Corso di Laurea in Business Informatics
Anno 2016/2017

Social Network Analysis

YouTube Network Analysis

Students Name
Armillotta Alessandro
De Franco Giuseppe
Di Sarli Leonardo
Pioli Giordano

Contents

1 Introduction

1.1 YouTube

YouTube è una piattaforma web, fondata il 14 febbraio 2005, che consente la condivisione e visualizzazione in rete di video (video sharing). Gli utenti possono anche votare e commentare i video. Sul sito è possibile vedere videoclip, trailer, video divertenti, notizie, slideshow e altro ancora. Nel novembre 2006 è stato acquistato dall'azienda statunitense Google per circa 1,7 miliardi di dollari. Attualmente secondo Alexa, è il secondo sito web più visitato al mondo, alle spalle solamente di Google.

Abbiamo deciso di studiare le collaborazioni tra canali YouTube (youtuber), estraendo dalla tab "Feature Channel" i canali che hanno delle collaborazioni con altri canali. Si è deciso di partire dal canale PewDiePie e da lì si è costruito il network da analizzare.

1.2 PewDiePie

Il nostro network è stato realizzato partendo dal canale PewDiePie, con ID UClHJZR3Gqxm24_Vd_AJ5Yw. PewDiePie, pseudonimo di **Felix Arvid Ulf Kjelberg**, è uno youtuber svedese. Il suo canale YouTube, creato nel 2010, ha raggiunto il milione di iscritti nel 2012. Dal 22 dicembre 2013 è quello con più iscritti in assoluto tra gli youtuber, così come dal 2014 è diventato il più visualizzato. Il 6 settembre 2015 PewDiePie ha raggiunto i 10 miliardi di visualizzazioni, divenendo il primo youtuber a raggiungere tale cifra; al mese di giugno 2017 il suo canale YouTube conta oltre 55 milioni di iscritti.

Stupiti da questi numeri, abbiamo deciso di effettuare la nostra analisi partendo da questo canale, cercando di analizzare la rete collegata ad esso.

2 Crawling

Per ottenere il network dei canali di YouTube, attraverso le API di **YouTube Data** ed il supporto del linguaggio **PHP**, è stato realizzato un piccolo crawler. Il crawler è reperibile su GitHub o al seguente link.
Lo script realizzato richiede l'inserimento dell'ID del canale, reperibile nella parte finale dell'URL YouTube (es. www.youtube.com/user/lHJZR3Gqxm24_Vd_AJ5Yw). Oltre all'ID, lo script richiede l'inserimento di un valore numerico che indica il **depth**, ossia la profondità fino a cui il crawler deve spingersi per realizzare la rete. Lo script passa l'ID che reitera per depth-volte la chiamata API. Alla profondita 0, viene scaricato l'ID passato, alla profondità 1 vengono scaricati gli amici del canale, alla profondità 2 gli amici degli amici e così via.
La chiamata alle API genera una risposta (file JSON) caratterizzata da diverse informazioni circa il canale (ID, Titolo, Lista Canali Correlati). La profondità che a noi interessa è la seguente:
item[0]->brandingSettings->channel->featuredChannelsUrls.

Figure 1: Livello "Fetured Channel"

A questo punto del file JSON, risiede l'informazione circa i "Fetured Channel", cioè i canali consigliati dal canale stesso. I Featured Channel sono canali consigliati dall'utente e non da YouTube in maniera automatica. Questa Tab di YouTube permette di mettere in risalto determinati canali i quali possono avere delle collaborazioni con il canale in questione.
Le principali funzioni PHP sono due:

1. **makeNetworkFromIds($depth)**: Questa funzione prende in input il depth passato nel form iniziale, scorre tutti gli ID YouTube passati ed effettua per ognuno una richiesta API. Ad ogni richiesta, viene presa la profondità "Feature Channel" del canale, vengono estratti i canali correlati e vengono assegnati come nodi ed archi in array . Se il depth è completo si ferma, altrimenti continua a reiterare la funzione fino a quando non arriva alla profondità dichiarata all'inizio.

2. **renderNetwork()**: La funzione renderNetwork si occupa di trasformare il dati ottenuti dal crawler in un file Gephi con la lista di archi e nodi.

3 Network Analysis

3.1 Software

Si è ritenuto opportuno confrontare le statistiche ottenute con la libreria NetworkX in python e il software Gephi, al fine di valutare l'eventuale eterogeneità dei risultati. La rete ottenuta è risultata connessa, logicamente, in modo diretto. Per problematiche legate alla libreria networkX si è deciso di procedere modificando il grafo da diretto a non-diretto. A livello grafico si è scelto di utilizzare la library matplotlib e lo stesso tool gephi.

3.2 Results

La rete realizzata risulta essere come la seguente immagine:

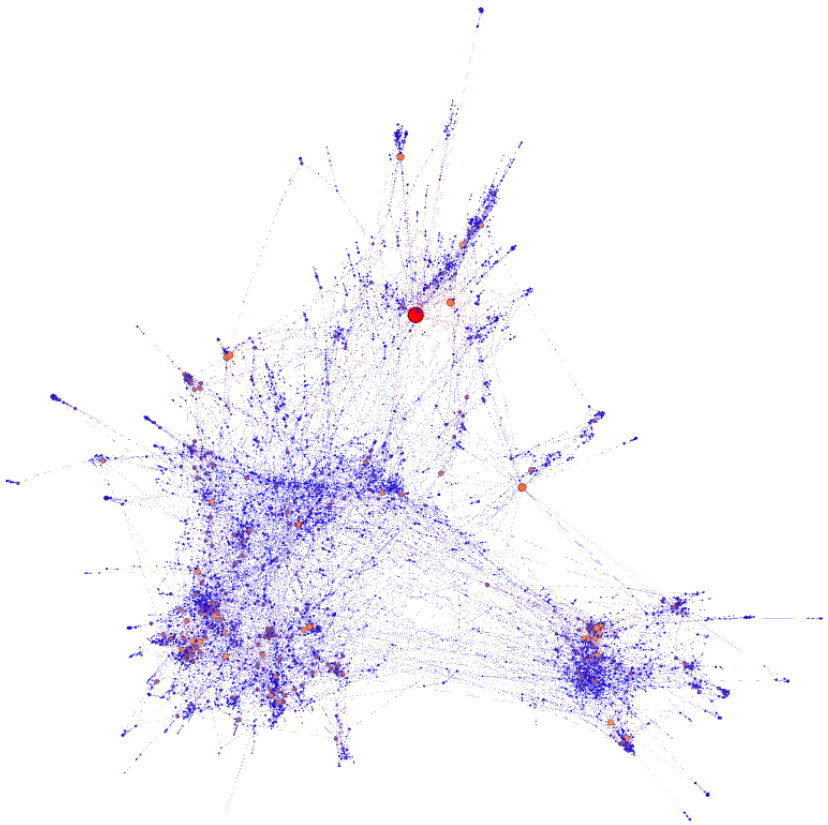

Figure 2: Grafo della rete dei canali YouTube

Il grafo in Figura 2 è stato ottenuto attraverso la libreria networkX ma è stato preso in considerazione come **indiretto**, formato da **6071 nodi** e **18815 archi**. La rete presenta una grande componente connessa (giant component) che implica l'assenza totale di nodi isolati (non connessi) e assenza di self-iteraction, visto che nessuno canale YouTube può inserire nei "Featured Channel" se stesso. Attraverso i plugin forniti da Cytoscape, Gephy ma sopratutto grazie a networkX, è stata effettuata un'approfondita analisi presentata di seguito.

Youtube Network	
Node	6071
Edges	18815
<k>	3.009
Components	1
Density	0.00102
Diameter	10
Avg. Clustering Coefficient	0.344
Avg. Shortest Path	5.75

Table 1: YouTube Network measures

3.2.1 Degree Distribution

L'analisi della Degree Distribution, ci permette di capire meglio la conformazione del network. In prima analisi possiamo dire che:

- il canale con il grado maggiore risulta essere "**Channel Frederator Network Members**", (degree 101);
- non esistono nodi isolati in quanto esiste solo una componente;

La rete, in particolare, ha un **Average Degree** pari a 3.009. Il grado medio ci permette di dire che in media un nodo è connesso a tre nodi. Ciò significa che in media un canale YouTube è collegato con almeno altri 3 canali . Questo risultato rispecchia la teoria **Small World**.
Guardando il grafico della distribuzione, notiamo che vi sono un alto numero di nodi con un grado basso e pochi HUB con grado elevato. I principali HUB rilevati nel network sono: Channel Frederator Network Members, Kin Community e ChannelFrederator. La distribuzione presenta un forte addensamento di nodi con basso grado ed una lunga coda caratterizzata da pochi hub con grado più alto. Questo andamento tipico delle reti reali, è definito **Power Law**.

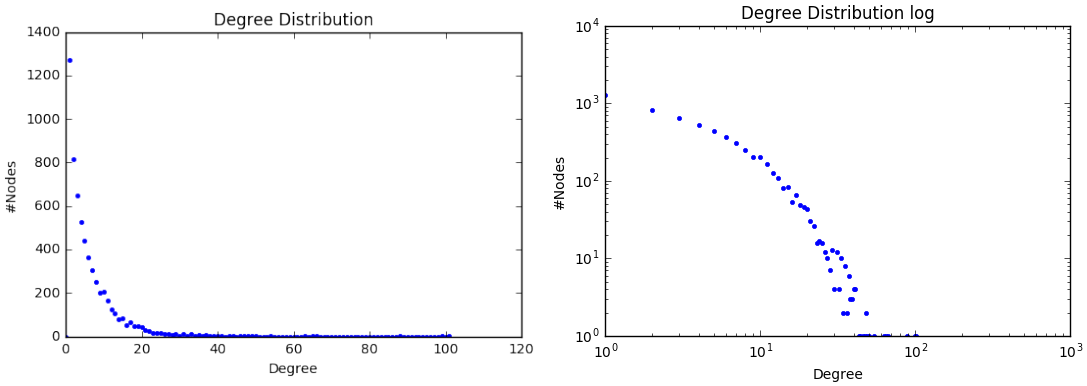

Figure 3: Degree Distribution

3.2.2 Paths and Distances

La distanza gioca un ruolo fondamentale nel determinare le iterazioni tra le componenti di un sistema. All'interno del network sono presenti 22940162 **Shortest Path**, ossia i cammini tra due nodi che hanno il minor numero di link. Il **Diameter**, ossia il massimo cammino minimo del network è pari a 10. L'**Average Shortest Path** è la media dei cammini minimi nel grafo, che è pari a 5.75.

3.2.3 Connected components

Proseguendo la nostra analisi, non abbiamo indivuduato all'interno del network più di una componente.

3.2.4 Clustering Coefficient, Density analysis

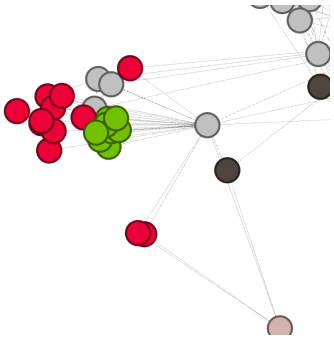

Figure 4: In rosso i nodi con alto CC

Il **Clustering Coefficient** misura la densità locale, ovvero la probabilità che due nodi adiacenti ad un nodo comune siano connessi tra loro. Il valore è pari a 0.344(34.4%), che rispecchia gli standard delle reti reali. La proprietà "small world" ha degli evidenti riscontri anche nel suddetto parametro, perché se gli archi sono inferiori rispetto al numero massimo di connessioni possibili tra i nodi e sono distribuiti in maniera sparsa, si viene a creare una fitta rete di connessioni con breve distanza tra i nodi. Per questa ragione i nodi tenderanno a chiudersi formando più triangoli. Di conseguenza il clustering coefficient risulterà essere più alto rispetto a quello di un random graph. E' un risultato in parte aspettato. Invece la **Density** del grafo ha un valore che si aggira intorno lo 0 (0.001). La densità di un grafo misura la probabilità che una qualsiasi coppia di nodi sia adiacente

3.2.5 Centrality Analysis

Le misure di **Centrality** aiutano ad identificare i più importanti nodi della rete, evidenziando cosa rende un nodo importante rispetto ad un altro nodo. Di seguito mostriamo i principali indicatori:

- **Degree Centrality**: La Degree centrality di un nodo è definita come il numero di archi incidenti ad esso. Misura la capacità immediata di un nodo di diffondere informazioni nella rete in base ai suoi vicini.

Degree Centrality	
Channel Frederator Network Members	0.0167
Kin Community	0.0163
ChannelFrederator	0.0145
SplayGaming	0.0109
This is Polaris	0.0107
CartoonHangover	0.0104

- **Closeness Centrality**: Questa misura esprime il valore in cui un nodo della rete è vicino a tutti gli altri, ovvero può essere rappresentato come l'inverso della somma degli Shortest Distance tra ogni nodo e ogni altro nodo nella rete.

Closeness Centrality	
This is Polaris	0.266
jacksepticeye	0.261
Cryaotic	0.260
PressHeartToContinue	0.259
GameGrumps	0.258
Markiplier	0.257

- **Betweennes Centrality**: Questa misura ci da informazioni circa l'importanza di un nodo. Più il valore della beetweennes è alto, più quel canale è importante in quanto ha più controllo nel network di YouTube. Il nodo o canale youtube con il più alto valore è "This is Polaris". Di seguito riportiamo i nodi più importanti del network:

Betweenness Centrality	
This is Polaris	0.0980
Rosanna Pansino	0.0733
Cryaotic	0.0511
LDShadowLady	0.0470
Fullscreen	0.0398
Monstercat	0.0366

- **Edge Betweennes**: Questa misura ci dà un'idea circa l'importanza di un determinato arco. Un arco con un valore elevato, rappresenta un **bridge** (ponte) tra due parti di una rete e la sua rimozione può influire sulla comunicazione tra diverse coppie di nodi. Di seguito elenchiamo i canali con il più alto valore:

Node A	**Node B**	**Value**
Kin Community	Rosanna Pansino	0.023
Cryaotic	Tasty	0.016
Markiplier	Rosanna Pansino	0.014
Cryaotic	Gamerbomb	0.012
Marzia	PewDiePie	0.0113
Mansl	SplayGaming	0.011

3.3 Network Comparison: Random Network e Barabasi Network

Nel paragrafo successivo, sono descritti i confronti con i modelli di ER e Barabasi.
Esistono due definizioni di reti random[1]:

1. **G(N, L) Model:** N nodi etichettati sono collegati con L link casuali. Questa è la definizione fornita da Erdòs e Rènyi;

2. **G(N, p) Model:** Ogni coppia di N nodi etichettati, è connessa con una probabilità p. Questo modello è introdotto da Gilbert ed approfondito da Erdòs–Rènyi;

Nella nostra analisi utilizzeremo il modello G(N, p), non solo per la facilità di calcolare le caratteristiche chiave della rete, ma anche perché in reti reali il numero di collegamenti raramente rimane fisso. Mentre nel modello 1 si assume che ci sia un numero fisso di nodi, nel Barabasi-Albert Network il numero di nodi cresce continuamente grazie all'aggiunta sistematica di nuovi nodi. Nel modello Erdòs Rènyi i nodi adiacenti al nuovo nodo vengono scelti casualmente; ciò non avviene nel modello 2, nel quale i nuovi nodi preferiscono connettersi a nodi con grado maggiore. Tale caratteristica è stata definita come Preferential Attachment. Dunque i nodi più grandi subiscono l'effetto del rich-gets-richer, divenendo così nodi hubs centrali per tutta la rete.

[1] http://barabasi.com/networksciencebook/chapter/3#random-network

3.3.1 Random Network Comparison (Erdòs–Rènyi)

Nella Network Science le reti Random vengono studiate per capire meglio le reti reali. Nella teoria il modello Random viene utilizzato per descrivere la casualità che determina la nascita di una rete reale.

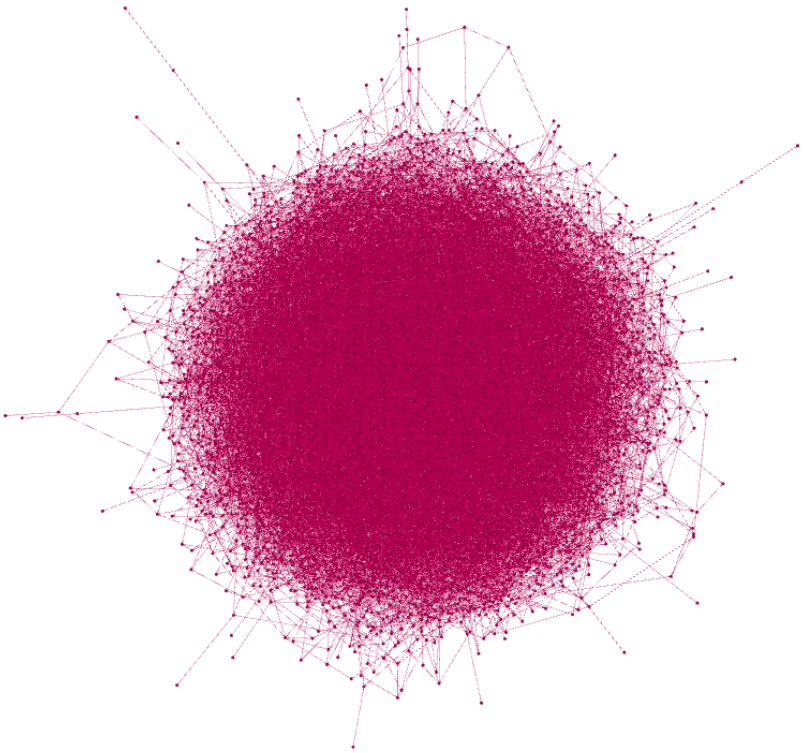

Figure 5: Random Network

Il primo confronto con il Network di YouTube si è effettuato con il modello **Random Erdos-Renyi**. Per questo modello, si parte da un numero N di nodi isolati, e successivamente, con una certa probabilità p, si collegano due coppie di nodi. Si ripete questo per ogni $\frac{(N-1)}{2}$ coppia di nodi. Una caratteristica fondamentale delle reti random è che la distribuzione del grado corrisponde ad una distribuzione di Poisson, con $p_k{=}e^{-<k>}\frac{<k>}{k!}$.
La rete random è stata generata tramite Gephi, con lo stesso numero di nodi della nostra rete e una probabilità p = 0.001012, pari alla densità del network iniziale. Di seguito vengono mostrati i risultati:

Index	**YouTube**	**Erdòs Rènyi**
Node	6071	6071
Edges	18815	18966
<k>	3.009	3.12
Components	1	14
Density	0.00102	0.00102
Diameter	10	$+\infty$
Avg. Clustering Coefficient	0.344	0.00097
Avg. Shortest Path	5.75	–

Table 2: Random Network vs. YouTube Network

Effettuando un'analisi delle misure su questa rete è stato possibile confrontare la rete Youtube con quella Random.
Il valore che rispetto alla rete originale è molto variato è il clustering coefficient che si avvicina molto allo 0. Ciò significa che dato un nodo, i link che connettono i propri vicini sono quasi 0. Infatti, differentemente dai grafi reali, nei grafi casuali raramente avviene la chiusura di tre nodi in triangoli di clustering. In un grafo casuale, se un utente x è amico degli utenti y e z, questa ipotesi non aumenta assolutamente la probabilità che y e z siano amici tra di loro, al contrario nei grafi reali questa considerazione può valere. In generale nei grafi reali il clustering coefficient risulterà essere sempre maggiore rispetto a quello presente nei grafi casuali. Nonostante il grado medio sia piuttosto simile, graficamente è possibile notare come la distribuzione del grado della rete YouTube non coincida con quella della rete random generata. Il grado medio della rete Random è pari a 3.012. Segnaliamo che il Diameter della rete random è pari a 15, ma questo è un risultato ottenuto da Gephi. Questo risultato si riferisce alla componente gigante, invece networkX ha giustamente calcolato il diametro della rete come $+\infty$; anche nel caso dello Shortest Path, non è possibile definirlo in quanto ci sono più componenti.
In linea con la teoria, questa rete rientra perfettamente nel regime **Supercritical** (<K> > 1). In questo regime abbiamo una grossa componente gigante con altre componenti più piccole, perfettamente in linea con il valore 14. Guardando

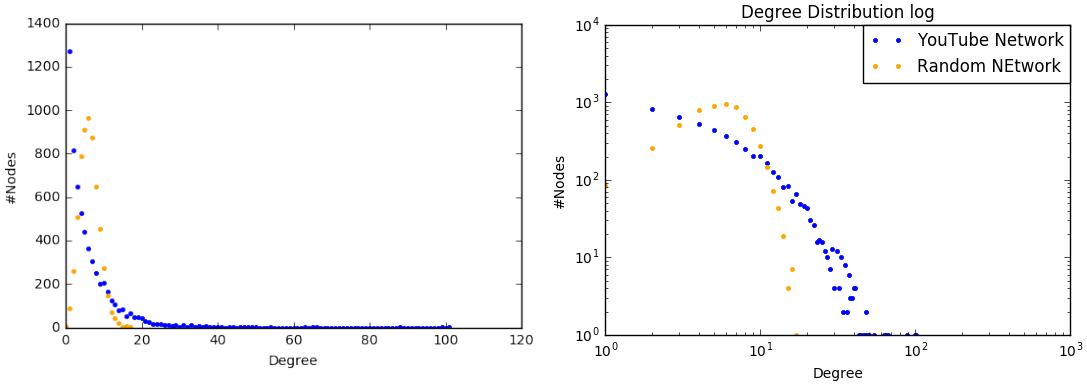

Figure 6: Random Network vs. YouTube Network

la distribuzione del grado possiamo notare che il Random Network ha l'esatta distribuzione di Poisson.

3.3.2 Barabasi-Albert Model (Preferntial Attachment Model)

Un secondo confronto è stato effettuato con il **Barabási-Albert Model**, modello di tipo scale-free.
Nello specifico, questo modello prevede la crescita della rete tramite l'aggiunta di un determinato numero di nodi. Questi nodi vengono aggiunti secondo il criterio del Preferential Attachment, secondo il quale i nodi tendono a collegarsi con i nodi maggiormente connessi. Una conseguenza diretta di ciò, è la nascita di HUB, che sono presenti nelle reti reali. Questo modello presenta una degree distribution che segue la power law.

Index	YouTube	Barabasi-Albert
Node	6071	6071
Edges	18815	18204
<k>	3.009	2.99
Components	1	1
Density	0.00102	0.001
Diameter	10	7
Avg. Clustering Coefficient	0.344	0.0087
Avg. Path Length	4.717	4.835

Table 3: YouTube Network vs. Barabasi-Albert

La Table 3 ci mostra i risultati ottenuti dal Barabasi-Albert model. In python si sono utilizzati i parametri N=6071 ed m=3. Come valore di m si è deciso di utilizzare il grado medio del network originario. Dai risultati notiamo che le componenti connesse non sono cambiate. Entrambi i modelli hanno una sola componente. Il valore che è realmente cambiato è il clustering coefficient che risulta essere molto più basso nel modello di Barabasi ma migliore rispetto al modello Erdòs Rènyi.

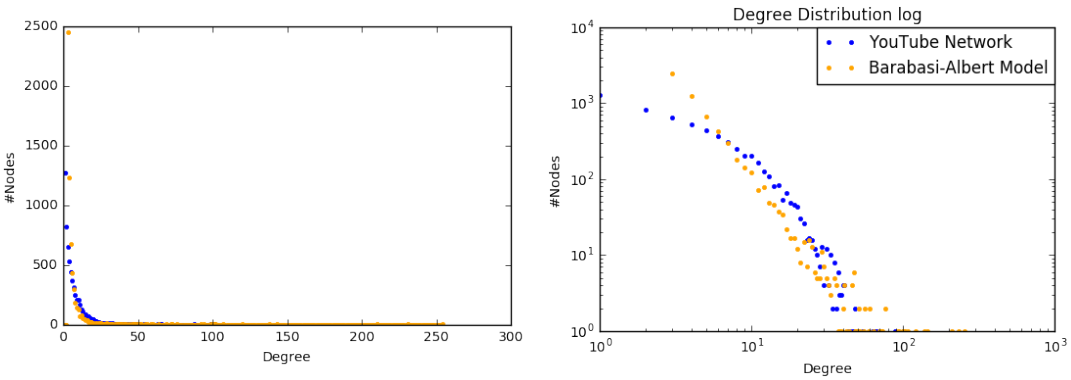

Figure 7: YouTube Network vs. Barabasi-Albert

4 Analytical Tasks

4.1 Community Discovery

L'obiettivo di questa fase è quello di individuare differenti communities esistenti nel network reale. Una community non è altro che un gruppo di nodi strettamente connessi tra loro, per vicinanza o similarità, rispetto a nodi appartenenti ad altri insiemi. Per far ciò sono stati applicati tre algoritmi differenti: **DEMON**, **k-Clique**, **CFinder**.
Per valutare le migliori partizioni abbiamo utilizzato quattro misure: Grado medio, Densità Interna, Conduttanza e Modularità. Il **grado medio** indica il grado medio di ogni nodo all'interno delle comunità.La **densità interna** indica quanti arichi esistono all'interno delle comunità rispetto al numero reale di archi. La **conduttanza** è la probabilità che un *random walker* esca dalla comunità. La **modularità** indica quanto sono connessi i nodi nella comunità rispetto ai restanti nodi della rete.

4.1.1 K-Clique

L'algoritmo è basato sul *percolation method* al fine di individuare le communities all'interno di una rete sociale, a partire da k-cliques. Una *clique* non è altro che un sotto-grafo completo di k nodi totalmente connessi. L'idea di base dell'algoritmo può essere riassunta da tale esempio: 2 k-cliques sono considerate adiacenti se condividono k-1 nodi e conseguentemente una community è definita come la massima unione di k-cliques che possono essere raggiunte da tutte le altre attraverso una serie di k-clique adiacenti.
Nel caso in esame si sono valutati diversi valori di k, appartenenti al range [2,20]. I risultati vengono riportati nel grafico sottostante.

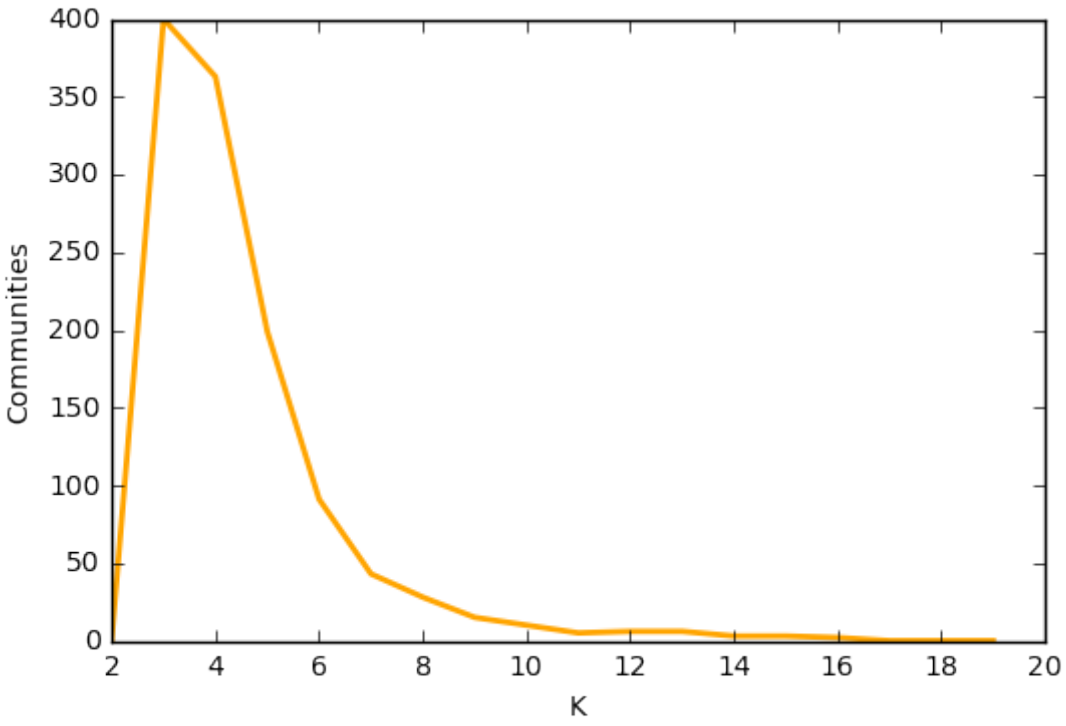

Figure 8: Communities for 2≤k≤20

Nel grafico è riportato il numero di comunità in base al valore di k. Per k=17,18,19,20 sono state ottenute 0 comunità. Per k=3 sono state trovate 400 comunità.
La tabella sottostante evidenzia i risultati dei vari valori k.

K	N°.Communities	Nodes	Edges	<k>	Internal Density	Conductance	Modularity
2	1	6071	18815	6.198	0.0002	0.0	0.0
3	400	4816	15506	2.919	0.213	0.561	0.2463
4	363	3080	10069	4.417	0.214	0.551	0.2301
5	199	1688	6200	5.76	0.222	0.528	0.1281
6	91	878	3824	7.25	0.225	0.461	0.0844
7	43	495	2554	8.967	0.225	0.373	0.0582
8	28	339	1885	10.00	0.232	0.396	0.0410
9	15	209	1328	11.68	0.232	0.306	0.0318
10	10	147	996	12.66	0.235	0.327	0.0261
11	5	92	726	15.67	0.226	0.113	0.0250
12	6	97	721	14.55	0.240	0.209	0.0189
13	6	93	676	14.22	0.245	0.235	0.0170
14	3	50	393	15.56	0.248	0.187	0.0127
15	3	50	393	15.56	0.248	0.187	0.0127
16	2	35	288	16.34	0.247	0.040	0.0125

Table 4: Risultati K-clique

Come si può notare dalla tabella, man mano che aumenta il valore K , il grado medio tende ad aumentare, al contrario degli altri valori. Con valore k=7 vengono individuate 43 comunità con 495 canali che ne fanno parte, circa il 7% dei canali della rete reale. Con valore k=16 vengono individuate 2 comunità con 35 canali che ne fanno parte, circa il 5% dei canali della rete reale. In riferimento a k=3, vengono individuate 400 comunità contenenti 4816 canali pari a circa il 70% dei canali della rete reale iniziale. Con questo valore vengono individuati canali in overlapping. Il canale *Fullscreen* fa parte di 28 comunità, *TheASHfire06 - ASH s PSP Games!* di 10, d*ietblond* condivide 6 comunità, coì come *iJustine* e *Yohhamgambal*.
Con k=5 invece, vengono individuate 199 comunità con 1688 canali, circa il 23% della rete reale iniziale. All'interno del sotto grafo, sono stati individuati canali in overlapping, ossia che fanno parte di più comunità. Nello specifico *Girbeagly* fa parte di 21 comunità, *clothesencounters* condivide 15 comunità, *Claire Marshall* fa parte di 10 comunità ed infine *Markiplier*, *The Game Chasers*, *TwistedGrimTV* fanno parte di 6 comunità. Dall'analisi si evidenzia che i più grandi canali YouTube in overlapping, sono canali di "Gamer" e "Gameplay". Molto spesso questi canali sono linkati da canali di genere diverso perchè sponsorizzati da grossi marchi online che trattano prodotti differenti.
Osservando le statistiche possiamo affermare che la migliore partizione secondo il valore di conduttanza è k=16 , mentre per la modularità la partizione migliore è quella ottenuta con k=3.

4.1.2 DEMON

Successivamente per analizzare le comunità abbiamo utilizzato l'algoritmo DEMON [2] che ha la caratteristica di individuare le comunità tramite l'utilizzo dell'ego network. Questo algoritmo sfrutta l'ego network per ridurre la complessità del network in modo da poter applicare senza costi enormi la Label Propagation. Inizialmente viene selezionato un nodo della rete su cui viene estratto l'ego network, successivamente dopo aver rimosso il nodo di partenza scelto, viene applicata la Label Propagation sulla ego network ottenuta precedentemente. A questo punto il nodo viene inserito nuovamente nella rete e vengono individuate le comunità. Le comunità vengono inserite in un insieme e vengono unite tutte quelle comunità che risultano simili per un determinato parametro o valore di soglia ϵ. Nell'algoritmo in questione il parametro ϵ indica il la % di nodi condivisi tra le comunità. Utilizzando python e sfruttando la libreria Demon abbiamo testato l'algoritmo nella rete YouTube, ottenendo i seguenti risultati:

ϵ	N°.Communities	Avg.Degree	Internal Density	Conductance	Modularity
0.25	228	5.175	0.103	0.391	0.372
0.40	384	4.601	0.132	0.476	0.360
0.60	575	4.412	0.135	0.537	0.315
0.90	1662	4.319	0.151	0.640	0.259
1	1805	4.231	0.168	0.651	0.265

Table 5: Risultati DEMON

Dalla tabella si può notare come all'aumentare di ϵ, diminuisce il grado medio e aumentano le comunità trovate. Notiamo inoltre come le misure di partition quality aumentino con l'aumentare di ϵ. Osservando le statistiche possiamo affermare che in assoluto la migliore partizione si ottiene per ϵ=0.25 , poichè ha la minor conduttanza e l amggior modularità.

4.1.3 CFinder

Il software CFinder utilizza l'algoritmo chiamato **Clique Percolation Method** o *CFinder*, il quale definisce le comunità come l'unione di clique sovrapposte, cioè che condividono dei nodi.
In particolare, ogni clique è composta da un k numero di nodi, e ciascuna comunità è definita come l'unione delle dette k-clique con una serie di k-clique adiacenti, dove con adiacenti intendiamo quelle che condividono k-1 nodi. Una comunità k-clique, secondo questo algoritmo, è quindi il più grande grafo che si ottiene con l'unione di clique adiacenti. Tramite CFinder si è analizzata la struttura e l'overlap delle diverse clique. I risultati ottenuti sono stati analizzati attraverso networkX.

[2] Coscia, Michele; Rossetti, Giulio; Giannotti, Fosca; Pedreschi, Dino "Uncovering Hierarchical and Overlapping Communities with a Local-First Approach" ACM Transactions on Knowledge Discovery from Data (TKDD), 9 (1), 2014.
Coscia, Michele; Rossetti, Giulio; Giannotti, Fosca; Pedreschi, Dino "DEMON: a Local-First Discovery Method for Overlapping Communities" SIGKDD international conference on knowledge discovery and data mining, pp. 615-623, IEEE ACM, 2012, ISBN: 978-1-4503-1462-6.

K	N°.Communities	Nodes	Edges	<k>	Internal Density	Conductance	Modularity
3	398	4797	15448	2.910	0.213	0.562	0.257
4	363	3071	10021	4.405	0.214	0.554	0.229
5	199	1685	6175	5.750	0.222	0.532	0.126
6	92	880	3812	7.195	0.226	0.469	0.0817
7	43	492	2532	8.906	0.225	0.383	0.0568
8	28	339	1885	10.003	0.232	0.396	0.0402
9	15	209	1328	11.689	0.232	0.306	0.0324
10	10	147	996	9.60	0.235	0.327	0.0268
11	5	92	726	15.678	0.226	0.113	0.0263
12	6	97	721	14.557	0.240	0.209	0.0205
13	6	93	676	14.220	0.245	0.235	0.0177
14	3	50	393	15.561	0.248	0.187	0.0096
15	3	50	393	15.561	0.248	0.187	0.0096
16	2	35	288	16.342	0.247	0.040	0.0094

Table 6: Risultati CFinder

Per valutare le comunità, attraverso Python, sono stati calcolati : grado medio, densità interna, conduttanza e modularità. Il valori di archi, nodi e numero comunità, sono stati riportati per descrivere meglio i risultati e capirne le differenze. Attraverso il tool grafico CFinder, si è riusciti ad indentificare i canali più significativi all'interno del network di YouTube.

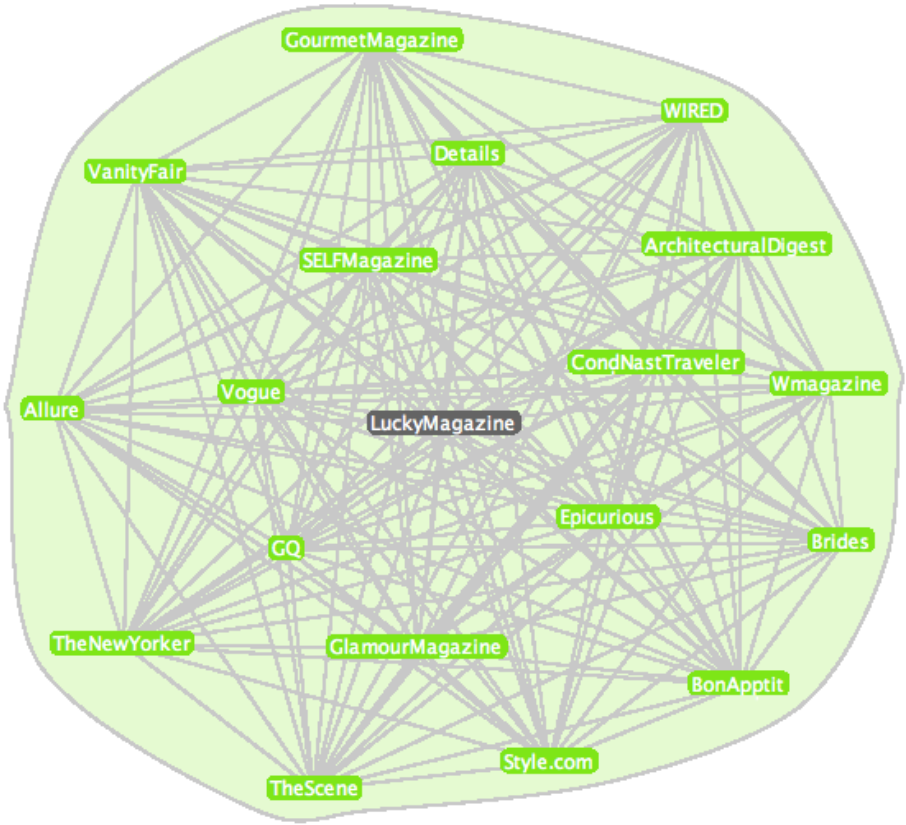

Figure 9: Canale Lucky Magazine

Il canale **Lucky Magazine** è il nodo che è più presente, ad ogni grado di k=3...16. Con k=3 il canale è condiviso da due comunità. In totale fa parte di 15 comunità. Lucky era un magazine fashion & lifestyle. Dal 2015 non è più in vendità ed i canali social non sono più attivi. Lucky era gestita da Advance Publications , che gestisce grandi magazine come Wired, GQ, Vanity Fair, Vogue. Si può affermare che questa rete compare a tutti i gradi di k perchè strettamente connessa, infatti tutti i canali gestiti da Advance Publications sono linkati tra di loro.

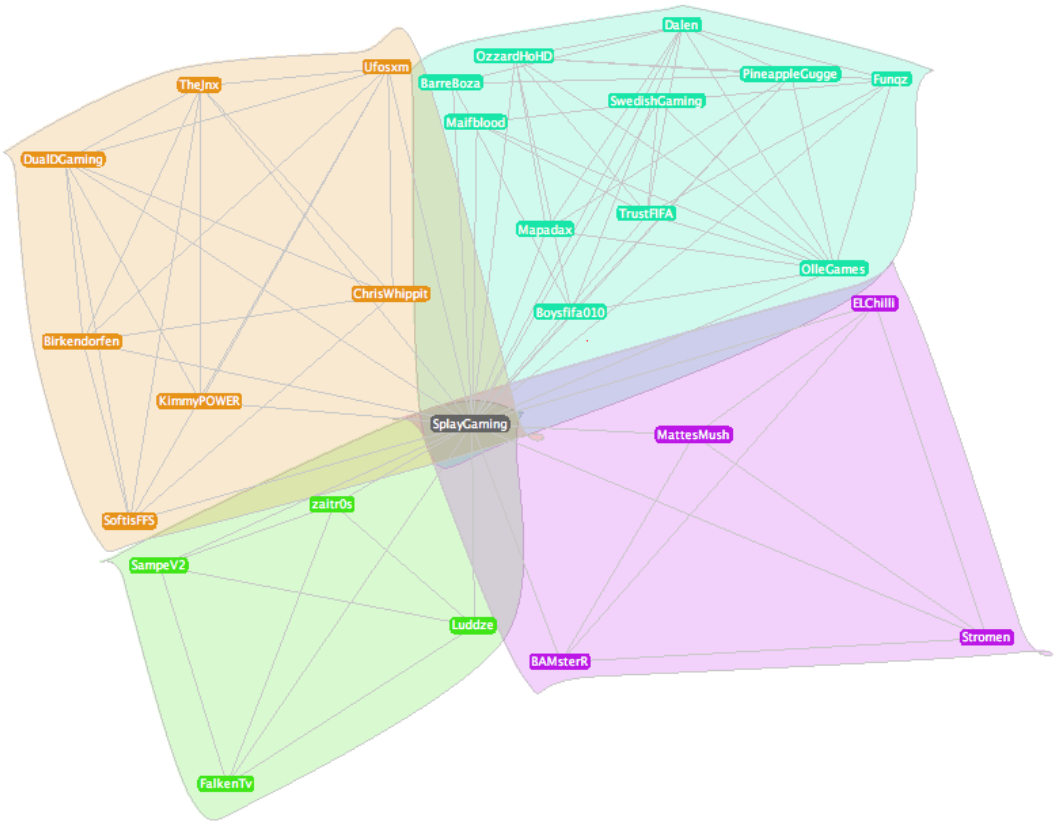

Figure 10: Comunità Splay Gaming

Il canale **Splay Gaming** fa parte di 15 comunità e come si evince dal nome, sono tutte comunità che riguardano videogame. In *Figure 10* sono riportate le comunià per k=5. Splay Gaming racchiude al proprio interno tutti gli youtuber di videogiochi. Ad esempio nella comunità in alto a destra sono racchiusi tutti gli youtuber di FIFA. In arancione , in alto a sinistra, sono racchiusi tutti gli youtuber di Minecraft. In base ai risultati ottenuti , possiamo affermare che in base alla conduttanza, la partizione migliore in assoluto riuslta essere k=16. Invece in base alla modularità la partizione migliore risulta essere k=3.

4.2 Tie Strength

4.2.1 Classificazione archi

Come già spiegato in precedenza, la natura della rete è quella di essere diretta: infatti uno YouTuber potrebbe includere un canale nella propria lista "Featured Channel", ma non essere a sua volta elencato nella lista di quest'ultimo. Per questo motivo un primo metodo per stabilire la forza dei legami può essere quello di etichettare come deboli i legami unilaterali, ovvero quelli in cui solamente uno dei due canali include l'altro nella propria lista, e come forti i legami bilaterali, ovvero quelli in cui la citazione è reciproca. Questo procedimento restituisce la seguente suddivisione:

Legami Deboli	Legami Forti
11657	14298

Table 7: Classificazione dei legami

Essendo impossibilitati a classificare i legami secondo una frequenza di interazione o una durata (ad esempio in una rete di telefonate), in quanto i link non sono pesati, come secondo metodo per sancire la forza dei legami è stato scelto quello di contare per ogni arco i nodi "amici" in comune per la coppia di nodi, ovvero il numero di triangoli "chiusi" per quel link nel grafo della rete.
Common Neighbors:

$$(u, v) = |\Gamma(u) \cap \Gamma(v)|$$

Dopodiché si è proceduto ad inserire i valori in una lista ordinata per poter calcolare una grandezza discriminante. In questo caso sono state considerate due alternative:

- il cinquantesimo percentile della lista, per cui ogni coppia di nodi formante un link, avente numero di "amici" in comune maggiore di questo valore, è considerata come legame forte e il resto delle coppie come legami deboli;
- la media dei valori della lista, per cui ogni coppia di nodi formante un link, avente un numero di "amici" in comune maggiore di questo valore, è considerata come legame forte e il resto delle coppie come legami deboli.

Le discriminanti ottenute sono le seguenti:

50° Percentile	2.0
Media	3.0

Table 8: Valori statistici con "Common Neighbors"

Una misura più raffinata per stabilire la forza dei legami è data dall'indice Jaccard, che normalizza il numero di nodi "amici" in comune per quanto i due nodi stessi sono "sociali".
Jaccard Index:

$$TS(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$$

Per definire una discriminante si è proceduto in maniera analoga rispetto al punto precedente e si sono ottenuti i seguenti valori:

50° Percentile	0.1
Media	0.1611

Table 9: Valori statistici con Jaccard

L'ultimo metodo sfruttato per classificare la forza dei legami è stato quello della edge betweenness centrality, ovvero il numero di cammini minimi che passano attraverso l'arco considerato, calcolato attraverso la funzione apposita della libreria NetworkX. Un arco con una edge betweenness centrality elevata rappresenta una connessione "ponte" tra due componenti di una rete, la cui rimozione può influenzare la comunicazione tra molte coppie di nodi attraverso i percorsi minimi , in quanto i legami deboli sono le scorciatoie che se eliminate disgregherebbero la rete. Per cui gli archi con alti valori di edge betweenness centrality vengono classificati come legami deboli e, viceversa, quelli caratterizzati da bassi valori di edge betwenness centrality sono legami forti.

4.2.2 Analisi dell'impatto dei legami nella rete

Per effettuare l'analisi dell'impatto dei diversi legami della rete, è stato deciso di prendere in considerazione le classificazioni ottenute reputando la rete come indiretta, in quanto fino a questo momento gli studi sulla rete sono stati effettuati in quest'ottica. Il metodo di classificazione prescelto è stato la edge betweenness centrality. Si sono inseriti i valori in due liste distinte, ordinate una in modo crescente e l'altra in modo decrescente. Successivamente si sono rimossi gli archi della rete, prima selezionando i nodi dalla prima lista, in ordine crescente di EB rimuovendo perciò prima i legami deboli, e successivamente selezionando i nodi dalla seconda lista, rimuovendo per primi i legami forti.

I dati estratti dalla rimozione dei link inclusi nella prima lista sono i seguenti:

% Archi Rimossi	N° CC	Dim CM / Dim Rete	Dist Media CM
0	1	1.0	5.7501
10	1	1.0	5.7505
20	4	0.9995	5.7551
30	10	0.9985	5.7747
40	50	0.9919	5.8221
50	306	0.9494	5.8919
60	1085	0.8194	5.9089
70	2173	0.6379	5.9250
80	3958	0.3472	5.0990
90	4843	0.2006	5.0264
100	6071	0.0	0

Table 10: CM: Componente Maggiore; CC: Componenti Connesse

Mentre quelli estratti dalla rimozione dei link inclusi nella seconda lista (decrescente) sono i seguenti:

% Archi Rimossi	N° CC	Dim CM / Dim Rete	Dist Media CM
0	1	1.0	5.7501
10	76	0.9530	7.7609
20	161	0.9077	9.6917
30	1561	0.6880	9.4078
40	1738	0.6094	11.2191
50	2043	0.4911	15.9034
60	2611	0.2478	17.6887
70	3248	0.0736	16.1069
80	4030	0.0186	6.06
90	5000	0.0047	3.2512
100	6071	0.0	0

Table 11: CM: Componente Maggiore; CC: Componenti Connesse

Come possiamo dedurre dal grafico seguente e come ci si può aspettare dalla distribuzione power law della nostra rete, citando il concetto espresso nel celebre articolo di Mark Granovetter "The Strength of Weak Ties", si può affermare che i legami più deboli possono avere un grande influsso sulla connessione della rete. Infatti rimuovendo per primi i link con più alta EB la dimensione del componente maggiore rispetto al totale dei nodi, decresce molto più rapidamente.

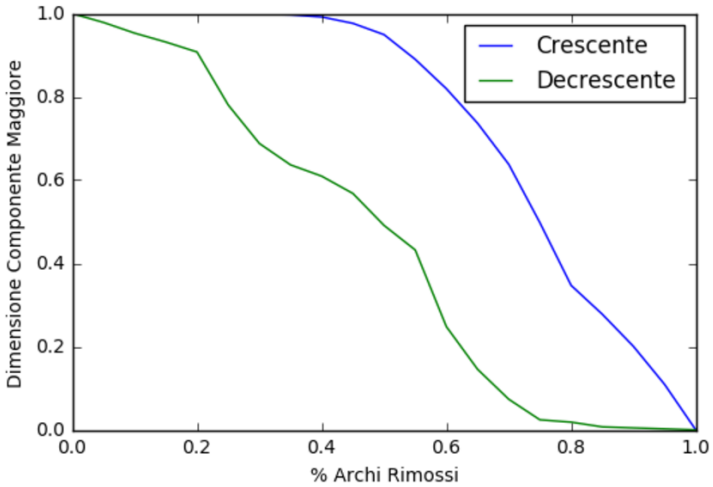

Figure 11: rapporto tra la dimensione del componente maggiore e la dimensione della rete all'aumentare degli archi rimossi

Quelli che Granovetter definisce "bridge", non sono solo ponti verso un altro nodo, ma anche ponti verso componenti lontane della rete, che sarebbero altrimenti del tutto estranee. Rimuovere dalla rete un legame forte non avrebbe quasi nessun effetto sui cammini minimi, in quanto pur sembrando indispensabili a tenere insieme la rete, non lo sono per ciò che riguarda i gradi di separazione.
Di seguito si possono apprezzare le differenze nella rimozione di link appartenenti alle due diverse liste, rispetto al numero di componenti connesse della rete e alla distanza media della componente maggiore.

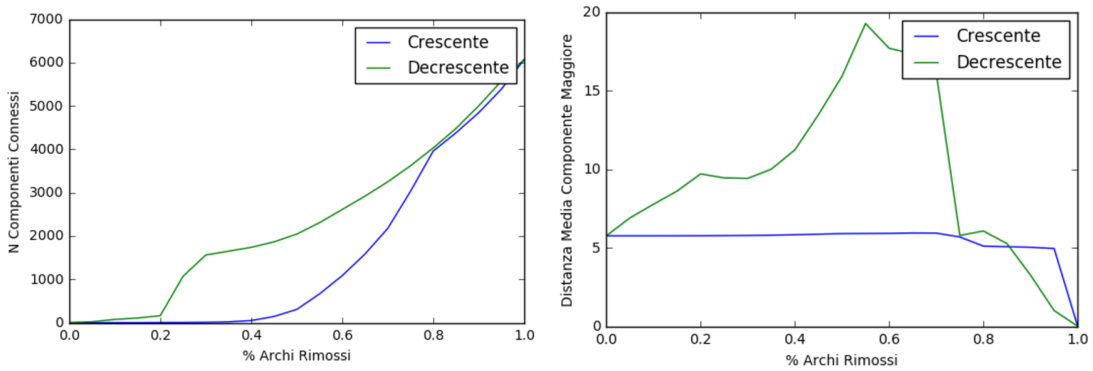

Figure 12: Cambiamenti in rete all'aumentare degli archi rimossi con BC

Per quanto riguarda l'utilizzo della edge betweenness centrality per ordinare le liste bisogna considerare che maggiore è il valore di un link, più è debole il legame. Viceversa utilizzando l'indice Jaccard, minore è il coefficiente, più è debole il legame. Per cui rimuovendo per primi i link presenti nella lista ordinata in modo crescente sono rimossi per primi i legami deboli e si ha una frammentazione più immediata della rete rispetto alla rimozione dei link in ordine decrescente.
Utilizzando questo metodo, specie nella fase iniziale, la rete si disgrega ancora più rapidamente rispetto al caso precedente. Già dopo poco meno di 5000 archi rimossi ci sono ben 2000 componenti circa (su un totale di 6071 nodi), partendo dai legami forti si ottiene un valore simile eliminando circa tre quarti degli archi.

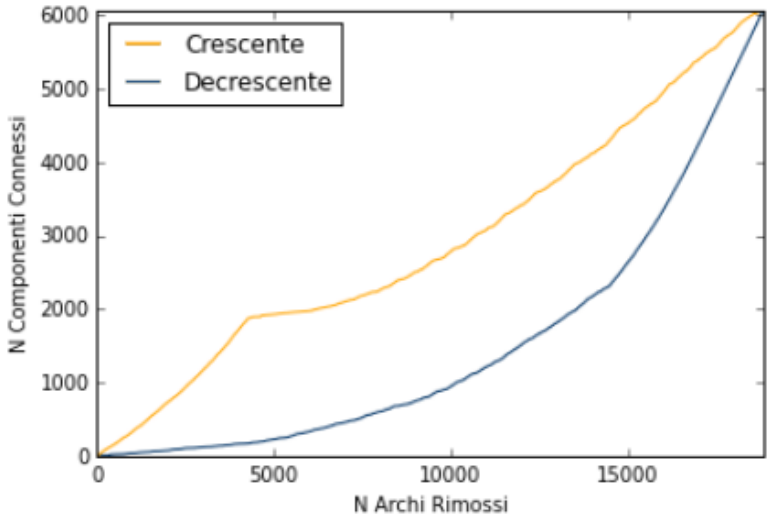

Figure 13: Effetti della rimozione di archi con Jaccard sul numero di componenti connesse

4.3 Spreading

La diffusione delle epidemie nelle reti complesse ultimamente è sempre più oggetto di studio e di ricerca. A seconda della natura della malattia e della rete esistono tipologie di epidemie differenti, come ad esempio:

- Epidemie in ambito Biologico: basate su malattie che si propagano via aerea (influenza, SARS, tubercolosi), malattie trasmesse per contatto (parassiti, peste), malattie trasmesse attraverso fluidi corporei (ebola, HIV), malattie infettive, etc...
- Epidemie in ambito Digitale: si tratta principalmente di virus informatici, programmi che si auto-riproducono, che si propagano copiandosi da un computer all'altro. La diffusione ricorda molto quello degli agenti patogeni (epidemia biologica), ma si differenziano soprattutto per ciò che sta alla base del virus.
- Epidemie in ambito Sociale: il concetto di epidemia nelle reti sociali assume i connotati di diffusione e assimilazione di conoscenze, innovazioni, comportamenti, etc...

La nostra analisi si concentrerà sullo studio della diffusione delle epidemie, attraverso quattro modelli (SIR,SIS,SI,threshold),inizialmente applicati sulla rete YouTube precedentemente scaricata, poi su una rete random e successivamente su una rete BA per poter effettuare così un confronto.

Verificheremo se il numero dei nodi infetti sarà destinato ad aumentare o calare durante l'epidemia, partendo dal presupposto che essa sia influenzata da due fattori:

- Struttura della rete.
- Probabilità che un nodo venga contagiato.

4.3.1 SI (Susceptible - Infected) Model

Il primo modello utilizzato è il SI caratterizzato da due stati, Suscettibile (S) e Infetto (I), e da un tasso di infezione . In questo modello il tasso di epidemia ha una crescita esponenziale dovuta al fatto che tutti i nodi della rete sono considerati suscettibili, per cui in questo modello tutti i nodi saranno destinati ad infettarsi. Proseguendo con l'analisi possiamo calcolare il tempo caratteristico (una stima dell'ordine di grandezza su scala temporale di reazione di un sistema) tramite la seguente formula:

$$\Gamma = \frac{1}{\beta < k >}$$

Da essa si capisce che più il grafo è connesso e il grado medio è elevato, più l'epidemia si diffonderà velocemente. Abbiamo testato il modello su python sfruttando la libreria NDLIB, ottenendo i seguenti risultati:

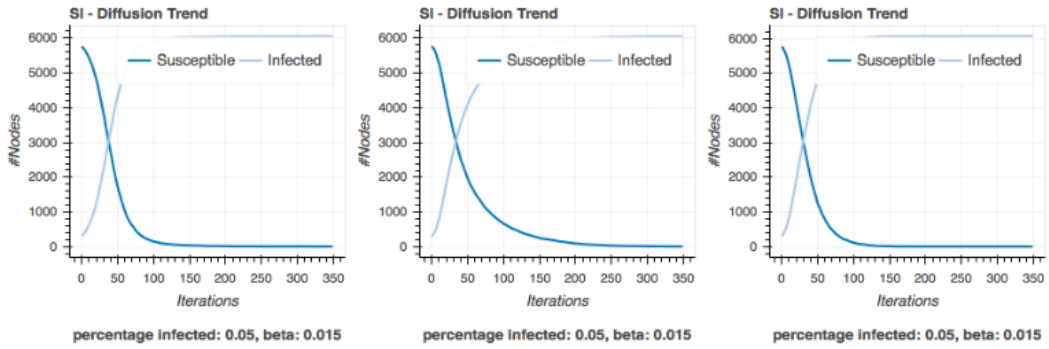

Figure 14: (A)-Rete Random, (B)-Rete YouTube, (C)- Rete BA.

Osservando i grafici si nota come il numero di nodi infetti abbia una crescita esponenziale e come per il grafico random servano più iterazioni per la la diffusione del virus.

4.3.2 SIS (Susceptible - Infected - Susceptible) Model

Il secondo modello utilizzato è il SIS caratterizzato dagli stati: Suscettibile e Infetto (come il SI), dove però un nodo infetto può tornare suscettibile. Oltre al tasso di virulenza β , abbiamo il tasso di guarigione μ. Con questi due valori possiamo calcolare il tasso di riproduzione del virus o Basic reproductive number (λ) , una variabile che ci dice se il virus è destinato ad "esplodere" o a sparire in base ai vari paramentri di infezione e guarigione.

$$\lambda = \frac{\beta}{\mu}$$

Ogni rete ha una propria soglia , che se maggiore di λ ci fornisce l'informazione che il virus in questa rete tenderà a scomparire, nel caso contrario tenderà ad esplodere generando un'epidemia. Le soglie per le varie reti possono essere calcolate come:

- Random Network : $\lambda(ra) = \frac{1}{<k>+1}$
- Scale Free : $\lambda(sf) = \frac{<k>}{<k^2>}$

La nostra analisi è proseguita testando il modello SIS nelle solite tre reti , prima per un valore di λ che provocasse un'epidemia in queste (Figura 15), successivamente per un valore che essendo più piccolo della soglia delle reti portasse l'epidemia a sparire nel tempo (Figura 16). Le soglie delle tre reti sono state calcolate come:

- Random Network : $\lambda(ra) = \frac{1}{<k>+1} = 0{,}245$
- YouTube: $\lambda(sf) = \frac{<k>}{<k^2>} = 0{,}321$
- Rete BA : $\lambda(sf) = \frac{<k>}{<k^2>} = 0{,}334$

Nella figura seguente si possono osservare i risultati dell'applicazione del modello SIS, con % di infetti=0.05 , β=0.007 e μ=0.01, sulle tre reti. Il Basic Reproductive Number con questi specifici paramentri equivale a: $\lambda = \frac{\beta}{\mu} = \frac{0.007}{0.01} = 0.7$ che come si può notare è un valore maggiore di tutte e tre le soglie delle rispettive reti.

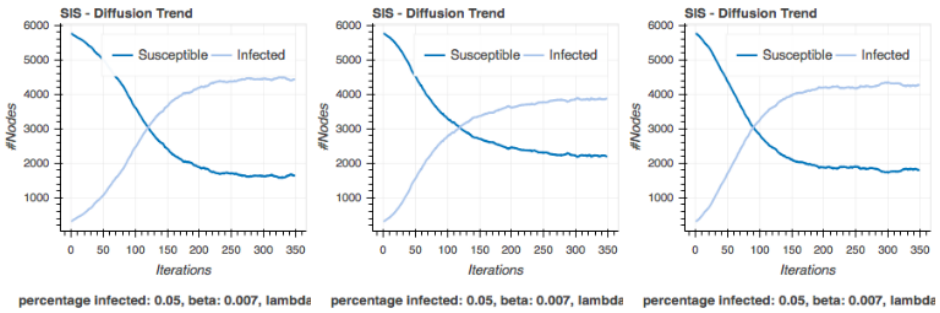

Figure 15: SIS con percentuale di infetti=0.05 e β=0.007 e μ=0.01. A-Rete Random,B-Rete YouTube,C- Rete BA

Successivamente abbiamo testato il modello SIS, con parametri differenti (% di infetti=0.25 e β=0.007 e μ=0.045), notando che anche con una percentuale di infetti iniziali del 25% quindi molto più ampia del caso precedente , l'epidemia non riesce a diffondersi attraverso la rete.
Questo è dovuto al fatto che $\lambda = \frac{\beta}{\mu} = \frac{0.007}{0.045} = 0.155$ del virus è minore della soglia delle tre reti. Notiamo subito che nella rete random la % di infetti diminuisce con il numero di iterazioni mentre per la rete YouTube e la rete scale free rimane pressochè costante nel tempo.

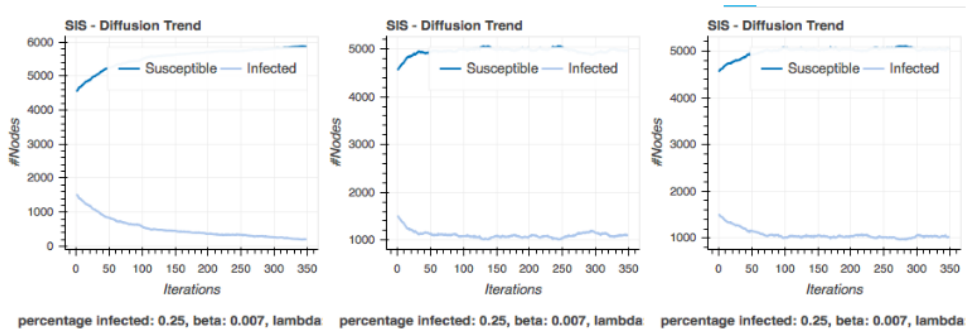

Figure 16: SIS con percentuale di infetti=0.25 e β=0.007 e μ=0.045.
A-Rete Random,B-Rete YouTube,C- Rete BA

4.3.3 SIR (Susceptible - Infected - Removed) Model

Il terzo modello utilizzato è il SIR caratterizzato da tre stati: Suscettibile e Infetto (come il SI,SIS), e da un nuovo stato, Removed che indica che un nodo precedentemente infetto può divenire immune o decedere. Nella figura seguente si può notare il grafico relativo al modello SIR, con % di inodi infetti del 0.05 , β=0.01 e μ=0.006.

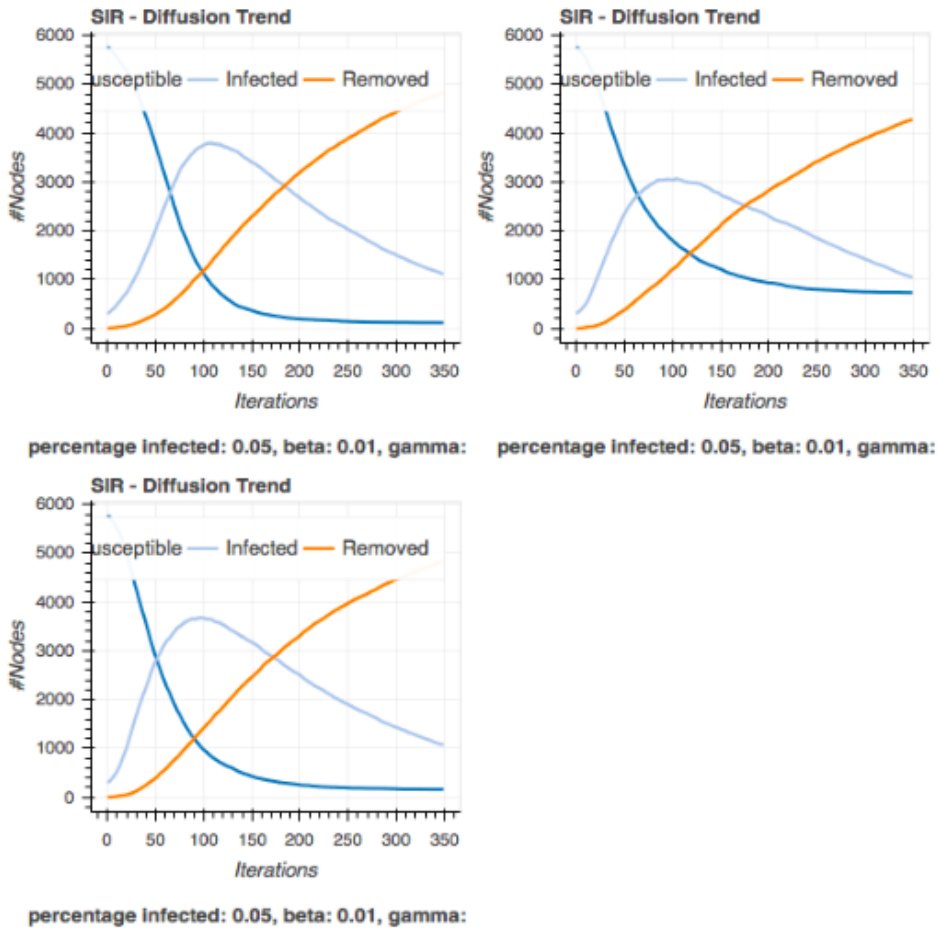

Figure 17: SIR con % di inodi infetti del 0.05 ,β=0.01 e μ=0.006.
A-Rete Random, B-Rete YouTube, C- Rete BA

Possiamo notare come nella rete random l'infenzione dei nodi procede molto più lentamente rispetto alle restanti due reti. Nelle restanti reti, essendo scale free, una volta infettato un nodo hub, l'infenzione esplode in maniera esponenziale.

4.3.4 Threshold

Infine abbiamo terminato l'analisi sulla diffusione delle epidemie applicando alle tre reti il modello a threshold. In questo modello ogni nodo possiede una certa soglia,oltre la quale il nodo verrà infettato, inoltre abbiamo già una % di nodi iniziali già infetti, che saranno l'incipit della diffusione dell'epidemia.

I nodi iniziali infetteranno i nodi vicini solo se il numero di nodi infettati collegati ad un vicino sarà maggiore della soglia di questo. testando il modello sulle tre reti abbiamo ottenuto il seguente risultato:

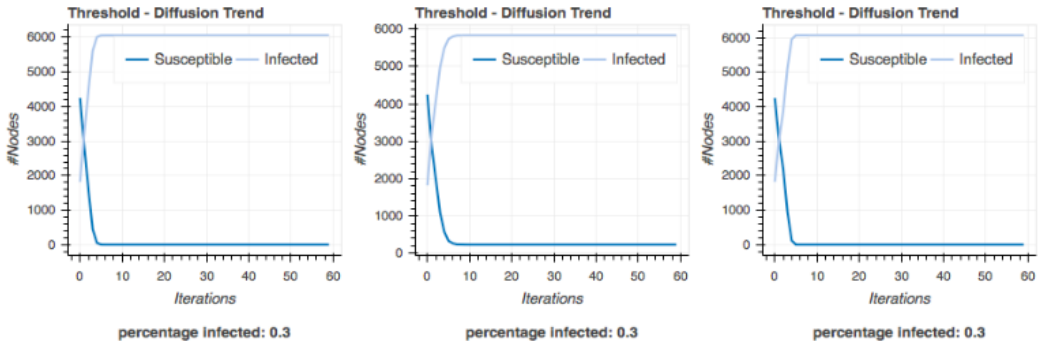

Figure 18: threshold con % di nodi infetti iniziali del 0.30 e con threshold=0.40. A-Rete Random,B-Rete YouTube,C- Rete BA

Applicando il modello threshold con una percentuale di nodi iniziali infetti pari al 30% e un threshold pari a 0.40, notiamo che in tutti e tre i modelli si verifica la diffusione dell'epidemia.

4.3.5 Conclusioni

Lo studio della diffusione delle epidemie attraverso le tre tipologie di reti ci ha portato alla conclusione (che concorda con le nozioni teoriche), che le epidemie, che siano poi malattie o solo diffusioni di notizie-informazioni hanno una maggiore rapidità di espansione nelle reti scale-free (come la BA e la nostra rete di YouTube), grazie alla presenza degli HUB che una volta raggiunti e infettati, riescono a contagiare un numero molto elevato di altri nodi.