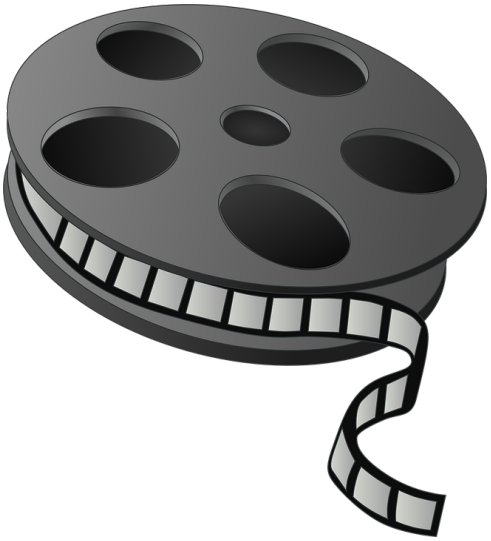
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

NETWORK TOUR OF DATA SCIENCE

EE-558

Team 04

Evolution of the movie industry



Julien BERGER
Jérémy JAYET
Hana SAMET
Mathieu SHIVA

January 18, 2019

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

1 Aim of the project and problem definition

The idea of our project is to use a subset of the IMDB movie dataset, taken from Kaggle [1], to make an analysis of the evolution of the movie industry throughout the years. More specifically, we want to have have an economy-oriented approach, by looking at properties such as the budget or the return on investment, and see if trends can be determined from these. The final goal is to see if we can determine certain communities based on the budget in the movie industry, with actors that always play in large budget movie, or if on the contrary, actors tend to plays in different types of movies.
We will make multiple graphs, one per decade, to analyze the evolution of the industry. The dataset contains information about average rating, popularity, revenue, budget and genre of the movies (among other things), which we will use to see if we can find trends or communities.

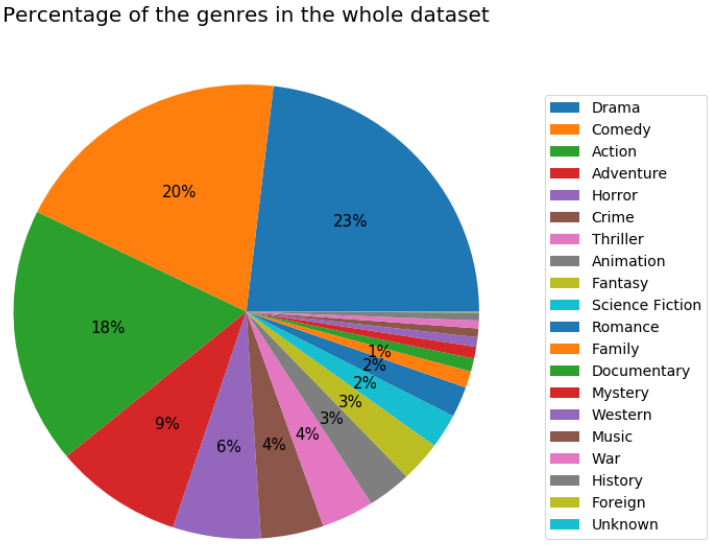
Additional graphs and information about the different sections of this report can be found in the jupyter notebook on our Github.

2 Dataset

2.1 Exploration

2.1.1 Genres

The initial dataset contains information 4803 movies, and 87'106 people (either part of the cast or of the crew). However as we want to use revenues and budget in our analysis, we removed the movies that were missing these data. Therefore we were left with a smaller dataset that contains 3229 movies. We also kept only the people that worked on at least 5 movies. There are 20 movie genres present in the dataset. The most represented genres in the dataset are drama, comedy and action, which account for 61% of the movies.



(a) The percentage of the movies of each genre in the dataset

Drama	747
Comedy	634
Action	588
Adventure	288
Horror	197
Crime	141
Thriller	118
Animation	99
Fantasy	93
Science Fiction	79
Romance	70
Family	38
Documentary	30
Mystery	27
Western	22
Music	20
War	18
History	18
Foreign	1
Unknown	1

(b) The count of the movies of each genre

Figure 1: The genres present in the filtered dataset

To build a graph that connects the movies between them, there are many possibilities. The one we initially chose was to link them if they had at least 2 actors in common. Of course, this method left some movies unconnected, as seen in Figure 2, so we worked with the giant component.

[1] https://www.kaggle.com/tmdb/tmdb-movie-metadata

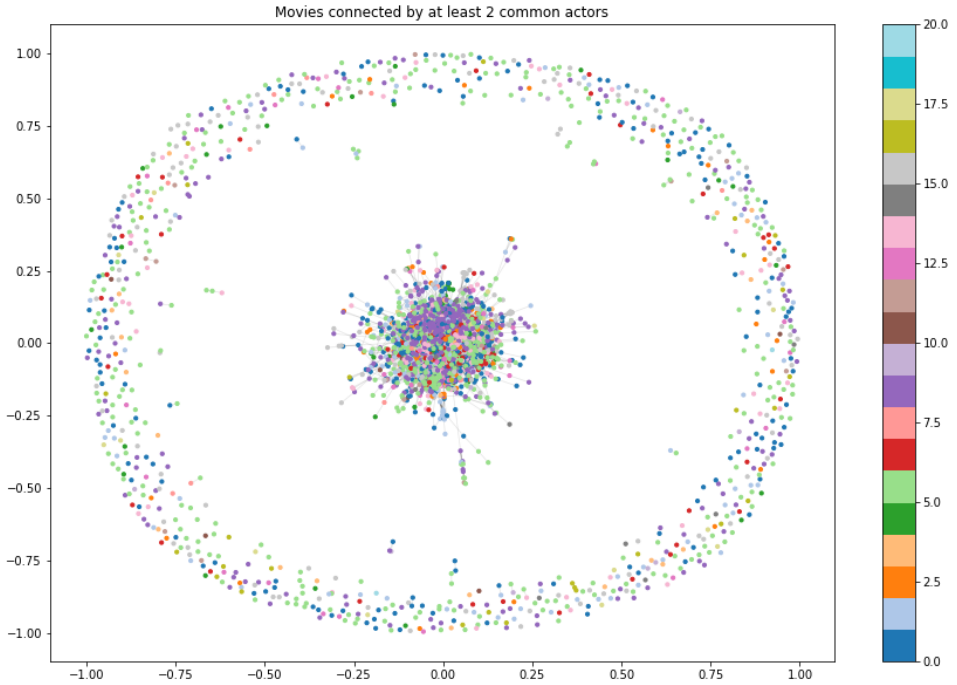
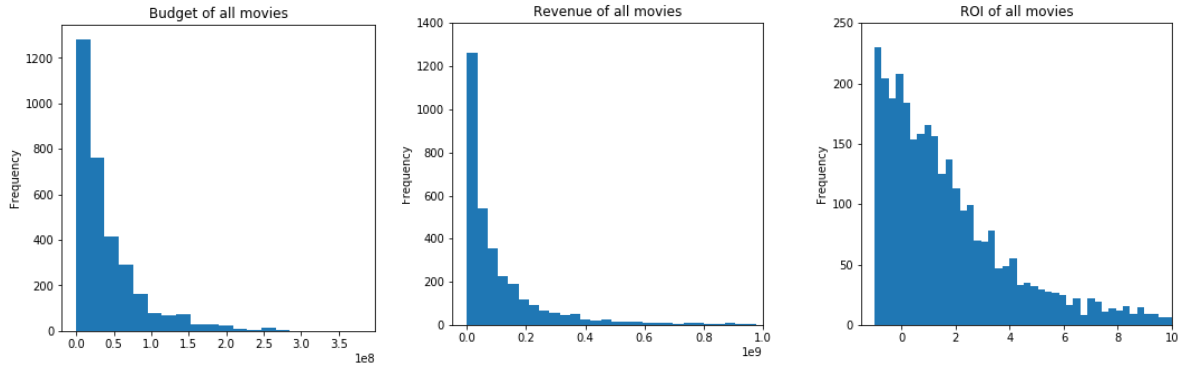


Figure 2: The graph that links movies if they have at least 2 actors that played in the same movie. The color represents the genre. In dark blue you have action movies, in light green you have drama and in light purple comedy, the three most represented genres.

2.1.2 Budget and revenue

The budget and revenue of each movie was given in the dataset, and the return on investment (ROI) for each movie was calculated as $ROI = \frac{(Revenue - Budget)}{Budget}$. The histograms of the budgets, the revenue and the return on investments can be seen in Figure 3. However, 23 movies had very high ROIs, and thus were not represented on this graph.



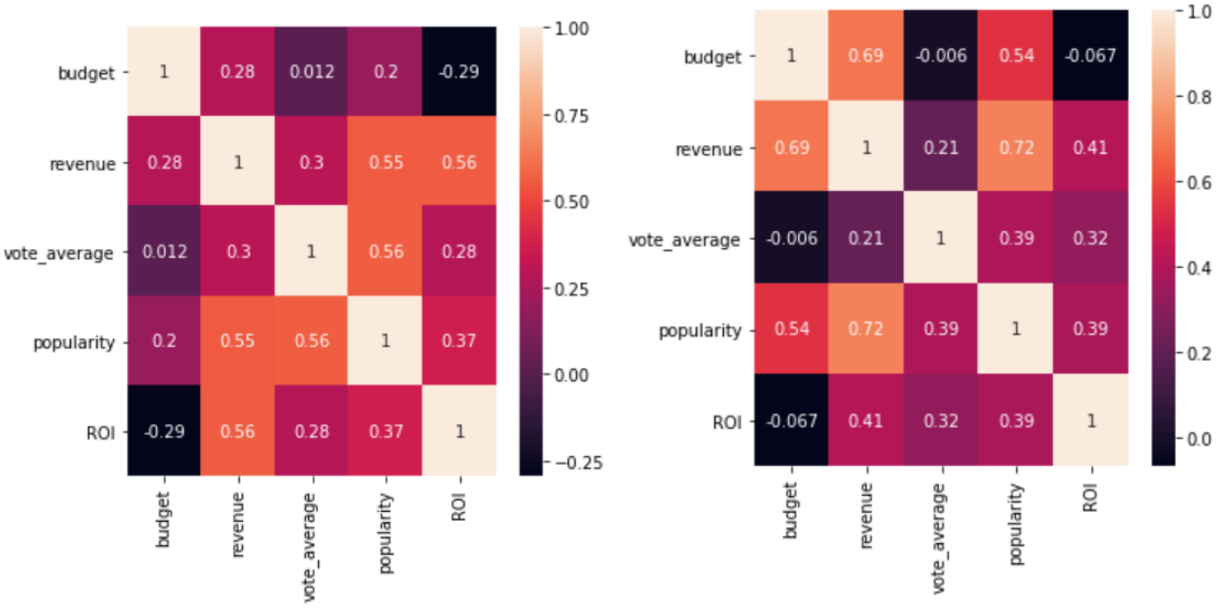
(a) The budgets (b) The revenues (c) The ROI

Figure 3: The financial histograms of the 3229 movies.

2.2 Exploitation

2.2.1 Decade separation and correlation

We split the dataset in 6 different decades, from 1960 to 2020, and created a graph for each decade (in similar manner than the one presented in Section 2.1.1). For each decade, we looked at the correlation between different parameters, as seen in Figure 4.



(a) The heatmap for the years 1980-1990

(b) The heatmap for the years 2000-2010

Figure 4: The heatmaps of correlation for two different decades

What can be seen from these maps, is that over the years the correlation between the budget and the popularity increased. However, it is not the case between the budget and the vote average. This points that the trend of the movie industry over the years is to make high-budget movies which will be very popular, thus improving the ROI. At the same time, this points that the best-rated movies are not necessarily the most popular ones.

We then tried to see if there was a strong correlation between the movie industry and the world economy. We collected (via Yahoo finance) the average yearly evolution of the S&P500 (an index representing the 500 largest US companies). We find a correlation of 75% over the last 50 years. Therefore we conclude that the movie industry is affected by the economic crisis/expansions thus adding a bias over the data. However the below graph and data analysis, are be based on decades and we assume that these effects would average out over the decade.

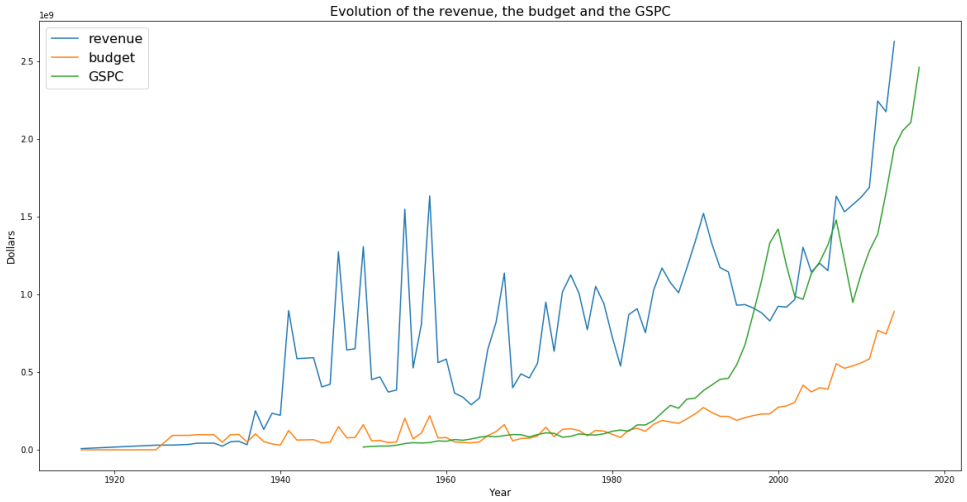


Figure 5: The evolution of the average budget and revenue, as well as of the S&P500 (GSPC)

We built additional graphs for each decade by using the euclidean distance of the revenues and the budget. The graphs obtained in this manner have elongated shapes, as shown in Figure 6.

Using them allows to better visualize if there are communities present, as well as the correlation between different parameters. Additional graphs can be found in our jupyter notebook.

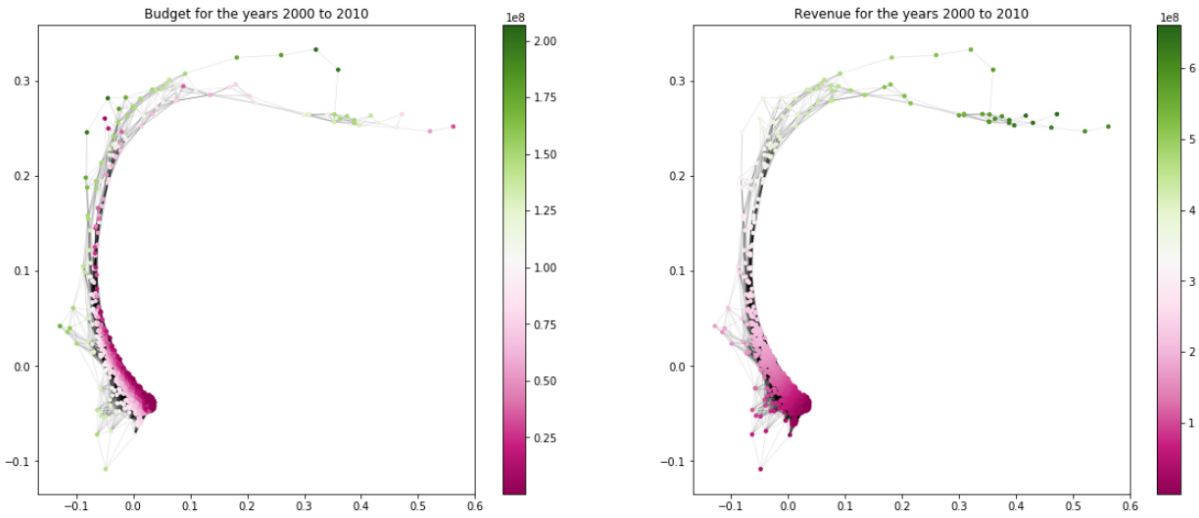


Figure 6: Graphs of the budget and the revenue for the years 2000 to 2010

2.2.2 Research for segregation of the actor community based on the budget category of the movies

In this section, we want to find out if there is some kind of segregation in the people of the movie industry. We focused on a segregation based on the "budget" of each movie, meaning that there would be communities of actors and crew playing and working only in the same class of movie, regarding the budget.

Our approach was to combine the two graphs we made before. We used the graph based on the budgets (see Figure 6) to determine a position for each node by using the laplacian eigenmaps. Then, we plotted the graph of the movies connected by the actors (see Figure 2), using the coordinates previously found. To visualize the interconnection across the graph, we put a Dirac impulsion on one node and used a heat filter to see the propagation of the signal through the edges.

We used different nodes to localize the impulse. If the node used has a large enough degree, meaning that it has a lot of actors in common with other movies, you can systematically see that the signal spreads across the whole graph, and not only in regions with a high or a low budget.An improvement to this method would be to quantify systematically the similarity of the edges between the two graphs, budget-based and actors-based.

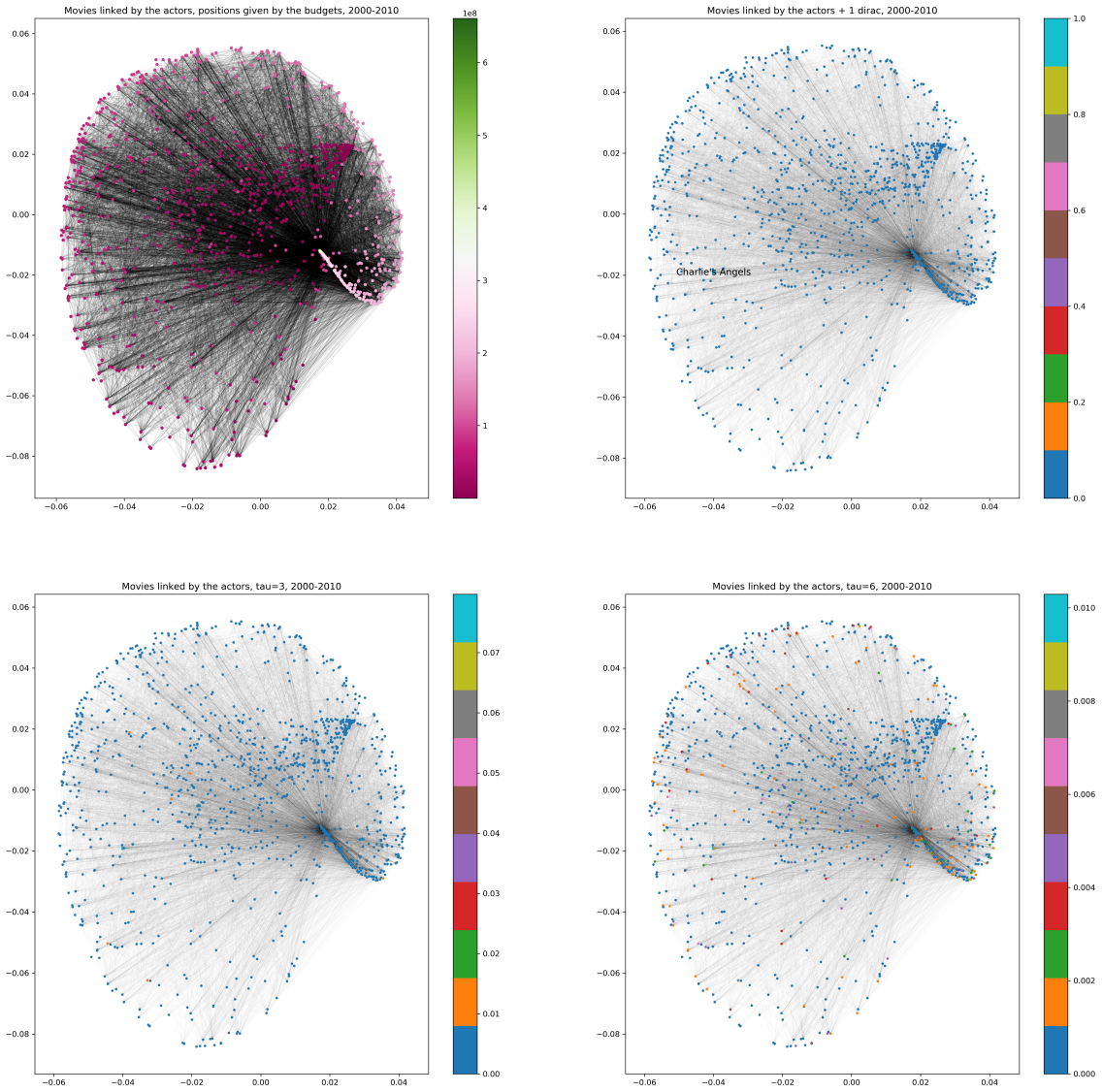


Figure 7: Graphs of the actors for the decade 2000-2010, with the node coordinates calculated with the graph based on the budget.

On the Figure 7, we can see four graphs. On the second row, the graphs show how the Dirac impulse spreads onto the graph. Over different iterations of the process we always obtained the same kind of results: the impulse spreads over nodes that are very far from each other in regard of their budget. We do not see any segregation. This would mean that the actors and the crew are not limited to play in one type of movie. You could find common people in movies with very different budgets.

3 Conclusion

In conclusion, the budgets and revenues of the movies industry increased over the years, in correlation with the global economy. The most represented genres are always drama, comedy and action, and represented more than half of the movies. As time passes, the popularity of movies with high budget increased, leading to larger return on investments. However, the ratings of the movies are not correlated with their budgets. Finally, we observed that there is no segregation based on the budget for the actors. They play in very different type of movie, and the high budget movies are not a closed-circle.