Analisi dell'affinità musicale degli artisti di Last.fm

Calogero Giudice
c.giudice@studenti.unipi.it
Student ID: 530155

ABSTRACT

Questo progetto analizza una rete di artisti musicali del social network *last.fm* in base a dei coefficienti di similarità (Sample e Jaccard Ratios) che potrebbero rappresentare delle affinità a livello stilistico o di pubblico. Sono state esaminate sia le **caratteristiche strutturali** della rete, procedendo a svolgere una previsione, in base ai Ratios menzionati in precedenza e al coefficiente di popolarità, su possibili aumenti o cali di popolarità futuri.

Questi risultati possono rappresentare un **modello** che potrebbe riflettersi nello scenario musicale odierno, fornendo nuovi studi e strumenti che possono essere utili a prevedere eventuali nuovi scenari. Infatti, possono esserci artisti emergenti che potrebbero godere di maggiore popolarità futura se ben collegati, con artisti di popolarità alta e collegati con ratio alti; allo stesso tempo, si potrebbero notare alcuni artisti ritenuti popolari che potrebbero avere un trend opposto e quindi in declino.[1]

KEYWORDS

Social Network Analysis, musica, *Last.fm*, affinità, similarità, popolarità

ACM Reference Format:
Calogero Giudice. 2025. Analisi dell'affinità musicale degli artisti di Last.fm. In *Analisi dell'affinità musicale degli artisti di Last.fm.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1]**Project Repositories**
Data Collection: https://github.com/Kalo9603/2024_Giudice/tree/main/data_collection
Network Analysis: https://github.com/Kalo9603/2024_Giudice/tree/main/network_analysis
Open Problem: https://github.com/Kalo9603/2024_Giudice/tree/main/open_problem
Report: https://github.com/Kalo9603/2024_Giudice/tree/main/report

SNA '25, 2024/25, University of Pisa, Italy

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$0.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUZIONE

Nell'epoca dell'affermazione dell'industria musicale digitale, piattaforme come *last.fm* offrono tanti dati sugli artisti musicali, sulle abitudini di ascolto degli utenti e sulle relazioni tra i singoli profili musicali (gli utenti, gli artisti, gli album, etc.). Gli artisti presenti nella piattaforma sono stati analizzati sotto una prospettiva di rete sociale: si possono ipotizzare così degli scenari, dei fenomeni interessanti: la similarità tra artisti, la popolarità degli stessi, l'affermazione di artisti emergenti o l'effetto contrario.

L'analisi parte dalla costruzione di una rete basata su coefficienti di similarità; di questi ne sono stati individuati due:

- Il *Sample Ratio*, ovvero l'intensità del legame tra due artisti in base ai simili condivisi.
- Lo *Jaccard Ratio*, ovvero l'affinità relativa all'insieme totale di simili.

Matematicamente, è molto più probabile che lo *Jaccard* sia inferiore del rispettivo *Sample*, poiché il denominatore del primo è sempre maggiore del secondo (a meno che tali denominatori non coincidono).

In parallelo si è stimata la **popolarità** degli artisti presenti nel dataset: per far ciò ci si è basati sullo **Z-Index**, che mette in relazione il numero di ascoltatori e quello di ascolti. Ciò ha permesso di classificare gli artisti in tre classi: ***mainstream***, ***medio*** ed ***emergente***; a questi ne sono stati affiancati altri tre: ***possibile mainstream***, ***possibile medio*** e ***ritorno emergente*** in base al trend di (de)crescita calcolato.

L'obiettivo principale della seconda parte progetto è quello di identificare le caratteristiche strutturali degli artisti mainstream e di quelli emergenti ed, eventualmente, di prevedere un cambio di status in base a determinate metriche.

2 DATA COLLECTION

La raccolta dei dati da analizzare si è concentrata su tre fasi specifiche:

(1) La generazione di un elenco di artisti in base agli artisti simili;
(2) La creazione di una lista di archi che collegano due artisti se entrambi i coefficienti di similarità superano

la soglia del 50%. Il *Sample Ratio* è calcolato rispetto al numero di simili condivisi, mentre lo *Jaccard Ratio* rispetto all'insieme totale dei simili;

(3) Il successivo accostamento del numero di ascolti, di ascoltatori e il calcolo delle popolarità degli artisti.

Gli elenchi degli artisti e degli artisti simili sono stati ottenuti per mezzo dell'API di *last.fm* e, tramite programmazione ad oggetti, sono state generate le altre liste. Nel complesso vi sono quattro file:

- *artists.csv*, una lista di 28 870 artisti;
- *links.csv*, un elenco di 98 890 archi che collegano 11 745 artisti (il 40.68% del totale). A causa delle lunghe tempistiche è stato elaborato il dataset sul 64.26% dei possibili archi totali (sono stati esaminati più di 267 milioni di possibili archi su oltre 416 milioni totali: $\frac{28\,870 \cdot 28\,869}{2} = 416\,724\,015$);
- *popularity.csv*: la medesima lista di *artists.csv* con inclusi il numero di ascoltatori, quello di ascolti, il coefficiente di popolarità e il suo logaritmo;
- *popularity_z.csv*: la medesima lista di *popularity.csv* a cui si accoda lo Z-Index.

In particolare, i coefficienti di correlazione citati sono calcolati come segue.

Indice	Formula
Sample Ratio	$SR(a,b) = \frac{\lvert S_a \cap S_b \rvert}{k}$
Jaccard Ratio	$JR(a,b) = \frac{\lvert S_a \cap S_b \rvert}{\lvert S_a \cup S_b \rvert}$
Popolarità	$pop_i = \frac{ascolti_i}{ascoltatori_i + 1}$
Popolarità logaritmica	$pop_{\log,i} = \log(pop_i + 1)$
Media	$\mu = \frac{1}{N} \sum_{i=1}^{N} pop_{\log,i}$
Deviazione standard	$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (pop_{\log,i} - \mu)^2}$
Z-Index	$Z_i = \frac{pop_{\log,i} - \mu}{\sigma}$

Table 1: Formule utilizzate per la classificazione e l'analisi strutturale

Dati selezionati

Al fine di avere un dataset iniziale di artisti si è pensato, partendo da quattro artisti provenienti da quattro generi musicali diversi (Annalisa per il pop italiano; Giorgio Vanni per le sigle, Katy Perry per la quota internazionale e Ado per il mercato giapponese).

Quindi, partendo da essi, viene creata la lista di artisti simili a cominciare da quello in input: se non è presente, l'artista trovato viene aggiunta in lista. La medesima operazione viene svolta, alla prima iterazione, sui primi quattro artisti; poi, dalla seconda, si considerano anche tutti quelli aggiunti in coda. In questo modo è stato ottenuto un dataset di 28 870 artisti. Relativamente all'ordine di inserimento nel dataset, a ciascun artista è stato assegnato un ID autoincrementale.

Al termine si è passati, basandosi sull'elenco degli artisti, alla generazione del dataset dei *links* del grafo. Per far ciò per ciascun artista della lista, sfruttando le proprietà di un grafo non orientato:

- Si considera l'artista con ID immediatamente successivo;
- Per la coppia in esame si prendono i rispettivi insiemi di artisti simili e si calcolano i ratio, le cui formule sono specificate nel formulario precedente;
- Se entrambi il Sample e lo Jaccard Ratio raggiungono la soglia del 50%, l'arco viene registrato nel dataset;
- Si prosegue il calcolo con l'artista successivo con il medesimo algoritmo fino a raggiungere la fine della lista degli artisti.

In questo modo, ad esempio, il primo nodo (Annalisa) farà 28 869 iterazioni, il secondo (Giorgio Vanni) 28 868, e così via. Quindi gli ultimi nodi faranno molte meno iterazioni rispetto ai primi perché le altre possibili coppie vengono esaminate dai nodi precedenti.

Per migliorare ulteriormente la velocità del processo è stato adottato il *multiprocessing*, in modo che più processi lavorino in parallelo sullo stesso set di artisti. Dati i tempi relativamente lunghi del processo, l'analisi del dataset è iniziato non appena il numero di artisti all'interno dei dataset dei *links* ha raggiunto la soglia richiesta. Al termine sono stati collegati 11 745 artisti con 98 890 archi.

3 NETWORK CHARACTERIZATION

Si esamina qui di seguito il grafo ottenuto dalla fase di Data Collection.

Come anticipato in precedenza, partendo da 28 870 artisti è stato ottenuto un grafo, analizzando il 64,26% dei possibili archi, formato da 11 745 artisti a 98 890 archi.

La densità quindi è molto bassa, ovvero di 1.433×10^{-3}; questo è dovuto principalmente al fatto di aver adottato una soglia minima di Jaccard Ratio più alta. Infatti, matematicamente, la probabilità che lo Jaccard Ratio di un arco abbia valore minore del rispettivo Sample è nettamente maggiore, in quanto il denominatore del rapporto (vedasi la tabella nel

paragrafo precedente) risulta maggiore di un valore fisso come nel Sample.

Riassumendo, le prime generalità del grafo sono le seguenti:

Artisti totali	28 870
Artisti nel grafo	11 745 (40,68%)
Archi	98 890
Densità	1.433×10^{-3}

Table 2: Prime generalità del grafo

Distribuzione di grado

Il grado medio del grafo è di 8.42.
Per calcolarlo sono stati utilizzati, per completezza, due metodi:

- Dividendo il numero di righe del dataset per il numero di artisti connessi;
- Dividendo la somma dei gradi dei singoli artisti per il numero degli stessi.

Nota che l'errore tra i due metodi è dell'ordine di 10^{-4}. Dal grafico si evince la forte asimmetria del grado medio degli artisti: ve ne sono molti con grado basso e meno con grado più alto.

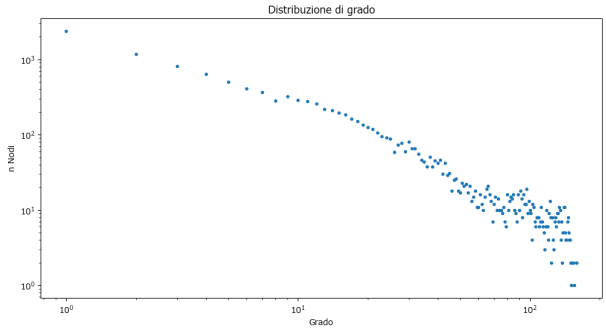

Figure 1: Distribuzione del grado degli artisti

Questa asimmetria dei gradi del grafo si mostra maggiormente dalle distribuzioni CDF e CCDF: la probabilità di trovare un artista con grado elevato decresce.

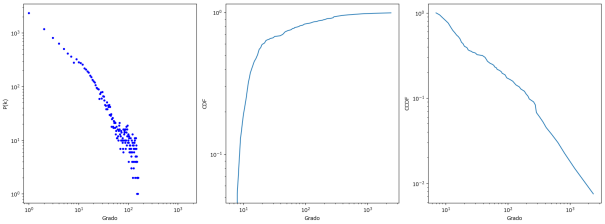

Figure 2: Distribuzione del grado degli artisti - CDF e CCDF

Dato che si tratta di un grafo pesato su due livelli, è interessante mostrare come si comportano i rispettivi pesi con i medesimi grafici. Per cui si illustrano i medesimi grafici ma pesati: qui è evidente come lo Jaccard Ratio risulta più traslato indietro rispetto al Sample corrispondente. dal grafico corrispondente CDF e CCDF si nota come il range dei gradi pesati Jaccard sia molto più variabile rispetto al Sample. Risulta altresì interessante la continua sovrapposizione di entrambi i gradi nel modello CCDF:

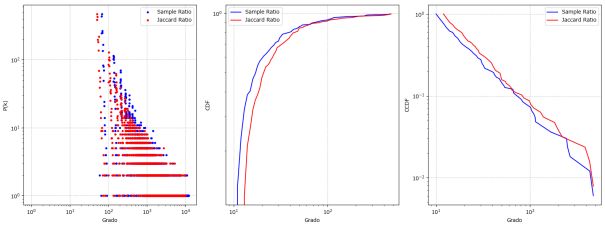

Figure 3: Distribuzione del grado pesato degli artisti - CDF e CCDF

Confronto con il grafo ER

Il grafo è stato confrontato con un corrispettivo Erdős-Rényi (ER) con lo stesso numero di archi e la stessa densità. Come da definizione, la distribuzione è simil-parabolica e i valori, data la densità relativamente bassa, poco appaiati.

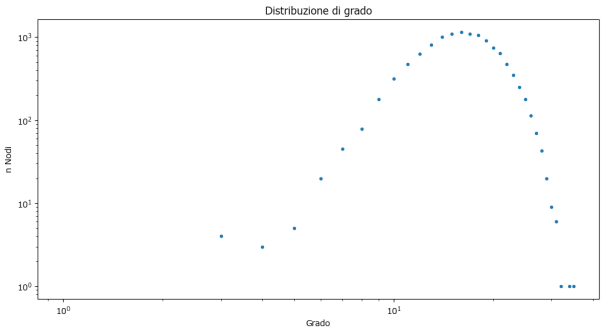

Figure 4: Grafo ER corrispondente

Confronto con il grafo BA

In maniera speculare è stato creato anche un corrispettivo Barabási-Albert (BA), formato da un unico componente, con lo stesso numero di archi e come parametro m, si è scelto il valore di default ($m = 3$); la sua densità è minore del grafo principale e ha più di un terzo degli archi. Il regime è *Supercritical*.

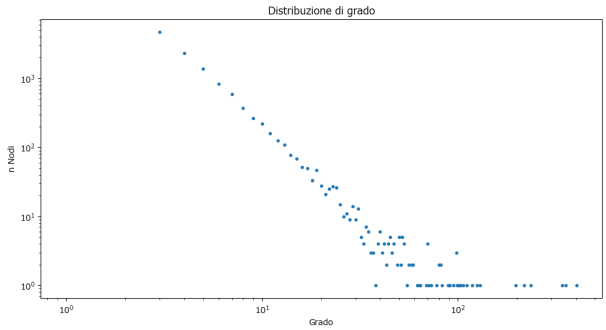

Figure 5: Grafo BA corrispondente

Dunque sono state messe a confronto le tre distribuzioni: per grado $n \approx 10$ tutti e tre tendono a coincidere. Le generalità dei tre grafi sono le seguenti:

	Originale	ER	BA
Nodi	11 745		
Archi	98 890	98 479	35 226
Densità	1.433×10^{-3}	1.427×10^{-3}	5.107×10^{-4}
Clustering medio	0.516	1.46×10^{-3}	5.163×10^{-3}
Grado medio	8.42	16.77	6
Regime	*Connected*		*Supercritical*

Table 3: Confronto tra le metriche del grafo e dei modelli ER e BA

La tabella sopra è accompagnata dal grafico sovrapposto seguente che mostra come la distribuzione del grafo si avvicini a quello BA:

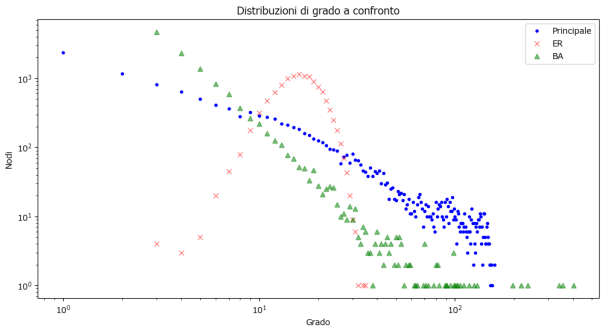

Figure 6: Confronto tra le tre distribuzioni

Si può constatare che il grafo, per quanto possa somigliare a un grafo BA, non si può prendere perfettamente a modello, complici il numero di archi e la densità nettamente minori del BA rispetto all'originale. Invece a occhio, se si sovrapponesse la medesima distribuzione pesata, si può notare che essa risulti molto più simile al BA rispetto al corrispettivo *inweighted*. Invece l'ER, sebbene alcuni valori di riferimento sono abbastanza vicini, non si avvicina minimamente alla distribuzione in analisi.

Distribuzione dei Ratio

Stabilito il criterio per la creazione degli archi si è pensato di svolgere una breve analisi anche sulla distribuzione normale dei ratio del dataset. Mettendo in relazione il ratio con la rispettiva frequenza, si possono mettere a confronto le gaussiane. Tenendo conto che si sta visualizzando nell'intervallo $[50, 100]$, la curva del Sample Ratio si mostra più "alta" e più "centrata", mentre quella dello Jaccard più bassa, omogenea, ma più dislocata a sinistra. Questo anche perché, appunto, lo Jaccard medio (57.39%) risulta nettamente inferiore del corrispettivo Sample medio (72.69%).

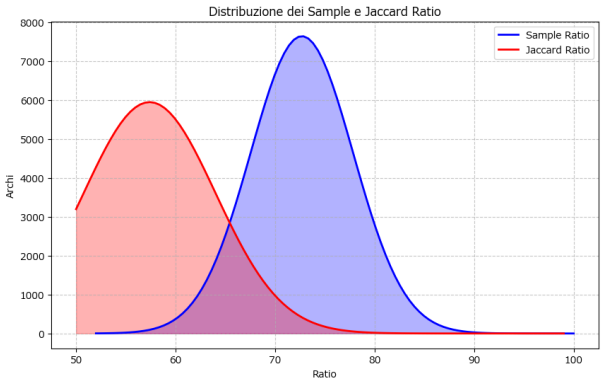

Figure 7: Distribuzione dei Sample e Jaccard Ratio

Componenti connesse

A causa delle soglie stabilite è stato previsto che il grafo in analisi sia molto frammentato: con l'analisi svolta sono state individuate 800 componenti connesse. Il componente maggiormente connesso è formato da ben 920 nodi: esso è formato da musicisti britannici o, più in generale, di molti artisti degli anni Settanta e Ottanta. Il sottografo menzionato, sia con gli archi Sample Ratio che con quelli Jaccard Ratio, è il seguente:

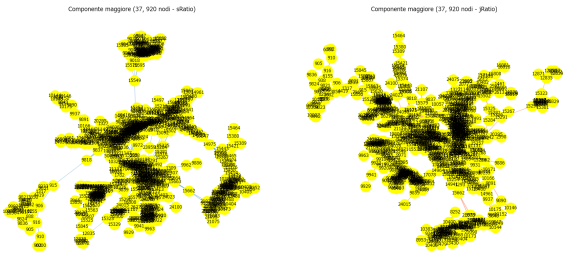

Figure 8: Sottocomponente maggiore

Path Analysis

Il numero dei sottografi è molto alto, per cui si è pensato di studiare come ne sono distribuiti i diametri. Dalla figura sotto si evince che più della metà dei sottografi sono formati da al più due nodi connessi da un singolo arco e, di conseguenza, ci sono pochi, pochissimi sottografi di diametro elevato. Il diametro medio è poco più che 2: quindi, in media, i sottografi sono formati maggiormente da 3-4 nodi:

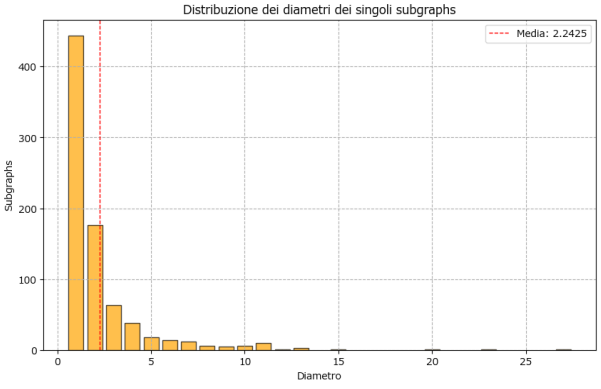

Figure 9: Distribuzione dei diametri dei subgraphs

Clustering e Densità

Il clustering medio è di 0.51646. In media, quindi, più vi sono poco più della metà di vicini collegati tra di loro, per ciascun nodo. Questo valore suggerisce che nella rete ci sono tante triadi o, comunque, gruppi piccoli di artisti; per cui gli artisti simili tendono a connettersi tra di loro in maniera abbastanza moderata.

Parallelamente a ciò, come è stato specificato in tabella, si è esaminato come sono distribuite le densità dei sottografi. In maniera speculare al grafico precedente, tutti i sottografi con diametro 1 hanno densità massima. Al contrario, maggiore è il diametro del sottografo, minore è la probabilità di avere una densità alta: per questi motivi la densità media per sottografo è decisamente più alta rispetto a quella globale (0.7713 per sottografi, 1.433×10^{-3} globale)

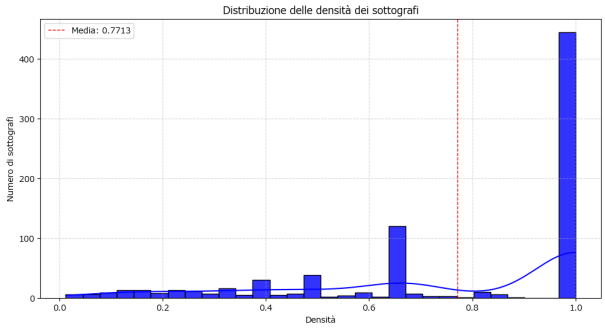

Figure 10: Distribuzione delle densità dei subgraphs

Centralities

Sono state esaminate sia la *Closeness* che la *Betweenness* Centrality. Su questi parametri salta nell'occhio che i valori massimi sono entrambi molto bassi, il che suggerisce una certa lontananza (in termini di Closeness) o di dominanza (per la Betweenness). Queste sono conseguenze nell'aver stabilito una soglia minima di Ratio relativamente centrale: infatti, la probabilità che due artisti siano legati da un ratio basso è ben maggiore.

I valori massimi di **Closeness** Centrality oscillano tra 1.347×10^{-2} e 1.429×10^{-2}: ciò suggerisce che, sebbene la soglia è relativamente bassa risulta coerente in quanto la rete è poco densa. La top 20 degli artisti per Closeness è dominata da artisti provenienti dal mondo country e country pop; inoltre può anche suggerire che all'interno del dataset gli artisti country sono collegati in maniera migliore rispetto agli altri generi musicali. Si mostra un grafico che mostra i 20 artisti con maggiore Closeness:

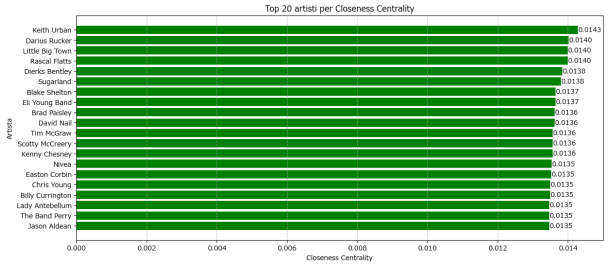

Figure 11: I primi 20 artisti per Closeness Centrality

I primi 20 valori di **Betweenness**, invece, si trovano in un range tra 8.6×10^{-4} e 2.95×10^{-3}: valori così bassi implicano una non dipendenza di artisti per mantenere la rete connessa, sia perché è formata da molte componenti connesse, sia perché il numero di archi è esiguo.

Può anche suggerire una struttura abbastanza distribuita e quindi vi sono molti percorsi alternativi. Si nota che gli artisti con maggiore Betweenneess sono meno mainstream e più eterogenei. Vi sono, infatti, artisti rock degli anni Ottanta, Classic Rock e artisti giapponesi, tra i quali di Alternative Rock/Metal; essi non sono al centro del grafo, ma fungono da ponti tra comunità diverse.

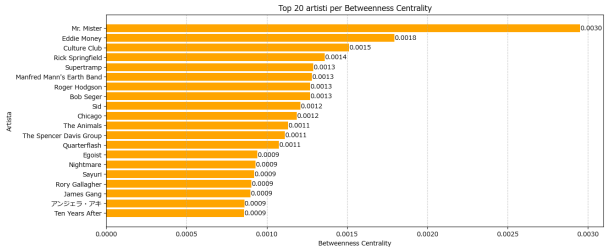

Figure 12: I primi 20 artisti per Betweenness Centrality

4 TASK: OPEN QUESTION

Come *Open Question* si è pensato di indagare su come è possibile distinguere gli artisti mainstream da quelli emergenti, in termini di centralità, densità e distribuzione delle connessioni. Inoltre si prova a prevedere se, in base ai Ratio e alla popolarità, un artista può avere la possibilità di diventare più conosciuto o, se le condizioni sono sfavorevoli, tornare poco conosciuto.

Generalità e definizione dei parametri

L'indagine comincia con il calcolo della *popolarità* di un artista. Come già specificato nella Table 1, la popolarità è stata calcolata come il rapporto tra il numero di ascolti e il numero di ascoltatori aumentato di 1 (in modo da non avere problemi di calcolo).

Per uniformare i *gap*, cioè per non avere popolarità troppo distanti l'una dall'altra e quindi per non agevolare molto gli artisti con pochi ascoltatori e molti ascolti, si è optato di utilizzare il suo logaritmo e, al fine di un riferimento con la media μ e la deviazione standard σ della popolarità, è stato calcolato lo Z-Index. Il valore è direttamente proporzionale alla distanza dalla media; il segno, invece, indica se l'artista si trova sopra/sotto la media stessa. Dalla Table 1 si riprendono le formule relative alla popolarità:

Indice	Formula
Popolarità	$pop_i = \frac{ascolti_i}{ascoltatori_i+1}$
Popolarità logaritmica	$pop_{\log,i} = \log(pop_i + 1)$
Media	$\mu = \frac{1}{N} \sum_{i=1}^{N} pop_{\log,i}$
Deviazione standard	$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (pop_{\log,i} - \mu)^2}$
Z-Index	$Z_i = \frac{pop_{\log,i} - \mu}{\sigma}$

Table 4: Formule utilizzate per la classificazione

Dai calcoli si evince che la popolarità logaritmica media, cioè il valore in cui lo Z-Index è nullo, è uguale a 2.768. I valori uscillano da 0 (quattro coppie di artisti senza ascolti, né ascoltatori) a 6.62 (Lyodra); questo vuol dire che la scala è abbastanza omogenea.

Definizione delle classi

Per poter definire se un artista è emergente, mainstream o nella via di mezzo, sono stati adottati i quantili. Se un artista ha uno Z-Index:

- minore o uguale al 15° percentile, allora è considerato *emergente*;
- maggiore o uguale all'85° percentile, allora è considerato *mainstream*;
- in tutti gli altri casi è *medio*.

Il seguente grafico mostra la dsitribuzione dello Z-Index con evidenziati i quantili di riferimento:

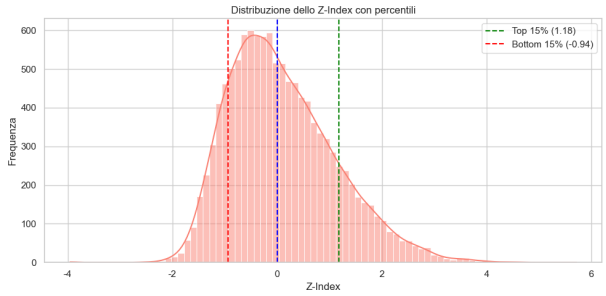

Figure 13: Distribuzione dello Z-Index con quantili

Gli artisti considerati emergenti si trovano al più *a sinistra* della retta in rosso, i mainstream al più *a destra* della retta in verde, infine i medi in mezzo le due rette citate.

Considerati questi parametri si è provato ad accostare tali classi agli archi del grafo: in questo modo si otterranno delle coppie del tipo *mainstream-mainstream*, *mainstream-medio*, ecc. Da ciò è stata creata una *heat map* che mostra come sono accoppiate le classi di artisti tra di loro:

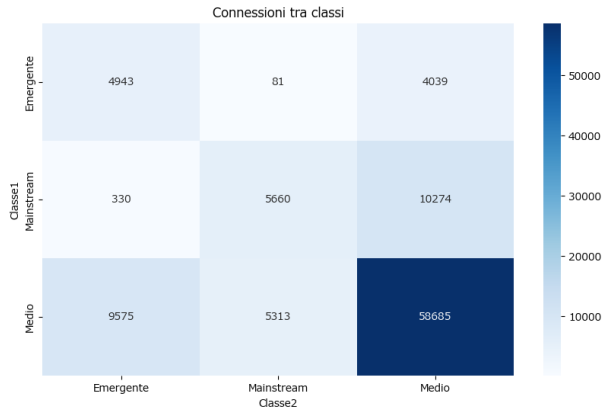

Figure 14: Connessioni tra le classi

Un dato interessante che salta all'occhio è che gli archi tra artisti mainstream ed emergenti è decisamente più basso tra gli altri ($330+81 = 411$ su 98 890, solo lo 0.41% del totale). Per cui è molto meno probabile che un emergente possa saltare direttamente al mainstream; dunque l'artista in questione dovrà necessariamente passare dalla classe intermedia.

Andando più nel dettaglio si è creato un *box plot* che mostra come sono distribuiti i Ratio in base alle classi sopra:

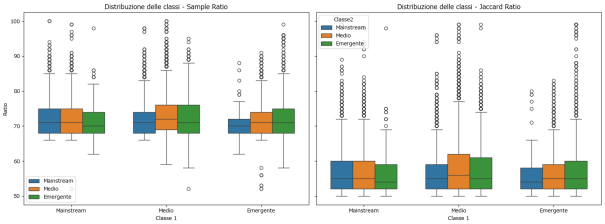

Figure 15: Distribuzione degli archi per Ratio e classe

Sul *Sample* Ratio, il numero di artisti simili condivisi (in proporzione) è relativamente omogeneo tra le classi. Tuttavia, potrebbero esserci più outlier o interazioni più diverse quando sono coinvolti artisti emergenti. Invece, lo *Jaccard* Ratio, penalizza le coppie con ampio numero di artisti simili non condivisi. Gli artisti medi condividono una frazione maggiore del loro universo simile, mentre i mainstream hanno insiemi più grandi e meno sovrapposti.

In seguito si è preso in considerazione un subgraph del grafo (si considera il primo) e si rappresentano le classi di riferimento.

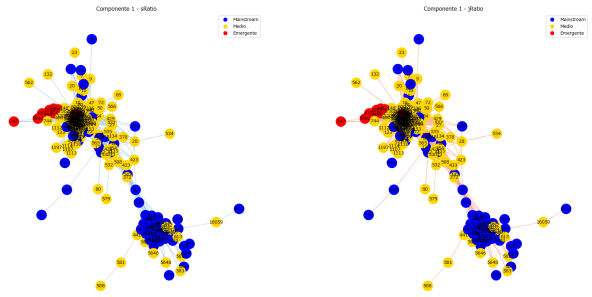

Figure 16: Subgraph 1 con classificazione

Di questo sottografo è interessante notare che esiste una community composta quasi totalmente da artisti mainstream; in più ne esiste un'altra formata per la maggiore da artisti medi. Infine, gli emergenti corrispondono a una piccola parte del *subgraph*.

Calcolo delle classi intermedie

Previsione del trend

5 DISCUSSION