A visual comparison between movie genres and gender portrayal: Insights regarding female representation in the film industry

Josefine Klintberg

Abstract—The current access to digital experiences and storytelling that appears in everything from movies to games and commercials highlights the occurrence of a large amount of stereotypes and equality issues. The goal of this project is to investigate and visualize connections between some of the most popular movies over time and how women are being represented in them, with main focus on weather the movie passes the *Bechdel test* or not. By visualizing the movies with regards to genre, income and production year, the implementation was performed with aim to investigate correlations between genres and female representation in movies in order to gain insight in how women are represented in the film industry.

Index Terms—Multivariate data, Information Visualization, Equality, Bechdel, Movies.

1 Introduction

With an increased access and encountering with digital representation of characters in games, movies and commercials, the importance of a diverse and including representation of gender is becoming more and more important. In order of adding a direct measure and minimum criteria for a better female representation in fiction, the Bechdel test can be used. The test originated from the comic strip "Dykes to Watch Out For" by Alison Bechdel in 1986 (Bechdel, 1986) and consist of three criterias; there must be at least two, named, female characters, they must talk to each other and they must talk about something that is not a man.

With main focus on the movie industry and how a gender asymmetric appearance and the lack of equal role models is forming the way we think and form our minds around gender, this project was implemented with focus on investigating if there is certain movie genres that suffers more from gender biases than others.

2 Background and Related Work

Within movies, women have throughout history been underrepresented and there exist many studies of gender biases in media but mostly focusing on content analysis and the automatization of the Bechdel test. A recent study made by Yang, Xu & Luo (2020) puts focus on the movie business and analyse some of the difficulties that women face in this industry, such as many movies that results in a high box-office gross have a very poor female representation.

The importance of the Bechdel test has been increasing with time as more and more people gain knowledge of this small measure of gender portrayal in fiction. The need of a concrete measurement is essential in order to highlight problematic behaviors and stereotypes. Agarwal, Zheng, Kamath, Balasubramanian & Ann Dey (2015) conclude in their study that women are portrayed as less important characters in movies that fail the Bechdel test than in those that pass the test.

Yang et al. (2020) highlights the under-representation of women both in cast and among directors as well as pointing out problematic issues regarding female stereotypes in certain genres, even though the movie may pass the Bechdel test. With many articles focusing on measurements and plotting relationships, our main goal is to create an interactive visualization that provides the user with the tool to from their own perceptual image of how women are portrayed in some of the most popular movies over time.

- *Josefine Klintberg, Linköping University, Sweden, e-mail: joskl841@liu.se. Project implemented in collaboration with Ester Lindgren, Linköping University, Sweden, e-mail: estli612@liu.se.*

3 Data

Two data sources have been used for this project: the Bechdel Dataset[1] and a subset of the Internet Movie Database (IMDb) [2]. Both of the datasets were acquired through Kaggle [3].

3.1 Data Acquisition

In order to implement the visualization, the two datasets were processed and combined into one with focus on dimensional reduction. A Python script was implemented that processed the data into a resulting json-file and it was also used to output statistics of the data being processed in the console, as a way to conclude that a good variety of data had been chosen.

The IMDb dataset of 5000 movies produced between 1970 and 2013 was acquired from the Kaggle website and consisted of 28 fields. This data was processed and the relevant fields that was kept was title, rating, year, genre, type and imdbid (which is used to join the datasets). Additionally the writer, director and award fields were kept in order for a possible extension of the visualization. The Bechdel Dataset consisted of 15 fields and the relevant fields kept was the binary test value, the imdbid and the gross income of the movie.

3.2 Data Preprocessing

The reduced data from the two datasets was paired with the use of the imdbid and then sorted in descending order according to the rating of the movie. TV series was removed from the list and the genre for each movie was set to be the primary genre, if multiple existed. Each movie got an individual ID in the visualization as well as a field for the movie poster address which resulted in 13 fields for each data item.

Finally, the data was ordered into a hierarchical structure with the use of the movie genre as classifier and outputted in a json-file to be used for the visualization. The script allows for easy modification of the amount of movies that are chosen and for the current visualization, the 100 movies with highest rating are used.

4 Method

The main goal of this project consisted of investigating and evaluating if there is a connection between genres and the female representation in movies. The method was formed with this in mind and the visualization was created with three steps of analysis. The first step was the data processing step as described under Section 3, where the dimensional reduction and classification created a selection of movies.

[1] https://www.kaggle.com/fivethirtyeight/fivethirtyeight-bechdel-dataset
[2] https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset
[3] https://www.kaggle.com/

In order to determine that sufficient data for the visualization had been chosen, the data statistics was evaluated with the following requirements:

- There must be at least 10 different genres in the selected movies
- Bechdel data must exist for at least 60 % of the selected movies

The second step was the implementation that was focusing on interaction design and characterized by analysis of how the data is visualized. With the help of Norman's design principles (Norman, 2002), the visualization was designed in order to achieve easy understanding and usability.

The last step consisted of user-centered evaluations that was performed with the aim of concluding weather the visualization had been implemented in a successful way that both allows for interaction as well as investigations of correlations between genres and female representation in movies. The evaluation consisted of six questions and was a mix of qualitative and quantitative measures that further was resulting in a evaluation report and connected back to the implementation step in order to fix potential issues.

5 Implementation

The data mining was implemented as a Python-script and gave a json-file as output. The implementation of the visualization graph was performed using the Javascript library *d3*, version 3, and displayed through an HTML-file and styled using CSS. The usage of circles to represent the years and the base for the Javascript implementation came from the Starburst approach by Srivastava [4].

The main approach for designing and implementing the visualization was performed with regards to the 'visualization seeking mantra' proposed by Shneiderman (1996); overview first, zoom and filter, then details on demand.

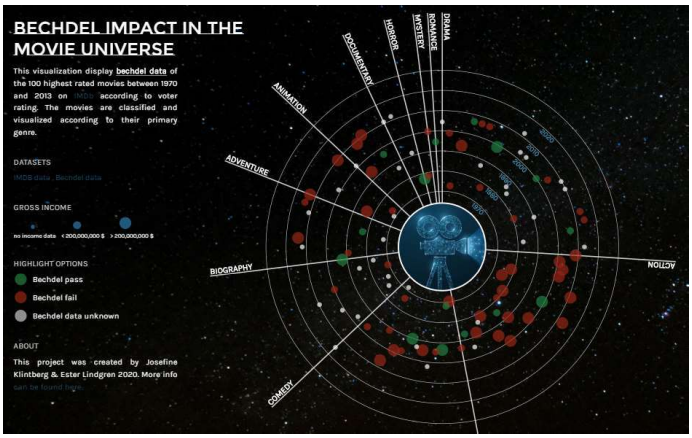

Fig. 1. Overview of the resulting visualization.

6 Results

The visualization was successfully implemented and can be observed at the Github Repositry for the project [5]. An overview of the visualization can be seen in Figure 1, where all the movies have been displayed as circles that are varying in size according to their gross income. The placement from the center is corresponding to the year of production and the placement between the pie chart pieces is mapped to the corresponding genre. It is noted that 33 movies are mapped according to the highest gross income range in the visualization and among these, 27 movies fail the Bechdel test and 6 movies pass the Bechdel test.

[4]https://observablehq.com/@mksrivastava/d3-starburst-from-sunburst

[5]https://jklintan.github.io/Visualizing-Women-In-Movies/

6.1 Data Selection

The resulting selection when choosing the 100 highest rated movies according to IMDb can be seen in Table 1. The selection consisted of 11 different genres and Bechdel data exist for 65 % of the movies. The selection of movies was therefor considered to be sufficient since the two requirements stated in Section 4 was fulfilled. Note that every individual genre has existing Bechdel data for more than 60 % of the selected movies except Drama that has 54 %.

Table 1. Statistics over the 100 chosen movies, their genres and the Bechdel test data.

Genre	Movies	Passing	Failing	Data Unknown
Action	20	3	16	1
Adventure	6	0	4	2
Animation	6	1	3	2
Biography	8	2	3	3
Comedy	10	0	2	8
Crime	16	3	10	3
Documentary	3	0	0	3
Drama	26	4	10	12
Horror	2	1	1	0
Mystery	1	0	1	0
Romance	1	1	0	0

6.2 Interaction Design

The implementation of the visualization was performed with regards to 'the visualization seeking mantra' as explained in Section 5 and also with focus on easy understanding for a user. This meant that the multivariate data needed to be mapped in different, but evident, ways.

After displaying the overview, as can be seen in Figure 1, the filtering of movies could be performed according to the Bechdel data as can be seen in Figure 2. The filtering allowed the user to highlight movies according to Bechdel data and form an opinion of the overall distribution of movies that pass and fail the Bechdel test.

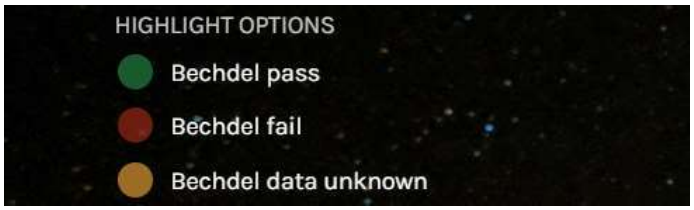

Fig. 2. Filtering options according to the Bechdel data.

The income data was mapped through the size of the circles that represent an individual movie. By this mapping, the user could easily see both the popularity of a movie but also investigate correlations between high box-office gross for a movie and the Bechdel data.

Fig. 3. Mapping of gross income for a movie with the use of circle radius.

The last step of 'the visualization seeking mantra' is the details on demand. This was displayed as a legend when hovering over a circle corresponding to an individual movie, as can be seen in Figure 4. When a user hovers over a movie, title, release year, genre and Bechdel data is displayed. Further more, the movie poster is displayed in the middle of the visualization.

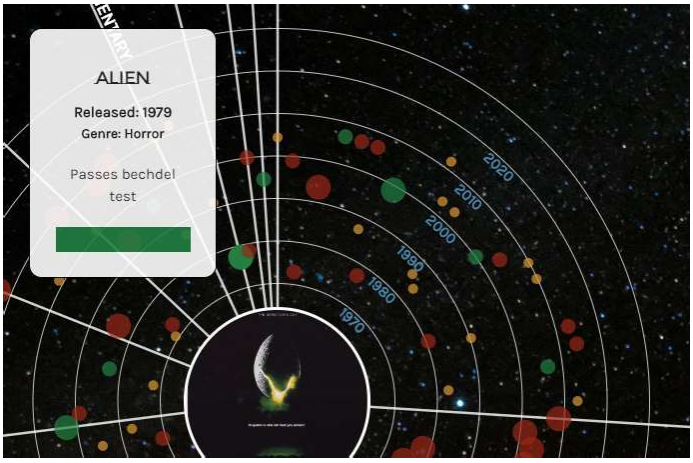

Fig. 4. Legend that display details on demand when hovering over an individual movie circle.

All design elements in the visualization was related to Norman's design principles (Norman, 2002) and especially hover effects were added to the movie circles and to the filtering buttons in order to provide affordance to the user of how to interact with the visualization. Additionally, the colors in green and red of the movie circles were styled with regards to a natural mapping between the passing or failing of the Bechdel test.

6.3 Evaluation

The survey of six questions was conducted with four participants. The qualitative questions were open questions where the participant were asked to think out loud. The quantitative questions measured how long time the user took when performing a task and deliver an answer to questions related to the visualization.

The user-centered evaluation confirmed that the visualization was a good application to encourage interaction since all the participants managed to use and understand both the filtering and the mapping of data within a short amount of time. All the participants claimed that they learned something new after interaction with the visualization during the survey. However, there was one issue regarding the mapping that was highlighted during the evaluation and that was the perception of which genre the movies corresponded to.

In order to further emphasize the mapping of data for the genres, an additional filtering function was implemented that causes all movies corresponding to the current genre to be highlighted when the user hovers over the name of that genre, or the line dividing this genre from the others. The result of this can be seen in Figure 5.

7 Conclusions and Future Work

The visualization is considered to be successful in order of providing an insightful tool that allows for interaction and individual reflection when evaluated with regards to comments from the participants in the user-centered evaluation. Questions that appeared during the evaluation also sparked possible extensions of the visualization to include data that can be used to investigate how the influence of female directors relate to the passing or failing of the Bechdel test.

Fig. 5. Highlighting of movies within the current genre when hovering.

The overview appearance of the visualization is the most striking image, where it can be seen that the majority of the movies do not pass the Bechdel test. Some of the movies offer a surprising Bechdel data, since many of the failing movies do have more than one named female character appearing on screen, but the movies are not passing the Bechdel test. This means that only the appearance of women is often not enough in order to result in an inclusive and diverse movie.

The failing of the Bechdel test can be seen as a flag for issues regarding an unequal cast and screenplay but could also be an indicator for more extensive problems regarding how the female characters in the movies are portrayed. This becomes evident in certain genres in this data, such as the genre Action, where 15 out of the 20 movies fail the Bechdel test, and the genre Crime, where 10 out of 16 movies fail the Bechdel test. Another issue when looking at the data is the fact that among the most successful movies, according to gross income, 27 out of 33 movies are failing the Bechdel test, which can be problematic if arguments for an unequal cast are based on the idea of how high box-office movies have looked like in the past.

This project concludes that some genres, and especially high box-office movies, suffer more from gender biases than others, but also that inequality issues in movies exist in more than just one way. In order to emphasize the need of change, the underlying factors need to be analyzed with regards to multiple factors, and not just one. The Bechdel test is a good way to flag issues, but an extension of the test to also consider things such as stereotypes within genres, how female characters are portrayed in terms of sexuality and trait of character could be a starting point in order to create a more diverse and inclusive film industry.

References

[1] A. Agarwal, J. Zheng, S. Kamath, S. Balasubramanian, and S. Ann Dey. Key female characters in film have more to talk about besides men: Automating the Bechdel test. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 830–840, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[2] A. Bechdel. *Dykes to watch out for*. Firebrand Books, Ithaca, New York, 1986.

[3] D. A. Norman. *The Design of Everyday Things*. Basic Books, Inc., USA, 2002.

[4] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343, Sep. 1996.

[5] L. Yang, Z. Xu, and J. Luo. Measuring women representation and impact in films over time. 2020.