Analyzing Book Checkout Trends at the Seattle Public Library

Sarvesh Fotedar, Ekrem Kizilkaya, Abbas Shaikh, Cody VanZandt, Andy Wang

Contents

1 Introduction

From classic literature to contemporary bestsellers, the Seattle Public Library's book collection offers a rich window into its surrounding community's literary tastes and interests. By analyzing checkout trends and book movement, we can uncover which novels and authors are most beloved among library patrons, giving us remarkable insight into the community's unique cultural and social context.

This report conducts a preliminary examination of approximately 41 million checkouts from the Seattle Public Library system between 2005 and 2022. We intend to combine this information with Goodreads review data from the UCSD Book Graph to answer general questions about the contours of contemporary American readership and even perform specific analyses on the competitive dynamics of the book authorship and publishing industry.

From this point forward, the following questions to guide our exploration and report:

1. Why do some books enjoy widespread acclaim while others fizzle?
2. Do popular books explode onto the scene or accumulate readers more gradually?
3. How do the popular success and critical acclaim of novels differ?
4. How has recent conglomeration of publishers changed the literary marketplace?
5. And, ultimately, to what degree can the success or failure of a book be predicted?

From the most popular genres and titles to the unique factors influencing readers' choices, book checkout data offers a rich window into the complex relationship between readers, libraries, and the larger cultural context. In this paper, we will examine the book checkout trends at the Seattle Public Library in detail, drawing on both statistical and literary analysis to shed light on the unique reading culture of this vibrant community.

2 Datasets

2.1 Seattle Public Library

Our primary dataset comes directly from the Seattle Public Library and has approximately twelve million rows for book checkouts spanning the five years between 2018 and 2022. Specifically, each row corresponds to a monthly count of checkouts for the physical or electronic version of an item. Items are not just limited to books; the Seattle Public Library also makes video discs, e-books, and sound discs available to patrons, to name a few. Each row is described by eleven variables, which are described in Table 1.

Variable	**Description**
UsageClass	Whether the item was physical or digital
CheckoutType	The tool or vendor that was used for checkout
MaterialType	The item type (ex: book, video disc, etc)
CheckoutMonth	The four digit checkout year
CheckoutYear	The month of checkout
Checkouts	The number of times that the item was checked out within the checkout month
Title	The full title and subtitle
Creator	The author or entity responsible for creating
Subjects	The subjects as they appear in the library catalog
Publisher	The publisher of the title
PublicationYear	The year that the item was published, printed, or copyrighted

Table 1: Seattle Public Library Variables

2.2 UCSD Book Graph

Our supplementary dataset comes from the UCSD Book Graph initiative and contains over fifteen million reviews for approximately two million books from 465,000 users. Each row of the dataset is a JSON object which represents a single review and its associated metadata. The nine attributes of each JSONified review are described in Table 2

Attribute	Description
user_id	The UUID of the reviewing user
book_id	The unique numerical id for the reviewed book
review_id	The UUID of the review itself
rating	The numerical rating of the book out of 5
review_text	The text content of the review
date_added	The date the review was published
date_updated	The date the review was updated, if applicable
n_votes	The number of votes endorsing the review
n_comments	The number of comments for the review

Table 2: UCSD Book Review Object Attributes

3 Exploratory Analysis

3.1 Total Checkouts per Month

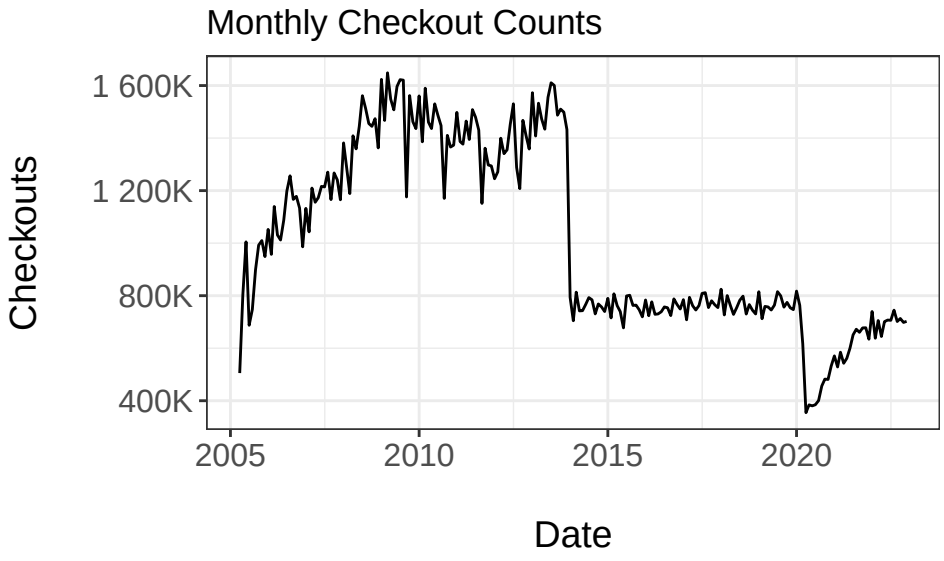

Figure 1: Total Checkouts per Month

Figure 1 shows the change in the total number of checkouts per month over the entire period from 2005 to 2022. The initial increase in checkouts from around 2005 to 2009 is most likely due to the adoption of the online checkout system which enabled data collection. There is also a steep drop in checkouts during 2020, which we are certain is due to the COVID-19 pandemic. Since then, the plot shows that checkouts have been recovering quickly and are almost at pre-pandemic numbers.

3.2 Total Checkouts per Month, by Medium

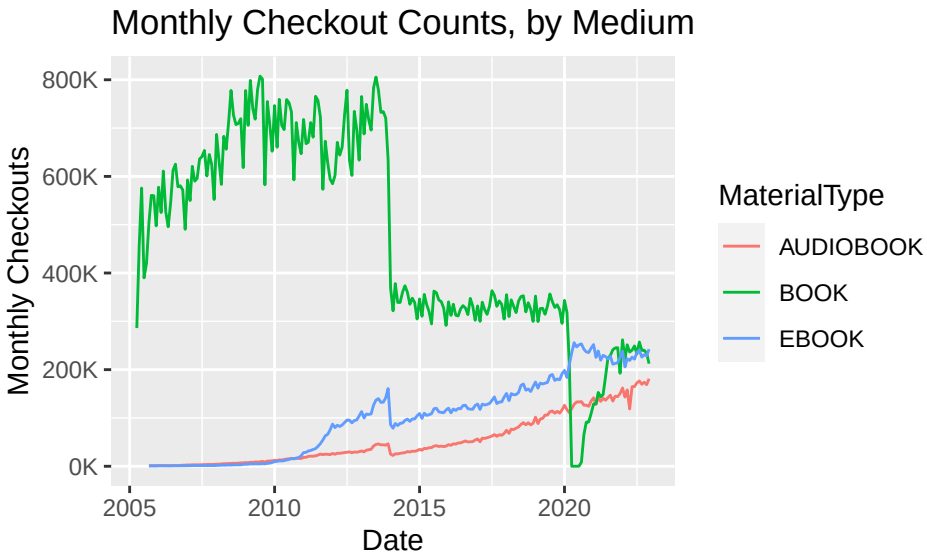

Figure 2: Total Checkouts per Month

Figure 2 is similar to Figure 1, but is broken down by the three most popular mediums: physical books, audiobooks, and e-books. The variation in checkout counts of physical books is almost identical to the pattern exhibited in Figure 1, including the steep drop in 2020. Interestingly, the checkouts of audiobooks and e-books were completely unaffected by the pandemic, most likely due to the ability to check them out online without physically visiting the library.

3.3 Relative Popularity of Mediums

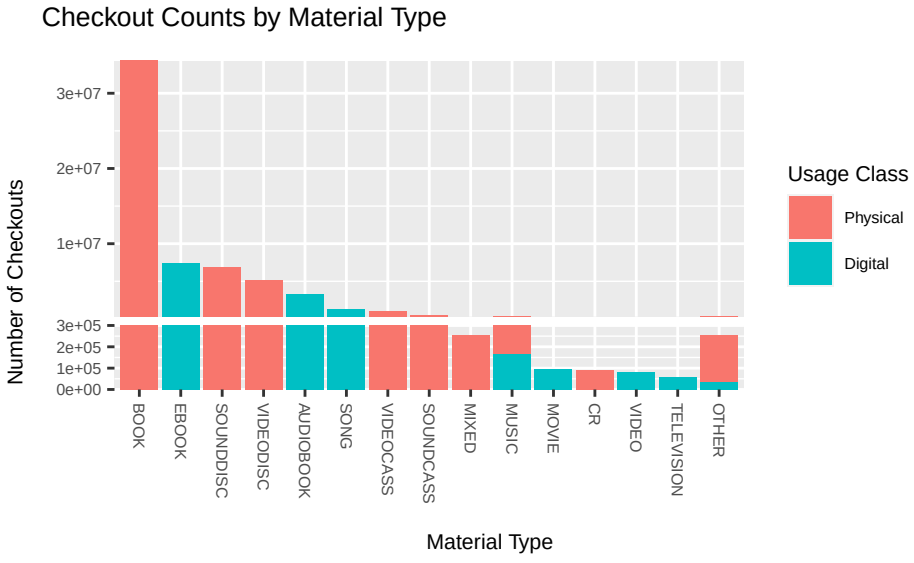

Figure 3: Checkout Counts by Medium

Figure 3 shows the relative popularity of the types of items that the Seattle Public Library makes available for checkout. Physical books are by far the highest, the checkouts of the next highest two categories (ebooks and sound discs) being around a quarter of the checkouts of physical books. This explains why in Figure 2, the pattern of physical book checkouts follows the total number so closely.

3.4 Genre Popularity Among Physical Books

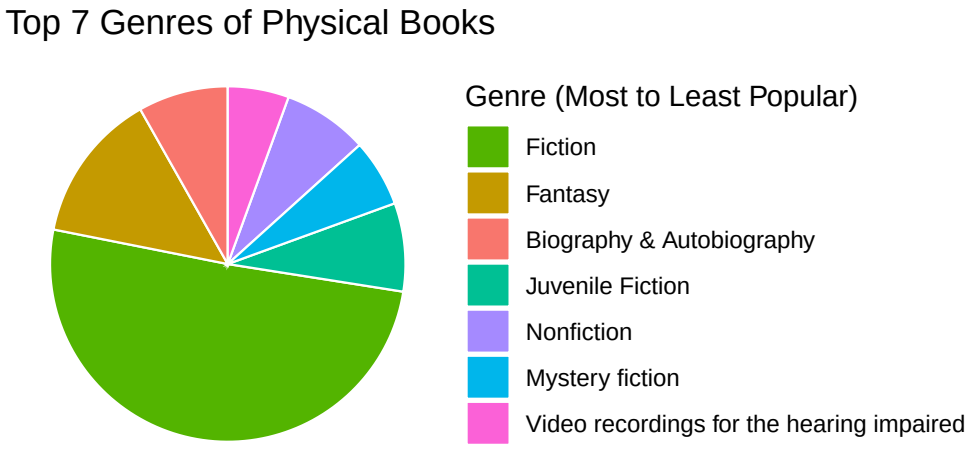

Figure 4: Genre Popularity

Figure 4 shows the relative popularity of the genres of physical books at the Seattle Public Library. Fiction is overwhelmingly popular, accounting for the genre of over 50% of the books that were checked out over the period.

4 Data-Driven Insights

4.1 Publisher Conglomeration

We aim to explore how the conglomeration of the publishing industry affects the literary marketplace, particularly in the microcosm of the Seattle Public Library. As such, we investigate the number of checkouts attributed to the top 5 percent of publishers every month by number of checkouts. We hypothesize that his metric will represent conglomeration within the publishing industry since a small portion of publishers controlling a disproportionately large part of the industry indicates greater conglomeration and less diversification of publishers used by the library. Figure 5 displays the number of checkouts per month accounted for from the top 5 percent of publishers overall, as well as separated by book, e-book, and audiobook publishers. What we find is that the proportion of checkouts controlled by top book publishers has been steadily decreasing, indicating a possible greater diversification in the publishers for physical books, whereas the number of checkouts for audio books and e-books from top publishers has increased consistently to around 90 percent currently. This suggests that large conglomerate publishers have grown to control the digital publishing industry as digitization of library content has become more common and more profitable in recent years.

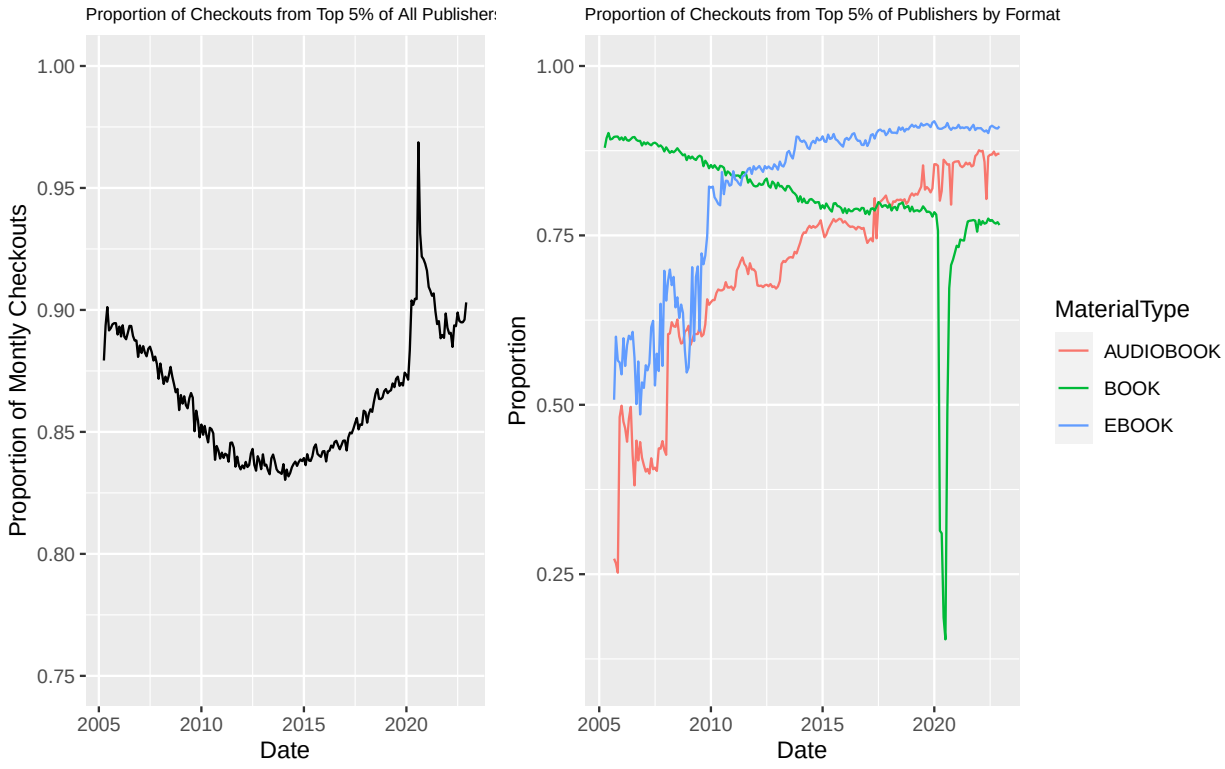

Figure 5: Measure of Publisher Conglomeration

4.2 Reading Lines

As the leaves began to turn in 2018, Sally Rooney debuted her second novel, *Normal People*. Rooney, an Irishwomen of scarcely 28 years, had already enjoyed widespread critical acclaim on both sides of the Atlantic with her breakout novel, *Conversations with Friends*. In much the same fashion, *Normal People* exploded out of Rooney's native Dublin, enthralling London readers by the thousands before crossing the ocean to pick up tens of thousands more in New York and beyond. New readers, however, were not the only souvenir of the book's transatlantic trek. It also gained a longer title. Styled on the continent simply as *Normal

People*, the book was marketed by publisher Hogarth Press to American audiences as *Normal People: a Novel*. The addition, known in the trade as a reading line, has been deployed by printers and publishers for the last 300 years, all to ensure their books end up – and stay – in the hands of readers. Of reading lines, we ask this: was Hogarth Press right? Do books with reading lines sell better than those without? Or to ask an even stronger version of the question: given two books that are otherwise *identical*, does the public prefer versions with or without readline lines?

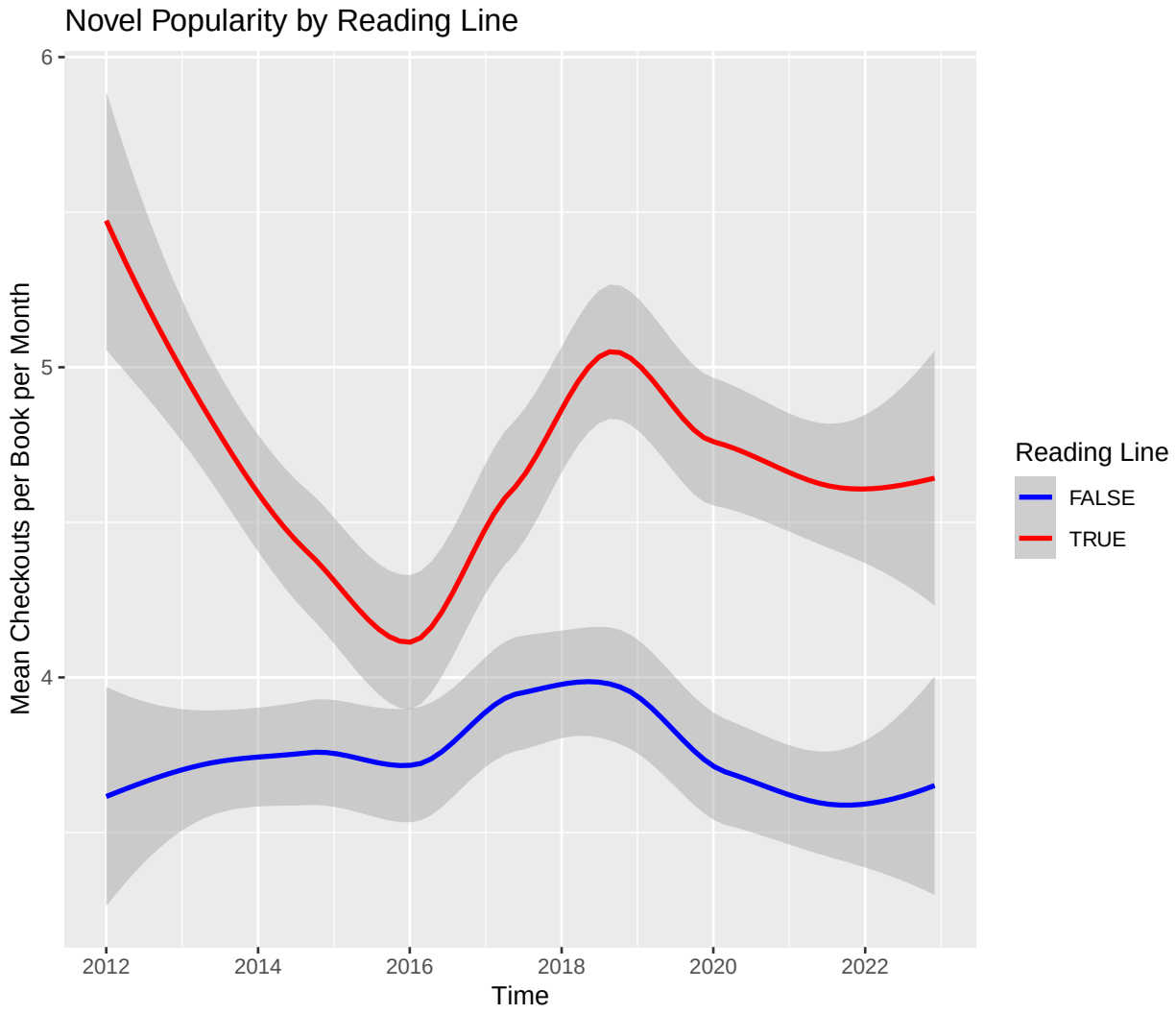

Figure 6: Are Books with Reading Lines More Popular than those Without?

The evidence – at least for the last decade – is compelling: books with reading lines outsell books without, even when the books themselves are identical. But the story is a bit more complicated. While the average lined book is checked out more than its unlined counterpart, that difference is largely driven by a handful of super-seller mega novels that dominate the marketplace for years. Think *Where the Crawdads Sing: a Novel*, *Less: a Novel*, and Sally Rooney's own offerings. When one considers total checkouts rather than average checkouts, the story reverses and unlined books lead checkouts by 50 percent. It is more accurate, then, to say that reading lines help the *most popular* books scale the heights, while doing little for books and languish further down bestseller lists – an interpretation corroborated when one notes that lined books make up 70 percent of top-20 bestsellers but only 40 percent of the top 1000. Reading lines, then, are a tool of

momentum: they help great books sell even better while offering comparatively little to more marginal titles.

Having established the value of reading lines, an author who aspires to the top of charts might next wonder which publishers are the best at affixing ": a Novel" and therefore bringing a book one step closer to superstardom. We can identify those elite literary purveyors by comparing the proportion of lined books to the average checkout numbers for each publisher.

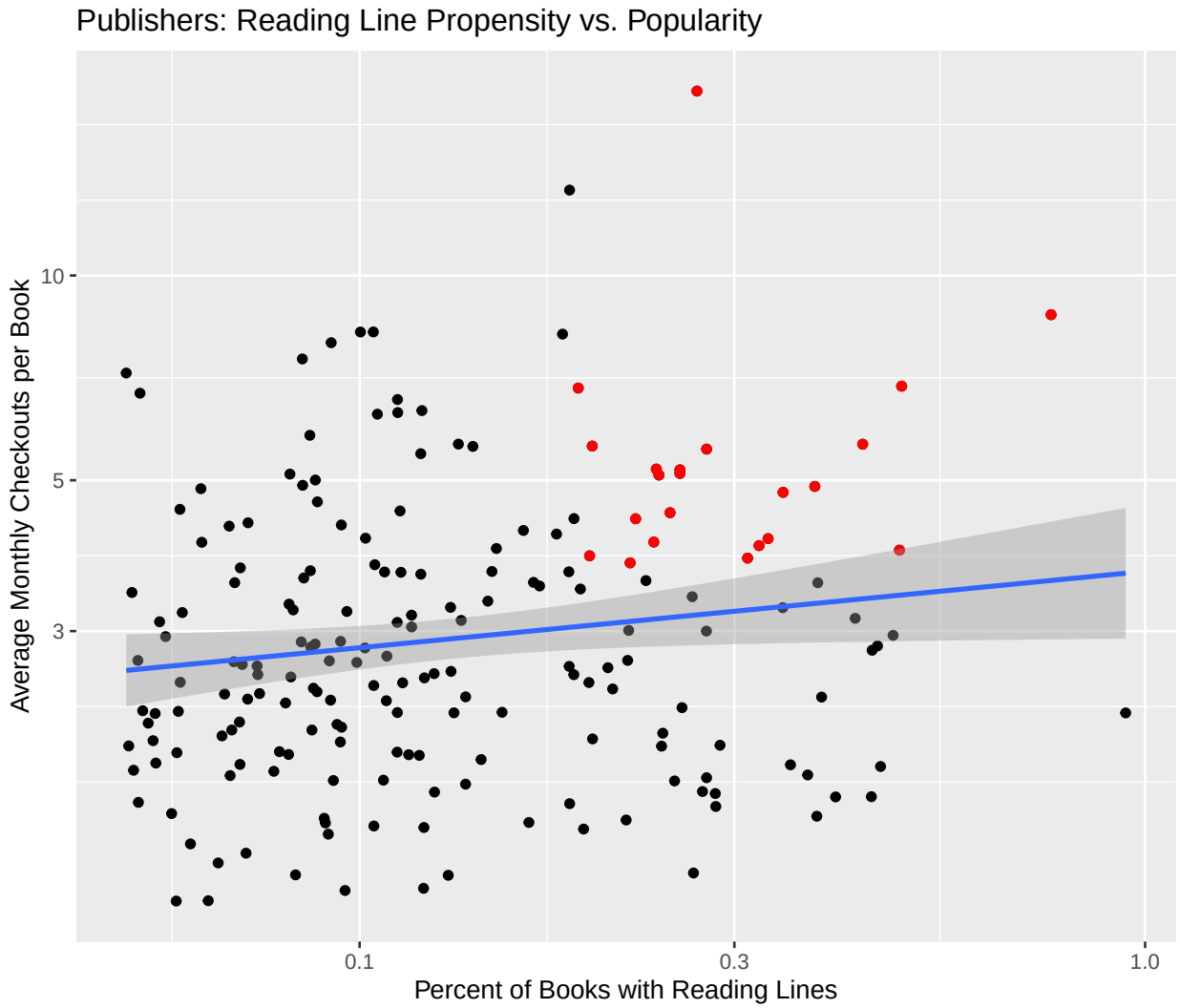

Figure 7: Publisher Tendency Towards Reading Lines vs. Book Popularity

The data – which incidentally affirms a positive relationship between reading lines and popular success – identifies a cadre of publishers who find tremendous popular success in affixing ": a Novel" to their titles. Certainly, books can certainly succeed with publishers who eschew lined novels, but an aspiring Sally Rooney may not wish to take that chance and instead prefer one of publishers the previous graph highlights in red.

Figure 8: Publisher Tendency Towards Reading Lines vs. Book Popularity

Among this list are the world's most successful literary publishers – small, elite specialists each in the kind of capital-L Literary fiction that routinely tops critical and popular lists. Crown's space-thriller *The Martian: a Novel*, Scribner's mega-hit *All the Light We Cannot See*, Atria's *Anxious People*, and leading the pack in combined popularity and propensity for reading lines: Hogarth Press, the American publisher of Sally Rooney's bestsellers.

In short, reading lines matter. Books with ": a Novel" affixed are more successful on average and make up a far more than their due share of bestsellers and critical darlings. But indiscriminately appending a reading line is unlikely to be a winning strategy. Indeed, one is best served by leaving that particular editorial decision to the literary minds behind the anglophone world's most successful independent publishers. It did, after all, work for Sally Rooney.

4.3 Controversy

Some books we love, some books we hate, and still others we *love* to *hate*. In this section, we present a novel method for identifying the most controversial books in our corpus of Goodreads review data. Among these

books are the political polarizing, the socially stigmatized, and chart-toppers that get everyone talking to each other, if perhaps not very nicely.

We first highlight our simple formula for controversy and division: $\frac{pn}{|p-n+1|}$ where p is the number of 5-star or **p**ositive reviews and n the number of 1-star or **n**egative ones. The idea behind the formula is simple: the numerator rewards raw popularity, while the denominator punishes any book with an imbalance of positivity and negativity. After all, a truly controversial book, needs to be both wildly popular *and* deeply divisive, inspiring love and hate in near equal measure among its dedicated devotees and determined detractors.

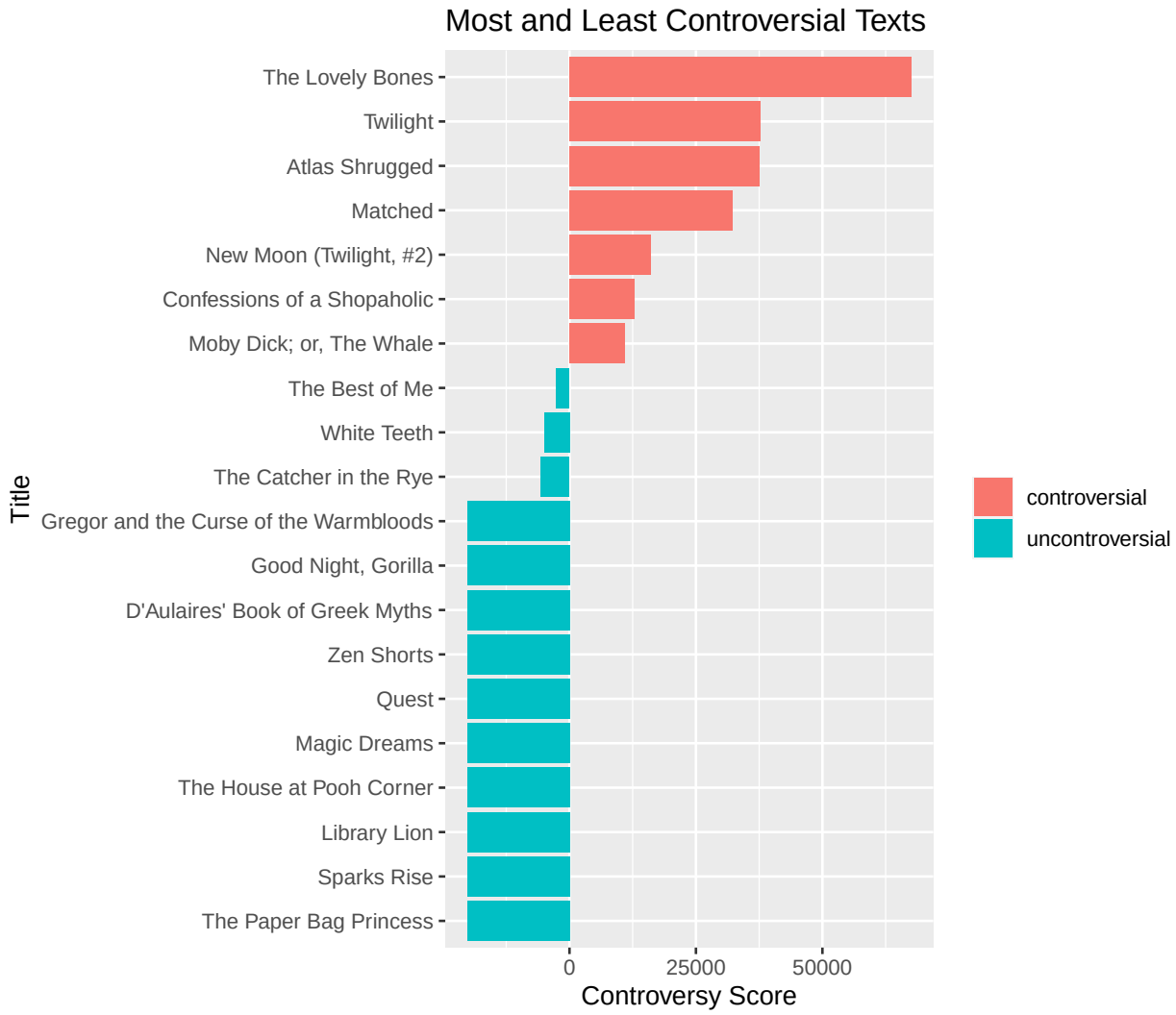

Figure 9: Publisher Tendency Towards Reading Lines vs. Book Popularity

Topping the list is Alice Sebold's *The Lovely Bones*, a mainstay on banned-books lists that chronicles the fictional story of a teenage girl who after being raped and murdered observes her family's grief from the afterlife. The second is Stephanie Meyer's young adult vampire romance, Twilight. A book that, for better or worse, needs no introduction. Even so, Stephanie Meyer wins the dubious honor of being the only author to feature twice in the top five (and indeed, the top 50) most controversial titles. By a completely different token, the third book is Ayn Rand's divisive opus of Objective philosophy and libertarianism. In truth,

the analysis writes itself, although it is still worth pausing the admire the diversity of controversy. Between philosophical treatises, young adult fiction, and classic literature, controversy seems to cut across taste, time, and genre. Controversy, then, is a slippery term that inhere in authors, genres, or writing styles, but in concepts. To dig deeper into *which* concepts, we present the results of an SVM regression that predicts our custom controversy score from the text of individual book reviews from the Goodreads review data set.

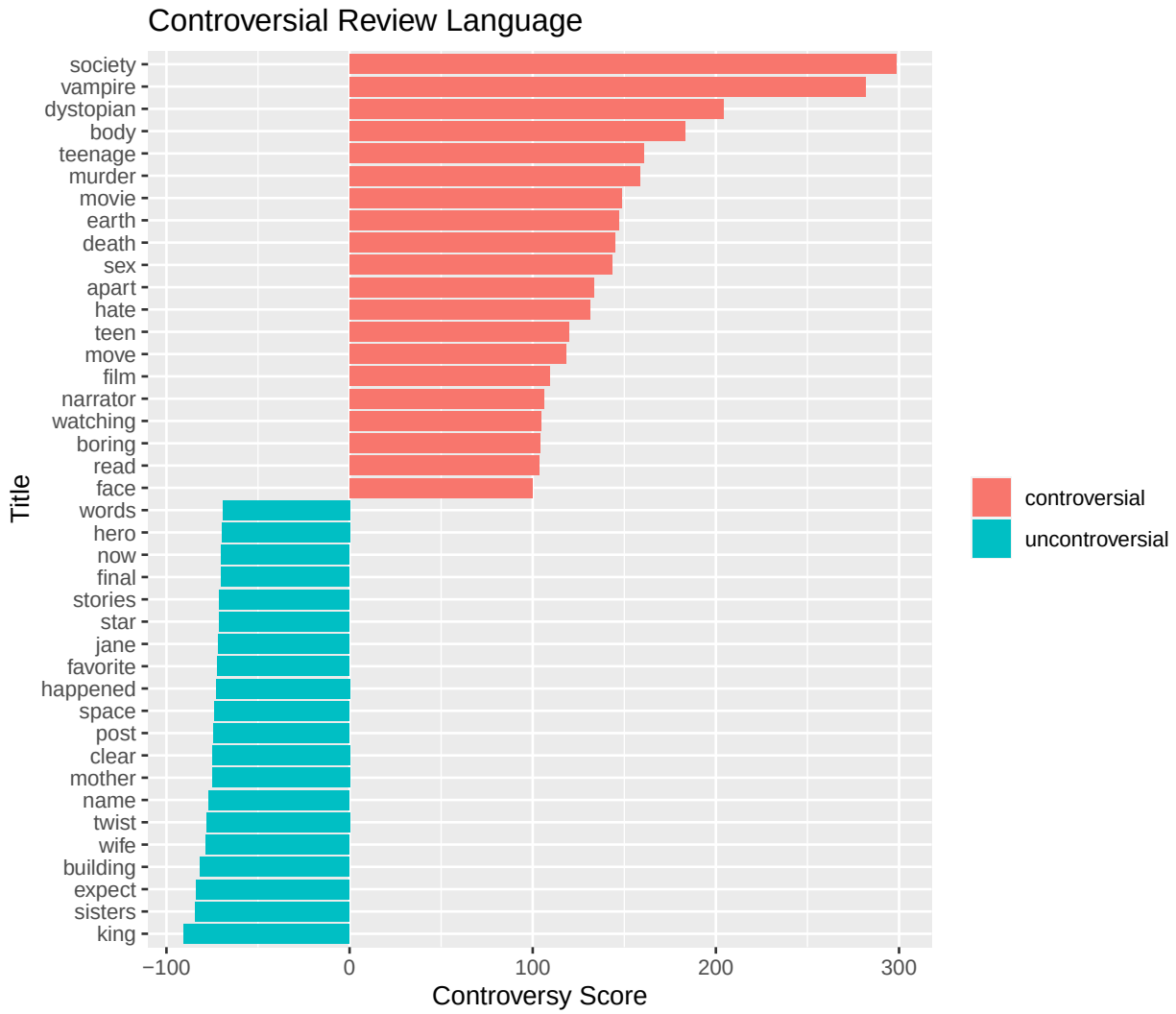

Figure 10: Publisher Tendency Towards Reading Lines vs. Book Popularity

While vampire *does* make its rather predictable appearance, the majority of the list is dominated by classic hallmarks of banned books and controversial texts: murder and death, sex and bodies, and dystopian society. More interesting, though, are the meta-narrative terms that do not so much characterize the content of the book, but the experience of reading the book. As the words "movie," "watching," and "film" imply, wherever movie adaptations go, controversy soon follows. Adaptation run the risk of dividing fans, and the fraught process of translating literary works for the silver screen often invites the most bitter of debates. Even more curiously, "narrator," and "read" suggest that reviewers often find themselves divided on the mechanics by which books make meaning. The uncontroversial narrator fades into the background, helping a reader to forget that they are, in fact, *reading*. By contrast, the controversial narrator stands out and draws comment, foregrounding a reader's own readership and thereby opening up the book's narrative structure to comments

both adoring and despising.

In summary, while a certain kind of content will always stir up controversy, readers also have strong and often opposing views on how books ought to go about the business of telling a story. For every division over teenage vampires, it seems there is another argument boiling just below the surface about adaptation, narrative voice, and very mechanics of literary meaning-making.

5 Killer Plot

Our killer plot is a novel visualization that provides insight into both the popularity and sentiment of a particular novel over time. At its most fundamental level, the killer plot shows a time series of the number of monthly checkouts of a particular book at the Seattle Public Library over a given range of months. At each month within this range, we display a specific word that is most distinctive among all words in the reviews of the novel from this month. We calculate the distinctiveness of this word using a weighted log odds metric. The size of the word is relative to its distinctiveness and the color corresponds to the rating of the review (i.e. green words correspond to 5-star reviews, yellow to 3 or 4-star reviews, and red to 1 or 2-star reviews).

Below, figures 11 and 12 are two examples of our killer plot for Suzanne Collins' *Catching Fire* and Paula Hawkins' *The Girl on the Train: A Novel*, respectively.

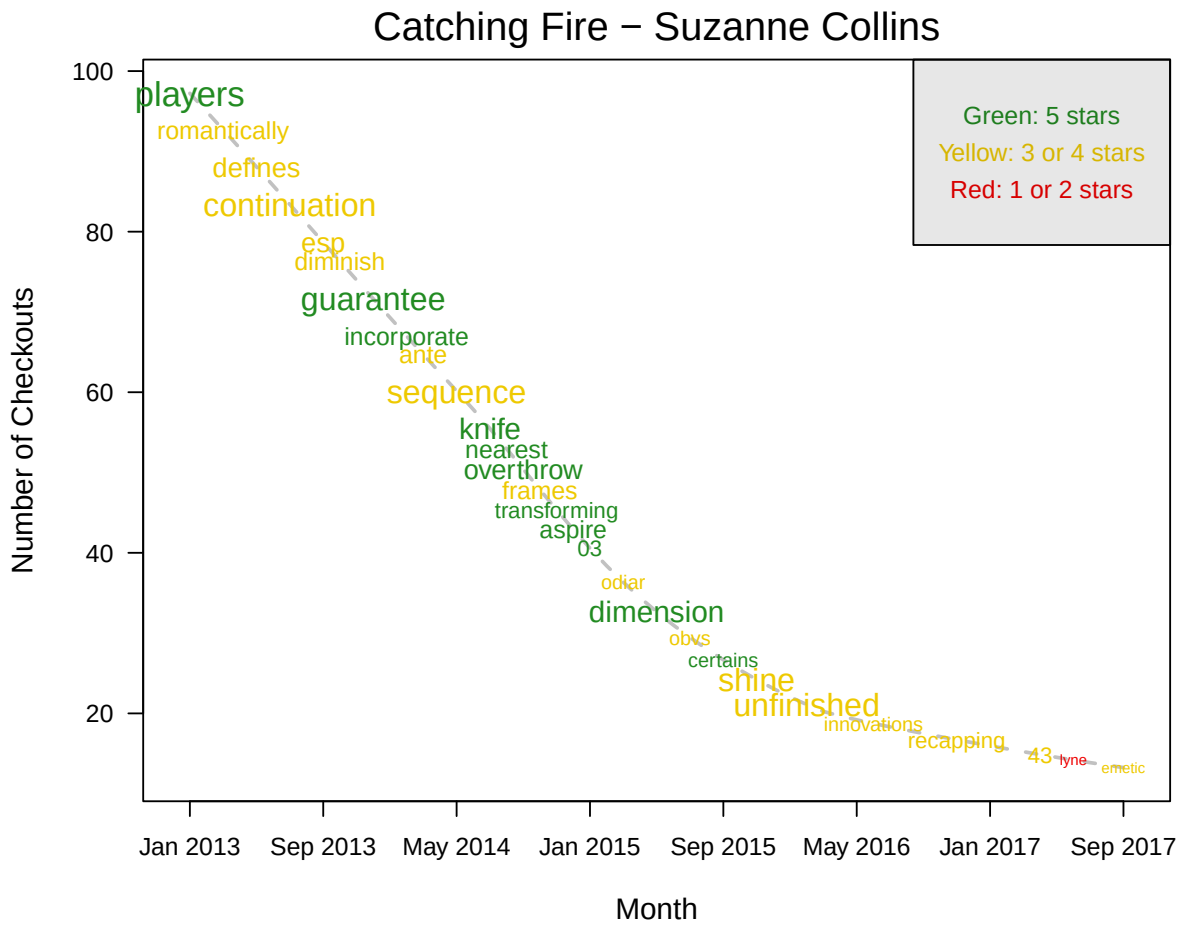

Figure 11: Killer Plot for Catching Fire by Suzanne Collins

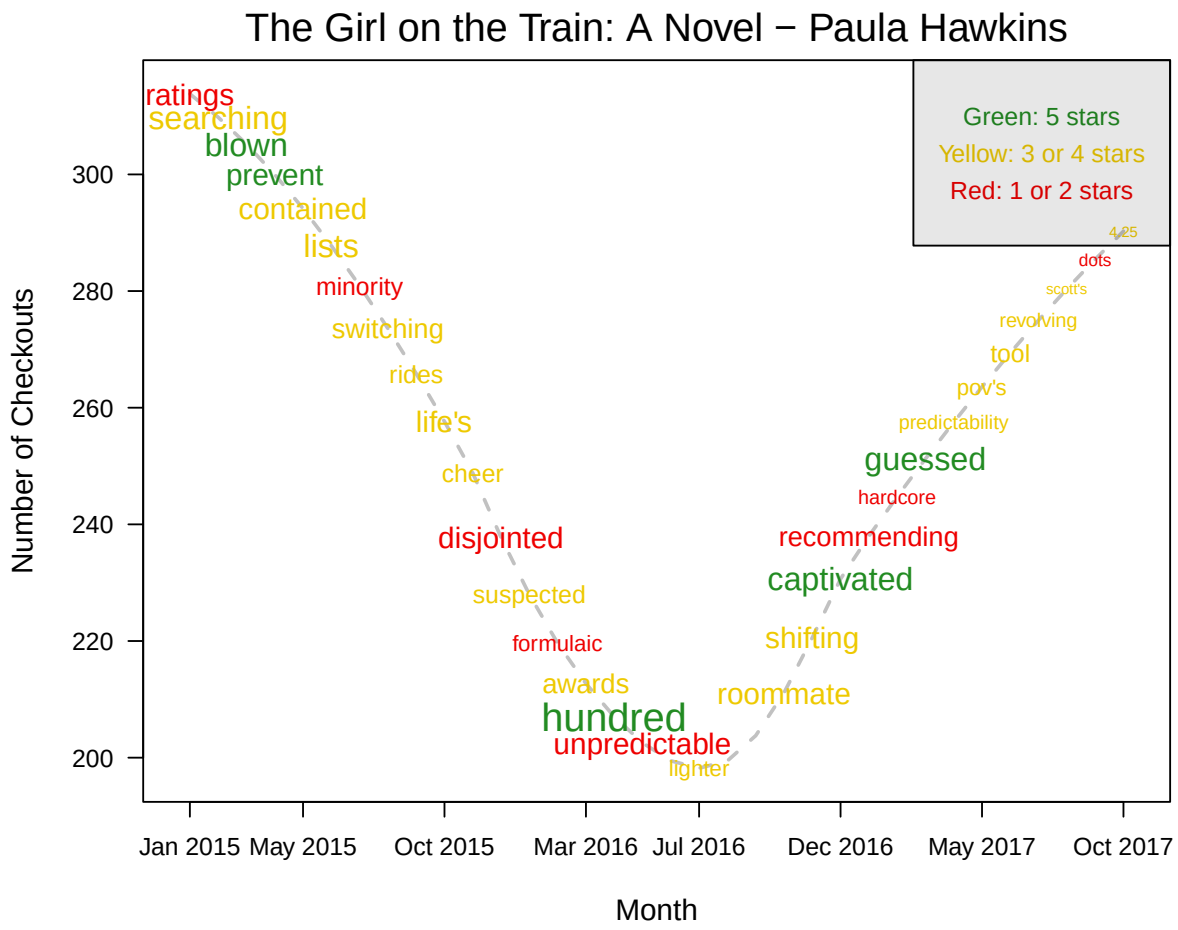

Figure 12: Killer Plot for The Girl on the Train by Paula Hawkins

With this representation, we hope to provide a unified representation of both the popularity of the novel in terms of checkouts from the library, but also the larger public sentiment towards the novel as is revealed through the review data. If, for example, a book initially achieves widespread popularity and then experiences another later resurgence, this plot allows us to visualize and interpret the sentiment surrounding its initial reception as well as its later resurgence. This can be particularly useful in understanding potential reasons behind why a particular novel may be experiencing a revival in the literary marketplace.

6 Conclusion

6.1 Results Discussion

Our work has allowed us to answer some of the guiding questions that we posed in the introduction.

Do popular books explode onto the scene or accumulate readers more gradually?

Our results suggest that while popular books initially explode onto the scene,subsequent surges in popularity are more gradual. Figure 12, our killer plot for The Girl on the Train, is a great example. After its initial decline in popularity, it resurged in popularity for around a year before declining again.

How do the popular success and critical acclaim of novels differ?

Our results suggest that popular success can vary based on the intensity of opinions. Figure 12, our killer plot for The Girl on the Train, is a great example. After its initial decline, reviews during its resurgence in popularity are largely neutral to negative. In the killer plot for Suzanne Collins' *Hunger Games: Mockingjay* that was not shown in the paper, reviews during a resurgence in popularity were overwhelmingly positive.

How has recent conglomeration of publishers changed the literary marketplace?

We find that for physical books, the publishing industry has actually seem to have undergone greater diversification in recent years. This has led to the availability of books from a wide variety of publishers. The opposite is true for e-books and audiobooks. Our results suggest that large conglomerate publishers have grown to control the digital publishing industry, which is unsurprising given the rapid rise of digital formats following pandemic and shutdown of the physical checkout in the library.

6.2 Limitations

Our investigation has several limitations that should be taken into account when interpreting the results. First, our primary dataset only includes book checkout information from the Seattle Public Library, which may not be representative of the reading habits and preferences of the wider American population.

Second, our supplementary review dataset contains reviews that were voluntarily submitted. This constitutes a small subset of the reading population and also is not necessarily representative of broader opinion. Furthermore, the quality and relevance of reviews may vary widely, which definitely impacts the interpretability and accuracy of our analyses.

Finally, similar to the limitations of our supplementary dataset, the book/item-specific information such as the subjects were manually entered by library staff. Especially considering the large size of the dataset, it is possible that there were errors or inconsistencies in the data that we did not catch during the data cleaning stage. This could also potentially impact our conclusions.

6.3 Future Improvements

To address the limitations mentioned above, we can expand our primary dataset to include additional libraries to get a more comprehensive look at reading habits across the country. We can also conduct more rigorous data cleaning and validatino procedures to minimize the potential impact of response bias in the reviews or inconsistencies in the book reporting.

To refine our analysis, we can consider investigating e-books and audiobooks more carefully. Our results already suggest that both digital formats are increasing in popularity, so it could be useful to incorporate analysis on those mediums in order to validate the conclusions we've already made.

6.4 In Closing

In an effort to draw these apparently disparate results together, we present the following conclusion. For although recent conglomeration has reduced the audiobook and ebook markets to a handful of major publishers, the independent book publishing market continues to deepen and diversify. The independent publishers – especially of literary fiction – maintain a rare skill at identifying and marketing transatlantic commercial

successes, which tend more often than not to bear the publishers' trademark reading line: ": a Novel." Lastly, any author looking to draft a novel worthy of the literary elite's reading tag would be well-advised to step lightly around sex, murder, vampires, or especially explicit narrative structures. And while we make no guarantees that following this advice will make anyone the next Sally Rooney, we do hope that any reader leaves this article with a deeper appreciate of the intricacies and idiosyncrasies of the American literary publishing industry.

7 Appendix

7.1 Fuzzy Title Matching

To make best use of our UCSD Book Graph Goodreads data, we need to match unique book IDs from Goodreads reviews to book checkouts from the Seattle Public Library. The Seattle data, however, is messy. Author information is often encoded in the title field, which is itself often full of misspellings and other typographical oddities. Dataset alignment, then, is a challenging problem in its own. As a first attempt, we employ a fuzzy string matching algorithm that computes pairwise similarity between the Goodreads and Seattle book titles and reports a match when that similarity crosses a given threshold. Our current similarity measure is Jaro-Winkler, which we selected primarily because it weighs matches at the beginning of a string more heavily than matches towards the end. Given the erroneous additions frequently appended to the end of Seattle library book titles, Jaro-Winkler ought to perform well. After parameter optimization, we report that approximately 60 percent of Seattle titles have been matched to Goodreads book IDs.

7.2 Popularity Curve Clustering

Beyond string matching, we also report some initial research into time series dimensionality reduction and clustering. In considering book popularity over time, it is natural to ask if there are certain popularity patterns that recur across books, publishers, and genres. After consulting the literature on time series clustering, we present a pipeline that we believed could identify clusters of books that exhibit similar popularity patterns. This pipeline computes monthly popularity time series curves and transforms them through z-normalization, discrete cosine transform, dynamic time warping, and k-medoids clustering.

Z-normalization makes the shape of popularity curves comparable across books with dramatically different raw checkout numbers. Discrete cosine transform reduces the dimensionality by decomposing and recomposing the time series using a smaller number of cosine waves. Dynamic time warping defines a similarity score for popularity curves with potentially non-overlapping time domains. And finally, k-medoids clustering modififes the more familiar k-medoids clustering by replacing average-computed centroids with median-computed centroids. K-medoids is a less common choice than k-means, so the selection is worth commenting upon. Compared with k-means, k-medoids is less sensitive to outliers and replaces the creation of artificial average-based centroids with median-based ones This sidesteps the need to average different popularity curves together during centroid creation – a questionable practice to be sure.

Disappointingly, if perhaps not quite unfortunately, the checkout timeseries did not meaningfully cluster even after optimization. We were able to identify groups will good intra- and inter-cluster metrics, but they were not interpretively interesting.