UNIVERSITÀ DEGLI STUDI DI UDINE

DIPARTIMENTO DI SCIENZE MATEMATICHE, INFORMATICHE E FISICHE

CORSO DI LAUREA TRIENNALE IN INTERNET OF THINGS, BIG DATA, MACHINE LEARNING

ANNO ACCADEMICO 2024/25

PROGETTO DI SOCIAL COMPUTING

ALLIEVI

Enrico Francesco CONTESSI

Roland GJOPALAJ

Giovanni PANTAROTTO

Cristian TOMASS

DOCENTI

Prof.ssa Hafsa AKEBLI

Prof. Michele LIZZIT

Prof. Stefano MIZZARO

Prof. Michael SOPRANO

---

email: contessi.enricofrancesco@spes.uniud.it **matricola:** 162738

email: gjopalaj.roland@spes.uniud.it **matricola:** 157277

email: pantarotto.giovanni@spes.uniud.it **matricola:** 157707

email: tomass.cristian@spes.uniud.it **matricola:** 147813

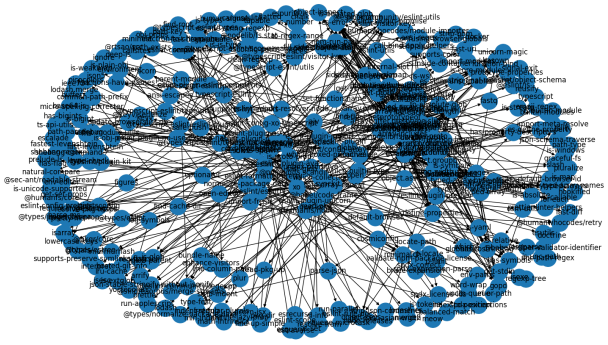
SOMMARIO

Questo progetto di **social computing** si concentra sull'analisi di due reti di dipendenze software, generate a partire dal pacchetto seed **xo**. I dati sono stati scaricati utilizzando l'API di NPM e rappresentano le dipendenze dirette e indirette dei pacchetti. È stato, poi, creato un grafo diretto per visualizzare le relazioni di dipendenza, seguito da un grafo indiretto espanso con l'aggiunta di 350 nuovi nodi, tramite la tecnica del preferential attachment. Entrambe le reti sono state analizzate sotto diversi aspetti, tra cui centralità, coefficiente di clustering, e indici di small-worldness.

Il grafo indiretto si è rivelato essere più robusto e ben connesso rispetto a quello diretto, con il nodo seed che, pur avendo un basso pagerank, gioca un ruolo centrale nella rete. Il grafo diretto, d'altra parte, ha visto una perdita di centralità per il nodo seed a causa dell'assenza di nodi entranti. Entrambe le reti risultano sparse e con bassa tendenza a formare cluster, suggerendo una struttura di pacchetti software poco modulare, che potrebbe compromettere la manutenibilità e comprensibilità del sistema.

METODOLOGIA

Per raggiungere gli obiettivi dello studio, è stato scelto come pacchetto seed **xo** (link), da cui sono stati scaricati i dati, in formato JSON, relativi alle dipendenze dirette e indirette, tramite l'utilizzo dell'API ufficiale di NPM. Dopo di ciò, è stato creato il primo grafo (orientato) delle dipendenze con il layout di Fruchterman-Reingold, che fornisce una visione globale di tutte le dipendenze, incluse quelle indirette, cioè quelle tra pacchetti da cui il pacchetto seed dipende.

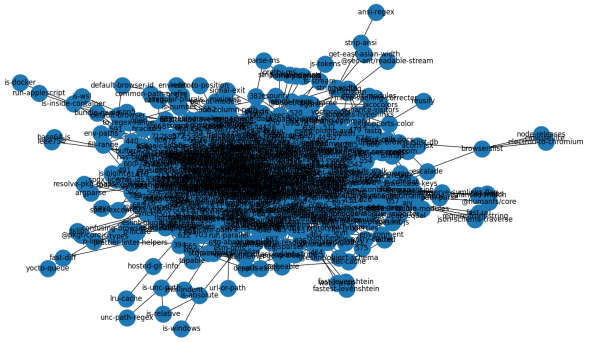


Successivamente, è stato creato il secondo grafo (non orientato) a partire dal primo, che è stato trasformato in indiretto, poi, espanso con 350 nuovi nodi, ognuno dei quali con 3 archi verso altri nodi del grafo, secondo una certa probabilità. Per fare ciò, è stata utilizzata la tecnica del **preferential attachment**, che consiste nel calcolare la probabilità per ogni nodo che quest'ultimo venga "scelto" da un nuovo nodo:

$$P(v_i) = d_i / \Sigma_j d_j$$

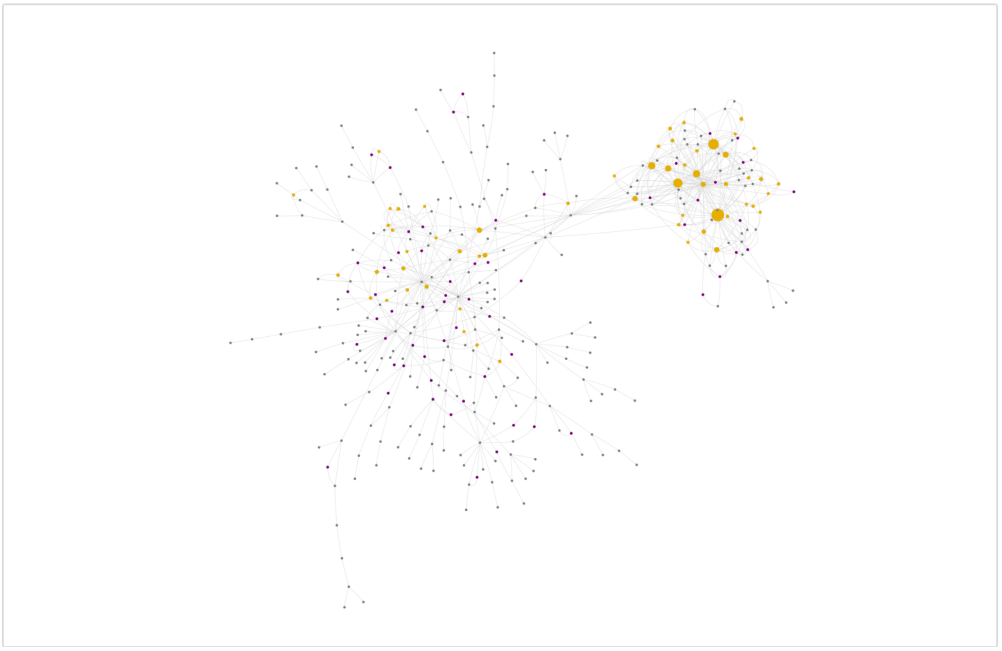
Dove al numeratore abbiamo d_i degree, cioè numero di vicini (neighbors), del nodo i-esimo e al denominatore $\Sigma_j d_j$, cioè la somma dei degree di tutti i nodi del grafo (compreso il nodo i-esimo).

A questo punto, è stato creata la rappresentazione visuale del secondo grafo, utilizzando, però, lo spring layout, in quanto il Fruchterman-Reingold layout non produce grafi indiretti.

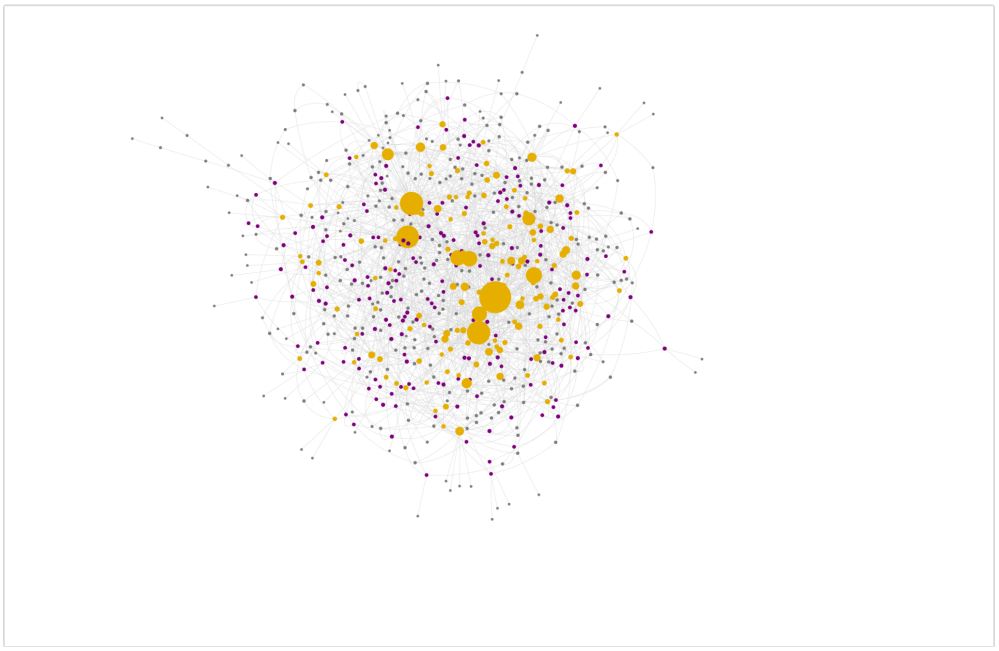


Essendo, però, che le due visualizzazioni statiche sono poco leggibili e interpretabili, sono state prodotte due illustrazioni interattive, che permettono una migliore esplorazione del grafo, utilizzando la libreria di Python PyVis (link).

Per il primo grafo la visualizzazione interattiva prodotta è la seguente:



Mentre per il secondo è la successiva:



NB: per una migliore visualizzazione ed esplorazione, aprire i file HTML situati nella cartella html del progetto.

RISULTATI

Durante la fase di analisi dei dati, è stato osservato che nel grafo diretto il coefficiente di clustering medio assume valore 0,079581, ciò indica che i nodi del grafo non tendono a formare dei cluster, cioè gruppi di nodi strettamente connessi tra di loro. Nel grafo indiretto, invece, è stato registrato un valore di 0,066198, che è più basso rispetto a quello del grafo diretto. Questi due risultati, inoltre, sono supportati dai valori inferiori allo 0,05, che sono stati osservati per la transitività, in entrambi i grafi, cioè il rapporto tra il numero di triadi chiuse e quelle possibili, e ciò dimostra che i due grafi sono dei grafi sparsi.

Nel grafo indiretto è stato possibile trovare i nodi centrali, tra cui è presente il nodo seed. Di quest'ultimo sono state calcolate la betweenness centrality e il PageRank, che assumono, rispettivamente, un valore di 0,21182 e di 0,014754. Il primo valore indica che il seed è "in mezzo" in circa il 21% dei cammini minimi fra tutte le coppie di nodi del grafo, mentre il valore molto basso di pagerank indica che il nodo è connesso ad altri nodi con un pagerank altrettanto basso. Inoltre, la closeness centrality del nodo seed assume valore 0,39250, posizionandolo in cima, rispetto agli altri centri del grafo, per vicinanza agli altri nodi. Infine, per la degree centrality del seed è stato annotato un valore normalizzato di 0,080057, che mostra come il seed, nonostante sia un nodo centrale, è connesso soltanto all'8% dei nodi del grafo, cioè è connesso a circa 61 nodi su un totale di 764 (414 di partenza e 350 aggiunti con il preferential attachment).

Nel grafo diretto non è stato possibile individuare i centri del grafo, a causa dei limiti della libreria di Python NetworkX (link), ma è stato possibile osservare i valori di centralità del nodo seed. Infatti, i valori che quest'ultimo assume per la betweenness e closeness centrality sono entrambi 0, questo perché per la prima metrica il nodo ha solo archi uscenti, mentre per la seconda non esiste alcun cammino che da un qualsiasi altro nodo porti al seed. Inoltre, anche per la in-degree centrality è stato annotato il valore 0, e ciò è dovuto dal fatto che il seed è un nodo gregario, cioè con soli archi uscenti. Per quanto riguarda l'ultima metrica, cioè il pagerank, invece, è stato osservato un valore inferiore rispetto a quanto annotato dal grafo indiretto, questo perché il nodo non ha archi entranti da nodi importanti e distribuisce il suo basso pagerank agli altri nodi.

Nel grafo indiretto, infine, è stato possibile analizzare due indici di small-worldness, sigma e omega. Il primo misura la frazione di nodi connessi alla componente gigante, rispetto a una rete casuale equivalente, mentre il secondo confronta il comportamento del grafo tra una rete a struttura casuale e una a struttura regolare. Per il primo indice è stato osservato un valore di 1,540083, che indica che il grafo ha una connettività migliore rispetto a una rete casuale con le stesse proprietà, e ciò vuol dire che c'è un alto livello di robustezza in caso di rimozione di nodi casuali. Per il secondo indice, invece, è stato annotato un valore di -0,042093, che suggerisce che il grafo ha una struttura intermedia tra una rete casuale e una regolare, e questo indica che la rete ha una tendenza a comportamenti regolari od organizzati nella sua struttura.

CONCLUSIONI

In conclusione, è stato possibile affermare che, in base ai risultati ottenuti, il grafo indiretto è robusto e ben connesso, con il nodo seed che risulta strategico per la sua posizione, nonostante il pagerank basso. Nel grafo diretto, invece, il seed perde rilevanza, a causa del fatto che non possiede nodi entranti, e ciò riduce la sua centralità complessiva e rende la rete meno robusta. Inoltre, è stato osservato che entrambe le reti sono sparse e con scarsa tendenza a formare cluster, e questo può essere dovuto alla scarsa modularizzazione dei pacchetti software, che può portare, a sua volta, a una difficile manutenibilità e comprensibilità del sistema.