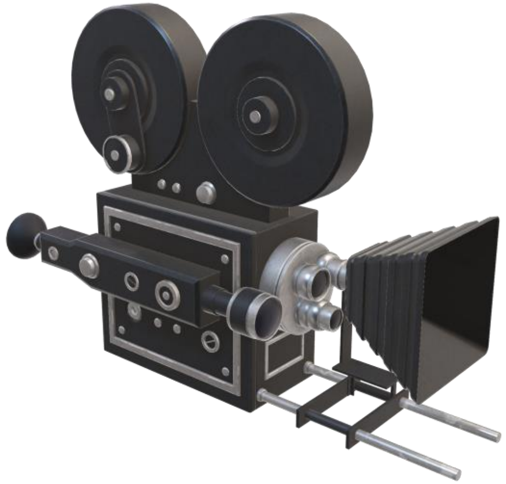
Movie Recommendation Using Graph Neural Networks

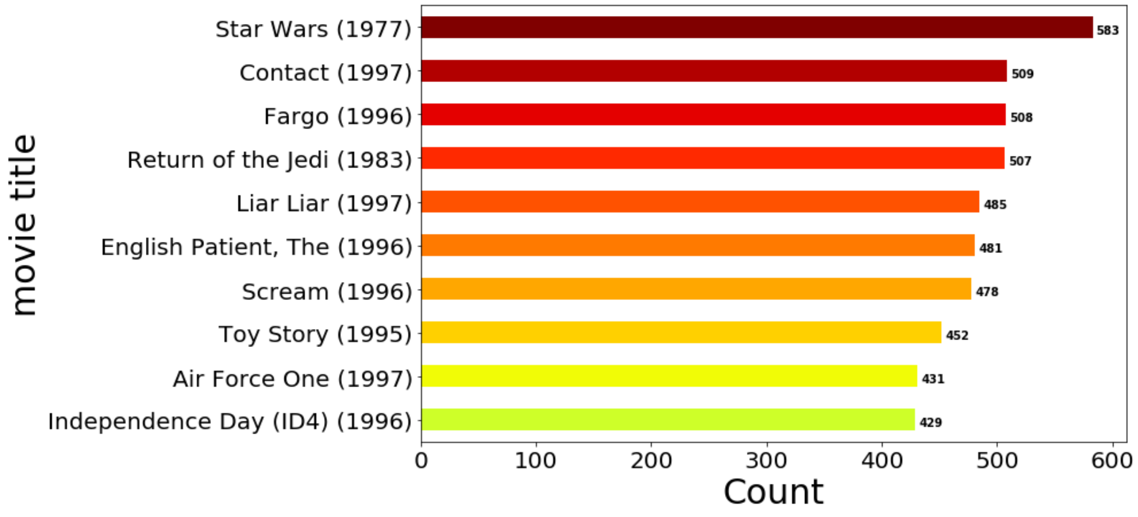
Kuan Tung, Chun-Hung Yeh, Hiroki Hayakawa and Jinhui Guo

Introduction

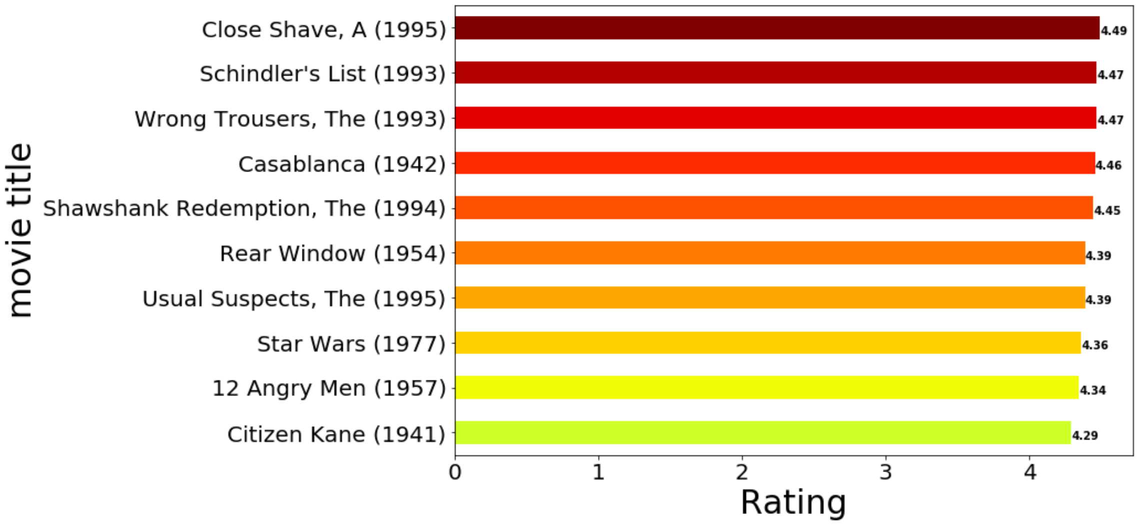


- Dataset: Movielens 100k
  - 100,000 ratings (1-5) from 943 users on 1682 movies
  - Each user has rated at least 20 movies.
  - Simple demographic info for the users (age, gender, occupation, zip)
- Network structures of movies and users
  - Movies: top rated, highest rating score, genres
  - Users: genders, occupation
- Similar Movies according to Genres
- Graph Convolutional Matrix Completion

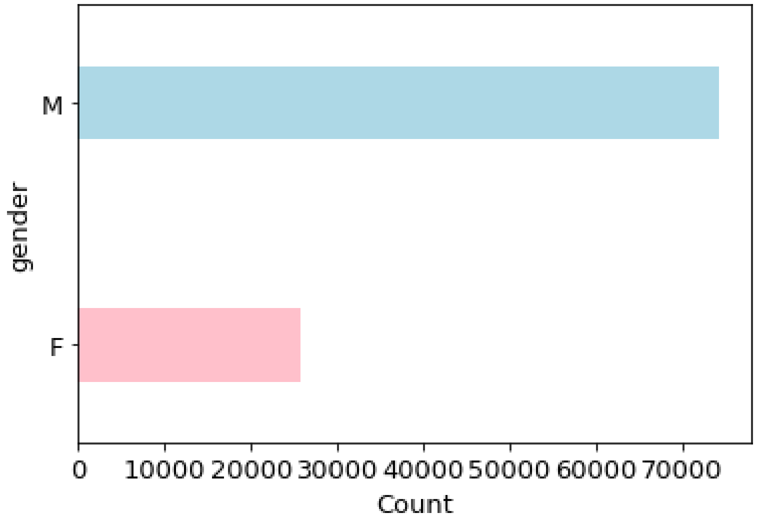
Top rated movies

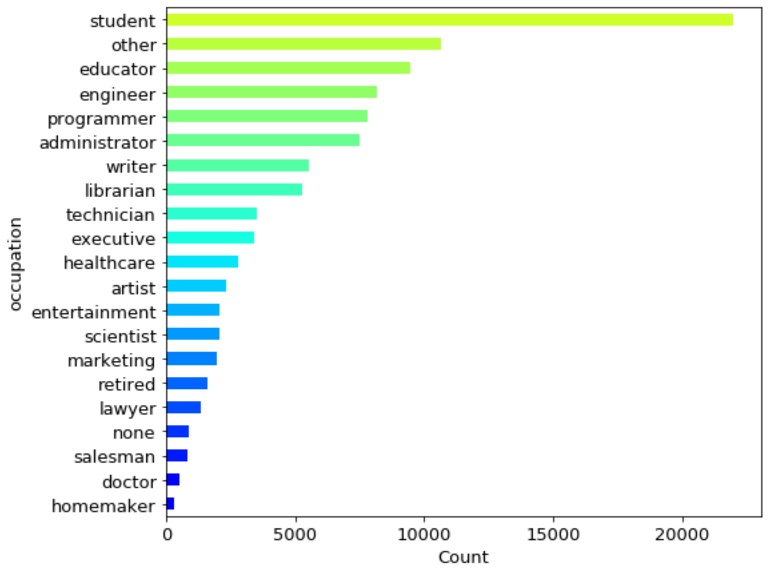


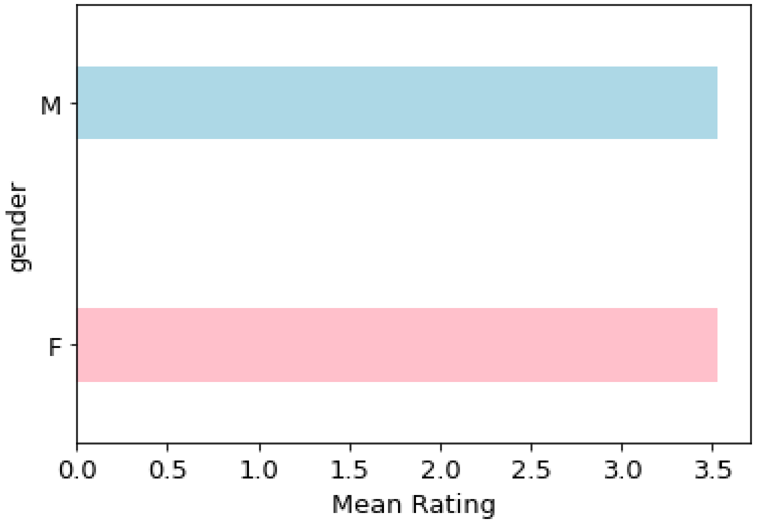
Top rating movies

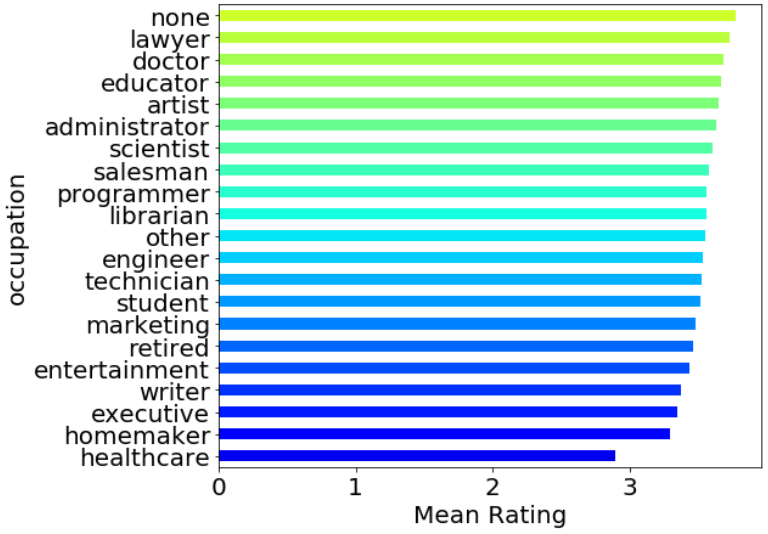


Gender and occupation









Movies and Genres

Network structure

Dimension reduction and clustering

Overview of Genres

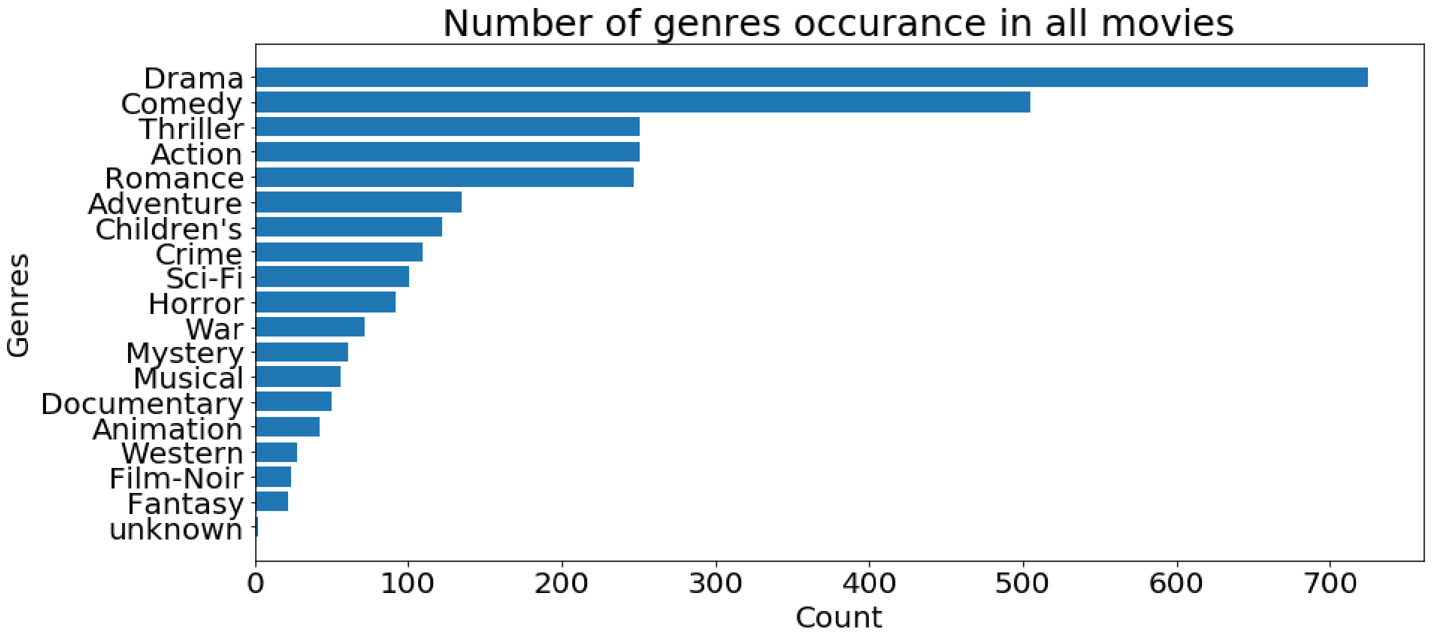
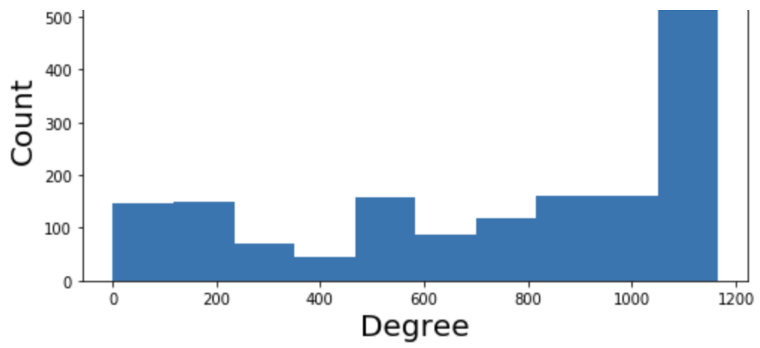
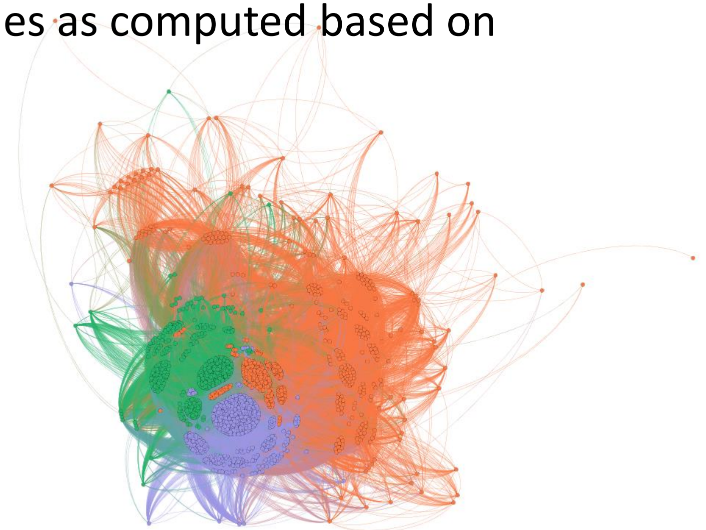


TABLE I
MOVIE GRAPH STATISTICS

Statistics	**Value**
Nodes	1682
Edges	642422
Average Degree	763.879
Diameter	6
Density	0.454
Modularity	0.168
Clustering Coefficient	0.797
Connected Component	1

- Nodes: movies
- Edges: the distance between two movies as computed based on one-hot vectors of their genres





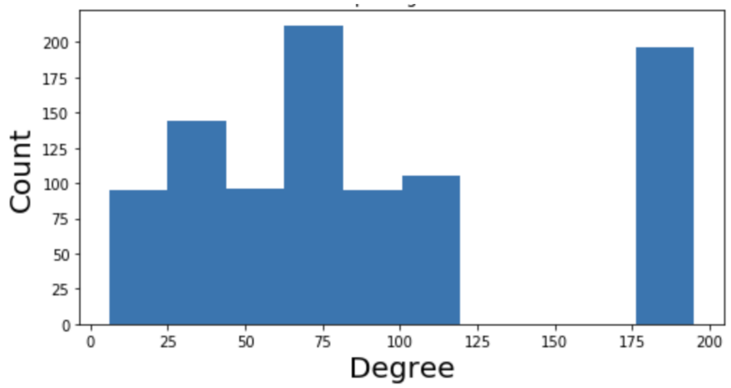
Movie graph

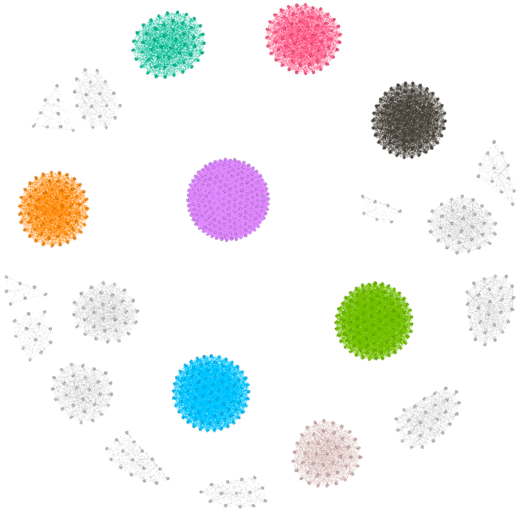
- Degree distribution
- Visualization
  - Link: https://dinotuku.com/ntds-2019-project-team-7/exploration/figure/genres-graph/

TABLE II
USER GRAPH STATISTICS

Statistics	Value
Nodes	943
Edges	41326
Average Degree	87.648
Diameter	1
Density	0.093
Modularity	0.744
Clustering Coefficient	1.0
Connected Component	21

- Nodes: User
- Edges: the distance between two users as computed based on one-hot vectors of their jobs



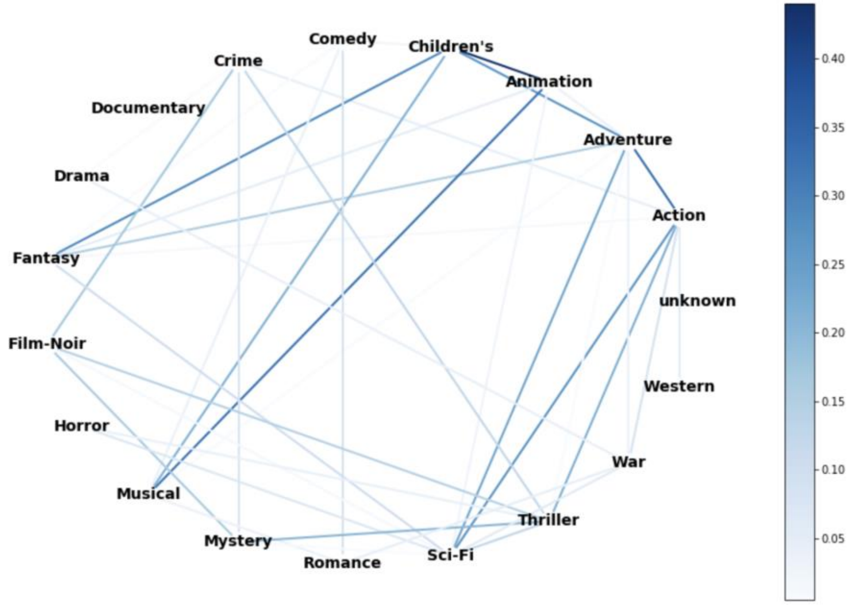


User graph

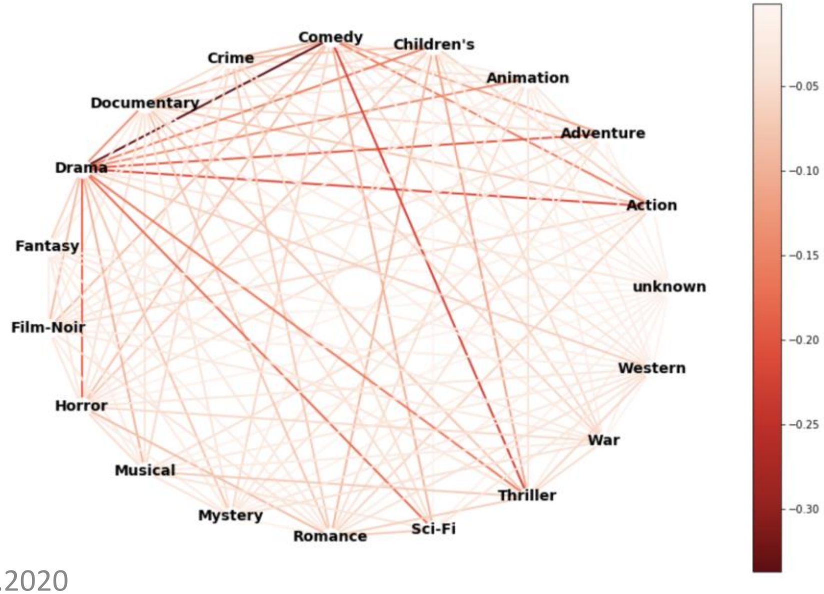
- Degree distribution
- Visualization
    - Link: https://dinotuku.com/ntds-2019-project-team-7/exploration/figure/genres-graph/

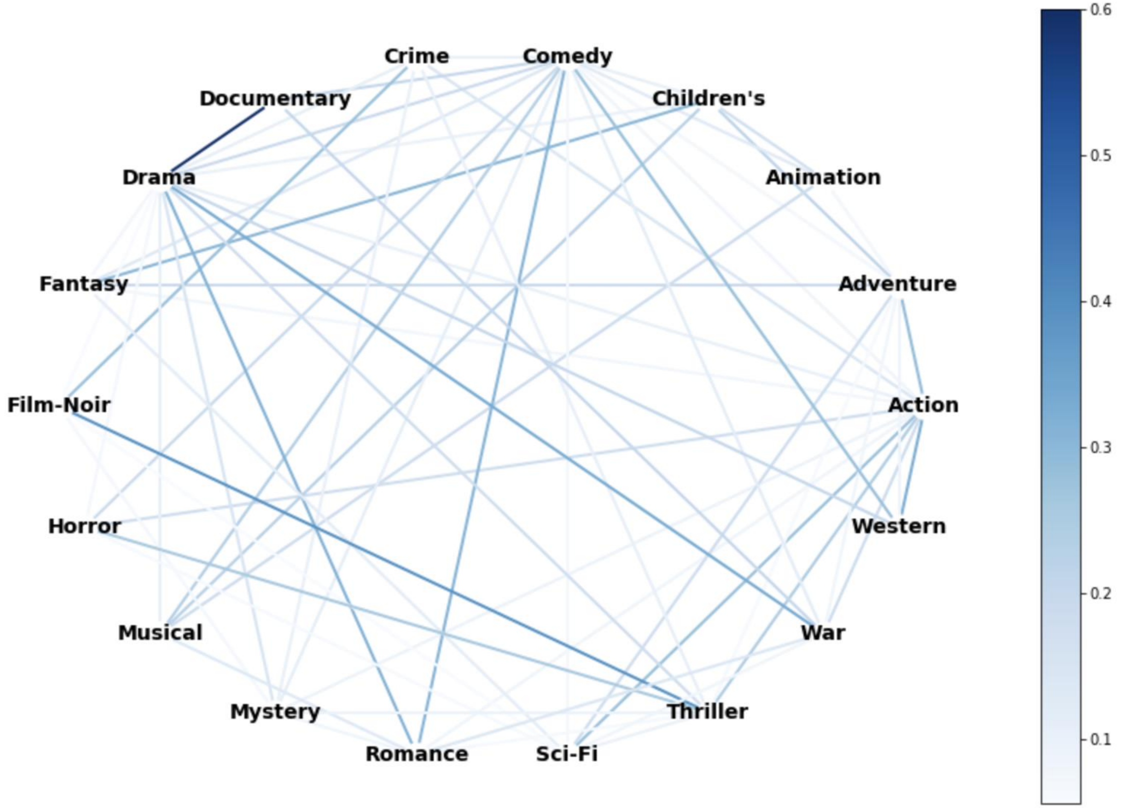
Genres Graph Using Co-occurrence and Correlation Matrix

Positive weights only



Negative weights only

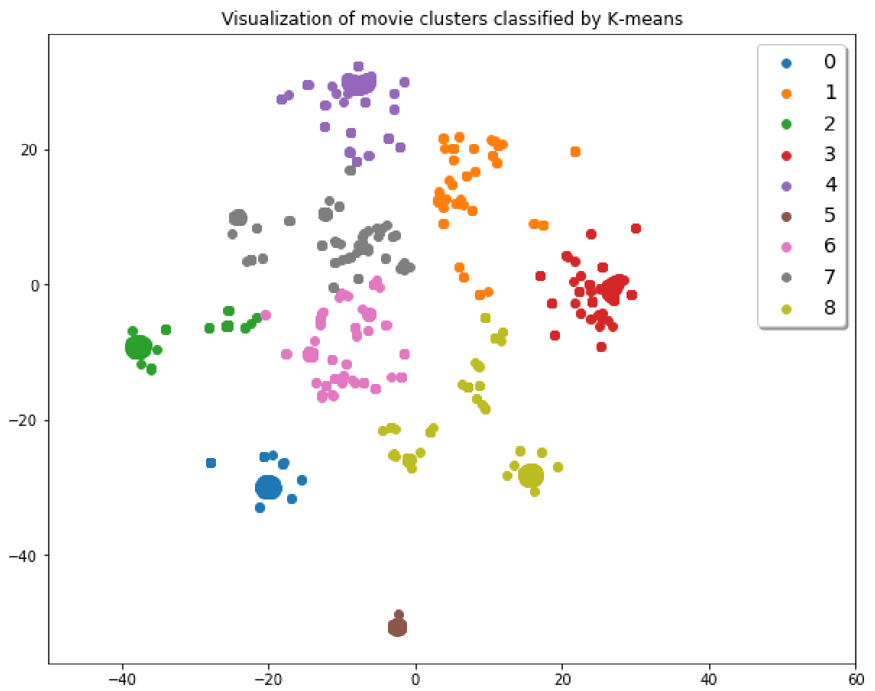


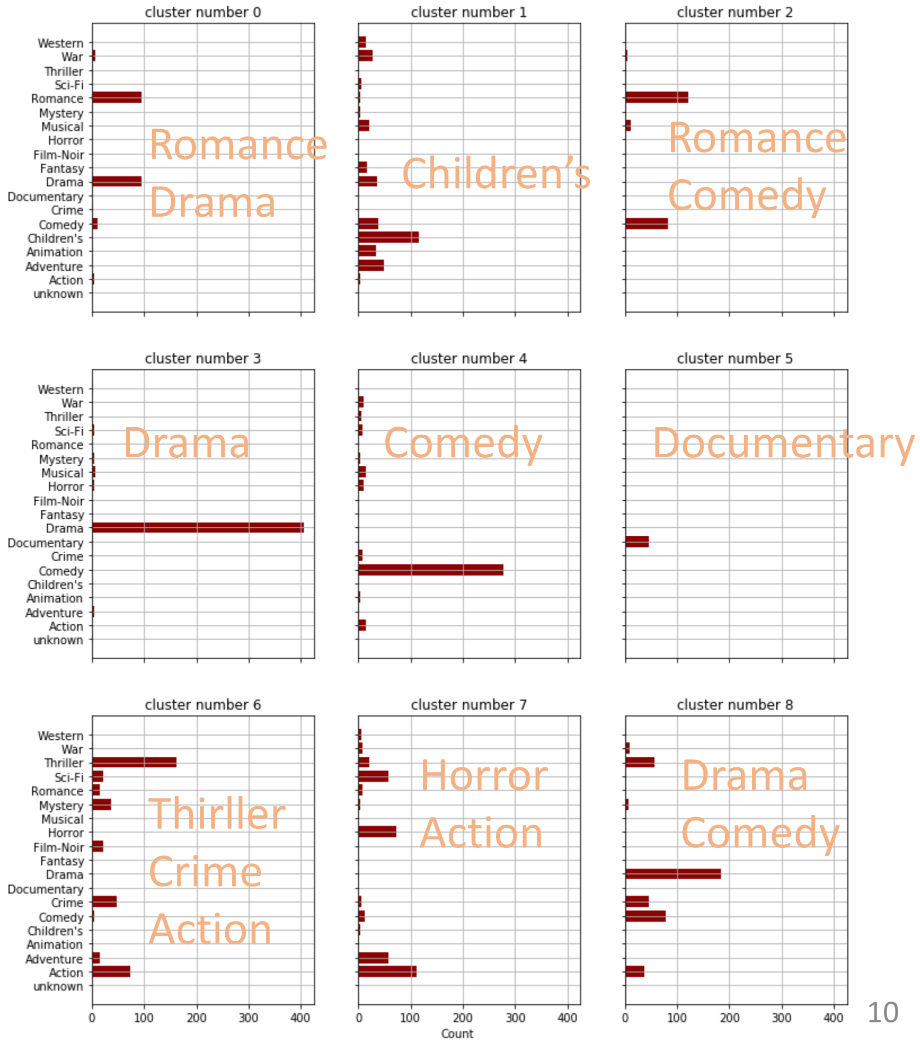


The most apparent pair that two genres co-occur is **Documentary and Drama** (edge weight: 0.6)

Similar Movies According to Genres

- Dimension reduction: t-Distributed Stochastic Neighbor Embedding (t-SNE)
- Clustering: K-means





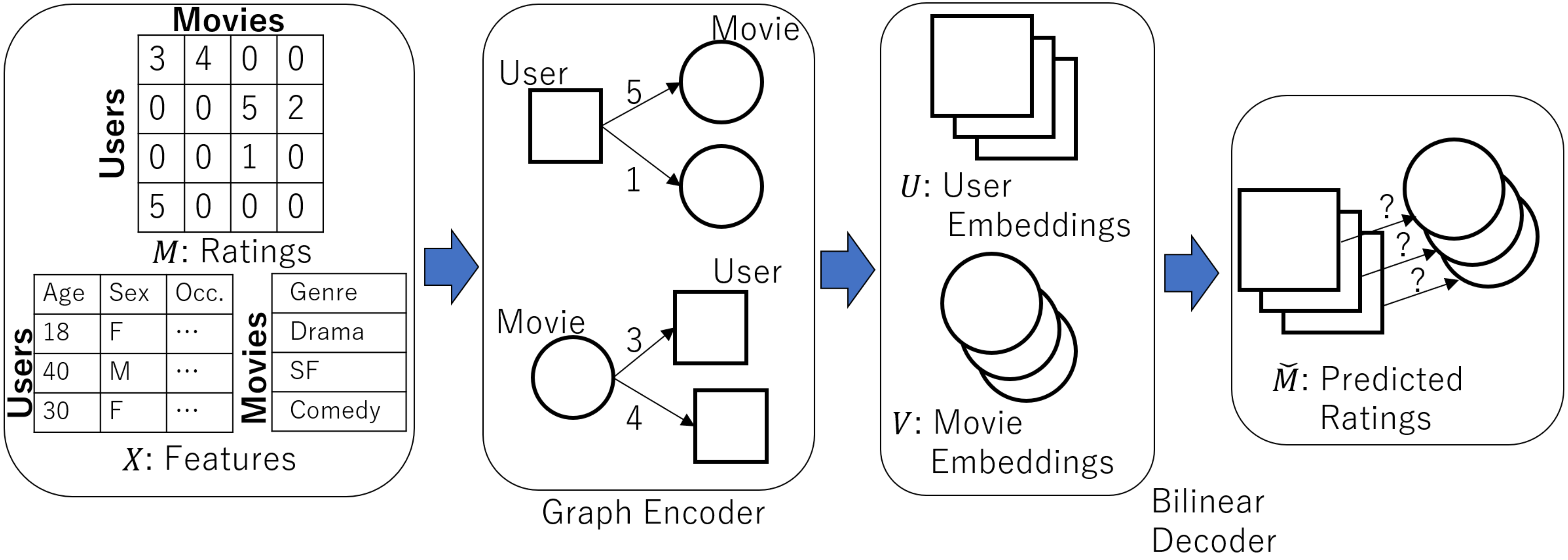
Prediction of Movie Ratings

Graph convolutional matrix completion (GC-MC)

Matrix factorization

Model comparison

Graph Convolutional Matrix Completion Method



Graph Convolutional Matrix Completion (Cont.)

❑Graph Convolutional Encoder

$$\mu_{j\rightarrow i,r} = \frac{1}{c_{ij}} W_r x_j, \quad c_{ij} = |\mathcal{N}_i|$$ (Info. Transfer: movie j to user i)

$$h_i = \sigma \left[\text{accum} \left(\sum_{j\in\mathcal{N}_{i,1}} \mu_{j\rightarrow i,1}, \cdots, \sum_{j\in\mathcal{N}_{i,R}} \mu_{j\rightarrow i,1} \right) \right]$$ (Accumulation)

$u_i = \sigma(Wh_i)$ or $u_i = \sigma(Wh_i + W_2^f f_i)$ with $f_i = \sigma(W_1^f x_i^f + b)$,

(Activation to obtain user embedding vector)

❑Bilinear Decoder

$$\check{M}_{ij} = \mathbb{E}_{p(\check{M}_{ij}=r)}[R] = \sum_{r\in R} r p(\check{M}_{ij} = r), \quad p(\check{M}_{ij} = r) = \frac{e^{u_i^T Q_r v_j}}{\sum_{r\in R} e^{u_i^T Q_r v_j}}$$

(Calculate expectation of rate from user i to movie j)

Models

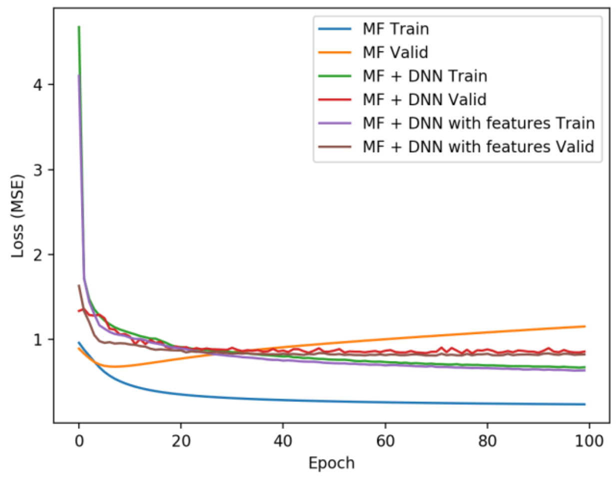
- Two based on graph convolutional matrix completion (GC-MC) methods
  - The hidden sizes of the stack graph convolution layer and dense layer were set as 500 and 75
  - Dropout (dropout rate = 0.7) was applied
- Three based on matrix factorization:
  - Sum of the dot value of the two latent factors and the two bias terms
  - Concatenate the two latent factors and feed into DNN
  - Add user and movie features
- For both:
  - The train/validation/test split was 72/8/20
  - Adam was chosen as the optimizer

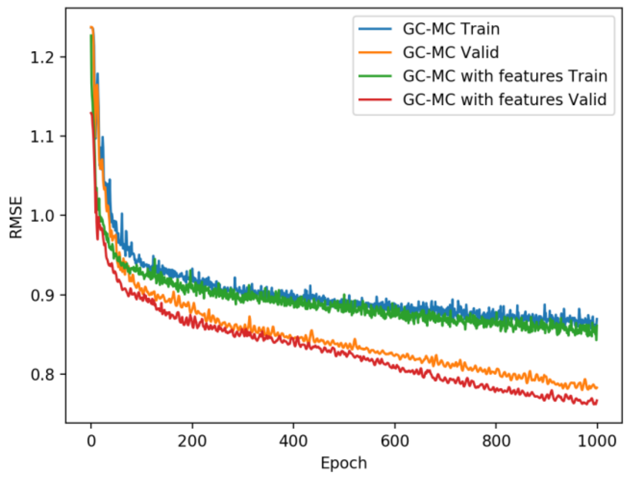
Results

- ❑GC-MC with features is the best model with the lowest RMSE (0.901)
- ❑The validation losses of both graph-based models are surprisingly lower than the training losses, and they keep decreasing

TABLE III
RECOMMENDATION SYSTEM RESULTS

Model	RMSE
MF	0.954
MF + DNN	0.942
MF + DNN with features	0.923
GC-MC	0.908
GC-MC with features	0.901





Conclusions and future prospective

Network structures:

Movie ratings by users with different gender and occupation

Movie graph and user graph (hubs and giant component)

Co-occurrence and correlation between movie genres

Dimension reduction and clustering:

t-SNE and K-means

9 meaningful clusters

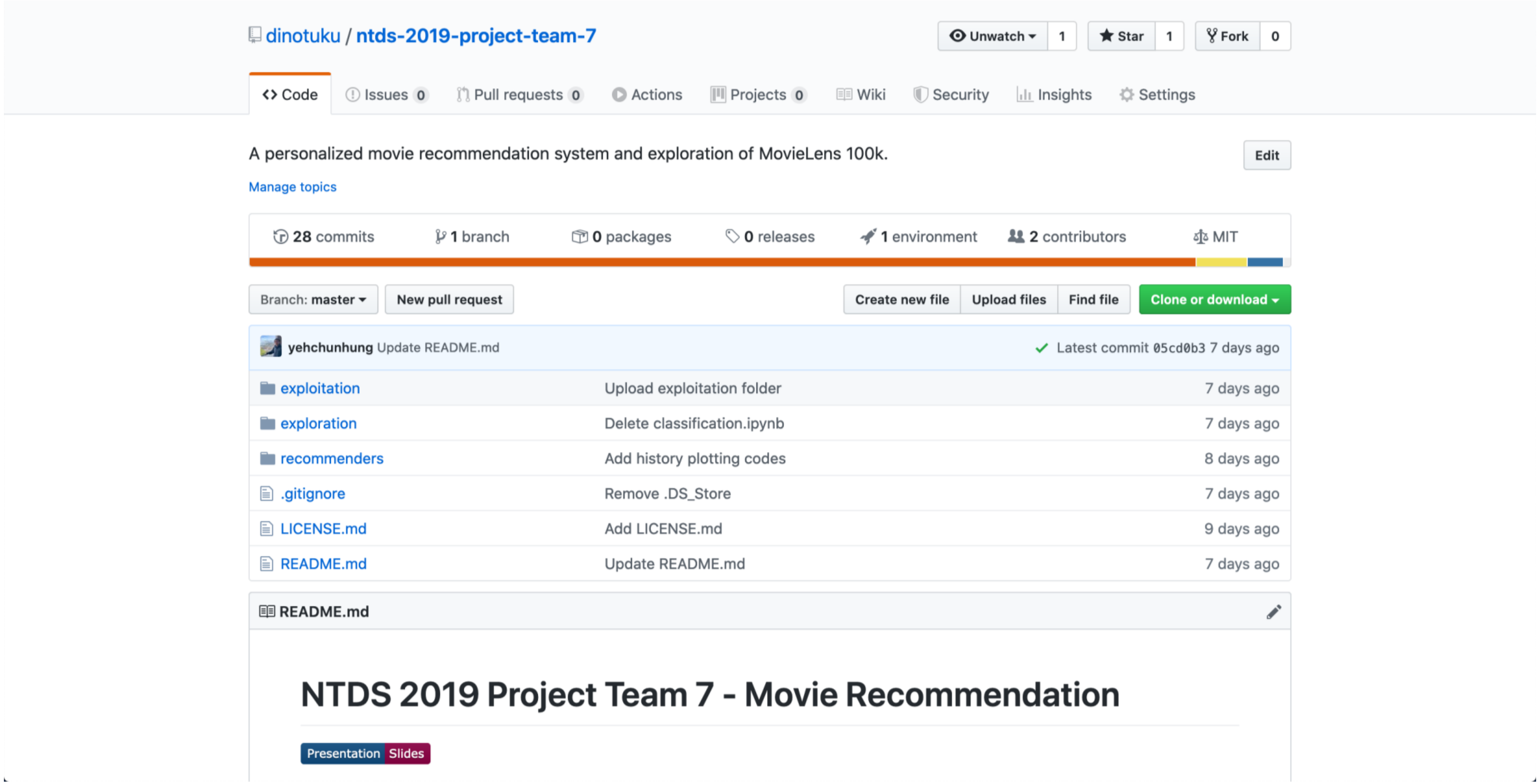
Similar new movie recommendation

Rating prediction

Five models: MF, MF- DNN, MF-DNN with features, GC-MC, and GC-MC with features

GC-MC with features scored the lowest RMSE value of 0.901

Tuning model configurations such as size of hidden layers, accumulation method, the number of training epochs, and optimization methods (e.g. SGD)



Check out our GitHub repository for more information

Thank you for listening !

Any question?

Reference

- Toy Story Poster, https://www.imdb.com/title/tt00114709/
- Harper, F. Maxwell, and Joseph A. Konstan. "The movielens datasets: History and context." *Acm transactions on interactive intelligent systems (tiis)* 5.4 (2015): 1-19.
- MovieLens 100K Dataset, https://grouplens.org/datasets/movielens/100k/
- Berg, Rianne van den, Thomas N. Kipf, and Max Welling. "Graph convolutional matrix completion." *arXiv preprint arXiv:1706.02263* (2017).
- Graph Convolutional Matrix Completion (GitHub repository), https://github.com/riannevdberg/gc-mc
- Matrix Factorization (lecture given by Hung-yi Lee), http://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2017/Lecture/MF.pdf