Movie-Genre Analysis

Selcuk Gulcan
Cs-529 Project

27.12.2021

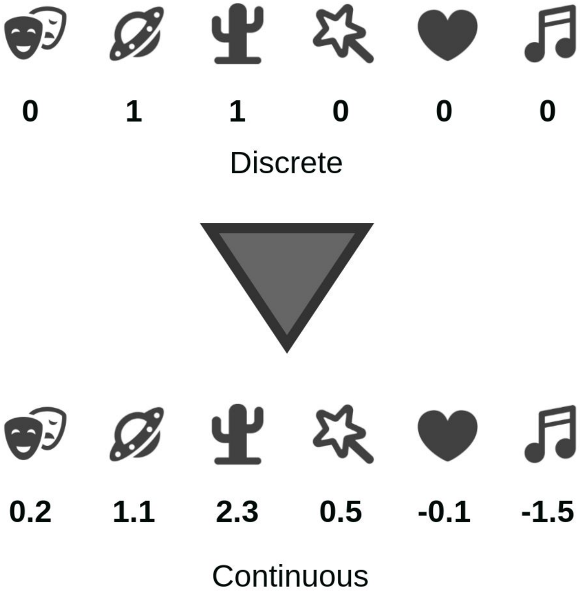
Outline

Problem

- Traditional representation of genres is categorical
- A movie contains a genre label or not (Binary)
- Can we convert categorical genre labels into a continuous vector?
- Each value in genre vector shows the effect of the corresponding particular genre on the movie

Title: Dark Tower

Genre: Western, Sci-fi

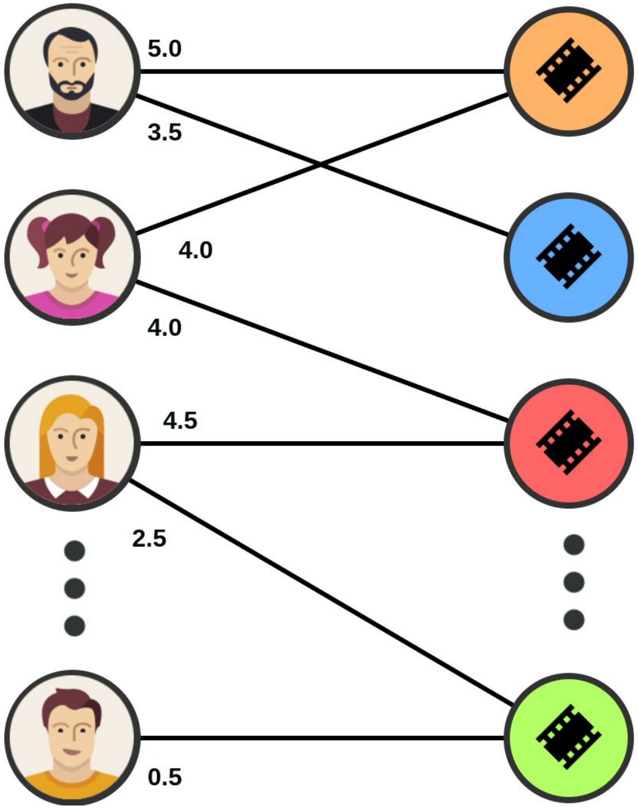


Continuous Vector Representation Benefits

- This representation allows us to do:
  - in-movie genre comparison
  - in-genre movie comparison
- Correctly labeling genres has a marketing value
- Recommender systems use genre information
  - Better genre information = Better recommandation

	0.2	1.1	2.3	0.5	-0.1	-1.5
	0.4	-3.4	-1.0	-1.3	0.3	-2.8
	-1.2	0.3	-2.4	0.6	0.6	-1.5
	0.8	0.5	0.3	0.5	0.7	-1.0
	0.6	0.7	0.3	5.1	-0.1	0.3
	1.6	0.4	0.7	0.0	0.0	-1.3

Dataset - User-Movie Rating Graph

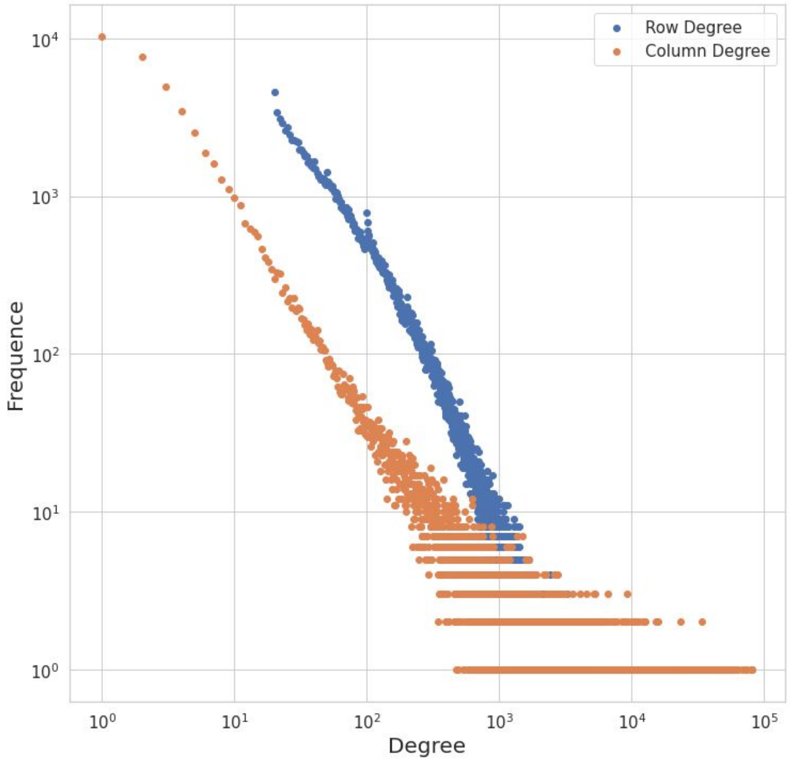


- Rating dataset known as movielens-25m[1]
- 2-mode bipartite user-movie graph
- Genre labels of the movies are also known

1. F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. Acm transactions on interactive intelligent systems (tiis) 5, 4 (2015), 1–19.

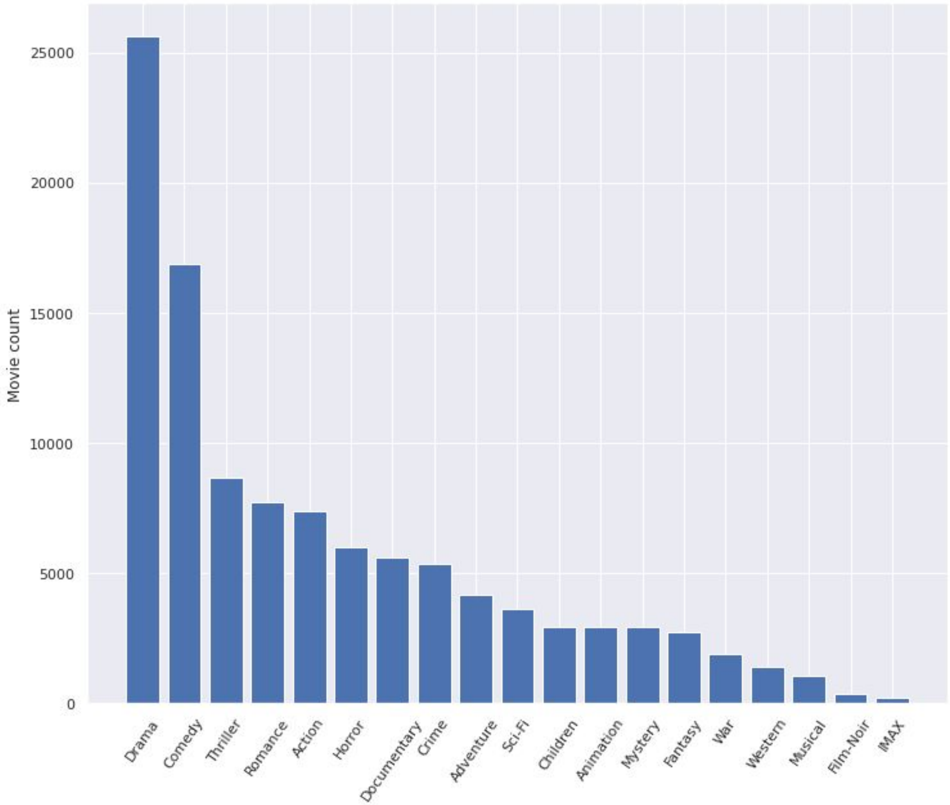
Dataset - Graph Statistics

- User count: 162,541
- Movie count: 59,047
- Rating count: 25,000,095
- Minimum user degree: 20
- Minimum movie degree: 1



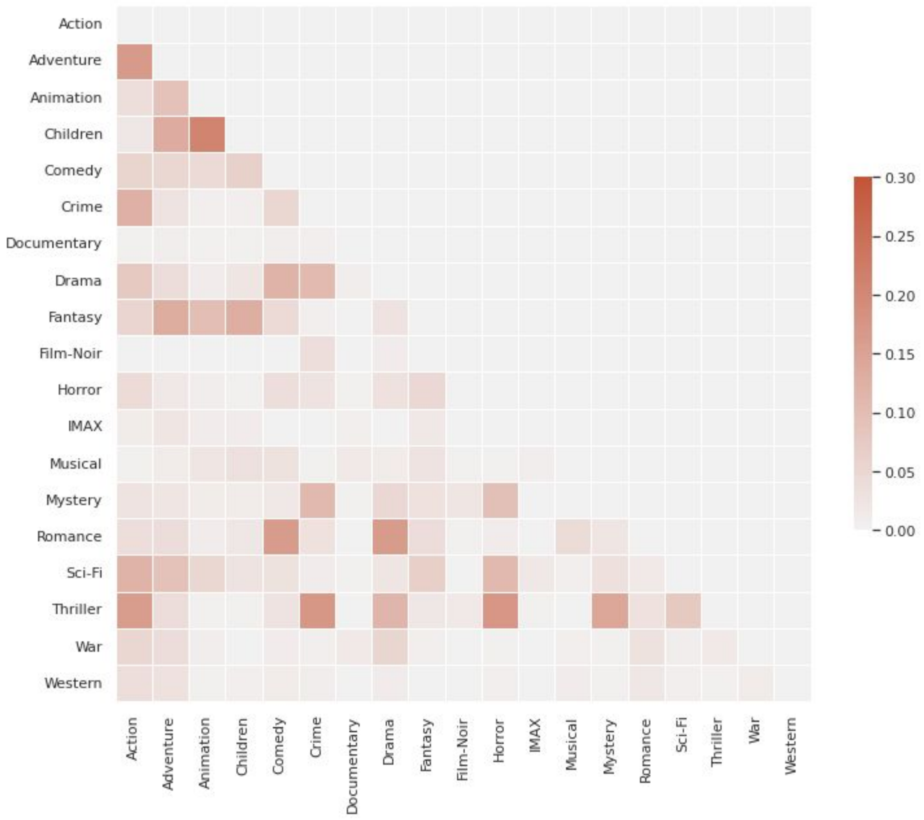
Dataset - Genre Statistics

- 19 genres
- 4 genres removed:
  - Documentary
  - Animation
  - Film-Noir
  - IMAX
- Non-uniform distribution



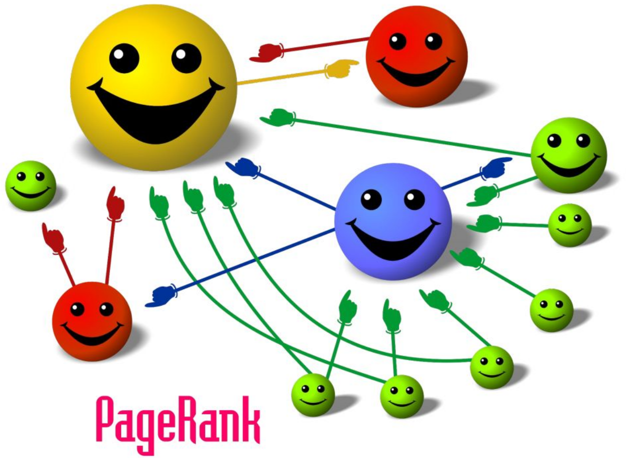
Dataset - Genre Statistics

- Noticeable correlations between some genres
  - Children - Animation
  - Thriller - Horror
  - Thriller - Crime
  - Adventure - Action



Background - Topic Specific Pagerank[2]

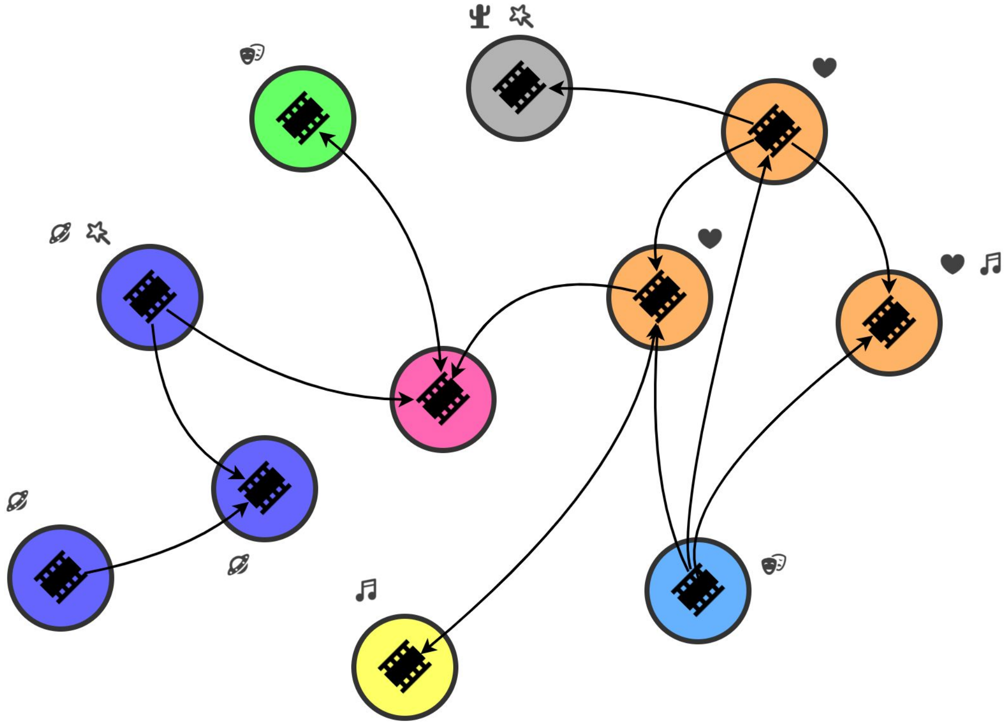
- Calculates the probability that a random surfer will land on a page
- Surfer does not always follow the edges
- Biased teleportation = Topic specific pagerank



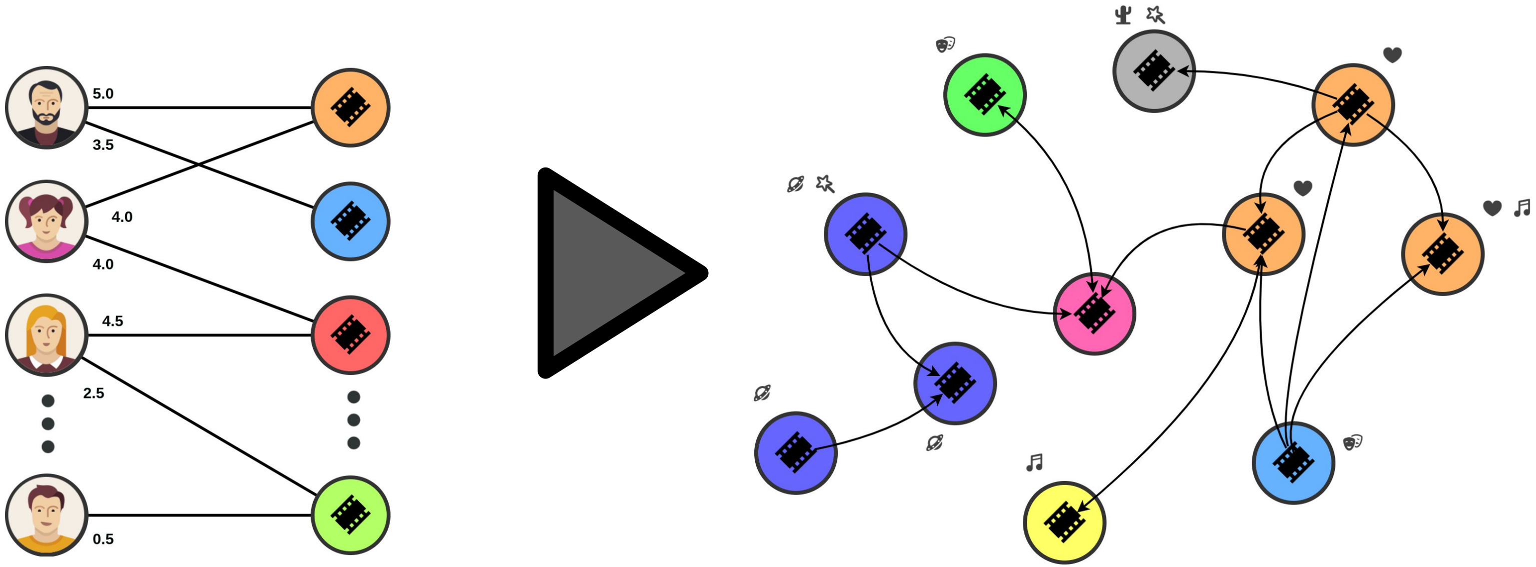
2. Taher H Haveliwala. 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. IEEE transactions on knowledge and data engineering 15, 4 (2003), 784–796.

Methodology - Movie Influence Graph

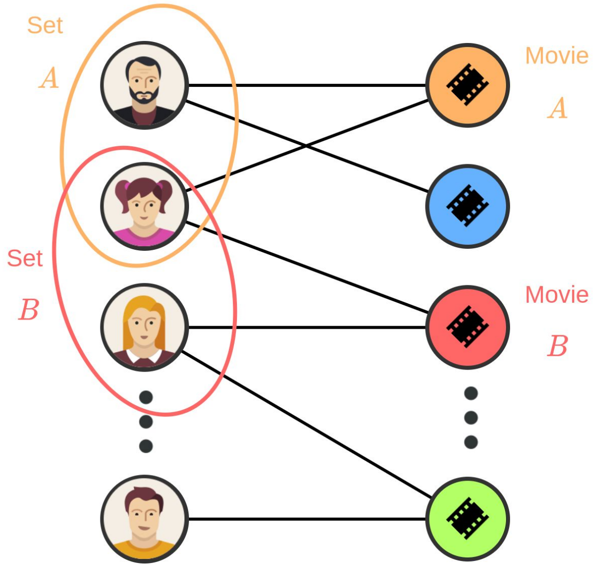
- Weighted and directed movie-movie graph
- Weight shows how much a movie influences another
- Each movie node is in one or more teleportation sets
- Run pagerank for each genre



Methodology - Movie Influence Graph



Methodology - Movie Influence Graph Generation

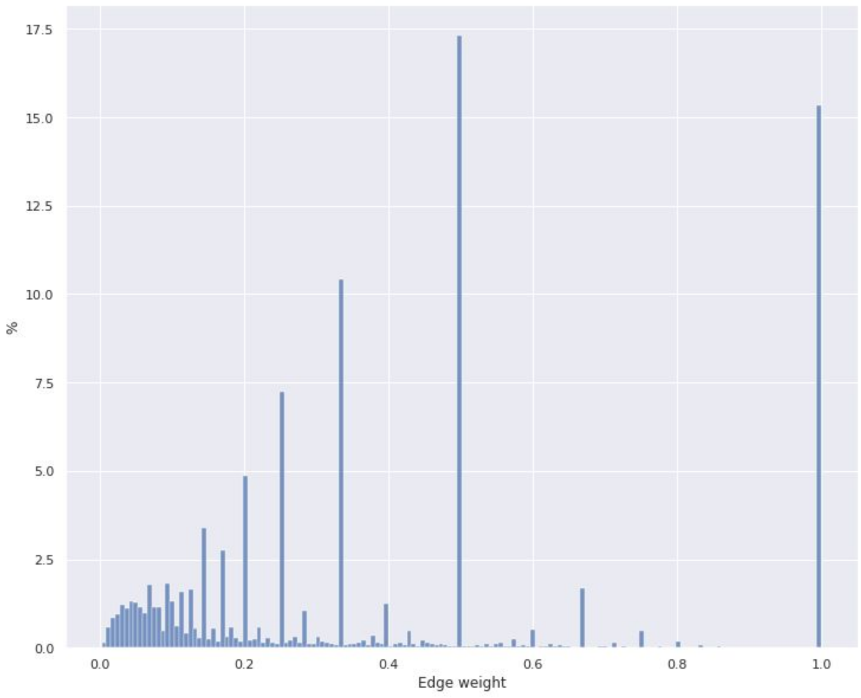


- Each movie has a user set
- There is an edge between two movie vertices if their user sets intersect
- In other words, if at least one person watched both movies.
- How about weights?

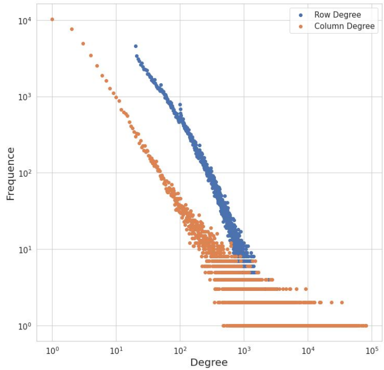
Methodology - Weight Formula Possibilities

1. Intersection $|A \cap B|$
2. Jaccard similarity $\dfrac{|A \cap B|}{|A \cup B|}$
3. Sorenson similarity $\dfrac{2 \times |A \cap B|}{|A| + |B|}$
4. Asymmetric weight $\dfrac{|A \cap B|}{|A|}$
5. Intersection over min $\dfrac{|A \cap B|}{min(|A|, |B|)}$

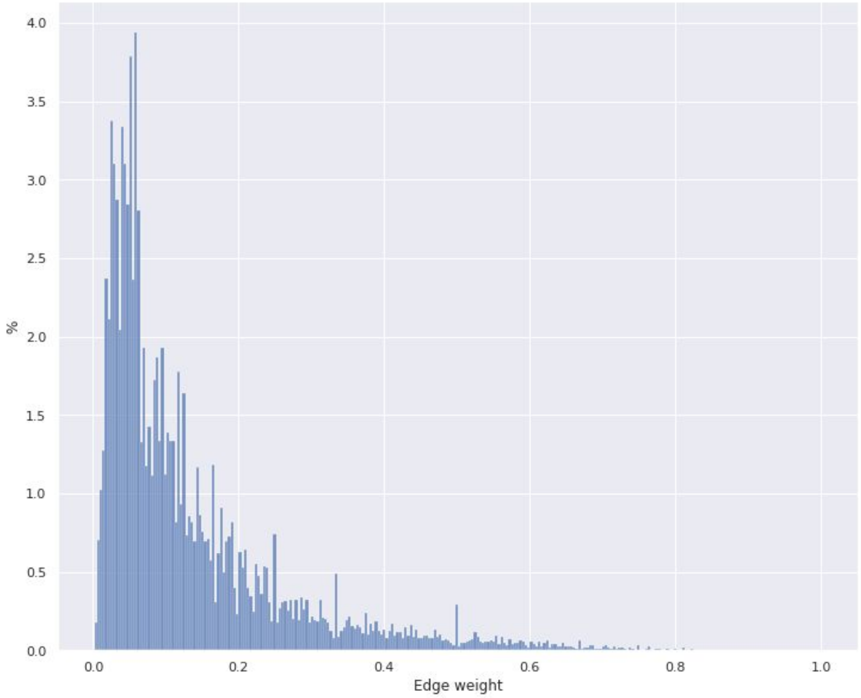
Methodology - Movie Influence Graph Statistics



- 1,364,630,530 edges
- Density: 0.35
- Edge weight frequencies are a bit odd
- Removing movies with less than 15 viewers

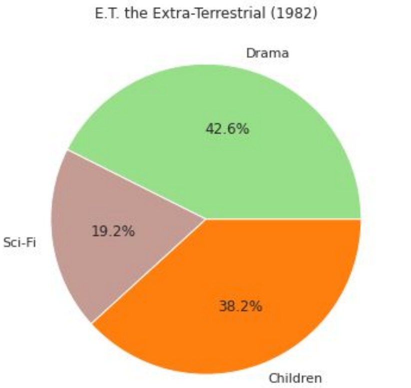


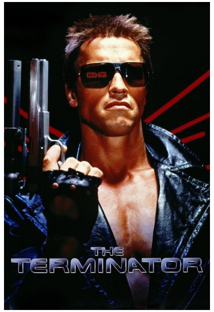
Methodology - Movie Influence Graph Statistics

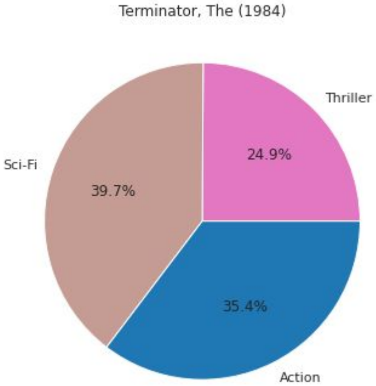


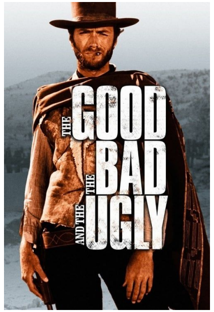
- Vertex count: 20,034
- Edge count: 56,090,310
- Density: 0.14

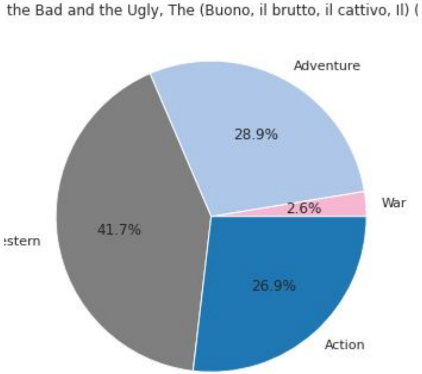
Results - In-Movie Genre Comparison

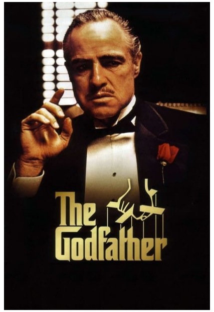


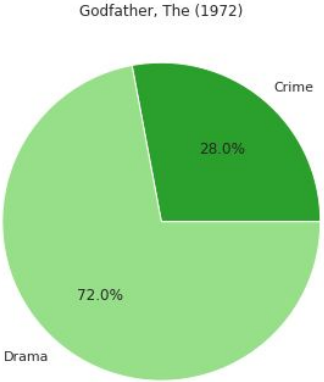




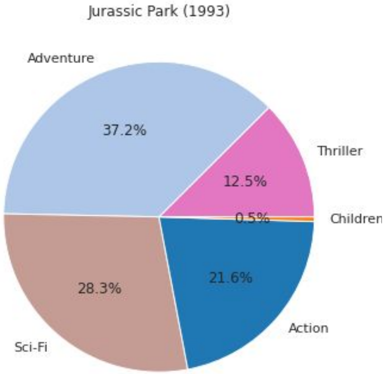


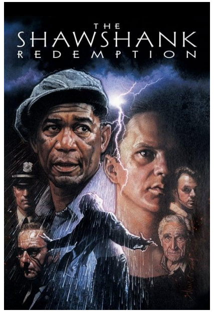


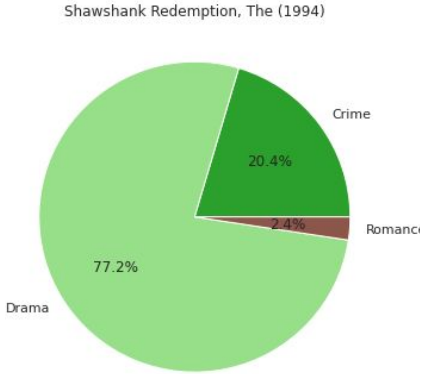


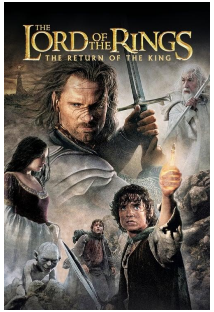


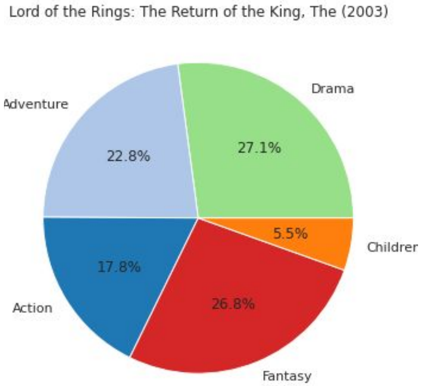
Results - In-Movie Genre Comparison

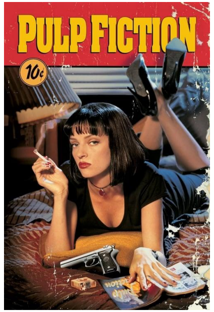


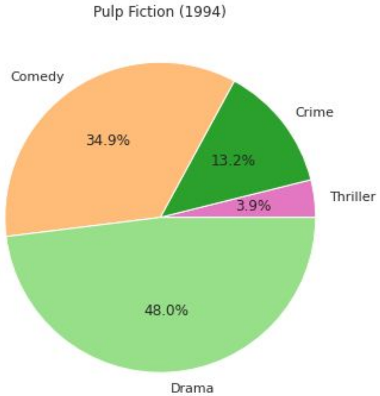




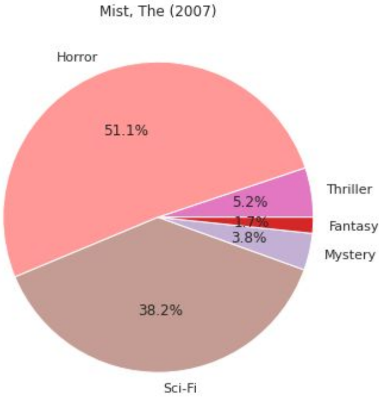


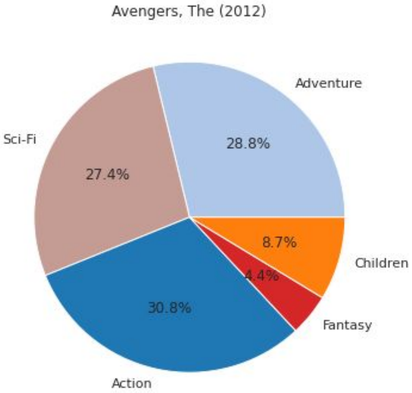


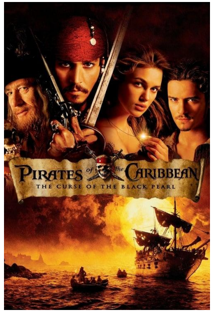


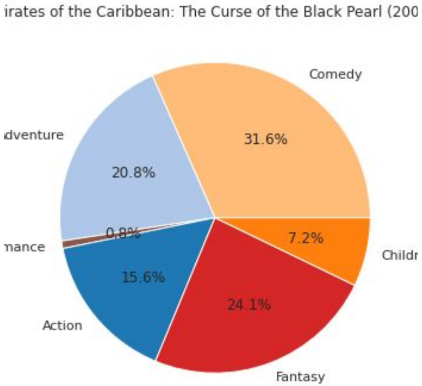


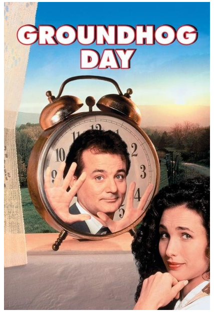
Results - In-Movie Genre Comparison

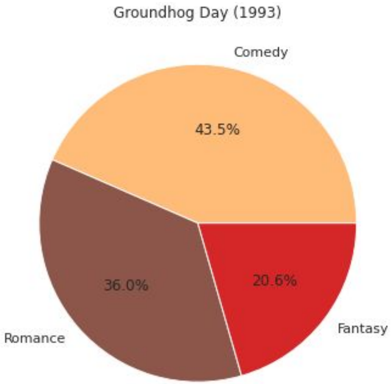




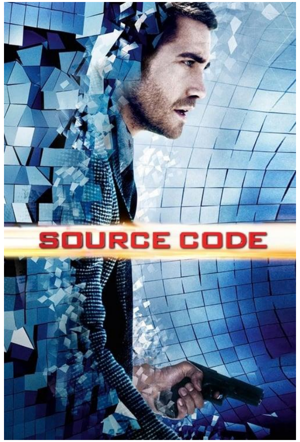


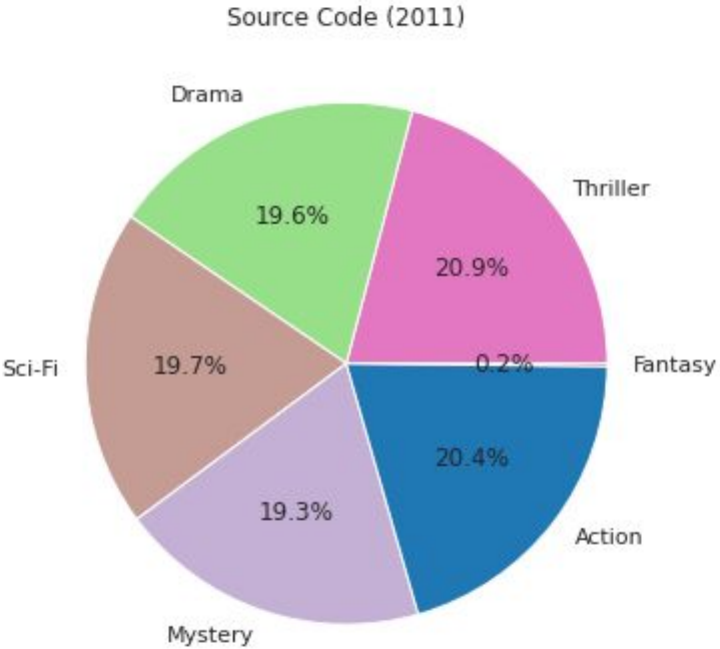






Results - In-Movie Genre Comparison





movielens

Genres

Thriller , Science Fiction , Mystery

IMDb

Action | Drama | Mystery

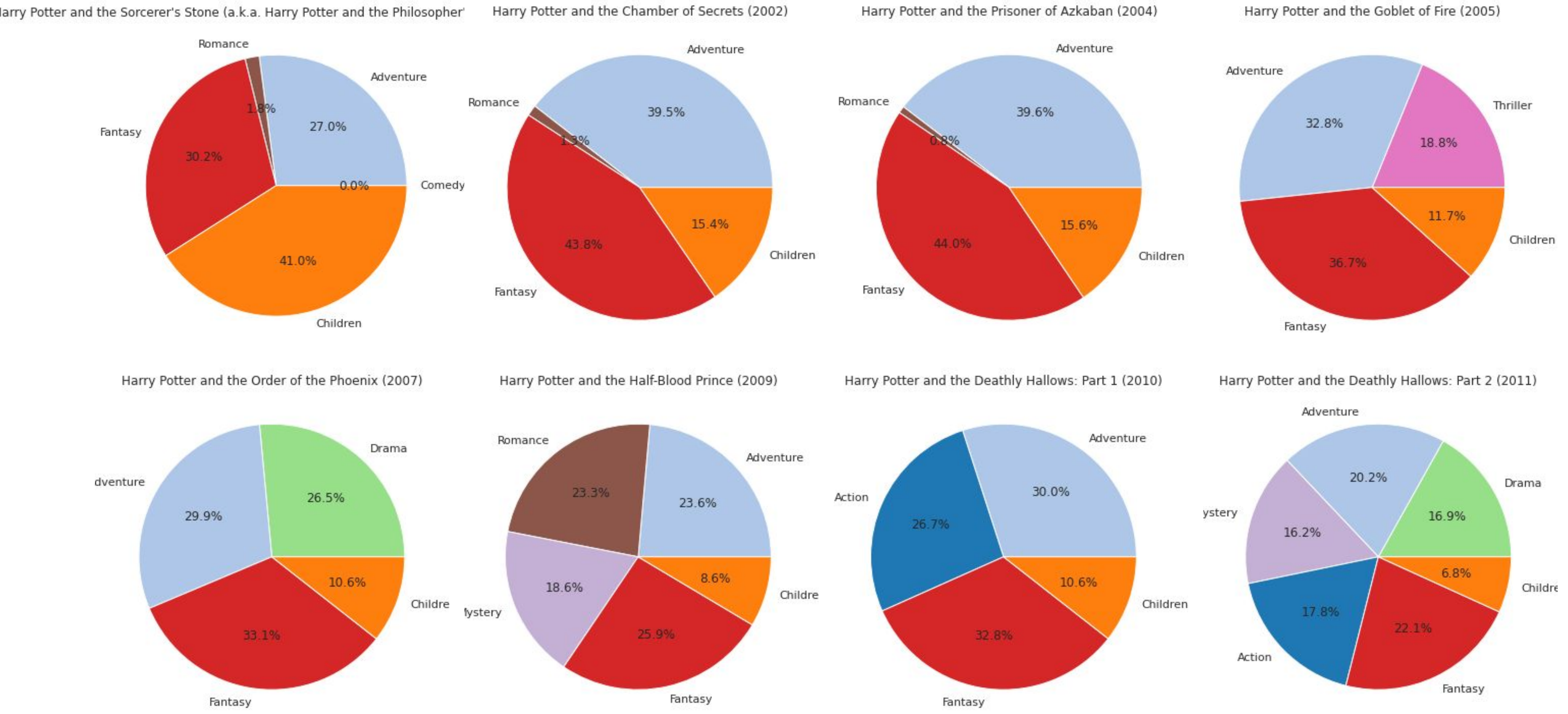
prime video

Science Fiction, Suspense

VUDU – FANDANGO –

Action I Sci-Fi I 2011

Results - In-Genre Movie Comparison

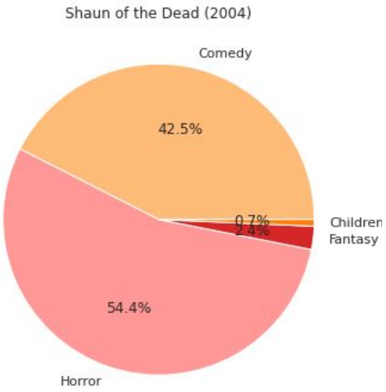


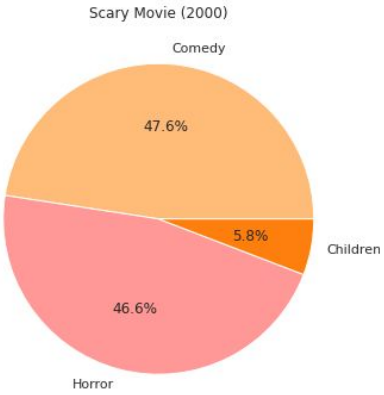
Thanks…

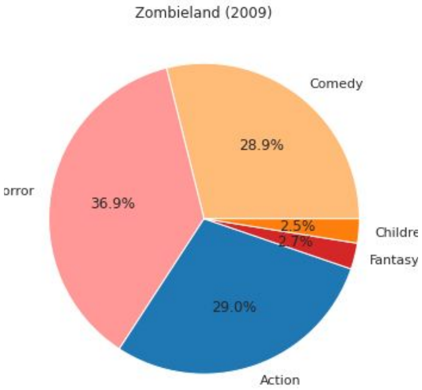
References

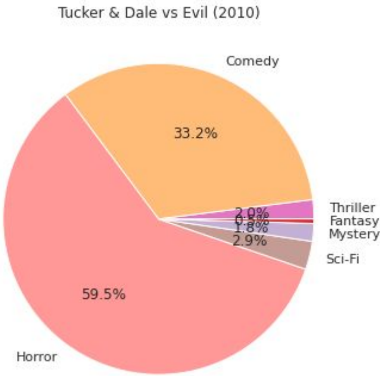
1. F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. Acm transactions on interactive intelligent systems (tiis) 5, 4 (2015), 1–19.
2. Taher H Haveliwala. 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. IEEE transactions on knowledge and data engineering 15, 4 (2003), 784–796.

Extra - Parody Movies (Bad Examples)



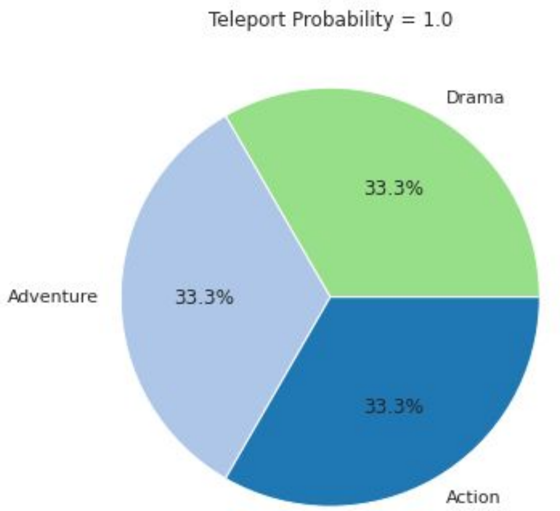


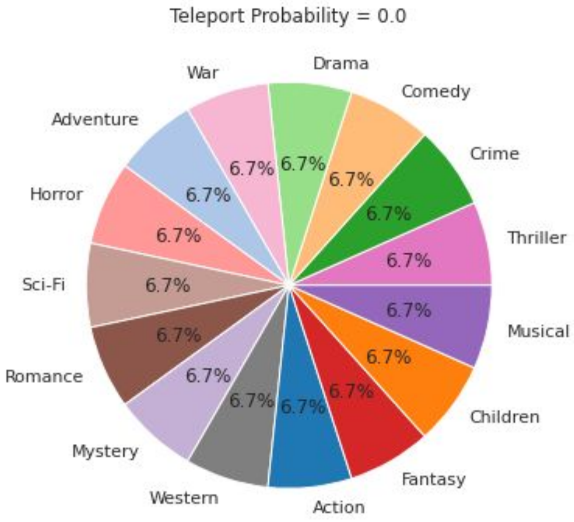




Extra - Teleport Probability

- Teleport probability shows our trust on the initial genre labels
- Good results when it's between 0.15 and 0.30





Extra - Implementation Details

- Problem: Pagerank score for genre x is not comparable to pagerank score of genre y
- Reason: Movie count imbalance between genres
- Solution 1: Different teleport probabilities for different genres
- Solution 2: Genre scores are calculated as:

  (genre pagerank / expected genre pagerank) - 1