Hacking into the Movie Industry

Team 9: Jimin Wang, Rui Chen, Shichao Jia, Zhuoyue Wang

Introduction

Current online movie databases store have harvested and stored a huge amount of information covering almost every aspect of movies, and such a myriad of data can be analyzed to provide insights into the movie industry. Having chosen the TMDB data set, we expect to build a network for the purpose of investigating and visualizing how different roles cooperate and interact with each other in the movie industry.

Data Set

There are two data sets named "tmdb_5000_movies" and "tmdb_5000_credits", which can be downloaded directly as csv-files from Kaggle.

"tmdb_5000_movies" contains almost 5000 movies and each of them has 20 features. Several main features like *Movie title*, *budget*, *genres* and *overview* have been displayed in each movie.

"tmdb_5000_credits" has the information of filmmakers and actors.

Goal

Our goal is to explore the underlying relationship between objects that we are interested in behind the movies based on the networks the we construct. We are planning to graph processing methods to analyze the characteristics of the data set, e.g., learning from graph, graph signal processing.

Milestones

1. Data pre-processing
2. Data visualization
3. Relationship extraction
4. Network establishment
5. Network analysis

Network

Cooperation of Companies

Node: All the movie production companies.
Edge: Build edge when two production companies invested in the same movie.

Cast and Crew Network

Node: Use the casts and the crews as the node
Edge: When they contributed for the same movie, we build an edge between these two nodes, and the weight of the edge depends on the times that they have cooperated.