

A dive Into the music industry

A network tour of data science 2019

Sacha Leblanc, Etienne Caquot, Grégoire Mayrhofer, Alexis Mermet

A data story about the music industry


Everybody knows the big interprets in today's music landscape. But only few knows about the people helping them to create their bigger hits.

- ❑ Who are these persons behind the scenes?
- ❑ How are they connected one with another?
- ❑ What is the nature of these connections?



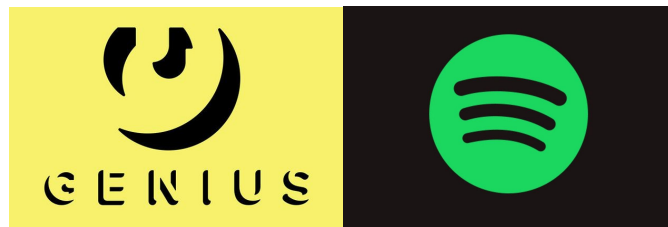
The project's goal

During this project we looked into multiple aspects of the musical industry using graph analysis methods.

- How are the big names of the industry connected with each others?
 - Genre analysis (Can we discern the genre in the dataset? Are they connected?)
 - Sub-Genre analysis (ex. can we find different genre of rap music?)
 - Does relationships have an influence on success?
- 

Creating the dataset

- ❑ We first download Spotify data from Kaggle: [Spotify Tracks DB](#)
- ❑ For each song in the dataset above:
 - ❑ Pull data from the Genius API
 - ❑ Obtain songwriters information
 - ❑ Obtain producers information
 - ❑ Obtain recording labels information



	track_name	artist_name	songwriters	producers	labels	genre	popularity
0	Be Without You	mary j. blige	[bryan michael cox, jason perry, johnt austin...]	[bryan michael cox]	[geffen records]	R&B	65
1	Desperado	rihanna	[banks, d. rachel, james fauntleroy, mick schu...]	[james fauntleroy, kuk harrell, mick schultz]	[westbury road, roc nation]	R&B	63

Data exploration (1)

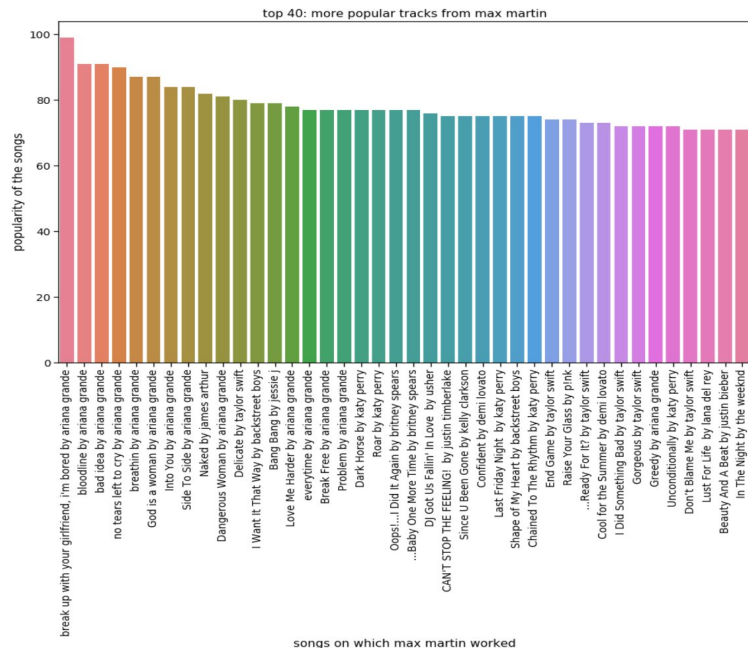
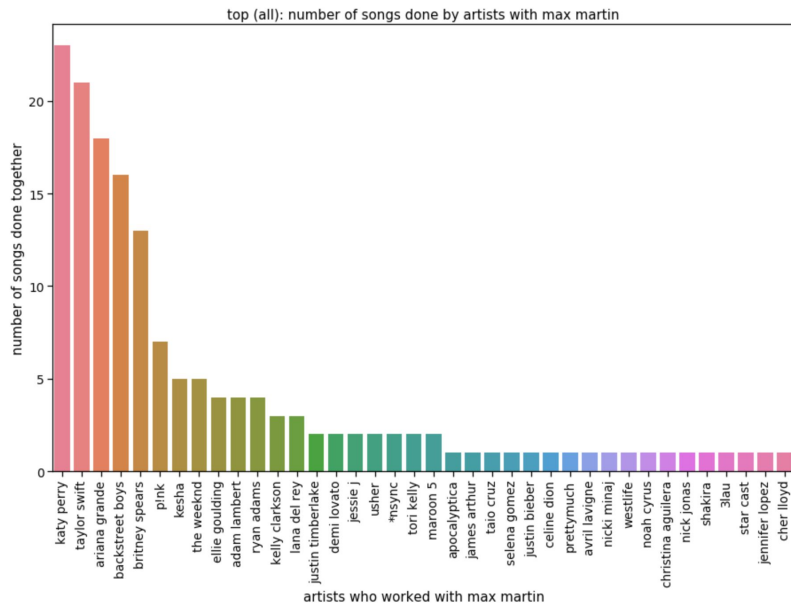
SOME NUMBERS ABOUT OUR DATASET:

- **Songs:** 70681
- **Genres:** 12 unique genres (Alternative, Blues, Country, Hip-Hop, Dance, Electronic, Folk, Indie, Pop, RNB, Rap, Reggae, Reggaeton)
- **Actors:** 22894
 - ◆ artists (interprets)
 - ◆ songwriters
 - ◆ producers
- **Popularity:** number from 0 to 100



Data exploration (2)

Let's look into one of the biggest songwriters of our dataset: **MAX MARTIN**



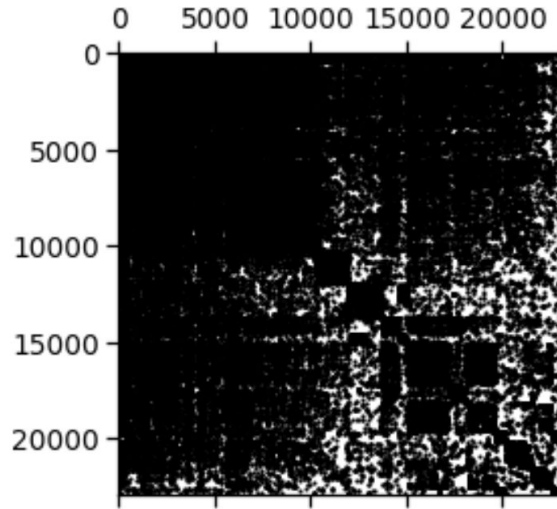
Create the adjacency matrix(1)

We created adjacency matrices in two different ways.

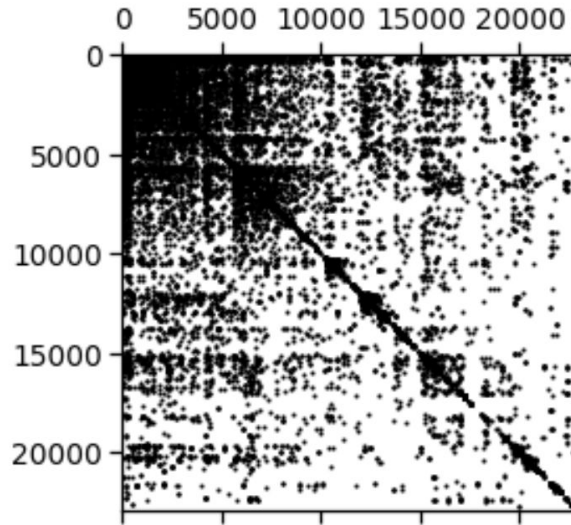
1. **Full collaborations matrix**: each entries in the matrix corresponds to the number of collaborations between any 2 actors.
2. **Strong collaborations matrix**: each entries in the matrix are set to one if 2 persons worked more than 3 times with each other. Else it is set to 0.



Create the adjacency matrix(2)



Full collaborations matrix

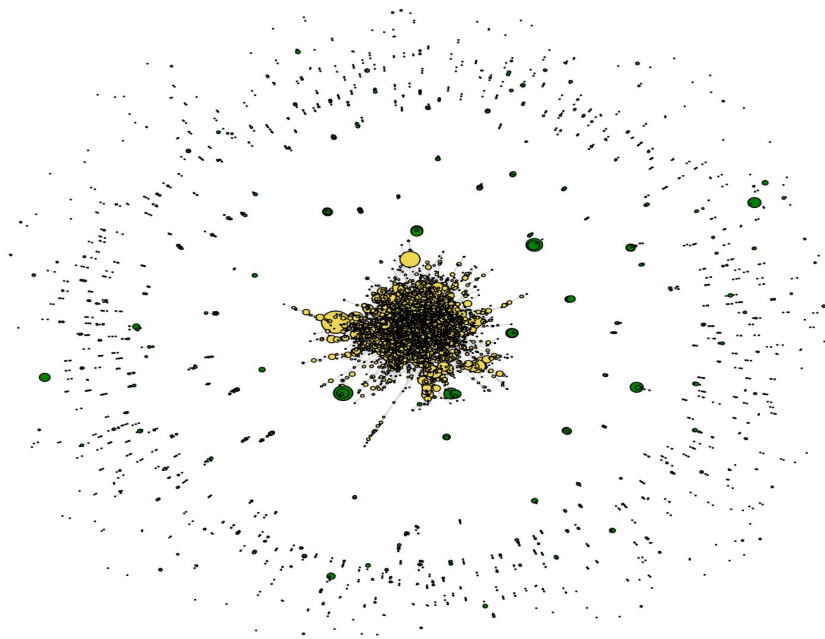


Strong collaborations matrix

Graph analysis: Visualization (1)

Adjacency matrix: created with the first method

data: 20% of total dataset

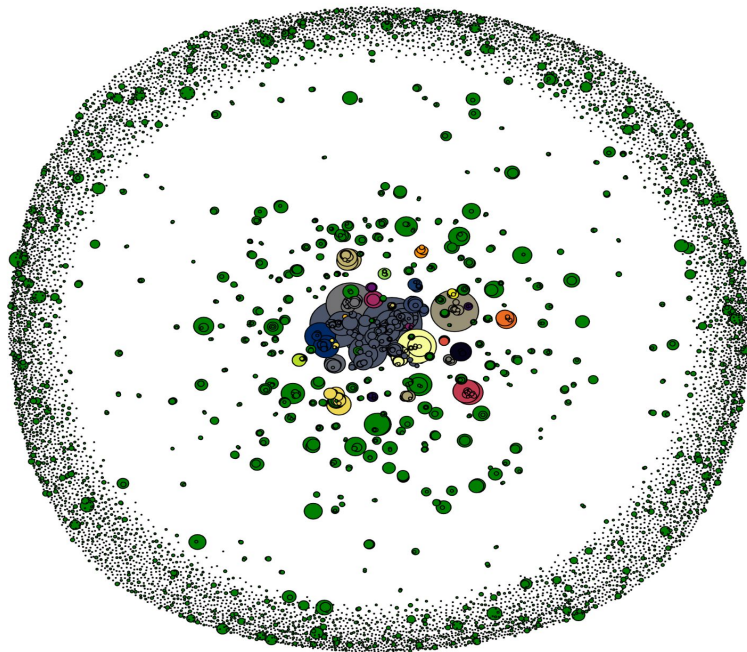


DEMO!

Graph analysis: Visualization (2)

Adjacency matrix: created with the second method

data: 20% of total dataset



DEMO!

The clustering algorithms

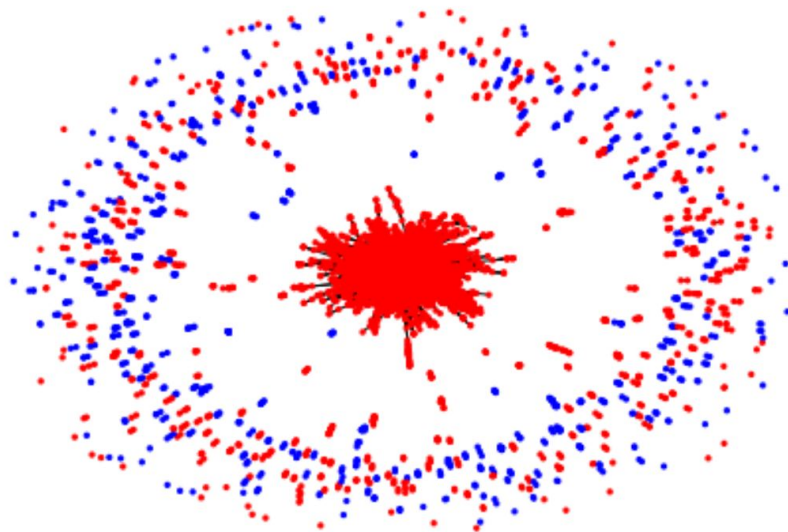
- ❑ **Spectral Clustering**: Uses information from the eigenvalues of special matrices to obtain the community of each node.
- ❑ **DBSCAN**: Given a set of points in some space, it groups together points that are closely packed together.
- ❑ **Louvain method**: Extract communities from large networks by optimization of the modularity value (which measures the strength of division of a network into modules).



Graph analysis: a world separated in two

Adjacency matrix: created with the first method

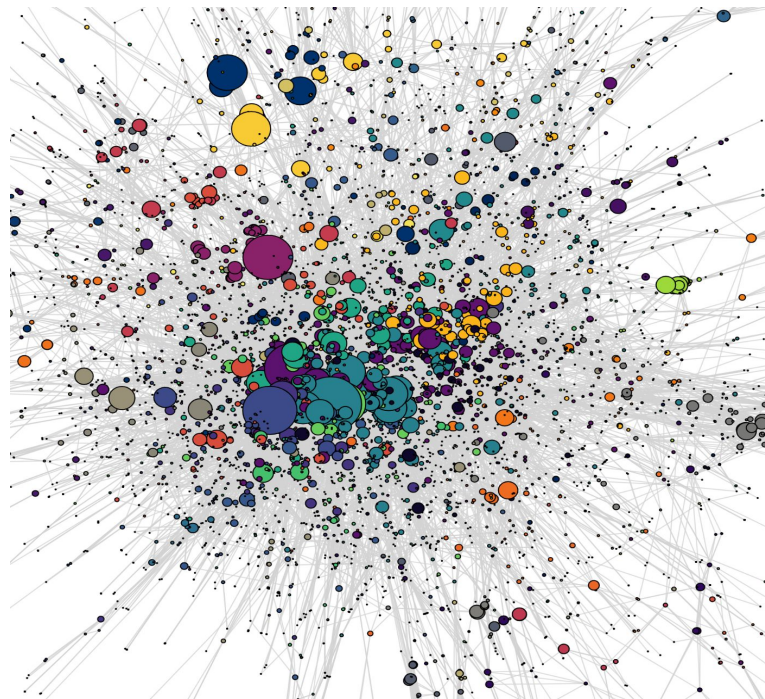
data: 20% of total dataset



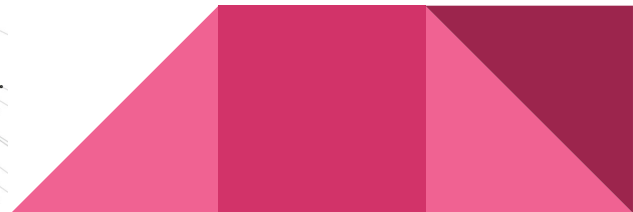
Graph analysis: Musical Genre detection (1)

Adjacency matrix: created with the first method

data: 20% of total dataset

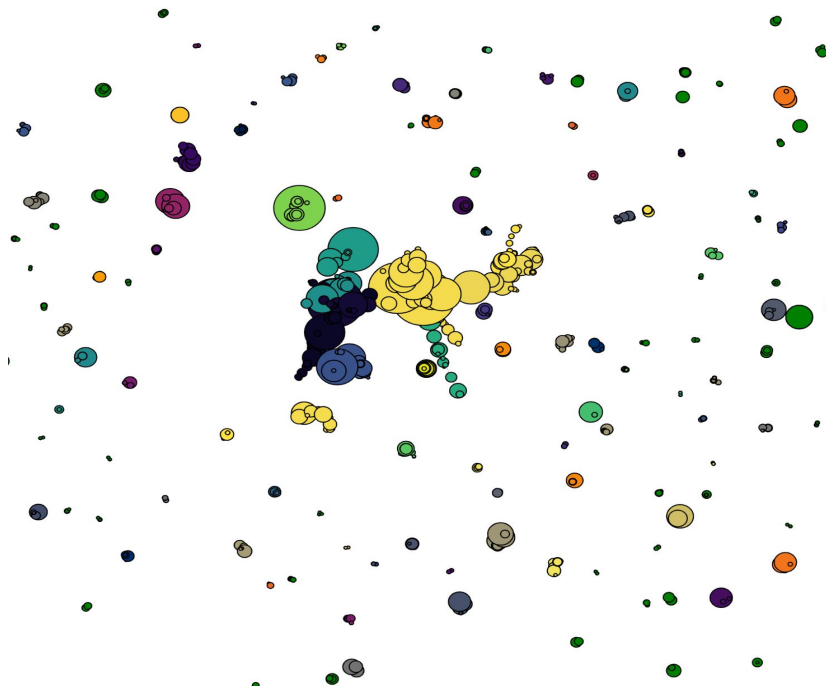


DEMO!



Graph analysis: Musical Genre detection (2)

Adjacency matrix: created with the second method
data: 20% of total dataset



DEMO!

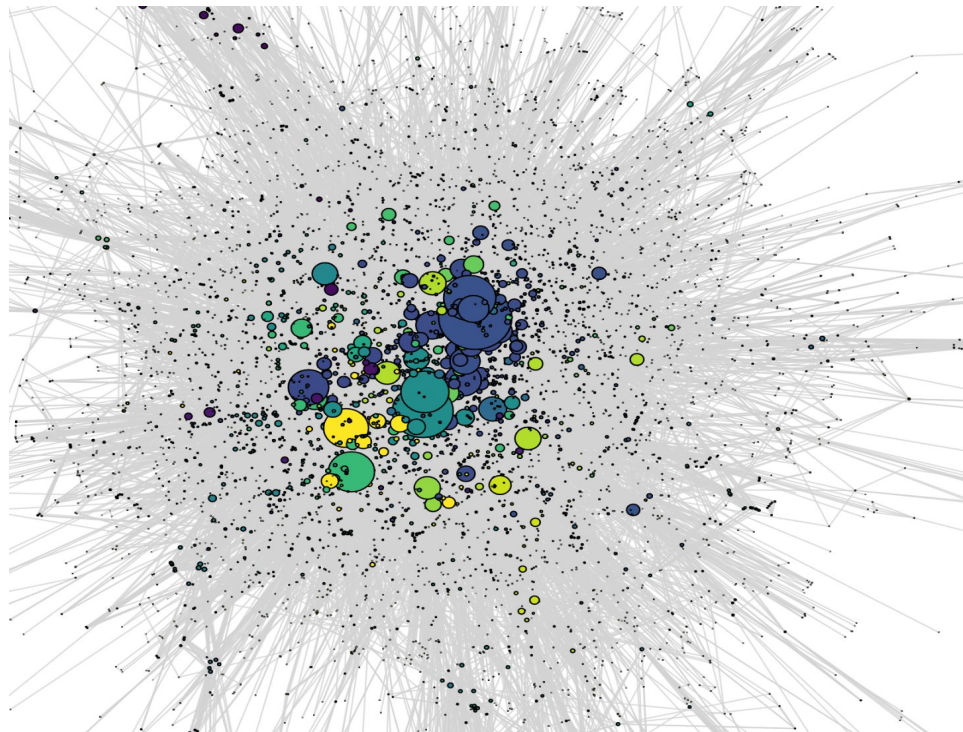
Graph analysis: Subgenre detection

Adjacency matrix: created with the first method

data: only songs from the RAP and HIP-HOP

DEMO!

Pop-rap2
Progressive hip hop
Psychedelic rap
Ragga5
Rap opera
Rap Rave
Rap rock
Rapcore4
Rap metal
Snap music
Stronda music
Trap
Trip hop1 4
Turntablism
Underground hip hop
Urban Pasifika5



Graph analysis: how to become popular

Adjacency matrix: created with the first method

data: only songs with popularity >70

DEMO!

	track_name	artist_name	songwriters	producers	labels	genre	popularity
8	Hero	mariah carey	[mariah carey, walter afanasieff]	[mariah carey, walter afanasieff]	[columbia records]	R&B	71
10	Tip Toe	jason derulo	[french montana, jason derulo, johnny mitchell...]	[bantu, pip]	[independent]	R&B	76
19	Dinero	jennifer lopez	[alexander prado, cardy b, danileigh, dj khale...]	[dj khaled, tbhits]	[independent]	R&B	70
20	Hate That I Love You	rihanna	[mikkel storleer eriksen, ne yo, rihanna, tor ...]	[stargate]	[def jam recordings]	R&B	72
26	OMG	usher	[usher, will.i.am]	[will.i.am]	[independent]	R&B	70



Limitations of the project

- The quantity of data we worked on is really small compared to the number of songs on Spotify (50M +)
 - BUT: we think the results we obtained are pretty well representative of the reality
- Lots of missing informations on Genius because of the collaborative nature of the platform
 - We had to make choice to replace missing informations
 - Some informations may be wrong



Can we go further?

- Extend the dataset as much as possible
 - Can be very long because Spotify limits the number of pull requests you can do in a day
 - Analysis will become even longer which can be problematic
 - plotting with Bokeh could lag a lot
- Do a time analysis of the data
 - extend the analysis on several years to see the evolution of the industry
 - focus on one actor and see how his collaborations evolved



Conclusion

This work has confirmed the prejudices we have about the music industry

1. it is a small well connected world
2. all successful songs are the produced by a small number of people

Graph clustering was a good approach to study the industry and allowed us to do subgenre detection.

Lots of measures about the graph couldn't be computed cause the graph is not fully connected.

