Essays in Empirical Industrial Organization and Corporate Finance

Thesis by

Ke Shi

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2025
Defended May 12, 2025

Ke Shi
ORCID: 0000-0002-1090-5976

ACKNOWLEDGEMENTS

This thesis grew out of many conversations, questions, and quiet hours. Along the way, I have been guided, challenged, and encouraged by people whose generosity of mind and spirit has left a lasting imprint. To them, I owe more gratitude than these few pages can hold.

I have been especially fortunate to be mentored by Jean-Laurent Rosenthal, whose breadth of knowledge and penetrating insight reflect both historical depth and real-world acuity. His ability to draw connections across time, theory, and practice consistently pushed me to think more clearly and expansively. I am grateful not only for his intellectual range but also for his kindness, spirit, and sharp wit. He exemplifies the kind of scholar and mentor I hope to become.

Matthew Shum first introduced me to the field of industrial organization, and his influence is deeply woven into how I think as a scholar. His blend of technical rigor, intellectual curiosity, and instinct for asking meaningful questions helped shape my understanding of what research can be. Our conversations always left me with sharper ideas and, paradoxically, even better questions. It is a rhythm of inquiry I will carry with me.

Yi Xin has been a steady and thoughtful presence throughout my time at Caltech. I have learned a great deal from her precise and careful approach to research, from her example as a generous and engaged mentor, and from the determination that animates both. She brings a spirit of quiet confidence and collegiality to every interaction: open, encouraging, and constructive. Her willingness to share both insight and experience of academic life has helped me navigate both its challenges and its promise.

Michael Ewens has been a generous and insightful guide in all things venture capital and entrepreneurship. He brings deep institutional knowledge, empirical discipline, and a good sense of humor, along with a professionalism that wears its experience lightly. His feedback was consistently clear and incisive, always pushing toward higher standards while remaining firmly grounded in substance.

Caltech is a rare and wonderful place: rigorous, tightly knit, and intellectually alive in a way that fosters focus and depth. The faculty have shaped my development not only through their teaching and research, but also through the seriousness and clarity with which they pursue ideas. I am especially grateful to Philip T. Hoffman

for his constant encouragement, perceptive feedback, and the clarity he brings to both writing and economic thinking. I thank Charles Sprenger, Omer Tamuz, and Marina Agranov for their generous support and guidance on the professional side of academic life, and Kirby Nielsen and Pawel Janas for organizing seminars and offering thoughtful feedback. I am also grateful to Tracy Dennison, Jakša Cvitanić, Federico Echenique, Lawrence J. Jin, Andrew J. Sinclair, Robert Sherman, Gabriel Lopez-Moctezuma, Alexander V. Hirsch, Michael Gibilisco, Jonathan N. Katz, and many others. Each has shaped my thinking and supported my learning in meaningful ways.

I gratefully acknowledge financial support from the James and Karen Gerard Fellowship in Social Sciences, the Ronald and Maxine Linde Institute of Economic and Management Sciences Research Grant, and the Clarence J. Hicks Scholarship. I am also thankful to Laurel M. Auchampaugh for her reliable and attentive support throughout my time at Caltech.

My fellow graduate students and friends have been a constant source of camaraderie and good humor. Their presence made the long and sometimes peculiar path to becoming researchers more navigable, and the existential puzzles of academic life a bit easier to interpret. Shunto J. Kobayashi has been both a valued collaborator and a generous friend. Kexin Feng and Fan Wu have been steadfast and supportive friends, with whom I shared countless hours of work, banter, and culinary escapades. My cohortmates, Peter Doe, Claudia Kann, and Polina Detkova, were essential companions in navigating both the curriculum and the broader demands of academic life. I also thank Aldo Lucia, Wanying (Kate) Huang, Danny Ebanks, Po-Hsuan Lin, Meng-Jhang (MJ) Fong, Zhuofang Li, Jack Adeney, Zhenlin Kang, Tharani Weerasooriya, Matthew Estes, Mitchell Linegar, Shiyu (Jake) Zhang, Jiatong Han, and many others.

My old friends from high school and college, despite the years and distance, have offered constancy and glimpses of an earlier self. I am grateful for their enduring presence.

Above all, I thank my parents for their love and strength.

ABSTRACT

This thesis consists of three chapters.

Chapter 1 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 complementarities. Applying this model to a major merger in the U.S. 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. While the merger alleviated capacity constraints, post-merger integration led to significant creative misalignment between authors and publishers. The merger also induced 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 in consumer markets. This research extends merger evaluation beyond consumer impact, offering a framework to analyze the broader consequences of mergers in labor markets characterized by worker-firm complementarities.

Chapter 2, coauthored with Miguel Alcobendas, Shunto J. Kobayashi, and Matthew Shum, studies the impact of online privacy protection, which has gained momentum in recent years and spurred both government regulations and private-sector initiatives. A centerpiece of this movement is the removal of third-party cookies, which are widely employed to track online user behavior and implement targeted ads, from web browsers. Using banner ad auction data from Yahoo, we study the effect of a third-party cookie ban on the online advertising market. We first document stylized facts about the value of third-party cookies to advertisers. Adopting a structural approach to recover advertisers' valuations from their bids in these auctions, we simulate a few counterfactual scenarios to quantify the impact of Google's plan to phase out third-party cookies from Chrome, its market-leading browser. Our counterfactual analysis suggests that an outright ban would reduce publisher revenue by 54% and advertiser surplus by 40%. The introduction of alternative tracking technologies under Google's Privacy Sandbox initiative would partially offset these losses. In either case, we find that big tech firms can leverage their informational advantage over their competitors and gain a larger surplus from the ban.

Chapter 3 examines how informal and formal networks shape performance in the venture capital (VC) industry. Using data on all U.S.-based VC investments from 1990 to 2009, supplemented with partner-level educational and employment histories from LinkedIn, I develop a structural framework that connects three types of networks: coinvestment ties, historical affiliations, and latent social connections. In the baseline model, VC performance is a function of peer performance, capturing network spillovers through a micro-founded production function. To address endogeneity in network formation, I extend the model using a two-step instrumental variables strategy that leverages variation in past professional and alumni ties. Finally, I introduce endogenous network formation where VCs strategically choose connections based on expected peer quality, allowing for the recovery of latent social networks from equilibrium outcomes. Across specifications, better-connected VCs exhibit significantly higher exit rates. Estimates from the endogenous model suggest that a 1% increase in social connectedness raises a VC's exit rate by 0.2 percentage points, while a 1% improvement in peer performance leads to a 0.74 percentage point increase in connection intensity. Informal relationships thus carry measurable economic weight, and the empirical approach developed here provides a new lens for identifying network effects in private capital markets.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

MERGERS AND MISMATCHES IN THE LABOR MARKET FOR CREATIVITY

1.1 Introduction

Mergers and acquisitions can reshape not only market structure but also the allocation of talent across firms (Naidu, Posner, and Weyl, 2018; Shapiro, 2019; Posner, 2021; Azar and Marinescu, 2024). In labor markets characterized by complex production relationships between firms and workers, consolidation can significantly influence who works where, how well workers and firms match, and how value is distributed. These labor-side consequences often hinge on factors beyond wages. In many settings, especially those involving high-skilled or creative work, successful employment depends on compatibility between workers and firms in terms of expertise, preferences, and collaborative potential. This two-sided matching process plays a critical role in determining productivity and job satisfaction. As consolidation alters the structure and composition of employers, it can disrupt existing matches, redirect talent, and generate uneven effects on worker outcomes. These dynamics raise important questions: How does ownership consolidation affect the quality of worker-firm matches? What are the implications for the distribution of opportunities and value across different types of workers? And how might such reallocations ripple through to the quality of output in affected industries?[1]

This paper addresses these questions by introducing an empirical framework designed to evaluate the consequences of mergers for labor market matching. The central contribution is to quantify both the efficiency and redistributive consequences of consolidation using a two-sided matching model with transferable utility. 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

[1]This paper focuses on the quality of matches and the distribution of opportunities and value across different types of workers. In the policy domain, heightened attention to labor market consolidation culminated in the U.S. Department of Justice (DOJ) and Federal Trade Commission's (FTC) release of the *2023 Merger Guidelines*, which outline how reduced competition can suppress wages, erode working conditions, and lower workplace quality. Recent empirical studies that engage more directly with antitrust concerns are discussed in the literature review below.

depends critically on the *complementarity* (or *compatibility*) between 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 framework distinguishes between two key channels through which mergers affect labor markets: a direct effect on compatibility and value creation, and an equilibrium effect that reshapes matching patterns across the market. While mergers may relax capacity constraints and enhance efficiency, the resulting reallocation of human capital can produce substantial welfare redistribution. Value flows disproportionately to the merged firm, while reduced employer competition can adversely affect workers' outcomes.

The empirical setting is the U.S. trade publishing industry, with a focus on the 2013 merger between Penguin and Random House, two of the largest publishers at the time. Publishing is an attractive empirical setting for several reasons. First, the relevant labor market is narrowly defined by the task of book writing, with limited overlap with other labor markets. This segmentation allows for a well-identified set of workers (authors) and employers (publishers). Second, the industry generates rich, observable data at the individual book level—information that is typically unavailable in other industries—which supports a fine-grained 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, successful book production depends heavily on intellectual and creative compatibility between authors and publishers, making publishing an especially suitable setting for studying labor matching in contexts where compatibility is central to value creation. Finally, the industry's high concentration, dominated by only a few major publishing companies, underscores the relevance of studying how mergers influence the allocation of creative talent. This relevance is exemplified by the recent DOJ action that blocked a merger attempt between Penguin Random House and Simon & Schuster, citing concerns over diminished competition for authors.[2]

The empirical analysis begins by documenting stylized facts that reveal strong patterns of assortative matching in the publishing market. Using constructed measures of compatibility based on author and publisher experience, genre focus, and literary style, the evidence shows that authors tend to match with publishers who share similar characteristics. Popular and high-quality authors, as measured by their publication histories, are more likely to partner with publishers of comparable stand-

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

ing. This pattern extends to genre and content preferences: authors and publishers demonstrate clear alignment in literary style and subject matter expertise.

The focus of the paper is on the redistributive consequences of the merger, with particular emphasis on how it reshapes the allocation of creative talent across firms. Because comprehensive data are available only for books published prior to the merger, and because equilibrium responses must be taken into account, a structural approach is adopted to estimate author-publisher match values and simulate merger outcomes through 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 utility created by the partnership. The market equilibrium is cleared through a transfer (typically from the publisher to the author), though this transfer mechanism is not modeled explicitly. Instead, the analysis focuses on the division of *post-transfer surplus* between the two sides of the market. To evaluate the merger's broader consequences, the model is extended to include downstream outcomes in the product market, measured by reader reception, which sheds light on consumer welfare. This component follows the logic of a selection model, in which only a subset of potential book projects materializes and becomes observable.

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, the analysis employs the partial equilibrium characterization proposed by Fox (2018), which significantly speeds up computation. Second, from an econometric point of view, the performance variables contain additional information on the match values and must be factored into the estimation, making it infeasible to directly apply the semiparametric approach in Fox (2018). To reconcile this, a parametric structure is imposed to link match formation with performance, enabling a likelihood-based estimation strategy. A third challenge arises due to the high dimensionality of the likelihood, which renders standard estimation techniques impractical. To overcome this, a Bayesian approach is adopted, extending the method of Sørensen (2007) from the non-transferable to the transferable utility framework.

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 values, as does prior

collaboration history. These patterns 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 just 15% under random assignment. In terms of book performance, an author's pre-existing success (measured by reader ratings and review counts) is the strongest predictor of future outcomes. By contrast, while editorial compatibility measures strongly affect initial matching decisions, they have a limited direct impact on book performance once selection effects are accounted for.

Building on the estimated match values and structural parameters, a series of counterfactual simulations is conducted to assess how the merger reshapes the allocation of author-publisher matches across the market. The merger is modeled as a complete integration of two previously independent companies, requiring new match values for the consolidated entity to replace those of its previously separate components. To capture a range of plausible post-merger dynamics, I consider three counterfactual scenarios: (1) synergistic integration, in which match values reflect the maximum of the two pre-merger values; (2) organic merger, where values are computed as a weighted average of the pre-merger values; and (3) Random House takeover, where the acquiring firm's values prevail.

The simulation results reveal divergent outcomes across post-merger scenarios and highlight trade-offs introduced by integration. Under synergistic integration, there is an efficiency improvement stemming from the merged entity's enhanced capacity to optimize author-publisher matches, a capability previously constrained when the companies operated independently. By contrast, 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. Prior to the merger, each publisher maintained a distinct editorial identity and allocated resources to specific author segments. The integration of operations limits this specialization, reducing the firm's ability to sustain compatibility across a diverse author pool and thereby diminishing the overall quality of matches.

The efficiency effects of the merger, if any, are distributed unevenly across market participants, with two notable distributional consequences. First, there is a substantial transfer of value from competing publishers to Penguin Random House, reflecting a shift in the allocation of high-value matches toward Penguin Random House. Second, profit gains for publishers are accompanied by welfare losses for

authors, raising concerns about the broader consequences of consolidation in this labor market. Authors' welfare losses emerge through two distinct mechanisms. The direct effect arises from reduced competition between the formerly separate companies, which puts downward pressure on author compensation, particularly for those previously contracted with either Penguin or Random House. The indirect effect operates through equilibrium sorting, leading to a redistribution of value among authors: while those selected by the merged entity may benefit, authors displaced to other publishers experience welfare losses. Market concentration in publishing, therefore, could simultaneously enhance efficiency and exacerbate inequality.

The distributional impact of the merger varies substantially across authors at different market positions. Industry debate has centered on which author segments would be most adversely affected. One view holds that debut and mid-list authors would be disadvantaged, as the merged firm might concentrate resources on high-profile titles. An alternative view, endorsed by the DOJ during its 2022 merger challenge, suggests that bestselling authors would be most negatively affected. The simulation results indicate that the direction and magnitude of welfare changes depend heavily on authors' post-merger movement between publishers. Among those who remain with Penguin Random House, welfare declines are observed across the board, with bestselling authors experiencing the largest losses, consistent with the DOJ's position. The outcomes differ, however, for authors who switch publishers. Debut and mid-list authors gain more than their bestselling counterparts when moving into Penguin Random House, whereas authors who exit the firm, particularly bestsellers, face pronounced welfare declines.

The consumer side of the market appears largely unaffected by the merger along observable dimensions of product quality. Books published by the merged firm exhibit no significant changes in average ratings or rating volumes, indicating that reader engagement and perceived quality remained stable following consolidation. This outcome is consistent 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 a noticeable decline in their book consumption experience. Therefore, relying solely on consumer-side metrics provides an incomplete view of merger impact and risks overlooking significant anticompetitive effects elsewhere, particularly in labor markets.

Related Literature

This paper engages with several strands of literature. First, it adds to the body of work on the impact of mergers (Asker and Nocke, 2021). While prior work has primarily focused 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. Building on 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, whereas most studies examine post-merger repositioning in product offerings or firm conduct (e.g., Fan (2013), Li et al. (2022), and Wollmann (2018)), this paper highlights the equilibrium consequences of re-sorting in labor markets, which arise from changes in firm-worker matching patterns. Third, this paper contributes to the literature on the impact of mergers on innovation, but through the lens of upstream labor inputs and their downstream effects. Past work, such as Igami and Uetake (2020) and Bonaimé and Wang (2024), has focused on firm-level innovation decisions rather than upstream labor inputs.

An emerging literature, in parallel with 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 body of work has devoted significant attention to explaining and estimating wage markdowns. Theoretical work follows three main 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 more general framework by conceptualizing employment as the joint production of value. It is among the first studies to structurally model and quantify the direct impact of market consolidation in labor markets at the micro level.

This paper also 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 research has examined the impact of intellectual property protection,

[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 contributions 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.

such as copyrights and patents, on creative and innovative work (Biasi and Moser, 2021; Giorcelli and Moser, 2020; Peukert and Reimers, 2022), as well as the effects of digitization in publishing (Reimers and Waldfogel, 2021; Peukert and Reimers, 2022; Nagaraj and Reimers, 2023). However, the role of market structure, as well as how changes in it affect creative labor, has received far less attention. This paper addresses that gap by offering a new empirical framework that conceptualizes the production of creative output from the matching between authors and publishers, a mechanism that exemplifies production in many high-skilled labor markets.

In terms of empirical methodology, this paper contributes to the study of matching markets with transferable utilities, with new emphasis on its implications for market structure and competition. This study builds on the theoretical foundations laid by seminal works such as Shapley and Shubik (1971), Becker (1973), Kelso and Crawford (1982), Roth (1984), and Sotomayor (1999), along with more recent advances by Azevedo and Hatfield (2018), among others.[4] While existing empirical applications typically focus on sorting patterns between the two sides of the market, this paper instead examines how matching frameworks can inform merger analysis and the distortions introduced by consolidation (Dupuy et al., 2017).[5] A main innovation of this paper is the full-fledged agent-level matching model with transferable utility and observed performance.[6] In contrast, most prior work aggregates individuals by coarse characteristics and estimates a two-sided random utility model (Choo and Siow, 2006). The use of observed match performance introduces a selection dimension that adds complexity to the model, akin to issues in sample selection. To address the resulting computational burden, this paper extends the Bayesian estimation technique in Sørensen (2007) to the transferable utility setting and adopts the semiparametric characterization in Fox (2010) and Fox (2018). Furthermore, the empirical framework recovers the post-transfer division of surplus based on the equilibrium characterization, allowing for welfare analysis on both sides of the market. Prior studies generally focus only on identifying the total joint surplus, without examining its distribution.[7]

[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). 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 closely related to the matching framework. The first builds on matching theory to analyze sorting in labor markets, e.g., Eeckhout and Kircher (2011), Eeckhout (2018), and Eeckhout and Kircher (2018), and aligns with this paper in emphasizing worker-

1.2 The Publishing Industry and Data Description

Trade publishing

Trade publishing refers to the sector of the publishing industry that produces books for general readership, sold through bookstores, retail outlets, and online sellers (Thompson, 2012).[8] The U.S. trade publishing industry is highly concentrated. Prior to 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, which further consolidated the market into what is now known as the "Big Five." Penguin Random House (PRH) became and remains the world's largest publisher. Together, the Big Five accounted for nearly 60 percent of the U.S. trade book sales market in 2021, and 91 percent of the market for publishing rights to "anticipated top sellers."[9] While growing industry concentration has often been justified on the grounds of economies of scale in terms of cost savings and enhanced bargaining power with downstream distributors, there are longstanding concerns about its competitive effects on authors. When Penguin Random House proposed acquiring Simon & Schuster in 2022, the merger was challenged and ultimately blocked on the basis that it would harm competition in the market for publishing rights, i.e., the labor market for authors.

Unlike other input markets, the labor market stands out due to the presence of match-specific preferences on both sides, extending beyond profit, wages, or non-pecuniary benefits. Both parties may value attributes unique to the relationship itself. For example, both publishers and authors may derive match-specific utility based on their shared interests, beliefs, or values, etc. This is especially true in the publishing industry, where the editorial match between authors and publishers (or editors) is a central priority for both sides. After acquiring a manuscript and before production-related services such as design, printing, and marketing, authors work closely with editors in a creative process to shape the final product. Publishers care deeply about

firm compatibility. The second concerns hedonic wage theory and workplace amenities, grounded in the theory of compensating differentials within a competitive equilibrium framework (Rosen, 1986; Hwang, Mortensen, and Reed, 1998; Manning, 2003; Card et al., 2018). Chiappori, McCann, and Nesheim (2010) formalizes 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), which emphasize the wage effects.

[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.

whether a manuscript aligns with their mission and editorial vision, while authors seek editors who understand and support their work. Although author compensation was the main focus in the merger case, it was repeatedly emphasized that authors value "editorial match, a feel the editor and [publishing] house understand[s] what they are writing." They want to collaborate 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."

Data and variables

The primary data for this study are drawn from Goodreads, a community-based online platform for book ratings, reviews, and social networking. The dataset was collected by Wan and McAuley (2018) and Wan et al. (2019) in late 2017.[10] The authors scraped users' "public shelves," virtual lists of books organized by themes and accessible without registration. The full dataset contains nearly 2.3 million books. Each book is associated with its author(s), publisher, and publication date. Additional fields include user-generated ratings and reviews, shelf labels, and brief textual descriptions.

This study focuses on a subsample of titles published between 2010 and 2016 (the last year of complete data), restricted to those with complete information on authorship, publisher, and publication year and month.[11] Reprints or new editions of existing titles are excluded, as they do not represent a new matching process between the author and the publisher. For convenience, each individual book (rather than each author) is treated as the unit of observation. In what follows, I use the terms "author" and "book" interchangeably to refer to the author side of the market. The final sample consists of over 140,000 books. Table 1.1 presents summary statistics of the books for the dataset.

On the publisher side, the analysis distinguishes among three categories: the Big Six (Penguin, Random House, Simon & Schuster, Hachette, HarperCollins, and Macmillan), a collection of notable, smaller houses grouped as "fringe publishers," and self-publishing.[12] Fringe publishers include several influential names

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

[11] Although records extend through 2017, books released closer to the collection date may have accumulated fewer ratings and reviews, introducing noise into performance metrics. Furthermore, while the merger between Penguin and Random House was completed in 2013, its effects likely unfolded gradually. To avoid confounding due to timing, the main estimation sample includes only books published between 2010 and 2013, prior to the merger. Counterfactual simulations are used to analyze post-merger effects.

[12] Thompson (2012) notes that the publishing industry is characterized by a peculiar market

such as Scholastic, Houghton Mifflin Harcourt, and Bloomsbury, among others. Self-publishing is treated as an outside option in the analysis. Because large publishing corporations often exhibit internal heterogeneity in content specialization, the analysis is disaggregated across ten genre categories to account for internal heterogeneity.[13]

The dataset contains only *observed matches* that are the equilibrium outcomes of the matching process. However, a full analysis requires information on all *potential matches* (hereafter also referred to as *"pairs"*) in the market. To account for these counterfactual matches, the data are augmented by constructing the Cartesian product of authors with published books in a given half-year and all publishers active during that period. In other words, the *matching market* is defined at the semiannual level, where authors with books released within the same half-year window are treated as a cohort facing a common set of potential publishers.[14] This semiannual structure reflects the seasonality of the publishing industry, which revolves around two main cycles: spring and fall.

Reader reception and book performance. Each book is associated with two key reader-side metrics: the *rating count*, measuring the number of user ratings, and the *average rating*, calculated across all editions up to the data collection date. The rating count serves as a proxy for a book's popularity, while the average rating reflects perceived quality (Cabral, 2012; Goldfarb and Tucker, 2019).[15] I take no normative stance on the value of a book and assume that popularity and quality, as reflected in these measures, capture reader utility. Because the distribution of rating

structure: a handful of dominant publishing firms and numerous small independent houses. Medium-sized publishers are rare and can therefore be reasonably excluded.

[13]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 often referred to as "genre fiction," popular styles typically treated as distinct from general literary fiction. Books may belong to multiple categories; the analysis considers each book's top two. Ideally, one would observe matching at the author-editor level and aggregate editors to their respective publishing houses. Because such information is not systematically available, publisher-genre pairings serve as a proxy for editorial experience within content areas.

[14]Although the data include publication dates, they do not record when contracts were signed. Nonetheless, because the publishing industry operates on a seasonal production calendar, it is reasonable to assume that books published within the same year were contracted around the same time.

[15]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 1.1: 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 based on all previously published books. For books with multiple authors, characteristics are averaged across co-authors. Publisher characteristics are aggregated from the same half-year in the prior year.

count is highly right-skewed, a logarithmic transformation is applied to reduce the influence of extreme values. Ratings are integer scores from 1 to 5, so the average rating falls within the range $[1, 5]$. This distribution is left-skewed: 1's and 5's are relatively rare and tend to occur in books with very few ratings. To address noise among low-count observations, a Bayesian smoothing method is used to compute

an adjusted average rating that incorporates population-level information.[16] The adjusted average rating centers slightly above 3.9.

Pre-match experience, expertise, and interaction. For each book, author-side covariates are constructed from their prior publication history. Two binary indicators capture the author's experience: *debut author* denotes a first-time author (approximately 40 percent of books), while *bestselling author* identifies those in the top 5 percent of cumulative rating counts, a measure consistent with industry estimates of concentration in profitability.[17] Authors who fall outside these groups are classified as *mid-list authors*, a publishing term referring to writers whose work is moderately successful but not blockbuster-level. The *author rating count percentile* and *author average rating* are proxies of popularity and quality, constructed based on the rating count and the average rating of all previous books.[18]

On the publisher side, the *share of debut authors* and *share of bestselling authors* are the corresponding measures of publishers' risk preferences, prioritization of commercial success versus literary exploration, and overall abilities to attract authors in either category.[19] In parallel, the *publisher rating count percentile* and the *publisher average rating* capture the publisher's recent performance in terms of popularity and perceived quality. All publisher variables are aggregated and averaged at the publisher-half-year-genre level. A key difference is that, whereas author-level measures reflect an author's full publication history, publisher-level variables are aggregated within a single half-year and genre. The difference reflects an underlying asymmetry in evaluation: while authors are typically assessed based on long-term reputations, current publisher performance is more likely to influence matching decisions in a given period.

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

$$BAR_i = \frac{AR_i \times RC_i + \overline{AR}_{pop} \times \overline{RC}_{pop}}{RC_i + \overline{RC}_{pop}}, \tag{1.1}$$

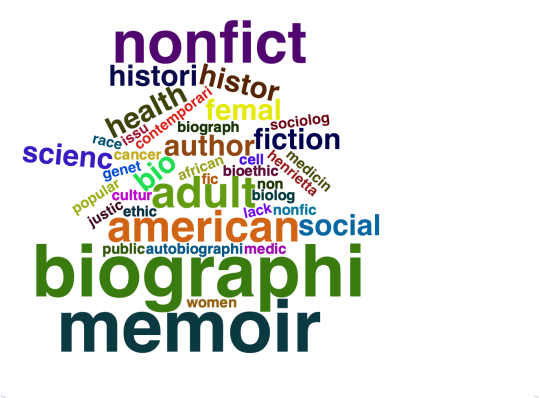
where AR is the average rating, RC is the rating count, and the overline indicates the population average. The population is defined at the half-year-genre level.

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

[18]Pre-match variables are constructed from all books published after 2000, ten years before the sample period. For the rating count, because books published in earlier dates tend to have accumulated more ratings, the temporal bias in the rating count is corrected by calculating each book's rating count percentile relative to books published in the same half-year. The variable is then computed as the cumulative average of these percentiles.

[19]These variables reflect outcomes of the matching equilibrium and should not be interpreted as exogenous traits. The analysis assumes that authors behave as price-takers, treating publisher characteristics as given when forming their preferences.

Figure 1.1: Example text: *The Immortal Life of Henrietta Lacks*, Rebecca Skloot, Pan Macmillan, 2010



(a) Genre (shelf labels)

(b) Content (book description)

Relational persistence is captured through two variables: *collaboration before*, indicating whether the author and publisher have worked together previously, and *number of collaborations*, reflecting the total count of prior joint publications. These variables describe the extent of prior interaction between an author and a publisher and serve as inputs for evaluating the role of match-specific history in the matching process.

Editorial compatibility. A distinctive feature of the publishing industry is the importance of editorial alignment between authors and publishers. To quantify this dimension, two variables are constructed from the text data associated with books: *genre similarity* and *content similarity*. The *genre* of a book is derived from user-generated shelf labels on Goodreads, which serve as crowd-sourced metadata reflecting genre, style, and thematic categories. The *content* of a book is based on the corpus of book descriptions (introductions) that summarize key narrative elements and thematic content. For illustration, Figure 1.1 presents word clouds for the 2010 bestseller *The Immortal Life of Henrietta Lacks* by science writer Rebecca Skloot. Note that words have been preprocessed and only word stems are displayed. Panel (a), based on shelf labels, shows the book associated with themes such as "biography," "nonfiction," "science," and "ethics." Panel (b), drawn from the book's description, indicates a storyline involving "cell," "immortal," "clone," and "research." Section 1.B provides additional examples of bookshelf labels and descriptions from books in the sample period.

To extract structured measures of editorial compatibility from the text data, latent

Figure 1.2: Example topics

(a) Example of a genre topic

(b) Example of a content topic

Dirichlet allocation (LDA), a widely used topic modeling technique, is applied for dimension reduction.[20] The LDA model is trained on a subsample of 6,000 books, with the number of topics K set to 50 for both the shelf label (genre) and book description (content) corpora. Each topic is estimated as a distribution over words in the corpus, capturing a distinct dimension of genre or thematic content. The resulting topic vectors summarize each book as a distribution over topics. Figure 1.2 shows example topics generated by the LDA model. Panel (a) shows an example genre topic with high probabilities on terms such as "religion," "philosophy," "science," and "psychology." Panel (b) shows an example content topic characterized by terms such as "parent," "disease," "help," and "medicine." Further examples can be found Section 1.B.

After estimating the topic-word distributions, each book is represented by two 50-dimensional vectors summarizing its genre and content profiles. These vectors correspond to the posterior distributions over the 50 topics, recovered separately for the shelf label and description corpora. To construct analogous measures for publishers, all books released by a publisher within a given genre-half-year combination are aggregated into a single document. Topic distributions are then estimated for this composite text, yielding genre and content vectors for each publisher. Editorial

[20] Topic modeling assumes that each text (document) from a corpus is generated from some K underlying "topics." Each topic is a probability distribution over the vocabulary present in the entire corpus. A document is then represented by a K-dimensional distribution over these topics. Topic modeling thus reduces the dimensionality from the vocabulary size to K topics. For methodological overviews, see Gentzkow, Kelly, and Taddy (2019) and Ash and Hansen (2023). For recent applications, see Hansen, McMahon, and Prat (2018), Bandiera et al. (2020), Djourelova, Durante, and Martin (2024), and Ash, Morelli, and Vannoni (2022).

compatibility between a book and a publisher is quantified using cosine similarity, a standard measure of vector alignment in document space.[21] A value of 1 indicates complete alignment, while 0 indicates orthogonality. This procedure produces two measures of editorial compatibility, *genre similarity* and *content similarity*, for each book-publisher pair.

1.3 Descriptive Evidence

Assortative matching

Matching markets are often characterized by positive assortative matching (Becker, 1973). In the context of publishing, this pattern can be assessed by examining whether authors and publishers align on observable dimensions such as experience, popularity, and quality. Table 1.2 presents regressions of an author's characteristics on those of the publisher with whom they are matched:

$$X_{ij}^{a} = \beta_0 + X_{ij}^{p\prime}\beta_1 + \varepsilon_{ij}, \tag{1.2}$$

where the unit of observation ij is a matched pair, X_{ij}^{a} denotes author characteristics, and X_{ij}^{p} denotes publisher characteristics. That is, conditional on a match, the regression estimates how publisher characteristics predict the characteristics of their matched authors.

There is a significant degree of positive assortative matching between measures of an author's experience and publishers' expertise. The coefficients on the diagonal entries in the table—linking similar characteristics on both sides—demonstrate that authors and publishers tend to match along the same dimensions. Authors with greater popularity (as measured by rating count percentile) and higher quality (average rating) are more likely to be matched with publishers of comparable strength. In addition, publishers' risk preferences, proxied by the historical shares of debut and bestselling authors in their catalogs, are positively correlated with corresponding characteristics on the author side. In contrast, publisher capacity and revenue are not strong predictors of author characteristics.

Second, there is also evidence of assortative matching along editorial compatibility, as measured by both genre and content similarity. Genre and content are represented as 50-dimensional vectors of topic weights for both books and publishers. Figure 1.3 plots the correlation matrices between the topic distributions of books and

[21] 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: $\frac{x \cdot y}{\|x\|\|y\|}$. For related applications, see Kelly et al. (2021), Cagé, Hervé, and Viaud (2020), and Bertrand et al. (2021).

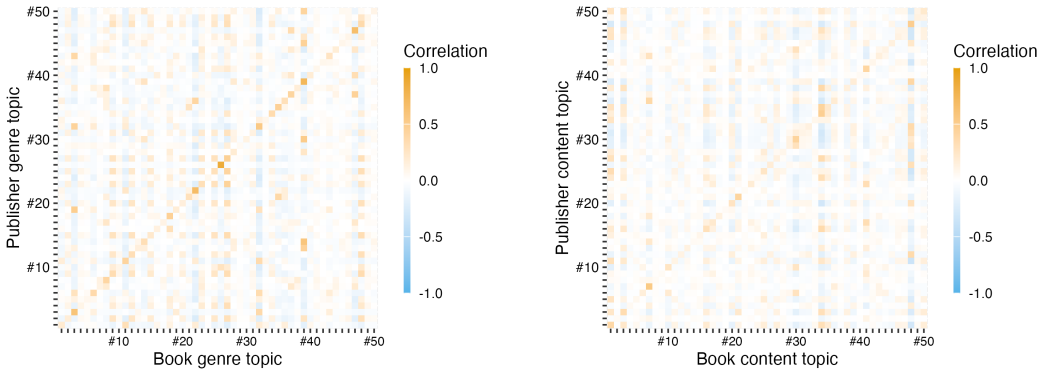
Table 1.2: Assortative matching

	log(Author ratings count per-centile)	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: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

those of their matched publishers. Panel (a) shows the correlation matrix for genre topic weights, and panel (b) for content topic weights. Color gradation indicates correlation magnitude, ranging from -1 to 1. The diagonal entries represent correlations for corresponding topics on the author and publisher sides. In the absence of editorial sorting, no systematic correlation would be expected. However, several patterns emerge. First, there is a positive correlation along the diagonals: if a book places high weight on certain topics, its matched publisher also tends to emphasize those same topics. Second, some book topics are negatively correlated with certain publisher topics, suggesting that authors and publishers in specific genres or content areas tend not to match across categories. Third, the correlations are stronger for

Figure 1.3: Topic correlation between the book and the publisher



(a) Genre (b) Content

Notes: This figure presents the correlation matrix between the topic distributions of books and their matched publishers. The horizontal axis lists the 50 book topics; the vertical axis lists the corresponding publisher topics. Yellow represents positive correlation and blue represents negative correlation. Details of the topic definitions are provided in Section 1.B.

genre than for content. This is expected, as genre topics are typically more clearly defined than content topics.

1.4 Structural Model and Estimation

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: firms $i \in I$ and workers $j \in J$. Firms can hire multiple workers, while each worker may be matched to only 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, including the outside option. Following the convention in the matching literature, the environment is full-information and frictionless in which all authors and publishers are potential match partners.[22] 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.[23] Note that $\mu_{0j} = 1$

[22]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.

[23]With a slight abuse of notation, I use the shorthand $\mu_{ij} = 1$ to denote the set of matched pairs $\{ij \in \tilde{I}\times J|\mu_{ij} = 1\}$, and similarly $\mu_{ij} = 0$ for unmatched pairs, in the indices of summation, product, maximum, and minimum operators.

means that worker j is unmatched. Finally, the model assumes that a firm's outside option of leaving positions unfilled carries an arbitrarily small utility, and that the number of workers 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. Under the assumption that $f(C)$ is additively separable across workers, the match-specific profit from a firm-worker pair ij is

$$\pi_{ij} = f_{ij} - t_{ij}, \tag{1.3}$$

where f_{ij} represents the output produced by firm i in collaboration with worker j.[24] Crucially, f_{ij} encompasses the total value generated in the match from the firm's perspective, including factors beyond immediate revenue or profit from production. For example, a firm may value the worker's alignment with its values, reputation, or long-term strategic fit, even if these attributes do not directly affect short-term profits.

Worker j's utility from working for firm i with a transfer t_{ij} is denoted by $u_j(i; t_{ij})$. Consistent with standard assumptions, u_j is additively separable in two components:

$$u_{ij} = a_{ij} + t_{ij}, \tag{1.4}$$

where a_{ij} captures the match-specific utility that worker j derives from working with firm i. This term reflects non-transferable preferences, such as the worker's valuation of the firm's culture, reputation, or work environment. The transfer t_{ij} encompasses more than just monetary wages; it also includes any contractually negotiated components, including non-monetary benefits. In the publishing context, these may include editorial attention, marketing support, or autonomy in the production process. Finally, let $u_{0j} = a_{0j}$ denote the utility from the outside option, which depends solely on worker j's type.

Let

$$v_{ij} = a_{ij} + f_{ij} = u_{ij} + \pi_{ij} \tag{1.5}$$

[24] A large part of the empirical matching literature assumes this functional form, in which $f(C)$ can be linearly decomposed into individual contributions f_{ij}, thereby ruling out complementarities and externalities in production. This assumption is reasonable in the publishing industry, where author-publisher relationships tend to be independent of other pairings. It also has theoretical appeal, as it ensures the existence of stable matching without further restrictions.

denote the *joint surplus* (or *value*) of the pair ij, which is independent of the transfer t_{ij}. For unmatched workers, the joint surplus is defined as $v_{0j} = u_{0j}$. Let $\boldsymbol{v} = (v_{ij})_{ij}$ denote the vector collecting joint surpluses of all potential matches. In the empirical literature of matching with transferable utility, the primary object of interest is this joint surplus v_{ij}. Intuitively, the first equality in (1.5) reflects that v_{ij} serves as a reduced-form summary of the total value generated by the match.[25] The second equality emphasizes the distributional perspective: u_{ij} and π_{ij} represent the net *surplus* received by the worker and firm, respectively. This post-transfer division of surplus is the primary focus of the analysis in 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 assignments to form a new match. Under this setup, the stability condition can be equivalently expressed as the following linear programming (LP) problem (Gretsky, Ostroy, and Zame, 1992; Galichon and Salanié, 2023):

$$\begin{aligned} \max_{\boldsymbol{\mu}} \ & \boldsymbol{v}'\boldsymbol{\mu} \\ \text{s.t.} \ & \sum_j \mu_{ij} = q_i \text{ for all } i \\ & \sum_i \mu_{ij} = 1 \text{ for all } j \\ & \mu_{ij} \in \{0, 1\}. \end{aligned} \tag{1.6}$$

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, transfers act as price signals that adjust to clear the market, analogous to prices in a competitive equilibrium.[26]

An inversion problem for estimation. From an empirical perspective, we face the inverse optimization problem: given an observed equilibrium matching $\boldsymbol{\mu}$, recover

[25] 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), the empirical 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 additional assumptions are imposed on preferences. For the purposes of this study, however, the distinction is unnecessary because only post-transfer utilities are relevant.

[26] Kelso and Crawford (1982) show that the stable matching can be achieved via a salary adjustment process that is a generalized version of deferred acceptance algorithm. This process resembles an ascending price auction in which firms take turns bidding for workers by offering increasing salaries.

the underlying values $\boldsymbol{v}$ that rationalize this matching. Formally, we seek the 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 (1.6).[27] This inversion involves solving for a vector of jointly consistent 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 justify the match; conversely, for an unmatched pair $i'j$, $v_{i'j}$ must lie below some upper bound $\overline{v}_{i'j}$ to ensure that it remains unmatched.

In the estimation, equilibrium bounds are computed by partially characterizing the matching equilibrium using the *two-pair-no-exchange* condition introduced in Fox (2010) and Fox (2018). This condition rules out profitable deviations in which two matched pairs, ij and $i'j'$, mutually abandon their current partners to form two new pairs, ij' and $i'j$, i.e.,

$$v_{ij} + v_{i'j'} > v_{ij'} + v_{i'j} \tag{1.7}$$

for all $\mu_{ij} = 1$, $\mu_{i'j'} = 1$, and $i \neq i'$.[28] 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 matched pairs where $\mu_{i'j'} = 1$ gives the highest lower bound of v_{ij}:

$$\underline{v}_{ij} = \max_{\substack{\mu_{i'j'}=1 \\ i'\neq i}} v_{ij'} + v_{i'j} - v_{i'j'}. \tag{1.8}$$

Conversely, for an unmatched pair $i'j$, 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 $v_{i'j}$:

$$\overline{v}_{i'j} = \min_{\mu_{i'j'}=1} v_{ij} + v_{i'j'} - v_{ij'}. \tag{1.9}$$

Division of surplus. While the pre-transfer preferences are not identified, the equilibrium characterization allows us to recover the post-transfer division of surplus,

[27] Mathematically, this corresponds to the dual cone (or polar cone, depending on the convention) of the set of feasible matchings $\tilde{\boldsymbol{\mu}}$ at $\boldsymbol{\mu}$.

[28] These conditions are necessary but not sufficient for the LP problem (1.6). In other words, the bounds $\underline{v}_{ij}$ and $\overline{v}_{i'j}$ are not tight. In principle, stability requires the absence of any profitable cycle of deviations (a notion of core stability), e.g., $v_{ij} + v_{i'j'} + v_{i''j''} > v_{ij'} + v_{i'j''} + v_{i''j}$. However, checking all such cycles is computationally intractable and unnecessary for the purposes of this paper. Fox (2018) demonstrate that the score estimator based on the inequality in (1.7) yields set identification. In my implementation, Monte Carlo simulations confirm that the parameters are identified. See details in Section 1.4.

u_{ij} for the worker and π_{ij} for the firm for all matched pairs $\mu_{ij} = 1$. Although the matching $\boldsymbol{\mu}$ is generically unique, the corresponding surplus division is not. In particular, Sotomayor (1999) shows that the set of post-transfer outcomes u and π forms a lattice. Accordingly, the estimation proceeds by first characterizing the set of feasible matchings and then identifying a firm-optimal allocation within this set. Any valid division of the joint surplus must support $\boldsymbol{\mu}$ as a stable matching by satisfying the pairwise stability condition. For a firm i and a worker j' who are not currently matched, the total value of their potential match must not exceed the combined utilities they each receive under the existing matching. That is,

$$v_{ij'} < u_{i'j'} + \pi_{ij}, \tag{1.10}$$

so that neither party has an incentive to deviate and form a new match. Substituting $\pi_{ij} = v_{ij} - u_{ij}$ and rearranging terms yields

$$u_{ij} - u_{i'j'} < v_{ij} - v_{ij'}. \tag{1.11}$$

Intuitively, this inequality states that worker j's utility (in ij) cannot exceed that of worker j' (in $i'j'$) by more than a threshold defined by the relative match values. Otherwise, j' could propose to i and achieve a mutually preferable deviation.

To further bound the utility u_{ij}, we impose the condition that workers in all matched pairs must receive a payoff strictly higher than their outside option:

$$u_{ij} > u_{0j} = v_{0j}. \tag{1.12}$$

On the firm side, because firms lack outside options in this framework, π_{ij} is not unbounded below by some reservation value, implying that u_{ij} is correspondingly unbounded above. Conveniently, no upper bound on u_{ij} is required for the firm-optimal allocation. In many labor markets, equilibrium outcomes arise from firms making sequential wage offers to workers, who choose whether to accept or reject them.[29] Kelso and Crawford (1982) show that this type of ascending, firm-proposing salary adjustment process results in the firm-optimal stable allocation. Thus, the unique lower bounds on u_{ij} in the firm-optimal outcomes are characterized by the

[29] In the publishing industry, for example, publishers frequently compete for manuscripts through bidding. The court record in *U.S. v. Bertelsmann SE & Co. KGaA, 646 F. Supp. 3d 1 (D.D.C. 2022)* documents numerous instances of such competitive bidding.

following linear program:

$$\begin{aligned} \min_{\boldsymbol{u}} \quad & \sum_{\mu_{ij}=1} u_{ij} \\ \text{s.t. } & u_{ij} - u_{i'j'} < v_{ij} - v_{ij'} \\ & u_{ij} > v_{0j} \\ & \text{for all } \mu_{ij} = 1, \mu_{i'j'} = 1, \text{ and } i \neq i'. \end{aligned} \tag{1.13}$$

Specification

Match value production. The value v_{ij} is parametrized as a linear function of observable pair-specific characteristics:

$$v_{ij} = X'_{ij}\beta + \varepsilon_{ij}, \tag{1.14}$$

where X_{ij} denotes firm-worker-specific characteristics, β is a vector of parameters to be estimated, and ε_{ij} is a random utility shock.[30] This specification captures how value arises from complementarities between firms and workers. The reservation value of the worker, v_{0j}, is specified as

$$v_{0j} = X'_{0j}\beta^{RV} + \varepsilon_{0j}, \tag{1.15}$$

where X_{0j} is worker-specific characteristics, and β^{RV} are the parameters to be estimated for the reservation value. The error term ε_{0j} is drawn from the same distribution as ε_{ij}. As in standard discrete choice models, β is identified only up to scale and location. Accordingly, the constant term is absent and the variance of the error term ε is normalized to 1 so that β is identified.

Match performance. In addition to the match value specification, two outcome equations are estimated to evaluate the performance of matched pairs where $\mu_{ij} = 1$. First, for popularity, the log-transformed rating count of the book is denoted by r_{ij}. This outcome is modeled as a function of a vector of pre-publication characteristics W_{ij}, which may differ from X_{ij} above:

$$r_{ij} = \log(RatingsCount_{ij}) = W'_{ij}\gamma^r + \eta_{ij}. \tag{1.16}$$

Similarly, let s_{ij} denote the book's average rating. This outcome is also modeled as a function of the book's pre-publication characteristics W_{ij}:

$$s_{ij} = AverageRating_{ij} = W'_{ij}\gamma^s + \zeta_{ij}. \tag{1.17}$$

[30] 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 (1.6) or (1.7), firm-specific and worker-specific characteristics do not affect matching outcomes.

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}). These two components of the model are complementary in structure and interpretation. On the one hand, incorporating the matching model in the performance outcomes is conceptually similar to Heckman's correction for sample selection bias (Heckman, 1979). As noted earlier, the set of observed books is not a random sample from the universe of possible author-publisher pairs; rather, it reflects endogenous matching decisions. Without correcting for this selection, direct estimation of equations (1.16) and (1.17) would yield biased coefficients because the observed matches represent a nonrandom subset of all possible matches. In this sense, the structural matching model functions as a control function approach to account for endogeneity due to selective matching.

On the other hand, book performance provides a channel to estimate sorting on unobservable characteristics. Unobserved factors specific to a match may influence both the latent value v_{ij} and the performance outcomes r_{ij} and s_{ij}. To the extent that these factors affect performance, the observed outcomes provide additional information about the underlying value of matched pairs. This approach is analogous to recovering latent heterogeneity from outcome variables in other structural models. A direct estimation of the matching model based on the equilibrium characterization (1.7), such as the semiparametric approach in Fox (2018), does not exploit the additional information embedded in the performance equations via correlated disturbances. This connection is made explicit in equation (1.D.2) in the Appendix, where the performance variables help inform the distribution of match values.

Errors. To link the match formation and performance components of the model, a parametric structure is imposed on the joint distribution of the unobserved error terms. The vector of errors $(\varepsilon, \eta, \zeta)$ is assumed to be independently and identically distributed across pairs ij and have a joint normal distribution with mean zero. To parameterize the covariance structure, it is convenient to decompose the errors into orthogonal components $(\varepsilon, \xi_1, \xi_2)$, each normally distributed with mean zero and variances $1, \sigma_1^2, \sigma_2^2$, respectively. Following the standard identification approach in probit models, the variance of ε is normalized to 1, ensuring that β is identified. To capture correlation between match formation and performance, ε is allowed to be correlated with the performance shocks. Specifically, let (ε, η) have covariance δ, and (ε, ζ) have covariance ω. Accordingly, $\eta = \delta\varepsilon + \xi_1$ and $\zeta = \omega\varepsilon + \xi_2$. This decomposition remains flexible; the only restriction is the normalization of the

variance of ε to 1. The resulting covariance matrix of $(\varepsilon, \eta, \zeta)$ is given by

$$\begin{pmatrix} 1 & \delta & \omega \\ \delta & \delta^2 + \sigma_1^2 & \delta\omega \\ \omega & \delta\omega & \omega^2 + \sigma_2^2 \end{pmatrix}. \tag{1.18}$$

Estimation

Let m index the matching market (cohort), each corresponding to a half-year period. Each market consists of two disjoint sets of publishers I_m and authors J_m. To simplify notation, I omit the subscript m when referring to quantities within a single market. Within a given market, each pair ij is characterized by the following variables: value-specific covariates X_{ij}, latent match value v_{ij}, observed equilibrium matching indicator μ_{ij}, performance-specific covariates W_{ij}, and performance variables r_{ij} and s_{ij} (observed only for matched pairs). Let italicized symbols $X, \boldsymbol{v}, \boldsymbol{\mu}, W, r, s$ denote the respective matrices or vectors collecting these variables across all pairs ij within a given market. Let upright bold symbols $\mathbf{X}, \mathbf{v}, \boldsymbol{\mu}, \mathbf{W}, \mathbf{r}, \mathbf{s}$ denote the corresponding stacked variables across all markets m in the dataset.

The parameters to be estimated include the valuation parameters β, β^{RV} from equations (1.14) and (1.15), the performance parameters γ^r, γ^s from equations (1.16) and (1.17), and the error covariance matrix $(\delta, \omega, \sigma_1^2, \sigma_2^2)$ in (1.18). Let $\boldsymbol{\theta}$ denote the full vector of parameters to be estimated.

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

$$\mathcal{L}_m(\beta|\boldsymbol{\mu}, X) = P(\boldsymbol{v} \in V_{\boldsymbol{\mu}}|\beta, X) = P(\varepsilon \in V_{\boldsymbol{\mu}} - X\beta) = \int \mathbf{1}(\varepsilon \in V_{\boldsymbol{\mu}} - X\beta) dF(\varepsilon). \tag{1.19}$$

Recall that $V_{\boldsymbol{\mu}}$ denotes the set of values that rationalize $\boldsymbol{\mu}$ as the observed equilibrium matching. In principle, β could be estimated by maximizing the total likelihood across all markets $\prod_m \mathcal{L}_m(\beta|\boldsymbol{\mu}, X)$. However, this likelihood is computationally intractable due to the high dimensionality of the integral. A key feature of the matching models is rivalry: agents do not act independently, and one firm's match with a worker excludes other firms from doing the same, and vice versa. As a result, the error terms within each market are interdependent and must be integrated jointly. This interdependence renders high-dimensional integration over market-specific errors computationally prohibitive.[31]

[31] To see this more explicitly, note that the latent values v_{ij} are not directly observed, so we cannot

To bypass the explicit evaluation of the likelihood function, I adopt a Bayesian approach to estimate the matching model, following Sorensen (2005) and Sørensen (2007). Specifically, estimation proceeds via Markov Chain Monte Carlo (MCMC) with Gibbs sampling, a data augmentation technique where the latent match values v_{ij} are treated as latent variables and sampled alongside structural parameters $\boldsymbol{\theta}$. The Markov chain proceeds by iteratively drawing from the full conditional distributions of each parameter, given the current values of all other parameters.[32]

Prior. Given the model specification, conjugate priors $f_0(\boldsymbol{\theta})$ are selected to ensure that the conditional posteriors remain within known parametric families, facilitating efficient Gibbs sampling. All parameter priors are assumed to be independent. For the parameters β, γ^r, γ^s, δ, and ω, normal priors are assigned: $f_0(\theta) \sim N(\theta_0, \Sigma_{\theta,0})$, with prior mean $\theta_0 = 0$ and covariance matrix $\Sigma_{\theta,0} = 10 \cdot I$, where I is the identity matrix of appropriate dimension. These choices reflect weakly informative priors centered at zero. For the variance parameters σ_1^2 and σ_2^2, inverse gamma priors are used: $f_0(\theta) \sim \text{Inv-Gamma}(\alpha_0, \beta_0)$ with shape and scale parameters $\alpha_0 = 1$ and $\beta_0 = 1$ (not to be confused with the coefficient vector β in the value function). These choices yield diffuse priors with heavy tails, reflecting minimal prior information.

Posterior. Given the error structure with mean zero and covariance matrix defined in (1.18), the conditional likelihood (or joint density) of the latent match values v and observed performance variables r and s in market m follows a multivariate normal distribution:

$$
\begin{aligned}
f_m(\boldsymbol{v}, r, s | X, W, \boldsymbol{\theta}) \propto \prod_{ij} \exp\left(-\frac{1}{2}\Big(v_{ij} - X_{ij}'\beta\Big)^2 \right) \times \\
\prod_{\mu_{ij}=1} \exp\left(-\frac{1}{2}\Big(\frac{r_{ij} - W_{ij}'\gamma^r - \delta(v_{ij} - X_{ij}'\beta)}{\sigma_1}\Big)^2 \right) \times \\
\prod_{\mu_{ij}=1} \exp\left(-\frac{1}{2}\Big(\frac{s_{ij} - W_{ij}'\gamma^s - \omega(v_{ij} - X_{ij}'\beta)}{\sigma_2}\Big)^2 \right). \qquad (1.20)
\end{aligned}
$$

construct observation-level likelihood contributions for individual matches. Instead, identification relies on inequalities implied by the equilibrium. For example, the pairwise stability condition (1.7) 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 the left-hand side can be treated as a random variable. However, this object is not an independent sample: the matched-pair errors (e.g., ε_{ij}) are sampled at a much higher rate than the unmatched terms (e.g., $\varepsilon_{ij'}$), further complicating inference.

[32] See Gelman et al. (2013) for Bayesian computation via MCMC. These methods are commonly used in discrete choice models and have been widely adopted in marketing and industrial organization research. See, for example, Rossi, Allenby, and McCulloch (2012).

Note that the normalization factor in the density function is omitted; only the kernel of the joint density is shown. As before, the summation condition $\mu_{ij} = 1$ denotes the set of observed matches.

The augmented posterior density f across all markets is proportional to the product of the prior distribution of parameters f_0, the conditional densities f_m from (1.20), and the boundary conditions that guarantee stability of the observed matching. That is,

$$f(\mathbf{v}, \mathbf{r}, \mathbf{s}, \boldsymbol{\theta} | \boldsymbol{\mu}, \mathbf{X}, \mathbf{W}) \propto f_0(\boldsymbol{\theta}) \times \prod_m \left[f_m(\boldsymbol{v}, r, s | X, W, \boldsymbol{\theta}) \times \prod_{\mu_{ij}=1} \mathbf{1}(v_{ij} \geq \underline{v}_{ij}) \times \prod_{\mu_{ij}=0} \mathbf{1}(v_{ij} < \overline{v}_{ij}) \right], \tag{1.21}$$

where $\underline{v}_{ij}$ and $\overline{v}_{ij}$ are the bounds defined in equations (1.8) and (1.9), respectively. The conditional densities of v_{ij} and θ are proportional to their respective components in the augmented posterior distribution in (1.21). Detailed expressions for the Gibbs samplers $f(v_{ij}|\cdot)$ and $f(\theta|\cdot)$ are provided in Section 1.D.

Estimation Results

The structural model is estimated using the subsample of books published between 2010 and 2013, with each half-year treated as a distinct matching market. Table 1.3 presents the parameter estimates from the structural model.

Value parameters. The estimates for the match value equation (1.14) and the reservation value specification (1.15) are reported in Table 1.3a. Since the parameters are identified only up to scale and location, their magnitudes are not directly interpretable. However, the estimated coefficients have the expected signs and are statistically significant. In particular, both genre similarity and content similarity—measures of editorial compatibility—have strong positive effects on match value. Past collaboration history also plays a significant role, indicating strong persistence in matching: once a match is formed, it tends to generate greater value and is likely to lead to repeated collaboration.

As in logit and probit models, the coefficients in the value equation are interpreted through their marginal effects. An analogous marginal effect is computed following the approach proposed by Sørensen (2007). Suppose two pairs of authors and publishers, ij and $i'j'$, have identical observed characteristics ($X_{ij} = X_{i'j'}$). Then the probability that ij is a match in equilibrium while $i'j'$ is not (or vice versa) is 0.5, holding all other matches fixed. 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

Table 1.3: 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)

one-unit change in X_{ij}.[33] For example, an increase of 0.01 in genre similarity (a continuous variable on $[0, 1]$) raises the probability of a match by approximately 0.5 percentage points.

Model fit. Model fit is assessed by comparing predicted matches with observed outcomes. Because the matching framework reflects two-sided decisions, there is no standard goodness-of-fit metric. However, from the authors' perspective—where each author is matched to a single publisher—the problem resembles a discrete choice setting. Accordingly, prediction accuracy is evaluated by checking whether the model correctly identifies the observed publisher for each author. The model achieves a prediction accuracy of approximately 67%. For comparison, prediction accuracy under random assignment subject to capacity constraints is only about

[33]The probability that ij is a match while $i'j'$ is not is $Pr(X'_{ij}\beta + \varepsilon_{ij} > X'_{i'j'}\beta + \varepsilon_{i'j'}) = \Phi((X'_{ij} - X'_{i'j'})\beta/\sqrt{2})$, which is equal to 0.5 when $X_{ij} = X_{i'j'}$. The marginal effect is the derivative of this probability with respect to X_{ij}, evaluated at $X_{ij} = X_{i'j'}$. For a binary variable, this effect is $\Phi(\beta/\sqrt{2}) - 0.5$. For a continuous variable, it is $\phi(0)\beta/\sqrt{2}$, where Φ and ϕ are the cumulative distribution function and probability density function of the standard normal distribution, respectively.

Table 1.3: 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		

15%.

The strength of the structural matching framework is further demonstrated by comparing it to alternative models of match formation, reported in Table 1.C9 in the appendix. In these alternative specifications, the unit of observation is a book-publisher pair, and the outcome is a binary indicator for whether the pair forms a match. The explanatory variables are the same as those used in the structural model.

Table 1.3: 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)

The key distinction is that these alternative models treat each book-publisher pair as an independent observation, whereas the structural matching model accounts for rivalry and the equilibrium interdependence across pairs. As expected, these models have less prediction accuracy, at about 52%-54%, compared to 67% under the structural matching model. Moreover, failing to account for the underlying matching process leads to systematically overstated coefficient estimates.

Performance parameters. The estimates of the performance equations are reported in Table 1.3b. The coefficients on pre-publication author characteristics—specifically, rating count percentile and average rating—are positive and statistically significant across both performance outcomes. This indicates temporal correlation in author success: an author's prior popularity and quality are strong predictors of future book performance, suggesting author ability as the primary driver of book success. By contrast, measures of compatibility, such as content similarity and prior collaborations, do not have statistically significant effects on performance once selection into matches is accounted for. This stands in contrast to their strong influence on the matching decision itself.

As with the match value equation, the performance equation estimates from the structural model are compared to those obtained from simpler specifications. Table 1.C11 in the Appendix reports results of direct OLS regressions of the performance variables on the same set of covariates, but without accounting for the endogenous matching process. The OLS estimates differ meaningfully from those obtained under the structural approach. For example, a one-percentile increase in a publisher's rating count increases the book's rating count by 4.9% in the structural model, but the OLS estimate is overstated at 5.2%. This difference reflects the bias introduced by unobserved selection into matches, which arises because equilibrium sorting influences observed performance outcomes.

1.5 Merger Simulation

The primary interest of this paper is the impact of mergers on the labor market and worker welfare. As discussed in Section 1.2, the 2013 merger between Penguin and Random House substantially consolidated the market for authors. To analyze its impact, a counterfactual simulation is conducted assuming the merger had occurred in 2010 rather than 2013. In this simulated environment, Penguin and Random House are treated as a single merged entity beginning in 2010. This methodology follows the simulation-based approach of Fan (2013), Wollmann (2018), and Li et al. (2022), enabling comparisons within a consistent cohort of agents prior to the observed merger. Throughout the discussion, I refer to this simulated scenario as the "post-merger" case. The subscripts P, RH, and PRH denote Penguin, Random House, and Penguin Random House, respectively.

To begin the counterfactual analysis, the post-merger primitives must be specified. If the merger were modeled simply as the removal of one firm from the market, authors would be weakly worse off due to the loss of a bidder on the employer side (Crawford and Knoer, 1981). However, a merger involves the integration of two firms into a single entity, which has three key implications in the context of my empirical model: market participants, capacity constraints, and match values. First, it is assumed that the set of market participants remains unchanged. In particular, the merger does not induce entry or exit of authors or other publishers. Second, drawing on the empirical observation that the total number of books published did not materially change following the merger, firm capacity is assumed to be constant. Accordingly, the capacity of the merged firm, q_{PRH}, is set equal to the sum of the original firms' capacities: $q_P + q_{RH}$. Third, match values are held constant for all firms except the merged entity. Only the match values involving Penguin Random House, v_{PRH}, will be different in the post-merger counterfactual and will be outlined in detail in Section 1.5.

To implement counterfactual experiments, the updated primitives $\boldsymbol{v}$ are used to simulate counterfactual equilibrium matching $\boldsymbol{\mu}$ for each semiannual matching market. This is done by applying the first LP characterization in (1.6). Once the simulated matches are obtained, the equilibrium division of surplus $\boldsymbol{u}$ is computed by solving the second LP problem in (1.13). Finally, the model generates implied book performance outcomes, i.e., rating count and average rating, based on the outcome equations (1.16) and (1.17).

The simulated post-merger outcomes are compared to a simulation of the pre-

merger baseline, allowing for a consistent evaluation within the same equilibrium framework. Although the model permits full adjustment by all market participants, many authors are expected to remain with their original publishers, as their match values are likely to dominate under both scenarios. To capture heterogeneity in the effects of the merger, the analysis focuses on two groups of authors: those who remain with Penguin or Random House, and those who switch publishers due to changes in equilibrium sorting. For each group, the impact of the merger is evaluated in terms of both total surplus and its allocation between authors and publishers. Three outcome metrics are analyzed to assess these effects: (1) the transfer of value from other publishers to Penguin Random House, (2) the shift in surplus from authors to publishers, and (3) the redistribution of surplus among authors with different market positions.

Counterfactual assumptions

The merged company's match values, relative to those of its predecessors, raise more nuanced empirical questions. For a given author j, the predecessor match values $v_{P,j}$ and $v_{RH,j}$ are replaced in the post-merger setting by a single value $v_{PRH,j}$. How this new value compares to the previous ones depends on the internal repositioning of Penguin Random House after the merger. The literature provides substantial evidence that mergers can reshape the strategic positioning of both acquiring and acquired firms. For instance, Sweeting (2010) presents reduced-form evidence of product repositioning after mergers, while Fan (2013) models product characteristics as endogenous to merger dynamics. Furthermore, Eliason et al. (2020) shows that acquired firms often converge toward the operational style or behavior of their new parent firms. Because internal changes within Penguin Random House are not directly observable in my data, three distinct post-merger scenarios are simulated to capture alternative integration structures: (1) synergistic integration, (2) organic merger, and (3) Random House takeover.

First, the synergistic integration scenario represents a best-case outcome in which the post-merger match value is set to the better of the two pre-merger values. This specification captures the idea that Penguin and Random House each bring complementary strengths and editorial expertise to the merged entity. Publishing is a highly individualized industry on the publisher's side and depends heavily on the expertise of specific editors. Since editorial staff largely remained in place following the merger, it is reasonable to expect that they continued to apply their specialized knowledge and relationships in the new organizational setting. Under this scenario,

the merged firm is assumed to draw on the strongest editorial match for each author, regardless of whether it originated from Penguin or Random House.

Alternatively, the organic merger scenario is motivated by the repositioning literature and assumes that Penguin Random House functions as a unified organization with characteristics defined as a weighted average of its two predecessor firms. This counterfactual reflects a case in which the merging companies reconcile their differences and proceed as a single, cohesive organization.[34] To implement this scenario, counterfactual publisher characteristics X and W are constructed by averaging the characteristics of books published by Penguin and Random House in each genre and time period, based on the prior year's publications.

Finally, the Random House takeover scenario assumes that the post-merger entity adopts only the characteristics of Random House. Although the 2013 merger began as a joint venture between Bertelsmann (the parent company of Random House) and Pearson (the parent company of Penguin), Bertelsmann held a majority stake from the outset and eventually acquired full ownership.[35] Given this trajectory, in which Bertelsmann gradually assumed full control, it is reasonable to infer that Random House exerted a dominant influence over post-merger decision-making and editorial strategy. This scenario thus treats the merger as a phased acquisition, in which the merged firm ultimately operates under Random House's editorial philosophy and organizational model.

Overall market effects

We now turn to the results of the simulation exercises. Table 1.4 summarizes the overall effects under the three counterfactual scenarios. Table 1.5 and Table 1.6 report the detailed results of the second counterfactual under organic merger. Results of other scenarios are presented in Section 1.E. All figures report changes in value from the pre-merger to the post-merger state. Because the match values are identified only up to a monotonic transformation, their absolute magnitudes are not directly interpretable. However, comparisons of relative magnitudes remain meaningful and informative. Each table is organized into six columns: the first three report

[34] Anecdotal accounts suggest that Penguin and Random House had distinct corporate cultures. Penguin, particularly under CEO John Makinson, was known for its innovation and willingness to take risks, often publishing experimental or controversial titles. In contrast, Random House was recognized for its scale and commercial focus, often producing blockbuster titles with broad market appeal.

[35] In 2013, Bertelsmann owned 53% of the joint venture, while 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 1.4: 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

aggregate changes, and the latter three report average changes per author. In both panels, the reported metrics include the total change in joint surplus, along with the respective changes accruing to authors and publishers.

Table 1.4 reports the change in total social surplus, $\sum v_{ij}\mu_{ij}$, as well as the corresponding shares accruing to authors and publishers under each of the three counterfactual scenarios. The results show a net increase in social surplus under the synergistic integration scenario, but a net loss under the organic merger and Random House takeover scenarios. This divergence reflects the tension between two opposing forces introduced by the merger: efficiency gains from improved matching capacity, and losses from diminished compatibility or creative misalignment.

Efficiency gains. On the one hand, the net gain under the synergistic integration scenario is expected. This counterfactual assumes both the combined capacities of the two firms and, for each author, the higher of the two pre-merger match values. Recall that equilibrium matching maximizes total social welfare. Under this assumption, the merger expands feasible allocations and thus must weakly increase total surplus. This mechanism is illustrated in the example provided in Figure 1.4, which 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 one, and all reservation values are negative, so that all authors strictly prefer to be matched. The table displays the match values for each author-publisher pair. The pre-merger equilibrium is straightforward to identify. In the post-merger scenario, Penguin Random House adopts the better of Penguin's or Random House's match values for each author. This expanded flexibility allows it to reallocate capacity and match with Coleridge, a higher-value pairing that was

Figure 1.4: Example of efficiency gains under synergistic integration

	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

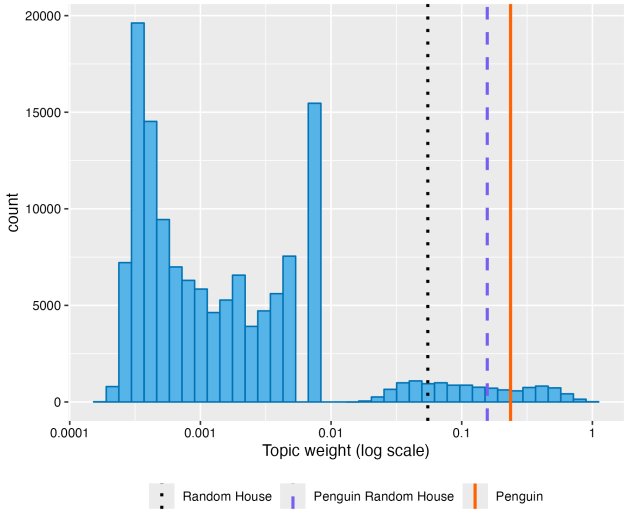
Notes: Rows represent publishers and columns represent authors. Each cell contains the match value for a specific publisher-author pair. All outside option values are assumed to be negative, ensuring that every author strictly prefers to be matched. Blue-colored cells indicate the equilibrium matches.

infeasible in the pre-merger equilibrium due to market tightness.

Creative misalignment. On the other hand, the net loss observed under the organic merger and Random House takeover scenarios reflects the effects of mismatch arising from market homogenization and the resulting decline in match values. The two merging publishers historically maintained distinct editorial philosophies and catered to different segments of the author pool. In both scenarios, the merged entity is forced to consolidate into a single post-merger identity, which leads to a loss of compatibility with some authors whose preferences were better aligned with one of the original firms. This erosion of editorial fit reduces the overall quality of matches, even if capacity remains constant.

To illustrate this mechanism, Figure 1.5 presents a histogram of authors in the genre of literary fiction, based on topic weights derived from the language model. Recall that each genre is measured on a continuous scale between 0 and 1. The concentration on the left represents books with low relevance to this genre, while the right-hand tail consists of books more closely associated with it. The solid and dotted vertical lines mark the positions of Penguin and Random House, respectively, on this genre dimension. By construction, the dashed line indicates the position of Penguin Random House under the organic merger scenario, where the post-merger characteristic is the average of its two predecessors. Because compatibility declines with the distance between an author and a publisher in topic space, this averaging leads to reduced alignment for some authors. Under both the organic merger and Random House takeover scenarios, more authors are positioned farther from the merged firm's post-merger identity, resulting in a decline in match quality.

Figure 1.5: Distribution of authors along some example genre (literary fiction)



Notes: The genre shown corresponds to topic #37 in the genre topic model described in Section 1.B. This topic approximately captures the characteristics of literary fiction.

Impact on authors

Penguin Random House accounts for approximately one-quarter of the market. Following the merger, about 8% of authors switched to a different publisher. Figure 1.6 illustrates the migration patterns of authors who changed publishers in the post-merger equilibrium. A notable feature of these movements is the frequent "exchange" of authors between Penguin Random House and other publishing houses. This pattern is expected, given the modeling assumption that match values for all non-merging publishers remain fixed, thereby preserving their relative rankings.

Column (2) of Table 1.4 reports changes in authors' share of total surplus across the three counterfactual scenarios. In each case, authors receive a smaller portion of the surplus in the post-merger equilibrium. While overall social surplus increases under synergistic integration, authors still experience a decline in utility—even when total market value improves. This indicates that consolidation affects not only efficiency but also the distribution of value. The resulting redistribution highlights the asymmetric impact of the merger, with publishers capturing a greater share of the gains while authors absorb a relative loss.

To understand the mechanism behind the changes in author welfare, the analysis considers two groups of authors: those who remain with Penguin Random House and those who switch publishers. Panel C of Table 1.5 focuses on the first group and shows consistent losses for these authors across scenarios. Even under the

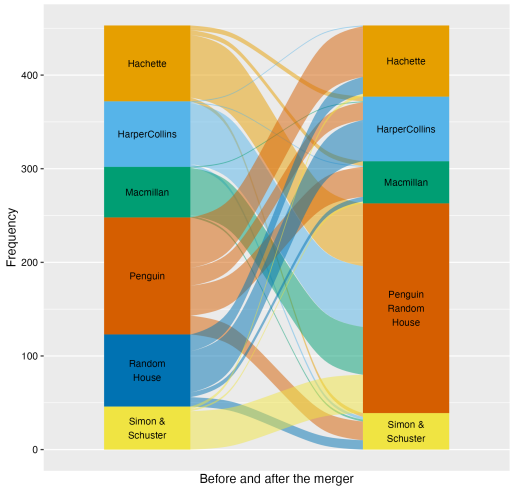
Table 1.5: Simulation results of organic merger

	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

synergistic integration scenario, where total surplus for these matches rises slightly post-merger, the authors' utilities decline substantially, while the publisher's share increases. This outcome is driven primarily by the loss of competition between previously independent firms. Before the merger, Penguin and Random House competed to attract authors, placing upward pressure on compensation. After the merger, this dynamic disappears, resulting in weakened bargaining power for authors. This mechanism aligns with concerns raised in the 2022 merger case, which emphasized the loss of competition between previously independent entities.

Panel D of Table 1.5 shows that authors who switched publishers experienced

Figure 1.6: Movement of authors after the merger



Notes: Movement of authors before and after the merger. Approximately 8% of authors switched publishers in the post-merger equilibrium. Authors who remained with their original publishers are not shown in the figure.

substantial changes in utility, with outcomes varying by the direction of movement. This re-sorting, largely consisting of exchanges between Penguin Random House and other publishers, is driven by the merged entity's expanded capacity. Authors who moved to Penguin Random House experienced notable welfare gains, while those who were displaced to other publishers faced losses, in general. The result is a polarization of outcomes across the author pool, with most of the surplus reallocation borne by authors. In effect, we observe a transfer of utility from authors previously matched with other publishers to those now matched with Penguin Random House, further highlighting the merger's uneven distributional consequences.

Heterogeneity by author market position. Given the centrality of distributional effects in this setting, the analysis further disaggregates outcomes by author market position. Three groups are considered: bestselling, mid-list, and debut authors. The results of this decomposition are reported in Table 1.6.

Column (4) of Panel A in Table 1.6 reports outcomes for authors who remained with Penguin Random House. While authors across all three market segments experienced utility losses, the impact is not evenly distributed. On average, bestselling and mid-list authors suffered greater losses than debut authors in absolute terms. This pattern holds consistently across the other two counterfactual scenarios. This disparity arises because bestselling authors were the most competitively sought after in the pre-merger market. The elimination of competition between Penguin and Random

Table 1.6: Simulation results of organic merger by author market position

	Aggregate change			Average change per author		
	Joint surplus (1)	Author (2)	Publisher (3)	Joint surplus (4)	Author (4)	Publisher (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

Notes: In Panel B, publishers refer to authors' destination assignments in the post-merger equilibrium.

House reduced their bargaining power, resulting in the largest utility decline. This finding provides further empirical support for the DOJ's position in the 2022 merger case that top-selling authors are especially vulnerable to harm from consolidation. Although debut authors were often expected to be the most vulnerable group, the analysis suggests that the loss of competition most severely affected those who had the most to lose, namely, established authors who benefited from pre-merger bidding

dynamics.

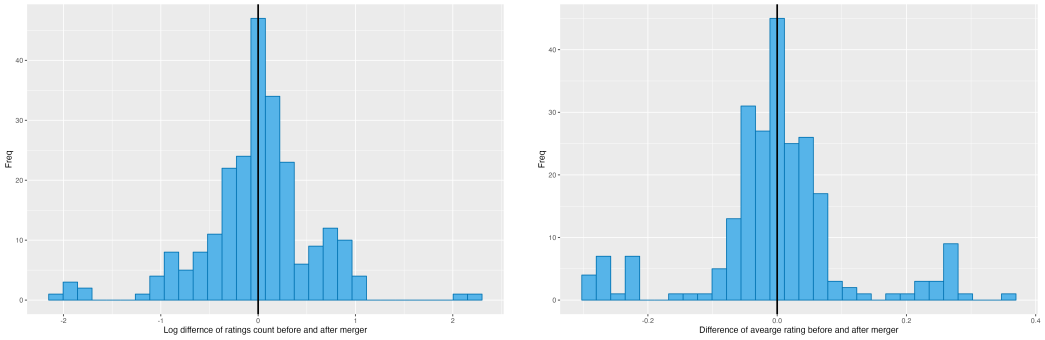
Column (4) of Panel B in Table 1.6 presents average changes in authors' share of surplus for those who were re-sorted to different publishers following the merger. The results are consistent with earlier findings of a utility transfer from authors at other publishers to those matched with Penguin Random House, but they also reveal heterogeneous effects across author segments. Bestselling authors who exited Penguin Random House experienced the largest declines in surplus share, while those who entered gained relatively little. In contrast, debut authors saw smaller losses when moving away from Penguin Random House, but the largest gains when moving into it. Mid-list authors again exhibited intermediate patterns, with modest losses at competing publishers and the highest gains at Penguin Random House. These patterns highlight Penguin Random House's central role in shaping the distribution of value within the post-merger publishing landscape, reflecting its dominant market position.

Additional results

Differentiated impact among publishers. Column (3) of Panel B in Table 1.5 shows a clear redistribution of surplus from competing publishers to Penguin Random House. While rival firms experience declines in both joint surplus and their share of value, Penguin Random House realizes a net gain. Panel C further isolates the internal effect of merging Penguin and Random House, showing that a substantial portion of this gain results from the elimination of competition between Penguin and Random House. Additionally, Panel D highlights the role of equilibrium sorting in driving this redistribution. Penguin Random House improves its market position by attracting authors who were previously matched with other publishers, thereby increasing its share of value at rivals' expense.

Impact on reader reception. The analysis also considers the consumer side by examining changes in reader reception for books affected by the merger. Figure 1.7 presents changes in both rating count and average rating for books whose authors were reallocated to different publishers in the post-merger equilibrium. The results indicate negligible changes in either metric, and t-tests confirm that the differences are not statistically significant. After accounting for publisher reassignment, books did not experience meaningful changes in popularity or perceived quality. This result aligns with the prevailing industry view that the merger's primary effects would not be felt by readers. Notably, had the merger been evaluated solely on the

Figure 1.7: Changes in reader reception



(a) Changes in rating count

(b) Changes in average rating

basis of consumer welfare, as is conventional in antitrust analysis, it would have appeared benign.

Alternative counterfactual assumptions. The results of the other two counterfactual simulations, synergistic integration and Random House takeover, are reported in Section 1.E. These simulations yield qualitatively similar patterns to the first scenario in terms of the redistribution effects. While the magnitudes of the effects differ, the core patterns persist: redistribution from authors to publishers, heterogeneous impacts across author segments, and the central role of Penguin Random House in reshaping market dynamics. This consistency across alternative assumptions strengthens the conclusion that the merger fundamentally reshaped the allocation of creative talent and surplus in the publishing industry, even when the mechanism of integration varies.

1.6 Conclusion

This paper examines how consolidation reshapes the allocation of human capital in the creative sector, using a structural two-sided matching framework to study the publishing industry. In markets where productivity depends on compatibility between workers and firms, matching patterns play a central role in determining both efficiency and welfare outcomes. The publishing industry offers a clear example of such a setting, characterized by strong assortative matching between authors and publishers along dimensions of editorial fit, reputation, and content specialization.

To evaluate the consequences of consolidation, this study simulates a major merger in U.S. publishing using estimated structural match values. The results indicate that while integration may alleviate capacity constraints and generate efficiency gains, the benefits accrue disproportionately to the merged firm. The merger redistributes

surplus away from rival firms and from workers to firms more broadly, altering who matches with whom and how value is divided. These shifts are not uniform. The reallocation of talent disproportionately affects experienced and commercially successful authors, who lose the most when competitive pressures are removed. Debut and mid-list authors face smaller adjustments in absolute terms but remain subject to sorting dynamics that reflect changes in the broader equilibrium.

These findings provide context for the DOJ's intervention in the 2022 proposed merger between Penguin Random House and Simon & Schuster. While the agency focused on potential reductions in author compensation, the results here suggest broader effects that extend beyond wages: shifts in match quality, surplus division, and opportunity structure. They also raise the possibility that even the 2013 Penguin-Random House merger may have altered the allocation of talent in meaningful ways. Notably, such outcomes would not have been captured under a conventional consumer-welfare lens, which emphasized the merger as a response to downstream pressure from retailers like Amazon. A narrow focus on prices or output risks overlooking substantial consequences for the flow and value of human capital.

The implications extend beyond publishing. Many high-skilled labor markets—such as consulting, academia, and other creative industries—share two critical features: employment relationships are match-specific, and value creation depends heavily on compatibility between workers and firms. In such settings, human capital is the primary input, and the structure of opportunities shapes both productivity and career outcomes. When mergers alter equilibrium match patterns, they do more than shift firm boundaries; they reconfigure the flow of talent, concentrating value within dominant organizations while weakening opportunities elsewhere. These changes can reshape career trajectories, redistribute bargaining leverage, and affect the institutional conditions under which creative and specialized work is produced. Capturing these dynamics requires tools that reflect the relational and two-sided nature of employment. Where complementarities drive value, mergers have consequences not just for firms, but for the movement and deployment of human capital itself.

References

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

Agarwal, Nikhil and Eric Budish (2021). "Market Design". In: *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". In: *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". In: *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". In: *Journal of Labor Economics* 28.2, pp. 203–210.

Asker, John and Volker Nocke (2021). "Collusion, Mergers, and Related Antitrust Issues". In: *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". In: *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". In: *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". In: *Journal of Political Economy* 128.4, pp. 1325–1369.

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

Berger, David, Kyle Herkenhoff, and Simon Mongey (2022). "Labor Market Power". In: *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*". In: *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". In: *American Economic Journal: Microeconomics* 13.4, pp. 218–260.

Boal, William M. and Michael R Ransom (1997). "Monopsony in the Labor Market". In: *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". In: *Management Science* 50.11, pp. 1587–1602.

Bonaimé, Alice and Ye (emma) Wang (2024). "Mergers, Product Prices, and Innovation: Evidence from the Pharmaceutical Industry". In: *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". In: *Journal of Labor Economics* 31.1, pp. 83–117.

Cabral, Luis (2012). "Reputation on the Internet". In: *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". In: *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". In: *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". In: *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". In: *Journal of Labor Economics* 36.S1, S13–S70.

Chade, Hector, Jan Eeckhout, and Lones Smith (2017). "Sorting through Search and Matching Models in Economics". In: *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". In: *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". In: *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". In: *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". In: *Economic Theory* 42.2, pp. 317–354.

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

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

Crawford, Vincent P. and Elsie Marie Knoer (1981). "Job Matching with Heterogeneous Firms and Workers". In: *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". In: *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". In: *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". In: *Annual Review of Economics* 10.1, pp. 1–29.

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

Eeckhout, Jan and Philipp Kircher (2018). "Assortative Matching With Large Firms". In: *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". In: *The Quarterly Journal of Economics* 135.1, pp. 221–267.

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

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

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

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

Galichon, Alfred and Bernard Salanié (2023). "Structural Estimation of Matching Markets with Transferable Utility". In: *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". In: *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". In: *Journal of Political Economy* 128.11, pp. 4163–4210.

Goldfarb, Avi and Catherine Tucker (2019). "Digital Economics". In: *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". In: *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". In: *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*". In: *The Quarterly Journal of Economics* 133.2, pp. 801–870.

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

Hwang, Hae-shin, Dale T. Mortensen, and W. Robert Reed (1998). "Hedonic Wages and Labor Market Search". In: *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". In: *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". In: *American Economic Review: Insights* 3.3, pp. 303–320.

Kelso, Alexander S. and Vincent P. Crawford (1982). "Job Matching, Coalition Formation, and Gross Substitutes". In: *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". In: *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". In: *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". In: *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". In: *ILR Review* 74.1, pp. 3–26.

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

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

Mindruta, Denisa (2013). "Value Creation in University-Firm Research Collaborations: A Matching Approach". In: *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". In: *Strategic Management Journal* 37.1, pp. 206–231.

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

Nagaraj, Abhishek and Imke Reimers (2023). "Digitization and the Market for Physical Works: Evidence from the Google Books Project". In: *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". In: *Harvard Law Review* 132.2, pp. 536–601.

Peukert, Christian and Imke Reimers (2022). "Digitization, Prediction, and Market Efficiency: Evidence from Book Publishing Deals". In: *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". In: *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". In: *American Economic Review* 111.6, pp. 1944–1971.

Rosen, Sherwin (1986). "The Theory of Equalizing Differences". In: *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". In: *Econometrica* 52.1, pp. 47–57.

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

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

Shapley, L. S. and M. Shubik (1971). "The Assignment Game I: The Core". In: *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". In: *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". In: *International Journal of Game Theory* 28.4, pp. 567–583.

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

Sweeting, Andrew (2010). "The Effects of Mergers on Product Positioning: Evidence from the Music Radio Industry". In: *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". In: *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". In: *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". In: *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". In: *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". In: *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". In: *Marketing Science* 28.6, pp. 1095–1111.

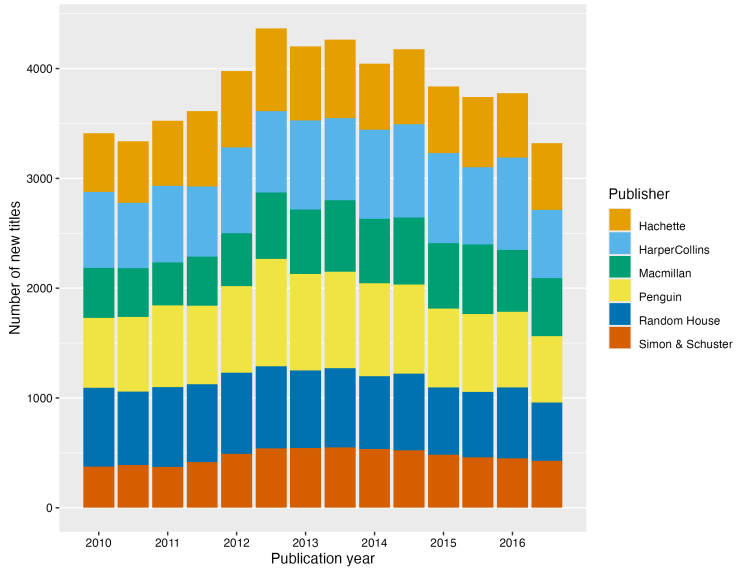
Yeh, Chen, Claudia Macaluso, and Brad Hershbein (2022). "Monopsony in the US Labor Market". In: *American Economic Review* 112.7, pp. 2099–2138.

1.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 1.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.[36] 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.

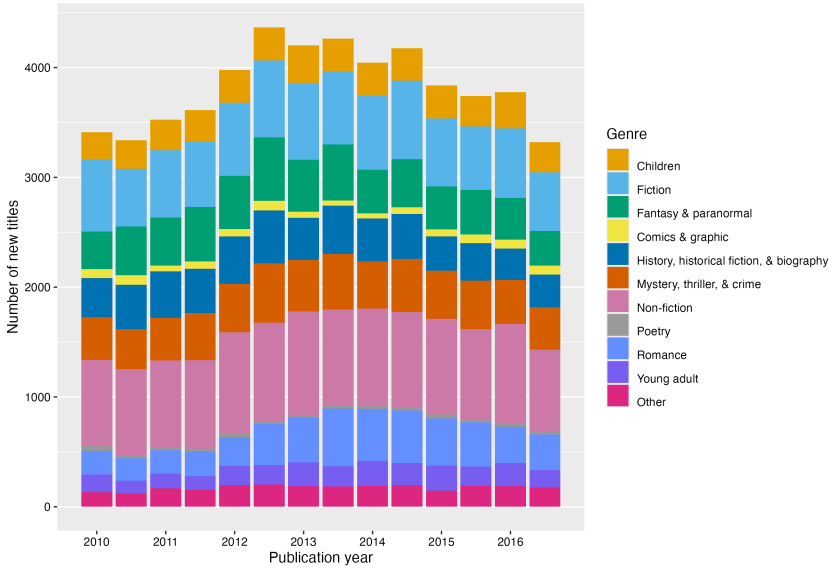
Figure 1.A8: Number of new titles in each half-year



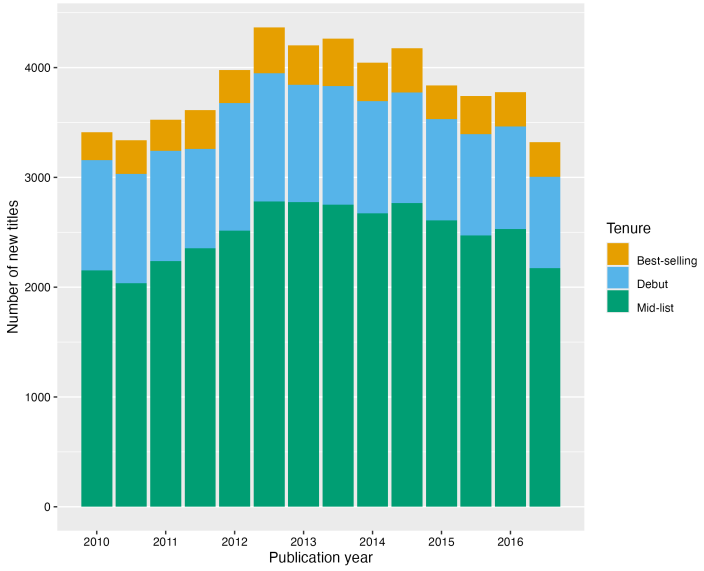
(a) By publisher

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

Figure 1.A8: Number of new titles in each half-year (cont.)



(b) By genre



(c) By author market position

1.B Topic Modeling

Text documents

Each book in the data has two associated text documents: bookshelf labels and a description. Figures Figure 1.B9 and Figure 1.B10 show word clouds of shelf labels and descriptions for four bestsellers between 2010 and 2012. Text documents are preprocessed using standard procedures before topic modeling, including tokenization, lowercasing, stemming, and stop-word removal.

Figure 1.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

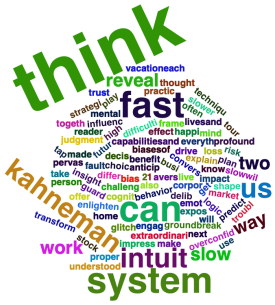
(c) *Thinking, Fast and Slow*, Daniel Kahneman, Farrar, Straus and Giroux, 2011

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

Figure 1.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



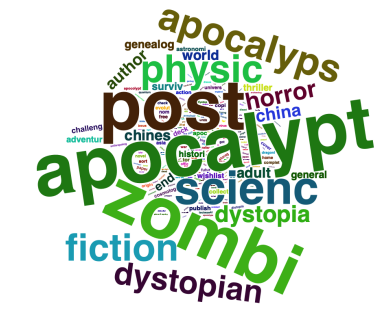
(c) *Thinking, Fast and Slow*, Daniel Kahneman, Farrar, Straus and Giroux, 2011

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

Genre topics

Figure 1.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 1.B12 shows the word probabilities of the most frequent words in all 50 topics.

Figure 1.B11: Examples of genre topic word clouds



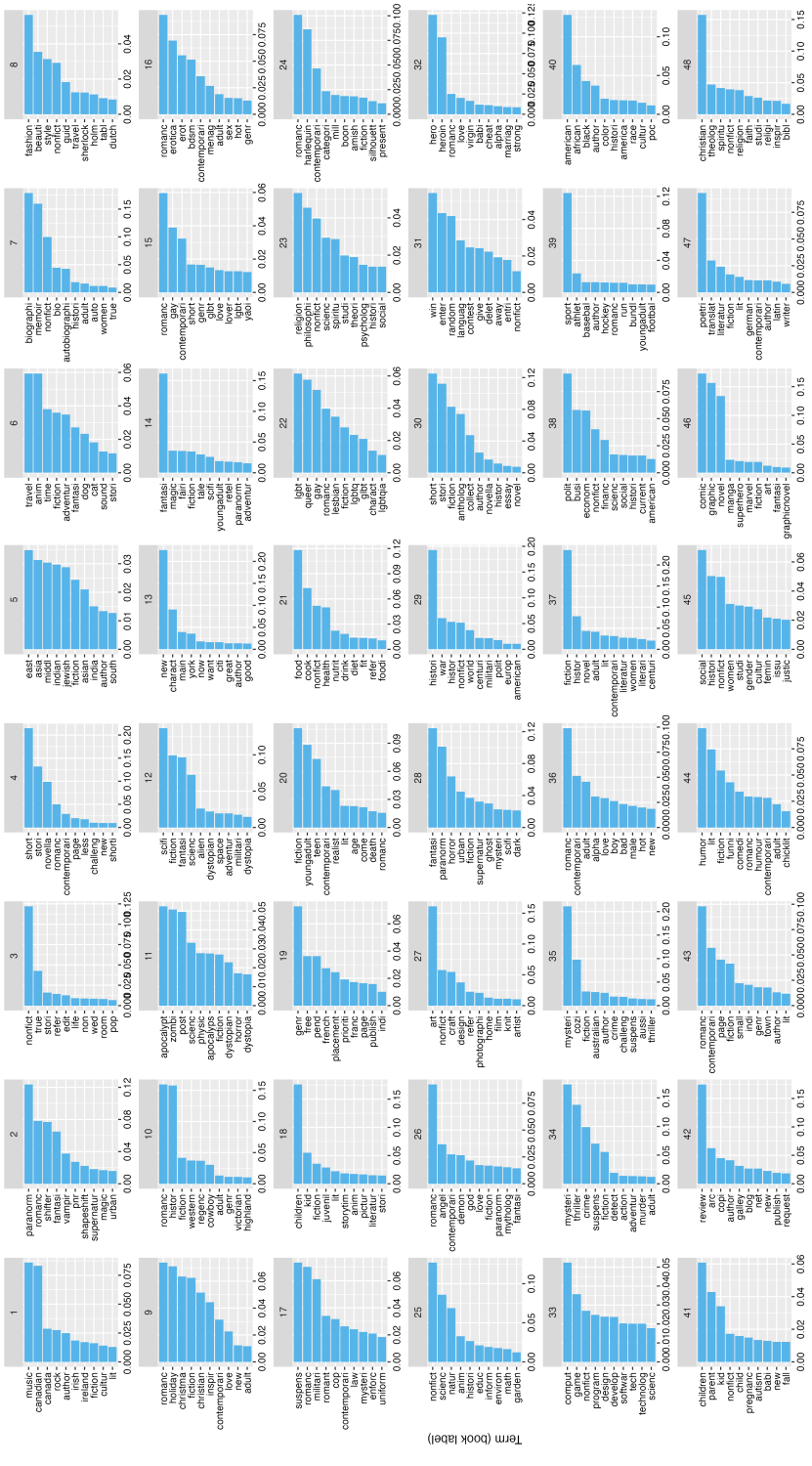
(a) Topic No. 11

(b) Topic No. 23

(c) Topic No. 33

(d) Topic No. 45

Figure 1.B12: Genre topic word probabilities from the LDA model



Content topics

Figure 1.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 1.B14 shows the word probabilities of the most frequent terms in all 50 topics.

Figure 1.B13: Examples of content topic word clouds

(a) Topic No. 1

(b) Topic No. 3

(c) Topic No. 21

(d) Topic No. 38

Figure 1.B14: Content topic word probabilities from the LDA model

1.C More Descriptive Evidence

Event study of the merger

Table 1.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 × $Year_{2010.5}$	−0.011	0.016*
	(0.006)	(0.008)
PRH × $Year_{2011}$	−0.013*	0.005
	(0.006)	(0.008)
PRH × $Year_{2011.5}$	0.002	0.012
	(0.006)	(0.008)
PRH × $Year_{2012}$	−0.003	0.007
	(0.006)	(0.008)
PRH × $Year_{2012.5}$	−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 × $Year_{2014.5}$	−0.008	−0.024**
	(0.006)	(0.008)
PRH × $Year_{2015}$	−0.010	−0.032***
	(0.006)	(0.008)
PRH × $Year_{2015.5}$	−0.004	−0.022**
	(0.006)	(0.008)
PRH × $Year_{2016}$	−0.010	−0.040***
	(0.006)	(0.008)
PRH × $Year_{2016.5}$	−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 1.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 × $Year_{2010.5}$	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 × $Year_{2011.5}$	−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 × $Year_{2012.5}$	−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 × $Year_{2014.5}$	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 × $Year_{2015.5}$	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 × $Year_{2016.5}$	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. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

More specifications of match formation

Table 1.C9 and Table 1.C10 present alternative specifications of match formation. In Table 1.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 1.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 1.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 1.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: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Direction regression of book performance

Table 1.C11: Book performance

	r (log(Ratings count)) (1)	s (Average rating) (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: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

1.D Estimation Details

Gibbs samplers

The prior distributions of parameters $f_0(\boldsymbol{\theta})$ and the augmented posterior are given in Section 1.4. 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 (1.21). For each variable, I collect terms and obtain a kernel that is in the same parametric family as their prior distributions.

Conditional distributions of $\boldsymbol{v}$

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

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

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:

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

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 (1.21) yields the following general form:

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

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,

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

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

For γ^r,

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

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

For γ^s,

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

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

For δ,

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

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

And for ω,

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

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

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,

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

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

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

And for σ_2^2,

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

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

1.E Additional Counterfactual Simulations

Refer to Section 1.5 for the assumptions and implementation of the counterfactual simulations.

Counterfactual 1: synergistic integration

Table 1.E12: Simulation results of synergistic integration

	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

Table 1.E13: Simulation results of synergistic integration by author market position

	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

Notes: In Panel B, publishers refer to authors' destination assignments in the post-merger equilibrium.

Counterfactual 3: Random House takeover

Table 1.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 1.E15: Simulation results of Random House takeover by market position

	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

Notes: In Panel B, publishers refer to authors' destination assignments in the post-merger equilibrium.

Chapter 2

THE IMPACT OF PRIVACY PROTECTION ON ONLINE ADVERTISING MARKETS

2.1 Introduction

Privacy protection is a key topic in the current policy discussions in the digital landscape. Much of the debate surrounds the use of third-party cookies, a device long employed by internet companies to track user behavior across the web, collect user information, and target them with highly personalized ads. Heightened concerns surrounding digital privacy have spurred policy debates and initiatives to curb the pervasive use of third-party cookies. A wave of data privacy legislation has been introduced or proposed in the European Union and across the United States to limit the use of third-party cookies.[1] In the private sector, browsers such as Apple's Safari and Mozilla Firefox have disabled third-party cookies by default, while Google had also planned to follow suit by phasing out third-party cookies in Chrome, currently the market-leading web browser. The potential removal of third-party cookies, sometimes dubbed "Cookiepocalypse" in the industry, sparked widespread resistance from industry stakeholders and was postponed several times because it undermines the foundation of the online advertising market. Moreover, removing third-party cookies—a decentralized protocol—could lead to industry concentration in the online ad supply chain, triggering antitrust sirens from legislators and government agencies.[2] Ultimately, these competitive concerns appear to have played a significant role in Google's decision to pause, and possibly reconsider, the full deprecation of third-party cookies.[3]

---

*This chapter is based on joint work with Miguel Alcobendas, Shunto J. Kobayashi, and Matthew Shum, whose contributions are gratefully acknowledged.

[1] See the General Data Protection Regulation (GDPR) of the European Union, the California Consumer Privacy Act of 2018 (CCPA), the Colorado Privacy Act (CPA), and the Virginia Consumer Data Protection Act (VCDPA).

[2] The EU has launched an antitrust probe into Google's plan to ban third-party cookies in Chrome. In the United States, federal lawmakers have also voiced antitrust concerns over the plan in a 2020 report by the US House Subcommittee on Antitrust.

[3] As of April 2025, Google announced it would no longer proceed with a standalone phase-out of third-party cookies in Chrome, citing industry feedback, regulatory changes, and advances in privacy technology. These developments suggest a recalibration rather than a full abandonment of Google's privacy initiatives, as competitive and regulatory pressures continue to shape the landscape. See `https://privacysandbox.com/news/privacy-sandbox-next-steps/`.

In this paper, we investigate the welfare consequences of the once-planned removal of third-party cookies and introduction of alternative tracking technologies under its "Privacy Sandbox" initiative. Although the plan to phase out third-party cookies has been paused indefinitely, studying its potential effects remains valuable: the proposed changes offer a rare counterfactual thought experiment for analyzing both the design of privacy policies and the economic dynamics of the online advertising market. Our key contribution is to quantify the unequal distributional effects on the demand side of the online advertising market, which encompasses advertisers and their intermediaries who purchase advertising opportunities and match them with advertisers. Although framed as a potential advance for consumer privacy, the removal of third-party cookies carries negative spillover risks, including the strengthening of information monopoly and the entrenchment of anti-competitive practices of large companies.[4] Curtailing the use of third-party cookies undermines firms' ability to target consumers, thereby reducing the surplus of advertisers and their intermediaries. Notably, major tech companies, such as Google, can continue to gather rich behavioral data directly through their widely used products (e.g., the Google search engine, Gmail, and YouTube), whereas other smaller intermediaries have no such recourse. Although the proposed new technology might partially offset the loss, we demonstrate that this is insufficient to diminish the information advantage enjoyed by large players.

To this end, we analyze a large sample of detailed bid-level data of online banner ad auctions from Yahoo, a prominent online news and media publisher. Online ads are sold via auctions: online publishers offer ad spaces when users access their websites, and advertisers bid to determine whose ad is shown. To streamline the process, advertisers use *demand-side platforms (DSPs)* to participate in auctions and bid on their behalf. Third-party cookies enter the process by allowing DSPs to retrieve information associated with the user and more accurately evaluate the ad opportunity. Our first set of empirical results confirms the value of third-party cookies to advertisers. We find that bidders are more likely to submit a bid and bid a higher amount in auctions with third-party cookies. Comparing DSPs' bidding decisions for users with third-party cookies to those without, we find that third-party cookies increase DSPs' bids by around 30% on average. The highest bid, which translates into the publisher's revenue, increases by as much as 80% on average.

Our primary interest is the revenue and welfare effects after Google blocks third-

[4]In January 2023, the U.S. Department of Justice brought an antitrust case against Google, alleging monopolization of the publisher ad server and ad exchange markets.

party cookies on Chrome and introduces alternative tracking technologies on the browser. Because the plan is yet to transpire and the bidders' underlying valuations are not observed, we adopt a structural approach to recover valuations and compute the counterfactual revenue and welfare for players in the market. Our empirical model is a first-price auction model with asymmetric bidders. We enrich the model with two essential features of the advertising market: bidder heterogeneity and auction heterogeneity. We characterize the equilibrium as a system of differential equations and adopt a numerical approach to approximate the bidding functions. The recovered valuation distributions and bidding strategies are consistent with the intuition that bidders value impressions with cookies more and bid for those more aggressively.

We then simulate the effect of "Cookiepocalypse," a third-party cookie ban on Chrome without any alternative means to track users. We consider two counterfactual specifications: a baseline *symmetric* ban in which all bidders are affected by the cookie ban and no longer receive cookie information, and an *asymmetric* ban in which one privileged bidder continues to observe cookie information for Chrome users. The second scenario emulates the information advantage enjoyed by a "Big Tech" player in the market. In the absence of third-party cookies, large firms like Google still have first-party access to user information inaccessible to other online advertising businesses. For each simulation, we solve the auction model under the counterfactual valuation distributions without third-party cookies.

We find a large negative effect worthy of the name Cookiepocalypse: in the baseline symmetric specification, such a ban would reduce the publisher's revenue by 54% and advertiser surplus by 40%. The asymmetric specification illustrates the egregiously unequal welfare distribution and anti-competitive impact of the cookie ban. The privileged bidder with exclusive access to Chrome users' data wins auctions twice as often and earns even more surplus compared to the no-ban status quo. Our results confirm and justify the antitrust concerns raised by Google's plan.

Our second counterfactual builds upon the first and introduces an alternative tracking technology that provides limited behavioral information on Chrome users. Google is developing a set of tools under the "Privacy Sandbox" initiative to replace third-party cookies. The spirit of its proposed technologies is to generate groups of users with similar interests, giving advertisers a way of targeting them without exposing details on individual users. We find that such a more privacy-friendly tracking technology would indeed soften the impact of "Cookiepocalypse" in terms of both

welfare and concentration.[5] The revenue loss decreases to 13% from 54% in the first counterfactual and that advertiser surplus falls from 40% to 8%. Furthermore, although the informationally advantageous bidder still gains more surplus compared to the status quo, other bidders' performance is only mildly impacted. Our results demonstrate the importance and benefits of providing advertisers with an alternative means to target users in order to mitigate the revenue and competitive impacts of the ban.

A limitation of this study is that, due to data constraints, we primarily examine the supply-side dynamics of the online advertising market and do not directly evaluate the implications for consumer welfare. A comprehensive assessment of consumer welfare effects would require additional data and modeling along several dimensions. First, targeted advertising reduces consumer search costs and improves product-consumer matching; consequently, eliminating third-party cookies could diminish consumer welfare.[6] Second, restricting cookies may reduce the availability of free online content financed by advertising, with potential consequences for content availability, quality, and variety. Analyzing this channel would require data on publisher behavior and consumer browsing patterns.[7] Third, while consumers may value greater privacy, estimating the magnitude of these preferences would require experimental or stated preference data on how users trade off privacy against other benefits.[8] We view our study as complementary to these analyses and as an important first step toward understanding the broader welfare consequences of changes to digital advertising technologies.

[5]There are additional antitrust implications over Google's becoming the dominant data vendor for its Privacy Sandbox product. For instance, these concerns led to antitrust investigations by the UK and EU regulatory authorities (`https://www.wsj.com/articles/google-chrome-privacy-plan-faces-u-k-competition-probe-11610119589`). These implications, while interesting, are outside the scope of the present paper.

[6]See, for example, Jeziorski and Segal (2015), Honka, Hortaçsu, and Vitorino (2017), and Tuchman, Nair, and Gardete (2018) for analyses of consumer-ad interactions and purchase decisions to quantify the welfare effects of advertising.

[7]This would involve modeling publishers' product offerings and potential exit decisions under reduced advertising revenues. Furthermore, much of ad-sponsored content is free. For analyses of how changes in free content influence consumer welfare, see Allcott et al. (2020) and Brynjolfsson, Kim, and Oh (2024), who quantify the economic value of free online services.

[8]Privacy valuation is further complicated by the "privacy paradox," wherein users' stated privacy concerns often diverge from their observed online behavior (Barth and Jong, 2017). For empirical efforts to estimate privacy preferences, see Athey, Catalini, and Tucker (2018) and Lee and Weber (2024).

Related literature

Our paper contributes to several existing strands of literature. First, our article contributes to the literature on targeting in advertising.[9] Many empirical studies find positive effects of targeting for advertisers and publishers (Rutz and Bucklin (2012), Lewis and Reiley (2014), Ghose and Todri-Adamopoulos (2016), Wernerfelt et al. (2024)). Our first set of empirical results is consistent with this strand of literature. Levin and Milgrom (2010), on the other hand, discuss trade-offs in narrower versus broader (or "conflated") targeting and argue that the former thins out markets and reduces competition and prices. Rafieian and Yoganarasimhan (2021) empirically confirm this prediction and show that the optimal level of targeting is not necessarily the finest level. Our results suggest that third-party cookies do not suffer from the problem of market-thinning.

Methodologically, our empirical approach connects with the structural empirical literature on auctions.[10] We model ad auctions via a first-price auction model with a binding reserve price, and we incorporate observed heterogeneity as well as unobserved heterogeneity (Krasnokutskaya, 2011; Hu, McAdams, and Shum, 2013; Haile and Kitamura, 2019). In addition, similarly to Athey, Levin, and Seira (2011), Krasnokutskaya and Seim (2011), and Kong (2020), we allow the valuation distributions to differ across bidders to capture the observed difference in their bidding behaviors.[11] To overcome the complexities introduced by auction and bidder heterogeneity, for both estimation and counterfactual analysis, we employ Mathematical Programs with Equilibrium Constraints (MPEC) developed by Hubbard and Paarsch (2009), Hubbard, Kirkegaard, and Paarsch (2013), and Hubbard and Paarsch (2014) to obtain equilibrium bidding strategies numerically.

Our work also contributes to the growing literature on the economics of privacy and data protection policies.[12] Several papers study the effect of restricting third-party cookies in online advertising and find a loss ranging from 4 percent to 66 percent

[9] See Goldfarb (2014) and section 6 of Goldfarb and Tucker (2019) for reviews of this literature on targeting in online advertising.

[10] There are a number of surveys of this literature, including Hong and Paarsch (2006), Athey and Haile (2007), and Perrigne and Vuong (2019).

[11] While our study takes the existing auction format (first-price) as given, in the particular context of online ad auctions, there is a strand of theoretical literature studying auction design (Celis et al. (2014), Abraham et al. (2020)).

[12] See Acquisti, Taylor, and Wagman (2016) and Brown (2016) for reviews of the economics of privacy and Goldfarb and Que (2023) for a review of the economics of digital privacy. Several authors (Goldberg, Johnson, and Shriver, 2019; Aridor, Che, and Salz, 2020) study the impact of the European Union's General Data Protection Regulation (GDPR) on web traffic and ad revenue. See Johnson (2022) for a survey of studies on the economic consequences of GDPR.

(Beales and Eisenach, 2014; Marotta, Abhishek, and Acquisti, 2019; Johnson, Shriver, and Du, 2020). The industry estimate is closer to the upper end, where a study by Google finds that disabling third-party cookies results in an average loss of 52% (Ravichandran and Korula, 2019). While most of these papers are retrospective studies using historical data, our paper provides a counterfactual scenario of the much-discussed Chrome cookie ban which, while planned, has yet to take place.

Finally, this article also connects with the emerging literature on the anti-competitive practices of big tech firms, particularly through the channel of data collection and privacy policy. Consent requirements may favor large firms (Campbell, Goldfarb, and Tucker (2015), Goldberg, Johnson, and Shriver (2019), Kesler, Kummer, and Schulte (2019)). Johnson, Shriver, and Goldberg (2022) and Peukert et al. (2022) show that the GDPR has led to a greater market concentration in the media tech industry, with Google emerging as a clear winner from the policy. Our article is the first to structurally evaluate the impact of Chrome's plan to remove third-party cookies from an antitrust point of view, connecting privacy policy with competition and demonstrating the skewed distribution of profits due to information monopoly.

2.2 Market Background

Online ad auctions

Our analysis focuses on real-time auctions of banner ad space shown to users when they browse web pages. Banner ads are displayed in rectangular boxes between or on the side of the main text. In industry parlance, the ad space for sale is called an *impression*—each time an ad is displayed on the user's screen, it is counted as one impression. The seller is the *publisher* whose web page is browsed by the user and who has an ad space for offer (Yahoo, in our case). The bidders are *advertisers* who compete for the ad space to impress the user. The auctions are mediated through an *ad exchange*, the "auction house" for ad spaces. Auctions at the Yahoo ad exchange, which are the focus of this paper, are in the *first-price sealed-bid* format.

The process of online ad auctions can involve many parties interacting automatically in real time. The auction is triggered when the user opens the web page through her browser. The publisher packages the offer of an ad space along with information about the user and sends it to the ad exchange.[13] The ad exchange then sends out

[13]The offer is usually made through a supply-side platform server that acts on behalf of the publisher. This step is not relevant to our purpose. A data management platform could also be involved to retrieve stored information of the user that may be of interest to the advertisers. The supply-side platform packages the ad space offer with all relevant information and sends it to the ad

a bid request to potential bidders (DSPs), inviting them to submit a bid. Given the large volume of auctions and the complexity of online bidding, advertisers do not participate directly in these auctions, but rather via *demand-side platforms (DSPs)*, which bid on behalf of their advertiser clients.[14] Using information about the user ready to view the ad, the DSP selects the most suitable advertiser for that impression and calculates the optimal bid for the ad space, considering competition from other DSPs. In any auction, DSPs typically submit only one bid on behalf of one of their advertiser clients.[15] In what follows, we use the terms advertisers and DSPs interchangeably and abstract away from the distinction between the DSPs and their advertiser clients.

DSPs are heterogeneous based on their purpose, specialty, and scope, and in this paper, we highlight that such heterogeneity is reflected in their bidding behavior. DSPs fall into three categories: general-purpose DSPs, rebroadcasters, and specialized DSPs. *General-purpose DSPs* provide a wide range of targeting options and optimization tools to help advertisers reach their target audience. They are typically used by large and medium-sized advertisers with sizable budgets and broad campaign objectives. *Rebroadcasters*, as the name implies, rebroadcast advertising opportunities to their own ad exchanges and consolidate bids from multiple DSPs participating in them, acting as intermediaries that increase market thickness. Rebroadcasters often provide additional services to help other DSPs target users. *Specialized DSPs* focus on reaching potential customers who have indicated specialized interests or previously interacted with a brand or website. They are particularly valuable for e-commerce advertisers looking to re-engage potential customers as well as subscription-based services to retain existing subscribers.

Anticipating our empirical implementation, we further categorize general-purpose DSPs and specialized DSPs by their size as either large or small. The size of the DSP captures the budget, experience, and sophistication of the DSPs. These aspects are relevant to their valuation distributions of impressions as well as bidding strategies, which are crucial in our empirical exercise below.

---

exchange.

[14]Many major internet companies, e.g., Amazon, Facebook, and Google, own DSP services. These DSPs bid for ad spaces on their own companies' and other publishers' websites. Yahoo also maintains its own DSP.

[15]Decarolis, Goldmanis, and Penta (2020) and Decarolis and Rovigatti (2021) study the potential anti-competitive effects of the delegation between the advertisers and the DSPs.

Cookies and behavioral targeting

To make their ads more effective, advertisers use *cookies* to track user activities and implement behavioral targeting. Cookies are small files of data created by a web server and stored on a user's device when they browse a website. They facilitate personalized experiences by associating user activity with an identifier stored in a database. For example, if a user visits a news website for the first time and selects English as her preferred language, the website stores this information in its server and saves a cookie file on the user's device. The next time the user visits the website, it will read the local cookie file, identify the user with the information in the database, and automatically select English as the preferred language. This type of cookie is accessible only by this specific news website and is called a *first-party cookie* because it is hosted and used exclusively by the website. They enhance usability by remembering settings like language preferences or login credentials. These cookies, which do not enable cross-site tracking, are generally uncontroversial.

In contrast, *Third-party cookies*, which originate from external entities embedded in a website (e.g., ad servers or social media widgets), are the subjects of intense scrutiny because of their role in user activity tracking and behavioral targeting. To continue the example above, alongside its own content, the news website also contains elements embedded by third-party servers, such as banner ads or social media share buttons. These third-party servers can store their own cookies to identify users and track their activity on the news website. What distinguishes third-party cookies is their ability to track user activities across a range of websites. For instance, if a user visits a retail website and browses for headphones, the third-party server would store this information and later recognize the same user when she returns to the news website, serving her an ad for those headphones.This cross-site tracking enables advertisers to construct detailed user profiles beyond basic demographic information (e.g., gender, age) by tracking browsing history, shopping behavior, and content engagement. As a result, third-party cookies play a central role in behavioral targeting in online advertising, but are also widely viewed as a threat to consumer privacy.

While users can voluntarily provide their demographic information like age and gender to web service providers and contextual details like geo-location, this data is generally less valuable to advertisers compared to behavioral tracking enabled by third-party cookies. As noted in Shiller (2020), highly personalized pricing based on web-browsing histories has a significantly greater impact on advertising

effectiveness compared to broad demographic segmentation. This distinction is critical from a policy perspective: while demographic and contextual data will likely remain accessible even under stricter privacy regulations, the ability to track users across websites via third-party cookies is subject to current and potentially further regulatory intervention, such as browser-imposed restrictions or legal constraints on cross-site tracking (see the following subsection). Our study focuses on this latter form of data collection, analyzing its role in online advertising and the implications of potential policy interventions.

Privacy protection

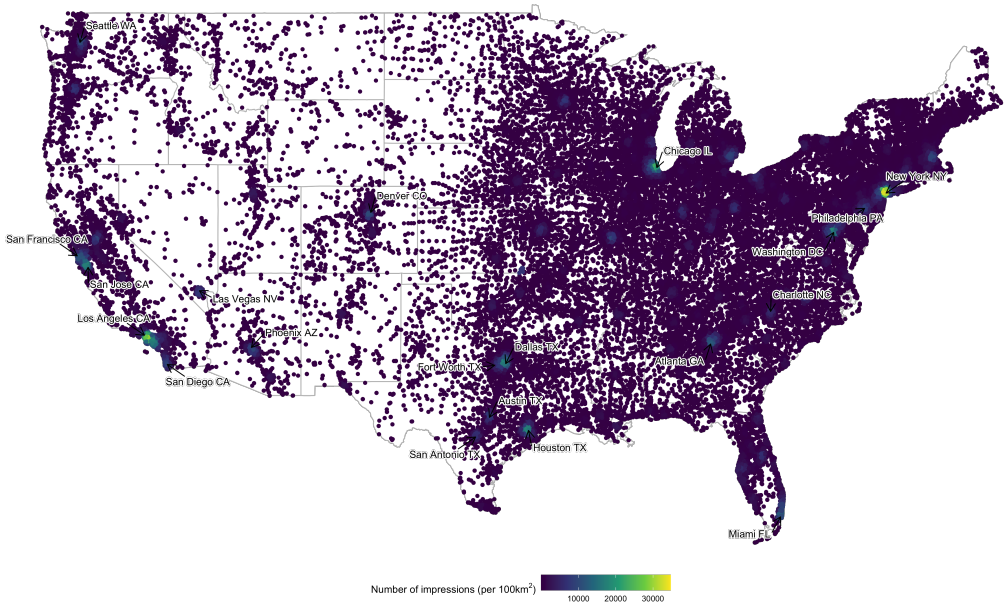
Given the controversial nature of third-party cookies and the growing concern over privacy breaches, many internet entities have either eliminated or curtailed third-party cookies in recent years. Web browsers have been at the forefront of this move. Safari and Firefox (which we refer to as the *blocked* browsers) have already blocked third-party cookies for their users and effectively shut down behavioral targeting by blocking the execution of scripts embedded by third-party servers. Third-party cookies are mostly unavailable for users of blocked browsers. On the other hand, as of 2022, Chrome, together with a few other browsers including Microsoft Edge (the *allowed* browsers), still enables third-party cookies by default. Third-party cookies are generally available on these browsers but could still be absent for a host of reasons.[16]

In addition to private-sector initiatives, the CCPA and other similar privacy regulations require large websites like Yahoo to implement a "Do Not Sell My Personal Information" link that enables users to opt out of the sale of their personal information. Under such an opt-out arrangement, publishers are not allowed to monetize the user's personal information (cookie, IP address, or precise geo-location data) by sharing it with third parties.[17] When cookies are no longer employed, DSPs have significantly less information about users and cannot engage in accurate behavioral ad targeting. In our empirical analysis below, we will exploit the variation in third-party cookie availability to evaluate the effect of behavioral ad targeting.

[16]For example, third-party cookies could be unavailable if the user chooses to block third-party cookies in their browser settings, or browses in private (incognito) mode, or has recently cleared cookies in her browser.

[17]Internet companies can still use broad geographical location (e.g., city) and contextual information of ad opportunities coming from these users for targeted ads at a broader stroke.

Figure 2.1: Geographical distribution of impressions



Note: Each dot represents the number of impressions originating within the 10 by 10 km^2 area around the dot during the week.

2.3 Data and Descriptive Statistics

We employ bidding data from banner ad auctions on sixteen websites of Yahoo, including Homepage, News, Finance, etc. We focus on a specific display ad format known as medium rectangular (MREC) units, which has the dimension 300×250 and is displayed to the right of the main content. This is one of the most popular ad formats, and the fixed size and position help us eliminate potential heterogeneity arising from these aspects. We consider a sample of user impressions from the United States during one week in May 2022. Figure 2.1 shows the geographical distribution of our sample, which roughly coincides with the population density of the US. The dataset consists of over 5.5 million bids from about 740,000 auctions.

Table 2.1 presents summary statistics of key variables in the dataset. The variable *bid* is the submitted bid price of an individual DSP. For reasons of confidentiality, we normalize the submitted bids to have a sample mean equal to 1. For every auction, we observe the *number of bidders* (out of a total of 33 DSPs) who entered the auction and submitted a bid, as well as the *winning (highest) bid*. There is substantial variation in the number of actual bidders for each auction, with a mean of 7.5 bidders and a standard deviation of 4.7. Our empirical model will factor in

Table 2.1: Summary statistics

Variable	No. observations	Pct. missing	Mean	Std. Dev.	Min	Median	Max
Auction:							
Bid	5,529,489	0.000	1.000	1.692	0.064	0.589	275.760
No. bidders	736,745	0.000	7.505	4.745	1.000	7.000	26.000
Winning (highest) bid	736,745	0.000	2.052	3.206	0.064	1.211	275.760
Cookie availability:							
Pct. cookie matched	736,745	0.000	0.577	0.404	0.000	0.800	1.000
Cookie matched	736,745	0.000	0.689	0.463	0.000	1.000	1.000
Privacy:							
Opt-out	736,745	0.000	0.089	0.284	0.000	0.000	1.000
Blocked	736,745	0.000	0.215	0.400	0.000	0.000	1.000
Device:							
Computer	736,745	0.000	0.968	0.177	0.000	1.000	1.000
Demographics:							
Female	736,745	0.000	0.125	0.331	0.000	0.000	1.000
Male	736,745	0.000	0.146	0.353	0.000	0.000	1.000
Gender unknown	736,745	0.000	0.729	0.444	0.000	1.000	1.000
Age 24 and below	736,745	0.000	0.001	0.031	0.000	0.000	1.000
Age 25 to 44	736,745	0.000	0.053	0.225	0.000	0.000	1.000
Age 45 to 64	736,745	0.000	0.120	0.325	0.000	0.000	1.000
Age 65 and above	736,745	0.000	0.064	0.245	0.000	0.000	1.000
Age unknown	736,745	0.000	0.761	0.426	0.000	1.000	1.000
Proxies for user information:							
Interest segments (10,000s)	736,745	0.581	2.558	1.100	0.000	2.551	8.741
Months monetized	736,745	0.580	29.118	24.225	0.000	32.000	55.000
Total revenue (normalized)	736,745	0.580	0.000	1.000	-0.685	-0.370	94.183
Average revenue (normalized)	736,745	0.580	0.003	0.063	-26.035	0.000	1.712
Days in database (10,000s)	736,745	0.725	1.742	0.510	0.000	1.912	1.912

this important behavioral pattern and account for bidders' entry decisions.

Two key variables describe the availability of third-party cookies for each impression. The variable *percentage of cookie matched* is the share of DSPs in each auction that successfully matched the user with a profile in their database. Lower values indicate that less user information is available.[18] The variable *cookie matched* is a binary variable indicating whether the percentage of cookie matched is nonzero for the impression. In other words, it indicates whether at least one bidder has a cookie identifier for the user. For ease of interpretation, our empirical analysis will primarily focus on this variable. In what follows, we refer to impressions with cookie matched = 1 as "cookie impressions" and those with cookie matched = 0 as "cookieless impressions."

The variable *opt-out* is a binary variable indicating if the user opts out of behavioral targeting. The variable *blocked* is a binary variable indicating if a browser blocks third-party cookies by default. It is equal to 1 for Safari or Firefox and 0 for other

[18]For two of the DSPs in our sample, data on whether they matched users with profiles using third-party cookies is unavailable. Therefore, we aggregate cookie availability variables at the user level for each impression rather than at the user-DSP level.

browsers. About 9% of auctions are for opt-out impressions, while 20% of auctions involve impressions using browsers that block third-party cookies.

We include additional characteristic variables indicating the amount of information available on the user. Yahoo's database of user profiles (including those without Yahoo accounts) contains its best guess (based on machine learning procedures) of the user's characteristics and proxies well for the user-specific information that can be inferred from third-party cookies. These include *gender* and *age* categories. The variable *interest segments* (in 10,000s) tallies the total number of interest segments that the user belongs to, where each segment is a prediction of the user's likely interest in a particular subject (e.g., automobile, basketball, gardening, etc.) The variable *months monetized* is the number of months that the user has been monetized by Yahoo, and the *total revenue* and *average revenue* are the total and average monthly revenue derived from the user, respectively, where total revenue is normalized with mean 0 and standard deviation 1. Finally, the variable *days in database* (in 10,000s) is the number of days for which the user profile has existed in Yahoo's database. A smaller number of days may imply that less information is available for the user.[19]

In addition to the user-specific characteristics, we observe variables associated with the origination of the impression. These include the *time (hour)* and the *city* of the impression, the *website* (a total of 16 including Yahoo Homepage, News, Finance, etc.) that published the impression, the device (*computer*) which indicates the user browsed with either a computer or a smartphone/tablet, and the *browser* (Safari, Firefox, Edge, Chrome, and others) with which the user accessed the web page.

Because our analysis focuses on the impact of Google's plan to terminate third-party cookies on Chrome, in Table 2.2, we show the mean statistic of key impression

[19]We note two caveats of these user-specific variables. First, while these variables quantify the user information observed by Yahoo's DSP, in our empirical analysis, we use these variables to proxy for what *any DSP* knows about these users, i.e., we assume all the DSPs observe the same information as the Yahoo DSP. Without data from other DSPs, it is impossible to validate this assumption; however, since many of the users in our dataset have registered Yahoo accounts, we believe that the information that Yahoo has on these users represents a "best case" (upper-bound) on the information that any DSP might have on these users.

Second, we observe a large incidence of missing data: about 70% of the users have unknown age and gender information. As age and gender are typically inferred indirectly from users' internet activities using machine learning algorithms, missing values for these variables typically imply that not enough tracking information is known about these users to permit reliable inference. Furthermore, the variables interest segments, months monetized, and revenue are unknown for around 60% of the analyzed users. The lack of such information is often due to users opting out or using browsers that block third-party cookies. To address this problem and as a robustness check, we have also implemented our empirical analysis on the subsample of users with a Yahoo account, for which the overall incidence of missing data is lower, and confirmed the robustness of our results.

Table 2.2: Summary statistics of impression characteristics by browser

	Chrome	Edge	Safari	Firefox	Other
Proportion	0.576	0.199	0.109	0.106	0.009
Cookie matched	0.869	0.847	0.000	0.000	0.842
Opt-out	0.088	0.097	0.056	0.121	0.033
Female	0.137	0.139	0.000	0.000	0.049
Male	0.154	0.170	0.000	0.000	0.065
Gender unknown	0.709	0.691	1.000	1.000	0.886
Age 24 and below	0.001	0.001	0.000	0.000	0.000
Age 25 - 44	0.074	0.050	0.000	0.000	0.015
Age 45 - 65	0.153	0.150	0.000	0.000	0.055
Age 65 and above	0.068	0.117	0.000	0.000	0.048
Age unknown	0.703	0.681	1.000	1.000	0.881

characteristics, broken down by browser. Importantly, Chrome accounts for almost 60% of the impressions in our data and dominates the browser industry, suggesting a substantial impact of Google's plan on the market. Impressions from Safari and Firefox, the two browsers that ban third-party cookies by default, account for roughly 20% of impressions. Accordingly, impressions from Safari and Firefox are missing third-party cookie information, i.e., *cookie matched* = 0, and gender and age information is unavailable for these impressions as well.

Bidding patterns

Table 2.3 presents a comparison of summary auction statistics for impressions with and without cookies. For either category, we calculate the averages and standard deviations (in parentheses) of the bid, the winning bid, the number of bidders, and the entry probability. Impressions with third-party cookie identifiers fare better for all variables of interest. In particular, submitted bids on average are about 25% higher (1.0 versus 0.76) for cookie impressions, and winning bids for cookie impressions are over two times higher (2.5 versus 1.2) than cookieless impressions. The difference arises from both higher submitted bids and a larger number of participating bidders, with bidders more than twice as likely to enter auctions for cookie impressions. Finally, the standard deviations of bids and winning bids are higher for cookie impressions. This is consistent with the idea that DSPs have more information about these users, enabling more precise targeting and, in turn, leading to greater variation in advertisers' valuations and bids.

Figure 2.2a shows the empirical CDFs of submitted bids in the dataset for five

Table 2.3: Comparison between auctions with and without third-party cookies

Variable	Cookie impressions	Cookieless impressions
Bid	1.041	0.764
	(1.715)	(1.566)
Winning bid	2.454	1.166
	(3.487)	(2.278)
No. bidders	9.283	3.558
	(4.315)	(2.933)
Entry probability	0.265	0.102
	(0.441)	(0.302)

Notes: The mean values are reported in both columns and standard deviations are in parentheses below. The bid is averaged at the bid level. The winning bid and the number of bidders are averaged at the auction level. Entry probability is calculated by first constructing a binary variable *Entry* for every auction-bidder pair. It is equal to 1 if the bidder submitted a bid in the auction.

categories of DSPs in the dataset by their type and size (as discussed in Section 2.2): 5 large general-purpose, 10 small general-purpose, 9 rebroadcaster, 3 large specialized, and 6 small specialized DSPs. Consistent with the results in Table 2.3, DSPs tend to bid higher for cookie impressions. In fact, the distribution of bids for cookie impressions first-order stochastically dominates that for cookieless impressions. Figure 2.2a also shows heterogeneity in submitted bid distributions among different groups of DSPs. The differences are driven by a few factors: Large DSPs generally have better access to user information, have more budget and experience, and are more sophisticated in matching advertisers with impressions. Specialized DSPs could focus on some areas of advertising, such as retailing or reconnecting with existing customers (e.g. retargeting). In terms of auction participation, Figure 2.2b displays the frequencies with which the five groups of DSPs participate in auctions for impressions with and without third-party cookies, and it also highlights heterogeneity in entry behavior across DSPs. The observed heterogeneity among the bidding DSPs motivates us to adopt an auction model with asymmetric bidders in the structural estimation exercise discussed below.

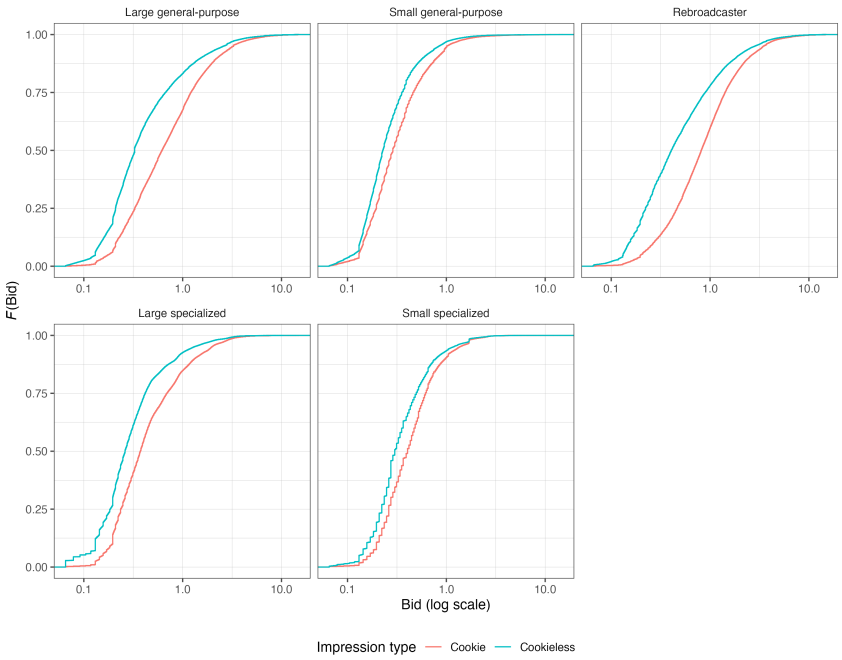
Evidence of the value of third-party cookies

Next, we present reduced-form evidence of the value of third-party cookies to advertisers. Specifically, we run regressions of the following form:

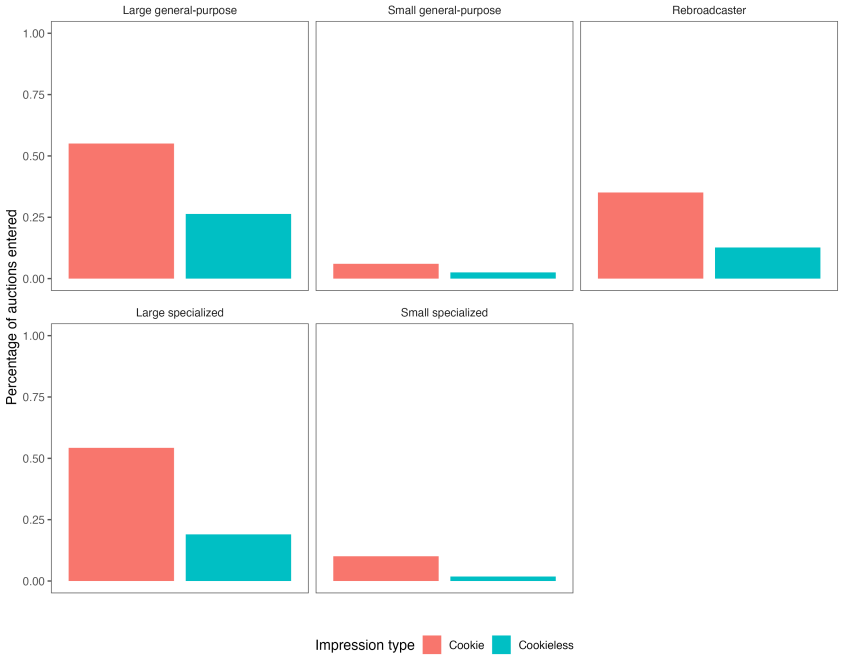
$$y_i = \beta_c \text{Cookie}_i + \mathbf{x}_i'\boldsymbol{\beta} + \alpha_i + \epsilon_i, \tag{2.1}$$

Figure 2.2: Cookie vs. cookieless: observed bidders' behavior by DSP group

(a) Empirical CDFs of submitted bids (log scale)



(b) Average entry frequencies



where i indexes a bidder or an auction depending on the model, y_i is the outcome variable to be specified later, Cookie_i indicates if third-party cookies are available for the impression, $\mathbf{x}_i$ is a vector of covariates that include gender and age information as well as proxies for the amount of information available on the user, α_i includes fixed effects of the hour in the day, the city, the website, and the browser. For models at the bidder level, we also include a DSP fixed effect to capture bidder heterogeneity. Standard errors are clustered by the hour, the city, and the website to account for potential correlations. The variable of interest is Cookie_i, where a positive and significant estimate of β_c would indicate the value of third-party cookies to the advertisers.

We first analyze the effect of cookie availability on submitted bids by taking the outcome variable $y_i = \log(\text{Bid}_i)$ for every bid i in equation 2.1. Table 2.4 columns (1)-(3) report the results of three alternative specifications. Column (1) includes only cookie availability and fixed effects; column (2) adds additional covariates; column (3) further adds a DSP fixed effect to account for bidder heterogeneity. We find quantitatively similar results in these models: having third-party cookies increases submitted bids by around 30%.

Next, we take the outcome variable $y_i = \log(\text{Winning bid}_i)$ for each auction i in equation 2.1 to examine the effect of cookie availability on the highest bid, which translates to the revenue for the publisher (Yahoo). Table 2.4 columns (4) and (5) report the results of two alternative specifications. We find that having third-party cookies increases the highest bids, and consequently Yahoo's revenue, by a substantial 75%. Observe that this effect is more than double the effect in columns (1)-(3). The difference can be attributed to the fact that the bid regression does not account for entry; it only captures submitted bids.

An important feature of the online ad market is that bidders participate in auctions selectively. Recall that table 2.1 showed substantial variation in the number of bidders for different auctions, with a mean of 7.5 bidders and a standard deviation of 4.7. Therefore, we run regression 2.1 with the outcome variable y_i as the number of bidders in each auction i. Table 2.5 columns (1) and (2) report the results of two alternative specifications. We find that, on average, an auction with third-party cookie identifiers induces about 5 more bidders (out of 33) to participate in the auction compared to an impression without. This is broadly consistent with some DSPs' strategies, who simply only enter auctions with third-party cookie identifiers.

Lastly, we examine the effect of cookie availability on the entry decision of bidders

Table 2.4: Regression results for submitted bids and winning bids

Dependent Variables:	log(Bid)			log(Winning bid)	
	(1)	(2)	(3)	(4)	(5)
Cookie	0.335***	0.318***	0.314***	0.887***	0.783***
	(0.028)	(0.046)	(0.031)	(0.018)	(0.042)
Opt-out		0.013	-0.004		-0.021
		(0.026)	(0.020)		(0.024)
Computer		-0.231***	-0.177***		-0.397***
		(0.028)	(0.030)		(0.055)
Gender female		0.097***	0.095***		0.260***
		(0.012)	(0.009)		(0.018)
Gender male		0.069***	0.064***		0.221***
		(0.010)	(0.007)		(0.014)
Age 24 and below		0.066***	0.050***		-0.054
		(0.008)	(0.009)		(0.033)
Age 25 to 44		0.015*	0.009		-0.100***
		(0.008)	(0.007)		(0.016)
Age 45 to 64		-0.010	-0.015**		-0.141***
		(0.006)	(0.006)		(0.019)
Age 65 and above		-0.022**	-0.031***		-0.178***
		(0.009)	(0.008)		(0.016)
Interest segments		-0.002	0.002		0.044***
		(0.001)	(0.001)		(0.005)
Months monetized		0.003***	0.002***		0.003***
		(0.000)	(0.000)		(0.000)
Total revenue (normalized)		-0.037***	-0.027***		-0.043***
		(0.002)	(0.002)		(0.003)
Days in database		-0.051***	-0.044***		-0.053***
		(0.004)	(0.004)		(0.005)
Fixed-effects					
Time (hour)	Yes	Yes	Yes	Yes	Yes
City	Yes	Yes	Yes	Yes	Yes
Website	Yes	Yes	Yes	Yes	Yes
Browser	Yes	Yes	Yes	Yes	Yes
DSP			Yes		
Observations	5,529,489	5,529,489	5,529,489	736,745	736,745
Adjusted R^2	0.10623	0.11052	0.31362	0.24918	0.26361

Notes: The base levels for age and gender are both Unknown. Standard errors are clustered by the hour of the day, the city, and the website and are heteroskedasticity-robust. ***, **, and * indicate statistical significance at the 1, 5, and 10% levels, respectively.

Table 2.5: Regression results for the number of bidders and entry decision

Dependent Variables:	No. bidders		Entry			
	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	OLS	OLS	OLS	OLS	Logit
Cookie	5.796***	5.220***	0.161***	0.145***	0.145***	0.184***
	(0.281)	(0.259)	(0.008)	(0.007)	(0.007)	(0.013)
Opt-out		0.098		0.003	0.003	0.018*
		(0.133)		(0.004)	(0.004)	(0.010)
Computer		-0.926***		-0.024***	-0.024***	-0.045***
		(0.103)		(0.003)	(0.003)	(0.006)
Gender female		0.768***		0.021***	0.021***	0.035***
		(0.046)		(0.001)	(0.001)	(0.003)
Gender male		0.327***		0.009***	0.009***	0.021***
		(0.054)		(0.002)	(0.002)	(0.004)
Age 24 and below		-0.182*		-0.005*	-0.005*	-0.018***
		(0.090)		(0.003)	(0.003)	(0.004)
Age 25 to 44		-0.498***		-0.014***	-0.014***	-0.020***
		(0.071)		(0.002)	(0.002)	(0.004)
Age 45 to 64		-0.627***		-0.017***	-0.017***	-0.023***
		(0.094)		(0.003)	(0.003)	(0.005)
Age 65 and above		-0.805***		-0.022***	-0.022***	-0.028***
		(0.099)		(0.003)	(0.003)	(0.005)
Interest segments		0.432***		0.012***	0.012***	0.012***
		(0.059)		(0.002)	(0.002)	(0.002)
Months monetized		0.019***		0.001***	0.001***	0.000***
		(0.002)		(0.000)	(0.000)	(0.000)
Total revenue (normalized)		-0.180***		-0.005***	-0.005***	-0.006***
		(0.012)		(0.000)	(0.000)	(0.000)
Days in database		-0.256***		-0.007***	-0.007***	-0.008***
		(0.021)		(0.001)	(0.001)	(0.001)
Fixed effects						
Time (hour)	Yes	Yes	Yes	Yes	Yes	Yes
City	Yes	Yes	Yes	Yes	Yes	Yes
Website	Yes	Yes	Yes	Yes	Yes	Yes
Browser	Yes	Yes	Yes	Yes	Yes	Yes
DSP					Yes	
Observations	736,745	736,745	26,522,820	26,522,820	26,522,820	2,652,282
Adjusted R^2	0.44635	0.46616	0.04701	0.04908	0.26756	

Notes: The base levels for age and gender are both Unknown. Column (6) reports the marginal effects of the logit model at the mean or mode values of the explanatory variables using a 10% sample of the dataset. The raw estimates are reported in table 2.A9 of the appendix. Standard errors are clustered by the hour of the day, the city, and the website and are heteroskedasticity-robust. ***, **, and * indicate statistical significance at the 1, 5, and 10% levels, respectively.

in the auctions. In model 2.1, the outcome variable y_i is Entry_i, a binary variable constructed for each auction-bidder pair that is equal to 1 if the bidder submitted a bid in the auction. Table 2.5 columns (3)-(5) report the results of three alternative specifications of such a linear probability model. We find that, on average, bidders are about 14% more likely to participate and submit a bid if the impression has third-party cookie identifiers. Assuming independence between the 33 DSPs, the increase in entry probability translates to an average increase in the number of bidders by $33 \times 0.14 \approx 5$, which is consistent with the estimation above. As a robustness check, we estimate a logit model for auction participation, $\text{Entry}_i = \mathbf{1}\{\beta_c \text{Cookie}_i + \mathbf{x}_i'\boldsymbol{\beta} + \alpha_i + \epsilon_i \geq 0\}$, where ϵ_i follows the standard logistic distribution. Table 2.5 column (6) reports the estimated marginal effects at the mean or mode values of the explanatory variables. The magnitude of the effect of cookies is comparable to those of the linear models. In the appendix, we report the point estimates of the logit model. The estimated coefficient on cookie availability translates into an odds ratio of $e^{1.19} = 3$; that is, the probability that a bidder participates in an auction for a cookie impression is three times higher than that for an auction for a cookieless impression.

2.4 Structural Estimation

Auction model and equilibrium characterization

Our empirical model is an independent private-value auction model with asymmetric bidders and binding reserve price (Krishna, 2009; Hubbard and Paarsch, 2014). We adopt the independent private-value assumption to reflect how users' impressions are horizontally differentiated; for instance, an impression from a male user is more valuable for male fashion brands but less valuable for female fashion brands. In light of the descriptive evidence showing substantial variation in bidding behavior across bidders (Figure 2.2a), we explicitly allow for bidder heterogeneity in the valuation distributions. Finally, our descriptive evidence shows that bidders enter only a fraction of auctions, and auctions in our data have reserve prices that vary across different websites.[20]

Consider an auction of an impression with a reserve price r and $i = 1, 2, \cdots, N$ potential buyers. Suppose each bidder i draws an independent private value v_i from a distribution $F_i(v_i)$ that is differentiable with a density function $f_i(v_i)$. We suppress

[20] An alternative approach is to introduce an entry stage where bidders endogenously decide if they would participate in an auction by comparing the expected profit to the bid preparation cost. This is not applicable in our context because the bid preparation cost in terms of computation and communication with the ad exchange is minimal compared to the reserve price.

the dependency on auction characteristics now and will allow them to depend on both observed and unobserved auction characteristics later. Assume that all valuation distributions across all auctions have a common, compact support $[0, \overline{v}]$. If no one bids above the reserve price, then the impression is not sold. Otherwise, the auction is resolved by the first-price mechanism where the bidder with the largest bid wins the auction and pays his bid b_i.

Suppose that all bidders are in equilibrium and use a bidding strategy $\beta_i(v_i)$ that is differentiable and monotone increasing in his valuation v_i. If the submitted bid b_i is less than the reserve price r, he loses the auction and receives zero profits. Otherwise, the expected profit of bidder i given his bid b_i is

$$\pi_i(b_i) = (v_i - b_i) \prod_{j \neq i} F_j\big(\varphi_j(b_i)\big), \tag{2.1}$$

where, for simplicity, $\varphi_j(b) = \beta_j^{-1}(b)$ denotes the inverse bid function.[21] The first-order condition of the profit maximization problem yields the following equilibrium condition:

$$\frac{1}{\varphi_i(b_i) - b_i} = \sum_{j \neq i} \frac{f_j\big(\varphi_j(b_i)\big)}{F_j\big(\varphi_j(b_i)\big)} \varphi_j'(b_i) \tag{2.2}$$

for $i = 1, 2, \cdots, N$. Equation (2.2) is a system of nonlinear ordinary differential equations in the inverse bid functions $\varphi_1, \cdots, \varphi_N$ that characterizes the Bayes-Nash equilibrium.[22]

In addition to the characterization above, we require two additional boundary conditions in order to solve the system. The lower boundary condition requires that any bidder who draws the reserve price r would bid the reserve price. That is, for $i = 1, 2, \cdots, N$,

$$\varphi_i(r) = r. \tag{2.3}$$

The upper boundary condition requires that all bidders will submit the same bid $\overline{b}$ when they draw the highest valuation $\overline{v}$. In terms of the inverse bid function φ_i, we

[21] Observe that the probability of winning is

$$\Pr(i \text{ wins} | b_i) = \prod_{j \neq i} \Pr\big(b_i > \beta_j(v_j)\big) = \prod_{j \neq i} \Pr\big(v_j < \beta_j^{-1}(b_i)\big) = \prod_{j \neq i} F_j\big(\beta_j^{-1}(b_i)\big) = \prod_{j \neq i} F_j\big(\varphi_j(b_i)\big).$$

[22] The existence and uniqueness of such an equilibrium are generally guaranteed under mild conditions. See Appendix G of Krishna (2009) for a discussion on the existence of such an equilibrium. See Lebrun (1999) for the conditions for the uniqueness of the equilibrium.

have for $i = 1, 2, \cdots, N$,

$$\varphi_i(\overline{b}) = \overline{v}. \tag{2.4}$$

Specifications

In every auction, there is a constant number of $N = 33$ potential bidders who are both qualified and ready to submit a bid.[23] As explained earlier, we model auction interaction at the DSP level rather than the thousands of advertisers that the DSPs bid on behalf of. This assumption stays close to reality and also simplifies the computation. We maintain the assumption that auctions in our sample are independent of one another, abstracting away from potential dynamic considerations of the DSPs.

Consider an auction t. Let x_t denote the observed characteristics known to all DSPs (such as the user's cookie availability, opt-in/opt-out status, browser type, and other characteristics including gender and age.) We let the valuation distribution of each bidder i, $F_{it}(\cdot)$, depend on both observed and unobserved auction characteristics. Specifically, we assume that the log of valuation, $\log(v_{it})$, follows a normal distribution with mean $x_t'\gamma + \alpha_i + u_t$ and variance σ_i^2 and is truncated to the interval $[0, \log(\overline{v})]$, where $\overline{v}$ denotes the maximum possible valuation across all auctions that any bidder could assign to an impression.[24] α_i and u_t are bidder-specific heterogeneity and auction-specific heterogeneity, respectively, and are key features of online ad auctions. We discuss each in turn.

There are two features in our specification that are integral to online ad auctions. First, we account for bidder heterogeneity by allowing asymmetric bidder valuation distributions through α_i and σ_i. We let each bidder i fall into five distinct groups

[23] These DSPs have registered and established a business relationship with Yahoo's ad exchange, and all of them were actively participating in the ad exchange during the sample period.

[24] The parametric approach follows earlier empirical literature of auctions with high-dimensional auction characteristics (Athey, Levin, and Seira, 2011; Krasnokutskaya and Seim, 2011). A nonparametric approach would be computational intractable in our context due to the curse of dimensionality. Moreover, because bidders do not enter auctions when their values fall short of the reserve price, our method parametrically recovers the valuation distributions below the reserve price as well as and the distribution of unobserved auction heterogeneity. These components are necessary for the counterfactual simulations.

The log-normal specification is motivated both by the empirical distribution of bids and by theoretical considerations. First, bids are highly right-skewed, reflecting the fact that some DSPs may place substantially higher value on certain impressions when they believe the user is a strong match. The logarithmic transformation also allows variance of the bid distribution to increase with more information, consistent with the observation that bid variance is higher for cookie impressions (see Table 2.3). Second, we capture the fact that an impression's value is often determined by multiplicative factors. For instance, factors such as interest segments, months monetized, and total revenue may interact multiplicatively in predicting the value of an impression.

according to their type and size: large general-purpose, small general-purpose, rebroadcaster, large specialized, and small specialized (see section 2.2). With slight abuse of notation, the subscript i of the parameters α_i and σ_i denotes the group to which the bidder belongs. As explained, different types of DSPs cater to advertisers of different budgets, objectives, and targeted consumers, which may lead to an ex-ante difference in their valuations for impressions. The size of DSPs is a key dimension that captures their experience and expertise in matching advertisers with impressions.[25]

Second, the term u_t captures the unobserved heterogeneity of the auction and is assumed to take a normal distribution with mean 0 and standard deviation σ_u. It essentially has a multiplicative effect on valuations as in Krasnokutskaya (2011). This allows for bids within an auction to be correlated conditional on observable characteristics, accounting for hidden characteristics commonly observed by the DSPs but not the econometrician. These include geo-location, contextual information from the publisher's page, and time-sensitive market conditions, among others, many of which are available to bidders even in the absence of third-party cookies.

Identification in our model relies on variation both within and across auctions, as well as across bidders. First, because we observe multiple bids within each auction, variation in bids b_{it} across bidders within the same auction t reveals relative differences in valuations v_{it}, which helps identify the shape of the valuation distribution. Second, the unobserved auction-specific component u_t, which is common to all bidders in a given auction but varies across auctions, is identified from shifts in the overall level of bids across auctions. Third, because bidders participate in multiple auctions, this within-bidder variation across auctions allows us to separately identify individual-level heterogeneity α_i.

Before proceeding to the estimation procedure, we highlight two key modeling assumptions that make estimation tractable but may shape the interpretation of our counterfactual simulations. First, we assume that all DSPs observe the same set of auction-level characteristics X_t as we do not have data on DSP-specific information about users (see Footnote 19). In reality, proprietary data partnerships and platform integrations may lead to asymmetries in information access, which could amplify differences in targeting precision across DSPs. Second, we adopt a pure private-value

[25] A fully asymmetric version of the model with a distinct valuation distribution for every bidder is not desirable in our empirical setting. This alternative information structure would require that bidders know all their competitors' exact valuation distributions—a very strong assumption. It is more realistic to assume that bidders only know their competitors' group-specific parameters.

framework, in which each DSP's valuation reflects its own private assessment of how well an impression matches its targeting criteria. This assumption is motivated by the horizontal differentiation of impressions: what is valuable to one advertiser may not be valuable to another. However, if a common value component were introduced and each DSP's valuation also depends on the signals of other bidders, then informational asymmetries could lead to classic winner's curse dynamics, where less-informed DSPs have noisier signals and would shade bids more aggressively or participate less often to avoid overpaying. Under such conditions, restricting access to user data under Cookiepocalypse could paradoxically increase competition or platform revenue by narrowing the informational gap between stronger and weaker DSPs. That said, because large DSPs could retain access to user-level information and continue to enjoy a substantial informational advantage, such a policy change would likely further strengthen their strategic position in the market.

Estimation procedure

We adopt a nested estimation procedure in which the inner loop solves for the inverse bidding strategies $\varphi_{it}(b)$ using the equilibrium characterization (2.2) and the outer loop estimates the valuation parameters using maximum likelihood.

For the outer loop, the valuation parameters are estimated parametrically with maximum likelihood. Specifically, let s_{it} be an indicator variable equal to 1 if bidder i submits a bid in auction t and 0 otherwise. The likelihood of bidder i's observed bidding behavior s_{it} and b_{it} in auction t given u_t is

$$\mathcal{L}_{it}(s_{it}, b_{it}, x_t, u_t; \gamma, \alpha_i, \sigma_i) = \big(F_{it}(r_t)\big)^{1-s_{it}} \big(f_{it}(\varphi_{it}(b_{it}))\varphi'_{it}(b_{it})\big)^{s_{it}}, \tag{2.5}$$

where the first component $F_{it}(r_t)$ corresponds to the probability of non-participation due to valuation below the reserve price r_t, and the second component $f_{it}(\varphi_{it}(b_{it}))\varphi'_{it}(b_{it})$ is the density function of bids obtained by change of variable using the inverse bidding function φ_{it}. Then the joint likelihood of all bidders in auction t is given by

$$\mathcal{L}_t(\boldsymbol{s}_t, \boldsymbol{b}_t, x_t; \gamma, \boldsymbol{\alpha}, \boldsymbol{\sigma}, \sigma_u) = \int \big(\prod_{i=1}^{N} \mathcal{L}_{it}\big)\phi(u_t)du_t, \tag{2.6}$$

where the unobserved heterogeneity is integrated out with respect to its normal density function $\phi(u_t)$ with mean 0 and variance σ_u^2. Since this integral does not have a closed-form solution, we approximate it using Gaussian-Hermite quadrature. We estimate the structural parameters by maximizing the sum of $\log(\mathcal{L}_t)$ over the auctions t in the data.

The inner loop solves for the inverse bidding functions $\varphi_{it}(b)$ for every auction. Because the equilibrium characterization (2.2) admits no closed-form solutions, we adopt a numerical approach to solve the system. Following Hubbard and Paarsch (2009), Hubbard, Kirkegaard, and Paarsch (2013), and Hubbard and Paarsch (2014), we use Mathematical Programs with Equilibrium Constraints (MPEC) to solve for the equilibrium of the first-price auction model with asymmetric bidders. We approximate the inverse bidding functions $\varphi_{it}(b)$ as a linear combination of the first K Chebyshev polynomials scaled to the interval $[r_t, \overline{b}_t]$:

$$\varphi_{it}(b) = \sum_{k=0}^{K} c_{k,it} T_k(b), \tag{2.7}$$

where $T_k(b)$ is the Chebyshev polynomial of degree k scaled to the interval $[r_t, \overline{b}_t]$.[26]

Then, we use the MPEC approach to discipline the Chebyshev coefficients $\boldsymbol{c}_t$ so that the first-order conditions defining the inverse equilibrium bid functions are approximately satisfied, subject to the boundary conditions (2.3) and (2.4). In addition, we impose rationality (players must bid less than their valuation) and monotonicity (bid functions are increasing) as shape constraints on the Chebyshev approximations (Hubbard and Paarsch, 2009; Hubbard, Kirkegaard, and Paarsch, 2013). Specifically, from equation (2.2), we define the deviation from equilibrium

$$G_{it}(b; \boldsymbol{c}_t, \overline{b}_t) = 1 - (\varphi_{it}(b) - b) \sum_{j \neq i} \frac{f_{jt}\big(\varphi_{jt}(b)\big)}{F_{jt}\big(\varphi_{jt}(b)\big)} \varphi'_{jt}(b), \tag{2.8}$$

where $\boldsymbol{c}_t$ are the coefficients of the linear combination of Chebyshev polynomials. Let $\mathcal{B}$ be the set of Chebyshev nodes in $[r_t, \overline{b}_t]$. For every auction t, we solve the following constrained optimization problem to obtain φ_{it} and $\overline{b}_t$:[27]

$$\min_{\boldsymbol{c}_t, \overline{b}_t} \quad \sum_{i=1}^{N} \sum_{b \in \mathcal{B}} \big[G_{it}(b; \boldsymbol{c}_t, \overline{b}_t) \big]^2 \tag{2.9}$$

$$\text{s.t. } \varphi_{it}(r_t) = r_t,\ \varphi_{it}(\overline{b}_t) = \overline{v},\ \varphi_{it}(b) \geq b,\ \varphi'_{it}(b) \geq 0, \text{ for } i = 1, \ldots, N \text{ and } b \in \mathcal{B}.$$

Table 2.6: Estimated parameters of valuation distributions

Parameter	Estimate
γ	
Cookie	1.288***
	(0.004)
Opt-out	-0.047***
	(0.007)
Gender female	-0.121***
	(0.009)
Gender male	0.003
	(0.009)
Age 44 and below	0.030***
	(0.010)
Age 45 to 64	-0.005
	(0.010)
Age 65 and above	-0.037***
	(0.011)
Interest segments	0.063***
	(0.001)
Months monetized	0.004***
	(0.000)
Total revenue (normalized)	0.000
	(0.000)
Days in database	0.402***
	(0.025)
Website fixed effects	Yes
Browser fixed effects	Yes
α	
Large general-purpose	-2.972***
	(0.007)
Small general-purpose	-7.490***
	(0.010)
Rebroadcaster	-4.144***
	(0.007)
Large specialized	-3.185***
	(0.007)
Small specialized	-6.111***
	(0.008)
σ	
Large general-purpose	1.931***
	(0.002)
Small general-purpose	2.587***
	(0.004)
Rebroadcaster	2.342***
	(0.002)
Large specialized	1.383***
	(0.001)
Small specialized	2.089***
	(0.002)
σ_u	0.626***
	(0.001)

Notes: Parameter estimates of the log of valuation, $\log(v_{it})$, which follows a normal distribution with mean $x_t'\gamma + \alpha_i + u_t$ and variance σ_i^2, where u_t is the unobserved auction heterogeneity that is distributed normally with mean 0 and variance σ_u^2. Estimates of website and browser fixed effects are not reported in the table. ***, **, and * indicate statistical significance at the 1, 5, and 10% levels, respectively.

Estimation results

Table 2.6 reports the estimated parameters of the valuation distributions. The estimates are of the expected sign and magnitude. In particular, the estimated coefficient of cookie availability is positive and significant, confirming that third-party cookies increase bidders' valuations. An impression with third-party cookies available raises the mean valuation by as much as 129 percent compared to an impression without cookies. Given bid shading in first-price auctions, the estimate is consistent with the reduced-form estimate of the effect on submitted bids. The estimated intercepts α_i and standard deviation σ_i show substantial differences in the mean and variance parameters of the valuation distribution across different DSP groups, where both small general-purpose and small specialized DSPs have low valuation distributions, reflecting their resource constraints. Lastly, the estimated variance of unobserved auction heterogeneity σ_u, while smaller in comparison to group-specific variances, remains statistically significant and positive. This suggests the presence of unobserved variations in auctions that are not accounted for by group-specific differences in the data.

We next present each bidder group's bidding pattern in response to third-party cookie availability in terms of their valuation distribution, entry probability, and bidding strategy. Following the group classification outlined in section 2.1, we organize the plots by large general-purpose, small general-purpose, rebroadcaster, large specialized, and small specialized DSPs. Each figure shows the outcome variables for both impressions with and without cookies. For illustration, the valuation distribution and the bidding strategy are evaluated at the average values of the covariates.

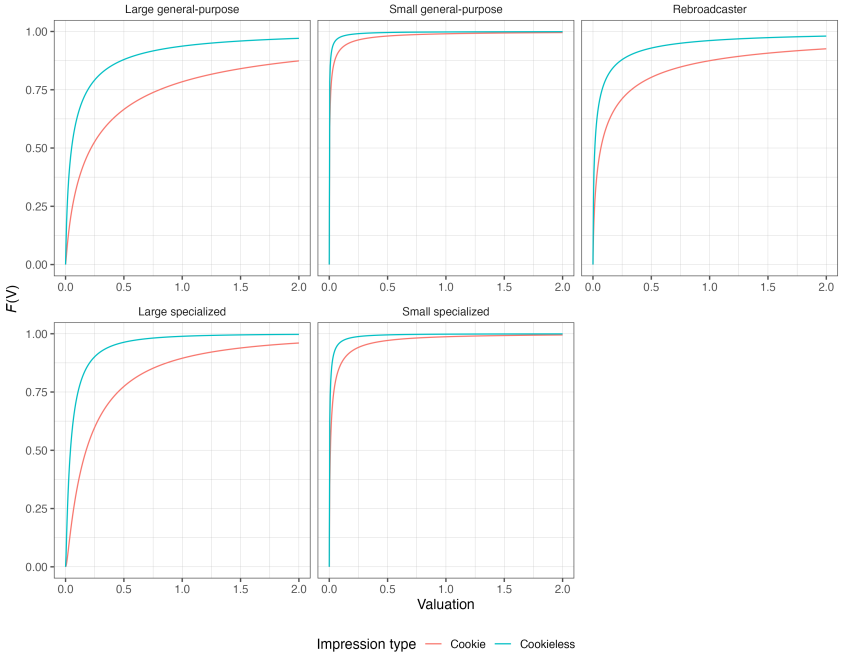
Figure 2.3a shows the cumulative distribution functions (CDFs) of recovered valuation distributions. Figure 2.3b presents the empirical density of fitted entry probability. It is the empirical distribution of the ex-ante probability of a bidder submitting a bid ("entry") prior to the realization of the valuation draw, i.e., $P(v_{it} > r_t) = 1 - F_{it}(r_t)$, the probability that the valuation exceeds the reserve price. For either figure, we observe a clear dominance relationship of cookie impressions over cookieless ones across different DSP groups. Bidders are more likely to place

[26] In the estimation, we set $\overline{v}$ based on the maximum observed bid across all auctions. At the same time, we allow the maximum bid $\overline{b}_t = \beta_{it}(\overline{v})$ to vary and estimate it separately for each auction t. This auction-specific $\overline{b}_t$ ensures that the inverse bidding functions φ_{it} remain sufficiently flexible to capture heterogeneity in bidding strategy across auctions.

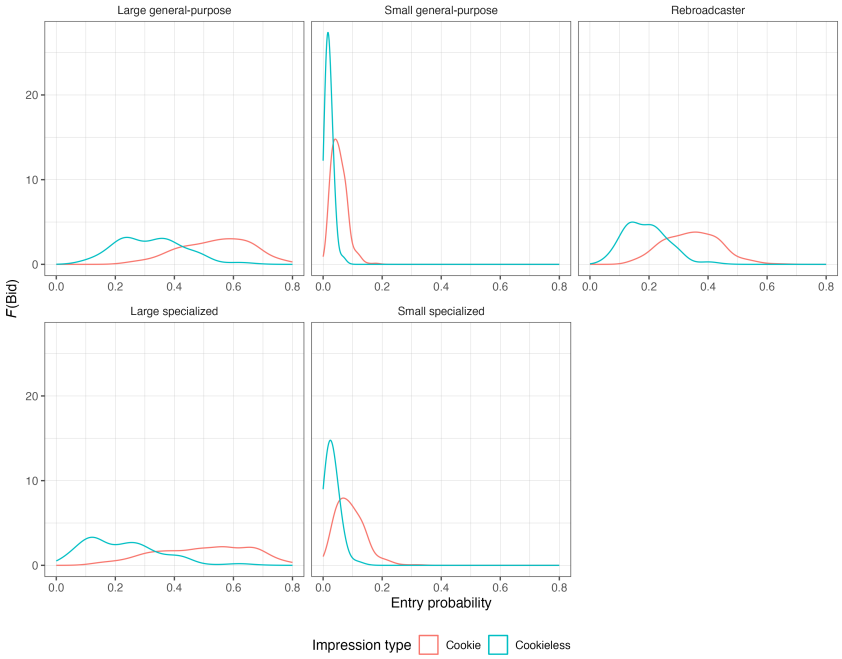
[27] In the implementation, we use the first $K = 5$ order Chebyshev polynomials and 20 Chebyshev nodes for $\mathcal{B}$ to numerically approximate the inverse bid functions. These specifications are sufficiently flexible for approximations in our setting.

Figure 2.3: Cookie vs. cookieless: estimated bidders' behavior by DSP group

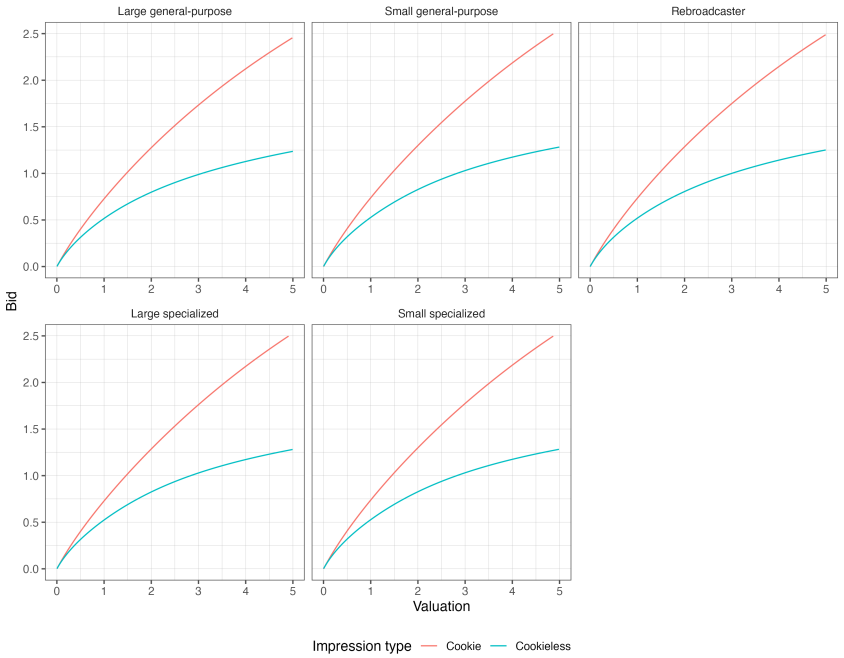
(a) CDFs of valuation distributions



(b) Density of entry probability



(c) Bid functions



Notes: Plots of bidder behavior by DSP groups with estimated parameters. The DSPs are grouped according to their purpose, specialty, and size. See section 2.1 for more details on the classification of DSPs. Subplot (a) shows the cumulative distribution function F_i of valuations at average auction characteristics. (b) shows the empirical density of entry probability, i.e., how likely the valuation exceeds the reserve price and the bidder submits a bid in an auction. (c) shows the bid function β_i at average auction characteristics. See Figure 2.A4 of the appendix for the bid functions of big and small general-purpose DSPs on the same plot.

a higher value and submit a bid in an auction with third-party cookies. There is also substantial heterogeneity across bidder groups. Notably, the effect of cookie availability is more pronounced for large DSPs.

Figure 2.3c presents the bidding strategy β_i and shows that bidders bid more aggressively for cookie impressions.[28] Observe that, for the same valuation, bidders on average place bids on a cookie impression that are about twice as much as those on a cookieless impression. The difference can be attributed to the competition intensity between the two types of auctions, where fewer bidders participate in auctions for

[28]Given the relatively large number of bidders (33 in this market), the average bidding strategies are expected to be similar among bidders, though some differences remain. The system of nonlinear ordinary differential equations (2.2) that characterizes the equilibrium shows that the amount of bid shading of a DSP i depends on *all* its competitor j's valuation distributions and bidding strategies. This strategic interdependence implies that any two DSPs face nearly identical competitive environments, which tends to align their bidding behavior. That said, we do observe some meaningful variation across DSPs. In particular, as shown in Figure 2.A4 in the Appendix, smaller bidders adopt more aggressive bidding strategies to compete against larger bidders, who tend to have higher valuations.

cookieless impressions. Overall, our estimated structural results demonstrate that the difference between the average revenue from the two types of auctions comes from differences in valuations, entry behavior, and bidding strategies.

2.5 Counterfactual Simulations

Using the structural estimates and the MPEC equilibrium solver, we simulate counterfactual scenarios to investigate the welfare redistribution of (1) Cookiepocalypse, the planned removal of third-party cookies from Chrome, and (2) Privacy Sandbox, the implementation of alternative tracking technologies. We show that the proposed changes have significant anti-competitive implications in terms of welfare distribution among advertisers.

For each scenario, we consider three specifications. First, we simulate a status quo scenario as the benchmark (a less noisy version of the status quo in the data), to which we will compare the counterfactual scenarios. We will see that the results from the status quo are comparable to the summary statistics from the actual data. Second, we simulate a *symmetric* ban in which the cookie ban applies to all bidders, and none of them observe the cookie information. Third, we simulate an *asymmetric* ban by designating one bidder from the large general-purpose DSP group as the "Big Tech" DSP who retains access to Chrome users' third-party cookie information, but none of the other bidders observe any cookie information for Chrome users.

The asymmetric ban mirrors concerns raised by antitrust authorities, whereby certain DSPs may have alternative ways to gather and use ad-relevant information about users even when third-party cookies are blocked. For instance, DSPs affiliated with prominent publishers may have extensive user information through *first-party* cookies, which are typically enabled even by browsers that block third-party cookies by default. They may be able to leverage this rich first-party information about users for placing ads not only on their own websites but also on third-party websites, thereby obtaining a large information advantage over DSPs without similar capabilities.[29] A prominent example is Google, which possesses large amounts of first-party information on many internet users via its extensive web ecosystem encompassing the Google search engine, Gmail, YouTube, and more. This unique access to first-party information may allow Google to circumvent the effects of the Chrome third-party

[29]This alternative information collection can be implemented with "digital fingerprinting" methods that track users via IP addresses or device IDs, thus sidestepping cookies altogether. Peukert et al. (2022) observe that the drop in third-party cookie requests after the enactment of GDPR in the European Union was accompanied by a rise in first-party cookie requests.

cookie ban and perhaps even to benefit from such a ban.[30]

To implement the counterfactual simulations, we draw a random sample of 10,000 auctions of impressions from the data. Importantly, this sample includes impressions from all browsers because we want to investigate the market-wide impact on the advertising market. For Chrome impressions (about 58% in the drawn pool), we manipulate their impression characteristics to emulate scenarios of the Cookiepocalypse. For each Chrome auction and each specification, we draw valuations based on the true user characteristics for each bidder and, depending on the scenario and the bidder, mask any third-party cookie information for each user to simulate the effects of the ban. (That is, the cookie availability variable is set to zero. Other user characteristics associated with third-party cookies are set to zero or marked as unknown.) Given the counterfactual valuation distributions, we compute the bidding strategies by solving the system of ordinary differential equations (2.2) that characterizes the equilibrium.

A limitation of our counterfactual analysis is that we adopt a partial equilibrium perspective by holding fixed the set of advertisers associated with each DSP. In practice, however, improvements in Google's targeting capabilities enabled by better access to user data could lead advertisers to shift spending toward DSPs like Google that benefit disproportionately from cookie deprecation. This reallocation could increase Google's market share and potentially allow it to raise prices, further reinforcing competitive advantages. Modeling such dynamics would require additional data on advertiser multi-homing, switching costs, and cross-platform substitution patterns, which are not observed in our dataset. Because our model does not capture these dynamic adjustments, we therefore caution that the counterfactual estimates do not capture the full general equilibrium effects of cookie policy changes and should be interpreted within the partial equilibrium framework. The competitive implications we document may understate the extent to which platform advantages could amplify over time.

Cookiepocalypse, blocking third-party cookies on Chrome

We first investigate the effect of Cookiepocalypse on submitted bids, the number of bidders, the winning bid (which translates into the publisher's revenue), and bidders' surplus. The results of this counterfactual simulation are presented in Table 2.7a.

[30]The anti-competitive implications of Google's plan on the ad supply chain have been closely scrutinized by government agencies. See Jeon (2020) for a more detailed discussion on the market power of Google in the online advertising markets.

Table 2.7: Counterfactual simulation of Cookiepocalypse

(a) Simulated outcome

	Status quo	Symmetric ban	Asymmetric ban
Bid	0.917	0.558	0.588
	(1.487)	(0.761)	(0.831)
No. bidders	7.383	4.771	4.918
	(3.965)	(2.879)	(2.865)
Publisher revenue	2.433	1.101	1.208
	(2.765)	(1.250)	(1.399)
Bidder surplus	3.703	2.234	2.465
	(5.604)	(4.367)	(4.629)

(b) Welfare distribution

	Status quo	Symmetric ban	Asymmetric ban
Winning frequency			
Big Tech DSP	-	-	0.152
Large general-purpose	0.083	0.082	0.076
Small general-purpose	0.003	0.003	0.003
Rebroadcaster	0.048	0.048	0.045
Large specialized	0.028	0.026	0.024
Small specialized	0.004	0.004	0.003
Surplus			
Big Tech DSP	-	-	48,900
Large general-purpose	31,800	18,700	17,600
Small general-purpose	928	559	476
Rebroadcaster	20,200	12,800	12700
Large specialized	5,030	2,150	1920
Small specialized	875	420	369
Full-information surplus			
Big Tech DSP		-	48,900
Large general-purpose		31,000	29,300
Small general-purpose		749	645
Rebroadcaster		18,500	18,000
Large specialized		4,890	4,260
Small specialized		651	548

Notes: Simulated results are based on 10,000 auctions randomly drawn from the data. The Big Tech DSP is drawn from the large general-purpose DSP group. For Chrome impressions, auction characteristics are masked for all bidders in the symmetric ban scenario and are available exclusively to the Big Tech DSP in the asymmetric ban scenario. For each scenario, valuations are updated according to counterfactual characteristics, and outcomes are recomputed using the equilibrium characterization.

We find that the average bid falls from $0.92 in the benchmark to $0.56, representing a 39% decrease, and the number of bidders decreases from 7.4 to 4.8. Altogether, this results in about a halving (-54%) of the average publisher revenue from $2.4 down to $1.1. This estimate is consistent with several studies investigating the potential effect of removing third-party cookies including industrial studies.[31] On the buyer side, advertisers acquiring impressions through DSPs suffer a substantial 40% reduction in their surplus (the difference between valuation and bid), from an average of $3.7 in the benchmark to $2.2 in the first counterfactual.

We next investigate the distributional effect among bidders in terms of their winning frequency and surplus to highlight the unequal impact of the Cookiepocalypse. In Table 2.7b, we report the outcome variables in the asymmetric ban counterfactual scenario separately for the Big Tech DSP and the other five bidder groups. (Recall that the Big Tech DSP is drawn from the large general-purpose DSP group in the benchmark.) In terms of winning frequency, the Big Tech DSP wins twice as often (15.4%) in this scenario compared to the benchmark (8.3%), thanks to its informational advantage of having sole access to the behavioral information of Chrome users. Its total surplus also increases accordingly from $31,800 in the status quo to $48,900 under the asymmetric ban, a 54% increase. At the same time, all the other bidders are impacted negatively by the ban, winning less frequently and receiving lower surpluses compared to the status quo and symmetric ban scenarios. Our results demonstrate that the third-party cookie ban leads to divergent experiences for the informationally advantaged and disadvantaged bidders, where the former benefit from the ban at the cost of the latter.

To further decompose this redistributive effect, we also calculate the "full-information" surplus, that is, the difference between the valuation under cookie availability and the bid in the counterfactual scenario. The gap between the full-information and limited-information surpluses quantifies the loss in bidder welfare due to the inability to make precise matches when DSPs lose the ability to accurately evaluate and target users following the cookie ban. Comparing this difference in Table 2.7b, we see that welfare loss stems primarily from the diminished ability of affected DSP to effectively target users post-cookie ban. The primary factor responsible for the welfare redistribution is the inability of disadvantaged bidders to match with the

[31] Several papers study the effect of restricting third-party cookies in online advertising and find a loss ranging from 4 percent to 66 percent (Beales and Eisenach, 2014; Marotta, Abhishek, and Acquisti, 2019; Johnson, Shriver, and Du, 2020). The industry estimate is closer to the upper end, where a study by Google finds that disabling third-party cookies results in an average loss of 52% (Ravichandran and Korula, 2019).

most appropriate advertisements, rather than the Big Tech DSP monopolizing all the valuable impressions in the market.

Privacy Sandbox, alternative tracking technologies

In the second counterfactual, we replace third-party cookies with an alternative privacy-friendly tracking technology that allows bidders to acquire some behavioral information on the users, albeit without the precision and granularity of the cookie-generated information. Google has proposed a few alternative tracking technologies under its *Privacy Sandbox* initiative since 2021, shortly after its announcement of a third-party cookie ban. A prominent proposal is the *Topic API*.[32] With Topics, the browser will infer a handful of recognizable, interest-based "categories" for the user (such as automotive, literature, rock music, etc.) based on recent browsing history to help sites serve relevant ads. However, the specific sites the user has visited are no longer shared across the web like they might have been with third-party cookies. In essence, this new method allows for tracking and targeting but in a more privacy-conscious and less precise manner than traditional third-party cookies.

In our implementation, because the exact alternative technology has not been finalized and we do not observe the user's interest categories, we follow the overarching principle of these proposed technologies that seek the best of the two worlds. On the one hand, users are afforded some degree of privacy; on the other hand, advertisers continue to observe user characteristics, albeit coarser ones. Specifically, we model this compromise between privacy and personalization by replacing Chrome users' behavioral characteristics with the average characteristics for each Yahoo website (e.g., Yahoo Mail, Yahoo Finance, Yahoo News, etc.). For example, the gender information of a Chrome user visiting Yahoo Finance is replaced by the website's proportions of male and female users. The Big Tech DSP, on the other hand, continues to observe Chrome users' exact characteristics.

The rightmost column of Table 2.8a contains the summary outcomes of the asymmetric ban under the Privacy Sandbox counterfactual. We find that the average bid has fallen from \$0.92 in the benchmark to \$0.82 in the counterfactual, and the number of bidders has decreased from 7.4 to 6.9. Altogether, this results in

[32] See `https://privacysandbox.com/`. Several techniques have been or are being proposed, developed, and experimented with. Google initially experimented with the Federated Learning of Cohorts (FLoC) in 2021 and "received valuable feedback from regulators, privacy advocates, developers and industry. The new Topics API proposal addresses the same general use case as FLoC, but takes a different approach intended to address the feedback received for FLoC. Chrome intends to experiment with the Topics API and is no longer developing FLoC."

Table 2.8: Counterfactual simulation of Privacy Sandbox

(a) Simulated outcome

	Status quo	Symmetric ban	Asymmetric ban
Bid	0.917	0.815	0.824
	(1.487)	(1.276)	(1.298)
No. bidders	7.383	6.838	6.872
	(3.965)	(3.636)	(3.636)
Publisher revenue	2.433	2.061	2.099
	(2.765)	(2.309)	(2.363)
Bidder surplus	3.703	3.378	3.442
	(5.604)	(5.380)	(5.475)

(b) Welfare distribution

	Status quo	Symmetric ban	Asymmetric ban
Winning frequency			
Big Tech DSP	-	-	0.094
Large general-purpose	0.083	0.083	0.082
Small general-purpose	0.003	0.003	0.003
Rebroadcaster	0.048	0.048	0.048
Large specialized	0.028	0.028	0.028
Small specialized	0.004	0.004	0.003
Surplus			
Big Tech DSP	-	-	36,000
Large general-purpose	31,800	28,700	28,600
Small general-purpose	928	878	826
Rebroadcaster	20,200	18,700	18,800
Large specialized	5,030	4,230	3,940
Small specialized	875	745	715
Full-information surplus			
Big Tech DSP		-	36,000
Large general-purpose		32,600	32,500
Small general-purpose		951	898
Rebroadcaster		20,500	20,500
Large specialized		5,400	5,030
Small specialized		843	805

Notes: Simulated results are based on 10,000 auctions randomly drawn from the data. The Big Tech DSP is drawn from the large general-purpose DSP group. For Chrome impressions, auction characteristics are averaged at the website level for all bidders in the symmetric ban scenario. Exact characteristics are available exclusively to the Big Tech DSP in the asymmetric ban scenario. For each scenario, valuations are updated according to counterfactual characteristics, and outcomes are recomputed using the equilibrium characterization.

a 13% drop in the average revenue per auction from \$2.4 to \$2.1, and the bidder (advertiser) surplus drops by 8% from \$3.7 to \$3.4. In a word, the Privacy Sandbox still results in sizable welfare losses for both the publisher and the advertiser–an expected consequence given the coarser information in the market. On the other hand, the impact is a lot more cushioned compared to that of the Cookiepocalypse counterfactual under which the publisher and the advertiser bear a much heavier loss of 54% and 40%, respectively.

Table 2.8b presents the differentiated impact on DSP groups. Compared to the Cookiepocalypse counterfactual in table 2.7b, Privacy Sandbox alleviates the anti-competitive redistribution as well as the rising market concentration in favor of the Big Tech DSP. For the Big Tech bidder under the asymmetric ban, both its winning frequency (9.4%) and total surplus (\$36,000) increase compared to the benchmark (8.3% and \$31,800, respectively), representing a more than 10% gain, though the advantage is substantially attenuated compared to that under Cookiepocalypse. The disadvantaged bidders also experience noteworthy improvement compared to Cookiepocalypse. Their metrics under either symmetric or asymmetric ban are much closer to the status quo level: Under the asymmetric ban, for example, large general-purpose DSPs enjoy a surplus of \$28,600, below the status quo level of \$31,800, but a substantial alleviation compared to \$17,600 under Cookiepocalypse. Although still a heavy 10% loss from the advertiser's perspective, this set of results suggests that advertising surplus and user privacy may not be fundamentally at odds. DSPs can rely on privacy-friendly technologies and coarser information to implement targeted ads without severely hurting their bottom lines. The anticompetitive redistribution effect, although much ameliorated compared to Cookiepocalypse, is still present and significant.

2.6 Conclusion

We study the impact of privacy protection on online advertising markets. As privacy concerns have mounted in recent years, internet browsers are increasingly moving away from third-party cookies, a widely-used tool to track online user behavior across the web and implement targeted ads. In this paper, we investigate the impact of a third-party cookie ban by analyzing online banner ad auctions using a detailed bid-level dataset from Yahoo. We find that auction participation, submitted bids, and revenue are higher when third-party cookies are available. This initial set of results demonstrates the pivotal role of third-party cookies in facilitating online advertising.

We next construct an empirical auction model, analytically characterize the equilibrium, and structurally recover valuation distributions from observed bids in the dataset. To evaluate the impact of the planned phasing-out of third-party cookies from Google Chrome, we perform counterfactual analyses based on the recovered structural parameters. Our results indicate that an outright ban—Cookiepocalypse—would reduce publisher revenue by 54% and advertiser surplus by 40%. However, the introduction of alternative, privacy-conscious tracking technologies under Google's Privacy Sandbox initiative, which delivers coarser user information to advertisers, would mitigate these losses.

We also quantify the redistribution of welfare resulting from the third-party cookie ban in which some large, informationally advantaged bidders could leverage their rich information over their competitors in online ad auctions. We find that these advantaged bidders stand to reap a larger surplus from the ban, whereas other bidders have no such recourse. Because of big tech firms' substantial presence in the ad supply chain and their abundant user information, the plan to eliminate third-party cookies raises antitrust concerns regarding competition and monopoly power in online advertising markets.

References

Abraham, Ittai, Susan Athey, Moshe Babaioff, and Michael D. Grubb (2020). "Peaches, lemons, and cookies: Designing auction markets with dispersed information". en. In: *Games and Economic Behavior* 124, pp. 454–477.

Acquisti, Alessandro, Curtis Taylor, and Liad Wagman (2016). "The Economics of Privacy". en. In: *Journal of Economic Literature* 54.2, pp. 442–492.

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow (2020). "The Welfare Effects of Social Media". In: *American Economic Review* 110.3, pp. 629–676.

Aridor, Guy, Yeon-Koo Che, and Tobias Salz (2020). *The Effect of Privacy Regulation on the Data Industry: Empirical Evidence from GDPR*. Working Paper.

Athey, Susan, Christian Catalini, and Catherine E. Tucker (2018). *The Digital Privacy Paradox: Small Money, Small Costs, Small Talk*. en. SSRN Scholarly Paper. Rochester, NY.

Athey, Susan and Philip A Haile (2007). "Nonparametric approaches to auctions". In: *Handbook of econometrics* 6. Publisher: Elsevier, pp. 3847–3965.

Athey, Susan, Jonathan Levin, and Enrique Seira (2011). "Comparing Open and Sealed Bid Auctions: Evidence from Timber Auctions". In: *The Quarterly Journal of Economics* 126.1, pp. 207–257.

Barth, Susanne and Menno D. T. de Jong (2017). "The privacy paradox – Investigating discrepancies between expressed privacy concerns and actual online behavior – A systematic literature review". en. In: *Telematics and Informatics* 34.7, pp. 1038–1058.

Beales, Howard and Jeffrey A. Eisenach (2014). *An Empirical Analysis of the Value of Information Sharing in the Market for Online Content*. en. SSRN Scholarly Paper. Rochester, NY.

Brown, Ian (2016). "The economics of privacy, data protection and surveillance". In: *Handbook on the Economics of the Internet*. Edward Elgar Publishing, pp. 247–261.

Brynjolfsson, Erik, Seon Tae Kim, and Joo Hee Oh (2024). "The Attention Economy: Measuring the Value of Free Goods on the Internet". In: *Info. Sys. Research* 35.3, pp. 978–991.

Campbell, James, Avi Goldfarb, and Catherine Tucker (2015). "Privacy Regulation and Market Structure". en. In: *Journal of Economics & Management Strategy* 24.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jems.12079, pp. 47–73.

Celis, L. Elisa, Gregory Lewis, Markus Mobius, and Hamid Nazerzadeh (2014). "Buy-It-Now or Take-a-Chance: Price Discrimination Through Randomized Auctions". In: *Management Science* 60.12. Publisher: INFORMS, pp. 2927–2948.

Decarolis, Francesco, Maris Goldmanis, and Antonio Penta (2020). "Marketing Agencies and Collusive Bidding in Online Ad Auctions". In: *Management Science* 66.10. Publisher: INFORMS, pp. 4433–4454.

Decarolis, Francesco and Gabriele Rovigatti (2021). "From Mad Men to Maths Men: Concentration and Buyer Power in Online Advertising". In: *American Economic Review* 111, pp. 3299–3327.

Ghose, Anindya and Vilma Todri-Adamopoulos (2016). "Toward a Digital Attribution Model: Measuring the Impact of Display Advertising on Online Consumer Behavior". In: *MIS Quarterly* 40.4. Publisher: Management Information Systems Research Center, University of Minnesota, pp. 889–910.

Goldberg, Samuel, Garrett Johnson, and Scott Shriver (2019). *Regulating Privacy Online: An Economic Evaluation of the GDPR*. en. SSRN Scholarly Paper. Rochester, NY.

Goldfarb, Avi (2014). "What is Different About Online Advertising?" In: *Review of Industrial Organization* 44.2. Publisher: Springer, pp. 115–129.

Goldfarb, Avi and Verina F. Que (2023). *The Economics of Digital Privacy*. Working Paper.

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

Haile, Philip A and Yuichi Kitamura (2019). "Unobserved heterogeneity in auctions". In: *Econometrics Journal* 22.1. Publisher: Oxford University Press, pp. C1–C19.

Hong, Han and Harry Paarsch (2006). *An introduction to the structural econometrics of auction data*. MIT Press.

Honka, Elisabeth, Ali Hortaçsu, and Maria Ana Vitorino (2017). "Advertising, Consumer Awareness, and Choice: Evidence from the U.S. Banking Industry". In: *The RAND Journal of Economics* 48.3, pp. 611–646.

Hu, Yingyao, David McAdams, and Matthew Shum (2013). "Identification of first-price auctions with non-separable unobserved heterogeneity". en. In: *Journal of Econometrics* 174.2, pp. 186–193.

Hubbard, Timothy P., René Kirkegaard, and Harry J. Paarsch (2013). "Using Economic Theory to Guide Numerical Analysis: Solving for Equilibria in Models of Asymmetric First-Price Auctions". en. In: *Computational Economics* 42.2, pp. 241–266.

Hubbard, Timothy P. and Harry J. Paarsch (2009). "Investigating bid preferences at low-price, sealed-bid auctions with endogenous participation". en. In: *International Journal of Industrial Organization* 27.1, pp. 1–14.

Hubbard, Timothy P. and Harry J. Paarsch (2014). "On the Numerical Solution of Equilibria in Auction Models with Asymmetries within the Private-Values Paradigm". en. In: *Handbook of Computational Economics*. Ed. by Karl Schmedders and Kenneth L. Judd. Vol. 3. Handbook of Computational Economics Vol. 3. Elsevier, pp. 37–115.

Jeon, Doh-Shin (2020). *Market power and transparency in open display advertising – a case study*. Tech. rep. EU Observatory on the Online Platform Economy.

Jeziorski, Przemyslaw and Ilya Segal (2015). "What Makes Them Click: Empirical Analysis of Consumer Demand for Search Advertising". In: *American Economic Journal: Microeconomics* 7.3, pp. 24–53.

Johnson, Garrett (2022). *Economic Research on Privacy Regulation: Lessons from the GDPR and Beyond*. Working Paper.

Johnson, Garrett, Scott Shriver, and Samuel Goldberg (2022). *Privacy & Market Concentration: Intended & Unintended Consequences of the GDPR*. en. SSRN Scholarly Paper. Rochester, NY.

Johnson, Garrett A., Scott K. Shriver, and Shaoyin Du (2020). "Consumer Privacy Choice in Online Advertising: Who Opts Out and at What Cost to Industry?" In: *Marketing Science* 39.1. Publisher: INFORMS, pp. 33–51.

Kesler, Reinhold, Michael Kummer, and Patrick Schulte (2019). *Competition and Privacy in Online Markets: Evidence from the Mobile App Industry*. en. SSRN Scholarly Paper. Rochester, NY.

Kong, Yunmi (2020). "Not knowing the competition: evidence and implications for auction design". In: *RAND Journal of Economics* 51.3. Publisher: Wiley Online Library, pp. 840–867.

Krasnokutskaya, Elena (2011). "Identification and Estimation of Auction Models with Unobserved Heterogeneity". In: *The Review of Economic Studies* 78.1. Publisher: [Oxford University Press, Review of Economic Studies, Ltd.], pp. 293–327.

Krasnokutskaya, Elena and Katja Seim (2011). "Bid Preference Programs and Participation in Highway Procurement Auctions". en. In: *American Economic Review* 101.6, pp. 2653–2686.

Krishna, Vijay (2009). *Auction Theory*. en. Google-Books-ID: qW1128ktG1gC. Academic Press.

Lebrun, Bernard (1999). "First Price Auctions in the Asymmetric N Bidder Case". In: *International Economic Review* 40.1. Publisher: [Economics Department of the University of Pennsylvania, Wiley, Institute of Social and Economic Research, Osaka University], pp. 125–142.

Lee, Yi-Shan and Roberto A. Weber (2024). "Revealed Privacy Preferences: Are Privacy Choices Rational?" In: *Management Science*.

Levin, Jonathan and Paul Milgrom (2010). "Online Advertising: Heterogeneity and Conflation in Market Design". en. In: *American Economic Review* 100.2, pp. 603–607.

Lewis, Randall and David Reiley (2014). "Online ads and offline sales: measuring the effect of retail advertising via a controlled experiment on Yahoo!" In: *Quantitative Marketing and Economics (QME)* 12.3. Publisher: Springer, pp. 235–266.

Marotta, Veronica, Vibhanshu Abhishek, and Alessandro Acquisti (2019). *Online Tracking and Publishers' Revenues: An Empirical Analysis*. en.

Perrigne, Isabelle and Quang Vuong (2019). "Econometrics of Auctions and Nonlinear Pricing". In: *Annual Review of Economics* 11.1. _eprint: https://doi.org/10.1146/annurev-economics-080218-025702, pp. 27–54.

Peukert, Christian, Stefan Bechtold, Michail Batikas, and Tobias Kretschmer (2022). "Regulatory Spillovers and Data Governance: Evidence from the GDPR". In: *Marketing Science* 41.4. Publisher: INFORMS, pp. 318–340.

Rafieian, Omid and Hema Yoganarasimhan (2021). "Targeting and Privacy in Mobile Advertising". In: *Marketing Science* 40.2. Publisher: INFORMS, pp. 193–218.

Ravichandran, Deepak and Nitish Korula (2019). *Effect of disabling third-party cookies on publisher revenue*. Tech. rep. Google.

Rutz, Oliver and Randolph Bucklin (2012). "Does banner advertising affect browsing for brands? clickstream choice model says yes, for some". en. In: *Quantitative Marketing and Economics (QME)* 10.2. Publisher: Springer, pp. 231–257.

Shiller, Benjamin Reed (2020). "Approximating Purchase Propensities and Reservation Prices from Broad Consumer Tracking". In: *International Economic Review* 61.2, pp. 847–870.

Tuchman, Anna E., Harikesh S. Nair, and Pedro M. Gardete (2018). "Television Ad-Skipping, Consumption Complementarities and the Consumer Demand for Advertising". In: *Quantitative Marketing and Economics* 16.2, pp. 111–174.

Wernerfelt, Nils, Anna Tuchman, Bradley T. Shapiro, and Robert Moakler (2024). "Estimating the Value of Offsite Tracking Data to Advertisers: Evidence from Meta". In: *Marketing Science*.

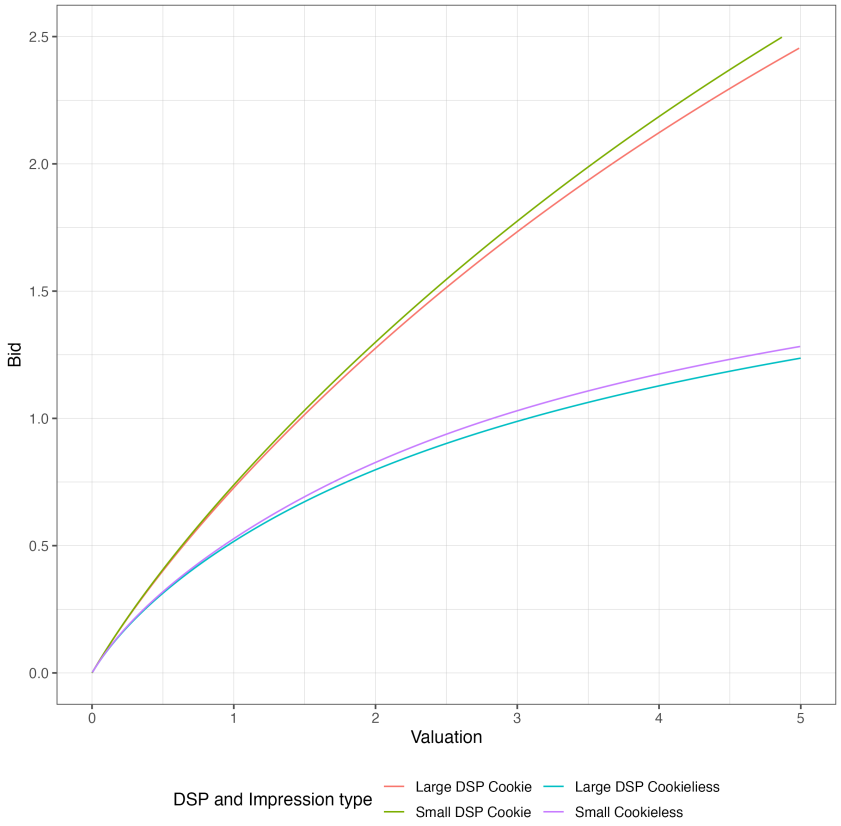
2.A Additional Tables and Figures

Table 2.A9: Regression results of logit model of entry decision

Dependent Variable:	Entry (1)
Cookie	1.191***
	(0.053)
Opt-out	0.084*
	(0.046)
Computer	-0.207***
	(0.028)
Gender female	0.164***
	(0.010)
Gender male	0.098***
	(0.015)
Age 24 and below	-0.086***
	(0.021)
Age 25 to 44	-0.099***
	(0.020)
Age 45 to 64	-0.113***
	(0.021)
Age 65 and above	-0.140***
	(0.023)
Interest segments	0.057***
	(0.008)
Months monetized	0.002***
	(0.000)
Total revenue (normalized)	-0.030***
	(0.002)
Days in database	-0.037***
	(0.004)
Fixed effects	
Time (hour)	Yes
City	Yes
Website	Yes
Browser	Yes
Observations	2,652,282

Notes: Estimation results of auction participation using logit model with 10% of the data. The base levels for age and gender are both Unknown. Standard errors are clustered by the hour of the day, the city, and the website and are heteroskedasticity-robust. ***, **, and * indicate statistical significance at the 1, 5, and 10% levels, respectively.

Figure 2.A4: Bidding functions of large and small general-purpose DSPs



Notes: Bid functions of large and small general DSPs for cookie and cookieless impressions using estimated parameters at average auction characteristics.

Chapter 3

VENTURE CAPITAL: A TALE OF THREE NETWORKS

3.1 Introduction

The venture capital (VC) industry plays a central role in financing innovation and entrepreneurship. Yet access to deals, capital, and follow-on support is often governed not just by investment strategy or firm performance, but by personal relationships built through shared education, early career ties, or social circles (Da Rin, Hellmann, and Puri, 2013). While it is well-documented that VC firms frequently collaborate extensively through coinvestment (syndication), the informal ties that underlie these partnerships are far less visible and not adequately understood. These social connections may facilitate trust, information sharing, and access to deals, but they may also entrench a small group of insiders ("the old boy network") and restrict entry into an already concentrated industry.[1] Understanding how these relationships shape investment outcomes is critical to evaluating both the efficiency and equity of the VC ecosystem (Hochberg, Ljungqvist, and Lu, 2010; Ewens, 2023).

This paper studies the causal impact of VC networks on fund performance, with a particular focus on the informal, personal connections that traditional data sources overlook. While prior research has found that better-connected VCs perform better (Hochberg, Ljungqvist, and Lu, 2007; Tian, 2011), it remains difficult to determine whether networks drive performance or simply reflect it (Da Rin, Hellmann, and Puri, 2013). Coinvestment-based measures capture only formal, observed relationships and cannot account for latent social ties. Moreover, existing empirical approaches often rely on coarse measures of network centrality without a clear economic interpretation. To address these challenges, I develop a structural network framework that connects three networks and estimate how VC performance improves through these connections: (1) formal *coinvestment* ties formed through joint startup funding; (2) *historical* (alumni and professional) connections based on shared education and prior employment; and (3) informal *social* networks that emerge from these past affiliations and shape ongoing collaboration and information flow.[2] By leveraging quasi-exogenous variation in VC partners' past affiliations and

[1] See, for example, `https://www.forbes.com/sites/oliversmith/2019/02/03/new-industry-report-exposes-british-vc-industry-as-an-old-boys-club`.

[2] The three networks refer to different layers of relationships in the VC industry. (1) The

endogenizing network formation, I identify the causal effects of social connections on investment success and recover the structure of informal networks that shape outcomes in venture capital.

At the heart of this paper is a simple intuition: in a networked environment, a VC's productivity depends not only on its own effort and capabilities, but also on the productivity of its connected peers. This interdependence arises because venture capital is an information-intensive business where relationships facilitate the flow of soft information, expertise, and reputational signals. Networks allow VCs to reduce uncertainty and improve decision-making at two critical stages of the investment process: screening and value creation. During screening, VCs benefit from shared signals, referrals, and joint due diligence with trusted peers, which improves selection quality and mitigates adverse selection risk. In the post-investment phase, networks expand the resources available to portfolio companies such as strategic advice, hiring support, and operational contacts and increase the likelihood of securing follow-on funding. Moreover, in the two-sided matching process between startups and investors, networks serve as a signal of reputation and credibility. Well-connected VCs are more attractive to high-quality entrepreneurs, not only because of their resources but also because their connections reflect market validation (Sørensen, 2007; Nahata, 2008). The structural model formalizes this mechanism by allowing a VC's performance to depend on the expected performance of its network neighbors, capturing how information and influence propagate through the network to shape investment outcomes.

The structural network model presented in Section 3.3 builds on the framework of Battaglini, Patacchini, and Rainone (2021) and is implemented in two stages. The first stage, described in Section 3.3, introduces a baseline model in which VCs are endowed with a fixed set of connections, and performance arises through information diffusion across the network. At its core is a simple production function: a VC's performance depends on both its own effort and the performance of its connected peers. This specification captures the idea that connected VCs share information

coinvestment network consists of formal ties established when VC firms jointly invest in startups, a common practice that connects nearly all major VCs in the U.S. market (Lerner, 1994; Brander, Amit, and Antweiler, 2002; Lerner, Shane, and Tsai, 2003; Hochberg, Ljungqvist, and Lu, 2007). (2) The historical network captures long-standing connections between VC partners formed through shared educational and professional backgrounds, such as attending the same universities or working at the same firms before entering venture capital (Rider, 2012; Shue, 2013; Huang, 2022). See also `https://news.crunchbase.com/data/venture-capitalists-go-college/`. (3) The social network reflects informal and personal relationships among VC partners built upon these historical ties that facilitate mutual support and information exchange, even without formal coinvestment ties.

and thereby improve each other's outcomes. A key innovation relative to previous VC network literature is the incorporation of a micro-founded mechanism linking networks to performance. Since effort is chosen in anticipation of peer outcomes, and performance is itself shaped by the network, the model yields a system of interdependent equations in which all VCs' performances are jointly determined. This formulation allows for a direct quantification of social spillovers and provides a structural interpretation of performance as an equilibrium-based measure of network centrality.

To address the endogeneity of network formation, I extend the model to allow for endogenous link choice among VCs. In this formulation, detailed in Section 3.3, VCs select their social connections in a first stage based on rational expectations about the equilibrium performance of their peers. The model is structured as a two-period game: in period one, agents choose links in anticipation of future benefits; in period two, they select effort levels given the realized network. Connection costs depend on observed compatibility, proxied by shared professional and educational history as well as characteristic similarity. Agents internalize the benefits of information diffusion and the costs of forming connections, resulting in an equilibrium network shaped by strategic behavior.

This structure allows the model to jointly identify the magnitude of peer spillovers and the elasticity of link formation. In doing so, it recovers latent social ties, informal relationships not directly observed in coinvestment data, by leveraging variation in performance, historical affiliations, and cross-sectional differences in characteristics. The intuition is straightforward: if two VCs are highly similar, share extensive past connections, and both perform well, a strong underlying social link is likely; if performance diverges despite those similarities, the strength of their social tie is likely limited. Crucially, this approach does not rely on observed coinvestment data, allowing for a conceptual and empirical distinction between formal investment ties and informal social networks.

The results in Section 3.5 follow the structure of the modeling framework. In the baseline model, performance is systematically related to the performance of a VC's connected peers. A 10 percentage point increase in a coinvestor's exit rate is associated with a 0.1 percentage point increase in the VC's own exit rate. The magnitude of this estimate is comparable to the reduced-form regression relating performance to centrality measures. This peer effect is robust across alternative specifications that use professional and alumni networks in place of coinvestment

ties, suggesting that informal connections can carry similar informational value and play a comparable role in driving performance.

The baseline model is then extended using a two-step IV approach. This method leverages professional and alumni networks, constructed from LinkedIn profiles of VC partners, as sources of exogenous variation. These historical affiliations serve as proxies for prior relationships that are unlikely to be influenced by current fund performance. In the first step, coinvestment links are explained using shared educational and professional backgrounds as well as characteristic similarity, under the assumption that VCs tend to partner with peers who are similar to themselves. The residuals from this link formation model capture unobserved factors affecting connection decisions. In the second step, these residuals are used as controls in the main performance equation to address potential selection and simultaneity. Preliminary results from this specification show that network spillovers remain positive and statistically significant, with magnitudes comparable to those in the baseline model. Moreover, the insignificance of the residual term suggests that professional and alumni networks account for much of the unobserved heterogeneity in network formation. These findings reinforce the view that long-standing social ties play a meaningful role in structuring VC networks and shaping fund outcomes.

Building on the baseline and IV models, the final specification allows for endogenous network formation, where VCs strategically select their connections in anticipation of performance gains. This approach jointly estimates both the impact of social ties on outcomes and the responsiveness of network formation to peer quality. The results show that a 1% increase in a VC's social connectedness, whether through forming new links or strengthening existing ones, is associated with a 0.2 percentage point increase in its own exit rate. At the same time, a 1% increase in a peer's performance leads to a 0.74 percentage point increase in connection intensity, indicating that VCs actively reconfigure their networks in response to the quality of their peers. Unlike the previous specifications, which rely on observed coinvestment or historical affiliations, this model recovers the latent social network directly from performance outcomes, past ties, and characteristic similarity. The recovered network shares many features with the coinvestment network but also reveals distinct patterns of informal connectivity. These differences suggest that personal and unobserved relationships play an important and independent role in shaping performance in venture capital.

The remainder of the paper proceeds as follows. Section 3.2 describes the data

and establishes reduced-form evidence. Section 3.3 presents the structural network model following Battaglini, Patacchini, and Rainone (2021) and Section 3.4 presents the details of the estimation. Section 3.5 presents the estimation results of the structural models, and Section 3.6 concludes.

Related literature

This paper contributes to three strands of literature. First, it advances research on the determinants of venture capital performance. Unlike traditional asset classes, VC investing requires intensive screening, monitoring, and value-added engagement under high uncertainty. A growing literature highlights the importance of partner characteristics, fund size, stage specialization, and experience in shaping returns (Kaplan and Schoar, 2005; Cochrane, 2005). In addition to fund-level characteristics, another distinctive feature of the VC industry is syndication. These coinvestment partnerships serve not only financial purposes but also function as channels for information exchange and strategic alignment. Hochberg, Ljungqvist, and Lu (2007) show that centrality in the coinvestment network is positively associated with fund success, suggesting that better-connected VCs benefit from enhanced deal flow and information. Similarly, Sørensen (2007) models VC-startup matching as a two-sided process, emphasizing how relationships shape selection. However, most of this literature relies on reduced-form methods and observable investment ties, which limit causal interpretation and overlook the role of latent social connections.

This paper provides the first structural estimation of VC networks, capturing how performance is endogenously shaped by the productivity of peers. The model formalizes the information-based mechanisms emphasized in the literature, both screening and value-adding, and extends earlier insights on syndication and matching (Sørensen, 2007; Hochberg, Ljungqvist, and Lu, 2007; Sorenson and Stuart, 2001; Sorenson and Stuart, 2008; Das, Jo, and Kim, 2011). Crucially, the structural framework identifies both how networks influence performance and how performance, in turn, affects network formation. This dual direction of influence is often missing from prior work, which typically focuses on either the value of network position (Sorenson and Stuart, 2001; Sorenson and Stuart, 2008) or the determinants of link formation (Lerner, 1994; Du, 2016; Bubna, Das, and Prabhala, 2020). By modeling both sides simultaneously, this paper offers a more comprehensive view of how networks shape outcomes in venture capital.

Second, this paper is related to a broader literature on social capital and informal

networks in finance. In public markets, personal connections have been shown to affect trading patterns, capital flows, and corporate decisions. Cohen, Frazzini, and Malloy (2008) document that mutual fund managers with shared educational ties exhibit similar trading behavior, while Engelberg, Gao, and Parsons (2012) show that social relationships influence capital allocation decisions among fund managers. In the corporate sphere, Shue (2013) finds that CEO networks, particularly those based on educational background, affect corporate policy choices. These studies highlight the role of informal relationships in shaping economic behavior, even in relatively transparent markets. In contrast, venture capital is a private, opaque market in which trust and repeated interaction are even more critical, yet the role of informal social networks remains underexplored. This paper extends the logic of social capital into the VC context by quantifying the causal effects of unobserved social ties not captured by coinvestment data on fund performance.

Third, this paper builds on structural approaches to modeling networks and peer effects in economics and finance (Allen and Babus, 2009). Foundational work by Acemoglu et al. (2012) and Elliott, Golub, and Jackson (2014) models how network-based spillovers contribute to aggregate outcomes and systemic risk. In terms of empirical implementations (Graham, 2020), more recent contributions by Battaglini, Sciabolazza, and Patacchini (2020), Battaglini, Patacchini, and Rainone (2021), and Lewbel, Qu, and Tang (2023) develop structural frameworks that allow for endogenous peer effects and link formation. I adapt this approach to the venture capital setting, estimating both how performance depends on peers and how relationships are formed in equilibrium. A key innovation is the use of historical biographical data (shared education and prior employment) to instrument for unobserved social ties. This enables the recovery of latent social networks and the identification of their causal effect on performance. Despite widespread belief in their importance, such informal networks remain largely unmeasured in the VC literature and are rarely incorporated into models of financial intermediation.

3.2 Data Description

VC data

The VC data cover all recorded deal flows involving U.S.-based venture capital firms between 1990 and 2009. The cutoff year of 2009 is selected to ensure that the performance of each VC fund can be meaningfully assessed, given the typical life cycle of a VC fund is approximately ten years. Each funding round involves a target company and a syndicate of VCs, although the composition of the syndicate may

change across rounds. Startups may receive multiple rounds of financing prior to an exit, which is classified as either an initial public offering (IPO), an acquisition by another firm, or a failure. Exit dates and modes are observed for completed cases. For firms still listed as active, a company is assumed to have failed if it has not received a new round of financing within the past five years, consistent with evidence that the operational life cycle of most startups does not exceed a decade.

The sample is restricted to traditional VC firms, defined as small partnerships focused exclusively on early-stage investing. Institutions such as investment banks, large corporate investors, and healthcare companies are excluded due to their scale and diversified operations, which obscure meaningful identification of inter-firm connections. Furthermore, the analysis includes only those VC firms with at least one partner who has a publicly accessible LinkedIn profile, as professional and alumni networks constructed from these data serve as key sources of exogenous variation. The final sample comprises 15,777 funding rounds involving 670 VC firms. Summary statistics are reported in Table 3.1.

VC performance

Following prior studies (Das, Jo, and Kim, 2011; Du, 2016; Lindsey, 2008; Hochberg, Ljungqvist, and Lu, 2007), VC performance is defined as the proportion of a firm's portfolio companies that have successfully exited the market through either an initial public offering (IPO) or an acquisition. Throughout the paper, the terms "exit rate" and "VC performance" are used interchangeably. While direct data on fund-level returns would provide a more precise measure of financial performance, such information is generally unavailable due to the absence of regulatory disclosure requirements for private VC firms. Despite this limitation, exit rate serves as a credible proxy, as successful exits are a key determinant of realized returns in the industry. Moreover, exit rate is bounded between zero and one, which facilitates model estimation and improves numerical stability. Summary statistics on VC exit rates are reported in Table 3.1.

Coinvestment networks

The coinvestment network is constructed using observed VC deal activity. For each round of funding, the data identify all participating VCs. The adjacency matrix $\mathbf{G}$ is defined such that for any pair of VCs i and j, the element g_{ij} records the total number of coinvestments between them over the observed period. This formulation

Table 3.1: Summary statistics of VC firms

	N	Mean	SD	Min	Max
No. rounds	6713	25.38	81.87	1	2373
No. startups	6713	13.80	37.39	1	989
Experience (years)	6713	6.35	7.66	0.00	39.23
No. coinvestments	6713	82.23	272.02	1	8075
No. coinvestors	6713	35.25	72.59	1	1327
Performance					
No. IPOs	6713	1.69	7.24	0.00	186.00
No. acquisitions	6713	5.09	15.87	0.00	375.00
No. write-offs	6713	4.29	10.91	0.00	285.00
No. private companies	6713	2.73	7.68	0.00	148.00
IPO rate	6713	0.082	0.19	0.00	1.00
Exit rate	6713	0.41	0.38	0.00	1.00
Attributes					
Pct business & financial services	6713	0.19	0.28	0.00	1.00
Pct consumer goods & services	6713	0.13	0.24	0.00	1.00
Pct healthcare	6713	0.22	0.34	0.00	1.00
Pct information technology	6713	0.38	0.36	0.00	1.00
Pct female	6713	0.022	0.11	0.00	1.00
Pct Asian	6713	0.023	0.13	0.00	1.00
Centrality					
Degree	6713	35.25	72.59	1.00	1327.00
Betweenness	6713	0.00029	0.0014	0.00	0.059
Harmonic	6713	0.34	0.068	0.00015	0.57
Eigenvector	6713	0.039	0.081	0.00	1.00

Notes: Summary statistics of VC characteristics based on VC deals data.

results in a weighted network, where g_{ij} reflects the intensity or strength of the connection. By convention, self-links are excluded, so $g_{ii} = 0$ for all i. For empirical implementation, two alternative measures of connection intensity are also considered: a binary indicator equal to one if i and j have ever coinvested, and a log-transformed version of the raw coinvestment count. Table 3.2 reports summary statistics for the coinvestment network. The unit of observation is an ordered VC pair, yielding $n(n-1)$ dyads in total for n VCs in the sample.

Given the adjacency matrix $\mathbf{G}$, four centrality measures are computed to characterize the position of each VC within the coinvestment network, following standard concepts from network and graph theory. Each VC is treated as a vertex, and each connection as an edge. (1) The degree centrality of a vertex is defined as the number

of distinct connections it has to other vertices. Since the network is undirected in this application, no distinction is made between in-degree and out-degree.[3] In the weighted version of the network, degree centrality can also incorporate the number of coinvestments as edge weights. (2) The betweenness centrality measures the number of shortest paths between all pairs of nodes that pass through a given vertex. For each pair of VCs in the network, there exists at least one shortest path that minimizes the number of intermediate steps (or the total edge weight in the case of weighted networks). Betweenness thus captures the extent to which a VC serves as a bridge within the network. (3) The harmonic centrality, closely related to closeness centrality, is the inverse of the average shortest path length from a node to all other reachable nodes in the network. This metric reflects how easily a VC can access the broader network of peers. (4) The eigenvector centrality measures a VC's influence based on the principle that connections to highly connected nodes contribute more to one's centrality. Formally, this metric is derived from the eigenvector associated with the principal eigenvalue λ in the linear system $\lambda \mathbf{x} = \mathbf{G}\mathbf{x}$. Summary statistics for these centrality measures are reported in Table 3.1.

It is also useful to introduce the concept of alpha centrality (sometimes also named after Katz, 1953; Bonacich and Lloyd, 2001), which will serve as a foundation for several structural formulations discussed later. Alpha centrality generalizes eigenvector centrality by incorporating external sources of influence. Formally, it is defined as the solution to the linear system:

$$\mathbf{x} = \delta \mathbf{G}\mathbf{x} + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $\mathbf{x}$ is the centrality vector, $\mathbf{G}$ is the adjacency matrix, $\boldsymbol{\varepsilon}$ is a vector of exogenous influence, and δ determines the relative weight of endogenous network effects versus external shocks. When $\boldsymbol{\varepsilon}$ is set to zero, this formulation reduces to eigenvector centrality. Alpha centrality can also be interpreted as a generalized form of degree centrality, where the influence of more distant nodes is discounted. The structure in equation (3.1) will reappear in later sections with different behavioral and economic interpretations.

Covariates

Several VC-level characteristics are included as covariates in the analysis. These variables are selected based on their potential influence on performance and their

[3] See Hochberg, Ljungqvist, and Lu (2007) for a discussion of directionality in the context of VC networks.

prominence in the venture capital literature.

First, performance is expected to correlate with fund size and industry specialization. Although direct observations of fund size are unavailable, two proxies are constructed: the number of startups backed by a VC and the number of funding rounds in which the VC has participated. These measures serve as reasonable indicators of investment capacity under the assumption that larger firms typically engage in more deals.

Second, VCs often concentrate their investments within one or a few sectors to leverage expertise and avoid the inefficiencies associated with over-diversification. Four major sectors are identified in the data: business and financial services, consumer goods and services, healthcare, and information technology.

Third, demographic composition is captured by two variables: the share of female partners and the share of Asian partners within each VC firm. The venture capital industry remains predominantly white and male, making it important to understand the role of gender and racial diversity in shaping outcomes. Gender and ethnicity are imputed using first and last names extracted from LinkedIn profiles. The classification algorithm is conservative in the sense that it minimizes false positives, resolving ambiguous cases in favor of male and non-Asian designations. The analysis focuses on the Asian versus non-Asian distinction for two reasons. Asian names are more reliably identified using this method, and the Asian presence in the industry is large enough to offer meaningful variation. Approximately 10 percent of partners in the sample are identified as Asian. Summary statistics for these covariates are reported in Table 3.1.

VC Partner Data

The VC data are supplemented with firm-level information on professional and alumni networks, constructed from the LinkedIn profiles of VC partners. LinkedIn is an online platform for professional networking where individuals voluntarily disclose their career history, educational background, and other credentials. The underlying dataset was assembled by a private data provider in 2017 through large-scale web scraping of publicly available LinkedIn profiles, capturing a range of attributes including employment history and education.

For the present study, the dataset is filtered to include individuals identified as partners or directors at the VC firms in the main sample. While LinkedIn data are self-reported and may contain inaccuracies, such concerns are mitigated by the

Table 3.2: Summary statistics of pairwise connection intensities

	N	Mean	SD	Min	Max
Coinvestment, **G**					
Having coinvested	45064369	0.01	0.07	0.00	1.00
No. coinvestments	236628	2.33	3.49	1.00	129.00
log(No. coinvestments)	236628	0.51	0.69	0.00	4.86
Professional connections, $\mathbf{H}_p$					
Having professional connections	45064369	0.00	0.04	0.00	1.00
No. professional connections	87758	23.28	657.18	1.00	88977.00
log(No. professional connections)	87758	0.70	1.17	0.00	11.40
Alumni connections, $\mathbf{H}_a$					
Having alumni connections	45064369	0.01	0.09	0.00	1.00
No. alumni connections	357024	9.92	113.69	1.00	20628.00
log(No. alumni connections)	357024	0.95	1.14	0.00	9.93

Notes: The unit of observation is a VC-VC pair. The number of coinvestments is calculated based on the common funding round that both VCs participated in. Professional connections and alumni connections are calculated at the individual level and aggregated at the VC level. For example, if partner A from VC 1 and partner B from VC 2 have both worked at the same company prior to joining their respective VCs, this is one professional connection.

incentives for senior professionals to maintain accurate public profiles. In addition, manual screening was conducted to remove spurious or clearly inconsistent entries.

A remaining limitation is that not all individuals maintain LinkedIn profiles, a gap more pronounced among smaller VC firms with fewer listed partners. As a result, the coverage of historical networks may be incomplete, potentially attenuating the estimated effect of alumni and professional ties. Although the LinkedIn data are at the individual level, all professional and alumni networks are aggregated to the firm level for empirical analysis.

Professional networks

Professional networks are constructed using work history data from the LinkedIn profiles of VC partners. An adjacency matrix $\mathbf{H}_p$ is defined such that each element h_{ij} represents the number of shared work experiences between partners at VCs i and j. A shared experience is defined as a case in which at least one partner from each VC has worked at the same company at some point in time. The value of h_{ij} is computed as the total number of such pairwise overlaps across all partners of the two firms.

It is acknowledged that not all shared affiliations reflect actual interpersonal rela-

tionships, as individuals may not have worked together directly. Consequently, the measure h_{ij} should be interpreted as a proxy for the potential basis of professional ties, rather than a direct measure of existing social links. Shared employment history lowers the barrier to future interaction and thus serves as a plausible foundation for informal networking. Summary statistics for pairwise professional connections are presented in Table 3.2. Three forms of the variable are reported: the raw count of shared affiliations, a log-transformed version, and a binary indicator equal to one if at least one shared connection exists.

Alumni networks

Alumni networks are constructed analogously, based on educational background. The adjacency matrix $\mathbf{H}_a$ is defined such that h_{ij} denotes the number of alumni connections between VCs i and j. A connection is counted when one partner from each VC has attended the same educational institution. The final value of h_{ij} is the sum of all such pairwise overlaps across partners from both firms. This measure captures potential affinity or ease of networking that may arise from shared educational backgrounds. As with professional networks, the alumni network is used as a proxy for the potential basis of informal ties. Summary statistics for these connections are also reported in Table 3.2.

Evidence of network on performance

To establish a benchmark, the correlation between VC performance and network position is examined, following the empirical strategy of Hochberg, Ljungqvist, and Lu (2007). The econometric model is

$$ExitRate_i = \gamma Centrality_i + X_i\beta + \varepsilon_i, \tag{3.2}$$

where the dependent variable $ExitRate_i$ denotes the proportion of a VC's portfolio companies that successfully exited through either an IPO or an acquisition. The central explanatory variable is a measure of network centrality, constructed using various definitions outlined above. While Hochberg, Ljungqvist, and Lu (2007) address endogeneity by constructing time-lagged centrality measures based on coinvestment activity in the five years preceding each fund's vintage year, we abstract from this dimension to focus on baseline correlations prior to structural estimation. The purpose of this reduced-form model is to document the strength of the correlation between performance and network position. Endogeneity concerns are addressed in subsequent sections using a structural framework.

Table 3.3: Reduced-form evidence of network effect on VC performance

	Dependent variable:								
	Exit rate								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Degree	0.002***					0.002***			
	(0.0002)					(0.0004)			
Betweenness		9.857***					15.972***		
		(2.068)					(4.092)		
Harmonic centrality			8.517***					6.517***	
			(0.850)					(0.952)	
Eigenvector centrality				0.637***					0.929***
				(0.057)					(0.121)
No. startups					0.001***	−0.001***	0.002***	0.001***	−0.001***
					(0.0001)	(0.0003)	(0.0003)	(0.0001)	(0.0003)
Percent business & finance					0.306***	0.277***	0.289***	0.335***	0.261***
					(0.093)	(0.090)	(0.092)	(0.090)	(0.089)
Percent consumer G&S					0.190*	0.164*	0.177*	0.184*	0.150
					(0.099)	(0.097)	(0.098)	(0.096)	(0.095)
Percent healthcare					0.425***	0.394***	0.392***	0.419***	0.385***
					(0.090)	(0.088)	(0.089)	(0.087)	(0.086)
Percent info tech					0.405***	0.364***	0.370***	0.379***	0.330***
					(0.092)	(0.090)	(0.091)	(0.089)	(0.089)
Percent female					0.014	0.019	0.016	0.010	0.021
					(0.039)	(0.038)	(0.038)	(0.038)	(0.037)
Percent Asian					−0.016	−0.017	−0.016	−0.026	−0.019
					(0.034)	(0.033)	(0.033)	(0.032)	(0.032)
Constant	0.339***	0.398***	0.100***	0.332***	0.047	0.035	0.064	−0.168*	0.049
	(0.013)	(0.012)	(0.033)	(0.013)	(0.083)	(0.080)	(0.082)	(0.086)	(0.079)
Observations	670	670	670	670	670	670	670	670	670
R^2	0.133	0.033	0.131	0.156	0.143	0.190	0.163	0.200	0.214
Adjusted R^2	0.131	0.031	0.129	0.155	0.134	0.181	0.153	0.190	0.204

Notes: Estimates of equation (3.2) of various specifications are presented. Columns (1)-(4) only use centrality measure as the explanatory variable. Columns (5)-(8) include additional covariates. *, **, and *** indicate statistical significance at the 10, 5, and 1% levels.

Table 3.3 reports the estimation results from equation (3.2) and serves as a baseline for comparison with later structural estimates. All coefficients on the centrality measures are positive and statistically significant, consistent with theoretical expectations. For example, an additional connection—corresponding to a one-unit increase in degree centrality—is associated with a 0.2 percentage point increase in exit rate, holding other factors constant. A one-standard-deviation increase in betweenness centrality (approximately 0.005) corresponds to a 5 percent increase in the exit rate. These magnitudes are comparable to those reported in Hochberg, Ljungqvist, and Lu (2007), reinforcing the robustness of the centrality-performance relationship.

Before turning to the structural framework, several limitations of the reduced-form approach merit discussion. First, the analysis is subject to significant endogeneity concerns arising from both omitted variables and reverse causality. Unobserved characteristics, such as partner ability or reputation, may influence both network

position and fund performance. For instance, highly capable VCs may be more effective at securing attractive deals and cultivating strategic relationships. Additionally, reverse causality is plausible: successful funds are more likely to attract coinvestors, and better-performing VCs may be selectively targeted for syndication by their peers. As a result, observed centrality may reflect performance outcomes rather than cause them.

Second, centrality measures, aside from degree, are abstract summaries of network position that are nonlinear in connections and difficult to interpret economically. These metrics do not provide insight into the marginal effect of adding a connection or improving the quality of an existing one. For example, while it is possible to estimate the effect of a one-standard-deviation increase in betweenness or eigenvector centrality, such changes do not map clearly onto intuitive or policy-relevant interpretations. Even degree centrality, which counts direct connections, captures only local network effects and ignores the broader influence of indirect ties. It also conflates the number and quality of connections. As a result, it is difficult to assess whether ties are formed with high- or low-performing peers. A more precise interpretation of network effects requires a model with explicit micro-foundations that links performance directly to the characteristics and outcomes of connected agents. This is the focus of the next section.

3.3 Structural Network Model

Following Battaglini, Patacchini, and Rainone (2021), this section presents a condensed version of the structural model and its econometric specification. Full details are provided in the appendix and the referenced papers. The central feature of the model is the equilibrium determination of VC performance, where each VC's outcome depends on the networked interactions with peers. Two network models are introduced in sequence: an exogenous model, in which VCs are endowed with predetermined connections, and an endogenous model, in which VCs strategically form links in anticipation of their performance implications. The analysis begins with a micro-founded production function that captures how social connections contribute to performance, drawing on foundational ideas from information economics.

Baseline exogenous networks

Production function

Financial intermediary markets, including venture capital, are characterized by high levels of uncertainty, risk, and information asymmetry. To mitigate these

frictions, VCs play two critical roles: screening and value-adding. The startup landscape is saturated with ventures, but only a small subset will generate outsized returns. VCs must carefully screen opportunities before committing capital and, once invested, actively support portfolio companies through strategic guidance, operational expertise, and access to networks—until exit via acquisition or IPO becomes viable. Both functions rely heavily on information: screening hinges on the ability to identify high-potential startups amid noisy signals, while value-adding depends on the VC's access to relevant resources and connections. In a networked environment, the quality and reach of information are shaped by whom a VC is connected to, both in terms of the number (extensive margin) and productivity (intensive margin) of its peers. This motivates a simple structural framework in which a VC's performance depends on the performance of its connected peers, in addition to its own effort and characteristics.

Consider a market that consists of n VCs, indexed by $\mathcal{N} = \{1, \ldots, n\}$. Each VC $i \in \mathcal{N}$ wants to maximize its *performance* P_i that follows the Cobb-Douglas production function with two inputs,[4] *social connectedness* s_i and own effort l_i:

$$P_i = \rho s_i^{\alpha} l_i^{1-\alpha} + \varepsilon_i, \tag{3.1}$$

where ε_i is an idiosyncratic shock and $\rho > 0$ is a productivity constant. Social connectedness s_i is defined as a weighted average of the performance of VC i's network peers:

$$s_i = \sum_{j \in \mathcal{N}} g_{ij} P_j, \tag{3.2}$$

where $g_{ij} \geq 0$ denotes the intensity of the unilateral social link from VC i to VC j. Let the matrix $\mathbf{G} = (g_{ij})$ denote the structure of the social network, with g_{ij} either binary (denoting the presence of a link) or continuous (capturing connection strength). For now, let us assume that the industry is exogenously endowed with the network $\mathbf{G}$. We will relax this assumption later. In the Cobb-Douglas specification, α is the elasticity of P_i with respect to s_i, that is, the responsiveness of a VC's performance with respect to its social connectedness s_i. Intuitively, g_{ij} captures the quantity of VC i's social ties (the extensive margin), while P_j captures the quality of these ties (the intensive margin). Thus, performance depends not only on

[4]The performance term P_i represents the effectiveness of VC i in generating successful investment outcomes. Empirically, this can be proxied by the fund's historical exit rate—the proportion of portfolio companies that achieve a successful IPO or acquisition.

individual effort but also on the productivity of connected peers and the structure of the underlying network.

This structure captures three key economic mechanisms linking a VC's performance P_i to the performance of its peers through network-based interactions. First, in the screening stage, P_i can be interpreted as the strength or precision of VC i's signal about a startup's quality. Connections to other informed VCs improve signal accuracy through shared information and referrals. Second, in the value-adding phase, P_i captures the social and human capital that enables a VC to support portfolio companies—primarily through targeted advice, strategic connections, and operational expertise. Because VCs rarely engage in day-to-day operations, much of their contribution stems from informational and reputational capital. Third, in the matching process between startups and investors, P_i can also reflect a VC's reputation. Well-connected VCs are more visible and credible to entrepreneurs, and association with high-performing peers signals competence and enhances the likelihood of being selected by top startups.

Exogenous network equilibrium

To close the model, let the cost of effort be linear, given by l_i, so that the VC maximizes net performance $P_i - l_i$. Under mild regularity conditions, there exists a unique equilibrium in which all VCs simultaneously choose optimal effort levels. The resulting equilibrium performance vector $\boldsymbol{P}$ satisfies the following autoregressive system:

$$\boldsymbol{P} = \delta \mathbf{G} \boldsymbol{P} + \boldsymbol{\varepsilon}, \tag{3.3}$$

where $\delta = \rho^{\frac{1}{\alpha}}(1-\alpha)^{\frac{1-\alpha}{\alpha}}$ is the social spillover (Battaglini, Sciabolazza, and Patacchini, 2020).

While the derivation is relegated to the appendix, equation (3.3) is intuitive: being connected to high-performing peers enhances one's own performance. Comparing equation (3.3) with the definition of alpha centrality in equation (3.1), we see that the equilibrium performance vector $\boldsymbol{P}$ corresponds exactly to the alpha centrality measure, with network weight δ and exogenous influence $\boldsymbol{\varepsilon}$. The key distinction is that, in this model, centrality is not an externally computed summary statistic used to explain performance—it is the equilibrium outcome of the model itself. In this sense, performance is not a consequence of centrality; it is centrality. The structural formulation thus provides a deeper behavioral interpretation: being effective is

equivalent to being central in the flow of information and influence. Compared to the reduced-form model in equation (3.2), which regresses performance on precomputed centrality measures, the structural model explains performance as an equilibrium outcome shaped by peer interactions.

Endogenous network formation

A key limitation of the exogenous network approach is that it treats link formation as driven solely by observable similarity (e.g., homophily), while ignoring strategic considerations in network formation. If a VC's performance improves due to exogenous factors, it is natural to expect other VCs to seek closer ties, anticipating spillovers from high-performing peers. This reverse causality is not captured in the baseline model (3.3) and is instead often absorbed into an endogeneity correction. Moreover, the analysis so far assumes that observed coinvestment ties fully reflect relevant connections, but informal and personal relationships—often unobserved—may also influence performance. These latent social ties introduce an additional layer of endogeneity that cannot be addressed through standard correction techniques. Together, these concerns motivate a structural model in which social networks are formed endogenously in equilibrium.

The structure of the model remains similar, but it now unfolds over two periods. In the first period, VC i chooses its network connections $\boldsymbol{g}_i = (g_{i1}, \ldots, g_{in})$ in anticipation of how these links will affect its future performance. In the second period, given the realized network, it then selects effort level l_i, and performance outcomes are determined in equilibrium. Forward-looking VCs optimize their network formation decisions in the first period, internalizing the effect of their connections on equilibrium effectiveness. The equilibrium is defined by the pure strategy profile $(\boldsymbol{g}_i, l_i)$, where $\boldsymbol{g}_i$ maps the VC i's type to a vector of connections, and l_i maps both type and network structure to the chosen effort level.

The final component of the model is the cost of forming social links. In the first period, VC i incurs a cost $c(g_{ij}, \theta_{ij}; \lambda)$ to establish a connection of intensity g_{ij} with VC j. This cost is increasing in the connection strength g_{ij} and decreasing in the pairwise compatibility θ_{ij}, which captures how naturally VCs i and j are able to form a tie (to be specified below). I assume the following isoelastic cost function:

$$c(g_{ij}, \theta_{ij}; \lambda) = \frac{\lambda}{1+\lambda} \left(\frac{g_{ij}}{\theta_{ij}} \right)^{1+\frac{1}{\lambda}}, \tag{3.4}$$

where $\lambda > 0$ captures the curvature of the cost function. As will become clear, λ

provides a convenient measure of the elasticity of link formation with respect to peer performance ε_{g_{ij},P_j}, i.e., the responsiveness of VC i's optimal connection intensity g_{ij} to the performance of its peer j, P_j.

Endogenous network equilibrium

Given the setup, Battaglini, Patacchini, and Rainone (2021) defines a *network competitive equilibrium* $(\boldsymbol{l}, \boldsymbol{P}, \mathbf{G})$ that satisfies three conditions: (1) In period 1, each VC chooses a vector of connections $\boldsymbol{g}_i = (g_{i1}, \ldots, g_{in})$ optimally given $\boldsymbol{P}$ (VCs are price-taking); (2) In period 2, each VC chooses own effort l_i optimally given $\boldsymbol{P}$ and $\boldsymbol{g}_i$; and (3) Performance P_i satisfies the production function given l_i and $\boldsymbol{g}_i$ (price must clear the market). Under mild regularity conditions, a unique pure-strategy equilibrium exists with interior solutions. The equilibrium performance $\boldsymbol{P}$ is characterized by

$$P_i = \varphi \sum_j (\theta_{ij} P_j)^{1+\lambda} + \varepsilon_i \tag{3.5}$$

for all i, where φ is a function of the structural parameters ρ, α, and λ.[5] In equilibrium, the social connectedness $\mathbf{G}$ is given by

$$g_{ij} = \theta_{ij}^{1+\lambda} (\alpha\delta P_j)^{\lambda} \tag{3.6}$$

for all $i \neq j$.

Equation (3.5) states that the resulting equilibrium performance is governed by a system of nonlinear equations. The parameter φ captures the strength of network spillovers. Comparing the endogenous system in equation (3.5) with the exogenous network equilibrium in equation (3.3), performance can be interpreted as a generalized form of alpha centrality, augmented by a nonlinear component driven by λ. This nonlinearity arises from endogenous network formation: because VC i optimally chooses its connection with j, g_{ij}, proportional to P_j^{λ} in equation (3.6), its own performance P_i becomes a function of $P_j^{1+\lambda}$. In this endogenous framework, centrality and performance are no longer separable; being central in the network reflects both connection strength and peer quality, jointly determined through forward-looking strategic behavior.

[5] The closed-form expression is $\varphi = \alpha^{\lambda}\delta^{1+\lambda}$, where $\delta = \rho^{\frac{1}{\alpha}}(1-\alpha)^{\frac{1-\alpha}{\alpha}}$ is a shorthand parameter in the model identical to that in equation (3.3). Details are provided in the appendix.

Finally, under the parametric specification in equation (3.4), the elasticity of link intensity g_{ij} with respect to peer performance P_j is exactly equal to λ.[6] This gives λ a convenient and intuitive interpretation: it captures both the sensitivity of link formation to peer quality and the strength of the feedback between performance and network structure. When $\lambda = 0$, equations (3.5) and (3.6) reduce to the baseline model (3.3), in which the network is exogenously given and fixed. Thus, λ provides a structural measure of how much active, performance-driven network formation occurs in the VC industry, and how much the endogenous model improves upon the exogenous benchmark in explaining observed performance.

3.4 Estimation

Baseline specifications

For estimation, the unobserved component of performance is assumed to depend linearly on observable VC-level characteristics. Let $\mathbf{X} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n]'$ denote the matrix of covariates. The baseline empirical model takes the form:

$$\boldsymbol{P} = \delta \mathbf{G}\boldsymbol{P} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.1}$$

which corresponds to a spatial autoregressive (SAR) model commonly used in the network literature. This system can be estimated via maximum likelihood and allows for direct inference on δ, the reduced-form parameter capturing the strength of peer spillovers.[7] If $\delta = 0$, network spillovers are absent, and the model reduces to a standard linear regression on individual characteristics.

Estimation is implemented using multiple specifications of the adjacency matrix $\mathbf{G}$. The baseline specification relies on observed coinvestment ties, which reflect formal collaboration between VC firms. While standard in the literature, this approach is vulnerable to endogeneity arising from simultaneity and omitted variables. To address this, I also consider alternative network matrices based on historical affiliations—specifically, alumni and professional networks, denoted $\mathbf{H}_a$ and $\mathbf{H}_p$, respectively—constructed from biographical information. These past connections are plausibly exogenous to current performance and help mitigate concerns related

[6] The elasticity of a link g_{ij} with respect to the effectiveness of j is

$$\varepsilon_{g_{ij},P_j} = \frac{\partial g_{ij}}{\partial P_j}\frac{P_j}{g_{ij}} = \theta_{ij}^{1+\lambda}(\alpha\delta P_j)^{\lambda-1}\alpha\delta\frac{P_j}{g_{ij}} = \lambda.$$

[7] Recall that δ is a composite of the structural parameters ρ and α from the Cobb-Douglas production function in equation (3.1), which cannot be separately identified.

to unobserved heterogeneity. However, they do not capture the influence of ongoing, contemporaneous interactions. To reconcile this limitation, a two-step estimation procedure is introduced, linking past and current networks while accounting for selection into coinvestment relationships.

Instrumental variable (IV) approach

The baseline specification in equation (3.1) is subject to endogeneity concerns due to simultaneity and omitted variables. For example, a VC partner's intrinsic ability or socioeconomic background may influence both performance and network formation, leading to biased estimates of peer effects. To address these issues, a two-step Heckman-style correction is introduced to account for selection into network links. Historical affiliations—specifically, alumni and professional ties—serve as proxies for unobserved individual heterogeneity.

In the first stage, the probability of a coinvestment tie between VC i and VC j is modeled as a function of past connections between their partners, based on shared educational and employment backgrounds. This step controls for selection driven by characteristics correlated with both performance and network structure. The residual from this regression captures unobserved factors influencing link formation and is included in the second stage as a control function. The performance equation is then re-estimated, incorporating this correction to isolate the causal effect of peer performance while mitigating endogeneity bias.

The identification strategy using alumni and professional networks relies on two conditions: relevance and exogeneity. Relevance is relatively uncontroversial. Shared educational and occupational experiences are well-documented predictors of long-term professional relationships. In the venture capital setting, such historical affiliations plausibly influence the formation of coinvestment ties, a pattern confirmed in the first-stage regression.

The exogeneity condition is more demanding. It assumes that historical connections affect current performance only through their influence on network formation, not through any direct or unobserved channels. This assumption may be difficult to satisfy in environments like venture capital, where informal and persistent social ties often operate alongside formal partnerships. While these instruments help mitigate endogeneity concerns related to unobserved ability and background characteristics, they may not fully account for latent social relationships. For this reason, alumni and professional networks are best viewed as useful but partial instruments. The struc-

tural model introduced in the next section addresses this limitation by endogenizing network formation directly.

The first step estimates a standard dyadic model of link formation, in which the presence or intensity of a coinvestment connection between VC i and VC j is explained by their historical ties and the distance between their observable characteristics. The specification is given by

$$g_{ij} = \gamma_0 + \gamma_1 h_{ij} + \sum_{\ell} \gamma^{\ell} d(X_i^{\ell}, X_j^{\ell}) + \eta_{ij}, \tag{3.2}$$

where h_{ij} denotes a past connection through shared educational or professional affiliation, and $d(X_i^{\ell}, X_j^{\ell})$ is a distance metric between VCs i and j along characteristic ℓ. The error term η_{ij} captures unobserved determinants of link formation. Intuitively, the probability or strength of a coinvestment tie increases with prior affiliation and decreases with dissimilarity in key attributes, reflecting homophily in network formation.

In the second step, two estimation strategies are available: a standard two-stage least squares (2SLS) instrumental variable (IV) approach using the predicted links $\hat{g}_{ij}$ from the first stage, or a control function method based on the residuals $\hat{\eta}_{ij}$, analogous to a Heckman correction. The second approach requires an assumption on the covariance of the residuals $\boldsymbol{\varepsilon}$ and η_{ij}, which are outlined in the appendix. In particular, it assumes that the correlation is the same between unobserved characteristics determining link formation η_{ij} and the unobserved characteristics driving the outcome ε_i for all VCs. Under this correction, the equilibrium performance equation is augmented as follows:

$$\boldsymbol{P} = \delta \mathbf{G}\boldsymbol{P} + \mathbf{X}\boldsymbol{\beta} + \psi\boldsymbol{\xi} + \boldsymbol{\varepsilon}, \tag{3.3}$$

where $\xi_i = \sum_{j \neq i} \eta_{ij}$ aggregates the residuals from the first-stage link formation equation. The additional term $\psi\boldsymbol{\xi}$ captures unobserved individual heterogeneity that influences both network formation and performance, correcting for selection bias in the estimation of peer effects.

Endogenous network formation

For econometric implementation of endogenous network formation, individual heterogeneity is assumed to be a linear function of observable characteristics. The resulting equilibrium condition yields the main estimating equation:

$$P_i = \varphi \sum_{j} (\theta_{ij} P_j)^{1+\lambda} + X_i\beta + \varepsilon_i, \tag{3.4}$$

where φ and λ are structural parameters, X_i denotes the vector of characteristics for VC i, and ε_i is the idiosyncratic error term. The term θ_{ij} captures the compatibility between VCs i and j and governs the strength of their latent social connection.

Compatibility θ_{ij} is modeled as a Bernoulli random variable, where the probability of a tie is given by a logistic function of χ_{ij}, a latent connectivity index that depends on historical ties and pairwise distances in observable characteristics:

$$\Pr(\theta_{ij} = 1|\chi_{ij}) = \frac{\exp(\chi_{ij})}{1 + \exp(\chi_{ij})}, \tag{3.5}$$
$$\text{with } \chi_{ij} = \kappa_0 + \kappa_1 h_{ij} + \sum_{\ell} \kappa^{\ell} d(X_i^{\ell}, X_j^{\ell}),$$

where h_{ij} denotes a prior connection (e.g., shared educational or employment background), and $d(X_i^{\ell}, X_j^{\ell})$ represents the distance between VCs i and j along characteristic ℓ. This formulation links observed historical data to the latent structure of social ties that influence equilibrium performance.

A final note distinguishes equation (3.5) from the link formation model used in the first step of the two-step network model in equation (3.2). In equation (3.5), the outcome variable is θ_{ij}, which represents latent compatibility between VCs and governs the probability of a social tie. In contrast, equation (3.2) models g_{ij}, the observed coinvestment connection itself, as a function of past ties and characteristic distances. While the two specifications are similar in form, their interpretation is fundamentally different. In the network formation model, social networks g_{ij} are not observed directly but are endogenously recovered from the equilibrium in equation (3.6). As will be shown in the empirical results, the recovered social networks share important features with observed coinvestment ties but also reveal distinct patterns of informal connectivity.

Estimation method

The estimation is implemented using Bayesian methods. The main estimation equations are the baseline network model (3.1), the two-step procedure (3.3), and the structural model with endogenous network formation (3.4). In each case, VC performance $\boldsymbol{P}$ appears on both sides of the equations through network interactions, introducing simultaneity that renders classical estimation approaches infeasible or inconsistent. Instead, all models are estimated via a Bayesian framework that accommodates the recursive structure of equilibrium and facilitates inference on the full posterior distribution of parameters.

Specifically, estimation proceeds via Approximate Bayesian Computation (ABC), a simulation-based method particularly suited to structural models with intractable likelihoods. The algorithm builds on the classic Metropolis-Hastings framework (see Metropolis et al., 1953; Hastings, 1970) and follows the implementation in Marjoram et al. (2003) and Battaglini, Patacchini, and Rainone (2021). Starting from an initial value of the parameters ω, the algorithm proposes a candidate parameter vector ω' from a pre-specified transition kernel. If the proposed parameter ω' fits the observed data $\boldsymbol{P}$ better according to the equilibrium condition than the current parameter ω does, then the algorithm moves to the proposed parameter ω' with some probability. The algorithm generates a Markov chain with a limiting stationary distribution, allowing consistent posterior inference on structural parameters despite the model's complexity. Under the assumption that the model is correctly specified, the posterior distribution coincides with the true conditional distribution of the parameter $P(\omega|\boldsymbol{P})$, the object of our interest.

3.5 Results

Baseline specifications

Table 3.4 reports estimation results from the baseline exogenous network model. This specification relates VC performance to observed network connections, without correcting for endogeneity or accounting for latent social ties. Column (1) provides a benchmark OLS regression of exit rates on VC characteristics alone, ignoring network effects. As expected, performance is positively associated with fund size, measured by the number of supported startups. This finding aligns with prior work showing that larger VCs tend to outperform, likely due to greater resources and broader deal access. Industry specialization also plays a role: VCs focused on healthcare and information technology exhibit higher exit rates, reflecting strong growth in these sectors over the sample period.

Columns (2) through (4) of Table 3.4 present results from the baseline network model in equation (3.1). Column (2) uses observed coinvestment ties $\mathbf{G}$ as the network matrix, while Columns (3) and (4) use professional and alumni networks, $\mathbf{H}_p$ and $\mathbf{H}_a$, respectively. Across all specifications, the estimated coefficient on the peer effect parameter δ is positive and statistically significant, consistent with the presence of network spillovers in VC performance.

The estimated magnitude of δ admits two potential interpretations. At the intensive margin, a 10 percentage point increase in the exit rate of a coinvestor is associated

Table 3.4: Estimation results of the baseline network model

	Dependent variable:			
	Exit rate			
	(1) No networks	(2) Coinvestment networks	(3) Professional networks	(4) Alumni networks
δ (Social spillover)		0.00934***	0.00127**	0.0107***
		(0.00126)	(0.000548)	(0.000605)
No. startups	0.000997***	-0.00098***	0.000799***	-0.000469*
	(0.000131)	(0.000323)	(0.000159)	(0.000269)
Percent business & finance	0.306***	0.265***	0.295***	0.0694
	(0.0927)	(0.0882)	(0.0925)	(0.166)
Percent consumer G&S	0.19*	0.151	0.179*	-0.0839
	(0.0993)	(0.0945)	(0.0991)	(0.179)
Percent healthcare	0.425***	0.388***	0.415***	0.14
	(0.0898)	(0.0845)	(0.0896)	(0.162)
Percent information tech	0.405***	0.354***	0.392***	0.0845
	(0.0918)	(0.0873)	(0.0917)	(0.165)
Percent female	0.0143	0.0198	0.0097	-0.00277
	(0.0388)	(0.0374)	(0.0387)	(0.0695)
Percent Asian	-0.0163	-0.0212	-0.0195	0.049
	(0.0335)	(0.032)	(0.0334)	(0.06)
Constant	0.047	0.037	0.0444	-0.0701
	(0.0826)	(0.0782)	(0.0823)	(0.148)
Observations	670	670	670	670

Notes: Column (1) reports the OLS of exit rate on VC characteristics. Columns (2) through (4) report the estimates of equation (3.1), with the coinvestment networks, professional networks, and alumni networks, respectively. *, **, and *** indicate statistical significance at the 10, 5, and 1% levels.

with a 0.1 percentage point increase in the VC's own exit rate, holding the network fixed. At the extensive margin, forming a new connection with a VC whose exit rate is 10% yields a similar improvement in expected performance. These effects are broadly consistent with the reduced-form estimates in Table 3.3, where an additional coinvestor is associated with a 0.2 percentage point increase in exit rate based on degree centrality. While the structural estimates are somewhat smaller in magnitude, they offer a more nuanced interpretation by jointly capturing both the quantity and quality of connections, rather than aggregating ties through centrality alone.

IV approach

First step: link formation

As a first step in the instrumental variable (IV) strategy, I assess whether historical networks are predictive of current coinvestment ties. Table 3.5 reports OLS estimates from the dyadic link formation model in equation (3.2). Columns (1) and (2) use binary indicators for the presence of a coinvestment tie, while Columns (3) and (4)

Table 3.5: First step in the IV model: coinvestment network formation

	Dependent variable:			
	If coinvest		No. coinvestments	
	(1)	(2)	(3)	(4)
Professional connections	0.131***		0.0278***	
	(0.00125)		(0.000582)	
Alumni connections		0.0461***		0.0154***
		(0.00065)		(0.000214)
No. startups (absolute distance)	-0.00367**	-0.00349*	-0.013	-0.011
	(0.00178)	(0.00179)	(0.0134)	(0.0134)
Percent business & finance (absolute distance)	-0.0124***	-0.0128***	-0.0595***	-0.0556***
	(0.00106)	(0.00106)	(0.00799)	(0.00796)
Percent consumer G&S (absolute distance)	-0.0189***	-0.019***	-0.083***	-0.0864***
	(0.000877)	(0.000883)	(0.00662)	(0.0066)
Percent healthcare (absolute distance)	-0.0251***	-0.0242***	-0.107***	-0.0994***
	(0.000782)	(0.000788)	(0.0059)	(0.00588)
Percent information tech (absolute distance)	-0.0102***	-0.011***	-0.0416***	-0.041***
	(0.0013)	(0.00131)	(0.00981)	(0.00978)
Percent female (absolute distance)	-0.0182***	-0.0169***	-0.129***	-0.113***
	(0.000565)	(0.000575)	(0.00425)	(0.00424)
Percent Asian (absolute distance)	-0.015***	-0.0156***	-0.108***	-0.0946***
	(0.000572)	(0.000578)	(0.00429)	(0.00429)
Constant	0.052***	0.0461***	0.268***	0.241***
	(0.000456)	(0.000517)	(0.00332)	(0.00334)
Observations	448900	448900	448900	448900

Notes: Results from the estimation of equation (3.2). Columns (1) and (2) use the binary outcome of coinvestment as the outcome variable. Columns (3) and (4) use the number of coinvestments as the outcome variable. *, **, and *** indicate statistical significance at the 10, 5, and 1% levels.

use the raw number of connections between VC pairs as the outcome variable.

The results provide strong evidence of homophily in network formation. Coefficients on the pairwise distances in fund size, industry specialization, and demographic characteristics are consistently negative and statistically significant, indicating that VCs are more likely to coinvest with similar peers. Demographic similarity appears particularly salient: VCs with greater representation of female or Asian partners are more likely to syndicate with others sharing these traits. This pattern may reflect preferences for in-group trust and collaboration, or alternatively, structural segmentation in an industry where informal networks shape access and opportunity.

Of particular interest for the IV strategy are the coefficients on professional and alumni networks. Both variables are positive and statistically significant across specifications, confirming that historical ties are predictive of current coinvestment behavior. In Columns (1) and (2), the presence of an alumni connection increases

Table 3.6: Estimation results of the IV model

	Dependent variable:					
	Exit rate					
	(1) No networks	(2) Baseline model	(3) IV (professional, binary)	(4) IV (alumni, binary)	(5) IV (professional, count)	(6) IV (alumni, count)
δ (Social spillover)		0.00934***	0.00934***	0.00934***	0.00742***	0.00744***
		(0.00126)	(0.00126)	(0.00126)	(0.000394)	(0.000394)
No. startups	0.000997***	-0.00098***	-0.000979***	-0.00098***	-0.00466***	-0.00465***
	(0.000131)	(0.000323)	(0.000323)	(0.000323)	(0.000263)	(0.000262)
Percent business & finance	0.306***	0.265***	0.265***	0.266***	0.328***	0.332***
	(0.0927)	(0.0882)	(0.0884)	(0.0885)	(0.115)	(0.115)
Percent consumer G&S	0.19*	0.151	0.15	0.152	0.307**	0.309**
	(0.0993)	(0.0945)	(0.0947)	(0.0948)	(0.123)	(0.123)
Percent healthcare	0.425***	0.388***	0.388***	0.389***	0.155	0.159
	(0.0898)	(0.0845)	(0.0847)	(0.0847)	(0.108)	(0.108)
Percent information tech	0.405***	0.354***	0.353***	0.355***	0.361***	0.365***
	(0.0918)	(0.0873)	(0.0877)	(0.0877)	(0.113)	(0.114)
Percent female	0.0143	0.0198	0.0198	0.0198	0.0754	0.0748
	(0.0388)	(0.0374)	(0.0374)	(0.0374)	(0.0491)	(0.0491)
Percent Asian	-0.0163	-0.0212	-0.0211	-0.0212	0.0691*	0.0689*
	(0.0335)	(0.032)	(0.032)	(0.032)	(0.0405)	(0.0406)
Constant	0.047	0.037	0.0376	0.0361	0.0934	0.09
	(0.0826)	(0.0782)	(0.0785)	(0.0785)	(0.103)	(0.103)
Observations	670	670	670	670	670	670

Notes: Column (1) reports the OLS of exit rate on VC characteristics. Column (2) reports the estimates of equation (3.1). Columns (3) to (6) report the estimates of equation (3.3), with the professional networks and alumni networks, either binary or raw count, respectively. *, **, and *** indicate statistical significance at the 10, 5, and 1% levels.

the probability of a coinvestment tie by approximately 5 percentage points, while a professional connection increases it by 13 percentage points. The stronger predictive power of professional ties holds in the continuous specifications as well: in Columns (3) and (4), each additional professional connection is associated with an increase of 0.03 coinvestments, compared to 0.015 for alumni connections. These results suggest that prior work experience plays a more substantial role than shared educational background in shaping current collaborative behavior among VCs.

Second step

Table 3.6 reports the second-stage estimation results using the IV strategy to address endogeneity in the baseline network model. As benchmarks, Column (1) presents OLS estimates of exit rates on VC characteristics alone, and Column (2) replicates the baseline network model using coinvestment ties from Table 3.4. Columns (3) through (6) report estimates from equation (3.3), incorporating a control function term derived from the first-stage link formation model.

Across all specifications, the estimated network spillover effect δ remains positive and statistically significant, consistent with the presence of peer effects in VC performance. The magnitudes are similar to those in the baseline model, suggesting that the initial results are not driven by omitted variable bias. The coefficients on

the control function term $\boldsymbol{\xi}$, capturing unobserved individual-level factors affecting both performance and network formation, are statistically insignificant. This finding implies that the professional and alumni networks used as instruments do not contain additional explanatory power beyond what is already captured by observed coinvestment ties.

Endogenous network formation

This section presents the estimation results from the endogenous network formation model. Table 3.7 and Table 3.8 report the median values of the posterior distributions. Table 3.7 summarizes the posterior medians for the key structural parameters of the network competitive equilibrium in equation (3.4), including the peer spillover parameter φ, the link formation elasticity λ, and the coefficients β on VC characteristics. Table 3.8 reports posterior medians of the parameters γ from the first-stage network formation equation (3.5).[8]

Before turning to the results, it is important to emphasize that both VC performance and network structure are jointly determined in equilibrium. The following interpretations therefore rely on the assumption that small changes in peer performance or network links have limited general equilibrium effects, i.e., they do not meaningfully alter the broader network architecture.

The analysis begins with the parameter φ in equation (3.4), which captures the strength of peer spillovers in the endogenous network setting. The estimate is positive and statistically significant, consistent with the presence of social externalities in performance. While the magnitude of φ is not directly interpretable due to the nonlinear structure of the model—specifically, the dependence on $P_j^{1+\lambda}$—the implied effects are economically meaningful. For example, an increase in the exit rate of a connected peer from 10% to 20% raises a VC's own expected exit rate by approximately 0.09 percentage points. Similarly, forming a new connection with a peer whose exit rate is 10% increases the VC's own performance by roughly 0.08 percentage points.

The parameter λ admits a more direct interpretation given the structure of the model. As shown in equation (3.6), λ represents the elasticity of connection intensity g_{ij}

[8]Instead of standard errors, the tables report empirical p-values for the null hypothesis that the parameter equals zero. These are computed as the proportion of posterior draws on the opposite side of zero from the posterior median. A p-value close to 0 or 1 indicates strong posterior support for a parameter being strictly negative or strictly positive, respectively. For instance, a p-value of 1 implies that the entire posterior support lies above zero, suggesting statistical significance at conventional levels.

Table 3.7: Results from the endogenous network model

	Dependent variable:
	Exit rate
φ (Social spillover)†	0.0002***
	[1.0000]
λ (Elasticity of network formation)†	0.7411***
	[1.0000]
No. startups	0.0010***
	[1.0000]
Percent business & finance	0.3395***
	[1.0000]
Percent consumer G&S	0.2165***
	[1.0000]
Percent healthcare	0.4897***
	[1.0000]
Percent information tech	0.4705***
	[1.0000]
Percent female	0.0148***
	[1.0000]
Percent Asian	-0.0232***
	[0.0000]
Pseudo-R^2	0.8352
Penalized pseudo-R^2	0.8341
MSE	0.1648
MASD	0.4320
Observations	670

Notes: λ is the elasticity of link g_{ij} with respect to the performance of j, E_j. φ is calculated based on the estimates of ρ, α, and λ. Estimates of parameters in equation (3.4) are reported in column (1). The median of the posterior distribution estimated with the ABC algorithm is reported for each parameter. The empirical p-value of zero on the estimated posterior is reported in the brackets. p-value is equal to 1 if the entire posterior distribution lies above zero and 0 if it lies below zero. *, **, and *** indicate statistical significance at the 10, 5, and 1% levels based on empirical p-values.

with respect to peer performance P_j. That is, a 1% increase in a peer's exit rate leads to a 0.74 percentage point increase in the intensity of the connection to that peer, holding all else constant and assuming negligible general equilibrium feedback. This result suggests that VCs are highly responsive to peer quality and strategically adjust their networks to strengthen ties with more effective partners.

Table 3.8 reports the estimated determinants of social connections in the VC industry based on the posterior medians from the first-stage network formation model.

Table 3.8: Results of link formation in the endogenous network model

	Dependent variable:
	Compatibility
Professional connection	1.3400***
	[1.0000]
No. startups (absolute distance)	0.0039***
	[1.0000]
Percent business & finance (absolute distance)	-4.1258***
	[0.0000]
Percent consumer G&S (absolute distance)	-3.1104***
	[0.0000]
Percent healthcare (absolute distance)	-0.8625***
	[0.0000]
Percent information tech (absolute distance)	-1.9955***
	[0.0000]
Percent female (absolute distance)	-0.1731***
	[0.0000]
Percent Asian (absolute distance)	-0.3480***
	[0.0000]
Constant	-1.8462***
	[0.0000]
Observations	448,900

Notes: Estimates of parameters in equation (3.5) are reported in column (1). The median of the posterior distribution estimated with the ABC algorithm is reported for each parameter. The empirical p-value of zero on the estimated posterior is reported in the brackets. p-value is equal to 1 if the support of the empirical posterior distribution is greater than zero, whereas p-value is equal to 0 if the support of the empirical posterior distribution is less than zero. *, **, and *** indicate statistical significance at the 10, 5, and 1% levels based on empirical p-values.

Consistent with prior results, there is strong evidence of homophily: VCs exhibit a pronounced tendency to form ties with peers who share similar demographic characteristics and industry specializations.

One notable exception is the positive coefficient on the distance in the number of supported startups, suggesting that VCs may be more likely to form connections with partners of different fund sizes. This result points to a potential complementarity in network formation. VCs may seek to diversify their information sets or mitigate risk by engaging with firms that differ in scale. Smaller VCs might benefit from the reach

and experience of larger funds, while larger VCs may gain access to niche expertise or localized knowledge from smaller peers. These heterogeneous connections could enhance the value of social ties beyond what would be expected from homophily in fund size alone.

A key finding from Table 3.8 is the positive and statistically significant coefficient on past connections. While the magnitude is not directly interpretable due to the logistic specification in equation (3.5), the direction and significance of the estimate are noteworthy. Importantly, these parameters are inferred from VC performance data within the structural model, rather than being derived from observed coinvestment networks. This alignment with intuitive patterns underscores the model's ability to recover meaningful latent social structures. Notably, the social networks inferred from the model differ from the observed coinvestment networks, suggesting that the model captures latent relational dynamics not immediately evident in direct investment ties. This distinction opens avenues for further research into the nature and implications of these latent social networks in the VC industry.

Comparison between models

To benchmark the endogenous network formation model, I compare it with two benchmark specifications in Table 3.9. The first is a benchmark without network effects, in which VC performance depends solely on observable characteristics. This is equivalent to imposing $\rho = 0$ in the production function (3.1), and consequently $\varphi = 0$ in equation (3.4). The second benchmark allows performance to depend on networks, but treats connections as exogenously given, i.e., VCs do not choose their links strategically. This corresponds to setting the network formation elasticity $\lambda = 0$, such that g_{ij} becomes equal to θ_{ij} in equation (3.6), and equation (3.4) reduces to the baseline exogenous network model in equation (3.1).

Several key findings emerge. First, the estimated social spillover parameter φ is positive and statistically significant in both the exogenous and endogenous models, confirming that peer performance influences a VC's own success. However, the magnitude is notably smaller in the endogenous case, likely reflecting the model's adjustment for strategic link formation. Second, the elasticity of network formation λ is estimated at 0.7411 and is highly significant, indicating that VCs actively respond to peer quality when forming social connections. Across all models, fund size and industry specialization are strong predictors of exit performance, particularly in healthcare and information technology. Interestingly, the coefficient on percent

Table 3.9: Comparison between the main estimation and two benchmarks

	Dependent variable:		
	Exit rate		
	(1) No networks	(2) Exogenous networks	(3) Endogenous networks
φ (Social spillover)†	-	0.0012*** [1.0000]	0.0002*** [1.0000]
λ (Elasticity of network formation)†	-	-	0.7411*** [1.0000]
No. startups	0.0010*** [1.0000]	0.0011*** [1.0000]	0.0010*** [1.0000]
Percent business & finance	0.3539*** [1.0000]	0.3300*** [1.0000]	0.3395*** [1.0000]
Percent consumer G&S	0.2403*** [1.0000]	0.2298*** [1.0000]	0.2165*** [1.0000]
Percent healthcare	0.4730*** [1.0000]	0.4323*** [1.0000]	0.4897*** [1.0000]
Percent information tech	0.4546*** [1.0000]	0.4581*** [1.0000]	0.4705*** [1.0000]
Percent female	0.0154 [0.6571]	0.0108*** [1.0000]	0.0148*** [1.0000]
Percent Asian	-0.0158 [0.3185]	-0.0172*** [0.0000]	-0.0232*** [0.0000]
Observations			

Notes: λ is the elasticity of link g_{ij} with respect to the performance of j, E_j. φ is calculated based on the estimates of ρ, α, and λ. Estimates of parameters in equation (3.5) are reported in column (3). Column (1) reports the estimates with the constraint $\lambda = 0$. Column (2) reports the estimates with the constraint $\rho = 0$. The median of the posterior distribution estimated with the ABC algorithm is reported for each parameter. The empirical p-value of zero on the estimated posterior is reported in the brackets. p-value is equal to 1 if the support of the empirical posterior distribution is greater than zero, whereas p-value is equal to 0 if the support of the empirical posterior distribution is less than zero. *, **, and *** indicate statistical significance at the 10, 5, and 1% levels based on empirical p-values.

Asian becomes increasingly negative and significant as network structure is more fully modeled, suggesting potential segmentation in network-driven access to high-quality deals.

3.6 Conclusion

The venture capital industry operates in an environment defined by high uncertainty, asymmetric information, and limited transparency. In such settings, networks, both formal and informal, play a critical role in reducing informational frictions, shaping investment decisions, and influencing performance outcomes. While it is widely acknowledged that networks matter in VC, most empirical work has focused on observed coinvestment ties using reduced-form methods, often leaving unresolved the underlying endogeneity and the role of latent social relationships.

This paper introduces a structural approach to studying VC networks, offering a unified framework that links network formation, information flow, and fund performance. To the best of my knowledge, this is the first study to estimate VC networks using a structural equilibrium model rooted in microeconomic foundations. The results provide robust evidence that VCs with stronger connections to high-performing peers achieve better outcomes, measured by the proportion of successful portfolio exits. Importantly, much of this effect appears driven by unobserved social connections not captured by formal coinvestment data.

The paper makes three main methodological contributions. First, I develop a microfounded production model that connects peer performance to own performance through information diffusion, grounding the analysis in theories of financial intermediation and organizational learning. Second, I use historical professional and alumni networks to construct instruments that help address endogeneity in link formation. This approach highlights the persistent influence of background affiliations and suggests that the VC industry is shaped by relationship-driven dynamics that may restrict entry for less-connected participants. Third, I propose an endogenous network formation model that recovers latent social networks directly from performance outcomes, past affiliations, and firm-level characteristics. This final contribution offers a novel way to study informal networks and shows that the recovered social structure, while overlapping with observed coinvestments, contains meaningful and distinct differences.

Overall, this study demonstrates that VC success depends not only on capital and skill, but also on access to and integration within the right networks. Informal ties and shared histories appear to be as influential as formal partnerships in determining access to deals and resources. While the model captures key aspects of social connectivity, it abstracts from dynamic changes in network composition and potential two-way causality in reputational development. Future work could build on this

framework by examining the dynamics of network evolution, the role of geographic and sectoral clustering, or the interaction between social capital and innovation outcomes. Understanding these forces sheds light on the deeper social architecture driving performance in entrepreneurial finance.

References

Acemoglu, Daron, Vasco M. Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi (2012). "The Network Origins of Aggregate Fluctuations". In: *Econometrica* 80.5, pp. 1977–2016.

Allen, Franklin and Ana Babus (2009). "Networks in Finance". In: *The Network Challenge: Strategy, Profit, and Risk in an Interlinked World*. Vol. 367.

Battaglini, Marco, Eleonora Patacchini, and Edoardo Rainone (2021). "Endogenous Social Interactions with Unobserved Networks". In: *The Review of Economic Studies*, rdab058.

Battaglini, Marco, Valerio Leone Sciabolazza, and Eleonora Patacchini (2020). "Effectiveness of Connected Legislators". In: *American Journal of Political Science* 64.4, pp. 739–756.

Bonacich, Phillip and Paulette Lloyd (2001). "Eigenvector-like Measures of Centrality for Asymmetric Relations". In: *Social Networks* 23.3, pp. 191–201.

Brander, James A., Raphael Amit, and Werner Antweiler (2002). "Venture-Capital Syndication: Improved Venture Selection vs. The Value-Added Hypothesis". In: *Journal of Economics & Management Strategy* 11.3, pp. 423–452.

Bubna, Amit, Sanjiv R. Das, and Nagpurnanand Prabhala (2020). "Venture Capital Communities". In: *Journal of Financial and Quantitative Analysis* 55.2, pp. 621–651.

Cochrane, John H. (2005). "The Risk and Return of Venture Capital". In: *Journal of Financial Economics* 75.1, pp. 3–52.

Cohen, Lauren, Andrea Frazzini, and Christopher Malloy (2008). "The Small World of Investing: Board Connections and Mutual Fund Returns". In: *Journal of Political Economy* 116.5, pp. 951–979.

Da Rin, Marco, Thomas Hellmann, and Manju Puri (2013). "A Survey of Venture Capital Research". In: *Handbook of the Economics of Finance*. Ed. by George M. Constantinides, Milton Harris, and Rene M. Stulz. Vol. 2. Elsevier, pp. 573–648.

Das, Sanjiv R., Hoje Jo, and Yongtae Kim (2011). "Polishing Diamonds in the Rough: The Sources of Syndicated Venture Performance". In: *Journal of Financial Intermediation* 20.2, pp. 199–230.

Du, Qianqian (2016). "Birds of a Feather or Celebrating Differences? The Formation and Impacts of Venture Capital Syndication". In: *Journal of Empirical Finance* 39, pp. 1–14.

Elliott, Matthew, Benjamin Golub, and Matthew O. Jackson (2014). "Financial Networks and Contagion". In: *American Economic Review* 104.10, pp. 3115–3153.

Engelberg, Joseph, Pengjie Gao, and Christopher A. Parsons (2012). "Friends with Money". In: *Journal of Financial Economics* 103.1, pp. 169–188.

Ewens, Michael (2023). “Gender and Race in Entrepreneurial Finance”. In: *Handbook of the Economics of Corporate Finance*. Ed. by B. Espen Eckbo, Gordon M. Phillips, and Morten Sorensen. Vol. 1. Private Equity and Entrepreneurial Finance. North-Holland, pp. 239–296.

Graham, Bryan S. (2020). “Network Data”. In: *Handbook of Econometrics*. Ed. by Steven N. Durlauf, Lars Peter Hansen, James J. Heckman, and Rosa L. Matzkin. Vol. 7. Handbook of Econometrics, Volume 7A. Elsevier, pp. 111–218.

Hastings, W. K. (1970). “Monte Carlo Sampling Methods Using Markov Chains and Their Applications”. In: *Biometrika* 57.1, pp. 97–109.

Hochberg, Yael V., Alexander Ljungqvist, and Yang Lu (2007). “Whom You Know Matters: Venture Capital Networks and Investment Performance”. In: *The Journal of Finance* 62.1, pp. 251–301.

Hochberg, Yael V., Alexander Ljungqvist, and Yang Lu (2010). “Networking as a Barrier to Entry and the Competitive Supply of Venture Capital”. In: *The Journal of Finance* 65.3, pp. 829–859.

Huang, Can (2022). *Networks in Venture Capital Markets*. SSRN Scholarly Paper. Rochester, NY.

Kaplan, Steven N. and Antoinette Schoar (2005). “Private Equity Performance: Returns, Persistence, and Capital Flows”. In: *The Journal of Finance* 60.4, pp. 1791–1823.

Katz, Leo (1953). “A New Status Index Derived from Sociometric Analysis”. In: *Psychometrika* 18.1, pp. 39–43.

Lerner, Josh, Hilary Shane, and Alexander Tsai (2003). “Do Equity Financing Cycles Matter? Evidence from Biotechnology Alliances”. In: *Journal of Financial Economics* 67.3, pp. 411–446.

Lerner, Joshua (1994). “The Syndication of Venture Capital Investments”. In: *Financial Management* 23.3, pp. 16–27.

Lewbel, Arthur, Xi Qu, and Xun Tang (2023). “Social Networks with Unobserved Links”. In: *Journal of Political Economy* 131.4, pp. 898–946.

Lindsey, Laura (2008). “Blurring Firm Boundaries: The Role of Venture Capital in Strategic Alliances”. In: *The Journal of Finance* 63.3, pp. 1137–1168.

Marjoram, Paul, John Molitor, Vincent Plagnol, and Simon Tavaré (2003). “Markov Chain Monte Carlo without Likelihoods”. In: *Proceedings of the National Academy of Sciences* 100.26, pp. 15324–15328.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller (1953). “Equation of State Calculations by Fast Computing Machines”. In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092.

Nahata, Rajarishi (2008). "Venture Capital Reputation and Investment Performance". In: *Journal of Financial Economics* 90.2, pp. 127–151.

Rider, Christopher I. (2012). "How Employees' Prior Affiliations Constrain Organizational Network Change: A Study of U.S. Venture Capital and Private Equity". In: *Administrative Science Quarterly* 57.3, pp. 453–483.

Shue, Kelly (2013). "Executive Networks and Firm Policies: Evidence from the Random Assignment of MBA Peers". In: *The Review of Financial Studies* 26.6, pp. 1401–1442.

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

Sorenson, Olav and Toby E. Stuart (2001). "Syndication Networks and the Spatial Distribution of Venture Capital Investments". In: *American Journal of Sociology* 106.6, pp. 1546–1588.

Sorenson, Olav and Toby E. Stuart (2008). "Bringing the Context Back In: Settings and the Search for Syndicate Partners in Venture Capital Investment Networks". In: *Administrative Science Quarterly* 53.2, pp. 266–294.

Tian, Xuan (2011). "The Causes and Consequences of Venture Capital Stage Financing". In: *Journal of Financial Economics* 101.1, pp. 132–159.

3.A Details of the Structural Model

This section reproduces results in Battaglini, Sciabolazza, and Patacchini (2020) and Battaglini, Patacchini, and Rainone (2021).

Exogenous network equilibrium

The production function for VC performance is given by

$$P_i = \rho s_i^{\alpha} l_i^{1-\alpha} + \varepsilon_i, \tag{3.A.1}$$

where s_i denotes the social connectedness of VC i, defined as

$$s_i = \sum_{j \in \mathcal{N}} g_{ij} P_j. \tag{3.A.2}$$

g_{ij} is the intensity of the social link between VCs i and j, and P_j represents the performance of peer j.[9]

Each VC chooses effort l_i to maximize performance net of effort cost:

$$\max_{l_i} \rho s_i^{\alpha} l_i^{1-\alpha} + \varepsilon_i - l_i. \tag{3.A.3}$$

The first-order condition yields the optimal effort level:[10]

$$l_i^* = (\rho(1-\alpha))^{\frac{1}{\alpha}} s_i. \tag{3.A.4}$$

Substituting this optimal choice back into the production function gives the equilibrium performance:

$$P_i^* = \delta \sum_{j \in \mathcal{N}} g_{ij} P_j^* + \varepsilon_i, \tag{3.A.5}$$

where $\delta = \rho^{\frac{1}{\alpha}}(1-\alpha)^{\frac{1-\alpha}{\alpha}}$ is a reduced-form parameter that summarizes the strength of social spillovers. Because the system in (3.A.5) is linear in $\boldsymbol{P}$, it admits a unique closed-form solution. Letting $\mathbf{G}$ denote the matrix of link intensities and $\boldsymbol{\varepsilon}$ the vector of individual shocks, the equilibrium is given by

$$\boldsymbol{P}(\mathbf{G}, \boldsymbol{\varepsilon}; \delta) = [\mathbf{I} - \delta \mathbf{G}]^{-1} \boldsymbol{\varepsilon}. \tag{3.A.6}$$

[9] The model imposes the following parameter restrictions. Effort is bounded such that $l_i \in [0, \bar{l}]$ for some $\bar{l} > 0$, and the cost of effort is normalized to l_i. Link intensity is similarly bounded, with $g_{ij} \in [0, \bar{g}]$ for some $\bar{g} > 0$, and self-connections are ruled out by assumption, i.e., $g_{ii} = 0$ for all i. Individual heterogeneity enters additively through ε_i, which is assumed to lie in the interval $[\underline{\varepsilon}, \bar{\varepsilon}]$ with $\underline{\varepsilon} > 0$ and $\bar{\varepsilon} \in (0, 1)$. Finally, assume that $\rho \bar{g}^{\alpha} \bar{l}^{1-\alpha} + \bar{\varepsilon} < 1$. This provides a sufficient condition that guarantees $P_i \in (0, 1)$.

[10] Assume that $\bar{l} > ((1-\alpha)\rho)^{\frac{1}{\alpha}}$. This guarantees interior solutions of $l_i < \bar{l}$.

Exogenous network equilibrium

The cost of establishing a social connection between VCs is modeled by the following functional form:[11]

$$c(g_{ij}, \theta_{ij}; \lambda) = \frac{\lambda}{1+\lambda} \left(\frac{g_{ij}}{\theta_{ij}} \right)^{1+\frac{1}{\lambda}}, \tag{3.A.7}$$

where g_{ij} is the intensity of the social connection from i to j, θ_{ij} captures compatibility or ease of forming the link, and $\lambda > 0$ governs the curvature of the cost function. The parameter λ thus determines the elasticity of connection formation with respect to peer performance and plays a key role in shaping equilibrium link choices.

The model is set in two periods. In period 1, VCs choose their network connections; in period 2, they select effort levels conditional on the realized network. Each VC is characterized by a type $\omega_i = (\varepsilon_i, (\theta_{ij})_j, \mathcal{M}_i)$, where ε_i represents idiosyncratic heterogeneity, θ_{ij} denotes compatibility with each potential peer VC j, and $\mathcal{M}_i$ is the set of VCs such that $\theta_{ij} > 0$. Let Ω denote the space of types.

A strategy profile consists of a pair of functions (g, l). The connection strategy $g : \Omega \to [0, \bar{g}]^{n-1}$ maps each VC's type to a vector of connection intensities, specifying how strongly they link to each other peer. The effort strategy $l : \Omega \times G \to [0, \bar{l}]$ maps each VC's type and the realized network G to an effort level in period 2. A pure-strategy equilibrium is defined as a fixed point (g, l) in which all VCs optimize given their expectations over peer performance, network structure, and the cost of forming and maintaining social connections.

We solve the game by backward induction. In period 2, VC i chooses its own effort l_i to maximize its performance net of effort cost. This problem is identical to the baseline model analyzed earlier, with equilibrium performance $\boldsymbol{P}^*$ determined by the autoregressive system in equation (3.A.5). Ignoring discounting and substituting the period-2 optimal effort into the production function, the continuation value for VC i is given by

$$P_i^*(\mathbf{G}, \boldsymbol{\varepsilon}) - l_i^*(\mathbf{G}, \boldsymbol{\varepsilon}) = \alpha\delta \sum_{j \in \mathcal{N}} g_{ij} P_j^*(\mathbf{G}, \boldsymbol{\varepsilon}) + \varepsilon_i. \tag{3.A.8}$$

[11] The cost of link formation c_{ij} is incurred solely by VC i, implying an asymmetric cost structure. That is, c_{ij} is borne by i alone, while c_{ji} is borne by j. This assumption simplifies the exposition and can be generalized to a symmetric or shared-cost formulation without affecting the core results.

In period 1, VC i chooses its connections $\boldsymbol{g}_i = (g_{i1}, \ldots g_{in})$ to maximize its expected continuation value net of connection costs. Using the parametric cost function from equation (3.A.7), the link formation problem becomes

$$\max_{\boldsymbol{g}_i} \sum_{j \in \mathcal{N}} \left(\alpha \delta g_{ij} P_j^*(\mathbf{G}, \boldsymbol{\varepsilon}) + \varepsilon_i - \frac{\lambda}{1+\lambda} \left(\frac{g_{ij}}{\theta_{ij}} \right)^{1+\frac{1}{\lambda}} \right). \tag{3.A.9}$$

The first-order condition of equation (3.A.9) yields the following characterization:

$$g_i^* \leq \theta_{ij}^{1+\lambda} (\alpha \delta P_j^*)^{\lambda}. \tag{3.A.10}$$

Together, equations (3.A.5) and (3.A.10) characterize the network competitive equilibrium $(\boldsymbol{l}^*, \boldsymbol{P}^*, \mathbf{G}^*)$. If an interior solution exists, then the two conditions collapse to the following system:

$$P_i^* = \varphi \sum_{j \in \mathcal{N}} (\theta_{ij} P_j^*)^{1+\lambda} + \varepsilon_i, \tag{3.A.11}$$

where $\varphi = \alpha^{\lambda} \delta^{1+\lambda}$. In other words, the equilibrium performance $\boldsymbol{P}^*$ is characterized by a system of nonlinear equations.[12]

Control function approach

To account for selection bias using the control function approach in the second stage, assume the unobserved components (ε, η) have the following joint distribution. $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)'$ and $\eta_i = (\eta_{i1}, \cdots, \eta_{in})'$ are jointly normally distributed with mean zero. The covariance has the following structure: $E(\varepsilon_i^2) = \sigma_\varepsilon^2$, $E(\eta_{ij}^2) = \sigma_\eta^2$, $E(\varepsilon_i \eta_{ij}) = \sigma_{\varepsilon\eta}$ for all $i \neq j$, and $E(\eta_{ij}\eta_{ik}) = 0$ for all $j \neq k$. Under these assumptions, the expected value of the second-stage error conditional on the first-stage residuals is given by $E(\varepsilon_i | \eta_{i1}, \cdots, \eta_{in}) = \psi \sum_{j \neq i} \eta_{ij}$, where $\psi = \sigma_{\varepsilon\eta}/\sigma_\eta^2$. Incorporating this selection term yields the corrected model:

$$\boldsymbol{P} = \delta \mathbf{G} \boldsymbol{P} + \mathbf{X}\boldsymbol{\beta} + \psi \boldsymbol{\xi} + \boldsymbol{\varepsilon}, \tag{3.A.12}$$

where $\xi_i = \sum_{j \neq i} \eta_{ij}$ captures unobserved factors influencing the likelihood of forming links.

---

[12] Assume that $\bar{g} > (\alpha\delta)^{\lambda} \bar{\theta}^{1+\lambda}$, where $\bar{\theta} = \max \theta_{ij}$. If $\delta \leq \frac{1}{\bar{\theta}} \left(\frac{1}{(1+\lambda)\alpha^{\lambda} \bar{m}} \right)^{\frac{1}{1+\lambda}}$, then the equilibrium is unique.