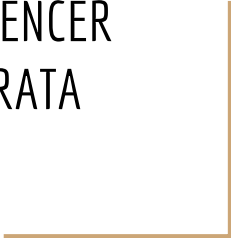




# MOVIELENS ANALYSIS

MICHAEL PAQUETTE, SPENCER  
RUDNICK, NOBUKI HARATA



# ABOUT THE DATA

- MovieLens (University of Minnesota)
- 1 million ratings
- 6000 users
- 4000 movies
- Released in 2003

# OG schema

## Movies

ID  
Title (Year)  
Genres

## Ratings

MovieID  
UserID  
Rating

## Users

ID  
Age  
Sex  
Occupation  
ZIP



# The Cleaning

# New schema

## Movies

ID

Title

Year

Genres

+ RatingsCount

+ AverageRating

+ \*Score

+ \*Genre Indicators

# New schema

## Movies

ID  
Title  
Year  
Genres  
+ RatingsCount  
+ AverageRating  
+ \*Score  
+ \*Genre Indicators

## Users

ID  
Age  
Sex  
Occupation  
Zip  
+ State  
+ Ratings

# New schema

## Movies

ID  
Title  
Year  
Genres  
+ RatingsCount  
+ AverageRating  
+ \*Score  
+ \*Genre Indicators

## Users

ID  
Age  
Sex  
Occupation  
Zip  
+ State  
+ Ratings

## Genres

Genre  
AvgMaleRating  
AvgFemaleRating  
Diff  
AvgScore

# Computing a movie's *score*

- `multiplier <- log(RatingsCount)Q1`
  - `adjustedRating <- AvgRating - 1`
  - `score <- multiplier * adjustedRating`
-



# Movies (TOP 10)

MovieID	Title	Genres	RatingsCount	AvgRating	Score	Year
260	Star Wars: Episode IV - A New Hope	c("Action", "Adventure", "Fantasy", "Sci-Fi")	2991	4.453694	6.597482	1977
318	Shawshank Redemption, The	Drama	2227	4.554558	6.539916	1994
1198	Raiders of the Lost Ark	c("Action", "Adventure")	2514	4.477725	6.499175	1981
527	Schindler's List	c("Drama", "War")	2304	4.510417	6.487183	1993
858	Godfather, The	c("Action", "Crime", "Drama")	2223	4.524966	6.483959	1972
2858	American Beauty	c("Comedy", "Drama")	3428	4.317386	6.445074	1999
2762	Sixth Sense, The	Thriller	2459	4.406263	6.347642	1999
1196	Star Wars: Episode V - The Empire Strikes Back	c("Action", "Adventure", "Drama", "Sci-Fi", "War")	2990	4.292977	6.290205	1980
50	Usual Suspects, The	c("Crime", "Thriller")	1783	4.517106	6.284346	1995
593	Silence of the Lambs, The	c("Drama", "Thriller")	2578	4.351823	6.284002	1991

# Users

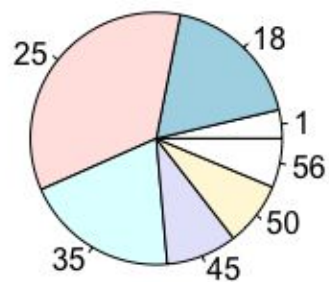
Sex	Age	Occupation	ZIP	State	Ratings
M	25	clerical/admin	00918	PR	c(2987+2i, 1259+3i, 3004+2i, 593+5i, 1271+4i, 778+...

# Genres

	genre	avgMaleRating	avgFemaleRating	diff	avgScore
10	Film-Noir	4.099784	4.022361	0.077423028	4.226584
17	War	3.895079	3.893101	0.001977704	3.672856
13	Mystery	3.664253	3.689506	-0.025253254	3.463338
18	Western	3.658620	3.553551	0.105069037	3.416520
6	Crime	3.719157	3.693331	0.025825870	3.412779

# Stories

### User Ages



#### Key

(age in years)

1: Under 18

18: 18 - 24

25: 25 - 34

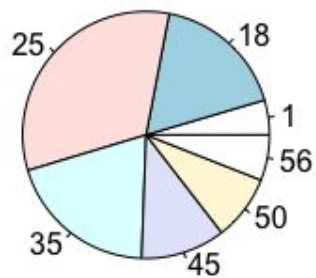
35: 35 - 44

45: 45 - 49

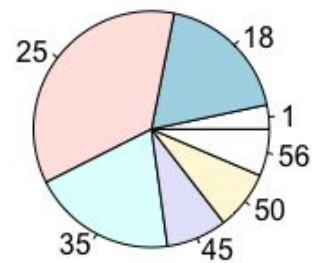
50: 50 - 55

56: 56+

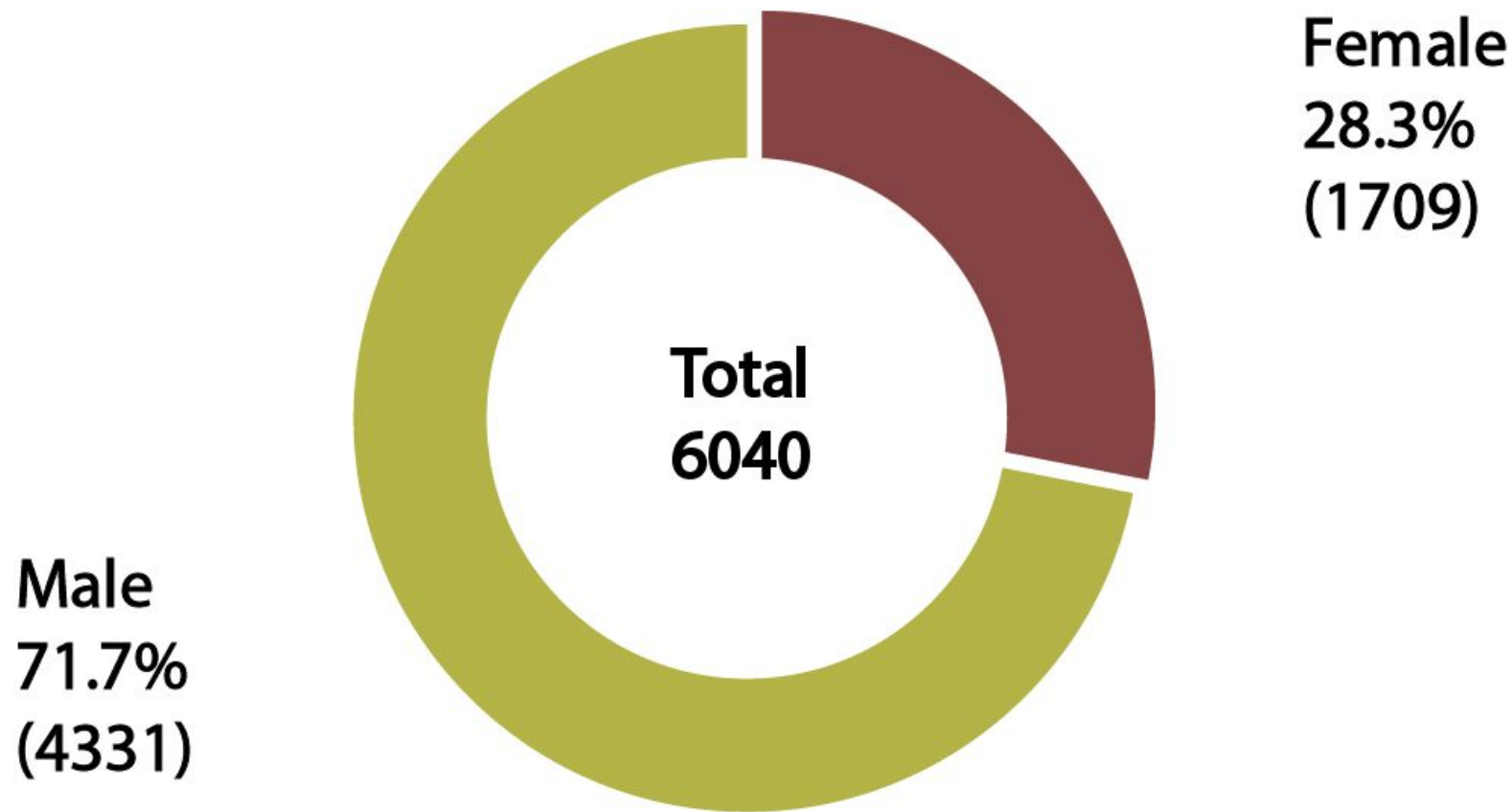
### Female User Ages



### Male User Ages

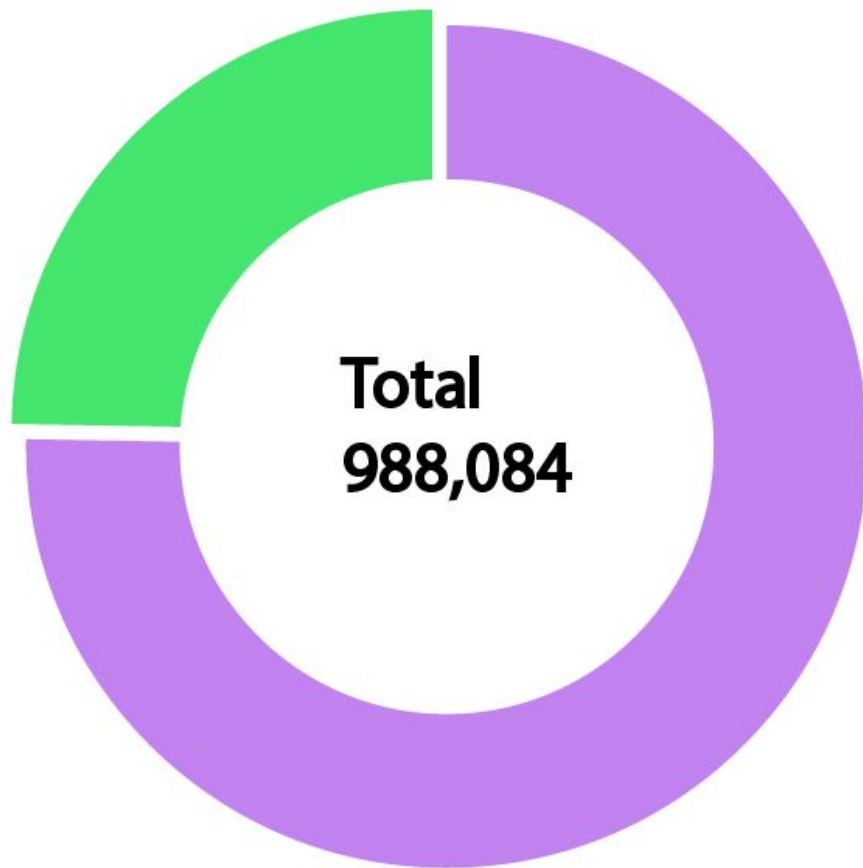


## Number of Users by Sex



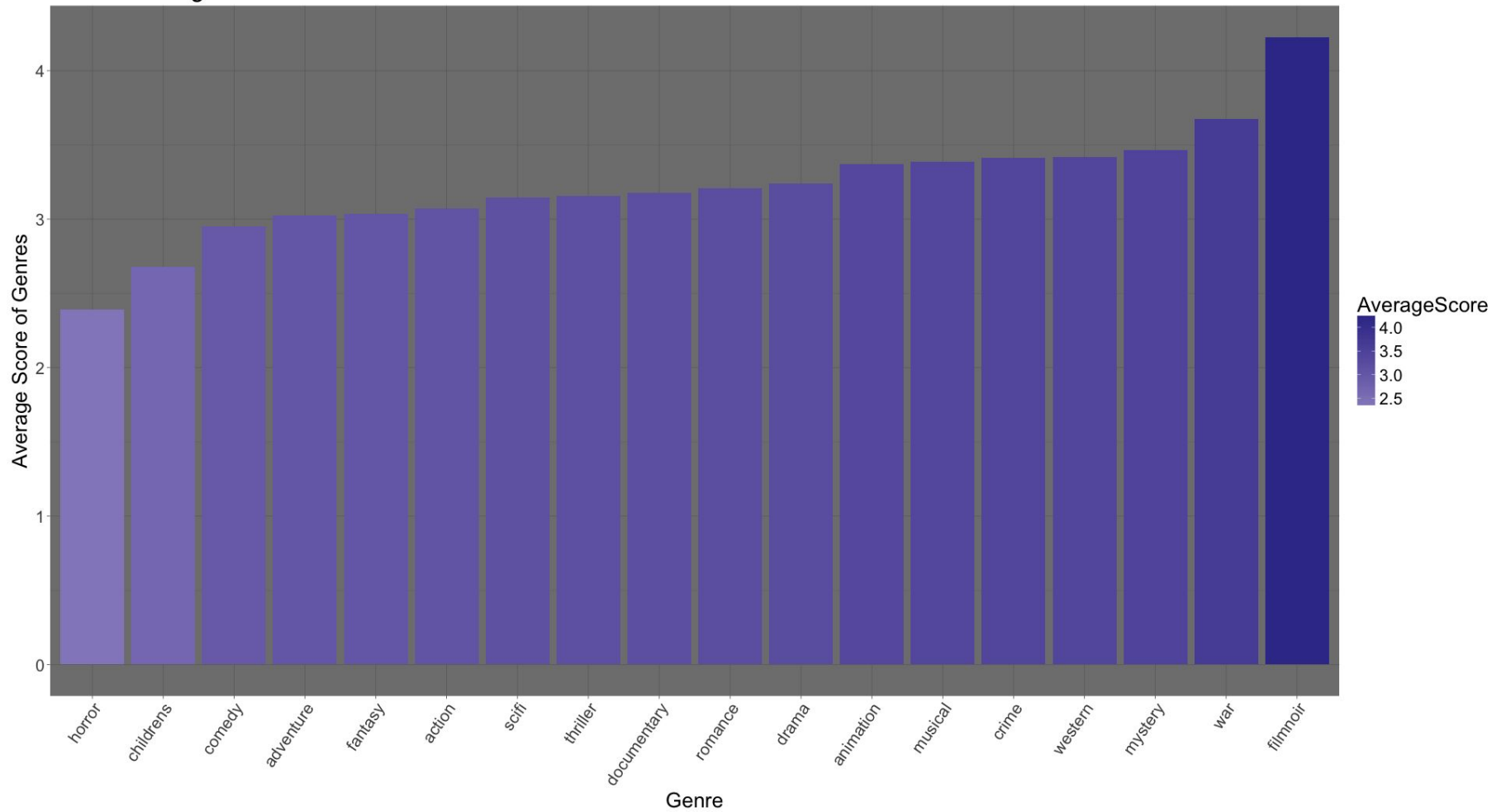
## Percent of Ratings by Sex

Female  
24.6%  
(243,108)



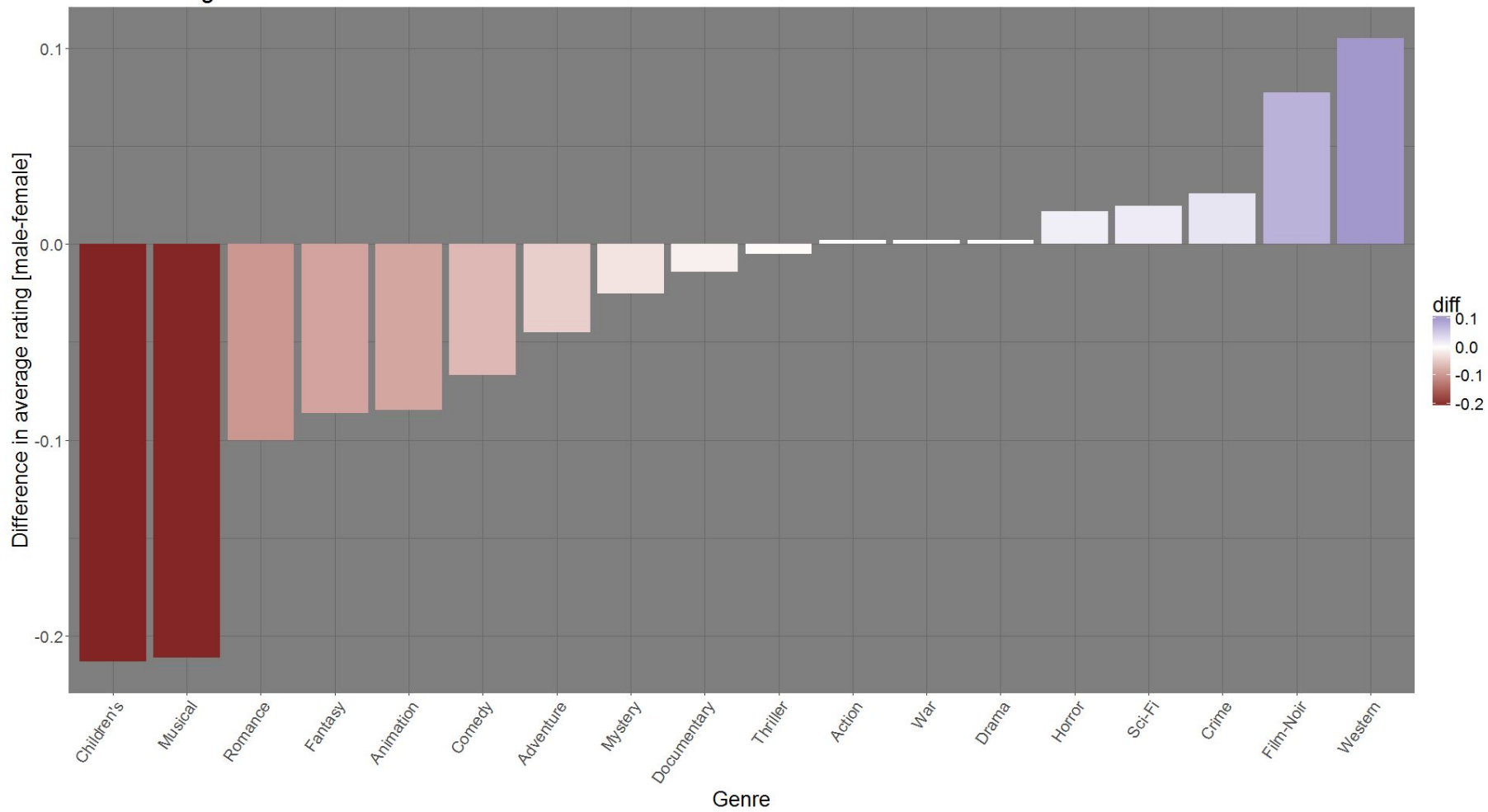
Male  
75.3%  
(744,976)

Genre Average Score





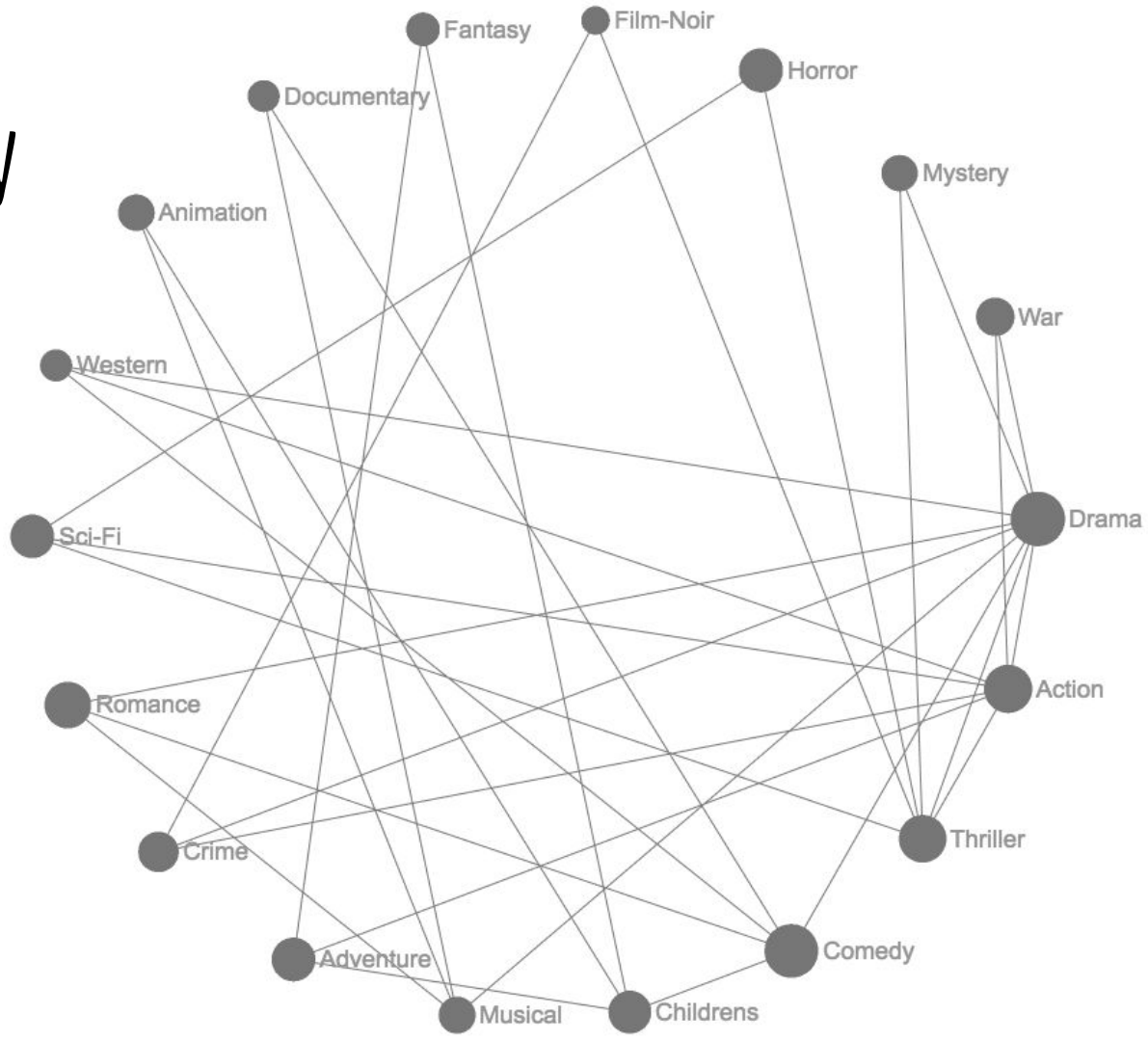
Genre Rating Differential Based on Sex



# Genre Connectivity

Connections indicate  
strongly-tied genres

Size indicates the number  
of movies of that genre



# If we had more time...

- Explore more connections between genre preferences and demographics
  - Age
  - Occupation
  - Location

# If we had more time...

- Explore more connections between genre preferences and demographics
  - Age
  - Occupation
  - Location
- Develop predictive model to suggest movies based on demographics and personal movie preferences

# If we had more time...

- Explore more connections between genre preferences and demographics
  - Age
  - Occupation
  - Location
- Develop predictive model to suggest movies based on demographics and personal movie preferences
- Profit??? \$\$\$

# Works Cited

*F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages. DOI=<http://dx.doi.org/10.1145/2827872>*



<https://GitHub.com/therudnick/movies>





**UHRRR AHHRR AAARGH**  
**(THANK YOU)**