

---

# social characters ,

the hierarchy of gender in  
contemporary fiction ;

Eve Kraicer



McGill

.txtLAB

COLLABORATIONS

---

## Introduction

Gender bias is ubiquitous in contemporary studies on cultural inequality. Which mediums and fields are particularly extreme in this bias? How does the dominance of male-driven narratives, of male faces in [movies](#) and male voices in [newsrooms](#), [boardrooms](#) and [lecture halls](#) effect the rest of us? Are there places where this skew towards men breaks? If so, what can they tell us?

In this study, we ask these questions of contemporary English language fiction. Much research has been devoted to looking at large scale inequalities in the production and evaluation (both [culturally](#) and [economically](#)) of the publishing industry, yet these most often look outside of the novelistic world - [authors](#) and [reviewers](#). Instead, we add to a small but growing area of analyses that looks at the genders *within* these works. In studying the characters inside the works, we examine the ways women characters are relegated to certain positions in fiction. To do so, we study two types of gender representation:



### visibility

1. How many women are there in our dataset?
2. How often are women the leading characters of novels (the protagonists)?
3. What is the most common gender pairing of leading pairs in a novel (whether as companions or antagonists)?

### connectivity

1. Are women more likely to interact with other women, or with men?
2. Are interactions between two women more emotionally negative or positive than between two men?
3. Is there a gendered aspect to the construction of social balance within novels?

For each of these measures, we look both at how trends emerge across genres, and across author genders. In doing so, we try to account for variables that may influence the visibility and connectivity of characters in these works. Ultimately, what we find is troubling. **Women take up less social space in our dataset, and when they do appear, rarely connect to other women.**

Why does this matter? We take this to be an issue of cultural inequality. The hierarchical structures that fictional women or girls are written into are intertwined with our conceptions of what womanhood and girlhood mean, and how much space and attention they are meant consume. Looking at the quantitative ways these inequalities manifest themselves in books gives us a deeper understanding of the persistence of certain forms of gender bias.

## Data

For this study, we used a collection of 1,333 novels from seven distinct genres. Our genres include thematic categories (Mystery, YA, Science Fiction and Romance) as well as social distinction categories (Prizewinners, Reviewed in the New York Times and Bestsellers). For each genre, we also record the number of unique authors, and the percentage of works written by women. A breakdown of our titles and authors by gender is shown in **Table 1**. The full essay upon which this paper is based can be found **here**. For a complete explanation of the data collection, and methods used to control for selection bias, see **Appendix**.

Genre	Code	Novels	# of Authors	% Women Authors
Science-Fiction	SF	192	155	31
Prizewinners	PW	208	188	41
Bestsellers	BS	195	96	42
NYT Reviewed	NYT	180	179	49
Mystery	MY	188	140	52
Young Adult	YA	174	144	85
Romance	ROM	196	172	98

**Table 1:** Summary of dataset by genres


## Limitations

Before discussing our findings, we want to bring attention to a few limitations and simplifications we make throughout our research.

**All of our assessments of character gender are based on predictions,** and therefore contain some degree of error. We discuss our handling of this in the Appendix and in the [full study](#).

**We study gender bias without looking at other identity positions.** This means our model flattens the category of “women” to encompass the intersecting identities and privileges that that label can contain. More research is needed to parse out how gender bias interacts with issues of [race or ethnicity](#), [class](#), [ability](#), and [sexuality](#).

**We use the gender binary.** We do not believe this is indicative of the diversity of gender identity and expression, either in fiction or reality. But given the high rates of pronouns that signal the binary in our dataset (about 75% of all character mentions), we believe it is an important site of study.



*For a fuller discussion of our data preparation and error correction, see the **Appendix**.*

## Visibility

How visible are women in contemporary fiction? It seems like a straightforward question of counting characters, and comparing the number of women to those of men. This would give us some traction on equality.

But there are other ways that we can understand visibility - namely, how often these women appear. Not every character in a novel, after all, is equally visible. To understand this better, we look at ranks of characters by their mention counts. Are there any books where we actually see more women than men? How prominent are women in novels when looking at how often they occur in the text? This section seeks to answer these questions.

### All Characters

**37.8%** of the 26,450 characters we tested were *predicted to be women*

This number is based on measures looking only at the top 20 characters of each novel. When we look instead at all characters, we see the percentage of women drop to under **20%**, but given the high percentage of unlabelled characters past the twentieth position, we do not put as much weight on that number.

Not only does this skew occur overall, but also at the level of rank. To get ranks, we order every character according to their number of mentions per 100,000 words. The characters mentioned the most in every novel (the protagonists) occupy rank position one. With the exception of protagonists (which we will discuss next), there are fewer women than men at every rank position (**Figure 1**). Not only this, but the number of women decreases proportionally with the average number of mentions of a character. **In other words, as we move farther and farther down our character rank lists, we see fewer and fewer women.**

### Genre and Author Gender

**Genre has no significant effect on the percentage of women.** That's not to say there is no variation - **Table 2** shows the range across genre, with YA having the highest percentage of women. But statistically, the variation is slight enough that these numbers are not meaningfully different.

It's seemingly a rule of fiction. Despite appealing to very different readerships - people who like prizewinners or YA or Romance - books, on average, write just under  $\frac{2}{3}$  of their characters as women. It approximates the 2-for-1 rule: write two men for every one woman.

Genre	BS	PW	NYT	MY	SF	ROM	YA
% Women	36.1	36.5	36.9	36.9	37.7	38.2	40.1

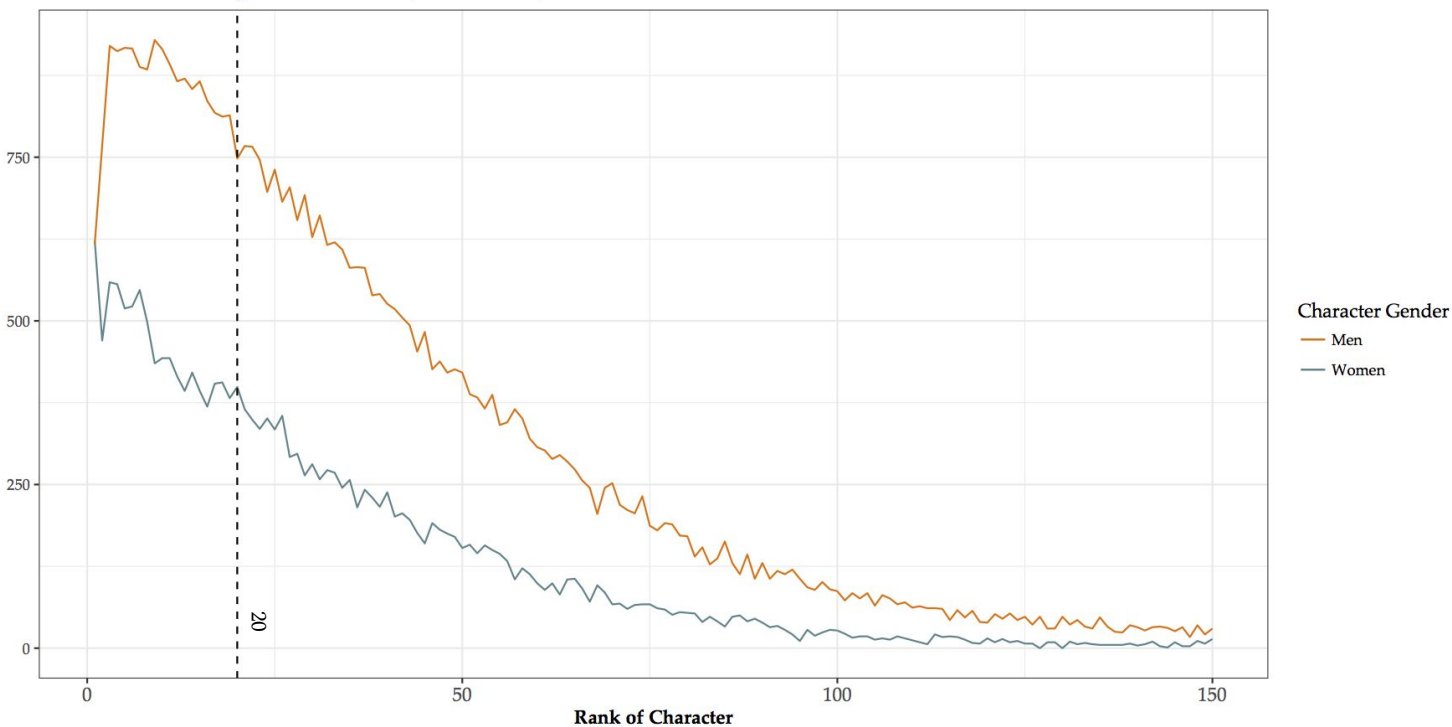
**Table 2:** Percent of women in the top twenty characters by genre

**Women authors are 1.18 times more likely to use women characters in their stories.** Although this indicates that increasing the number of women authors can shrink gender bias among characters, it can not equalize representation, at any rank position except the protagonist (**Figure 2**).

---

---

**Men and Women by Rank Position (All Authors)**

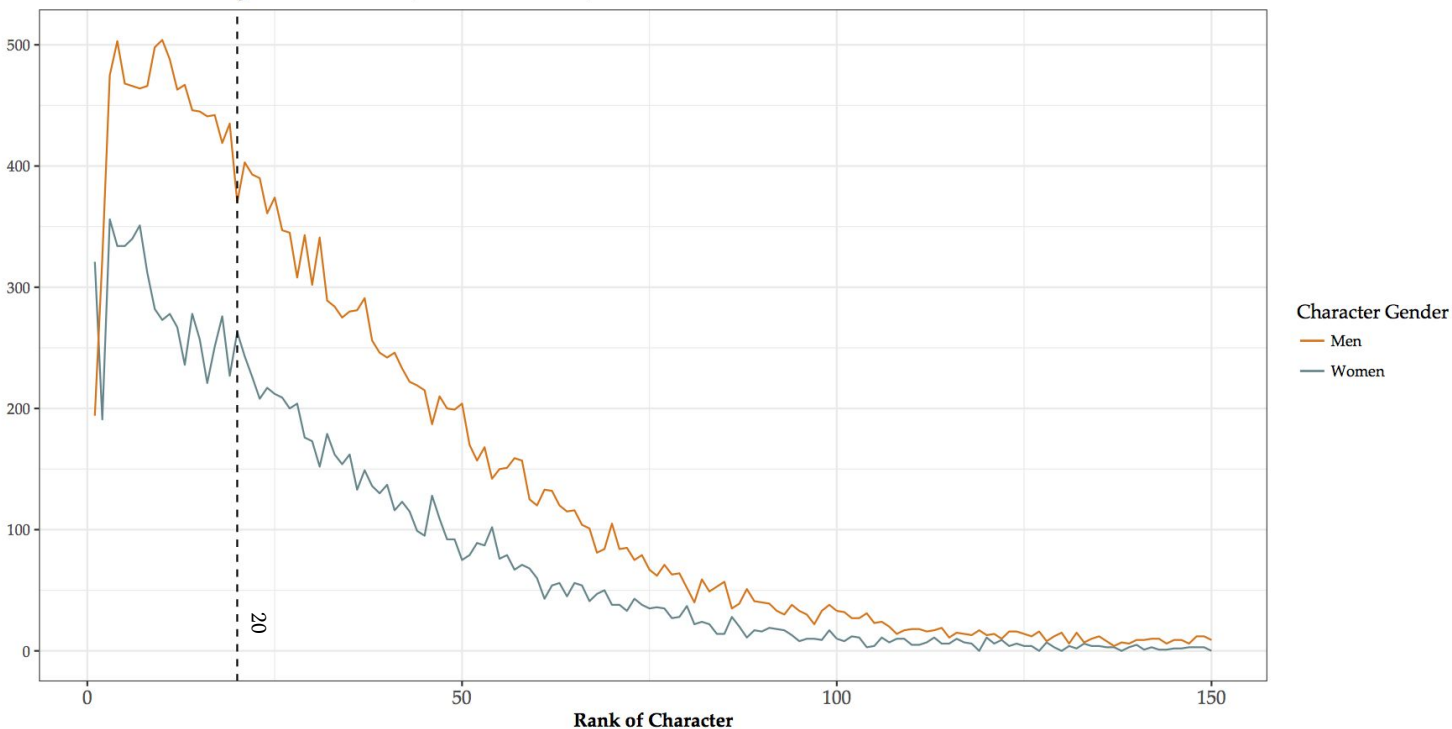


**Figure 1: Number of men and women characters by rank position(all authors)**

---

---

**Men and Women by Rank Position (Women Authors)**



**Figure 2: Number of men and women characters by rank position (women authors)**

---

---

---

---

# Even within books written by only women, there are *more characters that are men*

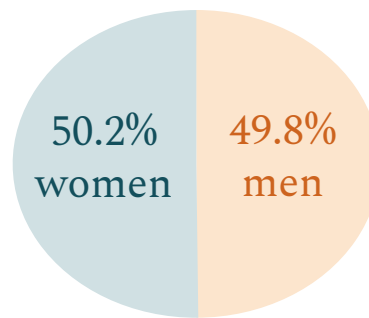
---

---

## Top Pairs

The overall counts of men and women, however, doesn't fully capture the experience of reading - there are always going to be certain characters that appear considerably more often and more centrally in narratives than others. To look at this, we looked into the gender distribution of the protagonist, and the gender relationship of that character to the next most mentioned character (the antagonist or companion).

**Across protagonists, we found almost a perfect 50/50 split between men and women in our dataset<sup>1</sup>.** In fact, there is actually a slightly higher rate of women than men. Although we take this as a positive gesture in terms of gender equality, this finding is surprising. On the one hand, it suggests that women are often the main voices in these stories. On the other, given that at all other rank positions women fall below men, we wonder if this might indicate some form of tokenization. If she is centralized, do the inequalities in the characters surrounding her become harder to see?



One way we can get traction on this notion of tokenization is to see how gender distributions operate among pairs of protagonists and antagonists. Do men and women pairs appear at equal rates here, too? No. **Books in our dataset are 1.6 times more likely to use two men than two women as the central pair in a novel:**

50.9%

of top character pairs are made up of **one woman and one man**

30.4%

of top character pairs are made up of **two men**

18.7%

of top character pairs are made up of **two women**

When looking across genre, we again found **no significant difference in the ratios of men to women as protagonists**. We also do not find that genre influences the ratio of man-man to woman-woman pairs at the center of the novel.

Here we find more seeming rules of gender in fiction - women characters can be half the protagonists, but despite this, women are significantly less likely to occupy both the position of protagonist and antagonist / companion compared to men.

	all books	only books by women
women protagonists	50.2%	62.3%
women top pairs	18.7%	25.6%

In both measures of protagonist and antagonist pairs, women are more likely to write women. This is the same for men, but in both cases, **men are significantly less likely to centralize women than women are to centralize men**.

## Connectivity

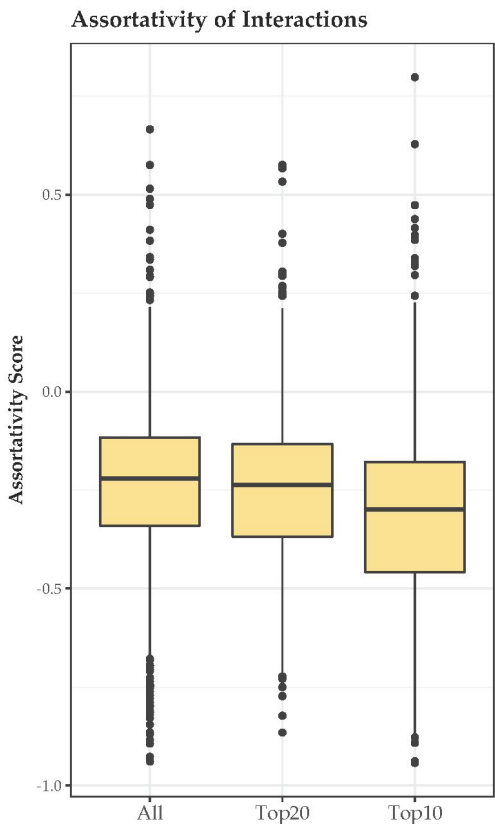
After identifying these strong patterns across our dataset in terms of gender visibility, we wanted to see if any gender patterns held across connections between characters. To do this, we looked at the interactions between men and women, which we define as co-occurrence between the bounds of a sentence.<sup>2</sup> How often are these interactions happening in different genres and across author genders? What is the general sentiment of interactions between women compared to men? How does gender influence social balance?

Through these measures, we try to better understand how gender plays out between characters, and what social roles women and men play in fiction.

## Assortativity

The first way we quantify this question is through the concept of gender assortativity. Assortativity looks at the ratio of same-gender pairs to mixed-gender pairs in a given novel. To calculate this, we look at all of the interactions in a book (the mean is 2,038 per novel). We then look at the number of those which are either between characters of the same (positive scores) or different genders (negative scores). On average, whether looking at all characters, characters in the top twenty, or in the top ten, we find only negative scores (**Fig. 3**)<sup>3</sup>.

In other words, men and women are much more likely to interact with one another than either men with men or women with women<sup>4</sup>. This mirrors our findings among protagonist-antagonist pairs, which also skew towards mixing men and women. Beyond this, we also find three other significant findings:



**Figure 3:** Average assortativity by character subsets, with increasing negativity

1. The average assortativity score decreases with more central characters, meaning **the preference for mixed gendered interactions is more pronounced with more important characters (Fig. 3).**
2. **Women authors write more strongly disassortative (mixed gender) networks than do men.** While women authors increase the number of women in the novel, they also increase the number of mixed-gender interactions, especially among the less prominent characters.
3. With the exception of Romance (which has significantly lower scores at every subset) **there was no significant difference across genres when looking at all interactions.**

This last finding is of particular interest to us. **What it suggests is that, as with visibility, the rules and patterns around gender structure in novels operate separately from genre.** This occurs despite a huge range in the number of interactions across genres; for example, on average, Bestsellers have 2,768 interactions, and YA novels have only 1,417. That is, these patterns of assortativity are seemingly not controlled by what a novel is ostensibly about, but rather by the fact that it is a novel, at all. Given the consistency of these negative scores, we identify this trend as a **heteronormative bias** in fiction.

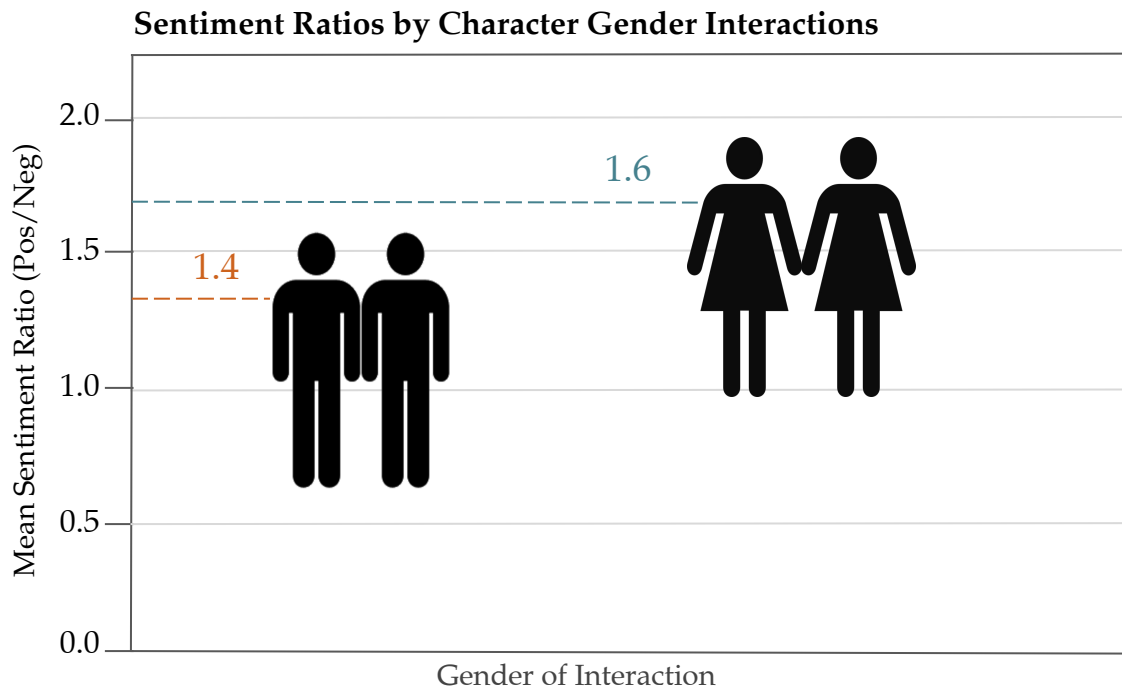
**On 'Heteronormativity':** Technically, [heteronormativity](#) describes the promotion of a worldview where heterosexuality is the proper sexuality. In addition to sexual identity, it is predicated on the denial of non-binary identities, including gender non-conforming and trans\* folks. We choose to use the term because we feel that the strength of the assortativity patterns in our novels are indicative of a particular worldview in fiction - men and women should be paired together. We do not have any data on the characterization of these interactions - and of course not all interactions between two characters are sexual, but the notion of what is 'normal' for novels is signalled, nonetheless.



## Sentiments

Whereas mixed gendered interactions are ubiquitous in fiction, and therefore have the capacity to take many forms, interactions between only men, and especially between only women (which is, perhaps unsurprisingly, the least common interaction type) are much more likely to be tokenized.

Given this, we wanted to get some traction on the sentiments behind the interactions between only men and only women. Using the [Bing Sentiment Dictionary](#), we compared the words co-occurring with a character in a sentence across all same-gendered interactions. **We found that for characters, across all genres women have a significantly higher ratio of positive to negative sentiments when interacting with other women (1.6) than do men with other men (1.4) (Fig. 4)**<sup>5</sup>. Author gender does not significantly impact this pattern, one way or the other.

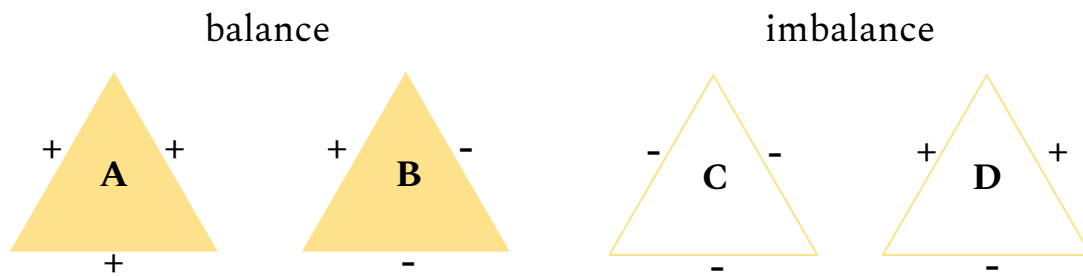


**Figure 4:** Mean sentiment ratios of interactions between two men or two women among the top twenty characters in a novel

How do we interpret the difference in sentiment scores between men and women? The answer here does not seem so straightforward. On the one hand, women pairs seem to be portrayed in more positive ways towards one another. This, conversely, fits into a stereotype of womanhood, one that is conflict averse.<sup>6</sup> This latter finding could also be connected to visibility of women - conflict drives narrative, and women alone are not as conflictual. **Is it possible, then, that amiable sentiments among women drive their underrepresentation?**

## Structural Balance

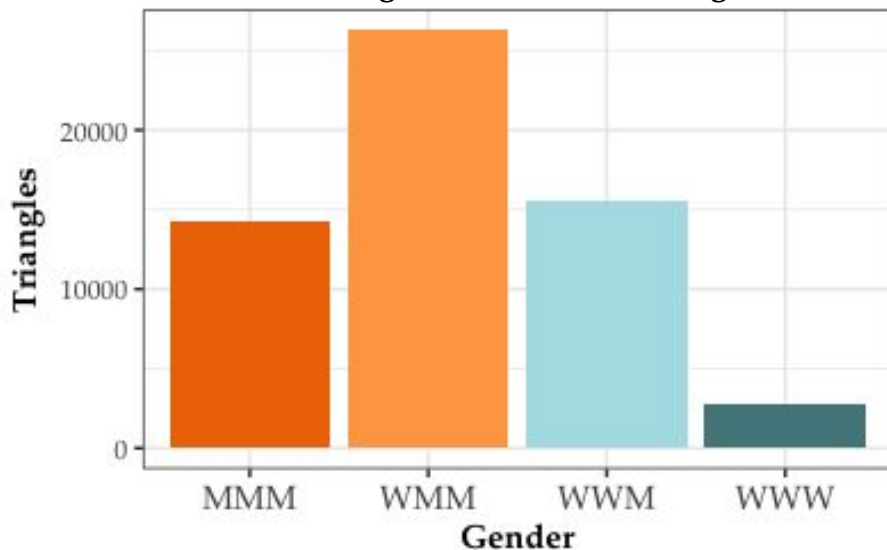
Our final question on connectivity is one of structural balance. That is, how does gender in a transitive triangle (three characters, where every possible pair is connected), influence its structure? We take this measure from sociological studies that describe these triangles as structured in one of two states - it is either balanced or imbalanced, depending on the sentiments between the three sets of pairs (Fig. 5). To do this, first we calculated if the majority of interactions between each pair in a triangle was more positive or more negative, and labelled the three sides accordingly. Second, we classified the triangle as either balanced or imbalanced using the sentiment results. Finally we classified the triangle by its gender configuration. From this, we compared the ratio of balanced to imbalanced triangles by the number of women involved.



**Figure 5:** Configurations of triangles by state. Balanced triangles consist of all positive relationships (A) or one positive relationship, also called 'mutual enemy' (B). Imbalanced triangles consist of all negative relationships (C) or one negative relationship, also called 'mutual friend' (D).

As it turns out, we **did not** find that the number of women in a triangle has a significant effect on whether or not it is balanced or imbalanced. We did, however, find solidification of both the skew towards men, and towards heteronormativity. **The most common triangle by far was constructed with one woman and two men (Fig. 6).** This encapsulates nearly all of our findings - for every one women, use two men. Always pair that one woman between men. But even more striking from this graphic is this necessary inverse to masculine dominance in literature: the underrepresentation of women not only as individuals, but social groups. **We studied 58,831 triangles, and only 2,781 were depictions of three interconnected women.**

**Transitive Triangles and Gender Configuration**



3 men make up **24% of triangles.**

3 women make up **less than 5%.**


**Figure 6:** Number of triangles in our dataset by gender. Instances of three women are exceedingly rare in our dataset.

## Conclusion

Throughout this research, we find three troubling trends.



**First**, we find that fictional women characters overall are underrepresented in contemporary novels. This is ameliorated slightly by the protagonist, who is just as likely to be a woman as a man. We wonder to what extent this may be tokenizing, and allow for the systemic inequalities in the rest of the character list to go unchallenged.

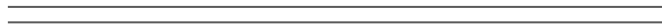


**Second**, we find that women are very unlikely to be paired with other women, whether as a top pair, or across social interactions in general; instead, we find fiction to skew heavily towards heteronormative patterns of connectivity. While research finds that real-world networks become more disassortative as people age (starting highly gendered in childhood and adolescence), they tend to mostly remain same-gendered. Interestingly, this is especially true for women.<sup>6</sup> Taken together, this suggests that social connections in our data set are conditioned by a deliberate effort to signal and foreground fictional heteronormativity.

**Third**, we find that both these structures of gender bias, are relatively genre independent - they operate regardless of the content or structure of the narrative. They are, it seems, formulas of contemporary publishing.

So how do we respond to these findings? Author gender is part of the picture - encouraging more stories from writers who are not men does start to equalize gender representation, but even among women-authored books, we still see this pattern of many characters that are men, talking to fewer characters that are women.

Beyond this, being able to quantify these norms allows us to better understand what is actually happening at the level of characterization. It can provide a concrete template for advocacy and for change. Making fiction less "real" - having men not dominate fictional space - may not necessarily change reality. But it at least gives us an alternative, an imaginative landscape where things might be otherwise.



## Data Preparation

To collect titles for our thematic categories, we sorted by popularity on Amazon.ca with respect to each of the genre tags provided by Amazon. For YA, we combined our Amazon-selected books with Readers' Choice Awards from Goodreads.com. To collect titles for our social distinction categories, we used three methods: for bestsellers, we collected the top 200 titles that had appeared for the most number of weeks on the New York Times bestsellers list since 2000 in descending order. For prizewinners, we collected all shortlisted and winning titles from five major literary awards in the US, UK, and Canada (including the Giller Prize, Governor General's Award, Pen/Faulkner, National Book Award, and Man Booker Award). Finally, for NYT Reviewed, we collected novels reviewed by the New York Times during our time frame.

To account for biases in this data collection, we use bootstrapping methods throughout, and all numbers we report are means based on samplings of 10,000 for visibility measures, and 1,000 for connectivity measures.

### Nodes and Edges

---

To transform our data into nodes and edges tables, we use BookNLP, developed by David Bamman. This allows us to:

1. Identify characters
2. Map characters and their aliases (nicknames as well as pronouns) to a single character ID
3. Generate a **nodes list** of predicted characters ranked by their number of mentions within every novel and a predicted gender, which can include one of three possible options of M, F or ? for unidentified characters
4. Mark character occurrences by sentence boundaries to build an **edges list** of all interactions, including sentiment scores based on the [Bing Sentiment Dictionary](#) for each co-occurrence based on verbs in sentence.

The nodes list provides the raw data for our measures on visibility, whereas the edges list provides the raw data for our measures on connectivity.

### Error Correction

---

Because we are using predictions of character gender, we take two steps to address potential error within our data. First, for all but one of our measures we focus only on the top twenty characters for a given novel. Second, the percentage of non-assigned genders rises considerably as we descend in character rank. Within the top twenty characters, only 11.5% of characters (or just slightly more than 2 per novel) are not labeled. Within the top ten it drops to 5.9%.

Even within the top-twenty characters in a book, however, some gender assignments are inaccurate. To validate this, we manually verified Book NLP's predicted genders. Within our validation data, we found a sensitivity of 83.75% / 78.57% for women in the top two / twenty respectively and a rate of 96% / 96.9% for men, suggesting that women characters are potentially being undercounted in our data. In order to account for this error, we modify our estimates of gender throughout using Rogan and Gladen's approach which elevates our estimates of the presence of women in novels from the raw counts produced by our BookNLP data.

For a full discussion of our methods, see **Kraicer and Piper (2019)**.

## Notes

1. We measure top pairs on a subset of 1,239 books. This subset contains all books we had a predicted gender for, after using separate processes to identify the protagonist in third-person and first person novels.
2. We define an interaction as sentence co-occurrence. This may occasionally capture instances where characters are referenced together without physically or audibly interacting. Although it's imperfect, it is much less subjective than other methods for identifying interactions within novels.
3. For this test, we only use characters with assigned genders.
4. This significance is drawn from the a permutation test. We keep the social structure, the number of women and the number of men in a given book, and then randomly shuffle their positions in the network 1,000 times. After doing this, we find that for interactions, the assortativity score of the actual network is always lower (more heteronormative) than the mean of the random interactions. This suggests there is something deliberate about the ways authors are connecting genders, that extends even beyond the number of characters they choose to write and in the ways they structure their network.
5. Although all genres use more positive sentiments among women than among men, there is a significant difference in the range of the difference. Bestsellers have the strongest difference between women and men (0.33), while NYT exhibit only a difference of 0.07.
6. See Kristine J. Ajrouch, Alysia Y. Blandon, and Toni C. Antonucci. "Social networks among men and women: The effects of age and socioeconomic status." *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 60. 6 (2005): 311-317.

## Acknowledgements

This research exists because of the efforts of many people and organizations. I'd like to thank both McGill University and SSHRC (Social Sciences and Humanities Research Council) for funding this research.

I'm entirely indebted to the brilliant and supportive team at txtLAB who I've had the opportunity to work with. To Beata and Alayne for keeping me on track, to the RA's for pushing me and inspiring me and talking me through coding errors / bad graphs / that time we ran out of coffee beans, and to Prof Piper for encouraging me to Take On a Piece of The Hetero-Patriarchy (and for supporting my overuse of that term throughout this process): thank you.

## Works Cited

- Anderson, Hanah and Matt Daniels. "Film Dialogue." *The Pudding*. April 2016, [Link](#).
- Chant, Sylvia. "The links between gender and poverty are over-simplified and under-problematised." *LSE Politics and Policy*. March 10, 2011, [Link](#).
- Cima, Rosie. "Bias, She Wrote." *The Pudding*. June 2017, [Link](#).
- Cochrane, Kristen. "Why Heteronormativity is a Bad Thing." *Teen Vogue*. September 1, 2016. [Link](#).
- Crenshaw, Kimberlé and Abby Dobson. (2016, October) *The Urgency of Intersectionality*. TEDWomen 2016. [Link](#).
- Griffith, Nicola. "Books about women don't win big awards." May 26, 2015. [Link](#).
- Grove, Jack. "Proportion of female professors up, but still below a quarter." *Times Higher Education*. February 28, 2015, [Link](#).
- Ibarra, Herminia and Morton T. Hansen. "Women CEOs: Why So Few?" *Harvard Business Review*. December 21, 2009, [Link](#).
- Piper, Andrew and Richard Jean So. "Women Write About Family, Men Write About War." *The New Republic*. April 8, 2016, [Link](#).
- Turlock, Amanda. "Women with Disabilities." *Stats Canada*. May 29, 2017, [Link](#).
- Weinberg, Dana B., and Adam Kapelner. "Comparing gender discrimination and inequality in indie and traditional publishing." *PloS one* 13, no. 4 (2018): e0195298.
- York, Catherine. "Women dominate journalism schools, but newsrooms are still a different story." *Poynter Institute*, September 18, 2017, [Link](#).