# Analisi del traffico su Twitter per le elezioni europee 2019

**Daniele Gambetta**
gambetta@studenti.unipi.it
daniele.gambetta7@gmail.com
Student ID: 465137

## ABSTRACT

Questo breve lavoro di ricerca prende in esame una rete di contenuti e interazioni su Twitter durante il pomeriggio di sabato 25 maggio inerenti alle elezioni europee. Oltre ad un grafo rappresentante vari tipi di nodi e di relazioni riguardo l'attività sul social network, sono presenti informazioni riguardo i vari nodi a seconda della tipologia. Quindi si è prima fatta un'analisi della topologia di rete e poi si è confrontato questi risultati con le informazioni aggiuntive, focalizzandosi su analisi di Community Detection, Link Prediction e analisi della polarizzazione dei contenuti.
[1]

## KEYWORDS

Social Network Analysis, Twitter, Polarizzazione

**ACM Reference Format:**
Daniele Gambetta. 2019. Analisi del traffico su Twitter per le elezioni europee 2019. In *Social Network Analysis '19*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUZIONE

In questa breve ricerca è stata condotta un'analisi sul traffico di tweet pubblicati nel pomeriggio di sabato 25 maggio 2019, penultimo giorno di elezioni europee. Il grafo analizzato rappresenta il traffico complessivo su Twitter nell'arco di due ore, nel quale i nodi possono essere di cinque tipi diversi (Utenti, Tweet, Hashtag, Media, Link), con significato dei vari link possibili dipendente da origine e arrivo. Inoltre ad ogni nodo sono associate varie proprietà a seconda del tipo.

## 2 DATA COLLECTION

L'acquisizione iniziale dei dati è stata effettuata tramite Gephi con il plugin Twitter Streaming Importer, che usando le API del social network permette di importare i tweet che soddisfano certi parametri e rappresentarli in un grafo, con varie possibili modalità.

### Origine dei dati

Nel nostro caso, sono state selezionate alcune parole identificative dei tweet inerenti alle elezioni europee impostando l'opzione "Full Network". Quindi ad essere interessati sono tutti i tweet contenenti determinati termini ("elezioni", "europee", "europawahl", "ue", "europe", "elections"), e il grafo restituito è composto a nodi che possono rappresentare:

- Tweet nei quali queste parole sono state usate
- Utenti che ne hanno fatto uso
- Hashtag contenuti nei tweet selezionati
- Media contenuti nei tweet selezionati
- Link contenuti nei tweet selezionati

Per quanto riguarda i link, tutti diretti, si caratterizzano quindi in:

- Link "Utente -> Utente": possono corrispondere a una risposta o ad un retweet fatti da un utente verso un altro (o verso sè stesso)
- Link "Utente -> Tweet": indica che un utente ha pubblicato un certo tweet
- Link "Tweet -> Hashtag": indica che il tweet conteneva quell'hashtag
- Link "Tweet -> Link"
- Link "Tweet -> Media"

In definitiva, la configurazione è modellizzabile con il seguente grafico:

Infine, il plugin consente di ottenere per ogni nodo alcune caratteristiche a seconda della tipologia. Tra quelle fornite dal plugin, sono state mantenuti il numero di followers e di friends (utenti che segue l'account) per i nodi utenti e la lingua per i nodi tweet.

Lo streaming è stato effettuato dalle 16:00 alle 18:00 (ora italiana) di sabato 25 maggio, collezionando così un grafo

[1] **Project Repositories**
Data Collection: https://github.com/sna-unipi/data-collection
Analytical Tasks: https://github.com/sna-unipi/analytical-tasks
Report: https://github.com/sna-unipi/project-report


*SNA '19, 2018/19, University of Pisa, Italy*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$0.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

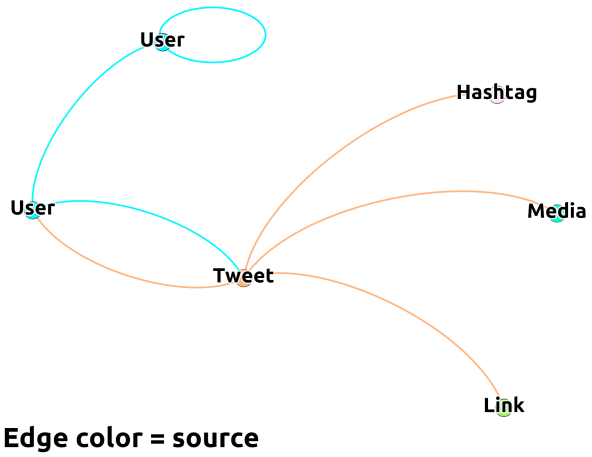



**Figure 1: Distribuzione dei nodi per tipo**

complessivo di 18325 nodi e 34714 link. A scopo puramente qualitativo è stata creata quindi con Gephi una rappresentazione del grafo.

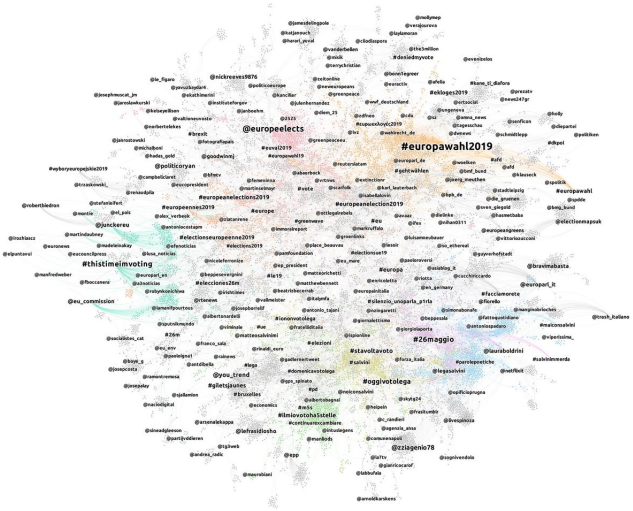



**Figure 2: Visualizzazione qualitativa del grafo ottenuta con Gephi**

Successivamente, per consentire un trattamento tramite librerie di Python, il grafo è stato esportato in formato gdf, dal quale è stato facile estrarre due file csv, uno per i nodi, ognuno con le proprie caratteristiche, e uno per i link. Infine questi csv sono stati importati in Notebook Jupyter per essere analizzati tramite librerie Python.

## 3 NETWORK CHARACTERIZATION

La rete quindi ottenuta è un grafo multidiretto con presenza di loop. Questo perchè un utente può rispondere ad un altro utente più volte (e in tal caso compariranno un link per ogni interazione), e inoltre può retwittare o rispondere a sè stesso, generando un link con origine e arrivo allo stesso nodo. Nonostante questo, per specifiche task è stata considerata la rete come non diretta (Link Prediction e Community Detection). Analizzando il grafo con la libreria NetworkX otteniamo le informazioni quantitative basilari

Nodi: 18325
Link: 34714
Selfloop: 119
Connessa: False
Bipartita: False
Diretta: True
Nodi isolati: 0
Componenti fortemente connesse: 18251
Componenti debolmente connesse: 313

I nodi sono inoltre suddivisi nei vari tipi con le percentuali rappresentate in figura:

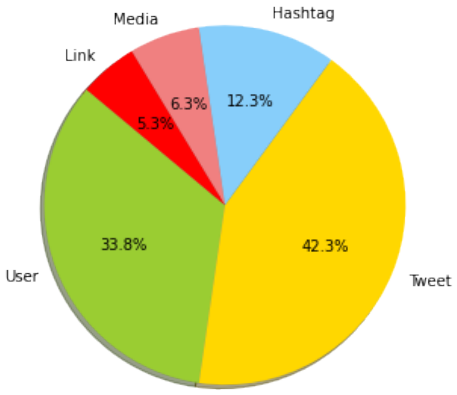



**Figure 3: Distribuzione dei nodi per tipo**

Per quanto riguarda la connessione, esistono 313 componenti debolmente connesse. Analizzandole, risulta che le cinque componenti più grandi hanno le seguenti dimensioni: [16644, 40, 30, 20, 19]. Quindi la componente più grande contiene il 91% circa dei nodi complessivi. Considerando che è un grafo diretto, la densità dei link è pari a $1.03 * 10^{-4}$, mentre quella della componente più grossa è $1.19 * 10^{-4}$. Inoltre,

provando a calcolare la densità della rete eliminando i link multipli si ottiene una densità di $1.02 * 10^{-4}$, rivelando quindi che il numero di link multipli non è significativo.

### Analisi del grado

Essendo un grafo diretto, è stata condotta un'analisi considerando il grado di ingresso, di uscita e quello complessivo, paragonando nel caso degli utenti queti valori con numero di followers e friends. Dal grafico pare che non ci sia correlazione particolarmente forte tra numero dei followers e viralità del contenuto, in effetti nello specifico risulta che gli account più seguiti non sono i più virali.

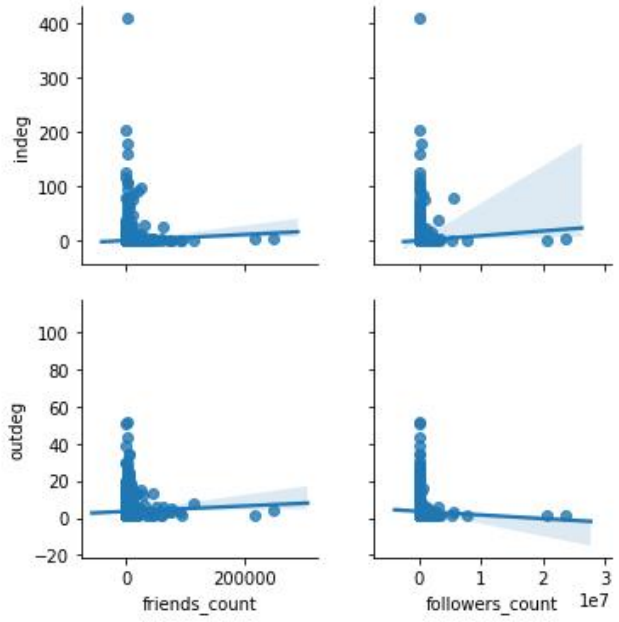



**Figure 4: Confronto tra grado e numero di friends e di followers per i nodi utente**

Per studiare il grado totale, prendiamo il grafico della distribuzione con lo stesso in scala log-log.

Come si può facilmente vedere, la distribuzione in scala log-log è molto prossima a quella di una retta, avvalorando l'idea intuitiva che la rete si comporti come un modello ad attaccamento preferenziale.

Per effettuare altre analisi è stato considerato il sottografo costituito solo dai noti Utente, e dai link tra questi. Il sottografo risulta così composto.

Nodi: 5442
Link: 6551
Selfloop: 119
Connessa: False
Bipartita: False

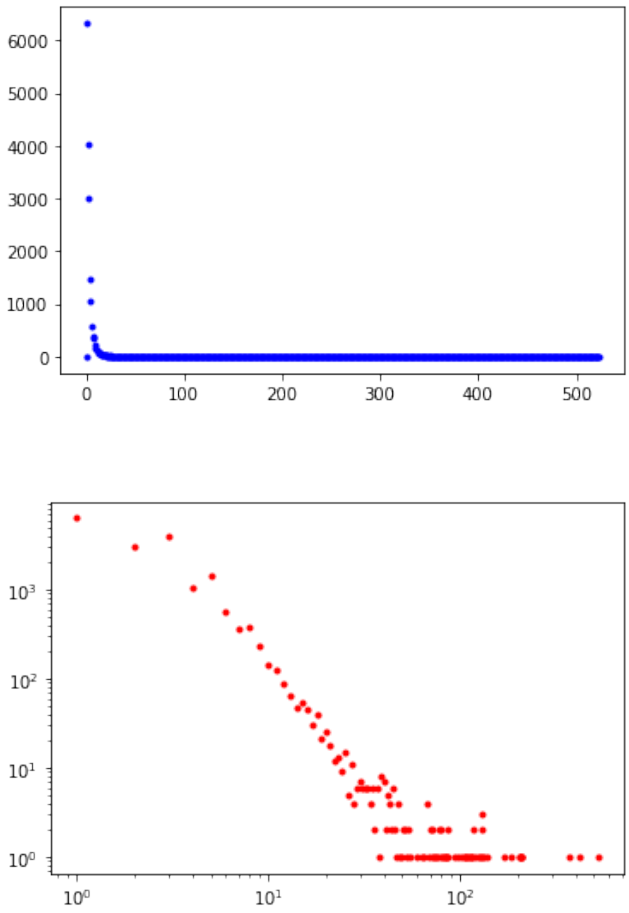


**Figure 5: Distribuzione del grado in scala lineare e loglog**


**Figure 6: Il tweet che rappresenta i nodi con grado più alto**

Diretta: True
Nodi isolati: 0
Componenti fortemente connese: 5415
Componenti debolmente connese: 593

Come ultime statistiche si è osservata la presenza di triangoli di utenti nel sottografo utente, con i seguenti utenti che mostrano la maggior presenza di triangoli:
@mollymep: 156
@the3million: 131
@verajourova: 126
@fahraht: 80
@markruffalo: 75

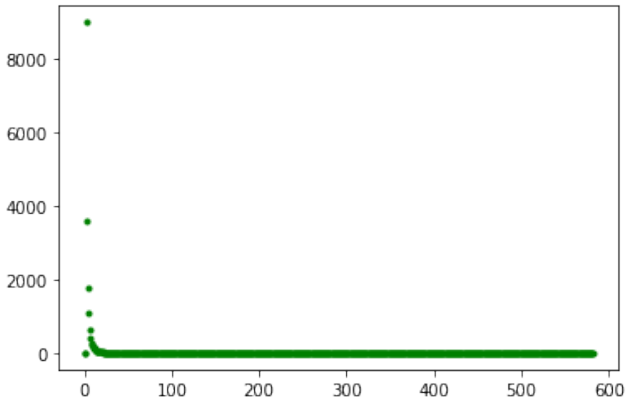
Inoltre per entrambe le reti (complessiva e utente) si è calcolato il clustering coefficient mentre per il grafo utente si è anche calcolato il diametro della componente connessa più grande.

average clust rete 0.1350886765534672
average clust user 0.2124084737220594

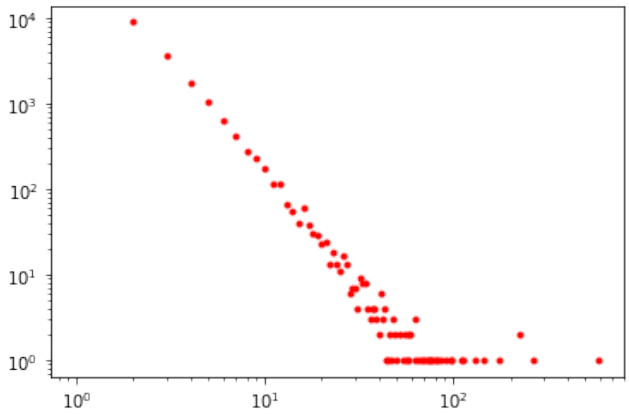
diametro rete utenti 18

### Confronto con BA e ER

Si vuole confrontare la topologia della rete generale e quella della rete utenti con reti di pari dimensioni e pari numero di link con topologie Barabasi-Albert e Erdos-Renyi, Quindi tramite libreria NetworkX si sono cercati parametri tali da ottenere le reti cercate. Nel caso del confronto con la rete generale, oltre ai 18000 nodi, si sono impostati come parametri m=2 per la rete BA e probabilità pari a 0.00022 per ER, ottenendo rispettivamente 35996 e 35802 link, numero molto prossimo a quello della nostra rete. A seguito si è analizzata la distribuzione del grado in queste reti, riscontrando come presumibile i caratteristici andamenti noti dalla teoria. Nel caso della rete BA si mostra anche il grafico in scala loglog.




Inoltre, con nuovi parametri pari n=544, m=1 e p=0.0004 si sono attenuti grafi con quantità di link paragonabile a quella del sottografo utente. Confrontando le varie statistiche, come già era stato accennato, risulta chiaro che la struttura



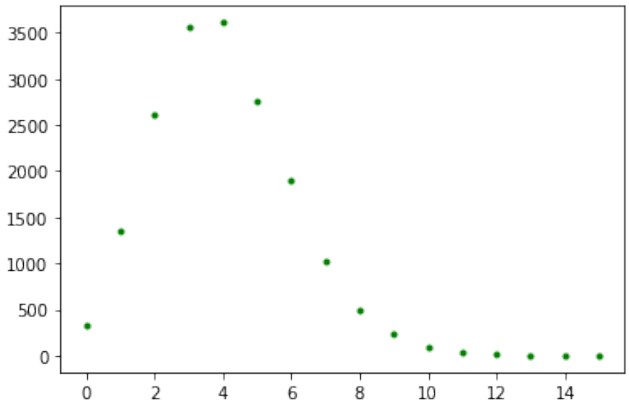




**Figure 7: Dall'alto verso il basso, i grafici delle distribuzioni del grado nel caso BA, BA in scala loglog e ER**


**Figure 8: Il tweet che rappresenta i nodi con grado più alto**

è più simile ad una rete ad attacamento preferenziale, cosa evidente dal confronto della distribuzione del grado. Nel caso della rete "piccola", il diametro di una rete BA generata è esattamente 18, pari al diametro della rete utente. Ad essere invece particolarmente differente, tra le reti "reali" di Twitter

ed entrambi i casi generati, è il cluestering coefficent medio, che si dimostra particolarmente più alto.

## 4 TASK 1: COMMUNITY DETENCTION

Come primo task si è analizzata la suddivisione in community della rete complessiva, considerandola come grafo non diretto e utilizzando la libreria cdlib. Sono stati applicati gli algoritmi Demon, Eigenvector e Louvain, che hanno fornito rispettivamente, 492, 314, 381 community.

Per cercare un confronto sulla validità della suddivisione si è utilizzata la classificazione dei tweet rispetto alla lingua. Come accennato all'inizio, il plugin Twitter Streaming Importer di Gephi ottiene per alcuni tweet anche la lingua in cui è stato scritto. Le cinque lingue più diffuse compaiono in questo ordine:
inglese 3905
italiano 2429
francese 282
spagnolo 267
tedesco 179

Con la lingua inglese che compare un numero sostanzialmente maggiore di volte rispetto alle altre. Per ogni algoritmo di community detection utilizzato si è quindi osservato quanto le comunità trovate siano adatte a distinguere le lingue. Il risultato più interessante riguarda il confronto basato su Louvain, che viene qui riportato.

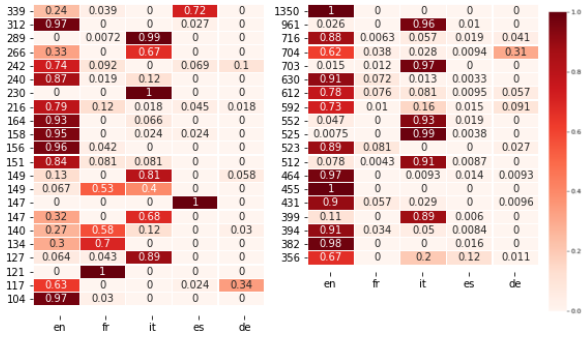



**Figure 9: Suddivisione lingua**

La tabella mostra, per le community più grandi, la composizione interna rispetto la lingua del tweet. Come si può osservare ovviamente in molte community la percentuale di lingua inglese è maggiore, ma alcune community di dimensione considerevole riportano una alta percentuale di lingue presenti in quantità complessivamente minore, indicando che il riconoscimento di community ha portato anche ad una buona seperazione rispetto la lingua dei tweet.

## 5 TASK 2: LINK PREDICTION

Per quanto riguarda la link prediction, utilizzando la libreria linkpred di Python, si è scelto di eliminare dalla rete un certo numero di link pari al 20% del totale, pari quindi a 4747 link. Quindi si sono applicati vari algoritmi, confrontando poi la previsione con i link appositamente rimossi. L'accuracy è stata calcolata inoltre anche valutando le 200 previsioni più probabili.
Common neighborous
Numero corretti: 96
Accuracy: 2.02
Numero corretti(200): 47
Accuracy(200): 23.5

Adamic Adar
Numero corretti: 409
Accuracy: 8.62
Numero corretti(200): 63
Accuracy(200): 31.5

Jaccard
Numero corretti: 70
Accuracy: 1.47 Numero corretti(200): 0
Accuracy(200): 0.0

Katz
Numero corretti: 143
Accuracy: 3.01
Numero corretti: 46
Accuracy(200): 23.0

## 6 TASK 3: POLARIZZAZIONE: GRAFO PROIEZIONE DEGLI HASHTAG

Come problema aperto si considera la ricerca della polarizzazione dei contenuti, cioè escogitare metodi basati sulla topologia della rete capaci di rilevare i diversi posizionamenti politici all'interno di una rete di interazioni tra utenti. Si pone inoltre in considerazione il fatto che in una rete come quella analizzata le connessioni tra i nodi rappresentano interazioni quindi risposte e retweet, i quali possono identificare non solo endorsement ma anche critiche o risposte in contrapposizione. Quindi, se l'obiettivo è studiare la polarizzazione dei contenuti, un metodo di community detection potrebbe essere inadeguato sul grafo totale. Quello che si tenta di fare quindi è basarsi su un altro grafo ottenuto dla principale, quello degli hashtag, tramite una proiezione dal grafo bipartito costituito dai nodi Tweet e Hashtag. Lo scopo, quindi, è costruire un grafo di nodi Hashtag dove due Hashtag sono connessi tra loro se compaiono in uno stesso tweet, dato che questo criterio solitamente determina una loro vicinanza

politica, e queesto grafo si procede con un algoritmo di community detection, testando su un campione scelto di termini la loro appartenenza alle comunità ottenute.

### Grafo bipartito e proiezione

Nella teoria dei grafi e delle reti, un grafo bipartito è un grafo tale che l'insieme dei suoi nodi si può partizionare in due sottoinsiemi tali che ogni nodo di una di queste due parti è connesso solo a nodi dell'altra.

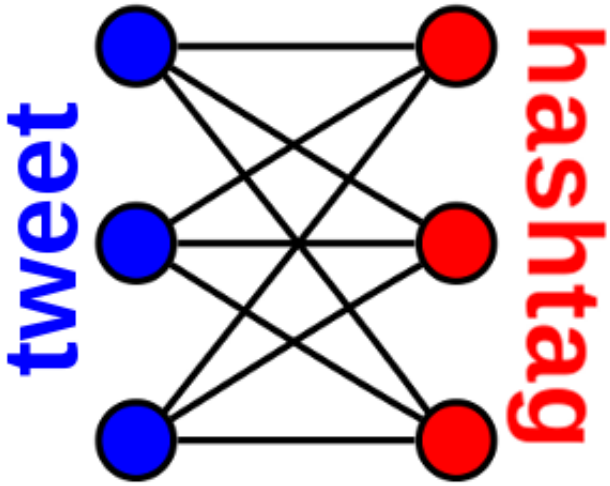



**Figure 10: Rappresentazione schematica del grafo bipartito con Tweet e Hashtag**

Da un grafo bipartito, costituito da due componenti A e B, è possibile ricavare due ulteriori grafi, chiamati proiezioni su A e su B, connettendo due nodi di un sottoinsieme se questi sono connessi ad uno stesso elemento nel grafo iniziale. Nel nostro caso, consideriamo il sottografo costituito dai nodi di tipo Tweet e Hashtag. Per come è stato ottenuto il grafo, non esistono link diretti tra nodi Tweet ad altri nodi Tweet, nè tra nodi Hashtag e altri nodi Hashtag. Di conseguenza il sottografo Tweet/Hashtag è bipartito. Allora, il grafo proiezione Hashtag, sarà costituito da Hashtag connessi se sono connessi nel grafo originale a uno stesso Tweet, quindi se compaiono in uno stesso tweet. A questo punto, è stato eseguito l'algoritmo Louvain sul grafo degli Hashtag, e sono state scelti degli hashtag noti per essere rappresentativi di due diverse e opposte opinioni politiche, quelle legate al Partito Democratico e quelle della Lega.
PD = ["facciamorete", "pd", "lasinistra", "zingaretti","nzingaretti"]
lega = [ "iostoconsalvini", "salvininonmollare", "matteosalvini", "primagliitaliani", "salvinipremier", "legagiovani", "legaumbria","stavoltavotolega"]

A questo punto, si è osservato in quale community comparissero i vari termini, riscontrando che effettivamente Louvain considerava tutte le parole legate al PD nel cluster numerato come 0 e tutte le parole legate ala Lega nel cluster 7, rilevando così la diversa polarizzazione. In una fase eventualmente successiva a questa operazione quindi, nel caso si volesse considerare la polarizzazione degli utenti o dei tweet, si potrebbe valutare questa in funzione della polarizzazione degli hashtag contenuti nel tweet o usati da un utente, eventualmente assegnando un valore numerico da 0 ad 1 per valutare eventuali polarizzazioni non univoche.

## 7 CONCLUSIONI

Le analisi compiute in questo breve lavoro rappresentano solo alcune delle possibili linee di ricerca nell'ambito dell'interazione online. In particolare, una rete caratterizzata da differenti tipi di nodi, ognuna con le proprie caratteristiche, se da un lato rischia di creare confusione, offre dall'altro la possibilità di rigenerare nuove reti secondo definizioni utili, confrontando poi le statistiche della topologia di rete con altre proprietà o informazioni inerenti al contesto. Seguendo questa linea di ricerca è forse allora possibile escogitare strategie mirate di analisi sulla polarizzazione, sulla previsione di link e sulla community detection rinforzate dall'uso di informazioni ulteriori rispetto alla semplice struttura della rete.