Progetto di Social Computing

Relazione sulla prima parte

Borsoi Allison
mat. 147566
borsoi.allison@spes.uniud.it

Bosoppi Stefano
mat. 153845
bosoppi.stefano@spes.uniud.it

Canciani Mauro
mat. 150260
canciani.mauro@spes.uniud.it

Costa Claudia
mat. 153256
costa.claudia@spes.uniud.it

a.a. 2022-2023

1 Introduzione

In questa relazione mostreremo come, attraverso lo scaricamento del maggior numero di informazioni significative sulla rete sociale in esame, ci è stato possibile confermare come la rete di utenti di Twitter che si dirama dal profilo di *@KevinRoitero* sia a tutti gli effetti una rete sociale reale.
Questo verrà dimostrato prima spiegando come abbiamo raccolto, quindi scaricato ed immagazzinato, i dati sugli utenti d'interesse di Twitter e poi analizzando, giustificando opportunamente le assunzioni fatte, la struttura e la topologia delle reti sociali da questi derivate.

2 Raccolta dei dati

In questa sezione verranno illustrate le nostre scelte implementative nella fase, da noi definita, di *raccolta dei dati*, ossia il reperimento, attraverso la libreria Tweepy, e la successiva serializzazione delle informazioni sugli account Twitter d'interesse.

2.1 Reperimento completo e distribuito

In aggiunta a quanto prescritto dal punto 3 della consegna, abbiamo scaricato i dati su tutti i follower dei follower di *@KevinRoitero*. Nella sezione 3, vedremo come questa decisione ci ha permesso di avere una visione completa della rete sociale dell'utente in esame nella sezione 3.

Quantità	130
Media	4 633
Deviazione standard	33 759
Minimo	0
Primo quartile	125
Mediana	292
Terzo quartile	918
Massimo	373 113

Tabella 1: Indici statistici sui seguaci dei follower con profilo pubblico di *@KevinRoitero*.

Per semplificare la trattazione, definiamo come *follower di primo grado* gli utenti che seguono *@KevinRoitero* e come *follower di secondo grado* coloro che seguono uno o più follower di primo grado. Se un seguace è sia di primo grado, in quanto follower di *@KevinRoitero*, che di secondo grado, perché segue di almeno un altro follower di primo grado, sarà da noi considerato come follower di primo grado.
Considerato che, per le informazioni riportate nella tabella 1 la maggioranza dei follower di primo grado possiede un numero di seguaci inferiore a mille, per la stragrande maggioranza delle richieste inviate alla Twitter API, non è necessaria la paginazione dei dati[1]. Per al più il 25% di tali seguaci, tuttavia, questa funzionalità è indispensabile.
Abbiamo quindi approfondito l'analisi di quel quarto di follower di primo grado con un alto numero di follower. I risultati, esibiti nella tabella 2, mostrano che la quantità di follower di primo grado "popolari", ossia con un numero di follower superiore a 1 000, è marginale rispetto al totale considerato, circa il 23%, e che almeno la metà di essi presenta un numero di seguaci inferiore a 2 000. Tuttavia, è degno di nota il fatto che un quarto di questa minoranza presenta un numero di follower superiore a 5 000.
Queste informazioni, combinate ai tempi di attesa dell'endpoint utilizzato[2], ci hanno portato a distribuire il reperimento dei dati tra i diversi membri del gruppo. Abbiamo ritenuto, infatti, troppo oneroso, per un singolo

[1] Ogni pagina per la richiesta dei follower di un utente può contenere al massimo 1 000 risultati.[2]

client, eseguire l'intero processo di scaricamento autonomamente, in quanto, per via dei limiti alla frequenza delle richieste[2], avrebbe impiegato un tempo eccessivo.
La distribuzione effettuata consiste nella partizione dei follower di primo grado in quattro segmenti contingui. Il risultato sono quindi quattro porzioni, di cui tre hanno per dimensione il quoziente della divisione, tra il numero totale di follower di primo grado e il numero di membri del gruppo, e una la somma tra il quoziente ed il resto. Identificando ogni membro con un numero intero compreso tra 0 e 3, si possono concatenare i quattro risultati parziali per ottenere la lista con tutti i follower di secondo grado scaricati.
Il limite principale di questa suddivisione è che non è propriamente equa da un punto di vista di carico di lavoro: avere un numero di follower di primo grado simile non garantisce un numero di follower di secondo grado altrettanto omogeneo e, di conseguenza, è probabile che vi siano discrepanze nella quantità di richieste fatte da ogni client.
Nonostante ciò, abbiamo comunque ritenuto questa soluzione appropriata perché, per le osservazioni ricavabili dalle tabelle 1 e 2, nella maggior parte dei client, lo scaricamento dei profili dai follower di primo grado "popolari" non ha rallentato eccessivamente il medesimo processo per quei follower di primo grado non "popolari".

Quantità	30
Media	19 124
Deviazione standard	69 188
Minimo	1 029
Primo quartile	1 165
Mediana	1 879
Terzo quartile	5 300
Massimo	373 113

Tabella 2: Statistiche sui seguaci dei follower di *@KevinRoitero*, con profilo pubblico, aventi più di 1 000 follower.

2.2 La serializzazione finale

In questa sottosezione, approfondiremo il passaggio da un'organizzazione *ad albero* del JSON ad una *a tabella hash*. Le conclusioni a cui giungeremo in seguito sono valide ed applicabili per ambedue le raccolte considerate[2].
Per formattazione ad albero, intendiamo un file JSON contenente le informazioni sull'utente principale, *@KevinRoitero*, e sui suoi follower di primo e secondo grado[3].
In questa struttura: l'utente di partenza, nel nostro caso *@KevinRoitero*, ricopra il ruolo di radice; i follower di primo grado siano "figli" dell'utente radice e i follower di secondo grado sono a loro volta "figli" di quelli di primo.
Questa organizzazione si è rivelata particolarmente comoda per lo scaricamento descritto nella sottosezione 2.1, ma non può essere mantenuta come versione finale per due motivi: in primo luogo non è conforme a quanto richiesto nel punto 3 della consegna, inoltre presenta una forte ridondanza delle informazioni dei follower di primo grado.
Per risolvere sia il problema della non conformità con la richiesta, che della ridondanza, abbiamo adottato una struttura che abbiamo definito *a tabella hash*. Questa formattazione, oltre ad essere conforme a quanto richiesto, riduce notevolmente la quantità di informazione duplicata: solamente l'identificativo univoco, *id*, viene ripetuto. Abbiamo denominato questa organizzazione *a tabella hash* perché minimizza i tempi di accesso alle informazioni di un qualsiasi utente: basta essere in possesso della chiave, l'*id*, e si può accedere quasi istantaneamente ai dati di un qualsiasi utente memorizzato. Proprietà particolarmente utile per lo svolgimento del punto 4.

3 Costruzione ed analisi dei grafi

In questa sezione discuteremo della creazione del grafo con attaccamento preferenziale (*preferential-attachment*) e del calcolo di alcune metriche rilevanti per i grafi in esame. Faremo inoltre un'analisi per confrontare i quattro grafi prodotti a partire dai dati descritti nella sezione 2, seguendo i punti 4 e 5 della consegna.
Al fine di facilitare la lettura, definiamo: il *grafo diretto parziale* come il grafo costruito a partire dai dati raccolti nel punto 3 della consegna, seguendo quanto prescritto dal punto 4; il *grafo diretto totale* analogo al precedente, ma generato dai dati reperiti nella sottosezione 2.1; il *grafo indiretto parziale* come il grafo realizzato dal grafo diretto parziale, in ottemperanza al punto 5 della consegna ed, infine, il *grafo diretto totale* che differisce dal precedente solo per il fatto che è stato realizzato a partire dal grafo diretto totale.

3.1 Il grafo con attaccamento preferenziale

Per costruire il grafo di cui al punto 5 della consegna, abbiamo utilizzato il modello di Barabàsi-Albert[1] in quanto abbiamo ipotizzato un attaccamento preferenziale lineare. Non avendo vincoli sul tipo di crescita della rete sociale, ci siamo limitati a modellare la situazione più semplice.

[2]Si intendono la raccolta di follower prevista dal punto 3 dalla consegna e quella descritta nella sottosezione 2.1.
[3]Per la definizione di follower di primo e di secondo grado si faccia riferimento alla sottosezione 2.1.

3.2 Le metriche di distanza

Essendo i due grafi diretti non fortemente connessi, come si vedrà nella sottosezione 3.3, vi sono dei nodi u e v per cui non esiste un cammino che li connetta. Da questo ricaviamo che l'eccentricità, ossia la massima distanza di un nodo da qualsiasi altro vertice[3], dei due nodi è infinita.
Siccome la determinazione del centro, del raggio e del diametro si basa sul valore dell'eccentricità di ciascun nodo di un grafo[4][3], abbiamo preferito calcolare queste metriche sulla più grande componente fortemente connessa di ciascun grafo diretto, in modo da ottenere valori non degeneri.

3.3 Analisi comparativa dei grafi

Nella figura 1 si può osservare un confronto tra la distribuzione dei gradi dei grafi diretti ed indiretti. Notiamo che i nodi dei grafi totali presentano, in generale, un grado lievemente più alto rispetto alle loro controparti parziali, fenomeno dovuto al fatto che i grafi totali, per costruzione[4], hanno un maggior numero di archi. Ciò che accumuna tutti i grafi è, però, la presenza di una distribuzione dei gradi molto simile ad una distribuzione *power-law*.

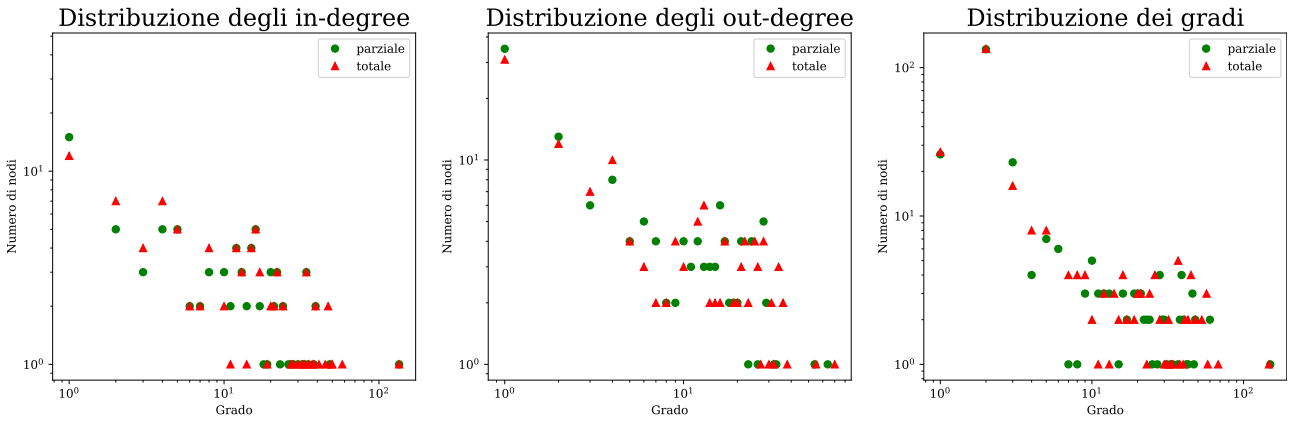

Figura 1: Rappresentazione delle distribuzioni degli *in-degree* e *out-degree* per i grafi diretti e dei gradi per i grafi indiretti.

L'aumento del numero di archi ha anche delle ricadute sulla massima componente fortemente conessa (componente gigante). La figura 2 evidenzia che per i grafi diretti, la dimensione di tale componente del grafo completo, 96 nodi, è maggiore rispetto a quella del grafo parziale, 89 nodi. Infatti, il cogliere tutte le relazioni di following tra i seguaci di primo grado[5] si traduce, nel grafo diretto totale, nella formazione di nuovi cammini tra due nodi e, di conseguenza, può rendere mutualmente raggiungibili due utenti che nella rete parziale non lo erano, aumentando quindi le dimensioni della componente gigante.

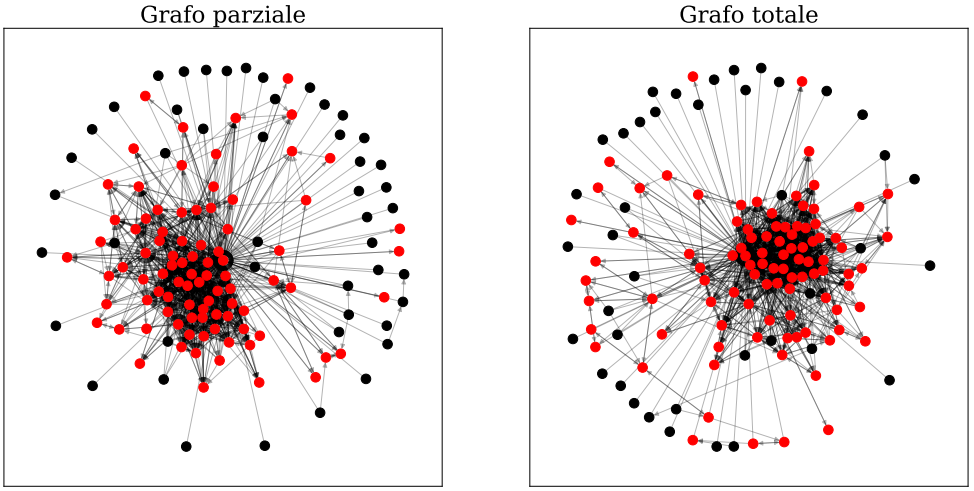

Figura 2: Rappresentazione delle più grandi CFC, i cui nodi sono colorati in rosso, dei grafi diretti.

[4]Si veda la sottosezione 2.1 ed il punto 4 della consegna.
[5]Si veda la sottosezione 2.1 per la definizione.

Per quanto riguarda le distanze dei quattro grafi, dai dati raccolti nella tabella 3, possiamo osservare che, per i grafi diretti, sono state confermate le osservazioni precedenti, in quanto l'aumento del raggio e del diametro dalla componente gigante del grafo parziale a quella del grafo totale sono una diretta conseguenza dell'aggiunta nuovi cammini tra nodi. Ciò che può sorprendere, però, è l'aumento della distanza media per il grafo diretto totale. Tale fenomeno può essere dato dal fatto che, rendendo più nodi raggiungibili, al grafo totale si aggiungono più cammini tra nodi e, in particolare, più cammini lunghi. Si transiziona quindi da una situazione in cui due nodi non sono raggiungibili, e quindi la loro distanza non è inserita nel calcolo della distanza media, nel grafo parziale, alla formazione di un cammino, anche lungo, tra questi nel grafo totale, questa volta incluso nel calcolo della media. Questa osservazione sembrerebbe anche giustificare la diminuzione del numero di raggi.
Nei grafi indiretti, invece, non c'è differenza tra raggio e diametro, mentre cala sensibilmente la distanza media e, nel grafo indiretto totale, il numero di centri è aumentato, diventando all'incirca un quarto dei nodi del grafo. Quest'ultimo dato è indicativo, in quanto mostra che nel grafo indiretto totale i nodi sono più "vicini" tra loro.

	Numero di centri	**Raggio**	**Distanza media**	**Diametro**
Grafo diretto parziale	11	3	1,370	4
Grafo diretto totale	4	4	1,547	7
Grafo indiretto parziale	48	3	2,560	5
Grafo indiretto totale	70	3	2,545	5

Tabella 3: Misure delle distanze dei quattro grafi considerati, dove raggio, distanza media e diametro sono misurati in nodi.

Passando ad un'analisi della centralità dei nodi dei quattro grafi, si osserva omogeneità delle metriche nel considerare *@KevinRoitero* come nodo centrale. In aggiunta, gli indici di posizione calcolati sulle metriche computate, sono abbastanza simili tra le coppie di grafi diretti ed indiretti. Ciononostante, riteniamo opportune alcune osservazioni: mentre nei grafi diretti i nodi con valori di *closeness* alti sono quelli appartenenti alla componente gigante, nei grafi indiretti sono quelli collegati con *@KevinRoitero*, in particolare quei nodi che sono collegati ad altri vertici che condividono un arco con lui, e, per quanto concerne *i punteggi di authority* e *di hubness*, anche qui, in tutte e quattro le reti, sono i nodi collegati a *@KevinRoitero* e, a loro volta, connessi tra loro ad avere il punteggio maggiore. Queste due peculiarità potrebbero trovare spiegazione nel fatto che l'utente esaminato ed i nodi precedentemente citati facciano parte di una cricca.

	ω	σ
Grafo indiretto parziale	0,139	1,339
Grafo indiretto totale	0,027	1,347

Tabella 4: Valori dei coefficienti ω e σ per i due grafi indiretti.

Siccome la presenza di cricche è una caratteristica delle reti *piccolo-mondo*, abbiamo, infine, calcolato i coefficienti ω e σ sui due grafi indiretti[6] per verificare questa ipotesi. La misura di "small-worldness"[5] (ω) è definita come

$$\omega = \frac{L_r}{L} - \frac{C}{C_l}$$

dove L e C sono, rispettivamente, la distanza media ed il coefficiente di clustering della rete che andiamo a considerare, mentre L_r è la distanza media di una rete casuale equivalente e C_l è il coefficiente di clustering di una rete regolare equivalente. Questo valore si basa sulla definizione di rete piccolo-mondo, infatti tende a 0 quando la rete ha una distanza media simile a quella di una rete casuale, ossia ha una distanza media bassa, ed un coefficiente di clustering simile a quello di una rete casuale, ossia ha un coefficiente di clustering alto.
Il coefficiente σ[5] è invece dato da

$$\sigma = \frac{\frac{C}{C_r}}{\frac{L}{L_r}}$$

dove C_r è il coefficiente di clustering di una rete casuale equivalente. Questo indice, a differenza del precedente, confronta la rete di partenza esclusivamente con una rete casuale equivalente e tende a valori maggiori di 1 quando la rete in esame è piccolo-mondo: questo si verifica, infatti, se ha un coefficiente di clustering alto e la distanza media è simile a quella della rete casuale equivalente.

[6] I coefficienti ω e σ possono essere calcolati solo su grafi indiretti in quanto il modello *piccolo-mondo* assume che le connessioni tra individui siano simmetriche.

Nella tabella 4, vediamo come tutti i coefficienti indicano che entrambe le reti, quella totale a maggior ragione, siano reti piccolo-mondo. Riteniamo che anche il valore di σ sia affidabile, in quanto la rete non è di grandi dimensioni, situazione in cui è nota la sua inattendibilità[5].

4 Conclusioni

Siamo quindi riusciti a dimostrare che le reti totali sia contengono delle informazioni che non sono di facile individuazione nella loro controparti parziali, ma soprattutto che comportamenti deducibili da queste ultime sono ben definite e facilmente osservabili nelle prime.
Ciò significa che il punto 3 della consegna non degrada eccessivamente la qualità dell'informazione fornita dalla rete ma, grazie a quanto fatto nella sottosezione 2.1, con una visione totale abbiamo potuto osservare la vera componente gigante della rete, e vedere le imprecisioni della controparte parziale sui centri, raggio e diametro, e come i nodi del grafo indiretto totale siano più "vicini" tra loro rispetto a quelli della rete parziale.
Si può quindi concludere, con alto grado di sicurezza, che la rete parziale e, ancor più definitivamente, la rete totale formata dai follower di primo e secondo grado di *@KevinRoitero* sia a tutti gli effetti una rete sociale reale, caratterizzata da una distribuzione dei gradi *power-law* e dalla conformità al modello piccolo-mondo.

Riferimenti bibliografici

[1] A. L. Barabási, R. A.: Emergence of scaling in random networks. In: *Science* 286 (1999), S. 509–512

[2] Twitter: *GET /2/users/:id/followers*. `https://developer.twitter.com/en/docs/twitter-api/users/follows/api-reference/get-users-id-followers`. Version: 2021. – Ultima visita nel 4 dicembre 2022

[3] Wikipedia: *Distance (graph theory)*. `https://en.wikipedia.org/wiki/Distance_(graph_theory)#Related_concepts`. Version: 2022. – Ultima visita nel 7 dicembre 2022.

[4] Wikipedia: *Graph center*. `https://en.wikipedia.org/wiki/Graph_center`. Version: 2022. – Ultima visita nel 7 dicembre 2022.

[5] Wikipedia: *Small world network*. `https://en.wikipedia.org/wiki/Small-world_network`. Version: 2022. – Ultima visita nel 7 dicembre 2022.