Network Analysis
Sending information through a network

Samuele Crea

June 2024

1 The Study

In questo Assignment discutiamo come un'informazione può diffondersi all'interno di un grafo reale in modo graduale, passando da un nodo ad un'altro attraverso gli edges.

L'idea dell'esperimento sta nel fatto che, all'inizio, vengono scelti un numero di **Gossipier** che contengono l'informazione e questi la inviano agli altri con un tasso di "contagio" (threshold) che varia da simulazione a simulazione.

Per semplicità i Gossipier verranno rappresentati dal colore verde mentre i nodi non ancora "informati" dal colore blu.
Esistono anche nodi di colore rosso, ovvero nodi "malicious" che contengono un'informazione falsa che differisce da quella originale.

La simulazione può finire in modi diversi (soprattutto a seconda del threshold scelto):

1. l'informazione raggiunge tutti i nodi possibili
2. lo spreading dell'informazione si blocca perchè i Gossipier non riescono a raggiungere la soglia di "contagio"

E' interessante nel nostro studio vedere come l'informazione di diffonde nel grafo "reale" e che ruolo hanno i nodi malicious.

2 The Graph

Per lo studio ho utilizzato lo stesso grafo del primo assignment, quello rappresentante le connessioni del social network GPlus.
Tuttavia quest'ultimo era davvero troppo grande, tanto da non permettermi di provare la simulazione di trasmissione dell'informazione.

Ho decisio di ridurre il grafo attraverso uno **snowball sampling**.
Questo tipo di sampling inizia con un set iniziale di nodi e aggiunge ricorsivamente i vicini fino a raggiungere il numero desiderato di nodi.

Il set di nodi iniziali dal quale ho iniziato il sampling non è casuale, infatti ho scelto i nodi più importanti basandomi sulle caratteristiche di degree, betweenness e closeness centrality.

('2300', '8306', '1876','2622') è quindi il set di nodi dal quale l'algoritmo è partito, arrivando fino al valore di 6000 nodi (un quarto di quelli iniziali).

Attraverso questo tipo di sampling è stato possibile ridurre notevolmente il grafo ma allo stesso tempo mantenerne invariate o simili tutte le metriche.
Nelle tre figure qui sotto sono rappresentate le metriche principali del grafo.

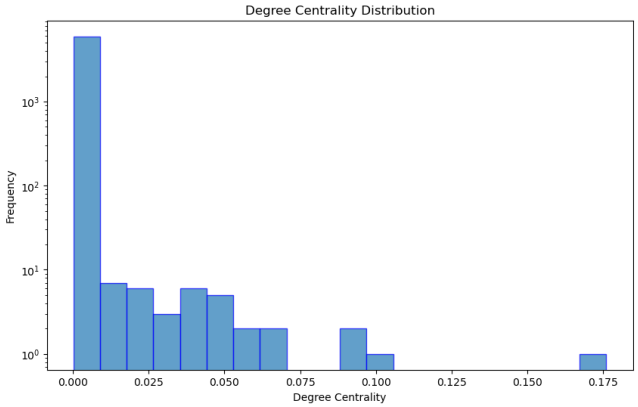

Figure 1: Degree centrality

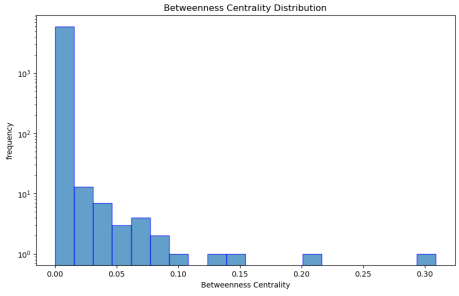

Figure 2: Betweenness centrality

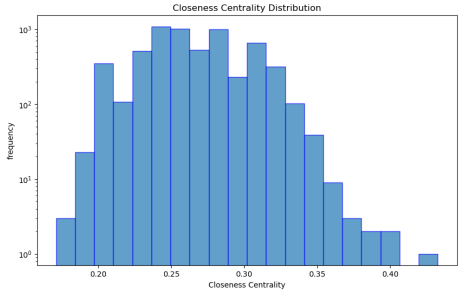

Figure 3: Closeness centrality

Come si può vedere dai grafici, le metriche sono rimaste molto simili al grafo iniziale, apparte ovviamente la dimensione dei nodi che è diminiuta drasticamente.

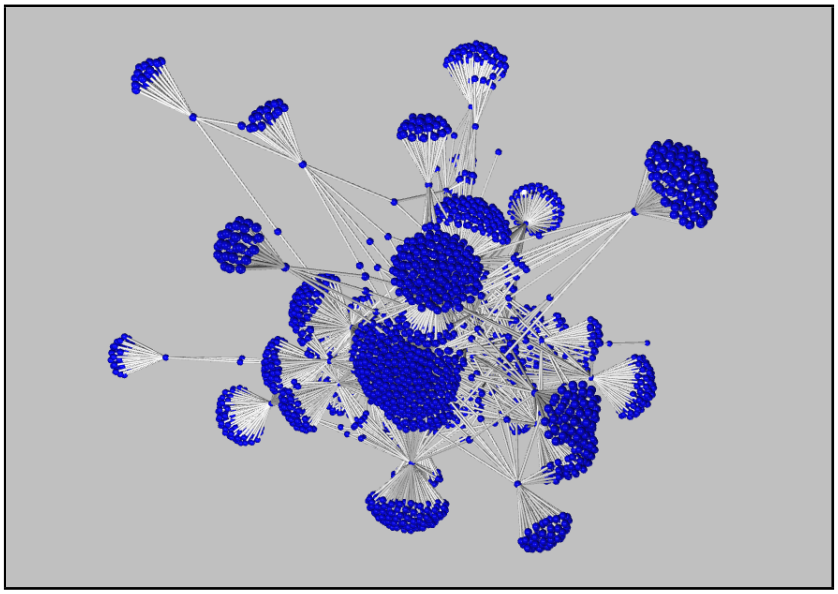

Figure 4: A better visualization of the graph

Come possiamo vedere dalla figura, il nostro grafo mantiente una struttura simile a quello di partenza e risulta solamente più "spoglio" rispetto a quello originale.

Con questo risultato è possibile adesso iniziare la nostra indagine sulla trasmissione dell'informazione nel nostro grafo.

3 Sending information through a network

3.1 Overview

Come già detto, la simulazione inizia con tutti i nodi di colore blu, il che significa che non possiedono alcun tipo di informazione.
Nel grafo vengono scelti poi dei nodi **Gossipier** e dei nodi **Malicious**, i primi verdi e i secondi all'inizio grigi e successivamente rossi.

Il ruolo dei nodi Malicious è particolare: in un primo momento questi sono a tutti gli effetti dei nodi "non informati" (grigi) come tutti gli altri ma nel

momento in cui un gossipier li informa, diventano nodi malevoli (rossi) che comunicano un'informazione modificata.

Come si potrà notare quindi l'aspetto iniziale del grafo è quello di una massa quasi completamente blu che lentamente, step dopo step, diventa verde e in parte anche rosso.

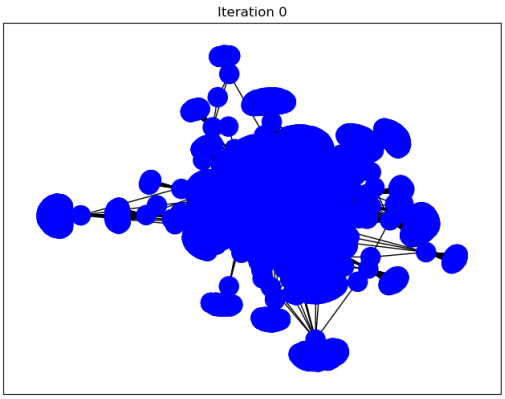

Figure 5: Initial Graph

3.2 Parameters

I parametri utilizzati dalla nostra simulazione sono principalmente 3:

- threshold: è un valore compreso tra 0 e 1, rappresenta la soglia di trasmissione dell'informazione. In altre parole, è la proporzione di vicini informati necessaria affinché un nodo non informato cambi stato e diventi informato.
- num gossipier: è il numero di nodi iniziale che contiene l'informazione e cerca di diffonderla
- num malicious: è il numero di nodi iniziale che contiene un'informazione malevola e cercherà di diffonderla dopo che verrà informato da un gossipier.

In più ogni ad nodo verrà associata una "informazione" che per i gossipier sarà la stringa "Hello world!" mentre per i nodi malicious sarà una versione della stringa stessa che però differisce di una lettera.

Il valore di threshold ha la sua funzione principalemente in una delle funzioni, quella che serve a trasmettere l'informazione.

Durante ogni iterazione del ciclo di diffusione dell'informazione, ogni nodo non informato verifica la proporzione di vicini che sono Gossipier. Se la proporzione di vicini informati supera la soglia (threshold), allora il nodo non informato cambia stato: se la maggioranza dei vicini informati sono malicious, allora il nodo diventa malicious altrimenti il nodo diventa gossipier.

3.3 Test with random gossipier

Volendo testare diverse situazioni all'interno del grafo ho deciso di run the script with different parameters.
Una prima simulazione è stata eseguita un threshold estremamente piccolo (0.01), che permetteva ad ogni round di inviare l'informazione a colpo sicuro.

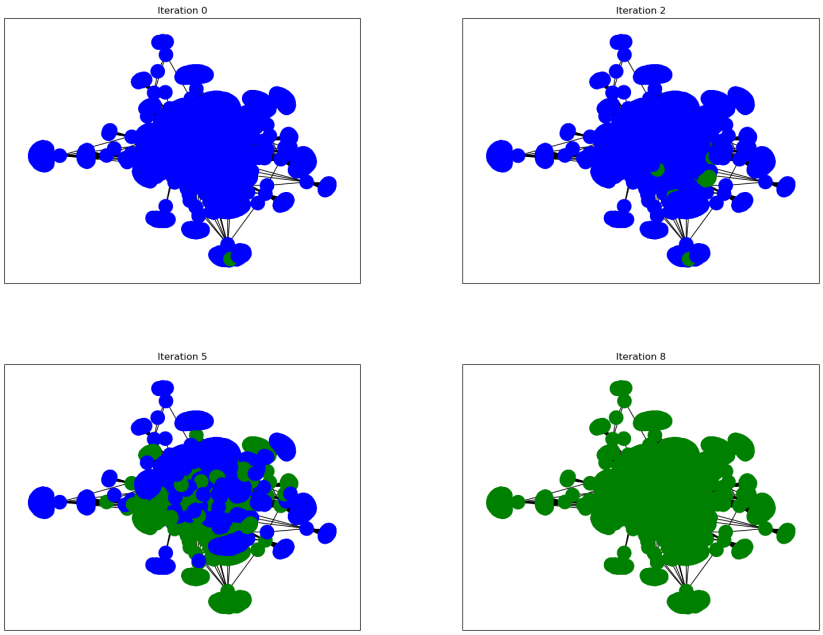

Figure 6: Spreading of the information

In questa simulazione il numero di nodi Gossipier e Malicious è davvero basso: rispettivamente 10 e 8.
Nonostante questo dopo 8 round, complice il threshold davvero basso, tutto il grafo è stato informato correttamente.
E' interessante come il numero di nodi malicious è aumentato poco durante la diffusione dell'informazione, probabilmente perchè come possiamo vedere i gossipier si sono concentrati in una piccola componente e i nodi malicious erano lontani da loro.

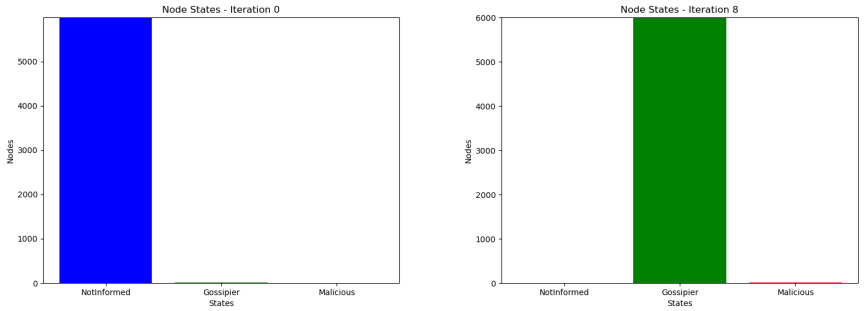

Figure 7: Number of Nodes

Dal grafico in figura 7 si nota meglio come il numero di Nodi Malicious sale, anche se di poco, durante la simulazione.

Passando ad un test più reale ho deciso di utilizzare un numero più sensato di gossipier e malicious, rispettivamente 600 (il 10 percento del nostro grafo) e 500.

Ho poi deciso di ripetere l'esperimento con due threshold diversi: 0.1 e 0.5

Come possiamo vedere in figura 8 qui sotto, nonostante la soglia di trasmissione piuttosto bassa non si è riuscito a diffondere l'informazione in tutto il grafo.

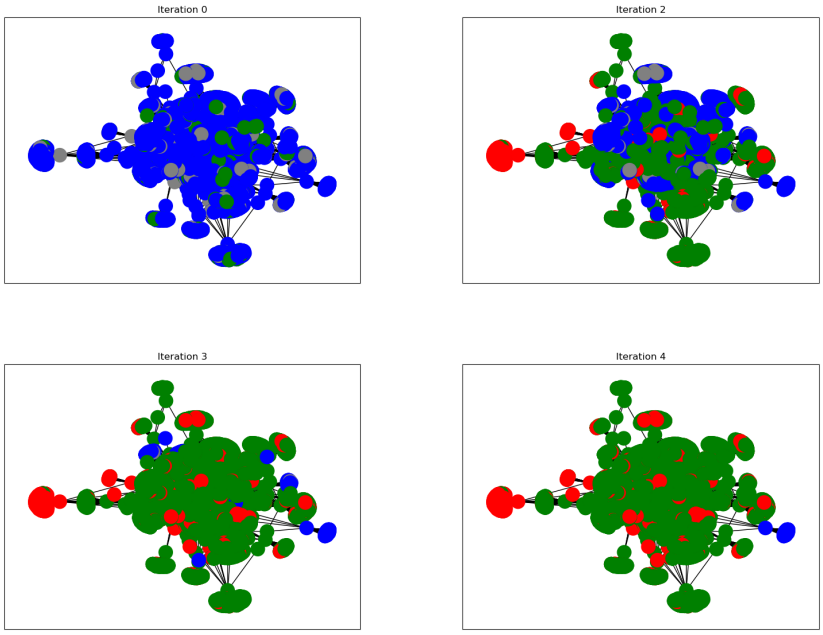

Figure 8: Spreading of the information

E' interessante come questa volta la diffusione di informazioni false è stata molto più marcata, ottenendo intere componenti del grafo come quella più a sinistra.

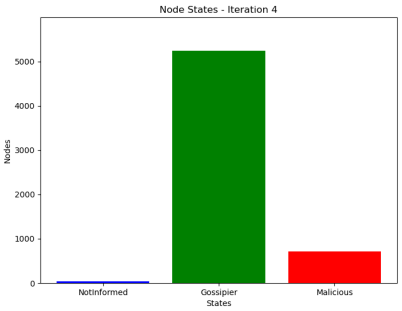

Figure 9: Number of Nodes

Come possiamo constatare dalla situazione alla fine della diffusione , questa volta i nodi malcious sono cresciuti notevolmente rispetto al primo esperimento.

Nel secondo esperimento è stato utilizzato un thresholld di 0.5, rendendo la diffusione più difficile.

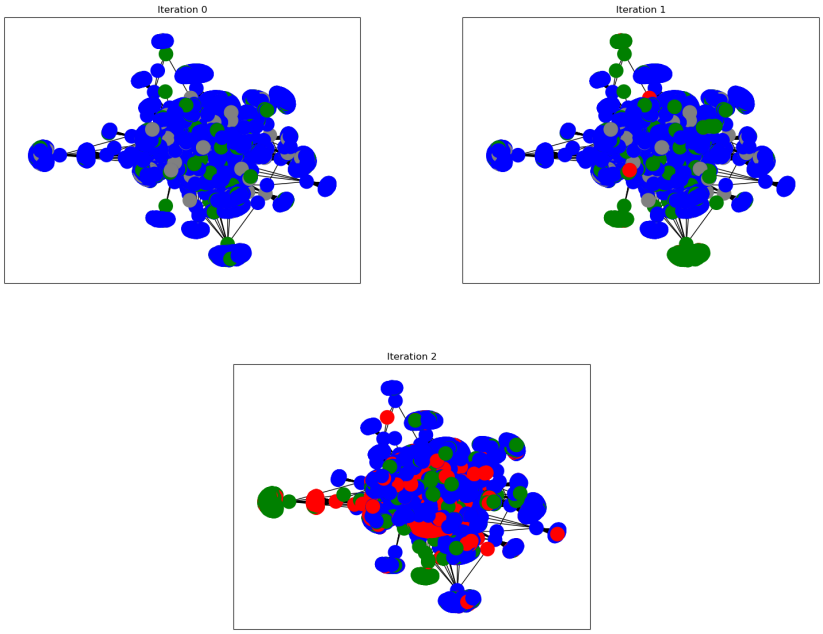

Figure 10: Spreading of the information

Come era immaginabile, con un threshold più altro i gossipier hanno fatto molta più fatica a diffondere l'informazione per poi fermarsi solo dopo 3 iterazioni.
Il risultato in figura 10 mostra come la maggior parte del grafo è rimasto non informato.

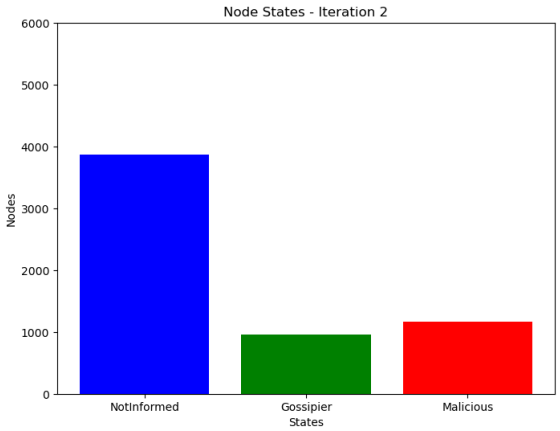

Figure 11: Number of Nodes

Un'altra informazione che traspare dal grafico in figura 11 è che, in questo caso il numero di nodi malicious finale ha superato il numero di gossipier "buoni".
Questo potrebbe essere un fatto casuale, ma è comunque interessante riportarlo.

3.4 Test using a choosen node

Un ulteriore test che ho deciso di fare è stato quello con un unico nodo e zero gossipier.

Questo tipo di test non sarebbe significativo se non fosse per la scelta del nodo, ovvero il 2300.
Questo nodo è risultato il più influnte nel grafo dall'analisi compiuta nel primo assignment.
Il nodo 2300 è quello con più collegamenti ed è quello con maggior degree, betweenness e closeness centrality.

Vediamo adesso come si comporta come unico "vettore" dell'informazione, con un threshol di 0.1 (piccolo ma come abbiamo visto non troppo).

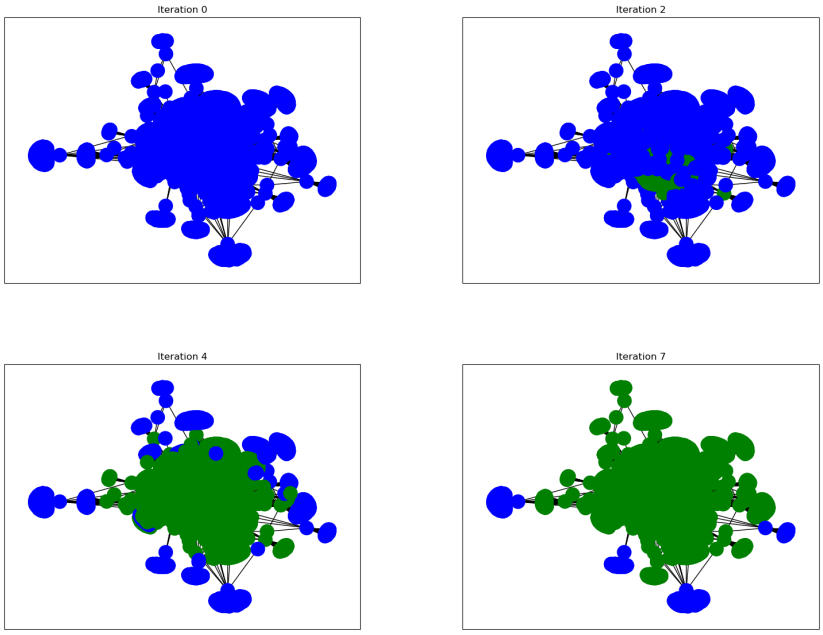

Figure 12: Spreading of the information

Come possiamo vedere nonostante la presenza di un solo gossipier la diffusione del messaggio è stata quasi completata, proprio perchè il nodo 2300 è quello in assoluto più collegato e influente nel grafo.

3.5 Cosine similarity

La cosine similarity è una misura di similarità tra due vettori in uno spazio multidimensionale.
Viene calcolata come il coseno dell'angolo tra i due vettori.

$$\text{cosine similarity}(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$

Nel contesto del nostro assignment, la cosine similarity viene utilizzata per misurare la similarità tra le informazioni (stringhe) conosciute dai nodi della rete. Le informazioni testuali vengono trasformate in vettori TF-IDF che rappresentano la frequenza e l'importanza delle parole.

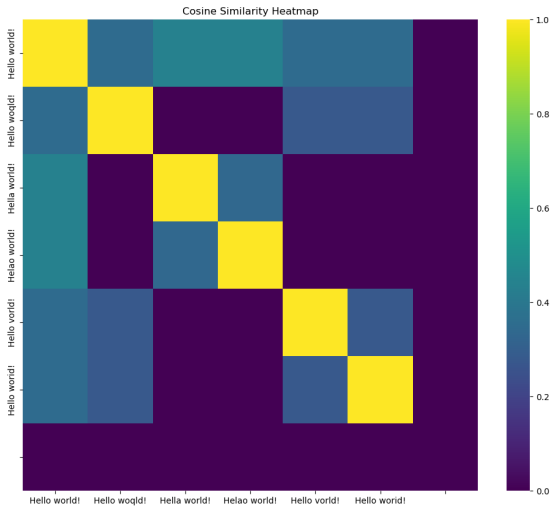

Figure 13: Cosine Similarity

La cosine similarity in figura è quella dell'esperimento con 600 gossipier e threshold di 0.1.

Le frasi "Hello world!" sono state riconosciute come identiche, evidenziando la massima similarità tra loro.
Ciò significa che ogni nodo che conosce questa frase avrà una rappresentazione vettoriale esattamente identica.

Le frasi con errori tipografici sono considerate meno simili alla frase originale, come indicato dai valori di cosine similarity inferiori, indicando che i nodi che conoscono queste varianti hanno rappresentazioni vettoriali meno simili a "Hello world!".

Questo tipo di rappresentazione permette di visualizzare chiaramente come l'informazione si è diffusa nel grafo e come le variazioni nell'informazione influenzano la similarità tra i nodi.