido calcolo avremo con ad esempio sette centralità, un totale di $\binom{5}{2}$ test differenti.

Codice B.4: Statistica di Wilcox

```
wilcox.test(dataset$Centrality1,dataset$Centrality2,
alternative="two.sided",exact=TRUE,paired=TRUE)
wilcox.test(dataset$Centrality1,dataset$Centrality2,
alternative="greater",exact=TRUE,paired=TRUE)
wilcox.test(dataset$Centrality1,dataset$Centrality2,
alternative="less",exact=TRUE,paired=TRUE)
```

Appendice C

Analisi integrativa grafo monetario

Si fornisce un'analisi integrativa alla descrizione statistica del grafo minore. Le centralità da integrare alle cinque già descritte, sono la centralità closeness e la centralità secondo Katz. I motivi per cui non sono stati integrati è per creare un giusto filo del discorso e una giusta descrizione dei diversi grafi che sia coerente.

C.1 Centralità closeness

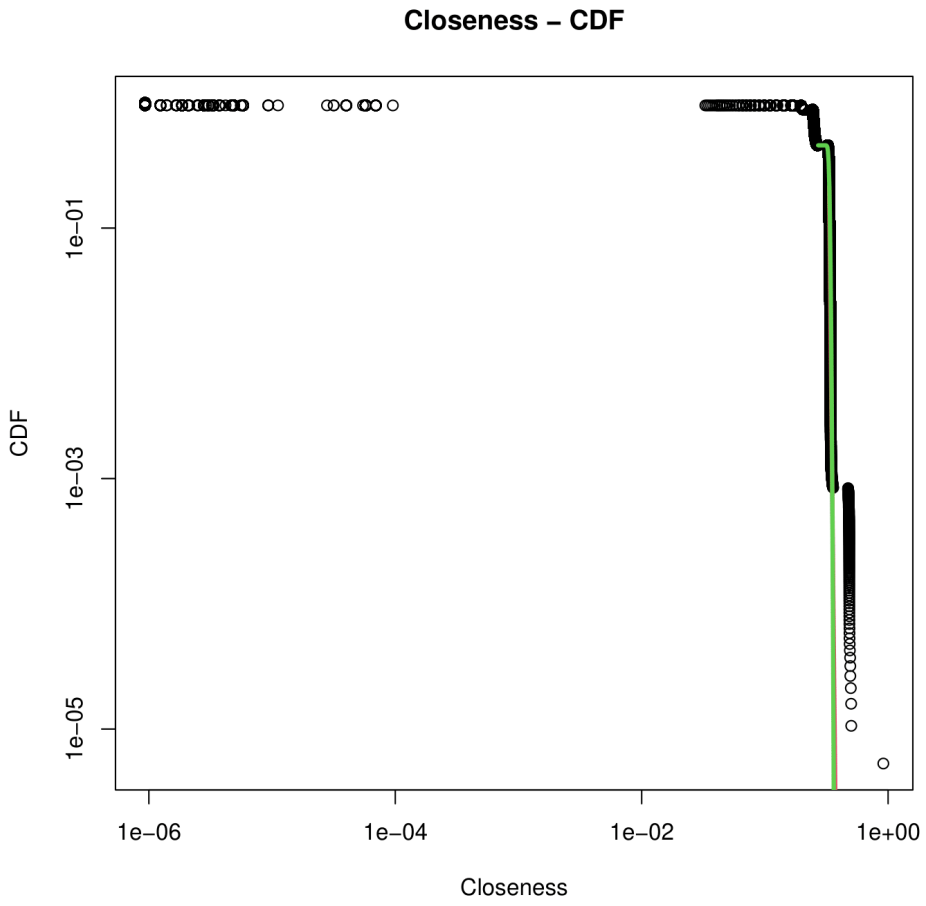
Per il grafo monetario è stato possibile calcolare la centralità closeness che fornisse una misura precisa dei nodi. Questa centralità è stata possibile calcolarla per la grandezza non troppo esagerata del grafo, spendendo circa 3 giorni per il calcolo rispetto alla closeness calcolata sugli altri tre grafi che all'incirca richiede due mesi di calcolo.

Si fornisce quindi uno stem-and-leaf, i parametri statistici e la CDF.

```
The decimal point is 1 digit(s) to the left of the |

0 | 000000000000000000000000000000000000000000000000000000000000000000000000000000+895029
0 | 55555555666667777777788888888899999
1 | 000000011111111122222222223444444444444444444444444444444444444444444444444444+10
1 | 55566666666666666666666666666677777777777777777777777777777777777777+356
2 | 000000000000000000000000000000000000000000000000000000000000000000000000000000+14769
2 | 55555555555555555555555555555555555555555555555555555555555555555555555555555+77880
3 | 22222222222222222222222222222222222222222222222222222222222222222222222222222+85923
3 | 55555555555555555555555555555555555555555555555555555555555555555555555555555+1
4 |
4 | 77777778888888888888888888888888888888888888888888888888888888888888888888888+74
5 | 000
5 |
6 |
6 |
7 |
7 |
8 |
8 |
9 | 1
```

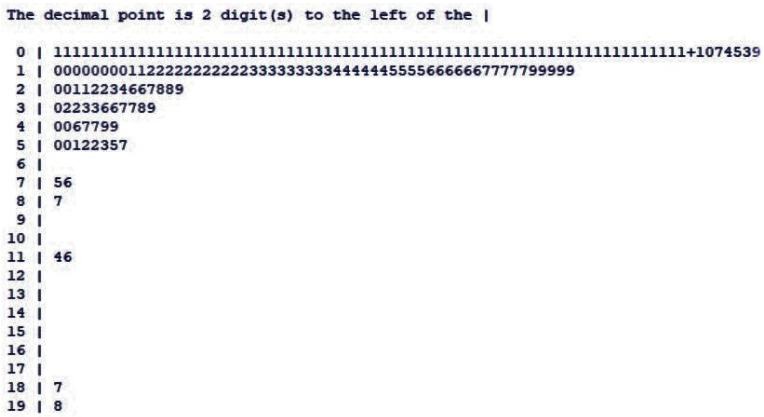
Parametro	Valore
Minimo	0
Q_1	0
Mediana	0
Media	$0,048$
Q_3	0
Massimo	$0,91$
Asimmetria	$1,91$
Curtosi	$4,83$
Dev. Standard	$0,11$
Coeff. Variazione	$2,27$



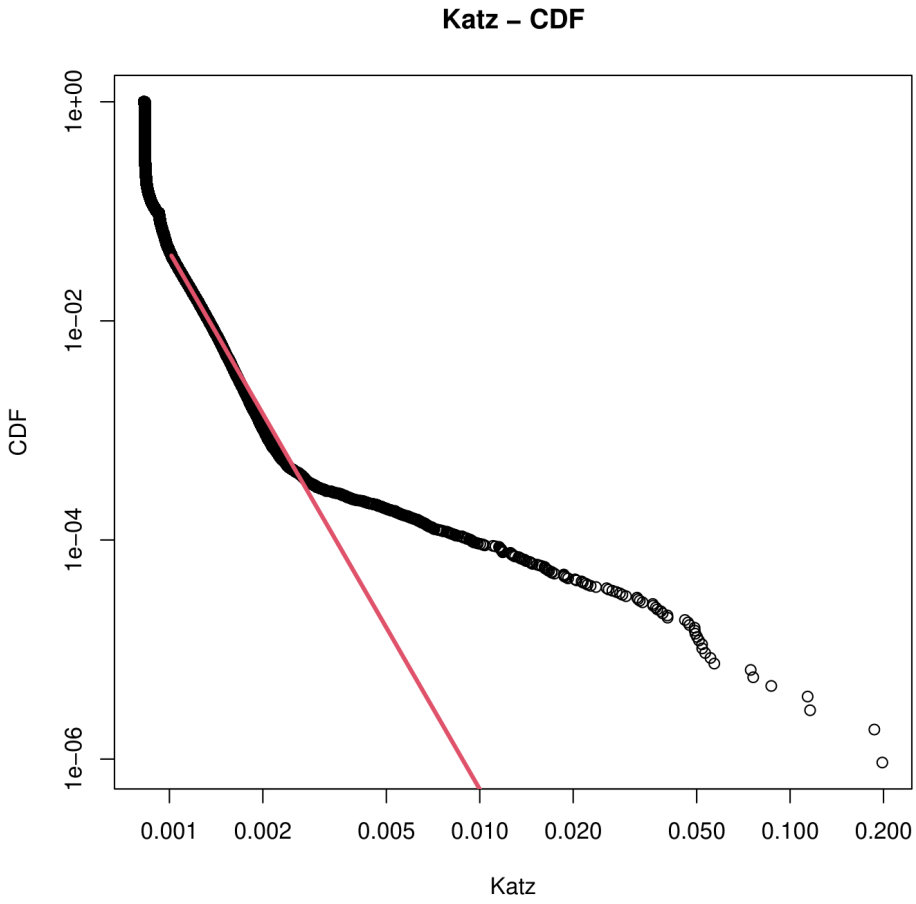
C.2 Centralità secondo Katz

Un problema diverso è stato riscontrato per la centralità secondo Katz. I tempi di calcolo sono stati relativamente brevi per tutti i grafi, ma le uniche misure consistenti sono state riscontrate nel grafo monetario, nei grafi maggiori tutte le misurazioni rislutavano con centralità uguale a zero.

Quindi, gli unici dati per questa centralità sono riguardanti il grafo più piccolo, e si offrono di conseguenza le misure analitiche e grafiche del suddetto grafo, per un parametro $\alpha = 0,0005$.



Parametro	Valore
Minimo	$0,00083$
Q_1	$0,00083$
Mediana	$0,00083$
Media	$0,00086$
Q_3	$0,00084$
Massimo	$0,20$
Asimmetria	$246,92$
Curtosi	89.206
Dev. Standard	$0,00043$
Coeff. Variazione	$0,50$


Katz – CDF
CDF
1e+00
1e−02
1e−04
1e−06
0.001
0.002
0.005
0.010
0.020
0.050
0.100
0.200
Katz

Bibliografia

[1] David Austin. How google finds your needle in the web's haystack. *American Mathematical Society Feature Column*, 10(12), 2006.

[2] Albert-László Barabási and Eric Bonabeau. Scale-free networks. *Scientific american*, 288(5):60–69, 2003.

[3] Alex Bavelas. Communication patterns in task-oriented groups. *The journal of the acoustical society of America*, 22(6):725–730, 1950.

[4] Peter M Blau. *Structural contexts of opportunities*. University of Chicago Press, 1994.

[5] ANTONIO M Chiesi. L'analisi dei reticoli sociali: teoria e metodi. *Rassegna Italiana di sociologia*, 2:291–310, 1980.

[6] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[7] Christophe Croux and Catherine Dehon. Influence functions of the spearman and kendall correlation measures. *Statistical methods & applications*, 19(4):497–515, 2010.

[8] Thomas H Davenport and John C Beck. The attention economy. *Ubiquity*, 2001(May):1–es, 2001.

[9] Paola Di Nicola. *La rete: metafora dell'appartenenza. Analisi strutturale e paradigma di rete*, volume 4. FrancoAngeli, 1998.

[10] Robert A Hanneman and Mark Riddle. Introduction to social network methods, 2005.

[11] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[12] Alan Kazdin. The token economy: A review and evaluation. 2012.

[13] Christian FA Negre, Uriel N Morzan, Heidi P Hendrickson, Rhitankar Pal, George P Lisi, J Patrick Loria, Ivan Rivalta, Junming Ho, and Victor S Batista. Eigenvector centrality for characterization of protein allosteric pathways. *Proceedings of the National Academy of Sciences*, 115(52):E12201–E12208, 2018.

[14] Mark EJ Newman. The mathematics of networks. *The new palgrave encyclopedia of economics*, 2(2008):1–12, 2008.

[15] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[16] Sergio Porta and Vito Latora. Multiple centrality assessment. centralità e ordine complesso nell'analisi spaziale e nel progetto urbano. *Territorio*, 39:189–202, 2006.

[17] Matteo Riondato and Evgenios M Kornaropoulos. Fast approximation of betweenness centrality through sampling. *Data Mining and Knowledge Discovery*, 30(2):438–475, 2016.

[18] Lamberto Soliani. *I test non-parametrici più citati nelle discipline scientifiche*. Uni. Nova, 2008.

[19] Christian Staudt, Aleksejs Sazonovs, and Henning Meyerhenke. Networkit: An interactive tool suite for high-performance network analysis. *arXiv preprint arXiv:1403.3005*, 2014.

[20] Steemit.io. "steem an incentivized, blockchain-based, public content platform.". 2018.

[21] Giuseppe Toscani. Sulle code di potenza di pareto. *Matematica, Cultura e Società. Rivista dell'Unione Matematica Italiana-Serie*, 1:21–30, 2016.

[22] v. Montagna and P. Pulieri. Analisi della topologia del grafo delle transazioni di steem. Master's thesis, Dipartimento di Informatica Corso di Laurea in Informatica, 2019.

[23] Sebastiano Vigna. Spectral ranking. *Network Science*, 4(4):433–445, 2016.

[24] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[25] Stanley Wasserman, Katherine Faust, et al. Social network analysis: Methods and applications. 1994.

[26] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[27] Barry Wellman and Stephen D Berkowitz. *Social structures: A network approach*, volume 2. CUP Archive, 1988.

[28] Michele Zenga. *Lezioni di statistica descrittiva: Seconda edizione*. G Giappichelli Editore, 2014.

[29] Donald W Zimmerman and Bruno D Zumbo. Relative power of the wilcoxon test, the friedman test, and repeated-measures anova on ranks. *The Journal of Experimental Education*, 62(1):75–86, 1993.