Social Network Analysis

Analisi di una porzione di rete sociale del social network musicale LASTFM.
https://github.com/LoreDema/Social_network_analysis_project

UNIVERSITÀ DI PISA

De Mattei Lorenzo: 469944
Meini Andrea: 453027
Rizza Vincenzo: 468707

Lo scopo dell'analisi è stato quello di valutare le caratteristiche della rete sociale e le abitudini musicali degli utenti.

La prima parte comprende la fase di data understanding e data cleaning che ci ha permesso di rilevare, correggendoli sin da subito, alcuni errori di impostazione dei dati da analizzare, e impostare concettualmente le analisi che abbiamo successivamente effettuato.

La rete sociale "reale", ottenuta da LastFM, è stata successivamente analizzata in rapporto alla rete casuale generata utilizzando la libreria Python NetworkX.
L'ultima fase del lavoro svolto ha riguardato l'analisi delle abitudini musicali degli utenti anche in relazione alla centralità degli utenti stessi all'interno del network.

1. Data understanding

LastFM è un social network che prevede collegamenti univoci tra i propri utenti. Per questo motivo la rete costituisce un grafo *non direzionato*.

Partendo da un utente seed abbiamo eseguito il crawling, ottenendo una rete composta da 5098 nodi e 32812 archi. L'utente da noi scelto come seed è "Maxxtallica" (http://www.lastfm.it/user/maxxtallica).

- Seed(utente iniziale): "Maxxtallica"
- Nodi(utenti): 5098
- Archi(amicizie): 32812

I dati crawlati a nostra disposizione sono formati da 4 file:

- Network_cleaned.csv: lista delle coppie di utenti collegati.
- Listenings.csv: contiene, per ogni utente, la lista delle canzoni ascoltate (al più 200 per ogni utente)

- Hotttnnesss.csv: rappresenta il grado di popolarità degli artisti in una scala che va da 0,0 a 1. Il file è composto da 8229 artisti sugli effettivi 48015 presenti nel file listening. Gli artisti che apparivano nel file listening.csv, con frequenza minore di 10 non sono stati presi in considerazione. I dati sono stati ottenuti tramite il servizio API di The Echonest (http://the.echonest.com/)
- Genre.csv: comprende la lista dei termini con i quali ogni artista viene descritto. Per ogni artista abbiamo una lista di termini ai quali sono associati i seguenti valori:
    - frequency, indica la frequenza di utilizzo di quel termine nelle descrizioni dell'artista
    - weight, indica la rilevanza di quel termine nelle descrizioni dell'artista.

  D'ora in poi utilizzeremo il termine "generi" riferendoci ai termini che descrivono un artista. Anche questi dati sono stati ottenuti tramite il servizio Api di The Echonest

Nell'analizzare il file listening.csv abbiamo notato che presentava numerose ripetizioni, ovvero talvolta erano presenti più volte le medesime canzoni ascoltate da uno stesso utente alla stessa ora. Il problema dipende dai risultati restituiti dal Database di LastFm, in quanto lo stesso problema è stato riscontrato visualizzando i dati tramite l'interfaccia grafica di LastFm. Abbiamo quindi eliminato dal file listening.csv tutti gli ascolti ripetuti.

2. Preprocessing

Per ogni artista a abbiamo calcolato l'appartenenza percentuale $A(a,g)$ ad ogni genere g basandoci sui pesi dei generi riferiti a quell'artista nel seguente modo:

$$(1)\ A(a,g) = \frac{W(a,g)}{\sum_{i=1}^{|GA(a)|} W(a,i)},$$

dove $W(a,g)$ è il peso del genere g per l'artista a e $GA(a)$ è la lista dei generi riferita a quell'artista.

Quindi abbiamo calcolato la frequenza negli ascolti per ogni genere $F(g)$:

$$(2)\ F(g) = \sum_{i=1}^{|L|} A(artist(i),\ g),$$

dove L è la lista degli ascolti della rete, e $artist(i)$ è l'artista corrispondente all'ascolto i.

Abbiamo infine calcolato la hotness media per ogni genere $HG(g)$:

$$(3)\ HG(g) = C \cdot \frac{\sum_{i=1}^{|Art|} A(Art(i),g)\cdot H(Art(i)}{\sum_{i=1}^{|Art|} A(Art(i),g)} + F(g),$$

dove $Art(i)$ è l'i-esimo artista, $H(a)$ è l'hotness dell'artista a e C è una costante data da

$$(4)\ C = K \cdot \frac{\sum_{i=1}^{|G|} \frac{F(G(i))}{\sum_{i=1}^{|Art|} A(Art(i),G(i))*H(Art(i)}}{|G|},$$

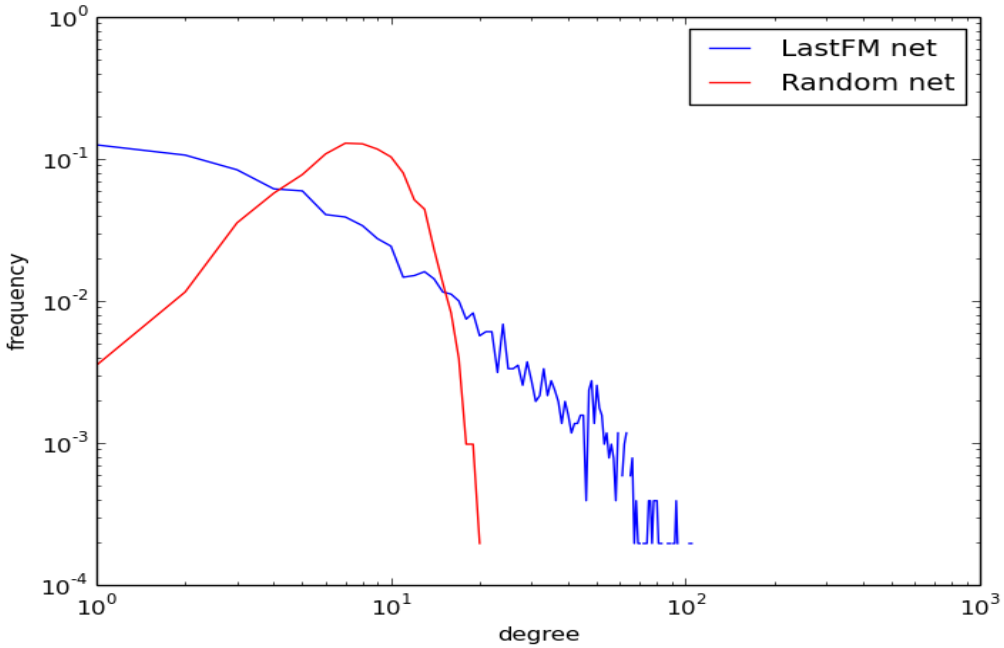
con G lista di tutti i generi, $G(i)$ il genere i-esimo e K costante pari a 10.
Abbiamo deciso di sommare la frequenza del genere in (3) per normalizzare il dato della hotness di genere in quanto erano presenti nel dataset generi che comparivano solamente una volta ma appartenevano ad un artista con una buona hotness. Questi generi avevano una hotness molto alta ma rappresentavano un dato di scarso valore. Per dare dunque più importanza al dato della hotness degli artisti rispetto al dato sulla frequenza del genere negli ascolti, abbiamo inserito la costante C, che è data dal rapporto medio per tutti i generi tra la hotness di genere calcolata senza l'utilizzo della frequenza e la frequenza stessa, moltiplicata per un costante K che indica l'importanza della hotness rispetto alla frequenza.

3. Social network analysis

Per l'analisi della rete abbiamo creato utilizzando la libreria per Python NetworkX una rete casuale utilizzando il metodo *gnm_random_graph(n, m)*. I parametri *n* e *m* indicano rispettivamente il numero di nodi e archi. Sono stati fissati in modo tale da essere corrispondenti al numero di nodi e archi della rete di LastFm. Quindi abbiamo confrontato le due reti secondo diversi aspetti. Segue la descrizione di queste analisi.

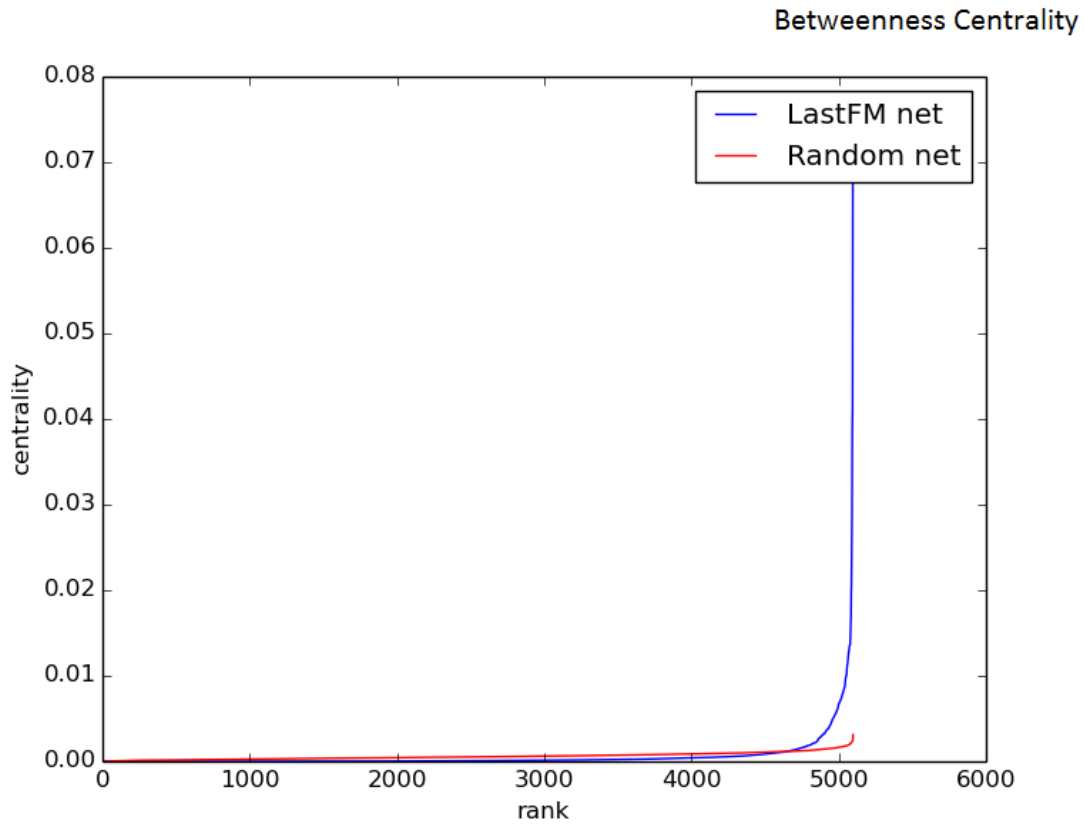
3.1 Degree distribution

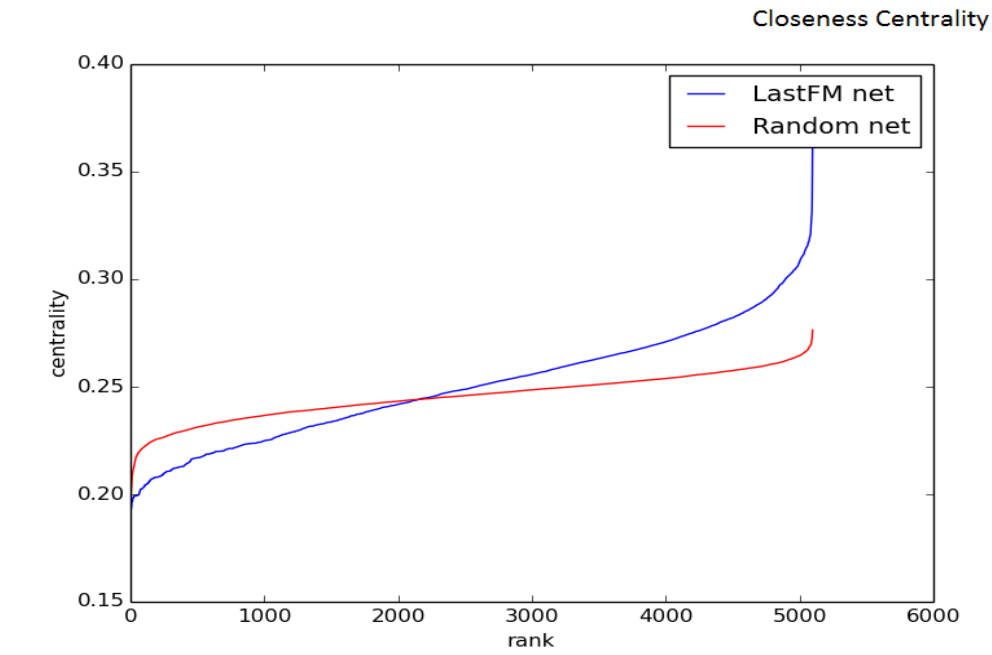
Nell'analizzare il rapporto nella distribuzione dei gradi (numero di connessioni di un nodo) per le due reti abbiamo riscontrato una notevole differenza di distribuzione. Nella rete casuale infatti la maggior parte dei nodi hanno un grado di distribuzione medio, abbiamo un andamento di tipo gaussiano. Nella rete "reale" al contrario troviamo una grossa quantità di nodi aventi un grado di distribuzione molto basso e pochi nodi aventi valori elevati. Il dato indica una maggiore uniformità tra i collegamenti all'interno della rete random rispetto alla rete LastFm.

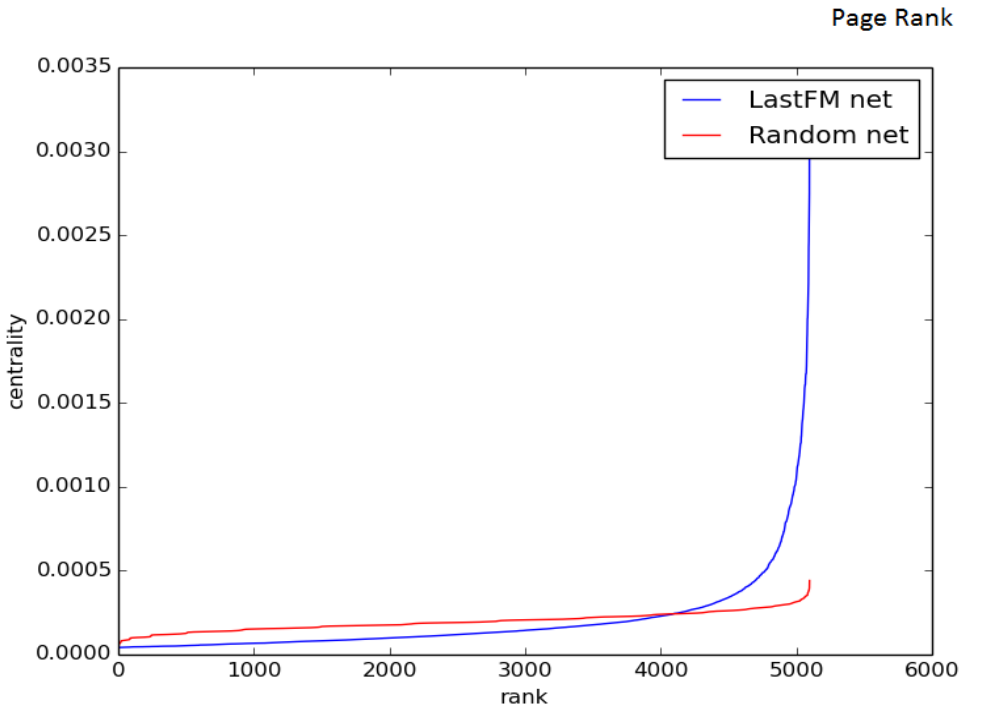


3.2 Misure di centralità

L'andamento delle tre misure di centralità per entrambe le reti è simile. Per il PageRank e la betweenness osserviamo la presenza di molti nodi con un valore molto basso e pochi nodi con un valore alto. Per la closeness è invece osservabile un ampio numero di nodi con valore medio e pochi nodi con valore basso e alto. Confrontando però l'andamento della rete LastFm con la rete random osserviamo che i valori della rete LastFm hanno un range molto più ampio con tutte e tre le misure e l'andamento per questa rete si discosta molto di più dall'andamento lineare rispetto alla rete random. È evidente che nella rete reale esiste una gerarchia di utenti molto più definita rispetto alla rete random.







3.3 Connected components

Le reti restituiscono un valore di connected components uguale a 1, ovvero non sono presenti sottoreti.

E’ interessante notare che questo dato, per la rete random, è legato alla scelta del metodo di generazione *gnm_random_graph*.

3.4 Average shortest path

La media dei cammini minimi per tutte le possibili coppie della rete LastFm e di quella random si attesta su circa 4 passi; entrambe le reti sono abbastanza compatte e la differenza tra le due sotto questo aspetto è trascurabile.

```
LastFM network average shortest path:
4.03868897873
Random network average shortest path:
4.08111095127
```

3.5 Average cluster coefficient

Abbiamo invece una differenza netta relativa al loro coefficiente medio di clustering, un dato che evidenzia l’attitudine della rete in analisi a formare sottoreti fortemente connesse tra loro.

```
LastFM network average cluster
coefficient: 0.145379112823
Random network average cluster
coefficient: 0.00180493834902
```

Dai dati ottenuti abbiamo notato che nella rete casuale il coefficiente di clustering è nettamente inferiore rispetto alla rete reale, il che dimostra la maggiore tendenza alla formazione di sottoreti coese di utenti all’interno del network di lastFm. Il dato è interessante e attraverso un’analisi più approfondita dei cluster di nodi si potrebbe valutare quanto questi gruppi condividano gli stessi gusti musicali.

La differenza è sottolineata oltretutto dal valore del “ Network Heterogeneity”, nonchè la tendenza della rete alla creazione di Hub.

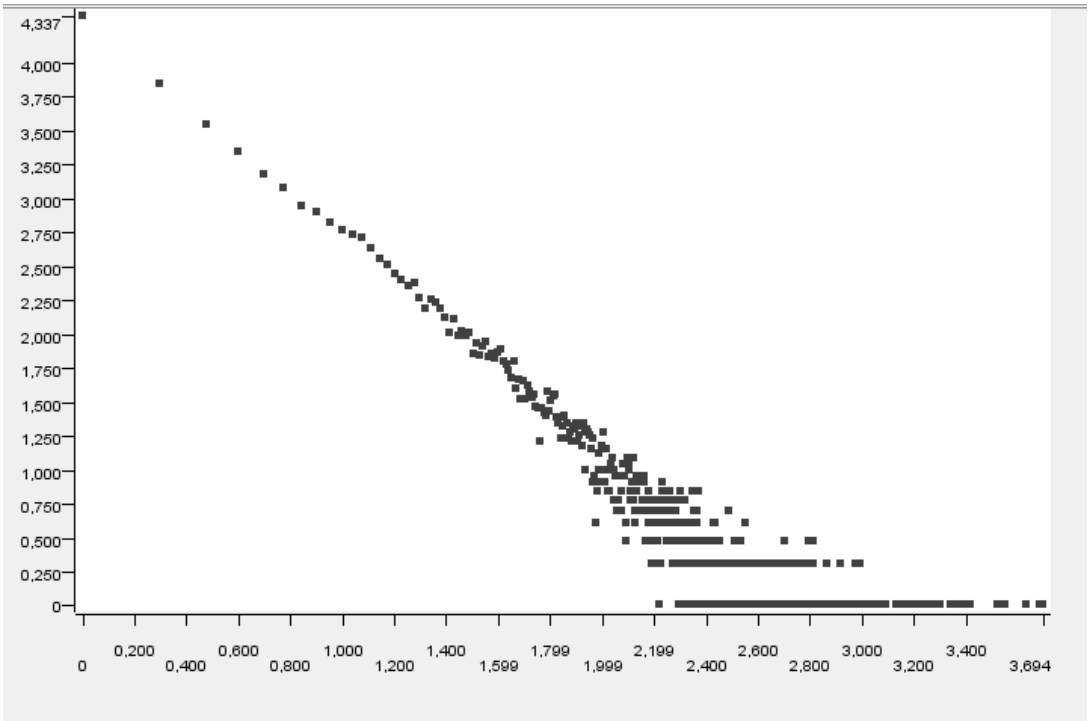
Random Network Heterogeneity: *0.329*
Real Network Heterogeneity: *1.474*

4. Listening analysis

Nell'analizzare i dati dei file aggiuntivi a nostra disposizione abbiamo cercato di mettere in relazione alcuni indicatori, nell'intento di vedere se ci fosse qualche interazione e collegamento tra i valori di hotnesss nonchè i valori di genere con le tendenze di ascolto riscontrate all'interno della nostra porzione significativa di utenti esaminati.

4.1 Distribuzione ascolti/artisti

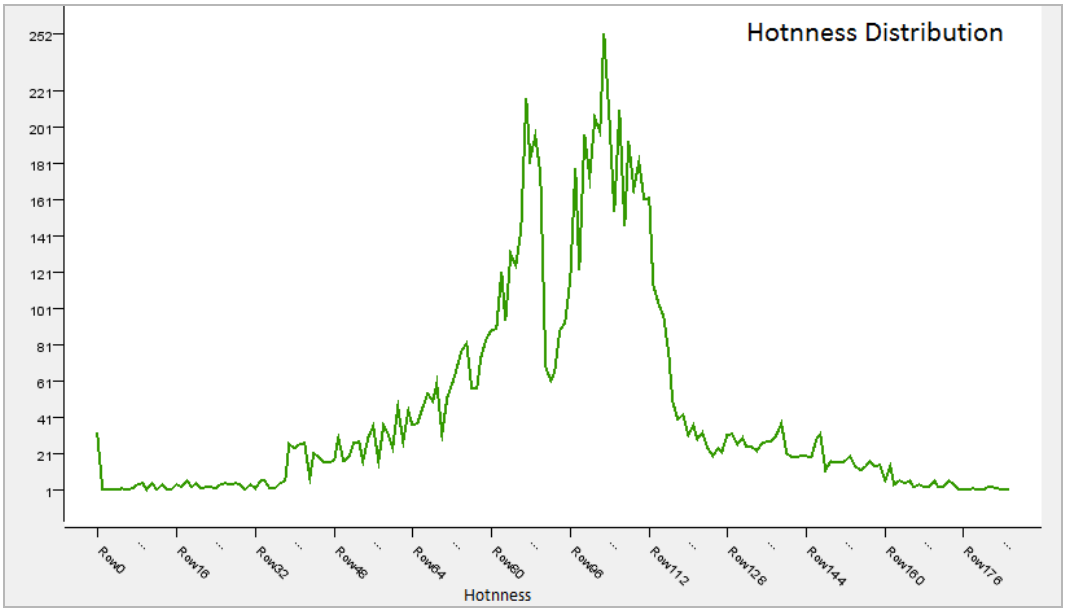
Il grafico sottostante mette in relazione, in scala logaritmica, la distribuzione dei numeri di ascolti raggruppati per artista.



Osservando il grafico è chiaro come vi sia una grande quantità di artisti ascoltati pochissime volte e, al contrario, un numero ristretto di artisti le cui tracce vengono ascoltate molto frequentemente.
Il dato è interessante in quanto solamente gli artisti principali hanno un grosso numero di utenti che ascoltano la loro musica a differenza della moltitudine di artisti meno conosciuti per i quali gli ascolti si riducono a poche decine di utenti.

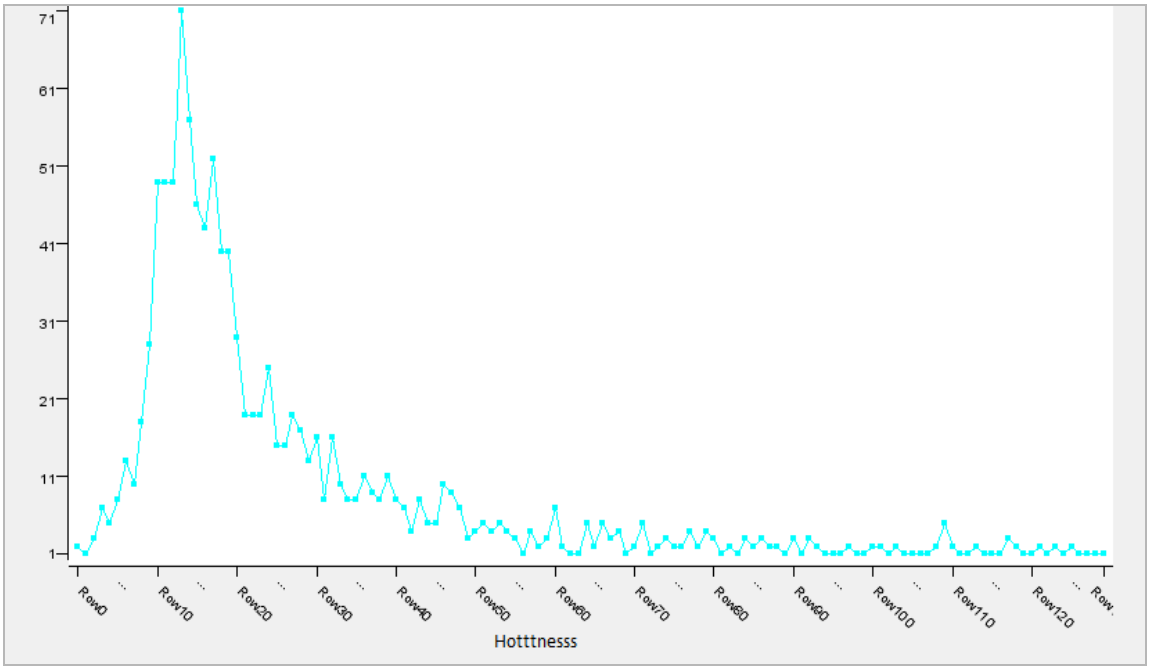
4.2 Distribuzione della Hotness in relazione agli artisti

Il plot sottostante mostra la distribuzione delle hotness (parametro che indica quanto un artista è popolare). Risulta evidente come all'interno della nostra rete siano più frequenti artisti aventi valori di hotness media mentre artisti con hotness bassa e alta risultano essere poco frequenti. Ci saremmo aspettati un alto numero di artisti con hotness bassa, questo dato potrebbe essere influenzato dal fatto che molti artisti con hotness ridotta non vengono assolutamente ascoltati all'interno della rete, o comunque gli ascolti relativi a questi artisti non raggiungo la soglia minima di 10 e quindi vengono scartati in fase di crawling.



4.3 Distribuzione della Hotness media dei generi

Dal plot ottenuto si evince che una grande percentuale di generi ascoltati dagli utenti del nostro dataset ha bassi valori di hotness. I generi musicali con valori di hotness medi o alti sono meno presenti tra gli ascolti degli utenti della nostra rete.



4.4 Distribuzione generi in relazione alla hotness degli artisti

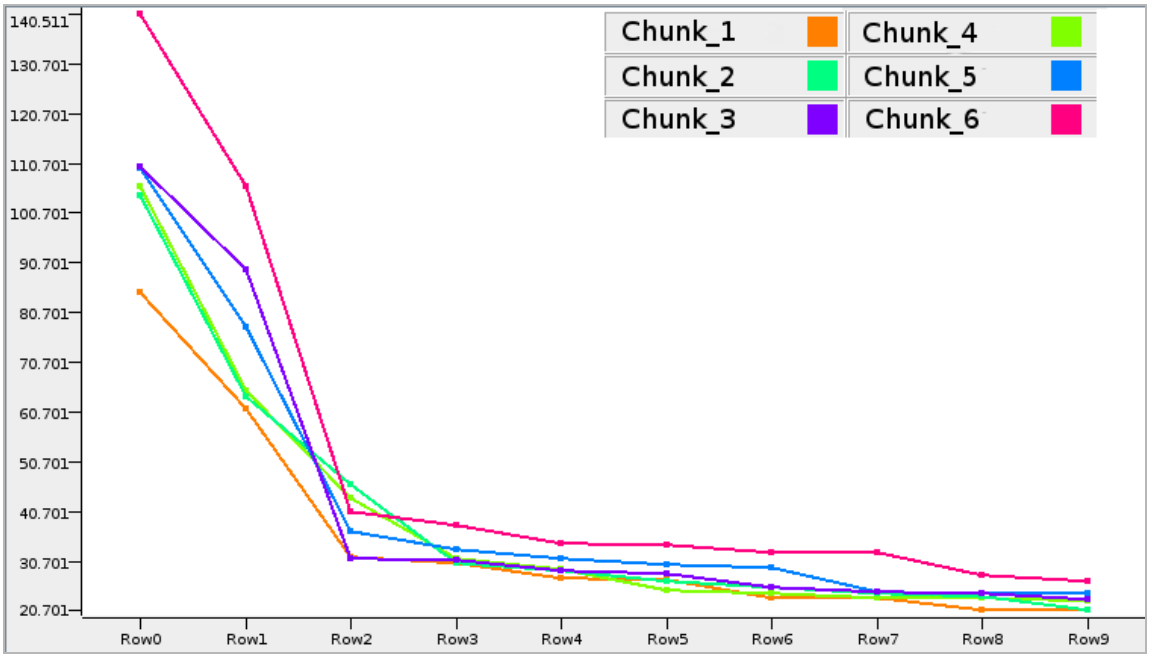
Un'ulteriore analisi è stata fatta dividendo in sei gruppi (*chunks*) la lista degli artisti ordinati per hotness. Dopodiché, per ogni chunk è stata calcolata la somma dei valori di appartenenza percentuale di ciascun genere per ogni artista, mostrando così quanto ciascun genere è frequente nel chunk in cui è presente.
Ordinando i valori ottenuti in senso ascendente e mostrando la Top10 dei generi per ciascun chunk (dove Chunk_1 è il gruppo di generi derivati dagli artisti con hotness bassa e Chunk_6 è il gruppo derivante dagli artisti con hotness alta), si ottengono i

seguenti risultati:

Row ID	Chunk 1 ...	Chunk 2 ...	Chunk 3 ...	Chunk 4...	Chunk 5 ...	Chunk 6...
Row0	rock	rock	rock	rock	rock	pop
Row1	metal	metal	pop	pop	pop	rock
Row2	heavy metal	pop	metal	metal	electronic	hip hop
Row3	jazz	heavy metal	jazz	electronic	indie	dance
Row4	pop	hard rock	electronic	indie	alternative	rap
Row5	hard rock	electronic	indie	alternative	metal	r&b
Row6	folk	folk	punk	jazz	soul	electronic
Row7	punk	death metal	80s	soundtrack	jazz	soul
Row8	blues	jazz	alternative	punk	alternative ...	alternative
Row9	death metal	indie	heavy metal	folk	hip hop	indie

Si nota che alcuni generi si distribuiscono abbastanza uniformemente all'interno di ogni chunk: *rock* è quasi sempre il genere che meglio descrive gli artisti, indipendentemente dalla loro hotness, così come *pop* e *indie* sono quasi sempre presenti.
Ci sono però alcune interessanti differenze: i generi *metal* e *heavy metal* sono molto prevalenti nei gruppi di artisti di hotness bassa e media ma non sono per niente presenti nei gruppi di artisti con hotness alta; *jazz* e *folk* subiscono lo stesso andamento ma in maniera più lieve; *hip hop* è in terza posizione nel Chunk_6 ma negli altri chunk è quasi assente.

Altri dati interessanti emergono dalla comparazione delle distribuzioni dei valori di appartenenza percentuale dei Top10 generi dei sei chunk.



Si nota che i primi due generi del Chunk_6 hanno un valore nettamente più alto dei primi due generi del Chunk_1 e, più in generale, i valori di appartenenza percentuale dei generi che meglio descrivono ciascun gruppo aumenta con l'aumentare della hotness degli artisti presenti in quel gruppo. Ciò significa che, ad esempio, i generi *pop* e *rock* descrivono in media molto più efficacemente gli artisti con hotness alta, mentre il gruppo di artisti con hotness bassa è descrivibile in maniera meno netta rispetto ai propri generi di riferimento.

4.5 Analisi degli ascolti per gruppi di utenti ordinati per centralità

Abbiamo suddiviso la lista degli utenti ordinati per grado di centralità in 4 gruppi, utilizzando come misure di centralità prima la closeness e poi la betweenness.
Per ogni gruppo abbiamo calcolato la frequenza degli ascolti per ogni artista, quindi abbiamo ottenuto una tabella di artisti e relativa frequenza di ascolto per ogni gruppo di utenti ordinata per frequenza.
In seguito abbiamo aggiunto alle tabelle due colonne, una contenente *per ogni artista* la frequenza di ascolti su tutta la rete, un'altra contenente *per ogni artista* la hotness corrispondente.

Similarmente per ogni gruppo abbiamo calcolato la frequenza degli ascolti *per ogni genere* utilizzando la formula (2) del capitolo 2, quindi abbiamo ottenuto una tabella di generi e relativa frequenza di ascolto per ogni gruppo di utenti ordinata per frequenza.
Abbiamo quindi aggiunto alle tabelle due colonne, una contenente *per ogni genere* la frequenza di ascolti su tutta la rete, un'altra contenente per ogni genere la hotness media corrispondente.

Per calcolare le frequenze abbiamo utilizzato gli ultimi 100 ascolti per ogni utente e abbiamo scartato gli utenti che non avevano almeno 100 ascolti.

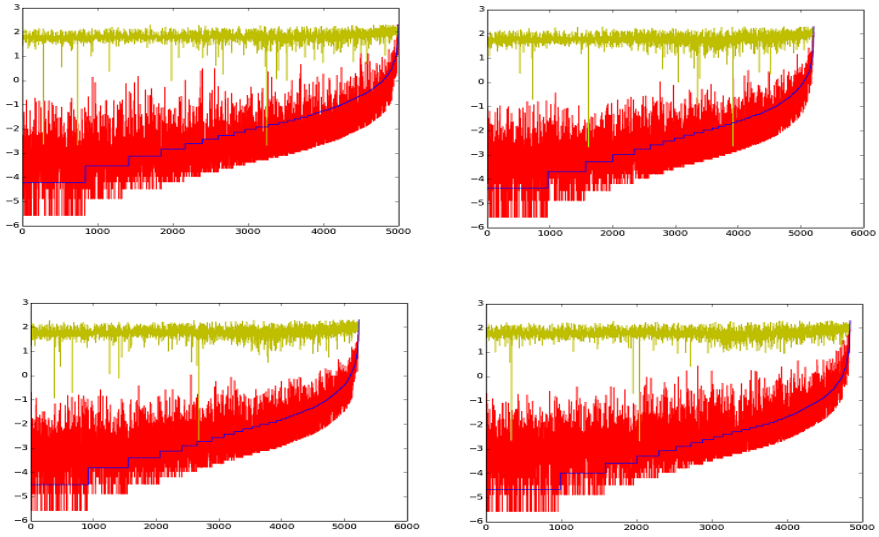
Per rendere confrontabili gli andamenti li abbiamo normalizzati. Per ogni colonna delle tabelle abbiamo applicato la seguente funzione ad ogni elemento:

$$N(x) = log_e\left(x * \frac{10}{max(X)}\right) ,$$

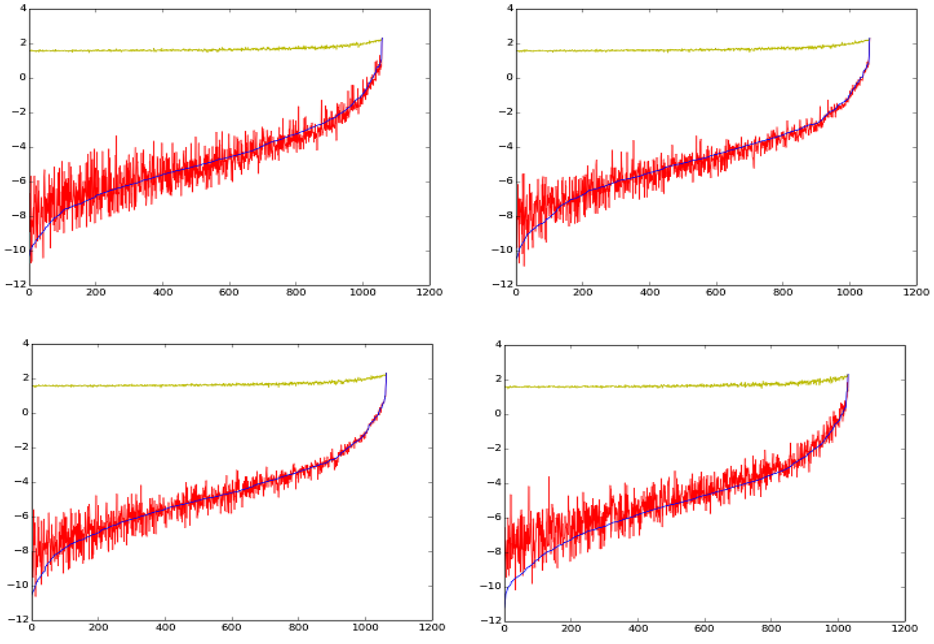
dove X corrisponde alla lista degli elementi della colonna. In tal modo gli elementi di tutte le colonne avranno come valore massimo $log_e(10)$. Quindi abbiamo stampato i diagrammi relativi. In blu possiamo osserva il valore della colonna relativa alle frequenze del gruppo di utenti, in rosso le frequenze globali, in giallo le hotness.

Dall'analisi dei diagrammi risulta evidente una scarsa correlazione tra la frequenza di ascolto e la hotness degli artisti. Emerge una lieve correlazione tra la frequenza di ascolto dei generi e hotness media dei generi, ma questo dato è scontato e non rilevante, in quanto come spiegato nel paragrafo 2, per calcolare la hotness media per genere si considera anche la frequenza di ascolto stessa. Emerge quindi che nessuno dei sottogruppi della rete è rappresentativo del trend generale degli ascolti musicali all'infuori della rete. Questo potrebbe essere determinato dal fatto che la rete in analisi non ha dimensioni grandissime o dalla scelta del seed in fase di crawling. Inoltre emerge che l'andamento della frequenza generale è abbastanza correlato con l'andamento della frequenza dei gruppi, sia per quel che riguarda i generi musicali

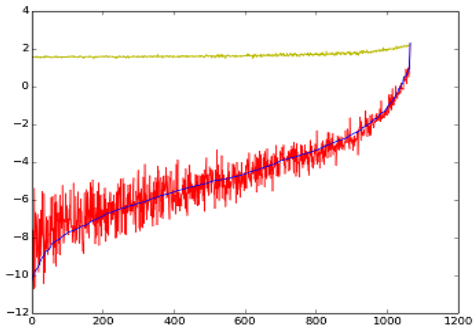
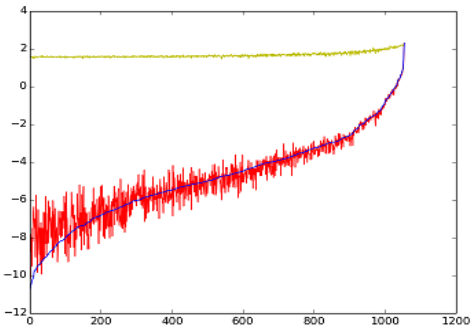
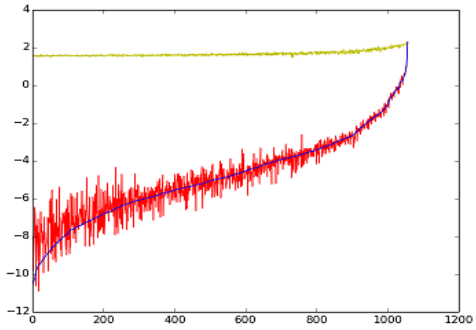
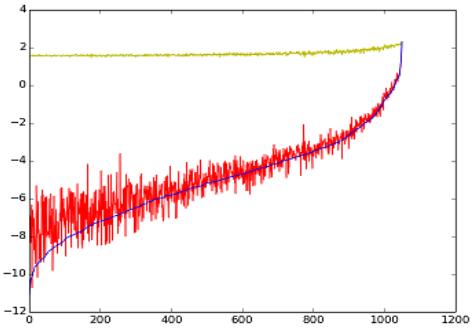
che gli artisti. Ciò indica che non ci sono variazioni significative nell'abitudine degli ascolti tra gli utenti più periferici, quelli più centrali e quelli con una centralità media.



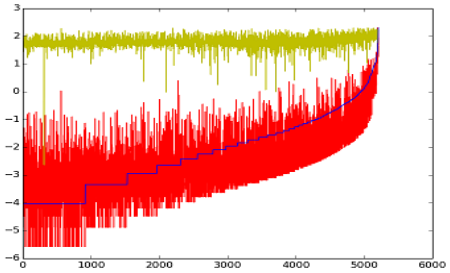
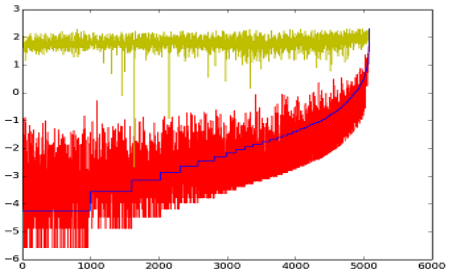
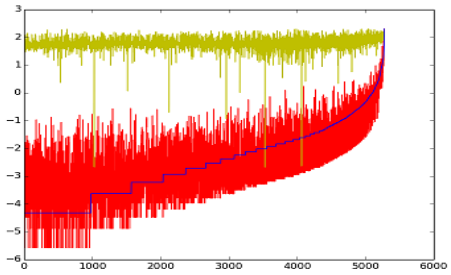
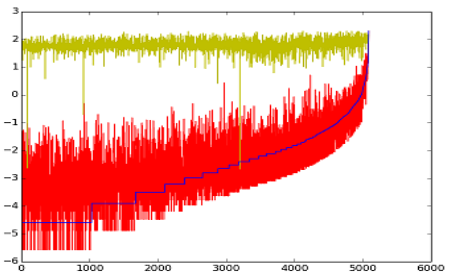
Artist chunk closeness



Genre chunk closeness

Genre chunk betweenness

Artist chunk betweenness