Keeping Up with the Kardashian

Lorenzo Filippo Maria Colombo[1], Beatrice Fumagalli[2], Matteo Porcino[3]

[1]l.colombo71@campus.unimib.it, 763581, *LM Data Science, Università degli studi di Milano-Bicocca, Milano*
[2]b.fumagalli9@campus.unimib.it, 784549, *LM Data Science, Università degli studi di Milano-Bicocca, Milano*
[3]m.porcino1@campus.unimib.it, 748876, *LM Data Science, Università degli studi di Milano-Bicocca, Milano*

30 giugno 2019

1 Introduzione

Con *Social Media Analytics* o *Social Network Analysis* si intende quell'insieme di tecniche e metodologie volte allo studio di strutture sociali mediante l'uso di **reti** e della **teoria dei grafi** a partire dagli inizi del 20° secolo. Questa disciplina, già importante per lo studio delle dinamiche, dei comportamenti e degli effetti che può avere un evento all'interno delle reti sociali, ha visto un repentino sviluppo grazie all'avvento dei **social network**, mondi digitali in cui ognuno, tramite il proprio alter-ego virtuale, è in grado di creare relazioni e di interagire con gli altri soggetti connessi. Studiare questi collegamenti attraverso la *Social Network Analysis* è diventato di fondamentale importanza, essa è oramai applicata all'interno di molteplici contesti, dalla sicurezza al marketing e dalla medicina alla biologia.

1.1 La Famiglia Kardashian-Jenner

I **"Kardashian"** nascono nel 1978, quando *Kris Houghton*, giovane hostess californiana, sposa *Robert Kardashian*, avvocato e uomo d'affari di spicco nella scena Californiana, divenuto famoso per lo stretto legame d'amicizia con *O. J. Simpson* e successivamente per esserne stato l'avvocato difensore nel famosissimo processo del 1995. Da questo matrimonio, finito a seguito del divorzio con *Robert* nel 1991, *Kris* ha quattro figli *Kourtney*, *Kim*, *Khloé* e *Rob*, rispettivamente nel '79, '80, '84 e '87. Un mese dopo il divorzio con *Robert*,

Kris sposa *Bruce Jenner*, ex atleta olimpico Americano, da cui ha due figlie, *Kendall* nel 1995 e *Kylie* nel 1997. A seguito della transizione di *Bruce*, resa pubblica nel 2015, ora conosciuta come *Caitlyn*, l'ex signora Kardashian divorzia nuovamente nel 2013. Lasciando da parte i "drama" della famiglia *Kardashian-Jenner*, essi sono altresì noti per l'enorme potere mediatico, alcuni degli account *Twitter* con più followers al mondo sono posseduti dai membri della famiglia, e per le grandi doti imprenditoriali. I "Kardashian-Jenner" valgono più di un miliardo e mezzo di dollari, *Kylie* con la sua compagnia *Kylie Cosmetics* vale da sola un miliardo di dollari e segue *Kim* con un valore di 350milioni di dollari.

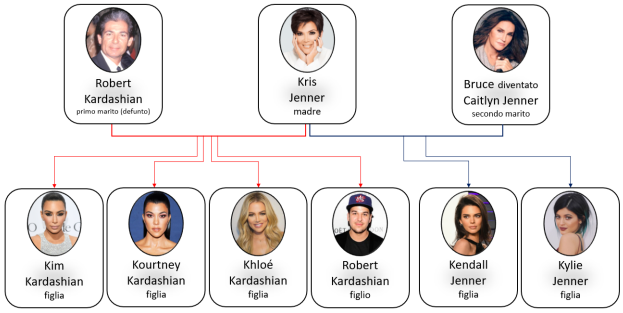

Figura 1: *La famiglia Kardashian-Jenner*

1.2 Keeping Up with the Kardashian

Con un background simile poteva questa *media-mogul-family* tirarsi indietro dal richiamo della televisione? È il 2007, e su **E! Entertainment**, canale americano dedicato allo showbiz, inizia *Keeping up with the Kardashians*, un reality che segue le vicende della vita quotidiana della famiglia "famosa per essere famosa". Ne fanno parte la mamma Kris, da allora ribattezzata *"momager"*, crasi tra mamma e manager, Caitlyn Jenner, ma soprattutto ci sono le tre ambiziose sorelle Kourtney, Kim e Khloé, con le loro vicende amorose, i matrimoni, le gravidanze, le liti e la loro vita da milionarie. Kendall e Kylie, che all'inizio erano poco più che bambine, hanno avuto modo di trarre beneficio anche loro dalla fama globale derivata dallo show televisivo. Oggi i social network fanno più di uno show televisivo e le sorelle Kardashian-Jenner li stanno sfruttando nel migliore dei modi. Basti pensare che quattro posizioni della top 10 degli account più seguiti su Instagram sono occupate proprio da loro (Kim è prima, Kendall è settima, Kylie è ottava, Khloé è decima). Mediaticamente non hanno rivali. Attualmente il programma televisivo è arrivato alla sua 16° edizione, senza contare i numerosi spin-off tra cui *Kourtney and Khloé Take Miami*, *Kourtney and Khloé Take New York*, *Kourtney and Khloé Take The Hamptons* e molti altri incentrati sulle vite dei membri più giovani di questa famiglia.

2 Obiettivo dell'analisi

Lo scopo di questo progetto è osservare il mondo di Twitter che gravita attorno a queste personalità di spicco e come la rete di followers si comporta e reagisce nei loro confronti. È stata infatti effettuata una prima indagine, nel corso del mese di Febbraio 2019, per poter capire quale fossero i topic principali che suscitavano l'interesse degli utenti nei confronti di questa famiglia. Da questa analisi preliminare è risultato evidente che era proprio il loro programma televisivo l'argomento di maggiore interesse, era inoltre programmata per Marzo 2019 l'uscita della nuova edizione del programma, la 16°. Per queste ragioni si è analizzato il flusso di tweet ad essa collegato nel corso della premiere e della settimana puntata. Nelle sezione successive di questa relazione verranno mostrate più nel dettaglio i diversi step di questa analisi, andando poi a presentare i risultati ottenuti.

3 Twitter

L'utilizzo di Twitter, a differenza di altre piattaforme social, si presta particolarmente bene per scopi come questi. La differenza sostanziale rispetto ad altri social risiede nel fatto che un utente può interagire con chiunque all'interno della rete senza la necessità di instaurare una relazione "bidirezionale" come quella di amicizia che deve sussistere tra gli utenti di Facebook. Al suo interno è infatti possibile creare, condividere e commentare contenuti istantaneamente.

Following e Followers Come detto sopra, all'interno di Twitter non esiste una relazione di "amicizia" come su Facebook, le relazioni che si possono instaurare tra gli utenti sono quelle di *following* e *followee*. Si instaurano nel momento in cui un utente A **segue** un utente B e A diventa un **follower** di B.

Retweet e Quote Questo tipo di relazioni non possono sussistere tra utenti di Twitter, ma si possono creare tra i *tweet* che gli utenti pubblicano. Nel caso del *retweet*, un utente B può decidere di *retweettare* il tweet di un utente A condividendolo all'interno della propria bacheca così come l'utente originale lo ha pubblicato. La *quote* invece è vista più generalmente come un commento al contenuto creato da altri, sia esso un commento diretto sotto ad un *tweet* o un commento ad un *tweet* che è stato ritweettato.

Hashtag e Mention Ultimi, ma non per importanza, si trovano *hashtag* e *mention*. Il primo è nato proprio all'interno di questa piattaforma, attraverso gli *hashtags* è possibile etichettare i propri tweet e aggiungerli all'insieme di tweet che parlano di quell'argomento, cliccando su di essi è altresì possibile recuperare il contenuto che si riferisce a quel particolare topic. Le

mention servono invece per taggare altri utenti all'interno di un tweet o di un commento, richiamandone così l'attenzione.

4 Descrizione Tecnica dell'ambiente di lavoro

In questa sezione verrà brevemente mostrato l'ambiente di lavoro che è stato utilizzato nel corso delle diverse fasi di questo progetto, partendo dallo streaming dei tweet riguardanti i *Kardashian* fino ai tools utilizzati per effettuare le visualizzazioni e rappresentare la rete Twitter risultati dai dati scaricati. I linguaggi utilizzati all'interno del progetto sono **R** e **Python**. Per la parte dedicata all'analisi dei grafi relativi alle puntate è stato utilizzato il servizio di notebook messo a disposizione da **Google Colab**.

4.1 Catturare lo streaming

Per poter scaricare il flusso di dati streaming proveniente da Twitter nelle 3 diverse fasi di download è stata utilizzata una macchina virtuale *Azure*. All'interno di questa VM, messa a disposizione dall'università, è presente un ambiente *Hortonworks* che rende disponibili i servizi di **Apache Kafka** e **MongoDB**.

4.1.1 Apache Kafka

Kafka, per semplicità, è una piattaforma software open-source di *stream processing*. Esso implementa un'architettura fortemente scalabile di code di messaggi attraverso i suoi due elementi fondamentali, i *publisher* o *producer* e i *subscriber* o *consumer*. Il compito del primo è quello di catturare il flusso in stream dei dati e di salvare i diversi messaggi intercettati all'interno di uno o più *topic*, il secondo invece fa una "subscribe" a uno o più topic e ne consuma i messaggi al suo interno, il tutto in maniera totalmente asincrona. L'utilizzo di questa piattaforma garantisce un altissimo throughput dei dati, permettendo di catturare un numero più elevato di tweet in questo specifico caso.

4.1.2 MongoDB

MongoDB è un servizio open-source di memorizzazione dei dati di tipo non relazionale. A differenza dei DBMS relazionali classici, di tipo SQL, i cui dati sono memorizzati in tabelle il cui schema logico viene determinato a priori, un DMBS di tipo non relazionale, o NoSQL, non ha uno schema logico definito a priori permettendo così di memorizzare dati con diversi formati. Questa sua caratteristica si presta molto bene per salvare i tweet catturati tramite *Kafka*, in quanto l'oggetto che viene scaricato ha una struttura variabile, a seconda delle informazioni che contiene.

4.1.3 Pipeline per la cattura e il salvataggio

Sono stati definiti due script in linguaggio **Python**, successivamente caricati sulla macchina virtuale, per poter assolvere al compito di cattura dei tweet.

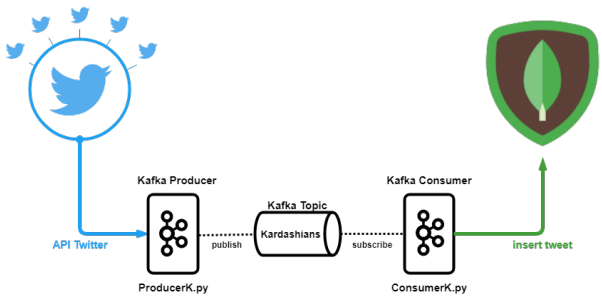

Figura 2: *Pipeline per il salvataggio dello streaming*

`Producer.py` All'interno di questo script, attraverso la libreria `tweepy`, il *Producer* si connette alle API di Twitter e attraverso il metodo `stream(filter=...)` inizia ad intercettare lo streaming di tweet, filtrandoli per le keywords passate come parametro. In questo caso le keywords utilizzate sono state i nomi e gli hashtag degli account ufficiali della famiglia *Kardashian*. Una volta intercettato il tweet viene incapsulato in un messaggio che viene pubblicato all'interno del topic "*Kardashian*".

`Consumer.py` Il *Consumer*, una volta fatta la subscribe al topic "*Kardashian*", inizia a scaricare i messaggi presenti all'interno del topic, li trasforma in formato `JSON` e li memorizza all'interno di un'istanza di *MongoDB*.

4.1.4 File scaricati

Alla fine di ogni sessione di streaming, una volta chiusa la connessione del `producer`, è stato possibile scaricare il contenuto presente all'interno della collection **kuwtk19** in cui venivano salvati i tweet.
Sono stati prodotti così i seguenti file:

- `stream_preliminare.json`
- `stream_1_episode.json`
- `stream_7_episode.json`

4.2 Data pre-processing

Il file `Data cleaning and pre-processing.ipynb` è il notebook utilizzato per trasformare i file `json` iniziali in formato `csv`, più maneggevole per le analisi successive. Fatta eccezione per alcune trasformazioni applicate al campo `text` dei tweet quali:

- rimozione dei link presenti
- rimozione delle emoji
- estrazione del testo del tweet per esteso se troncato perchè troppo lungo

le altre operazioni fatte su questi file sono state per la maggior parte di decodifica del file in formato `json` e di estrazione da alcuni campi complessi di informazioni che sarebbeto tornate utili successivamente nell'analisi, come:

- la data di creazione del tweet che è stato retweettato
- la data di creazione del tweet che è stato commentato
- gli handle degli account degli utenti che sono stati, retweettati o mezionati

per citarne alcune. Dopo aver creato i nuovi campi con all'interno le informazioni appena estratte, sono state eliminate le colonne in cui non era presente alcun tipo di dato ed infine sono stati scritti i file csv contenenti i dati dei tweet scaricati durante la prima e la settimana puntata.

- `first_episode.csv`
- `seventh_episode.csv`

4.3 Data extraction and analysis

Successivamente alla creazione dei file in formato `csv` sono stati scritti quattro notebook:

- `graph-data-1-episode.ipynb`
- `graph-data-7-episode.ipynb`
- `sma-1-episode.ipynb`
- `sma-7-episode.ipynb`

Come si può evincere dai nomi dei file, le operazioni che effettuano sono le medesime sui dati della prima e della settimana puntata, la necessità di duplicare il codice è stata necessaria per via delle dimensioni del file inerente ai tweet della prima puntata, pesando all'incirca 1GB l'esecuzione del codice era estremamente difficoltosa, sono stati creati così dei file separati per fare le analisi sulle due puntate.

4.3.1 I File `graph-data-x-episode`

All'interno di questi notebook i dati scaricati vengono processati per estrarre tutti gli utenti e le relazioni presenti. I dati riguardanti gli utenti di twitter si trovano all'interno del campo `user` presente all'interno dell'oggetto `tweet`. Al fine di estrarre tutti gli utenti coinvolti, per poi successivamente visualizzare il grafo della puntata, è stato necessario estrarre i dati degli utenti contenuti anche all'interno dei campi:

- `retweeted_status`: Se il tweet scaricato è un retweet, all'interno di questo campo è presente il tweet originale al cui interno, nel suo campo `user` sono presenti i dati dell'utente il cui tweet è stato retweettato
- `quotes_status`: Similmente al caso appena spiegato sopra, se il tweet scaricato è un commento ad un tweet o un retweet, al suo interno sarà presente il tweet che è stato commentato così come i dati dell'utente che ha pubblicato in origine quel tweet
- `entities`: All'interno di questo campo sono presenti gli eventuali utenti che vengono menzionati all'interno del testo del tweet scaricato, da questo campo è possibile estrarne il nome utente, `screen_name` e l'id dell'utente, `id_str`. Successivamente tramite la libreria `tweepy` sono stati scaricati i dati riguardanti gli utenti che comparivano solamente all'interno di questo campo.

Allo stesso modo sono state estratte tutte le interazioni presenti all'interno dei dati scaricati. Le interazioni considerate sono state: i retweet, le quote o commenti e le menzioni degli utenti. Terminata l'estrazione delle informazioni vengono scritti due file:

- `user_data_x_episode.csv`
- `relation_data_x_episode.csv`

Questi file saranno poi utilizzati all'interno dello script `R` per analizzare il grafo della puntata, visualizzare il grafo delle interazioni e all'interno del programma *gephi* per avere una visualizzazione globale del grafo della puntata.

4.3.2 I File `sma-x-episode`

Così come per i file descritti nella 4.3.1, anche per la parte di analisi sono stati creati due notebook, uno per ciascuna puntata. Nei paragrafi a seguire verrano brevemente descritte le analisi effettuate.

Delta e Boxplot per prima cosa viene calcolata la differenza di tempo tra la creazione del tweet e quando avviene l'interazione, sia essa un retweet o una quote. Successivamente al calcolo di questa differenza di tempo, denominata *Delta*, vengono generati i boxplot dei tempi di risposta, espressi in minuti, per ogni componente della famiglia considerato, sia dei retweet che delle quote.

Calcolo coverage Sfruttando il *delta* calcolato precedentemente, viene assegnato ad ogni tweet un "gruppo". Per calcolare il gruppo da assegnare viene utilizzato il quoziente della divisione: $delta/5$. Ogni gruppo rappresenta quindi l'insieme di nodi che interagiscono con il tweet originale dopo $gruppo * 5$ minuti dalla sua pubblicazione. Sotto l'ipotesi che la rete a disposizione corrisponda a quella reale, è possibile così calcolare il **coverage**, ad intervalli di 5 minuti, che hanno i singoli tweet.

Sentiment Analysis Dal momento che il testo contenuto all'interno dei tweet rappresenta l'informazione principale in essi contenuta, viene effettuata una *sentiment analysis* su di essi. In questo caso specifico, l'analisi non viene effettuata su tutti tweet della puntata, ma

solamente su quei tweet che vengono individuati come quote. Essi sono infatti considerabili come i commenti degli utenti, esprimono perciò un'opinione soggettiva ed è pertanto interessante studiarne il sentiment per avere un'idea di come reagisce il pubblico, che segue *KUWTK*, ai contenuti postati dalla famiglia. La polarità (positiva, neutrale, negativa) è stata calcolata utilizzando la libreria *TextBlob* disponibile in Python. Il modulo *textblob.sentiments* contiene due implementazioni di *sentiment analysis*: *PatternAnalyzer*, di default e *NaiveBayesAnalyzer*. Si è scelto di utilizzare il primo in quanto mostrava una migliore classificazione del sentiment dei quotes. Quest'ultimi vengono poi visualizzati con un istogramma che mostra la porzione di tweet classificati come positivi e quelli classificati negativi, i neutrali sono stati omessi in quanto privi di significato.

Wordcloud Similmente a come fatto durante l'analisi preliminare viene generata una wordcloud utilizzando come dati il testo delle quote della puntata, mostrando così gli argomenti che hanno suscitato più movimento all'interno della comunità, la stessa viene poi visualizzata considerando prima le quote etichettate come positive dalla sentiment ed infine quelle negative.

4.3.3 Colab - SMA_graph_k.ipynb

Come per il file illustrato sopra, anche questo notebook utilizza i dati in formato `csv` relativi agli utenti e alle interazioni, durante lo stream, per effettuare le analisi sulla rete di queste ultime. Grazie alla libreria `networkX`, alle funzioni che implementa e alle risorse messe a disposizione da **Google Colab** è stato possibile creare i grafi completi delle due puntate. Per ogni grafo sono stati calcolati:

- betweenness centrality (su un campione degli utenti)
- centrality rispetto agli autovalori
- grado dei nodi
- ponti
- edge betweenness
- modularità
- community
- assortatività

4.4 Data visualization tool

La maggior parte degli output visivi dell'analisi sono stati generati o direttamente all'interno del codice, utilizzando le opportune librerie sia per il linguaggio `Python` che per il linguaggio `R` o utilizzando **Tableau**, un programma apposito per la creazione di visualizzazioni. Nessuno di questi però è stato in grado di produrre un output che mostrasse il grafo completo degli utenti e delle loro interazioni durante ciascuna puntata, questo per via della dimensione dello stesso.

4.4.1 Gephi

Gephi è un software open-source nato appositamente per effettuare analisi e creare visualizzazioni di reti sociali. È in grado di gestire ed elaborare reti formate da un grande numero di nodi ed archi, sono inoltre disponibili diversi plug-in sviluppati dalla comunità che utilizza il software, alcuni dei quali specializzati nel calcolo del layout ottimale per la visualizzazione di grafi di grandi dimensioni.

5 Analisi preliminare

Questa analisi preliminare, come è stato detto precedentemente nella sezione 2, è stata effettuata scaricando in maniera totalmente casuale, senza uno scopo preciso, i tweet riguardanti i *Kardashian*. Come filtri per la cattura dello stream sono stati utilizzati i nomi dei loro account personali e i rispettivi hashtag:

- @KimKardashian e #KimKardashian
- @KrisJenner e #KrisJenner
- @KylieJenner e #KylieJenner
- @khloekardashian e #KhloeKardashian
- @kourtneykardash e #KourtneyKardashian
- @KendallJenner e #KendallJenner

Non sono stati inclusi altrimenti della famiglia come *Caitlyn* e *Rob* in quanto la loro presenza su Twitter è pressochè nulla. Il download è stato effettuato dal 13 al 14 Aprile e sono stati scaricati un totale di 20407 tweet.

5.1 Wordcloud

Per capire quali fossero i temi di dscussione più ricorrenti all'interno dei testi dei tweet scaricati si è deciso di ricorrere alla visualizzazione nota come **wordcloud** o **tagcloud**. In questa particolare visualizzazione le parole sono collocate all'interno di una "*nuvola*" e viene assegnata ad ogni parola una dimensione in relazione alla frequenza o alla significanza della stessa. Dovendo semplicemente indagare quali fossero le parole più ricorrenti all'interno dei testi è stata utilizzata la variante dove si utilizza la frequenza dei termini per stabilire la dimensione delle parole.

Come è possibile notare nella figura 3 la parola più ricorrente risulta essere "*KUWTK*", ovvero il loro programma televisivo, *Keeping Up With The Kardashian*. Esso infatti compare quasi nel 10% dei tweet scaricati. Effettuando una breve ricerca online è risultato che la forte presenza dell'hashtag **#KUWTK** fosse giustificata dal fatto che da li a breve sarebbe andata in onda la prima puntata della nuova stagione della trasmissione. A seguito di ciò si è deciso di scaricare i tweet durante la premiere della nuova stagione e durante un'ulteriore puntata nel mezzo della stagione, la settimana più precisamente, per poter effettuare un confronto.

Figura 3: *Wordcloud generata dai testi dei tweet scaricati*

6 Prima puntata

Lo streaming della prima puntata copre il periodo dal 27 marzo 2019 al 3 aprile 2019: Riportiamo nella tabella sottostante alcuni dati in merito ai tweet scaricati:

	1° Puntata
Numero di utenti	69626
Numero di tweet	103837
Numero di tweet distinti	25481
Numero di retweet totali	77135
Percentuale di retweet sul totale	74.28 %
Numero di retweet dei K	26976
Percentuale di retweet dei K sul totale	25.98 %
Percentuale di retweet dei K sui retweet	34.97 %
Numero di quote totali	48059
Numero di quote dei K	30172
Percentuale di quote dei K sul totale	29.06 %
Percentuale di quote dei K sui retweet	62.78 %

Tabella 1: *Dati sulla prima puntata*

È possibile notare come i retweet costituiscano quasi 75% dei tweet totali scaricati, essi sono infatti circa 78mila tweet su 103mila scaricati. Di questi solo il 35% sono retweet di tweet scritti diretattamente da uno dei componenti della famiglia. Differente la situazione considerando lo quote, i commenti ad un tweet di un *Kardashian* costituiscono infatti il 63% dei commenti totali scaricati. In figura 4 è possibile vedere il flusso dei tweet durante lo streaming. Nel grafico è possibile notare due picchi principali di attività in corrispondenza dell'inizio dello stream e della messa in onda della puntata. Questo picco iniziale di attività non prevista è dovuto al fatto che *Kim Kardashian* ha *spammato* l'uscita, a breve, della nuova stagione del programma.

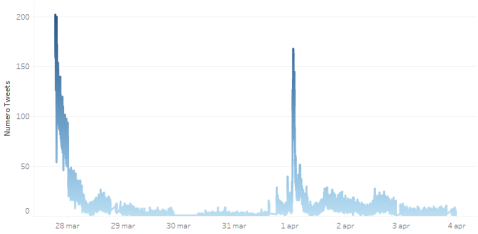

Figura 4: *Flusso dei tweet durante lo streaming*

Mentre in figura 7 è possibile vedere il flusso delle interazioni degli utenti con gli account della famiglia *Kardashian*. In questo grafico è possibile vedere come il primo picco di attività sia stato effettivamente guidato dalle interazioni con *Kim Kardashian* mentre la maggior parte delle interazioni durante la puntata è dovuta principalmente a *Khloe* e *Kris*.

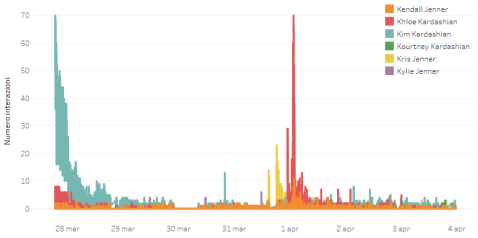

Figura 5: *Flusso delle interazioni con i Kardashian durante lo streaming*

6.1 Coverage

Come detto nella 4.3.2, per ogni tweet è stato calcolato un indice, a cui è stato dato il nome di **coverage**. Questo indice è calcolato come il seguente rapporto:

$$\frac{N_{u,i}}{N_{u,tot}}$$

Dove $N_{u,i}$ rappresenta il numero di utenti che interagiscono direttamente con il tweet i dell'utente u e $N_{u,tot}$ il numero totale di utenti che hanno direttamente interagito con u. Quando è pari a 1 indica che il tweet dell'utente ha raggiunto tutti gli altri utenti a cui è collegato, un valore superiore a 1 significa che il tweet ha raggiunto più utenti rispetto a quelli con cui aveva già interagito. Vengono riportati nella tabella seguente i 5 utenti con più interazioni e i rispettivi valori di *coverage media* calcolati.

User	Coverage media
KimKardashian	0.27
EpicMovieClips	0.30
KUWTK	0.01
BarstoolRia	0.37
KrisJenner	0.16

Tabella 2: *Top 5 utenti per interazioni e loro coverage*

Risulta immmediatamente molto basso il valore di coverage dell'account ufficiale del programma *KUWTK*. Indagando sulle motivazioni per un valore così basso è risultato che durante lo stream la maggioranza degli utenti non ha interagito direttamente con questo account ma bensì con quello di *Kim*, che risulta comunque avere una coverage bassa, intorno a 0,27. Naturalmente questi valori sono puramente "indicativi" in quanto la rete sottostante non è la rete di interazioni reale presente su Twitter.

7 Settima puntata

Lo streaming della settima puntata copre il periodo dal 25 maggio 2019 al 28 maggio 2019. Riportiamo nella tabella 3 alcuni dati in merito ai tweet scaricati:

	7° Puntata
Numero di utenti	22945
Numero di tweet	39153
Numero di tweet distinti	21626
Numero di retweet totali	19556
Percentuale di retweet sul totale	49.95 %
Numero di retweet dei K	3128
Percentuale di retweet dei K sul totale	7.99 %
Percentuale di retweet dei K sui retweet	16 %
Numero di quote totali	4581
Numero di quote dei K	492
Percentuale di quote dei K sul totale	1.26 %
Percentuale di quote dei K sui retweet	10.74 %

Tabella 3: *Dati sulla settima puntata*

Partendo dal numero complessivo di tweet scaricati, notevolmente inferiore rispetto al primo streaming, 39153 tweets contro poco più di 100mila, è possibile notare come anche il numero di retweet sia calato a quasi il 50% del totale. Di quest'ultimi, solamente 3128 rappresentano i retweet ufficiali delle Kardashians. La situazione per quanto concerne i commenti ai tweet è ulteriormente in ribasso. Infatti il numero di quotes è di 4581, circa il 12% rispetto al numero di tweet e dei quali solo 492, l'1,26%, sono stati pubblicati da un appartenente alla famiglia.

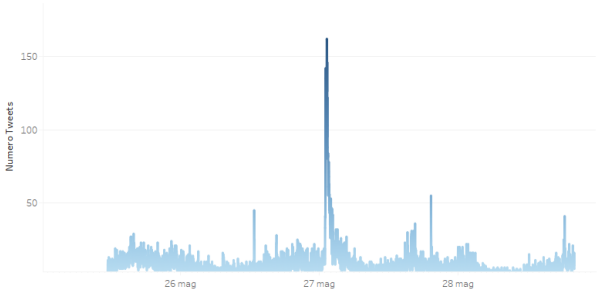

Figura 6: *Flusso dei tweet durante lo streaming*

In figura 6 è possibile vedere il flusso dei tweet durante lo streaming della 7° puntata. Nel grafico è possibile notare un picco principale di attività in corrispondenza del 27 maggio, giorno della messa in onda. Pertanto, nonostante l'attesa della puntata non sia stata fortemente sentita, la puntata di per sè ha generato molti *rumors* sul *social network*. Mentre in figura 7 è possibile vedere il flusso delle interazioni degli utenti con gli account della famiglia *Kardashian*. Come si può notare le tre sorelle più attive a livello di interazioni sono Kim, Khloè e Kendall.

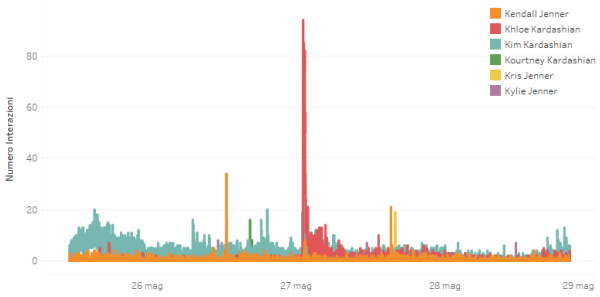

Figura 7: *Flusso delle interazioni con i Kardashian durante lo streaming*

7.1 Coverage

Vengono riportati nella tabella seguente i 5 utenti con più interazioni e i rispettivi valori di *coverage media* calcolati rispetto alla settimana puntata.

User	Coverage media
PhotosOfKanye	0.06
_KayyyBeeee	0.87
KUWTK	0.02
jerardlb	0.27
khloekardashian	0.08

Tabella 4: *Top 5 utenti per interazioni e loro coverage*

Anche in questo caso risulta basso il valore di coverage dell'account ufficiale *KUWTK*, è invece presente tra i nodi con più interazioni l'utente *_KayyyBeeee* che

attraverso i suoi tweet durante la puntata è riuscito a coprire in media l'87% dei nodi con cui era connesso.

8 Analisi e Confronto

8.1 Boxplot

Grazie ai Delta precedentemente creati si è scelto di utilizzare il grafico *Box plot* per analizzare i tempi di retweet e di commento ai tweet ufficiali di ciascuna delle Kardashians.

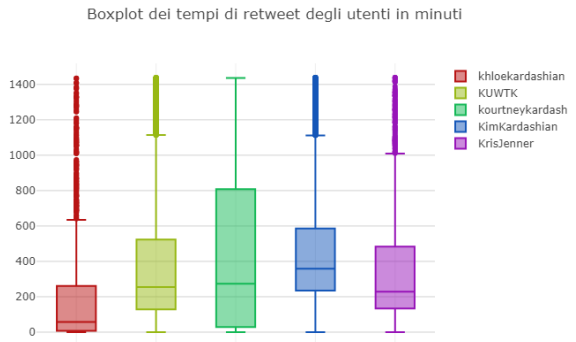

Figura 8: *Boxplot dei tempi di retweet degli utenti (1° puntata)*

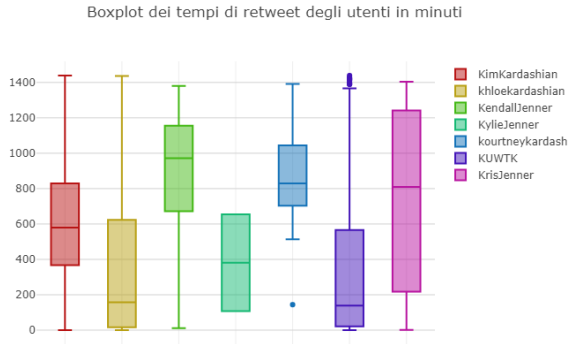

Figura 9: *Boxplot dei tempi di retweet degli utenti (7° puntata)*

La fig. 8 e la fig. 9 rappresentano i Box plot relativi al tempo di retweet rispettivamente della prima e della settima puntata. Si ricorda che in entrambi i grafici il tempo lungo l'asse Y è espresso in minuti. È interessante notare come Khloé Kardashian, che si ricorda essere la terza genita della coppia Kardashian-Jenner, abbia avuto il tempo medio di retweet più basso in assoluto nella prima puntata e sia stata preceduta solamente dalla pagina ufficiale KUWTK per pochi secondi nella settima puntata, nonostante sia quarta per numero di follower su Twitter rispetto all'intera famiglia. Il basso tempo di retweet dei post di Khloé Kardashian può dipendere fatto che è estremamente attiva nei confronti dei propri followers, sia attraverso retweet che attraverso risposte dirette. Dopo Khloè K. si trovano la madre Kris Jenner, in seconda posizione nella prima puntata e la sorellastra Kylie Jenner, in terza posizione nella settima puntata, la quale, assieme a Kourtney, sono le due ad avere avuto il maggior tempo minimo di retweet. E' importante sottolineare che Kylie Jenner, così come Kendall Jenner, appare solamente nel Box plot della settima puntata, poiché nel periodo di streaming dei dati relativi alla prima puntata non hanno scritto nessun tweet. Inoltre si sottolinea che i tweet di Kylie sono inerenti la sua linea di prodotti di bellezza e non il programma. Il profilo ufficiale KUWTK, come già detto primo per tempo medio di retweet nella settima puntata, si posiziona tra i primi anche nel corso della prima puntata. Questo dato è deducibile in quanto è il profilo maggiormente attivo sotto un punto di vista di "sponsorizzazione" del programma tramite il social network Twitter. Kim, nonostante sia la Kardashian con il maggior tempo medio di retweet, è stata *retweettata* per breve tempo, come indica il terzo quartile sotto i 600 minuti (10 ore) e il massimo tempo di retweet a circa 18 ore. Infatti Kourtney, con un tempo medio inferiore a Kim, presenta però il massimo picco di tempo di retweet della prima puntata, oltre le 23 ore. Ciò a indicare che ha fatto parlare di sé per lungo tempo a differenza delle altre Kardashians, probabilmente a causa, o si potrebbe dire per merito, del fatto che gran parte della puntata sia stata incentrata sull'annuncio della rottura con il suo fidanzato e il successivo "ritiro spirituale" con le sorelle ed amiche nella bella casa di Palm Spring.

Figura 10: *Boxplot dei tempi di risposta degli utenti (1° puntata)*

La figure 10 e 11 rappresentano invece i tempi di risposta ai tweet ufficiali delle Kardashians. In entrambe le puntate Khloé K. è la protagonista che ha ottenuto nel più breve tempo medio possibile quotes sotto i propri post. Come già specificato, ciò può dipendere

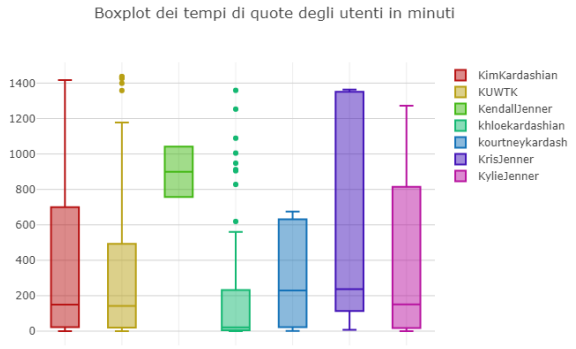

Figura 11: *Boxplot dei tempi di risposta degli utenti (7° puntata)*

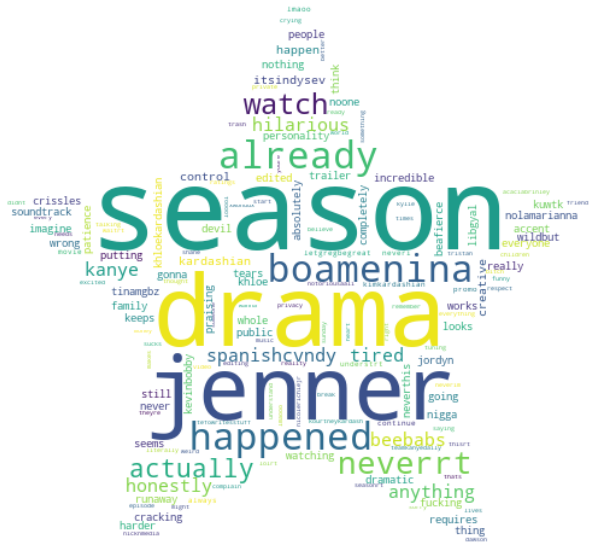

Figura 12: *Wordcloud generata dai testi dei tweet scaricati*

Figura 13: *Wordcloud generata dai testi dei tweet scaricati*

dall'alto tasso di risposta del personaggio nei confronti dei propri fan. Se non si considera il profilo ufficiale del programma, in seconda posizione per la prima puntata, come per il Box plot dei retweet, compare Kris Jenner. Nella settima puntata compaiono invece Kim e Kylie, le quali, come si può notare dal loro terzo quartile (700 e oltre 800), vengono commentate nel lungo periodo. Kris Jenner rimane indubbiamente la regina indiscussa della puntata, che viene commentata durante l'intero periodo di streaming, con un tempo medio molto basso. Questa reazione dei followers è probabilmente dovuta alla scelta della *momager* di portare nella casa di Palm Spring il nuovo fidanzato di soli 38 anni, due anni in meno della primogenita Kourtney. Caso "anomalo" è Kendall Jenner, che non ha utilizzato Twitter né durante la prima puntata, né durante la settima, ricevendo retweet e commenti a tweet di vecchia data solamente durante il periodo di streaming della settima puntata; d'altra parte è noto il fatto che prediliga altri *social networks*.

Le fig. 12 e 13 rappresentano le due Word Cloud relative ai commenti rispettivamente alla prima e settima puntata. Come si può notare, in entrambe le rappresentazioni appare "Kanye", che è stato notevolmente citato durante entrambe le puntate in quanto di forte presenza all'interno di questa stagione del programma. Infatti, oltre alla doppia intervista con la moglie Kim, Kanye appare numerose volte nel corso della prima puntata, parte della quale è proprio dedicata alla visita dello studio di Kanye a Chicago da parte di Khloé, Kim e la figlia North. Anche durante la settima puntata Kanye ha un ruolo da protagonista nella vicenda che lo vede coinvolto con il nuovo fidanzato della capostipite Kris. Tuttavia, le parole più semanticamente significative della Word Cloud inerente la prima puntata sono "drama" e "happened", accompagnate da diverse sentiment words, tra le quali: hilarious, honestly, laugh, creative. Questo insieme fa riferimento ai fatti principali accaduti durante l'apertura del reality, quasi tutti drammatici, ma dall'effetto divertente: la rottura tra Kourtney e il suo fidanzato, il dissing del rapper Rhymefest verso Kanye West e infine la falsa notizia detta da North a scuola circa la presunta gravidanza della mamma Kim. Per quanto riguarda la settima puntata invece, le due parole più evidenti sono proprio le due sorelle, Kim e Khloè, le quali sono state le prime a conoscere il nuovo fidanzato di mamma Kris e, inizialmente, a non condividerne la scelta, in quanto di soli 38 anni.

8.2 Sentiment analysis

Nonostante i drammi accaduti in entrambe le puntate, i fans del reality si sono espressi in maniera positiva

circa le due puntate, come emerge dai due istogrammi in fig. 14 e 15 che rappresentano la distribuzione della polarità ottenuta attraverso la Sentiment Analysis effettuata sui quotes ai tweet delle due puntate. Escludendo i commenti classificati come neutrali, si può notare che i commenti in entrambe le puntate sono complessivamente per lo più positivi, esattamente per il 41% nella prima e il 42% nella seconda rispetto al totale dei commenti analizzati.

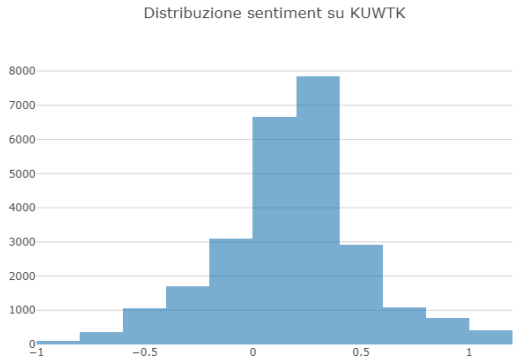

Figura 14: *Boxplot dei tempi di risposta degli utenti (1° puntata)*

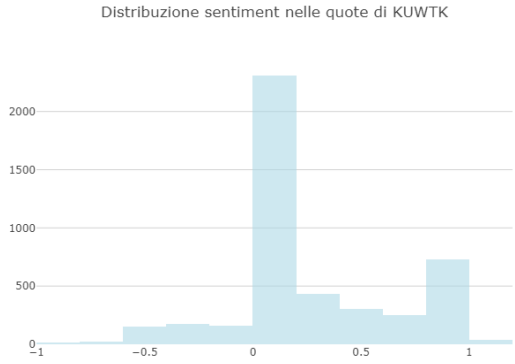

Figura 15: *Boxplot dei tempi di risposta degli utenti (7° puntata)*

9 Network Graph

L'ultimo *task* affrontato consiste nella creazione del grafo della prima e della settima puntata. Per quanto riguarda la prima viene generato un grafo orientato con le seguenti caratteristiche:

- Nodi: 69625
- Archi: 71578
- Grado medio: 2.0561
- Densità: 0.00002953

Per quanto riguarda invece la settima puntata viene generato un grafo orientato, di dimensioni inferiori rispetto a quello della prima, con le seguenti caratteristiche:

- Nodi: 22944
- Archi: 13194
- Grado medio: 1.1501
- Densità: 0.00005013

Da entrambi i grafi originali sono stati estratti due sottografi, formati dalla componente più grande del grafo di partenza, al fine di permettere un'analisi più accurata dato che le eccessive dimensioni dei primi lo impedivano. Estraendo la componente più grande ed eliminando i nodi isolati, ossia i nodi di grado pari a 0, si ottengono così i seguenti due sottografi, rispettivamente della prima (sx) e settima puntata (dx), visualizzati nelle figure 16 e 17.

- Nodi: 4798
- Archi: 11952
- Grado medio: 4.9821
- Densità: 0.00003513
- Diametro: 16
- Assortatività: -0.4216867

- Nodi: 9397
- Archi: 11079
- Grado medio: 2.3580
- Densità: 0.00025096
- Diametro: 15
- Assortatività: -0.31131265

Per quanto riguarda la *Community Detection* del grafo viene utilizzato l'algoritmo divisivo di Girvan-Newman, che esegue un processo di scomposizione gerarchica dove vengono rimossi gli archi con il maggior valore di edge betweenness. La rimozione progressiva degli archi comporterà una partizione della rete in due o più sottoreti molto dense di relazioni, formerà cioè le communities. Per la valutazione della qualità della divisione della rete in comunità operata dall'algoritmo viene osservata la modularità.
Prendendo in considerazione solamente i gruppi formati da almeno tre nodi, vengono identificate 174 communities nel grafo della prima puntata, di cui la maggiore formata da 24207 nodi, e 269 nella settima, di cui la maggiore formata da 1961 nodi.

Sono stati poi estratti i 10 nodi con grado maggiore, identificabili come nodi *Hubs*, ossia i nodi con il maggior numero di collegamenti all'interno della rete in esame, di gran lunga superiore al numero di collegamenti degli altri nodi nella medesima rete. Come visibile anche nelle rappresentazioni 16 e 17 i primi cinque nodi Hub della prima puntata sono:

- KimKardashian, grado: 26547
- EpicMovieClips, grado: 16978
- KUWTK, grado: 8832

- BarstoolRia, grado: 6162
- KrisJenner, grado: 2301

I nodi hubs della settima sono invece:

- PhotosOfKanye, grado: 2028
- _KayyyBeeee, grado: 1420
- KUWTK, grado: 832
- jerardlb, grado: 808
- khloekardashian, grado: 309

L'importanza di questi nodi è sottolineata dal fatto che essi rappresentano anche i nodi con il maggior numero di ponti, ossia archi la cui rimozione disconnette un'intera componente del grafo. Infatti per quanto riguada la prima puntata KimKardashian e EpicMovieClips presentano rispettivamente 24101 e 15664 ponti, nella settima invece si hanno PhotosOfKanye e _KayyyBeeee con 1880 e 1164 ponti. Come si può notare la pagina KUWTK rappresenta in entrambe le puntate uno tra i nodi più importanti di tutta la rete, probabilmente anche a causa delle azioni di sponsorizzazione effettuate dalla pagina ufficiale. Nella prima puntata tra le Kardashians le due di maggiore rilievo sono Kim e Kris, nella settima invece ha un grado maggiore Khloè. Si ricorda che quest'ultima è estremamente attiva nei confronti dei propri fans e followers.

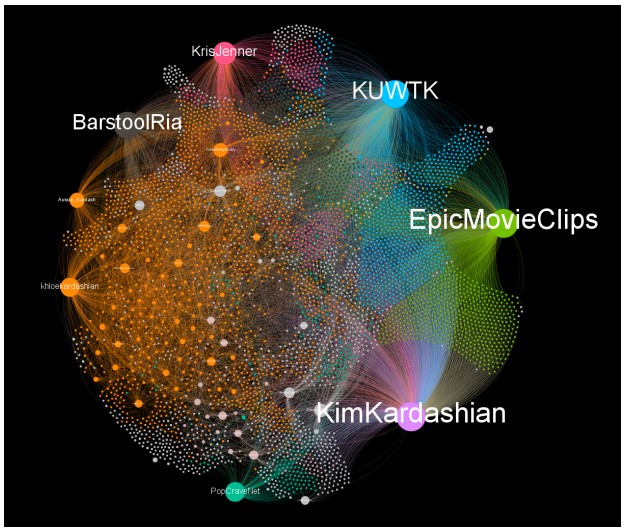

Figura 16: *Grafo prima puntata*

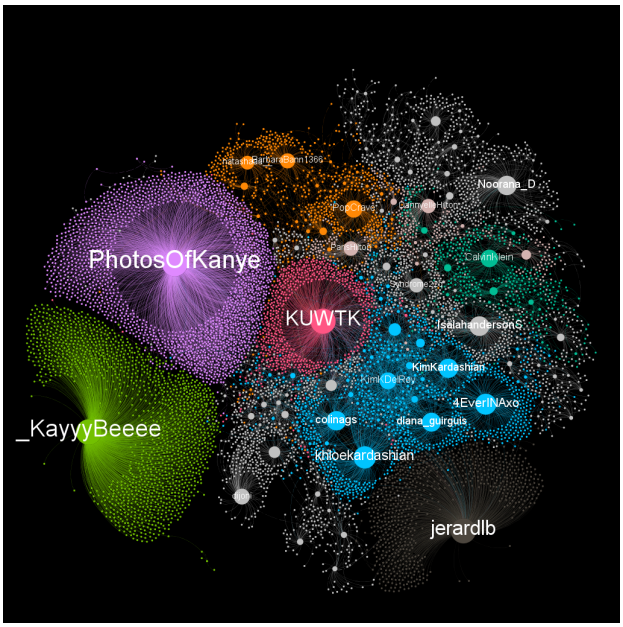

Figura 17: *Grafo settima puntata*

10 Conclusioni

Quando si pensa a un gruppo di persone che ha conquistato ogni aspetto dei *social media*, viene in mente solo un *clan*: le Kardashians (+ Jenners). Per conquistare la vetta dei *social media* hanno usato più tecniche, ma hanno principalmente abusato del concetto di *business intelligence* e *online visual marketing*. Tutte le sorelle sono considerate influenti, ma a parte questo sono tutte molto strategiche nei loro sforzi imprenditoriali. Hanno trovato un modo per esemplificare la loro proposta di valore e indirizzare il loro pubblico specifico. Hanno analizzato le persone che guardano principalmente i loro video, li seguono sui social media, acquistano i loro prodotti e li usano per creare modi innovativi per catturare il pubblico a lungo termine. Nel complesso, con il loro uso strategico dei *social media*, le sorelle Kardashian sono senza alcun dubbio le imprenditrici di maggior successo di questa generazione e "Keeping up with the Kardashian" rimane il fulcro dell'impero. Infatti, come si è analizzato nel corso di questo progetto, il loro potere mediatico è senza precedenti e di conseguenza la loro influenza che, come sopra analizzato, appare positiva tra gli utenti. Si ricorda comunque che la regola è sempre "purchè se ne parli" e sicuramente le Kardashians sono le regine del *gossip*, motivo del successo del loro *reality*.