Giambattista Amati,
Simone Angelini
Fondazione Ugo Bordoni, Roma
Francesca Capri,
Giorgio Gambosi,
Gianluca Rossi,
Università di Roma Tor Vergata
Gianmarco Fusco,
Giuseppe Pierri
ISCOM-Istituto Superiore delle Comunicazioni e delle Tecnologie dell'Informazione, Roma
Paola Vocca
Università della Tuscia, Viterbo

Comparazione tra retweet graph cumulativi e dinamici in twitter

Comparison between cumulative and dynamic retweet graphs on twitter

Sommario: Le proprietà topologiche derivate da reti sociali, come Twitter, forniscono importanti intuizioni sulla natura delle attività sociali o sul modo in cui si diffonde l'informazione attraverso la rete. Esse potrebbero avere anche un impatto rilevante sulla progettazione di nuove applicazioni e sul miglioramento di servizi già esistenti. A tale scopo, in [8,9], i cui contenuti sono riportati in questo articolo, si sono studiati due tipi di Retweet Graph: Cumulative Evolutionary Retweet Graph e Dynamic Retweet Graph (DRG). I Retweet Graph cumulativi tengono conto di come le relazioni tra gli utenti Twitter evolvono nel tempo, una volta che un arco è inserito nel grafo, non verrà mai cancellato [6,7]. I Dynamic Retweet Graph tengono conto del dinamismo degli utenti di Twitter, una volta che un tweet è stato retwittato l'ultima volta tutti gli archi che rappresentano quel tweet sono cancellati. Si analizzano le caratteristiche di questi grafi usando tre differenti flussi di Twitter, costruiti su tre contesti diversi: due sono basati sugli eventi (Black Friday 2015 e World Series 2015) e uno sul campionamento di tutto il flusso di Twitter filtrato per lingua, come ad esempio l'italiano (Italian Sampling). Infine, si sono analizzate alcune metriche standard delle reti sociali per comparare le proprietà strutturali dei Cumulative Evolutionary Retweet Graph e Dynamic Retweet Graph (DRG).

Abstract: Topological properties of graphs derived from social network platforms, like Twitter, give important insights on the nature of the social activities or on the way information spreads over the network. It may have also a relevant impact on designing new applications and improving already existing services. For this reason, we study the Retweet Graphs and the work, presented in this paper, concerns the construction of two types of Retweet Graph: Cumulative Evolutionary Retweet Graph e Dynamic Retweet Graph (DRG). The Cumulative Evolutionary Retweet Graph take into account how users relations evolve over time, once an edge is inserted, it is never deleted [6,7]. The Dynamic Retweet Graph take into account the dynamics of Twitter users, once a tweet has been retweeted the last time all the edges representing this tweet are deleted. We analyze the

characteristics of this graph using three different Twitter streams, built on three different contexts: two are event based (the 2015 Black Friday and the 2015 World Series), and one on the sampling of the whole Twitter stream filtered by language, such as the Italian (Italian Sampling). We use some standard social network analysis metrics to compare the structural properties of the Cumulative Evolutionary Retweet Graph e Dynamic Retweet Graph (DRG).

1. Introduzione

Twitter è una piattaforma di *micro-blogging* che ha caratteristiche specifiche che la rendono diversa da altri reti sociali, come ad esempio *Facebook*, anche a causa della sua apertura verso le interazioni tra utenti.

Infatti, *Twitter* permette diverse azioni da parte degli utenti, ognuna delle quali produce una interazione tra di essi ed induce diversi tipi di rete [1].

Il grafo dei *following/follower* è il tipo di rete più studiato, rappresenta un'interazione statica, ottenuto associando gli utenti ai nodi e inserendo un arco diretto dal nodo *a* al nodo *b* se *a* segue *b*.

Questa è la rete più naturale ed intuitiva per rappresentare l'universo di *Twitter* (*Twitterverse*) ed è stata ampiamente studiata negli anni.

Dal primo studio qualitativo del grafo dei *follower* di *Twitter* [2] è emerso che la distribuzione dei *follower* non seguisse una *power law;* studi successivi hanno osservato che tale grafo [3] esibisse delle caratteristiche strutturali simili sia a quelle delle reti sociali che a quelle delle reti informative; infine si è utilizzato il grafo dei *follower* per identificare *account* autorevoli [4].

Sfortunatamente, i dati del grafo dei *following/follower* sono più difficili da ottenere a causa delle politiche restrittive di accesso a *Twitter,* in più tale grafo potrebbe non essere utile a descrivere il comportamento di *Twitter,* in quanto non esprime completamente come le informazioni si diffondono.

Un altro importante tipo di grafo che deriva da *Twitter* è il *Retweet Graph*. In *Twitter* un account può sia inviare un messaggio (un *tweet*) di massimo 140 caratteri, sia inoltrare un tweet di un altro *account*. Quest'ultimo tipo di messaggio si chiama *retweet*. Un *retweet* è a volte accompagnato dai commenti dei *retweeter*.

Il *Retweet Graph* è definito quindi come un grafo orientato, dove i vertici sono gli *account* e l'arco tra due account A e B esiste se A *retwitta* un *tweet* di B.

Sorprendentemente, questo tipo di rete è stata studiata solo per rilevare lo spam [5].

I *Retweet Graph* potrebbero essere rilevanti anche per studiare la propagazione dell'informazione, poiché essi codificano meglio gli argomenti rilevanti e le relazioni tra gli *account* rispetto ai grafi dei

follower, modellando quindi con più precisione la rete di informazione alla base di *Twitterverse*.

Inoltre, a causa delle limitazioni delle API di *Twitter*, il grafo dei *follower* è difficile da costruire, e solo reti parziali possono essere fisicamente derivate.

In [8,9] si è fatto riferimento a diversi tipi di *Retweet Graph* e si è studiata la loro evoluzione temporale.

Più precisamente, si sono analizzati due classi di *Retweet Graph*: (1) E*vent-driven Retweet Graph* derivati monitorando specifici eventi limitati nel tempo; e (2) *Sampling Retweet Graph* che contengono l'attività di *tweeting-retweeting* della rete di *Twitter* per un periodo di tempo, limitata per lingua, area geografica e / o parole chiave.

Gli *Event-driven Retweet Graph* sono costruiti filtrando dei *tweet* in base a parole chiave (come *hashtag e account* rilevanti tra le tante) specifici dell'evento e seguendone i *retweet*. Differentemente, nel *Sampling Retweet Graph*, i filtri sono rappresentati da una lista di stop word (parole più comuni di uno specifico linguaggio) che permettono di individuare quasi tutto il flusso della rete in una particolare lingua.

Negli articoli [8,9], per gli *Event-driven Retweet Graph*, si sono studiati il *Black Friday* e le *World Serie*s del 2015 e, per quanto riguarda i grafi con *Sampling*, si sono filtrati i *tweet* utilizzando un elenco di stop word italiane, che hanno consentito di isolare le attività di *tweeting-retweeting* italiane, ottenendo così l'*Italian Sampling Retweet Graph*.

Gli esperimenti hanno rivelato una sostanziale differenza tra le due classi di grafi, sia confrontando i loro stati finali, sia la loro evoluzione. Si ipotizza che ci siano due possibili interpretazioni che spieghino questa differenza. Un motivo potrebbe essere una duplice natura della rete di *Twitter*: rete sociale e di informazione. Infatti, l'*Italian Sampling* è più simile ad una rete sociale, poiché mostra un diametro minore, un coefficiente di *clustering* maggiore e una massima componente connessa più grande rispetto ai due grafi di tipo *Event-driven*.

Il lavoro, presentato in [8,9], riguarda la costruzione di due tipi di *Retweet Graph: Cumulative Evolutionary Retweet Graph* e *Dynamic Retweet Graph (DRG).*

Nel caso cumulativo, si è costruita una sequenza di grafi, derivata da delle istantanee del *Retweet Graph* prese ogni 4 ore; in tali grafi quando un arco è inserito non verrà mai cancellato.

Questo approccio ci ha permesso di studiare l'evoluzione temporale dei *Retweet Graph,* di valutare la crescita della rete di Twitter e di scoprire discontinuità nell'evoluzione, causate dall'intervallo temporale di esistenza degli eventi, nel caso di *Event-driven Retweet Graph.*

Mentre, i *Dynamic Retweet Graph* sono una variante dei *Retweet Graph* e vengono costruiti considerando il dinamismo degli utenti di *Twitter.*

In un *DRG,* una volta che un *tweet* è stato retwittato per l'ultima volta, tutti gli archi che rappresentano quel *tweet* saranno cancellati, in questo modo si riesce a modellare la fine di un *tweet* in un flusso di un mezzo sociale.

Si è analizzata l'evoluzione dei *DRG,* su un periodo di due mesi. Infine, si è fatta una comparazione tra i *Cumulative Evolutionary Retweet Graph* ed i *Dynamic Retweet Graph* attraverso le principali misure strutturali usate, generalmente, per caratterizzare la natura dei grafi: distanza media, coefficiente di clustering e distribuzione dei gradi entranti ed uscenti.

Inoltre, le misure, appena citate, permettono di valutare le proprietà topologiche di questi grafi e sono di fondamentale importanza per valutare la struttura e per prevedere l'evoluzione della rete di *Twitter*, sia dal punto di vista sociale che dal punto di vista dell'informazione.

I risultati in [8,9], ottenuti dalla comparazione, hanno mostrato una significativa differenza tra i grafi cumulativi ed i corrispondenti *DRG,* sia nel modo in cui essi crescono e sia nel modo in cui le misure, citate in precedenza, si evolvono.

Dalla comparazione, è emerso che il *DRG dell'Italian Sampling* mantiene le stesse proprietà strutturali del corrispettivo grafo cumulativo, mentre questo non accade per gli *Event-driven Retweet Graph.*

2. Dataset e costruzione dei grafi

2.1 Descrizione del dataset

In [8,9] si è eseguita un'analisi dell'evoluzione dei *Retweet Graph* utilizzando tre differenti collezioni di *Twitter*, che sono state costruite attraverso il monitoraggio delle attività in tre diversi contesti. Due di queste collezioni sono *Event-driven*, ossia grafi ottenuti osservando gli eventi del *BlackFriday 2015* e le *World Series 2015*, mentre il terzo, *Italian Sampling*, è un grafo ottenuto dal campionamento del flusso italiano, senza osservazione di un particolare evento.

Questi flussi sono stati ottenuti come di seguito descritto:

1. **Italian Sampling.** L'estrazione dei *tweet* si è basata su un insieme di stopword tipiche della lingua italiana: articoli, preposizioni, congiunzioni, avverbi ecc. I *tweet* sono stati quindi filtrati per lingua, selezionando solo quelli italiani

2. **World Series.** Le parole chiave utilizzate per l'estrazione dei *tweet* pertinenti all'evento sono state: #worldseries, world series, #Royals, #Mets, @Mets, #Royals, @Royals, @KCRoyals. I *tweet* sono stati anche filtrati sulla lingua, selezionando solo quelli inglesi. L'evento è iniziato il 27-10-2015 e si è concluso il 01-11-2015. Dalla tabella 1, si può vedere che il periodo di osservazione è iniziato dopo l'inizio dell'evento.

3. **Black Friday 2015.** Le parole chiave utilizzate per estrarre i *tweet* pertinenti sono state: blackfriday, black friday,@blackfriday_fm. L'evento si è verificato il 27-11-2015, e rientra nel periodo di osservazione.

Tabella 1:Descrizione delle collezioni e relative caratteristiche principali.

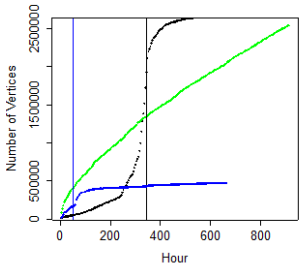
Tabella 1: Descrizione delle collezioni

	Inizio	Fine	# tweets
Italian	10-26-2015	12-12-2015	74,749,330
Sampling	11:27:12AM	05:50:45AM	
World series	10-30-2015	11-27-2015	1,932,330
	2:07:07PM	11:06:06AM	
Black Friday	11-13-2015	12-18-2015	9,891,4
	2:38:16PM	1:32:17PM	

Tabella1. Descrizione delle collezioni e relative caratteristiche principali.

2.2 Costruzione del Cumulative Evolutionary Retweet Graph

I grafi sono stati costruiti come di seguito descritto: l'insieme dei vertici rappresenta gli utenti e vi è un arco diretto da un utente A verso un utente B, se A "*retweetta*" un *tweet* di B. Ogni arco (A, B) è etichettato inoltre con il *timestamp* del primo *retweet* che A ha effettuato verso B. Si usano i *timestamp* per costruire le sequenze temporali dei grafi. Ogni quattro ore, si costruiscono i grafi fino a quell'istante di tempo.
Si noti che la grandezza dei grafi non diminuisce, poiché solo i nuovi vertici (*account*) e archi (*retweet*) vengono aggiunti e non ci sono cancellazioni.
In Figura 1, si mostra l'evoluzione delle dimensioni dei tre dataset durante il periodo di osservazione. È interessante notare che, mentre, nel grafico dell'*Italian Sampling* la crescita è pressoché lineare nel periodo di osservazione, per i grafici *Event-driven* questo non accade, in quanto entrambe le dimensioni raggiungono un punto di saturazione dopo la fine del relativo evento (dopo circa 100 ore per le *World Series*, 400 ore per il *Black Friday*).



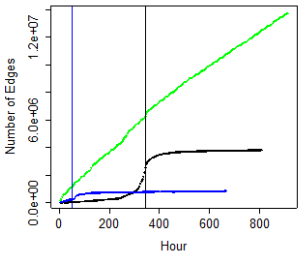


Figura 1. *Numero di vertici e numero di archi di Italian Sampling (verde), World Series (blu), and Black Friday (nero) in funzione delle ore. Le linee verticali blu e nere indicano la fine dell'evento corrispondente.*

Figura 2. Numero di vertici (blu) e numero di archi (verde) di: Italian Sampling, World Series e Black Friday in funzione delle ore.

2.3 Costruzione del Dynamic Retweet Graph

Il grafo DRG G=(V,E,l) è definito come segue:

V è l'insieme dei nodi che rappresenta gli account di Twitter ed e∈ E rappresenta l'interazione (retweet) tra due account.

In particolare, c'è un arco diretto da un account a verso un account b, se a ha retwittato almeno un tweet di b, che può già essere un retweet.

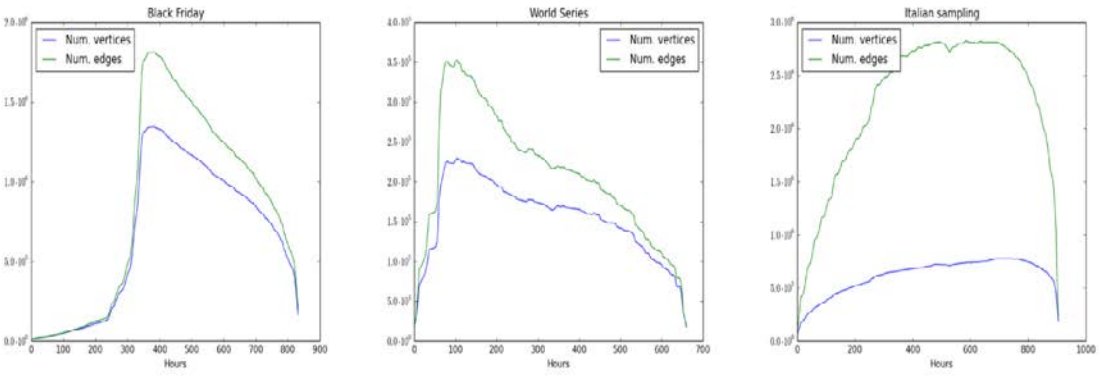
Si osservi, che a potrebbe retwittare più tweet di b. Quindi, per riuscire a distinguere ogni arco e = (a,b), si associa ad esso una lista l(e) che contiene la coppia (i,t) dove i è l'id del tweet originale e t è il timestamp in cui a retwitta i da b.

Le coppie di l(e) sono ordinate per timestamp in ordine non decrescente.

Dai dati in G, è possibile definire per ogni tweet i, la data di nascita, ovvero il timestamp del primo retweet di i e la data di morte di i, ossia il timestamp dell'ultimo tweet di i.

Con queste definizioni, è possibile costruire nel tempo una serie di sottografi DRG $G_t = (V_t, E_t)$ al tempo t dove V_t contiene l'insieme di nodi vivi ed E_t contiene un qualsiasi arco e in E di un tweet vivo.

In figura 2 si mostra l'evoluzione delle dimensioni dei tre dataset durante il periodo di osservazione.



E' interessante notare che gli Event-driven Retweet Graph ed il Sampling Retweet Graph evolvono in due modi diversi: gli Event-driven mostrano una rapida crescita vicina all'evento e quindi, un lento declino; il Sampling graph ha una crescita lenta ed un rapido declino.

Riguardo Event-driven Retweet Graph, la rapida crescita in prossimità dell'evento è giustificata dall'interesse per quell'evento, anche la graduale perdita di interesse spiega il lento declino.

3. DESCRIZIONE DELLE MISURE

Nel seguente paragrafo si descriveranno i risultati in [8,9], ottenuti dalla comparazione delle misure strutturali eseguite sui due tipi di grafi: Cumulative Evolutionary Retweet Graph e Dynamic Retweet Graph .

3.1 Distanza media

Il cammino minimo tra due nodi a e v del grafo è il numero minimo di passi a partire da a per raggiungere v. La distanza media è definita come la media aritmetica dei cammini minimi tra tutte le coppie di nodi raggiungibili del grafo. L'andamento della distanza media nel tempo è mostrata in Figura 3, dalla quale si può notare che nell'Italian Sampling la grandezza della distanza media è uguale in entrambi i grafi; al contrario gli Event-driven DRG Graph sono molto instabili.

Essi non convergono e la crescita ed il decadimento sono molto rapidi. In più, la grandezza della distanza media degli Event-driven DRG Graph è molto più piccola rispetto al corrispettivo cumulativo.

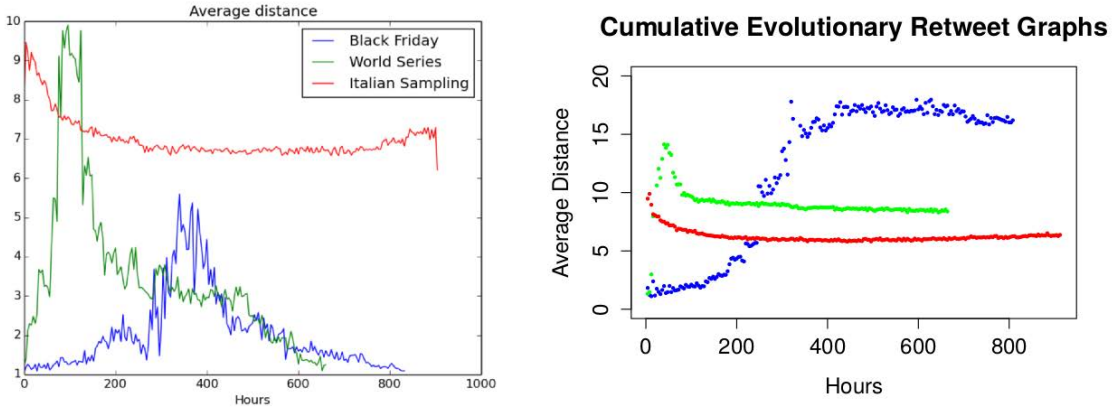


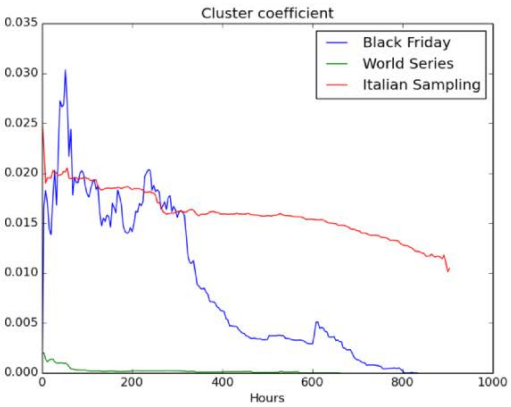
Figura 3. Sulla sinistra si vede l'andamento della distanza media nei, sulla destra quello nei corrispondenti grafi cumulativi.

3.2 Coefficiente di clustering

Come seconda caratteristica, si è considerata l'evoluzione del coefficiente di clustering globale.

Il coefficiente di clustering globale quantifica la probabilità che se un vertice A è connesso al vertice B e il vertice B è collegato al vertice C, allora anche il vertice A sarà collegato al vertice C. In altre parole, la probabilità che l'amico del tuo amico sia anche tuo amico.

In Figura 4, si mostra l'andamento del coefficiente di clustering nei due retweet graph costruiti.



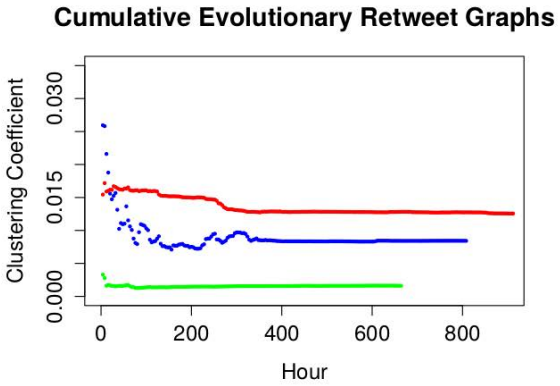


Figura 4. Sulla sinistra si vede l'andamento del coefficiente di clustering nei DRG, sulla destra quello nei corrispondenti grafi cumulativi.

3.3 Massimo grado uscente/entrante

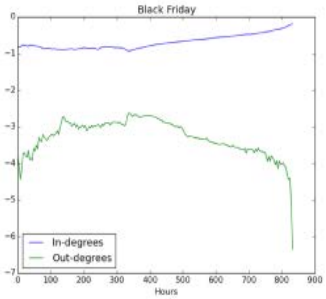
Il grado uscente di un nodo v è il numero di archi uscenti dal nodo considerato.

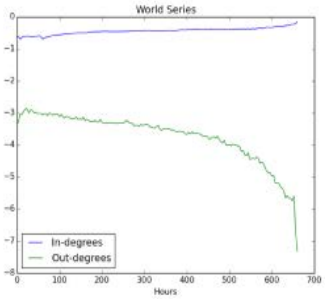
Di conseguenza, il massimo grado uscente è il più grande grado uscente per ogni nodo v appartenente al grafo.

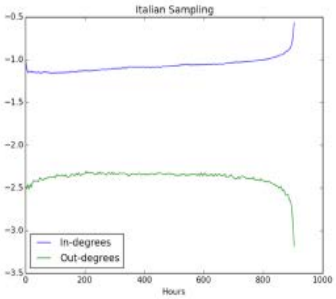
In maniera simile si definisce il massimo grado entrante.

Dalla figura 5 si deduce che le distribuzioni dei gradi entranti nei tre grafi seguano una distribuzione power-law. Stesso comportamento è osservato nelle distribuzioni dei gradi uscenti dei tre grafi.

Figura 5. Distribuzione dei gradi entranti (blu) e distribuzione dei gradi uscenti (verde) di: Italian Sampling, World Series e Black Friday in funzione







CONCLUSIONI

Grazie alle tecnologie Big Data, si è effettuata un'analisi approfondita su due tipi di Retweet Graph: Cumulative Evolutionary Retweet Graph e Dynamic Retweet Graph (DRG), relativamente a tre differenti flussi provenienti da Twitter, due di tipo Event-driven (Black Friday, World Series) e uno basato su campionamento (Italian Sampling).

Il grafo relativo all' Italian Sampling sembra avere caratteristiche più simili ad una rete sociale, rispetto ai grafi Event-driven, infatti mostra una distanza media più piccola ed un alto coefficiente di clustering.

Questo è uno dei primi articoli che studia sistematicamente l'evoluzione temporale di grafi generati da una rete sociale e il tempo di vita di un tweet ed il suo naturale decadimento.

Dall'analisi eseguita, si è visto che i DRG relativi all'Italian Sampling mantengono le stesse proprietà strutturali dei corrispettivi grafi cumulativi.

Al contrario, i DRG relativi ai grafi Event-driven sono diversi dai corrispettivi grafi cumulativi, inoltre mostrano degli effetti di bordo su tutte le misure strutturali, crescono e decadono in modo simile e raggiungono un picco a metà del loro tempo di vita.

RINGRAZIAMENTI

Questo lavoro è stato condotto presso il Laboratorio di Big Data di ISCOM-MISE (Istituto Superiore delle Comunicazioni e delle Tecnologie dell'Informazione, Ministero dello Sviluppo Economico).

BIBLIOGRAFIA

[1] Amati G., Angelini S., Bianchi M., Fusco G., Gambosi G., Gaudino G., Marcone G., Rossi G. and Vocca P. 2015. Moving Beyond the Twitter Follow Graph*. In proceeding of DART 2015,* DOI: 10.5220/0005616906120619.

[2] Kwak H., Lee C., Park H., and Moon S. (2010). What is twitter, a social network or a news media? *In Proceedings of the 19th International Conference on World Wide Web, WWW '10*, pages 591–600, New York, NY, USA, 2010. ACM.

[3] Myers S.A:, Sharma A., Gupta P., and Lin J. (2014). Information network or social network?: The structure of the twitter follow graph. *In Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion, WWW Companion '14*, pages 493–498, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.

[4] Java A., Song X., Finin T., and Tseng B. (2007) Why we twitter: Understandingmicroblogging usage and communities. *In Proceedings of the 9th WebKDD and 1st SNAKDD 2007 Workshop on Web Mining and Social Network Analysis, WebKDD/SNA-KDD '07*, pages 56–65, New York, NY, USA, 2007. ACM.

[5] Bild D.R., Liu Y., Dick R. P., Z. Morley Mao, and Dan S. Wallach. 2015. Aggregate characterization of account behavior in Twitter and analysis of the Retweet Graph. *ACM Trans. Internet Technol*., 15(1):4:1–4:24, March 2015.

[6]Jure Leskovec, Deepayan Chakrabarti, Jon Kleinberg, Christos Faloutsos, and Zoubin Ghahramani. Kronecker graphs: An approach to modeling networks. J. Mach. Learn. Res., 11:985–1042, March 2010.

[7] Jurij Leskovec, Deepayan Chakrabarti, Jon Kleinberg, and Christos Faloutsos. Realistic,mathematically tractable graph generation and evolution, using kronecker multiplication. In Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD'05, pages 133–145, Berlin, Heidelberg, 2005. Springer-Verlag.

[8] Amati G., Angelini S., Capri F., Gambosi G., Rossi G. and Vocca P. 2016. Twitter Temporal Evolution analysis: comparing event and topic driven retweet graphs. *In proceeding of the* International Conference on Big Data Analytics, Data Mining and Computational Intelligence (BigDaCI2016), 2-4 July 2016 Funchal, Madeira, Portugal

[9] Amati G., Angelini S., Capri F., Gambosi G., Rossi G. and Vocca P. 2016. *On the Retweet Decay of the Evolutionary Retweet Graph. In proceeding of the 2nd* International Conference on Smart Objects and Technologies for Social Good (GoodTechs2016), november 30-december 1, 2016 Venice,Italy.