# Understanding Film Characters and their Social Networks through a Gender Lens

**Weizhen Sheng**
University of Pennsylvania
wsheng@seas.upenn.edu

**Advised by:**
**Ani Nenkova**
University of Pennsylvania
nenkova@seas.upenn.edu

## Abstract

In this paper, we analyze film characters and their social networks, and compare differences observed across gender. We present approaches to identifying characters and their gender, and introduce topic modeling to find character archetypes. We also use a graph model to represent social structures and to quantify both the size of a character's social network and the character's importance in a network. Our experiment results indicate that gender does indeed impact representation, character portrayal, and character networks.

## 1 Introduction

Stories permeate our society and form an important part of our culture. Much of the media that we consume in our daily lives is driven by stories; one such prominent form of media is film. Narrative films may contain complex narrative structures, as outlined by a *screenplay* or script. These stories are often made up with complex characters and character social networks.

This paper will perform a character-driven analysis of film screenplays to better understand how characters are portrayed in media. In addition, given the increasing popularity and relevance of social network analysis due to social media and the connected nature of today's world, this study will also analyze the relationships and social networks of film characters. This is inspired by the popular concept of "character-driven" stories, where the focus is on the characters and their internal development.

Lastly, this paper will analyze both characters and their social networks through a gender lens. There has been much ongoing research in the area of media and gender, investigating how gender is represented in different forms of media. Inequalities in society may be reflected in the media and film that we consume, and so we will analyze observed differences across gender in the context of a film. Ultimately, we will attempt to extract real-world applicable insights about how characters are portrayed and what this may say about gender representation.

## 2 Related Work

There has been past work both on parsing screenplays, identifying main characters in films, extracting social networks from media (e.g., films, television programs, literary fiction), and identifying gender of names, of which several provide insightful observations that guide this study.

For parsing screenplays, Agarwal et al. provide a formalized approach for parsing screenplays (Agarwal et al., 2014). Their model identifies various scene boundaries and character names through an NLP and ML based approach; though we do not use their model directly, we use their insights on fixed levels of indentation and other boundaries in our own parsing of screenplays. Gorinski and Lapata studied the task of summarizing movie scripts, and in doing so, compiled the ScriptBase movie script corpus that is used in this study (Gorinski and Lapata, 2015).

There has been some previous work done in identifying main characters. Agarwal et al. found main characters by identifying those with the highest betweenness centrality scores (Agarwal et al., 2015), but this approach mainly captures importance by social network relevance. In this paper, we will introduce an approach that aims to classify main characters in terms of their importance to the story.

In terms of extracting social networks, Bost et al. suggest that there are key differences in extracting static networks over a whole time period (i.e., the entire film) versus dynamic networks over smaller time-slices (Bost et al., 2016). Their work suggests that cumulative, static networks can be

useful for analyzing fictional works with well-defined communities, such as films. Bost et al. also provide various rules for estimating verbal interactions by parsing when speech turns start and end (Bost et al., 2016). Agarwal et al. investigate a similar topic of identifying interaction networks of characters and compared two approaches (Agarwal et al., 2015). The first defined an interaction to exist between all characters within a singular scene. The second defined an interaction to exist only between characters with adjacent lines, disregarding other characters also appearing in the scene but who did not have a consecutive line. They found that the second approach, with adjacent lines, performs significantly better in identifying interactions. We intend to expand on this work of extracting social networks by examining how these structures are impacted by gender.

Lastly, Agarwal et al. automated the Bechdel test on film screenplays, which required labeling characters with a gender (Agarwal et al., 2015). In determining the gender of a name, they compared three different methods. The first used the Internet Movie Database (IMDB) to map characters to their corresponding actors/actresses, which then allowed them to associate characters with the actors/actresses' gender. The second method used the Social Security Administration (SSA) of the United States' publicly available list of baby first names, which includes gender. Lastly, they used named entity recognition (NER) techniques and assigned gender based on the gender of associated third person pronouns (e.g., *she*, *he*). They ultimately found that the NER approach yields the best results.

## 3 The Dataset

This study uses the 1,276 films compiled in The ScriptBase Corpus (Gorinski and Lapata, 2015). This data set includes various metadata (e.g., IMDB rating, date of release, etc...), the screenplay, user-written summaries, and more. The films span years 1927-2013 and comprise 23 genres.

### 3.1 Parsing Screenplays

The dataset provides films whose screenplays are formatted in varying ways, which we aim to parse for character names, character dialogue, and character directions if possible using primarily regex. Through manual combing, we found that 1,259 films can be categorized under six parsable formats, with the remaining 17 films having unparsable formats (e.g., the screenplay is written entirely in terms of director narration and has no dialogue). The 17 unparseable films are discarded from this study. Among the remaining 1,259 films, 1,178 of them share a common format and the other 81 films are distributed among the remaining five screenplay formats; these films are all parsed accordingly (Sheng, 2020). Note that due to formatting inconsistencies, we were unable to parse for character directions in the 81 films not sharing the common screenplay format, and instead only extract character dialogue.

## 4 Characters

### 4.1 Identifying Main Characters

As we are interested in understanding media representation, we choose to focus on analyzing the *main characters* of a film who typically have more nuanced and interesting characterizations. The definition of a main character is subjective and ambiguous in and of itself, so identifying these main characters is a rather non-trivial task. For the purposes of this study, we define main characters in a film to be ones whose actors/actresses are listed as "Starring" cast members on the film's infobox in its Wikipedia page; we will use these labels as our "gold" labels when comparing different labeling approaches later. All other characters will be denoted as *minor characters*.

From parsing the screenplay, we have each character's name and the number of lines they have in the screenplay. Among this list of characters, we discard ones with only one line; these are usually either exceedingly minor and irrelevant characters, or are a result of parsing errors for title cards or the like. We begin the task of identifying main characters by looking at how many lines a character has in the screenplay. We expect that the majority of a screenplay's lines will be spoken by a small subset of characters, who likely are the main characters. As seen in Figure 1 and Figure 2, there is indeed a sharp drop-off in the distribution of number of lines, with a very long right-tail of minor characters who only have one line. This confirms our hypothesis that the bulk of a screenplay's lines likely belong to the main characters of the film. Note that the drop-off is is steeper in some films compared to others, suggesting that some films have larger discrepancies between their main and minor characters.

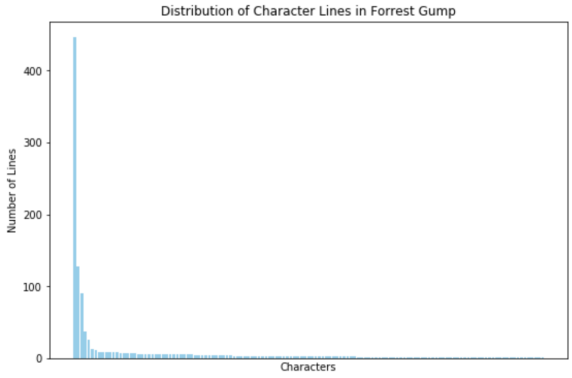


Figure 1: Distribution of number of character lines in the film *Forrest Gump*. Note that there is a steep cut-off after the character with the most number of lines.

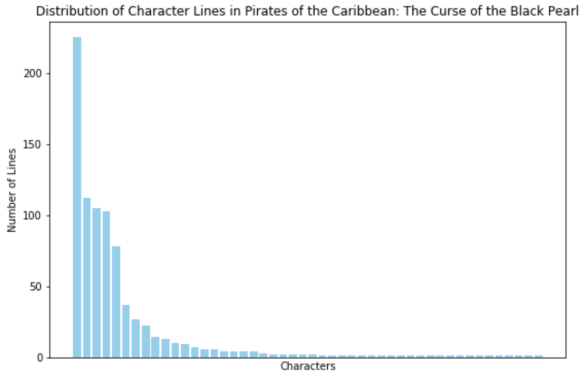


Figure 2: Distribution of number of character lines in the film *Pirates of the Caribbean: The Curse of the Black Pearl*. Note that there is still a steep cut-off after the character with the most number of lines, though it tapers off more than in Figure 1.

We define the primary main character as the *protagonist* and label them as the character with the most number of lines in the screenplay. We first check the validity of identifying the protagonist in this way by comparing results with manually identifying the protagonists from 12 films; these films are picked such that it may be difficult even for humans to identify who the protagonist is. We find that there is a 75% accuracy, although we note that all of the incorrectly identified protagonists belong to either films with an ensemble cast of main characters where a central protagonist character is difficult to agree upon (e.g., *The Princess Bride*), or films where the protagonist does not speak (e.g., *WALL-E*). We are satisfied with the results of this approach and continue.

We now attempt to identify the remaining main characters. To this end, we try two different approaches, outlined below:

1. *Comparing with protagonist*: For each character in our list, we compare their number of lines with the number of lines of the protagonist, denoted as $n$. We label a character as a main character if they have $\geq 0.3n$ lines. This approach will be referred to as the *protag* approach.

2. *Comparing with previous*: We consider characters in decreasing order of number of lines. For each character with $m$ lines, we label them as a main character if $m \geq 0.3m_{prev}$, where $m_{prev}$ is the number of lines of the previously identified main character. When finding the second main character after the protagonist, $m_{prev}$ is the number of lines of the protagonist. This approach will be referred to as the *prev* approach.

We compare the results of using each approach in trying to identify main characters of the same 12 films used for identifying the protagonist, whose "gold" label lists of main characters were manually extracted from their Wikipedia pages. The averaged results across these 12 films are summarized below:

| Approach | Precision | Recall | F-score |
|---|---|---|---|
| *protag* | 0.972 | 0.548 | 0.679 |
| *prev* | 0.870 | 0.562 | 0.569 |

Table 1: Averaged precision, recall, and F-scores across 12 films for the *protag* and *prev* approaches of identifying main characters.

Given the results, we choose to use the *protag* approach moving forward. Using this approach, we identify 4,250 main characters across our entire dataset; the results are summarized in Table 2 and the distribution is plotted in Figure 3.

| mean | mode | min | max | $Q_1$ | $Q_2$ | $Q_3$ |
|---|---|---|---|---|---|---|
| 3.37 | 2 | 1 | 16 | 2 | 3 | 4 |

Table 2: Statistics for number of main characters, as identified using the *protag* approach.

### 4.2 Identifying Gender

For the purposes of this study, we will use the simplifying assumption that characters have one of two genders (male or female). We present a novel approach for labeling a character's gender,

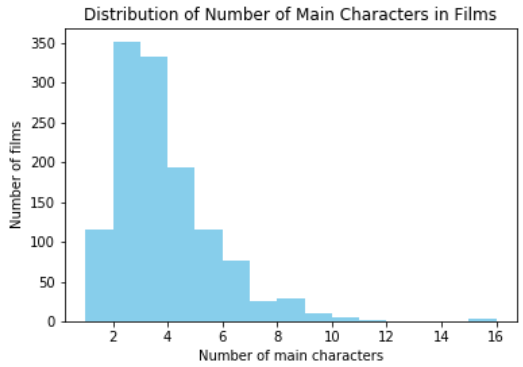


Figure 3: Distribution of number of main characters, as identified using the *protag* approach.

different from previously presented methods discussed above (see Section 2). This method is more parsimonious and uses GloVe word embeddings trained on a Wikipedia 2014 corpus (Pennington et al., 2014). We convert character names to vector representations before computing cosine similarity scores between the name and two labels corresponding to gender (e.g., "male" and "female"). The character is then classified as whichever gender the name had a higher similarity score with. We experimented with different sets of gender labels across 47 characters from 15 films, with results summarized below:

| Gender Labels | Precision | Recall | F-score |
|---|---|---|---|
| "Male" vs. "Female" | 0.345 | 0.769 | 0.476 |
| "Man" vs. "Woman" | 0.611 | 0.846 | 0.710 |
| "Boy" vs. "Girl" | 0.476 | 0.333 | 0.392 |
| "Masculinity" vs."Femininity" | 0.667 | 0.923 | 0.774 |

Table 3: Results of labeling gender using word embeddings and comparing cosine similarity scores.

We found using the terms "masculinity" and "femininity" yielded the highest F-score of 0.774. Although this approach isn't perfect at the labeling task, we conclude the results are satisfactory and we proceed to label all films' main characters with this approach. We find that 164 of the character names are not in the corpus and so these names are discarded. We use the remaining 4,086 characters moving forward.

| Gender | Count |
|---|---|
| Male | 2,361 |
| Female | 1,725 |
| Unidentified | 164 |

Table 4: Gender breakdown of labelled characters.

### 4.3 Female Representation in Films

Using the results from identifying gender, we were interested in seeing the amount of female representation in films. Perhaps unsurprisingly, almost two-thirds of films have male protagonists.

| Gender | Count |
|---|---|
| Male | 807 |
| Female | 429 |
| Unidentified | 23 |

Table 5: Gender breakdown of protagonists.

We also consider the representation of female characters in the set of main characters of a film – even if she isn't the primary protagonist, a female character could still be a significant character. However, as we can see in Figure 4, most films have only two or fewer female main characters, suggesting that a very small number of female characters are significant in their respective films. Furthermore, compared to the male distribution of number of main characters, the female distribution is shifted left, peaks at a lower value of one female main character, and has a shorter right-tail.

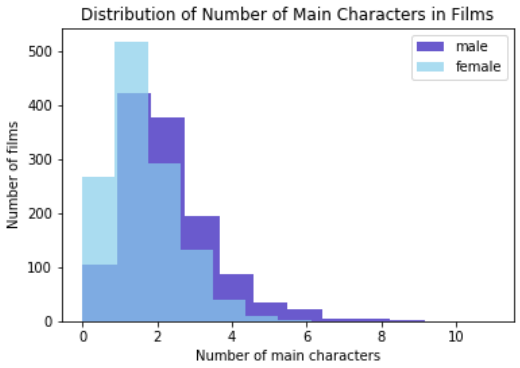


Figure 4: Distribution of number of main characters by gender.

That being said, if a film has only two or three main characters, having one female main character can be considered good representation; in other words, the ratio of main characters who are female would be an informative metric to consider. We vi-

sualize the distribution of this ratio across films in Figure 5. The shape of the distribution quite noisy, with several peaks and troughs. As we can see, the highest peak is at a ratio of 0.5, followed closely by a ratio of 0.0. This suggests that while many films equally represent their male and female characters in the main cast, there are also several that don't represent female characters at all. It's also worthwhile to note that on the other end of this spectrum, there is a small subset of films whose entire cast of main characters is female.

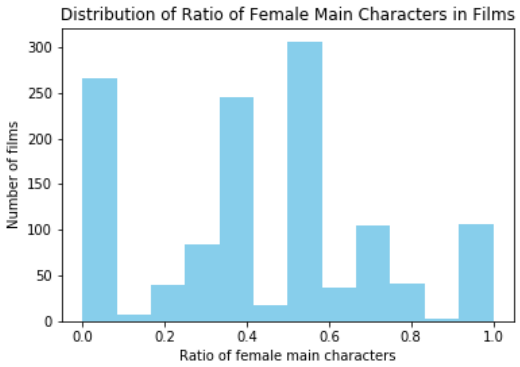


Figure 5: Distribution of ratio of female main characters across films.

### 4.4 Sentiment in Female Portrayal

After investigating the amount of female representation in films, we now turn towards understanding *how* females are portrayed. Many gender stereotypes typically associate females with positive traits such as "gentle," "calm," "kind." We are curious to see if these stereotypes carry over into media portrayals.

To investigate this, we use sentiment analysis to characterize main characters' lines. Using the simple rule-based model VADER (Hutto and Gilbert, 2014), we compute a "compound" score that sums normalized lexicon ratings ranging between -1 and 1 (which represent negative and positive sentiment respectively). This score is computed for each line of a character's dialogue, and the scores for all the lines of a character are averaged to calculate a character's overall sentiment score. These sentiment scores are summarized below:

Sentiment scores skew ever so slightly more positive for female main characters, as seen in Table 6 and Figure 6. (Note that Figure 6 is normalized within each gender so that the higher number of male main characters does not skew the results.) This seems to suggest that female characters are portrayed slightly more positively and have more dialogue with positive meaning, such as "They seemed so romantic and daring" or "It's beautiful." That being said, female characters can also say lines with negative sentiment, such as "It's poisoned!" or "I can't breathe." Overall, both male and female characters have positive sentiment scores near 0, representing neutral-positive sentiment.

| Gender | Mean Senti. | Median Senti. |
|---|---|---|
| Male | 0.009 | 0.008 |
| Female | 0.016 | 0.013 |

Table 6: Mean and median sentiment scores across main characters, broken down by gender.

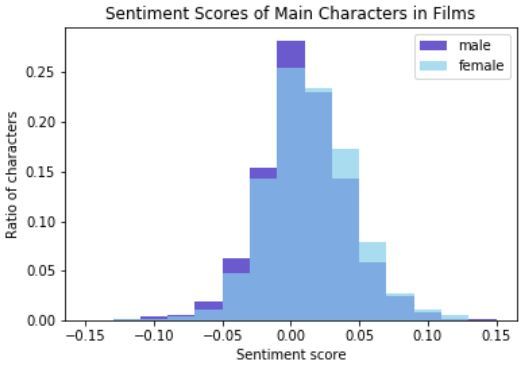


Figure 6: Sentiment score distribution by gender.

### 4.5 Character Archetypes

To better understand exactly how characters are portrayed in films beyond sentiment analysis, we look into grouping common topics spoken by characters and relating these to *character archetypes*. Character archetypes were first introduced by Swiss psychologist Carl Jung to explain images and themes with universal meanings. Jung identified 12 character archetypes, such as the caregiver (e.g., a mother motif) or the jester (e.g., a trickster motif) (Mark and Pearson, 2001). The idea of character archetypes has been used in the literary field to write better characters, and other authors have also expanded on the initial set of 12 Jungian archetypes. For example, Victoria Schmidt offers eight heroine and eight hero archetypes in her book *45 Master Characters*, comparing each archetype to a figure from Greek or Egyptian mythology (Schmidt, 2001).

To this end, we ran Latent Dirichlet Allocation (LDA) models with 25 topics on character lines

using MALLET (McCallum, 2002), and then labelled each character with their principal topic. Each generated topic has 20 keywords that characterize it. Across the 25 topics, there is a similar distribution of principal topics for male and female characters, as seen in Figure 7. For example, Topic 6 is the most common principal topic for both genders, with 28.1% of male characters and 25.5% of female characters labeled with it. We also note that a few of the generated topics have close to 0% of characters labelled with it, suggesting that these are insignificant topics. Some top topics for both genders are summarized in Table 7.

| Topic | F (%) | M (%) | Keywords |
|---|---|---|---|
| 6 | 25.5% | 28.1% | yeah hey gonna... |
| 8 | 19.2% | 17.2% | door smiles sits... |
| 9 | 17.9% | 11.4% | good love home... |
| 25 | 16.0% | 16.1% | body wall reaches... |

Table 7: Topics that are common among both female (F) and male (M) characters. A sample of each topic's keywords is displayed.

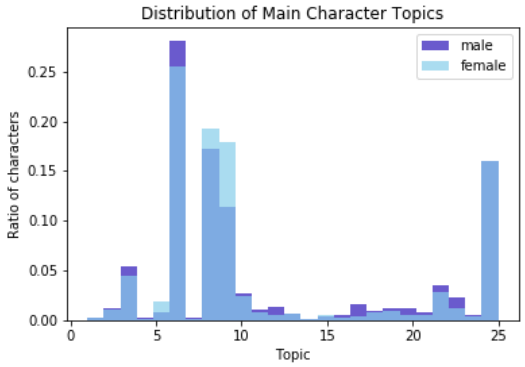


Figure 7: Principal topic distribution by gender.

Nonetheless, there are still gender differences in terms of topic distribution. We find the proportion within each gender that is labeled with each topic (e.g., 1.86% of females versus 0.81% of males are labeled with Topic 5), and then identify the topics where there are the largest differences in proportion. This is found by calculating $\frac{f-m}{f}$, where $f$ and $m$ are the proportions of females and males labelled with the topic, respectively. Some select topics with high discrepancies are summarized below:

In order to better interpret what each of the 25 topics corresponds to, we compare their topic keywords with archetype text descriptions and compute cosine similarity scores. We then map each topic to its top-scoring Jungian archetype, hero archetype, and heroine archetype. Some sample topic-archetype mappings are shown in Table 9.

| Topic | Diff | Keywords |
|---|---|---|
| 5 | +56.6% | love phone night kiss... |
| 9 | +36.0% | good time love home... |
| 12 | -167.85% | sir control commander... |
| 17 | -247.0% | money office people ... |

Table 8: Topics with the highest gender discrepancies. The "Diff" column shows how female proportion compares to male proportion (e.g., Topic 5 female proportion is 56.5% higher). A sample of each topic's keywords is displayed.

From these elementary mappings, we can draw some initial observations on how different characters are portrayed. The most commonly mapped Jungian archetype across both genders is "The Everyman," an archetype representing the common person who fears being alone. For females, they are much less likely than males to be characterized as "The Ruler" or "The Rebel" archetypes, as these correspond to topics with very low female proportions (e.g., Topics 12, 20). This perhaps indicates that female characters tend to be portrayed as less powerful or aggressive individuals.

The hero and heroine archetypes reveal some interesting initial thoughts and topic associations as well. For example, the Topic 5 mapping suggests that the "Persephone" archetype is quite common relative to the corresponding male one; this archetypes characterizes a woman who is "childlike" and carefree, perhaps indulging in parties and romance. The "Dionysus" archetype characterizes a man who is associated with women and sensuality. Although both archetypes have similar topic associations, we see from the gender representation that characters associated with this topic are more likely to be female. Another interesting observation is that the "Athena" archetype seems to be quite rare, both in absolute and relative terms. This archetype characterizes an intelligent woman who aims to prove she is as capable as male peers. This archetype is mapped to Topic 12, which has very low female representation. In contrast, the mapped male archetype of "Zeus" is of a man who is a powerful leader, and we note that male characters have high representation relative to females in this topic.

Generally speaking, these archetypes are difficult to map and interpret due to their inherently

| Topic | Keywords | Jungian | Hero | Heroine |
|---|---|---|---|---|
| 5 | love phone night kiss happy music beautiful... | The Everyperson | Dionysus: "The Woman's Man" | Persephone: "The Maiden" |
| 9 | good time love home... back pause make... | The Jester | Hermes: "The Fool" | Hera: "The Matriarch" |
| 12 | sir control commander ship captain star power... | The Ruler | Zeus: "The King" | Athena: "The Father's Daughter" |

Table 9: Select topics and their mapped Jungian, hero, and heroine archetypes.

subjective nature. These initial experiments suggest that topic associations do exist and differ between gender, but this area can certainly be explored further in the future (see Section 6). More examples of the generated LDA model topics and mapped archetypes can be found in the Appendix.

## 5 Social Networks

We now move onto studying a character's relations with other characters in the film. We employ social network analysis to understand the underlying social network of film characters and to analyze how gender plays a role in these social networks.

### 5.1 Identifying Character Interactions

We must first extract characters' interactions before building entire social networks. We define a character interaction to be when two characters have adjacent dialogue, as discussed in Section 2. Note that the lines are non-transitive, adjacent lines; in the situation where character A speaks with character B, and then character B speaks with character C, A does not have an interaction with C.

### 5.2 Understanding Social Networks

We are interested in visualizing the underlying social networks of films and seeing if there are any trends or differences across films. We extract complete network structures via parsing of character interactions. In a graph structure, we represent characters as nodes (with main characters colored according to their gender) and character interactions as edges connecting nodes. An edge's thickness corresponds to the number of interactions between two characters; as such, thicker edges represent more interactions between a pair of characters. Example social network graphs are displayed in Figure 8 and Figure 9.

The network structure differs across films. Visually, we can see that there a small subset of all characters in a film dominates the

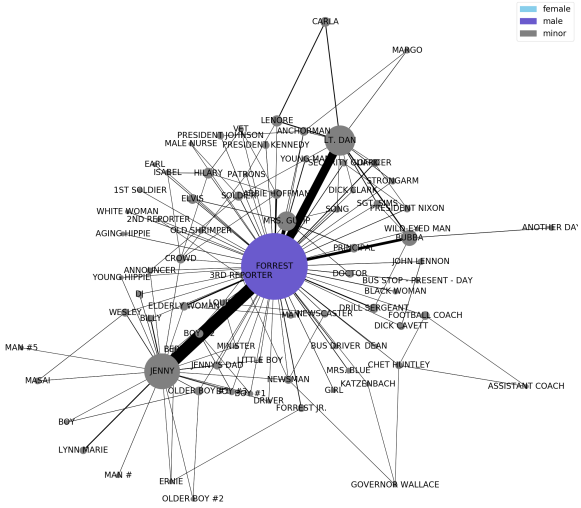


Figure 8: Social network for the film *Forrest Gump*.

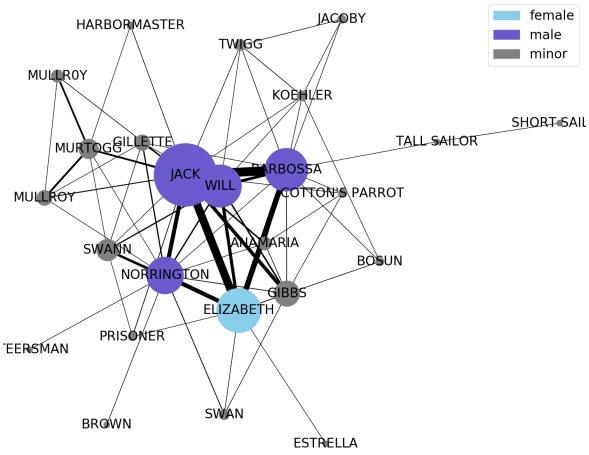


Figure 9: Social network for the film *Pirates of the Caribbean: The Curse of the Black Pearl*.

social network. This is particularly evident in films like *Forrest Gump*, where there is one central protagonist whom the story revolves around. On the other hand, films like *Pirates of the Caribbean: The Curse of the Black Pearl* have a more complex network clustered around a set of main characters. We also note that

in both social network visualizations, there are only a handful of weighted edges, indicating that similar to characters and lines, a small handful of interactions make up the majority of all character interactions that occur in a film. More examples of extracted film social networks can be found in the Appendix.

### 5.3 Quantifying Network Importance

Looking at these social network visualizations, it appears that many social networks of films are dominated by male characters, both in terms of sheer number and in terms of their "importance" to the network. We begin with quantifying this "importance" by looking at the proportion of interactions that a main character has with other main characters, versus with minor characters.

| Gender | Main | Minor |
|---|---|---|
| Male | 56.7% | 43.3% |
| Female | 61.5% | 38.5% |

Table 10: Averaged proportions of male and female main characters' interactions that are with other main characters, versus with minor characters.

Although main characters across both genders have most of their interactions with other main characters, it is perhaps surprising that male main characters interact more with minor characters. This reveals an interesting phenomenon where male characters tend to have more overreaching relationships and wider social networks, even with less important characters. On the flip side, female characters are relegated to a small handful of relationships that constitute their entire social networks. This results in less opportunities for a character to interact with other characters, which could lead to fewer opportunities for character development. This phenomenon is visible when we plot the individual social networks of male versus female main characters. We compare two main characters in *Pirates of the Caribbean: The Curse of the Black Pearl* – Jack (male) and Elizabeth (female) – in Figure 10 and Figure 11. Elizabeth's social network is visually sparser than Jack's is.

To better quantify this phenomenon, we compute and compare two graph centrality values:

1. **Degree centrality:** A measure of the number of neighbors a node has. In a film's social network, this can be interpreted as how well-connected the character is. We will denote this as $C_D$.

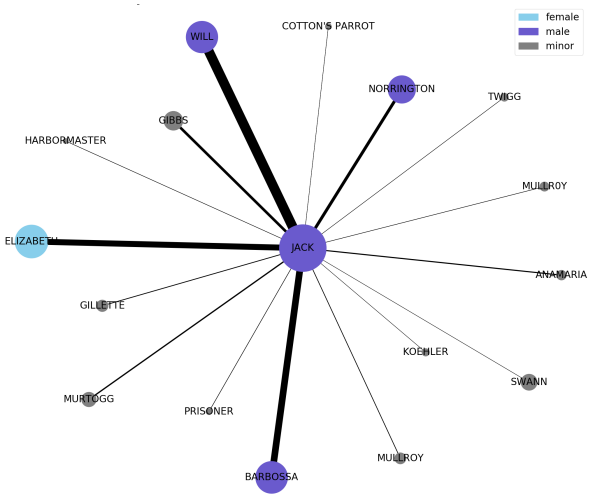


Figure 10: Social network of the character "Jack" in the film *Pirates of the Caribbean: The Curse of the Black Pearl*.

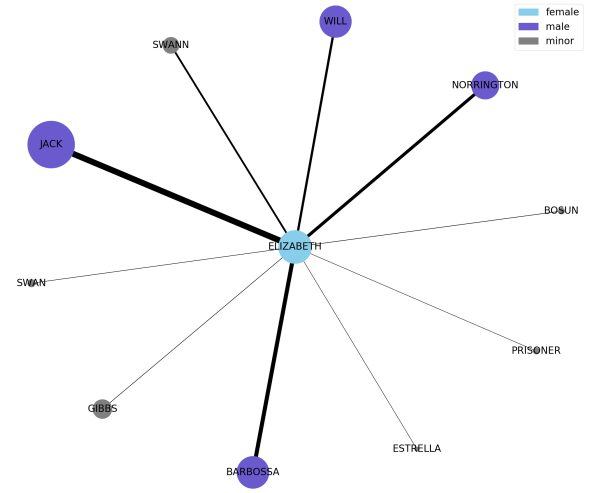


Figure 11: Social network of the character "Elizabeth" in the film *Pirates of the Caribbean: The Curse of the Black Pearl*.

2. **Betweenness centrality:** A measure of how often a node acts as a bridge along the shortest path between other nodes. In a film's social network, this can be interpreted as the influence a character has on the communication between other characters. We will denote this as $C_B$.

We calculate the $C_D$ and $C_B$ values for each main character in each film. We then normalize both centrality scores within each film; for a given main character with a degree centrality score $d$, the score is normalized as $\frac{d-\mu}{\sigma}$, where $\mu$ is the mean centrality score of all characters in the film (in-

cluding minor characters) and $\sigma$ is the standard deviation of these scores. Betweenness scores are normalized in the same way. We normalize scores to better interpret a character's relative network importance compared to other characters in the same film. These normalized centrality scores are summarized below:

| Gender | mean($C_D$) | median($C_D$) |
|---|---|---|
| Male | 2.51 | 2.30 |
| Female | 2.16 | 1.87 |

Table 11: Mean and median $C_D$ values across male and female main characters.

| Gender | mean($C_B$) | median($C_B$) |
|---|---|---|
| Male | 2.48 | 2.11 |
| Female | 2.01 | 1.50 |

Table 12: Mean and median $C_B$ values across male and female main characters.

The summary statistics indicate that male main characters have higher degree and betweenness centrality scores than female main characters. A higher degree centrality shows that male characters have interactions with more characters, and thus have wider social networks. A higher betweenness centrality illustrates that male characters are more influential in their networks and act as more of a connecting bridge between other characters in the film. We see a visualization of differences in centrality score distributions in Figure 12 and Figure 13. These histograms are normalized within each gender so that the higher number of male main characters does not skew the results. For both degree and betweenness centrality, the female distribution is shifted more to the left, with higher peaks at lower score values.

## 6 Conclusion & Future Work

In this paper, we analyzed film characters and their social networks, presenting novel methods of identifying main characters, labeling gender, considering character portrayals, and quantifying a character's importance in their film's social network. We used a wide range of tools to do so, and employed various social network analysis metrics. Ultimately, we found that females are indeed underrepresented in films, often portrayed with different topic associations than male characters are,

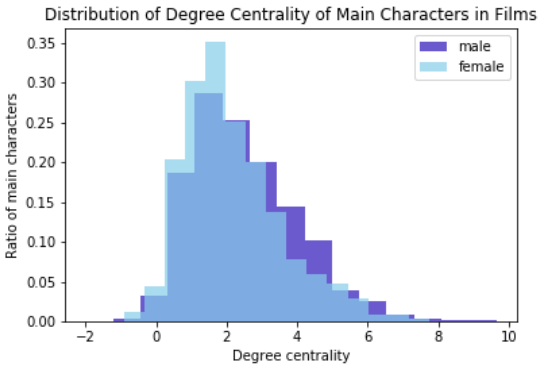


Figure 12: Distribution of normalized degree centrality scores by gender.

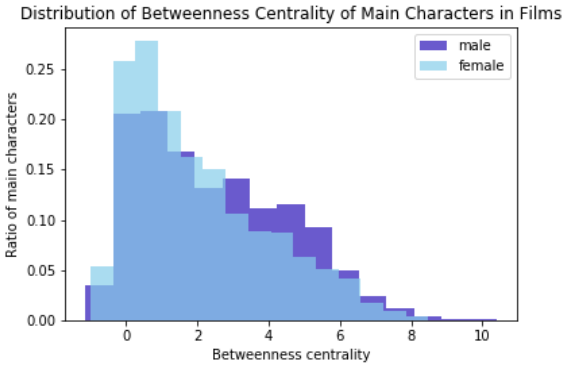


Figure 13: Distribution of normalized betweenness centrality scores by gender.

and usually have narrower social networks and less influence in their social networks.

In addition to the ones outlined above, we performed numerous other experiments. Unfortunately, many of these did not yield interesting results. One of these experiments looked to find a relation between a film's success (e.g., ratings) and how it portrays and develops its characters, but we were unable to find any significant results. Despite this, we believe there are still several directions this study can take in the future. We did some initial experiments to find trends in a film's success (e.g., ratings) across genres (e.g., Drama, Thriller, Comedy). Future work can extend this to find differences in character portrayals across genres.

In terms of studying characters, future work can also continue investigating character archetypes by using more complex techniques than computing text similarity on the generated LDA topics. Instead, a crowdsourcing approach can be taken, where workers (e.g., on Amazon Mechanical Turk) can identify which, if any, archetype a topic most closely maps to. This could provide

additional insights on how characters conform to existing archetypes. Another extension is looking at combinations of a character's top topics, as opposed to only their singular principal topic.

Lastly, it could be interesting to study the *qualitative* aspect of character interactions in addition to the quantitative (i.e., the centrality measures). Similar to how we ran topic modeling on character dialogues, future work can analyze the quality and semantic meaning of the dialogue lines that form character interactions. This can lead to a better understanding of which interactions are most significant in a film.

Overall, we were able to successfully run initial experiments exploring the field of films and characters through a gender lens. Our results indicate that we as consumers should remain mindful of the media we consume and be conscious of how it may represent gender. We hope that with continued work, more insights in the area of media representation can be gathered in the future.

## 7 Acknowledgements

Thanks to Professor Ani Nenkova for her technical advice and feedback in guiding the direction of this independent study, to Professor Clayton Greenberg for his continued support, and to Professor Jonah Berger for his product virality expertise.

## A References

Apoorv Agarwal, Sriramkumar Balasubramanian, Jiehan Zheng, and Sarthak Dash. 2014. Parsing Screenplays for Extracting Social Networks from Movies. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLfL) @ EACL 2014*, Gothenburg, Sweden.

Apoorv Agarwal, Jiehan Zheng, Shruti Vasanth Kamath, Sriram Balasubramaniann, and Shirin Ann Dey. 2015. Key Female Characters in Film Have More to Talk About Besides Men: Automating the Bechdel Test. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, Denver, Colorado.

Xavier Bost, Vincent Labatut, Serigne Gueye, and Georges Linares. 2016. Narrative Smoothing: Dynamic Conversational Network for the Analysis of TV Series Plots. In *DyNo: 2nd International Workshop on Dynamics in Networks, in conjunction with the 2016 IEEE/ACM International Conference ASONAM.*, San Francisco, California.

Philip John Gorinski and Mirella Lapata. 2015. Movie Script Summarization as Graph-based Scene Extraction. In *Proceedings of NAACL-HLT 2015*, Denver, Colorado, USA.

C. J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *AAAI Publications, Eighth International AAAI Conference on Weblogs and Social Media*, Ann Arbor, Michigan.

Margaret Mark and Carol S. Pearson. 2001. *The Hero and the Outlaw: Building Extraordinary Brands Through the Power of Archetypes*. McGraw Hill Professional.

Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. Available at http://mallet.cs.umass.edu.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.

Victoria Lynn Schmidt. 2001. *45 Master Characters*. F+W Media.

Weizhen Sheng. 2020. Film Screenplay Analysis. https://github.com/w-sheng/Film-Screenplay-Analysis.

# B Appendix

| Topic | Keywords | Jungian | Hero | Heroine |
|---|---|---|---|---|
| 5 | love phone night kiss happy music beautiful... | The Everyperson | Dionysus: "The Woman's Man" | Persephone: "The Maiden" |
| 8 | door back room sits smiles hand face... | The Everyperson | Hephestus: "The Inventor" | Hera: "The Matriarch" |
| 9 | good time love home back pause make... | The Jester | Hermes: "The Fool" | Hera: "The Matriarch" |
| 12 | sir control commander ship captain star power... | The Ruler | Zeus: "The King" | Athena: "The Father's Daughter" |
| 22 | father god love man give lord great queen young... | The Everyperson | Hephestus: "The Inventor" | Demeter: "The Nurturer" |
| 25 | back eyes open wall body reaches suddenly... | The Everyperson | Zeus: "The King" | Hera: "The Matriarch" |

Table 13: Select topics and mapped archetypes.

| Archetype | Description |
|---|---|
| The Everyperson | This archetype embodies the good old boy, the girl next door... |
| The Jester | The Jester offers an element of humor. The Jester lightens up tense situations... |
| The Ruler | This archetype is also called the Leader or the King... |
| Dionysus | A fun-loving, sensual man who can't relate to masculine pastimes... |
| Hephestus | A brilliant genius who has the greatest inventions that he uses to support... |
| Hermes | A playful, carefree soul who enjoys his freedom and doesn't worry... |
| Zeus | A powerful leader, even a bit of a control freak (loss of power is death to him)... |
| Persephone | A carefree, childlike woman who prefers to let others handle the details of life... |
| Hera | A strong, supportive, committed woman who sticks by her family no matter what... |
| Athena | A studious and intelligent woman who furthers her career by aligning herself... |
| Demeter | A kind and compassionate woman who sacrifices much in order to help others... |

Table 14: Select archetypes and description snippet.

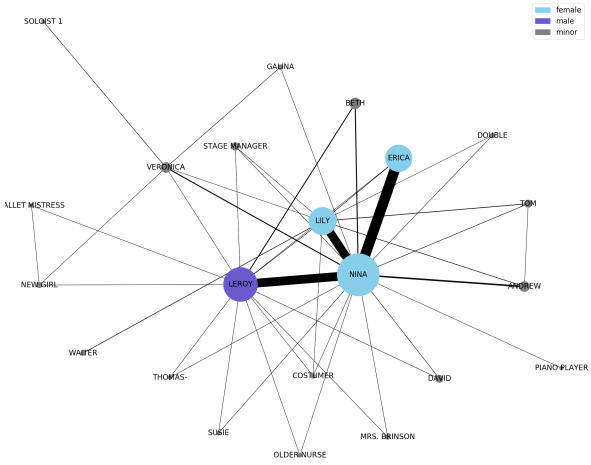


Figure 14: Social network for the film *Black Swan*.

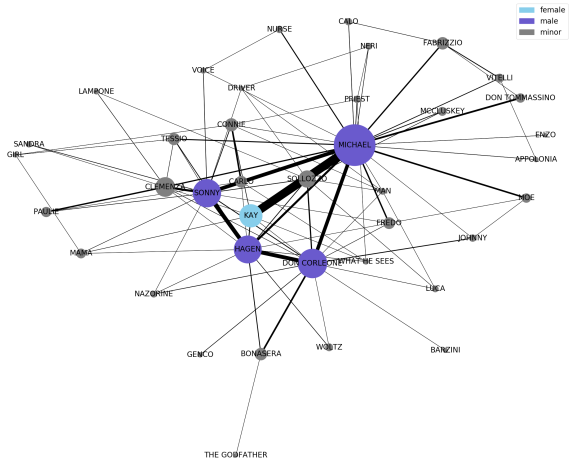


Figure 15: Social network for the film *The Godfather*.