



SAPIENZA
UNIVERSITÀ DI ROMA

La canzone nostra

Presentato da:
Emanuele Ruggeri
Andrea Palazzi

Data science & complexity - A.A. 2022/2023



Indice

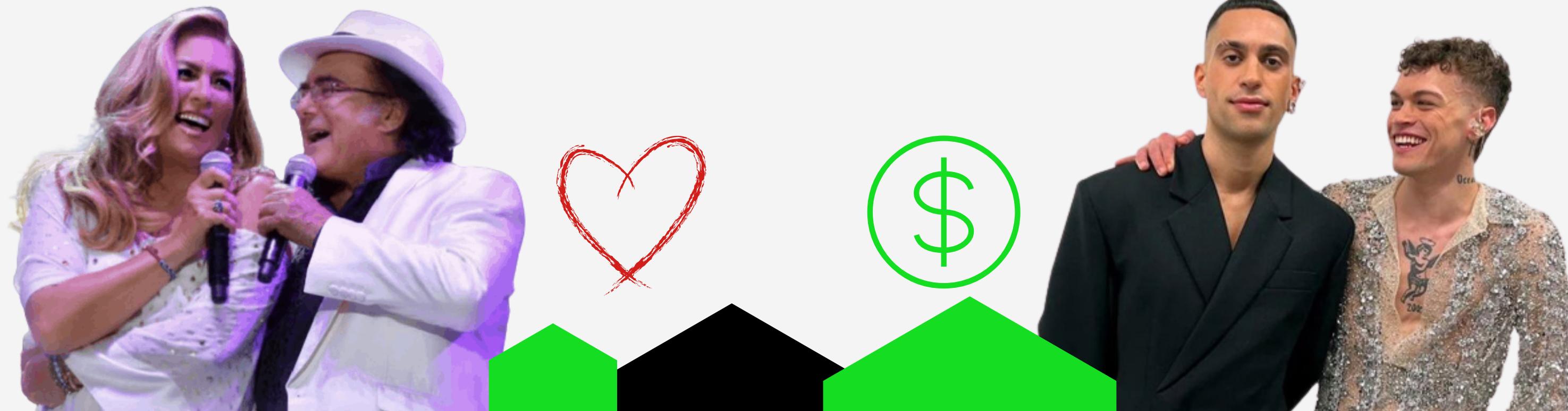


- Introduzione e domanda di ricerca
- Analisi del network
- Misure di centralità
- *Song attributes*
- Correlazioni
- Regressioni
- Analisi delle componenti principali
- Conclusioni e spunti di ricerca

Introduzione e domanda di ricerca

Le **misure di centralità** della rete, combinate con gli attributi dei brani, permettono una **previsione accurata** o generano solo **rumore** inutile?

Per un **artista emergente**, quali potrebbero essere gli **artisti migliori** con cui **collaborare** e gli **attributi** su cui lavorare per ottimizzare la **compatibilità** del *featuring*?



Il nostro punto focale

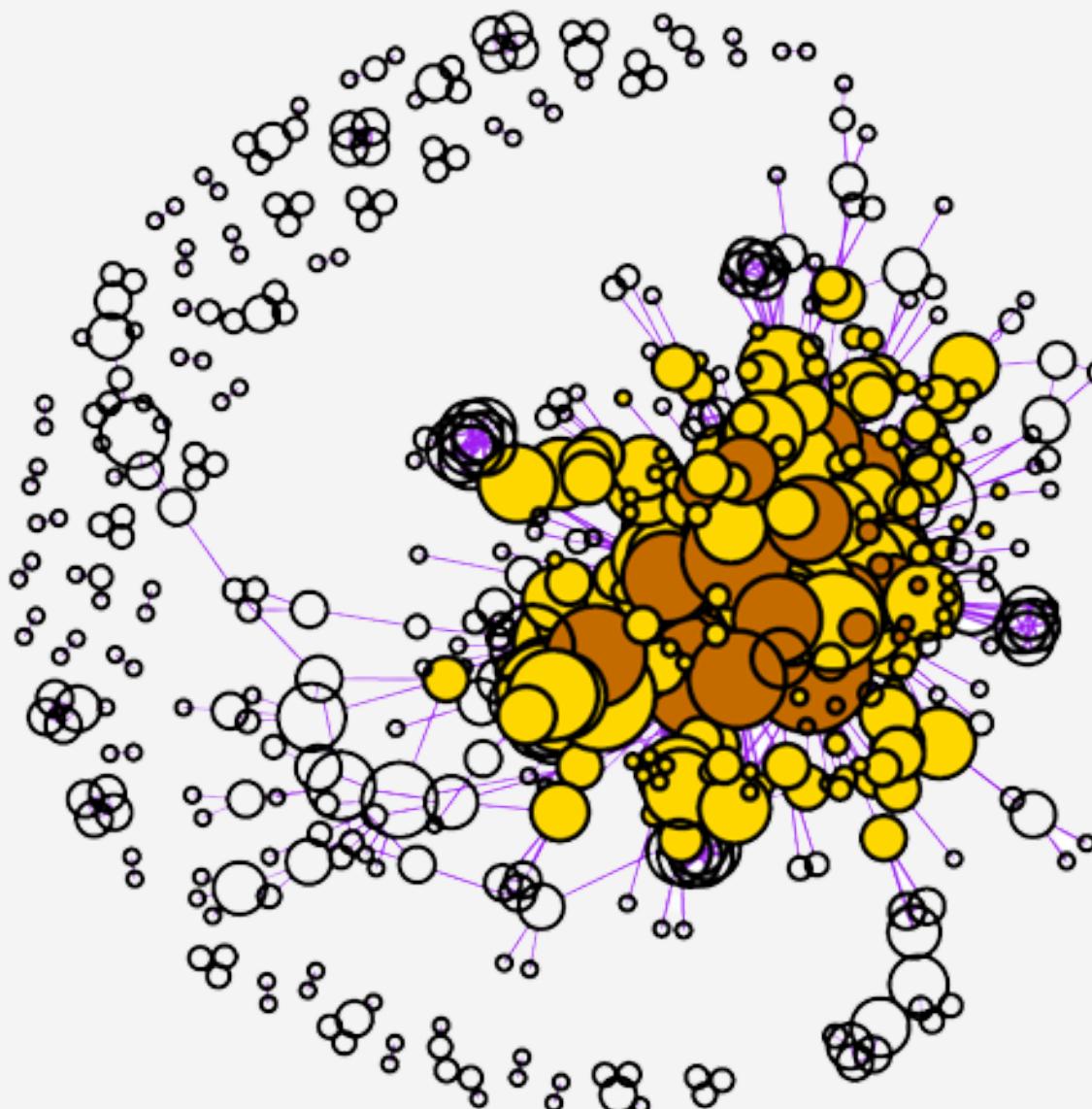
Lorenzo Lepore è un **artista emergente** romano che ha già alcune pubblicazioni all'attivo di cui una in collaborazione con un altro artista poco ascoltato.

Per il suo **prossimo singolo** la sua etichetta potrebbe dargli la possibilità di **collaborare** con **Francesco De Gregori**.

Riusciamo a prevedere come si **evolverebbe** la sua **rete** e il suo **successo** sulle piattaforme?



Analisi del network [1]



Grafo del network italiano.

Grandezza dei nodi in funzione log del grado;
colori dei nodi in funzione della distanza dal
nodo con grado più alto.

track_name
Length:1787999



A
Length:1698 B
Length:1698

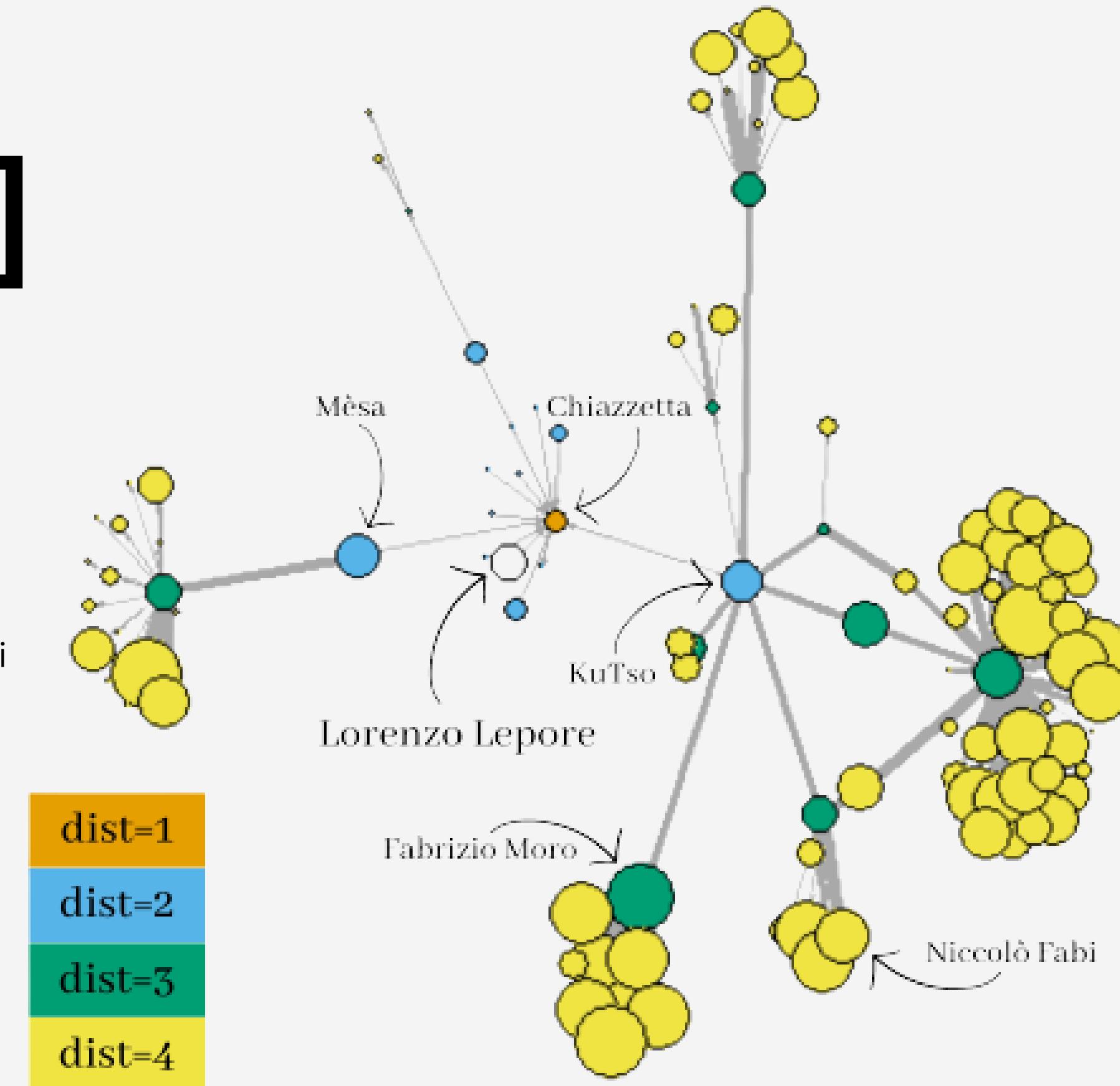


Analisi del network [2]

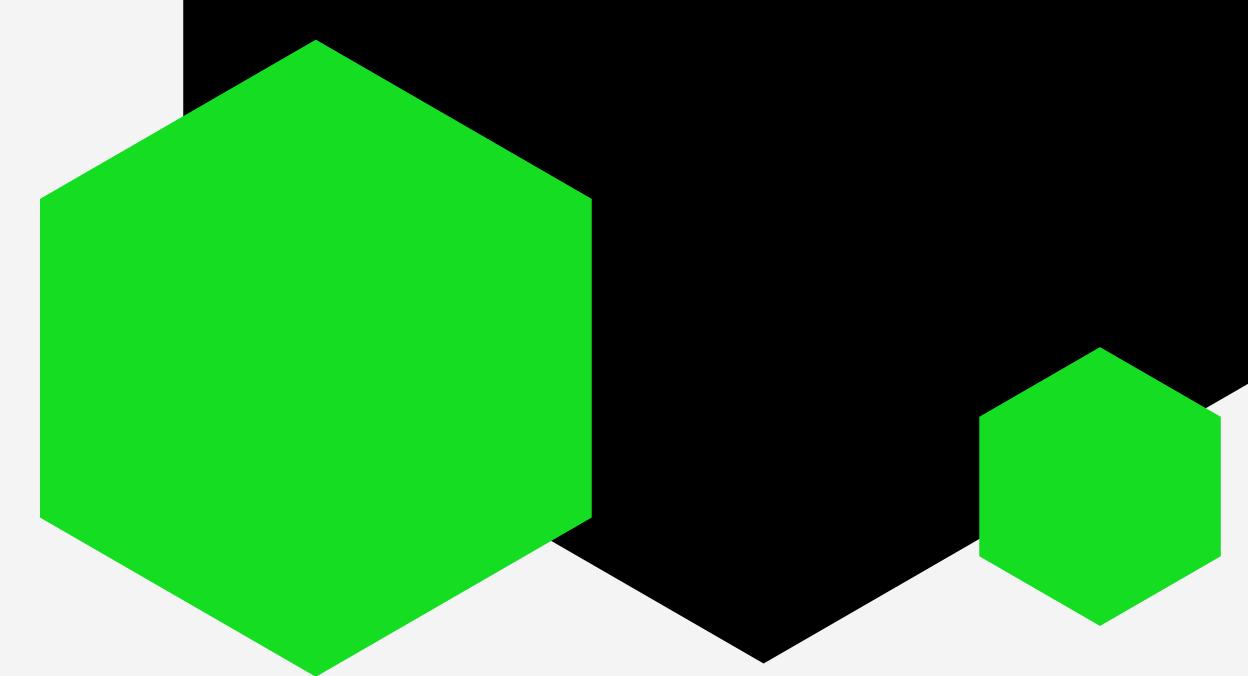
Grafo del punto focale all'istante zero

Spessore degli archi in funzione log degli ascolti

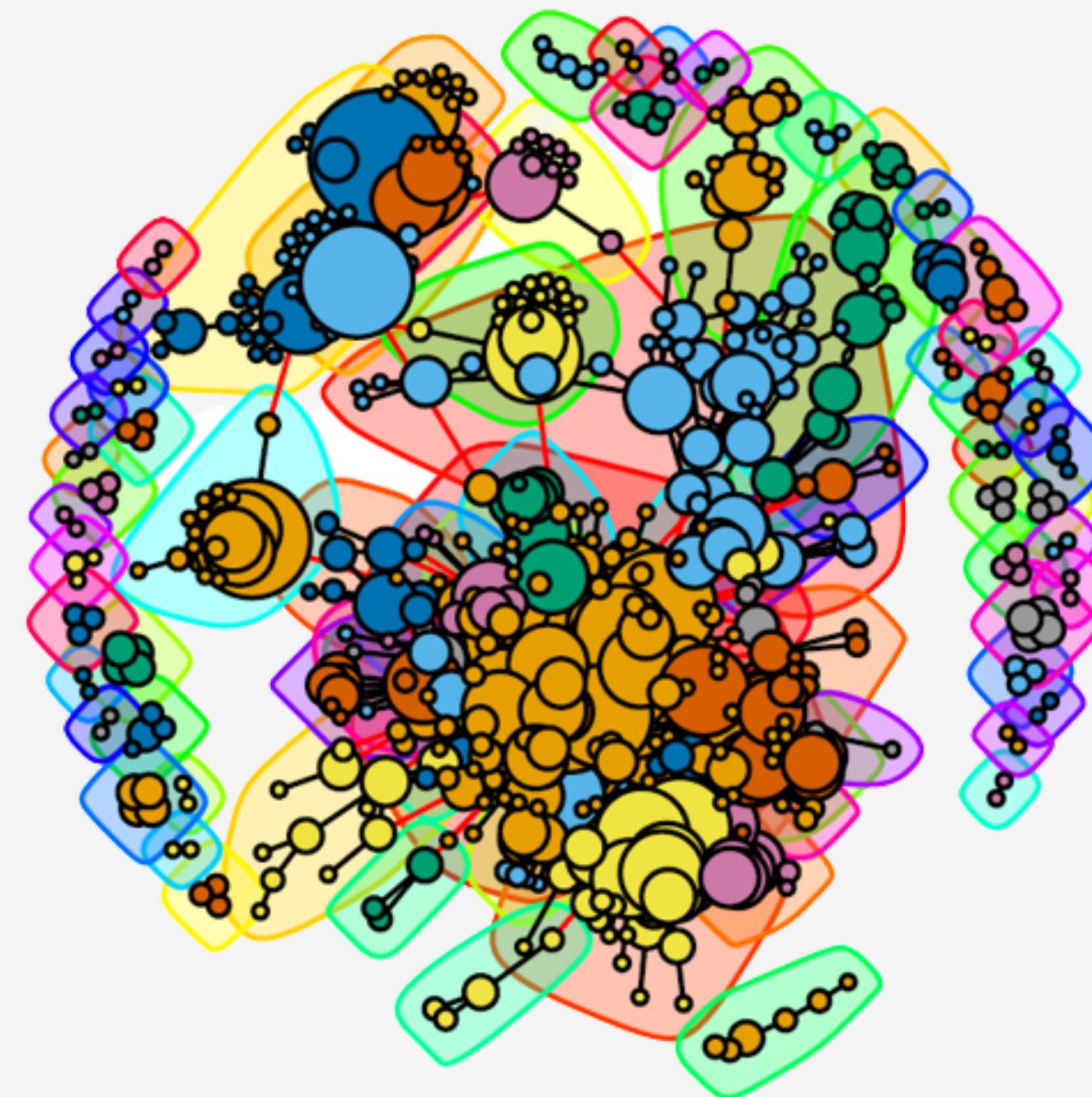
Dimensione dei nodi in funzione log degli ascoltatori mensili



Analisi del network [3]



Grafo con struttura in cluster
che massimizza la modularità

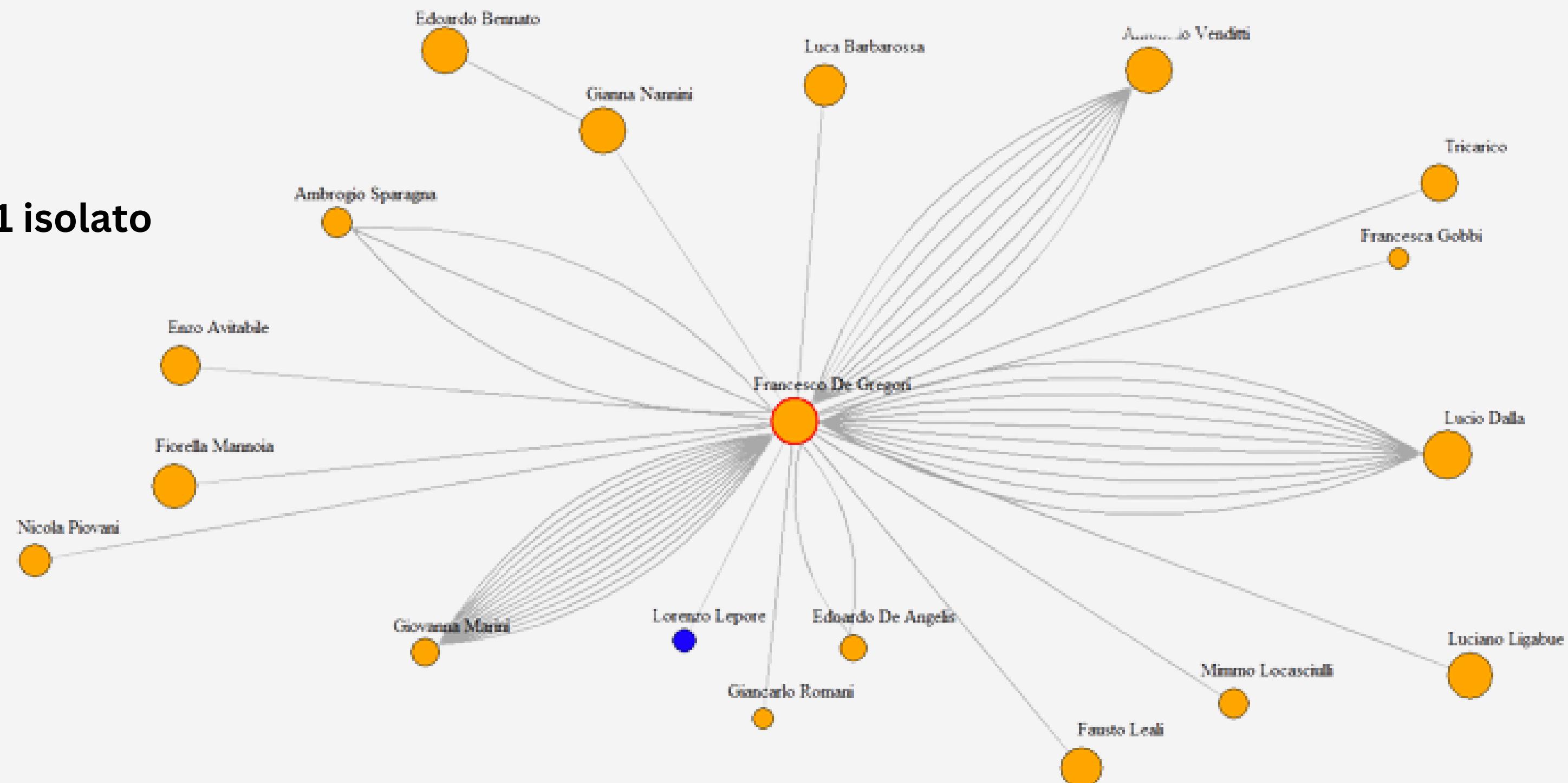


Grandezza dei nodi in funzione log del grado

Il punto focale si inserisce
all'interno del cluster 41

Analisi del network [4]

Cluster 41 isolato



Misure di centralità della rete [1]

Per ogni **nodo** del grafo sono state individuati i valori di:

- Degree

```
> mean(degree)
[1] 4.914616
```

- Betweenness

```
> mean(betweenness)
[1] 570.3046
```

- Closeness

```
> mean(closeness)
[1] 0.1112197
```



nodeId	nodeName	cMember	degree	betweenness	closeness
246	Lorenzo Lepore	41	2	3111.114	0.0003534818

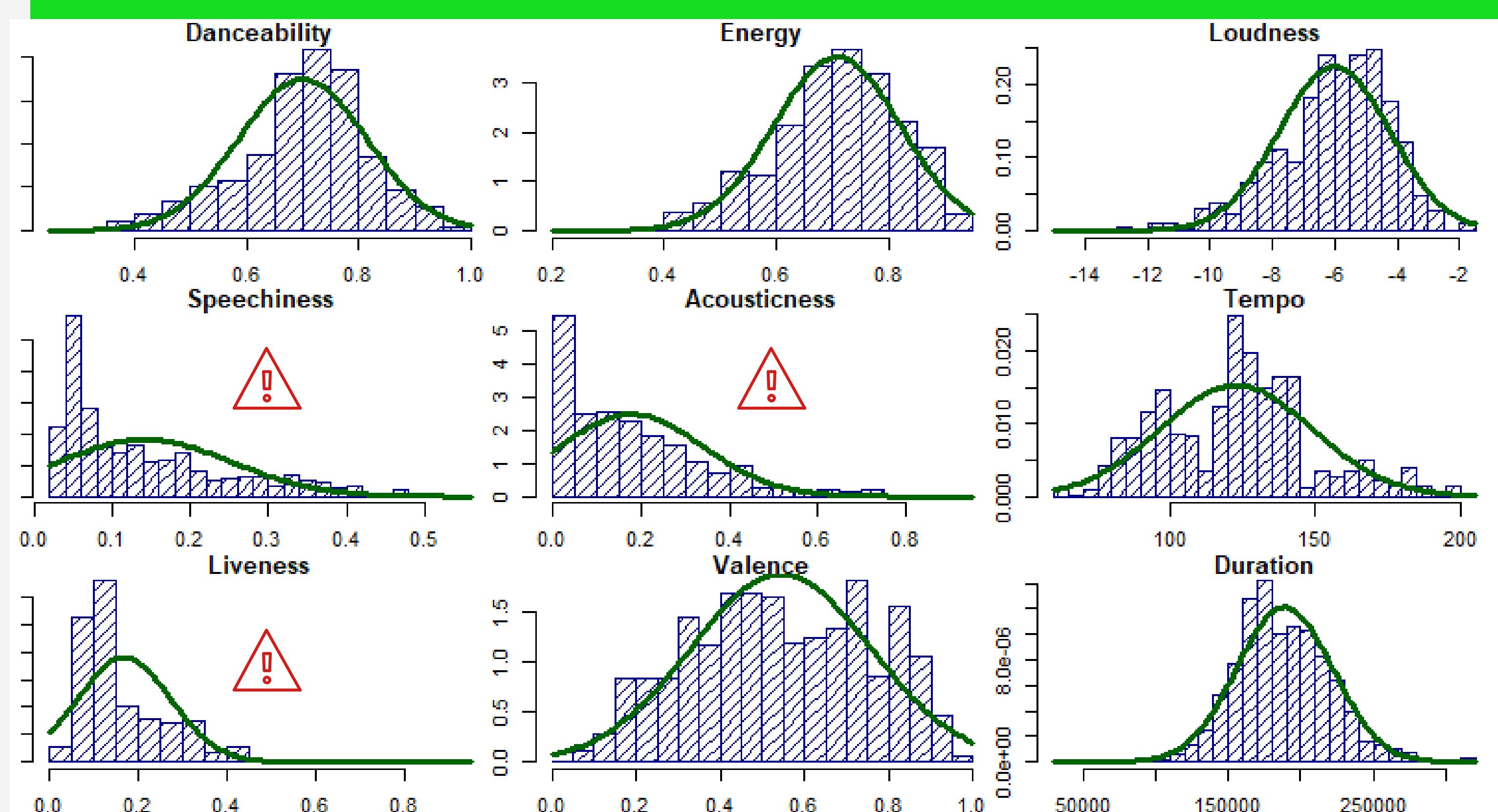
Misure di centralità della rete [2]

	asc. A	deg. A	bet. A	clos. A	asc. B	deg. B	bet. B	clos. B
ascoltatori A	1,00	-0,30	-0,14	0,22	0,68	-0,25	-0,06	0,21
degree A	-0,30	1,00	0,80	-0,32	-0,26	0,18	0,17	-0,31
betweenness A	-0,14	0,68	1,00	-0,22	-0,09	0,09	0,19	-0,21
closeness A	0,22	-0,32	-0,22	1,00	0,16	-0,24	-0,15	1,00
ascoltatori B	0,75	-0,26	-0,09	0,16	1,00	-0,20	-0,02	0,15
degree B	-0,25	0,18	0,09	-0,24	-0,20	1,00	0,75	-0,24
betweenness B	-0,06	0,17	0,19	-0,15	-0,02	0,75	1,00	-0,15
closeness B	0,21	-0,31	-0,21	1,00	0,15	-0,24	-0,15	1,00

Dalla tabella emergono delle **correlazioni forti** e delle **correlazioni deboli** tra alcune misure di centralità.

Di queste, alcune sono di **immediata spiegazione**, altre sono più interessanti.

Song attributes [1]



Alcuni attributi sono stati
scartati a priori in quanto
non pertinenti all'analisi

Song attributes [2]

Esistono già molti **studi** che cercano di mettere in **relazione** gli **attributi** delle canzoni per formulare ipotesi e **previsioni**...

	peak rank	weeks on chart	streams	danceability	energy	loudness	sp.ness	ac.ness	inst.ness	liveness	valence	tempo	duration
peak rank	1,00	-0,36	-0,29	-0,07	-0,05	-0,04	0,10	-0,01	0,02	-0,12	-0,10	-0,08	-0,13
weeks on chart	-0,36	1,00	-0,12	0,07	-0,02	0,00	-0,08	-0,04	-0,05	-0,05	0,14	0,05	0,09
streams	-0,29	-0,12	1,00	0,01	0,03	-0,01	0,21	0,00	0,01	-0,08	0,05	0,00	-0,01
danceability	-0,07	0,07	0,01	1,00	-0,04	0,01	0,06	-0,11	0,10	-0,13	0,20	-0,11	-0,17
energy	-0,05	-0,02	0,03	-0,04	1,00	0,66	-0,05	-0,32	0,03	0,09	0,40	0,02	-0,08
loudness	-0,04	0,00	-0,01	0,01	0,66	1,00	-0,23	-0,22	0,03	0,02	0,42	-0,05	-0,01
speechiness	0,10	-0,08	0,21	0,06	-0,05	-0,23	1,00	0,01	0,01	-0,04	0,00	-0,05	-0,33
acousticness	-0,01	-0,04	0,00	-0,11	-0,32	-0,22	0,01	1,00	-0,13	-0,05	-0,09	0,00	0,15
instrumentalness	0,02	-0,05	0,01	0,10	0,03	0,03	0,01	-0,13	1,00	-0,10	-0,04	0,07	-0,04
liveness	-0,12	-0,05	-0,08	-0,13	0,09	0,02	-0,04	-0,05	-0,10	1,00	0,06	-0,10	-0,10
valence	-0,10	0,14	0,05	0,20	0,40	0,42	0,00	-0,09	-0,04	0,06	1,00	0,03	-0,17
tempo	-0,08	0,05	0,00	-0,11	0,02	-0,05	-0,05	0,00	0,07	-0,10	0,03	1,00	-0,21
duration	-0,13	0,09	-0,01	-0,17	-0,08	-0,01	-0,33	0,15	-0,04	-0,10	-0,17	-0,21	1,00

Development Economics and Management Research Studies (JDMS), A Peer-reviewed International Journal, ISSN 2582 5119 (Online), 09 (11), 10 -19, January-March 2022

can also be observed in the confusion matrix and ROC Curve given above.

Conclusion and Future Work

This paper presents a methodology to predict whether a song will be popular or not using data collected from music metrics. The data was collected from Kaggle and trained and compared with the help of three classification algorithms namely, Random Forest Classifier, K-Nearest Neighbour and Linear Support Vector Classifier, that can make predictions of Song before its popularity. Among these three models, we have found that Random Forest results and accuracy which was up to 89%. Hence, we conclude that the model is good for future songs popularity predictions. This research includes an analysis of loudness, acousticness, energy, key, etc. that could help predict a song's popularity. The research also includes the popularity predictions of the song using python and song data from Spotify API through Kaggle.

The main limitation of this model is that the dataset which is used for the prediction of songs is only available on the Spotify platform. However, for future scope, we can also use data from other platforms such as YouTube, Google Play, etc. to predict the popularity of songs and also for different app platforms with help of various machine learning models.

Kokita P (2017). Predicting Popularity of Online Videos Using Machine Learning. IEEE Transactions on Multimedia. 2017;19(11):2561-2569.

Chiu D, Zhu Z (2016). Modeling Dynamics of Online Video Popularity. IEEE Transactions on Multimedia. 2016;18(9):1882-1895.

Cong G (2013). On predicting the popularity of new media content on Twitter. Journal of the American Society for Information Science and Technology. 2013;64(7):1399-1410.

Estratto di PAREEK, Prashant; SHANKAR, Poorna; SAKARIYA, N. **Predicting music popularity using machine learning algorithm and music metrics** available in spotify. J. Dev. Econ. Manag. Res. Stud. JDMS, 2022, 9: 10-19.

Correlazioni [1]

Ma è possibile affinare la previsione
aggiungendo ai dati caratteristici delle
canzoni i **dati dei nodi** di provenienza?

	asc. A	deg. A	bet. A	clos. A	asc. B	deg. B	bet. B	clos. B	peak rank	weeks on chart	streams	danceability	energy	loudness	sp.ness	ac.ness	inst.ness	liveness	valence	tempo	duration
ascoltatori A	1,00	-0,30	-0,14	0,22	0,68	-0,25	-0,06	0,21	-0,01	0,10	0,03	0,16	0,02	0,12	-0,14	-0,14	-0,02	-0,07	0,22	-0,04	0,14
degree A	-0,30	1,00	0,80	-0,32	-0,26	0,18	0,17	-0,31	-0,29	0,00	0,04	-0,07	0,11	0,10	0,18	0,11	-0,04	0,11	-0,10	0,00	0,08
betweenness A	-0,14	0,80	1,00	-0,22	-0,09	0,09	0,19	-0,21	-0,31	0,06	0,07	0,11	0,08	0,10	0,22	0,04	-0,07	0,04	-0,01	-0,01	0,01
closeness A	0,22	-0,32	-0,22	1,00	0,16	-0,24	-0,15	1,00	-0,15	0,16	0,12	-0,02	-0,03	-0,17	-0,09	-0,09	0,06	0,01	0,06	0,15	-0,04
ascoltatori B	0,68	-0,26	-0,09	0,16	1,00	-0,20	-0,02	0,15	-0,08	0,10	0,02	0,27	0,02	0,11	-0,20	-0,20	0,01	-0,05	0,23	-0,04	0,14
degree B	-0,25	0,18	0,09	-0,24	-0,20	1,00	0,75	-0,24	-0,02	-0,14	0,14	-0,03	-0,11	-0,04	0,12	0,04	0,03	-0,01	-0,20	-0,19	0,06
betweenness B	-0,06	0,17	0,19	-0,15	-0,02	0,75	1,00	-0,15	-0,14	-0,03	0,13	0,16	-0,10	0,02	0,01	0,09	-0,08	-0,09	-0,09	-0,14	0,11
closeness B	0,21	-0,31	-0,21	1,00	0,15	-0,24	-0,15	1,00	-0,16	0,16	0,12	-0,03	-0,03	-0,17	-0,10	-0,09	0,06	0,01	0,06	0,15	-0,05
peak rank	-0,01	-0,29	-0,31	-0,15	-0,08	-0,02	-0,14	-0,16	1,00	-0,36	-0,29	-0,07	-0,05	-0,04	0,10	-0,01	0,02	-0,12	-0,10	-0,08	-0,13
weeks on chart	0,10	0,00	0,06	0,16	0,10	-0,14	-0,03	0,16	-0,36	1,00	-0,12	0,07	-0,02	0,00	-0,08	-0,04	-0,05	-0,05	0,14	0,05	0,09
streams	0,03	0,04	0,07	0,12	0,02	0,14	0,13	0,12	-0,29	-0,12	1,00	0,01	0,03	-0,01	0,21	0,00	0,01	-0,08	0,05	0,00	-0,01
danceability	0,16	-0,07	0,11	-0,02	0,27	-0,03	0,16	-0,03	-0,07	0,07	0,01	1,00	-0,04	0,01	0,06	-0,11	0,10	-0,13	0,20	-0,11	-0,17
energy	0,02	0,11	0,08	-0,03	0,02	-0,11	-0,10	-0,03	-0,05	-0,02	0,03	-0,04	1,00	0,66	-0,05	-0,32	0,03	0,09	0,40	0,02	-0,08
loudness	0,12	0,10	0,10	-0,17	0,11	-0,04	0,02	-0,17	-0,04	0,00	-0,01	0,01	0,66	1,00	-0,23	-0,22	0,03	0,02	0,42	-0,05	-0,01
speechiness	-0,14	0,18	0,22	-0,09	-0,20	0,12	0,01	-0,10	0,10	-0,08	0,21	0,06	-0,05	-0,23	1,00	0,01	0,01	-0,04	0,00	-0,05	-0,33
acousticness	-0,14	0,11	0,04	-0,09	-0,20	0,04	0,09	-0,09	-0,01	-0,04	0,00	-0,11	-0,32	-0,22	0,01	1,00	-0,13	-0,05	-0,09	0,00	0,15
instrumentalness	-0,02	-0,04	-0,07	0,06	0,01	0,03	-0,08	0,06	0,02	-0,05	0,01	0,10	0,03	0,03	0,01	-0,13	1,00	-0,10	-0,04	0,07	-0,04
liveness	-0,07	0,11	0,04	0,01	-0,05	-0,01	-0,09	0,01	-0,12	-0,05	-0,08	-0,13	0,09	0,02	-0,04	-0,05	-0,10	1,00	0,06	-0,10	-0,10
valence	0,22	-0,10	-0,01	0,06	0,23	-0,20	-0,09	0,06	-0,10	0,14	0,05	0,20	0,40	0,42	0,00	-0,09	-0,04	0,06	1,00	0,03	-0,17
tempo	-0,04	0,00	-0,01	0,15	-0,04	-0,19	-0,14	0,15	-0,08	0,05	0,00	-0,11	0,02	-0,05	-0,05	0,00	0,07	-0,10	0,03	1,00	-0,21
duration	0,14	0,08	0,01	-0,04	0,14	0,06	0,11	-0,05	-0,13	0,09	-0,01	-0,17	-0,08	-0,01	-0,33	0,15	-0,04	-0,10	-0,17	-0,21	1,00

Correlazioni [2]

La matrice delle correlazioni appare evidentemente divisa in tre sezioni:

1. Dati dei **nodi**
2. Misure di **successo**
3. **Attributi** dei brani

Solo una correlazione (debole) si inquadra al di fuori di tali sezioni.

	asc. A	deg. A	bet. A	clos. A	asc. B	deg. B	bet. B	clos. B	peak rank	weeks on chart	streams	danceability	energy	loudness	sp.ness	ac.ness	inst.ness	liveness	valence	tempo	duration
ascoltatori A	1,00	-0,30	-0,14	0,22	0,68	-0,25	-0,06	0,21	-0,01	0,10	0,03	0,16	0,02	0,12	-0,14	-0,14	-0,02	-0,07	0,22	-0,04	0,14
degree A	-0,30	1,00	0,80	-0,32	-0,26	0,18	0,17	-0,31	-0,29	0,00	0,04	-0,07	0,11	0,10	0,18	0,11	-0,04	0,11	-0,10	0,00	0,08
betweenness A	-0,14	0,80	1,00	-0,22	-0,09	0,09	0,19	-0,21	-0,31	0,06	0,07	0,11	0,08	0,10	0,22	0,04	-0,07	0,04	-0,01	-0,01	0,01
closeness A	0,22	-0,32	-0,22	1,00	0,16	-0,24	-0,15	1,00	-0,15	0,16	0,12	-0,02	-0,03	-0,17	-0,09	-0,09	0,06	0,01	0,06	0,15	-0,04
ascoltatori B	0,68	-0,26	-0,09	0,16	1,00	-0,20	-0,02	0,15	-0,08	0,10	0,02	0,27	0,02	0,11	-0,20	-0,20	0,01	-0,05	0,23	-0,04	0,14
degree B	-0,25	0,18	0,09	-0,24	-0,20	1,00	0,75	-0,24	-0,02	-0,14	0,14	-0,03	-0,11	-0,04	0,12	0,04	0,03	-0,01	-0,20	-0,19	0,06
betweenness B	-0,06	0,17	0,19	-0,15	-0,02	0,75	1,00	-0,15	-0,14	-0,03	0,13	0,16	-0,10	0,02	0,01	0,09	-0,08	-0,09	-0,09	-0,14	0,11
closeness B	0,21	-0,31	-0,21	1,00	0,15	-0,24	-0,15	1,00	-0,16	0,16	0,12	-0,03	-0,03	-0,17	-0,10	-0,09	0,06	0,01	0,06	0,15	-0,05
peak rank	-0,01	-0,29	-0,31	-0,15	-0,08	-0,02	-0,14	-0,16	1,00	-0,36	-0,29	-0,07	-0,05	-0,04	0,10	-0,01	0,02	-0,12	-0,10	-0,08	-0,13
weeks on chart	0,10	0,00	0,06	0,16	0,10	-0,14	-0,03	0,16	-0,36	1,00	-0,12	0,07	-0,02	0,00	-0,08	-0,04	0,05	-0,05	0,14	0,05	0,09
streams	0,03	0,04	0,07	0,12	0,02	0,14	0,13	0,12	-0,29	-0,12	1,00	0,01	0,03	-0,01	0,21	0,00	0,01	-0,08	0,05	0,00	-0,01
danceability	0,16	-0,07	0,11	-0,02	0,27	-0,03	0,16	-0,03	-0,07	0,07	0,01	1,00	-0,04	0,01	0,06	-0,11	0,10	-0,13	0,20	-0,11	-0,17
energy	0,02	0,11	0,08	-0,03	0,02	-0,11	-0,10	-0,03	-0,05	-0,02	0,03	-0,04	1,00	0,66	-0,05	-0,32	0,03	0,09	0,40	0,02	-0,08
loudness	0,12	0,10	0,10	-0,17	0,11	-0,04	0,02	-0,17	-0,04	0,00	-0,01	0,01	0,66	1,00	-0,23	-0,22	0,03	0,02	0,42	-0,05	-0,01
speechiness	-0,14	0,18	0,22	-0,09	-0,20	0,12	0,01	-0,10	0,10	-0,08	0,21	0,06	-0,05	-0,23	1,00	0,01	-0,04	0,00	-0,05	-0,33	
acousticness	-0,14	0,11	0,04	-0,09	-0,20	0,04	0,09	-0,09	-0,01	-0,04	0,00	-0,11	-0,32	-0,22	0,01	1,00	-0,13	-0,05	-0,09	0,00	0,15
instrumentalness	-0,02	-0,04	-0,07	0,06	0,01	0,03	-0,08	0,06	0,02	-0,05	0,01	0,10	0,03	0,03	0,01	-0,13	1,00	-0,10	-0,04	0,07	-0,04
liveness	-0,07	0,11	0,04	0,01	-0,05	-0,01	-0,09	0,01	-0,12	-0,05	-0,08	-0,13	0,09	0,02	-0,04	-0,05	-0,10	1,00	0,06	-0,10	-0,10
valence	0,22	-0,10	-0,01	0,06	0,23	-0,20	-0,09	0,06	-0,10	0,14	0,05	0,20	0,40	0,42	0,00	-0,09	-0,04	0,06	1,00	0,03	-0,17
tempo	-0,04	0,00	-0,01	0,15	-0,04	-0,19	-0,14	0,15	-0,08	0,05	0,00	-0,11	0,02	-0,05	-0,05	0,00	0,07	-0,10	0,03	1,00	-0,21
duration	0,14	0,08	0,01	-0,04	0,14	0,06	0,11	-0,05	-0,13	0,09	-0,01	-0,17	-0,08	-0,01	-0,01	-0,33	0,15	-0,04	-0,10	-0,17	1,00

Correlazioni [3]

Appare subito evidente che la domanda posta in precedenza ha una **risposta** piuttosto chiara e deludente.

Con i **dati** a nostra **disposizione** non è possibile aggiungere precisione alla previsione, anzi andremmo solamente ad aggiungere **rumore**.

Provando una **regressione lineare multipla** su un campione, infatti, osserviamo le seguenti percentuali di **accuratezza*** delle previsioni:

- Utilizzando solo dati sugli **attributi** dei brani: **78%**
- Combinando i dati sugli **attributi** e i dati dei **nodi**: **27%**

L'**analisi** dovrà a questo punto **escluderà** la sezione degli **attributi** per **concentrarsi** esclusivamente sulla sezione dei **nodi** e delle **misure di successo**.

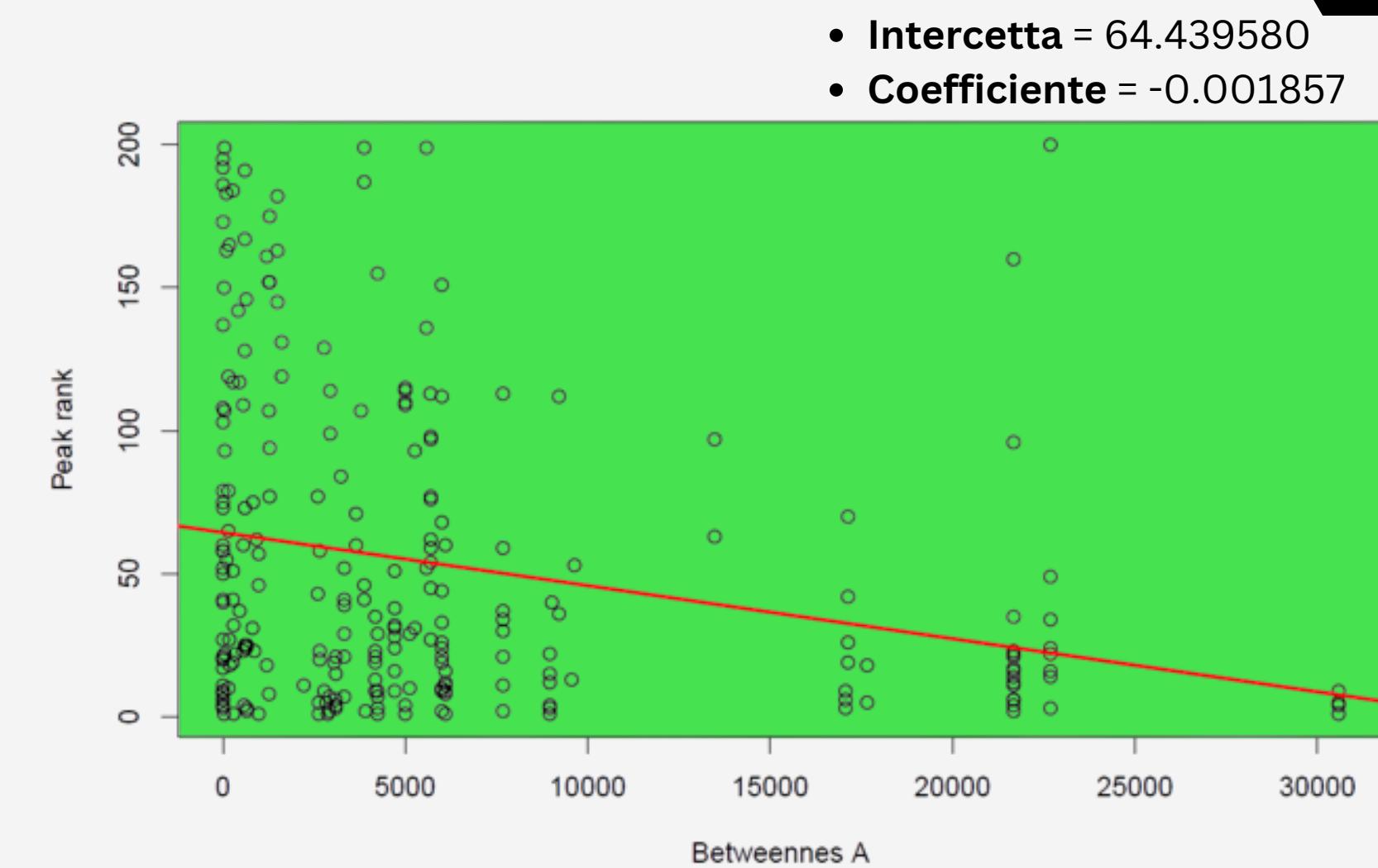
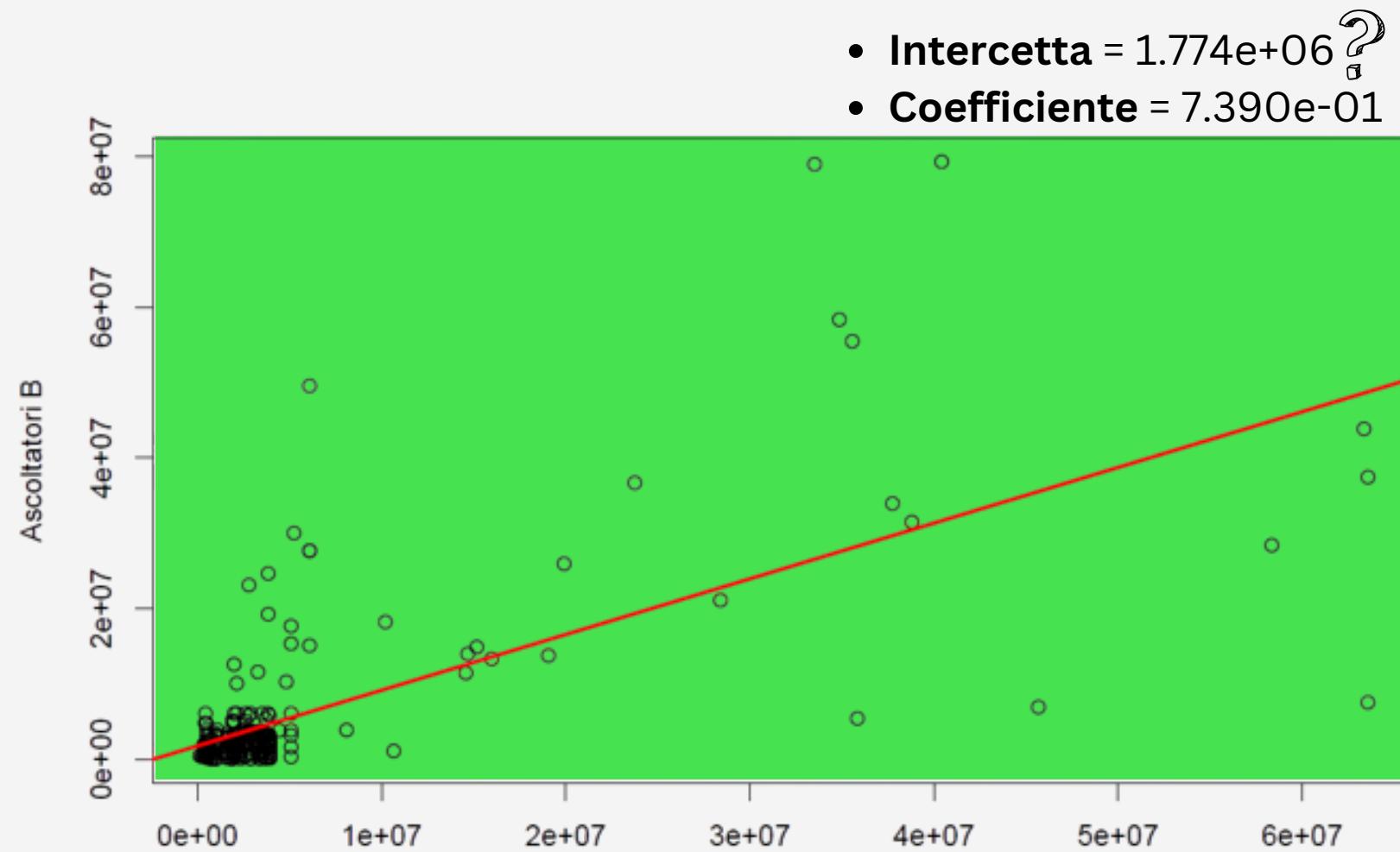


Frame estratto dal videoclip ufficiale
del brano "Malincomia"

*% di previsioni che centrano un intorno del 5% dal dato fattuale

Regressioni [1]

Proviamo a dare un valore ad alcune delle **correlazioni evidenziate** del quadrante dei nodi.



Siamo quindi riusciti a **quantificare** le due correlazioni più interessanti inquadrate all'interno della matrice.

Tuttavia, dagli scatterplot (e non solo) possiamo notare una forte **asimmetria** nelle distribuzioni dei valori delle variabili, con conseguenze sull'accuratezza delle previsioni e l'**interpretazione** dei risultati.

Regressioni [2]

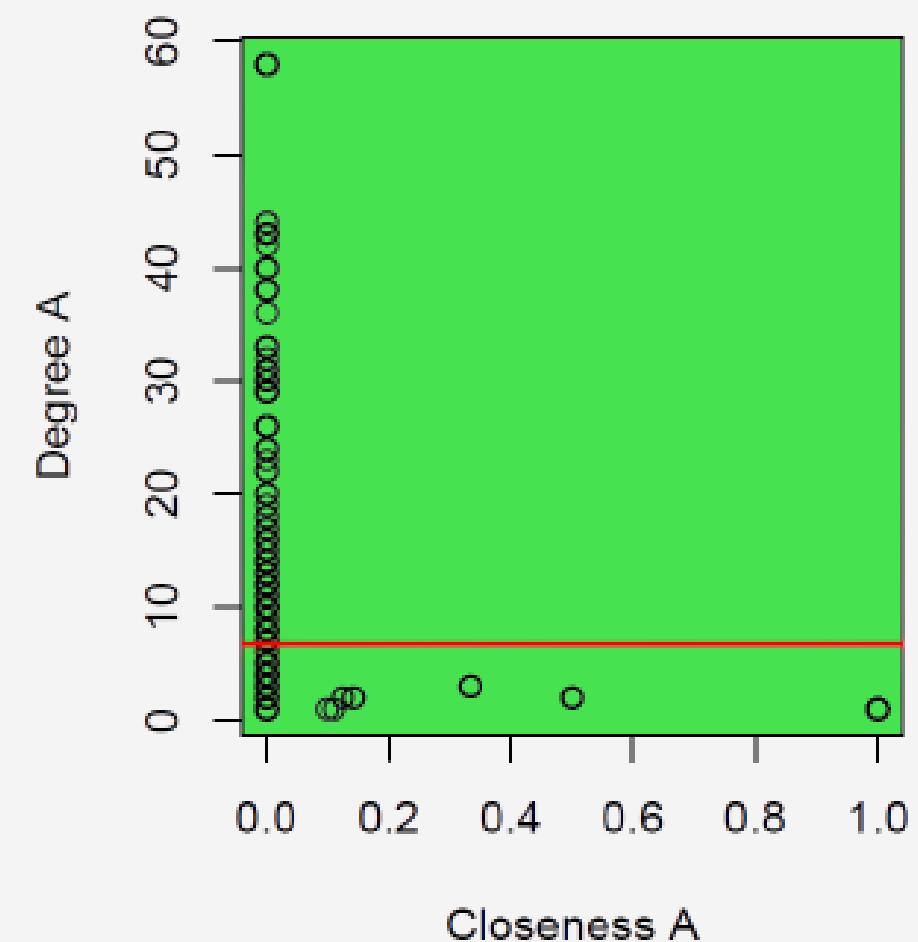
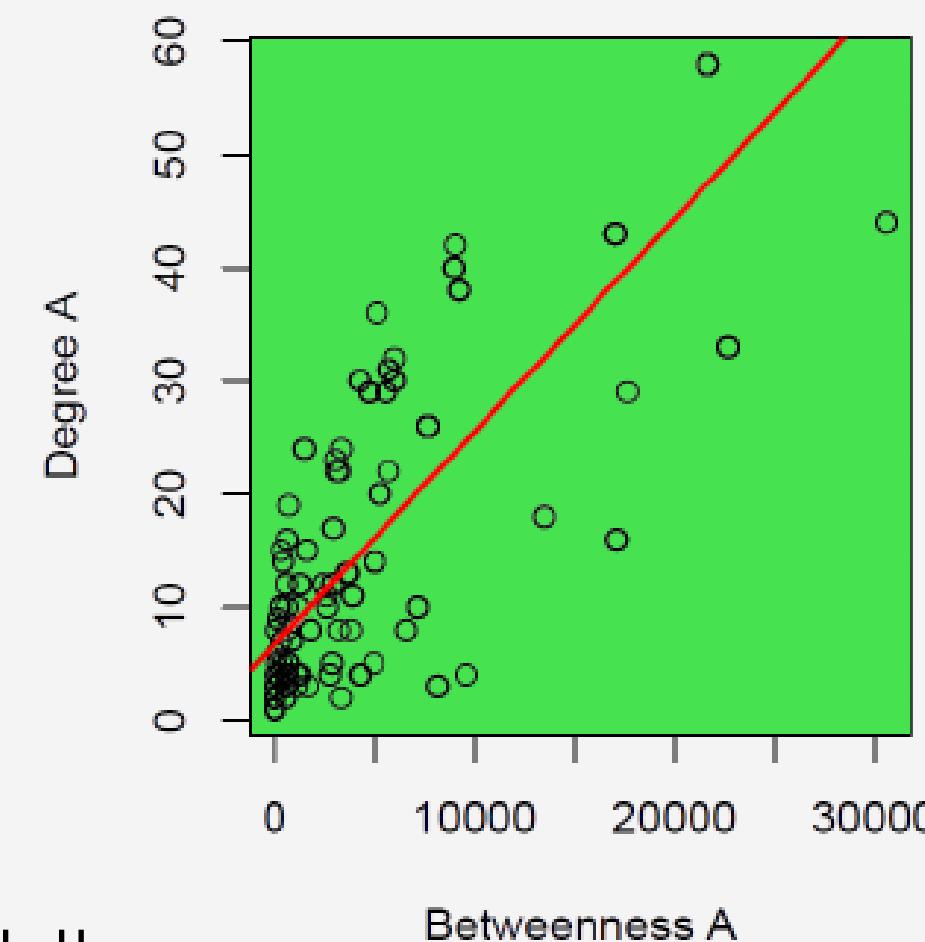
Dalla matrice delle correlazioni notiamo che la variabile ***degree A*** è correlata alle variabili ***betweenness A***, ***closeness A*** e ***closeness B***; questo potrebbe spingerci verso l'impostazione di una **regressione lineare multipla**.

Le variabili ***closeness A*** e ***closeness B*** sono **strettamente correlate** (anche per definizione) e di conseguenza consideriamo solo la prima.



Il **problema**, tuttavia, risiede ancora una volta nella fortissima **asimmetria** nelle distribuzioni dei valori.

Non è in alcun modo **possibile** svolgere delle analisi attendibili in questa direzione (**accuratezza** della previsione = **6%**).



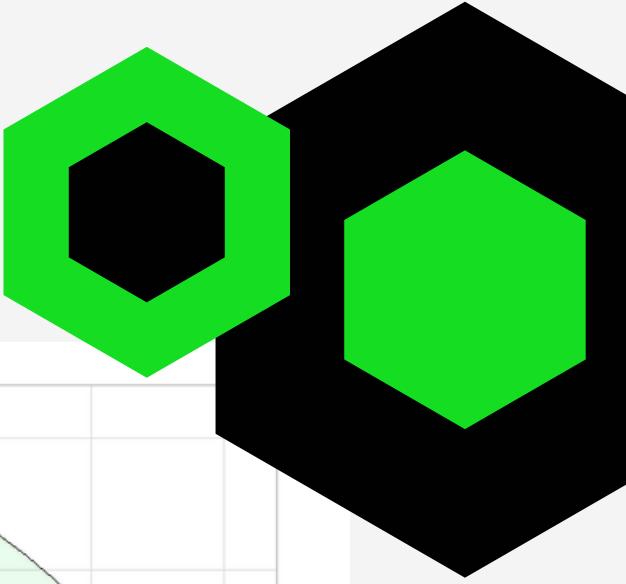
Risultati?

Le **misure di centralità** della rete, combinate con gli attributi dei brani, permettono una **previsione accurata** o generano solo **rumore** inutile?

Per un **artista emergente**, quali potrebbero essere gli **artisti migliori** con cui **collaborare** e gli **attributi** su cui lavorare per ottimizzare la **compatibilità** del *featuring*?

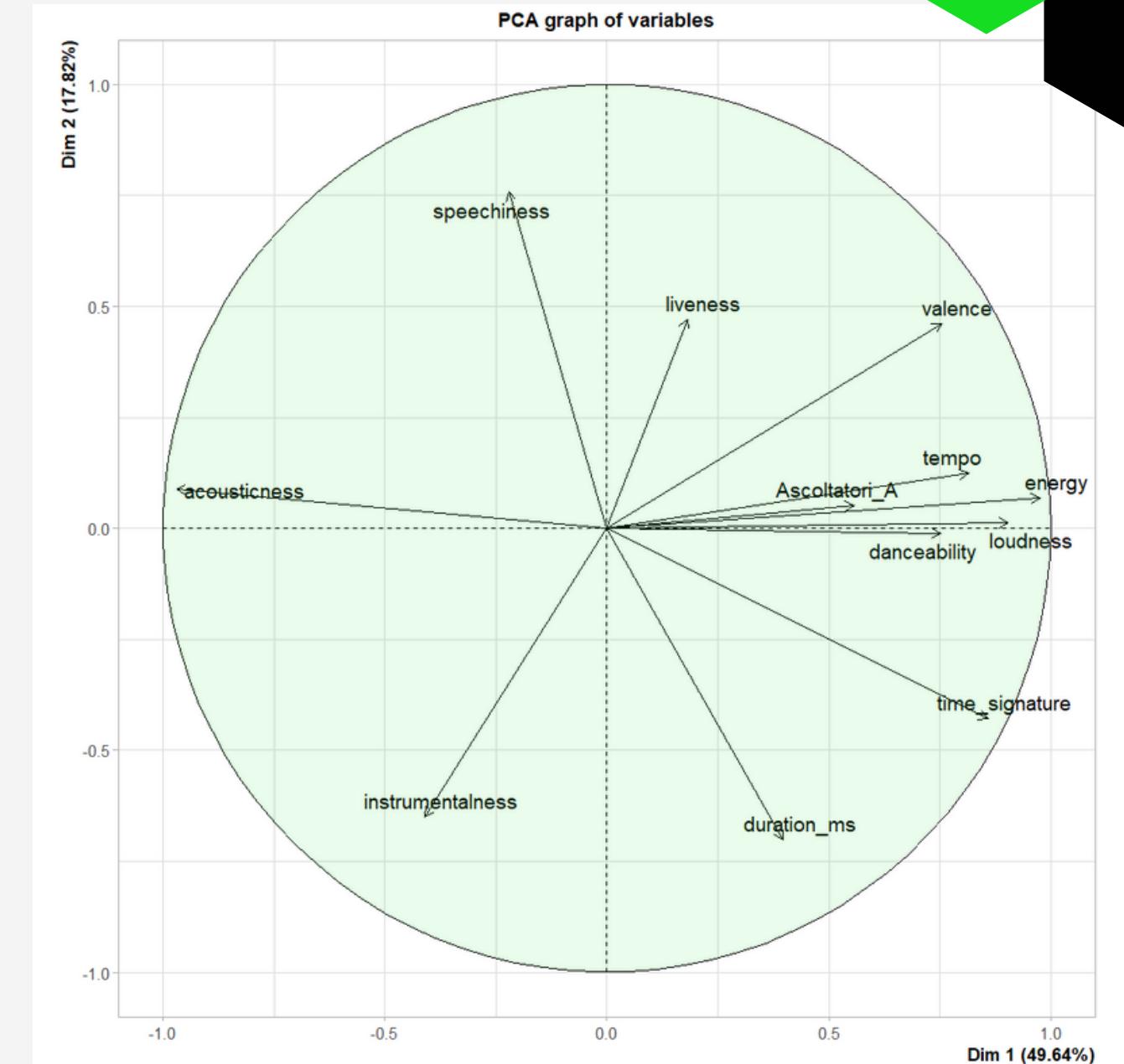


PCA [1]

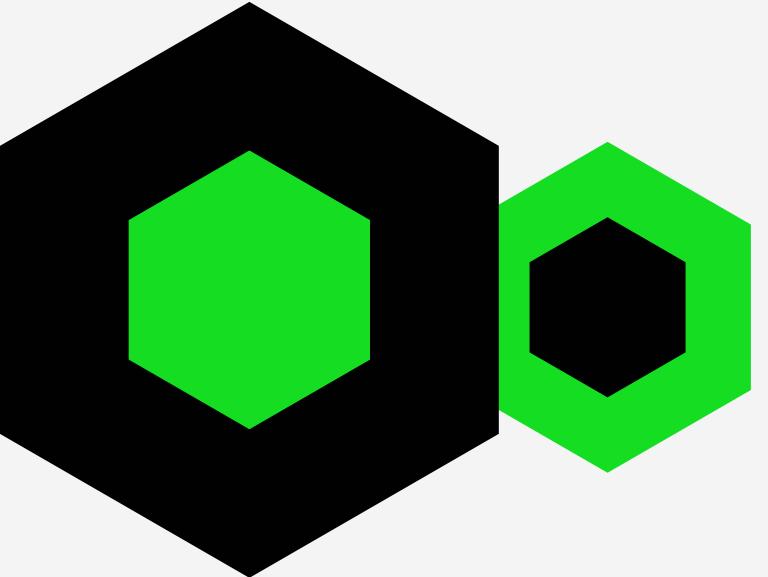


L'analisi si basa su un dataset che mette in risalto gli **attributi** tipici dei **brani** realizzati dagli artisti inquadrati nel **cluster 41**.

	Commerciabilità	Incombenza	Fedeltà
	Dim.1	Dim.2	Dim.3
danceability	0.7594734	0.02733826	-0.50159196
energy	0.9746265	0.07105890	0.08622700
loudness	0.9034735	0.02492198	-0.04276009
speechiness	-0.2274784	0.76335149	-0.14558875
acousticness	-0.9690526	0.08630714	-0.08660399
instrumentalness	-0.3975351	-0.64280248	-0.19639091
liveness	0.1668523	0.43053135	0.73599362
valence	0.7558137	0.48396596	-0.26636034
tempo	0.8208516	0.13982708	-0.15214098
time_signature	0.8622348	-0.42022003	-0.05956000
duration_ms	0.3907106	-0.71386897	0.21287680
Ascoltatori_mensili	0.4204806	-0.05219110	0.74202737
	alta	indifferente	alta



Dalle componenti principali risulta che si desiderano **alti valori** per la **prima componente** e **alti valori** nella **terza**.

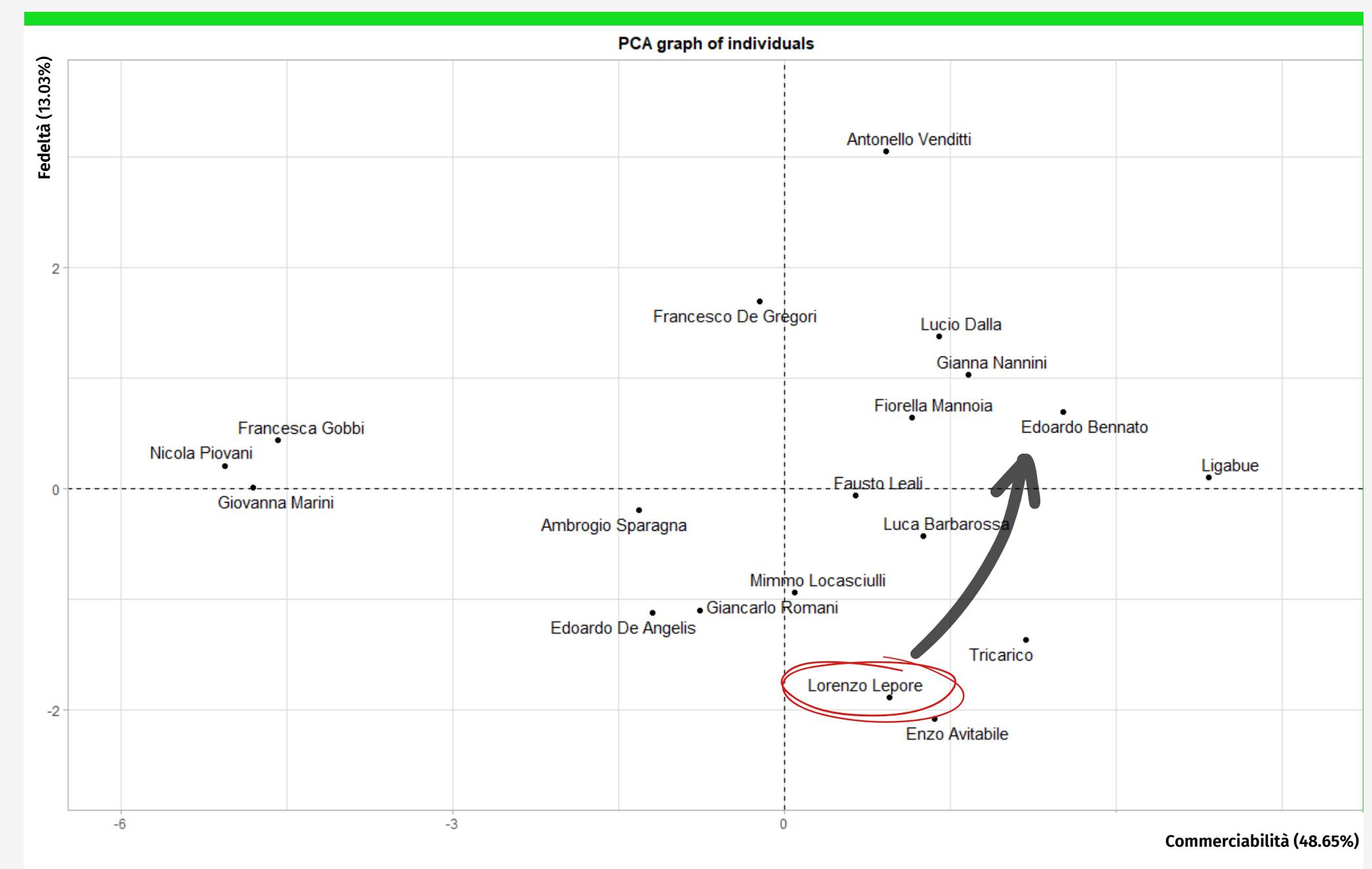


PCA [2]

Dal grafico degli individui con la **prima e terza componente** principale, risulta che i **migliori artisti** con cui collaborare (in riferimento al punto focale) sono presenti nel **primo quadrante**.

In particolare Antonello Venditti, Gianna Nannini, Fiorella Mannoia, Edoardo Bennato e Ligabue.

Abbiamo dunque trovato una sorta di **risposta** per il **caso Lorenzo Lepore**, identificando gli **attributi** sui quali puntare maggiormente per creare una **compatibilità** con gli artisti del quadrante, afferenti a **dimensioni positive**.



Future linee di ricerca

L'attuale ricerca condotta non ci ha permesso di giungere a risultati o indicazioni attendibili: l'esplorazione del **dataset** e della struttura della **rete** del mercato discografico italiano da sole **non risultano sufficienti**.

Facendo leva sull'**opinione diffusa** dalla "dottrina" secondo la quale una **collaborazione** porta quasi sempre dei **benefici**, sarebbe opportuno inquadrare e soprattutto valutare **altri parametri** che attengono...

- alla sfera del **marketing**, come l'immagine e la promozione combinata
- ai **booster**, che abbiamo definito come tutti **picchi di visibilità** dati da importanti nodi e/o eventi
- alla **viralità**, intendendola come capacità di essere accostati ad un top trend e sfruttare l'**ascolto accessorio** e non protagonista

Per ultimo, è importante dare dei **confini** di tipo quantitativo al **concetto di successo** ed, eventualmente, **identificare** delle **varianti** che possono essere perseguitate singolarmente o congiuntamente dai management.

Grazie
per l'attenzione!



Emanuele Ruggeri



Andrea Palazzi



SAPIENZA
UNIVERSITÀ DI ROMA