Fridays-For-Future's hashtag network

Silvia Cosmo
s.cosmo@studenti.unipi.it
Student ID: 628299

Marco D'Arrigo
m.darrigo2@studenti.unipi.it
Student ID: 563441

Naomi Esposito
n.esposito3@studenti.unipi.it
Student ID: 572797

ABSTRACT

Il Fridays for Future[1] non è più solo uno sciopero scolastico portato avanti da un gruppo di studenti, bensì oggi rappresenta uno dei movimenti ambientalisti più importanti nella scena internazionale e Twitter ne costituisce la piazza di scambio di informazioni ed interazioni.
Mediante la costruzione di un grafo basato sulla relazione degli hashtag e per mezzo di tecniche di network analysis, si tenta di analizzare l'evoluzione di una rete, che ingloba internazionalmente le principali tematiche politiche dell'attualità.

KEYWORDS

Social Network Analysis, Twitter, hashtags, FridaysForFuture

ACM Reference Format:
Silvia Cosmo, Marco D'Arrigo, and Naomi Esposito. 2022. Fridays-For-Future's hashtag network. In *Social Network Analysis '22*. ACM, New York, NY, USA, 12 pages.

1 INTRODUCTION

La piattaforma che abbiamo scelto per effettuare le analisi oggetto del presente lavoro è Twitter[2].
Social network di microblogging, costituisce un mezzo di diffusione di notizie ed informazioni.
L'attività di ciascun utente è principalmente costituita dalla condivisione di brevi messaggi (280 caratteri), caratterizzati dalla presenza di uno o più **hashtag**: parole chiave o combinazione di parole concatenate precedute dal simbolo del cancelletto .
Tipo di tag metadato, l'hashtag è il principale strumento di diffusione e ricerca delle tematiche sull'intera piattaforma. Rappresenta, infatti, un mezzo che implementa l'indicizzazione, la categorizzazione e la ricerca dei contenuti.
In Twitter l'uso degli hashtag non è obbligatorio, tuttavia l'uso consente la diffusione dei tweet pubblicati, aumentando la possibilità di ottenere più interazioni con un pubblico che vada anche oltre la cerchia dei propri seguaci (followers).
In virtù di ciò l'obiettivo della presente ricerca è quello di effettuare un'analisi degli hashtag di Twitter, riconducibili alle tematiche del **Fridays For Future** (FFF), al fine di ottenere informazioni circa la diffusione di argomenti affini all'interno della comunità degli utenti, rappresentando gli hashtag in un grafo.
I nodi sono costituiti degli hashtag, tra i quali esiste un edge in comune se esiste almeno un tweet in cui sono presenti insieme. Il numero di volte in cui due hashtag co-occorrono, costituisce il peso dell'edge.
Nel corso dell'analisi vengono studiate le peculiarità di questa rete, confrontandola con alcuni modelli sintetici quali: **Barabasi-Albert, Erdos-Renyi e Watts-Strogatz**.
La ragione della scelta di questa tematica, risiede nel fatto che ad essa sono collegati altrettanti argomenti di attualità, offrendo una panoramica su una community che trova spazio in uno scenario globale.
La relazione si sviluppa descrivendo innanzitutto le modalità d'acquisizione dei dati, con annesse le principali informazioni statistiche e la costruzione della rete.
Si passa in seguito all'analisi delle caratteristiche della rete reale e il confronto con i principali modelli sintetici sopra elencati. Successivamente si espongono gli approcci di community discovery, spreading e link prediction.
Su quest'ultimo in conclusione si espone un approfondimento attraverso il quale ci si propone di aggiungere ai nodi-hashtag i rispettivi significati, andando a creare un sistema di previsione di link per Twitter.

2 DATA COLLECTION

La rete oggetto dell'analisi è stata costruita per mezzo di dati estratti dalla piattaforma *Twitter*, utilizzando la libreria **TWINT**[3] (Twitter Intelligence Tool), uno scraping tool scritto in Python che permette di ottenere i dati senza l'uso dell'API.
Si è optato per l'uso di questo tool, in quanto esso permette di non avere rilevanti limitazioni nell'attività di scraping e crawling data. Inoltre, è risultato agevole nell'impostazione dei parametri per l'identificazione del campione oggetto del nostro studio.
Nella fase iniziale di crawling data infatti, si è indicato un limite massimo di tweet da scansionare (50000) nelle date comprese tra il 2019 e il 2022. A queste indicazioni sono state aggiunte ulteriori condizioni riguardanti le caratteristiche dei tweet.
Al fine di ottenere una rete ben connessa, si è optato di escludere i tweet che non presentassero alcun tipo di interazione (*likes, replies, retweet*).

[1] **Project Repositories**
Data Collection:
Data Collection link
Analytical Tasks:
Analytical Tasks link
Report:
Report link

[2] https://twitter.com/

[3] https://github.com/twintproject/twint/wiki/Basic-usage

SNA '22, 2021/22, University of Pisa, Italy

Questi ultimi parametri sono stati successivamente modificati ulteriormente, aumentandone i valori al fine di ottenere un ridimensionamento della quantità di tweet da prendere in analisi. Infine, lo studio è stato effettuato utilizzando come keyword *"FridaysForFuture"*.
La scelta di svolgere la ricerca sulla base di una parola chiave, anziché di un hashtag, ci ha permesso di ottenere tweet contenenti due o più tag afferenti all'argomento "FFF" ed ad altri affini.
É, infatti, quanto emerso dall'analisi linguistica svolta sui tweet del nostro dataset, attraverso la quale sono stati estrapolati i primi venti trigrammi più frequenti; dati riportati in Figura [1].

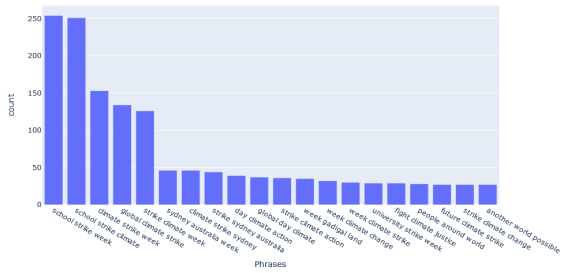


Figure 1: Trigrammi più frequenti nei Tweet

Dai dati ottenuti è stata creata una rete **Tag-to-Tag**, i cui nodi sono hashtag collegati da un edge se co-occorrenti in almeno un tweet. É stato inserito come peso dell'arco il numero di tweet in cui i tags erano presenti contemporaneamente.
Nella seguente Tabella [1] vengono riportate le descrizioni dei dati estratti e, a seguire, in Tabella [2] è possibile visualizzare le informazioni statistiche di partenza.

Hashtag	Lista di tag presenti in un tweet
User id	Id associato a ciascun utente autore del tweet
Username	Username di ciascun utente autore del tweet
Tweet	Testo dei tweet
Language	Lingua di ciascuno dei tweet
Mention	Menzioni presenti in ciascun tweet
Replies count	Numero di replies
Retweet count	Numero di retweet
Likes count	Numero di likes
Link	Link di ciascun tweet

Table 1: Descrizione dei dati scaricati

Numero totale di tweet	23504
Numero di hashtags singoli	10361
Numero medio di hashtag per tweet	3.07

Table 2: Statistiche del Dataset

Già dalle iniziali informazioni statistiche emerge un'ampia varietà di hashtags, come viene raffigurato nella WordCloud in Figura [2], sebbene alcuni di questi emergano per la loro frequenza d'uso. Nello specifico gli hashtags "**climatestrike**", "**climateaction**", "**climatecrisis**", "**climatestrikeonline**", presentano una frequenza che varia in un range tra 3000 e 1000 occorrenze.

Figure 2: WordCloud dei Tags

L'unico hashtag con frequenza superiore risulta essere "fridaysforfuture" con 10524 occorrenze: dato condizionato dalle modalità di crawling e scraping effettuate per mezzo di twint.
L'elevata pluralità degli hashtags viene confermata dal fatto che circa il 99% appare, infatti, meno di 100 volte, suggerendo la presenza di una distribuzione pronunciatamente distorta.

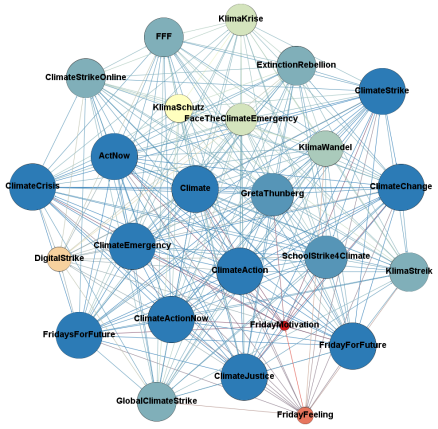


Figure 3: FridaysForFuture's Network

In Figura [3] è possibile visualizzare il grafo della rete. Al fine di ottenere una rappresentazione più limpida del grafo, ed avere una panoramica della sua struttura, sono stati filtrati i nodi aventi un degree inferiore a 300.
Da ciò, ne è risulta una rete bidirezionale con 10361 nodi e 63960

archi, come è possibile osservare in Tabella [3], nella quale sono riportate le caratteristiche più rilevanti della rete.

In particolare, è possibile notare che si tratta di un grafo sparso dal momento che:
- L « LMAX[4];
- <k>[5] « N-1;
- d(G)[6] « 1.

Infatti, se il grafo fosse stato completo si avrebbe avuto i seguenti parametri: **L = LMAX**, grado medio <**k**> pari a **N - 1 = 10360** e densità di rete **d(G) = 0.001**.
Inoltre, la rete rientra nel regime **connected**, in quanto il grado medio è maggiore di **ln(N) = 9.25**.
Ulteriori analisi ed informazioni saranno presentate nelle sezioni successive.

Number of nodes N	10361
Number of edges L	63960
LMAX	53669980
Average Degree <k>	12.3
Average Clustering Coefficient	0.808
Density d(G)	0.001

Table 3: Statistiche della rete

3 NETWORK CHARACTERIZATION

In questa sezione è stata confrontata la rete costruita (RW - *Real World Network*) con tre modelli sintetici, creati con algoritmi predefiniti, i cui parametri sono stati impostati per avere lo stesso numero di nodi della rete originale e una quantità simile di edges, analizzandone similitudini e differenze con lo standard, al fine di dare una prospettiva più profonda della composizione della rete.
Sono riportati in Tabella [4] i valori di nodi ed archi dei modelli considerati:
- **Barabási-Albert (BA)** con **m** = 6.99.
- **Erdös–Rényi (ER)** con **p**[7] = 0.001027.
- **Watts-Strogatz (WS)** con **k** = 12 e **p** = 0.1

	BA	ER	WS
Number of nodes	10361	10361	10361
Number of edges	72478	55123	63960

Table 4: Reti sintetiche

[4] $LMAX = \frac{N*(N-1)}{2}$
[5] $< k >= \frac{2*L}{N}$
[6] $d(G) = \frac{L}{LMAX}$
[7] Il valore di p è scelto pari alla densità della rete iniziale

3.1 Degree Distribution Analysis

Osservando la degree distribution delle reti in Figura [4] è possibile affermare che la rete è abbastanza simile al modello Barabási-Albert, mentre appare molto diversa dalle reti Erdös-Rényi e Watts-Strogatz.
É stata riscontrata la presenza di hub nella rete: di fatti, la creazione di collegamenti tra i tags segue un processo di **preferential attachment**.
I 6 tag con grado maggiore, superiore a 1000, sono risultati essere *fridaysforfuture* (8976), *climatestrike* (2081), *climateaction* (1302), *climatecrisis* (1077), *climatechange* (1034) e *fridayforfuture* (1032), inerenti la questione del cambiamento climatico e le azioni del movimento in risposta al problema.
Il grado minimo è risultato essere pari ad 1 per tags estremamente rari: probabilmente perché legati a tematiche particolarmente specifiche, o potenzialmente legati ad un errore durante l'inserimento, o dovuti alla creazione di nuovi tag. Non è stata quindi riscontrata la presenza di nodi isolati, cioè tag con grado pari a 0.
I tags della rete seguono una distribuzione **power law**, poiché vi è un rapporto inversamente proporzionale tra il numero dei nodi e il loro grado.
Inoltre, considerando i valori del grado massimo e del grado minimo, è risultato un **gamma** compreso tra 2 e 3 (**gamma = 2.016**[8] permettendo di collocare la rete nell'**ultra-small world** del regime **scale-free**.

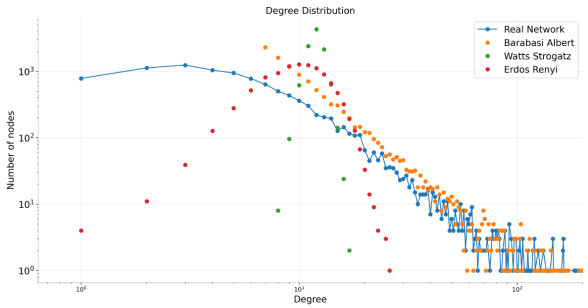


Figure 4: Degree Distribution

3.2 Connected Components Analysis

La rete analizzata è composta da un totale di 48 componenti connesse, tra le quali è emerso una **giant component** composta da 10257 nodi (99% dell'intera rete), mentre ogni altra componente comprendeva al massimo 3 nodi. I modelli sintetici hanno ottenuto una singola componente connessa, tranne il modello *Erdos-Renyi* che aveva un nodo non connesso.

3.3 Path Analysis

Per la componente più grande di ogni rete sono stati calcolati **average shortest path** e **diametro**, riportati in Tabella [5].

[8] $kmax = kmin * N^{\frac{1}{\gamma-1}}$

La rete utilizzata ha riportato valori di diametro simili alle reti sintetiche (il maggiore, pari ad 8, è stato riscontrato nell'analisi della rete Watts-Strogatz), mentre l'average shortest path ha un valore inferiore a quello delle altre reti, e inferiore a $ln(N) = 9.25$.
Considerando i pesi sui nodi(count) per la rete reale, il percorso minimo medio è risultato essere pari a 2.79.

	RW	BA	ER	WS
Diameter	7	5	7	8
Average	2.23	3.37	4.17	5.54

Table 5: Statistiche del Path

3.4 Clustering coefficient - Density analysis

Il coefficiente di clustering globale del grafo è 0.8, valore più alto di quelli ottenuti nelle reti sintetiche (come osservabile in Tabella [6]).
Il coefficiente di clustering medio della rete ER è risultato pari alla probabilità p come atteso, con un ordine di magnitudine decisamente inferiore al valore per la rete reale.
Il modello random quindi fallisce nel catturare la **degree distribution** della rete reale e anche il coefficiente di clustering[9].
Inoltre, osservando il rapporto tra coefficiente di clustering locale e il grado dei nodi è stato evidenziato come i nodi con un grado superiore presentavano un coefficiente di clustering molto basso, mentre i nodi di grado inferiore tendono ad assumere valori di coefficiente di clustering in tutto lo spazio di valori (Figura [8]).
La densità delle reti sintetiche BA e ER, è risultata essere simile (con variazioni dopo la 4 cifra decimale) alla densità della rete di Twitter, riportata in Tabella [3]. Il valore molto basso della densità (prossimo a 0.001) per tutte le reti è giustificato dalla presenza di un numero di archi molto inferiore al numero massimo possibile **N(N-1)/2**, prossimo a 53 milioni.

RW	BA	ER	WS
0.808	0.008	0.001	0.498

Table 6: Clustering Coefficients

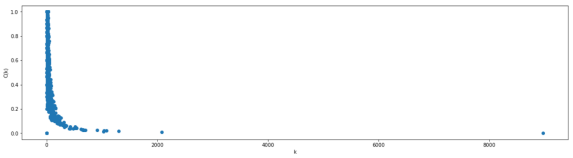


Figure 5: Local clustering coefficient vs Node degree

[9]Nelle reti random per posizionare gli archi ci si basa unicamente sulla probabilità p ed ogni arco è posizionato indipendentemente dagli altri, quindi la probabilità di chiudere un triangolo nella rete è molto bassa.

3.5 Centrality analysis

La centrality della rete è stata analizzata con metodi basati su diverse definizioni di **centralità**: **degree based**, **connectivity based (Eigenvector, PageRank)**, **geometric based (closeness, harmonic, betweenness)**.
In Figura [6] è possibile osservare i risultati di tali analisi per i primi 15 nodi ed, in particolare, la differenza tra i due metodi **geometric based** (Closeness e Harmonic) e gli altri metodi in termini di tags più centrali.
Quasi tutti i metodi hanno riportato *'fridaysforfuture', 'climatestrike' e 'climateaction'* come nodi più centrali, rappresentanti, quindi, i tag con maggior numero di *likes, retweet e commenti.*
Una differenza è stata riscontrata con il metodo Eigenvector, il quale ha riportato i tag 'climateemergency' e 'climatestrikeonline' con una centrality maggiore rispetto agli altri metodi, come riportato in Figura [6].
L'**Eigenvector Centrality**, infatti, misura la reputazione di un nodo e il riconoscimento della stessa che gli viene dato dagli altri nodi. In questo contesto i nodi come quelli sopra citati forniscono maggior riconoscimento al gruppo *Action* poiché si tratta di tags legati a questo tema.

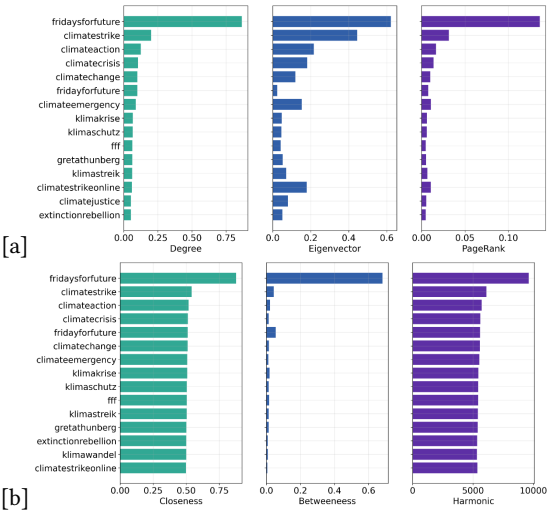


Figure 6: Primi 15 tags per centralità

4 STATIC COMMUNITY DISCOVERY

L'obiettivo di questa sezione è quello di provare ad identificare le comunità nascoste all'interno della rete, servendosi dell'utilizzo di sei metodi: **Louvain**[10], **Label Propagation**[11], **Demon**[12], **Infomap**[13], **Greedy Modularity**[14] **e K-Clique**[15].
Inoltre, è stato applicato anche l'algoritmo K-Cliquee[0] ma a causa dell'alto costo computazionale si è optato per eseguirlo su un campione random di 9000 nodi della rete originale.
Grazie alla libreria cdlib, è stato possibile effettuare per alcuni metodi una fase di ottimizzazione degli iperparametri attraverso una random search, avente come obiettivo la massimizzazione della **Modularità**.
Bisogna sottolineare che i medoti Demon e K-Clique, in quanto individuano **overlapping communities**, danno la possibilità ad ogni singolo nodo di appartenere a più di una comunità; a differenza degli altri metodi che, al contrario, individuano comunità di tipo **crisp**.
All'interno della Tabella [7] è possibile visionare i range contenenti i valori utilizzati durante la fase di ottimizzazione e il risultante valore ottimo trovato.

Algoritmo	Iperparametri	Val. ottimale
Louvain	*resolution* $\in [0.1, 1]$	0.9
K-clique	$k \in [2, 8]$	k = 4
Demon	*epsilon* $\in [0.1, 0.6]$	epsilon = 0.5
	min_com_size $\in [3, 5]$	min_com_siz = 4

Table 7: Scelta iperparametri

La ricerca del livello ottimale per ogni iperparametro è stata effettuata con l'obiettivo di massimizzare **Average Internal Degree**[16] e **Internal Edge Density**[17] e minimizzare la **Conductance**[18]; eccetto per l'algoritmo Infomap che, al contrario rispetto agli altri metodi, utilizza la massimizzazione della Conductance.
Per la scelta del valore migliore per il parametro *resolution* dell'algoritmo Louvain, come è possibile osservare nella Tabella [8], il valore ottimo risulta essere 0.9.
Infatti, a parità di numero di comunità ottenute, tale valore riesce a raggiungere complessivamente livelli maggiormente performanti inerenti i seguenti valori: AID, EID, Modularity e Conductance.

Resolution	0.1	0.3	0.5	0.7	0.9
Mododularity	0.083	0.163	0.256	0.338	0.4801
Cononductance	0.790	0.769	0.761	0.499	0.2435
AID	0.816	1.124	1.281	2.644	3.202
IED	0.414	0.470	0.510	0.936	0.834
N° Community	1449	1601	1839	933	185
Max. dim. community	7348	6309	4777	5546	3798

Table 8: Risultati Louvain

	Louvein	LabelP.	Demon	Infomap	GM
Mod.	0.48	0.028	0.19	0.39	0.41
Cond.	0.24	0.30	0.91	0.46	0.37
AID	3.202	2.097	8.936	3.220	2.966
IED	0.834	0.956	0.655	0.634	0.683
N°com.	185	172	230	706	236
Max_d._com.	3798	9802	6955	594	1002

Table 9: Valori principali degli algoritmi

All'interno della Tabella [9] è possibile osservare i risultati ottenuti per tutti gli algoritmi utilizzati, eccetto per il K-clique per i motivi sopra citati.
Infatti, il peggior valore della modularità è stato riscontrato con l'algoritmo K-Clique, anche se non è confrontabile direttamente con quello degli altri modelli a causa delle semplificazioni apportate prima della sua esecuzioni e con il quale è stato ricoperto solo il 78% del totale dei nodi; va inoltre sottolineato che anche il modello Demon, essendo un algoritmo con overlapping come K-Clique, ha ricoperto il solo l'87% dei nodi.

In Figura [8] viene riportato il boxplot della distribuzione della dimensione delle comunità prodotte dagli algoritmi.
Label Propagation è risultato avere una comunità di grandi dimensioni (comprendente il 94,6% dei nodi) e molte comunità di piccole dimensioni, ottenendo un risultato sbilanciato.
Dalla Figura [8], è possibile fare alcune osservazioni più dettagliate.
L'algoritmo Demon risulta aver prodotto più comunità di grosse dimensioni, un risultato analogo è stato prodotto anche dall'algoritmo Greedy Modularity, rispetto agli altri algoritmi.
L'infomap, invece, è risultato avere non solo il maggior numero di communities, ma il 63% di esse sono di dimensioni che vanno da 1 a 10 nodi.

[10] L'algoritmo Louvain fonde iterativamente comunità al fine di massimizzare come metrica obiettivo la modularità.
[11] L'algoritmo Label Propagation assegna ad ogni nodo un'etichetta, la quale, propagandosi attraverso la rete, viene aggiornata sulla base dell'etichetta di maggioranza dei vicini del nodo.
[12] L'algoritmo Demon si basa sul concetto di "Ego Network", cercando di creare comunità tramite l'ottenimento per ogni nodo della sua "Ego Network", ovvero la rete collegata direttamente al nodo, escludendo successivamente il nodo stesso.
[13] L'algoritmo Infomap si basa sull'idea di usare un approccio bottom-up, dove ogni iterazione computa la descrizione di un percorso identificata da un percorso random nel grafo, dove è possibile clusterizzare il percorso incontrando un community e poi facendo partire da capo la simulazione del percorso random. Differentemente dagli altri metodi, cerca di massimizzare la Condunctance, poiché un basso valore indica la presenza di pochi ponti all'interno della community.
[14] Utilizza la modularità per trovare le strutture delle comunità. Ad ogni passo dell'algoritmo vengono unite due comunità che contribuiscono al massimo valore positivo alla modularità globale.
[15] L'algoritmo K-Clique, dato un certo k, individua le clique presenti nel grafo e unisce in un unico insieme le clique che possiedono k-1 nodi in comune.
[0],
[16] Misura la bontà di una community a seconda del grado medio dei nodi appartenenti a quella community
[17] Determina la bontà di una community in base alla maggiore densità di link all'interno della stessa
[18] Definisce la bontà in base alla probabilità di uscire dalla community,eseguendo un cammino casuale tra un nodo i e un nodo j di una stessa community

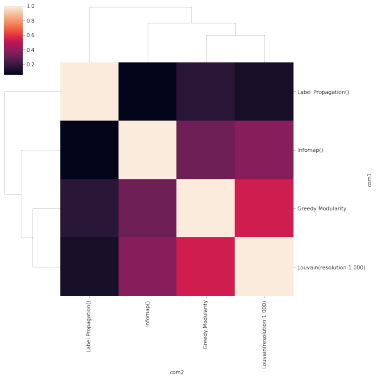


Figure 7: Heatmap per confrontare gli algoritmi migliori

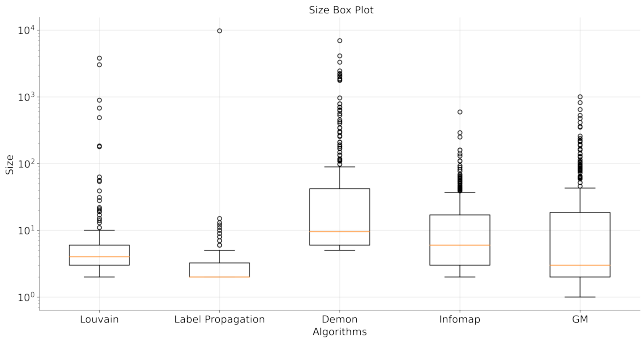
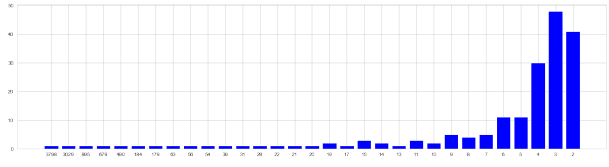
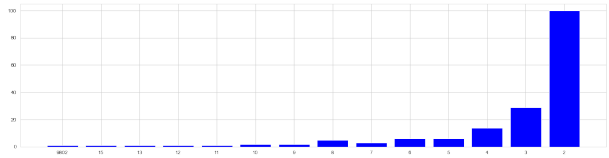


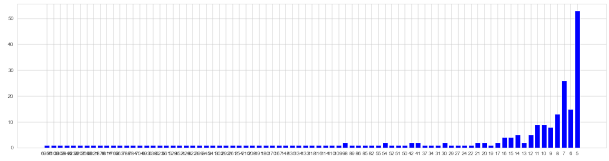
Figure 8: Box plot comunità



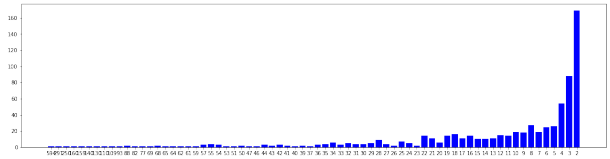
(a) *Louvain*



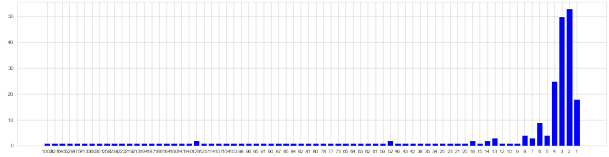
(b) *Lalbel Propagation*



(c) *Demon*



(d) *Infomap*



(e) *GM*

Figure 9: Cardinalità delle comunità

L'algoritmo Louvain, come è possibile osservare dalla Figura [9], è stato quello che, a parità dei valori risultati dalle metriche e dal numero e dalla qualità di comunità prodotte, ha perseguito il risultato più soddisfacente in termini di Modularity e Conductance, riportando meno partizionamenti di grosse dimensioni e di dimensioni inferiori, grazie alla peculiarità di non sovrapposizione del modello, ma anche partizionamenti semanticamente molto significative.

Nella Figura [10] è possibile vedere i WordClouds di 4 comunità ottenute dal modello Louvain, nei quali è possibile notare divisioni semantiche tra le comunità: - La comunità 0 (Figura [10a]) riguarda i più importanti hashtags in lingua inglese legati al Friday for Future, con tags quali fridaysforfuture, fossilfuels, climateaction;
- La comunità 2 (Figura [10b]) riguarda gli hashtags più rilevanti inerenti la comunità tedesca;
- La comunità 4 (Figura [10c]) riguarda la transizione ecologica con termini come solar, energy, wind, green;
- La comunità 9 (Figura [10d]) è legata a temi di attualità come la guerra in Ucraina (standwithurkaine o stopwar) o la povertà di alcuni ceti sociali in Germania (IchBinArmutsbetroffen).

Un'ulteriore analisi sulle caratteristiche principali riguardanti le 4 comunità sopracitate è stata fatta trattando queste come sottografi della rete originale. In Tabella [10] sono riportati i risultati dove è possibile notare come il numero di nodi ed archi tende a diminuire passando da una comunità con un indice basso (ad esempio la comunità 0) ad una comunità con un indice alto (ad esempio la comunità 9), con una conseguente diminuzione dei parametri ad essi legati (LMAX, Avg. degree, Avg. Clustering Coefficient).
Questo dimostra che l'algoritmo Louvain non identifica comunità con al suo interno un'equa distribuzione dei nodi, ma bensì comunità sempre più piccole e dense.

Sulle comunità sono state effettuate le stesse analisi di centralità riportate in Sezione [3].
Per ogni comunità, osservando i nodi più centrali con le varie misure, di cui viene riportato un estratto in Figura [11], non sono state riscontrate sostanziali differenze in termini di hashtags centrali, tranne che nel loro ordine.
Da questi risultati è stato però possibile validare le comunità rispetto

(a) *Community 0* (b) *Community 1*

(c) *Community 4* (d) *Community 9*

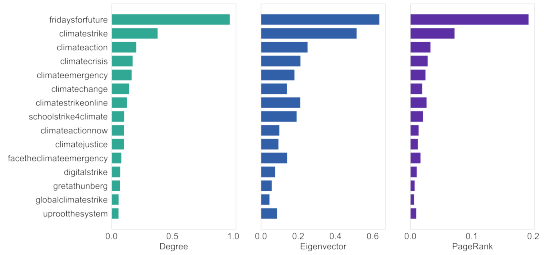
Figure 10: WordClouds comumunity Louvain

Community	0	1	4	9
Number of nodes N	3798	3029	490	54
Number of edges L	18220	12984	2889	383
LMAX	7210503	4585906	119805	1431
Avg Degree <k>	9.5945	8.5731	11.7918	14.1852
Avg Clustering Coefficient	0.7856	0.6994	0.7472	0.8341
Density d(G)	0.0025	0.0028	0.0241	0.2676

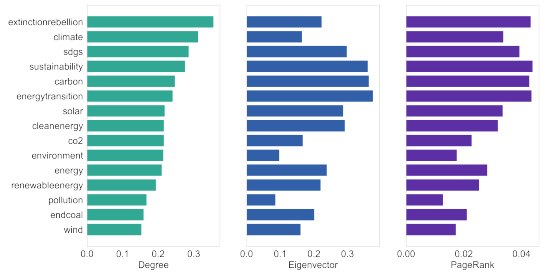
Table 10: Statistiche delle community

alle divisioni semantiche effettuate nell'analisi precedente.
Ad esempio, per Comunità 0 (Figura [11a]) i nodi più centrali sono risultati essere fridaysforfuture, climatestrike, climateemergency, schoolstrike4climate, gretatunberg, mentre per la Comunità 4 (Figura [11b]) sono stati extinctionrebellino, climate, sustainability, solar, energy.

Per effettuare un ulteriore confronto tra le comunità ottenute, sono stati utilizzati i punteggi NMI (Normalized Mutual Information) e NF1 (Normalized F1 score).
Dato che NMI necessita di una copertura totale dei nodi, è stato applicato solo ai metodi che hanno rispettato questo vincolo, dunque escludendo Demon e K-cliques.
I risultati peggiori sono stati ottenuti dal confronto Infomap e Label Propagation (pari a 0.13), Label Propagation e GM (pari a 0.14) e Louvain e Label Propagation (pari a 0.17), indicando che le tre coppie di comunità forniscono risultati quasi completamente scorrelati tra loro.
Il risultato maggiore è stato raggiunto dal confronto fra Infomap e Greedy Modularity (pari a 0,65): infatti, come è possibile vedere nella Tabella [9], i due algoritmi risultano ottenere valori simili di Modularity, Conductance, AID e IED, anche se presentano un numero di community ampiamente diverso, derivante dal diverso metodo di creazione dei cluster per i diversi algoritmi.
Di fatti, per l'Infomap il partizionamento si basa sul flusso indotto dal modello di connessioni in una determinata rete, mentre per il Greedy Modularity il partizionamento consiste nell'unione, in



(a) *Community 0*



(b) *Community 4*

Figure 11: Centrality communities 0 e 4

modo iterativo, di coppie di comunità in presenza di un aumento della Modularità.
La massimizzazione della Modularità, però, porta l'algoritmo del Greedy Modularity, all'inserimento forzato di piccole comunità all'interno di comunità più grandi, questo spiega il perché di un numero così diverso di comunità fra un algoritmo e l'altro.
Infine, dal momento che, come è possibile vedere nella Tabella [9], i valori ottenuti per Louvain, Infomap e Greedy Modularity risultano essere più simili rispetto al confronto con Label Propagation, il confronto a coppie fra Louvain e Infomap e Louvain e Greedy Modularity ha prodotto rispettivamente un punteggio di 0.459, per la prima coppia, e 0.455, per la seconda coppia, indicando che i tre gruppi di comunità forniscono dei risultati simili fra loro (visibile anche in Figura [8]).

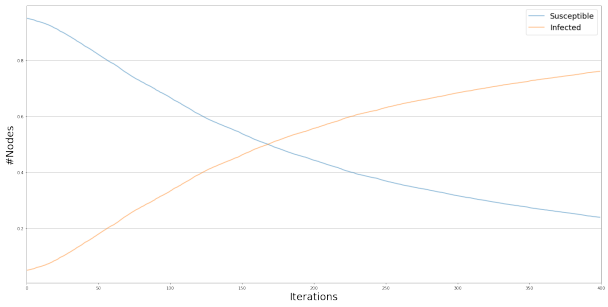
Non sono stati ottenuti risultati migliori nemmeno calcolando i punteggi NF1, per i quali è stato scelto di utilizzare come partizionamento ground truth le comunità scoperte da Louvain, identificato come migliore in precedenza.
Come possibile vedere in Tabella [11], i valori molto bassi di NF1 confermano come gli algoritmi di Community Discovery non trovano corrispondenze in termini di modalità di partizionamento della rete.

Louvain vs Label P.	0.04
Louvain vs Demon	0.05
Louvain vs Infomap	0.23
Louvain vs Greedy M.	0.35

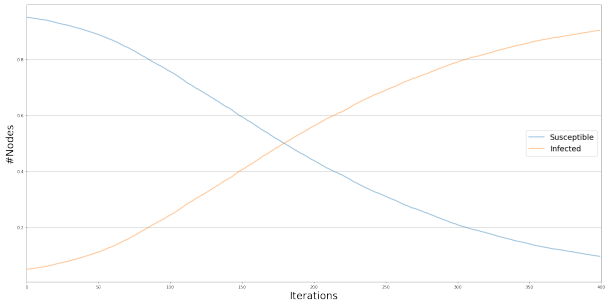
Table 11: NF1

5 SPREADING

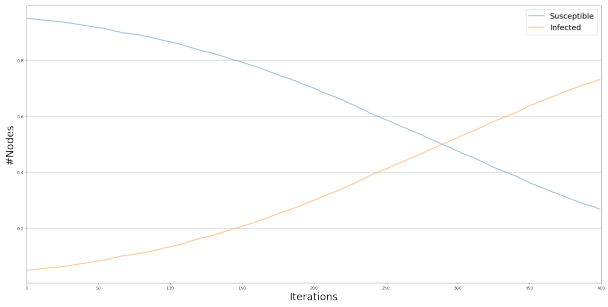
In questa sezione viene analizzato lo **spreading**, grazie al quale è possibile andare a determinare le caratteristiche della rete e riuscire ad effettuare la valutazione del ruolo che alcuni nodi svolgono all'interno di questa; di fatti, l'idea è quella che particolari nodi possano accelerare o rallentare la diffusione di un'idea/infezione.
Nel caso specifico, si è scelto di utilizzare due modelli matematici di diffusione, **SI** e **Profile**, andando a confrontare i risultati ottenuti sulla rete *Twitter - Fridays for Future* - con quelli ottenuti sui modelli sintetici presentati in Sezione [3].



(a) ***SI: Twitter***



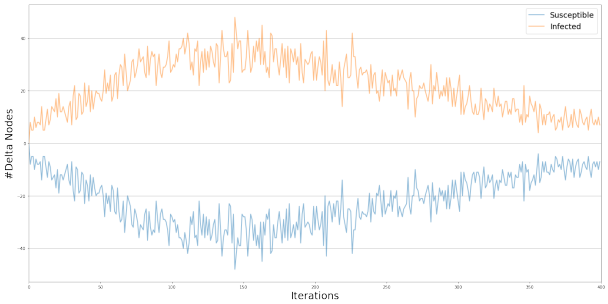
(b) ***SI: BA***



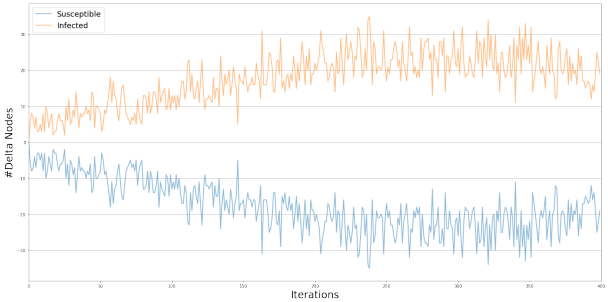
(c) ***SI:ER***

Figure 12: SI: Diffusione

A causa del tipo di rete utilizzata, non è stato possibile considerare lo *spreading* come la classica diffusione epidemica, ma piuttosto come la diffusione di un'informazione quale ad esempio al verificarsi di un evento in un determinato luogo (sciopero a livello locale), inizialmente legato solo ad alcuni hashtags, collegati agli utenti verificati nella community ed ai loro tweet creati, e via via legato ad altri hashtags attraverso il verificarsi dell'evento in altre parti del mondo (estensione dello sciopero a livello globale). Per questo motivo, è stato preferito il modello SI[19] (*susceptible-infected*) rispetto alle sue estensioni SIS e SIR, poiché gli stati di guarigione e rimozione sarebbero stati privi di significato in questa analisi. Il modello prevede per un individuo la possibilità di trovarsi o nello stato **Suscettibile (S)** o nello stato **Infetto (I)**, non prevedendo quindi la possibilità di trovarsi nello stato Rimosso (R).
Il modello è stato eseguito utilizzando una percentuale iniziale di infetti pari al 5% della popolazione con β[20] = **0.001**.



(a) ***SI: Delta BA***



(b) ***SI: Delta ER***

Figure 13: SI: grafici Delta

All'interno della Figura [12] viene riportato come la velocità di diffusione in una rete random, delle stesse dimensioni della rete Twitter, risulti essere minore rispetto ad una rete Scale-free, ottenendo un rapporto tra sani e infetti di circa il 50% all'iterazione 303, mentre per la rete Twitter questo rapporto è stato ottenuto all'iterazione 175.
La dinamica di diffusione per la rete *Barabàsi-Albert* è simile al comportamento atteso per la rete random, ma molto più veloce, come visibile anche in Figura[13] dalla somiglianza tra i delta ottenuti

[19] SI, SIS e SIR sono modelli probabilistici basati sulla diffusione di un agente patogeno su una rete composta da N nodi, in cui ogni nodo può essere infettato con una certa probabilità

[20] Il parametro β indica la probabilità che la malattia si trasmetta da un individuo malato ad uno sano per ogni unità di tempo. L'incremento medio dei malati nel tempo è dato da: $\frac{di}{dt} = \beta si = \beta i(1 - i)$, dove s indica la percentuale di individui sani, ovvero (S/N); mentre i indica la percentuale di individui malati, ovvero (I/N).

dall'esecuzione del modello per le due reti.
Come osservabile in Figura [12] l'informazione si diffonde per la rete BA più velocemente rispetto alla rete ER, ma più lentamente rispetto alla rete Twitter, raggiungendo un rapporto 50-50 tra sani ed infetti all'iterazione 180.
Proprio come ci aspettavamo, è stato possibile osservare che dall'aumento di β derivi un aumento della velocità di diffusione.
Di fatti, come è possibile osservare in Figura [14], l'esecuzione del modello sulla rete Twitter con β = 0.01 porta al raggiungimento di una percentuale di circa il 50% tra sani e infetti già all'iterazione 18 e lo stato di saturazione[21] della rete all'iterazione 340.

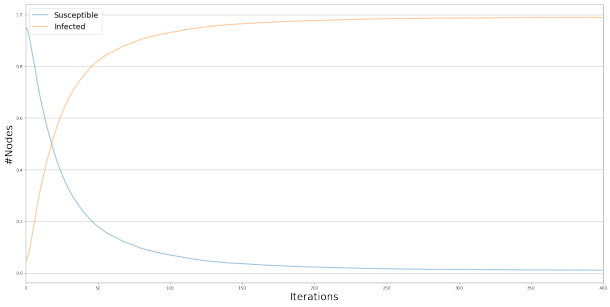


Figure 14: SI: Twitter β = 0.01

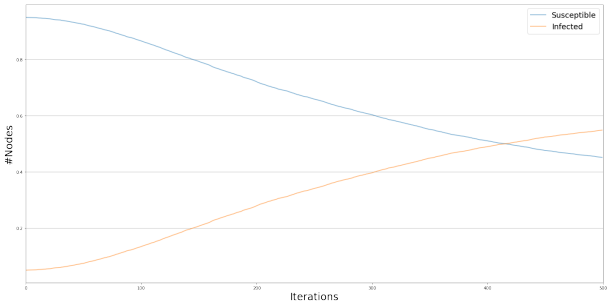


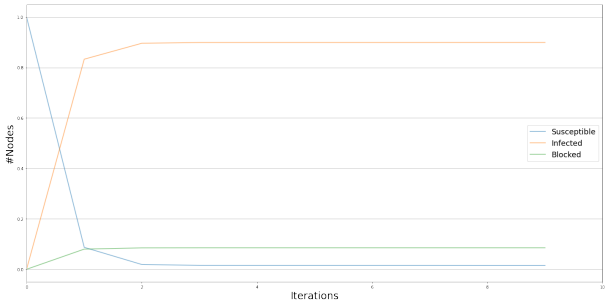
Figure 15: SI: Twitter senza hubs

Altra considerazione da fare riguarda gli hubs presenti all'interno della rete; di fatti, questi influenzano significativamente la velocità di diffusione dell'informazione, proprio perché questi, dal momento che hanno maggiori opportunità di entrare in contatto con un individuo infetto, allora posseggono anche una maggiore probabilità di ammalarsi e di contagiare a loro volta altri individui.
Nella rete presa in esame, gli hubs sono rappresentati dagli hashtags principali quali ad esempio *"fridaysforfuture", "climatestrike", "climateaction"*, che più probabilmente sarebbero i primi ad essere utilizzati dagli utenti nella creazione di nuovi tweet o nella ricondivisione di tweet già esistenti.
Nella Figura [15] viene riportato il risultato della diffusione dell'informazione nella rete Twitter se non vi fossero hubs all'interno, dimostrando come la rimozione di quest'ultimi (a parità di β) determini una velocità di diffusione inferiore rispetto all'implementazione in Figura [12a], portando ad una situazione di parità tra l'insieme dei sani e l'insieme degli infetti all'iterazione 404, quindi 229 iterazioni dopo rispetto la rete con gli hubs.

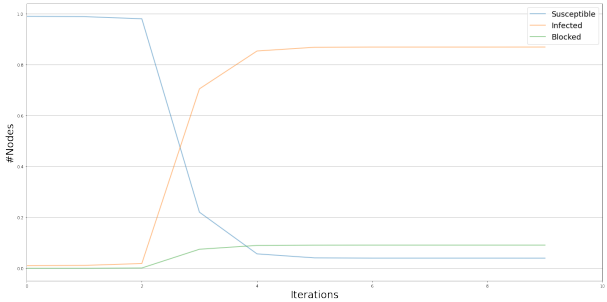
Un'altra tecnica utilizzata per l'analisi della rete è stata il **Profile model**[22], procedura che assume che il processo di diffusione sia solo apparente, dando ad ogni nodo la possibilità di scegliere se adottare o meno un determinato comportamento, sulla base dei propri interessi, quindi scostandosi dall'idea di diffusione epidemica classica (non adatta alla rete).
Poiché il processo di diffusione, in questo contesto, parte da un insieme di nodi infected_nodes che hanno già adottato un determinato comportamento S, sono state effettuate quattro prove, differenti nel modo in cui questo insieme viene scelto:
- selezionando i nodi più centrali emersi in Sezione [3], come riportato in Figura [16a]; - selezionando i nodi marginali, cioè non appartenenti alla giant component emersa in Sezione [3], come riportato in Figura [16b]; - selezionando i nodi random (scegliendo una frazione iniziale di nodi pari a 0.06% del campione), come riportato in Figura [16c]; - selezionando il nodo più centrale da ogni community identificata dall'algoritmo Louvain in Sezione [4], come riportato in Figura [16d]. Per ogni nodo suscettibile in prossimità di un nodo in S viene lanciata una moneta sbilanciata, il cui sbilanciamento in questa implementazione è pari al 10%, per ogni nodo. Il nodo adotterà il comportamento nel caso in cui il lancio darà esito positivo. Se l'adozione viene rifiutata, un nodo entra in stato blocked con una probabilità fissata a 10%. L'adopter_rate è stato settato a 0.001, quindi ad ogni iterazione lo 0.1% di nodi adotta il comportamento spontaneamente per effetti endogeni. É stata scelta questa configurazione di valori poiché, aumentando lo sbilanciamento sarebbe diminuita la velocità con cui i nodi avrebbero adottato il comportamento, mentre aumentando l'adopter_rate la diffusione del comportamento sarebbe risultata più veloce. In Figura [16] vengono riportati i comportamenti di diffusione per le quattro prove effettuate. Come considerato anche per il modello di diffusione precedente, è stato osservato anche per il Profile model che l'aumento dei nodi che adottano il comportamento e la conseguente diminuzione dei nodi suscettibili sia più lento quando si parte da nodi casuali (Figura [16c]) rispetto a quando l'insieme S viene inizializzato con nodi centrali (Figura [16a]). Le altre due implementazioni (Figura [16b - 16d]) hanno rallentato ulteriormente la diffusione del comportamento (rispetto alle due implementazioni precedenti) ottenendo grafici di diffusione pressoché identici. I grafici di diffusione dei modelli sintetici, in questo caso, non sono stati riportati, poiché si sono comportati in modo molto simile alle prove presenti in Figura [16b - 16d].

[21] Secondo il modello SI l'epidemia terminerà quando tutti gli individui saranno stati infettati e quindi la rete si troverà nello stato di saturazione, in particolare si ha che: $\lim_{t\to\infty} i(t) = 1, \lim_{s\to\infty} s(t) = 0$

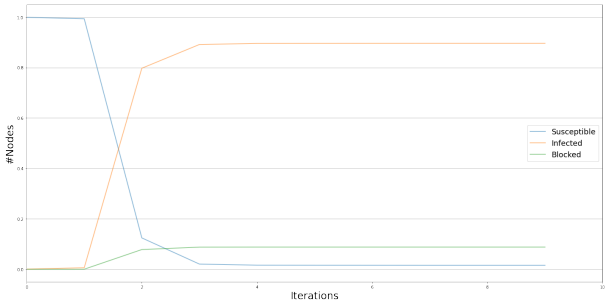
[22] https://ndlib.readthedocs.io/en/latest/reference/models/epidemics/Profile.html



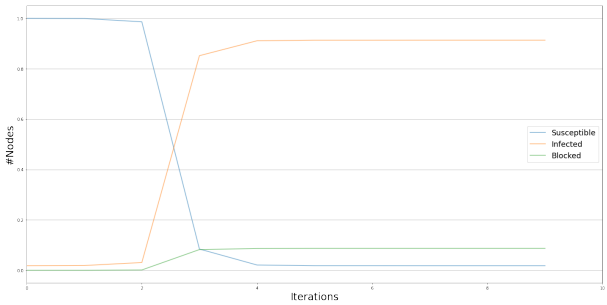
(a) *Profile: nodi centrali*



(b) *Profile: nodi marginali*



(c) *Profile: nodi random*



(d) *Profile: nodi community Louvain*

Figure 16: Profile: Diffusione

6 LINK PREDICTION

In questa sezione si cerca di prevedere i link tra i nodi della rete, per necessità computazionali, lo svolgimento di questa task è stata effettuata attraverso l'utilizzo di due sottografi della rete.
I due sottografi G1 e G2 sono stati ottenuti selezionando rispettivamente nodi con grado maggiore di 50 e 30.
In G1 si hanno 257 nodi e 5901 archi, mentre in G2, 511 nodi e 10458 archi.
É stata effettuata questa scelta di sampling poichè mantenendo solo i nodi più importanti, aventi maggiore probabilità di essere connessi, si riduce la quantità di falsi positivi riscontrabile.
I metodi di previsione non supervisionata eseguiti, sono stati valutati attraverso la divisione dei sottografi in training set e test set con uno split 80% - 20%.
In Figura [17 - 18] sono riportate le curve ROC dei metodi utilizzati, divisibili in tre tipologie:
- **Neighborhood-based**: le cui misure assegnano un punteggio tra due nodi x ed y, determinato dai vicini che hanno in comune i due nodi. Di questa categoria sono stati testati CommonNeighbours, AdamicAdar e Jaccard.
- **Path-based**: applicando di questa categoria Katz, misura che considera il numero di percorsi tra i due nodi.
- **Ranking**: considerando la misura SimRank, la quale assegna lo score basandosi sulla similarità dei vicini.

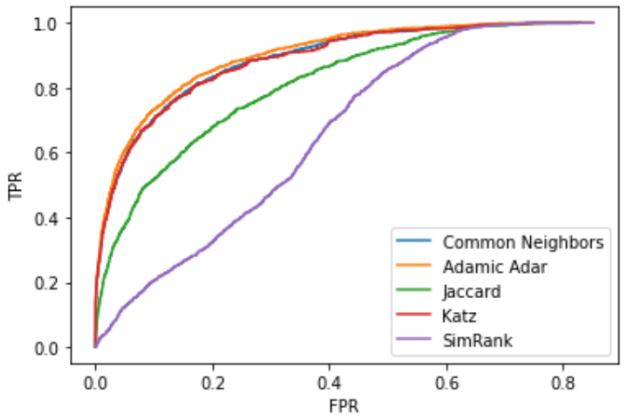


Figure 17: Curve ROC su G1

L'AUC della curva roc ottenuta dai metodi per i due grafi è riportata in Tabella [12], nella quale è possibile notare come i predittori basati sul **neighborhood** a parte Jaccard abbiano ottenuto valori migliori passando dal campione G1 al campione G2.

	AUC_G1	AUC_G2
Common Neighbours	0.73	0.79
Adamic Adar	0.74	0.80
Jaccard	0.66	0.65
SimRank	0.55	0.59
Katz	0.75	0.86

Table 12: Statistiche del Path

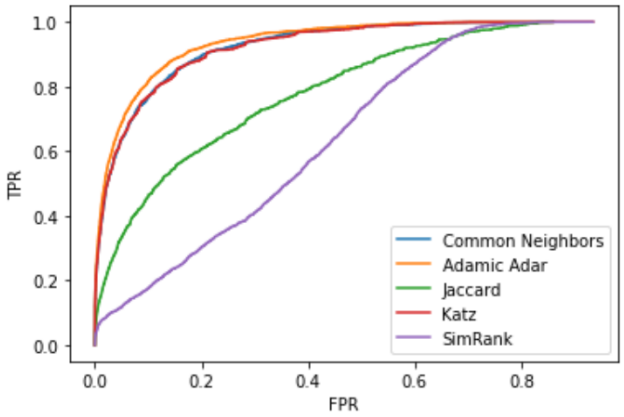


Figure 18: Curve ROC su G1

Gli altri due metodi, basati su paths e su ranking hanno ottenuto per G2 valori superiori alla prova precedente, in particolare Katz.
Complessivamente sono stati ottenuti ottimi risultati, probabilmente a causa della caratteristiche della rete di essere a regime connesso, come discusso in Sezione [3.1].
Ulteriori esplorazioni, quali ad esempio l'intersezione dei primi 100 link previsti da ogni metodo, non hanno portato ad informazioni aggiuntive.
Gli stessi metodi di Link Prediction sono stati testati sui sottografi ottenuti dalle prime 8 comunità riportate dal metodo Louvain (discusse in Sezione [4]) considerando per ogni comunità un campione.
Sono stati ottenuti risultati di AUC nettamente inferiori alle prove precedenti, come visibile in Tabella [13], nelle quali i nodi non venivano campionati in modo random, ma considerando il loro grado.
Tuttavia, anche in questo caso il metodo Katz ha ottenuto valori superiori agli altri metodi, dimostrandosi nuovamente il migliore per la tipologia di dati considerati.

Community	0	1	2	3	4	5	6
Common Nbr	0.004	0.003	0.12	0.20	0.19	0.11	0.03
Adamic Adar	0.005	0.004	0.14	0.22	0.21	0.12	0.03
Jaccard	0.001	0.003	0.08	0.13	0.14	0.09	0.03
Katz	0.010	0.005	0.49	0.49	0.74	0.30	0.10
SimRank	0.008	0.006	0.47	0.37	0.65	0.28	0.10

Table 13: AUC of Communities

7 OPEN QUESTION: EMBEDDING WORD2VEC

In questa sezione si propone una versione modificata della task svolta in Sezione [6], nella quale vengono valutati modelli di previsione dei link tra i nodi della rete, considerando anche il significato di tali nodi, attraverso approcci di text mining.
L'obiettivo dello studio era quello di valutare se l'incorporamento nella rete tag-to-tag della semantica di ogni tag avrebbe portato ad aumentare o meno i valori di previsione riportati nello studio precedente.
A tal fine ci siamo ispirati a social network nei quali l'utilizzo degli hashtags è regolato dalla presenza di un database, contenente i significati e le definizioni di ciascun tag, sviluppando una proposta analoga per Twitter.
La fase di raccolta dei significati dei tags è stata svolta andando ad estrarre direttamente da Wikipedia le prime righe di testo corrispondenti al tag cercato, per fare ciò abbiamo usato il pacchetto Wikipedia-API il quale è wrapper di Python per API di Wikipedia.
Sono quindi stati tenuti in conto solo i tag riguardanti pagine esistenti di wikipedia, abbiamo cosi ottenuto un documento contenente circa 300 tag in inglese e tedesco, e i rispettivi paragrafi contenenti la possibile spiegazione, verificando manualmente l'attendibilità di tale corrispondenza.
Al fine di migliorare l'efficienza dei modelli considerati, le definizioni dei tag acquisite sono state preelaborate.
In particolare, sono stati sostituiti gli urls e i numeri, rimossa la punteggiatura e gli insiemi di caratteri non codificati in fase di collezione.
Successivamente, il testo è stato convertito in minuscolo, tokenizzato, rimosse le stopwords e trasformato attraverso il metodo word2vec per l'incorporamento con i rispettivi tags.
È stata preferita la metodologia word2vec ad altre a causa dei problemi computazionali legati all'applicabilità sul dataset.
Infine, è stato creato un dataset formato dai link, trasformati concatenando i valori corrispondenti a coppie di nodi, ed una nuova variabile target binaria, la quale assumeva valore 1 per i collegamenti esistenti e valore 0 per i collegamenti non esistenti.
Sono stati considerati i modelli DecisionTree, Regressione Logistica, LightGBM24 (algoritmo di boosting basato su DecisionTree) e Linear SVM, effettuando (dove possibile) una scelta degli iperparametri ottimi attraverso una GridSearchCV, nella quale il numero di folds è stato ottenuto tramite una StratifiedKFold.
In Tabella [14] vengono riportati i gruppi di iperparametri testati, dai quali sono stati scelti i migliori attraverso l'ottimizzazione dell'AUC (metrica preferita per poter effettuare confronti più accurati con quanto riportato in Sezione [6]).

DecisionTree
Criterion: gini, entropy
Max depth: [2, 5, 10, 15, None]
Min Samples Split: [2, 5, 10, 20]
Min Samples Leaf : [1, 5, 10, 20]
Linear SVM
C: [0.001, 0.05, 0.01, 0.1, 1.0, 10.0, 50, 100.0]
tol: [1.0, 1e-1, 1e-2, 1e-3, 1e-4, 1e-5, 1e-6]

Table 14: Iperparametri testati

Infine, sono state delineate e testate tre architetture Neural Network.
La prima, chiamata Model1 costituita da 8 livelli caratterizzati da 32 neuroni ciascuno, aventi funzione di attivazione di tipo *tanh*.
La seconda, Model2 contiene un livello formato da 128 neuroni e 4 livelli da 32 neuroni, aventi come funzione di attivazione relu.

La terza (Model3) composta dai parametri elencati in Tabella [15] alla quale è stata infine applicata un'ottimizzazione di tipo Stochastic gradient descent.

Units	10
Hidden layers	1
Momentum	0.9
Learning rate	0.261
Regularizer	0.0001

Table 15: Parametri Model3

I modelli sono stati testati sui dati generati dal sottografo G1 ottenuti in precedenza (Sezione [6]) e su una prova bilanciata (G1_Balance) attraverso Random Undersampling della classe maggioritaria (target = 0), poichè il rapporto tra link esistenti e nessun link era vicino al 29% (riflettendo la condizione reale del grafo in cui il numero di collegamenti esistenti è inferiore al numero di collegamenti che avrebbero potuto esistere).
É stato escluso dalle prove il sottografo G2 (Sezione [6]) a causa di problemi computazionali.
In Tabella [16] vengono riassunti i risultati della metrica AUC ottenuti per i modelli sui dati.

	G1	G1_Balance
DecisionTree	0.826	0.802
Regressione	0.802	0.807
LightGBM	0.910	0.890
Linear SVM	0.879	0.862
NN Model1	0.634	0.703
NN Model2	0.536	0.500
NN Model3	0.828	0.833

Table 16: Risultati AUC

Come osservabile dalla Tabella [16] il bilanciamento non ha portato ad ottenere risultati simili alla prova non bilanciata, in certi casi addirittura peggiori.
In generale i risultati migliori sono stati ottenuti con LightGBM, I quali risultano anche superiori a quelli riportati in Tabella [12] frutto dell'applicazione standard del task di Link Prediction.

8 CONCLUSIONI

In conclusione, l'intento di analizzare gli Hashtags usati su Twitter per descrivere il fenomeno del Fridays for Future è stato conseguito con successo. Grazie alle analisi riportate nelle Sezioni [4] e [6], è stato possibile effettuare un'esame più approfondito attraverso l'utilizzo di tecniche di Data Mining e di Analisi del testo. Tale lavoro, potrebbe essere utilizzato come base per uno studio più approfondito, al fine di poter integrare all'interno della piattaforma Twitter un sistema di suggerimento e auto-completamento degli Hashtags, così da consigliare agli utenti quegli hashtags che potrebbero consentire maggiori interazioni fra loro, portando un vantaggio sia a quest'ultimi che alla piattaforma stessa. Infatti, come visto all'interno della sezione [7], è stata creata una ipotetica raccolta di hashtags più importanti, ai quali è stata aggiunta una breve descrizione. L'implementazione di un dizionario, contenente solo gli hashtags più importanti, potrebbe essere d'aiuto sia a chi ancora non conosce un fenomeno in tendenza e vorrebbe comprenderlo, sia potrebbe essere d'aiuto agli *account Twitter verificati*, allo scopo di poter lanciare il proprio hashtag da collegare direttamente al proprio account. Tale lavoro, infine, potrebbe essere utilizzato, anche con il fine di evitare il problema della creazione di hashtags contenenti refusi linguistici.

REFERENCES

(1) Barabási Albert-László, Network Science, Cambridge University Press (2016). http://networksciencebook.com/

(2) Lancichinetti, Andrea, Santo Fortunato, and Filippo Radicchi. "Benchmark Graphs for Testing Community Detection Algorithms." Physical Review E 78.4 (2008). https://doi.org/10.1103/PhysRevE.78.046110