Progetto di Social Network Analysis

Andrea Sghedoni
Laurea Magistrale in Informatica, Università degli Studi di Bologna
A.A. 2015/2016

Abstract **:** L'obiettivo del progetto è quello di investigare, capire e comprendere quali sono le caratteristiche che rendono determinati nodi più influenti, rispetto ad altri, all'interno di un Social Network.
Per nodi "più influenti", si intendono quegli utenti che risultano assumere una posizione privilegiata nella rete, riuscendo, di conseguenza, ad avere sempre un certo successo nel propagare informazione e nel ricevere attenzione nei contenuti che intende diramare.

1. INTRODUZIONE

L'enorme sviluppo tecnologico a cui il mondo si sta sottoponendo porta inevitabilmente alla nascita e creazione di nuovi rapporti sociali. L'esplosione di Social Network Online, come ad esempio Twitter e Facebook, rappresenta una grossa opportunità per la collezione di nuove informazione, per quanto riguarda lo studio delle relazioni, rapporti tra persone, sotto forma di account utente e del comportamento umano.
Sempre più persone considerano queste piattaforme come uno strumento fondamentale per mantenere relazioni, intrattenere rapporti, diffondere ed apprendere notizie.
Nella società vi sono soggetti che influenzano il comportamento di altri soggetti tramite il loro comportamento, proprio come succede all'interno di un Social Network Online, dove determinati utenti possono comportarsi in modo tale da influenzare altri utenti, entrando in contatto/relazione.
Il successo, in termini di propagazione dell'informazione, è una concomitanza di fattori come ad esempio la popolarità(decisamente il più importante), frequenza di propagazione, originalità e tipo dei contenuti, momento storico/sociale.
La perdita ed il guadagno di influenza sulla rete di un utente, in realtà, è ancora un problema aperto, in quanto risulta piuttosto difficile formulare principi per cui una determinata azione può portare ad un guadagno/perdita di influenza nella rete che si sta considerando.
Lo scopo del progetto è, come già accennato nell'abstract, di investigare e studiare quei nodi che dovrebbero risultare i più influenti sulla rete.
Capire sostanzialmente i parametri fondamentali che caratterizzano questi utenti e che li rendono i più interessanti, in termini di popolarità ed influenza.

2. TWITTER

Il Social Network che sarà preso in considerazione nello studio è Twitter.
Questo viene definito un microblog, dove l'utente può postare un messaggio di massimo 140 caratteri sulla rete. Le relazioni tra gli utenti si basano sul concetto di Followers/Following, dove un utente può seguire un altro utente, venendo notificato così di tutti i contenuti che posta.
Il meccanismo sopra citato, porta così alla creazione di reti asimmetriche, in quanto le relazioni orientate, possono non essere corrisposte dagli utenti seguiti.

Le relazioni possono sembrare una parte concettuale abbastanza semplice da comprendere, ma in realtà è determinante avere ben chiaro come esse vengano gestite dal Social Network.
Facebook, ad esempio, gestisce le relazioni in modo differente in quanto i legami sono non orientati e i rapporti sono di conseguenza simmetrici.
Risulta quindi necessario capire la natura delle relazioni, in quanto la progettazione e lo studio della rete potrebbe variare in modo consistente.

3. GOAL

L'obiettivo del progetto è quello di dimostrare le caratteristiche fondamentali che un profilo Twitter deve avere, per essere considerato influente all'interno della rete a cui appartiene ed efficiente in termini di spreading dell'informazione.
Si vogliono approvare le ipotesi che un profilo sia predisposto ad ottimizzare la propagazione di informazione quando ha un numero elevato di Followers, quando è posizionato strategicamente all'interno della network e la qualità dei legami in uscita è tale da aumentare la visibilità dei contenuti in una buona fetta di rete.

4. DATASET

I dati sono stati reperiti in open source da un sito web di nome Stanford Network Analysis Project. Questo fornisce numerose risorse, molto utili, per affrontare progetti di social network analysis. Per lo scopo del progetto è stato necessario accedere a queste risorse per quanto riguarda la parte di dataset(https://snap.stanford.edu/data/higgs-twitter.html). Il sito fornisce appunto dataset di sottoreti Twitter in file testuali, i quali hanno la struttura come segue:

...
User-A User-B
User-C User-D
…

Il significato è di facile comprensione, in quanto ogni riga identifica un collegamento unidirezionale della nostra rete. In questo caso "User-A follows User-B" e "User-C follows User-D".
Gli identificativi degli utenti sono stati resi anonimi, per problemi di privacy, e pubblicati come stringhe numeriche di 8-10 cifre.
Inoltre le risorse forniscono anche file relativi alle attività degli utenti Twitter sopracitati, come ad esempio retweet, menzioni, post e altro.
Per lo scopo del progetto è stato necessario considerare anche il file relativo ai retweet, in modo tale da riconoscere i nodi con più successo in tale senso.
Questi file hanno una struttura del tutto analoga a quelli citati in precedenza:

…
User-A User-B n
User-C User-D n
…
dove gli utenti A e C retwettano contenuti, rispettivamente di B e D. Il valore n invece sta ad indicare il numero di retweet, effettuati da A o C verso B e D.
Il file relativo alla rete di Follower/Following originale contiene 456.626 nodi e 14.855.842 collegamenti tra essi, con una dimensione totale del file di circa 170 Mb.
Maneggiare file di questa dimensione nel software Ucinet ed Excel risulta praticamente impossibile, dato che Ucinet riesce a lavorare al meglio con reti che non superino i 5000 nodi. Di conseguenza si è circoscritta una rete di dimensioni idonee facendo riferimento al file dei

retweet, ovvero si sono identificati i 400 nodi che hanno ricevuto il numero maggiore di retweet ed è stata estrapolata la rete, a partire dal file originale, sulla base di questi attori.
Quindi, la rete estrapolata su cui si baseranno le analisi è di 400 nodi, con 4731 collegamenti asimmetrici tra gli utenti.
L'analisi dunque si soffermerà sui nodi che non solo sono i più influenti nella rete di partenza, ma sono i più influenti anche nella sottorete estrapolata.
Il file relativo ai retweet è stato utilizzato per estrapolare i 400 attori con più retweet nella rete generale e per ricavare, conseguentemente, il numero di retweet ricevuti all'interno della sottorete estrapolata.

4.1. PERIODO E TEMA

I dati raccolti fanno riferimento all'attività Twitter dal 1/7/2012 al 7/7/2012.
Gli utenti, i post ed i conseguenti retweet sono relativi alla rilevazione, presso il CERN di Ginevra, del Bosone Higgs, avvenuta il 4/7/2012. Le informazioni quindi circoscrivono tutti quegli utenti che hanno propagato informazione e si sono interessati a questa fondamentale scoperta nel mondo della Fisica.

4.2. SCRIPTING

Una delle prime fasi di progetto è stata quella di pulire e strutturare le informazioni che i tipi di file, descritti in precedenza, ci forniscono.
Per questo è stato necessario implementare una importante fase di scripting, la quale ha permesso appunto di creare quelle risorse che poi verranno importate nei software di analisi, quali UciNet e NetDraw.
Come linguaggio di scripting si è scelto PHP 5, in quanto risulta di facile implementazione la scrittura su un file, la lettura, lo splitting di stringhe e la gestione di array.
Questa parte è stata consistente ed ha richiesto la creazione di svariati script PHP per riuscire ad avere i dati prefissati.
Come prima cosa si è dovuto estrapolare i 400 nodi con maggior retweet, tra i 456.626 nodi totali. Successivamente, avendo ricavato gli attori della nostra network, è stato necessario cercare i link di interazione, tra i 14.855.842 link totali.
A questo punto è stato possibile costruire la matrice di adiacenza 400x400, ponendo 1 alla presenza di un collegamento asimmetrico.
Infine si è anche calcolato il numero di retweet che ogni nodo ha ricevuto dagli attori della sottorete dei 400.

5. ANALISI

Per l'analisi dei dati si è fatto riferimento allo strumento Ucinet 6 ed a NetDraw, per una corrispondenza grafica delle informazioni generate.
Ucinet da la possibilità di importare le network in svariati modi, la soluzione adottata in questo progetto sta nell'importazione di file testuali all'interno del foglio di lavoro di Ucinet, per poi salvarlo/esportarlo nei formati ##h e ##d.

5.1. ANALISI DEGREE

Ipotesi 1: *I nodi più influenti vengono individuati in base all'alto numero di followers, ovvero dai loro collegamenti in entrata. Questi, paragonandoli con l'OutDegree (followings), dovrebbero essere suddivisi in due sottocategorie: quelli intrinsecamente influenti che non hanno bisogno di sollecitare legami per essere seguiti e quelli che hanno un alto OutDegree per cercare di aumentare sempre più i propri followers.*

La prima analisi riguarda una valutazione dei valori di out-Degree ed in-Degree, in quanto essendo una rete con collegamenti direzionali è necessaria la distinzione tra legami in entrata e legami in uscita.
L'analisi dei nodi con più legami, ovvero quelli che riscuotono più successo in termini di followers attivi, porta a pensare a due diverse situazioni.
La prima, paragonando i valori di inDegree con OutDegree, riguarda il fatto che molti utenti abbiano ricevuto la corrispondenza di followers perché abbiano loro, prima di tutti, iniziato a farsi conoscere sulla rete ed instaurare legami. Questi sono quegli utenti che hanno un numero elevato di followers, ma dualmente anche un grande numero di following.
Con tutta probabilità il fatto che questi abbiano un certo successo sulla rete, lo si deve ad un loro precedente sforzo nel seguire anticipatamente, i propri futuri followers.
Gli esempi più evidenti di questo genere di utente possono essere ritrovati nella tabella sottostante, identificati dagli id 1988, 5226, 519, 408, 5549.
La seconda situazione riguarda quei nodi che hanno un elevato numero di followers, ma un numero contenuto di followings. Questi, con tutta probabilità, sono utenti che non hanno bisogno di sforzo per essere popolari nella rete, ne sono interessati all'informazione degli altri utenti.
Questo comportamente è tipico di agenzie giornalistiche, siti di informazione che in questo caso avranno citato le prime indiscrezioni riguardo la rilevazione della particella.
È normale pensare che questi siano la maggioranza, avendo estratto dalla rete originale quei nodi con maggior successo di retweet.

Tuttavia, considerando ora il tema dello spreading dell'informazione, non è detto che un numero elevato di Followers sia matematica conseguenza di una propagazione efficiente dei contenuti. Ad esempio, avere la maggior parte dei seguaci mal connessi alla rete può non portare nessun beneficio al processo di propagazione. Quindi, è necessaria un'analisi più approfondita.

	OutDegree	InDegree	Retweeted
88	31.000	91.000	59
1988	56.000	69.000	32
220	9.000	68.000	1
1503	2.000	65.000	5
206	0.000	65.000	-
8	8.000	63.000	1
677	14.000	63.000	12
138	1.000	60.000	-
301	20.000	59.000	2
1276	12.000	55.000	7
1852	6.000	51.000	1
349	14.000	50.000	16
5226	56.000	50.000	13
519	55.000	49.000	6
352	22.000	48.000	-
1062	20.000	46.000	-
465	8.000	44.000	-
383	9.000	43.000	-
463	0.000	43.000	-
408	45.000	42.000	-
3549	25.000	41.000	1

tab 1. *In e Out Degree dei nodi con maggior n° followers*

Da notare che i Retweet possono sembrare pochi, ma in realtà, dato che si sta considerando solo la sottorete dei 400 nodi più retweettati, della rete totale di circa 400 mila nodi, ottenere anche pochi retweet significa dare una grossa visibilità al contenuto postato.
Inoltre i nodi che non hanno ottenuto retweet nella sottorete considerata, potrebbero anche aver postato poco o nulla.
Quest'ultimo punto, purtroppo, rimane un punto aperto in quanto le risorse SNAP non forniscono la quantità di post che ogni profilo ha effettuato.

5.2. ANALISI RECIPROCITÀ

Ipotesi 2 : *I nodi con più followers saranno maggiormente corrisposti nei loro ingressi in uscita, rispetto ai nodi intrinsecamente più influenti, i quali non avranno una grossa necessità di corrispondere i loro collegamenti in ingresso.*

Uno studio molto interessante riguarda la reciprocità dei legami che interconnettono i nodi, dove per ogni utente è possibile investigare il grado di simmetria e non-simmetria dei collegamenti in entrata(followers) e dei collegamenti in uscita(following).
Ucinet, data una rete in input, riesce a ricavare indici relativi alla simmetria e non-simmetria dei legami, la porzione di legami uscenti corrisposti(colonna Sym/Out) ed infine la porzione dei legami entranti corrisposti(colonna Sym/In).
Come prima operazione è bene soffermarsi brevemente sulla simmetria generale della rete, la quale risulta, come mostrato nella tab.2, relativamente bassa(0.34).
Questo risultato è previsto e preventivato dal fatto che all'interno della network, vi siano molti profili Twitter che non necessitino, in linea di massima, di corrispondere i propri followers(come anticipato alla fine del paragrafo 5.1).

		Measures
1	Recip Arcs	1596
2	Unrecip Arcs	3135
3	All Arcs	4731
4	Arc Reciprocity	0,337
5	Sym Dyads	798
6	Asym Dyads	3135
7	All Dyads	3933
8	Dyad Reciprocity	0,203

tab 2. *Simmetria della Network*

L'ipotesi 1 preventivata, nel precedente studio dei degree, e l'ipotesi sopracitata trova riscontro nell'analisi delle simmetrie.
Gli utenti che si sono "costruiti" in termini di followers, sfruttando le logiche di Twitter, presentano un numero elevato di followings(derivato dallo studio precedente) e caratteristiche di simmetria differenti rispetto a quegli utenti che vengono seguiti, indipendentemente, dal loro comportamento.
I risultati sottostanti, portano a pensare che gli utenti della prima categoria, abbiano in generale una simmetria più alta nei legami rispetto alla seconda categoria, in quanto corrisposti nelle numerose richieste di legami effettuati nel corso del tempo.
Gli utenti della seconda categoria invece, presentano una simmetria nel legami notevolmente più bassa, in quanto vengono seguiti, con tutta probabilità, per la natura dei propri post e delle informazioni che propagano e non per le richieste di legami che effettua.
Per entrare più nello specifico, si considerino i nodi 519, 5226, 408 come rappresentanti della prima categoria ed i nodi 220, 1503, 349, 88, 677 come rappresentanti della seconda categoria.
Come si può comprendere dalla tabella 3, gli utenti che si sono costruiti sfruttando i legami Twitter presentano, una simmetria nei loro legami in input abbastanza alta(0.5-0.6), rispetto agli altri nodi(0.1-0.2).
Questo risultato deriva dal fatto che questi utenti hanno bisogno di instaurare legami, per raggiungere una posizione privilegiata all'interno della network.
Al contrario, gli utenti intrinsecamente interessanti non necessitano di corrispondere i propri followers, presentando, valori di simmetria nei propri legami in uscita(Sym/out in tabella), generalmente più alti degli altri nodi.
Questo perché, essendo profili molto influenti nella rete, verranno quasi sempre

corrisposti nelle loro richieste di collegamenti esterni.

	Symmetric	Non-Symme	Sym/Out	Sym/In	OutDegr	InDegr
220	0.069	0.931	0.556	0.074	9.000	68.000
1503	0.031	0.969	1.000	0.031	2.000	65.000
349	0.164	0.836	0.643	0.180	14.000	50.000
88	0.258	0.742	0.806	0.275	31.000	91.000
677	0.203	0.797	0.929	0.206	14.000	63.000
1988	0.225	0.775	0.411	0.333	56.000	69.000
408	0.381	0.619	0.533	0.571	45.000	42.000
5226	0.325	0.675	0.464	0.520	56.000	50.000
519	0.284	0.716	0.418	0.469	55.000	49.000

tab 3. *Analisi delle simmetrie degli utenti vip*

Nodi appartenenti ad entrambe le categorie hanno comunque ottenuto successo, in termini di retweet, sulla rete. Quest'analisi è servita anche per comprendere che un nodo, anche se non è un profilo noto a priori, può comunque ottenere dei risultati in termini di spreading, mantenendo e creando link.

5.3. CORE/PERIPHERY

Ipotesi 2 : *I nodi ritenuti influenti risiedono nel Core della rete.*

La struttura Core/Periphery si può ritrovare in molti studi riguardanti la Social Network Analysis, in quanto permette di identificare i componenti di core, formato da un nucleo denso e coeso di nodi, ed i componenti di Pheriphery, composto da nodi sparsi e mal connessi.
Nella rete considerata nel progetto, 85 nodi fanno parte del Core e 315 della Pheriphery.
In questo caso la nascita di un Core, può essere interpretata come il fatto che gli utenti più influenti tendano, sempre più, a stringere legami con altri nodi influenti, trascurando quelli con la periferia.
Entrando nello specifico, questa analisi è stata fondamentale per avere una conferma sui nodi considerati come i più influenti ed efficaci nel propagare informazione, in quanto tutti i nodi individuati nel paragrafo 5.1 e tutti i nodi con maggior retweet risiedono nel Core, accertando che essi occupano una posizione di privilegio nella network.

5.4. BETWEENNESS CENTRALITY

Considerando l'analisi di Betweenness Centrality si misura il potere di un nodo, in termine di brokeraggio.
Questa caratteristica è definita come espressione della posizione di intermediazione fra coppie di nodi, ovvero conta il numero di volte che questo compare in un percorso tra tutte le coppie di nodi.
Come prima cosa si pone attenzione alle misure della rete, in generale.
La media di Betweenness Centrality degli attori è ricavata a 792.47, con un indice di centralizzazione del 13.69%.
Questo indica che la rete è abbastanza distribuita nei collegamenti, ma che probabilmente alcuni nodi deterranno valori di centrality molto elevati dati i valori massimi(22482.18) e minimi(0) ottenuti.

		Betweenness	nBetweenness
1	Mean	792.470	0.499
2	Std Dev	1858.335	1.170
3	Sum	316988.000	199.612
4	Variance	3453408.000	1.369
5	SSQ	1632566656.000	647.380
6	MCSSQ	1381363200.000	547.767
7	Euc Norm	40405.031	25.444
8	Minimum	0.000	0.000
9	Maximum	22482.180	14.157
10	N of Obs	400.000	400.000

Network Centralization Index = 13.69%

tab 4. *Betweenness Centrality della network*

Nella tabella 5, seguente, vengono mostrati i nodi con i valori maggiori di Betweenness Centrality.
Questi nodi broker sono in uno status invidiabile per chi dovesse propagare informazione, in quanto i collegamenti instaurati fanno sì che essi siano in una posizione nevralgica della rete.

	Betweenness	nBetweenness	Retweeted
88	22482.180	14.157	59.000
37532	15974.508	10.059	-
5226	11086.049	6.981	13.000
6940	7844.147	4.940	-
1988	6124.776	3.857	32.000

tab 5. *Nodi con alta Betweenness Centrality*

Ricordando che la media di centrality della rete è di 792.47, si nota come i nodi elencati in Tabella 5, abbiano valori molto superiori alla media.
Tra questi, ritroviamo tre dei nodi che hanno ricevuto un importante consenso nel loro processo di spreading(nodo 88, 5226 e 1988).
Questo risultato è importante, perché tende a confermare, che un altro parametro fondamentale per propagare efficacemente informazioni sta nella posizione del nodo all'interno della rete e della qualità dei collegamenti instaurati.

6. CONCLUSIONI

Il progetto riguardava la dimostrazione dei fattori fondamentali che un profilo Twitter deve avere, per ritenersi efficace nel processo di spreading dell'informazione.
Una prima analisi ha verificato che il numero di Followers è, ovviamente, un parametro fondamentale per avere successo.
Non a caso il nodo più retwettato nella rete(88) è quello con più collegamenti in entrata.
Successivamente, si è dimostrato che un profilo, se si dimostra attivo nel creare, corrispondere e mantenere collegamenti, ha la possibilità di rientrare tra quelli più influenti.Lo studio di Core/Periphery prova che i nodi maggiormente retweettati e con alto numero di Followers fanno parte del Core e tendono sempre più ad associarsi tra loro.
L'analisi di centralità, invece, va ad integrare l'ipotesi che non solo il numero di followers è un fattore importante per lo spreading, ma anche la posizione dell'utente rispetto agli altri e alla qualità dei legami che è riuscito ad instaurare nel tempo.

7. BIBLIOGRAFIA E RIFERIMENTI

- *[DATASET]Stanford Network Analysis Project (SNAP), https://snap.stanford.edu/data/higgs-twitter.html , Stanford University*
- Ucinet Software, https://sites.google.com/site/ucinetsoftware/home
- M.Ruffino, http://cs.unibo.it/~ruffino/Dispense%20SNA/Slides%20SNA.pdf , dispense @ Università di Bologna
- PHP Doc, http://php.net/docs.php , per fase di scripting
- A.J. Morales, J. Borondo, J.C. Losada, R.M. Benito, ***Efficiency of human activity on information spreading on Twitter***, Università di Madrid 2014