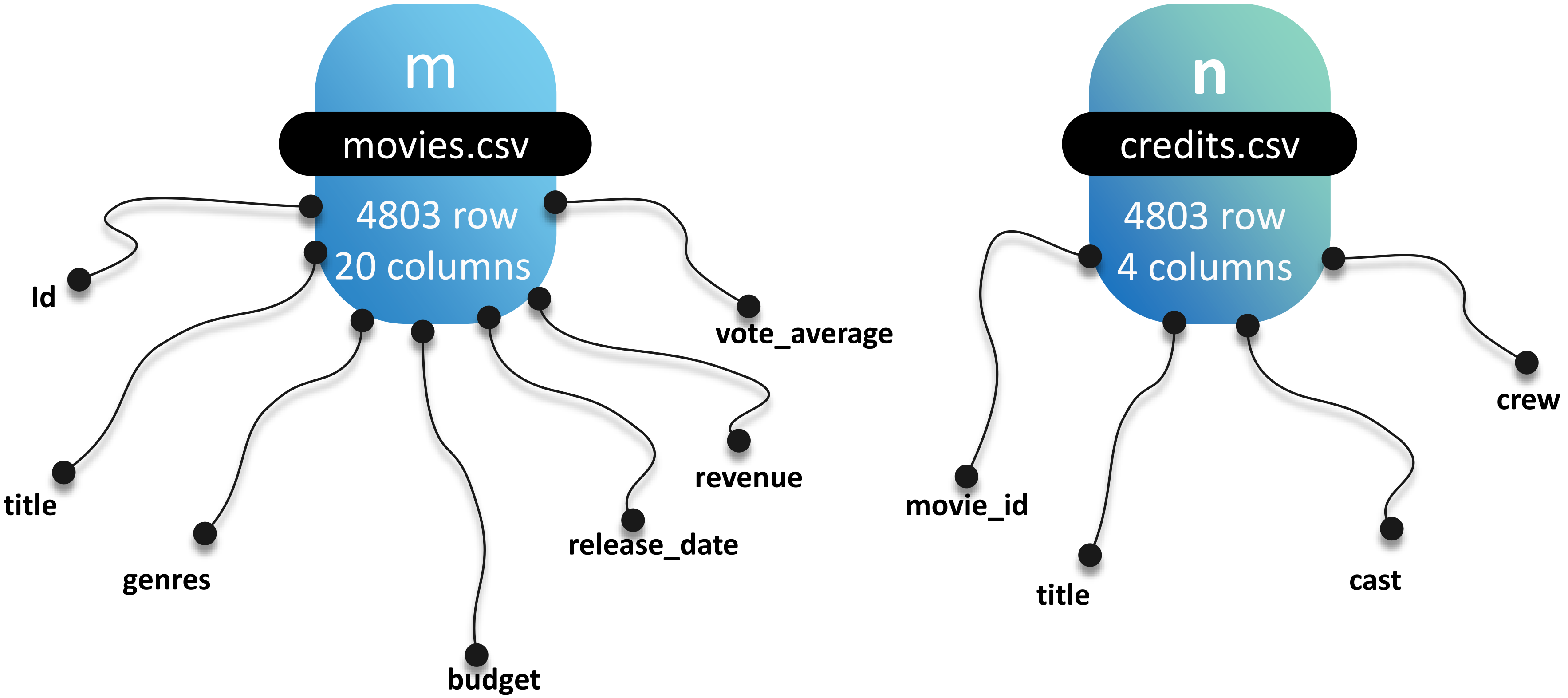
Evolution of the movie industry

Team 04

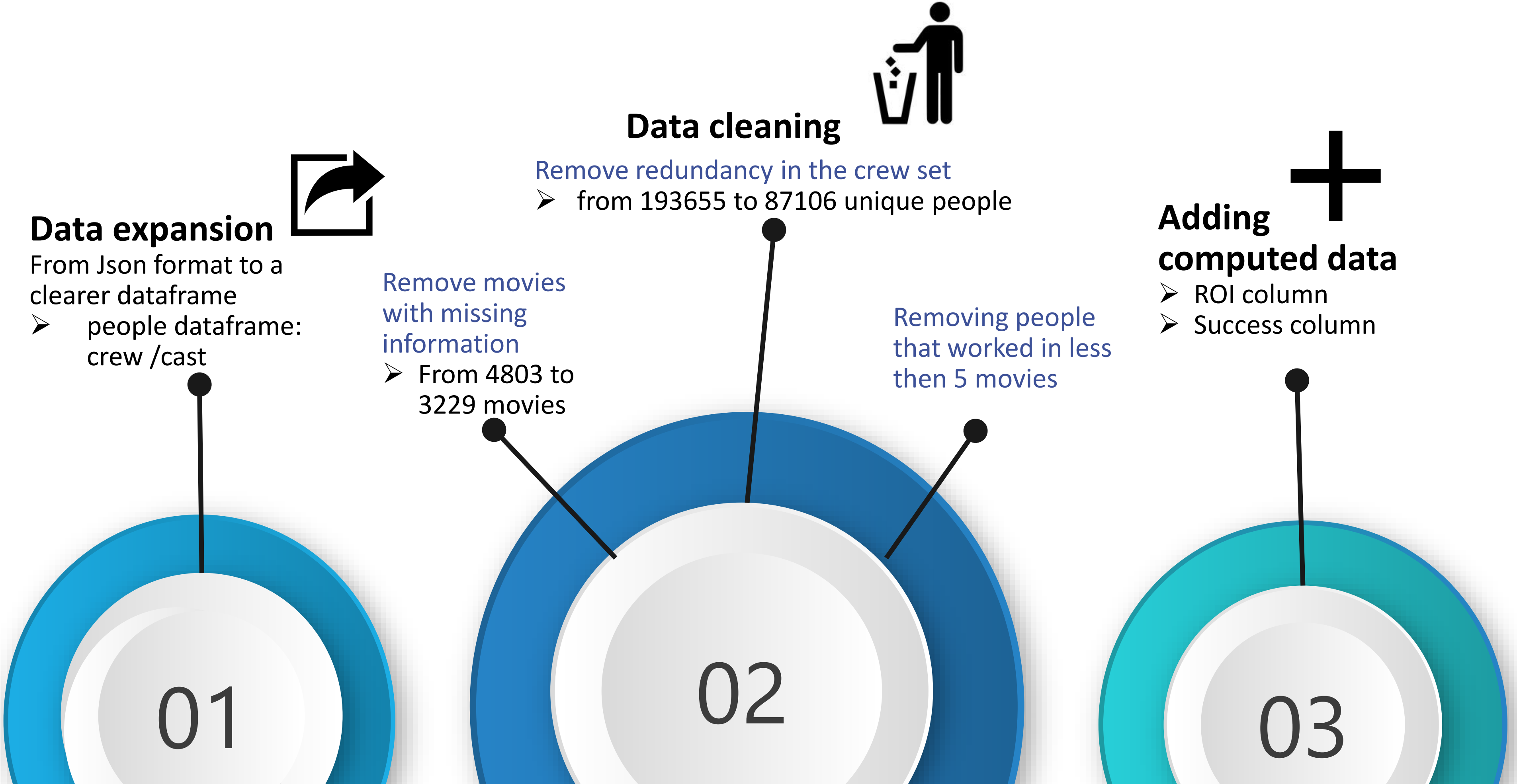
Julien Berger - Jérémy Jayet - Hana Samet - Mathieu Shiva

The initial dataset

Two csv files



Data cleaning



Data expansion

From Json format to a clearer dataframe

- ➢ people dataframe: crew /cast

Data cleaning

Remove redundancy in the crew set

- ➢ from 193655 to 87106 unique people

Remove movies with missing information

- ➢ From 4803 to 3229 movies

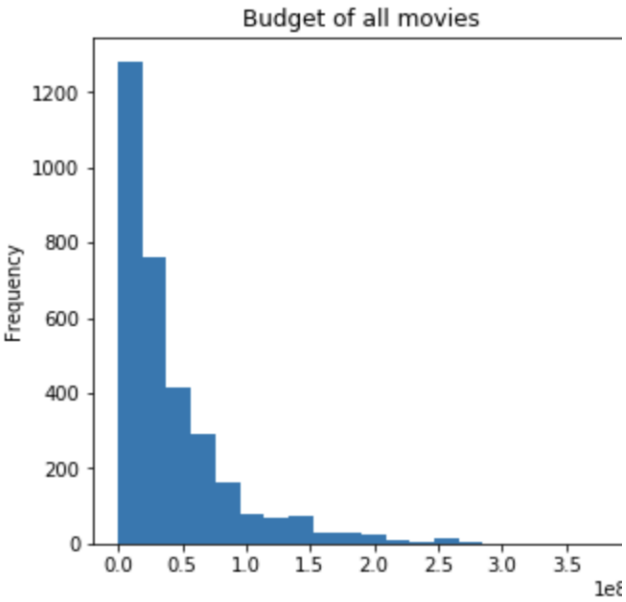
Removing people that worked in less then 5 movies

Adding computed data

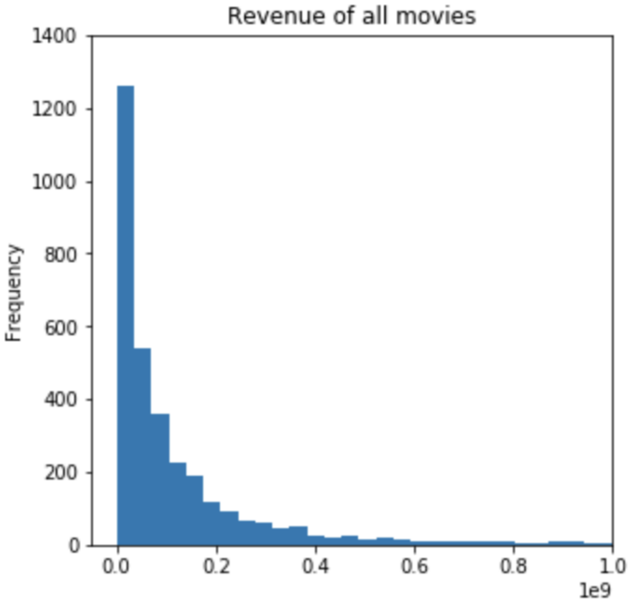
- ➢ ROI column
- ➢ Success column

Budget/revenue Analysis

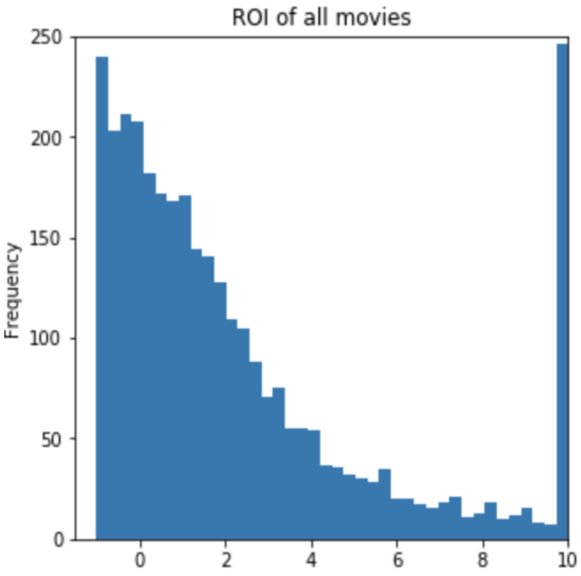
The financial histograms of the 3229 movies


Budget of all movies

Tails-like distributions


Revenue of all movies

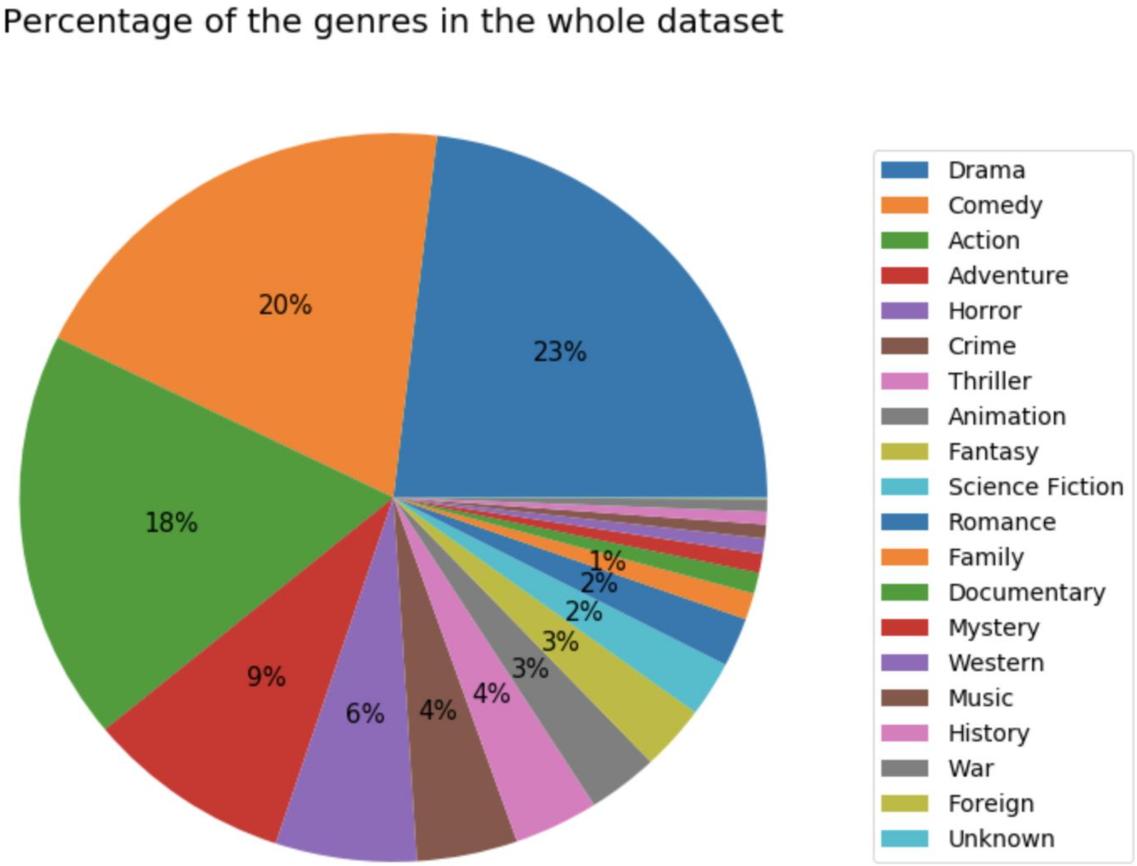
A lot of movies had revenues that were smaller than their budget.


ROI of all movies

Only 23 movies had very high ROIs

Genres

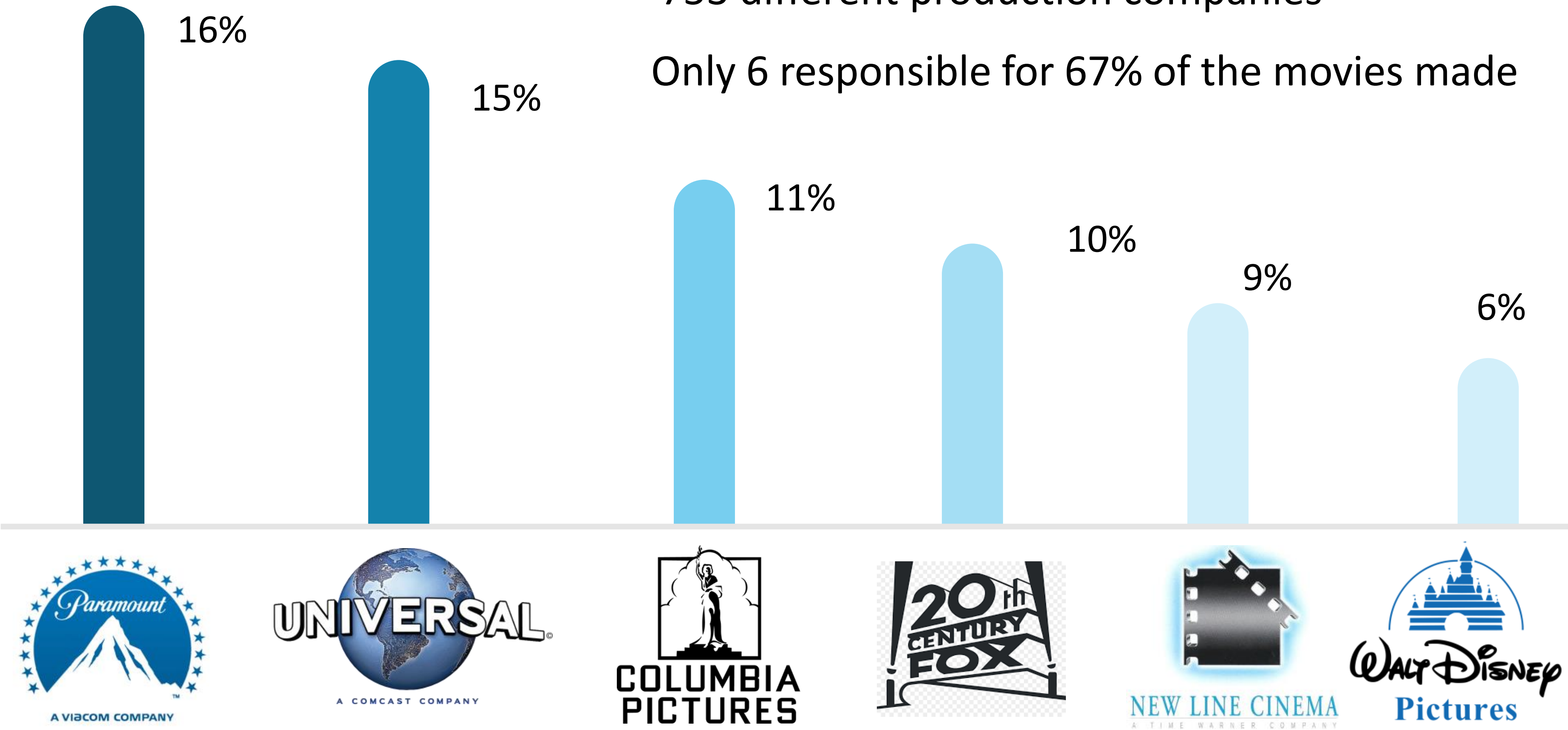
Drama, Comedy and Action represent 61% of the movies


Percentage of the genres in the whole dataset

Production companies

753 different production companies

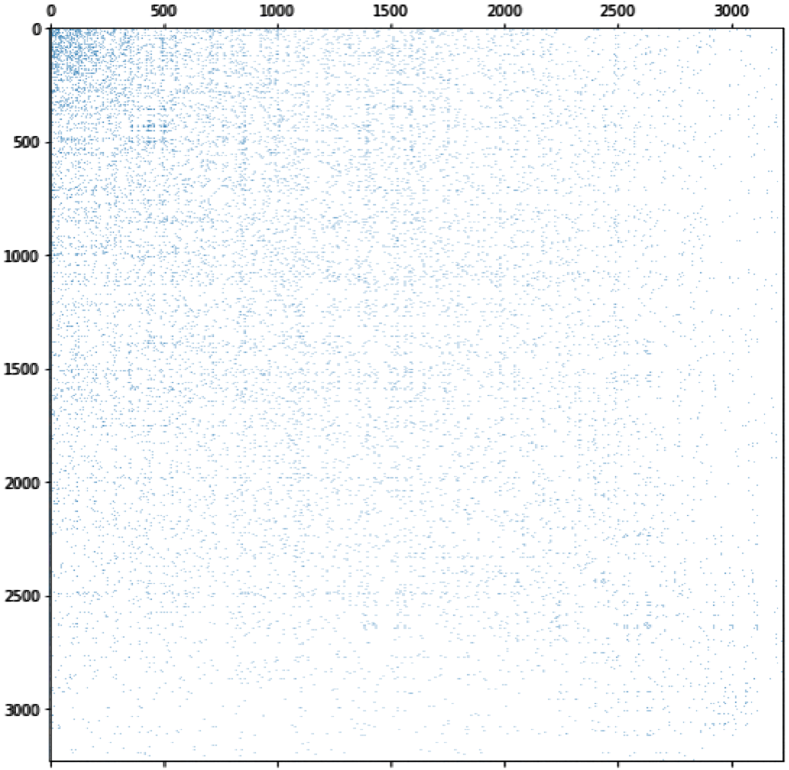
Only 6 responsible for 67% of the movies made

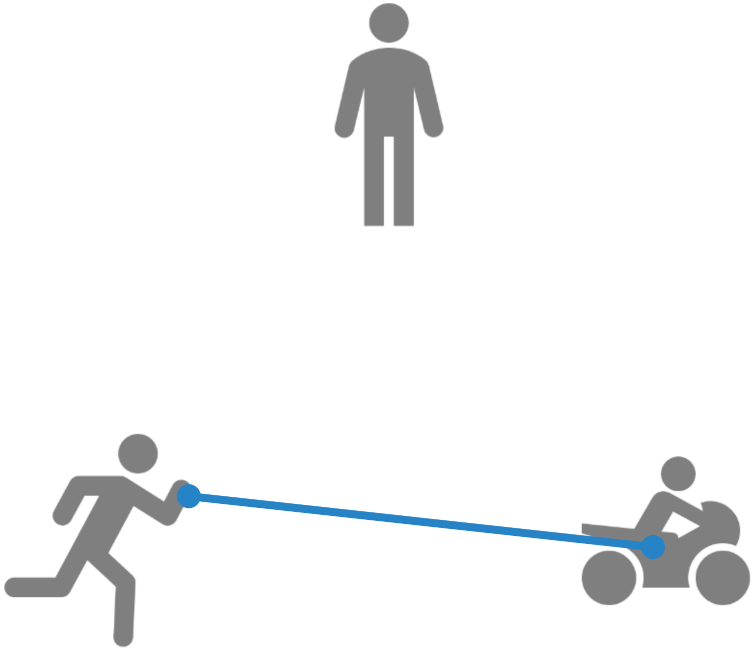


Adjacency matrix - People based

$<k> = 4$

Actors only

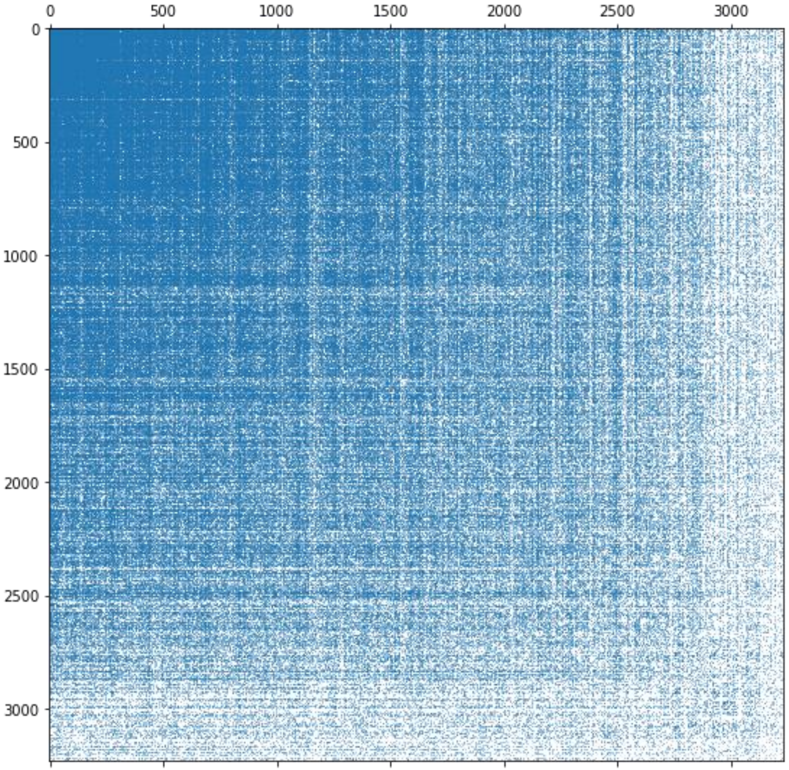


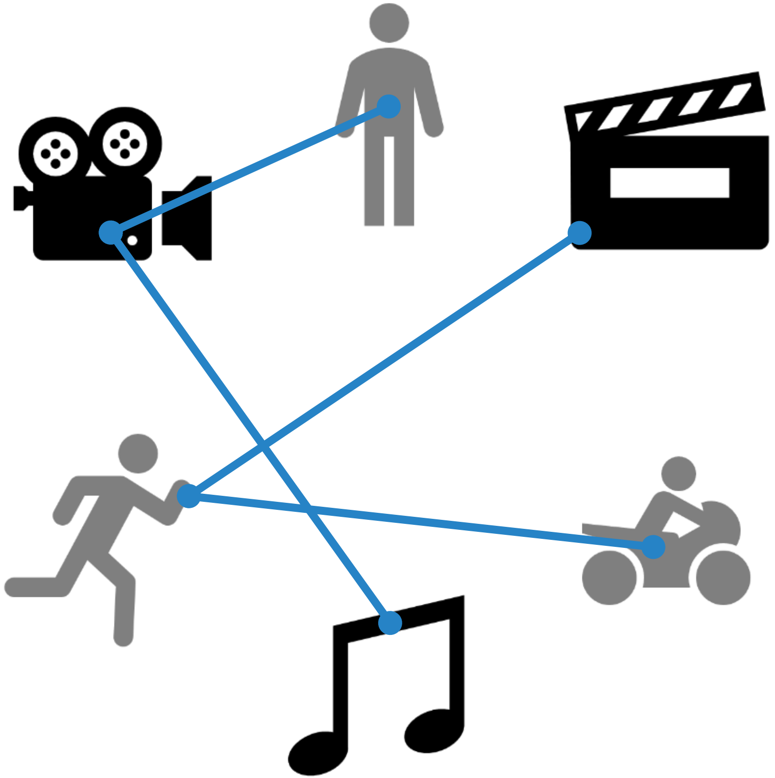


Adjacency matrix - People based

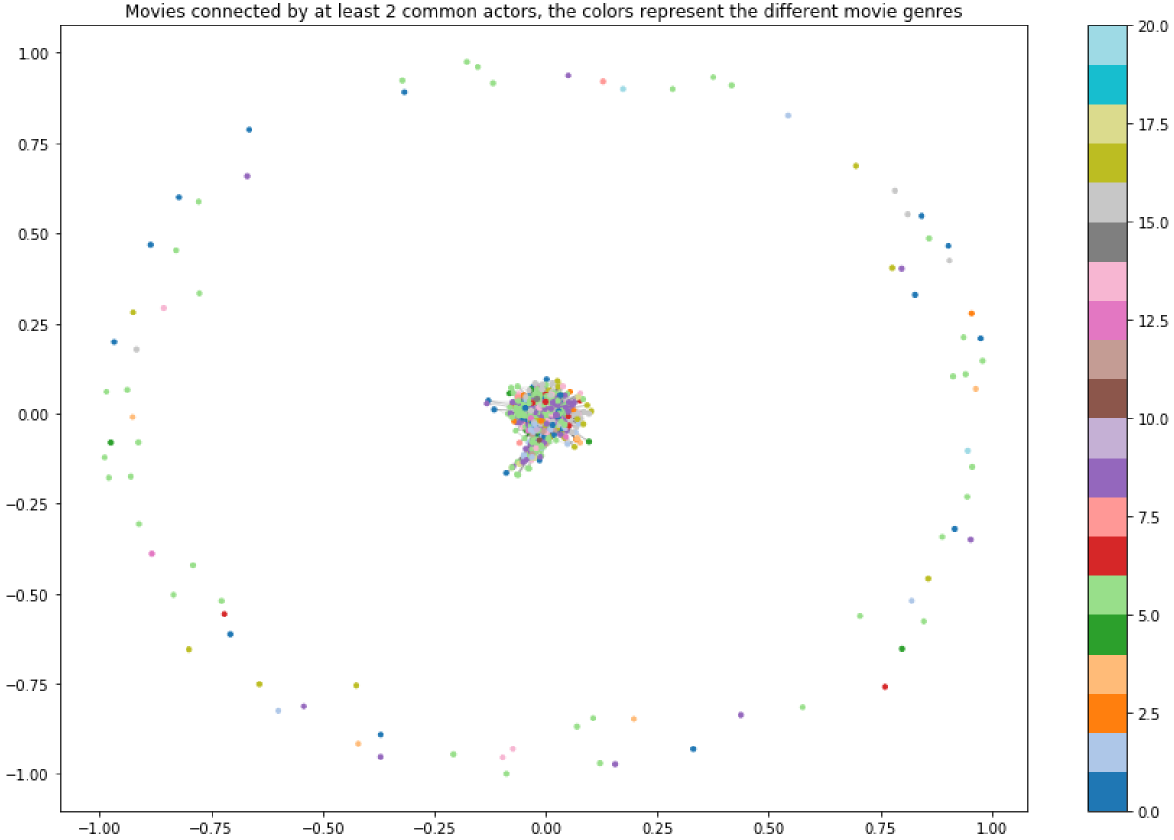
$\langle k \rangle = 188$

Actors and crew



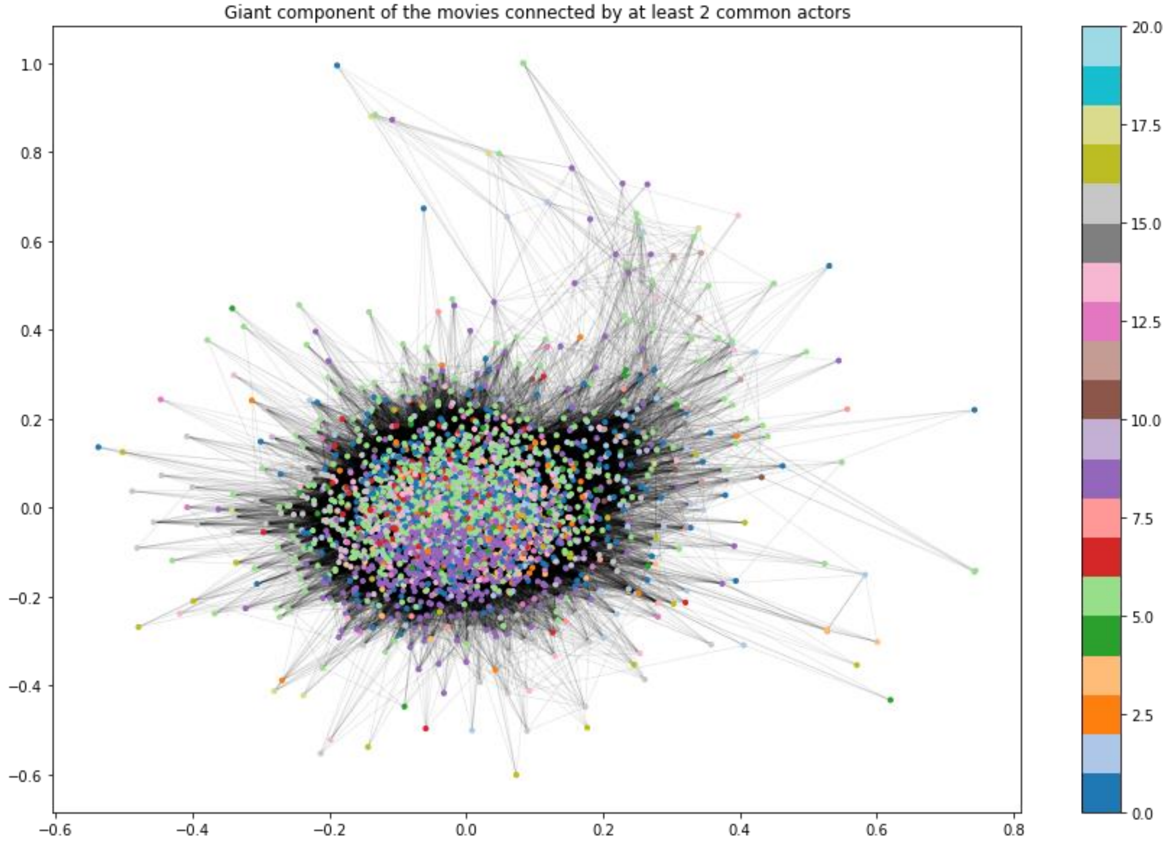


Adjacency matrix - People based



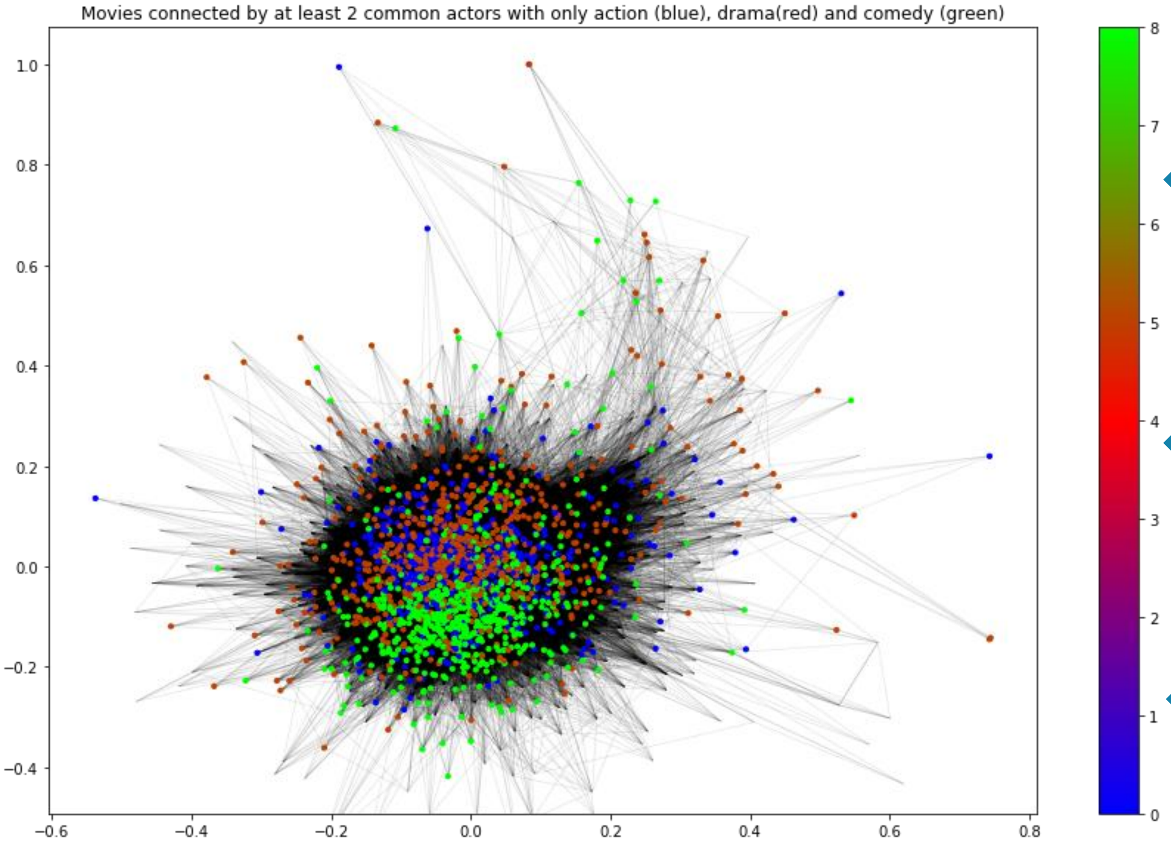
2883 movies are connected

Adjacency matrix - People based



2883 movies are connected

Adjacency matrix - People based



Movies connected by at least 2 common actors with only action (blue), drama(red) and comedy (green)

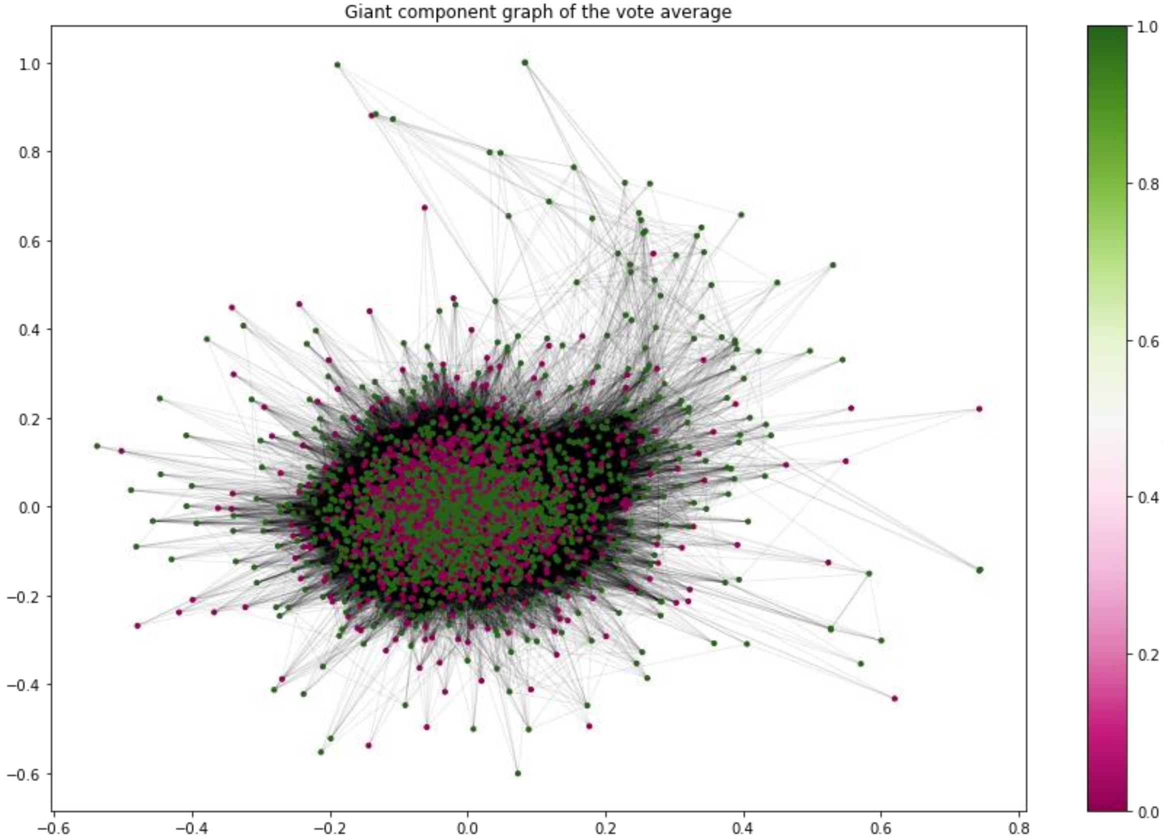
2883 movies are connected

No communities can be seen

Separation of the graph in subgraphs is required

Adjacency matrix - People based

Success rate based on the ROI



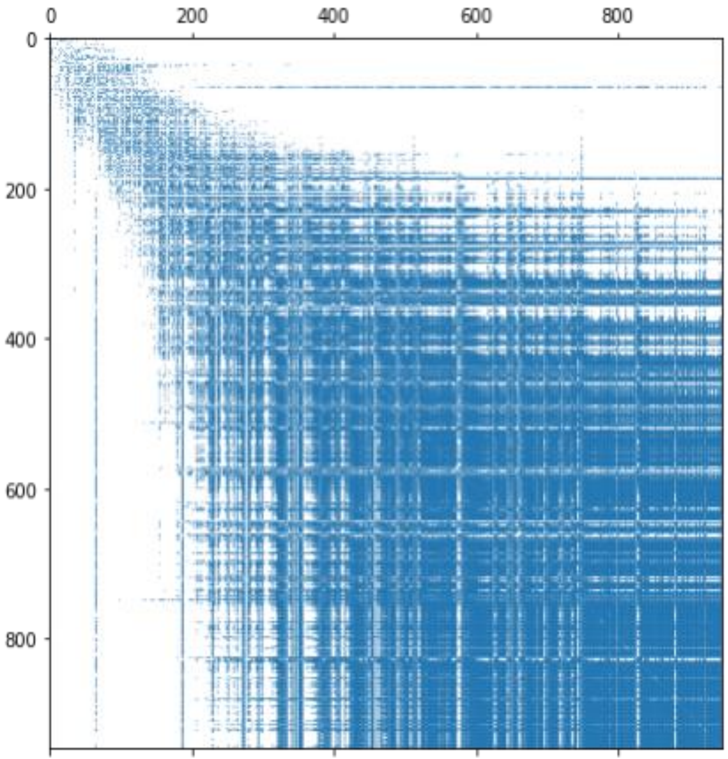
2883 movies are connected

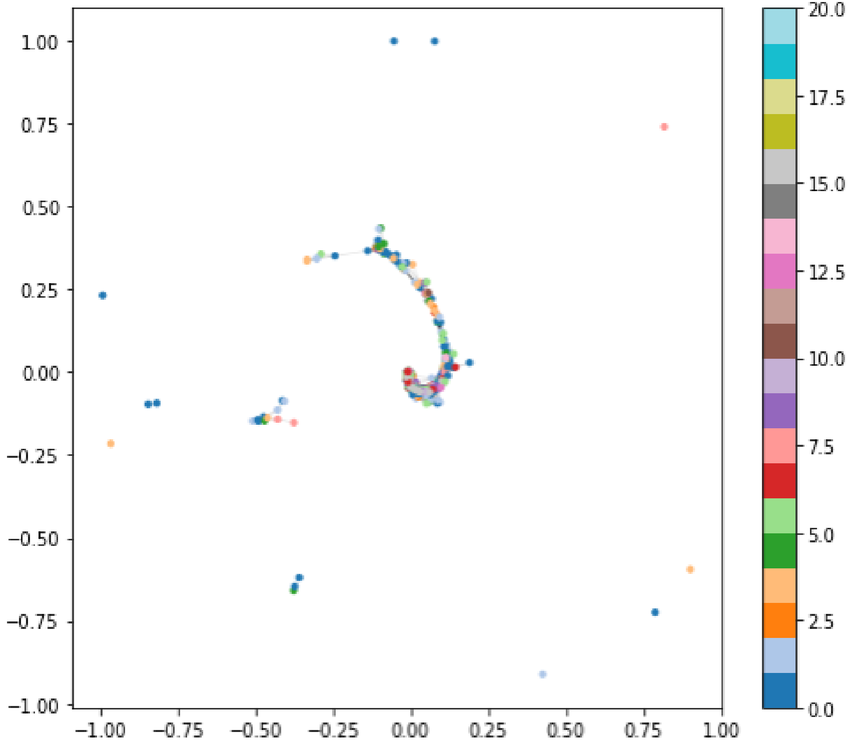
No communities can be seen

Separation of the graph in subgraphs is required

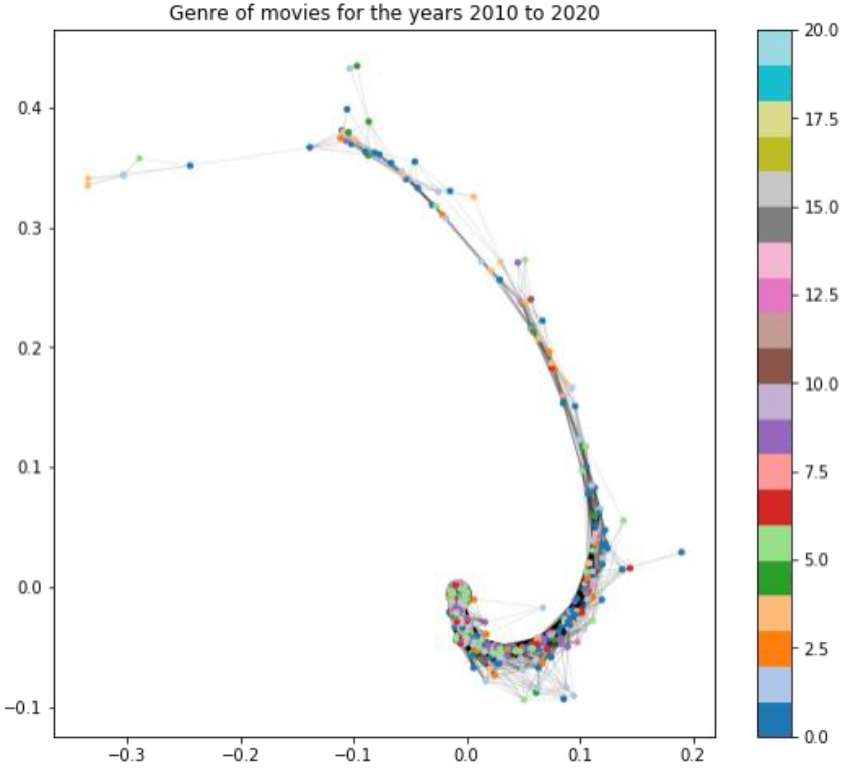
Adjacency matrix - Budget based

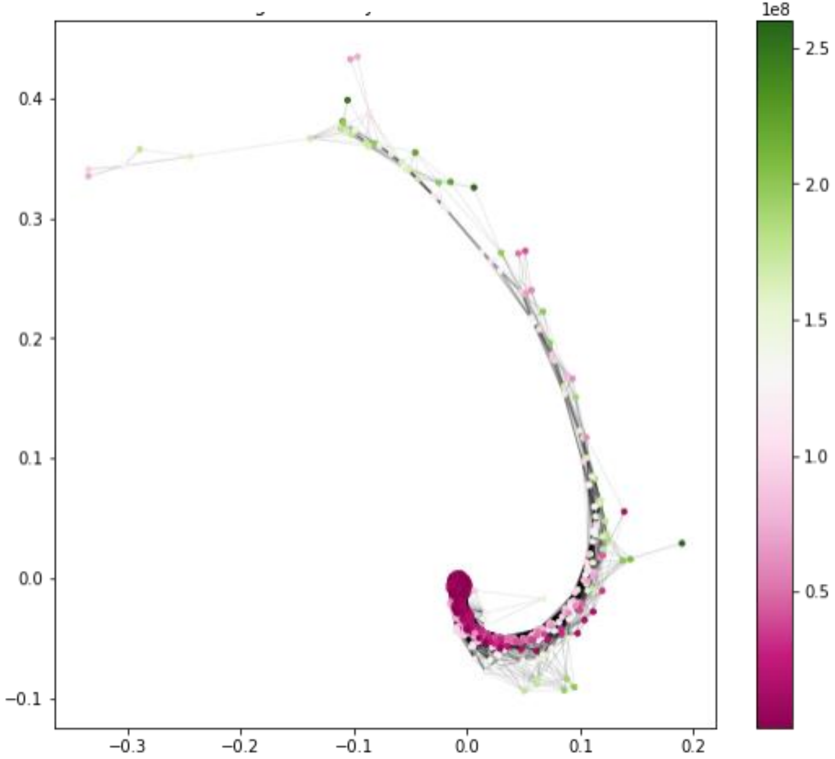
Euclidean distances between movies based on the budgets and revenues



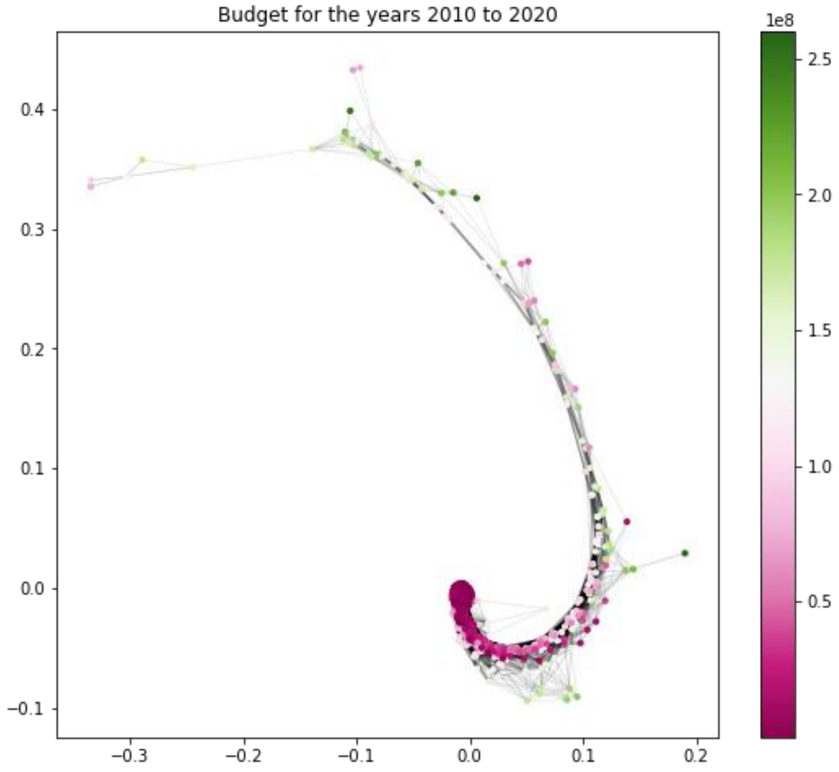


Adjacency matrix - Budget based

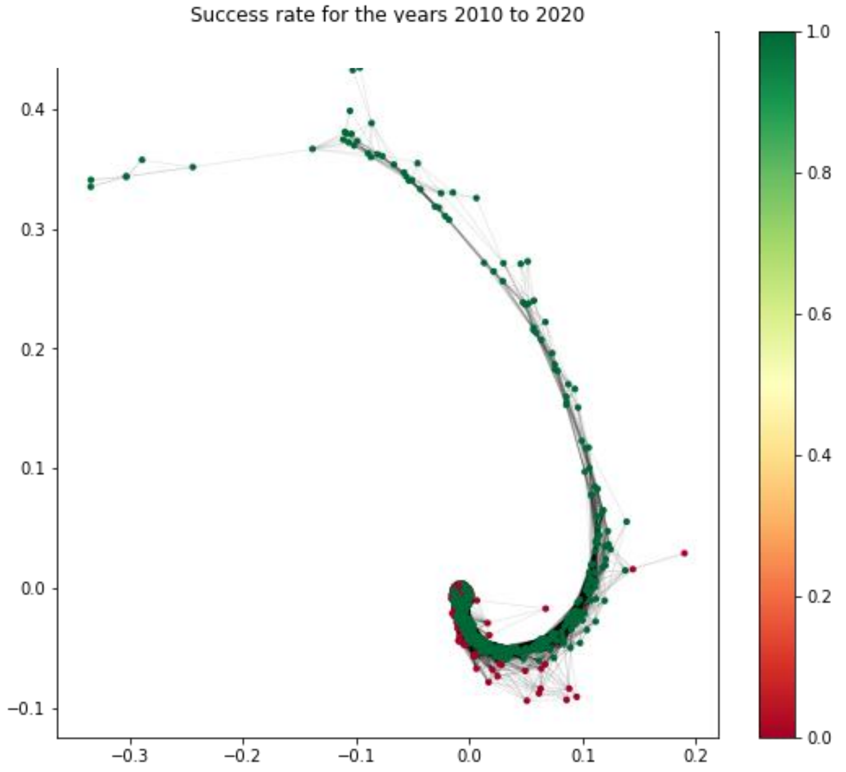




Adjacency matrix - Budget based



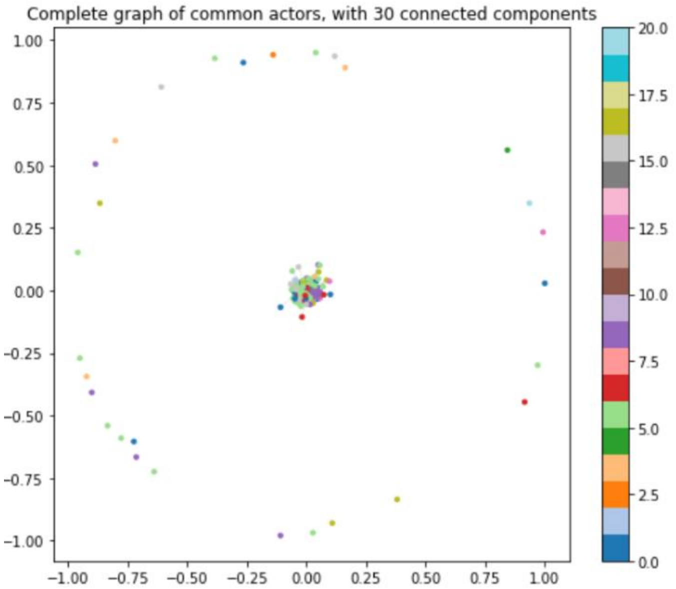
Budget for the years 2010 to 2020



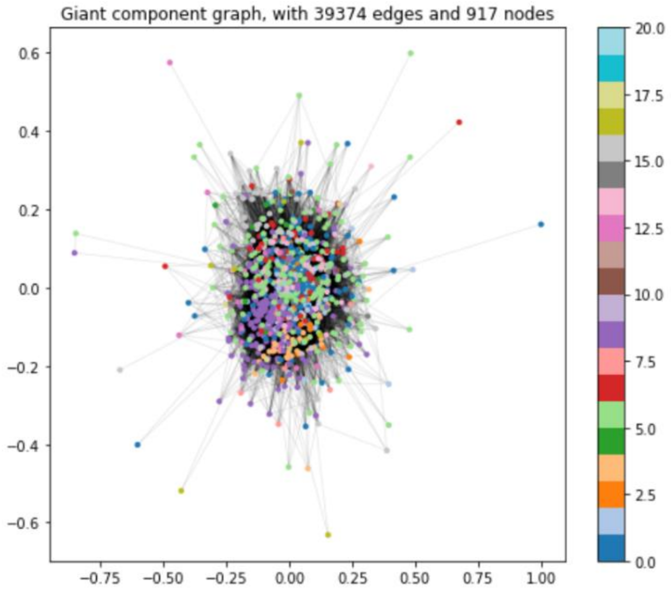
Success rate for the years 2010 to 2020

Separation into decades

- 6 decades: 1960-2020
- Movies adjacency

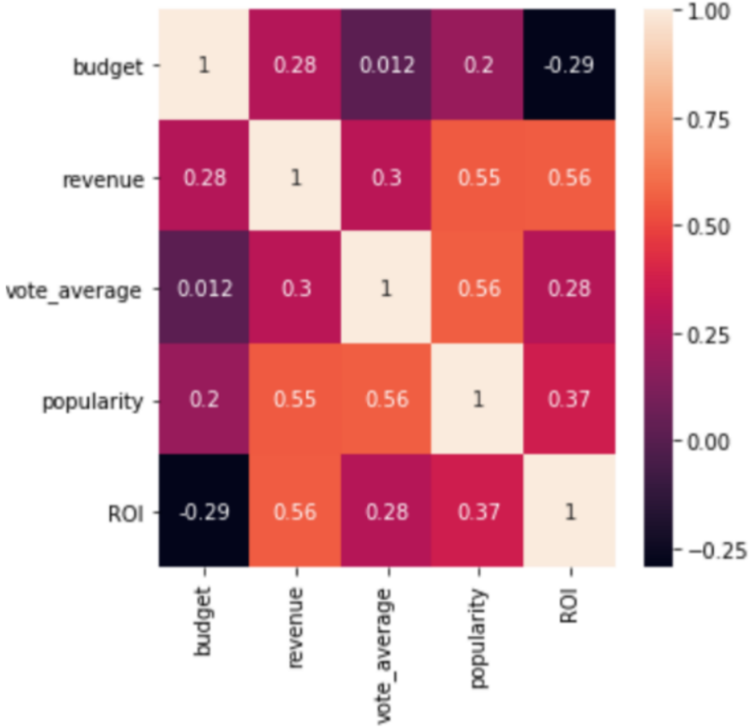


Complete graph of common actors, with 30 connected components

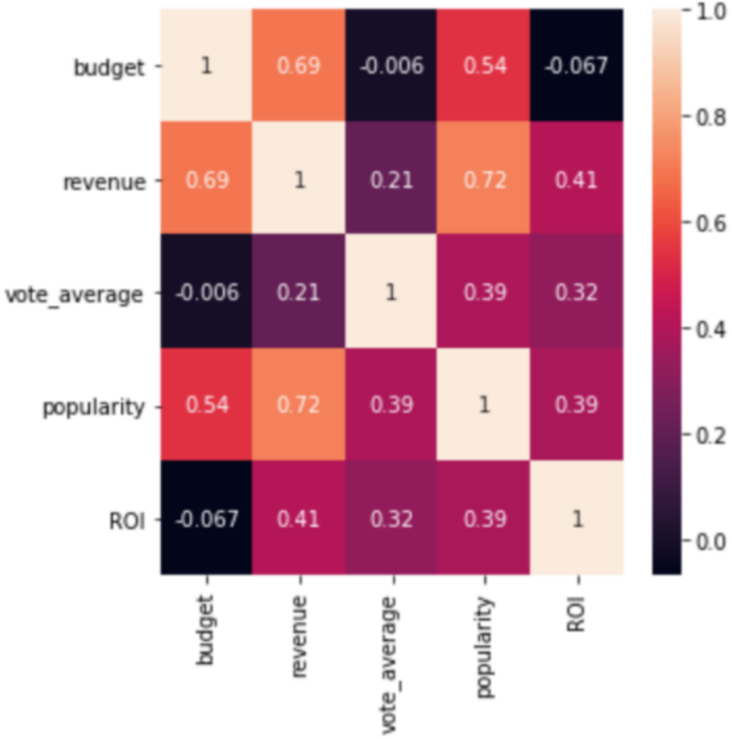


Giant component graph, with 39374 edges and 917 nodes

Correlation Heat Maps

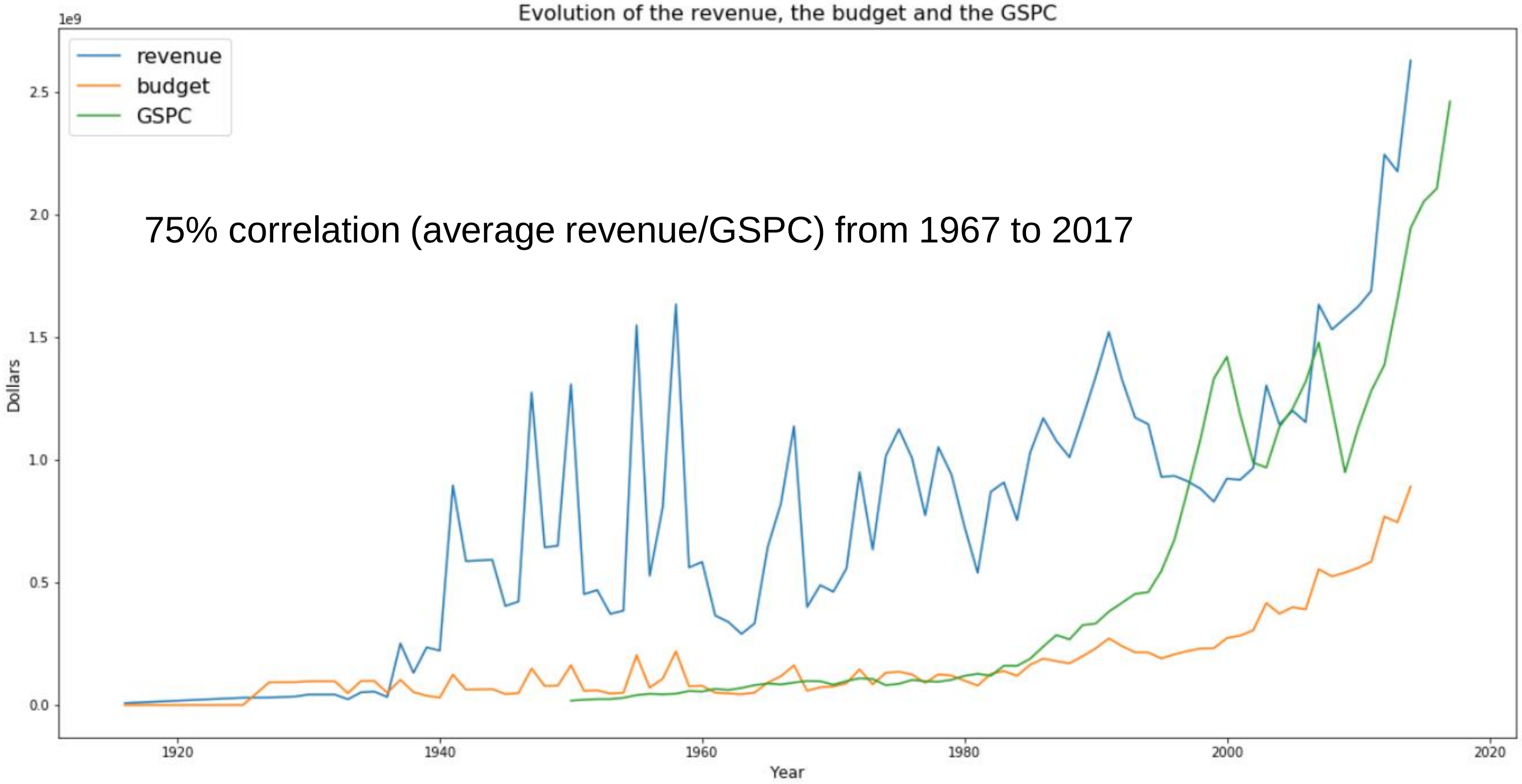


The heatmap for the years 1980-1990



The heatmap for the years 2000-2010

Correlation with the global economy

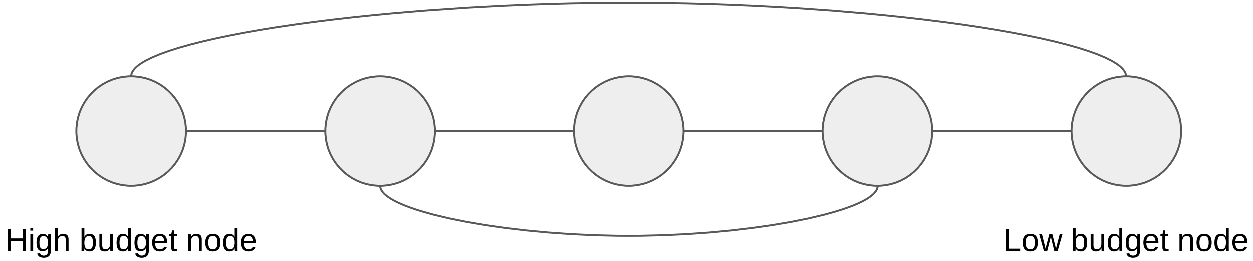


75% correlation (average revenue/GSPC) from 1967 to 2017

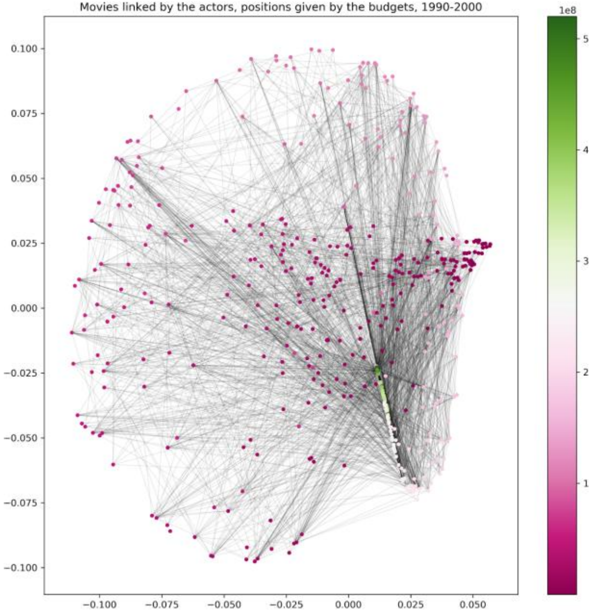
Budget-based graph vs actors-crew based graph

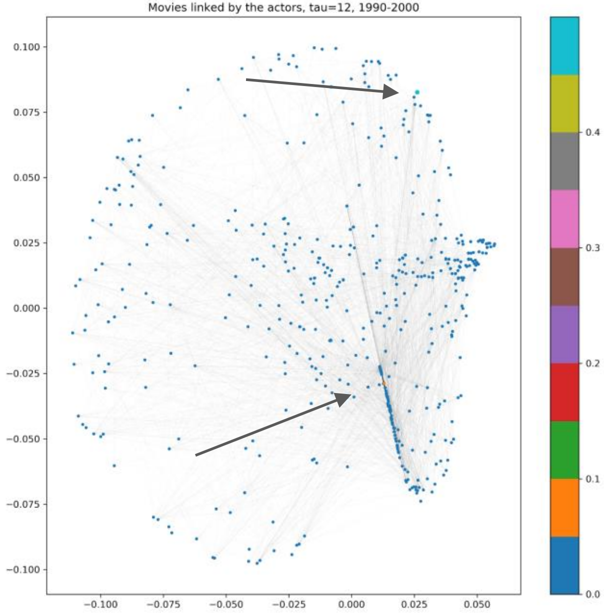
The actors-crew consists mostly in its giant components which is connected.

But how is it connected ?



Diffusion of a Dirac impulse on the graph



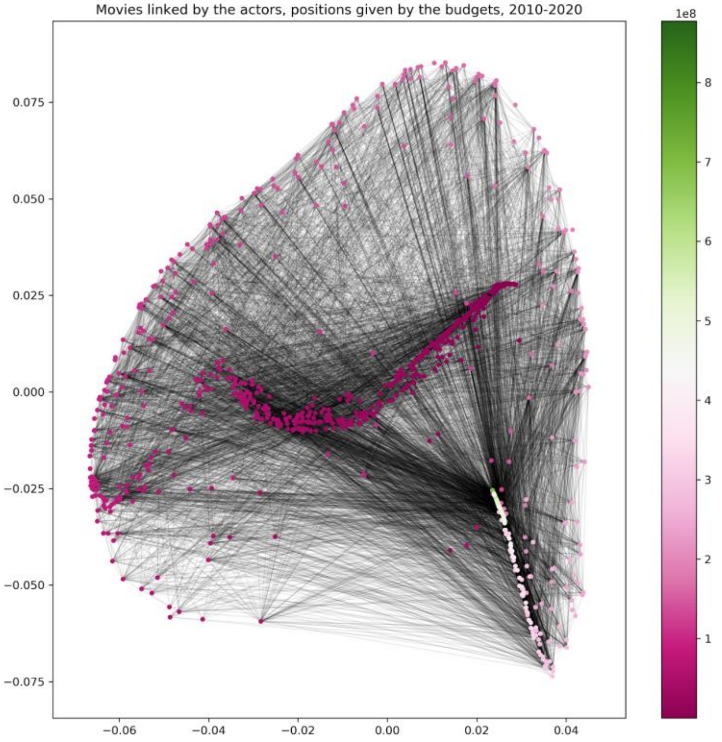


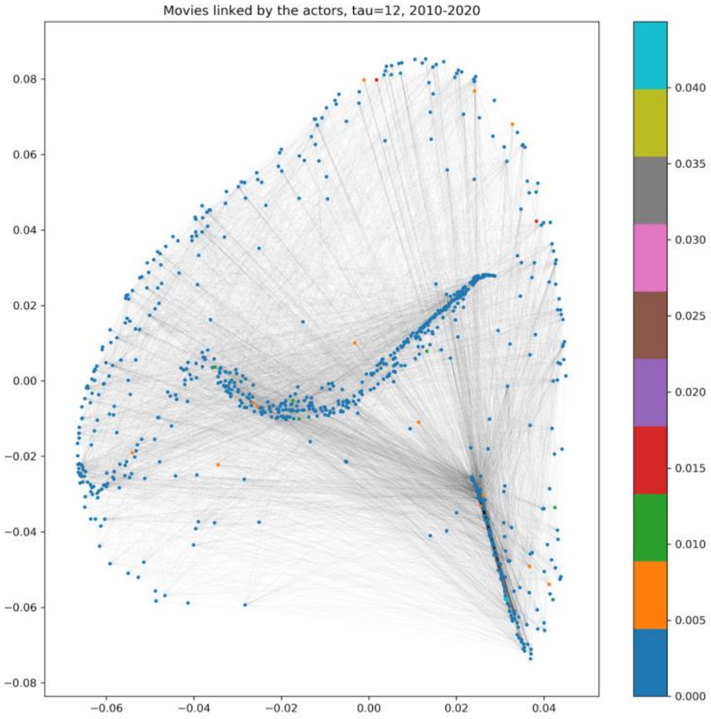
Qualitative analysis

T

Sr
gra

Th
o



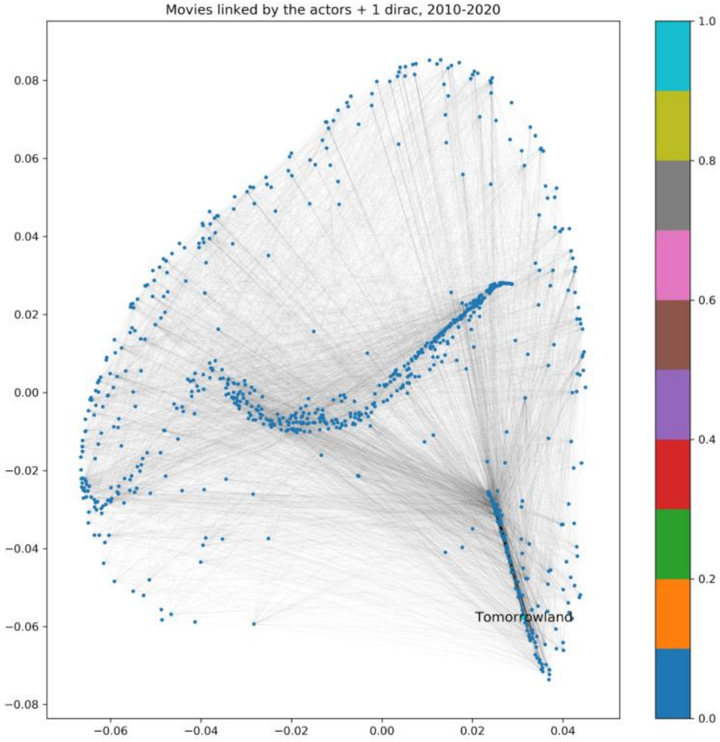


Qualitative analysis

The signal spreads across the whole graph

Smoothness on the actors based graph does not mean smoothness on the budget based graph.

There are no communities based on the budget class of the films



Conclusion

- We did not find any communities in the actors-based graph.
- The budget-based graph does not show any community either and the budget appears as a very smooth signal over the graph.
- There is a high correlation between the popularity and the ROI.
- The budget based graph has a very different structure than the actors based graph. Therefore we can conclude that there is no segregation between actors playing in low budget movies and those playing in high budget movies.