Report delle attività svolte

- Weighted Network Analysis

Ho svolto un'analisi (superficiale) del network pesato contenuto nel file *weighted_academic_graph.txt*. Nel network, ogni nodo rappresenta uno stato, mentre ogni arco rappresenta uno spostamento da uno stato ad un altro. Inoltre, ogni arco contiene informazioni riguardo all'anno in cui lo spostamento è avvenuto e a quante volte in quell'anno è stato condotto uno spostamento tra i due stati. Quest'ultima informazione rappresenta la variabile peso caratterizzante il network. Eseguendo il notebook *weighted_network_analysis.ipynb* è possibile osservare la mia analisi, nel quale conduco uno studio sulla distribuzione dei gradi generali, in entrata e in uscita dei nodi componenti il network, sulla connectedness del network, sulla distribuzione del clustering coefficient tra i nodi componenti il network e sulla lunghezza dei cammini.

- Estrazione dati dai dump di Zenodo

Ho estratto e processato i dati contenuti nei dump di Zenodo (reperibili qui). Tali dati sono principalmente di due tipi, uno riguardante le organizzazioni coinvolte nel finanziamento di attività inerenti alla ricerca, e l'altro riguardante i progetti presentati all'interno di un determinato framework. Le organizzazioni posso essere di varia natura, come, per esempio, università o finanziatori privati. Il procedimento di estrazione consiste in uno script Python (*zenodo_parser.py*) che decomprime, decripta e converte i record contenuti nei dump in un unico file XML. Successivamente, tramite un secondo script Python (*zenodo_xml_miner.py*), i dati più interessanti contenuti nel file XML vengono selezionati e salvati in un file JSONL. I dati relativi alle organizzazioni consistono in una divisione per Paesi degli ID caratterizzanti ogni organizzazione, mentre quelli relativi ai framework sono organizzati in modo tale che ogni riga contiene le informazioni relative a un determinato ricercatore comparso all'interno del framework come, per esempio, nome, cognome, id MAG, ORCID o URL (se presenti), e una lista dei progetti al quale egli ha partecipato. Ogni progetto è individuato da un nome, ossia il titolo che è stato dato alla sua pubblicazione, e una serie di keywords. Per motivi di spazio, i file JSONL risultanti dall'intero processo vengono salvati su un'apposita macchina remota, ospitata dal CNR.

- Analisi del Coauthors Network

Per ognuno dei network pesati dei coautori, che consistono in network non diretti nei quali i pesi rappresentano il numero di volte che un coppia di ricercatori hanno collaborato durante un particolare anno, ho calcolato le statistiche di base per quanto riguarda le analisi delle reti. Ho riportato in dei CSV le informazioni per quanto riguarda il numero di nodi, il numero di archi, il coefficiente di clustering medio, la transitivity, il diametro e il raggio. Oltre a queste statistiche di

base, ho fornito anche delle visualizzazioni riguardo la distribuzione dei gradi dei nodi, la distribuzione dei pesi degli archi, la degree centrality dei nodi, la closeness centrality dei nodi e la betweenness dei nodi.

- Costruzione dei Networks a partire dai Dump di Zenodo

Ho scritto uno script Python che, per ognuno dei file XML generati a partire dai Dump di Zenodo riguardanti i risultati dei vari framework, genera un Collaboration Network nel quale ogni nodo corrisponde a un ricercatore dotato di identificatore MAG, e ogni arco rappresenta una collaborazione tra due ricercatori all'interno di un progetto contenuto nel framework rappresentato dal file XML. Per ogni progetto vengono presi in considerazione soltanto i ricercatori dotati di identificatore MAG. Vengono inoltre salvati il nome del progetto, la data di accettazione, e, se disponibili, l'identificativo del progetto e il suo acronimo. Le informazioni correlate al progetto, ossia nome, identificativo e acronimo, vengono infine salvate come un file JSONL distinto. Al termine dello script viene generato un file JSONL rappresentante la lista degli archi contenuti nel network, nel quale ogni riga contiene le informazioni relative alla collaborazione tra due ricercatori per un determinato anno e progetto.
Un altro script Python si occupa di costruire e analizzare i network prodotti al passo precedente. Per prima cosa viene generato un DataFrame a partire dalla lista degli archi generata al termine del passo precedente e, successivamente, per ogni anno distinto presente nel DataFrame, viene generato il network corrispondente. Infine, per ogni network, vengono raccolte le statistiche relative al numero dei nodi, al numero degli archi, alla densità, al coefficiente di clustering medio, alla transitività, al raggio e al diametro. È da sottolineare che le statistiche relative al coefficiente di clustering medio, alla transitività, al raggio e al diametro sono ottenute a partire dall'analisi delle prime 10000 componenti connesse composte da un minimo di 3 nodi fino a un massimo di 1000 nodi. Questo vincolo è necessario perché altrimenti lo script impiegherebbe un tempo proibitivo per la terminazione. Oltre alle statistiche menzionate precedentemente, vengono generate anche dei grafici rappresentanti la distribuzione del grado dei nodi e del peso degli archi per ognuno degli anni distinguibili all'interno dei vari DataFrame.

- Entropia Binaria sulle Affiliazioni

Per ogni network costruito a partire dai frameworks recuperabili dai dump di Zenodo, ho calcolato l'entropia binaria relativa alla distribuzione di probabilità ottenuta a partire dai paesi di affiliazione dei ricercatori contenuti nella ego network di ogni ricercatore presente nel network. Il procedimento per il calcolo delle entropie binarie è svolto nel modo seguente: dal momento che ogni framework possiede un dump nel quale sono presenti pubblicazioni di papers svoltesi in anni diversi, si procede al calcolo delle entropie iterando per ogni possibile anno recuperabile.
Per ogni network creato per una specifica coppia framework/anno, si procede iterando sui nodi, ossia, sui ricercatori, contenuti al suo interno. Per prima cosa viene recuperata la ego network del ricercatore protagonista dell'iterazione corrente, la quale contiene coloro che hanno collaborato alla stesura di un paper con il ricercatore, e, successivamente, viene creata una lista contenente i paesi di affiliazione di ognuno dei collaboratori presenti nella ego network. L'entropia binaria viene quindi calcolata sulla lista ottenuta al passo precedente. Viene inoltre tenuto conto del fatto che l'affiliazione più popolare tra i collaboratori del ricercatore corrente sia dello stesso paese del

ricercatore, oppure di un paese diverso e, in tal caso, l'entropia ottenuta viene moltiplicata per -1. Per ogni coppia framework/anno viene infine salvato un file JSONL contenente, per ogni ricercatore, il valore dell'entropia binaria e la classe di appartenenza di tale entropia per ognuno degli anni in cui il ricercatore ha collaborato alla stesura di un paper. Tutti i file JSONL generati dalla procedura vengono quindi aggregati in un file JSONL rappresentativo di uno specifico framework.