Mergers and Mismatches in the Labor Market for Creativity

Ke Shi*

Caltech

May 5, 2025

Abstract

This paper introduces a novel empirical framework to assess the impact of ownership consolidation on labor markets, addressing growing concerns about labor market power. I develop a two-sided matching model tailored to the creative labor force, a segment characterized by strong worker-firm compatibility. Applying this model to a major merger in the US publishing industry, I leverage rich text data to analyze its effects on the author labor market. Counterfactual merger simulations reveal a trade-off between efficiency gains, creative misalignment, and redistributive effects. Although the merger alleviated capacity constraints, post-merger integration resulted in significant creative misalignment between authors and publishers. The merger also triggered substantial value transfers from competing publishers and authors to the merged entity, with established authors bearing the heaviest losses. Notably, the merger's anticompetitive effects manifested primarily in labor markets rather than consumer markets. This research extends merger evaluation beyond consumer impact, providing a framework for analyzing the broader consequences of mergers on labor markets characterized by worker-firm complementarities.

*California Institute of Technology. Email: keshi@caltech.edu.

1 Introduction

Recent years have witnessed intensified government scrutiny of labor market power (Naidu, Posner, and Weyl 2018; Shapiro 2019; Posner 2021; Azar and Marinescu 2024). This heightened oversight culminated in the U.S. Department of Justice (DOJ) and Federal Trade Commission's (FTC) joint release of the *2023 Merger Guidelines*, which explicitly addresses how diminished labor market competition can depress wages, degrade working conditions, and reduce workplace quality.[1] While labor markets are subject to the same antitrust principles as product markets, they possess distinct features that both intensify and complicate competition concerns (Naidu, Posner, and Weyl 2018). A key distinction is their two-sided nature, where successful employment requires matching between workers and firms based on factors extending beyond wages.[2] This matching dynamic is particularly crucial in high-skilled and creative industries, where both parties value non-monetary aspects of the relationship and where compatibility significantly influences productivity. This raises critical questions: How does market consolidation affect worker-firm matching and compatibility? What are the broader implications for worker welfare beyond compensation? And how do these labor market dynamics ultimately impact the quality of goods and services delivered to consumers?

In this paper, I address these questions by developing a new empirical framework to analyze the impact of consolidation on labor markets. Given the distinctive nature of labor markets, the key contribution of this paper is to quantify the trade-offs and redistributive effects of mergers with a two-sided matching model. This conceptual framework recognizes that employment transcends simple transactions: it is a complex human relationship in which both workers and firms are driven by factors beyond monetary incentives. Employment represents a *joint production* of *value* (or *surplus*) that is shared between the two parties. This value creation crucially depends on the *complementarity* (or *compatibility*) be-

[1]For other policy initiatives, see, for example, the 2021 Executive Order on Promoting Competition in the American Economy, which recognizes that "consolidation has increased the power of corporate employers, making it harder for workers to bargain for higher wages and better work conditions" (*Executive Order No. 14036, 3 C.F.R. 36987*, 2021). For antitrust cases in labor markets, see, for example, *U.S. v. Bertelsmann SE & Co. KGaA, 646 F. Supp. 3d 1 (D.D.C. 2022)* and *Federal Trade Commission v. Kroger Company, 3:24-Cv-00347 (D. Or.)*

[2]The *2023 Merger Guidelines* notes that "finding a job requires the worker and the employer to agree to the match. Even within a given salary and skill range, employers often have specific demands for the experience, skills, availability, and other attributes that they desire in their employees. At the same time, workers may seek not only a paycheck, but also work that they value in a workplace that matches their own preferences, as different workers may value the same aspects of a job differently."

tween the two sides. The benefits each side receives reflect their total welfare gain from the partnership, a metric that captures more than just wages or profits alone. The equilibrium framework also allows me to disentangle two key effects of a merger: the direct effect on compatibility and value creation, and the equilibrium effect on market-wide matching patterns. As I will show, while mergers may improve efficiency through reduced capacity constraints, they generate significant redistribution and equity concerns via equilibrium effects. Value flows from other firms to the merged company, while workers suffer from diminished employer competition.

To implement this framework empirically, I focus on the US trade publishing industry and examine the 2013 merger between Penguin and Random House, two major publishing companies at the time. Publishing is an attractive empirical setting for several reasons. First, this labor market is clearly defined by the specific task of writing books, relatively segmented from other labor markets, and provides a well-identified universe of workers (authors) and employers (publishers). Second, the industry generates rich, observable data at the individual book level, information typically unavailable in other industries, that enables detailed analysis of author-publisher sorting patterns. Each author's labor product is well defined, and their performance is quantifiable through reader reviews and ratings. Third, book production hinges on strong intellectual and creative compatibility between authors and publishers, making the publishing industry particularly suitable for studying labor matching, where such compatibility concerns are most consequential compared to other industries. Finally, the industry's high concentration, dominated by only a few major publishing companies, makes it particularly relevant for studying labor market power, as evidenced by the recent successful blocking of a merger attempt between Penguin Random House and Simon & Schuster on monopsony grounds.

I begin the empirical analysis by establishing stylized facts that demonstrate assortative matching in the publishing market and reveal the effect of the merger on the equilibrium matching between authors and publishers. First, using constructed measures of compatibility in experiences and tastes between authors and publishers, I find strong evidence that they tend to display similar characteristics. Popular and high-quality authors—measured by their publication track records—are more likely to partner with publishers of comparable caliber. This matching extends to genre and content preferences, and authors and publishers demonstrate clear alignment in their literary styles and subject matter expertise.

My primary focus is on the redistributive effects of the merger. Since comprehensive data are available only for books published before the merger and equilibrium effects must be accounted for, I adopt a structural approach to recover author-publisher match values and simulate merger outcomes through a counterfactual analysis. The empirical model is a two-sided, many-to-one matching framework with transferable utilities, based on the canonical work of Shapley and Shubik (1971), Kelso and Crawford (1982), and Sotomayor (1999). The model captures the surplus or value generated by an author-publisher match, encompassing all the utilities created by the partnership. The market equilibrium is cleared through a transfer (typically from the publisher to the author), though this transfer mechanism itself is not explicitly modeled; instead, I focus on the division of post-transfer surplus allocated to each side of the market. Further, to assess the merger's downstream effects on readers, I incorporate book performance data, measured by reader reception, which sheds light on the product market's response and its implications for consumer welfare. This second component of the model operates similarly to a selection model, where only a subset of books is observed.

Estimation in matching models with transferable utilities and observed match performances presents three main challenges. First, characterizing the equilibrium is computationally intensive. To mitigate this, I adopt the partial equilibrium characterization from Fox (2018), which significantly speeds up the computation. Second, from an econometric point of view, the performance variables contain additional information on the match values and must be factored into the match value, making it infeasible to directly apply the semiparametric approach in Fox (2018). To bridge this gap, I adopt a parametric approach to connect the two parts of the model, which allows for a likelihood-based estimation procedure. However, the high dimensionality of the likelihood makes direct inference infeasible. To overcome this, I implement a Bayesian approach, extending the method of Sørensen (2007) from non-transferable to transferable utility models.

The structural estimation reveals several key findings about the publishing industry. First, editorial compatibility measures—including genre and content similarity between authors and publishers—significantly influence match value, as do past collaboration histories. These results suggest strong relationship stickiness in the industry: once a successful match forms, it tends to generate more value and lead to subsequent collaborations. The model demonstrates strong predictive power, correctly forecasting 67% of author-publisher matches compared to 15% under random assignment. Regarding book performance, I find that an author's pre-existing success (measured by ratings and review counts) is the strongest predictor of future book performance. Interestingly, while editorial compatibility measures strongly affect initial matching decisions, they have a limited direct impact on book performance after accounting for selection effects. This finding highlights the importance of properly accounting for the endogenous matching process when studying market outcomes.

Using the estimated parameters and recovered match values, I conduct several counterfactual simulations to analyze the merger's impact on the market. The merger represents a complete integration of two companies, requiring new match values for the consolidated entity to replace those of its previously separate components. I consider three counterfactual scenarios regarding the merged firm's value creation capabilities: (1) synergistic collaboration, where post-merger values reflect the stronger of the two pre-merger values; (2) organic merger, maintaining the weighted average of pre-merger values; and (3) Random House takeover, where the acquiring firm's values dominate. These scenarios represent a spectrum of post-merger outcomes and reveal an important trade-off between two opposing forces. Under synergistic collaboration, there is an efficiency improvement that stems from the merged entity's enhanced capacity to optimize author-publisher matches, a capability previously constrained when the companies operated independently. On the other hand, under the latter two scenarios, there is a substantial value loss that arises from post-merger integration and creative misalignment between affected authors and publishers. Intuitively, whereas before, each publisher could maintain specialized focus and resources to target different author segments, the merged entity's integrated operations reduce its compatibility with a broad spectrum of authors.

In addition, the efficiency changes from the merger are distributed highly unequally among market participants. My analysis reveals two concerning distributional effects. First, there is a substantial transfer of value from competing publishers to Penguin Random House, suggesting consolidation of market power. Second, publishers' profit gains come at the expense of authors' welfare, validating antitrust concerns about the harm of consolidation in this market. The authors' welfare losses stem from two distinct mechanisms. The direct effect occurs through reduced competition between the formerly separate companies, which puts downward pressure on author compensation, particularly affecting those previously contracted with either Penguin or Random House. The indirect effect operates through equilibrium sorting, creating a redistribution of value among authors: while those selected by the merged entity may benefit, authors displaced to other publishers experience welfare losses. These findings support the argument that market concentration in publishing can simultaneously enhance efficiency and exacerbate inequality.

Next, we examine how the merger's impact varies across authors at different career stages. The industry debate centered on which author segments would bear the greatest burden: some argued that *debut* authors and *mid-list* authors (those with moderate but not bestselling success) would suffer the most, as the merged entity would prioritize commercial blockbusters. Others, including the DOJ in the 2022 merger case, maintained that bestselling authors would face the most severe impact. My analysis reveals that the magnitude and direction of effects depend critically on authors' movement patterns post-merger. Among authors remaining with Penguin Random House, all experience welfare losses, with bestselling authors taking the largest hit, supporting the DOJ's position. However, the effects differ for authors changing publishers. When moving to Penguin Random House, debut and mid-list authors realize larger gains compared to their bestselling counterparts. In contrast, when authors leave Penguin Random House, bestseller authors experience the steepest welfare losses. These findings suggest that market power affects different author segments through distinct mechanisms, with implications for both industry practices and antitrust policy.

Finally, my analysis reveals that the consumer side of the market remained largely unaffected by the merger along observable quality dimensions. The books published by the merged entity did not show significant changes in rating volume or average ratings, suggesting that reader engagement and perceived quality remained stable despite market consolidation. This finding aligns with industry expectations that the merger's primary effects would manifest outside the reader experience. Although a complete assessment of consumer impact would require pricing data, the stability of quality metrics suggests that readers did not experience obvious degradation in their book consumption experience. These results highlight why merger evaluations must look beyond traditional consumer-side metrics: such a narrow lens may miss substantial anticompetitive effects in other dimensions, particularly in labor markets.

1.1 Related Literature

This paper contributes to several strands of literature. First, it contributes to the literature on the impact of mergers (Asker and Nocke 2021). The previous literature on mergers has focused mostly on the effect of mergers on product markets and consumer welfare. This paper is among the first to investigate the impact on labor markets, a growing field with important policy implications. In contrast to existing studies in this field, e.g., Arnold (2019), Prager and Schmitt (2021), Rubens (2023), Montag (2023), and Arnold et al. (2023), a key innovation of this paper is the characterization of labor markets as two-sided markets with preferences and compatibility between firms and workers. Second, existing studies generally focus on the effect of post-merger repositioning on product choice and firm conduct, e.g., Fan (2013), Li et al. (2022), and Wollmann (2018), whereas this paper considers the equilibrium impact of re-sorting that stems from the matching between firms and workers. Third, my paper speaks to the literature on merger's effect on innovation, but from the perspective of upstream labor input that has downstream spillover effects. Past work such as Igami and Uetake (2020) and Bonaimé and Wang (2024) has focused on firm choices themselves.

An emerging literature, in parallel to increasing policy concerns, examines monopsony power in labor markets (Naidu, Posner, and Weyl 2018; Marinescu and Hovenkamp 2019; Marinescu and Posner 2020; Berger, Herkenhoff, and Mongey 2022; Berger et al. 2023). This literature has devoted significant attention to explaining and estimating wage markdowns. Theoretical work has developed along three approaches: classic oligopsony, job differentiation, and search (Azar and Marinescu 2024).[3] While this paper aligns with the second strand by considering non-wage job characteristics, it offers a significantly more general framework by conceptualizing employment as the joint production of value. It is among the first studies to structurally model and analyze the direct impact of market consolidation events in labor markets at the micro-level.

This paper further contributes to the literature on creativity and its associated labor force, with a focus on the publishing industry (Canoy, van Ours, and van der Ploeg 2006). Past work has focused on the impact of intellectual property protection, such as copyrights and patents, on creative and innovative work (Biasi and Moser 2021; Giorcelli and Moser 2020; Peukert and Reimers 2022) or the effect of digitization in the publishing industry (Reimers and Waldfogel 2021; Peukert and Reimers 2022; Nagaraj and Reimers 2023), but little has been said about the impact of market structure and the changes thereof. This paper fills in this gap by offering a new empirical framework that conceives the production of creative output from the matching between the author and the publisher that exemplifies production in many high-skilled labor settings.

In terms of empirical methodology, this paper contributes to empirical studies of match-

[3]Research on monopsony power in labor markets dates back to Boal and Ransom (1997) and Manning (2003). See recent surveys by Ashenfelter, Farber, and Ransom (2010), Manning (2011), and Manning (2021). Recent empirical investigations include Azar et al. (2020), Treuren (2022), Yeh, Macaluso, and Hershbein (2022), Rubens (2023), Delabastita and Rubens (2023), Azar, Berry, and Marinescu (2022), and Fisher (2024), among others.

ing markets with transferable utilities, with a new emphasis on its implications on market structure and competition. I draw on the theoretical foundation in the seminal works of Shapley and Shubik (1971), Becker (1973), Kelso and Crawford (1982), Roth (1984), and Sotomayor (1999) with recent progress by Azevedo and Hatfield (2018), among others.[4] Existing empirical applications generally focus on the sorting patterns between the two sides, whereas this paper considers its implication on merger analysis and the additional layer of distortion it introduces (Dupuy et al. 2017).[5] In terms of empirical framework, a main difference is the full-fledged agent-level matching model of transferable utility with observed performance.[6] Past work often aggregates individuals by characteristics and estimates a two-sided random utility model (Choo and Siow 2006) because observable characteristics tend to be coarse. The observed performances are akin to a selection model and introduce additional complexity into the model. To overcome computational strain, I extend the Bayesian computation technique in Sørensen (2007) to a transferable context and adopt the semiparametric characterization in Fox (2010) and Fox (2018).[7] Further, I recover the post-transfer division of surplus based on the equilibrium characterization to analyze the welfare impact on both sides. Past works generally focus on inference on the joint surplus only.

[4]See survey by Chade, Eeckhout, and Smith (2017).

[5]See, for example, Yang, Shi, and Goldfarb (2009), Mindruta (2013), Mindruta, Moeen, and Agarwal (2016), Akkus, Cookson, and Hortaçsu (2016), and Chen et al. (2021), among others. Two closely related papers in labor matching are Boyd et al. (2013) and Agarwal (2015)

[6]See surveys and empirical methods by Chiappori and Salanié (2016), Graham (2011), Agarwal and Budish (2021), and Galichon and Salanié (2023).

[7]There are two strands of labor literature that are closely related to the matching literature. First, drawing on the theory of matching is a body of work that emphasizes sorting patterns in the labor market, e.g., Eeckhout and Kircher (2011), Eeckhout (2018), and Eeckhout and Kircher (2018). This paper is closely related in the sense that it emphasizes the compatibility between the publisher and the author. A second strand is the literature on hedonic wage and workplace amenities based on the theory of compensating differentials—a competitive equilibrium framework—in Rosen (1986), Hwang, Mortensen, and Reed (1998), Manning (2003), and Card et al. (2018). Chiappori, McCann, and Nesheim (2010) identifies the equivalence between hedonic models and stable matching. Recent empirical applications in this framework include Taber and Vejlin (2020) and Lamadon, Mogstad, and Setzler (2022), among others, and emphasize the wage effect.

2 The Publishing Industry and Data Description

2.1 Trade publishing

Trade publishing refers to books intended for general readership and sold through bookstores, retail outlets, and online sellers (Thompson 2012).[8] The US trade publishing industry is concentrated. Before the 2013 merger, there were six major publishing companies (the "Big Six"): Penguin, Random House, Simon & Schuster, Hachette, HarperCollins, and Macmillan. Penguin and Random House announced their merger in October 2012, and completed the process in July 2013. The merger further consolidated the market into the "Big Five." Penguin Random House (PRH) became and continues to be the world's largest publisher. Together, the Big Five held nearly 60 percent of the market for the sale of trade books in 2021, and 91 percent of the market for publishing rights to "anticipated top sellers."[9] While the growing concentration of the industry has long been justified on the basis of economies of scale in terms of cost savings and bargaining power with respect to downstream distributors, there have been competitive concerns about its impact on authors. When Penguin Random House proposed to acquire Simon & Schuster in 2022, it was challenged and enjoined on the ground that the merger would compromise competition in the market for publishing rights, i.e., the labor market of the authors.

Unlike other input markets, the labor market stands out due to the presence of match-specific preferences on both sides, beyond just profit, wages, and non-pecuniary benefits. Both parties may value factors unique to their relationship.[10] This is particularly evident in the publishing industry, where the editorial match between authors and publishers (or editors) is a key priority for both parties. After acquiring a manuscript and before production-related services such as design, printing, and marketing, the authors collaborate closely with the editors in a creative process to shape the final product. Publishers are concerned with whether the author's work aligns with their mission and literary vision, while authors seek editors who truly understand their work. Although author compensation was the primary concern in the merger case, it was emphasized repeatedly that

[8] As opposed to professional (e.g., tax manuals), educational (e.g., assessment materials), or academic publishing.

[9] Figures and quotes in this section, unless otherwise noted, are from court records in *U.S. v. Bertelsmann SE & Co. KGaA, 646 F. Supp. 3d 1 (D.D.C. 2022)*. "Anticipated top sellers" are books that meet the $250,000 advance threshold, a key definition in the case.

[10] For example, both publishers and authors may derive match-specific utility based on their shared interests, beliefs, or values, etc.

authors value "editorial match, a feel the editor and [publishing] house understand[s] what they are writing." They want to work with editors who "share their vision for the book" and who can help them to "bring the book into the world" and "create an audience for it."[11]

2.2 Data and variables

The main data for this study come from Goodreads, a community-based online platform for book rating, review, and social networking. The dataset was collected by Wan and McAuley (2018) and Wan et al. (2019) in late 2017.[12] The authors scraped users' "public shelves," a virtual list of books organized by themes accessible to anyone without registration. The complete dataset consists of nearly 2.3 million books. Each book is associated with its author(s), publisher, and publication date. In addition, I observe the book's rating and reviews, user-generated shelf labels, and a description.

For the study, I focus on a subsample of titles published between 2010 and 2016 (the last year of complete data) with complete information of authorship, publisher, and publication year and month.[13] Reprints or new editions of existing titles are discarded because they do not involve a new matching process between the author and the publisher. For convenience, each individual book (rather than the author) is treated as a unit of observation. In what follows, I use the terms author and book interchangeably to refer to the author side of the market. The sample consists of more than 140,000 books. Table 1 presents summary statistics of the books in the data.

On the publisher side, I consider the Big Six (Penguin, Random House, Simon & Schuster, Hachette, HarperCollins, and Macmillan), a group of notable publishing houses collected under "fringe publishers," and self-publishing.[14] Fringe publishers include some

[11]See footnote 9.

[12]Available at `https://mengtingwan.github.io/data/goodreads.html`.

[13]Although book data are available through 2017, the nature of a review platform means that books published earlier would have received more ratings and reviews by the time of data collection. For this reason, books published close to the data collection date have noisier information. Second, although the merger was completed in 2013, its effect could take years to realize. Although some preliminary evidence below would suggest that immediate changes did take place, a clear cut-off is unlikely and later dates are too close to the data collection date to allow meaningful inference. Therefore, for the main estimation, I will only use books published in 2010-2013 prior to the merger date and conduct a counterfactual merger simulation.

[14]Thompson (2012) notes that the publishing industry is characterized by a peculiar market structure: a handful of dominant publishing cooperates and numerous small independent houses. Medium-sized

key players such as Scholastic, Houghton Mifflin Harcourt, and Bloomsbury, among others. Finally, self-publishing is treated as an outside option in the analysis. Because publishers are big corporations and can have strengths and weaknesses in different areas of publication, the analysis is broken down using 10 genre categories to account for the internal heterogeneity of each publisher.[15]

The dataset contains only *observed matches* that are the equilibrium outcomes of a matching process, and a full analysis requires information on all *potential matches* (also referred to as *"pairs"* throughout the text) in the market. To this end, I construct an augmented dataset of all potential matches by taking the Cartesian product of the set of authors with books published in the half-year and the set of publishers. The matching market is defined at the semiannual level. That is, authors with a book published in the same half-year are considered one cohort up for matching with publishers.[16] This corresponds to the seasonality of the publishing industry which has a spring and a fall season.

Reader reception and book performance. For each book, I observe the number of ratings it has received (*rating count*) as well as the *average rating* across all versions of the book up to the date of data collection. I use the rating count to proxy the popularity of the book and the average rating to proxy its quality (Cabral 2012; Goldfarb and Tucker 2019).[17] Note that I do not take a normative stance on the value of a book and assume that popularity and quality reflect the utility of readers. Because the distribution of rating count is right-skewed with a long tail for bestsellers, I use the log transformation of rating count to dampen the long tail. The rating is an integer score from 1 to 5. Therefore, the average rating is in the range $[1, 5]$. The distribution of the average rating is left-skewed.

publishers are rare, so it is safe to ignore some of these.

[15] The 10 categories are (1) children, (2) comics & graphic, (3) fantasy & paranormal, (4) fiction, (5) history, historical fiction, & biography, (6) mystery, thriller, & crime, (7) non-fiction, (8) poetry, (9) romance, and (10) young adult. Categories (3), (6), and (9) are known as "genre fiction," popular styles that are often treated as distinct categories as opposed to generic literary fictions. Note that a book might belong to multiple categories. In the analysis, I consider the top two categories of each book whenever it belongs to multiple. An ideal dataset would observe the match at the author-editor level and then aggregate editors at their respective publishing houses. This information is unavailable, so I use the publisher-genre as a rough proxy for the editorial experience in the genre.

[16] I only observe the publication date, but not when the contracts are signed. The publishing industry abides by the lifetime of books. It is reasonable to assume that books that are published in the same year have been contracted around the same time.

[17] The literature on ratings and reviews shows that an effective rating and reputation system reflects the quality of goods and services and generally improves welfare by directing consumers to more desirable choices. For example, see Cabral and Hortaçsu (2010), Bolton, Katok, and Ockenfels (2004), Chen and Xie (2008), Chevalier and Mayzlin (2006), Dellarocas (2003), Deng et al. (2021), Sun (2012), and Wu et al. (2015), among others.

Table 1: Summary Statistics

Variable	N	Mean	SD	Min	Med	Max
Book characteristics						
E-book	136731	0.30	0.46	0	0	1
Part of a series	136731	0.25	0.44	0	0	1
Reader reception and book performance						
log(Ratings count)	136731	4.36	2.26	0.69	4.23	14.76
Ratings count percentile	136731	0.57	0.30	0.032	0.62	1.00
Average rating (Bayesian adjusted)	136731	3.93	0.34	1.41	3.92	5.00
Author characteristics						
Debut author	136731	0.36	0.48	0	0	1
Bestselling author	136731	0.047	0.21	0	0	1
log(Num prior books)	136731	1.16	1.19	0.00	0.69	5.40
Author ratings count percentile	136731	0.41	0.37	0.00	0.45	1.00
Author average rating	136731	2.48	1.87	0.00	3.69	5.00
Publisher characteristics (by genre, of previous half-year)						
log(Capacity)	136731	6.02	1.29	1.10	5.73	8.66
Revenue (in $B)	136731	0.51	0.94	0.00	0.00	3.84
Share of debut author	136731	0.37	0.21	0.00	0.38	1.00
Share of bestselling author	136731	0.049	0.057	0.00	0.027	0.40
Publisher ratings count percentile	136731	0.59	0.21	0.13	0.64	0.89
Publisher average rating	136731	3.89	0.13	3.24	3.86	4.40
Author-publisher characteristics						
Collaboration before	136731	0.24	0.40	0.00	0.00	1.00
log(Num past collaborations)	136731	0.42	0.73	0.00	0.00	4.87
Book-publisher characteristics						
Genre similarity	136731	0.45	0.31	0.00	0.46	1.00
Content similarity	136731	0.43	0.27	0.00	0.47	0.96

Notes: Author characteristics are aggregated over all previous books. For books with multiple authors, author characteristics are average across all authors. Publisher characteristics are aggregated for the same half-year in the previous year.

The occurrences of 1's and 5's are relatively rare and arise mostly for books with few ratings. Because books with few ratings are noisy, I use the Bayesian average to adjust it by the population average rating count and average rating.[18] The adjusted average rating is slightly above 3.9.

Pre-match experience, expertise, and interaction. For each book, pre-match characteristics are constructed from the author's track record of popularity and quality prior to the publication of the current book. The *debut author* and *bestselling author* are two variables that measure the author's experience. A debut author is one who publishes his or her very first book, which takes up about 40% of the books in the data. A bestselling author, on the other hand, is one who has published extensively and been widely recognized. In the data, I take the top 5% of the authors by their cumulative average number of ratings.[19] The remaining authors are called *mid-list authors* (a publishing jargon, from the publisher's "list" of books under management), whose books sell reasonably well, but not at a blockbuster or bestseller level. The *author rating count percentile* and *author average rating* are proxies of popularity and quality, constructed based on the cumulative rating count and the average rating of all previous books.[20] The average rating is the cumulative average of the average ratings of all previous books. For the rating count, because books published in earlier dates tend to have accumulated more ratings, to make books comparable across cohorts, I use the percentile of the books' rating count among books published in the same half-year. The variable is the cumulative average of the rating count percentile.

On the publisher side, the *share of debut authors* and *share of bestselling authors* are the corresponding measures of publishers' risk preferences, priority of commercial successes versus literary exploration, and overall abilities to attract authors in either category.[21] Similarly, the *publisher rating count percentile* and the *publisher average rating* measure the publisher's publication record. All publisher variables are aggregated and averaged for the publishers at the publisher-half-yearly-genre level. The distinction is that, whereas the

[18]The Bayesian (adjusted) average rating (BAR) of a book is

$$BAR_i = \frac{AR_i \times RC_i + \overline{AR}_{pop} \times \overline{RC}_{pop}}{RC_i + \overline{RC}_{pop}}, \tag{1}$$

where AR is the average rating, RC is the rating count, and the overline indicates the population average. Population is defined at the half-year-genre level.

[19]This is based on the observation that "the top 4 percent of profitable titles generate 60 percent of profitability".

[20]Pre-match variables are constructed from all books published after 2000, ten years before the sample period.

[21]These variables are the equilibrium outcomes. I assume a "price-taking" behavior from the authors.

authors' entire publication history is taken into account, for publishers, the aggregation is only for a half-year period and subdivided into genres. The reason is that, while authors are evaluated based on their track records, the publishers' recent publications are more likely to impact the matching.

The variables *collaboration before* and *number of collaborations* are constructed from the historical interactions between the author and the publisher. Because authors are likely to stay with publishers they have already known, these variables account for the status quo and the dynamic dependence on previous relationships.

Editorial compatibility. A key feature of the publishing industry is sorting on editorial match between the two sides. I construct two variables from the text data associated with the books, *genre similarity* and *content similarity*, to measure the editorial compatibility between authors and publishers. First, the *genre* of a book is generated from the corpus of shelf labels, community-generated text for the genre, style, topic, and other categorical features of the book. Second, the *content* of a book is taken from the corpus of book description (introduction) that contains information related to the content and story of the book. For example, the word clouds in Figure 1 show the labels and description of the 2010 bestseller *The Immortal Life of Henrietta Lacks* by American science writer Rebecca Skloot. Note that words have been preprocessed and only word stems are shown. Panel (a) shows that the book is of the genres and themes "biography," "nonfiction," "science," and "ethic," and panel (b) shows that the book tells a story around "cell," "immortal," "clone," and "research." Appendix B includes additional examples of bookshelf labels and descriptions of books from the sample period.

Given the text data, I use latent Dirichlet allocation (LDA), a common technique in topic modeling, for dimension reduction.[22] I use a subsample of 6000 books to train each model and assume $K = 50$ topics for both corpora. The model estimates a distribution of vocabulary (word frequency) for each topic. Detailed results of the LDA models can be found in Appendix B. Figure 2 shows some example topics generated by the LDA model. Panel (a) shows an example genre topic that has high probabilities over terms such

[22] In broad strokes, topic modeling assumes that each text (document) is generated by some K common "topics." Each topic is represented by a distribution over the vocabulary present in the entire corpus, which can be loosely interpreted as the word frequency under the topic. In turn, each text is characterized by a K-dimensional distribution over topics. Topic modeling reduces the dimension from the dimension of the vocabulary to K topics. See Gentzkow, Kelly, and Taddy (2019) and Ash and Hansen (2023) for details of topic modeling. See Hansen, McMahon, and Prat (2018), Bandiera et al. (2020), Djourelova, Durante, and Martin (2024), and Ash, Morelli, and Vannoni (2022) for some recent application.

Figure 1: Example text: *The Immortal Life of Henrietta Lacks*, Rebecca Skloot, Pan Macmillan, 2010

(a) Genre (shelf labels)

(b) Content (book description)

Figure 2: Example topics

(a) Example of a genre topic

(b) Example of a content topic

as "religion," "philosophy," "science," and "psychology." Panel (b) shows an example content topic that has high probabilities over terms such as "parent," "disease," "help," and "medicine." Additional example topics are also found in Appendix B.

After estimating the topic-word distributions, I can recover the posterior of each book's probabilistic distributions over the shared set of topics, yielding two vectors each of length 50 that summarize the genre and content of the book. I do so for publishers' publication history at the publisher-genre-half-yearly level as well by aggregating the books of the publisher as one text and recovering two vectors of genre topic and content. I measure the editorial compatibility using cosine similarity, a common measure of document distance.[23] A magnitude of 1 means that the two completely overlap, and 0 means that the two have no similarities. This results in two measures of document distances: genre similarity and content similarity between every pair of book and publisher.

3 Descriptive Evidence

3.1 Assortative matching

Matching markets are characterized by positive assortative matching (Becker 1973). I first verify sorting on observable characteristics between the publisher and the author in terms of their experience, popularity, and quality. Table 2 presents regressions of an author's characteristic on the characteristics of the publisher with which she is matched.

$$X_{ij}^{a} = \beta_0 + X_{ij}^{p\prime}\beta_1 + \varepsilon_{ij}, \tag{2}$$

where the unit of observation ij is a matched pair, X_{ij}^{a} is the characteristics of the author, and X_{ij}^{p} is the characteristics of the publisher. In other words, the regression says conditional on being a match, given the publisher's characteristics, what are the author's characteristics likely to be.

There is a significant degree of positive assortative matching between measures of an author's experience and that of a publisher's expertise. The diagonal entries in the regressions are similar characteristics from both sides and demonstrate that authors and publishers match based on characteristics along the same dimensions. Authors of greater popularity (rating count percentile) and quality (average rating) tend to match with publishers of similar strength. I also find that publishers' risk preferences, measured by the past share of debut and bestselling authors they work with, are positively correlated with these features on the author side. The capacity of the publishers and their revenue, on the other hand, are not strong predictors of the authors' characteristics.

Second, there is also assortative matching along editorial compatibility measured by

[23] Given two n-dimensional vectors of topic distributions, x and y, their cosine similarity is the dot product normalized by the product of their magnitudes: $\frac{x \cdot y}{\|x\|\|y\|}$. See Kelly et al. (2021), Cagé, Hervé, and Viaud (2020), and Bertrand et al. (2021) for discussions of document distance.

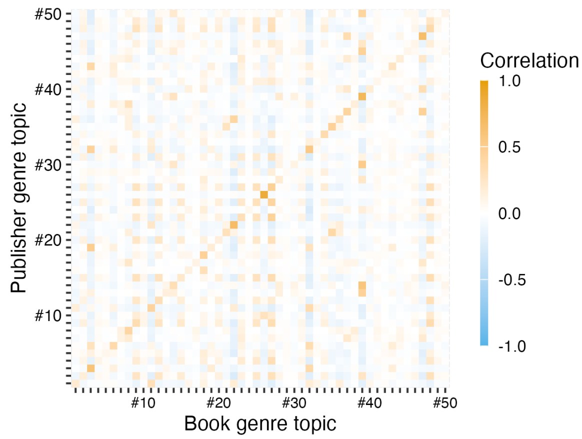
Table 2: Assortative matching

	log(Author ratings count percentile)	log(Author average rating)	Debut author	Bestselling author	log(Num prior books)
	(1)	(2)	(3)	(4)	(5)
Publisher ratings count percentile	0.479***	0.370***	−0.116***	0.093***	−0.323***
	(0.018)	(0.085)	(0.022)	(0.014)	(0.060)
Publisher average rating	0.085***	0.701***	−0.051*	0.065***	0.773***
	(0.020)	(0.095)	(0.025)	(0.015)	(0.067)
Share of debut authors	−0.615***	−3.011***	0.786***	−0.011	−3.092***
	(0.012)	(0.057)	(0.015)	(0.009)	(0.040)
Share of bestselling authors	0.447***	0.610***	−0.143***	0.830***	0.654***
	(0.029)	(0.138)	(0.036)	(0.022)	(0.097)
log(Capacity)	−0.015***	−0.085***	0.020***	−0.003*	−0.038***
	(0.002)	(0.010)	(0.003)	(0.002)	(0.007)
Revenue	0.003	0.005	−0.002	−0.005	0.000
	(0.004)	(0.020)	(0.005)	(0.003)	(0.014)
Constant	0.068	1.043**	0.258*	−0.272***	−0.696*
	(0.081)	(0.392)	(0.102)	(0.063)	(0.276)
Year fixed effects	Yes	Yes	Yes	Yes	Yes
Publisher fixed effects	Yes	Yes	Yes	Yes	Yes
R^2	0.164	0.111	0.115	0.060	0.244
Observations	87111	87111	87111	87111	87111

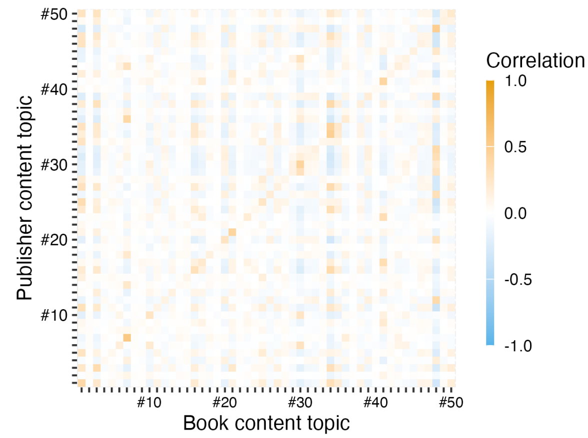
Notes: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

both genre and content. Recall that the genre and content of books and publishers are summarized in vectors of distribution over 50 topics. Figure 3 plots the correlation matrices between the book's topic weights and those of their matched publishers. Panel (a) is the correlation matrix of the genre topic weights and panel (b) is that of content topic weights, where the color gradation indicates the magnitude of the correlation from -1 to 1. The diagonal entries are the corresponding topics for the author and the publisher. If books were randomly assigned to publishers with no respect for editorial compatibility, we would expect to see no correlation pattern. However, several patterns emerge. First, I find positive correlation along the diagonal entries. That is, if a book has large weights on certain topics, then it is likely that their matched publisher shares larger weights over the same topics. Second, certain book topics display negative correlations with other topics

Figure 3: Topic correlation between the book and the publisher



(a) Genre



(b) Content

Notes: Correlation matrix of the topic distribution of books and publishers. The horizontal axis is the 50 topics of the book and the vertical axis is the corresponding topics of the publisher. Yellow represents positive correlation and blue represents negative correlation. Details of the topics are found in Appendix B.

on the publisher side, suggesting that authors and publishers of certain genres or content do not work across categories. Third, the two sides display a stronger correlation in terms of genre compared to in terms of content. This is not surprising because genre topics are more clearly defined compared to content topics.

4 Structural Model and Estimation

4.1 Two-Sided Matching Model

The structural model is based on a two-sided many-to-one matching framework with transferable utility (Kelso and Crawford 1982). Consider a market consisting of two disjoint sets of firms $i \in I$ and workers $j \in J$. Firms can hire multiple workers, while each worker can only be employed by one firm. Let q_i be the hiring capacity of firm i. Workers may remain unmatched (or "matched" to an outside option with index 0). Let $\tilde{I} = I \cup \{0\}$ denote the augmented set of firms. Following the convention in the matching literature, I assume a full-information, zero-friction environment where all authors and publishers enter the market as potential matches.[24] A *matching* $\boldsymbol{\mu} \in \{0,1\}^{|\tilde{I}\times J|}$ is a binary vector, where $\mu_{ij} = 1$ indicates that firm i is matched with worker j and 0 otherwise.[25] Note that $\mu_{0j} = 1$ means that worker j is unmatched to any firm. Lastly, the model assumes that a firm's outside option of leaving positions unfilled carries an arbitrarily small utility, and the number of workers far exceeds the number of firms, ensuring that all firms' capacity constraints are binding in equilibrium.

Firm i's profit from employing a set of workers $C_i \subseteq J$ and offering a vector of *transfers* (wages) t_{ij} is $\pi_i(C_i; t_{ij}) = f(C_i) - \sum_{j \in C_i} t_{ij}$, where $f(C)$ is the production function. Assuming the production function is linearly separable in workers, the match-specific profit from a pair ij is

$$\pi_{ij} = f_{ij} - t_{ij}, \tag{3}$$

where f_{ij} represents the output produced by firm i in collaboration with worker j.[26] Crucially, f_{ij} encompasses all value produced in the firm-worker pairing from the firm's perspective, which includes factors beyond just the immediate revenue or profit from production. For example, a firm might place value on qualities such as the worker's alignment with its values, reputation, or long-term strategic goals, even if these factors do not directly impact short-term financial outcomes.

Worker j's utility from working for firm i with a transfer t_{ij} is $u_j(i; t_{ij})$. Following convention, I assume that u_j is linearly separable in two components

$$u_{ij} = a_{ij} + t_{ij}, \tag{4}$$

where a_{ij} is the match-specific utility that j derives from working with firm i, which reflects how much the worker personally values the firm, such as their preferences for the firm's culture, reputation, or work environment. The transfer t_{ij} encompasses more than

[24]This assumption is realistic in the context of the publishing industry, which is relatively small and interconnected. Literary agencies, in particular, play an important role in facilitating matches between authors and publishers by providing information and reducing search frictions. However, for simplicity, the model abstracts from the role of these intermediaries.

[25]For simplicity, with a slight abuse of notation, I will use the shorthand $\mu_{ij} = 1$ to denote the set of matched pairs $\{ij \in \tilde{I} \times J | \mu_{ij} = 1\}$ (and $\mu_{ij} = 0$ for unmatched pairs) in the indices of summation, product, maximum, and minimum.

[26]A large part of the empirical matching literature assumes this functional form where $f(C)$ can be linearly decomposed into f_ij, which rules out complementarities and externalities in production. This is a reasonable assumption in the publishing industry where the relationship between authors and publishers tends to be independent of others. This assumption is also important from a theoretical standpoint, as it helps guarantee stable matching without further restrictions.

just the wage; it includes anything negotiated as part of the contract, such as non-monetary benefits. In the publishing context, this could include factors like the level of attention and support the author expects from the publisher. Finally, let $u_{0j} = a_{0j}$ denote the value of the outside option, which depends solely on the worker's type j.

Let

$$v_{ij} = a_{ij} + f_{ij} = u_{ij} + \pi_{ij} \tag{5}$$

denote the *joint surplus* (or *value*) of the pair ij, which does not depend on the transfer t_{ij}. By definition, for unmatched workers, $v_{0j} = u_{0j}$. Let $\boldsymbol{v} = (v_{ij})_{ij}$ denote the vector of joint surpluses for all potential matches. In the empirical literature of matching with transferable utilities, the focus is primarily on this joint surplus, v_{ij}. Intuitively, the first equality in (5) indicates that v_{ij} is a joint production function that captures the total value produced by the match between i and j in a reduced form.[27] The second equality pertains to the distribution of the joint surplus, where u_{ij} and π_{ij} represent the *net welfare* (or *surplus*) that the worker and the firm each receive from the match. This post-transfer split of surplus will be the primary focus of this study.

Equilibrium. The standard solution concept is *pairwise stability*. A matching $\boldsymbol{\mu}$ is pairwise stable if for any unmatched pair $\mu_{ij'} = 0$, we have $v_{ij'} < u_{i'j'} + \pi_{ij}$. In other words, no unmatched pair has an incentive to deviate from their current matches to form a new one. Given the setup, the stable matching condition can be reformulated as the following linear programming (LP) problem (Gretsky, Ostroy, and Zame 1992; Galichon and Salanié 2023):

$$\begin{aligned} &\max_{\boldsymbol{\mu}} \ \boldsymbol{v}'\boldsymbol{\mu} \\ &\text{s.t.} \ \sum_j \mu_{ij} = q_i \text{ for all } i \\ &\qquad \sum_i \mu_{ij} = 1 \text{ for all } j \\ &\qquad \mu_{ij} \in \{0, 1\}. \end{aligned} \tag{6}$$

[27] Although the exposition so far has assumed that value production is separable into two preference components, a_{ij} and f_{ij}, as is commonly assumed in the literature (Kelso and Crawford 1982), empirically, the distinction between "preference" and "transfer" is not always clear. Furthermore, these components are not empirically identified unless the transfer is explicitly defined (e.g., the wage) and observed, or strong assumptions are made about preferences. However, for the purposes of this study, such distinctions are unnecessary because only post-transfer utilities are relevant.

The solution to this LP always exists and is generically unique. Furthermore, the LP formulation suggests that a matching is stable if and only if it maximizes total social welfare (Sotomayor 1999; Azevedo and Hatfield 2018). Intuitively, the transfer serves as a price signal that adjusts to clear the market in a competitive equilibrium.[28]

An inversion problem for estimation. From an empirical point of view, we face the inverse optimization problem: given an observed equilibrium matching $\boldsymbol{\mu}$, recover the underlying values $\boldsymbol{v}$ that generate such a matching. Formally, we want to compute a set of values $V_{\boldsymbol{\mu}}$ that can rationalize the observed matching, i.e.,

$$V_{\boldsymbol{\mu}} = \{\boldsymbol{v} \in \mathbb{R}^{|\tilde{I} \times J|} | \boldsymbol{v}'\boldsymbol{\mu} > \boldsymbol{v}'\tilde{\boldsymbol{\mu}} \text{ for all feasible } \tilde{\boldsymbol{\mu}} \neq \boldsymbol{\mu}\},$$

where a *feasible* matching $\tilde{\boldsymbol{\mu}}$ is one that satisfies the constraints in the LP problem (6).[29] The problem requires solving for a vector of bounds on v_{ij} that are mutually consistent: For a matched pair ij, the value v_{ij} must exceed some lower bound $\underline{v}_{ij}$ to maintain a match. Conversely, for an unmatched pair ij', the value $v_{ij'}$ must remain below some upper bound $\overline{v}_{i'j}$ to ensure that it remains unmatched.

In the estimation, I compute these bounds by partially characterizing the equilibrium using a *two-pair-no-exchange* condition in Fox (2010) and Fox (2018). This condition rules out a single deviation from equilibrium where two matched pairs, ij and $i'j'$, mutually abandon their current partners to form two new pairs, ij' and $i'j$, i.e.,

$$v_{ij} + v_{i'j'} > v_{ij'} + v_{i'j} \tag{7}$$

for all $\mu_{ij} = 1$, $\mu_{i'j'} = 1$, and $i \neq i'$. For a matched pair ij, this implies that $v_{ij} > v_{ij'} + v_{i'j} - v_{i'j'}$ for all other matched pairs $i'j'$. Taking the maximum of the right-hand side over all other

[28] Kelso and Crawford (1982) show that the stable matching can be reached from a salary adjustment algorithm that is a generalized version of deferred acceptance algorithm. This algorithm is in spirit similar to an ascending price auction in which firms take turns to bid for workers, competing in an upward salary adjustment process.

[29] Mathematically, This is the dual cone (or polar cone, depending on the convention) of the set of feasible matchings $\tilde{\boldsymbol{\mu}}$ at $\boldsymbol{\mu}$.

matched pairs where $\mu_{i'j'} = 1$ gives the highest lower bound of v_{ij}:[30]

$$\underline{v}_{ij} = \max_{\substack{\mu_{i'j'}=1 \\ i' \neq i}} v_{ij'} + v_{i'j} - v_{i'j'}. \tag{8}$$

Conversely, for $i'j$ that is not a match, we have $v_{i'j} < v_{ij} + v_{i'j'} - v_{ij'}$ where i is the firm that j is actually matched with. This condition holds for all workers j' that are matched to firm i'. Taking the minimum of the right-hand side over all such pairs where $\mu_{i'j'} = 1$ yields a least upper bound of $v_{i'j}$:

$$\overline{v}_{i'j} = \min_{\mu_{i'j'}=1} v_{ij} + v_{i'j'} - v_{ij'}. \tag{9}$$

Division of surplus. While the pre-transfer preferences are not identified, the equilibrium characterization allows us to recover the post-transfer division of surplus—u_{ij} for the worker and π_{ij} for the firm for all matched pairs $\mu_{ij} = 1$. Although the equilibrium matching $\boldsymbol{\mu}$ is generically unique, the split of surplus is, however, not. In particular, Sotomayor (1999) shows that the set of post-transfer outcomes u and π form a lattice structure. Therefore, we first characterize this set and then determine a firm-optimal allocation. The equilibrium division of the surplus must justify $\boldsymbol{\mu}$ as a stable matching by satisfying the pairwise stability condition. For a firm i and a worker j' who are not currently matched, the value of their potential match cannot exceed the sum of their current utilities. In other words,

$$v_{ij'} < u_{i'j'} + \pi_{ij}, \tag{10}$$

so that there is no incentive to break off current matches and form a new match. Substituting $\pi_{ij} = v_{ij} - u_{ij}$ and rearranging terms, we obtain

$$u_{ij} - u_{i'j'} < v_{ij} - v_{ij'}. \tag{11}$$

Intuitively, this inequality states that in order to prevent ij' from forming a match, the utility of worker j (in ij) cannot exceed that of worker j' (in $i'j'$) more than some upper

[30] Note that these conditions are only necessary but not sufficient for the LP problem (6). In other words, the bounds $\underline{v}_{ij}$ and $\overline{v}_{i'j}$ are not tight. In principle, we also require a no-exchange condition for all cycles of matched pairs—a notion of core stability—e.g., $v_{ij} + v_{i'j'} + v_{i''j''} > v_{ij'} + v_{i'j''} + v_{i''j}$, to fully satisfy the LP problem. This characterization is computationally intractable and unnecessary for our purpose. Fox (2018) demonstrate that the score estimator based on the inequality in (7) is set-identified. In my implementation, Monte Carlo simulations confirm that the parameters are identified. See details in Section 4.3.

bound. Otherwise, j' could propose to i and achieve a mutually preferable deviation.

To further bound the utilities u_{ij}, workers in all matched pairs must receive a payoff higher than that of their outside option, i.e.,

$$u_{ij} > u_{0j} = v_{0j}. \tag{12}$$

On the firm side, because I assume that firms do not have outside options, π_{ij} is not constrained below by some reservation value, which implies that u_{ij} is not bounded from above. Conveniently, we do not need this upper bound condition. In many labor markets, firms often take turns offering wages to workers, who then decide whether to accept or decline.[31] Kelso and Crawford (1982) show that this ascending, firm-proposing salary adjustment mechanism results in the firm-optimal outcome in the set of stable allocations. Thus, the unique lower bounds of u_{ij} in the firm-optimal outcomes are characterized by the following LP:

$$\begin{aligned}
&\min_{\boldsymbol{u}} \sum_{\mu_{ij}=1} u_{ij} && (13)\\
&\text{s.t. } u_{ij} - u_{i'j'} < v_{ij} - v_{ij'} \\
&\qquad u_{ij} > v_{0j} \\
&\qquad \text{for all } \mu_{ij} = 1, \mu_{i'j'} = 1, \text{ and } i \neq i'.
\end{aligned}$$

4.2 Specification

Match value production. I parameterize the value v_{ij} linearly in the pair's observable characteristics

$$v_{ij} = X_{ij}'\beta + \varepsilon_{ij}, \tag{14}$$

where X_{ij} are firm-worker-specific characteristics, β is a vector of parameters to be estimated, and ε_{ij} is a random utility shock.[32] This is the value production function that depends on the complementarities between the two sides. Second, the reservation value

[31] In the publishing industry, for example, publishers frequently bid competitively for an author's manuscript. The court record in *U.S. v. Bertelsmann SE & Co. KGaA, 646 F. Supp. 3d 1 (D.D.C. 2022)* contains numerous examples of such competitive bidding among publishers.

[32] Under the matching with transferable utility framework, X_{ij} must vary across both i and j for identification of β. Observe that in the equilibrium characterization (6) or (7), firm-specific and worker-specific characteristics do not affect equilibrium matching.

of the worker v_{0j} is specified as

$$v_{0j} = X'_{0j}\beta^{RV} + \varepsilon_{0j}, \tag{15}$$

where X_{0j} is the characteristics of the worker. Because the explanatory characteristics are different from the main specification of the values in equation (14), I denote the parameters β^{RV} where RV stands for the reservation value. Note that ε_{0j} is drawn from the same distribution as ε_{ij}. As in discrete choice models of random utility, β is identified up to scale and level. Therefore, the constant term is absent. I fix the variance of the error term ε at 1 so that β is identified.

Match performance. In addition, we also have two additional performance equations to measure the success of the book for matched pairs $\mu_{ij} = 1$. First, for popularity, I log-transform the rating count to r_{ij} and let it depend on the set of characteristics W_{ij}. (Note that W_{ij} is potentially different from X_{ij}.)

$$r_{ij} = \log(RatingsCount_{ij}) = W'_{ij}\gamma^{r} + \eta_{ij}. \tag{16}$$

Similarly, I let s_{ij} denote the average rating and also let it depend on the book's pre-publication characteristics

$$s_{ij} = AverageRating_{ij} = W'_{ij}\gamma^{s} + \zeta_{ij}. \tag{17}$$

A key feature of the structural model is that the matching and performance are related through the correlation between ε_{ij} and (η_{ij}, ζ_{ij}). The two parts of the model complement each other in the following sense. On the one hand, incorporating the matching model in the performance outcomes is in a spirit similar to Heckman's correction (Heckman 1979). As alluded to earlier, the books that are published are not a random sample, but the results of the matching between authors and publishers described above. In the absence of the matching model, a direct estimation of equations (16) and (17) will produce biased estimates because the observed matched pairs are a selected sample of all the potential pairs. The matching framework is equivalent to the two-step control function approach to correct bias arising from nonrandomly selected samples.

On the other hand, book performance provides a channel to estimate sorting on unobservable characteristics. There could be unobservable match-specific characteristics that affect both the value v_{ij} and the performance r_{ij} and s_{ij}. To the extent that they enter the performance, the performance provides additional information on the values of matched pairs. This is similar to recovering unobservable heterogeneity from the observed outcomes common in many other settings. A direct estimation of the matching model based on the equilibrium characterization (7) (such as the semiparametric approach in Fox (2018)) loses information because the performance equation contains additional information through the correlated error terms. In the appendix, this is made explicit in equation (D.2) where the performance variables enter the distribution of the values.

Errors. To relate the two parts of the model, I take a parametric approach by specifying the distribution of the error terms. I assume that errors $(\varepsilon, \eta, \zeta)$ are independently and identically distributed across pairs ij and have a joint normal distribution with mean 0. For the covariance matrix, it is convenient to decompose the error terms into orthogonal components $(\varepsilon, \xi_1, \xi_2)$, all normally distributed with mean 0 and variances $1, \sigma_1^2, \sigma_2^2$, respectively. As in probit models, by fixing the variance of ε at 1, β is identified. I let (ε, η) have covariance δ and (ε, ζ) have covariance ω and decompose η and ζ respectively such that $\eta = \delta\varepsilon + \xi_1$ and $\zeta = \omega\varepsilon + \xi_2$. Note that this is still flexible and the only restriction is that the variance of ε is 1. Then the covariance matrix of $(\varepsilon, \eta, \zeta)$ is given by

$$\begin{pmatrix} 1 & \delta & \omega \\ \delta & \delta^2 + \sigma_1^2 & \delta\omega \\ \omega & \delta\omega & \omega^2 + \sigma_2^2 \end{pmatrix}. \tag{18}$$

4.3 Estimation

Let m index the matching markets that correspond to each half-year. Each market consists of two disjoint sets of publishers I_m and authors J_m. Following the notation in the structural model, at the individual market level, I omit the subscript m to simplify the notation. Within a given market, every pair of agents ij is characterized by the following variables: value-specific characteristics X_{ij}, latent match value v_{ij}, equilibrium matching μ_{ij}, performance-specific characteristics W_{ij}, and performance variables r_{ij} and s_{ij} for matched pairs. Let italic $X, \boldsymbol{v}, \boldsymbol{\mu}, W, r, s$ be the respective matrices or vectors that collect variables over all pairs ij in a given market. Let bold upright $\mathbf{X}, \mathbf{v}, \boldsymbol{\upmu}, \mathbf{W}, \mathbf{r}, \mathbf{s}$ collect these same variables in all matching markets in the dataset.

The parameters to estimate are the valuation parameters β, β^{RV} in (14) and (15), the performance parameters γ^r, γ^s in (16) and (17), and the covariance matrix of the error terms $(\delta, \omega, \sigma_1^2, \sigma_2^2)$ in (18). Let $\boldsymbol{\theta}$ collect all parameters.

A direct estimation is infeasible in this context. Observe that the likelihood function of the matching $\boldsymbol{\mu}$ in market m (ignoring the performance equations for now) is

$$\mathcal{L}_m(\beta|\boldsymbol{\mu}, X) = P(v \in V_{\boldsymbol{\mu}}|\beta, X) = P(\varepsilon \in V_{\boldsymbol{\mu}} - X\beta) = \int \mathbf{1}(\varepsilon \in V_{\boldsymbol{\mu}} - X\beta)dF(\varepsilon). \tag{19}$$

Recall that $V_{\boldsymbol{\mu}}$ is the set of values that rationalize $\boldsymbol{\mu}$ as the observed equilibrium matching. β can in principle be estimated by maximizing the total likelihood across all markets $\prod_m \mathcal{L}_m(\beta|\boldsymbol{\mu}, X)$. However, the likelihood function is difficult to evaluate given the dimension of the integral. A key feature of the matching models is rivalry, that the agents do no act in isolation and one firm's matching with a worker precludes another firm's matching therewith and vice versa. Therefore, the error terms within the same market must be simultaneously integrated out, but this is too computationally costly to be feasible.[33]

To bypass the explicit evaluation of the likelihood function, I use a Bayesian approach to estimate this matching model as in Sorensen (2005) and Sørensen (2007). Specifically, I adopt Markov Chain Monte Carlo (MCMC) simulations with Gibbs sampling, a data augmentation technique where the latent variables v_{ij} are treated as auxiliary parameters to be sampled alongside other parameters $\boldsymbol{\theta}$. The Markov Chain is constructed by iteratively sampling from the conditional distributions of parameters given the previous draws of other parameters.[34]

Prior distributions. To set up the estimation, I first specify the prior distributions of the parameters before deriving the conditional distributions (samplers). Given the model specification, I choose the following conjugate prior distributions $f_0(\boldsymbol{\theta})$ so that the conditional posteriors will be in the same family of parametric distributions. All parameter prior distributions are independent. The prior distributions f_0 of $\beta, \gamma^r, \gamma^s$ and δ, ω are normal distributions $N(\theta_0, \Sigma_{\theta,0})$. I use fairly uninformative priors with mean $\theta_0 = 0$ and covariance $\Sigma_{\theta,0} = I \times 10$, where I is the identity matrix of compatible dimensions. The prior distributions f_0 of σ_1^2, σ_2^2 are inverse gamma distributions with shape and scale parameters

[33]To see this more explicitly, notice that v_{ij} is not observed in the data so that I cannot directly establish the likelihood for every individual observation ij. The equilibrium characterization only yields information on the relationship between the error terms. In particular, the equilibrium characterization (7) implies that $-\varepsilon_{ij} - \varepsilon_{i'j'} + \varepsilon_{ij'} + \varepsilon_{i'j} < X'_{ij}\beta + X'_{i'j'}\beta - X'_{ij'}\beta - X'_{i'j}\beta$ so that we can treat the left-hand side $-\varepsilon_{ij} - \varepsilon_{i'j'} + \varepsilon_{ij'} + \varepsilon_{i'j}$ as a random variable. However, notice that it is no longer an independent sample. The matched terms ε_{ij} are sampled at a much higher rate compared to the unmatched terms $\varepsilon_{ij'}$.

[34]See Gelman et al. (2013) for an introduction to this class of methods. MCMC is a popular tool in discrete choice models and has been widely adopted in marketing research. See, for example, Rossi, Allenby, and McCulloch (2012).

α_0 and β_0 (not to be confused with the parameter β). I let $\alpha_0 = 1$ and $\beta_0 = 1$.

Posterior. Given the specification of the error distribution with mean 0 and covariance matrix (18), the likelihood function (or conditional density) of the latent variable v and performance variables r, s in market m is a normal distribution.

$$f_m(\boldsymbol{v}, r, s | X, W, \boldsymbol{\theta}) \propto \prod_{ij} \exp\left(-\frac{1}{2}\left(v_{ij} - X'_{ij}\beta\right)^2\right) \times \\ \prod_{\mu_{ij}=1} \exp\left(-\frac{1}{2}\left(\frac{r_{ij} - W'_{ij}\gamma^r - \delta(v_{ij} - X'_{ij}\beta)}{\sigma_1}\right)^2\right) \times \\ \prod_{\mu_{ij}=1} \exp\left(-\frac{1}{2}\left(\frac{s_{ij} - W'_{ij}\gamma^s - \omega(v_{ij} - X'_{ij}\beta)}{\sigma_2}\right)^2\right). \tag{20}$$

Note that the normalization factor in the density functions is omitted and only the kernel of the density function is given. Recall also that the index of summation $\mu_{ij} = 1$ is a shorthand for the set of observed matches.

The augmented posterior density f in all markets is proportional to the product of the prior distribution of parameters f_0, the conditional densities f_m in (20), as well as the boundary conditions that characterize stable matching.

$$f(\mathbf{v}, \mathbf{r}, \mathbf{s}, \boldsymbol{\theta} | \boldsymbol{\mu}, \mathbf{X}, \mathbf{W}) \propto f_0(\boldsymbol{\theta}) \times \prod_m \left[f_m(\boldsymbol{v}, r, s | X, W, \boldsymbol{\theta}) \times \prod_{\mu_{ij}=1} \mathbf{1}(v_{ij} \geq \underline{v}_{ij}) \times \prod_{\mu_{ij}=0} \mathbf{1}(v_{ij} < \overline{v}_{ij}) \right], \tag{21}$$

where $\underline{v}_{ij}$ and $\overline{v}_{ij}$ are defined in equations (8) and (9).

The conditional densities of v and $\boldsymbol{\theta}$ are proportional to the respective components in the augmented posterior (21). See Appendix D for details of the Gibbs samplers $f(v_{ij}|\cdot)$ and $f(\theta|\cdot)$.

4.4 Estimation Results

I estimate the structural model on the sub-dataset from 2010 to 2013, where each half-year is treated as a distinct matching market. Table 3 presents the parameter estimates from the structural model.

Value parameters. The parameters of the value equation (14) and the reservation value

Table 3: Estimates from structural model

(a) Value parameters

Parameter	Mean	Median	Marginal Effect	SE
β				
Ratings count percentile interaction	-5.999^{***}	-6.052	-1.692	(0.290)
Average rating interaction	2.435^{***}	2.318	0.687	(0.343)
Debut interaction	1.923^{***}	1.922	0.542	(0.144)
Bestselling interaction	5.525^{***}	5.525	1.558	(0.766)
Genre similarity	1.829^{***}	1.824	0.516	(0.067)
Content similarity	1.159^{***}	1.164	0.327	(0.083)
Collaboration before	2.030^{***}	2.029	0.424	(0.060)
log(Num prior collaborations)	0.881^{***}	0.882	0.249	(0.039)
β^{RV}				
Debut author	3.251^{***}	3.292	0.489	(0.192)
log(Num prior books)	-0.085^{*}	-0.084	-0.024	(0.042)
Author average rating	1.799^{***}	1.780	0.508	(0.104)
Author ratings count percentile	-5.139^{***}	-5.177	-1.450	(0.261)

Notes: .

(15) are presented in Table 3a. Because the parameters are identified up to scale and level, the magnitudes of the estimates are not immediately interpretable. But the signs are of expected sign and are statistically significant. In particular, I find that the genre similarity and content similarity, two measures of editorial compatibility, strongly influence the match value. Past collaboration history also heavily influences match value, suggesting strong stickiness in the industry that once a match is formed, it is likely to generate more value and result in subsequent collaborations.

As in logit and probit models, the coefficients are interpreted by calculating their marginal effects. I compute an analogous marginal effect with the following definition a la Sørensen (2007). If two pairs of authors and publishers, ij and $i'j'$, have identical attributes, $X_{ij} = X_{i'j'}$, then in equilibrium, the probability of one pair being a match but not the other is one half, assuming that capacity constraints are not interfering. The marginal effect of a characteristic is defined as the change in the probability of ij being a match but not $i'j'$ that results from a unit change in the characteristic X_{ij}.[35] For example, an increase of 0.01

[35] The probability of ij being a match but not $i'j'$ is $Pr(X_{ij}'\beta + \varepsilon_{ij} > X_{i'j'}'\beta + \varepsilon_{i'j'}) = \Phi((X_{ij}' - X_{i'j'}')\beta/\sqrt{2})$. This is one half when $X_{ij} = X_{i'j'}$. The marginal effect is the derivative evaluated with respect to X_{ij} at $X_{ij} = X_{i'j'}$.

Table 3: Estimates from structural model (cont.)

(b) Performance parameters

Parameter	Mean	Median	SE
γ^r			
Debut author	5.126***	5.123	(0.305)
Bestselling author	1.011***	1.010	(0.069)
log(Num prior books)	0.008	0.008	(0.026)
Author ratings count percentile	4.112***	4.109	(0.096)
Author average rating	0.736***	0.736	(0.078)
Capacity	0.155***	0.154	(0.023)
Revenue	0.011	0.011	(0.015)
Publisher ratings count percentile	4.883***	4.887	(0.171)
Publisher average rating	−0.139	-0.139	(0.201)
Genre similarity	1.170***	1.170	(0.068)
Content similarity	0.075	0.077	(0.077)
Collaboration before	−0.024	-0.025	(0.074)
log(Num prior collaborations)	0.046	0.046	(0.044)
γ^s			
Debut author	2.338***	2.338	(0.048)
Bestselling author	0.030**	0.030	(0.011)
log(Num prior books)	0.012**	0.012	(0.004)
Author ratings count percentile	0.103***	0.103	(0.015)
Author average rating	0.600***	0.600	(0.012)
Capacity	−0.003	-0.003	(0.003)
Revenue	−0.001	-0.001	(0.002)
Publisher ratings count percentile	−0.221***	-0.221	(0.026)
Publisher average rating	0.562***	0.562	(0.032)
Genre similarity	−0.049***	-0.049	(0.010)
Content similarity	0.030*	0.030	(0.012)
Collaboration before	−0.011	-0.011	(0.011)
log(Num prior collaborations)	0.018**	0.018	(0.007)
Year fixed-effect	Yes		

in genre similarity (a continuous variable in the range of $[0, 1]$) increases the probability of being a match by 0.5%.

Model fit. I next investigate the model fit by comparing the predicted matching against

For binary variables, this is $\Phi(\beta/\sqrt{2}) - 0.5$. For continuous variables, this is $\phi(0)\beta/\sqrt{2}$, where Φ and ϕ are the cdf and pdf of the standard normal distribution.

Table 3: Estimates from structural model (cont.)

(c) Covariance matrix

Parameter	Mean	Median	SE
δ	0.329***	0.328	(0.041)
ω	−0.002	-0.002	(0.006)
σ_1^2	2.076***	2.075	(0.034)
σ_2^2	0.052***	0.052	(0.001)

Notes: .

the observed matching. Because the matching framework involves two-sided choices, there is no readily available goodness-of-fit measure. However, from the perspective of the authors, who are only matched to a single publisher, the problem resembles a choice problem. Therefore, I calculate the prediction accuracy from the authors' perspective by examining if the model correctly predicts their matched publisher. I find that the prediction accuracy is about 67%. Compare this with the prediction accuracy of the random assignment at only about 15%.[36]

The strength of the matching framework is further substantiated by comparing this to other model specifications of match formation in Table C9 in the appendix. In these models, the unit of observation is a book-publisher pair, and the outcome is a binary variable indicating if the pair is a match. This is regressed on the same set of explanatory variables as in the structural model. The difference is that these alternative models treat each book-publisher pair as an independent observation, but the matching framework incorporates rivalry and the equilibrium dependence among observations. As expected, these models have less prediction accuracy, at about 52%-54%, compared to the matching framework. Without accounting for the matching, most estimates are overestimated.

Performance parameters. The estimates of the performance parameters are presented in Table 3b. The coefficients on pre-publication author rating count percentile and average rating are positive and statistically significant in both performance metrics. This suggests a temporal correlation among the author's works, and the author's ability is the most important factor in deciding the book's success. Interestingly, compatibility measures such as content similarity and past collaborations do not significantly affect the book's perfor-

[36]Randomly assign books to publishers subject to their capacity constraints.

mance after accounting for selection, compared to how they directly affect matching.

Like the value equation, I compare the estimates of the performance equation in the structural model against simpler specifications. Table C11 in the Appendix presents results of direct OLS regressions of the performance variables on the same set of regressors, without accounting for equilibrium matching. I find that these estimates are different from the structural estimates. For example, a one-percentile increase in a publisher's rating count raises the book's rating count by 4.9% under the matching estimation, but it is overstated at 5.2% under direct OLS. The difference between these estimates is the indirect effect of sorting on the performance of the books.

5 Merger Simulation

The primary interest of this paper is the impact of mergers on the labor market and worker welfare. As discussed in Section 2.1, the 2013 Penguin Random House merger significantly consolidated the market for authors. Given the available data, I perform a counterfactual analysis, assuming the merger took place in 2010 instead of 2013, treating Penguin and Random House as a single publisher in a counterfactual fashion. This method follows the simulation approaches used in Fan (2013), Wollmann (2018), and Li et al. (2022) to evaluate mergers, allowing a comparison of the same cohort of agents. In the following discussion, I use the term "post-merger" to refer to this simulated merger. The subscripts P, RH, and PRH denote the companies Penguin, Random House, and Penguin Random House, respectively.

To start, the post-merger primitives must be specified. If the merger simply involved the removal of one firm from the market, workers would be weakly worse off as there would be one fewer bidder on the buyer side (Crawford and Knoer 1981). However, a merger involves the combination of two firms into one, which has three key implications for the market in my empirical model: participants, capacity constraints, and match values. First, I assume that all other market participants remain unchanged, meaning that authors will not enter or exit the market as a result of the merger. Second, based on the observation that there were no significant changes in the number of books published post-merger, I assume there is no capacity adjustment. Consequently, the capacity constraint of the new company, q_{PRH}, will be the sum of the two firms' capacities, $q_P + q_{RH}$. Third, I assume that the match values of all other publishers remain the same but only those of the merged publisher, v_{PRH}, are affected. I will discuss this assumption more in Section 5.1.

To implement counterfactual experiments, I use the updated primitives $\boldsymbol{v}$ to simulate counterfactual equilibrium matching $\boldsymbol{\mu}$ for every matching market by applying the LP characterization in (6). After simulating the matches, I compute the equilibrium division of surplus $\boldsymbol{u}$ using the LP problem in (13). In addition, I calculate the realized book performance metrics, including the rating count and the average rating in (16) and (17).

I then compare the simulated post-merger counterfactual outcomes to a simulated version of the pre-merger outcomes. Although the equilibrium framework assumes all participants are involved in the counterfactual, I still expect that a significant portion of authors will remain with their original publisher, as their match values will likely dominate in both scenarios. Therefore, I focus on two distinct groups of authors: those who stay with Penguin or Random House and those who switch publishers due to changes in sorting. For both groups, I examine the impact on the total surplus and the division of that surplus between authors and publishers. To assess these effects, I analyze three key metrics: (1) the transfer of value from other publishers to Penguin Random House, (2) the shift of surplus from authors to publishers, and (3) the redistribution of surplus among authors with varying levels of tenure. This approach allows for a nuanced understanding of how the merger affects both the market dynamics and the welfare of different participants.

5.1 Counterfactual assumptions

The match values of the merged company, compared to those of its predecessors, present more nuanced empirical questions. For an author j, the predecessors' $v_{P,j}$ and $v_{RH,j}$ are now replaced by $v_{PRH,j}$. How $v_{PRH,j}$ changes in relation to $v_{P,j}$ and $v_{RH,j}$ depends on the post-merger repositioning of Penguin Random House. The literature has found substantial evidence that mergers affect the positioning of both acquiring and acquired firms. For example, Sweeting (2010) provides reduced-form evidence of product repositioning after mergers, while Fan (2013) endogenizes product characteristics to analyze mergers along this dimension. Furthermore, Eliason et al. (2020) shows that acquired firms tend to converge toward the behavior of their new parent companies. Because internal changes within Penguin Random House are not directly observable, I perform three merger simulations under different scenarios: (1) synergistic collaboration, (2) organic merger, and (3) Random House takeover.

First, in the synergistic collaboration scenario, I assume a best-case outcome where the merger value reflects the better of the two merging companies. This assumption captures the idea that Penguin and Random House could each contribute their respective strengths and expertise post-merger. Publishing is highly individualized on the publisher's side and relies heavily on the expertise of individual editors. Since the editors remained with Penguin Random House after the merger, as was the case, it is reasonable to expect that they would continue to apply their specialized knowledge and skills in the post-merger environment.

Second, under the organic merger scenario, I draw from insights in the repositioning literature and assume that Penguin Random House operates as a single entity, with its characteristics being a weighted average of its predecessors. This counterfactual simulates a scenario in which the two merging companies must reconcile their differences and move forward as one cohesive organization.[37] To implement this, I use the publishers' characteristics, X and W, which are computed per genre-period by averaging the characteristics of books published in the genre from the previous year. The counterfactual characteristics of a unified Penguin Random House for each period were computed by combining previously published books from both companies. Using this set of new characteristics, I then recompute the potential match values, $v_{PRH,j}$, for Penguin Random House.

Third, in the Random House takeover scenario, I assume that the post-merger entity reflects only the characteristics of Random House. Although the 2013 merger initially involved shared ownership between Bertelsmann (Random House's parent company) and Pearson (Penguin's parent company), Bertelsmann held a majority stake, while Pearson controlled the remainder. Over time, Pearson sold its shares to Bertelsmann, leaving Penguin Random House as a wholly owned subsidiary of Bertelsmann.[38] Given this trajectory, where Random House gradually gained full control, it is reasonable to assume that Random House's influence dominated decision-making and likely shaped Penguin's publishing strategies post-merger. Thus, this scenario models the merger as a step-by-step acquisition, with the newly merged company essentially operating under Random House's philosophy and editorial approach. Note that this assumption is the least generous among the three. Intuitively, authors who were previously paired with Random House are stripped of their first-best and must settle for some other publishers.

[37] Anecdotal evidence and account suggests that Penguin and Random House had vastly different corporate culture. Penguin, particularly under CEO John Makinson, was known for its innovation and independence and recognized for risk-taking in publishing more experimental and controversial works. Random House, on the other hand, had a reputation for its size and market strength. It was known for its focus on commercial publishing, often producing blockbuster titles with a broader appeal.

[38] In 2013, Bertelsmann owned 53% of the joint venture, and Pearson held 47%. In 2017, Pearson sold 22% of its shares to Bertelsmann, and in 2020, it sold the remaining shares, making Penguin Random House a wholly owned subsidiary of Bertelsmann.

Table 4: Changes to total social surplus

	Aggregate change		
	Joint surplus	Author share	Publisher share
	Panel A: Total social change		
Synergistic collaboration	6.44	-286.51	292.95
Organic merger	-22.67	-319.40	296.73
Penguin Takeover	-102.72	-365.61	262.88

5.2 Simulation results of synergistic collaboration

I now discuss the impact of the merger on the matching between authors and publishers. Table 4 presents the overall impact under the three counterfactual scenarios. Table 5 and Table 6 present the results of the first counterfactual under synergistic collaboration. All figures represent changes in value from pre- to post-merger states. It is important to note that because values are identified up to a monotone transformation, their absolute magnitudes cannot be directly interpreted; however, their relative magnitudes can be meaningfully compared. The table is structured with six columns: the first three show aggregated changes, while the latter three display average changes per author. In both cases, the table presents the total change to the joint surplus, as well as the author's and publisher's respective shares of this change.

Overall effect. Table 4 illustrates the change in total social surplus $\sum v_{ij}\mu_{ij}$ as well as the author and publisher shares after the merger under the three counterfactual assumptions. The results indicate a net increase in social surplus under synergistic collaboration but a net loss under the other two scenarios. The difference is a result of two opposing forces of the merger.

On the one hand, the net gain is expected in synergistic collaboration. The counterfactual assumes a combination of capacities and the better value between the two merging entities after the merger. Recall that equilibrium maximizes total social welfare. Therefore, the total social surplus must weakly increase post-merger. This mechanism is illustrated in an example in Figure 4. The example features three publishers (Penguin, Random House, and Publisher 3) and three authors (Austen, Byron, and Coleridge). For simplicity, assume that each publisher has a capacity of exactly 1, and all reservation values are negative, ensuring that all authors prefer to be matched. The table displays the matched values for

Figure 4: Example of redistribution

	Aus.	Byr.	Col.
Penguin	10	0	5
RH	0	3	0
Publisher 3	0	1	2

(a) Pre merger

	Aus.	Byr.	Col.
Penguin RH	10	3	5
Publisher 3	0	1	2

(b) Post merger

Notes: Rows represent publishers and columns represent authors. Each cell contains the match value for a specific publisher-author pair. All outside option values are negative, ensuring every author prefers being matched. Blue-colored cells indicate the equilibrium matches.

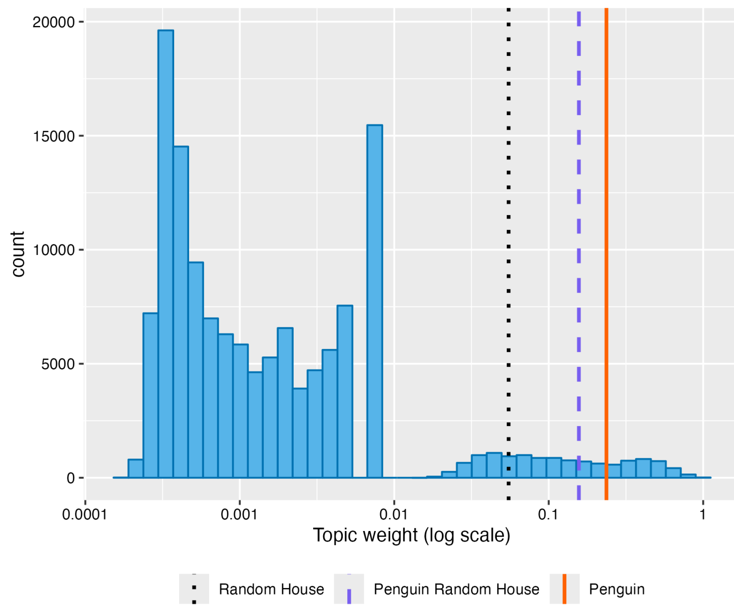
each author-publisher pair. The pre-merger equilibrium outcome is readily apparent. In the post-merger scenario, we assume Penguin Random House's match values are the better of the two merging companies for each author. Notably, the merger allows Penguin Random House to match with Coleridge, a higher-value pairing that was not realized in the pre-merger scenario due to capacity constraints.

On the other hand, the net loss under organic merger and Random House takeover represents the effect of mismatch arising from market homogenization and consequently deteriorating match values. The two merging publishers have different editorial philosophies and serve authors of their respective strengths. Under either organic merger or Random House takeover, the merging entity must integrate and settle at a unique post-merger position, resulting in a loss of compatibility between some authors and publishers.

To illustrate this, Figure 5 shows an example of the histogram of authors along the genre of literary fiction from the language model. Recall that the genre is measured by a weight between 0 and 1. The heavy distribution on the left consists of books that are not related to this topic, whereas the fraction on the right consists of books that are related. The solid and dotted vertical lines indicate the positions of Penguin and Random House, respectively, over this particular characteristic. By construction, the dashed line represents that of Penguin Random House under organic merger, where the post-merger position is the average of its predecessors. The distance between the author and the publisher is negatively related to their compatibility. Under either organic merger or Random House takeover, more authors moved away from the realized post-merger entity, leading to a loss in compatibility.

Differentiated impact among publishers. Column (1) of Panel B reveals a redistribution of value from other publishers to Penguin Random House. While other publishers suffer a loss in joint surplus, Penguin Random House experiences an increase in value. To

Figure 5: Distribution of authors along some example genre (literary fiction)



Notes: The genre is topic #37 in the genre topic model in Appendix B. It corresponds roughly to the genre of literary fiction.

decompose this change, Panel C shows that the internal changes within Penguin Random House from combining the two companies were relatively small at 1.3. This is because authors who remained with Penguin Random House did not see a significant rise in their match value post-merger. Panel D, on the other hand, demonstrates that the redistribution among publishers was primarily driven by sorting, with Penguin Random House gaining welfare at the expense of other publishers.

Decrease in author welfare.

About a quarter of the market is owned by Penguin Random House. Approximately 8% of authors transitioned to a different publisher following the merger. Figure 6 illustrates the migration patterns of writers who changed their publisher post-merger. Notably, the majority of these movements involve an "exchange" of authors between Penguin Random House and other publishing houses. This pattern is expected, given the assumption that the match values of all other publishers remain unchanged post-merger, preserving their relative orders.

a notable redistribution of value is observed across all three scenarios: while overall social surplus increases, there is a shift from authors to publishers. Specifically, authors

Table 5: Simulation results of synergistic collaboration

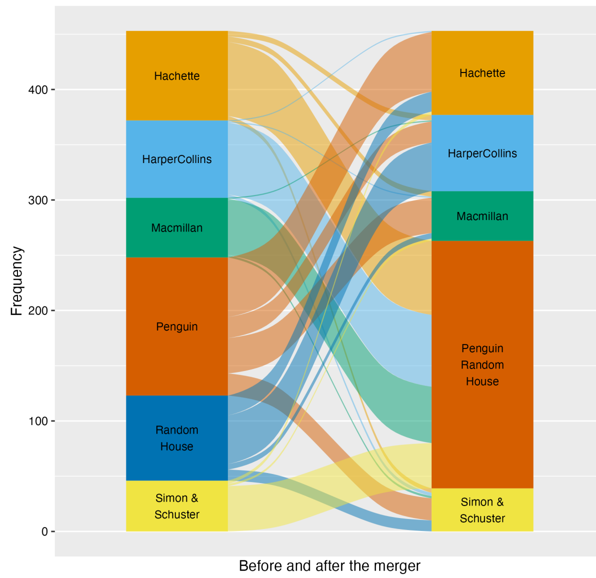
	Aggregate change			Average change per author		
	Joint surplus (1)	Author share (2)	Publisher share (3)	Joint surplus (4)	Author share (5)	Publisher share (6)
	Panel A: Total social change					
Social	6.44	-286.51	292.95			
	Panel B: Publisher total change					
Hachette	-8.39	-13.90	5.51			
HarperCollins	-5.48	-10.93	5.44			
Macmillan	-7.11	-14.11	7.00			
Penguin Random House	46.05	-197.18	243.23			
Simon & Schuster	-9.80	-15.63	5.83			
	Panel C: PRH's internal change					
Penguin Random House	1.26	-236.15	237.40	0.001	-0.102	0.102
	Panel D: Changes from sorting					
Hachette	-8.39	-8.30	-0.09	-0.262	-0.260	-0.003
HarperCollins	-5.48	-5.80	0.31	-0.228	-0.242	0.013
Macmillan	-7.11	-7.41	0.31	-0.395	-0.412	0.017
Penguin Random House	44.79	38.96	5.83	0.487	0.423	0.063
Simon & Schuster	-9.80	-10.08	0.28	-0.700	-0.720	0.020

experience a decrease in their utility post-merger, despite the overall market gains. This redistribution highlights the differential impact of the merger on the two sides within the publishing industry. I now discuss this distribution impact in detail.

Column (2) reveals changes to authors' share of the surplus. Panel A demonstrates that despite an overall net gain in social welfare, this improvement accrues to publishers at the expense of authors, who as a group suffer a net loss.

A closer look at Penguin Random House authors in Panel C reveals an even more pronounced inequality. Authors who remained with Penguin Random House experienced substantial utility losses, despite a slight increase in total value post-merger. Meanwhile, the publisher saw a notable increase. This loss stems primarily from weakened competition, a direct consequence of the merger. Pre-merger, Penguin and Random House had to

Figure 6: Movement of authors after the merger



Notes: Movement of authors before and after the merger. Note that 8% of authors have moved. Authors who have stayed with their original publishers are not show.

compete to match with desired authors, creating upward bidding pressure. Post-merger, this competitive dynamic disappears, aligning with concerns raised in the 2022 merger case about reduced competition between the formerly separate entities.

The impact on author welfare extends beyond those staying with Penguin Random House. Panel D shows that authors who moved between publishers experienced significant welfare changes, with the direction of movement determining gains or losses. This re-sorting, primarily involving exchanges between Penguin Random House and other publishers, results from Penguin Random House's expanded capacity post-merger. Authors moving to Penguin Random House saw substantial welfare gains, while those moving away suffered losses. This process creates a polarization among authors, with most of the value changes borne by the authors themselves. Essentially, we observe a transfer of utility from authors of other publishers to those of Penguin Random House, further illustrating the uneven distribution of merger effects across the industry.

Heterogeneity by author tenure. Given that the distributional effect on authors is a key concern in this market, I break down the analysis along author tenure. Specifically, I examine three groups of authors: bestselling, mid-list, and debut. The analysis focuses on two subsets: authors who remained with Penguin Random House and those who were matched with a different publisher post-merger. The results of this decomposition are

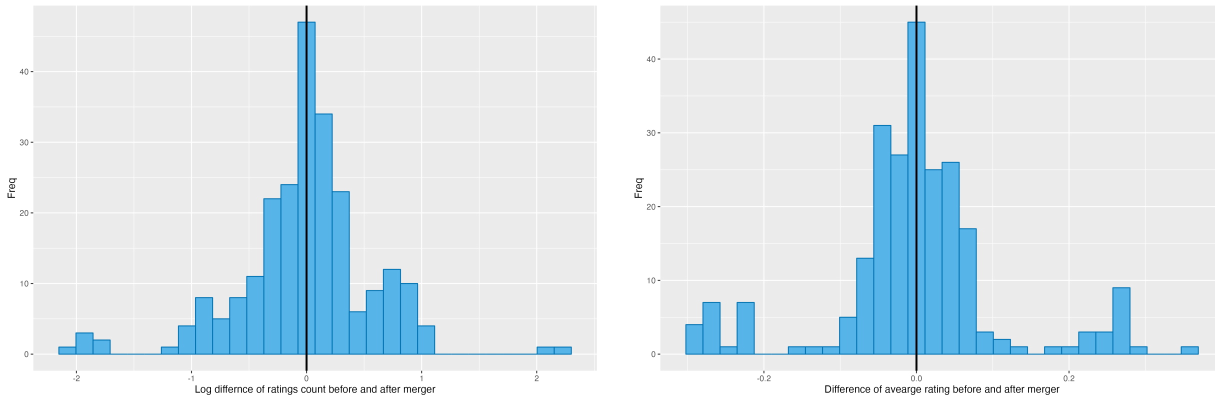
Table 6: Simulation results of synergistic collaboration by author tenure

	Aggregate change			Average change per author		
	Joint surplus (1)	Author share (2)	Publisher share (3)	Joint surplus (4)	Author share (4)	Publisher share (5)
	Panel A: PRH's internal change					
Best-selling	0.00	-22.53	22.53	0.000	-0.110	0.110
Mid-list	0.86	-208.83	209.68	0.001	-0.131	0.132
Debut	0.40	-4.79	5.19	0.001	-0.009	0.010
	Panel B: Changes from sorting					
Best-selling						
Hachette	0.05	-0.05	0.10	0.015	-0.017	0.032
HarperCollins	-0.15	0.00	-0.15	-0.076	0.001	-0.077
Macmillan	0.23	-0.00	0.23	0.227	-0.001	0.228
Mid-list						
Hachette	0.28	-0.09	0.37	0.031	-0.010	0.041
HarperCollins	0.10	-0.07	0.18	0.010	-0.007	0.018
Macmillan	1.20	-0.30	1.50	0.150	-0.037	0.187
Penguin Random House	0.86	0.34	0.52	0.028	0.011	0.017
Simon & Schuster	0.89	-0.28	1.16	0.148	-0.046	0.194
Debut						
Hachette	-7.76	-0.42	-7.34	-0.388	-0.021	-0.367
HarperCollins	-0.45	-0.66	0.21	-0.038	-0.055	0.018
Macmillan	1.44	-0.29	1.73	0.160	-0.032	0.192
Penguin Random House	-12.56	1.31	-13.87	-0.206	0.022	-0.227
Simon & Schuster	-0.33	-0.15	-0.18	-0.042	-0.018	-0.023

presented Table 6.

Panel A shows changes for authors who remained with Penguin Random House. Authors across all three tenure categories experienced losses, but the impact is uneven. At the average author level, bestselling and mid-list authors suffered notably greater loss compared to debut authors in absolute terms. This disparity stems from bestselling authors being the most sought-after pre-merger; thus, the loss of competition post-merger resulted in the largest utility shock for them. This finding further supports the DOJ's argument in the 2022 merger case that top-selling authors stand to lose the most. Interestingly, while the writer community was justifiably concerned that debut authors would be worse off

Figure 7: Changes in reader reception



(a) Change in rating count

(b) Change in average rating

after the merger, the analysis shows that the transfer is largely from authors to publishers rather than among authors themselves.

Panel B illustrates changes for authors who were sorted to different publishers post-merger. The findings align with previous observations of a transfer from other publishers' authors to those of Penguin Random House but reveal heterogeneous effects across author tenure categories. At the per-author level, bestselling authors who left Penguin Random House suffered the most significant losses, while those joining gained little. In contrast, debut authors who left Penguin Random House experienced comparatively smaller losses, but those who joined reaped the most substantial gains. Mid-list authors fall between these two extremes. These patterns underscore Penguin Random House's pivotal role as the market leader in driving value distribution across the industry.

Impact on reader reception. Finally, I investigate the impact on the consumer side in terms of the reception of the reader of the books affected by the merger. Figure 7 shows changes in rating count and average ratings for books that were directly impacted and sorted to different publishers. The analysis reveals negligible changes in both metrics, with a t-statistic test confirming that the differences are not statistically significant. Books did not experience significant changes in popularity or perceived quality after accounting for publisher changes. This finding aligns with industry consensus that the merger's primary effects would not materialize on the reader side. Notably, if this merger were evaluated solely on consumer welfare grounds, as is conventionally done, it would appear harmless.

Alternative counterfactual assumptions. The results of the other two counterfactual simulations, "organic merger" and "Random House takeover," are presented in Appendix E.

These simulations generated qualitatively similar results to our main findings. Although the magnitude of effects varied, the overall patterns remained consistent: redistribution from authors to publishers, heterogeneous impacts across author tenure categories, and Penguin Random House's significant role in reshaping market dynamics. This consistency across counterfactuals strengthens the robustness of my conclusions about the merger's impacts on the publishing industry.

6 Conclusion

I study the impact of market consolidation on the labor market for creativity using a two-sided matching framework. As competition concerns in labor markets have grown in recent years, there is an increasing need for analytic tools tailored to their special characteristics. Using the publishing industry, which exemplifies two-sided market preferences, I find strong patterns of assortative matching, confirming compatibility as a crucial feature in analyzing this market.

I then develop an empirical matching model with transferable utilities and structurally recover match values from observed matches. To evaluate merger impacts, I perform counterfactual merger simulations based on these recovered structural parameters. My results reveal that while the merger generates positive efficiency gains, these benefits accrue primarily to publishers, particularly the merged firm. The merger redistributes welfare from other publishers to the merged firm and from authors to publishers generally. The impact varies across author tenure categories: bestselling authors are most negatively affected, particularly those previously working with Penguin Random House, who either stayed or moved away. In contrast, debut and mid-list authors experience relatively mild impacts in absolute terms.

My findings support the DOJ's intervention in the 2022 merger attempt between Penguin Random House and Simon & Schuster. While the agency's primary concern was authors' potential loss of compensation, my analysis systematically demonstrates the adverse impacts on authors in this market. Moreover, based on these results, one might argue that even the 2013 merger was anticompetitive. Notably, examining consumer welfare alone would not have raised concerns; the merger was even defended as necessary to strengthen the industry's bargaining power against downstream distributors, particularly Amazon. My analysis reveals how evaluating mergers solely on consumer welfare grounds may overlook significant anticompetitive effects in labor markets.

My analysis has implications for other industries where labor matching is critical. The publishing industry exemplifies two key features common to high-skilled labor markets: worker-firm compatibility matters, and employment relationships extend beyond mere transactions. In sectors such as consulting, academia, and the creative industries, workers and firms invest significantly in finding suitable matches, and the quality of these matches substantially affects productivity. These features suggest that traditional merger analysis that focuses solely on price effects or consumer welfare may miss important competitive dynamics in labor markets. Given the growing dominance of large firms across such industries, this raises concerns about workers' competitive disadvantage, particularly in sectors where relationship-specific investments and worker-firm compatibility are central to value creation.

References

Agarwal, Nikhil (2015). "An Empirical Model of the Medical Match". *American Economic Review* 105.7, pp. 1939–1978.

Agarwal, Nikhil and Eric Budish (2021). "Market Design". *Handbook of Industrial Organization*. Vol. 5. Elsevier, pp. 1–79.

Akkus, Oktay, J. Anthony Cookson, and Ali Hortaçsu (2016). "The Determinants of Bank Mergers: A Revealed Preference Analysis". *Management Science* 62.8, pp. 2241–2258.

Arnold, David (2019). *Mergers and Acquisitions, Local Labor Market Concentration, and Worker Outcomes*. SSRN Scholarly Paper. Rochester, NY.

Arnold, David, Kevin S. Milligan, Terry Moon, and Amirhossein Tavakoli (2023). *Job Transitions and Employee Earnings After Acquisitions: Linking Corporate and Worker Outcomes*. Working Paper.

Ash, Elliott and Stephen Hansen (2023). "Text Algorithms in Economics". *Annual Review of Economics* 15.Volume 15, 2023, pp. 659–688.

Ash, Elliott, Massimo Morelli, and Matia Vannoni (2022). *More Laws, More Growth? Evidence from U.S. States*. SSRN Scholarly Paper. Rochester, NY.

Ashenfelter, Orley C., Henry Farber, and Michael R Ransom (2010). "Labor Market Monopsony". *Journal of Labor Economics* 28.2, pp. 203–210.

Asker, John and Volker Nocke (2021). "Collusion, Mergers, and Related Antitrust Issues". *Handbook of Industrial Organization*. Ed. by Kate Ho, Ali Hortaçsu, and Alessandro Lizzeri. Vol. 5. Handbook of Industrial Organization, Volume 5. Elsevier, pp. 177–279.

Azar, José and Ioana Marinescu (2024). "Monopsony Power in the Labor Market: From Theory to Policy". *Annual Review of Economics* 16.Volume 16, 2024, pp. 491–518.

Azar, José, Ioana Marinescu, Marshall Steinbaum, and Bledi Taska (2020). "Concentration in US Labor Markets: Evidence from Online Vacancy Data". *Labour Economics* 66, p. 101886.

Azar, José A., Steven T. Berry, and Ioana Marinescu (2022). *Estimating Labor Market Power*. Working Paper.

Azevedo, Eduardo M. and John William Hatfield (2018). *Existence of Equilibrium in Large Matching Markets With Complementarities*. SSRN Scholarly Paper. Rochester, NY.

Bandiera, Oriana, Andrea Prat, Stephen Hansen, and Raffaella Sadun (2020). "CEO Behavior and Firm Performance". *Journal of Political Economy* 128.4, pp. 1325–1369.

Becker, Gary S. (1973). "A Theory of Marriage: Part I". *Journal of Political Economy* 81.4, pp. 813–846.

Berger, David, Kyle Herkenhoff, and Simon Mongey (2022). "Labor Market Power". *American Economic Review* 112.4, pp. 1147–1193.

Berger, David W., Thomas Hasenzagl, Kyle F. Herkenhoff, Simon Mongey, and Eric A. Posner (2023). *Merger Guidelines for the Labor Market*. Working Paper.

Bertrand, Marianne, Matilde Bombardini, Raymond Fisman, Brad Hackinen, and Francesco Trebbi (2021). "Hall of Mirrors: Corporate Philanthropy and Strategic Advocacy*". *The Quarterly Journal of Economics* 136.4, pp. 2413–2465.

Biasi, Barbara and Petra Moser (2021). "Effects of Copyrights on Science: Evidence from the WWII Book Republication Program". *American Economic Journal: Microeconomics* 13.4, pp. 218–260.

Boal, William M. and Michael R Ransom (1997). "Monopsony in the Labor Market". *Journal of Economic Literature* 35.1, pp. 86–112.

Bolton, Gary E., Elena Katok, and Axel Ockenfels (2004). "How Effective Are Electronic Reputation Mechanisms? An Experimental Investigation". *Management Science* 50.11, pp. 1587–1602.

Bonaimé, Alice and Ye (emma) Wang (2024). "Mergers, Product Prices, and Innovation: Evidence from the Pharmaceutical Industry". *The Journal of Finance* 79.3, pp. 2195–2236.

Boyd, Donald, Hamilton Lankford, Susanna Loeb, and James Wyckoff (2013). "Analyzing the Determinants of the Matching of Public School Teachers to Jobs: Disentangling the Preferences of Teachers and Employers". *Journal of Labor Economics* 31.1, pp. 83–117.

Cabral, Luis (2012). "Reputation on the Internet". *The Oxford Handbook of the Digital Economy*. Ed. by Martin Peitz and Joel Waldfogel. Oxford University Press, pp. 343–354.

Cabral, Luís and Ali Hortaçsu (2010). "The Dynamics of Seller Reputation: Evidence from Ebay". *The Journal of Industrial Economics* 58.1, pp. 54–78.

Cagé, Julia, Nicolas Hervé, and Marie-Luce Viaud (2020). "The Production of Information in an Online World". *The Review of Economic Studies* 87.5, pp. 2126–2164.

Canoy, Marcel, Jan C. van Ours, and Frederick van der Ploeg (2006). "The Economics of Books". *Handbook of the Economics of Art and Culture*. Ed. by Victor A. Ginsburg and David Throsby. Vol. 1. Elsevier, pp. 721–761.

Card, David, Ana Rute Cardoso, Joerg Heining, and Patrick Kline (2018). "Firms and Labor Market Inequality: Evidence and Some Theory". *Journal of Labor Economics* 36.S1, S13–S70.

Chade, Hector, Jan Eeckhout, and Lones Smith (2017). "Sorting through Search and Matching Models in Economics". *Journal of Economic Literature* 55.2, pp. 493–544.

Chen, Guoli, Sterling Huang, Philipp Meyer-Doyle, and Denisa Mindruta (2021). "Generalist versus Specialist CEOs and Acquisitions: Two-sided Matching and the Impact of

CEO Characteristics on Firm Outcomes". *Strategic Management Journal* 42.6, pp. 1184–1214.

Chen, Yubo and Jinhong Xie (2008). "Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix". *Management Science* 54.3, pp. 477–491.

Chevalier, Judith A. and Dina Mayzlin (2006). "The Effect of Word of Mouth on Sales: Online Book Reviews". *Journal of Marketing Research* 43.3, pp. 345–354.

Chiappori, Pierre-André, Robert J. McCann, and Lars P. Nesheim (2010). "Hedonic Price Equilibria, Stable Matching, and Optimal Transport: Equivalence, Topology, and Uniqueness". *Economic Theory* 42.2, pp. 317–354.

Chiappori, Pierre-André and Bernard Salanié (2016). "The Econometrics of Matching Models". *Journal of Economic Literature* 54.3, pp. 832–861.

Choo, Eugene and Aloysius Siow (2006). "Who Marries Whom and Why". *Journal of Political Economy* 114.1, pp. 175–201.

Crawford, Vincent P. and Elsie Marie Knoer (1981). "Job Matching with Heterogeneous Firms and Workers". *Econometrica* 49.2, pp. 437–450.

Delabastita, Vincent and Michael Rubens (2023). *Colluding Against Workers*.

Dellarocas, Chrysanthos (2003). "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms". *Management Science* 49.10, pp. 1407–1424.

Deng, Yipu, Jinyang Zheng, Warut Khern-am-nuai, and Karthik Natarajan Kannan (2021). *More than the Quantity: The Value of Editorial Reviews for a UGC Platform*. SSRN Scholarly Paper. Rochester, NY.

Department of Justice and Federal Trade Commission (2023). *2023 Merger Guidelines*.

Djourelova, Milena, Ruben Durante, and Gregory J Martin (2024). "The Impact of Online Competition on Local Newspapers: Evidence from the Introduction of Craigslist". *The Review of Economic Studies*, rdae049.

Dupuy, Arnaud, Alfred Galichon, Sonia Jaffe, and Scott Duke Kominers (2017). *Taxation in Matching Markets*. SSRN Scholarly Paper. Rochester, NY.

Eeckhout, Jan (2018). "Sorting in the Labor Market". *Annual Review of Economics* 10.1, pp. 1–29.

Eeckhout, Jan and Philipp Kircher (2011). "Identifying Sorting—In Theory". *The Review of Economic Studies* 78.3, pp. 872–906.

Eeckhout, Jan and Philipp Kircher (2018). "Assortative Matching With Large Firms". *Econometrica* 86.1, pp. 85–132.

Eliason, Paul J, Benjamin Heebsh, Ryan C McDevitt, and James W Roberts (2020). "How Acquisitions Affect Firm Behavior and Performance: Evidence from the Dialysis Industry". *The Quarterly Journal of Economics* 135.1, pp. 221–267.

Executive Order No. 14036, 3 C.F.R. 36987 (2021).

Fan, Ying (2013). "Ownership Consolidation and Product Characteristics: A Study of the US Daily Newspaper Market". *American Economic Review* 103.5, pp. 1598–1628.

Federal Trade Commission v. Kroger Company, 3:24-Cv-00347 (D. Or.) (N.d.).

Fisher, Jack (2024). *Monopsony Power in the Gig Economy*.

Fox, Jeremy T. (2010). "Identification in Matching Games: Identification in Matching Games". *Quantitative Economics* 1.2, pp. 203–254.

Fox, Jeremy T. (2018). "Estimating Matching Games with Transfers". *Quantitative Economics* 9.1, pp. 1–38.

Galichon, Alfred and Bernard Salanié (2023). "Structural Estimation of Matching Markets with Transferable Utility". *Online and Matching-Based Market Design*. Ed. by Federico Echenique, Nicole Immorlica, and Vijay V. Vazirani. 1st ed. Cambridge University Press, pp. 552–572.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin (2013). *Bayesian Data Analysis, Third Edition*. CRC Press.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019). "Text as Data". *Journal of Economic Literature* 57.3, pp. 535–574.

Giorcelli, Michela and Petra Moser (2020). "Copyrights and Creativity: Evidence from Italian Opera in the Napoleonic Age". *Journal of Political Economy* 128.11, pp. 4163–4210.

Goldfarb, Avi and Catherine Tucker (2019). "Digital Economics". *Journal of Economic Literature* 57.1, pp. 3–43.

Graham, Bryan S. (2011). "Econometric Methods for the Analysis of Assignment Problems in the Presence of Complementarity and Social Spillovers". *Handbook of Social Economics*. Vol. 1. Elsevier, pp. 965–1052.

Gretsky, Neil E., Joseph M. Ostroy, and William R. Zame (1992). "The Nonatomic Assignment Model". *Economic Theory* 2.1, pp. 103–127.

Hansen, Stephen, Michael McMahon, and Andrea Prat (2018). "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach*". *The Quarterly Journal of Economics* 133.2, pp. 801–870.

Heckman, James J. (1979). "Sample Selection Bias as a Specification Error". *Econometrica* 47.1, pp. 153–161.

Hwang, Hae-shin, Dale T. Mortensen, and W. Robert Reed (1998). "Hedonic Wages and Labor Market Search". *Journal of Labor Economics* 16.4, pp. 815–847.

Igami, Mitsuru and Kosuke Uetake (2020). "Mergers, Innovation, and Entry-Exit Dynamics: Consolidation of the Hard Disk Drive Industry, 1996–2016". *The Review of Economic Studies* 87.6, pp. 2672–2702.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy (2021). "Measuring Technological Innovation over the Long Run". *American Economic Review: Insights* 3.3, pp. 303–320.

Kelso, Alexander S. and Vincent P. Crawford (1982). "Job Matching, Coalition Formation, and Gross Substitutes". *Econometrica* 50.6, pp. 1483–1504.

Lamadon, Thibaut, Magne Mogstad, and Bradley Setzler (2022). "Imperfect Competition, Compensating Differentials, and Rent Sharing in the US Labor Market". *American Economic Review* 112.1, pp. 169–212.

Li, Sophia, Joe Mazur, Yongjoon Park, James Roberts, Andrew Sweeting, and Jun Zhang (2022). "Repositioning and Market Power after Airline Mergers". *The RAND Journal of Economics* 53.1, pp. 166–199.

Manning, Alan (2003). *Monopsony in Motion: Imperfect Competition in Labor Markets*. Princeton University Press.

Manning, Alan (2011). "Imperfect Competition in the Labor Market". *Handbook of Labor Economics*. Ed. by David Card and Orley Ashenfelter. Vol. 4. Elsevier, pp. 973–1041.

Manning, Alan (2021). "Monopsony in Labor Markets: A Review". *ILR Review* 74.1, pp. 3–26.

Marinescu, Ioana and Herbert Hovenkamp (2019). "Anticompetitive Mergers in Labor Markets". *Indiana Law Journal* 94.3, p. 1031.

Marinescu, Ioana Elena and Eric A. Posner (2020). "Why Has Antitrust Law Failed Workers?" *Cornell Law Review* 105.5.

Mindruta, Denisa (2013). "Value Creation in University-Firm Research Collaborations: A Matching Approach". *Strategic Management Journal* 34.6, pp. 644–665.

Mindruta, Denisa, Mahka Moeen, and Rajshree Agarwal (2016). "A Two-Sided Matching Approach for Partner Selection and Assessing Complementarities in Partners' Attributes in Inter-Firm Alliances". *Strategic Management Journal* 37.1, pp. 206–231.

Montag, Felix (2023). "Mergers, Foreign Competition, and Jobs: Evidence from the U.S. Appliance Industry". *SSRN Electronic Journal*.

Nagaraj, Abhishek and Imke Reimers (2023). "Digitization and the Market for Physical Works: Evidence from the Google Books Project". *American Economic Journal: Economic Policy* 15.4, pp. 428–458.

Naidu, Suresh, Eric A. Posner, and Glen Weyl (2018). "Antitrust Remedies for Labor Market Power". *Harvard Law Review* 132.2, pp. 536–601.

Peukert, Christian and Imke Reimers (2022). "Digitization, Prediction, and Market Efficiency: Evidence from Book Publishing Deals". *Management Science* 68.9, pp. 6907–6924.

Posner, Eric A. (2021). *How Antitrust Failed Workers*. Oxford, New York: Oxford University Press.

Prager, Elena and Matt Schmitt (2021). "Employer Consolidation and Wages: Evidence from Hospitals". *American Economic Review* 111.2, pp. 397–427.

Reimers, Imke and Joel Waldfogel (2021). "Digitization and Pre-purchase Information: The Causal and Welfare Impacts of Reviews and Crowd Ratings". *American Economic Review* 111.6, pp. 1944–1971.

Rosen, Sherwin (1986). "The Theory of Equalizing Differences". *Handbook of Labor Economics*. Vol. 1. Elsevier, pp. 641–692.

Rossi, Peter E., Greg M. Allenby, and Rob McCulloch (2012). *Bayesian Statistics and Marketing*. John Wiley & Sons.

Roth, Alvin E. (1984). "Stability and Polarization of Interests in Job Matching". *Econometrica* 52.1, pp. 47–57.

Rubens, Michael (2023). "Market Structure, Oligopsony Power, and Productivity". *American Economic Review* 113.9, pp. 2382–2410.

Shapiro, Carl (2019). "Protecting Competition in the American Economy: Merger Control, Tech Titans, Labor Markets". *Journal of Economic Perspectives* 33.3, pp. 69–93.

Shapley, L. S. and M. Shubik (1971). "The Assignment Game I: The Core". *International Journal of Game Theory* 1.1, pp. 111–130.

Sorensen, Morten (2005). "An Economic and Econometric Analysis of Market Sorting with an Application to Venture Capital". PhD thesis.

Sørensen, Morten (2007). "How Smart Is Smart Money? A Two-Sided Matching Model of Venture Capital". *The Journal of Finance* 62.6, pp. 2725–2762.

Sotomayor, Marilda (1999). "The Lattice Structure of the Set of Stable Outcomes of the Multiple Partners Assignment Game". *International Journal of Game Theory* 28.4, pp. 567–583.

Sun, Monic (2012). "How Does the Variance of Product Ratings Matter?" *Management Science* 58.4, pp. 696–707.

Sweeting, Andrew (2010). "The Effects of Mergers on Product Positioning: Evidence from the Music Radio Industry". *The RAND Journal of Economics* 41.2, pp. 372–397.

Taber, Christopher and Rune Vejlin (2020). "Estimation of a Roy/Search/Compensating Differential Model of the Labor Market". *Econometrica* 88.3, pp. 1031–1069.

Thompson, John B. (2012). *Merchants of Culture: The Publishing Business in the Twenty-First Century*. Plume.

Treuren, Leonard (2022). *Wage Markups and Buyer Power in Intermediate Input Markets*.

U.S. v. Bertelsmann SE & Co. KGaA, 646 F. Supp. 3d 1 (D.D.C. 2022) (2022).

Wan, Mengting and Julian McAuley (2018). “Item Recommendation on Monotonic Behavior Chains”. *Proceedings of the 12th ACM Conference on Recommender Systems*. RecSys ’18. New York, NY, USA: Association for Computing Machinery, pp. 86–94.

Wan, Mengting, Rishabh Misra, Ndapa Nakashole, and Julian McAuley (2019). “Fine-Grained Spoiler Detection from Large-Scale Review Corpora”. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2605–2610.

Wollmann, Thomas G. (2018). “Trucks without Bailouts: Equilibrium Product Characteristics for Commercial Vehicles”. *American Economic Review* 108.6, pp. 1364–1406.

Wu, Chunhua, Hai Che, Tat Y. Chan, and Xianghua Lu (2015). “The Economic Value of Online Reviews”. *Marketing Science* 34.5, pp. 739–754.

Yang, Yupin, Mengze Shi, and Avi Goldfarb (2009). “Estimating the Value of Brand Alliances in Professional Team Sports”. *Marketing Science* 28.6, pp. 1095–1111.

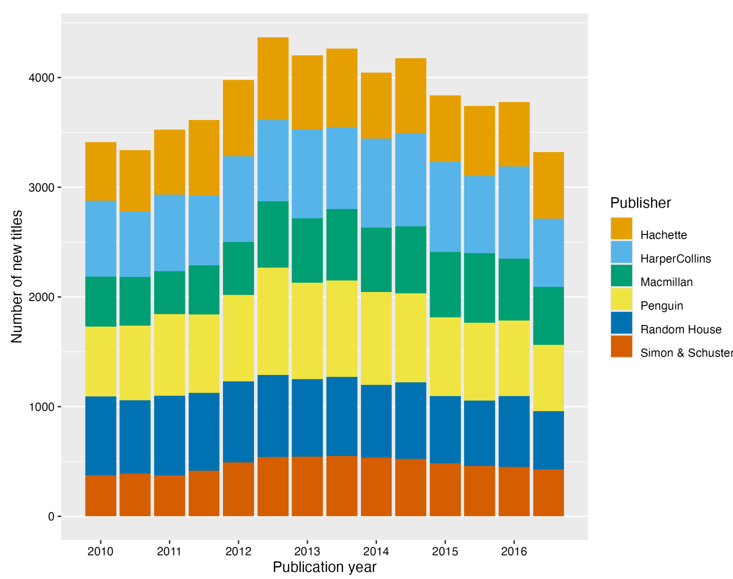
Yeh, Chen, Claudia Macaluso, and Brad Hershbein (2022). “Monopsony in the US Labor Market”. *American Economic Review* 112.7, pp. 2099–2138.

Appendix A Data Details

The data used for this paper is from Goodreads, and collected by Wan and McAuley (2018) and Wan et al. (2019). Figure Figure A8 shows the number of new titles books published by the "Big Six" in each half-year in the sample period 2010-2016 by publisher, genre, and author tenure. Reprints or new editions of existing titles are not included.

In the original dataset, either the imprint, division, or the publishing company is observed as the publisher for each book. *Imprints* are trade names under which books are published. A single publishing company may have many imprints, often the result of market consolidation. The imprint names have been kept to preserve unique editorial identities and serve specific reader segments. For example, Penguin Random House has more than 300 imprints as of 2020.[39] Some notable ones include DK, Alfred A. Knopf, Doubleday, Vintage, Viking, etc. Penguin and Random House are themselves imprint names, as well. I have manually coded the imprints to their parent publishers. Therefore, imprints that originally belong to Penguin or Random House can still be distinguished post-merger, but in the analysis are treated as a single entity.

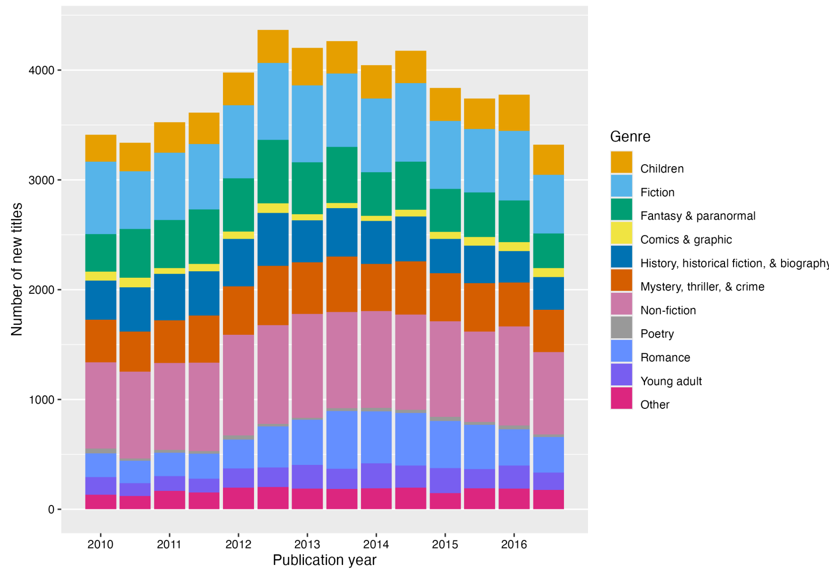
Figure A8: Number of new titles in each half-year



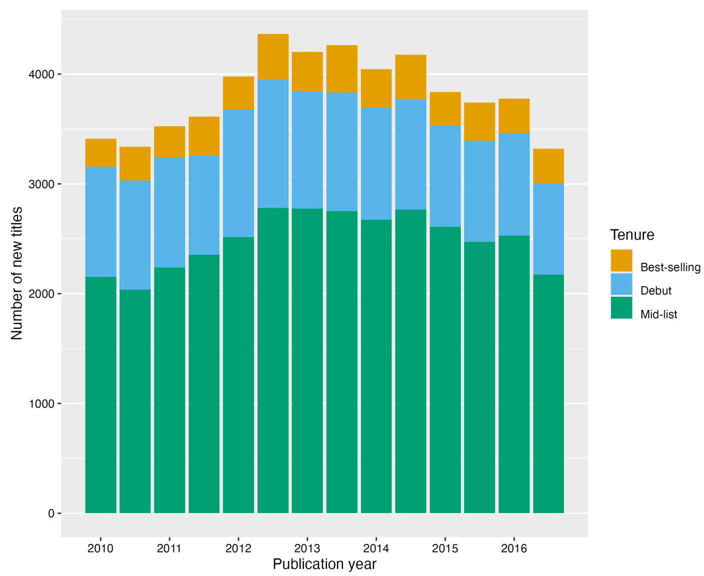
(a) By publisher

[39] See https://www.publishersweekly.com/pw/by-topic/industry-news/publisher-news/article/82901-bertelsmann-now-owns-100-of-prh.html.

Figure A8: Number of new titles in each half-year (cont.)



(b) By genre



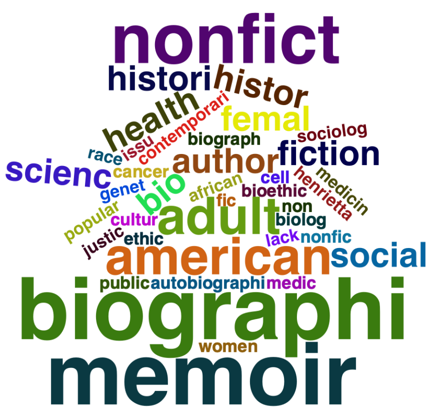
(c) By author tenure

Appendix B Topic Modeling

B.1 Text documents

Each book in the data has two associated text documents: bookshelf labels and a description. Figures Figure B9 and Figure B10 show word clouds of shelf labels and descriptions for four bestsellers from 2010–2012. Text documents are preprocessed using standard procedures before topic modeling, including tokenization, lowercasing, stemming, and stopword removal.

Figure B9: Examples of book shelf labels



(a) *The Immortal Life of Henrietta Lacks*, Rebecca Skloot, Pan Macmillan, 2010

(b) *Mockingjay (The Hunger Games, #3)*, Suzanne Collins, Scholastic Press, 2010

(c) *Thinking, Fast and Slow*, Daniel Kahneman, Farrar, Straus and Giroux, 2011

(d) *Dragons Love Tacos*, Adam Rubin, illustrated by Daniel Salmieri, Dial Books, 2012

Figure B10: Examples of book description

(a) *The Immortal Life of Henrietta Lacks*, Rebecca Skloot, Pan Macmillan, 2010

(b) *Mockingjay (The Hunger Games, #3)*, Suzanne Collins, Scholastic Press, 2010

(c) *Thinking, Fast and Slow*, Daniel Kahneman, Farrar, Straus and Giroux, 2011

(d) *Dragons Love Tacos*, Adam Rubin, illustrated by Daniel Salmieri, Dial Books, 2012

B.2 Genre topics

Figure B11 shows word clouds of some example genre topics from the LDA model trained on the corpus of book shelf labels. The most frequent terms in these topics are "apocalypse," "religion," "compute," and "social," respectively. Figure B12 shows the word probabilities of the most frequent words in all 50 topics.

Figure B11: Examples of genre topic word clouds

(a) Topic No. 11

(b) Topic No. 23

(c) Topic No. 33

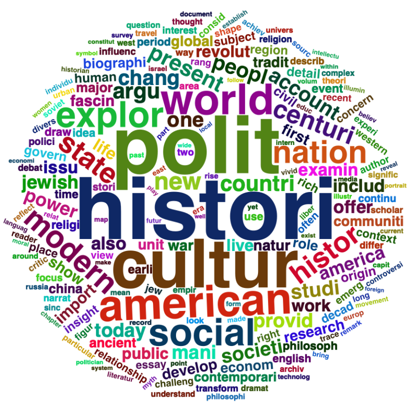
(d) Topic No. 45

Figure B12: Genre topic word probabilities from the LDA model

B.3 Content topics

Figure B13 shows word clouds of some example content topics from the LDA model trained on the corpus of book descriptions. The most frequent terms in these topics are "history," "life," "poem," and "children," respectively. Figure B14 shows the word probabilities of the most frequent terms in all 50 topics.

Figure B13: Examples of content topic word clouds



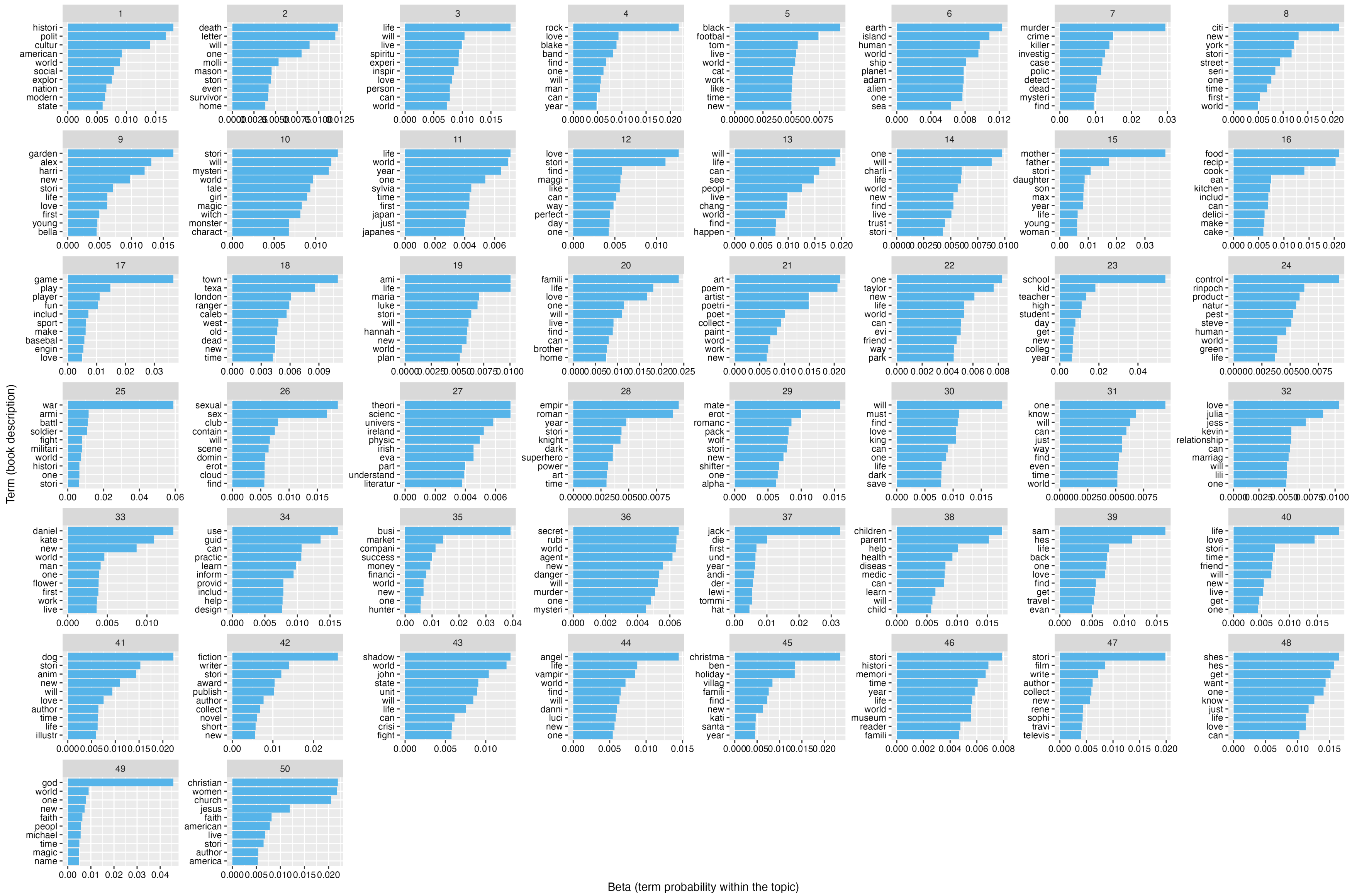
(a) Topic No. 1

(b) Topic No. 3

(c) Topic No. 21

(d) Topic No. 38

Figure B14: Content topic word probabilities from the LDA model



Appendix C More Descriptive Evidence

C.1 Event study of the merger

Table C7: Changes to pre-publication characteristics

	Author ratings count percentile	Author average rating
	(1)	(2)
PRH × Year_{2010}	−0.005	0.001
	(0.006)	(0.008)
PRH × $\text{Year}_{2010.5}$	−0.011	0.016*
	(0.006)	(0.008)
PRH × Year_{2011}	−0.013*	0.005
	(0.006)	(0.008)
PRH × $\text{Year}_{2011.5}$	0.002	0.012
	(0.006)	(0.008)
PRH × Year_{2012}	−0.003	0.007
	(0.006)	(0.008)
PRH × $\text{Year}_{2012.5}$	−0.002	0.005
	(0.006)	(0.007)
PRH × Year_{2013}	−0.015*	−0.001
	(0.006)	(0.007)
PRH × Year_{2014}	−0.012*	−0.020**
	(0.006)	(0.008)
PRH × $\text{Year}_{2014.5}$	−0.008	−0.024**
	(0.006)	(0.008)
PRH × Year_{2015}	−0.010	−0.032***
	(0.006)	(0.008)
PRH × $\text{Year}_{2015.5}$	−0.004	−0.022**
	(0.006)	(0.008)
PRH × Year_{2016}	−0.010	−0.040***
	(0.006)	(0.008)
PRH × $\text{Year}_{2016.5}$	−0.007	−0.028***
	(0.007)	(0.008)
Constant	0.457***	3.923***
	(0.006)	(0.008)
Book characteristics	Yes	Yes
Book-publisher characteristics	Yes	Yes
R^2	0.815	0.989
Observations	136731	136731

Notes: The reference year is 2013.5. Control variables are not reported. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Table C8: Changes to post-publication performance

	Ratings count percentile	log(Ratings count)	Average rating
	(1)	(2)	(3)
PRH × Year_{2010}	−0.005	−0.040	0.023
	(0.007)	(0.057)	(0.012)
PRH × $\text{Year}_{2010.5}$	0.000	0.005	0.028*
	(0.007)	(0.057)	(0.012)
PRH × Year_{2011}	−0.006	−0.027	0.021
	(0.007)	(0.055)	(0.011)
PRH × $\text{Year}_{2011.5}$	−0.007	−0.030	0.042***
	(0.007)	(0.055)	(0.011)
PRH × Year_{2012}	0.007	0.050	0.051***
	(0.007)	(0.054)	(0.011)
PRH × $\text{Year}_{2012.5}$	−0.011	−0.069	0.025*
	(0.007)	(0.052)	(0.011)
PRH × Year_{2013}	0.002	0.005	0.018
	(0.007)	(0.053)	(0.011)
PRH × Year_{2014}	0.005	0.012	−0.017
	(0.007)	(0.053)	(0.011)
PRH × $\text{Year}_{2014.5}$	0.010	0.049	−0.017
	(0.007)	(0.053)	(0.011)
PRH × Year_{2015}	0.009	0.026	−0.023*
	(0.007)	(0.055)	(0.011)
PRH × $\text{Year}_{2015.5}$	0.003	0.007	−0.025*
	(0.007)	(0.056)	(0.011)
PRH × Year_{2016}	0.004	−0.007	−0.026*
	(0.007)	(0.056)	(0.011)
PRH × $\text{Year}_{2016.5}$	0.000	−0.002	−0.035**
	(0.007)	(0.058)	(0.012)
Constant	−0.089***	−1.253***	1.729***
	(0.012)	(0.095)	(0.019)
Book characteristics	Yes	Yes	Yes
Book-publisher characteristics	Yes	Yes	Yes
R^2	0.649	0.622	0.307
Observations	136731	136731	136731

Notes: The reference year is 2013.5. Control variables are not reported. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

C.2 More specifications of match formation

Table C9 and Table C10 present alternative specifications of match formation. In Table C9, the unit of observation is an author-publisher pair and the outcome is a binary variable indicating if it is a match. In Table C10, the unit of observation is a book and the outcome variable is the publisher to which the book is matched. In other words, these are multinomial choice models from the perspective of the author. To be consistent with the structural estimation, the subsample of 2010-13 data is used. Note that only the Big Five and fringe publishers are used in the estimation because self-publishing is considered to be the outside option.

Table C9: Matching formation with binary outcomes

	LPM	Logit		Probit	
		Estimate	Marginal Effect	Estimate	Marginal Effect
	(1)	(2)	(3)	(4)	(5)
Ratings count percentile interaction	-0.118^{***}	-1.967^{***}	-0.116^{***}	-0.932^{***}	-0.110^{***}
	(0.002)	(0.040)	(0.002)	(0.019)	(0.002)
Average rating interaction	0.046^{***}	0.752^{***}	0.044^{***}	0.347^{***}	0.041^{***}
	(0.001)	(0.020)	(0.001)	(0.010)	(0.001)
Debut interaction	0.109^{***}	1.735^{***}	0.102^{***}	0.821^{***}	0.097^{***}
	(0.004)	(0.067)	(0.004)	(0.032)	(0.004)
Bestselling interaction	-0.045^{***}	-0.704^{***}	-0.042^{***}	-0.402^{***}	-0.048^{***}
	(0.013)	(0.212)	(0.013)	(0.107)	(0.013)
Collaboration before	0.315^{***}	1.969^{***}	0.116^{***}	1.146^{***}	0.136^{***}
	(0.003)	(0.031)	(0.002)	(0.018)	(0.002)
log(Num prior collaborations)	0.210^{***}	1.380^{***}	0.081^{***}	0.726^{***}	0.086^{***}
	(0.002)	(0.022)	(0.001)	(0.012)	(0.001)
Genre similarity	0.074^{***}	1.201^{***}	0.071^{***}	0.597^{***}	0.071^{***}
	(0.001)	(0.022)	(0.001)	(0.011)	(0.001)
Content similarity	0.067^{***}	1.170^{***}	0.069^{***}	0.539^{***}	0.064^{***}
	(0.002)	(0.029)	(0.002)	(0.014)	(0.002)
Constant	-0.022^{***}	-4.214^{***}		-2.251^{***}	
	(0.002)	(0.030)		(0.014)	
R^2	0.251				
Num. obs.	520968	520968		520968	
Log Likelihood		-117823.552		-117192.775	

Notes: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Table C10: Matching formation with categorical outcomes

	Multinomial logit						Multinomial probit					
	(1) Hachette	(2) Harper-Collins	(3) Mac-millan	(4) Penguin	(5) Random House	(6) Simon & Schuster	(7) Hachette	(8) Harper-Collins	(9) Mac-millan	(10) Penguin	(11) Random House	(12) Simon & Schuster
Ratings count percentile interaction	2.125^{***}	1.720^{***}	2.681^{***}	2.420^{***}	2.520^{***}	2.944^{***}	0.732^{***}	0.421^{***}	1.114^{***}	0.757^{***}	0.922^{***}	1.166^{***}
	(0.141)	(0.140)	(0.155)	(0.136)	(0.141)	(0.162)	(0.078)	(0.080)	(0.079)	(0.061)	(0.074)	(0.089)
Average rating interaction	-1.049^{***}	-1.304^{***}	-1.497^{***}	-1.444^{***}	-1.601^{***}	-1.501^{***}	-0.339^{***}	-0.480^{***}	-0.600^{***}	-0.469^{***}	-0.588^{***}	-0.541^{***}
	(0.077)	(0.072)	(0.081)	(0.071)	(0.072)	(0.087)	(0.042)	(0.042)	(0.040)	(0.030)	(0.038)	(0.042)
Debut interaction	1.886^{***}	1.091^{***}	1.373^{***}	1.059^{***}	0.997^{***}	1.445^{***}	0.769^{***}	0.314^{*}	0.476^{***}	0.198^{*}	0.257^{**}	0.408^{***}
	(0.239)	(0.222)	(0.241)	(0.218)	(0.214)	(0.261)	(0.143)	(0.124)	(0.110)	(0.080)	(0.096)	(0.115)
Bestselling interaction	9.071^{***}	9.661^{***}	0.404	5.527^{***}	3.274^{***}	7.521^{***}	3.979^{***}	4.344^{***}	-2.114^{***}	1.278^{***}	-0.251	2.041^{***}
	(0.754)	(0.751)	(0.931)	(0.769)	(0.818)	(0.786)	(0.350)	(0.376)	(0.429)	(0.293)	(0.357)	(0.344)
Collaboration before	13.239^{***}	13.255^{***}	13.396^{***}	13.210^{***}	13.573^{***}	13.100^{***}	4.286^{***}	4.305^{***}	4.386^{***}	4.079^{***}	4.374^{***}	3.978^{***}
	(0.295)	(0.295)	(0.297)	(0.294)	(0.295)	(0.297)	(0.085)	(0.087)	(0.068)	(0.064)	(0.058)	(0.076)
log(Num prior collaborations)	-3.489^{***}	-3.347^{***}	-3.641^{***}	-3.349^{***}	-3.645^{***}	-3.469^{***}	-1.128^{***}	-1.025^{***}	-1.252^{***}	-1.040^{***}	-1.225^{***}	-1.122^{***}
	(0.089)	(0.088)	(0.091)	(0.088)	(0.089)	(0.090)	(0.028)	(0.027)	(0.026)	(0.021)	(0.025)	(0.028)
Genre similarity	0.331^{***}	0.348^{***}	0.410^{***}	0.127	0.253^{***}	0.403^{***}	0.155^{***}	0.155^{***}	0.196^{***}	0.018	0.118^{***}	0.216^{***}
	(0.075)	(0.073)	(0.080)	(0.071)	(0.073)	(0.083)	(0.043)	(0.038)	(0.039)	(0.029)	(0.031)	(0.040)
Content similarity	-1.558^{***}	-1.443^{***}	-1.028^{***}	-1.152^{***}	-1.150^{***}	-1.281^{***}	-0.620^{***}	-0.537^{***}	-0.254^{***}	-0.274^{***}	-0.305^{***}	-0.317^{***}
	(0.091)	(0.089)	(0.098)	(0.086)	(0.088)	(0.101)	(0.055)	(0.056)	(0.046)	(0.034)	(0.037)	(0.048)
Constant	-0.982^{***}	-0.572^{***}	-1.111^{***}	-0.552^{***}	-0.524^{***}	-1.237^{***}	-1.261^{***}	-1.008^{***}	-1.114^{***}	-0.587^{***}	-0.621^{***}	-1.249^{***}
	(0.103)	(0.095)	(0.105)	(0.093)	(0.093)	(0.112)	(0.127)	(0.104)	(0.067)	(0.047)	(0.093)	(0.092)
Log Likelihood	−70889.694	−70889.694	−70889.694	−70889.694	−70889.694	−70889.694						
Num. obs.	45285	45285	45285	45285	45285	45285	45285	45285	45285	45285	45285	45285

Notes: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

C.3 Direction regression of book performance

Table C11: Book performance

	r (log(Ratings count))	s (Average rating)
	(1)	(2)
Author ratings count percentile	4.199***	0.103***
	(0.095)	(0.015)
Author average rating	0.753***	0.600***
	(0.079)	(0.012)
Debut author	5.311***	2.338***
	(0.307)	(0.048)
Bestselling author	1.001***	0.030**
	(0.068)	(0.011)
log(Num prior books)	0.039	0.012**
	(0.025)	(0.004)
Publisher ratings count percentile	5.225***	−0.222***
	(0.167)	(0.026)
Publisher average rating	0.034	0.561***
	(0.205)	(0.032)
log(Capacity)	0.088***	−0.003
	(0.021)	(0.003)
Revenue	0.039**	−0.001
	(0.015)	(0.002)
Collaboration before	−0.279***	−0.010
	(0.066)	(0.010)
log(Num prior collaborations)	−0.050	0.019**
	(0.042)	(0.007)
Genre similarity	1.009***	−0.049***
	(0.064)	(0.010)
Content similarity	−0.062	0.031**
	(0.075)	(0.012)
Constant	−4.826***	−0.544***
	(0.845)	(0.132)
Year fixed effects	Yes	Yes
R^2	0.520	0.310

Notes: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Appendix D Estimation Details

D.1 Gibbs samplers

The prior distributions of parameters $f_0(\boldsymbol{\theta})$ and the augmented posterior are given in Section 4.3. The conditional distributions of the latent variables v_{ij} and parameters $\boldsymbol{\theta}$ are proportional to the parts that they enter in the augmented posterior in equation (21). For each variable, I collect terms and obtain a kernel that is in the same parametric family as their prior distributions.

Conditional distributions of v

For a pair ij, the conditional distribution of the latent variable v_{ij} is proportional to the product of the conditional density and the equilibrium condition. Let $\boldsymbol{v}_{-ij}$ denote the values of all other pairs in the market. Notice that $\boldsymbol{\mu}$ and $\boldsymbol{v}_{-ij}$ enter the density through the bounds $\overline{v}_{ij}$ or $\underline{v}_{ij}$ in equilibrium characterization.

If the pair is not matched, i.e., $\mu_{ij} = 0$, then the conditional distribution is

$$f(v_{ij}|\boldsymbol{\mu}, \boldsymbol{v}_{-ij}, X_{ij}, W_{ij}, \boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\Big(v_{ij} - X'_{ij}\beta\Big)^2\right) \times \mathbf{1}(v_{ij} < \overline{v}_{ij}). \tag{D.1}$$

This is a normal distribution $N(X'_{ij}\beta, 1)$ truncated above at $\overline{v}_{ij}$. Note that because the pair is not matched, no performance variable enters the density.

Conversely, if the pair is matched, i.e., $\mu_{ij} = 1$, the conditional density is more complicated because of the additional information from the performance variables. Completing the square with respect to v_{ij} yields the following density:

$$\begin{aligned} &f(v_{ij}|\boldsymbol{\mu}, \boldsymbol{v}_{-ij}, s_{ij}, r_{ij}, X_{ij}, W_{ij}, \boldsymbol{\theta}) \propto \\ &\quad \exp\left(-\frac{1}{2}(1 + \frac{\delta^2}{\sigma_1^2} + \frac{\omega^2}{\sigma_2^2})\Big(v_{ij} - X'_{ij}\beta - (\frac{\delta}{\sigma_1^2}(r_{ij} - W'_{ij}\gamma^r) + \frac{\omega}{\sigma_2^2}(s_{ij} - W'_{ij}\gamma^s))\big/(1 + \frac{\delta^2}{\sigma_1^2} + \frac{\omega^2}{\sigma_2^2})\Big)^2\right) \\ &\quad \times \mathbf{1}(v_{ij} \geq \underline{v}_{ij}). \end{aligned} \tag{D.2}$$

This is a truncated normal distribution $N(X'_{ij}\beta + (\frac{\delta}{\sigma_1^2}(r_{ij} - W'_{ij}\gamma^r) + \frac{\omega}{\sigma_2^2}(s_{ij} - W'_{ij}\gamma^s))\big/(1 + \frac{\delta^2}{\sigma_1^2} + \frac{\omega^2}{\sigma_2^2}), 1\big/(1 + \frac{\delta^2}{\sigma_1^2} + \frac{\omega^2}{\sigma_2^2}))$ truncated below at $\underline{v}_{ij}$.

Conditional distributions of parameters $\beta, \gamma^r, \gamma^s, \delta, \omega$

Let θ denote the parameter of interest. For each parameter θ, collecting terms involving θ in the augmented posterior (21) yields the following general form:

$$f(\theta|\boldsymbol{\mu}, \mathbf{v}, \mathbf{s}, \mathbf{r}, \mathbf{X}, \mathbf{W}, \boldsymbol{\theta}_{-\theta}) \propto \exp\Bigg(-\frac{1}{2}\Big(\theta' M_\theta \theta + 2\theta' N_\theta\Big)\Bigg), \tag{D.3}$$

where $\boldsymbol{\theta}_{-\theta}$ denotes all other parameters, M_θ is a symmetric matrix, and N_θ is a vector, both of dimensions compatible with the length of θ. Completing the square with respect to θ gives the normal distribution $N(-M_\theta^{-1}N_\theta, M_\theta^{-1})$. The parameters of the prior distributions $f_0(\theta)$ have been specified in the main text.

For β,

$$M_\beta = \Sigma_{\beta,0}^{-1} + \sum_m \Bigg[\sum_{ij} X_{ij}X_{ij}' + \sum_{\mu_{ij}=1} (\frac{\delta^2}{\sigma_1^2} + \frac{\omega^2}{\sigma_2^2}) X_{ij}X_{ij}'\Bigg], \tag{D.4}$$

$$N_\beta = -\Sigma_{\beta,0}^{-1}\beta_0 + \sum_m \Bigg[\sum_{ij} -X_{ij}v_{ij} + \sum_{\mu_{ij}=1} \frac{\delta}{\sigma_1^2} X_{ij}(r_{ij} - W_{ij}'\gamma^r - \delta v_{ij}) + \sum_{\mu_{ij}=1} \frac{\omega}{\sigma_2^2} X_{ij}(s_{ij} - W_{ij}'\gamma^s - \omega v_{ij})\Bigg]; \tag{D.5}$$

For γ^r,

$$M_{\gamma^r} = \Sigma_{\gamma^r,0}^{-1} + \sum_m \sum_{\mu_{ij}=1} \frac{1}{\sigma_1^2} W_{ij}W_{ij}', \tag{D.6}$$

$$N_{\gamma^r} = -\Sigma_{\gamma^r,0}^{-1}\gamma_0^r - \sum_m \sum_{\mu_{ij}=1} \frac{1}{\sigma_1^2} W_{ij}(r_{ij} - \delta(v_{ij} - X_{ij}'\beta)); \tag{D.7}$$

For γ^s,

$$M_{\gamma^s} = \Sigma_{\gamma^s,0}^{-1} + \sum_m \sum_{\mu_{ij}=1} \frac{1}{\sigma_2^2} W_{ij}W_{ij}', \tag{D.8}$$

$$N_{\gamma^s} = -\Sigma_{\gamma^s,0}^{-1}\gamma_0^s - \sum_m \sum_{\mu_{ij}=1} \frac{1}{\sigma_2^2} W_{ij}(s_{ij} - \omega(v_{ij} - X_{ij}'\beta)); \tag{D.9}$$

For δ,

$$M_\delta = \Sigma_{\delta,0}^{-1} + \sum_m \sum_{\mu_{ij}=1} \frac{1}{\sigma_1^2}(v_{ij} - X'_{ij}\beta)^2, \tag{D.10}$$

$$N_\delta = -\Sigma_{\delta,0}^{-1}\delta_0 - \sum_m \sum_{\mu_{ij}=1} \frac{1}{\sigma_1^2}(r_{ij} - W'_{ij}\gamma^r)(v_{ij} - X'_{ij}\beta); \tag{D.11}$$

And for ω,

$$M_\omega = \Sigma_{\omega,0}^{-1} + \sum_m \sum_{\mu_{ij}=1} \frac{1}{\sigma_2^2}(v_{ij} - X'_{ij}\beta)^2, \tag{D.12}$$

$$N_\omega = -\Sigma_{\omega,0}^{-1}\omega_0 - \sum_m \sum_{\mu_{ij}=1} \frac{1}{\sigma_2^2}(s_{ij} - W'_{ij}\gamma^s)(v_{ij} - X'_{ij}\beta). \tag{D.13}$$

Conditional distributions of parameters σ_1^2 and σ_2^2

The conditional distributions of σ_1^2 and σ_2^2 are both inverse gamma distributions with the following shape and scale parameters.

For σ_1^2,

$$\alpha_{\sigma_1^2} = \alpha_0 + \frac{1}{2}\sum_m |J_m|, \tag{D.14}$$

$$\beta_{\sigma_1^2} = \beta_0 + \frac{1}{2}\sum_m \sum_{\mu_{ij}=1} (r_{ij} - W'_{ij}\gamma^r - \delta(v_{ij} - X'_{ij}\beta))^2, \tag{D.15}$$

where $|J_m|$ is the number of workers in market m;

And for σ_2^2,

$$\alpha_{\sigma_2^2} = \alpha_0 + \frac{1}{2}\sum_m |J_m|, \tag{D.16}$$

$$\beta_{\sigma_2^2} = \beta_0 + \frac{1}{2}\sum_m \sum_{\mu_{ij}=1} (s_{ij} - W'_{ij}\gamma^s - \omega(v_{ij} - X'_{ij}\beta))^2. \tag{D.17}$$

D.2 Initial values of MCMC

To speed up convergence, I precompute the parameters $\beta, \gamma^r, \gamma^s$ with reduced form approaches ignoring the interdependence through the error terms. Specifically, I run regressions in equations (16) and (17) directly and obtain estimates for γ^r and γ^s.

For β, I adopt a two-step procedure. First, I use the semiparametric approach in Fox (2018) with the maximum score estimator. The score function of market m is similarly defined using the two-pair-no-exchange characterization as in equation (7):

$$S_m(\mu, X; \beta) = \sum_{\substack{\mu_{ij}=1 \\ \mu_{i'j'}=1 \\ i' \neq i}} \mathbf{1}\big(X'_{ij}\beta + X'_{i'j'}\beta > X'_{ij'}\beta + X'_{i'j}\beta\big). \tag{D.18}$$

In other words, the maximum score estimator maximizes the number of correct inequalities in the LP characterization. The total score function is $S = \sum_m S_m(\mu, X; \beta)$. For estimation, I use simulated annealing to obtain an estimate of β. Denote it $\hat{\beta}$.

Note that this approach is semiparametric and makes no assumption on the distribution of the error term. Because of this, β is only identified up to a scale. To make it compatible with the parametric specification in equation (14) where ε_{ij} is normally distributed with mean 0 and variance 1, we need to recover the variance of the error term σ^2_ε implied by the data and the estimate $\hat{\beta}$, and then deflate $\hat{\beta}$ by σ_ε.

To do so, in the second step, I parametrically estimate σ^2_ε in the value equation (14) given the estimated $\hat{\beta}$. The equilibrium condition requires that the error terms satisfy the following inequality:

$$-\varepsilon_{ij} - \varepsilon_{i'j'} + \varepsilon_{ij'} + \varepsilon_{i'j} < X'_{ij}\beta + X'_{i'j'}\beta - X'_{ij'}\beta - X'_{i'j}\beta \tag{D.19}$$

for all $\mu_{ij} = 1$, $\mu_{i'j'} = 1$, and $i \neq i'$. The left-hand side is a random variable with distribution $N(0, 4\sigma^2_\varepsilon)$ and the right-hand side can be estimated with $\hat{\beta}$ from the first step. The likelihood function of market m is

$$\mathcal{L}_m(\sigma^2_\varepsilon | \mu, X, \hat{\beta}) = \prod_{\substack{\mu_{ij}=1 \\ \mu_{i'j'}=1 \\ i' \neq i}} \Phi(X'_{ij}\hat{\beta} + X'_{i'j'}\hat{\beta} - X'_{ij'}\hat{\beta} - X'_{i'j}\hat{\beta}; 0, 4\sigma^2_\varepsilon), \tag{D.20}$$

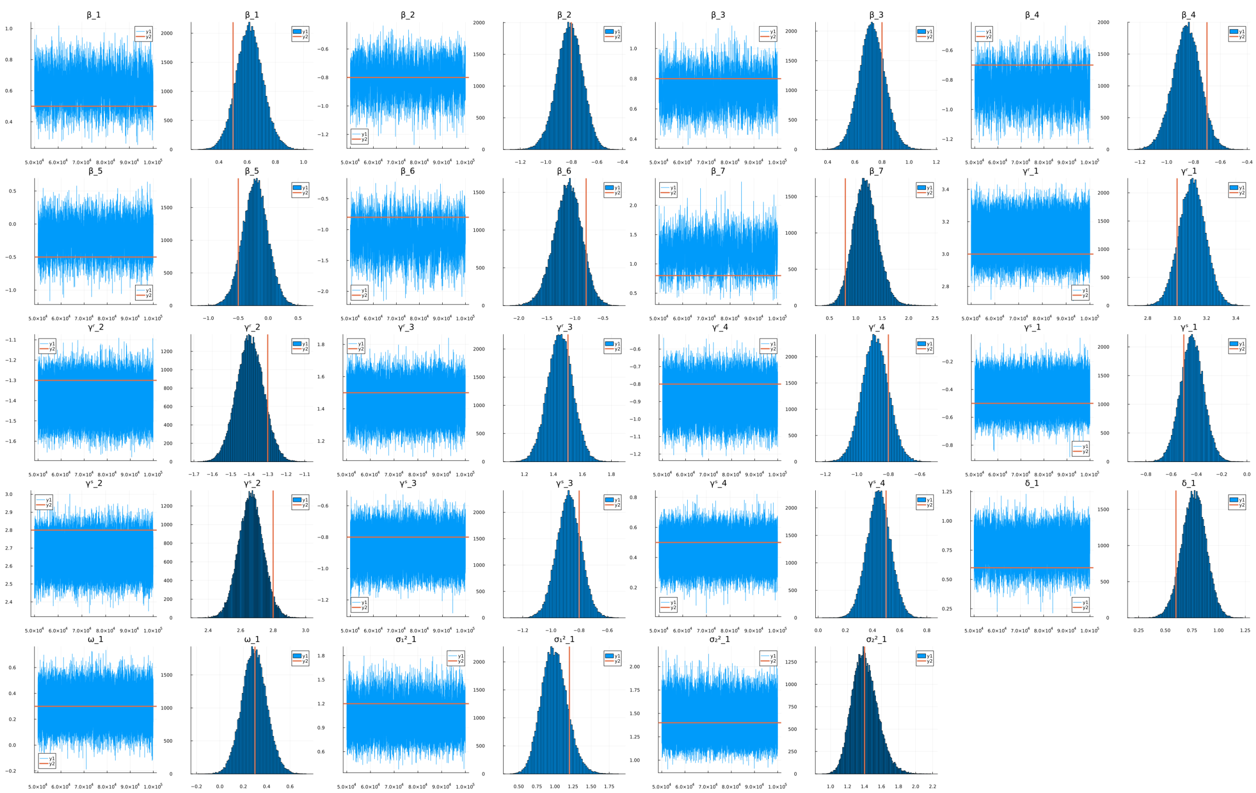
where $\Phi(\cdot; 0, 4\sigma^2_\varepsilon)$ is the CDF of the normal distribution $N(0, 4\sigma^2_\varepsilon)$. I then obtain an estimate of $\hat{\sigma}_\varepsilon$ by maximizing the likelihood function. The starting value in the MCMC is $\hat{\beta}/\hat{\sigma}_\varepsilon$.

Notice, however, that the estimate $\hat{\sigma}_\varepsilon$ is not unbiased because the sample is not independent. In particular, matched pairs ij and $i'j'$ are sampled repeatedly, but unmatched pairs ij' and $i'j$ are only sampled once. A correct likelihood function would have to simultaneously integrate out the joint distribution. For the purpose of generating an initial value for the MCMC, the bias can be safely ignored.

D.3 Estimation on generated dataset

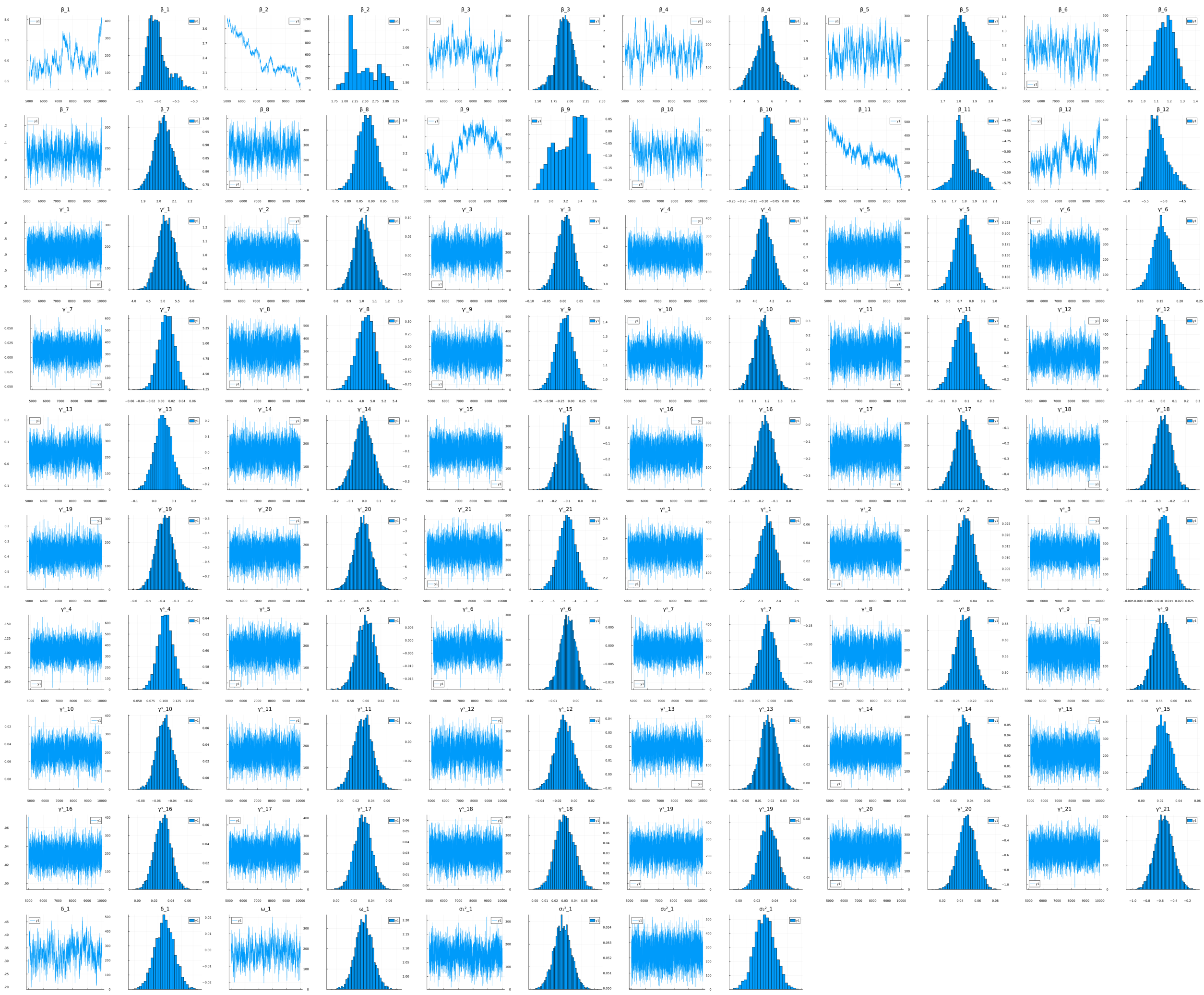
Figure D15 shows the estimation results on a dataset generated according to the structural model in Section 4. Both the trace and the posterior distribution of the MCMC are shown. The red line indicates the true value of the parameter.

Figure D15: Estimation results on a generated dataset



D.4 Details of estimation results

Figure D16: Estimation results



Appendix E Additional Counterfactual Simulations

Refer to Section 5 for the assumptions and implementation of the counterfactual simulations.

E.1 Counterfactual 2: organic merger

Table E12: Simulation results of organic merger

	Aggregate change			Average change per author		
	Joint surplus	Author	Publisher	Joint surplus	Author	Publisher
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Total social change						
Social	-22.67	-319.40	296.73			
Panel B: Publisher total change						
Hachette	-86.94	-82.45	-4.50			
HarperCollins	-3.04	19.16	-22.20			
Macmillan	-71.89	-63.31	-8.58			
Penguin Random House	38.46	-172.02	210.47			
Simon & Schuster	-3.58	-1.11	-2.46			
Panel C: PRH's internal change						
Penguin Random House	-62.91	-232.29	169.37	-0.030	-0.110	0.080
Panel D: Changes from sorting						
Hachette	-86.94	-81.29	-5.65	-0.977	-0.913	-0.064
HarperCollins	-3.04	9.37	-12.41	-0.041	0.127	-0.168
Macmillan	-71.89	-69.40	-2.49	-1.307	-1.262	-0.045
Penguin Random House	101.37	60.27	41.10	0.326	0.194	0.132
Simon & Schuster	-3.58	3.70	-7.28	-0.078	0.080	-0.158

Table E13: Simulation results of organic merger by author tenure

	Aggregate change			Average change per author		
	Joint surplus	Author	Publisher	Joint surplus	Author	Publisher
	(1)	(2)	(3)	(4)	(4)	(5)
			Panel A: PRH's internal change			
Best-selling	1.99	-21.89	23.88	0.010	-0.111	0.121
Mid-list	-11.36	-206.85	195.49	-0.008	-0.140	0.133
Debut	-53.54	-3.54	-50.00	-0.123	-0.008	-0.115
			Panel B: Changes from sorting			
Best-selling						
Hachette	-0.12	-0.41	0.29	-0.041	-0.137	0.096
HarperCollins	-0.61	-0.27	-0.35	-0.153	-0.066	-0.086
Macmillan	0.27	-0.00	0.28	0.274	-0.003	0.278
Penguin Random House	5.69	1.40	4.29	0.814	0.200	0.613
Mid-list						
Hachette	-1.91	-2.67	0.76	-0.040	-0.056	0.016
HarperCollins	-2.76	-1.13	-1.63	-0.099	-0.040	-0.058
Macmillan	3.35	-1.17	4.53	0.084	-0.029	0.113
Penguin Random House	61.93	40.84	21.09	0.350	0.231	0.119
Simon & Schuster	-0.74	-0.50	-0.23	-0.039	-0.027	-0.012
Debut						
Hachette	-8.47	-2.28	-6.19	-0.223	-0.060	-0.163
HarperCollins	-4.11	-2.61	-1.50	-0.098	-0.062	-0.036
Macmillan	1.64	-0.19	1.83	0.117	-0.014	0.131
Penguin Random House	-23.64	3.35	-26.99	-0.186	0.026	-0.213
Simon & Schuster	-1.67	-0.73	-0.94	-0.062	-0.027	-0.035

E.2 Counterfactual 3: Random House takeover

Table E14: Simulation results of Random House takeover

	Aggregate change			Average change per author		
	Joint surplus	Author	Publisher	Joint surplus	Author	Publisher
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A: Total social change			
Social	-102.72	-365.61	262.88			
			Panel B: Publisher total change			
Hachette	-10.61	-36.24	25.63			
HarperCollins	-76.90	-88.64	11.73			
Macmillan	-45.03	-50.02	5.00			
Penguin Random House	122.18	-46.49	168.67			
Simon & Schuster	24.76	13.13	11.62			
			Panel C: PRH's internal change			
Penguin Random House	-93.12	-258.53	165.41	-0.046	-0.128	0.082
			Panel D: Changes from sorting			
Hachette	-10.61	-9.60	-1.01	-0.114	-0.103	-0.011
HarperCollins	-76.90	-81.40	4.50	-0.487	-0.515	0.028
Macmillan	-45.03	-41.77	-3.25	-0.883	-0.819	-0.064
Penguin Random House	215.30	212.04	3.26	0.551	0.542	0.008
Simon & Schuster	24.76	24.89	-0.13	0.359	0.361	-0.002

Table E15: Simulation results of Random House takeover by author tenure

	Aggregate change			Average change per author		
	Joint surplus	Author	Publisher	Joint surplus	Author	Publisher
	(1)	(2)	(3)	(4)	(4)	(5)
			Panel A: PRH's internal change			
Best-selling	-11.98	-26.45	14.47	-0.062	-0.137	0.075
Mid-list	-45.22	-222.41	177.20	-0.031	-0.151	0.120
Debut	-35.92	-9.66	-26.26	-0.099	-0.027	-0.072
			Panel B: Changes from sorting			
Best-selling						
Hachette	-0.59	-0.84	0.25	-0.084	-0.119	0.035
HarperCollins	-2.25	-1.86	-0.39	-0.225	-0.186	-0.039
Penguin Random House	5.35	1.37	3.98	0.668	0.171	0.497
Simon & Schuster	0.16	-0.12	0.28	0.155	-0.122	0.277
Mid-list						
Hachette	-3.68	-3.77	0.08	-0.115	-0.118	0.003
HarperCollins	-12.42	-7.51	-4.91	-0.239	-0.145	-0.094
Macmillan	3.89	-1.90	5.79	0.114	-0.056	0.170
Penguin Random House	55.99	32.89	23.10	0.304	0.179	0.126
Simon & Schuster	-2.10	-1.49	-0.61	-0.105	-0.075	-0.030
Debut						
Hachette	-11.45	-4.64	-6.80	-0.212	-0.086	-0.126
HarperCollins	-24.38	-15.11	-9.27	-0.254	-0.157	-0.097
Macmillan	1.41	-0.96	2.37	0.083	-0.057	0.139
Penguin Random House	-34.68	1.99	-36.67	-0.174	0.010	-0.184
Simon & Schuster	-3.65	-2.66	-0.99	-0.076	-0.055	-0.021