THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 1 Capitole

Présentée et soutenue par

Luise EISFELD

Le 6 juillet 2023

Dissertation sur l'organisation industrielle empirique et l'économie de la numérisation

---

Ecole doctorale : **TSE - Toulouse Sciences Economiques**

Spécialité : **Sciences Economiques - Toulouse**

Unité de recherche :
TSE-R - Toulouse School of Economics - Recherche

Thèse dirigée par
Bruno JULLIEN

Jury

Mme Margaret KYLE, Rapporteure
Mme Isis DURRMEYER, Examinatrice
M. Christian BONTEMPS, Examinateur
M. Daniel ERSHOV, Examinateur
M. Bruno JULLIEN, Directeur de thèse
M. Alexandre DE CORNIERE, Co-directeur de thèse
M. Wilfried SAND-ZANTMAN, Président

The University neither endorses not condemns opinions expressed in this thesis.

Essays in Empirical Industrial Organization and the Economics of Digitization

Luise Eisfeld

Thèse
En vue de l'obtention du
Doctorat de l'Université de Toulouse
École doctorale: Sciences Economiques-Toulouse School of Economics

6 juillet 2023

Acknowledgement

Writing this thesis would not have been possible without support from all sides. First, I would like to thank for the extensive guidance received from my advisors, Alexandre de Cornière, Daniel Ershov, and Bruno Jullien. I am immensely grateful for their time and patience, their resourcefulness, their well-placed optimism, as well as their thoughtful questioning. Their encouragement and their confidence in my abilities gave me additional impetus to tackle an ambitious research question, and I have learned a lot from them over the years. I am moreover extremely thankful to Christian Bontemps, Isis Durrmeyer, Margaret Kyle, Marc Rysman, and Wilfried Sand-Zantman for agreeing to assist as members of my doctoral jury.

Throughout my Ph.D., I experienced TSE as a place where people have an open ear and where faculty are willing to share their advice on various aspects of the Ph.D. with me: this ranged from discussing preliminary ideas to providing deeper insights on my research and giving meaningful assistance during the Job Market year. I first want to thank Uli Hege for his mentorship and his support. His enthusiasm for my research was an important source of motivation for me, and his advice has been extremely valuable in navigating towards the field of finance over the past years. Further, I want to thank Patrick Coen, Fabrice Collard, Jacques Crémer, Olivier de Groote, Pierre Dubois, Isis Durrmeyer, Ana Gazmuri, Christian Hellwig, Nour Meddahi, Mathias Reynaert, Andrew Rhodes, François Salanié, and Paul Seabright for investing valuable time into me by answering questions, reading drafts, or sharing advice.

It has been a great pleasure to collaborate alongside Pascal Lavergne, Estelle Malavolti, François Poinas, and Patrick Rey as Teaching Assistant or Workshop Coordinator, as well as to contribute in minor ways by being Research Assistant to Patrick Rey during the last months of my Ph.D. I am extremely grateful to the administrative staff at TSE – in particular, to Caroline Tejedor, Ludmila Namolovan-Stephan, Christelle Fotso-Tatchum, Christelle Fauchié, and Alexandre Noceto – for their kind collaboration and for their effort.

With the generous support of TSE and UT1, I was able to spend six months of my Ph.D. in the United States. My visit to Boston University allowed me to tap into a stellar community of researchers and has shaped my academic thinking. I am immensely grateful to Tim Simcoe at BU Questrom, and to Jim Bessen at the TPRI, for making this possible and for getting me set up in Boston. I equally thank Marc Rysman for his warm welcome and introduction at the BU Econ Department. I sincerely thank Florian Ederer for enabling a further inspiring visit to Yale SOM, and Shane Greenstein for facilitating to cross-register with HBS. All of these academics enabled me to present in workshops or to attend classes and seminars. They spent time offering advice on my research, and made me feel a very welcomed guest at their respective institutions.

I am immensely grateful to have shared my time in Toulouse with such a fantastic and extraordinarily friendly cohort, some of which have ventured off to other places, but still kept in touch over the years: Alae, Alípio, Antoine, Dakang, Esteban, Fernando, Gosia, Jun, Ivan, Li, Matthieu, Philipp, Romain, Stefan, Thuy, Vatsala, Viri, and Yu Fu have all made Ph.D. life even more enjoyable, and I will bear fond memories of going through ups and downs together. I thank further fellow Ph.D. students and friends who accompanied me on this journey both academically and personally, and with whom

I shared many interactions that lifted me up. Among those are, in particular, Gökçe, Hanlin, Jacopo, Joana, Johanna, Kevin, Léa, Luca, Marine, Max, Nicolàs, Oscar, Rossi, Sarah, and Xin — it was great to be around each of you. There are numerous further individuals with whom I crossed paths, shared lunches, or went hiking with, and who were a true part of my life at TSE. I am extremely thankful to the Women at TSE association -– I am impressed by what they have built up and achieved, and it was a wonderful pleasure to have contributed in minor ways to the sense of community in this association.

I sincerely thank Chuqing, Erich, Felix, John, and Rubaiyat: you made my onboarding process at BU extremely smooth, and my stay in Boston both intellectually stimulating and fun. I would hope that any visiting student can have such a kind and welcoming crowd of people around them. I truly found soulmates in Boston by sharing a flat with Di, Konstantina, and Lisa — thank you for making me feel at home from day one onwards.

I also extend my thanks to the academic community whom I have met at conferences, workshops, and seminars. It has been a fantastic experience to meet and interact with academics that are struggling to find workarounds around very similar issues that I faced in my research. All three chapters have benefited from these interactions as well as from thoughtful comments provided by discussants and listeners.

Last but not least, I want to thank my partner, my sisters, my family, and dear friends for their unwavering support. These people have been a constant presence in my life, even if they were not always geographically close. They have patiently listened to my complaints and encouraged me, but also helped me distract from work. I especially thank my bunch of wonderful nieces and nephews, whose unique and innocent perspectives onto the world have helped me stay grounded in the realities of life.

Chapters 1 and 3 employ data from Crunchbase (www.crunchbase.com). Chapter 2 uses data provided via the Wharton Customer Analytics Initiative and sponsored by an anonymous online travel agent. I thank the TSE Digital Center and the Artificial and Natural Intelligence Toulouse Institute (ANITI) for their financial support.

Summary

This thesis studies questions in Industrial Organization arising in the context of new technology adoption and digitization. The decreasing costs of storing, accessing, and distributing information resulting from digitization have profoundly impacted the behavior of firms and the functioning of markets in many industries. Through empirical analyses, the chapters in this thesis seek to comprehend and to evaluate certain economic phenomena associated with digitization.

In Chapter 1, I focus on a topic that is highly debated in antitrust policy circles by shedding new light on startup acquisitions in the software industry. Given the substantial economies of scale and scope in software markets, the main competitive force in these markets is believed to come from new, innovative entrants. Therefore, I examine the incentives of young, venture capital-funded startups to enter the market in light of the numerous acquisitions that have been taking place. My contributions lie in (1) collecting and assembling new data that enable to identify competing firms; (2) in producing new, policy-relevant facts on startup acquisitions in software markets; and (3) in building and estimating a stylized dynamic structural model of startup entry. My findings suggest that acquisitions can, in general, spur the incentives for new, innovative entry. On the other hand, certain types of acquisitions, particularly those targeting very mature startups and conducted at high prices, are followed by fewer entrants, underscoring the importance of antitrust review of mergers on a case-by-case basis.

The decline in the costs of distributing information has given online platforms the role of "gatekeepers" that control what information users ultimately view. Chapter 2 therefore focuses on the implications of platform design for market competition, specifically on the algorithms that rank listed products on e-commerce websites. By examining pricing behavior and rankings of hotels on an online travel agent, I find that the ranking algorithm used by this platform tends to intensify price competition between hotels by displaying hotels more visibly at times at which they are priced lower. A structural model of consumer search allows to simulate market outcomes under counterfactual rankings, indicating that consumers would face somewhat higher prices if the ranking algorithm ranked hotels at random. Overall, the chapter highlights the significant impact online platforms have on competition within industries and ultimately consumer welfare.

As digitization has led to an increase in the amount of data collected on individuals, many jurisdictions have enacted privacy regulation targeted at protecting citizens' personal data. Chapter 3 explores the effects of such privacy regulation on startup acquisitions. In particular, I focus on unintended consequences of the General Data Protection Regulation (GDPR) implemented in 2018 in the European Union. Motivating anecdotal evidence suggests that the GDPR may have increased the burden during due diligence as well as the risk of conducting an acquisition. Analyzing acquisitions of startups conducted between 2014 and 2019, I find that the number of acquisitions of VC-funded European startups has indeed declined after enactment of the GDPR compared to startups based in the US or other non-European countries.

Résumé

Cette thèse étudie des questions en organisation industrielle qui surgissent dans le contexte de l'adoption des nouvelles technologies et de la numérisation. La diminution des coûts de stockage, d'accès et de distribution de l'information résultant de la numérisation a profondément impacté le comportement des entreprises et le fonctionnement des marchés dans de nombreuses industries. À travers des analyses empiriques, les chapitres de cette thèse cherchent à comprendre et à évaluer certains phénomènes économiques associés à la numérisation.

Au chapitre 1, je me concentre sur un sujet très débattu dans les cercles de politique antitrust en éclairant les acquisitions de startups dans l'industrie des logiciels. Étant donné les économies d'échelle et de champ substantielles dans les marchés des logiciels, la principale force concurrentielle dans ces marchés est censée provenir de nouveaux entrants innovants. Par conséquent, j'examine les incitations des startups jeunes et financées par du capital-risque à entrer sur le marché à la lumière des nombreuses acquisitions qui ont lieu. Mes contributions résident dans (1) la collecte et l'assemblage de nouvelles données qui permettent d'identifier les entreprises concurrentes; (2) la production de faits nouveaux et pertinents pour les politiques sur les acquisitions de startups dans les marchés des logiciels; et (3) la construction et l'estimation d'un modèle structurel dynamique stylisé de l'entrée de startup. Mes résultats suggèrent que les acquisitions peuvent, en général, stimuler les incitations à de nouvelles entrées innovantes. D'un autre côté, certains types d'acquisitions, en particulier celles visant des startups très matures et réalisées à des prix élevés, sont suivies par moins d'entrants, soulignant l'importance de l'examen des fusions antitrust au cas par cas.

La baisse des coûts de distribution de l'information a donné aux plateformes en ligne le rôle de « portiers » qui contrôlent les informations que les utilisateurs finaux voient. Le chapitre 2 se concentre donc sur la conception des plateformes, en particulier les algorithmes qui classent les produits répertoriés sur les sites de commerce électronique. En examinant le comportement des prix et les classements des hôtels sur un agent de voyage en ligne, je constate que l'algorithme de classement utilisé par cette plateforme a tendance à intensifier la concurrence des prix entre les hôtels en affichant les hôtels plus visibles à des moments où leur prix est plus bas. Un modèle structurel de recherche des consommateurs permet de simuler les résultats du marché sous des classements contrefactuels, indiquant que les consommateurs seraient confrontés à des prix légèrement plus élevés si l'algorithme de classement classait les hôtels au hasard. Dans l'ensemble, le chapitre met en évidence l'impact significatif des plateformes en ligne sur la concurrence au sein des industries et, en fin de compte, sur le bien-être des consommateurs.

Comme la numérisation a entraîné une augmentation de la quantité de données collectées sur les individus, de nombreuses juridictions ont promulgué une réglementation visant à protéger les données personnelles des citoyens. Le chapitre 3 explore les effets d'une telle réglementation sur les acquisitions de start-ups. En particulier, je me concentre sur les conséquences involontaires du règlement général sur la protection des données (RGPD) mis en place en 2018 dans l'Union européenne. Des éléments de preuve anecdotiques motivants suggèrent que le RGPD a pu augmenter la charge lors de la due diligence ainsi que le risque de réaliser une acquisition. En analysant les acquisitions de start-ups réalisées

entre 2014 et 2019, je constate que le nombre d'acquisitions de start-ups européennes financées par des sociétés de capital-risque a effectivement diminué après l'entrée en vigueur du RGPD par rapport aux start-ups basées aux États-Unis ou dans d'autres pays non européens.

Contents

Chapter 1

Entry and Acquisitions in Software Markets

Abstract

How do acquisitions of young, innovative, venture capital-funded firms (startups) affect firms' incentives to enter a market? I create a product-level dataset of enterprise software, and use textual analysis to identify competing firms. Motivated by new stylized facts on startup acquisitions in software, I build and estimate a dynamic model of startups' entry decisions in the face of these acquisitions. In the model, acquisitions can affect returns to entry (1) by affecting market structure, and (2) by providing an entry-for-buyout incentive to potential entrants. Using the parameter estimates, I simulate how startup entry would evolve over time if merger control was tightened. The simulations reveal that, if all startup acquisitions were blocked, entry would decline on the order of 8-20% in some markets. In contrast, I find suggestive evidence that blocking mergers between established industry players and more mature startups might increase entry. These findings indicate that case-by-case merger review can best foster sustained startup entry.

1 Introduction

Companies active in information technology – most famously, dominant incumbents such as Alphabet and Microsoft, but also much smaller players such as Dropbox and HubSpot – have acquired thousands of other firms over the past two decades. The majority of target firms in these transactions were *startups*: young, innovative, venture capital-funded firms. How do these acquisitions affect startups' incentives to initially *enter* into a given market? New, innovative entrants are thought to be a main competitive force that disciplines dominant incumbents in software markets (e.g., Crémer, de Montjoye, and Schweitzer (2019), Scott Morton, Jullien, Katz, and Kimmelman (2019)).

On the one hand, acquisitions provide an entry-for-buyout incentive if the returns from being acquired are higher than the returns from competing (Bisceglia, Padilla, Perkins, & Piccolo, 2021, Cabral, 2018, Mermelstein, Nocke, Satterthwaite, & Whinston, 2020, Phillips & Zhdanov, 2013, Rasmusen, 1988). In software markets, over 90% of successful, venture-backed startups are acquired by other firms, as opposed to being listed on public stock markets.[1] Survey results indicate that acquisitions are a major goal for startup founders.[2] This suggests that startup acquisitions can reward innovation efforts and encourage entry ex ante.

On the other hand, ex post, an acquisition affects market structure and the competitive environment that new entrants are facing. In some situations, an acquired company benefits from the acquirer's market power, or its complementary assets. The acquired product is then able to capture a larger share of the market. This *decreases* returns to entry for potential entrants, converting the market into a so-called "kill zone" in which entry is deterred (e.g., Denicolò and Polo (2021), Kamepalli, Rajan, and Zingales (2021), Whinston (1990)).

I study startups' entry incentives in the face of acquisitions (1) by collecting and assembling new data that enables to identify competing firms; (2) by producing new, policy-relevant facts on startup acquisitions in software markets; and (3) by building and estimating a stylized dynamic structural model of startup entry.

Answering the research question requires to accurately define *markets*, i.e., sets of companies that produce substitutable products and that interact strategically with each other. To obtain such a notion of competing firms, I construct a new dataset by web-scraping product-level data from *Capterra*, a vertical search engine for enterprise software.[3] As *Capterra*'s purpose is to assist consumer search, I take its product descriptions and categories and employ text-as-data methods to segment products into clusters of likely substitutes. Unlike previous literature that employs firm-level industry classification systems, this new approach enables the construction of markets at the *product* level. I merge these product data with information on firms' entry and acquisition decisions stemming from *Crunchbase*.

The data produce new, policy-relevant descriptive facts on startup acquisitions in software markets. I find that acquisitions of particularly young startups are very prevalent in enterprise software. Recent policy discussions have focused on a few dominant incumbents (in particular, on the so-called

[1] Author's computation, using a sample of enterprise software startups with successful exits in 2005-2020 from the data portal *Crunchbase*. In contrast, only 50% of startups in the biotech or pharmaceutical industry exit via acquisition. See Appendix 1.C.10.

[2] The survey results are not public but will be made public in a forthcoming working paper by Stephen Michael Impink.

[3] I thank the parent company *Gartner* for granting me the official permission to web-scrape the *Capterra* website.

GAFAM[4]). However, the data reveal that other acquirers show a similar pattern of acquisitions. At the same time, acquisition patterns in software are markedly different from observed patterns in biotechnology or pharmaceuticals, which motivates the industry-level focus of this study.

I distinguish between different types of acquirers along the dimensions of industry incumbency and measures of age, and argue that these acquirer types are likely driven by different motives. I call acquirers that, like the target, are active in enterprise software, "strategic" acquirers. I argue that strategic acquirers are the most likely to possess the capabilites and the incentives to bolster up an acquired firm's product, either via synergies or via market power, in a way that will deter follow-on entry. In contrast, industry outsiders tend to acquire enterprise software startups to vertically integrate new tools, and do not have incentives to affect a market niche's development in the long run. Financial acquirers like private equity firms tend to be transitional owners, and acquire in order to generate financial returns.

Presumably, all types of acquirers may generate an entry-for-buyout incentive, whereas only acquisitions conducted by strategic acquirers can lead to entry-deterring effects. I compare entry patterns in the quarters following major acquisitions conducted by either financial or strategic acquirers, akin to an event study framework. The results indicate that major acquisitions conducted by strategic, but not financial, acquirers tend to be followed by a decrease in new entry.

Studying how acquisitions affect startup entry requires to analyze and quantify both any entry-deterring effect that is transmitted via market structure, as well as the more long-run entry-for-buyout effect. This requires a structural model of startup entry that accounts for both channels of effect. I thus set up a dynamic model that can explain startups' entry decisions in a stylized way. In each period and in each market, a new set of forward-looking potential entrants considers whether to enter the market.[5] Firms that have entered obtain flow payoffs every period. These flow profits depend in a reduced-form way on market structure; in particular, on the number of competitors, as well as on large, strategic acquisitions of competitors in the past. In future periods, firms may moreover be acquired, or be listed on the public stock market. Whenever such a transition in ownership occurs, firms earn a single lump-sum return, and stop earning flow profits. Acquisitions and listing events are modeled as stochastic shocks that arrive upon the startup with varying frequencies across markets, and are assumed exogenous conditional on proxies for market size. When deciding whether to enter a given market, potential entrants on the one hand take into account the current and expected future market structure. On the other hand, the entrants form beliefs about the likelihood with which a change in ownership – being acquired, or going public – occurs. Using a revealed preference approach and a two-step estimation method (following, e.g., Aguirregabiria and Mira (2007) and Bajari, Benkard, and Levin (2007)), I estimate the parameters quantifying the importance of each of these channels for spurring or deterring entry.

The parameter estimates reveal that in markets in which firms are acquired at a higher frequency, startups are more inclined to enter, conditional on proxies for market size. Moreover, reflecting the findings of the event study, certain types of acquisitions – those conducted by major industry incumbents and targeting more mature startups – are followed by a decline in entry. The overall effects from banning

[4]This acronym refers to the firms *Google (Alphabet)*, *Amazon*, *Facebook (Meta)*, *Apple*, and *Microsoft*.

[5]The model therefore does not endogenize the decision regarding the timing of entry; nor the decision to exit a market; nor the decision which market to enter.

all or a subset of acquisitions are determined by the magnitudes of both channels. Based on preliminary counterfactual simulations, I find that startup entry may decline if all startup acquisitions were blocked. In particular, in markets in which the profits from competing are low relative to the returns from being acquired, entry drops in the order of 8 to 20% in the counterfactual. In those markets, the entry-for-buyout incentive is strong, and firms barely enter in order to compete. In contrast, if we allow for a causal interpretation of strategic acquisitions, blocking only mergers conducted by large, strategic acquirers would boost entry by over 4% in affected markets. Overall, this suggests that, in order to foster entry, competition authorities should continue reviewing mergers on a case-by-case basis.

Both the descriptive and the model-based findings that this paper generates are of first-order importance from an antitrust perspective. The types of acquisitions that are the focus of this paper are "stealth mergers" (Wollmann, 2019) that rarely meet merger notification thresholds, as transaction targets concern small, albeit highly innovative and potentially disruptive firms. The sheer number of these types of transactions has caught the attention of antitrust practitioners and academics worldwide.[6] At the same time, software is an industry where entry is highly valuable, as strong network effects often lead markets to "tip". The competitive forces ensuring that incumbents have sustained high rates of innovation therefore come from potential entrants competing *for* the market, instead of companies *within* the market. This has led antitrust regulators to claim that digital platforms could "buy their way out of competing", as Lina Khan, the current Chairperson of the US Federal Trade Commission, phrased it (Khan, 2021).

By studying innovative entry, this paper is linked to the long-running question of how firms' innovation incentives are affected by their competitive environment, going back to Schumpeter (1942) and Arrow (1962). Moreover, entry dynamics and the motives of acquisitions in software markets are still poorly documented and understood.[7] As these markets are bringing vast welfare gains in the years to come, understanding any frictions that entering startups face is economically important.

Literature. This paper has two main contributions to the literature. New findings on startup acquisitions and entry in software markets, of both descriptive and model-based nature, make up the first contribution. The first distinction with respect to existing literature is that, to my knowledge, I am the first to provide an industrial organization model that disentangles different channels through which acquisitions can increase startup entry. A second distinction is in respect to the level of focus on a particular firm or industry. Essentially, previous empirical literature studying acquisitions in innovative industry can be segmented into two strands: one strand has focused uniquely on the GAFAM firms and has characterized acquisition patterns (Affeldt & Kesler, 2021a, 2021b, Argentesi et al., 2021, Gautier & Lamesch, 2021). Another strand has looked at the link between M&As and innovation or business dynamism from a macro perspective, using cross-industry data and studying general equilibrium effects (Fons-Rosen, Roldan-Blanco, & Schmitz, 2023, Vaziri, 2022). My perspective lies in between those ap-

[6] According to a report by the US Federal Trade Commission, the GAFAM conducted 618 acquisitions (excluding patent acquisitions or hiring events) in 2010-2019 (Federal Trade Commission, 2021). 85% of those acquisitions took place below the reporting thresholds provided by the Hart-Scott-Rodino Act.

[7] For instance, it is not documented whether "killer acquisitions" (Cunningham, Ederer, & Ma, 2021) are more, or less, common in software as in pharmaceutical markets. It is ultimately also not clear whether acquisitions concern mostly targets that are active in the same core market as the acquired company, or in complementary markets.

proaches: I study acquirers beyond the GAFAM, which allows for comparisons of the big five acquirers to other acquirer types, and for more general conclusions. At the same time, however, I focus on acquisitions of startups active in a single industry – enterprise software – which is motivated by the fact that patterns of acquisitions are markedly different between this industry and pharmaceuticals, for instance. To my knowledge, this has only been done by G. Z. Jin, Leccese, and Wagman (2022) in technology markets, or by Cunningham et al. (2021) for pharmaceutical markets. My dataset thus covers an entire industry branch and tens of thousands of companies whose strategic actions I model. At the same time, I am able to follow acquisitions at the *product* level. I am thus able to characterize the effects of acquisitions in software at a higher level of granularity compared to earlier work, coming somewhat closer to data used in the context of the pharmaceutical industry, where project-level data is abundant (e.g., Cunningham et al. (2021), Khmelnitskaya (2021), Majewska (2022)).

To provide the new facts, I create a comprehensive dataset of software products that enables to distinguish competing firms in a novel way. I apply textual analysis to delineate which firms compete with each other. Such an approach has been pioneered by Hoberg and Phillips (2016), who employ textual analysis to 10-K reports of publicly listed firms. In contrast to those authors, I use a text base that covers private firms, and I account for multi-product firms that are active in more than one market.[8]

My descriptive findings shed new light on startups' commercialization strategies, as highlighted in Gans and Stern (2003). Also studying the enterprise software industry, Cockburn and MacGarvie (2011) study the relationship between patenting and entry into a market.

The second contribution is a model that allows to disentangle and to quantify two channels through which acquisitions can affect returns to entry. Previous empirical literature focuses on only one of these channels. Bauer and Prado (2021), G. Z. Jin et al. (2022), Kamepalli et al. (2021), Koski, Kässi, and Braesemann (2020) employ reduced-form regressions to study how GAFAM acquisitions correlate with measures of venture capital (VC) investment or entry. X. Wang (2018) and Warg (2021) find that startups "cater" to potential acquirers by investing into adjacent technology areas that may be useful for potential acquirers, which may be viewed as evidence for an entry-for-buyout effect. The counterfactual scenario in which all acquisitions are blocked, however, depends on both effects. A linear regression cannot disentangle the different channels of effect that are associated with acquisitions. The model I set up explicitly intends to quantify the two channels, which allows to simulate how entry would evolve under counterfactual antitrust regimes. The most related work that models (among other firms' decisions) entry decisions within a dynamic structural model is Igami (2017) and Igami and Uetake (2020). However, those papers ask different research questions and study a mature (as opposed to growing) industry with homogeneous products and fewer players, using a different data environment, which is why I choose a different modelling approach.

[8]Using textual analysis to distinguish industries or competitors seems to become more common in the literature. Decarolis and Rovigatti (2021) distinguish competing advertisers in the context of online ad auctions using textual analysis, and I follow their approach of vectorizing words. Foroughi and Stern (2019) use text-as-data methods to identify digital products in the medical device industry. The use of textual analysis for defining markets is also seemingly at the forefront of innovative approaches to competition practice, e.g., in analyses of merger cases. See `https://www.compasslexecon.com/the-analysis/using-natural-language-processing-in-competition-cases/03-22-2022/` and `https://www.compasslexecon.com/measuring-of-competition-using-natural-language-processing/`, both accessed 18/12/2022.

Prior theoretical research has pointed out a potential entry-deterring effect of acquisitions. Possible mechanisms by which an acquisition can deter entry are bundling (e.g., Whinston (1990)), entrenchment via cumulative acquisitions (Denicolò & Polo, 2021), the expectation of imitation (Motta & Shelegia, 2021), or a lack of early product adoption in the presence of switching costs and network effects (Kamepalli et al., 2021).[9] Whereas I empirically study any potential negative effect of acquisitions, my results cannot speak to the precise source of any decline in entry, or potential welfare effects.

I model entrepreneurs' entry decisions as dynamic strategic interactions with incomplete information. Methodologically, my paper leans on the dynamic discrete game literature (Aguirregabiria & Mira, 2007, Bajari et al., 2007) and employs forward simulation techniques used in Hotz, Miller, Sanders, and Smith (1994). A main difference to most dynamic models is that in my setting, agents only take a decision once (instead of every period). Nevertheless, the model is dynamic: agents in my model incur a sunk cost upon entering, are forward-looking, and their payoffs depend on state variables that evolve according to agents' decisions.

More broadly, this paper contributes to the open question on the link between market structure (including mergers and breakups) and innovation. This paper therefore relates to theoretical models, some of which emphasize the entry-for-buyout effect (Hollenbeck, 2020, Jullien & Lefouili, 2018, Mermelstein et al., 2020, Nocke & Whinston, 2010), as well as empirical investigations (Haucap, Rasch, & Stiebale, 2019, Igami & Uetake, 2020, Poege, 2022, Watzinger, Fackler, Nagler, & Schnitzer, 2020).

Roadmap. The paper is organized as follows. I cover the data construction in Section 2, and extensive descriptive analyses on acquisitions in enterprise software in Section 3. Section 4 provides motivating reduced-form evidence for the differential effects of different types of acquisitions. Section 3 introduces the model and covers its estimation. Section 6 presents the results and covers the counterfactual simulation. Section 7 discusses the findings and assumptions. Section 8 concludes.

2 Setting, Data, and Market Definitions

2.1 Setting: Startup Entry in Enterprise Software

I study firm entry in the *enterprise software* industry. Acquisitions of startups are an extremely prevalent feature in enterprise software. In fact, the companies acquiring the highest *number* of startups of any industry worldwide are mostly active in software. Strong industry-level differences in the rates of startups exiting the industry via acquisition, as opposed to via IPO, suggest that the motives of entry and of acquisitions might be software specific (see Appendix 1.C.10), and therefore warrant an industry-level investigation.

[9] For a more thorough elaboration of possible entry-deterring effects outlined by the abovementioned theory literature, see Appendix 1.B.1. Further literature has described conditions under which incumbents have an incentive to merge with a nascent competitor in order to discontinue the target's product and remove a future competitor (Cunningham et al., 2021, Motta & Peitz, 2021). Moreover, theoretical literature has proposed further implications of the acquisitions of nascent competitors, such as effects on the direction of innovation, which however I cannot speak to (Bryan & Hovenkamp, 2020, Cabral, 2018, 2021, Callander & Matouschek, 2020, Dijk, Moraga-Gonzàlez, & Motchenkova, n.d., Fumagalli, Motta, & Tarantino, 2022, Gilbert & Katz, 2022, Guéron & Lee, 2022, Hege & Hennessy, 2010, Katz, 2021, Letina, Schmutzler, & Seibel, 2021). Lemley and McCreary (2020) propose policy changes that provide alternatives to acquisitions as exit routes, thereby likely fostering startup entry.

I define as enterprise software any software product that can be used in a business environment. This definition captures both, products that are targeted specifically to business clients (i.e., B2B software such as customer relationship management software or accounting software), as well as products for use in both professional as well as private contexts (such as filesharing software).[10] The enterprise software industry is large and growing: between 2005 and 2020, enterprise software startups received more VC funding than all startups belonging to the biotechnology and pharmaceuticals industry (see Appendix 1.C.9). Enterprise software is likely to bring along important welfare gains in the years to come. Software enables the adoption of new technology in enterprises, such as cloud computing or analytics, which can substantially reduce costs or increase efficiency.[11]

I consider entry by *startups*, which are young, risky, very innovative, VC-backed, privately held companies. In industries with network effects where markets may "tip" in favor of a large incumbent, the threat of entry by these young firms is deemed essential for guaranteeing competition *for* the market (Furman, Coyle, Fletcher, Marsden, & McAuley, 2019). More broadly, startups play an important role for innovation and industry dynamics in the economy. In the past, startups have redefined markets and out-competed large incumbents in some industries. Startups tend to bring forward more inventions (Kortum & Lerner, 2000), as well as higher quality and more novel inventions (Schnitzer & Watzinger, 2022), than established companies. VC-funded startups have contributed to economic welfare in meaningful ways, as illustrated recently with *BioNTech*'s development of Covid-19 vaccines.

Upon being founded by entrepreneurs, startups obtain staged rounds of capital injections, primarily by groups (syndicates) of VC investors. These financial intermediaries are specialized in providing funding, as well as advice, to these risky, but potentially high-growth firms in exchange for an equity stake. VC investors manage closed-ended funds, which implies that they need to divest after a period of 7-10 years. Optimally, a startup makes a successful "exit", and is either listed on a public stock exchange (and thus becomes a public company), or (more commonly) is sold to another firm.[12] Both of these events are generally considered a success, and may yield a high return to investors and founders. However, roughly half of all startups end up failing, yielding no or little return.[13]

For anecdotal evidence of the entry-for-buyout and kill zone effect in the enterprise software industry, I refer the interested reader to Appendix sections 1.A.1 and 1.A.2.

2.2 Data

Answering the research question requires data on companies' actions – in particular, on entry and acquisition decisions – in clearly defined markets. I obtain information on firms' actions from the data portal *Crunchbase*. To distinguish which firms actually compete with each other, I additionally web-scrape data product characteristics and descriptions from *Capterra*, a vertical search engine for enterprise software.

[10]This definition therefore excludes software products for uniquely private use, such as gaming or social networks.

[11]Berman and Israeli (2022) for instance find that the adoption of analytics dashboards by e-commerce websites increases firms' weekly revenues by 4-10%.

[12]See Appendix 1.C.10 for exit rates of startups active in enterprise software, and biotech or pharma. More recently, some startups have been able to stay private for longer. In those cases, early investors sell their shares to investors specialized on later-stage companies (so-called crossover investors).

[13]The reader may refer to Gompers and Lerner (2001) for further institutional details on VC funding and startup growth.

The web-scraped product-level data allows to use text-as-data methods to classify products into distinct markets, and to produce new descriptive findings on startup acquisitions in software.

The final dataset used for the reduced-form analyses and structural model is a market-quarter panel detailing the number of entering firms, number of competitors, and types and number of acquisitions, in over 400 different markets.

Firm-level Panel: *Crunchbase*

Crunchbase is a data portal that tracks financial information on over a million public and private companies, in particular VC-funded firms. It records companies' founding dates, funding rounds, acquisitions, investments into other companies, initial public offerings (IPOs), and closures. Unlike other financial databases, having received a VC investment is not a pre-condition for being present on this database. *Crunchbase* is well-established in the empirical finance literature, and is believed to capture early-stage funding rounds and acquisitions of small sizes especially well compared to other data sources (Z. Jin, 2019, Yu, 2020).

As *Crunchbase* contains both, venture capital and other types of investments (such as private equity), I use *Crunchbase*'s "Glossary of Funding Types" (Crunchbase, 2022), industry reports and prior literature as guidance to know which types of investments to classify as venture capital.[14] I then define "startups" as companies that have received at least one such VC-type investment. I further define a startup's "entry" event as the first VC-funding round for a firm in my data.[15] *Crunchbase* itself defines acquisitions as majority takeovers.

Using information of all acquisitions in a company's lifetime, I reconstruct the parent-subsidiary structure of up to two levels of all firms over time.[16] I then construct a panel of company events.

Cross-section of Enterprise Software Products: *Capterra*

The *Crunchbase* dataset also contains information on a startup's industry in the form of industry labels and descriptive text. However, these labels are relatively broad: there are fewer than 800 labels describing the entire economy, which turns out to not enable determining which companies actually compete against each other.[17] Many of the labels are specific to an *industry*, but not to a *market* (e.g. the label "enterprise software" could in principle capture markets as distant as enterprise resource management

[14] I define investments of the following types as being VC investments: *Angel*, *Pre-Seed*, *Seed*, *Series A* to *Series J*, *Series Unknown*, *Corporate Round*, *Undisclosed* and *Convertible Note*. I consider VC investments as financial investments into very early-stage, high-risk companies. The listed investment types' descriptions in *Crunchbase*'s *Glossary of Funding Types* match these characteristics (Crunchbase, 2022). Thus, investment types such as, for instance, *Post-IPO Debt*, *Grant* or *Product Crowdfunding* are not considered as typical VC investments. See Appendix 1.C.2 for details.

[15] According to this definition, a firm that has had a "founding" event but that has not received any funding has not "entered" the market yet. I do not view this as a restriction, but rather a feature: this way, one can interpret entry decisions as being taken jointly by investors and entrepreneurs. Investors can be assumed to be rational, forward-looking agents that take into account the prospects of exit and the development of the market environment when deciding to fund a company in a given market. My decision of focusing on the first round of financing is moreover driven by the fact that it is unclear what a founding event really means on *Crunchbase*: in principle, it could signify, for instance, the date at which the founders first got together; the date of product launch; the date of company registration; etc., and it is unclear whether these capture market entry very well.

[16] This allows to associate acquisitions that were undertaken by e.g. LinkedIn after its acquisition by Alphabet as a GAFAM-acquisition. In general, the partent-subsidiary structure can go above two tiers; however, this is rare on *Crunchbase* and does not occur for the sample of firms considered.

[17] If one used these labels as markets, one would end up with over 1,300 firms per "market", which is unreasonably many. Note also that *Crunchbase*'s main purpose is not the precise categorization of startups into markets or areas of activity, but rather the documentation of startups and their funding round events.

and video advertising). Second, the labels given by *Crunchbase* vary on the firm level. However, many firms are multi-product firms. Amazon for instance is famously an e-commerce platform, a logistics company, and offers cloud computing services. Distinguishing which companies compete with each other in a given market requires a *product-level* definition of competitors. Moreover, I intend to consider both public and private firms. This prevents me from using standard industry classifications that are available for public firms only, or from using 10-K reports to distinguish competitors as Hoberg and Phillips (2016) pioneered. Lastly, it is not always clear from *Crunchbase* whether a given company is actually still active in producing a given product (see Appendix 1.C.5).

In order to obtain more accurate, product-level information, I web-scrape information from a platform called *Capterra*.[18] *Capterra* is a vertical search engine for enterprise software, and is thus designed to assist customers with comparing and finding suitable enterprise software products. It is one of the market leaders among platforms offering this service. The website classifies enterprise software products into at least one of 821 narrow categories – for example, "Audio Editing Software", "Conference Software" or "Spreadsheet Software". It provides descriptive text, information on the producing company, as well as user reviews and ratings for each product (see Figure 1.1 and 1.2).[19] The range of enterprise software products covered on *Capterra* is exhaustive and very up-to-date.[20] Overall, the product categories and descriptive text that assist consumers to find suitable business software offer a natural structure that can be used to identify competing products.

From *Capterra*'s product listings pages, I obtain over 70,000 links to product pages, which I query one-by-one in June and July of 2021. From each product page, I download and save, among others, product and company names; the categories that a product is assigned to; the company's web domain; a text describing the product; and the user rating as well as the cumulative number of user reviews.[21]

All in all, I make use of the *Capterra* data for the following purposes: first, the textual data allows to cluster products into groups of substitutable products, with the help of a pre-trained data set stemming from a machine learning model that allows me to vectorizes the textual data (see Section 2.2). Second, the data indicates which enterprise software startups' products are actually active and available as of July 2021. Third, the information on the number of reviews yields an indication of whether a given product is being used at all. This allows me to differentiate companies that are actually "relevant" competitors that act strategically in a given market, and which ones might be considered a non-strategic fringe.

[18]*Capterra* is owned by *Gartner*, a large public consulting and technological research company. I thank *Gartner* for allowing me to scrape this website.

[19]Reviews and ratings are pooled across the *Gartner Digital Markets network*, which comprises *Capterra* as well as two other subsidiary websites (*GetApp* and *Software Advice*).

[20]Based on comparisons with its competitors, information on reviews and ratings seem accurate and representative. *Capterra*'s main competitor is the platform *G2*, which provides a similar vertical search engine with reviews, categories and descriptions on enterprise software products. As of July 2021, the three *Gartner* owned websites had a somewhat larger number of monthly visits (over 10 Million) than the platform *G2* (8.5 Million), and it is available in over 30 countries and at least seven languages. Looking at individual products, the relative number of reviews - an indicator of demand - seemed comparable between *G2* and *Capterra*. Using the Internet Archive ("Waybackmachine"), I found at least anecdotally that products that were discontinued were removed earlier from the *Capterra* website than from *G2*.

[21]I also save, but do not currently use, a text describing the intended audience for the given product; pricing information; company headquarter location; the year in which the company was founded; and the time and date of each instance of scraping. See Appendix 1.C.3 for details on the web-scraping process.

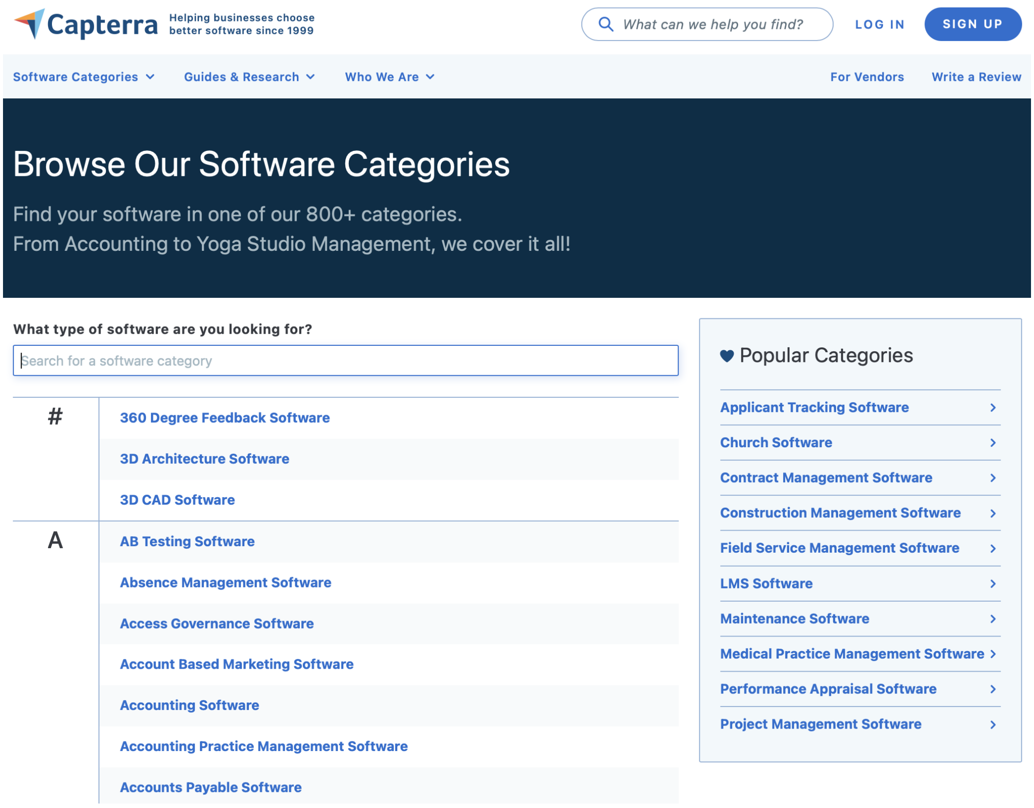


Figure 1.1: *Capterra*'s categories page

Matching *Capterra* to *Crunchbase* data

I match products on *Capterra* to their respective firms on *Crunchbase* using company URL and company name (see Figure 1.2). Certain products originate from one company, but are now associated with a different company's URL and name that has in the past acquired the originating firm. I therefore make sure to match products to their respective originating companies, instead of to the name of the new owner, by additionally using the names of all acquired companies for each firm that has been an active acquirer in the past.[22]

I end up merging 71% of all web-scraped products, accounting for 96% of products with over 100 reviews, to firms on *Crunchbase*. Almost all remaining non-matched products do not have many reviews, and are thus likely insignificant competitors that do not play a major role in this market. Manual checks confirm a very high accuracy of this matching procedure.

From the remaining firms in the *Crunchbase* data, I moreover include firms into my sample that (1) are enterprise software related based on their descriptive text, industry group and industry variable, *and* (2) that were acquired by a firm that was matched and thus owns a product on *Capterra*.[23] The products of these acquired, enterprise software related companies do not seem to be present on *Capterra*, and were therefore likely not continued (or not even developed) under the original name after the acquisition

[22] Details of the matching procedure are provided in Appendix 1.C.4.

[23] To identify (likely) enterprise software related firms on *Capterra*, I first manually divide a set of firms into either enterprise software related companies, or other firms. Based on this, I develop selection criteria that employ *Crunchbase*'s descriptive text, industry group and industry variable, and that allow me to select enterprise software related companies from *Crunchbase* systematically.

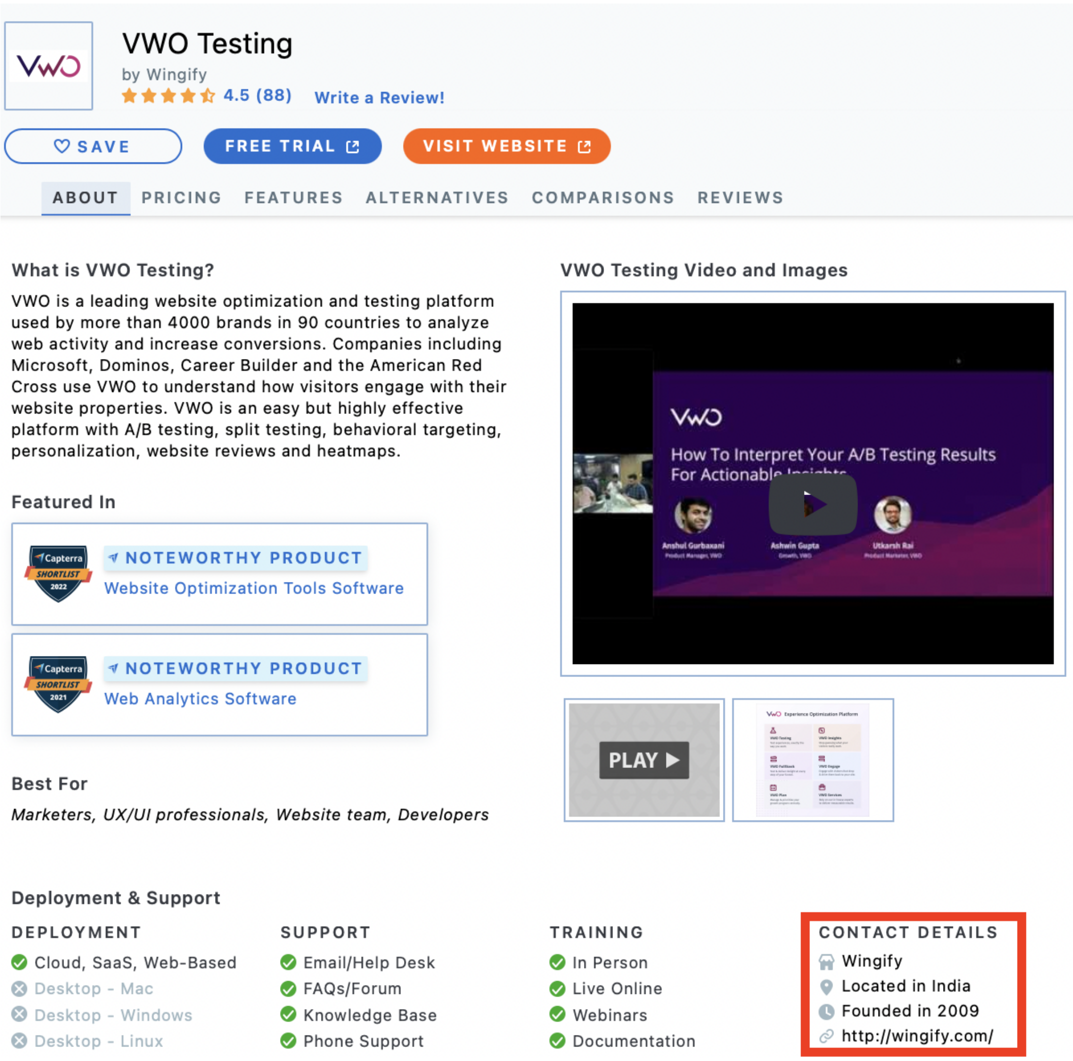


Figure 1.2: Example of product page on *Capterra*. The red frame highlights the company information (in particular, name and URL) available for all products on *Capterra*.

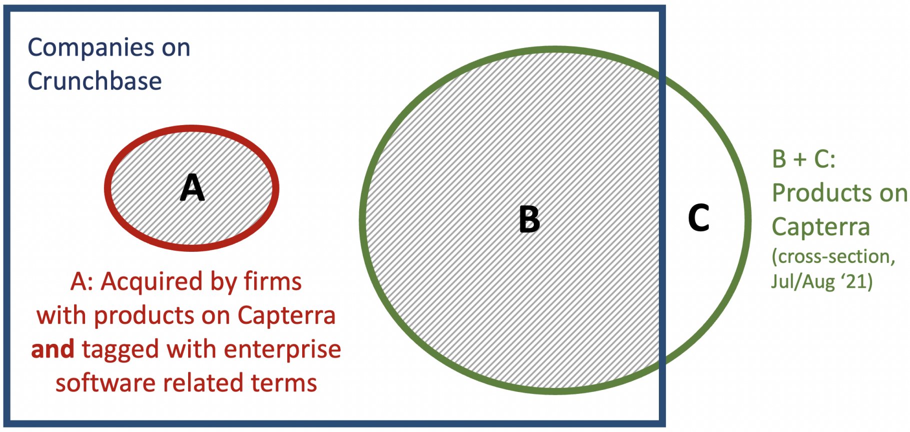


Figure 1.3: Illustration of the created sample. The hatched region (A and B) illustrates the sample used. Set B is obtained by the match of *Capterra* products to *Crunchbase* firms. Set A is added to account for significant companies acquired in the past, but shut down as of 2021. Set C is the (likely insignificant) set of products on *Capterra* that was not matched to companies on *Crunchbase*.

took place. Nevertheless, this set of acquired firms should likely be considered as relevant entrants and competitors, and thus as part of the market. Figure 1.3 summarizes the types of firms that are part of the sample. Appendix 1.C.5 provides a further discussion as well as evidence against potential selection issues.

The final dataset contains 46,186 currently existing products and their respective companies' events, as well as the events of 5,034 enterprise software companies that were acquired and whose products are not existing under the same name on *Capterra* any more.

2.3 Defining Markets using Textual Analysis

I create markets of substitutable products by using text-as-data methods.[24] Each product on *Capterra* is associated with a body of text. This text consists of the names of one or more software categories, and of the product description. As a given product may be associated to more than one category, one cannot create disjoint sets of products using *Capterra*'s categories alone. I therefore (1) extract keywords contained in the category name or in the descriptive text, (2) vectorize this textual data (following Decarolis and Rovigatti (2021)), and (3) use a clustering algorithm that creates non-overlapping groups of likely substitutable products.

The textual information on *Capterra* are informative about a product's functionalities, in the sense that companies present in the same (or similar) categories, and described with similar keywords, should be more substitutable: *Capterra*'s purpose is to guide consumers searching for specific enterprise software products. Just like other vertical search engines or product platforms, the company therefore has

[24] See Gentzkow, Kelly, and Taddy (2019) for a review on these methods.

an incentive to provide accurate product categories and descriptions.[25]

I build a dictionary of meaningful keywords by using all *category names* (e.g., "filesharing" for "File-sharing Software"), as well as additional keywords that are frequently occurring in *Capterra*'s product description. Details can be found in Appendix 1.C.6.

To cluster all products into disjoint markets, I first embed the textual information into a vector space that carries linguistic meaning. I follow the approach taken by Decarolis and Rovigatti (2021): I first match each keyword, for instance, "file-sharing" or "collaboration", to a pretrained word vector stemming from *GloVe*, an unsupervised learning algorithm for obtaining vector representations for words (Pennington, Socher, & Manning, 2014).[26] I thereby place each keyword at a certain location within a 300-dimensional vector space. Synonyms and terms that are linguistically close to each other tend to be located close to each other in this space. For each product, I then take the average of all its word vectors, so that each product is associated with a single location.

Next, I cluster products (based on their respective locations in the vector space) into distinct markets using a k-means clustering algorithm (see Appendix Section 1.C.7). Products whose vectors are located close to each other, and thus, whose descriptors are close in meaning, will be clustered into the same product group.

The k-means algorithm requires the researcher to provide a number of segments ex ante. I employ the silhouette score as guidance, which measures the goodness of a given clustering technique. I find that clustering into 500 to 600 markets maximizes the silhouette score, and results in reasonable market definitions based on various manual validation checks. For instance, when comparing my market definitions to the market definitions from merger decisions by the UK Competition and Markets Authority, I find that the majority of products are correctly categorized as substitutes (see Appendix Section 1.C.8).[27]

Table 1.1 shows basic descriptive statistics of the matched raw dataset for the period of 2012 to 2020. The dataset covers a sample of over 20,000 firms. The majority of these firms – 65% – are indeed VC-funded. In contrast, only 4.6% of producing companies are (at any point in the observation period) public firms, showing that a lot of relevant entry behavior would be missed if one were to focus on only public firms. Table 1.2 exhibits descriptives on the market-quarter panel. It becomes clear that the data tend to be right skewed.

[25]Previous research on the market for mobile applications has made use of the fact that app categories are meaningful for defining competitors (e.g., Affeldt and Kesler (2021b), Ershov (2022), Janssen, Kesler, Kummer, and Waldfogel (2022), or Yin, Davis, and Muzyrya (2014)). I essentially proceed in a similar way, although I cannot rely on the categories alone, as these are overlapping in my case. *Capterra* confirmed to me that categories and text are accurate. According to the company, new products are being placed into a single category when they are introduced on the website, upon which companies can request to be added to further categories. A dedicated catalog team will then review the request and approve the product if the additional category seems suitable. For enterprise software related companies acquired in the past and whose products are not available on *Capterra* – set A in Figure 1.3 – I extract the same keywords from *Crunchbase*'s descriptive text.

[26]The word vectors were trained on Common Crawl, i.e. textual data stemming from craweling the web, which is very suitable for my purpose.

[27]The textual analysis in principle allows for distance metrics between markets. The current version of this paper does not make use of this, as the purpose is to obtain disjoint sets of markets that are assumed to be independent to be able to conduct regression analysis and to estimate a structural model. In future research, I plan to create distance measures between products, and to conduct additional analyses.

Number of products	25,552
· Percent of products alive	80.9%
Number of companies	21,419
· Percent of companies ever VC-funded in 2012-20	63.9%
· Percent of companies ever public in 2012-20	4.5%
Number of acquisitions	6,778
· Percent in which target is VC-funded	42.4%
Number of IPOs	384
· Percent in which firm going public is VC-funded	54.4%

Table 1.1: Basic descriptives of entire matched data, 2012-2020. I exclude LBOs and management buy-outs from the acquisitions.

	Mean	St. Dev.	Pctl(25)	Median	Pctl(75)
"Pre-event" firms (count)	4.396	4.926	1	3	6
VC-funded, pre-exit startups (count)	15.362	15.377	5	10	20
Acquired & alive startups (count)	1.547	2.149	0	1	2
Public firms (count)	3.777	3.721	1	3	5
Startups entering (count)	0.651	1.033	0	0	1
Startups acquisitions (count)	0.161	0.542	0	0	0
Startups IPOs (count)	0.020	0.141	0	0	0
Startup acquisitions: transaction price	395.2	945.8	40	130	360
Startup IPOs: valuation	3,774.7	13,875.4	352	885.4	2,165.2

Table 1.2: Descriptives using market-quarter panel, comprising 474 markets (after dropping markets that I view as outliers), 2012-2020. Prices and valuations in million US$. Pre-event firms are companies that are less than 3 years old (based on their founding date) and have not recorded any other event yet (in particular, no funding round). I therefore do not consider these as startup firms.

3 Stylized Facts

This section lays out empirical facts that motivate the research question, guide the modeling assumptions, and are building blocks towards the model-based results. I distinguish and document different types of acquirers, along the dimensions of whether the acquirer is active in the industry sector of enterprise software, and based on measures of age (Section 3.1). The findings can be summarized as follows:

1. The different acquirer types acquire different types of targets, reflecting their heterogeneous motives (Section 3.2).

2. Many acquired products are discontinued (Section 3.3).

3. There is suggestive evidence that most acquisitions are nonhorizontal (Section 3.4).

3.1 Distinguishing Different Types of Acquirers

I identify three main types of acquirers.

- *Companies in enterprise software*: these companies have existing products on *Capterra* that do not stem from a previous acquisition, and are thus active producers of enterprise software.
  - Examples: the so-called GAFAM; Cisco; Oracle; Salesforce; VMware.
- *Financial companies*: I identify these companies as active in finance, based on *Crunchbase* information.[28] Among these are private equity firms.
  - Examples: Vista Equity Partners; TransUnion; Thoma Bravo.
- *Other industries, i.e. companies outside of Enterprise Software and Finance*: these companies do not have existing products on *Capterra*, and are thus mainly active in other industries.[29]
  - Examples: The We Company; Verizon; Dentsu International; Samsung Electronics; Ericsson.

The fractions of these three main acquirer types are displayed in Figure 1.4. Over 65% of acquisitions of exiting startups are conducted by other industry peers. 14% of acquisitions are carried out by financial firms, and 20% are carried out by firms that are neither active in enterprise software, nor in finance. Further characteristics on these three types of acquirers can be found in Appendix 1.D.

I divide enterprise software acquirers into further (non-exhaustive) sub-groups along the measures of age or firm maturity, and innovativeness (measured as having received VC funding in the past). Moreover, I segment the GAFAM firms from the others, as those have been the focus of attention by competition policy practitioners, and are deemed to be especially dominant in many markets. These

[28] To do so, I use *Crunchbase*'s industry tags. Moreover, *Crunchbase* tags companies that act as investors with an "investor type" variable (this may be, for instance, "Investment Bank" or "Private Equity Firm").

[29] Among these are also holding companies: I define these as all companies that do not produce software products themselves, but acquire software companies and seem to hold software products in a portfolio. Using *Crunchbase*'s industry tags, I find that over half of Industry Outsider acquirers are active in related industry sectors, such as (other) software (e.g. StackPath), advertising (e.g. Amobee), data/artificial intelligence (e.g. Amdocs), media/content (e.g. Groupon), or hardware/telecom (e.g. Verizon). The other half of Industry Outsider acquirers is active in less related industry sectors, such as transportation, consumer products, e-commerce, or biotech.

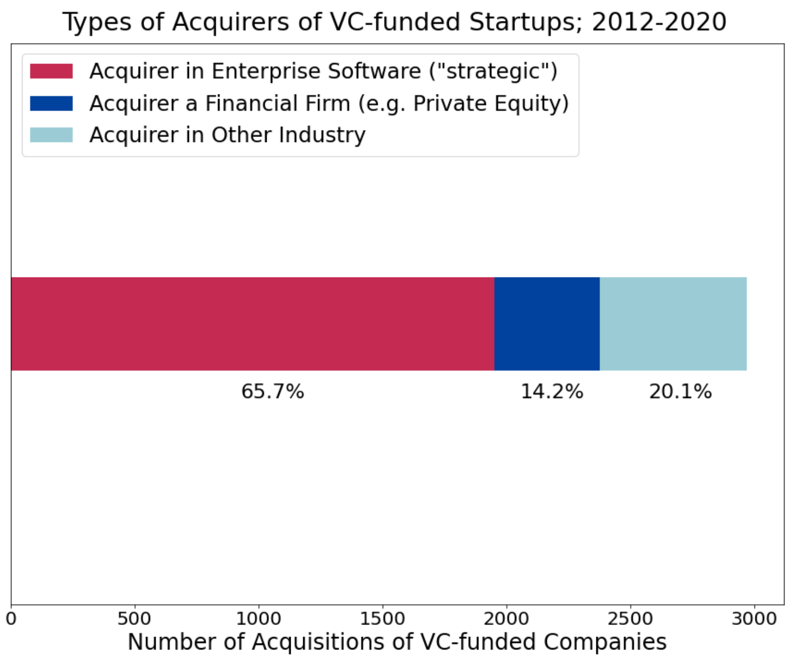


Figure 1.4: Types of acquirers for first-time acquisitions ("exits") of VC-funded startups worldwide in the domain of enterprise software, for acquisitions occurring between 2012 and 2020. The total number of such acquisitions is 2,973. Acquisitions are counted on the company (as opposed to the product) level. I exclude acquisitions of the types LBO or management buyout.

Acquirer type (# of startup acq)	Description	Examples
GAFAM (156 acq)	Google (Alphabet), Apple, Facebook (Meta), Amazon, Microsoft and their subsidiaries.	GAFAM, LinkedIn, AWS, GitHub.
Old tech (190 acq)	Public companies founded prior to 1995 with over 10,000 employees.	Cisco, Oracle, VMware, SAP, Dell EMC, HP Enterprises, IBM, Adobe.
New tech (174 acq)	Companies founded 1995 or later that started off as VC-funded companies, but that have exited.	Salesforce, Palo Alto Networks, Workday, Servicenow.
Pre-exit (630 acq)	VC-funded startups acquiring at a time at which they have not "exited" (been acquired / gone public) yet.	Sprinklr, Freshworks, Ignite Technologies, Dropbox, DataRobot, Stripe, Hootsuite.

Table 1.3: Definitions of subgroups of enterprise software acquirers. These groups are distinct, but not exhaustive. The number of acquisitions focuses on exiting VC-funded startup acquisitions that were carried out in the years of 2012-2020. (For the category "new tech", using only VC-funded companies avoids taking into account spin-offs from older companies that have a very recent founding date, such as Hewlett Packard Enterprise.)

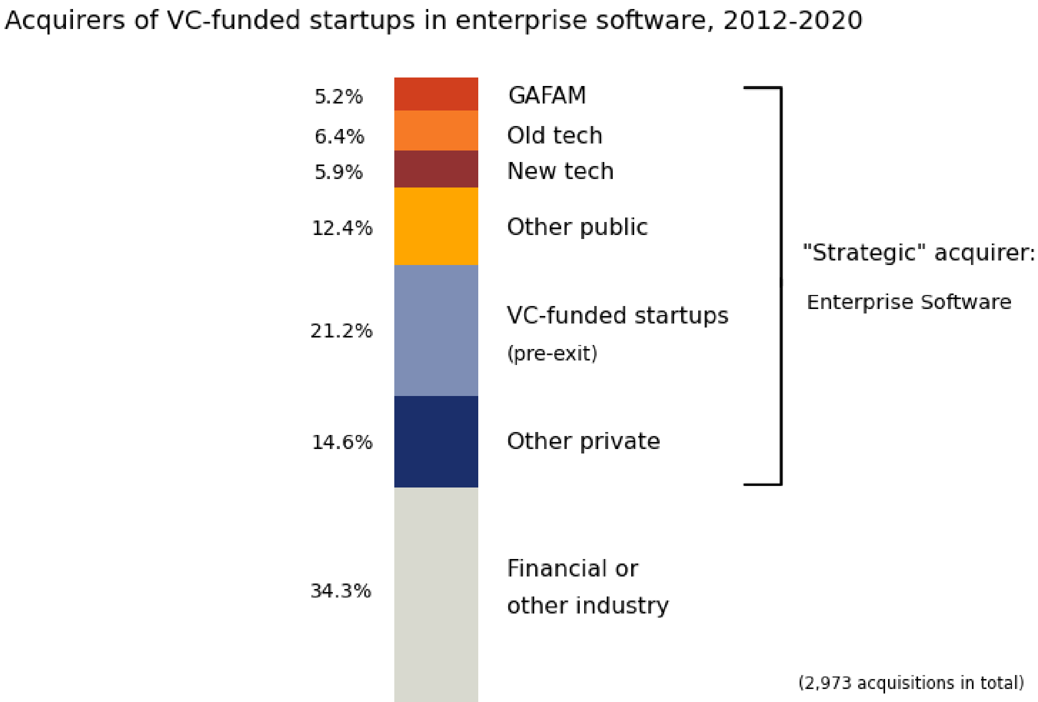


Figure 1.5: Subgroups of acquirers for first-time acquisitions ("exits") of VC-funded startups worldwide in the domain of enterprise software, for acquisitions occurring between 2012 and 2020. As in Figure 1.4, the total number of such acquisitions is 2,973; the numbers are on the company (as opposed to product) level; and I exclude acquisitions of the types LBO or management buyout.

sub-groups are detailed in Table 1.3, and their proportions are shown in Figure 1.5. Note that companies may switch between these categories as they grow: for instance, Dropbox acquisitons are contained in the category *pre-exit* for the years in which Dropbox had not exited yet, and are contained in the category *new tech* after Dropbox has become a public company.

An interesting and perhaps surprising fact is the scale at which *other* startups appear to be a major exit route for growing startups: companies within the groups GAFAM, Old Tech and New Tech conducted each roughly 150-200 startup acquisitions in the years of 2012-2020, whereas pre-exit firms account for over 600 startup acquisitions. Therefore, out of all startups exiting via acquisition in 2012-2020 in the domain of enterprise software, 21% were sold to other startups. In contrast, only somewhat more than 5% were sold to GAFAM firms.

3.2 Different Acquirer Types Acquire Different Sets of Targets

Next, I turn to the question: for each of these different sets of acquirers, what are the likely motives for acquiring? I try to reveal acquirers' motives by studying the characteristics of the acquired companies for each of these acquirer types.

I start with a more aggregate pattern, and find out to what extent target firms are *VC-funded startups*, as opposed to other, non VC-funded companies, for each of these acquirer types. The numbers are detailed in Table 1.4. Looking at Panel A, what is noteworthy is that roughly half (48.9%) of targets acquired by enterprise software firms are VC-funded, pre-exit startups. This number is much lower for

Panel A: "Broad" groups of acquirers (exhaustive, covering all observations) *(in %)*

Acquirer type	≤3y old and no VC funding (yet)	VC-funded, pre-exit ("startup")	VC-funded, post exit	Not VC-funded (and >3y old)	Total
Enterprise Software	6.4	48.9	3.6	38.7	100
Financial	2.9	29.1	5.0	62.1	100
Other Industries	4.4	36.7	4.6	53.3	100

Panel B: Looking at subgroups of enterprise software acquirers

Acquirer type	≤3y old and no VC funding (yet)	VC-funded, pre-exit ("startup")	VC-funded, post exit	Not VC-funded (and >3y old)	Total
GAFAM	9.8	72.0	2.8	12.6	100
New tech	5.5	76.4	4.7	11.8	100
Old tech	1.9	63.6	8.4	20.7	100
Pre-exit (VC-funded)	9.4	51.8	4.7	31.6	100

Table 1.4: Which types of companies are acquired by different types of acquirers? I use data from 2012-2020. I exclude leveraged buyouts and management buyouts, but otherwise place no restriction on the type of company acquired.

financial firms (29.1%) or firms in other industries (36.7%), which both tend to acquire non-VC-funded firms. Panel B shows a result that is particularly interesting from a policy perspective. Comparing the different subgroups of enterprise software acquirers, the pattern of firms acquired by the GAFAM is closest to New Tech firms. For both groups of firms, the share of targets that are VC-funded is very high, amounting to more than 70%. Similarly, the share of targets that are very young companies that have no prior funding history is also very high. Old Tech and pre-exit firms tend to be more active acquiring firms that are non-VC funded. Moreover, Old Tech firms very rarely acquire very young companies with no prior funding history. All in all, whereas the Old Tech firms tend to dominate markets in a similar fashion as the GAFAM, they apparently pursue an acquisition strategy that is quite different from the GAFAM.

For the remainder of this section, I consider only acquisitions in which the target was a VC-funded startup. I first compare the maturity of startups at the time of acquisition by different types of acquirers. In particular, I consider acquisition price and valuation (Table 1.5) and age (Table 1.6) at exit.[30] I observe the following pattern: enterprise software firms tend to acquire firms that are younger, and at lower prices and at lower valuations, compared to financial acquirers. Moreover, we observe a striking amount of heterogeneity between the sub-groups of enterprise software firms. Notably, Old Tech firms tend to acquire at a higher age, at the highest price, and high valuations. The same pattern is observed when looking at the amount of funding a startup has received at exit (see Appendix 1.D, Table 1.20). Startups acquired by Old Tech firms thus tend to be quite mature, and the Old Tech's acquisition pattern somewhat resembles that of financial firms. In contrast, the New Tech firms, but in particular the GAFAM, acquire VC-funded startups at lower prices, lower valuations, and at a much lower age. For pre-exit firms, the acquisition pattern instead points to the possibility that pre-exit firms might tend to acquire mainly financially distressed startups, as acquisition prices are either missing or very low (Kerr, Nanda, & Rhodes-Kropf, 2014).

Table 1.7 looks at the average time span between the last funding round raised, and the date of ac-

[30]One caveat of prices and valuations is that this amount is very often missing; most likely particularly at the low end. I therefore also report the percent of observations in which the price or valuation variables are not available.

Panel A: "Broad" groups of acquirers (exhaustive, covering all observations)

Acquirer type	**Acquisition price** (million USD; median)	**Valuation at exit** (million USD; median)	% acquisition price is not available	% valuation is not available
Enterprise Software	120.0	25.0	81.0	94.6
Financial	150.0	59.0	88.4	95.5
Other Industries	100.0	26.2	79.9	93.4
Panel B: Looking at subgroups of enterprise software acquirers, and at IPOs				
GAFAM	164.0	49.8	78.2	90.4
New tech	152.5	263.0	60.7	92.9
Old tech	400.0	476.0	74.2	92.6
Pre-exit	15.8	4.4	95.8	94.3
At IPO	-	-	1000.0	69.1

Table 1.5: Acquisition prices and valuations at exits of VC-funded startups (left columns) , as well as the percent of observations in which valuation or acquisition price are not available (right columns). 2012-2020. Excludes leveraged buyouts or management buyouts.

Panel A: "Broad" groups of acquirers (exhaustive, covering all observations)

Acquirer type	Age (median # of years since founding date)	Age (median # of years since first funding round)
Enterprise Software	6.6	4.5
Financial	9.8	5.5
Other Industries	7.2	5.2
Panel B: Looking at subgroups of enterprise software acquirers, and at IPOs		
GAFAM	4.6	3.4
New tech	4.9	3.7
Old tech	7.3	5.1
Pre-exit	5.6	3.8
IPO	10.6	7.7

Table 1.6: Age at exits of VC-funded startups, 2012-2020. Excludes leveraged buyouts or management buyouts.

Panel A: "Broad" groups of acquirers

Acquirer type	# of years since last funding round (mean)
Enterprise Software	2.7
Financial	3.5
Other Industries	3.3
Panel B: Subgroups of enterprise software acquirers, and IPOs	
GAFAM	1.8
New tech	1.8
Old tech	2.4
Pre-exit	2.4
IPO	2.2

Table 1.7: Time (in years) since last funding round at time of exits of VC-funded startups, 2012-2020. Excludes leveraged buyouts or management buyouts.

quisition, for the different types of acquirers. The rationale for doing so is that startups that have very recently raised new capital should not face strong financial constraints. These firms may have relatively higher bargaining power and presumably do not get sold due to a fire sale. This time span tends to be particularly low for GAFAM and New Tech acquirers. In contrast, acquisitions by financial and other acquirers happen nearly twice as long after the latest funding round. As acquisitions are negotiated between startups and acquirers, this could reflect entrepreneurs' preferences for selling to one of the GAFAM, as opposed to other firms, thereby indicating GAFAM's strong bargaining power. These numbers could also be a sign that GAFAM tend to acqurie pre-emptively, or have better information on the quality of the startups, compared to other acquirers.

A final finding concerning all acquirer types is that many acquirers are serial acquirers. For enterprise software and financial firms, the median number of acquisitions of any industry during the company's lifetime is 8 (5 for companies in other industries). This mirrors David (2021), who emphasizes that serial acquisitions are a ubiquitous feature in the economy.

3.3 Many Acquired Products are Discontinued After the Acquisition

As explained in Section 2.2, the data I created contain companies that were acquired in the past, but whose products are not available any more under the same brand name. For acquisitions of VC-funded, enterprise software related startups in 2012-2020, I find that in a majority – 57% – of acquisitions, the product has been discontinued under the same brand name after the acquisition, as of 2021. These numbers align with recent literature studying GAFAM-acquisitions: Affeldt and Kesler (2021a) consider over 50 GAFAM-acquired mobile apps and find that half of these apps are discontinued. Gautier and Lamesch (2021) find that the GAFAM shut down the companies in 60% of all cases. My results show that this carries over to other acquirers active in the software industry, and seems to be a widespread phenomenon in software.

Shutdown rates vary depending on the acquiring firm. Table 1.8 shows that shutdowns are especially

Panel A: "Broad" groups of acquirers (exhaustive)

Acquirer type	**Discontinuations,** percent	**Discontinuations,** count
Enterprise Software	67.1%	1322
Financial	36.1%	153
Other Industries	38.3%	231
All acquirers	56.9%	1706

Panel B: Subgroups of enterprise software acquirers

Old tech	72.1%	137
New tech	64.9%	109
GAFAM	80.8%	126
Pre-exit	66.8%	432

Table 1.8: Discontinuations of products post-acquisition, for different types of acquirers, and for startups acquired in 2012-2020.

	Age: years since founding (median)	**Age**: years since first funding round (median)	**Price** in US$ million (median)
Products discontinued	6.2	4.0	100.0
Products kept alive	7.8	5.2	136.8

Table 1.9: Heterogeneity in age at acquisition and in transaction price, for startups whose products were either discontinued (top), or kept alive (bottom).

prevalent for acquirers that are enterprise software firms themselves; these companies discontinue the acquired product in 67% of all acquisitions. Financial firms, in contrast, discontinue the acquired products in only 36% of all acquisitions.

The acquired companies whose products are shut down are at the median one to two years younger at the time of acquisition (Table 1.9), and acquired at 75% of the price, compared to continued products[31]. The shutdown rate is even higher and amounts to 75% for companies that were acquired at an age of less than 3 years and that have not received any funding yet (and are thus not considered startups based on my definition). All of this suggests that many of the shutdown products did not have a large share of demand at the time of acquisition, and possibly did not yet have a fully developed product. Appendix 1.E contains further details on these acquisitions.

3.4 Suggestively, Most Acquisitions Are Nonhorizontal

I call acquisitions "horizontal" if a startup supplies a product that competes with an acquirer's existing product in the same narrow market as of 2021. According to this definition, and using the above narrow market definitions, I find that only 8% of all acquisitions of VC-funded startups in 2012-2020 can be classified as horizontal.[32] Anecdotally, it seems that most acquisitions could instead be classified as either vertical, or conglomerate type.

[31] Not in Table. Prices are missing in 83% of shut-down acquisitions, and in 77% of continued acquisitions. As presumably low prices are missing more often (Kerr et al., 2014), the difference in median acquisition prices might therefore well be even higher.

[32] I also find variation in the number of horizontal mergers across different acquirer types. However, this variation is not very insightful, as it correlates by construction with the number of enterprise products supplied by the acquirer.

However, there are caveats to this observation. First, it is impossible to obtain information on products that are in the development stage within the acquirer's boundaries: an acquirer acquiring a target supplying a product that is complementary to its internal research efforts (which are unobserved) are therefore not classified as being horizontal, according to this definition. The second caveat is that I take market defintions to be static, whereas a product's market might in principle change over time. These caveats point to the importance of future research in this area.

3.5 Discussion

What are the motives behind the shutdown acquisitions that I find? Product shutdowns could in principle be so-called killer acquisitions[33]. The data however suggest that these types of acquisitions might be rare in the context of enterprise software. First, the vast majority of acquired firms in this industry are very young and sometimes have not even raised a single funding round. Thereby, the bulk of firms seem to not be very likely to be a serious threat to a major incumbent such as Google. Second, the finding that most acquisitions are nonhorizontal makes them less likely to be killer acquisitions. Moreover, as Table 1.8 shows, shutdowns are prevalent among companies with much less market power than Google and the likes. Even startups that have not exited yet and that are very young shut products down in 67% of the acquisitions they undertake. Preexit startups or "new" tech firms account for a much larger share of discontinued startups than the GAFAM.

What may the purpose of acquiring and discontinuing products be? Anecdotally, acquired products are oftentimes integrated into the acquirer's existing product as an additional feature or functionality, or to otherwise improve the existing product, if the acquirer is an enterprise software firm.[34] Some of the transactions seem to be so-called acqui-hires in which the acquired startup's employees are paid to become part of the acquiring company.[35] For financial acquirers, the motive of discontinuing product might be somewhat different. Anecdotally, it seems that financial acquirers more often merge (and possibly restructure) two companies in their portfolios, rather than entirely discontinuing or acqui-hiring target companies.[36] I have also found cases in which the product was rebranded. However, any rebranding seems to have gone along with a number of changes to the original product.[37]

The difference in the age profile of acquired startups between enterprise software and financial firms

[33] Killer acquisitions comprise 5.3 to 7.4 percent of acquisitions in the setting of pharmaceuticals studied in Cunningham et al. (2021).

[34] For instance, according to news reports, this may have been the case with Amazon's acquisition of the data warehousing company Amiato, see `https://techcrunch.com/2015/04/20/amazons-aws-acquired-amiato/`; Google's acquisition of app performance startup Pulse.io, see `https://venturebeat.com/2015/05/28/google-acquires-mobile-app-performance-startup-pulse-io/`; or Upskill's acquisition of Pristine, see `https://www.prnewswire.com/news-releases/augmented-reality-industry-leader-upskill-acquires-pristine-300453872.html` (all accessed 07/08/2022).

[35] Examples are *Dropbox-Verst*, *Google-Bebop*, *Apple-Union Bay Networks*, *Twitter-tenXer*, and *Box-Wagon*. In 3% of startup shutdown-acquisitions, the *Crunchbase* data in fact indicate that the acquisition is an acqui-hire. I believe the actual number of acqui-hires to be rather higher. For instance, whenever the acquirer announced the shutdown at the time of the acquisition, the acquisition may quite likely have been an acqui-hire. Note that, interestingly, Ng and Stuart (2021) find that a acqui-hired employees turn over at a much higher rate compared to organically hired employees.

[36] One example is the alternative data company *7Park Data*, which was acquired by *Vista Equity Partners* and later folded into *Apptio*, another one of *Vista Equity Partners*'s portfolio firms. Another example is *SCIO Health Analytics*, which was acquired by the holding group *ExlService Holdings* and is now part of its product *EXL Health*.

[37] An example is the acquisition of Acompli, a mobile email and productivity app, by Microsoft. The product was rebranded as Outlook Mobile a month after the acquisition; see, e.g., `https://www.theverge.com/2015/1/29/7936081/microsoft-outlook-app-ios-android-features` (accessed 07/08/2022).

is in line with the fact that financial firms acquire tested products, as presumably these firms are interested in obtaining cashflows. In contrast, enterprise software firms might even be interested in acquiring startups whose products do *not* yet have a customer base. As software is based on communication protocols and programming languages, different pieces of software are interoperable, and software can be created in a modular way. Moreover, a startup producing a tool that is in principle functioning, or that was created by a capable team, might be an interesting target for another software firm even if these products failed to attract demand. This aspect is very different in the pharmaceutical market and may thus be an explanation for why we do not see as many acquisitions of very young startups in the domain of pharmaceuticals or biotech, as shown in Appendix 1.C.10.

4 Reduced-form Evidence on Acquisitions and Entry

As pointed out in Section 3, acquirer types most likely differ in important ways in their respective motives when acquiring startups. Moreover, one may argue that only certain types of acquirers have the capabilities and the incentives to deter follow-on entry upon acquiring a startup in a market. In particular, only firms active in the same industry of enterprise software – which I call *strategic* – may possess complementary assets, resources, or market power that could fundamentally influence the acquired product's capabilities to compete in a given market. These types of acquirers may also have a strong incentive to fundamentally affect competition in their favor following the acquisition in a given market, as they may acquire to enter new markets, or to build a software ecosystem. These potential entry-deterring effects may be stronger if the acquirer is more dominant (e.g., Denicolò and Polo (2021), Kamepalli et al. (2021), Motta and Shelegia (2021)), or larger and thus more likely to possess resources to create a synergistic value.

This contrasts with the intentions and capabilities of acquirers in financial and other industries. Many the financial acquirers that I observe are private equity investors. These are typically transitional owners of the acquired firms, and tend to be focused on generating cashflows in the medium term by changing a companies' management, with the intention of later reselling the company. For acquirers active in other industries, acquisitions in enterprise software may often be vertical integrations of software products. I also count as other industry an acquirer who does not produce software itself, but may be a holding company that hold a portfolio of software products and that yield stable returns.[38] An acquisition by a non-enterprise software acquirer therefore is a transition in the ownership of a startup that should however not fundamentally affect market structure and competition in a way that deters follow-on entry.[39] Therefore, I pose the following hypothesis:

- **Hypothesis**: Acquisitions conducted by a strategic acquirer may subsequently decrease entry into a given market. This effect should be stronger if the strategic acquirer is dominant. The effect is absent for acquisitions undertaken by a acquirers active in other industries.

[38] Examples are *Valsoft* or *Ropers Technologies*.

[39] At best, the effect should be positive, for instance if the acquired product is subsequently used in-house, but discontinued to previous customers. New entrants should then expect more demand.

I attempt to shed light on this hypothesis with the help of an event study framework. I employ quarter-market panel data ranging from 2012-2020, and study this hypothesis using the following linear model:

$$num\_entrants_{m,t} = \beta \sum_{\tau=0}^{K} acquisition_{m,t-\tau} + \lambda_m + \lambda_t + \epsilon_{m,t} \tag{1.1}$$

$num\_entrants_{m,t}$ denotes the number of VC-funded startups entering in a given market m at quarter t. The variable $acquisition_{m,t-\tau}$ is a binary variable that takes on the value 1 if an acquisition of a certain type took place in market m and quarter $t-\tau$, and 0 otherwise. K is the event window, which I set to 4 in my preferred specification. The coefficient of interests is therefore β. λ_m and λ_t denote market and quarter fixed effects, and $\epsilon_{m,t}$ is an econometric error term.

Entry-deterring effects are expected to be more likely when the acquired startup is more valuable, and when the startup's product continues to be developed and marketed after the acquisition. I therefore study acquisitions of VC-funded, private startups at a transaction price above 100US$ million, and focus only on acquisitions in which the product has not been discontinued.[40] I drop LBOs or management buyouts. I consider broad, as well as more narrow definitions of "strategic" and "financial" acquirer types. The broadest definition of strategic acquirers considers all enterprise software acquirers; more narrow definitions consider subsets of these. Similarly, the broadest definition of financial acquirers considers both financial as well as industry outsider firms.[41]

Table 1.10 displays the results. In columns (1), (2) and (3), the acquirer is a strategic acquirer, whereas in columns (4) and (5), the acquirer is a financial acquirer or an acquirer from another industry. The results provide suggestive support for the hypothesis. Major startup acquisitions by strategic acquirers – both using wide as well as more narrow definitions – tend to be followed by a decline in entry. This pattern is less prevalent for financial acquirers. The result holds when decreasing the threshold of "major" acquisition to a transaction price of 50US$ million.

I perform test for a possible anticipation effect by asking: are acquisitions of these different acquirers *preceded* by a decline, or by an increase, in entry? Table 1.11 suggests that only major acquisitions by public enterprise software companies may be preceded by a significant drop in entry.

One concern might be that "treated" markets, i.e., markets in which a large acquisition took place at *any* point, might differ in terms of observables or unobservables compared with markets in which no such acquisition occurred. In Table 1.12, I perform the event study using only markets in which *any* major acquisition occurred. Even in this setting, which has a much smaller sample size, the coefficients retain the same sign as before.

Even though these regression results do not allow for a causal interpretation, they are interesting

[40]The median transaction price for these VC-funded startups with continued products is 168US$ million. I drop acquisitions that occurred in the first τ or the last τ quarters of the time period under study. In case there are multiple such acquisitions in a given market-quarter or just in adjacent time periods, I continue to set the indicator equal to 1. Finally, note that, as mentioned previously, acquisition prices are often not observed. However, I expect that prices are highly likely to be observed conditional on the price being high; thus, I expect to capture all major acquisitions in this analysis.

[41]To give examples of events used in these regressions: major acquisitions by enterprise software companies include *Dropbox-DocSend*, *Google-Looker*, *Microsoft-Yammer*, *Amazon-CloudEndure*, *Docusign-SpringCM*, or *Oracle-Moat*, for instance. Examples of major acquisitions by financial companies are *LiveU-Francisco Partners*, *Acquia-Vista Equity Partners*, or *Smartly.io-Providence Equity Partners*. Exmaples of major acquisitions by companies in other industries are *Rocke-Flatiron Health*, *McDonald's-Dynamic Yield*, *Continental-Zonar*, or *Dupont-Granular*.

Table 1.10: Event study: acquisitions and entry patterns, using an event window of 4 quarters. Market-quarter panel, 2012-2020.

	Dependent variable:				
	Number of entrants in market m, quarter t (Sample mean: 0.65)				
	Strategic acquirer			Financial acquirer	
	(1)	(2)	(3)	(4)	(5)
Major acq by enterprise software company	−0.112*				
(89 acquisitions)	(0.059)				
Major acq by public enterprise software company		−0.158**			
(59 acquisitions)		(0.075)			
Major acq by GAFAM or 'New Tech'			−0.401***		
(21 acquisitions)			(0.135)		
Major acq by company not in enterpr softw (incl. financial)				−0.101	
(40 acquisitions)				(0.072)	
Major acq by financial company					0.032
(13 acquisitions)					(0.119)
Market FE	✓	✓	✓	✓	✓
Quarter FE	✓	✓	✓	✓	✓
Adjusted R^2	0.299	0.299	0.3	0.299	0.299
Observations	17,064	17,064	17,064	17,064	17,064

Standard errors in parentheses, clustered at market level. *p<0.1; **p<0.05; ***p<0.01

Table 1.11: Testing for anticipation effects: are events preceded by more, or less entry? Market-quarter panel, 2012-2020.

	Dependent variable:				
	Number of entrants in market m, quarter t (Sample mean: 0.65)				
	Strategic acquirer			Financial acquirer	
	(1)	(2)	(3)	(4)	(5)
Major acq by enterprise software company	−0.089				
(89 acquisitions)	(0.075)				
Major acq by public enterprise software company		−0.136**			
(59 acquisitions)		(0.064)			
Major acq by GAFAM or 'New Tech'			0.103		
(21 acquisitions)			(0.179)		
Major acq by company not in enterpr softw (incl. financial)				0.038	
(40 acquisitions)				(0.105)	
Major acq by financial company					0.006
(13 acquisitions)					(0.149)
Market FE	✓	✓	✓	✓	✓
Quarter FE	✓	✓	✓	✓	✓
Adjusted R^2	0.299	0.299	0.299	0.299	0.299
Observations	16,590	16,590	16,590	16,590	16,590

Standard errors in parentheses, clustered at market level. *p<0.1; **p<0.05; ***p<0.01

Table 1.12: Same event study as in main text (Table 1.10), but as control group, use only markets in which a major acquisition of *any* type has occurred.

	Dependent variable:				
	Number of entrants in market m, quarter t (Sample mean: 0.65)				
	Strategic acquirer			Financial acquirer	
	(1)	(2)	(3)	(4)	(5)
Major acq by enterprise software company (89 acquisitions)	−0.108* (0.062)				
Major acq by public enterprise software company (59 acquisitions)		−0.159** (0.072)			
Major acq by GAFAM or 'New Tech' (21 acquisitions)			−0.383*** (0.138)		
Major acq by company not in enterpr softw (incl. financial) (40 acquisitions)				−0.078 (0.074)	
Major acq by financial company (13 acquisitions)					0.110 (0.129)
Market FE	✓	✓	✓	✓	✓
Quarter FE	✓	✓	✓	✓	✓
Adjusted R^2	0.252	0.252	0.253	0.251	0.251
Observations	3,420	3,420	3,420	3,420	3,420

Standard errors in parentheses, clustered at market level. *p<0.1; **p<0.05; ***p<0.01

and even surprising: as explained in Section 3.5, acquisitions by strategic acquirers seem to often be part of their innovative strategy. At least for some of the acquisitions observed in the data, the motive may be to acquire innovative capabilities in the form of strategic assets or human capital. One may thus have expected strategic acquirers to acquire in markets that experience a rise in demand, and thus an increase in entry. This goes against my findings in Tables 1.10 and 1.11, which both show that strategic acquisitions are not preceded by more entry, and even tend to be succeeded by a fall in entry.

Instead of conducting an event study, Table 1.25 in Appendix 1.G repeats the regression, this time using the cumulative sum of acquisitions of a certain type in a certain market as an explanatory variable. Again, entry is higher in markets that have been subject to many acquisitions by acquirers active in enterprise software; however, we also observe a significant decline in entry subsequently to financial or other-industry acquisitions.[42]

A concern may be that these results could be driven by the year of 2020 which was affected by the beginning of the Covid-19 epidemic, or by a trend. The results of GAFAM and New Tech acquisitions hold when studying the time period of 2014-2019, which is the time period under study in the model. In contrast, the coefficient for the broader groups of strategic acquirers become insignificant. The results moreover roughly hold for a longer event window of 5 quarters, but fade for a shorter event window of 3 quarters. I also consider an event study where I consider all acquisitions with a transaction price of above 50US$ million (as opposed to 100US$ million), with similar results. A Poisson instead of a linear

[42] I also tried employing the estimator suggested by Callaway and Sant'Anna (2021), which correctly accounts for the staggered nature of the events and which does not require treatment effects to be constant. Due to the fact that events are "too staggered" and too rare when using the market-quarter panel, I end up with too few observations per "group" – i.e. per treatment period – to allow for reliable estimates. When collapsing the data into a market-year panel and using the estimator suggested by Callaway and Sant'Anna (2021), I obtain a negative but imprecisely estimated coefficient on strategic acquisitions.

model obtains similar results as in the baseline. See Appendix 1.G for these robustness checks and a further placebo test. One issue in all versions seems to be that acquisitions conducted by companies in other or financial industries tends to be negative as well, albeit not significant, despite using quarter fixed effects that should control for time trends. This is especially prelavent when regressing on the cumulative sum of acquisitions, or using a standard difference-in-differences analysis.

Overall, these reduced-form results offer suggestive support for entry-deterring effects of major strategic acquisitions, subject to the caveat of endogeneity. They contribute to recent literature that has found mixed results on the presence of a "kill zone": whereas Affeldt and Kesler (2021b), Kamepalli et al. (2021), and Koski et al. (2020) are aligned with my findings, G. Z. Jin et al. (2022) and Bauer and Prado (2021) find an increase in VC funding after an acquisiiton by a large technology company took place.[43]

Note that any reduced-form approach will shed light on a mix of the short-run effect of an acquisition that is transmitted through market structure, and the more long-run entry-for-buyout effect. Studying both types of effects is only possible within a dynamic structural model of startup entry, which is the subject of Section 3.

5 Dynamic Model of Entry

In order to study and quantify the entry-for-buyout, as well as the market structure effect, I build a dynamic model that I can take to the data. The economic agents in this model are potential entrants deciding whether or not to enter into a given market. I model these entry decisions as a dynamic discrete game with imperfect information that captures the competitive effects of other firms' entry decisions. The framework leans on models of dynamic discrete choice (Aguirregabiria & Mira, 2007, Bajari et al., 2007), but is markedly different to the extent that agents only get one single chance to make their decision of entering, or staying out of the market, instead of taking a new decision every period.[44] Acquisitions and IPOs are assumed exogenous in this model conditionally on twenty market-category effects that control for unobserved variables in the 440 markets.[45]

5.1 Setup

Time is discrete and infinite, and each decision period is a quarter. We consider a finite number of independent markets. In every period and in every market, there is a new set entrepreneurs with ideas for a new product in that market. These entrepreneurs form an exogenously given, fixed set of potential entrants in every period and every market.[46] In each period, all potential entrants simultaneously decide

[43]Note that all mentioned papers study the effects of acquisitions conducted by the GAFAM firms only. With the exception of Affeldt and Kesler (2021b) who take a focused approach and study 50 acquisitions conducted by the GAFAM in the mobile app market, those papers focus on many industries and employs alternative, firm-level (and thus possibly less precise) market definitions.

[44]This is natural in this setting. The model is nevertheless dynamic: agents are forward-looking, they incur a sunk costs, and their actions affect state variables that change every period.

[45]See Section 7 for a discussion of this assumption.

[46]Other models of firm entry have fixed these potential entrants in a similar way, e.g., Perez-Saiz (2015) or Igami (2017). I run robustness checks with respect to this assumption.

whether to enter the market or not, so as to maximize their expected profits. The potential entrants are homogeneous, except for private i.i.d. shocks that each agent draws from a distribution.

If a potential entrant decides not to enter the market, there will be no future chance of entry, and she stays out forever.

If a potential entrant decides to enter, she will earn flow profits in each period. These flow profits depend on a vector of state variables that are common knowledge, $\boldsymbol{x}_{mt}$. The state variables capture in a stylized way aspects of market structure that are likely to influence firm profits.

In every period following the entry decision, companies may be able to "exit" (i.e., experience a transfer in ownership) by being target in an acquisition, or by listing on the public stock market. These exit events allow the entrepreneurs to cash out: once acquired or listed on the stock market, a firm stops earning flow profits, and instead earns a single lump-sum return. I model acquisitions and IPOs as stochastic shocks that arrive upon active startups. If no acquisition or IPO opportunity arrives in a given period, the firm continues earning flow profits, and transitions into the next period.

The timing within each period is as follows:

1. All potential entrants observe the vector of state variables $\boldsymbol{x}_{mt}$ that is common knowledge, and privately observe a cost shock $\boldsymbol{\epsilon}_{imt} = \{\epsilon^0_{imt}, \epsilon^1_{imt}\}$.
2. All potential entrants simultaneously decide: {enter, stay out}, so as to maximize their expected profits.
3. All companies on the market earn payoffs:
   - Firms that are acquired in this period earn R^{acq};
   - firms that are going public in this period earn R^{ipo};
   - all other firms, including the new entrants, earn flow profits that depend on the new vector of state variables and a vector of parameters, $\pi(\boldsymbol{x}_{mt+1}; \boldsymbol{\gamma})$.

The equilibrium concept of the game is a Markov perfect equilibrium in pure strategies (Ericson & Pakes, 1995). A condition of this equilibrium concept is that players' strategies are functions of only payoff-relevant state variables.

Without loss of generality, the value of staying out is normalized to zero plus the random shock. The shock can be viewed as components of a sunk costs associated with a given action. Let θ denote the set of all structural parameters. The choice-specific value functions for entering and for staying out, excluding the random cost shock, write:

$$U^0(\boldsymbol{x}_{mt}; \theta) = 0 \tag{1.2}$$

$$U^1(\boldsymbol{x}_{mt}; \theta) = \mathbb{E}\big[\pi(\boldsymbol{x}_{mt+1}; \boldsymbol{\gamma}) - \kappa + \beta V(\boldsymbol{x}_{mt+1}; \theta, \cdot) \mid \boldsymbol{x}_{mt}\big] \tag{1.3}$$

so that potential entrant i's decision problem is given by:

$$\max \quad \Big\{U^0(\boldsymbol{x}_{mt}; \theta) + \epsilon^0_{imt}, \quad U^1(\boldsymbol{x}_{mt}; \theta) + \epsilon^1_{imt}\Big\} \tag{1.4}$$

$\pi(\boldsymbol{x}_{mt};\boldsymbol{\gamma})$ denote the flow profits that the firm obtains in each period, which depend on the state variables, $\boldsymbol{x}_{mt}$, and a vector of parameters affecting these flow profits, $\boldsymbol{\gamma}$. κ is a parameter denoting the sunk cost of entry, which the potential entrant incurs only once upon entering. $\beta \in (0,1)$ is the discount factor. The expected payoffs in future periods can be expressed as follows:

$$\begin{aligned} V(\boldsymbol{x}_{mt};\theta,\cdot) \quad = \quad & \alpha^{ipo}\big(p_m^{ipo}\cdot R^{ipo}\big) + \alpha^{acq}\big(p_m^{acq}\cdot R^{acq}\big) \\ & + (1-p_m^{ipo}-p_m^{acq})\mathbb{E}\big[\pi(\boldsymbol{x}_{mt+1};\boldsymbol{\gamma}) + \beta V(\boldsymbol{x}_{mt+1};\theta,\cdot) \mid \boldsymbol{x}_{mt}\big] \end{aligned} \tag{1.5}$$

As stated above, in every period following the entry decision, a firm may receive an opportunity to "exit" in the form of an acquisition or an IPO at probabilities p_m^{acq} and p_m^{ipo}. Such an exit yields returns (either acquisition price, or firm value) R^{acq} or R^{ipo}, respectively. In the model's current version, R^{acq} and R^{ipo} enter as data moments into the model. α^{acq} and α^{ipo} are parameters that essentially measure the extent to which startups' profits are influenced by exit opportunities in their given market.[47] If the firm is not acquired nor listed on the stock market, which is the case at probability $(1-p_m^{acq}-p_m^{ipo})$, then the firm continues to earn flow profits in that period. p_m^{acq} and p_m^{ipo} are data moments, in particular, the observed frequency at which startups are acquired, or go public, in market m (more on this in Section 5.2). In the next period, any of the same set of events – {acquisition; IPO; continue} – may occur, and so on. The vector of structural parameters is given by $\theta = (\boldsymbol{\gamma}, \alpha^{acq}, \alpha^{ipo}, \kappa)$.

I assume that $(\epsilon^0_{imt}, \epsilon^1_{imt})$ are independently and identically distributed according to a type-1 extreme value distribution. These shocks are privately observed by firms, but unobserved by the econometrician.

I do not observe profits, nor demand, for the tens of thousands of firms observed in my dataset.[48] Therefore, I employ a semi-structural approach: I treat profits as a latent variable, as does previous literature that models firms' discrete choices (e.g., Bresnahan and Reiss (1991), Collard-Wexler (2013), Seim (2006)). This approach makes use of the fact that a firm's presence on a market indicates that it must have been profitable for the firm to enter, by revealed preference. Unobserved profits are modelled as depending on state variables that, according to economic theory, should influence profits. By relating firms' entry decisions to these state variables through the lens of the model, one can estimate the parameters "measuring" the extent to which these state variables affect the profitability of a given market in a given time period.

5.2 Parameterization and Laws of Motion

Per-period flow profits, $\pi(\boldsymbol{x_{mt}};\boldsymbol{\gamma})$, depend on a vector of common knowledge state variables that are relevant to firms' profits. They are parameterized as follows:

$$\pi(\boldsymbol{x_{mt}};\boldsymbol{\gamma}) = \gamma^N \log(N_{mt}) + \gamma^A A^{\text{strat}}_{mt} + \gamma^M_{l(m)} \tag{1.6}$$

[47] Note that the acquisition price or valuation of an IPO is not necessarily valued at face value: anecdotally, entrepreneurs may hold a small share of just 5-20% of the company at the time of exit (for instance, see https://blossomstreetventures.medium.com/saas-founder-and-vc-ownership-data-a2a7e940bbcb, accessed 02/12/2022).

[48] I do observe the number reviews which may be indicative of demand. However, I observe these only as a single cross-section, and only for products not discontinued before 2021.

N_{mt} denotes the number of competitors in market m at time t. It is thus an endogenous state variable that evolves according to firms' entry decisions, as well as to an exogenous component[49]. A^{strat}_{mt} denotes the cumulative number of major competing startups that have been acquired (and kept alive) by a strategic acquirer, and evolves exogenously. This state variable is assumed to be exogenous and captures in a heuristic way that, if a major startup competitor is acquired by a strategic acquirer, this affects competition in market m, and thus expected profits.[50] $\gamma^M_{l(m)}$ is the intercept coefficient, which differs by market type $l(m)$, with $l = 1, \ldots, L$. It captures *market-category* effects that are constant over time and only vary at the market level, and is thus an exogenous state variable. These can be interpreted as measuring baseline profits that can be earned of a given market, and are required to control for a market's unobserved size or profitability, following Y. Wang (2022) (see Section 6.1 for how I construct these). The state variables are summarized by $\boldsymbol{x}_{mt} = \{N_{mt}, A^{\text{strat}}_{mt}, l(m)\}$.

I use the logged number of competitors as affecting flow profits in order to capture that, empirically, going from one to two competitors affects firm profits more strongly than going from, say, ten to eleven competitors (see, for instance, Mazzeo (2002)). One can expect γ^N to be negative, capturing that baseline profits are declining in the number of competitors N_{mt}. γ^A can be expected to be negative as well (based on the reduced-form results in Section 4), i.e., the number of major strategic acquisitions of competitors A^{strat}_{mt} lowers returns to entry.

In the light of the research question, the key parameters of interest are γ^A and α^{acq}. γ^A measures the extent to which a major strategic acquisition may depress entry. In contrast, α^{acq} measures the extent to which companies have an incentive to enter a market because they face the prospect of being acquired themselves in the future.

I define competitors in a market m at time t, N_{mt}, as consisting of products with at least one review produced by the following firms: VC-funded startups; public companies; acquired startups whose products have been continued; "pre-event" firms that have been founded within the last three years; and non-acquired private firms.[51] The law of motion of N_{mt} writes as follows:

$$N_{mt} = N_{mt-1} + num\_entrants_{mt} - D^{exog}_{\text{exit}} + D^{exog}_{\text{entry}} \tag{1.7}$$

$num\_entrants_{mt}$ denotes the endogenous number of entrants that enter in period t. In contrast, D^{exog}_{exit} and D^{exog}_{entry} are *exogenous* variables that are included to match the data, as companies may leave or be added to N_{mt} in ways not modelled.[52] I model these as random variables that follow a Bernoulli distri-

[49] The exogenous component is required to rationalize the data; see Section 5.2.

[50] Previous research has modelled firms as heterogeneous agents, which enables to capture the effects of acquisitions on competition and entry incentives in more explicit ways. For instance, in Perez-Saiz (2015), the acquired firm obtains the acquiring firms' characteristics, which affects competition. Similarly, in Igami and Uetake (2020), a merger between firms affects competing firms' productivity profiles. As I do not model firm productivity or firm characteristics, I use this stylized variable to capture that the acquisition affects comeptition in the market. In an alternative specification covered in Appendix 1.H, I consider instead a state variable denoted $A^{\text{strat in } t-K}_{mt}$ that mirrors the event study indicator variables employed in Section 4, and is equal to 1 in the event of a strategic acquisition in the past K quarters, and 0 otherwise.

[51] I thus exclude companies whose products in a given market do not have any review. I moreover exclude acquired non-VC-funded private companies, as well as private companies that have been coded as "inactive" based on them not having recorded any "event" on *Crunchbase* for 5 years. This choice is supported by the better fit in the first stage, indicating that products without any reviews may be viewed as a competitive fringe. Adjusting this definition of competitors does not qualitatively affect final results.

[52] For instance, a firm may be acquired and shut down (which leads to a reduction in the number of competitors by 1). Alternatively, a firm that is not VC-funded may enter (which leads to an increase in the number of competitors by 1).

Parameter		Value
Discount factor	β	0.9
Number of potential entrants	N^{pe}	6
Number of market types	L	20
Return from IPO	R^{ipo}	768
Return from acquisition	R^{acq}	130

Table 1.13: Parameters that are calibrated or constructed using data moments.

bution with parameters p^{exog}_{exit} and p^{exog}_{entry}, respectively. I estimate these parameters in a first step using a frequency estimator.

As in the event studies, I estimate versions of the model using a broader, and a more narrow definition of strategic acquirers. The broad definition encompasses all enterprise software acquirers, whereas the narrow definition accounts for a subset of enterprise software acquirers, namely New Tech and GAFAM acquirers.

Whereas only major strategic acquisitions can affect A^{strat}_{mt}, both strategic as well as financial acquisitions can affect p^{acq}_m. Indeed, any startup acquisition typically yields revenues to the target firm's owners. Therefore, both strategic as well as outsider and financial acquisitions may generally be perceived as a successful exit, allowing entrepreneurs and investors to cash out.[53] I thus take p^{acq}_m and p^{ipo}_m as being the rates of acquisitions and IPOs of VC-funded startups that we observe in the data in each market from 2010 to 2020. Therefore, the entry-for-buyout parameter is identified by variation between markets in the long-run percentage of startups acquired (p^{acq}_m), and observed entry into a given market. The market structure parameter is identified by variation between and within markets in acquisitions conducted by strategic acquirers, and observed entry. I discuss potential endogeneity concerns in Section 7.2.

R^{acq} is the median acquisition price for acquisitions of startups (130US$ million in the data), and R^{ipo} the median valuation of startups going public (768US$ million), between 2010 and 2020. [54] I fix the set of potential entrants in each period, N^{pe}, to the maximum number of entrants ever observed in a given market-quarter, which is equal to six.[55] I fix the number of market categories, L, to 20, motivated by first-stage results (see Section 6.1). As the discount factor is not identified, I set it to $\beta = 0.9$ (see, e.g., Igami and Uetake (2020), who calibrate the discount factor to the same magnitude, also employing quarterly data). Table 1.13 details all calibrated parameters.

5.3 Estimation

The primitives of the model are the structural parameters, $\theta = (\gamma^N, \gamma^A, \{\gamma^M_{l(m)}\}^{20}_{l=2}, \alpha^{acq}, \alpha^{ipo}, \kappa)$. I employ a two-step estimation method (e.g. Aguirregabiria and Mira (2007), Bajari et al. (2007)), which is essentially an extension of Hotz and Miller (1993)'s conditional choice probability estimator. It circum-

[53]This is the case in particular for buyouts by private equity firms. Anecdotally, see Chopra (2018)'s article in the online news outlet TechCrunch: "In years past, stigma often accompanied private equity sales [...] Today, private equity buyout firms can provide a solid (and on occasion excellent) exit route —- as well as an increasingly common one".

[54]I have explored the idea of making R^{acq} and R^{ipo} dependent on the state space, which is complicated by the fact that we observe very few instances of IPOs and acquisition prices. Estimating the model making R^{acq} dependent on broader bins of state variables did not affect final results significantly. I am continuing to explore this.

[55]The rationale for fixing the number of potential entrants to the maximum number of entrants ever observed in the data is laid out in Igami (2017).

vents the need to solve a dynamic discrete game in over 400 independent markets, which would make the estimation computationally infeasible. Instead, agents' equilibrium beliefs are obtained from the data. This approach deals with the problem of multiple equilibria. The underlying assumption is that the data have been generated by the same equilibrium, conditional on market observables.[56]

First stage

In a first step, I use data on agents' choices and state variables to estimate reduced-form regressions – policy functions (or conditional choice probabilities) – that map the state space into potential entrants' actions:

$$num\_entrants_{mt} = \phi_1 N_{mt} + \phi_2 A_{mt}^{\text{strat}} + \delta_m + \eta_{mt} \tag{1.8}$$

δ_m may either be market fixed effects, or broader, somewhat less flexible market-category fixed effects that account for unobserved market size or profitability (in this case, $\delta_{l(m)}$). Transition probabilities of the exogenously evolving (components of) state variables are estimated nonparametrically using a frequency estimator. Note that this first stage is essentially model-free. Policy functions characterize agents' actions given the state space, and transition probabilities describe how the state space evolves.

Note that we are not ultimately interested in the parameter estimates from the policy function in equation 1.8, but in the set of structural parameters, θ, estimated in the second step. Nevertheless, the parameters of the policy function give us an initial insight into the drivers of entry decisions, and in particular into the competitive effects. However, the main purpose of the estimated policy functions and transition probabilities is to forward-simulate the state space in a next step. For each state variable, one can simulate S paths sufficiently far into the future, until discounting renders the payoffs of any additional periods insignificant. Taking the average across these paths, and summing up each period's expected flow profits, yields the expected payoffs of a discrete action, given a set of parameter values.[57]

Second stage

The second step estimates the structural parameters by imposing optimality on all agents' choices observed in the data. Under the assumption that error terms are type-1 extreme value distributed, one obtains the following conditional choice probabilities for entering:

$$\Psi^1(\boldsymbol{x}_{mt};\theta) = \frac{\exp\left(U^1(\boldsymbol{x}_{mt};\theta)\right)}{\exp\left(U^0(\boldsymbol{x}_{mt};\theta)\right) + \exp\left(U^1(\boldsymbol{x}_{mt};\theta)\right)} \tag{1.9}$$

These conditional choice probabilities incorporate agents' beliefs about the future, and in particular about their opponents' behavior in a Markov perfect equilibrium (Aguirregabiria & Mira, 2010, Arcidiacono & Ellickson, 2011). Based on the conditional choice probabilities and agents' observed decisions in the data, one can set up the likelihood function, folloing Aguirregabiria and Mira (2007). Maximizing

[56] See Aguirregabiria and Mira (2010) for a survey on this matter.

[57] It may occur that the simulated number of competitors in a future time period reaches a value below 0, due to the exogenous entry and exit rates. I found that this is the case in far less than 0.1% of simulated observations, and if it occurs, then only far in the future (at which, due to discounting, it would hardly matter for firms' decisions). In case the forward-simulated number of competitors does hit 0, I set these equal to 0.5 to be able to take logs.

the likelihood function yields the estimates of the structural parameters that are the most likely to have generated the data.

6 Results

I use market-quarterly data to estimate the model. After excluding a few markets that I regard as outliers, I end up with 440 markets in the years of 2014-2019 (24 periods), yielding 10,560 observations.

6.1 First Stage: Startups' Entry Decisions

The results for the first stage can be found in Table 1.14, using a "broad" definition of strategic acquirers, and Table 1.15, which reports analogous estimates using a "narrow" definition. I begin with a linear model with no fixed effects in columns (1) of both Tables. I retrieve a positive coefficient on $\log(N_{mt})$, which would imply that more competitors attract *more* entrants. This counterintuitive sign when examining strategic interaction effects is a very common result in the empirical industrial organization literature (e.g. Collard-Wexler (2013), Igami and Yang (2016), Y. Wang (2022)), and stems from unobserved market-specific factors that are not controlled for. In this context, market size and profitability would both lead to more competitors present on the market being correlated with more entry. To control for these unobserved factors, I estimate the model using market fixed effects in column (2). Reassuringly, the coefficient on the number of competitors becomes negative. The coefficient on major enterprise software acquisitions is negative, although insignificant when using the broad definition in Table 1.14. As the dependent variable is a count variable, I also employ a Poisson specification in column (3), which yields negative significant coefficients, albeit at somewhat lower magnitude.

One potential concern with the linear model might be the incidental parameters problem. I therefore employ a less flexible version of market fixed effects, which the literature has called market-category effects (Collard-Wexler, 2013, Y. Wang, 2022). These types of fixed effects equivalently control for unobserved heterogeneity of markets. I follow Y. Wang (2022) and Lin (2015), and first estimate the model with market fixed effects in column (2). From the estimated market fixed effects, I construct $L = 20$ quantiles. I then associate each market into one of 20 bins, or groups, according to the quantile which its fixed effect estimate falls into. I re-estimate the model, this time using dummies based on these L groups, as opposed to a dummy based on the market (as would be the case for market fixed effects). Just like market fixed effects, the group-level dummies control for unobserved heterogeneity between markets that is persistent over time. Column (4) shows that this procedure yields similar results.[58] Finally, I employ market fixed effects along with quarter fixed effects in column (5) to control for seasonal

[58] I have estimated the regression employing fewer or more groups; it seems that using 20 groups is just sufficient. The more groups I use, the closer the estimates to the results in column (2), but the more likely one will face an issue regarding the incidental parameters problem, and the more one will possibly absorb too much of the variation in p_m^{ipo} and p_m^{acq} in the second stage. I investigate which types of markets have a high, and which have a low estimated market-category effect. I find that markets with the lowest estimated market-category effect (and thus likely low profitability and/or size) tend to be markets that appeal to narrow customer segments, e.g. markets tagged with the keywords "church / accounting / membership / donation", "club / membership / fitness / business", "catering / event / business / food", or "call / predictive / dialer / call-center". In contrast, markets with the highest estimated market-category effect seem to be active in broader, more growing markets, for instance in markets tagged with the keywords "artificial-intelligence / platform / customer / business", "app / development / application / building", as well as markets related to business intelligence, CRM, and marketing.

effects which are present in the data. I again recover similar results; seemingly, the negative strategic effect is not driven by any seasonal effect.

Table 1.14: First stage, using a broad definition of "strategic" acquirers. Standard errors are clustered at the market level. In column (4), they are computed using block bootstrapping with 5,000 bootstraps to account for the estimated market-category fixed effect.

	Dependent variable:				
	Number of entrants in market m, quarter t				
			Poisson		
	(1)	(2)	(3)	(4)	(5)
# of competitors	0.021***	−0.165***	−0.120***	−0.069***	−0.163***
	(0.001)	(0.015)	(0.017)	(0.012)	(0.015)
Cumulative # of major Enterprise Software acquisitions	0.026	0.094	0.137	0.161**	0.107
	(0.056)	(0.088)	(0.086)	(0.070)	(0.088)
$\mathbf{1}${quarter=2}					−0.126***
					(0.020)
$\mathbf{1}${quarter=3}					−0.152***
					(0.019)
$\mathbf{1}${quarter=4}					−0.214***
					(0.019)
Market FE		✓	✓		✓
20 market-category FE				✓	
Adjusted R^2	0.11	0.34		0.24	0.35
Log Likelihood			−9,809.514		
Akaike Inf. Crit.			20,503.030		
Observations	10,560	10,560	10,560	10,560	10,560
Standard errors clustered at market level.			*p<0.1; **p<0.05; ***p<0.01		

Using any of these policy functions, and using frequency estimates of the parameters p^{exog}_{exit} and p^{exog}_{entry} ($\hat{p}^{exog}_{\text{exit}} = 0.061$ and $\hat{p}^{exog}_{\text{entry}} = 0.0076$), I can use the law of motion in equation 1.7 to forward simulate the endogenous state variable N_{mt}. I employ the estimates of column (2), and draw 200 paths of 100 time periods.

The remaining state variables are exogenous. In order to forward-simulate the state variable A^{strat}_{mt}, I estimate the empirical frequency with which a strategic acquisition occurs. I then forward-simulate occurrences of major strategic acquisitions by drawing from a Bernoulli distribution each period, and construct the forward simulated flow of A^{strat}_{mt} so that it reflects the cumulative number of competing firms acquired by a strategic acquirer.

Finally, I use the $L = 20$ estimated market-category fixed effects as the only market characteristic ($\gamma^M_{l(m)}$), which stay constant over time.

6.2 Second Stage: Structural Parameters

The estimates for the structural parameters can be found in Table 1.16. Column (1) shows the results using a broad definition of strategic acquisitions by considering all major acquisitions conducted by a strategic acquirer, using in the first stage column (2) from Table 1.14. All parameters have the expected sign. In particular, the competitive effect is significantly negative, and the effect of a strategic acquisiton is negative, albeit not significant. The returns from being acquired or doing an IPO in the future are

Table 1.15: First stage, using a narrower definition of "strategic" acquirers. Standard errors are clustered at the market level. In column (4), they are computed using block bootstrapping with 5,000 bootstraps to account for the estimated market-category fixed effect.

	Dependent variable:				
	Number of entrants in market m, quarter t				
			Poisson		
	(1)	(2)	(3)	(4)	(5)
# of competitors	0.022***	−0.163***	−0.117***	−0.066***	−0.161***
	(0.001)	(0.015)	(0.017)	(0.012)	(0.015)
Cumulative # of major New Tech or GAFAM acquisitions	−0.118	−0.083	−0.026	−0.080	−0.068
	(0.135)	(0.219)	(0.245)	(0.148)	(0.220)
1{quarter=2}					−0.125***
					(0.020)
1{quarter=3}					−0.151***
					(0.019)
1{quarter=4}					−0.212***
					(0.019)
Market FE		✓	✓		✓
20 market-category FE				✓	
Adjusted R^2	0.11	0.34		0.24	0.35
Log Likelihood			−9,813.854		
Akaike Inf. Crit.			20,511.710		
Observations	10,560	10,560	10,560	10,560	10,560

Standard errors clustered at the market level. *p<0.1; **p<0.05; ***p<0.01

both positive and significant, indicating that a higher expected acquisition or IPO in the future makes entry more profitable. Moreover, the market category fixed effects, which are supposed to account for unobserved heterogeneity in profitability or market size, are successively becoming larger.

Column (2) employs a more narrow way to define strategic acquirers by using all major acquisitions by New Tech or GAFAM firms, and employing column (2) of Table 1.15 in the first stage. Again, parameters have the expected sign. The strategic acquistion effect now becomes marginally significant, albeit only at the 10% level.

Interpretation. The estimate for α^{acq} essentially measures an entrepreneur's valuation for being more likely to be acquired at a given price measured in millions of dollars. One can therefore express entrepreneurs' sunk costs of entry in terms of these expected dollars by dividing the estimate of the parameter κ by the estimate of the parameter α^{acq}. Using the results from column (2), I find that the sunk costs of entry parameter is approximately equal to 78 million US$. This is less than the lifetime amount of funding that successfully exiting, later-stage enterprise software startups obtain, according to *Crunchbase* data. Further, I find that the lifetime costs of having one additional competitor in the market are equal to 6.8 million US$. Moving up from the least to the most profitable market, in terms of the 20 market-category fixed effects, is worth 322 million US$, which emphasizes the importance of market fixed effects. Moving up from the 50th to the 55th quantile is worth 12.8 million US$.

It is noteworthy that the prospect of being acquired is not valued very highly compared to the other parameters, and that a lot of value depends on the market-category effects. As mentioned above, this could result from the fact that entrepreneurs in fact likely receive only a fraction of the acquisition price,

Table 1.16: Estimates of structural parameters.

	(1)	(2)
Entry costs, κ	-3.008^{***}	-2.978^{***}
	(0.139)	(0.138)
log(# of competitors), γ^N	-0.246^{***}	-0.251^{***}
	(0.010)	(0.011)
Cumsum of strategic acq of competitor **by Enterprise Software acquirer**, γ^A	-0.011	
	(0.015)	
Cumsum of strategic acq of competitor **by GAFAM or New Tech**, γ^A		-0.068^{*}
		(0.038)
Own IPO in future, α^{ipo}	0.005^{***}	0.006^{***}
	(0.001)	(0.001)
Own acquisition in future, α^{acq}	0.038^{***}	0.038^{***}
	(0.003)	(0.003)
Market category 2, γ_2^M (5th-10th perc)	0.321^{***}	0.325^{***}
	(0.024)	(0.024)
Market category 3, γ_3^M (10th-15th perc)	0.392^{***}	0.397^{***}
	(0.025)	(0.025)
...	...	...
Market category 19, γ_{19}^M (90th-95thth perc)	1.140^{***}	1.156^{***}
	(0.046)	(0.046)
Market category 20, γ_{20}^M (95th-100th perc)	1.298^{***}	1.316^{***}
	(0.050)	(0.050)
Log-likelihood	-10619.51	-10613.52
Observations	10,560	10,560

Note: *p<0.1; **p<0.05; ***p<0.01

or of the valuation when going public, respectively. If one was to account for this, the estimates for γ^A could likely rise 5- to 20-fold. Moreover, this finding could indicate that entrepreneurs may place a high value on competing in a market, and do not rely on being bought out. It could also possibly reflect the highly probabilistic nature of being acquired in a given market and risk-aversion on the part of entrepreneurs.

It is not clear what the main driving force is behind the estimated parameters. For instance, prior literature has established the presence of "IPO peer effects", which could explain the positive coefficient of α^{ipo} (Aghamolla & Thakor, 2021). Similarly, the positive α^{acq} be due to the entry-for-buyout effect, but might also be partly driven by a hearding effect (Conti, Guzman, & Rabi, 2021) or an acquisition probability effect (Song & Walkling, 2000).

6.3 Counterfactual Simulations

Procedure

One of the purposes of the model is to answer the question: how would entry evolve if acquisitions by certain types of acquirers were blocked by competition authorities? The ultimate impact depends on the respective magnitudes of the estimated parameters for the entry-for-buyout effect, $\hat{\alpha}^{acq}$, and the estimated market-structure effect of acquisitions, $\hat{\gamma}_A$. As explained below, I currently do not solve for the equilibrium that equates agents' actions with agents' beliefs for computational reasons.

I study two counterfactual changes in the prevailing antitrust regime. In the first scenario, the competition authority blocks only major strategic startup acquisitions.[59] In the second scenario, the competition authority blocks all startup acquisitions altogether. In each scenario, I assume that the policy change takes place in the first quarter of 2014, i.e. the first period of observation of my data.

To conduct the simulation, I take the starting values of the state variables to be their respective values in this first period. I simulate the entry decisions of N^{pe} potential entrants in this period. Based on the simulated entry behavior, I can calculate the state variables for the next period, and iterate until the end of the sample period. To elaborate, I carry out the following steps:

1. Take $x_{m,2014Q1}$ from the data.

2. Adjust the transition probabilities according to the counterfactual that one is interested in: for instance, for the counterfactual in which no acquisitions are possible, set the probability of a future buyout to 0. Based on this, forward-simulate the state variables, drawing 200 paths for 100 time periods into the future.

3. Using the estimated parameters from Table 1.16, column (2), and the forward-simulated state variables, compute the expected discounted value of entering.

4. For each potential entrant, draw i.i.d. cost shocks ϵ^0_{ijt}, ϵ^1_{ijt} from a type-1 extreme value distribution.

[59] This reflects a recommendation by, for instance, US Congress Committee on the Judiciary (2022), see their recommendation for "Restoring Competition in the Digital Economy" on p.14: "Presumptive prohibition against future mergers and acquisitions by the dominant platforms".

5. Given the value of entering and the drawn cost shocks, compute the number of actual entrants (i.e. the number of potential entrants for which the value of entering is higher than the value of staying out).

6. Compute and simulate what the counterfactual state variables will be in the next period.

7. Repeat steps 2 to 6 until the last period of observation.

For the forward-simulation in step 2, I use the original policy function and transition probabilities. I thereby assume that startups hold onto their original beliefs of how state variables will evolve over time. This simplification can be viewed as an initial impulse by the agents, and an approximation to a full counterfactual simulation. If one were to account for the fact that startups' beliefs regarding the state space evolution were to adjust, one would have to solve for a fixed point that equates startups' beliefs to observed actions in the counterfactual world. Given the large number of observed markets, this is computationally infeasible.[60]

How would entry evolve under counterfactual merger policy regimes? – Results

I begin by examining the effects on entry and on the number of competitors in the average market. Table 1.17 displays the effects of blocking only certain, or all, startup acquisitions on the number of entrants and number of competitors across markets and periods. I first simulate the counterfactual in which only strategic acquisitions are blocked. This results in a very slight increase in entry and in competition in the average market.

I then simulate the counterfactual in which all acquisitions are blocked. Given the current values of the parameter estimates, in the average market, firms *prefer* competing on the market forever, rather than being acquired. This leads to the finding exhibited in the second row of Table 1.17: entry rates and the number of competitors increase in the counterfactual. In reality, however, it may be unlikely that firms competed forever in a situation in which acquisitions are not possible at all. Instead, there might be a substantial risk of profits going to zero, as there would be no opportunities to find VC funding due to the lack in exit opportunities.

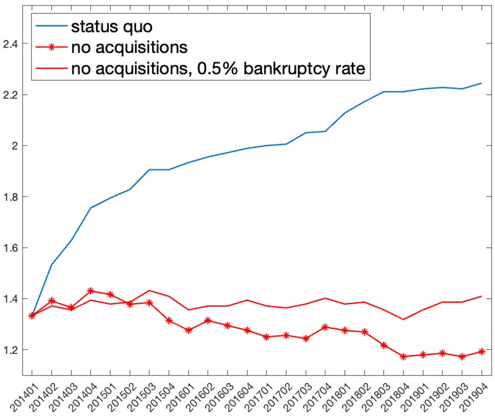
I therefore introduce a rate at which firms may obtain a negative shock that leads profits to go to 0 in the counterfactual with no acquisitions, akin to a bankruptcy rate. The results are displayed in rows 3 and 4. If firms have a 0.25% increased probability of having profits go down to 0 in every quarter in the counterfactual with no acquisitions will lead to a reduction in the number of entrants as well as in the number of competitors.[61] In currently ongoing work, I will verify to make sure these assumed rate of bankruptcies could be supported by scientific literature in empirical finance. I carry out fifteen simulations of each type, and take the average.[62]

As the data contain over 400 markets, I can explore how the effect of blocking startup acquisitions varies across markets of different types. In particular, by way of the market-category effects, the struc-

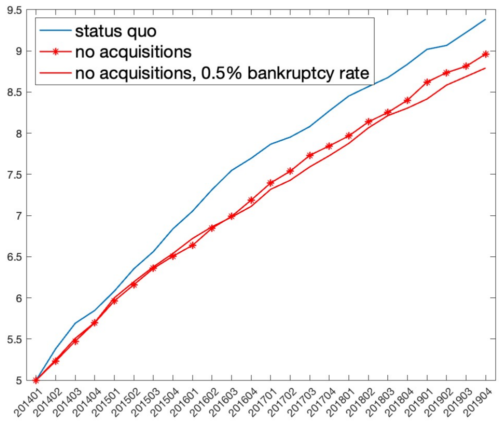
[60] In future iterations of the paper, I plan to either fully solve this dynamic problem in a small subset of markets. An alternative would be to consider an approximation based on Aguirregabiria and Ho (2012).

[61] Using *Crunchbase* data, I find that the actual quarterly bankruptcy rate for enterprise software startups is around 1.2%.

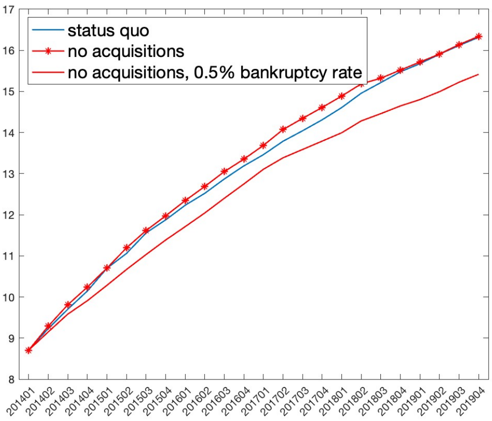
[62] I will increase this number in future iterations of the project. For the time being, the simulation exercise can be viewed as an approximation.



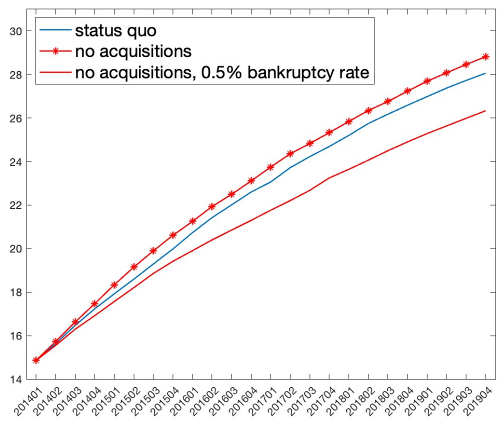
(a) Markets at 5th percentile of profitability



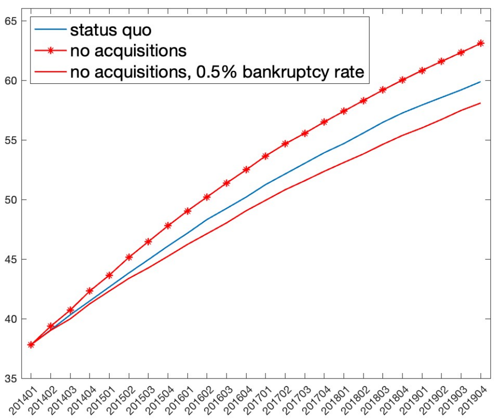
(b) Markets at 25th percentile of profitability



(c) Markets at 50th percentile of profitability



(d) Markets at 75th percentile of profitability



(e) Markets at 95th percentile of profitability

Figure 1.6: Heterogeneity in the effect of blocking startup acquisitions on the number of competitors, across markets of different unobserved profitability.

Counterfactual	Change in entry in numbers	in percent	Change in # of competitors in numbers	in percent
Blocking only New Tech & GAFAM acquisitions:				
· Effect on average market	0.002	0.44%	0.05	0.24%
· Effect on market affected by strategic acquisition	0.04	4.12%	0.56	1.36%
Blocking all acquisitions; startups earn profits forever in counterfactual:				
· Effect on average market	0.03	5.09%	0.36	1.85%
Blocking all acquisitions; 0.25% chance of profits going to 0 per quarter:				
· Effect on average market	-0.02	-4.18%	-0.35	-1.78%
Blocking all acquisitions; 0.5% chance of profits going to 0 per quarter:				
· Effect on average market	-0.05	-10.73%	-0.79	-4.03%

Table 1.17: Change in the mean number and percent of entrants, and in competitors, in counterfactual scenarios compared to the baseline.

tural model essentially groups markets according to their unobserved market size or inherent profitability. Figure 1.6 shows that effects do vary for markets of different profitability. For low-profitability markets – panels (a) and (b) – the number of firms decreases in the counterfactual, especially in later time periods. In contrast, entry tends to increase in markets with a very high inherent profitability, as in those markets, staying on the market – as opposed to being acquired – is very profitable. I intend to explore this heterogeneity and its plausibility further in future work.[63]

7 Discussion

7.1 Limitations regarding market definitions

The market definitions that I employ are more granular than standard industry classification systems used in previous literature, and thereby allow to make progress on our understanding of the effects of startup acquisitions in software markets. However, these new product-level market definitions are subject to some of the same caveats that more standard firm-level taxonomies suffer from. In software, startups at times change the focus of their products and pivot from one market into another one, which cannot be captured by static market definitions. The market definitions also cannot account for a possible interdependence between markets, which could arise from the bundling of products or the provision software ecosystems. Nor can the market definitions capture the distinction of markets for technology, as opposed to product markets (see Gans and Stern (2003)). Finally, consumer inertia and switching costs are thought to be important in digital markets (e.g., Scott Morton et al. (2019)), which may render

[63]My results are in fact plausible when compared to findings by Fons-Rosen et al. (2023), who find a decline in the startup rate of 14.9% if all startup acuqisitions were blocked in the entire US economy.

products within a market less substitutable than their product descriptions suggest.

These caveats are shared by all other market definitions that do not actually estimate substitution patterns from demand data. How to accurately define markets for software is a frontier research question itself.[64] The discussion highlights the need for future empirical advances in characterizing demand for software and competition between nascent software products. I am in the process of conducting a number of robustness checks, for instance by varying the number of markets created. In future work, I plan to work with proximity measures between different markets.

7.2 Endogeneity of acquisitions

I estimate the model parameters under the assumption that acquisitions are exogenous. Realistically, however, the decision to acquire another firm is driven by a multitude of considerations, including acquirers' expectations about the players in the market. First, I hope to be controlling for some market-level unobservables that might contribute to a higher frequency of startup acquisitions in one market versus another one with the help of the market-category effects. Second, by comparing acquisitions conducted by financial and strategic acquirers, the event study assumes that there is a random element in who acquires whom at what time in a given market, which is not implausible. This suggests that the estimated market structure effect is nevertheless meaningful.

To elaborate, I first turn to potential endogeneity concerns regarding the entry-for-buyout parameter. Each market observed is in a long-run equilibrium of startups entering the market, and startups being acquired in that market. The entry-for-buyout parameter is identified by between-market variation in the market-specific, long-run percentage of startups acquired in a given market (p_m^{acq}), and observed startup entry. One concern might be that both acquisitions and entry behavior are being driven by an unobserved variable, such as technological advances leading to a rise in demand and an increase in entry.[65] On the other hand, this may also have induced acquisitions to take place, as companies may have found it profitable to buy and integrate software producers. The market-category effects that I employ can control for this to some extent, as the estimation essentially only uses variation within the given profitability quantile a given market is in. In future iterations of this project, I hope to employ an instrumental variable strategy to provide causal evidence for the existence of this channel.[66]

The market structure parameter is identified by variation in the number of entrants around the time of a major acquisition by a strategic acquirer both between and within markets. As pointed out earlier, for same-industry acquirers, the motive of an acquisition is often to integrate a product or a product feature into the acquirers' existing product portfolio, or to enter a market. For these firms, the acquisition decision is therefore driven by higher expected profits when expanding into a certain complementary direction in the product space. Thus, the consideration of *which market* to acquire in is possibly endogenous to certain observed and unobserved market characteristics. However, there is a random element in

[64] See Aridor (2022), who estimates consumer substitution patterns across social media with the help of a field experiment.

[65] For instance, the ubiquitous collection of data requires all companies in the economy to collect and analyze consumer data in order to remain competitive. As a result, demand has risen for data analytics, data management, or dashboard software, for instance. This may on one hand have induced new startups to enter.

[66] Potential instruments are regulations that made acquisitions in a given market more burdensome. I found anecdotal evidence that in the European Union, the General Data Protection Regulation (GDPR) may have had such an effect.

which of the startups in a given market is ultimately purchased, *by whom*, and *at what time*. The match value between a target firm and an acquirer is affected by characteristics such as language, travel distance, whether the two firms happen to share the same technology stack, or sympathy, which are exogenous to new startups' entry decisions. Anecdotally, startups frequently turn down offers they obtain, seemingly for reasons exogenous to market or firm characteristics.[67] Overall, contrasting financial with strategic acquirers around the event of the merger announcement, akin to an event study – as done in the reduced form – may be as close as one can get to finding out about any potential entry-deterring effect of major strategic buyouts.

Endogenizing acquisitions, as is done in some prior research (e.g., Igami and Uetake (2020), Stahl (2011)), is not feasible in my setting due to computational and conceptual challenges.[68]

8 Conclusion

This paper studies the link between innovative entry and acquisitions, and thereby sheds light on a set of questions that is of an enormous importance for economic welfare. What drives the provision of new, innovative products in a market, and how does merger policy affect firms' incentives to do so? My data collection effort allows to make progress on this question in the context of startup acquisitions in the software industry. Merger policy in software markets is being fiercely debated in many jurisdictions, but our understanding of the motives as well as the implications of these mergers for competition and innovation is extremely limited.

I provide new data and descriptive evidence of the likely effects of the acquisitions of VC-funded startups in the enterprise software industry. I build and estimate a model of startup entry decisions that fleshes out, in a stylized way, an entry-for-buyout effect that fosters entry, and an effect via market structure that deters entry. I find that an overall ban of all startup acquisitions would decrease entry by 8-20% in markets that have a low baseline profitability. Nonetheless, acquisitions conducted by strategic acquirers appear to deter entry. If these acquisitions were banned, entry might be increased. These findings are highly relevant to the ongoing policy debates regarding startup acquisitions in technology sectors. More broadly, my results can contribute to the debate on the relationship between market structure and innovation, going back to Schumpeter (1942) and Arrow (1962).

The data I collected and the evidence I have found open up several avenues for future research. One important policy concern is not only that firms are *able* to enter, but also that firms are willing to enter and *remain independent* upon successful entry. Recent literature has provided evidence that startups wanting to be publicly listed might face barriers (Ederer & Pellegrino, 2023). Firms' decisions to agree to a buyout, as opposed to continued operation, is likely a function of startup age, funding,

[67] Snap, for instance, received an offer to be acquired by Google and Facebook, but eventually remained independent. An interview I conducted with a startup co-founder who shall remain unnamed revealed that their company received offers from three of the GAFAM firms, but eventually sold to another large software company.

[68] First, in model with endogenous acquisitions, there would be thousands of potential acquirers at any given time, as I study an entire industry with over 400 markets at once. Second, it is far from obvious how to write down a model that accurately describes acquiring firms' decision making in my setting: the motives that are driving acquisitions here seem to be very heterogeneous, and any attempt of writing down a stylized model would not capture those accurately enough. The setting that Igami and Uetake (2020) study is much "cleaner": products are homogeneous, and firms can be described by a single profitability parameter that is plausibly very influential for merger decisions in their context.

the number and types of alternative acquirers, the costs and risks associated with an IPO, and further determinants of startups' outside option. Future research could study what affects firms' willingness to remain independent in software markets, possibly with the help of a model endogenizing the decision to agree to a buyout.

Moreover, in future work I would like to exploit distance metrics between markets which one can obtain using the text-as-data methods. This would yield further evidence on the extent to which different types of acquirers buy startups in (dis)similar market niches, and inform about firms' innovation and expansion activities.

The paper's strength lies in generalizable results on an entire industry sector, comprising tens of thousands of companies. However, unless one is willing to make very strong assumptions, the lack of demand data precludes me from making any strong conclusions regarding welfare implications. In this respect, my findings invite a number of follow-up questions, such as: how much does new product entry contribute to welfare? What is the welfare consequence of the frequently observed discontinuation and integration of products? – I leave these questions for future research.

1.A Anecdotal evidence

1.A.1 Entry-for-buyout effect

Benedict Evans, VC investor & analyst, August 2022

In a blog post titled "The FTC versus tech M&A", Benedict Evans comments on the FTC's proposal to block Meta from acquiring Within, as well as on the FTC's approach towards M&As in Tech more generally:

> "M&A is a central part of the Silicon Valley ecosystem [...] How do you fund companies if both IPOs and M&A are off the table?"[69]

Bénédicte de Raphélis Soissan, founder of Clustree, July 2020

I transcribed the following quote from an interview by *Tech Off*, a podcast by *Les Echos Entrepreneurs*, with *Clustree* founder Bénédicte de Raphélis Soissan. She had previously successfully sold her startup *Clustree* to Cornerstone. In English (translated with help from deepl.com):

> "If I were to start a company again, it wouldn't be to see if I could raise the funds. It would rather be testing potential buyers with your idea, to see what the exits are. [...] From the start, the best way to test the market for me would be to go see potential competitors / buyers (even if it can be risky - so you have to see how you do it) to really test what the market is in terms of exit, what is [inaudible], is there interest for this type of product, this type of technology, etc."[70]

1.A.2 Kill zone effect

Jason Roberts, founder of Preezo, 2010

In "How I Screwed Up My Google Acquisition", the founder talks about his attempt at selling his startup *Preezo* to Google:

> "I heard nothing from Google until the following June when I read that they had acquired Zenter, a YCombinator startup working on the same problem. At that point my heart sank as it was obvious that the window of opportunity had closed and it wasn't a few months later that Google Presentations itself was released. While the Google version wasn't quite as powerful or polished as Preezo, being that it was free, solidly good enough and integrated into a complete productivity suite meant it was going to be very tough going for Preezo as a standalone product. To make matters

[69] See https://web.archive.org/web/20221027140238/https://www.ben-evans.com/benedictevans/2022/8/1/within-and-tech-mampa (accessed 07/04/2023).

[70] Original, in French: *"Si je devais remonter une boîte, ça serait pas voir si j'arrive à lever les fonds. Ça serait quasiment tester les acquéreurs potentiels avec ton idée, pour voir en fait quels sont les exits. [...] Dès le départ, la meilleur façon de tester le marché ça serait pour moi d'aller voir des compétiteurs / acquéreurs potentiels (même si ça peut être risqué - donc il faut voir comment tu le fais) pour vraiment tester quel est le marché en terme d'exit, quel est [inaudible], est-ce qu'il y a intérêt pour ce type de produit, ce type de techno, etc."* Source: see https://business.lesechos.fr/entrepreneurs/communaute/0603458127497-podcast-benedicte-de-raphelis-soissan-fondatrice-de-clustree-338661.php (accessed 07/04/2023).

> worse, Yahoo and Microsoft had continued to abstain from the web office race, shunting any hopes that acquisition offers might be soon forthcoming."[71]

1.B Entry-deterring effect of acquisitions

1.B.1 Summary of theoretical literature

I here elaborate somewhat more in depth on the theoretical papers that have established why an acquisition in technology markets can deter entry. The theory of bundling (e.g., Whinston (1990)) suggests that companies may leverage market power from one market into another, and thus foreclose rivals. Motivated by the acquisitions in digital markets, the model proposed by Denicolò and Polo (2021), a cumulative number of acquisitions can entrench a dominant position of an incumbent, leading to market power and less entry, even in the presence of an entry-for-buyout effect. Motta and Shelegia (2021) do not directly speak to a possibly entry-deterring effect of acquisitions, but study the interaction of incumbents' aggressive behavior and acquisitions. In their model, entrants expect aggressive behavior by the incumbent in the form of imitation, and therefore produce a complement instead of a substitute to the incumbent's product. Even in anticipation of possibly being bought out, entrants may stay away from directly challenging the incumbent. As acquisitions by large strategic buyers are often akin to market entry by the buyer, one might expect this "kill zone" effect in anticipation of imitation to take place upon an acquisition in a given market. Kamepalli et al. (2021) study a setting with network effects and consumer switching costs. In their model, consumers anticipate that startups' products will be acquired and integrated into the acquirer's product. To avoid switching costs, consumers are therefore reluctant to try out a new product, which leads to low adoption and low demand of the startups' products, and subsequently to a lack of willingness to fund new entrants.

1.C Supplementary information on data creation

1.C.1 Cleaning and construction of firm-event panel data using *Crunchbase*

Crunchbase comprises over a million public, private, as well as firms that existed in the past but have been closed. Companies may be located all over the world and may span all sectors of the economy, but people who have worked for the VC industry mentioned to me that *Crunchbase*'s coverage may be most accurate for firms located in North America and Europe. Information on *Crunchbase* are sourced using Machine Learning, an in-house data team, a venture program, and via crowdsourcing.

The *Crunchbase* data was obtained in a format that requires some handling of the data in order to make it useful for economic analyses. First, *Crunchbase* contains "organizations", which comprises companies, but also other institutions like schools; I therefore exclude the latter. I then create a "firm-event panel" in which each observation corresponds to a certain "event" that was happening in a given com-

[71] See `https://web.archive.org/web/20220728220203/http://www.codusoperandi.com/posts/how-i-screwed-up-my-google-acquisition` (accessed 07/04/2023).

pany's lifetime, as well as its characteristics. I obtain the following events from *Crunchbase*: *founded, getting funding, investment, being acquired, acquiring, IPO, inactive, closed*. In addition, I create the event "inactive" based on prior literature as the date five years after any kind of relevant event of a given private, non-acquired company.[72] From such a dataset, one can easily create quarterly data of, for instance, the number of acquisitions per quarter, or the number or volume of funding rounds.

I moreover create the parent-subsidiary structure for all firms. I consider parents up to two levels up of a given focal company, which is sufficient in all cases in my data.

1.C.2 Definitions of "startup" and "Venture Capital funding round"

Venture Capital funding round: Any funding round of the following type: *Angel, Pre-Seed, Seed, Series A to Series J, Unknown Series, Corporate Round, Convertible Note, Undisclosed*. I thus exclude, for instance, Post IPO funding rounds, Private Equity, or Secondary Market investment.
(Pre-exit) Startup: Any private company that has raised at least one Venture Capital funding round (i.e. prior to any recorded event of the type *acquisition, IPO, closed* or *inactive*).

I focus on startups, as startups have been found to be particularly innovative and disruptive. Startup acquisitions account for approximately 44% of all acquisitions observed in the matched data. This fact is reflected in my data showing that products supplied by VC-funded startups have more reviews, even when employing a range of controls for company characteristics and age (see Appendix 1.F, Table 1.22).

As pointed out in the text, *Crunchbase* defines acquisitions as majority takeovers, which may mean majority investments. This is very reasonable, as a majority investment allows startup founders and early investors to cash out, and transfers ownership and control into new hands[73].

1.C.3 Web-scraping *Capterra*

I first web-scrape the list of categories available on *Capterra* (see Figure 1.1). For each category, I then query the listings page, which I fully expand to obtain a list of all the products that are associated with that given category. For each product in that list, I download the hyperlink that directs to the specific product page (see Figure 1.2). I end up with 72,986 unique links to product pages on *Capterra*, which I query one-by-one in June and July of 2021.

In that process, I find that in some instances, a single product can have multiple URLs (and thus product pages) on *Capterra*. I therefore define unique products based on product name and the first sentences of the descriptive text. For each product, I collect all the categories it can be active in. I finally obtain approximately 70,000 unique product-level observations.

[72] I have found prior literature that codes companies that did not receive venture capital within 3, 5, or 7 years as inactive.

[73] See *TechCrunch* reporting on Vista Equity Partner's majority investment of Pipedrive: "[...] as is the case with these type of private equity buyouts, many of Pipedrive's early shareholders will have exited or partially exited, including employees/management and early backers. This is either voluntary or mandatory as part of a shareholder agreement "drag-along" clause." See `/web/20221105105842/https://techcrunch.com/2020/11/12/european-unicorns-are-no-longer-a-pipe-dream/`, accessed 05/11/2022. Another example is from the press statement from Francisco Partners regarding their majority investment of LiveU: "Francisco Partners, a global technology-focused private equity firm, together with co-investor IGP Capital, have acquired LiveU from its existing shareholders to accelerate further the company's global expansion.", see `/web/20221105112118/https://www.franciscopartners.com/news/liveu-announces-majority-investment-from-francisco-partners-to-accelerate-growth`, accessed 05/11/2022.

Product Name on Capterra	Name of Producing Company (from Capterra)	Matched to Crunchbase Company	Matched how?
A			
Jira	Atlassian	Atlassian	URL & company name
Adobe Acrobat Reader DC	Adobe	Adobe	URL & company name
ClickUp	ClickUp	ClickUp	URL & company name
Box	Box	Box	URL & company name
Safari	Apple	Apple	URL & company name
B			
AWS Cloud9	Amazon Web Services	Cloud9 IDE	Amazon acquired the company Cloud9 IDE
Widevine DRM	Google	Widevine	Google acquired the company Widevine
Yammer	Microsoft	Yammer	Microsoft acquired the company Yammer

Figure 1.7: Example of how existing products on *Capterra* were matched to firms on *Crunchbase*. Products in panel A were matched by company URL and name. Products in panel B were matched to the target that was acquired by producing firm in the past based on name similarity.

1.C.4 Merging *Capterra* products to *Crunchbase* companies

I first use company URL and name to match products on *Capterra* to their producing firms on *Crunchbase*.[74] Panel A in Figure 1.7 gives a few examples of products matched to companies by URL and name.

However, in cases where the product originated with a startup, but is now provided by the acquirer, the above matching algorithm will associate the product to its acquirer and current owner, not to its *originating* company. To trace products back to the startups that may have been the originators of a given product that was then acquired, I make use of the fact that young startups typically provide a single product whose name is the same as the company's name. Therefore, whenever a given product's producing firm (as indicated on *Capterra*) has previously acquired a company that shares any similarity with a given *product*'s name, I assume that it is the *acquired* firm that initially entered the market with this product; see panel B in Figure 1.7.

1.C.5 Checking for possible sample selection issues

As noted above, the product-level data obtained by *Capterra* is cross-sectional and covers enterprise software products available in June and July of 2021. One might be concerned that *Capterra* suffers from survival bias, and thus does not accurately capture all relevant entrants and competitors in the enterprise software industry in 2012-2020. However, I note that survival bias is likely not problematic, and possibly even wanted, as it allows to disregard likely irrelevant competitors. As the focus of this

[74] I first extract all firm URLs that are unique in both *Crunchbase* and *Capterra*, and match those products to firms based solely by URL. For the remaining firms with non-unique URLs on either *Crunchbase* or *Capterra*, I then employ a fuzzy matching algorithm to match the remaining firms: both their URLs must be equal, and additionally, firm names must at least share some similarity. Finally, somewhat less than 1% of all products are matched manually by looking up the company.

Table 1.18: Investigating the companies with enterprise software related tags and keywords on *Capterra* that are not part of my sample. These percentages are approximate and are recovered from both systematic investigation, as well as an additional manual investigation of a random sample of 20 firms out of all firms that could not systematically be associated to one of the below reasons.

likely reason why firm is not included in sample	% of companies
closed or inactive before 2021	31.5
many missing values in important variables (e.g. industry or country)	18.6
acquired (by non-enterprise software) and discontinued	14.8
likely not in enterprise software	14.1
founded very recently	10.6
located in China, Japan, or Korea (systematically under-represented on *Capterra*)	7.8
arguably should have been part of sample	2.6

paper is on firm entry (as opposed to closures), and as the data collection took place soon after the end of the sample period, I am likely capturing all actually relevant and actually viable entrants and competitors.

To investigate potential selection issues further, I compare the sample of firms used with the set of companies on *Crunchbase* that contain enterprise software related tags and keywords. I find that the latter is twice as large as the sample used, but at the same time, would capture only 60% of the currently included sample. I therefore conduct manual and systematic analyses of these likely enterprise software related companies that are not part of my sample, in order to get a sense of whether those firms should have been included. As shown in Table 1.18, I find that approximately 32% of these companies have likely been shut down as of 2021, and another 19% have missing data in usually well-covered variables and are thus likely not major or very active either. 14% of companies seem to actually be active in other industries (such as consulting, venture capital, business development, or actually provide add-ons to existing products), albeit being tagged with enterprise software related terms. I find that for only between two and three percent of these companies, one may argue that they should have been included into my sample. These companies are missing from *Capterra* for unknown reasons. However, this selection will likely be random across markets, and will likely not affect major competitors.

To conclude, I find that indeed, those companies that contain enterprise software related tags and keywords, but are not contained in my sample, should for the most part not have been included into my sample. Instead, the set of companies that contain enterprise software related tags and keywords will provide a *less* accurate definition of the industry: as written above, less than two thirds of the used sample would be considered as enterprise software when using *Crunchbase*'s tags and company description. This concerns, for instance, multi-product firms like Facebook or Apple, which are providing enterprise software, but are not tagged with related keywords, and would thus not be included into this alternative sample.

A final note regarding the geographic reach of my sample may be of interest. As of 2021, *Capterra* was available in many Western European languages as well as in Japanese. Accordingly, I find European and North American companies to be somewhat over-represented on *Capterra*, and companies from East

Asian countries and Russia to be under-represented.[75] I do not believe this to be problematic, as the products developed by companies in those countries might indeed not be available in English or other Western European languages, and might thus not be easily substitutable with the products covered by *Capterra*.

1.C.6 Building a dictionary and tagging products with keywords

Each product on *Capterra* can be associated to *more* than one category.[76] This precludes me from using the *Capterra* categories directly as market definitions. In order to place products into *unique* and *disjoint* markets, I essentially need to reduce the dimensionality of the categories. Aside the category names, I "tag" products with further meaningful keywords whenever those appear in the products' descriptive text.

I first clean some category names. I replace some acronyms in the category name (e.g. "Search Engine Optimization" instead of "SEO"), and I create bi-grams (e.g. by replacing "photo editing" by "photo-editing"). Moreover, I add a small number of further meaningful terms to that dictionary. I then "tag" each product with the respective keywords whenever they occur either in the category name, or in its descriptive text. Acquired companies whose products were shut down (and for which *Capterra* categories or product description are thus not available) are tagged with the respective keywords from the same dictionary whenever they occur in these companies' *Crunchbase* industry tag or descriptive text. For instance, if a given company on *Crunchbase* is described as providing spreadsheet software, this company's product will be associated with the term "spreadsheet".

1.C.7 K-means clustering

I employ a k-means clustering algorithm to partition products into disjoint sets. For a given number of clusters k, k-means clustering divides observations into groups in a way that minimizes the within-cluster variation summed over all k clusters. Within-cluster variation is defined to be the squared Euclidean distance.

Further clustering algorithms exist, but so far have resulted in less intuitive outcomes in my setting. In particular, using HDBSCAN yielded clusters that are less aligned with *Capterra*'s initial product categories, which might be a meaningful benchmark. Evaluating outcomes of a certain clustering outcome does in fact not seem to be straightforward. See Grimmer and King (2011) for how to evaluate the outcome of a clustering algorithm, and see Delgado, Porter, and Stern (2016), who evaluate different methods for detecting regional industry clusters. I intend to explore this issue more thoroughly in future work.

[75] It is known that *Crunchbase* is already mostly covering European and North American companies better that, but my analysis shows that *Capterra* is even more centered on North America and European companies. Even though *Capterra* is available in Japanese, Japanese firms are not very represented on *Capterra*.

[76] The average number of categories per product, for instance, is 1.9, the median is one. 29 products are associated to over 30 categories.

1.C.8 Validation of Market Definition

Using market definitions from recent merger decisions in the domain of enterprise software

In Figure 1.8, I conduct a validation of the market definitions by comparing my markets to markets distinguished by the UK Competition and Markets Authority in their decisions with respect to the Salesforce-Tableau merger, and the Google-Looker merger (see here: `bit.ly/3XhIE2T` and here: `bit.ly/3iemS0N`, both accessed 15/03/2022). I find that, when grouping products into 500 markets, twelve out of the 15 products (80%) are categorized as substitutes and thus into the same market. When grouping products into 400 markets, ten products are classified as substitutes.

Source: CMA reports: Salesforce-Tableau & Google-Looker

Table 1: Shares of supply in Modern BI by revenues worldwide in 2018[78]

Company	Share
Salesforce.com	[0-5]%
Tableau Software	[20-30]%
Parties combined	***[20-30]%***
Microsoft	[5-10]%
SAP	[5-10]%
Oracle	[5-10]%
IBM	[10-20]%
Qlik	[10-20]%
SAS	[0-5]%
MicroStrategy	[0-5]%
Information Builders	[0-5]%
TIBCO	[0-5]%
DOMO	[0-5]%
Looker	[0-5]%
Aggregate of all Other Companies	[10-20]%
Grand Total	**100.0%**

Sisense
Periscope

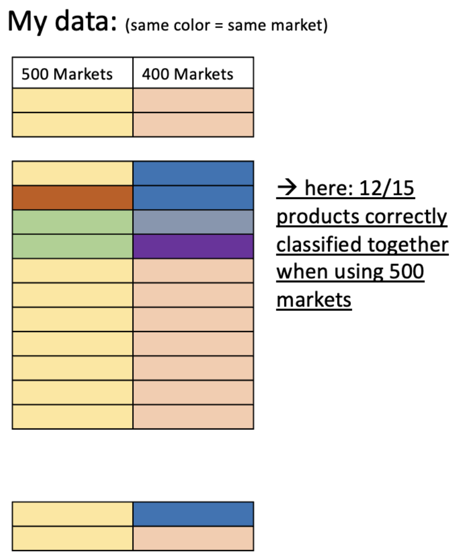


Figure 1.8: Validation of market definition.

Using known firms

Below is a list of products that are clustered together in the same market, as well as the three most commonly occurring keywords of products in those markets (after excluding the terms "software" and "management"):

- {filesharing, syncing, file}: Dropbox for Business, Box, Google Drive, OneDrive, etc.
- {presentation, presentations, tool}: PowerPoint, Google Slides, Slidebean, Pitch, etc.
- {development, application, build}: Github, Gitlab, Bitbucket, etc.
- {browser, internet, email}: Google Chrome, Firefox, Safari, Microsoft Edge, Yandex Browser, Tor Browser, etc.
- {customer, service, call}: Kustomer, Zendesk, Freshdesk, Hiver, Salesforce Service Cloud, etc.

1.C.9 Size of Enterprise Software Industry

These computations are based on full *Crunchbase* data (instead of only *Crunchbase* firms that are matched to *Capterra*), and thus separate of the main part of the paper. I compare enterprise software, and biotechnology / pharmaceuticals, as both of these industries are thought to be captured especially well on *Crunchbase*, and characterized by high innovation. As *Crunchbase* does not specifically distinguish industries, I define these industries as follows:

Definition of Enterprise Software. I define as belonging to *enterprise software* all *Crunchbase* organizations that have any of the following categories:

- Sales Automation, Enterprise Software, Advertising, Developer Tools, Web Development, SaaS, Digital Marketing, Analytics, SEO, Business Intelligence, CRM, Web Hosting, Cyber Security, Cloud

I then exclude all organizations that have any of the following categories:

- Biotechnology, Pharmaceutical, Hardware, Insurance, Physical Security, GreenTech, Oil and Gas, Farming, Wine and Spirits, Packaging Services, Solar, Air Transportation, Aerospace, Consulting, Robotics, Semiconductor, Wearables, Sensor, Power Grid, Audiobooks, Video Game, Medical Device

Definition of Biotech and Pharma. I define as belonging to *biotechnology and pharmaceuticals* all *Crunchbase* organizations in any of the following categories:

- Biotechnology, Pharmaceutical

I then exclude all organizations that have any of the following categories:

- Enterprise Software, SaaS, Machine Learning, Artificial Intelligence

I then look at only relevant VC funding rounds, with VC funding rounds defined as in 1.C.2. I find that between 2005 and 2020, enterprise software startups worldwide have raised US$237 billion, whereas pharmaceutical and biotechnology startups have raised US$177. Looking at all investments (not only VC investments), the enterprise software industry has received US$319, whereas the pharmaceutical and biotechnology industry has received US$278. (Note, however, that it is possible that R&D in pharmaceuticals and biotechnology is less likely to be VC funded.)

1.C.10 In Software Markets, Startup Acquisitions are Especially Prevalent

I first document the high prevalence of startup acquisitions in the software industry compared to other industries: firms active in software are among the most important *acquirers* of VC-backed startups (Section 1.C.10), and successful *targets* active in enterprise software predominantly exit via acquisition (Section 1.C.10). These findings suggests that the motives for these numerous startup acquisitions may be specific to the software industry, which provides a motivation for conducting the study *within* this industry.

Rank	Acquirer name	# startups acquired	Acquirer name	Billion US$
1	**Alphabet**	139	**Facebook**	24.3
2	**Microsoft**	75	Walmart	19.6
3	**Apple**	68	**Alibaba Group**	15.3
4	**Cisco**	67	**Cisco**	15.0
5	**Facebook**	66	**Alphabet**	12.8
6	**Dell EMC**	64	**Microsoft**	12.4
7	Vista Equity Partners	54	**eBay**	10.8
8	**Amazon**	53	**SAP**	8.7
9	**Yahoo**	49	Illumina	8.7
10	**Salesforce**	48	**Intuit**	8.5
11	**Twitter**	45	**Didi**	8.0
12	**Oracle**	38	**Amazon**	7.5
13	**Intel**	37	Johnson & Johnson	6.9
14	**eBay**	34	Merck	6.8
15	Thoma Bravo	32	**Dell EMC**	6.3
16	**IBM**	32	Investor AB	6.3
17	Walmart	29	Roche	6.3
18	**Alibaba Group**	26	**Uber**	6.0
19	**Groupon**	25	Bristol-Myers Squibb	5.9
20	**IAC**	22	AbbVie	5.8

Table 1.19: Largest acquirers of VC-funded startups of any industry (first exits only, excluding LBOs and management buyouts), in count (left) and transaction volume (right), 2005-2020. Companies active in digital technology or software in **bold**. Acquisition prices are missing in 82% of observations, most likely for smaller acquisitions and startups in financial distress ("fire sales", see Kerr et al. (2014)). I consider acquired startups worldwide, but startups located in North America or Europe are most likely over-represented on *Crunchbase*.

Acquirer side: in terms of numbers, the most important acquirers of startups of any industry are software firms

Table 1.19 shows the top twenty acquiring firms of VC-funded startups in 2005-2020.[77] For each acquirer, I sum up both the number of acquired firms, as well as the transaction prices.[78] Looking at the names of the top 20 acquirers in terms of the number of acquired firms (left column), what is striking is that most of the listed companies are producers of software. The GAFAM are among the top 10 acquiring firms, but many other digital technology firms are very active in startup acquisitions as well. Even relatively young and smaller companies like Groupon, Dropbox, or Twitter, are among the top 20 acquirers of VC-funded startups *of any industry*. Looking at top acquirers of VC-funded startups in terms of dollar volume, a different set of companies shows up, with financial and biotechnology firms appearing as top acquirers. Overall, this pattern hints at the idea that acquisitions of startups may be important for essentially all software firms. However, software firms tend to acquire companies at lower prices, but more of them, compared to companies active in finance or pharmaceuticals.

[77] Note that I here do not place any restriction on the type of industry or geographic location of acquirer or target firm, and use the entire *Crunchbase* database, as opposed to the *Crunchbase-Capterra* match.

[78] Acquisitions conducted by subsidiaries of a parent firm are counted as the parent firm's acquisition. This means: acquisitions conducted by Flipkart after Walmart purchased a majority stake in Flipkart are counted as acquisitions by Walmart, for instance. If I do not take into account these acquisitions by subsidiaries, the left column in fact contains only software firms.

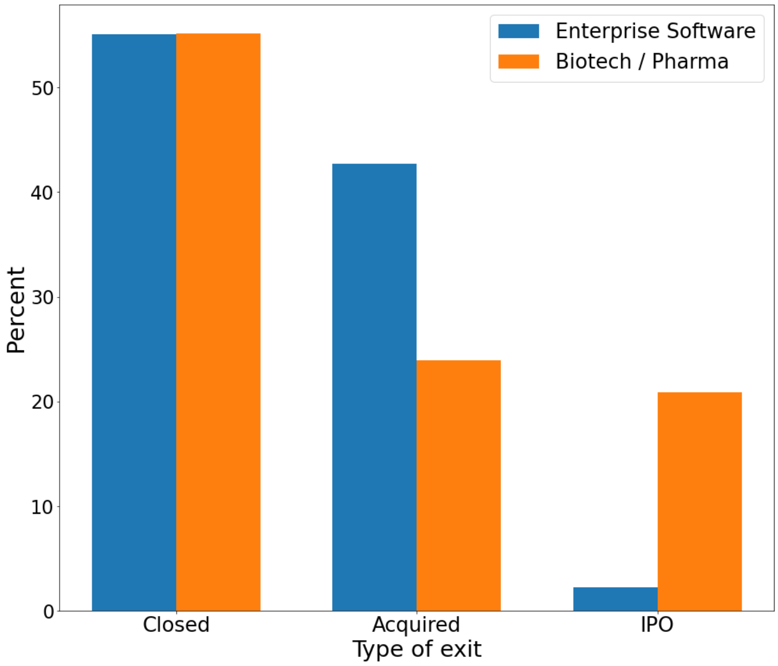


Figure 1.9: Types of exits of startups in biotechnology & pharmaceuticals, and enterprise software, in percent. I consider US-based startups founded after 2001 and exiting in 2005-2020. Details on the industry definition can be found in Appendix 1.C.9.

Target side: startups in software are more likely to exit via acquisition than startups in other industries

Next, I turn from acquirers to potential targets by comparing how startups in different industries "exit" the private financial market. I here juxtapose startups active in enterprise software with startups active in biotechnology and pharmaceutical industries. As explained in Section 2.1, startups can successfully exit either by being acquired, or by being listed as a public company on a stock exchange. Whereas failure rates are remarkably similar (55%) for startups active in both industries[79], I find that out of all successfully exiting startups in enterprise software, 95% exit by acquisition. In the biotechnology and pharmaceutical industry, the common exit routes are strikingly different: here, 53% of successful startups exit by acquisition. The finding highlights once again that motives for entry and acquisitions might be fundamentally different across industries (due to different production technologies etc.), and that within-industry studies are needed to fully comprehend the effects of startup acquisitions.

1.D Further Details on Different Acquirer Types

The three types of acquiring companies – enterprise software, financial, and other – not only vary by industry sector, but also in terms of other characteristics. Enterprise software acquirers are more likely to once have been VC-funded themselves (68%), tend to be somewhat younger than financial or other acquirers, and tend to be located in the US and California. Industry outsider firms are relatively more

[79] This rate is in line with empirical finance literature, e.g. Kerr et al. (2014), who find that 55% of startups that received VC funding were terminated at a loss.

Panel A: "Broad" groups of acquirers (exhaustive)

Acquirer type	Number of funding rounds (mean)	Volume of funding (million USD, median)	% funding volume is not available
Enterprise Software	2.7	7.4	12.0
Financial	2.6	10.0	12.0
Other Industries	2.8	8.5	14.9

Panel B: Looking at non-exhaustive groups of enterprise software acquirers, and IPOs

Acquirer type	Number of funding rounds (mean)	Volume of funding (million USD, median)	% funding volume is not available
GAFAM	2.7	10.0	9.6
New tech	2.9	10.8	12.5
Old tech	3.3	25.2	10.5
Pre-exit	2.5	3.4	14.6
IPO	4.5	101.0	4.1

Table 1.20: Number and volume of VC funding rounds at exits of VC-funded startups, 2012-2020. Excludes leveraged buyouts or management buyouts.

likely to be foreign to the target compared to the other groups. Financial companies tend to be much smaller than acquirers of the other types in terms of employment size, and are less likely to be public companies. I found that for only 35% of acquisitions, the acquirer is a public company as of 2021.

Table 1.20 shows the pattern of funding rounds received by different startups at the time of exit. It closely mirrors the patterns observed for startup age, price and valuation (Tables 1.5 and 1.6) at exit.

1.E Many Acquired Products are Discontinued After the Acquisition: Further Anecdotal Evidence

Acquihires. Acquisitions on *Crunchbase* may be tagged with an "acqui-hire" tag. I find that 2.6% of acquisitions of startups in which the product was shut down are recorded as acqui-hire events. In contrast, for products kept alive, only 0.7% of acquisitions are recorded as an acqui-hire.

Timing. I do not observe the timing of the shut-down in the data. However, anecdotally there are both, cases in which the shut-down was announced right at the time of the acquisition (e.g. Box-Wagon, Dropbox-CloudOn, Dropbox-Verst, Google-AppJet), or after a few years (e.g. Microsoft-Wunderlist, Dropbox-Mailbox, Qlik-DataMarket, or Oracle-Ravello Systems, whose products were shut down between two and four years after the acquisition).

For those startups that were acquired and kept alive, I can compile further descriptives using the web-scraped product-level data. I first look at the number of products produced by an acquired firm. I find that those startups that exited via IPO or via an acquisition by a financial acquirer have on average 2 or 1.4 products respectively, as of 2021. In contrast, companies exiting by GAFAM or pre-exit firms are always single-product. Next, I look at the number of reviews of products acquired and continued, which could be an indication for demand. Table 1.21 reveals that products acquired by the GAFAM tend to have many more reviews. However, it should however be born in mind that the GAFAM are also especially likely to discontinue products. Moreover, it is not clear whether high number in reviews indicates that the acquisition has boosted demand for these products, or whether these products were

Panel A:

Acquirer type	Number of reviews (median)	Number of reviews (mean)
Enterprise Software	2.0	152.5
Financial	1.0	44.8
Other Industries	1.0	48.8
Panel B: Looking at a subset of Enterprise Software acquirer groups:		
GAFAM	19.0	1234.5
New tech	8.0	26.5
Old tech	2.0	40.5
Pre-exit	2.0	13.0
IPO	12.0	572.4

Table 1.21: Number of reviews, VC-funded startups with continued products only, 2012-2020. For multi-product firms, I sum the reviews of all products supplied by a given firm.

previously successful ones.

1.F Products by VC-funded Startups Tend to Have More Reviews

Reviews could be interpreted as a proxy for product demand, or for product quality. In Table 1.22, columns (1) and (2) show the results of a regression of the number of reviews of a given product on firm characteristics; in particular, on the number of VC funding rounds (column (1)) and on whether or not the firm has received any VC funding round (column (2)). Columns (3) and (4) show the results of a regression of the average number of reviews of a given company's products on the same set of regressors. Note that both regressions use cross-sectional data only.

It is remarkable that funding rounds seem positively correlated with the number of reviews, even after accounting for company cohort, company employee size, and "status" (acquired, IPO, operating, inactive, closed). In general, however, there seem to be a lot of other factors explaining the number of reviews, as indicated by the low adjusted R^2.

Table 1.22: Regression using cross-sectional data: what explains product reviews?

	Dependent variable:			
	Product-level data: num_reviews		Company-level data: mean_reviews	
	(1)	(2)	(3)	(4)
# of VC funding rounds received by producing company	9.996**		10.167**	
	(4.119)		(4.496)	
1{Any VC funding round received by producing company}		12.035		25.460***
		(9.499)		(8.329)
as.factor(status)closed	−37.845***	−36.730***	−44.914***	−46.426***
	(10.975)	(11.478)	(11.161)	(11.612)
as.factor(status)inactive	−6.695	−11.255	−17.029	−24.692**
	(9.719)	(10.452)	(10.714)	(11.449)
as.factor(status)ipo	124.192***	126.652***	9.465	11.629
	(39.476)	(39.487)	(32.137)	(32.110)
as.factor(status)operating	−5.109	−0.711	−21.051	−18.699
	(11.793)	(11.765)	(12.923)	(12.678)
as.factor(employee_count)10000+	311.316***	317.951***	150.189***	160.931***
	(66.689)	(66.293)	(49.833)	(50.204)
as.factor(employee_count)1001-5000	185.764***	199.116***	255.708***	271.787***
	(40.302)	(43.357)	(70.471)	(75.298)
as.factor(employee_count)101-250	14.577	26.337***	22.658	34.902***
	(11.466)	(10.049)	(13.821)	(11.710)
as.factor(employee_count)11-50	−2.108	1.727	0.649	4.299*
	(2.944)	(2.498)	(3.045)	(2.445)
as.factor(employee_count)251-500	22.710***	34.788***	41.584***	55.727***
	(8.668)	(9.023)	(10.657)	(10.951)
as.factor(employee_count)5001-10000	89.000**	97.722***	102.069**	111.804**
	(36.779)	(37.013)	(51.939)	(51.964)
as.factor(employee_count)501-1000	102.530***	114.908***	129.627***	143.098***
	(27.788)	(28.096)	(33.180)	(34.255)
as.factor(employee_count)51-100	3.110	11.495***	6.789	14.910***
	(4.860)	(3.862)	(5.055)	(3.573)
as.factor(employee_count)unknown	24.615***	28.426***	17.991***	25.327***
	(7.214)	(7.771)	(5.864)	(6.506)
Company year-of-birth FE	✓	✓	✓	✓
Observations	20,432	20,432	16,374	16,374
Adjusted R^2	0.031	0.030	0.018	0.016

Standard errors are heteroskedasticity-robust. *p<0.1; **p<0.05; ***p<0.01

1.G Robustness: Event Studies

Table 1.23: Event study like in Table 1.10, 4 quarters, 2012-2020, but Poisson model instead of linear model.

	Dependent variable:				
	Number of entrants in market m, quarter t				
	(1)	(2)	(3)	(4)	(5)
Major acq by enterpr softw company	−0.097*				
	(0.058)				
Major acq by public enterpr software softw		−0.126*			
		(0.068)			
Major acq by GAFAM or 'New Tech'			−0.306***		
			(0.111)		
Major acq by company in other industry				−0.116	
				(0.081)	
Major acq by financial company					0.116
					(0.132)
Market FE	✓	✓	✓	✓	✓
Quarter FE	✓	✓	✓	✓	✓
Observations	17,064	17,064	17,064	17,064	17,064
Log Likelihood	−15,983.270	−15,982.910	−15,980.500	−15,983.920	−15,984.630
Akaike Inf. Crit.	32,986.550	32,985.820	32,981.010	32,987.850	32,989.270
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 1.24: Event study like in Table 1.10, 4 quarters, 2012-2020, but using all acquisitions above a transaction value of 50US$ million as events.

	Dependent variable:				
	Number of entrants in market m, quarter t				
	(1)	(2)	(3)	(4)	(5)
Major acq by enterpr softw company	−0.216*				
	(0.113)				
Major acq by public enterpr software softw		−0.282**			
		(0.122)			
Major acq by GAFAM or 'New Tech'			−0.361***		
			(0.093)		
Major acq by company in other industry				−0.098	
				(0.060)	
Major acq by financial company					−0.090
					(0.127)
Market FE	✓	✓	✓	✓	✓
Quarter FE	✓	✓	✓	✓	✓
Adjusted R^2	0.299	0.299	0.3	0.299	0.299
Observations	17,064	17,064	17,064	17,064	17,064
SEs clustered on market level.					*p<0.1; **p<0.05; ***p<0.01

Table 1.25: Cumulative sum of major acquisitions of a given type in a given market and startup entry. Market-quarter panel, 2012-2020.

	Dependent variable:				
	Number of entrants in market m, quarter t (Sample mean: 0.65)				
	(1)	(2)	(3)	(4)	(5)
Major acq by enterprise software company (ranges from 0 to 4)	-0.142^{***} (0.053)				
Major acq by public enterprise software company (ranges from 0 to 4)		-0.133^{**} (0.065)			
Major acq by GAFAM or 'New Tech' (ranges from 0 to 2)			-0.297^{***} (0.085)		
Major acq by company not in enterpr softw (incl. financial) (ranges from 0 to 2)				-0.141^{*} (0.072)	
Major acq by financial company (ranges from 0 to 1)					-0.129 (0.105)
Market FE	✓	✓	✓	✓	✓
Quarter FE	✓	✓	✓	✓	✓
Adjusted R^2	0.3	0.299	0.3	0.299	0.299
Observations	17,064	17,064	17,064	17,064	17,064

SEs clustered on market level. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 1.26: Standard diff-in-diff, using events as in Table 1.10, 2012-2020.

	Dependent variable:				
	Number of entrants in market m, quarter t				
	(1)	(2)	(3)	(4)	(5)
Major acq by enterpr softw company	-0.173^{**} (0.075)				
Major acq by public enterpr software softw		-0.144 (0.089)			
Major acq by GAFAM or 'New Tech'			-0.343^{***} (0.112)		
Major acq by company in other industry				-0.161^{*} (0.091)	
Major acq by financial company					-0.129 (0.105)
Market FE	✓	✓	✓	✓	✓
Quarter FE	✓	✓	✓	✓	✓
Adjusted R^2	0.3	0.3	0.3	0.299	0.299
Observations	17,064	17,064	17,064	17,064	17,064

SEs clustered on market level. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 1.27: Same event study as in main text (Table 1.10), using event window of 4 quarters, but this time using data from 2014-2019.

	Dependent variable:				
	Number of entrants in market m, quarter t				
	(1)	(2)	(3)	(4)	(5)
Major acq by enterpr softw company	−0.123				
	(0.075)				
Major acq by public enterpr software softw		−0.152			
		(0.093)			
Major acq by GAFAM or 'New Tech'			-0.539^{***}		
			(0.140)		
Major acq by company in other industry				−0.050	
				(0.107)	
Major acq by financial company					0.174
					(0.155)
Market FE	✓	✓	✓	✓	✓
Quarter FE	✓	✓	✓	✓	✓
Adjusted R^2	0.304	0.304	0.305	0.304	0.304
Observations	11,376	11,376	11,376	11,376	11,376
SEs clustered on market level.			*p<0.1; **p<0.05; ***p<0.01		

Table 1.28: Event window: 5 quarters. Market-quarter panel, 2012-2020.

	Dependent variable:				
	Number of entrants in market m, quarter t				
	(1)	(2)	(3)	(4)	(5)
Major acq by enterpr softw company	−0.088				
	(0.067)				
Major acq by public enterpr software softw		-0.138^{*}			
		(0.078)			
Major acq by GAFAM or 'New Tech'			-0.407^{***}		
			(0.139)		
Major acq by company in other industry				−0.079	
				(0.077)	
Major acq by financial company					−0.064
					(0.141)
Market FE	✓	✓	✓	✓	✓
Quarter FE	✓	✓	✓	✓	✓
Adjusted R^2	0.299	0.299	0.3	0.299	0.299
Observations	17,064	17,064	17,064	17,064	17,064
SEs clustered on market level.			*p<0.1; **p<0.05; ***p<0.01		

Table 1.29: Event window: 3 quarters. Market-quarter panel, 2012-2020.

	Dependent variable:				
	Number of entrants in market m, quarter t				
	(1)	(2)	(3)	(4)	(5)
Major acq by enterpr softw company	−0.093				
	(0.072)				
Major acq by public enterpr software softw		−0.114			
		(0.097)			
Major acq by GAFAM or 'New Tech'			-0.328^{**}		
			(0.151)		
Major acq by company in other industry				−0.100	
				(0.085)	
Major acq by financial company					−0.023
					(0.190)
Market FE	✓	✓	✓	✓	✓
Quarter FE	✓	✓	✓	✓	✓
Adjusted R^2	0.299	0.299	0.3	0.299	0.299
Observations	17,064	17,064	17,064	17,064	17,064
SEs clustered on market level.			*p<0.1; **p<0.05; ***p<0.01		

1.H Alternative Model Specification

The model covered in the main text contains $\boldsymbol{x}_{mt} = \{N_{mt}, A^{\text{strat}}_{mt}, l(m)\}$ as a vector of state variables. Here, I consider instead an alternative version of the model with $\boldsymbol{x}_{mt} = \{N_{mt}, A^{\text{strat in } t-K}_{mt}, l(m)\}$ as state variables. $A^{\text{strat in } t-K}_{mt}$ mirrors the event study indicator variables employed in Section 4, and is equal to 1 in the event of a strategic acquisition in the past K quarters, and 0 otherwise.

I set $K = 4$, as in the reduced-form regressions in Section 4. The first-stage results are displayed in Tables 1.30 and 1.31. The coefficients of the strategic acquisiiton affect increase somewhat in magnitude, as the variable is now a dummy (instead of the cumulative number), and remains mostly insignficiant. There is essentially no change in the fit of the first-stage regression model.

The second stage results are displayed in Table 1.32. Judging from the log-likelihood, the fit of the model is somewhat worse compared to the main results in the text, and the coefficients of γ^A are insignificant in both specifications.

Table 1.30: First stage, using a broad definition of "strategic" acquirers, and a moving average indicator variable with window length 4 quarters to capture strategic acquisition effect. Standard errors are clustered at the market level. In column (4), they are computed using block bootstrapping with 5,000 bootstraps to account for the estimated market-category fixed effect.

	Dependent variable:				
	Number of entrants in market m, quarter t				
			Poisson		
	(1)	(2)	(3)	(4)	(5)
# of competitors	0.022***	−0.163***	−0.118***	−0.066***	−0.161***
	(0.001)	(0.015)	(0.017)	(0.012)	(0.015)
Major Enterprise Software acquisition pre-4Q	−0.056	−0.002	0.050	−0.021	−0.007
	(0.069)	(0.067)	(0.079)	(0.078)	(0.067)
1{quarter=2}					−0.126***
					(0.020)
1{quarter=3}					−0.151***
					(0.019)
1{quarter=4}					−0.212***
					(0.019)
Market FE		✓	✓		✓
20 market-category FE				✓	
Adjusted R^2	0.11	0.34		0.24	0.35
Log Likelihood			−9,813.634		
Akaike Inf. Crit.			20,511.270		
Observations	10,560	10,560	10,560	10,560	10,560
Standard errors clustered at market level.			*p<0.1; **p<0.05; ***p<0.01		

Table 1.31: First stage, using a narrower definition of "strategic" acquirers, and a moving average indicator variable with window length 4 quarters to capture strategic acquisition effect. Standard errors are clustered at the market level. In column (4), they are computed using block bootstrapping with 5,000 bootstraps to account for the estimated market-category fixed effect.

	Dependent variable:				
	Number of entrants in market m, quarter t				
			Poisson		
	(1)	(2)	(3)	(4)	(5)
# of competitors	0.022***	−0.162***	−0.117***	−0.066***	−0.161***
	(0.001)	(0.015)	(0.017)	(0.012)	(0.015)
Major New Tech or GAFAM acquisition pre-4Q	−0.311*	−0.205	−0.195	−0.178	−0.203
	(0.174)	(0.162)	(0.176)	(0.210)	(0.163)
1{quarter=2}					−0.125***
					(0.020)
1{quarter=3}					−0.151***
					(0.019)
1{quarter=4}					−0.212***
					(0.019)
Market FE		✓	✓		✓
20 market-category FE				✓	
Adjusted R^2	0.11	0.34		0.24	0.35
Log Likelihood			−9,812.992		
Akaike Inf. Crit.			20,509.980		
Observations	10,560	10,560	10,560	10,560	10,560
Standard errors clustered at the market level.			*p<0.1; **p<0.05; ***p<0.01		

Table 1.32: Estimates of structural parameters, this time using $A_{mt}^{\text{strat in } t-K}$ as state variable with K=4.

	(1)	(2)
Entry costs, κ	-2.957^{***}	-2.984^{***}
	(0.137)	(0.138)
log(# of competitors), γ^N	-0.248^{***}	-0.248^{***}
	(0.011)	(0.011)
Strategic acq of competitor **by Enterprise Software acquirer**, γ^A	-0.032	
(Dummy indicating major such acquisition in past 4 quarters)	(0.027)	
Strategic acq of competitor **by GAFAM or New Tech**, γ^A		-0.085
(Dummy indicating major such acquisition in past 4 quarters)		(0.053)
Own IPO in future, α^{ipo}	0.006^{***}	0.005^{***}
	(0.001)	(0.001)
Own acquisition in future, α^{acq}	0.037^{***}	0.038^{***}
	(0.003)	(0.003)
Market category 2, γ_2^M (5th-10th perc)	0.319^{***}	0.320^{***}
	(0.024)	(0.024)
Market category 3, γ_3^M (10th-15th perc)	0.390^{***}	0.392^{***}
	(0.025)	(0.025)
...	...	...
Market category 19, γ_{19}^M (90th-95thth perc)	1.142^{***}	1.143^{***}
	(0.046)	(0.046)
Market category 20, γ_{20}^M (95th-100th perc)	1.301^{***}	1.301^{***}
	(0.050)	(0.050)
Log-likelihood	-10631.63	-10617.22
Observations: 440 markets, 24 quarters	10,560	10,560

Note: *p<0.1; **p<0.05; ***p<0.01

References

Affeldt, P., & Kesler, R. (2021a). Big Tech Acquisitions - Towards Empirical Evidence. *Journal of European Competition Law and Practice*, *12*(6), 471–478. doi: 10.1093/jeclap/lpab025

Affeldt, P., & Kesler, R. (2021b). Competitors' Reactions to Big Tech Acquisitions: Evidence from Mobile Apps. *Working Paper*.

Aghamolla, C., & Thakor, R. T. (2021). IPO peer effects. *Journal of Financial Economics*. doi: 10.1016/j.jfineco.2021.05.055

Aguirregabiria, V., & Ho, C. Y. (2012). A dynamic oligopoly game of the US airline industry: Estimation and policy experiments. *Journal of Econometrics*, *168*(1), 156–173. Retrieved from http://dx.doi.org/10.1016/j.jeconom.2011.09.013 doi: 10.1016/j.jeconom.2011.09.013

Aguirregabiria, V., & Mira, P. (2007). Sequential Estimation of Dynamic Discrete Games. *Econometrica*, *75*(1), 1–53. Retrieved from https://doi.org/10.1111/j.1468-0262.2007.00731.x doi: https://doi.org/10.1111/j.1468-0262.2007.00731.x

Aguirregabiria, V., & Mira, P. (2010). Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, *156*(1), 38–67. Retrieved from https://www.sciencedirect.com/science/article/pii/S0304407609001985 doi: https://doi.org/10.1016/j.jeconom.2009.09.007

Arcidiacono, P., & Ellickson, P. B. (2011). Practical Methods for Estimation of Dynamic Discrete Choice Models. *Annual Review of Economics*, *3*(1), 363–394. Retrieved from https://doi.org/10.1146/annurev-economics-111809-125038 doi: 10.1146/annurev-economics-111809-125038

Argentesi, E., Buccirossi, P., Calvano, E., Duso, T., Marrazzo, A., & Nava, S. (2021). Merger Policy in Digital Markets: An Ex Post Assessment. *Journal of Competition Law and Economics*, *17*(1), 95–140. doi: 10.1093/joclec/nhaa020

Aridor, G. (2022). Drivers of Digital Attention: Evidence from a Social Media Experiment. *Working Paper*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4069567

Arrow, K. (1962). Economic welfare and the allocation of resources for invention. *The rate and direction of inventive activity: Economic and social factors*, 609–626.

Bajari, P., Benkard, C. L., & Levin, J. (2007). Estimating Dynamic Models of Imperfect Competition. *Econometrica*, *75*(5), 1331–1370. Retrieved from https://doi.org/10.1111/j.1468-0262.2007.00796.x doi: https://doi.org/10.1111/j.1468-0262.2007.00796.x

Bauer, J. M., & Prado, T. S. (2021). Effects of Big Tech Acquisitions on Start-up Funding and Innovation.

Working Paper. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3787127&download=yes

Berman, R., & Israeli, A. (2022). The Value of Descriptive Analytics: Evidence from Online Retailers. *Marketing Science*, *forthcomin*. Retrieved from https://pubsonline.informs.org/doi/10.1287/mksc.2022.1352

Bisceglia, M., Padilla, J., Perkins, J., & Piccolo, S. (2021). Optimal Exit Policy with Uncertain Demand. *Working Paper*.

Bresnahan, T. F., & Reiss, P. C. (1991). Entry and Competition in Concentrated Markets. *Journal of Political Economy*, *99*(5), 977–1009.

Bryan, K. A., & Hovenkamp, E. (2020). Antitrust Limits on Startup Acquisitions. *Review of Industrial Organization*, *56*(4), 615–636. doi: 10.1007/s11151-020-09751-5

Cabral, L. (2018). Standing on the Shoulders of Dwarfs: Dominant Firms and Innovation Incentives. *CEPR Discussion Paper*. Retrieved from http://luiscabral.net/economics/workingpapers/innovation201808.pdf

Cabral, L. (2021). Merger policy in digital industries. *Information Economics and Policy*, *54*. doi: 10.1016/j.infoecopol.2020.100866

Callander, S., & Matouschek, N. (2020). The Novelty of Innovation: Competition, Disruption, and Antitrust Policy. *Working Paper*. Retrieved from https://sjc.people.stanford.edu/sites/g/files/sbiybj4051/f/innovation_13october2020.pdf

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230. Retrieved from https://doi.org/10.1016/j.jeconom.2020.12.001 doi: 10.1016/j.jeconom.2020.12.001

Chopra, A. (2018). Private equity buyouts have become viable exit options — even for early-stage startups. *TechCrunch*, *28 Decembe*. Retrieved from https://web.archive.org/web/20210615140407/https://techcrunch.com/2018/12/28/private-equity-buyouts-have-become-viable-exit-options-even-for-early-stage-startups/

Cockburn, I. M., & MacGarvie, M. J. (2011). Entry and Patenting in the Software Industry. *Management Science*, *57*(5), 915–933. Retrieved from http://www.jstor.org/stable/25835748

Collard-Wexler, A. (2013). Demand Fluctuations in the Ready-Mix Concrete Industry. *Econometrica*, *81*(3), 1003–1037. Retrieved from https://doi.org/10.3982/ECTA6877 doi: https://doi.org/10.3982/ECTA6877

Conti, A., Guzman, J., & Rabi, R. (2021). Herding in the Market for Startup Acquisitions.

Crémer, J., de Montjoye, Y.-A., & Schweitzer, H. (2019). Competition Policy for the Digital Era. *Final Report for the Commission Directorate-General for Competition*.

Crunchbase. (2022). *Glossary of Funding Types.* Retrieved 2022-03-12, from https://support.crunchbase.com/hc/en-us/articles/115010458467-Glossary-of-Funding-Types

Cunningham, C., Ederer, F., & Ma, S. (2021). Killer Acquisitions. *Journal of Political Economy*, *129*(3), 649–702.

David, J. M. (2021). The Aggregate Implications of Mergers and Acquisitions. *The Review of Economic Studies*, *88*(4), 1796–1830. doi: 10.1093/restud/rdaa077

Decarolis, F., & Rovigatti, G. (2021). From Mad Men to Maths Men: Concentration and Buyer Power in Online Advertising. *American Economic Review*, *111*(10), 3299–3327. doi: 10.1257/AER.20190811

Delgado, M., Porter, M. E., & Stern, S. (2016). Defining clusters of related industries. *Journal of Economic Geography*, *16*(1), 1–38.

Denicolò, V., & Polo, M. (2021). Acquisitions, innovation and the entrenchment of monopoly. *Working Paper*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3988255

Dijk, E., Moraga-Gonzàlez, J. L., & Motchenkova, E. (n.d.). How Do Start-up Acquisitions Affect the Direction of Innovation? *Journal of Industrial Economics*.

Ederer, F., & Pellegrino, B. (2023). The Great Startup Sellout and the Rise of Oligopoly. *AEA Papers & Proceedings*, *113*, 1–5. doi: 10.2139/ssrn.4322834

Ericson, R., & Pakes, A. (1995). Markov-Perfect Industry Dynamics: A Framework for Empirical Work. *The Review of Economic Studies*, *62*(1), 53–82. Retrieved from http://www.jstor.org/stable/2297841 doi: 10.2307/2297841

Ershov, D. (2022). Competing with Superstars in the Mobile App Market. *Working Paper*. Retrieved from https://drive.google.com/file/d/1xst6SPBR9taDLbiPa__XtaG5eF5ojTVZ/view doi: 10.2139/ssrn.3265662

Federal Trade Commission. (2021). Non-HSR Reported Acquisitions by Select Technology Platforms, 2010-2019: An FTC Study. *Technical Report*. Retrieved from https://www.ftc.gov/system/files/documents/reports/non-hsr-reported-acquisitions-select-technology-platforms-2010-2019-ftc-study/p201201technologyplatformstudy2021.pdf

Fons-Rosen, C., Roldan-Blanco, P., & Schmitz, T. (2023). The Aggregate Effects of Acquisitions on Innovation and Economic Growth. *Working Paper*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3785485 doi: 10.2139/ssrn.3785485

Foroughi, C., & Stern, A. D. (2019). Who Drives Digital Innovation?: Evidence from the US Medical Device Industry. *HBS Working Paper*. Retrieved from https://www.hbs.edu/ris/PublicationFiles/19-120_a1d7a285-f669-462f-a7bd-6643b73b1a7b.pdf

Fumagalli, C., Motta, M., & Tarantino, E. (2022). Shelving or Developing? Acquisition of Potential Competitors. *Working Paper*.

Furman, J., Coyle, D., Fletcher, A., Marsden, P., & McAuley, D. (2019). Un-locking digital competition. Report of the Digital Competition Expert Panel.

Gans, J. S., & Stern, S. (2003). The product market and the market for "ideas": Commercialization strategies for technology entrepreneurs. *Research Policy*, *32*(2), 333–350. doi: 10.1016/S0048-7333(02)00103-8

Gautier, A., & Lamesch, J. (2021). Mergers in the digital economy. *Information Economics and Policy*, *54*, 1–34. doi: 10.1016/j.infoecopol.2020.100890

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, *57*(3), 535–574.

Retrieved from https://www.aeaweb.org/articles?id=10.1257/jel.20181020 doi: 10.1257/jel.20181020

Gilbert, R. J., & Katz, M. L. (2022). Dynamic merger policy and pre-merger product choice by an entrant. *International Journal of Industrial Organization*, *81*. Retrieved from https://doi.org/10.1016/j.ijindorg.2021.102812 doi: 10.1016/j.ijindorg.2021.102812

Gompers, P., & Lerner, J. (2001). The venture capital revolution. *Journal of Economic Perspectives*, *15*(2), 145–168. doi: 10.1257/jep.15.2.145

Grimmer, J., & King, G. (2011). General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, *108*(7), 2643–2650.

Guéron, Y., & Lee, J. (2022). Innovation Learning, and Killer Acquisitions. *Working Paper*.

Haucap, J., Rasch, A., & Stiebale, J. (2019). How mergers affect innovation: Theory and evidence. *International Journal of Industrial Organization*, *63*, 283–325. Retrieved from https://doi.org/10.1016/j.ijindorg.2018.10.003 doi: 10.1016/j.ijindorg.2018.10.003

Hege, U., & Hennessy, C. (2010). Acquisition Values and Optimal Financial (In)Flexibility. *The Review of Financial Studies*, *23*(7), 2865–2899. Retrieved from https://doi.org/10.1093/rfs/hhq017 doi: 10.1093/rfs/hhq017

Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, *124*(5), 1423–1465.

Hollenbeck, B. (2020). Horizontal mergers and innovation in concentrated industries. *Quantitative Marketing and Economics*, *18*(1), 1–37. doi: 10.1007/s11129-019-09218-2

Hotz, J. V., & Miller, R. A. (1993). Conditional choice probabilities and the estimation of dynamic models. *Review of Economic Studies*, *60*(3), 497–529. doi: 10.2307/2298122

Hotz, J. V., Miller, R. A., Sanders, S., & Smith, J. (1994). A simulation estimator for dynamic models of discrete choice. *The Review of Economic Studies*, *61*(2), 265–289.

Igami, M. (2017). Estimating the innovator's dilemma: Structural analysis of creative destruction in the hard disk drive industry, 1981–1998. *Journal of Political Economy*, *125*(3), 798–847. doi: 10.1086/691524

Igami, M., & Uetake, K. (2020). Mergers, Innovation, and Entry-Exit Dynamics: Consolidation of the Hard Disk Drive Industry, 1996-2016. *Review of Economic Studies*, *87*(6), 2672–2702. doi: 10.1093/restud/rdz044

Igami, M., & Yang, N. (2016). Unobserved heterogeneity in dynamic games: Cannibalization and preemptive entry of hamburger chains in Canada. *Quantitative Economics*, *7*(2), 483–521. Retrieved from https://doi.org/10.3982/QE478 doi: https://doi.org/10.3982/QE478

Janssen, R., Kesler, R., Kummer, M. E., & Waldfogel, J. (2022). GDPR and the lost generation of innovative apps. *NBER Working Paper No. w30028*.

Jin, G. Z., Leccese, M., & Wagman, L. (2022). How Do Top Acquirers Compare in Technology Mergers? New Evidence from an S&P Taxonomy. *Working Paper*. doi: 10.2139/ssrn.4010491

Jin, Z. (2019). How do Large Companies Affect Entrepreneurship: Evidence From Amazon's HQ2 Search. *Working Paper*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm

?abstract_id=3280715

Jullien, B., & Lefouili, Y. (2018). Horizontal mergers and innovation. *Journal of Competition Law and Economics*, *14*(3), 364–392. doi: 10.1093/joclec/nhy010

Kamepalli, S. K., Rajan, R. G., & Zingales, L. (2021). Kill Zone. *Working Paper*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3555915 doi: 10.2139/ssrn.3555915

Katz, M. L. (2021). Big Tech mergers: Innovation, competition for the market, and the acquisition of emerging competitors. *Information Economics and Policy*, *54*. Retrieved from https://doi.org/10.1016/j.infoecopol.2020.100883 doi: 10.1016/j.infoecopol.2020.100883

Kerr, W. R., Nanda, R., & Rhodes-Kropf, M. (2014). Entrepreneurship as Experimentation. *Journal of Economic Perspectives*, *28*(3), 25–48. Retrieved from https://www.aeaweb.org/articles?id=10.1257/jep.28.3.25 doi: 10.1257/jep.28.3.25

Khan, L. M. (2021). *FTC Staff Presents Report on Nearly a Decade of Unreported Acquisitions by the Biggest Technology Companies.* Retrieved 2022-07-10, from https://www.ftc.gov/system/files/documents/public_statements/1596332/remarks_of_chair_lina_m_khan_regarding_non-hsr_reported_acquisitions_by_select_technology_platforms.pdf

Khmelnitskaya, E. (2021). Competition and Attrition in Drug Development. *Working Paper*. Retrieved from https://drive.google.com/file/d/19eJJBfJzakbkPasKy478At7O_R3Jro3p/view

Kortum, S., & Lerner, J. (2000). Assessing the Contribution of Venture Capital to Innovation. *The RAND Journal of Economics*, *31*(4), 674–692. Retrieved from http://www.jstor.org/stable/2696354 doi: 10.2307/2696354

Koski, H., Kässi, O., & Braesemann, F. (2020). *Killers on the Road of Emerging Start-ups – Implications for Market Entry and Venture Capital Financing* (ETLA Working Papers No. 81). The Research Institute of the Finnish Economy. Retrieved from https://ideas.repec.org/p/rif/wpaper/81.html

Lemley, M. A., & McCreary, A. (2020). Exit Strategy. *Working Paper*.

Letina, I., Schmutzler, A., & Seibel, R. (2021). Killer Acquisitions and Beyond: Policy Effects on Innovation Strategies. *Working Paper*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3673150 doi: 10.2139/ssrn.3673150

Lin, H. (2015). Quality Choice and Market Structure: A Dynamic Analysis of Nursing Home Oligopolies. *International Economic Review*, *56*(4), 1261–1290. Retrieved from https://doi.org/10.1111/iere.12137 doi: https://doi.org/10.1111/iere.12137

Majewska, G. (2022). Incentivizing Novelty in Antibiotic Development. *Working Paper*. Retrieved from https://gosia-majewska.github.io/files/Majewska_JMP.pdf

Mazzeo, M. J. (2002). Product Choice and Oligopoly Market Structure. *The RAND Journal of Economics*, *33*(2), 221–242. Retrieved from http://www.jstor.org/stable/3087431 doi: 10.2307/3087431

Mermelstein, B., Nocke, V., Satterthwaite, M. A., & Whinston, M. D. (2020). Internal versus external growth in industries with scale economies: A computational model of optimal merger policy. *Journal of Political Economy*, *128*(1), 301–341.

Motta, M., & Peitz, M. (2021). Big Tech Mergers. *Information Economics and Policy*, *54*. Retrieved from https://doi.org/10.1016/j.infoecopol.2020.100868 doi: 10.1016/j.infoecopol.2020.100868

Motta, M., & Shelegia, S. (2021). The "kill zone": Copying, acquisition and start-ups' direction of innovation. *Working Paper*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3846285

Ng, W., & Stuart, T. E. (2021). Acquired employees versus hired employees: Retained or turned over? *Strategic Management Journal*(October 2019), 1–21. doi: 10.1002/smj.3361

Nocke, V., & Whinston, M. D. (2010). Dynamic merger review. *Journal of Political Economy*, *118*(6), 1201–1251. doi: 10.1086/658161

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)*. Retrieved from http://nlp.

Perez-Saiz, H. (2015). Building new plants or entering by acquisition? Firm heterogeneity and entry barriers in the U.S. cement industry. *The RAND Journal of Economics*, *46*(3), 625–649. Retrieved from https://doi.org/10.1111/1756-2171.12100 doi: https://doi.org/10.1111/1756-2171.12100

Phillips, G. M., & Zhdanov, A. (2013). R&D and the Incentives from Merger and Acquisition Activity. *The Review of Financial Studies*, *26*(1), 34–78. Retrieved from https://doi.org/10.1093/rfs/hhs109 doi: 10.1093/rfs/hhs109

Poege, F. (2022). Competition and Innovation: The Breakup of IG Farben. *Working Paper*. Retrieved from https://www.felixpoege.eu/assets/publications/2021IGFarben/research_paper.pdf

Rasmusen, E. (1988). Entry for Buyout. *Journal of Industrial Economics*, *36*(3), 281–299.

Schnitzer, M., & Watzinger, M. (2022). Measuring the Spillovers of Venture Capital. *The Review of Economics and Statistics*, *104*(2), 276–292. Retrieved from https://doi.org/10.1162/rest_a_00937 doi: 10.1162/rest_a_00937

Schumpeter, J. A. (1942). Capitalism, Socialism and Democracy. *New York NY: Harper & Brothers*.

Scott Morton, F., Jullien, B., Katz, R., & Kimmelman, G. (2019). Committee for the Study of Digital Platforms Market Structure and Antitrust Subcommittee Subcommittee. *Stigler Center for the Study of the Economy and the State*.

Seim, K. (2006). An empirical model of firm entry with endogenous product-type choices. *RAND Journal of Economics*, *37*(3), 619–640. doi: 10.1111/j.1756-2171.2006.tb00034.x

Song, M. H., & Walkling, R. A. (2000). Abnormal returns to rivals of acquisition targets: A test of the 'acquisition probability hypothesis'. *Journal of Financial Economics*, *55*(2), 143–171. doi: 10.1016/S0304-405X(99)00048-3

Stahl, J. C. (2011). A Dynamic Analysis of Consolidation in the Broadcast Television Industry. *Working Paper*. Retrieved from https://1e91077d-a-62cb3a1a-s-sites.googlegroups.com/site/jessicacalfeestahl/home/JCStahl_TV_7_25_11.pdf?attachauth=ANoY7cqXI2rCcPH3MHm2Cl1BmsYSqqyBHi3-aYshrdfJzRsgClFYjBVs7nLtfppp8WN4S8afl7guuTUVYTr3WnxkIlA33bdd51de_811eqedyij_1oFoOkuggp8u7laa_VTPcDX6hHv doi: 10.17016/feds.2009.48

US Congress Committee on the Judiciary. (2022). Investigation of Competition in Digital Markets: Majority Staff Report and Recommendations - Part I. *House Committee Print*, *117-8*.

Vaziri, M. (2022). Antitrust Law and Business Dynamism. *2243 Cambridge Working Papers in Economics*. Retrieved from https://www.econ.cam.ac.uk/research-files/repec/cam/pdf/cwpe2243.pdf

Wang, X. (2018). Catering Innovation: Entrepreneurship and the Acquisition Market. *Working Paper*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3247274 doi: 10.2139/ssrn.3247274

Wang, Y. (2022). Competition and Multilevel Technology Adoption: A Dynamic Analysis of Electronic Medical Records Adoption in U.S. Hospitals. *International Economic Review*, *63*(3), 1357–1395. Retrieved from https://doi.org/10.1111/iere.12586 doi: https://doi.org/10.1111/iere.12586

Warg, K. (2021). The Acquisition Option and Startup Innovations. *Working Paper*. Retrieved from https://katarinawarg.com/wp-content/uploads/2021/12/Katarina-Warg-The-Acquisition-Option-and-Startup-Innovations.pdf

Watzinger, M., Fackler, T. A., Nagler, M., & Schnitzer, M. (2020). How Antitrust Enforcement Can Spur Innovation: Bell Labs and the 1956 Consent Decree. *American Economic Journal: Economic Policy*, *12*(4), 328–359. Retrieved from https://www.aeaweb.org/articles?id=10.1257/pol.20190086 doi: 10.1257/pol.20190086

Whinston, M. D. (1990). Tying, Foreclosure, and Exclusion. *The American Economic Review*, *80*(4), 837–859. Retrieved from http://www.jstor.org/stable/2006711

Wollmann, T. G. (2019). Stealth Consolidation: Evidence from an Amendment to the Hart-Scott-Rodino Act. *American Economic Review: Insights*, *1*(1), 77–94. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aeri.20180137 doi: 10.1257/aeri.20180137

Yin, P.-L., Davis, J. P., & Muzyrya, Y. (2014). Entrepreneurial Innovation: Killer Apps in the iPhone Ecosystem. *American Economic Review*, *104*(5), 255–259. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.104.5.255 doi: 10.1257/aer.104.5.255

Yu, S. (2020). How do accelerators impact the performance of high-technology ventures? *Management Science*, *66*(2), 530–552. doi: 10.1287/mnsc.2018.3256

Chapter 2

How Do Online Product Rankings Influence Sellers' Pricing Behavior?

Abstract

Products that are displayed more prominently on e-commerce platforms are more likely to be found and purchased by consumers. The algorithms ranking these products, however, may condition a product's position on a listings page on its price. Using web-scraped data from hotels displayed on Expedia and an instrumental variable identification strategy, I find that the ranking algorithm tends to display hotels at *less* favorable positions at times at which they are priced *higher*. I provide a framework that employs these estimates jointly with demand parameters obtained from a sequential search model. I simulate a counterfactual scenario, which reveals that Expedia's ranking algorithm tends to intensify price competition between sellers compared to a random ranking. This increases consumer welfare, but reduces seller profits by decreasing prices by 1.89€ on average. My finding has consequences for two-sided platforms' optimal design of ranking algorithms: in order to foster adoption, platforms should carefully trade off benefits arising to the two sides, and consider equilibrium effects.

1 Introduction

A fundamental concern of sellers distributing goods or services online is the visibility of their products to consumers. E-commerce websites such as Amazon or Expedia provide access to broad varieties of products offered by numerous third-party sellers. Evaluating the manifold products is costly for consumers. To facilitate product search and discovery, platforms display algorithmically ranked lists of relevant products once users enter a product query. These default rankings have been found to systematically influence the products that consumers become aware of and ultimately purchase (e.g., Ursu (2018), Glick, Richards, Sapozhnikov, and Seabright (2014)).

Although the algorithms used in practice are very complex, the algorithm may make a product's positioning dependent on its price, all else equal. Trying to guide consumers towards attractive offers, platforms may find it profitable to prioritize lower-priced products by displaying them more prominently.[1] However, with sellers being aware that lower prices translate into higher visibility, such an algorithm amplifies the competitive pressure between sellers and provides additional incentives to reduce prices. In other contexts, the platform may find it optimal to do the reverse, and may steer consumers to higher-margin products so as to extract more revenue per consumer visit (Hagiu & Jullien, 2011). This weakens price competition and enables sellers to charge higher markups (Dinerstein, Einav, Levin, & Sundaresan, 2018). Rankings can therefore affect sellers' price-setting decisions and influence market outcomes in meaningful ways.

This paper focuses on these supply-side implications of product rankings, which have been left relatively unexplored by empirical literature that has mostly dealt with demand-side effects of rankings. The empirical setting studied is hotel booking services offered by online travel agents (OTAs). Those intermediaries display hotels that are algorithmically ranked by recommender systems as consumers enter a query. Most user clicks and purchases occur under these default rankings. Using web-scraped data from Expedia, I find that offering a lower price leads to a better positioning of a given hotel on the platform's listings page. I then use consumer search and transactions data to estimate a model of sequential consumer search that closely follows Ursu (2018). The estimated ranking parameters and the demand parameters then enable to perform counterfactual simulations. Preliminary results show that hotels will set somewhat higher prices if the ranking does not take into account supply-side effects, although losses in consumer surpluses remain limited. Expedia's ranking algorithm thus seems to move hotels' pricing decisions into a more competitive equilibrium with lower prices, compared to a scenario where prices do not affect visibility.

Studying the effect of ranking algorithms on market outcomes within a structural framework is interesting and important for several reasons. First, it is a well-established fact that search frictions persist to exist even on the internet, contrary to expectations. A naturally arising question is therefore how these frictions affect sellers' pricing behavior and the functioning of markets, which this article attempts

[1] One example – which is the context studied in this paper – is given by online travel agents that provide hotel booking services. These platforms explicitly state that the competitiveness of hotels' offerings (in particular, prices) are taken into account in their rankings (see Section 2 as well as Appendix 2.A.1). A further example is provided by product listings on Amazon. L. Chen, Mislove, and Wilson (2016) find that both Amazon's ranking of sellers of a given product, as well as which seller gets to be displayed prominently in the "Buy Box", is to a large degree affected by the price the seller chooses.

to shed light on. Second, this question is relevant from a public policy perspective. Online platforms essentially act as gatekeepers that control what information users view and how market participants interact. The issue of prominence in particular was at the heart of a high-profile antitrust case in Europe on Google Search[2], but our knowledge on how visibility affects market outcomes is still limited. My results also inform about the strength of competition in the market for online booking services, which was subject to antitrust investigations in several European countries.[3] Third, understanding the effect of rankings on prices is of fundamental interest for platforms themselves. Online intermediaries are two-sided markets that crucially depend on attracting both sellers and consumers to be successful. When devising product rankings, these platforms need to assess the trade-offs on both sides. While platforms can evaluate how a particular ranking affects consumers' choices through A/B-testing, those experiments on subgroups of consumers are not going to shed light on the ranking's effect on *supply side* behavior. Instead, analysis of how rankings affect pricing decision requires a structural framework that incorporates demand and supply to shed light on such equilibrium effects. Finally, the hospitality industry is an especially intriguing setting for the study of product rankings, as consumer search is of great importance in this context.[4]

On the supply side, I thus propose a framework in which the economic agents are hotels deciding what prices to set. What differentiates this framework from standard supply models is that hotels set prices in anticipation of how these decisions affect their respective positions in the ranking. If a hotel marginally increases its price, it thus faces the common "direct" negative impact on demand (conditionally on being seen), as well as an "indirect" marginal impact on demand that is driven by the price's impact on visibility. My ultimate goal is to simulate what demand and hotel prices would be like under a counterfactual ranking of hotels. This requires, first, an estimate of how price and promotion affect a hotel's position. I obtain this estimate using a linear fixed effects regression with instrumental variables. On the demand side, one needs a model that accounts for the fact that consumers are not fully informed, but face search costs, and are more likely to click and book hotels that are displayed more visibly. I use Ursu (2018)'s model of consumer search, in which each search corresponds to a click on a given hotel. Estimating consumer search costs in an unbiased way, however, is not possible with the actual, relevance-based ranking of hotels that I have at hand. I thus directly calibrate the search cost parameter to the estimates Ursu (2018) obtains under a random ranking of hotels (assuming that search costs have not changed over time), and only estimate preference parameters for hotel characteristics. Given these estimates, I can then simulate demand for hotels that are displayed in a given fashion, and conduct counterfactual experiments.

[2]See `https://tinyurl.com/ycxpwncx` for the European Commission's press release on this case (accessed 03/05/2022).

[3]In a sequence of decisions, the German Bundeskartellamt, for instance, prohibited the use of 'best price' clauses by OTAs, emphasizing the restriction of competition between OTAs that such practices would entail: see `https://tinyurl.com/4e4j446j` and `https://tinyurl.com/55pkau9u` (accessed 02/04/2023). Also see European Commission (2016) for a report on a monitoring excercise on the hotel booking sector by the European Commission and several European antitrust authorities.

[4]Travellers typically have limited knowledge of hotels and local conditions in a given city, which is the reason why intermediation has been important in the travel sector even prior to the emergence of OTAs. Given the extent of differentiation of hotels and consumer tastes, and given that hotels may face thousands of potential competitors, hotels pay a lot of attention to their visibility on these platforms. See the 2017 Position Paper by HOTREC, the European hospitality association, which emphasizes the importance of visibility in this context: `https://www.hotrec.eu/wp-content/customer-area/storage/f5282293ec286d90ba33117497c7c2c6/HOTREC-position-on-the-mid-term-review-of-the-DSM-Strategy-10-October-2017.pdf` (accessed 07/05/2019).

Two datasets are used to carry out the analysis. The first dataset was web-scraped from Expedia over two months in early 2019 by carrying out daily queries for a diverse range of travel dates. It entails hotels' ranking positions as well as their pricing and promotion decisions, and thus enables to examine the effect of hotels' pricing decisions on positions. The identification of the marginal impact of prices and promotion on the ranking rests on two components: first, hotel fixed effects are used in order to control for any unobserved hotel-specific, time-invariant factors impacting a hotel's average position in the ranking (such as a hotel's unobserved "quality" or the amount of commission it pays to the platform). From the joint variation in positions, prices and promotions for a *given* hotel in the data, I can then estimate the correlation between a hotel's price with its ranking over query and travel dates. Second, the simultaneity of hotels' price-setting decisions with respect to the ranking are addressed with the help of instrumental variables for prices.

The second dataset obtained via the Wharton Customer Analytics Initiative (WCAI) details the entire search process of actual consumers who arrive on a travel website and search for a hotel in the same cities that I web-scraped. The data reveal that most clicks and purchases occur under the default ranking, and a remarkable 70% of all users recorded in the data only see search results that have been ranked by default. I moreover find a strong correlation of position and clicks (driven both by higher relevance of more visible hotels as well as by search costs), confirming prior literature. The stream of consumers' clicks and eventual purchases, and the characteristics of hotels observed in the data, serve to estimate the demand parameters for the counterfactual scenarios.

I find a significant and economically meaningful effect of a hotel's price on its position: my estimates reveal that a one-dollar increase in a given hotel's price implies a decrease (towards less visible positions) by roughly three positions on Expedia's listings page, ceteris paribus. This effect is qualitatively robust across specifications and across different sets of instrumental variables. This result is interesting in itself: Expedia's ranking algorithm appears to intensify price competition between hotels, so that hotels would be setting higher prices in a situation without such a ranking algorithm. The results moreover indicate a positive correlation of a hotel's rank position and its sales promotion decision (i.e., a sales promotion associated with being ranked more visibly), although a clear causal link cannot be established, likely due to a lack of sufficiently strong instruments. The consumer side results are qualitatively in accordance with Ursu (2018). Given the estimates on both the consumer side as well as the ranking side, one can back out hotels' marginal costs from the hotels' optimal pricing condition. I then consider exogenous perturbations to variables such as the ranking's elasticity with respect to prices, and simulate how this affects market outcomes. I find that hotels tend to set lower prices if prices are not influential for their respective rankings.

Literature. This paper relates to previous research dealing with platform design, consumer search, and the industrial organization of the market for online booking intermediaries. Over the past few years, there have been a few notable contributions that study supply-side effects of platform and search design in online contexts. Like myself, these contributions estimate demand models that can account for search frictions or the fact that consumers only consider a subset of available products.

Dinerstein et al. (2018) analyze the effects of a search design change on eBay that essentially decreased users' search costs of finding the cheapest product. The authors consider sellers of a very homogenous product, and estimate that the re-design led to a decrease in sellers' prices and markups. In contrast, in my setting, hotels are both horizontally and vertically differentiated, and I focus my attention on the ranking algorithm, as opposed to the search design more generally. Further, Lee and Musolff (2021) find that Amazon tends to make demand more elastic and thus intensifies price competition. This, however, comes at the expense of lower entry by third-party sellers in the long run. Teng (2022) studies a change in the search algorithm of Apple's App Store, which led to reduced prominence of Apple's own apps. Instead of prices, the author analyzes the updating frequency of mobile applications, and finds that the average app quality would be higher if Apple restrained from self-preferencing its own apps on its platform.[5] Ershov (2022) finds that a change in the search design of the Android app store triggered new entry (and thus more variety), but led to congestion externalities that dominated consumers' gains from variety. In contrast to the aforementioned authors, I consider a market in which prices seem to be sellers' main strategic variables, whereas entry tends to be stable.[6] Bar-Isaac and Shelegia (2022) take on the platform's perspective, and theoretically study the trade-offs a platform faces when deciding between allocating visibility to sellers by using an auction, or alternatively by using an algorithm, in different contexts. Further related papers investigate sellers' obfuscation strategies online (Ellison & Ellison, 2009), and the design of peer-to-peer platforms where matching is central (Fradkin, 2017).

Next, this research is related to a number of papers estimating structural models of consumer search in diverse settings, under different assumptions on search behavior (fixed sample search or sequential search), and with different types of data (aggregate or individual-level). The most relevant papers among these use consumer search data from online hotel platforms (Y. Chen & Yao, 2016, De los Santos & Koulayev, 2017, Koulayev, 2014, Ursu, 2018), and have found that search costs in the hotel search context are very significant, with important implications for hotels. The sequential search model used on the demand side of this model closely follows Ursu (2018). Search costs in this model are essentially defined as the costs of clicking on a given hotel, and are assumed to increase in positions.

Finally, this paper is related to studies focusing on competition issues on online hotel booking platforms. A paper that equally focuses on hotels' visibility on these platforms is a study by Hunold, Kesler, and Laitenberger (2020), and explores hotels' pricing decisions across booking channels and its impact on hotel rankings. The authors build a model showing that a ranking that maximizes an OTA's short-term profits is not necessarily in accordance with the ranking that maximizes the match value of consumers. An empirical finding in this paper is that setting a lower price on a hotel's *own* website will lead to a worse position in the OTA's rank, as it decreases the hotel's booking likelihood. By employing the ranking strategically, a platform can thus discipline hotels in their price setting decisions. My work is complementary to their study. Anecdotal evidence points out that prices *per se*, not only price differentials between booking channels, influence rankings. Having access to a dataset that details actual search behavior further allows me to account for the demand side and study counterfactual scenarios.

[5]In addition, both Lee and Musolff (2021) and Teng (2022) also study the prevalence and the welfare implications of self-preferencing on Amazon and on the Apple app store, respectively.

[6]Whereas hotels can in practice de-list from online platforms, this tends to happen extremely rarely. Hotel bookings have been commonly mediated by third parties even before the advent of the internet.

Further related papers study hotels' price-setting behaviors (Cho, Lee, Rust, & Yu, 2018, Li, Netessine, & Koulayev, 2017), buyer substitution across purchasing channels (Cazaubiel, Cure, Johansen, & Vergé, 2020), or the effect of online platform price parity clauses (Hunold, Kesler, Laitenberger, & Schlütter, 2018, Mantovani, Piga, & Reggiani, 2021). A paper that focuses on buyer substitution between hotels and Airbnbs and localized competition is Schaefer and Tran (2020).

Roadmap. Section 2 of this paper presents background information about the online hospitality sector and OTAs' rankings of hotels. Section 3 contains the model for the supply side, and Section 4 details the datasets used and provides descriptive statistics. The empirical strategy will be presented and discussed in Section 5, with the results presented in Section 6. Section 7 models counterfactuals. Section 8 concludes.

2 Background

The hotel sales on all major OTAs[7] take place under the so-called "agency model". Under this business model, the supplier (i.e., the hotel) sets the final price of the product that is sold through the intermediary. The platform merely mediates the transaction between the hotels and consumers by facilitating product search, by giving hotels exposure to potential customers, and by processing the payment.[8] OTAs receive a fixed ad-valorem commission rate for each purchase that is carried out. On Expedia, the standard rate is 15%, but can privately re-negotiated and tends to be somewhat lower for large chains (Cazaubiel et al., 2020). Industry reports quote rates between 10% and 25%.[9]

Consumers arriving on an OTA make a query by entering a city, the number of travellers, and arrival and departure dates, and are then confronted with what I call a "listings page". On Expedia, consumers then view an ordered list of typically 55 hotels per page that are ranked by default according to what Expedia calls a "Recommended" ranking. Depending on the number of hotels available at the given location, the ranked hotels can extend to several, possibly hundreds of pages. On the listings page, consumers can sort or filter[10] hotels, and browse to further pages. Clicking on a listed hotel takes a consumer to what is referred to as a "hotel page" with more detailed hotel and pricing information.

While OTAs do not disclose the exact functioning of their possibly highly complex default ranking algorithms, it is widely believed to be based on relevance. Moreover, Expedia states that a hotel's price competitiveness (current price, active promotions) are factored into the ranking order (see Figure 2.10 in the Appendix). Hunold et al. (2020) find evidence that the OTAs Expedia and Booking.com actively demote hotels if they set lower prices on their own or competing OTAs' websites.

Hotels' main strategic variable seems to be the nightly rate for a room, which is displayed very prominently on OTAs' listings page. Second, hotels are able to offer promotions on OTAs. This can be a discounted price for a given travel date and room type, or free extra services (such free cancellation

[7] For example, on Booking.com, Expedia, or HRS.

[8] At least some OTAs also provide revenue management and pricing or promotion tools to hotels.

[9] See, e.g., `https://prenohq.com/blog/ota-commission-rates-expedia-booking-com-more/` (accessed 02/05/2021).

[10] Sorting refers to sorting by price or by guest rating, for example. Filtering means that only hotels of, for example, a certain star rating are displayed.

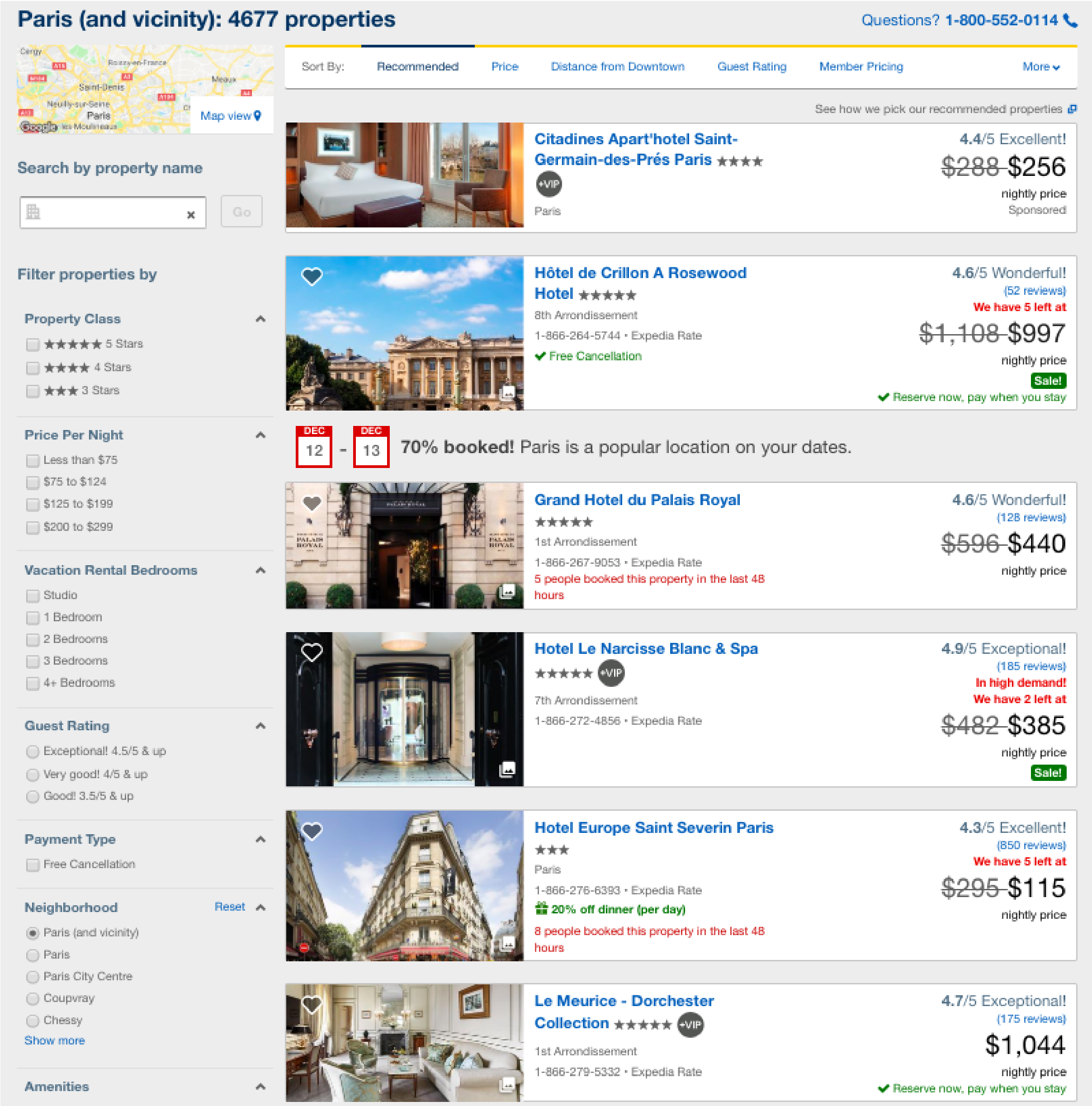


Figure 2.1: Exemplary screen shot of Expedia's listings page for hotels in Paris, as of February 2019.

or a rebate for meals). What is visible in Figure 2.1 is that many hotels moreover show both a price in black besides a grey strikethrough price. As of 2019, however, the strikethrough price was no discount given by the hotel as long as there is no "Sale!" flag.[11] Therefore, in this paper, I call a "sale" whenever a listing is flagged with a green "Sale!" flag (see Figure 2.1), and thereby displayed in a somewhat more salient way.[12] The price is the key explanatory variables in my analysis, and the key strategic decision my model incorporates.[13] More facts and insights on hotels' promotional decisions can be retrieved in Appendix 2.D.3.

[11]Instead, the strikethrough price is, according to Expedia, a comparison price, namely "the third highest price for this room type at this property (with the same length of stay and cancellation policy) that customers have found on our site during a 30 day window around your selected check-in date".

[12]See Expedia Partner Central for more info (last accessed 27/02/2020). The discount in percent that is offered by a given hotel is viewed when hoovering with the cursor above a hotel's price or sale flag.

[13]The sale flag indicator is considered in some specifications of the ranking analysis.

Finally, the listings page also features advertisements run by Expedia TravelAds (see first listing in Figure 2.1, which says "Sponsored"). Hotels displayed as such sponsored listings pay for every click that occurs and can even specify to target a particular audience. I will later exclude these sponsored listings from my analysis, and thus focus on the "organic" ranking that is determined by the algorithm only (see Appendix 2.C.2).

3 Model

The centerpiece of the model is a hotel's pricing decision. The hotels' first order conditions show that optimal prices depend on the magnitudes of two effects: the common direct effect on demand (i.e., as hotels raise prices, demand decreases, *ceteris paribus*), and, secondly, the indirect effect on a hotel's visibility (propagated via the ranking algorithm) on demand. On the consumer side, I take as given that consumer search follows the sequential search framework detailed in Ursu (2018). I do not explicitly model a platforms' decision of how to design the ranking algorithm, but discuss the tradeoffs and possible incentives that platforms face in Section 3.1.

3.1 Supply Model

I consider J differentiated hotels in a given market that sell their rooms on an intermediary's website. The intermediary charges an ad-valorem commission fee $\tau \in (0,1)$ for each purchase it mediates. On a given query date q and for travel date t (with $t \geq q$), hotels have marginal costs c_{qjt}.[14] Taking as given the pricing decisions of their competitors, at query date q, hotels set prices for room bookings for date t, such that the resulting decisions form a Bertrand Nash Equilibrium among all hotels. Hotel j's profit is (as always) the product of the markup and the demand. What is new is that demand for hotel j in turn is not only a function of the hotel's pre-promotion price p_{jqt}, but also of the position r_{jqt} which the hotel is by default displayed in on the OTA's results page. By employing a demand model in which demand depends on a product's position, the full model accounts for the stage in which consumers search for products. Products with a worse positioning in the ranking (i.e., a higher r_{jqt}) are less likely to be found by consumers, resulting in lower demand.[15]

Let $\mathbf{p_{qt}}$ and $\mathbf{r_{qt}}$ be the vectors of prices and rankings for all J hotels in the market. Hotel j's profit maximization problem writes

$$\max_{p_{jqt} \in \mathbb{R}+} \Big(p_{jqt} \cdot (1-\tau) - c_{jqt}\Big) M s_{jqt}\big(\mathbf{p_{qt}}, \mathbf{r_{qt}}(p_{jqt})\big).$$

Next, define $\tilde{c}_{jqt} \equiv \frac{c_{jqt}}{(1-\tau)}$. Deriving the first order conditions using with respect to price (using the

[14] These marginal costs are likely composed of the opportunity costs of not selling a room for a given night. See discussion in Section 7.2.

[15] In an alternative version of the model considered in Appendix 2.B, I extend the model such that hotels not only choose prices, but also discounts.

chain and the product rule) and re-arranging, one obtains:

$$p_{jqt} = \tilde{c}_{jqt} + -\frac{s_{jqt}(\cdot)}{\underbrace{\frac{\partial s_{jqt}(\cdot)}{\partial p_{jqt}}}_{\text{direct } (<0)} + \underbrace{\frac{\partial s_{jqt}(\cdot)}{\partial r_{jqt}} r'_{jqt}(p_{jqt})}_{\text{indirect } (<0)}} \tag{2.1}$$

When deciding on prices, hotels thus take into account two types of effects on demand. The "usual" (*direct*) demand effect $\frac{\partial s_{jqt}(\cdot)}{\partial p_{jqt}}$ expresses that, other things equal, a lower price will lead to higher demand. Second, prices and promotions affect the average position r_{jqt} which a given hotel is displayed in, which in turn affects how many users become aware of the hotel, click on it, and purchase it. Thus, the *indirect* effect $\frac{\partial s_{jqt}(\cdot)}{\partial r_{jqt}} r'_{jqt}(p_{jqt})$ is additionally going to be taken into account by hotels.

As a result, if a platform modifies the ranking algorithm – for example, by employing an algorithm that reacts more sensitively to prices, or by ranking hotels completely independent of their prices – hotels will take this into account and set prices differently. This affects hotels' profits and consumer surplus.

Discussion of Supply Model

The above model abstracts from the fact that hotels typically distribute rooms via multiple sales channels. According to a monitoring study conducted by European Commission (2016), which reports survey results of European hotels, chain hotels sold 35% of their rooms via OTAs, whereas for independent hotels this fraction is 42% in 2016. As Hunold et al. (2020) find, OTAs' ranking algorithms in fact seem to take price differentials between the price listed on the OTA and the price listed on the hotels' or other websites into account when ranking hotels, and thus may effectively punish hotels whenever they do not provide the lowest price to the OTA's website. When setting prices on Expedia, hotels are thus likely to take into account their prices and expected demand on other sales channels. I nevertheless believe that the trade-off captured in the above model is likely to be of first order for hotels. Evidence shows that consumers are not very likely to substitute between different distribution channels (Cazaubiel et al., 2020), so that it is adequate to consider the hotel's decision of which price to post on Expedia in isolation, as the model does. Moreover, the ranking algorithms of different OTAs seem to work similarly, therefore providing similar pressure on the price.

Relatedly, the model above also assumes that all hotels are shown and booked via Expedia's default "recommended" ranking. Instead, one could think that hotels differ in the extent to which they react to the ranking: While some hotels might be highly dependent on the Expedia default ranking sales channel, others might derive most bookings from regularly returning guests or via their own website, therefore being rather insensitive to the ranking. The assumption, however, is supported by the fact that consumer search data described below indeed shows that most consumers search for hotels via the default ranking - only around 34% of consumers ever decide to sort or filter any hotels. More than three quarters of all bookings occur under the default ranking.

Moreover, hotels' pricing decisions are in reality a high-dimensional, dynamic problem (see Cho et al. (2018)). Hotels are capacity constrained and need to set prices for a range of future dates, across multiple sales channels, and for different room types. I abstract from these dynamic considerations, as they would make the model too complex for my setting.

Possibly, promotions should also be a part of the hotel's decision problem because promotions on OTAs are very prevalent and likely to influence consumers' booking decisions. On the ranking side, Expedia explicitly states that it takes promotions into account when ranking hotels (see Figure 2.10 in the Appendix). On the consumer side, models of consumer search have consistently found positive utility parameter for a "promotion" dummy, meaning that consumers seem to derive positive utility from purchasing a hotel that is on promotion. This can be due to the fact that Expedia's red "Sale!" flag makes hotels appear in a more salient way if they are on sale (see Figure 2.1). Moreover, a higher prices in the hotel market may signal higher quality, so that high-quality hotels may find it optimal to keep baseline prices high but use promotions to sell available rooms. On the other hand, it is unclear how exactly hotels use promotions in this setting: sales promotions do not correlate very much with demand (see Figure 2.5), and the different types of promotions employed on Expedia tend to be opaque (see Appendix 2.D.3). To be sure, I do consider a model version that endogenizes hotels' promotion decisions in Appendix 2.B.

3.2 Demand Model

To obtain the effect of prices on demand, I employ the sequential search framework that is used in a hotel search context by Ursu (2018). Consumers' search behavior under optimal sequential search can be described by Weitzman (1979)'s Selection, Stopping, and Choice rules. Ursu (2018) shows how to estimate consumers' search and purchase decisions jointly in a way that the parameters are consistent with those rules, with the method partly based on work by Kim, Albuquerque, and Bronnenberg (2010).

In this sequential search model, consumers can extract basic hotel information (prices, stars, rating etc.) upon being confronted with a results page listing different hotels. Based on this listing page information, consumers can thus costlessly form an expected utility v_j for each of the hotels, which is parametrized as being linear in a hotel's stars, review score, location score, price, a promotion indicator, and a brand indicator. Consumers can then incur a costly search by clicking on any listing, which will lead to the hotel page where further information can be accessed and where the hotel can be booked. Any such click will reveal a part of a user's utility for a given hotel that is ex ante random, ϵ_j. Hotels ranked further down the page (in less positions) are more costly to click on, which can be interpreted as the costs of scrolling down a page. These search costs are parametrized as $c_j(r_j) = \exp(k + \gamma r_j)$, where r_j here stands for rank of hotel j in a consumer's search, and k is a mean level of search costs. Purchasing decisions are finally made based on the realized utility after search, $u_j = v_j + \epsilon_j$.

The model assumes that consumers' search and purchasing decisions observed in the data follow a model of optimal sequential search, which can be characterized by Weitzman's (1979) optimal sequential search rules. These rules in turn imply bounds on the parameter values for consumers' valuations for price, ranking, location etc. as well as search costs. All in all, the model thus reflects how the ranking impacts demand for a given hotel.

The framework is an adequate description of how hotel search on OTAs in practice: When searching for hotels on an OTA, a user first enters a query and is confronted with an ordered list of hotels, which I call “results page” or “listing page”. While basic hotel features (star rating, price, traveller review ratings etc.) are visible from this page, a consumer can click on a given hotel’s listing to be referred to what I call a “hotel page”. On this hotel page, the consumer finds out further details about the hotel as she can, for instance, view pictures, retrieve information about amenities, or read traveller reviews.

The framework does not endogenize a consumer’s decision to refine search results (e.g., by applying filters or by sorting them), or the decision to browse through results pages, or to alter the query. Instead, the model simply poses that, after entering a query, a user is confronted with a listing page containing all the hotels she will view during her complete search sequence. For now, I am thus focussing on estimating consumers’ clicks on hotel listings, and purchases only. The model is based on the assumption that every click (which in this framework is a "search") and purchase (or "choice") observed in the dataset has been generated by an optimal sequential search process.

Setup

A user i arrives on the OTA and makes a query, specifying location, travel date, and the number of travellers and rooms. She is then confronted with a listing page containing J_i hotels, where J_i is the total number of hotels which i ever views during her complete search sequence on the OTA. Each hotel j that is displayed on the results page has a position value r_{ij}, which is simply the position on the listing page which the hotel was displayed during user i’s search. Basic hotel features can be costlessly inferred from this results page. However, as in practice, a consumer needs to click on a given hotel’s listing to discover her final valuation for the hotel. Such clicks are costly, with clicks on less prominent and therefore "worse" positions being more costly.

The utility in this model that a given user i derives from hotel j consists of three components:

- The **expected utility prior to search** that can be inferred without incurring costs, v_{ij}, for $j \in \{0, 1, ..., J_i\}$.
- The **expected utility from clicking** on hotel j, $\epsilon_{ij} \sim N(0, \sigma_j^2)$, with $\sigma_j > 0$ for $j \in \{0, 1, ..., J_i\}$.
- The **search (click) costs** $c_{ij}(r_{ij})$, for $j \in \{0, 1, ..., J_i\}$.

The total utility that user i derives from purchasing hotel j is therefore $u_{ij} = v_{ij} + \epsilon_{ij}$. Moreover, I parametrize the search costs as follows to ensure that they are positive:

$$c_{ij}(r_{ij}) = exp(k + \gamma r_{ij})$$

I expect $c'_{ij}(r_{ij}) > 0$. As the data do not contain any information that a user obtains from the hotel page, ϵ_{ij} is unobserved, which is why I assume it follows a normal distribution.

Optimal Search

Sequential search means that after a given click that has been made, the user decides to either click on more options, or to stop searching. If she stops, she will decide which of the searched products (including the outside option) to buy.

A critical role for the optimal sequential search strategy is taken on by the **reservation value** z_{ij} of consumer i for hotel j, defined as:

$$c_{ij} = \int_{z_{ij}}^{\infty} (u_{ij} - z_{ij}) f(u_{ij}) du_{ij} \tag{2.2}$$

where $f(.)$ is the probability density function of u_{ij}. This implies that the reservation value is the hypothetical utility that would make user i indifferent between searching and not searching product j, given the search costs c_{ij}. According to Weitzman (1979), the optimal search strategy can be characterized by three simple rules.

1. **Selection Rule:** The options should be searched in descending order of the reservation utility.

2. **Stopping Rule:** The consumer should stop searching when the highest utility obtained so far is larger than any reservation value of the unsearched options.

3. **Choice Rule:** The consumer should choose the option that yields the highest utility (including the outside option).

Thus, the selection rule defines the order of the searches, while the stopping rule defines the length of the search. The rules imply a number of inequalities concerning the relationship between reservation values and utilities for the products: Using Ursu (2018)'s notation, assume that a user i searches a total of s hotels. Let $R_i(n)$ denote the identity of the hotel with the n-th highest reservation utility, and thus the n'th hotel that was searched. Thus $R_i = [R_i(1), .., R_i(n), ..., R_i(s)]$ is the set of searched hotels and the order in which they were searched. Moreover, let $R_i(0)$ and $j = 0$ denote the outside option.

From Weitzman (1979)'s selection rule, we know that, given that user i makes her n'th search, she will optimally pick the hotel that has the highest reservation utility out of all those hotels that have not been searched yet:

$$z_{iR_i(n)} \geq \max_{k=n+1}^{J_i} z_{iR_i(k)} \qquad \forall n \in \{1, ..., J_i - 1\}$$

From the stopping rule, one obtains two separate inequalities. First, user i will make an n'th search when the reservation utility of the product searched in the n'th search exceeds the utility that was revealed from all other searched products (including the outside utility):

$$z_{iR_i(n)} \geq \max_{k=0}^{n-1} u_{ik} \qquad \forall k \in \{0, ..., n-1\}$$

Second, given that user i searches s products, it must be that all hotels that are not searched have a reservation utility that is lower than the maximum of the utility of all searched alternatives, including the outside option:

$$z_{iR_i(m)} \leq \max_{k=0}^{s} u_{iR_i(k)} \qquad \forall m \in \{s+1, ..., J_i\}$$

Last, the choice rule implies that the product that is ultimately chosen must yield a larger utility than any of the other searched options, including the outside option:

$$u_{ij} \geq \max_{k=0}^{s} u_{iR_i(k)} \qquad \forall j \in R_i \cup \{0\}$$

These four inequalities define the probability that a given user i searches in order R_i and purchases product j, and put restrictions on the values for the utility parameters. Given the observed searches and choices for all users, one can derive the joint likelihood. Subsequently, one can estimate the utility parameters and the effect of position on search costs using simulated maximum likelihood estimation[16].

To precisely pin down the mean search costs k, another expression is needed. As Kim et al. (2010) show, from the definition of the reservation utility, one can obtain the following expression:

$$\frac{c_{ij}}{\sigma_j} = \left(1 - \Phi\left(\frac{z_{ij} - v_{ij}}{\sigma_j}\right)\right)\left(\frac{v_{ij} - z_{ij}}{\sigma_j} + \frac{\phi\left(\frac{z_{ij}-v_{ij}}{\sigma_j}\right)}{1 - \Phi\left(\frac{z_{ij}-v_{ij}}{\sigma_j}\right)}\right)$$

where $\phi(.)$ and $\Phi(.)$ are the probability density function and the cumulative distribution function of the standard normal distribution, respectively. Kim et al. (2010) further explain that for a given c_{ij} and σ_j (which will be normalized to 1), one can obtain a unique value of $\frac{z_{ij}-v_{ij}}{\sigma_j}$. By creating a look-up table, one can thus obtain the precise value for $\frac{z_{ij}-v_{ij}}{\sigma_j}$ outside the estimation loop, which allows to compute the exact reservation utility via the expression $z_{ij} = v_{ij} + \sigma_j \frac{z_{ij}-v_{ij}}{\sigma_j}$.

Based on the restrictions described above that result from Weitzman (1979)'s rules, one can derive the probability that a consumer searches in a given order and chooses a product j by integrating over the space of values of ϵ that result in the observed pattern of clicks and choices. From there, one can form the log-likelihood, which can finally be estimated I refer to Ursu (2018) for a more detailed discussion on estimation and identification, as I am using the exact same method.

3.3 The Platform's Incentives

The platform's ranking algorithm is a crucial element of the market design that fundamentally influences both sides of the market. On the consumer side, a platform may on the one hand want to assist consumers with finding valuable bargains or high-quality matches (and thus essentially reduce search costs), thereby increasing the website's overall attractiveness in the long run.[17] On the other hand, platforms have the short run objective of maximizing revenues per visit by essentially diverting consumer search (Hagiu & Jullien, 2011). In an OTA setting, the platform may for example find it profitable to display a few pricy (and higher quality) hotels very visibly and thereby induce consumers to buy a more expensive hotel.

On the hotel side, the OTA may want hotels to compete fiercely so as to induce them to price competitively and offer promotions, which again makes consumers willing to visit the OTA's website. However, an OTA would not want to encourage competition on prices too much, as hotels might otherwise engage

[16] The ϵ_{ij} of the searched options are not going to distributed normally any more.

[17] This point is also highlighted by Dinerstein et al. (2018) when mentioning that guiding consumer search is one of two key search design objectives, the other one being that a platform may want to foster stronger pricing incentives on sellers.

in obfuscation or not even be willing to join the platform in the first place.[18] Moreover, note that the platform's profit in the above framework is the ad valorem fee multiplied by the net price of all bookings incurred, similarly to the model by Hunold et al. (2020). It is not clear whether a platform would want to display a given hotel more visibly when its price is lower than usual, as opposed to it being higher than usual: With a higher price, the platform receives a higher revenue given that the hotel is being booked; however, a higher price would also decrease the likelihood of the hotel being booked, ceteris paribus. One can therefore identify a number of effects of a search platform's ranking algorithm decision on its revenues which may counteract or enforce each other. The empirical estimates below shed some light on how these platforms' (or Expedia's, at least) rankings currently operate.

4 Data

I use two main datasets. The first dataset is consumer search and transactions data stemming from consumer queries made in 2009 on one of the major online travel agencies in the US and in the world[19], and was obtained via the Wharton Customer Analytics Initiative (WCAI). As all observations are generated by actual consumer queries, this dataset does not yield sufficient variation in the rankings and prices of a given hotel over query and travel dates that would allow to estimate the relationship of hotels' prices and rankings. I therefore additionally web-scrape data from Expedia in 2019 that gives me variation in hotels' rankings and prices across query and travel dates. Both datasets are available for four cities – Manhattan, Budapest, Cancun, and Paris –, but I focus the analysis on Paris only.[20]

4.1 Demand Side: Consumer Search and Transactions Data

The consumer search data motivates the analysis, helps to verify the importance of the default ranking, and is used to obtain utility parameters on the demand side. This dataset records the search behavior of nearly 18,000 US-based users on a major OTA, and cover all searches for hotels at four local markets (Budapest, Cancun, Manhattan and Paris) that took place on this OTA's website between October 1st and October 15th, 2009. The data amount to roughly 1.3 million observations, where each observation corresponds to a given hotel that appeared during a given user's query. For each hotel that was viewed, I observe the precise query that was entered along with time and date, whether the hotel appeared in a query for which a refinement action was taken (i.e., sorting by price), browsing behavior (i.e., flipping through pages), and whether it was clicked on (to inquire further information on the hotel page) or purchased. Both clicks and purchases include a time stamp. A "query" is defined to be a combination of request (i.e., typically location and travel dates entered), refinement action, and page.[21] A hotel is

[18] Casual empiricism indicates that airlines, for example, have appeared to compete more intensely on price during the past years, with fees such as checked luggage or meals increasingly not included in the baseline price that is displayed on search aggregators.

[19] According to the travel research website tnooz, this website was one of the most visited travel agent websites as of October 2009. Moreover, it was ranked as being the most used platform worldwide over many years. See https://www.tnooz.com/article/us-travel-site-crunch-data-week-end-october-17-2009/ (accessed 03/03/2019).

[20] A benefit of focusing on Paris is that geolocation data is especially well-covered for the subset of web-scraped hotels in Paris. In contrast, for Cancun for instance, many addresses are misspecified and the geolocation cannot be recovered.

[21] Thus, browsing to the next listing page is a new "query", as is any re-ranking or filtering action.

considered to be "viewed" if it was displayed anywhere on the current listings page.[22] The basic dimensions are displayed in Table 2.1. The data contain basic hotel characteristics such as star rating, price, a binary "promotion" variable, brand and parent company it belongs to (if any), location and name of the hotel, and which position it appeared on. All in all, the dataset gives an insight of consumers' search behaviors at a very granular level and is exhaustive compared to datasets that have been used by other empirical literature[23].

Table 2.1: Basic data summary: WCAI consumer search data

Market	Observations	Users	Queries	Hotels	Clicks	Purchases
Budapest	353,051	4,946	14,794	276	9,337	262
Cancun	270,071	4,211	11,588	110	7,580	84
Manhattan	350,183	4,197	13,772	543	6,048	111
Paris	328,926	4,424	13,204	1,637	5,787	108
TOTAL	1,302,231	17,762	53,358	2,566	28,752	565

Number of observations, users, queries, hotels, clicks, and purchases, by market and in total. Note: I only use observations from Paris in the analysis.

Table 2.2: Search activities for buyers and non-buyers (averaged over users)

	Market	# Queries	# Hotels viewed	# Clicks	# Refinements
Buyers	Budapest	5.19	51.49	5.96	0.51
	Cancun	5.67	41.55	6.53	0.49
	Manhattan	7.07	81.46	7.57	0.64
	Paris	6.33	77.32	6.08	0.73
Non-buyers	Budapest	2.86	42.12	1.71	0.29
	Cancun	2.69	35.21	1.72	0.25
	Manhattan	3.19	51.96	1.31	0.43
	Paris	2.90	49.12	1.22	0.38

Average number of searches, hotels viewed, clicks made, and refinement options chosen in each of the markets are displayed. A new "query" is made whenever any request setting is changed, or a refinement method is chosen. Note that I only use observations from Paris in my subsequent analysis.

Exploring consumers' search strategies confirms the fundamental importance of the default ranking during the search process. On average, consumers seem to have small consideration sets, and rarely flip through results pages. Table 2.2 shows that in Paris, even eventual buyers of a hotel room end up viewing less than 80 out of 1,600 potentially available hotels, and click on six of them. Refinement actions (comprising filtering and sorting) are rarely used – on average, roughly three quarters of consumers who end up purchasing take one refinement action. Only 30% of all users (and 45% of users who end up buying) ever view hotels that are not ranked by default (not displayed).

In accordance with this, Table 2.3 shows that most views, clicks and purchases occur under the default ranking. Again, this stresses the importance of hotels' position in OTAs' default ranking for their revenues. Lastly, Figures 2.2 and 2.3 suggestively point to the importance of a hotel's rank for

[22] As is common in these datasets, I am not able to take into account whether a consumer actually scrolled to a given hotel and looked at it.

[23] For example, Koulayev (2014) does not observe purchases, and Ursu (2018) does not observe refinement decisions or the order of clicks that were made.

Table 2.3: WCAI data: sorting behavior and importance of default ranking

Ranking	% of users who ever view given ranking	Number of clicks	Number of purchases
Default ranking	97.48	20,523	440
Sort by hotel name	1.52	90	3
Sort by city name	0.56	51	1
Sort by distance	12.63	3,057	34
Sort by star rating	2.89	624	5
Sort by price	17.53	4,565	70
Sort by reviews	1.81	407	12

Left column shows the percentage of users who ever see search results that are ranked to a given ranking, using data from all four cities. (Since many users see the default ranking and may take refinement actions in addition to that, it is natural that the percentages for users does not add up to 100.)

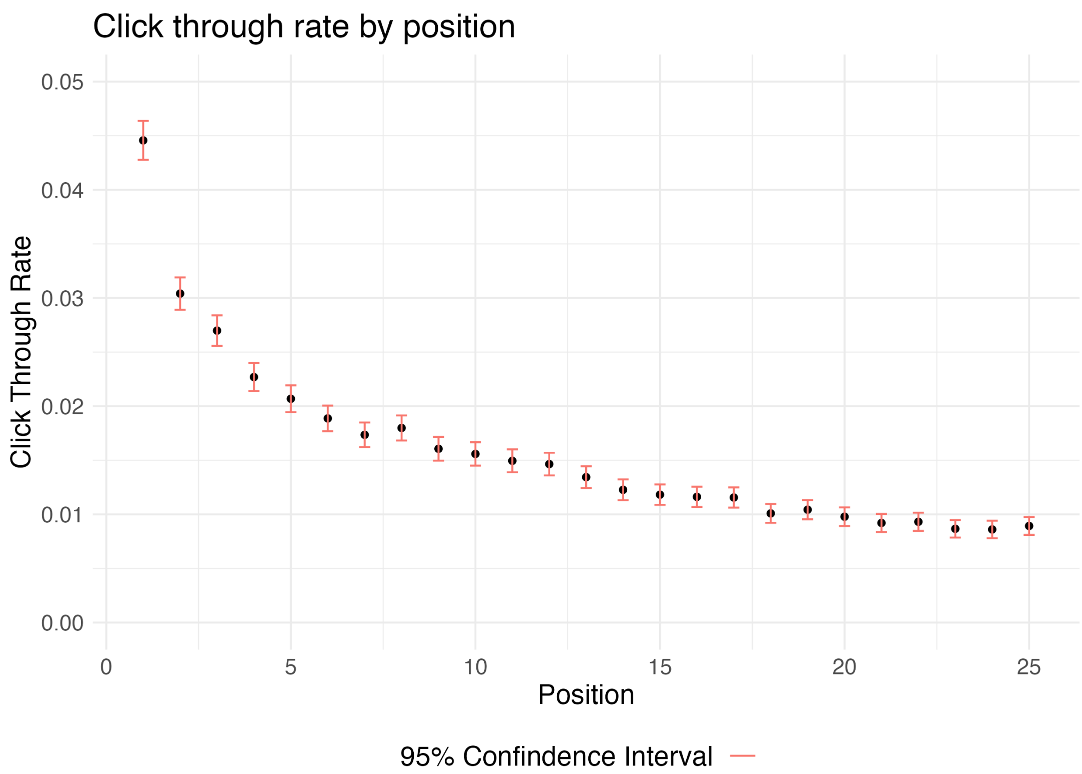


Figure 2.2: Average click through rate by position on a given results page. I exclude searches in which fewer than 20 hotels were displayed on a given page (which would occur if strict filtering methods are used, or if the query specifies the property name). This plot uses observations of clicks and positions occurring under any ranking: when focusing when focusing on data generated by the default ranking only, or alternatively any other than the default ranking, the results look qualitatively extremely similar (but are shifted downwards for the default ranking, and upwards for other rankings).

consumers' click and purchase decisions. However, note that these figures do not take into account the fact that better-ranked hotels are likely also of higher quality, and therefore do not allow for a causal interpretation.[24]

[24]In contrast, Ursu (2018) establishes a causal link using randomly ranked search results.

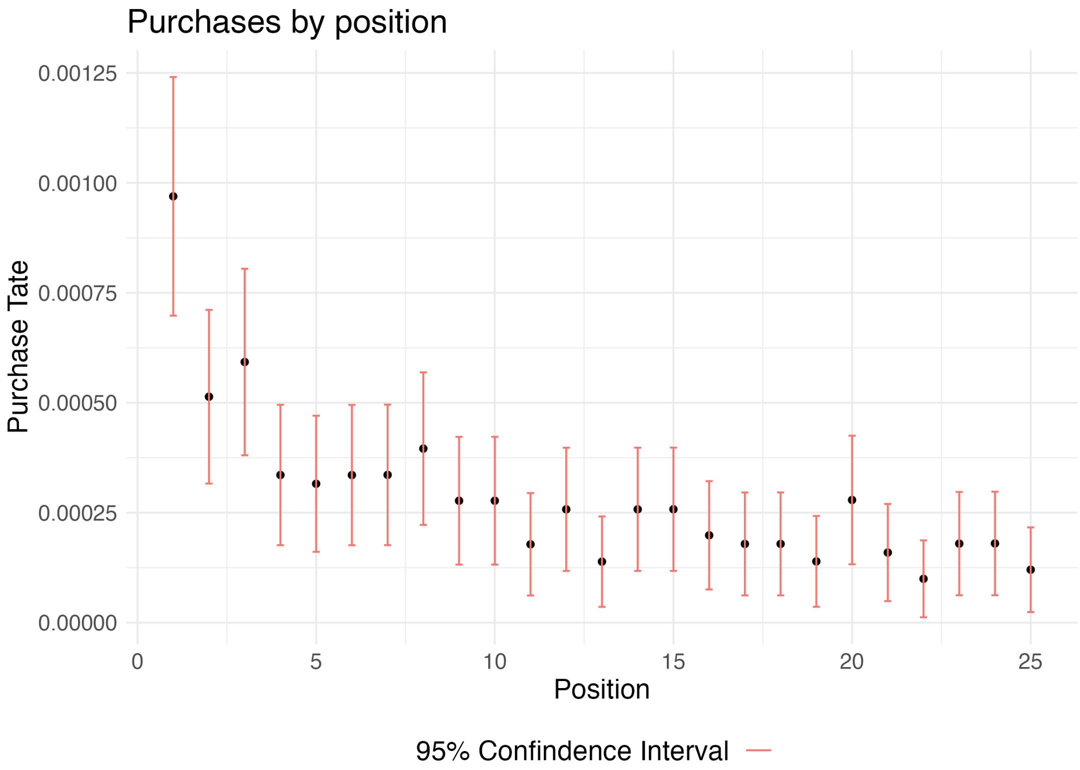


Figure 2.3: Average purchasing rate by position on a given results page. I exclude searches in which fewer than 20 hotels were displayed on a given page (which would occur if strict filtering methods are used, or if the query specifies the property name). This plot uses observations of clicks and positions occurring under any ranking: the findings look qualitatively extremely similar when focusing on default ranking. However, the downward trend becomes very noisy when considering only queries under alternative rankings, due to a low number of purchases ever occurring under alternative rankings.

Appendix 2.D.1 provides further descriptive facts on how users search for hotels online, focusing on Paris.

4.2 Ranking Side: Web-Scraped Data

The second dataset is web-scraped from Expedia, which is the dominant OTA in the US. In the months of February and March 2019, I carried out daily hotel queries for hotel stays in the cities covered in the search and transactions data (i.e., Budapest, Cancun, Manhattan and Paris). I conduct queries for trips beginning between one and 250 days after the query date (all in all, for 16 different travel dates or more each day per location), such that the latest travel dates are in November 2019.[25] During each such query, I obtain hotels' default position, pricing, promotions, and additional characteristics from all results pages that appear in a query.[26] I employ search parameters that are empirically most often being entered by consumers (see Appendix 2.D.1), and are thus the most relevant for hotels setting prices, namely weekend or single night stays for two persons in one room. All in all, the data cover over 6,000

[25] I do so to reflect the observation that users' booking windows are also very heterogeneous, as shown in the WCAI dataset, see Appendix 2.D.1.

[26] Cookies were cleared after each request. All requests were carried out using the same user agent. A few travel dates that should have been collected on a given query date are missing due to failing Internet connections. See Appendix 2.C.2 for further details.

distinct hotels which I identify as active on the platform, over 2,800 of them in Paris. As seen on the screen shot in Figure 2.1, hotels offer various extras to entice consumers (free cancellation, free offers, discounts, a "Sale!" flag, etc.). I collect all these information from the website, too, allowing me to potentially control for these factors in my analysis. In the resulting dataset, an observation is a hotel j that the "scraper" views at query date q for a potential single-night stay starting at travel date t.

The analysis focuses on hotels' decisions on which prices to set, which is a variable that seems influential for both a hotel's rank as well as the purchasing likelihood.[27] I find substantial variation in hotels' pricing decisions over time.

As seen in Table 2.4, the distribution of prices across observations seems to have a long right tail. Moreover, relatively many hotels display a "Sale!" flag at any point during the query period. Appendix 2.D.3 provides further descriptive results on hotels' promotion decisions. Figure 2.4 shows the variation in average prices across time by star rating of the hotel in Paris, and strikingly illustrates the strong seasonality in pricing. Especially those hotels that have four or five stars appear to show significant variation of average prices over travel dates, possibly reflecting changing demand patterns. Interestingly, the month of August – typically the travel month in France during which certain amenities as well as sights are closed – shows very little day-to-day variation in prices.

Table 2.5 displays the between and within variation in positions, prices, and sales in the Paris data (which are used for the estimations below). The within-hotel variation in these variables is very large: the within-variance of prices amounts to 65 euros, and the within-variance in position to 341. To illustrate this high variation within hotels further, I find that the median hotel's difference between the maximum and the minimum price across different travel dates amounts to 164 euros (not shown in table). The median hotel's difference between the maximum and the minimum rank across different travel dates amounts to almost 1,700 positions.

Table 2.4: Pricing behavior across cities

City	Active Hotels	Median price (€)	Avg. price (€)	% of hotels with "Sale" flag	% of hotels ever with "Sale" flag
Budapest	1,045	71	277.03	29.05%	36.17%
Cancun	972	85	180.02	18.11%	23.77%
Manhattan	1,222	259	300.87	18.80%	24.55%
Paris	2,842	135	168.09	28.55%	48.63%

Averages are computed over all query-travel date observations in each of the cities.

Figure 2.5 shows the fraction of hotels with a "Sale" flag at any travel date, separately for hotels of different star ratings.[28] Note that for observations in the month of February and March, the booking window is very short, as scraping took place in those months. In those months, the fraction of hotels with sales offers is very low across all star ratings, amounting to 10-20%. The fraction of hotels on sale subsequently grows and stabilizes from May onwards, and varies considerably between hotels of different star ratings: interestingly, both 1- and 5-star hotels rarely offer sales. In contrast, 4, 2, and 3-star

[27] Endogenizing a hotel's decision to offer a sale is difficult due to the opacity of the types of sales that can be offered, and a lack of understanding how hotels make sales decisions or how pricing and sales decisions interact; see the discussions in Appendices 2.B and 2.D.3.

[28] Note that 36% of hotels in Paris do not have any star rating and are thus excluded from the plot.

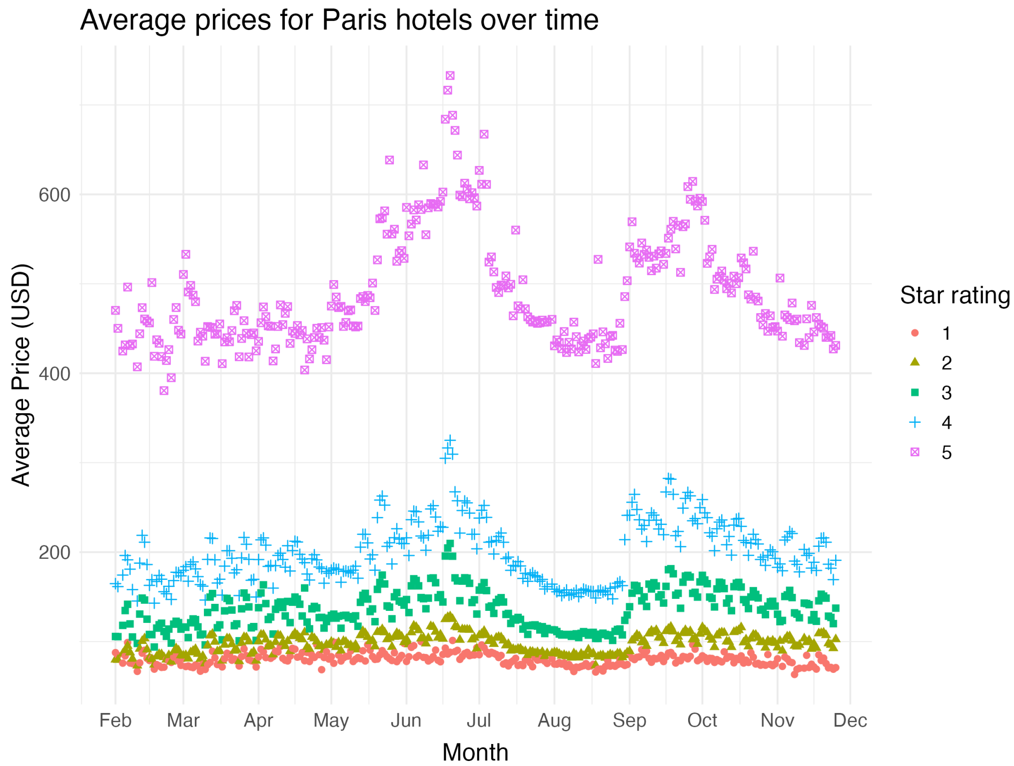


Figure 2.4: Prices of Paris hotels over travel dates.

hotels offer sales more often, amounting to up to 35% of observations for 3-star hotels. As pointed out previously, promotions on Expedia are relatively opaque, with the "Sale" flag possibly having different meanings that can only be apparent when hovering above it with a mouse, and thus potentially requiring greater effort by users. These promotions are studied in greater detail in Appendix .

Table 2.5: Within and Between Variance, Paris data

Variable		Mean	Std.Dev.	Min	Max	Observations
Position	overall	1020.6	608.6	1	2466.0	N = 1,202,289
	between		604.2	44.5	2396.8	n = 2,842
	within		340.7	-897.0	3006.4	T-bar = 423.0
Price	overall	168.1	271.9	27	13768.0	N = 1,202,289
	between		355.1	39.33	12460.1	n = 2,842
	within		65.3	-3020.1	12079.8	T-bar = 423.0
Sale	overall	0.286	0.451	0	1	N = 1,202,289
	between		0.346	0	1	n = 2,842
	within		0.271	-0.713	1.283	T-bar = 423.0

Ursu (2018)'s dataset and her specification for consumer indirect utility includes a hotel's location score (with better location being equivalent to a higher score), which is likely to be important given that location is likely to be a major dimension for consumers when booking hotel rooms. The scraped data does not include any such score; therefore, I construct a location score by computing hotels' distances to the Louvre museum (which I take as the center of the city) and then assign these distances to discrete

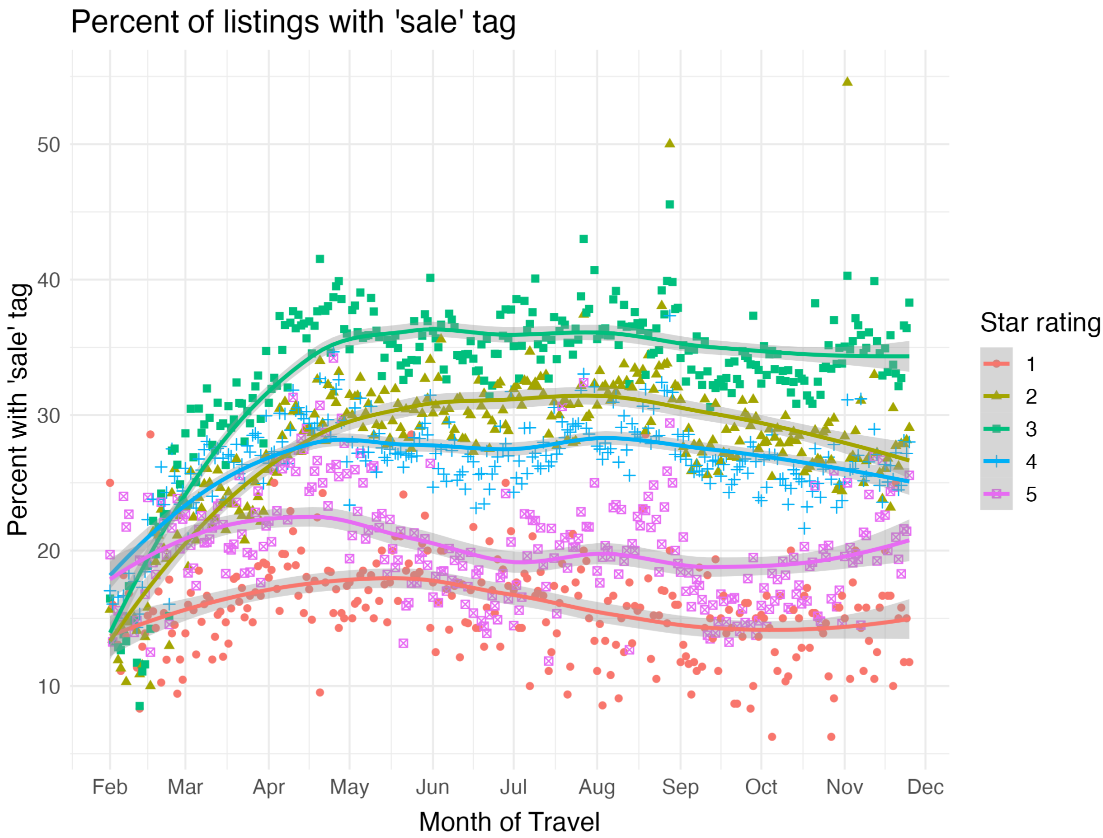


Figure 2.5: Average percent of listings with "Sale" flag at a given travel date for hotels in Paris. Locally estimated scatterplot smoothing is used for plotting the smoothed lines. Note that for observations in February and March, the booking window is very short. The plot looks very similar when looking at the average percentage of hotels offering any percentage discount greater than 0.

location scores that mirror the one of Ursu (2018) dataset.

Appendix 2.C.3 compares the WCAI to the web-scraped dataset.

4.3 Further Datasets

I obtain further datasets that allow to construct the instrumental variables. First, hotels' addresses are web-scraped from Expedia as well. I then use the Google Maps API to obtain hotels' geolocation (latitude and longitude). I find that precise geolocation data are available for 82% of the hotels in Paris (and for 94% after removing hotels that I identify as "apartments", as opposed to actual hotels).

Second, I obtain publicly available data on 58,067 Airbnbs in Paris from the website `insideairbnb.com`. For each Airbnb, the data contain information on the Airbnb's zip code it is located in, as well as price and availability from beginning Feburary 2019 for the following 365 days.[29]

5 Empirical Strategy

[29] I drop Airbnbs that have not had any booking for the past 12 months, as those seem to be inactive, and end up with 37,885 Airbnbs located in 50 zip codes. I then construct average prices at the zip code travel date level.

5.1 Effect of Prices on Ranking: Linear Model with Fixed Effects

The first step of the analysis is to estimate the effect of a given hotel's price on that hotel's position in the ranking, denoted $r'_{jqt}(p_{jqt})$ in the theoretical model of Section 3. To capture this relation in a high-level sense, I employ a linear model with hotel and query-travel date fixed effects. I use observations of hotels, their pricing decisions, and their respective rankings observed in the web-scraped data to identify the parameters of interest. The equation of interest writes:

$$r_{jqt} = \alpha p_{jqt} + \beta d_{jqt} + \delta past\_book_{jq} + \gamma_j + \lambda_{qt} + \epsilon_{jqt} \tag{2.3}$$

where r_{jqt}, p_{jqt} and d_{jqt} indicate hotel j's average position, price and an indicator variable for specifying whether a hotel is on sale[30] for travel date t and query date q. The coefficient of interest is the price coefficient, α, which can be viewed as an estimate for $r'_{jqg}(p_{jqg})$.[31] $past\_book_{jq}$ is a variable counting the number of bookings (for any travel date) that have been made within a 48-hour window prior to scraping a given hotel hotel and that is displayed on some listings, as seen in Figure 2.1. It serves to control (to some extent) for demand shocks that a hotel may experience and that the ranking algorithm may pick up. The promotion dummy d_{jqt} allows to control for whether or not a hotel's listing has a "Sale!"-flag, which may affect the hotel's ranking. Hotel fixed effects γ_j allow to control for any unobserved hotel characteristics that are fixed over time. Employing these hotel fixed effects is crucial: Expedia almost surely takes into account characteristics of a given hotel that are unobserved to the researcher, such as any unobserved "quality", or factors that affect the business relation between the hotel and the platform, such as the amount of commission the hotel pays. As long as commissions or quailty do not vary over time – and this is unlikely to be the case for the 2-month period during which the web-scraping took place –, the fixed effects can account for them. The focus of this analysis is therefore on how a *given* hotel's average ranking on Expedia – as observed by the web-scraper making the query (and which proxies for the average customer) – reacts to the price it sets, holding the hotel's identity fixed. λ_{qt} is a query date-travel date fixed effect accounting for seasonal demand patterns, for instance.

Endogeneity and Instrumental Variables

The hotels' positions, prices, and sales decisions observed in the data are equilibrium values that are simultaneously determined by both, a hotel's reaction to the behavior of the ranking algorithm, as well as the algorithm's elasticity with respect to hotels' prices. As a result, the variables of interest in Equation (2.3) are endogenous, and any ordinary least squares estimation would yield biased estimates.[32] I use instrumental variable techniques to solve for this endogeneity concern. Valid instruments are variables that exogenously shift price or promotion decisions of a given hotel across booking or query dates, and that are orthogonal to anything that may influence a hotel's position over time. Instruments for prices commonly used in the empirical IO literature are based on firms' marginal costs. However, hotel

[30] As indicated by having a "Sale!" flag, as shown in Figure 2.1. The sale flag typically denotes a percentage reduction in the price. See Appendix 2.D.3 for a discussion.

[31] If we were to endogenize hotels' promotion decision as well, we would also be interested in β.

[32] Under reasonable assumptions, one would expect the bias to be negative.

prices vary strongly even across days of a week, which cannot reflect marginal costs but rather changing demand across travel dates. Hotels' pricing policies are more likely driven by *opportunity* costs, and are thus higher for high-demand, and lower for low-demand days.

To understand hotels' pricing decisions in greater detail, I conducted an interview with a representative of the German hotel association (*Hotelverband Deutschland*) in February 2020. I was told that hotels often employ revenue management software. This software suggests prices and takes into account competitors' prices as well as major events taking place in the neighborhood, and provides a rationale for employing the below instruments. Overall, I provide estimates using different sets of instruments, each of which have their advantages and disadvantages.

Neighborhood Instruments. As a first instrument, I use the prices and sales indicator variables of hotels that are located within 500 meters distance of a focal hotel. The rationale why this should be a relevant instrument stems from both the anecdotal evidence from my abovementioned interview with the German hotel association, as well as academic literature. Li et al. (2017) find that hotels seem to closely match the prices of those hotels which they believe to be close competitors (i.e., hotels that are located in the same neighborhood and/or have the same star rating as the focal hotel). Cho et al. (2018) equally find that empirically, competing hotels' prices strongly co-move. Moreover, hotels located close to each other and possibly having the same star rating are likely to experience common demand shocks, for instance because of an event happening in that neighborhood, or due to changing patterns in business and leisure travellers. Indeed, Schaefer and Tran (2020) find a higher substitutability between hospitality businesses located close to each other in the same district, and emphasize the importance of localized competition between hotels. Building up on these empirical facts, the logic of this neighborhood-based instrument is therefore that prices of neighboring hotels constitute a measure of local, short-term demand. Prices of neighboring hotels should thus be correlated with the focal hotel's price.

Using this set of instruments, the identification rests on the assumption that local demand conditions influence hotels' price setting decisions, but that hotels have better knowledge about these conditions than Expedia does, so that local demand conditions do not affect the hotel's rank.[33] The instrument is not valid if Expedia was able to anticipate future bookings of consumers as well as hotels, and thus display hotels with a high demand shock – and thus at a time at which their prices are high – more visibly.

Airbnb Instruments. The next instrument is based on an intuition that is similar to the one of the neighborhood instruments. First, prices of Airbnbs that are located in proximity (here, in the same zip code) as the focal hotel are likely to be correlated with the focal hotel's price, but not with the focal hotel's position. To identify both the price as well as the sales parameter, I additionally employ the interaction of an indicator variable indicating certain weeks in August in which hotel prices are not very dispersed

[33] One might also think that, during times of high demand, hotels set prices to match each others' rates and have their rooms occupied at high prices, and might be less concerned about their ranking at those times. Also, note that over 2,500 active hotels are listed on Expedia for Paris that all appear in the ranked list, and hotels are capacity constrained and can only be allocated into one of the ranking's slots, so that there must be some randomness in how hotels are being ranked.

(see Figure 2.4) with the zip code.[34] The argument for the exogeneity of this instrument is similar to the one of the neighborhood instrument.

The Airbnb data do not contain any variation of prices across *query* dates. Therefore, I create a "panel" of hotels by collapsing multiple observation of a given travel date (stemming from multiple queries made for that travel date) to a *single observation per hotel per travel date*. The analysis is thus performed on the average values of price, position etc. for a given hotel and travel date, with averages computed over query dates. Denoting $\bar{p}_{jt}$ the average of p_{qjt} over all query dates q (and analogous for the other variables), I thus estimate the parameters of the following model:

$$\bar{r}_{jt} = \alpha^{avg\_q}\bar{p}_{jt} + \beta^{avg\_q}\bar{d}_{jqt} + \delta^{avg\_q}\overline{past\_book}_{jq} + \gamma_j^{avg\_q} + \lambda_t^{avg\_q} + \eta_{jt} \tag{2.4}$$

A drawback of this instrument is that the variation in Airbnb prices across travel dates less strong. As Figure 2.6 shows, while there is some seasonal variation across months, price variation for Airbnbs is still substantially smaller than for hotels. Moreover, I find that Airbnbs tend to be priced higher on Fridays or Saturdays, whereas hotels are priced highest on Tuesdays and Wednesdays, reflecting demand coming from different types of customers.

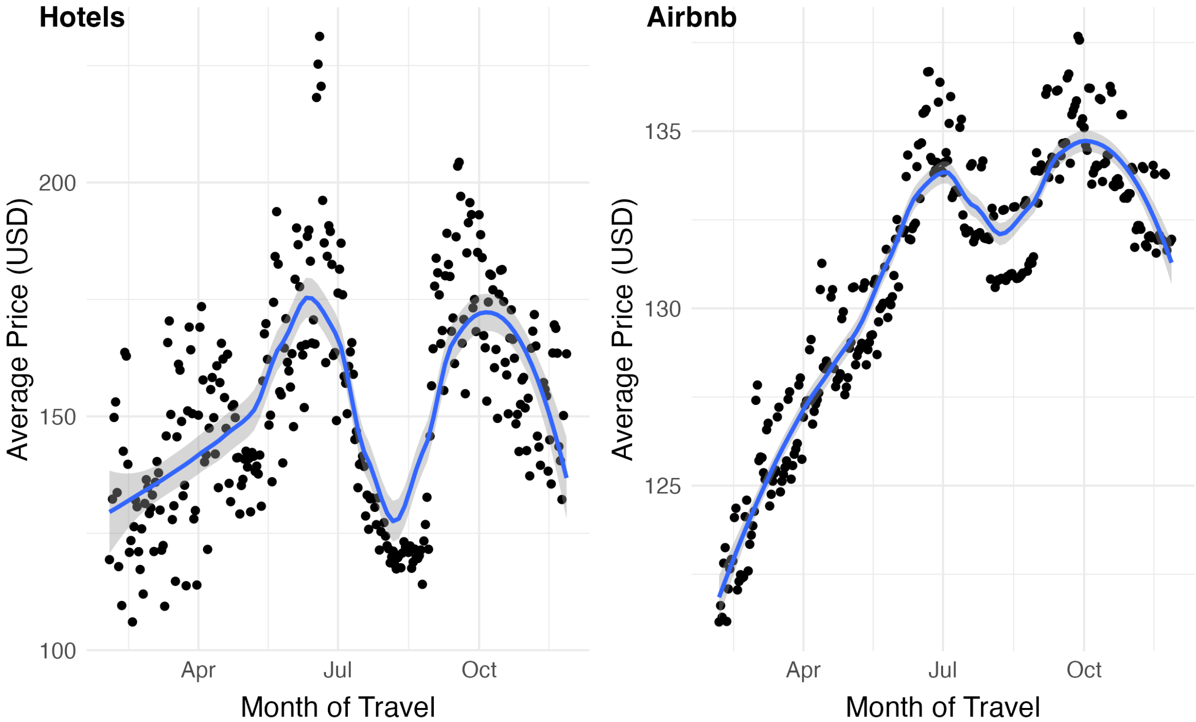


Figure 2.6: Average prices of hotels and Airbnbs over travel dates. Locally estimated scatterplot smoothing is used for plotting the smoothed lines. Beware that the y-axes have different scales. I drop 5-star hotels to reduce noise.

Brand-based Instruments. An observation that has been made in other market settings is that prices

[34] Anecdotally, August is a holiday month in France, with many attractions and even amenities in Paris being closed. Based on Figure 2.4, it does seem like competition between hotels works in some sense differently during those months. In the past, I have also employed indicator variables for certain well-known events (public holidays or the end of the Tour de France for Paris, for instance) happening in the city as an additional instrument, interacted with the zip code. As these instruments tend to not be very relevant, I have excluded them in my current version of the results.

tend to be set relatively uniformly across outlets of a certain brand. DellaVigna and Gentzkow (2019) show for US groceries and drug store chains that prices within a given brand show substantially less variation than prices between brands, even though consumer demographics may vary widely between regions. My web-scraped dataset contains observations of 39 different hotel brands with at least two outlets for Paris alone. I find that the variation of prices and sales within a given brand is much smaller than variation of prices between brands. I therefore use prices and sales indicator variables of hotels of the same brand as instrument for the focal hotel's pricing and sales decision, as these are correlated with the focal hotel's decisions, but should not causally influence the focal hotel's position in the ranking. The identification assumption is that pricing and sales decisions are being made on an at least city-wide level, and therefore less in response to the ranking. A drawback of this instrument is that its application is of course limited to those hotels which indeed are brand hotels, therefore accounting for a subset of only 25% of observations.

Dynamic Panel Instruments. A technique that has been employed in the literature especially for production function estimation (see Arellano and Bond (1991)) is the use of dynamic panel instruments. To use these instruments, I am again aggregating the hotel-query date-travel date panel, so that each observation corresponds to one hotel-travel date observation averaged across queries. Then, given J large and T small, under the assumption that $\mathbb{E}[\epsilon_{jt}] = \mathbb{E}[\epsilon_{jt}\epsilon_{js}] = 0$ for $t \neq s$ (i.e., absent serial correlation in the error terms), one can use the lagged difference of hotel prices as instrument for the current prices.

5.2 Demand Side: Utility Parameters

Ursu (2018) uses data from randomly ranked hotels (where a hotel's quality is independent from its position in the ranking) in order to estimate the relevant parameters. Since the listings pages observed in the WCAI search data are not randomly ranked, I employ the estimated search cost parameters from Ursu (2018) - denoted $\hat{k}$ and $\hat{\gamma}$ in her paper - to avoid any bias due to the relevance-based, default ranking. Thus, I will not estimate the search cost parameters k and γ, but will only employ her model for estimating the utility parameters of hotel observables such as stars or location score.[35] Just like Ursu (2018), I also do not instrument for prices, and instead assume that conditional on a given entered query, the observed price variation is unlikely to be correlated with the error term in the utility, and mostly captured by travel date and query characteristics. I refer to Ursu (2018) for details of the log-likelihood estimation. Overall, the demand side estimates enable me to simulate consumers' purchasing choices under any counterfactual ranking of hotels.

6 Results

Both the ranking as well as the consumer side estimations are based on data from Paris.

[35]The underlying assumption is that consumers' search costs do not depend which city a consumer is considering. However, I do allow for variation in consumers' valuations for hotel characteristics by estimating the hotel preference parameters using consumer search data from Paris directly.

6.1 Ranking

Tables 2.6 to 2.9 display the results of the effect of a hotel's price on the rank position which it is displayed in. Each table provides a specification that accounts for a sales indicator in addition to a hotel's price, or does not. Moreover, in each table, results using instruments are juxtaposed with the OLS results stemming from the same data. All specifications include hotel and query-travel-date fixed effects (or travel-date fixed effects, respectively, for Tables 2.7 and 2.9, which are based on the somewhat aggregated version of the data).

Across specifications, I find a robust positive coefficient for price that, using instruments, ranges between 3.4 and 3.8 for the query-level results in Tables 2.6 and 2.8, and between 1.5 and 2.3 for the results based on the more aggregate travel-date panel in Tables 2.7 and 2.9. The coefficient is significantly different from 0 in all specifications, and, with the exception of columns (3) and (4) in Table 2.7, increases in magnitude once instruments are employed. Its value implies that a 1 dollar increase in price would lead to a shift in the hotel's position by roughly one to four positions towards the lower end of the page (i.e., an increase in the position number). Given the extent to which hotels' prices vary in practice, this effect is not minor, and would imply that hotels may be quite substantially pressured to set lower prices by the default ranking. The ranking's elasticity with respect to hotels' prices therefore seems to be economically meaningful.

For the coefficient of the sales dummy, the results give a more blurry picture. While it is negative and varies between -74 and -104 in the OLS specifications, it jumps up (or down, respectively, when using the brand-based instruments in Table 2.8), to a substantial extent once instrumented for, peaking at -480 in Table 2.7. The reason for this is most likely that the instruments for sale are relatively weak, and that it is more challenging to identify parameters with two, instead of one, endogenous variable.

Interestingly, the number of bookings in a 48-hour window tends to increase visibility of a given hotel, indicating that the ranking algorithm may indeed take popularity into account.

All Tables report the F-test statistics for a test of relevance of the excluded instruments on price and sales respectively. All instruments pass the weak instruments rule of thumb, although the Arellano-Bond type instruments tend to be the weakest. The corresponding first stages are reported in Tables **??** to 2.22 in Appendix 2.E.

All in all, my results suggest that Expedia's hotel ranking intensifies price competition between hotels by pushing hotels to less visible positions if they offer a higher price. The implication is that the elasticity of demand that hotels face is higher compared to a situation in which OTAs rank hotels randomly, for instance. Consequently, hotel markups are lower in the situation with such a ranking algorithm. Another potential implication could be that hotels provide costly add-ons or may try to employ, to some extent, the obfuscation strategies that are explained in Ellison and Ellison (2009).

Robustness: Neighbors based on "donut"-shaped area around focal hotel. As noted above, the exogeneity of the neighborhood instrument relies on the assumption that the prices of neighboring hotels do not causally influence the position of the focal hotel. This assumption would not hold if Expedia is well informed about local demand shocks. If Expedia was able to anticipate a positive shift in demand for rooms in a given neighborhood (for example based on users' search and click behavior), then it might

Table 2.6: Results using neighborhood-based instruments

	Dependent variable:			
	position			
	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
price	2.007***	3.463***	1.973***	3.394***
	(0.216)	(0.146)	(0.214)	(0.150)
1{sale}			−74.041***	−290.785***
			(4.348)	(105.237)
# bookings past 48h	−0.945***	−0.820***	−0.894***	−0.617***
	(0.203)	(0.194)	(0.201)	(0.204)
Query×travel-date FE	✓	✓	✓	✓
Hotel FE	✓	✓	✓	✓
1st stage F-stat of excl. instruments on price		104.2		77.25
1st stage F-stat of excl. instruments on sale				85.58
Observations	914,814	914,814	914,814	914,814
Adjusted R^2	0.730	0.711	0.731	0.701
Standard errors clustered at query×travel-date level.			*p<0.1; **p<0.05; ***p<0.01	

Column (2) uses as instruments for price the number of neighboring hotels available on a given query and travel date, and the average price of neighboring hotels at a given travel and query date. Column (4) uses as instruments for price and the sale indicator the number of neighboring hotels and the average price and sales indicator. Neighboring hotels are defined as being located within 500 meters of the focal hotel.

have an incentive to display hotels in that neighborhood very visibly despite their increased prices. This would lead to a downward bias in the estimated price coefficient, as hotels would be ranked at visible positions (low position numbers) despite their high prices.[36] In this exercise, I therefore redefine the definition of a hotel's neighbors by determining that neighbors are hotels located within a ring-shaped area surrounding the focal hotel between the 500 meter and 1 kilometer radius. Prices of these hotels are possibly less likely to causally influence the focal hotel's position, but might still well be correlated with its prices. The results are displayed in Table 2.10. The instruments are somewhat weaker than in Table 2.6, as indicated by the lower first stage partial F-test of the excluded instruments, but still considered strong enough. The instrumented price coefficient in column (2) is very similar as in Table 2.6, whereas the coefficient on the instrumented sales indicator in column (4) increases substantially in magnitude, possibly due to weaker instruments.

Robustness: Aggregated panel. I estimate the parameters using the neighborhood-based as well as the brand-based instruments also on the more aggregate hotel-travel date panel (i.e., the panel used for the Airbnb and Arellano-Bond type instruments above). The results, displayed in Tables 2.23 and 2.24 in Appendix 2.F, are extremely similar to my findings above.

Robustness: Linear probability model. The above linear regressions assume that a hotel's position can be linearly related to its price. I explore alternative specifications by employing as dependent

[36] If hotels were more likely to be on sale when demand is low, and Expedia was informed about local demand shifts, the bias in the coefficient of the sale dummy would be biased upwards. However, from the data, the determinants of a hotel's decision to go on sale are not completely clear.

Table 2.7: Results using Airbnb and 1{August}×zip code instruments (based on more aggregated travel-date panel of hotels)

	Dependent variable:			
	average position across queries			
	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
price	1.908^{***}	2.066^{***}	1.880^{***}	1.477^{***}
(average across queries)	(0.338)	(0.171)	(0.336)	(0.198)
1{sale}			-76.533^{***}	-492.178^{***}
(average across queries)			(8.274)	(146.941)
# bookings past 48h	-1.177^{**}	-1.174^{**}	-1.143^{**}	-0.966^{**}
(average across queries)	(0.510)	(0.506)	(0.494)	(0.451)
Travel-date FE	✓	✓	✓	✓
Hotel FE	✓	✓	✓	✓
1st stage F-stat of excl. instruments on price		286.99		117.88
1st stage F-stat of excl. instruments on sale				72.91
Observations	447,378	447,378	447,378	447,378
Adjusted R^2	0.787	0.787	0.788	0.760
Standard errors clustered at travel date level.				$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

The dependent variable is hotel j's average position in Expedia's listings page for travel date t across web-scraped queries. Instruments used for mean_price in column (2) are the price and availability of Airbnbs located in the same zip code as a focal hotel. Instruments used for mean_price and mean_sale in column (4) are both the price and availability of Airbnbs located in the same zip code as a focal hotel, as well as an indicator variable for the weeks in August with little price variation interacted with the zip code.

Table 2.8: Results using brand-based instruments

	Dependent variable:			
	position			
	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
price	1.199^{*}	3.757^{***}	1.179^{*}	3.743^{***}
	(0.652)	(0.284)	(0.645)	(0.284)
1{sale}			-103.698^{***}	-37.077
			(16.042)	(33.557)
# bookings past 48h	-0.235	-0.197	-0.155	-0.169
	(0.271)	(0.254)	(0.262)	(0.250)
Query×travel-date FE	✓	✓	✓	✓
Hotel FE	✓	✓	✓	✓
1st stage F-stat of excl. instruments on price		54.28		77.93
1st stage F-stat of excl. instruments on sale				1126.23
Observations	258,011	258,011	258,011	258,011
Adjusted R^2	0.738	0.651	0.739	0.652
Standard errors clustered at query×travel-date level.				$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Column (2) uses as instruments for price the average price charged by other hotels in Paris of the same brand. Column (4) uses as instruments for price and the sales indicator the average price and average sales indicator employed by other hotels in Paris of the same brand.

Table 2.9: Results using Arellano-Bond instruments (based on more aggregated travel-date panel of hotels)

	Dependent variable:			
	average position across queries			
	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
price	1.813***	2.349***	1.788***	2.322***
(average across queries)	(0.361)	(0.436)	(0.358)	(0.435)
1{sale}			−76.228***	−95.585***
(average across queries)			(8.645)	(28.868)
# bookings past 48h	−0.341	−0.326	−0.320	−0.300
(average across queries)	(0.270)	(0.260)	(0.262)	(0.251)
Travel-date FE	✓	✓	✓	✓
Hotel FE	✓	✓	✓	✓
1st stage F-stat of excl. instruments on price		20.79		36.28
1st stage F-stat of excl. instruments on sale				21.62
Observations	451,366	451,366	451,366	451,366
Adjusted R^2	0.795	0.793	0.796	0.794
Standard errors clustered at travel date level.			*p<0.1; **p<0.05; ***p<0.01	

The dependent variable is hotel j's average position in Expedia's listings page for travel date t across web-scraped queries. Column (2) uses the lagged difference of prices as an instrument for a hotel's average price on travel date t. Column (4) uses the lagged differences of prices and sales indicator variable as instruments for a hotel's average price and sales indicator on travel date t.

Table 2.10: Results using neighborhood-based instruments ("neighbor" = hotels $\in$ [500m, 1,000m) from focal hotel)

	Dependent variable:			
	position			
	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
price	2.008***	3.521***	1.974***	3.316***
	(0.214)	(0.166)	(0.213)	(0.170)
1{sale}			−73.811***	−874.567***
			(4.314)	(134.050)
# bookings past 48h	−0.890***	−0.759***	−0.845***	−0.209
	(0.180)	(0.171)	(0.178)	(0.204)
Query×travel-date FE	✓	✓	✓	✓
Hotel FE	✓	✓	✓	✓
1st stage F-stat of excl. instruments on price		54.05		40.52
1st stage F-stat of excl. instruments on sale				94.99
Observations	927,322	927,322	927,322	927,322
Adjusted R^2	0.732	0.713	0.734	0.573
Standard errors clustered at query×travel-date level.			*p<0.1; **p<0.05; ***p<0.01	

Column (2) uses as instruments for price the number of neighboring hotels available on a given query and travel date, and the average price of neighboring hotels at a given travel and query date. Column (4) uses as instruments for price and the sale indicator the average price and sales indicator. Neighboring hotels are defined as being located in a "donut" shaped circle of 500 to 1,000 meters of the focal hotel.

variable an indicator variable specifying whether a given hotel is displayed at the top 10 (or top 5, respectively) positions. The results are shown in Appendix 2.F Tables 2.25 and 2.26 using neighborhood-based instruments. Clearly, the effect of price retains its sign across specifications, i.e., a higher price is associated with a decline in the probability of being displayed very visibly. For the parameter on the sales dummy, it looses significance once instruments are employed.

Robustness: Other travel destinations. Preliminary evidence shows that the results are quantitatively somewhat different across destinations, but qualitatively the same. It is plausible that they might be different, since the competitive environments across the different cities seem very different (many small hotels (Paris) vs. few large hotels (Cancun); strong vs. weak seasonality in demand; high number of business travellers vs. holiday makers, etc.). Expedia might to some extent adjust its ranking algorithm to the different types of travellers it expects for a given city; and hotels may focus on different segments of demand, and appear thus to engage in different pricing strategies.

6.2 Demand

Table 2.11 displays the demand estimates which are derived using Ursu (2018)'s sequential search model and the WCAI data on consumer searches in Paris. The estimation is based on data of only those searches that contain at least one click (which comprise in fact only 16% of all searches): searches with no clicks are likely to be carried out by customers that may not be "seriously" searching for a hotel or by bots scraping the page.[37] Moreover, as explained above, I use the search cost parameter estimates from Ursu (2018) instead of estimating them again (I use the results derived in column (1) of Table 8 in her paper). Thus, I set $\hat{\gamma} = 0.0044$ as a position parameter, and $\hat{k} = -1.0305$ as the search cost constant (see Table 2.12). Columns (1a) and (1b) use a location score defined by a hotel's distance to the Louvre in Paris, whereas columns (2a) and (2b) use a location score based on the minimum distances to a variety of touristy landmarks in Paris.[38] Columns (1a) and (2a) use different starting values for the simulated maximum likelihood estimation than columns (1b) and (2b), showing that the starting value does not affect results very much. The price coefficient has the expected sign, being significantly negative and is comparatively large in magnitude. Stars, the location score (with a high index meaning better location), and the outside option are all positive significant. The coefficient of the review score has a positive sign, but is only marginally significant. The sales parameter has a positive sign, but is insignificant, which reflects Ursu (2018)'s results where this coefficient is insignificant in three out of the four destinations. All in all, the results I obtain seem intuitive and are in line with the results of both Ursu (2018) as well as Y. Chen and Yao (2016).

[37] Information on the exact data cleaning process can be found in the Appendix, Section 2.C.1.

[38] Accounting for hotel stars and checkin date, the zip code of the Louvre – 75001 – persistently features the highest priced hotels, possibly indicating a premium consumers are willing to pay for the perceived centrality of that location.

Table 2.11: Search model estimates using WCAI Paris data. Results based on different starting values and different definitions of "location score".

	Using Louvre-based location score:		*Using more complex location score:*	
	(1a)	(1b)	(2a)	(2b)
price (\$100)	-0.176***	-0.176***	-0.179***	-0.179***
	(0.021)	(0.022)	(0.027)	(0.027)
stars	0.072***	0.072***	0.075***	0.075***
	(0.023)	(0.022)	(0.026)	(0.026)
review score	0.033**	0.033*	0.035*	0.035*
	(0.019)	(0.021)	(0.026)	(0.026)
chain dummy	-0.047*	-0.047*	-0.042*	-0.042*
	(0.030)	(0.030)	(0.031)	(0.031)
location score alt1	0.035***	0.035***		
	(0.011)	(0.012)		
location score alt2			0.044***	0.044***
			(0.013)	(0.013)
sale	0.052	0.052	0.049	0.049
	(0.046)	(0.047)	(0.055)	(0.055)
outside option	0.620***	0.622***	0.668***	0.668***
	(0.087)	(0.087)	(0.086)	(0.086)
Log-likelihood	-3,739.3	-3,739.3	-3737.8	-3737.8
# individuals	1,051	1,051	1,051	1,051
# observations	31,874	31,874	31,874	31,874
Note:			*p<0.1; **p<0.05; ***p<0.01	

The definition of the location score for the two columns on the left is based on the distance to the Louvre in Paris. The definition of the location score for the two columns on the right is based on the minimum distance to any of these locations: Sacré Coeur, Louvre, Les Invalides, Hotel de Ville. Results in the first and third column (columns a) are based on using the zero vector as a starting value. Results in the second and fourth column (columns b) are based on using $[0.3, -0.2, 0, 0, 0.2, -0.2, 0]$ as a starting vector.

Table 2.12: Fixed parameters, search model

search cost constant (k)	-1.0305
search cost position parameter (γ)	0.0044

Note: Parameters are taken from Ursu (2018) and are being held constant across specifications of the demand model.

7 Counterfactuals

The goal of the counterfactuals is to study hotels' price-setting decisions and consumers' welfare gains or losses under a counterfactual ranking algorithm. I only simulate changes in hotels' pricing decisions, and abstract from any changes in hotels' willingness to offer a sale. As above, this is motivated by the lack of understanding about what determines hotels' decisions to offer a discount.

Recall that I do not observe demand in my web-scraped dataset. To compute market shares, I therefore employ the structural demand parameters estimated on the 2009 consumer search and purchase dataset. I then simulate consumers that arrive on Expedia and are confronted with a list of hotels corresponding to the listings observed in the web-scraped dataset from 2019. The underlying assumption is that consumers' preference parameters have not changed over time.

7.1 Method

To conduct the counterfactual simulations, I make a simplification in that I carry out counterfactuals only for observations of hotels displayed on the first results page. This is motivated by the fact that consumers hardly ever click or purchase hotels on later results pages, leading to almost zero demand for hotels displayed on those pages.

I first show how the estimates from above (the demand parameters, and the ranking's elasticitiy with respect to prices and positions), along with observed prices, imply marginal costs for each hotel observation. Recall that $\hat{\alpha}$ denotes an estimate of $r'_{jqt}(p_{jqt})$, i.e., the effect of a hotel's price on a hotel's position obtained from the estimations above. Re-arranging the first order condition (2.1) yields:

$$\tilde{c}_{jqt} \quad = \quad p_{jqt} \quad + \quad \underbrace{\frac{s_{jqt}(\cdot)}{\frac{\partial s_{jqt}(\cdot)}{\partial p_{jqt}} + \frac{\partial s_{jqt}(\cdot)}{\partial r_{jqt}} \cdot \hat{\alpha}}}_{<0} \tag{2.5}$$

Prices p_{jqt} are observed in the data. To approximately match Expedia's commission rate, I set $\tau = 0.2$. As the web-scraped dataset does not contain demand data, market shares $s_{jqt}(\cdot)$ are computed by simulating 100,000 consumers that are being confronted with a given sequence of hotel listings, and choosing which hotels rooms to book, if any. The parameters $\frac{\partial s_{jqt}(\cdot)}{\partial p_{jqt}}$ and $\frac{\partial s_{jqt}(\cdot)}{\partial r_{jqt}}$ can only be derived by simulation, as no closed-form solution exists. Starting with the data of observed hotel listings for each given query-travel dates, I compute the following:

- For $\frac{\partial s_{jqt}(\cdot)}{\partial p_{jqt}}$, I compute what an increase in its price of 5% implies for its market share, by simulating 1000 consumers who view the given hotel in query q for date t.
- For $\frac{\partial s_{jqt}(\cdot)}{\partial r_{jqt}}$, I equivalently increase each hotel's position number by 1, and by simulating 1000 consumers I compute how this affects the hotel's market share.

For each hotel, I first average the simulated derivate across all query-travel date observations to obtain $\overline{\frac{\partial s_j(\cdot)}{\partial p_j}}$ and $\overline{\frac{\partial s_j(\cdot)}{\partial r_j}}$. Assuming that these estimates are the same across hotels, I then take the average again, obtaining a value for $\overline{\frac{\partial s(\cdot)}{\partial p}}$ and for $\overline{\frac{\partial s(\cdot)}{\partial r}}$, respectively. I can then back out costs using the following equation:

$$\tilde{c}_{jqt} \quad = \quad p_{jqt} \quad + \quad \frac{s_{jqt}(\cdot)}{\overline{\frac{\partial s(\cdot)}{\partial p}} + \overline{\frac{\partial s(\cdot)}{\partial r}} \cdot \hat{\alpha}} \tag{2.6}$$

Given all parameters and hotels' cost estimates, one can perturb $\hat{\alpha}$ and simulate prices and market shares under different scenarios. Markups can easily be derived by transforming Equation 2.6:

$$\frac{p_{jqt} - \tilde{c}_{jqt}}{p_{jqt}} \quad = \quad -\frac{s_{jqt}(\cdot)}{p_{jqt} \cdot \left(\overline{\frac{\partial s(\cdot)}{\partial p}} + \overline{\frac{\partial s(\cdot)}{\partial r}} \cdot \hat{\alpha}\right)} \tag{2.7}$$

7.2 Results for costs and interpretation

The simulations yield $\overline{\frac{\partial s(\cdot)}{\partial p}} = -0.014$ and $\overline{\frac{\partial s(\cdot)}{\partial r}} = -0.032$. Moreover, I use $\hat{\alpha} = 3.2$ to back out the costs.

Averaged across all hotel observations, I find average costs of 206.4 €, whereas average prices for the sample of hotels are 217.83 € (see Table 2.13). The substantial variance in backed-out costs is reflective of the substantial variation in prices across hotel observations.

Markups averaged across hotel observations are extremely low. In fact, the model implies markups equal to 0 for 63% of hotel observations. In other words, the model predicts that those hotels set prices equal to marginal cost, and an average markup across all hotel observations of close to 0. In contrast, other research like Cazaubiel et al. (2020) find average markups of 35-43%. The reason for my finding is the peculiarity of the search model generating very sparse demand. Indeed, the 63% of hotel observations for which markups are supposedly 0 also generate 0 demand in a given query, so that by construction marginal costs are predicted to be equal to prices (see equation 2.6). Only 5% of hotel observations derive a market share of more than 10%, and only 2% a market share of more than 20%. The relatively high search costs contribute to this finding, and possibly also be the lack in consumer heterogeneity in preferences of the sequential search model. In reality, it is unlikely that certain hotels obtain absolutely no demand on Expedia. Moreover, in my model, a hotel's position in a given query affects the hotel's market share, which in turn affects the hotel's marginal costs. A hotel that is displayed on the bottom of a page in a given query may thus derive no demand, and therefore my model predicts that it should have set prices equal to marginal cost. Finally, markups are low precisely because the model predicts that the algorithm itself intensifies price competition quite substantially through the additional term $\frac{\partial \bar{s}(\cdot)}{\partial r} \cdot \hat{\alpha}$ that enters the denominator in the first order condition.

These aspects, as well as the substantial price fluctuation over time that we saw in Figure 2.4, highlight once again that in this setting, marginal costs cannot be interpreted as physical marginal costs. In reality, hotels most likely face very high fixed costs and very low marginal costs of a certain occupation of a room. A hotel's physical marginal costs should not depend on the position in which it appears in a given query. Moreover, hotels' price-setting decisions are likely to be dynamic problems since hotels are capacity constrained. When setting prices, hotels optimally take into account their current occupancy and the expected demand of rooms. All in all, I believe that marginal costs should here rather be interpreted as the option value of having a room booked at a given date.

Table 2.14 shows the distribution of costs, markups and prices across hotels of a given star rating. It is perhaps noteworthy that markups are *higher* for hotels that with 2, 3 or 4 stars, compared to hotels with 5 stars. This is again a result of some very high-priced, luxury hotels having a zero or very small market share, and thus having a low value for $s_{jqt}(\cdot)$.

Table 2.13: Basic descriptives: prices, costs and markups

	Average	Median	Std Dev
Prices	217.83	150	213.18
Costs	206.4	140	215.0
Markups	8.4×10^4	0	0.003

Table 2.14: Average prices, costs and markups, by star rating

	1 star	2 stars	3 stars	4 stars	5 stars
Prices	120.44	100.88	120.83	151.6	423.83
Costs	114.19	89.36	109.62	142.48	417.24
Markups	3×10^{-7}	3.9×10^{-4}	4.2×10^{-4}	6.7×10^{-4}	9.3×10^{-5}
# of hotels	4	54	312	403	83

7.3 Simulating Counterfactuals

The model assumptions and estimates derived above allow to assess how the change in the ranking algorithm or other parameters impact equilibrium prices and market shares. I perform counterfactual simulations using 100,000 simulated consumers for those hotel observations with positive market shares (the other hotel observations would not adjust prices in response anyways, since their prices are being fixed equal to marginal costs).

I first simulate prices and market shares under the scenario in which the ranking's elasticity with respect to prices, $\frac{\partial r_j(\cdot)}{\partial p_j} = \alpha$, is set to 32, i.e. increased by a factor 10. I find that in response, as expected, all hotels react by decreasing their price. However, the price increase is very small and amounts to only 23 cents on average. This is a result of the market shares being very often small, so that the effect of a stronger ranking yields only in a limited reaction by hotels. Moreover, recall that hotels *already* are under a lot of pricing pressure given *any* effect from the ranking, which already lowers their markups in contrast to a situation without such a ranking.

In a next step, I therefore compute equilibrium prices and market shares in a scenario in which the ranking does not matter, i.e. $\alpha = 0$. I find much larger effects on prices. Whereas the average price increase amounts to only 1.89 euros, 7% of hotels increase their price by more than 10 euros. Overall, while there is significant heterogeneity in hotels' price adjustments. The price adjustments for the majority of hotels are still relatively low, however, due to the small market shares (see Figure 2.9).

Future work could go further and not only change the elasticity parameter, but also the actual displayed ranking of hotels (and thus show hotels in alternative orderings to the simulated consumers).

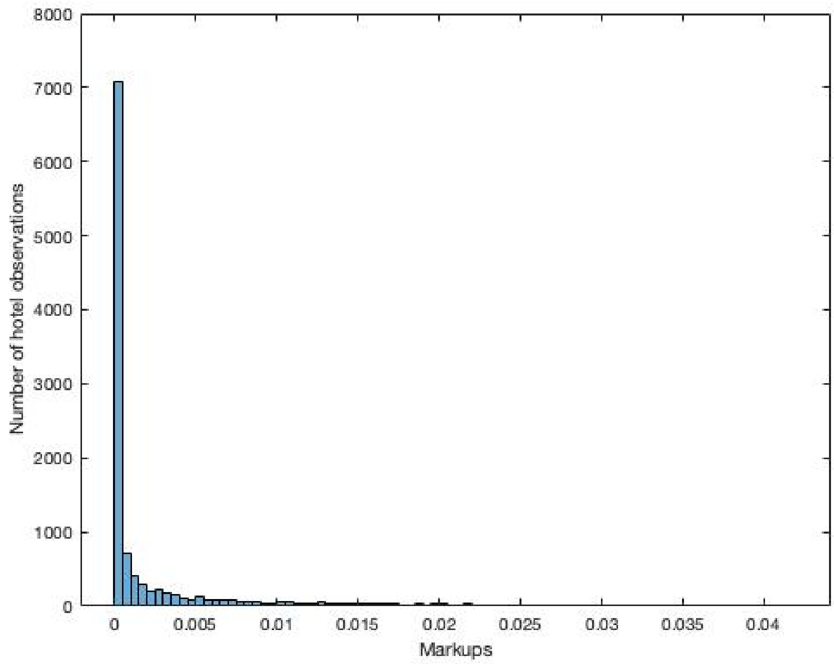


Figure 2.7: Distribution of markups across hotel observations, after removing the 63% of observations with markups = 0.

In contrast, above I vary only the parameter that governs the ranking effect on hotels' prices, without perturbing the ranking itself. This would yield substantially different market shares for hotels. Another interesting exercise would be to compare the current simulations and observed market shares with a scenario in which rankings are random. In addition, it would be interesting to compare the observed prices and market shares to a scenario in which consumers have perfect information (0 search costs) and do not search at all.

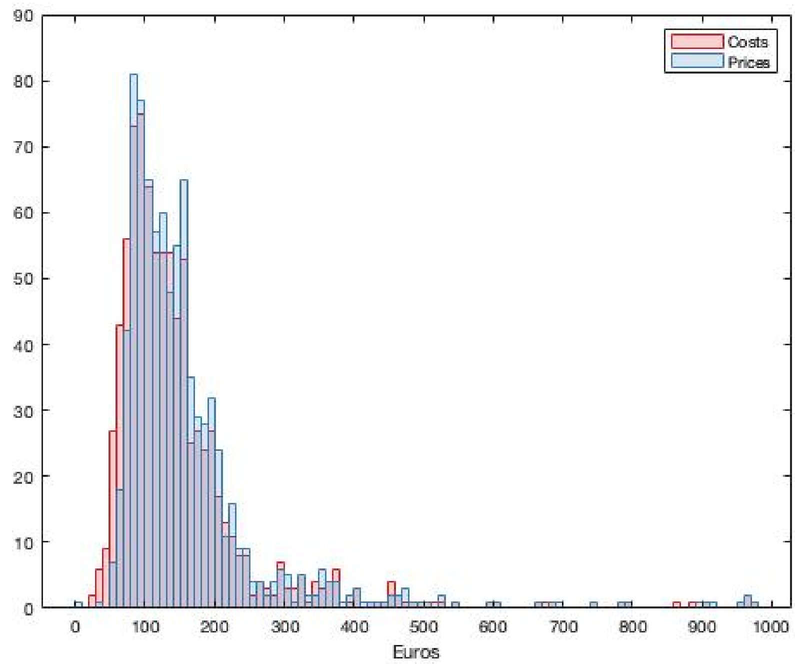


Figure 2.8: Distribution of costs and average prices across hotel observations.

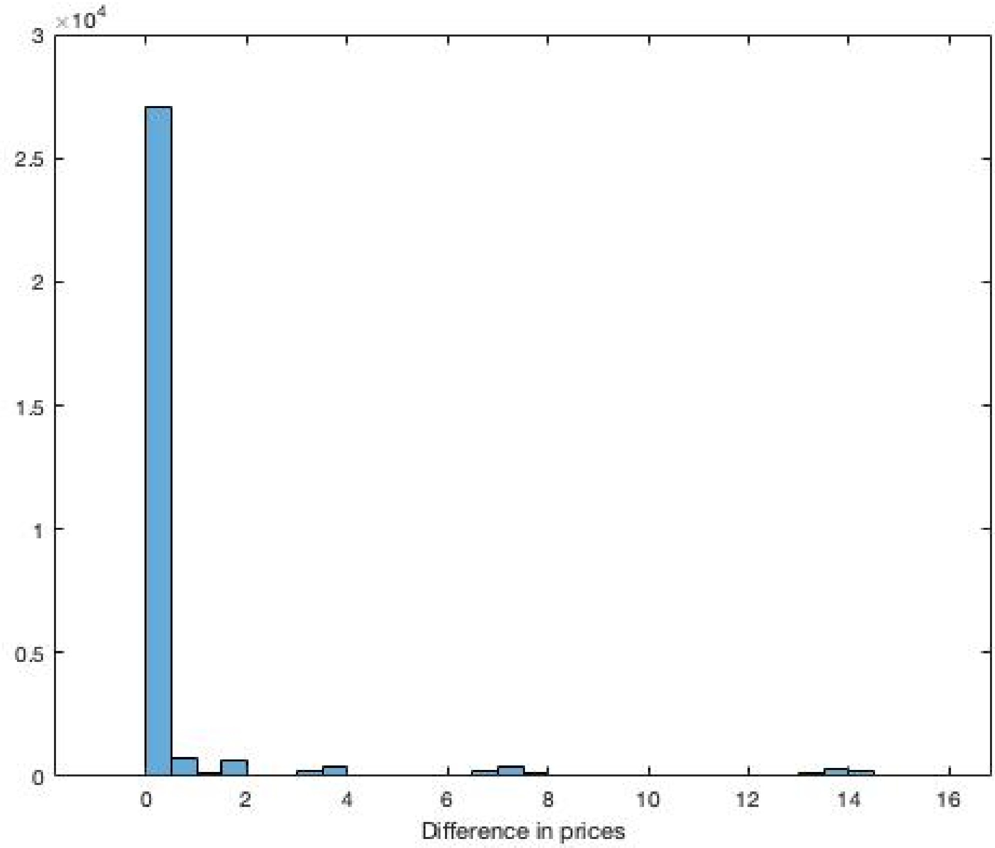


Figure 2.9: Change in prices (in euros) when effect of ranking is shut off.

8 Conclusion

The seemingly infinite virtual shelf space available online has given rise to a cornucopia of products competing for consumers' attention. As consumers face search costs, online platforms play an essential role in designing algorithms that make products discoverable and that guide consumer search. As visibility is important to generate demand, online sellers may adjust their strategic decision-making so as to be displayed more favorably by the algorithms.

In the context of Expedia, I find that the algorithms that these intermediaries employ tend to display hotels more visibly at times at which they are priced lower. This tends to intensify price competition among hotels, thus moving the market into a somewhat more competitive equilibrium. The structural model of the demand as well as the supply side allows to simulate how hotels would set prices under a counterfactual ranking algorithm. I find that, if prices did not affect hotels' visibility, hotels would on average increases prices by 1.89€. 7% of hotels increasing their prices by more than 10€.

These results and the tools I use are, first, important for managers of online platforms: the equilibrium effects my analysis sheds light on can *not* be traced out by merely using A/B experiments that randomize the hotel rankings across website visitors. Instead, investigating these effects requires a structural model of the buyer, as well as the seller side, allowing to predict how changes of the ranking algorithm affect sellers' equilibrium prices. My results suggest that sellers tend to be squeezed by Expedia's ranking algorithms, whereas consumers are enticed by providing higher competition. Recall that platforms, however, crucially need to attract both seller, as well as buyer-side in order to subsist. My finding therefore has managerial implications for platforms seeking to trade off between providing value to the two sides, while at the same time also being profitable in the short as well as long run.

Second, my results are interesting for antitrust authorities, which have in the past been concerned about online booking platforms' market dominance and behavior with respect to hotels. My finding that booking platforms intensify price competition provide suggestive evidence that online platforms provide additional value to the consumer side – at the expense of hotels, however. This is aligned with the complaints that hotel associations in many countries have voiced regarding the dominance of online platforms.

The analysis suffers from several limitations. The linear relationship between rank and hotels' prices is an abstraction of a platform's ranking algorithm, which is possibly nonlinear and very complex. The model of hotels' price-setting decisions is equally simplified: in practice, hotels engage in dynamic revenue management as their inventory is limited and expected demand may change as a given travel date comes closer. Promotional sales practices seem to be widely used, but how hotels choose promotions in addition to prices is difficult to explain with a model. My model also abstracts from the fact that hotels might attract consumers via different channels (for example via their own website, or via the non-default rankings on the online travel agent).

Future research could attempt at solving any of these issues. Moreover, this paper could benefit from a more explicit model of the platform's decision process in how to optimally trade off between providing benefits to the seller as well as the buyer side, and what its incentives are when designing the algorithm. Lastly, policymakers as well as consumer advocates have emphasized the idea of mandating platforms to increase the transparency of their rankings. I leave the investigation of the effects of such a policy change for future research.

2.A Expedia's hotel listings page

2.A.1 Information on the ranking

Expedia gives some guidance to hotels on what affects hotels' visibility online, but the precise workings of the ranking algorithms are proprietary and possibly very complex. As seen in the screen shot in Figure 2.10, according to Expedia, hotels can improve their "Offer Strength", which in turn affects "Quality Scores" determining how a hotel will be ranked on Expedia as of 2019. Further, the tourism industry website Skift notes that, according to Expedia, "Offer Strength" (which includes prices and promotions) is the most important dimension according to which hotels are ranked, while "Compensation" is the least important one (see `https://tinyurl.com/54k3k397`, accessed 03/05/2023). Finally, Hannak et al. (2017) find that, as of 2014, Expedia's ranking differs quite substantially from one user to the next, and find that a lot of this is due to A/B testing. Note that compensation is a factor that I control for by taking hotel fixed effects, as long as those do not vary over the 2-month period in which I scraped the data.

2.A.2 Further note regarding Figure 2.1

As can be seen in Figure 2.1, notifications such as "We have 2 rooms left!" or "8 people booked in the last 48 hours" might be displayed along with the hotel that is on promotion, possibly creating some kind of urgency with consumers. The crossed out price reflects either the standard rate (in case of a sale), or (if displayed without there being a sale flag), according to Expedia, "the third highest price for [the displayed] room type at this hotel with the same length of stay and cancellation policy that customers have found within a 30-day window around the selected check-in date." Some hotels offer free cancellation, membership prices or further benefits like free breakfast. Moreover, some hotels feature the number of people who booked the hotel during the last 48 hours for any travel date, as detailed in listings three and five on Figure 2.1. All of these information are gathered and can serve as controls.

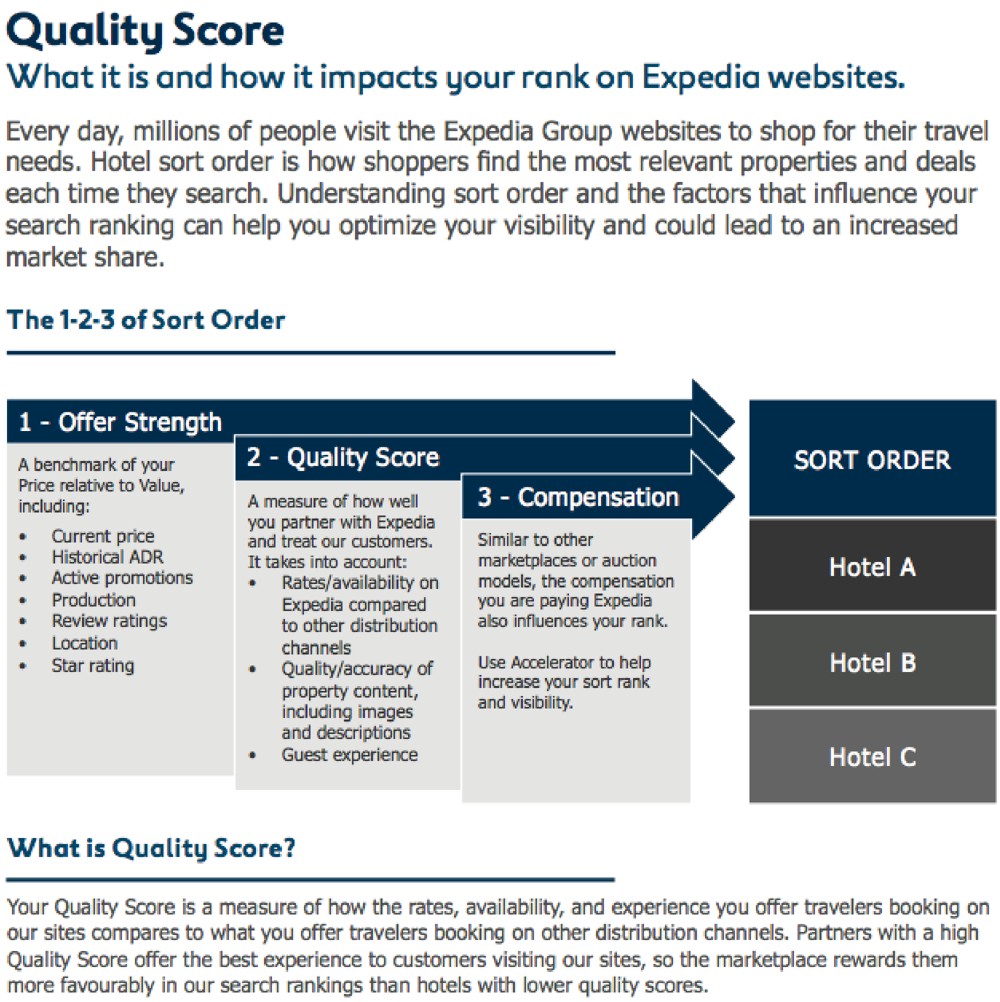


Figure 2.10: Determinants of a hotel's "Quality Score", which is a crucial input into a hotel's ranking, as of Expedia. Source: https://discover.expediapartnercentral.com/wp-content/uploads/2016/12/Expedia_Marketplace-White-Paper-April-2016.pdf (accessed 13/03/2019).

2.B Version of the model in which hotels decide on sales promotions

Previous literature has found that consumers do not only take into account prices, but also *in addition* value discounts that are offered by the hotel (Y. Chen & Yao, 2016, De los Santos & Koulayev, 2017, Koulayev, 2014, Ursu, 2018). At the same time, however, promotions on Expedia tend to be very opaque (see Appendix 2.D.3). Without a behavioral model of consumer choice, or a dynamic pricing problem in which promotions serve to make last minute changes to prices, it is difficult to understand why a hotel would want to offer a promotion *in addition* to a price. Moreover, if promotions are valuable to consumers, hotels would want to *always* show promotions, and raise prices somewhat so as to account for that.

Despite these conceptual difficulties, in this Section, I provide a version of the model in which hotels not only choose prices, but also have the ability to choose a binary variable that will give a discount on the price shown. Taking as given the pricing and promotion decisions of their competitors, at a given date q, hotels thus (i) set prices and (ii) decide whether or not to offer a discount for room bookings for date t, such that the resulting decisions form a Bertrand Nash Equilibrium.

I model the decision to provide a discount as a binary choice, i.e., $d_{jqt} \in \{0, 20\}$: the reason is that discounted hotels are highlighted (no matter how large the discount is), which might already have an effect on users, while the actual percentage reduction is usually only visible when hovering with the pointer over the "sale" button on the results page. Moreover, this fits to the consumer browsing data, in which promotions are also captured by a binary variable. I assume that offering a promotion comes at a cost of 20 euros to a hotel, which is empirically approximately the average discount that hotels offer for Paris on Expedia.

A discount has the potential to raise demand, but is costly for the hotel. As in the model in the main text, hotel j's profit is the product of the markup and the demand $Ms_{jqt}(\cdot)$ (market size $\times$ market share). What is new is that demand for hotel j in turn is not only a function of the hotel's pre-promotion price p_{jqt} and the promotional discount d_{jqt}, but also of the position r_{jqt} which the hotel is by default displayed in on the OTA's results page. By modelling demand as being dependent on a product's positioning, the model accounts for the stage in which consumers search for products. Products with a worse positioning in the ranking (i.e. a higher r_{jqt}) are less likely to be found by consumers, resulting in lower demand.

Let $(\mathbf{p_{qt}}, \mathbf{d_{qt}}, \mathbf{r_{qt}})$ be the vectors of prices, promotion decisions and rankings for all J hotels in the market. Hotel j's profit maximization problem then writes

$$\max_{p_{jqt} \in \mathbb{R}^+,\, d_{jqt} \in \{0,D\}} \Big(p_{jqt} \cdot (1-\tau) - \big(d_{jqt} + c_{jqt}\big)\Big) M s_{jqt}\big(\mathbf{p_{qt}}, \mathbf{d_{qt}}, \mathbf{r_{qt}}(p_{jqt}, d_{jqt})\big).$$

Next, define $\tilde{c}_{jqt} \equiv \frac{c_{jqt}}{(1-\tau)}$ and $\tilde{d}_{jqt} \equiv \frac{d_{jqt}}{(1-\tau)}$. Deriving the first order conditions using with respect to price (using the chain and the product rule) and re-arranging, one obtains:

$$p_{jqt} = \tilde{c}_{jqt} + \tilde{d}_{jqt} - \frac{s_{jqt}(\cdot)}{\underbrace{\frac{\partial s_{jqt}(\cdot)}{\partial p_{jqt}}}_{\text{direct } (<0)} + \underbrace{\frac{\partial s_{jqt}(\cdot)}{\partial r_{jqt}} r'_{jqt}(p_{jqt})}_{\text{indirect } (<0)}} \tag{2.8}$$

When deciding on prices, hotels thus take into account two types of effects on demand. The "usual" demand ("direct") effect $\frac{\partial s_{jqt}(\cdot)}{\partial p_{jqt}}$ expresses that, other things equal, a lower price will lead to higher demand. Second, prices and promotions affect the average position r_{jqt} which a given hotel is displayed in, which in turn affects how many users become aware of the hotel, click on it, and purchase it. Thus, the "indirect effect" $\frac{\partial s_{jqt}(\cdot)}{\partial r_{jqt}} r'_{jqt}(p_{jqt})$ is additionally going to be taken into account by hotels.

Moreover, as I model giving a discount as a discrete choice, a hotel is willing to offer a discount whenever

$$\Pi_{promo} \geq \Pi_{nopromo}$$

$$\Leftrightarrow \quad \Big(p_{jqt}(1-\tau) - c_{jqt}\Big) \cdot \Big(s_{jqt}\big(\mathbf{p_{qt}}, \mathbf{d_{qt}}, \mathbf{r_{qt}}(p_{jqt}, d_{jqt}=20)\big) - s_{jqt}\big(\mathbf{p_{qt}}, \mathbf{d_{qt}}, \mathbf{r_{qt}}(p_{jqt}, d_{jqt}=0)\big)\Big)$$
$$\geq 20 \cdot s_{jqt}\big(\mathbf{p_{qt}}, \mathbf{d_{qt}}, \mathbf{r_{qt}}(p_{jqt}, d_{jqt}=20)\big)$$

Note that again, $\mathbf{d}_{qt}$ enters the market share $s_{jqt}(\cdot)$ both directly and indirectly through affecting $r_{jqt}(\cdot)$. Thus, a hotel's willingness to offer a discount depends again not only on how the discount affects the demand directly, but also on how the discount affects the hotel's visibility on the platform.

As a result, if a platform modifies the ranking algorithm – for example by making the algorithm more sensitive to prices, or by ranking hotels completely independent of their prices – hotels will take this into account and set different prices and promotions. This affects hotels' markups and consumer surplus.

2.C Data Construction

2.C.1 Cleaning of WCAI Dataset

The analysis employs the WCAI dataset on consumer searches and purchases only in order to estimate the demand model's utility parameters for the Parisian data. Overall, my cleaning procedure is similar to the one by Ursu (2018), which she describes in the Appendix. The raw dataset contains 1,546,296 observations of 19,658 distinct user IDs making, all in all, 565 purchases in either Budapest, Cancun, Manhattan, or Paris.

- After removing a number of data errors (such as duplicated rows), outliers (in particular, observations of prices below 6$; queries for nine or more people; queries for 5 or more kids; a few observations of hotels with prices above 10,000US$), and unusual searches that are likely to be errors (such as searches with an unreasonable long length of stay or for dates in the past or in the far future). I end up with 1,361,377 observations of 17,880 distinct user IDs.

- Next, I try to remove all those observations that are likely to not have been made by actual humans, but by robots scraping the website. In those searches, all hotels were clicked on or all websites were looked at within a small time frame, with never any purchase made. This reduces the number of observations to 1,302,231 observations of 17,762 users who conduct 53,358 distinct searches.

Taking only searches for hotels in Paris, one obtains 328,926 observations of 4,424 users making 13,204 distinct searches.

For the purpose of estimating a model, I again use only a subset of this dataset, as I only include observations with the following characteristics:

- Searches within the default ranking of the platform, without any minimum star rating specified, and without an explicit hotel name specified (no sorting or filtering);

- Searches which contain at least one click.

In the end, the final dataset that I use for the estimation of the demand parameters contains 31,874 observations, stemming from 1,051 searches. To be able to estimate Ursu's (2018) model using the search data, I moreover create "effective" position numbers: hotels displayed on page 2 in position 3, for instance, are given the position number equal to: "(number of page 1)+position number on page 2". This is in slight contrast to the dataset that Ursu (2018) uses, which only contains observations on click and purchase behavior occurring on the first results page.

2.C.2 Cleaning of Web-Scraped Dataset

In the web-scraped dataset, each observation corresponds to a hotel that showed up in a certain query. There is a certain number of hotel observations in which no prices are displayed: this typically occurs for hotels that are fully booked, which are displayed on later pages of the results page. I remove all observations with no prices displayed.

Sponsored and organic listings. Expedia typically displays 55 hotels per results page.[39] Aside from the ranking algorithm that ranks all hotels, hotels on Expedia can pay to appear at a given position in its ranking: in this case, Expedia discloses that the given hotel is "sponsored", as can be seen in Figure 2.11. A sponsored hotel will in addition to that be displayed in the organic ranking as well. The organic listing will presumably continue to be a function of the hotel's price, whereas the sponsored listing will be determined by the hotel's willingness to pay for the slot. My analysis focuses on Expedia's organic ranking; therefore, I collect information about whether a given hotel appears in a sponsored listing, and subsequently exclude sponsored listings from my analysis. As seen in Figure 2.12, across all all results pages, sponsored listings are essentially limited to positions 1, 7, 35, 54 and 55, which are almost always occupied by sponsored hotel displays. In contrast, all other positions are almost always occupied by organically ranked hotels.

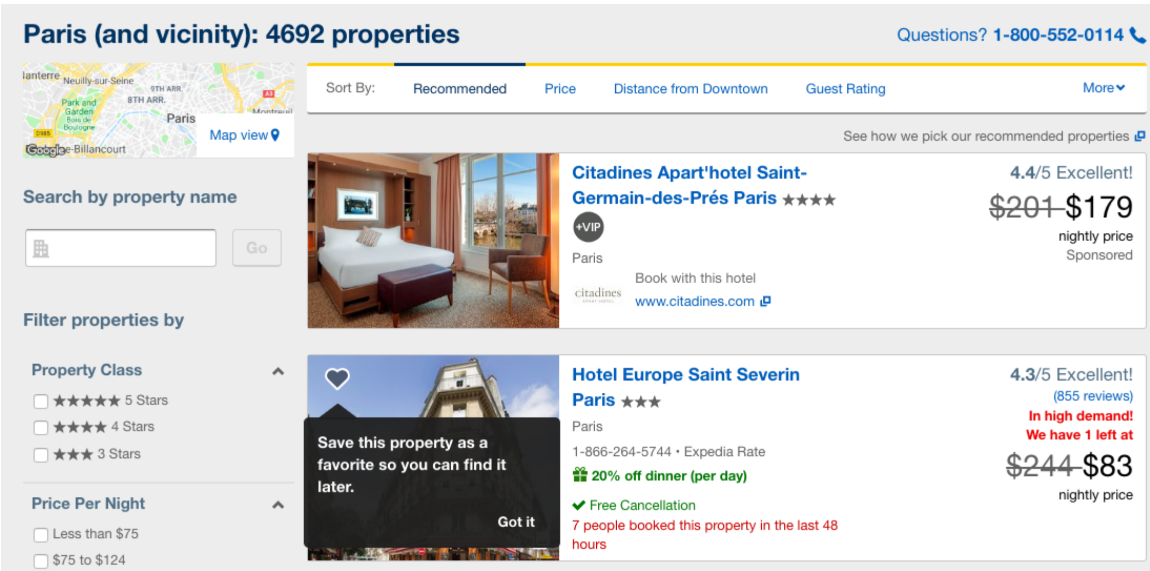


Figure 2.11: Screen shot taken from Expedia's listings page. It can be seen that the hotel in position 1 is sponsored (written below its price).

Table 2.15 shows the basic dimensions of the entire web-scraped dataset, after removing sponsored listings and listings without pricing information.

Table 2.15: Basic dimensions of web-scraped dataset, after removing (1) sponsored listings, and (2) listings with no price information. The number of observations for each city differs because the number of hotels per city differs a lot (see last column).

City	Observ	First query (date & time)	Last query (date & time)	# query dates scraped	# travel dates scraped	# hotels scraped
Budapest	319,796	2019-02-01 16:54:00	2019-04-01 15:40:00	784	310	1,045
Cancun	271,254	2019-02-02 14:50:00	2019-03-31 22:44:00	737	306	972
Manhattan	306,660	2019-01-30 18:39:00	2019-04-01 10:42:00	746	306	1,222
Paris	1,202,289	2019-02-01 18:04:00	2019-04-01 12:43:00	745	302	2,842

[39] In my data, a few listings pages have less than 55 hotels on the same page. My impression is that it is related to the quality of the Internet connection when accessing the page.

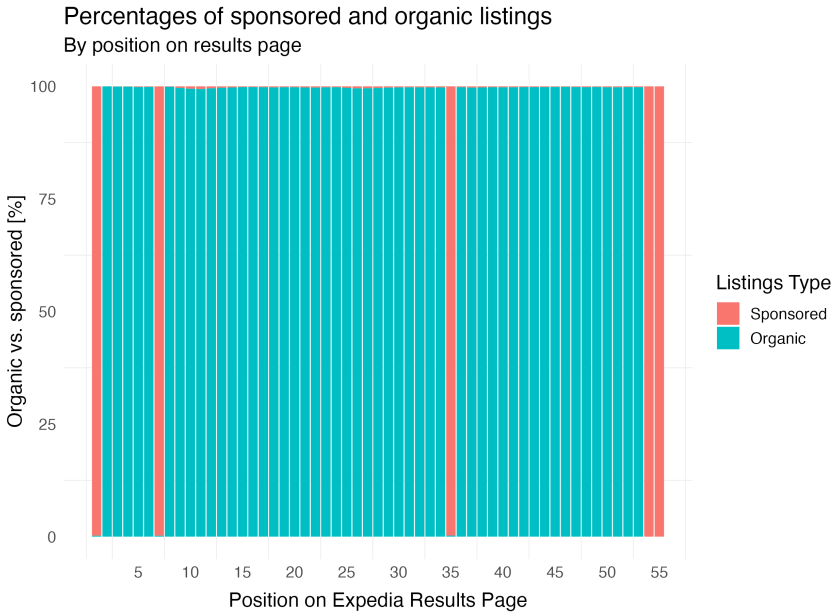


Figure 2.12: Percentage of "organic", as opposed to "sponsored", listings across observations of all 55 possible positions on Expedia's results pages.

2.C.3 Comparing hotels observed in the WCAI to hotels observed in the web-scraped dataset

Table 2.16 shows the market characteristics by city in the web-scraped, as well as in the WCAI dataset. There are notable differences in market structure across cities, but especially also when comparing the web-scraped to the WCAI dataset. The web-scraped and the WCAI dataset may differ for two reasons: (1) the markets have changed from 2009 to 2019; (2) the WCAI dataset is generated by actual consumer queries, and it is thus unlikely that the observations are giving an accurate picture of the full market. Indeed, the numbers of active hotels indicate that the web-scraped data captures a much higher number of hotels: in the case of Budapest, the number of hotels scraped in 2019 is over three times as high, for instance. The WCAI dataset moreover contains a much higher share of branded hotels.

Table 2.16: Market characteristics for the web-scraped dataset, as well as WCAI dataset.

	City	# Hotels	Avg. stars	Avg. review	Avg. # reviews	Brand (%)
Web-scraped (2019)	Budapest	1,045	3.25	4.11	333.20	10.02
	Cancun	972	2.95	3.92	588.22	7.49
	Manhattan	1,222	3.54	4.18	2,198.83	32.73
	Paris	2,842	3.27	3.93	377.78	25.93
WCAI (2009)	Budapest	276	3.25	4.15	17.44	25.21
	Cancun	110	3.31	4.17	252.91	31.63
	Manhattan	543	2.85	3.96	117.35	42.51
	Paris	1,637	2.56	3.77	15.98	40.91

There are two explanations formally why market characteristics differ across the datasets: (1) The market has changed over time. (2) The WCAI dataset is generated from actual consumer queries, and thus likely does not capture the entire market.

2.D Further Descriptive Analysis

2.D.1 Further aspects of users' search behavior, using data for searches in Paris only

The following descriptives provide further evidence of how users search for hotels online, and motivate my choice of search parameters employed when web-scraping. Figure 2.13 shows that the majority of users looking for accommodation in Paris are asking for stays of one to four nights. Note that as of 2009, this OTA did not necessarily require users to specify travel dates in order to see results in the listings page.

Figure 2.14 shows a large amount of heterogeneity in the booking window. The uptick at the booking window of 70 to 100 days may be driven by the popularity of travel dates around Christmas and New Year's in Paris (recall that all queries are made in the first 2 weeks of October).

Figure 2.15 displays the number of adult travelers for whom rooms are searched for, and Figure 2.16 the number of rooms. Note that as of 2009, 2 adults was the default in this OTA. In 84% of all searches made, the user did not specify the name of the property (in the remaining 16% of queries, the users tend to enter brand names like Best Western or Mercure, or landmarks like Louvre or Eiffel). As seen in Figure 2.17, in the vast majority of searches, there is no minimum star rating specified (i.e., no filtering of specific hotels takes place).

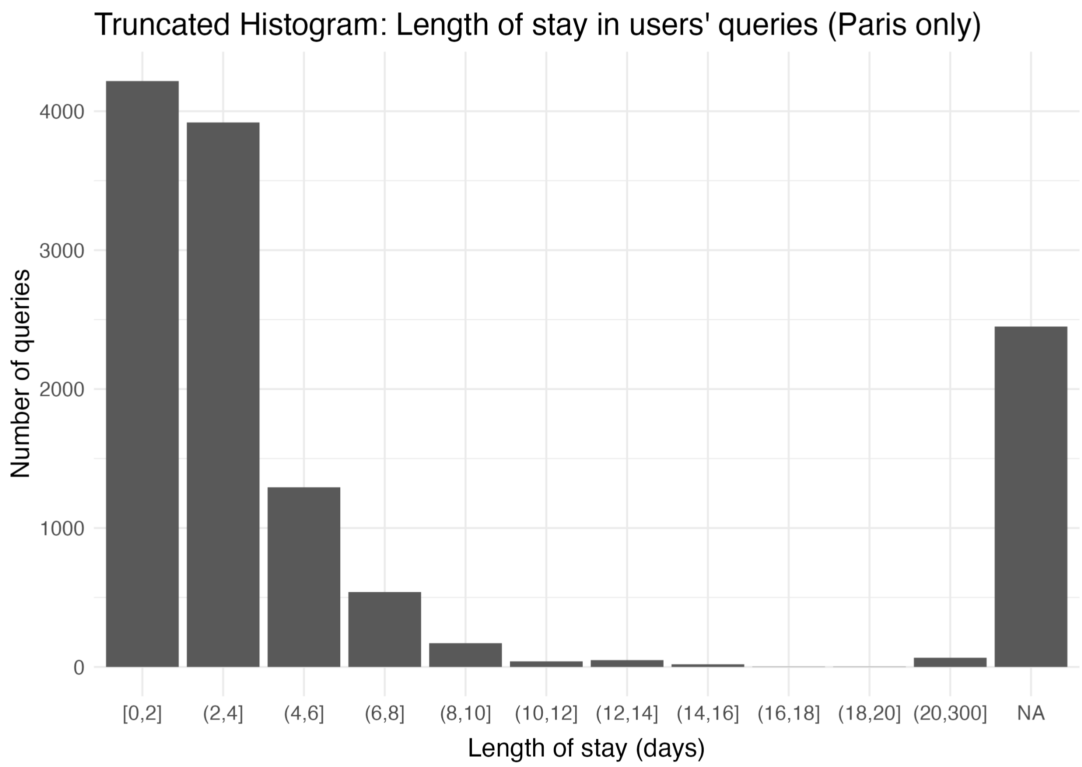


Figure 2.13: Length of stay asked for across user queries.

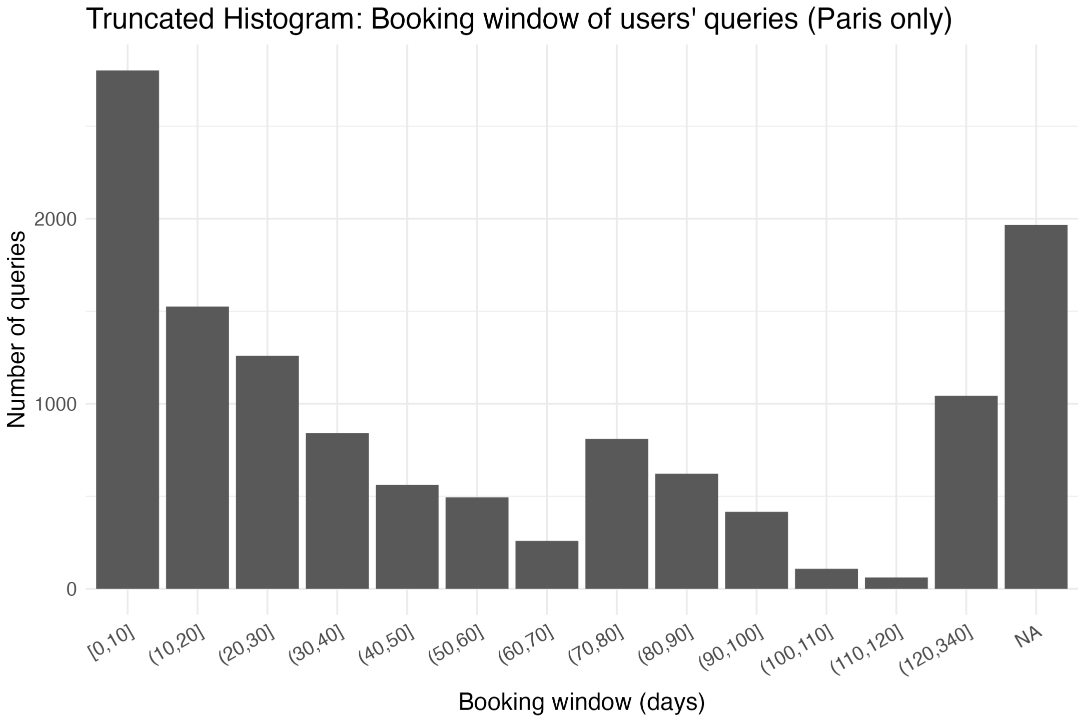


Figure 2.14: Booking windows across user queries.

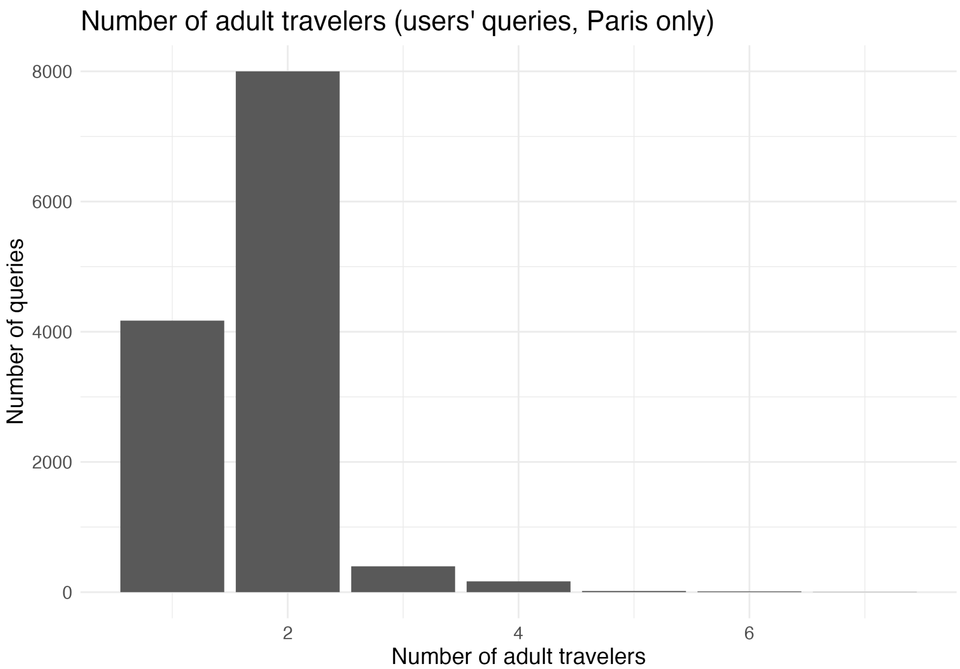


Figure 2.15: Number of adult travelers. Note that as of 2009, searching for 2 adults was the default option provided by this OTA.

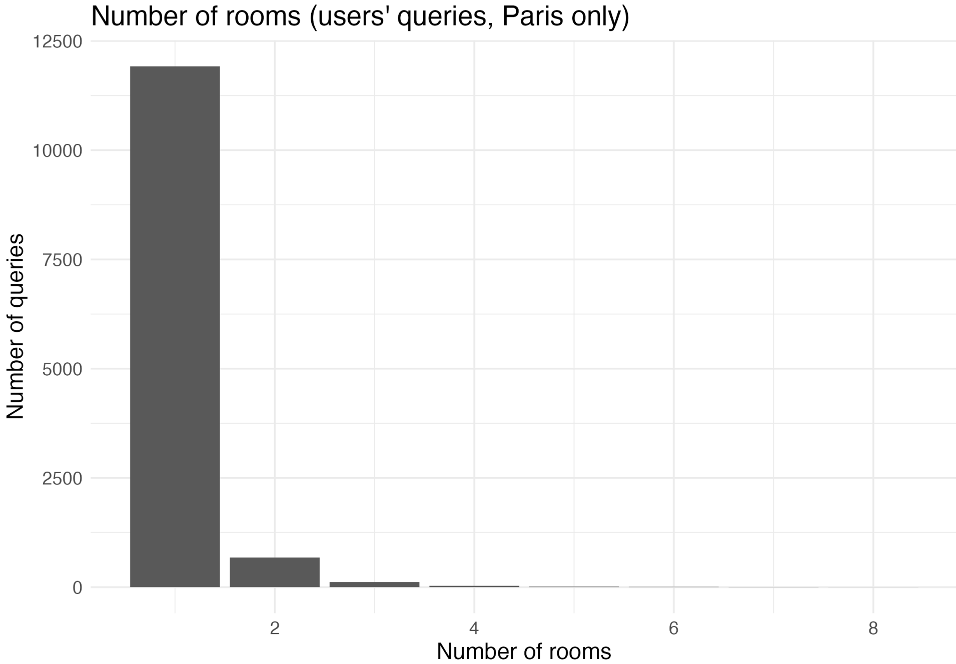


Figure 2.16: Number of rooms searched for.

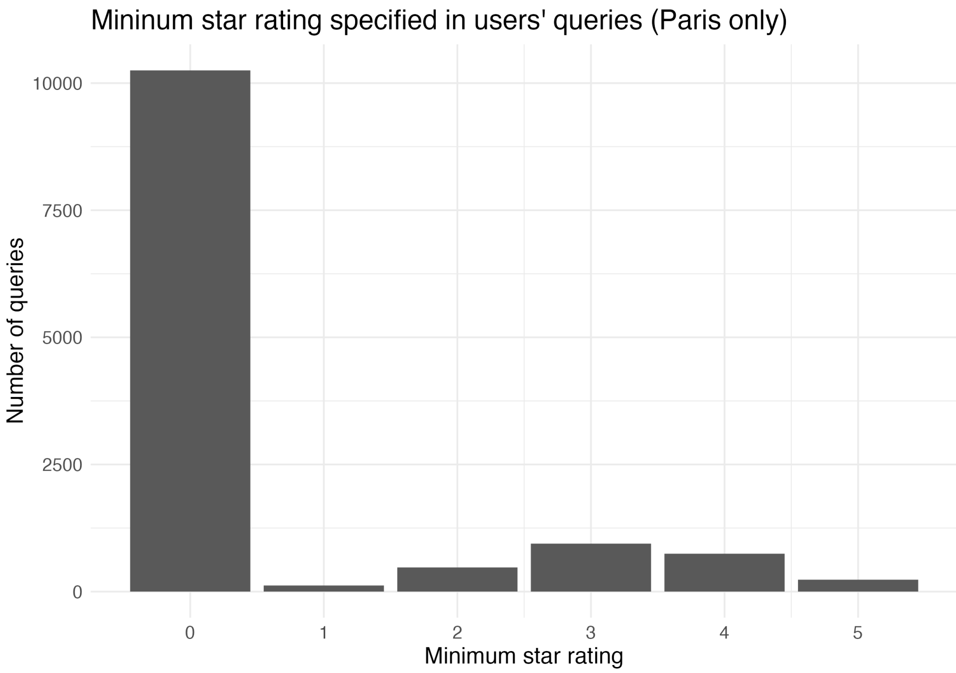


Figure 2.17: Minimum star rating specified by user.

2.D.2 Hotels' pricing decisions

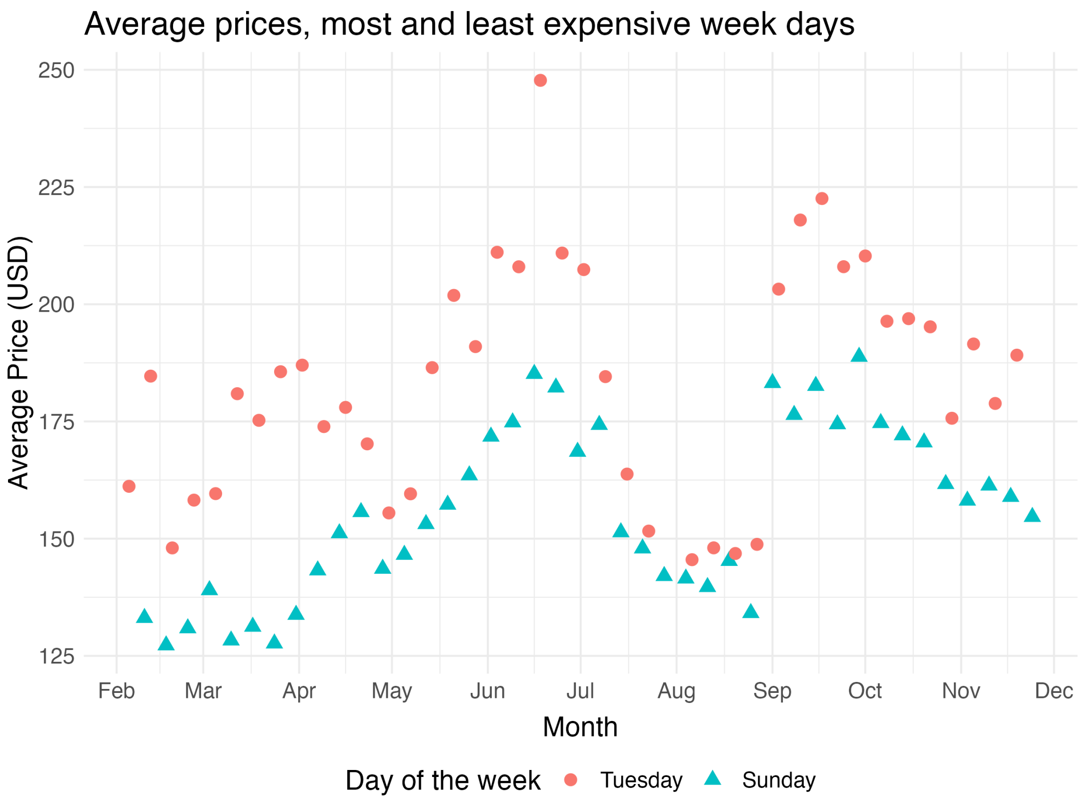


Figure 2.18: Average prices for most and least expensive days of the week.

2.D.3 Hotels' promotion decisions

Hotels on Expedia are able to offer various kinds of special promotions, some of which are opaque even after the investigating these. First, hotels can be displayed with a green "Sale!" flag (as seen in Figure 2.1). When hovering above the flag, a toll tip may *or may not* open up that describes the promotion in place. The promotion in place may for instance be a a discount given on the price, or a special "Expedia rate".[40] Aside from the "Sale!" flag, other flags or highlighted pieces of text can be seen, such as "Tonight only", special membership rates, or free extras. Table 2.17 displays the fraction of hotels in a given city that ever offer a given type of discount. Overall, 84% of hotels in Paris are observed offering *any* kind of promotion across queries.

Figure 2.19 considers the average fraction of hotels offering free cancellation. In general, 2-star hotels are the most likely to offer free cancellation, followed by 1-star hotels, although there is considerable variation for these types of hotels. 3-, 4-, and 5-star hotels are the least likely to offer free cancellation. Interestingly, for short booking windows, free cancellation seems to be offered *more* frequently for 4- and 5-star hotels and subsequently declines, but *less* often for 1-, 2-, and 3-star hotels; i.e., the gradient

[40]Expedia wrote, as of 2019: "Expedia Rate properties may qualify for special promotions and coupon redemption. Expedia Rate requires that your credit card be charged for the full payment upon reservation. Special offers may apply to specific room/unit types and have additional terms and conditions."

Table 2.17: Percentage of hotels that *ever* offer a given promotion in any given query.

City	# Hotels	% of hotels ever offering given promotion							
		"Sale" tag	Dicount (in %) >0	"Tonight only"	Membership rate	Free cancellation	Free extra	Expedia rate	*Any promo*
Budap	1,045	36.2	35.9	0.8	37.6	49.4	2.1	60.4	65.8
Canc	972	23.8	22.7	0.4	23.0	38.9	8.4	32.6	47.4
Manh	1,222	24.5	23.6	0.3	24.1	48.7	2.2	51.8	58.6
Paris	2,842	48.6	46.6	0.5	26.2	63.4	11.7	78.4	84.1

at the beginning of the observation period differs between the different groups of hotels.

Figure 2.20 shows that the percentage of hotels offering "Expedia rate" is smooth over time, and increasing in the number of stars. (The exception are hotels with 1-star ratings, for which this measure is very noisy due to a low number of observations, and thus not plotted.)

Figure 2.21 plots the average discounts offered by hotels, conditionally on offering any discount greater than zero. There is essentially no variation in the discount rate for hotels of 2 or more stars, neither over time nor between the groups. In contrast, 1-star hotels seem to behave very heterogeneously, although this finding could again be partly driven by a low number of observations.

All in all, these plots may give us some insight into how hotels engage in dynamic revenue management over booking windows. However, as mentioned in Appendix 2.B, in addition to the opacity, there are several conceptual difficulties regarding hotels' decisions of whether to offer any promotion. I therefore do not endogenize hotels' promotion decisions in my main model.

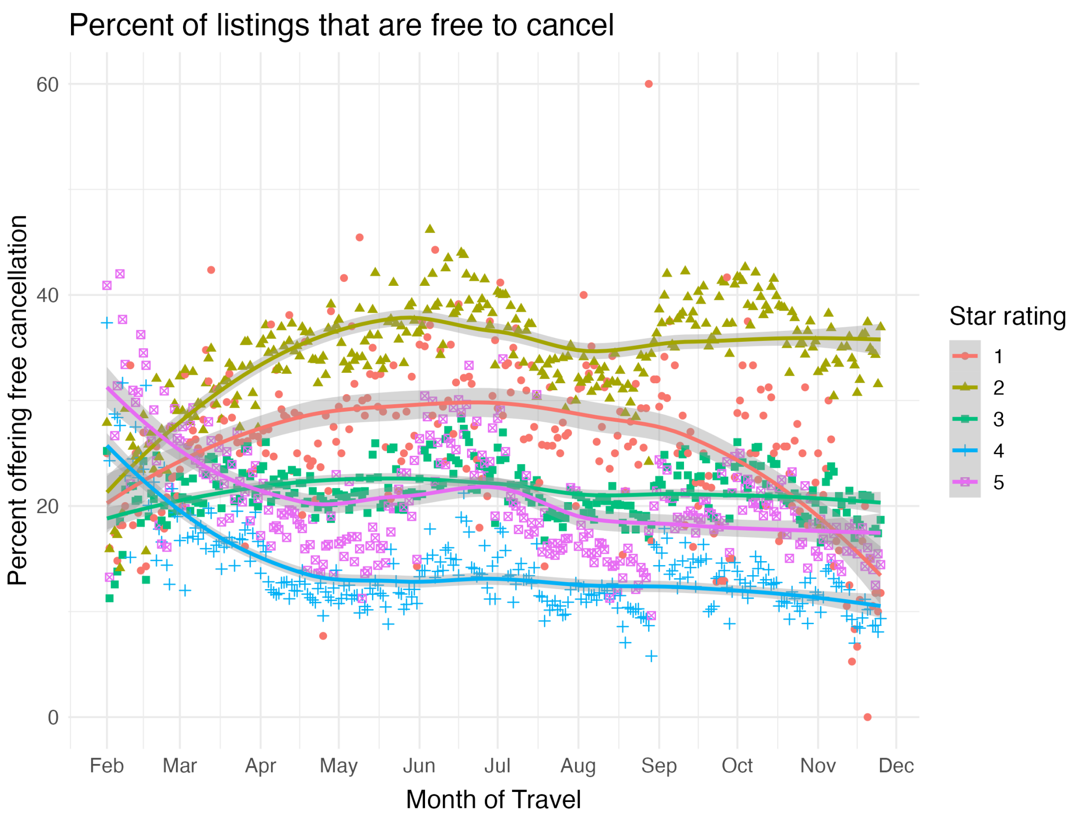


Figure 2.19: Average percent of listings offering free cancellation at a given travel date for hotels in Paris. Locally estimated scatterplot smoothing is used for the lines. Note that for observations in February and March, the booking window is very short.

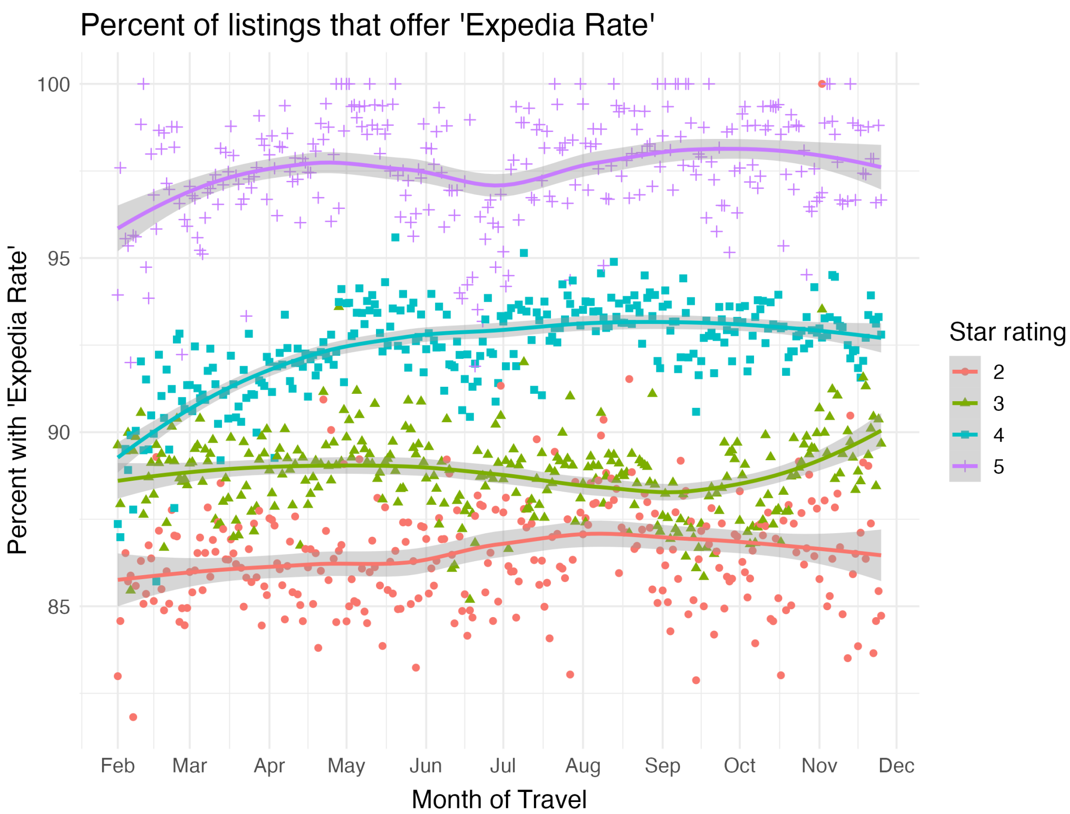


Figure 2.20: Average percent of listings with "Expedia Rate" at a given travel date for hotels in Paris. Locally estimated scatterplot smoothing is used for the lines. Note that for observations in February and March, the booking window is very short.

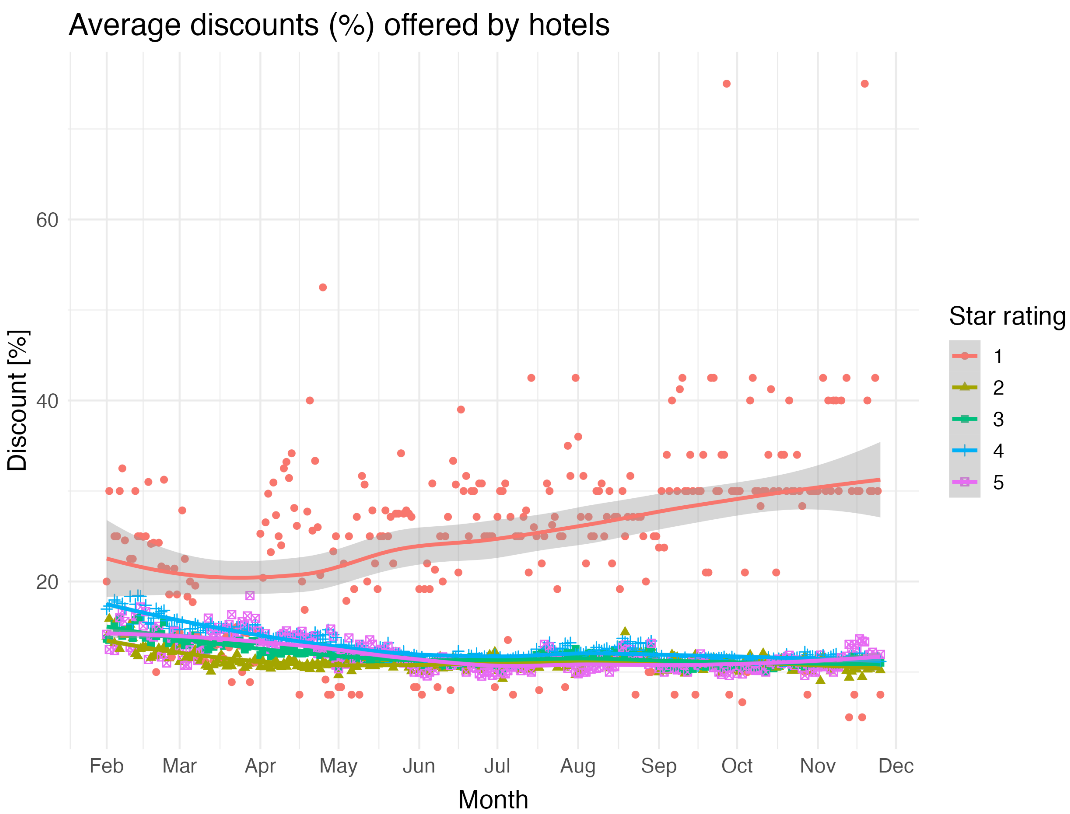


Figure 2.21: Average discount offered by hotels (in percent) for a given travel date for hotels in Paris. Locally estimated scatterplot smoothing is used for the lines. Note that for observations in February and March, the booking window is very short.

2.E First Stages

Table 2.18: First stages, neighborhood-based instruments

	Dependent variable:		
	price		sale
	(1)	(2)	(3)
avg_neighb_price	0.300***	0.301***	−0.0001***
	(0.065)	(0.066)	(0.00001)
avg_neighb_sale		1.945	0.052***
		(1.888)	(0.004)
neighb_count	−0.348***	−0.347***	0.0003***
	(0.054)	(0.054)	(0.0001)
pastbookings_count	−0.082***	−0.082***	0.001***
	(0.023)	(0.023)	(0.0001)
Query×travel-date FE	✓	✓	✓
Hotel FE	✓	✓	✓
F-stat excl. instruments	104.2	77.25	85.58
Observations	914,814	914,814	914,814
Adjusted R^2	0.821	0.821	0.659

Standard errors clustered at query×travel-date level. *p<0.1; **p<0.05; ***p<0.01

Neighboring hotels are defined to be within a 500m radius of focal hotel

Table 2.19: First stages, neighborhood-based instruments ("donut" definition)

	Dependent variable:		
	price		sale
	(1)	(2)	(3)
avg_neighb_price	0.287***	0.287***	−0.00004**
	(0.078)	(0.078)	(0.00002)
avg_neighb_sale		0.746	0.082***
		(1.387)	(0.005)
neighb_count	−0.165***	−0.165***	0.0003***
	(0.029)	(0.029)	(0.0001)
pastbookings_count	−0.081***	−0.081***	0.001***
	(0.022)	(0.022)	(0.0001)
Query×travel-date FE	✓	✓	✓
Hotel FE	✓	✓	✓
F-stat excl. instruments	54.05	40.52	94.99
Observations	927,322	927,322	927,322
Adjusted R^2	0.819	0.819	0.661

Standard errors clustered at query×travel-date level. *p<0.1; **p<0.05; ***p<0.01

Neighboring hotels are defined to be located within a donut-shaped ring ∈[500m, 1,000m) radius of focal hotel.

Table 2.20: First stages, Airbnb and 1{august}×zip code instrument

	Dependent variable:		
	mean_price		mean_sale
	(1)	(2)	(3)
airbnb_p_zip	2.050***	2.013***	−0.002***
	(0.086)	(0.082)	(0.0001)
airbnb_avail_zip	−0.015**	−0.013*	0.00002
	(0.007)	(0.007)	(0.00002)
mean_pastbooking	−0.022	−0.023	0.0004**
	(0.041)	(0.041)	(0.0002)
zip code×1{August}		✓	✓
Travel date FE	✓	✓	✓
Hotel FE	✓	✓	✓
F-stat excl. instruments	286.99	117.88	72.91
Observations	447,378	447,378	447,378
Adjusted R^2	0.827	0.827	0.752

Standard errors clustered at travel date level. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 2.21: First stages, brand-based instruments

	Dependent variable:		
	price	mean_price	sale
	(1)	(2)	(3)
avg_price_brand	0.259***	0.258***	−0.00003***
	(0.035)	(0.035)	(0.00001)
avg_sale_brand		−4.808***	0.425***
		(1.413)	(0.009)
pastbookings_count	−0.012	−0.011	0.001***
	(0.041)	(0.041)	(0.0001)
Query×travel-date FE	✓	✓	✓
Hotel FE	✓	✓	✓
F-stat excl. instruments	54.28	77.93	1126.23
Observations	258,011	258,011	258,011
Adjusted R^2	0.661	0.661	0.663

Standard errors clustered at query×travel-date level. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 2.22: First stages, Arellano-Bond type instruments

	Dependent variable:		
	mean_price		mean_sale
	(1)	(2)	(3)
mean_price_diff2	0.134***	0.134***	−0.00001
	(0.029)	(0.030)	(0.00001)
mean_sale_diff2		−0.706	0.096***
		(0.669)	(0.016)
mean_pastbooking	−0.025	−0.025	0.0003*
	(0.034)	(0.034)	(0.0002)
Checkin date FE	✓	✓	✓
Hotel FE	✓	✓	✓
F-stat excl. instruments	20.79	36.28	21.62
Observations	451,366	451,366	451,366
Adjusted R^2	0.967	0.967	0.773

Standard errors clustered at checkin date level. *p<0.1; **p<0.05; ***p<0.01

2.F Robustness Checks

Table 2.23: Results using neighborhood instruments, employing more aggregate hotel-travel date panel

	Dependent variable:			
	average position across queries			
	(1)	(2)	(3)	(4)
price	2.069***	3.124***	2.041***	3.191***
(average across queries)	(0.340)	(0.108)	(0.338)	(0.121)
1{sale}			−70.597***	454.136**
(average across queries)			(8.036)	(180.621)
# bookings past 48h	−0.736**	−0.710**	−0.712**	−0.864**
(average across queries)	(0.318)	(0.299)	(0.310)	(0.386)
Travel-date FE	✓	✓	✓	✓
Hotel FE	✓	✓	✓	✓
1st stage F-stat of excl. instruments on price		122.66		134.98
1st stage F-stat of excl. instruments on sale				45.65
Observations	508,956	508,956	508,956	508,956
Adjusted R^2	0.783	0.774	0.784	0.734

Standard errors clustered at travel date level. *p<0.1; **p<0.05; ***p<0.01

The dependent variable is hotel j's average position in Expedia's listings page for travel date t across web-scraped queries. Column (2) uses as instruments for price the number of neighboring hotels available on a given query and travel date, and the average price of neighboring hotels at a given travel and query date. Column (4) uses as instruments for price and the sale indicator the number of neighboring hotels and the average price and sales indicator. Neighboring hotels are defined as being located within 500 meters of the focal hotel.

Table 2.24: Results using brand-based instruments, employing more aggregate hotel-travel date panel

	Dependent variable:			
	average position across queries			
	(1)	(2)	(3)	(4)
price	1.783***	3.768***	1.758***	3.750***
(average across queries)	(0.312)	(0.176)	(0.310)	(0.177)
1{sale}			−75.236***	89.926**
(average across queries)			(7.705)	(35.480)
# bookings past 48h	−0.585**	−0.553**	−0.566**	−0.577**
(average across queries)	(0.280)	(0.249)	(0.273)	(0.260)
Travel-date FE	✓	✓	✓	✓
Hotel FE	✓	✓	✓	✓
1st stage F-stat of excl. instruments on price		99.98		50.28
1st stage F-stat of excl. instruments on sale				453.18
Observations	562,732	562,732	562,732	562,732
Adjusted R^2	0.798	0.765	0.799	0.764

Standard errors clustered at travel date level. *p<0.1; **p<0.05; ***p<0.01

The dependent variable is hotel j's average position in Expedia's listings page for travel date t across web-scraped queries. Column (2) uses as instruments for price the average price charged by other hotels in Paris of the same brand. Column (4) uses as instruments for price and the sales indicator the average price and average sales indicator employed by other hotels in Paris of the same brand.

Table 2.25: Linear probability model, using neighborhood-based instruments

	Dependent variable:			
	P(1{position ≤ 10)			
	(1)	(2)	(3)	(4)
price	−0.0001***	−0.00004***	−0.0001***	−0.00004***
	(0.00001)	(0.00001)	(0.00001)	(0.00001)
1{sale}			0.004***	−0.011
			(0.0003)	(0.009)
# bookings past 48h	0.001***	0.001***	0.001***	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Query-travel date FE	✓	✓	✓	✓
Hotel FE	✓	✓	✓	✓
1st stage F-stat of excl. instruments on price		104.2		77.25
1st stage F-stat of excl. instruments on sale				85.58
Observations	914,814	914,814	914,814	914,814
Adjusted R^2	0.173	0.172	0.173	0.170

Standard errors clustered at query×travel-date level. *p<0.1; **p<0.05; ***p<0.01

Table 2.26: Linear probability model, using neighborhood-based instruments

	Dependent variable:			
	P(1{position $\leq$ 5)			
	(1)	(2)	(3)	(4)
price	-0.00003^{***}	-0.00002^{***}	-0.00003^{***}	-0.00002^{***}
	(0.00000)	(0.00001)	(0.00000)	(0.00001)
1{sale}			0.002^{***}	-0.003
			(0.0002)	(0.007)
# bookings past 48h	0.0004^{***}	0.0004^{***}	0.0004^{***}	0.0004^{***}
	(0.00004)	(0.00004)	(0.00004)	(0.00004)
Query-travel date FE	✓	✓	✓	✓
Hotel FE	✓	✓	✓	✓
1st stage F-stat of excl. instruments on price		104.2		77.25
1st stage F-stat of excl. instruments on sale				85.58
Observations	914,814	914,814	914,814	914,814
Adjusted R^2	0.142	0.142	0.142	0.141

Standard errors clustered at query×travel-date level. *p<0.1; **p<0.05; ***p<0.01

References

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data:monte carlo evidence and an application to employment equations. *Review of Economic Studies*, *58*(2), 277–297.

Bar-Isaac, H., & Shelegia, S. (2022). Monetizing Steering. *Working Paper*, 1–25. Retrieved from https://www.dropbox.com/s/d3l0zv4kvhfmoox/Monetizing.pdf?dl=0

Cazaubiel, A., Cure, M., Johansen, B. O., & Vergé, T. (2020). Substitution between online distribution channels: Evidence from the Oslo hotel market. *International Journal of Industrial Organization*, *69*, 102577. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167718719301055

Chen, L., Mislove, A., & Wilson, C. (2016). An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace. *WWW '16: Proceedings of the 25th International Conference on World Wide Web*, 1339–1349.

Chen, Y., & Yao, S. (2016). Sequential Search with Refinement: Model and Application with Click-Stream Data. *Management Science*, *63*(12), 4345–4365. Retrieved from https://doi.org/10.1287/mnsc.2016.2557 doi: 10.1287/mnsc.2016.2557

Cho, S., Lee, G., Rust, J., & Yu, M. (2018). Optimal Dynamic Hotel Pricing. *Working Paper*. Retrieved from https://economics.sas.upenn.edu/index.php/system/files/2018-04/hp_final_update.pdf

De los Santos, B., & Koulayev, S. (2017). Optimizing click-through in online rankings with endogenous search refinement. *Marketing Science*, *36*(4), 542–564.

DellaVigna, S., & Gentzkow, M. (2019). Uniform Pricing in U.S. Retail Chains. *The Quarterly Journal of Economics*, *134*(4), 2011–2084.

Dinerstein, M., Einav, L., Levin, J., & Sundaresan, N. (2018). Consumer Price Search and Platform Design in Internet Commerce. *American Economic Review*, *108*(7), 1820–1859.

Ellison, G., & Ellison, S. F. (2009). Search, Obfuscation, and Price Elasticities on the Internet. *Econometrica*, *77*(2), 427–452.

Ershov, D. (2022). Variety-Based Congestion in Online Markets: Evidence from Mobile Apps. *Forthcoming in American Economic Journal: Microeconomics*. Retrieved from https://drive.google.com/file/d/1behYlPeLP0XVO6591U0oWm3KbwC4_6bU/view

European Commission. (2016). *Report on the monitoring exercise carried out in the online hotel booking sector by EU competition authorities in 2016* (Tech. Rep.). Retrieved from https://ec.europa.eu/

competition/ecn/hotel_monitoring_report_en.pdf

Fradkin, A. (2017). *Search, Matching, and the Role of Digital Marketplace Design in Enabling Trade: Evidence from Airbnb.* Retrieved from http://andreyfradkin.com/assets/SearchMatchingEfficiency.pdf

Glick, M., Richards, G., Sapozhnikov, M., & Seabright, P. (2014). How Does Ranking Affect User Choice in Online Search? *Review of Industrial Organization*, *45*(2), 99–119.

Hagiu, A., & Jullien, B. (2011). Why do intermediaries divert search? *The RAND Journal of Economics*, *42*(2), 337–362.

Hannak, A., Sapiezynski, P., Khaki, A. M., Lazer, D., Mislove, A., & Wilson, C. (2017). Measuring Personalization of Web Search. *Proceedings of the 22nd International Conference on World Wide Web*. Retrieved from http://arxiv.org/abs/1706.05011 doi: 10.1145/2488388.2488435

Hunold, M., Kesler, R., & Laitenberger, U. (2020). Rankings of Online Travel Agents, Channel Pricing, and Consumer Protection. *Marketing Science*.

Hunold, M., Kesler, R., Laitenberger, U., & Schlütter, F. (2018). Evaluation of best price clauses in online hotel bookings. *International Journal of Industrial Organization*, *61*, 542–571.

Kim, J. B., Albuquerque, P., & Bronnenberg, B. J. (2010). Online Demand Under Limited Consumer Search. *Marketing Science*, *29*(6), 1001–1023.

Koulayev, S. (2014). Search for differentiated products: identification and estimation. *The RAND Journal of Economics*, *45*(3), 553–575.

Lee, K. H., & Musolff, L. (2021). Entry Into Two-Sided Markets Shaped By Platform-Guided Search. *Working Paper*. Retrieved from https://lmusolff.github.io/papers/Entry_and_Platform_Guided_Search.pdf

Li, J., Netessine, S., & Koulayev, S. (2017). Price to Compete … with Many: How to Identify Price Competition in High-Dimensional Space. *Management Science*, *64*(9), 4118–4136. Retrieved from https://doi.org/10.1287/mnsc.2017.2820 doi: 10.1287/mnsc.2017.2820

Mantovani, A., Piga, C. A., & Reggiani, C. (2021). Online platform price parity clauses: Evidence from the EU Booking.com case. *European Economic Review*, *131*, 103625. Retrieved from https://www.sciencedirect.com/science/article/pii/S0014292120302555

Schaefer, M., & Tran, K. D. (2020). Airbnb, Hotels, and Localized Competition. *Working Paper*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3676907

Teng, X. (2022). Self-preferencing, Quality Provision, and Welfare in Mobile Application Markets. *Working Paper*.

Ursu, R. M. (2018). The Power of Rankings: Quantifying the Effect of Rankings on Online Consumer Search and Purchase Decisions. *Marketing Science*, *37*(4), 530–552. Retrieved from https://doi.org/10.1287/mksc.2017.1072

Weitzman, M. L. (1979). Optimal Search for the Best Alternative. *Econometrica*, *47*(3), 641–654. Retrieved from http://www.jstor.org/stable/1910412

Chapter 3

Privacy Regulation and Startup M&As

Abstract

This paper explores how privacy regulation can affect merger and acquisition (M&A) activity. In principle, privacy regulation can, on the one hand, affect the profitability mergers and acquisitions *negatively*: privacy regulation likely increases the cost of due diligence, and privacy regulation can prevent merging companies to combine their data post acquisition. On the other hand, mergers can also be thought of a means that enable data sharing between companies in environments with high privacy protection that make sharing data with third parties unlawful, so that privacy regulation can also *positively* affect the profitability of mergers and acquisitions. To investigate the link between privacy regulation and M&As, I study the likely effects of the General Data Protection Regulation (GDPR) on M&As of startup firms using a difference-in-differences framework. My preliminary findings suggest a decline in M&As of EU-based startups following the introduction of GDPR. Interestingly, this decline seems to be driven by domestic acquirers, thus contradicting anecdotal evidence. I propose follow-up research questions and further investigations.

1 Introduction

With the growing significance of data for the economy, safeguarding private data has become a prominent topic in public policy discussions.[1] Various policy initiatives to protect citizens' personal data have been proposed and implemented in several jurisdictions, including the European Union's General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), or Brazil's General Data Protection Law (LGPD). These policies impose specific duties and requirements on firms that store or handle personal data, and have thus affected firms' business environments.

This research aims at examining the impact of privacy regulation on major strategic decisions that firms take, namely, on merger and acquisition (M&A) activity. M&As by definition change ownership and control of companies. They are determinants of market structure and competition, and can therefore affect industry dynamics and consumer welfare in significant ways. For the time being, this paper focuses on M&As of venture capital (VC)-funded startups. For these highly innovative firms, the prospect of being acquired is likely a driver of new entry[2]. In the software industry for instance, a vast majority of successful exits for startups occur through acquisition (Eisfeld, 2023). At the same time, many startups' business models rely on collecting large amounts of data. Young startups may moreover lack the scale for more structured internal processes, or for teams of lawyers that may give guidance in navigating new regulations. Findings by Jia, Jin, and Wagman (2021) suggest that young firms may thus be particularly negatively impacted by privacy regulations.

Privacy regulation can affect M&A activity either positively or negatively. Firstly, upon an acquisition, the acquiring firm assumes the target firm's liabilities with respect to privacy laws. To limit any risk, the acquiring firm is required to conduct due diligence on all aspects of data protection. This increases the administrative burden and costs of a merger, reducing its net value to the acquiring firm. The risks associated with unknowingly taking on previous undetected liabilities via M&As are illustrated by the 18.4 million pound fine imposed on Marriott in 2020, following a data breach of a previously acquired target firm, Starwood, that had occurred in 2014.[3] Moreover, legal practitioners have in the past emphasized that privacy regulations are a burden on M&As, which is corroborated by survey evidence of merging parties (see below). Secondly, privacy regulations may restrict the merged entity's ability to combine different datasets post acquisition. The value of a merger, however, may be based on the data that a company has collected. If sharing data between the entities is deemed unlawful, or if sharing of data is only possible after implementing certain technical or organizational measures, this could negatively affect the value of the merger. However, thirdly, mergers can also be thought of as facilitating data sharing between companies in an environment of high privacy protection. When privacy regulations prohibit sharing of data with third parties, one may suppose that data can instead be shared legally within the confines of a firm, rather than across firms' boundaries. Viewed from this perspective, privacy regulation can in principle incentivize firms to conduct more data-driven mergers.

These opposing forces warrant an empirical investigation of the impact of privacy regulation on

[1] According to the Google Books Ngram Viewer, the terms "privacy" or "data protection" have never in history been employed in books more than nowadays, see `https://tinyurl.com/ytehwcm4` (accessed 05/04/2023).

[2] See literature on the entry-for-buyout effect.

[3] The merger between Marriott and Starwood took place only in 2016, i.e., the data breach occurred prior to the merger.

M&A activity, and this is the purpose of this paper. I do so by focusing on the GDPR, an impactful privacy regulation implemented in the European Union (EU). Enacted in April 2016 and enrolled in May 2018, the GDPR has harmonized data protection across the EU and imposes strict rules on the processing, collection, access and portability on any personal data from EU citizens. With the associated fines in case of infractions being very high, and many fines being levied as of now[4], the GDPR has had real bite, and thus can be viewed as a very relevant change to the business environment of firms with any such data. Previous research has shown that GDPR has impacted the economics of collecting and using data, in some cases to the detriment of especially smaller firms (for a survey, see Johnson (2022)).

To study the effect of the GDPR on startup M&As, I conduct a difference-in-difference analysis, comparing the number of acquisitions of EU-based startups to acquisitions of American or other non-EU startups pre and post enactment (and enrollment, respectively) of the GDPR. I find evidence for a decline in acquisitions of startups based in the EU after GDPR enrollment compared to the control groups. Further, motivated by evidence that privacy regulation has limited transfers of data outside the boundaries of a jurisdiction, I examine how the GDPR has affected mergers in which the acquirer is a domestic firms, as opposed to mergers where the acquiring firm is located outside the union (outside the EU, or USA or rest of world, respectively). Interestingly, my findings suggest that the decline may be driven by fewer domestic mergers.

I suggest follow-up questions for subsequent investigations. First, there is likely to be a significant amount of heterogneity in the extent to which industries are impacted by the GDPR. I therefore propose to investigate heterogeneous effects between industries, which would corroborate my findings. Second, not only might fewer M&As taking place: deals might also take longer to close, or may be conducted at lower prices. Answering this last question requires additional data, which I have only begun to analyze. Third, an important follow-up question is whether the GDPR has affected competition between firms and market dynamics in significant ways.

Related Literature

Prior research has highlighted that the GDPR has inhibited transfer and trade of data between firms; see, e.g., Johnson, Shriver, and Goldberg (n.d.), and Peukert, Bechtold, Batikas, and Kretschmer (2022). The fact that privacy laws make purchases of data more difficult is also reflected in Canayaz, Kantorovitch, and Mihet (2022). In the context of the Californian CCPA and voice-AI firms that heavily rely on data, these authors find that privacy regulation particularly hurt small firms with small consumer bases who are heavily reliant on purchasing external data.

Jia et al. (2021) provide the most closely related research to mine. The authors offer empirical evidence that the enrollment of the GDPR led to a relative decline in startup entry and VC funding, and that this effect is stronger for startups active in data related and business-to-consumer industries. In a working paper, Lambrecht (2017) studies the implications of the EU's e-Privacy Directive enacted in 2002 for startups active in the industries of online news, online advertisement, and cloud computing. She similarly finds that VC investment into startups based in Europe grew more slowly than investments

[4]See `https://www.enforcementtracker.com` (last accessed 02/05/2023).

into their US-based counterparts. As VC investors' willingness to fund startups depends on exit opportunities, the reason for the decline detected in those papers might in fact be a less optimistic prospect of a buyout, which motivates my analysis. In a further paper, Jia, Jin, and Wagman (2020) find that the GDPR led VC investment to become more local. The authors highlight information frictions and legal uncertainties with respect to the GDPR as a source. Those findings, as well as anecdotal evidence that the GDPR has negatively affected the legal circumstances of cross-border data flows, motivates my additional analysis of cross-border mergers.

De Cornière and Taylor (2021) as well as Condorelli and Padilla (2022) theoretically study data-driven mergers, where the main purpose of the merger activity is to gain access to a new dataset. De Cornière and Taylor (2021) find that the effect of data-driven mergers for consumer surplus depends on (1) whether or not data is pro-competitive or anti-competitive, and (2) whether data trade would be possible absent the merger. Condorelli and Padilla (2022) show conditions under which acquiring data confers an advantage that can inhibit follow-on entry. Note that in the setting studied in both of these papers, acquiring data is in fact the main *purpose* of the merger.

Surveys with startup founders and VC investors suggest GDPR compliance had become a key criterion for selecting suppliers and business partners (Martin, Matt, Niebel, & Blind, 2019).[5] This suggests that, in a similar vein, GDPR compliance may also be of importance during M&A transactions. For a thorough survey of economic research on the effects of the GDPR, see Johnson (2022).

This research is also related to the interaction of competition law and privacy law, as highlighted by the Facebook case in Germany (see Kerber and Zolna (2022)).

2 Background and anecdotal evidence

2.1 Anecdotal evidence that the GDPR likely increased costs & relevance of due diligence

Blog articles and reports from law firms or consultancies suggest that ensuring GDPR compliance during M&As is costly and complicated. First, to mitigate potential risks, the acquiring firm must ensure that the target firm complies with the GDPR. This likely increases the importance of conducting thorough due diligence, which may result in increased costs. Second, GDPR compliance must be ensured whenever information is exchanged between merging parties at any point during the merging process. Specific provisions may need to be taken, which again may increase the costs associated with the the transactions. The below quotes highlight these issues.

- "We are currently seeing that the EU's General Data Protection Regulation (GDPR) has become a major factor in mergers and acquisition (M&A) transactions, adding complexity to the due diligence process and sometimes even causing deals to have issues. [...] Data security has therefore quickly become a top priority of the due diligence process and suggestions are that the GDPR impact on M&A due diligence will, over the next five years, cause even greater scrutiny of data

[5] The same survey suggests that data privacy enforcement had become significantly tougher with GDPR, and non-compliance was not a viable strategy for startups.

protection policies and process of target companies by potential acquirers." – See https://tinyurl.com/2p89w2ra (accessed 06/04/2023).

- Written on April 4, 2022: "A 'standard market' approach has not yet been developed. [...] The issue of data protection is also present in transactions and, as the fines imposed show, is also highly relevant and explosive. However, there are still no clear guidelines on how to deal with these regulations in corporate transactions. In this respect, it will probably always depend on the individual case and be necessary to examine precisely which personal data may be processed to what extent and in what manner in order not to be exposed to an excessive liability risk." – See https://tinyurl.com/2p8hwa9x (accessed 06/04/2023).

- "Transferring personal data has always been subject to data protection law, but the GDPR brings increased penalties for breach of its provisions and more onerous requirements to demonstrate compliance. [...] GDPR is a complex area" – See https://tinyurl.com/37hhut67 (accessed 28/04/2023).

- "This increase in data combined with intensifying global data protection regulations is introducing new responsibilities and elevated risks around M&A activity. [...] sometimes concessions [requested by competition authorities] conflict with privacy law – if a transaction reaches that point, the parties may have a serious uphill battle in taking the deal across the finish line. [...] This potentiality reinforces the critical nature of a comprehensive due diligence before a deal is submitted to the authorities for approval. [...] The regulatory landscape is fast evolving, very complex and when intersected with data issues, the risks for M&A are significantly multiplied. Data can bring tremendous value, but it can also be a long-term privacy risk." – See https://tinyurl.com/yc6eazxt (accessed 06/04/2023).

- "To the extent that the data processing activities of the target company are not in line with the GDPR, buyers may face significant risks." – See https://tinyurl.com/ym7vdheh (accessed 06/04/2023).

- According to a survey of M&A professionals in Europe conducted by the Merrill Corporation and Euromoney, merger deals frequently do not progress because of concerns related to GDPR, see Appendix 3.A, Figure 3.4. A majority of these professionals believes that the GDPR will increase acquirers' scrutiny of data protection policies of target firms, see Appendix 3.A, Figure 3.5.

- A blog post on the Harvard Law School Forum on Corporate Governance suggests that privacy-related liabilities of acquisition targets can pose a significant risk, and suggests that pre-merger, due diligence and inquiry into privacy issues is important. See https://tinyurl.com/42k94xp5 (accessed 06/04/2023).

The case of Marriott and Starwood mentioned above illustrates the high risk involved of purchasing a company without having conducted extremely careful due diligence with regards to data protection. Additional legal advice on privacy-related matters may be advisable.

2.2 Evidence suggesting that the GDPR may have inhibited the sharing of data post-acquisition

Further, there is anecdotal evidence suggesting that the GDPR may have obstructed the exchange of data between firms that merged. The first piece of evidence comes from a blog post on the Harvard Law School Forum on Corporate Governance by Daniel Ilan[6]. The author highlights that due to privacy protection, a lawful transfer of data from one merging party to another can be made more costly or even be prevented. Problems can arise, on one hand, if the acquiring firm has a *less* robust privacy policy than the target firm, but wishes to integrate personal data collected by the target firm. In this case, the author writes "the purchaser may need to bring its own data privacy practices into compliance with the target's applicable privacy policy", and highlights the Facebook-WhatsApp case. Second, problems also arise if the acquiring firm has a *more* robust privacy policy than the target firm. The author notes that "The most reasonable approach will likely be for the purchaser to either (1) maintain the target as a separate entity/division that does not use purchaser's data or (2) bring the target's practices into compliance with purchaser's previous promises (though this could involve significant costs)." In the EU, data from the merging partner may not be used for a different purpose than what it was originally collected for.

In the Microsoft-LinkedIn decision, the European Commission notes that "[The GDPR] may further limit Microsoft's ability to have access and to process its users' personal data in the future since the new rules will strengthen the existing rights and empowering individuals with more control over their personal data (i.e. easier access to personal data; right to data portability; etc.)." – See `https://tinyurl.com/5n8zkwj5` (accessed 06/04/2023).[7]

2.3 M&As might enable data sharing

Mergers can also be viewed as a means of *facilitating* the exchange and use of data between two firms. Chen, Choe, Cong, and Matsushima (2022), Condorelli and Padilla (2022), and De Cornière and Taylor (2021) focus on the competitive effects of these so-called data-driven mergers. A report by the UK Competition and Markets Authority suggests that data can likely be shared easier within large entities than across firms[8]. If this was the case, then data protection regulation that limits the trade of data *across* firms' boundaries could in some circumstances lead to *more* mergers. Thus, with the advent of the GDPR, firms could have substituted from buying data collected by third parties to acquiring those data-collecting businesses.

[6]See `https://tinyurl.com/hrxrnb87` (accessed 06/04/2023)

[7]The European Commission makes similar remarks on the Apple-Shazam as well as other mergers in digital technology markets. See `https://tinyurl.com/4k8pp7j7` (accessed 07/04/2023).

[8]"Online platforms and digital advertising: Market study interim report", p.15, see `https://tinyurl.com/bdhhry3k` (accessed 06/04/2023).

3 Data and descriptive analysis

I obtain data from *Crunchbase*, a data portal that records all important events of both public and private firms worldwide.[9] VC investors and industry experts as well as academic scholars use this data portal to track all major events such as funding rounds and exits of VC-funded firms. From *Crunchbase*, I filter all acquisitions of VC-funded firms, hereafter *startup acquisitions*.[10] I extract variables such as announcement date, firms' headquarter locations, acquisition price (if available), number of employees (as of the date of data download), and two sets of industry tags (wide and narrow). I additionally construct variables indicating the number of previous funding rounds, the amount of money raised at time of acquisition, and company age at acquisition (using either the first event, or the recorded founding date, as possible "birth" dates). I obtain information on acquisitions of startups that have their headquarters in one of three geographic areas:

- Startups with headquarters in any of 30 countries in the European Economic Area (EEA) (including the United Kingdom, which was still part of the EU until the beginning of 2020), as well as Switzerland. For simplicity, I call this group "EEA".
- Startups with headquarters in any of the 50 states of the USA or the District of Columbia.
- Startups with headquarters in any of eight non-US OECD member states which recorded at least six startup acquisitions in 2014-2019: Canada, Korea, Japan, Israel, Australia, Turkey, Chile, and Mexico. For simplicity, I call this control group "Rest of the World" (RoW).

I construct a balanced panel at the month-state level for 2014-2019 with 6,408 observations. "State" here refers to countries for the EU or the RoW, and to federal States for data from the US. For each month and state, I count the number of startup acquisitions that occurred in that period.[11] All countries and their respective number of acquisitions are listed in Table 3.13 in Appendix 3.B.

Table 3.1 details the distribution of the number of acquisitions across three regions. Notably, startup acquisitions in the USA account for 66% of all startup acquisitions in the geographic regions under consideration.[12] It is striking that the number of startup acquisitions in the US is clearly more than twice as high as in the EEA. As the GDP of the US is not being twice as high as the GDP of the EEA, this highlights the comparably high level of VC-funded entrepreneurship in the USA.

Table 3.1: Number of startup acquisitions, for startups located in a given region, 2014-2019.

Region		Number of Acquisitions	Percent
EEA plus CHE	(treatment)	2134	25.58
USA	(control 1)	5498	65.92
RoW	(control 2)	709	8.50
Total		8341	100

[9]It is believed to focus in particular on US- and Europe-based startups.

[10]I use well-established notions of what is a VC-funding round; further details are explained in Appendix 3.B. *Crunchbase* only takes into accounts majority acquisitions.

[11]*Crunchbase* records the announcement date as "acquisition date".

[12]See Table 3.13 for a detailed overview of the countries contained in the sample.

Table 3.2 shows the distribution of the number of acquisitions across month-states for each of the three geographic areas. It is striking – and, in fact, concerning – that the number of acquisitions is highly right skewed in particular for the USA.

Table 3.2: Distribution of startup acquisitions across the month-state panel, 2014-2019.

Region	Min	25th perc	Median	Mean	75th perc	95th perc	Max	SD
EEA	0	0	0	0.96	1	6	14	2.04
USA	0	0	0	1.53	1	6	42	4.48
RoW	0	0	0	1.23	2	6	11	1.99

To get an initial idea of the M&A activity across the three regions and over time, Figures 3.1 shows the monthly total number of startup acquisitions over a longer horizon, summed across states (or countries) in a given region and month. All three regions display a somewhat upward trend until the end of 2019, where we observe a decline in acquisitions in particular for US-based startups.

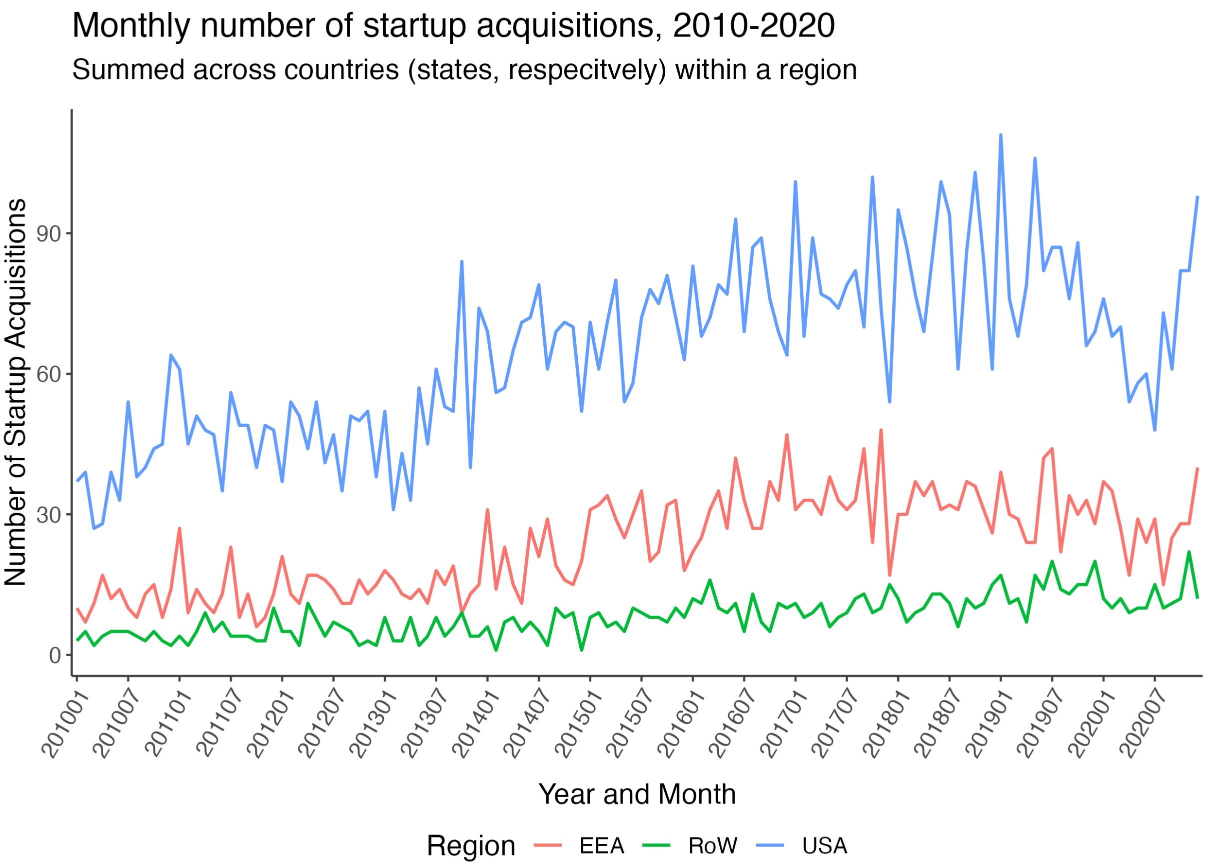


Figure 3.1: Monthly total number of startup acquisitions, 2010-2020.

In contrast to me, Jia et al. (2021) group all startups into four crude industry groups (finance, healthcare, information technology, and others), and construct a panel on the month-state-industry level. Given that startup acquisitions are very rare events, I keep the panel at the month-state level for the time being. Those authors use a similar period of observation to me (January 2014 until March 2019). They adjoin Crunchbase data with data from Venture Expert to obtain a more accurate insight of industries.

4 Difference-in-differences analysis

I carry out a difference-in-differences analysis of the effect of the GDPR on the number of startup acquisitions. I define as "treated" startups in the EEA geographic area. Similarly to previous literature, I use startups located in non-EU countries as a control group.[13] The assumption is that, first, startups located in non-EU countries are less likely to have European customers, and thus less likely to be affected by the GDPR.[14] Second, I make the standard parallel trends assumption that, in the absence of the GDPR, acquisitions in treated and control countries would have developed similarly.

I construct two control groups:

- Control group 1 consists of 50 states plus the District of Columbia in the USA.
- Control group 2 consists of the eight RoW countries.

I carry out the following specification, with subscripts indicating country (for the USA, state) g and month t:

$$y_{gt} = \beta_1 EEA_g \times gdpr\_enact_t + \beta_2 EEA_g \times gdpr\_enroll_t + \alpha_g + \alpha_t + \epsilon_{gt} \tag{3.1}$$

where

- $gdpr\_enact_t$ is an indicator equal to 1 from April 2016 onwards, and 0 otherwise;
- $gdpr\_enroll_t$ is an indicator equal to 1 from May 2018 onwards, and 0 otherwise.

The dependent variable y_{gt} may be the number of startup acquisitions in a given month and state, or functions thereof. α_g and α_t are state (or country, respectively) and time fixed effects: state fixed effects capture meaningful state-specific differences in the number of acquisitions that we see in the data. Equivalently, time fixed effects capture time The coefficients of interests are thus β_1 and β_2.[15]

Before going on to the results, it is worth investigating the average number of startup acquisitions in a given month-state for each of the three geographies in Figure 3.2. The Figure focuses on the time period used in the difference-in-difference framework. It also distinguishes the two separate "treated" time periods (post enactment as well as post enrollemnt) and indicates the average using dashed lines. There is an increase in acquisitions of VC-funded startups in all three regions as we go from the pre-enactment to the post-enactment, but pre-enrollment period. Notably, however, we see a slight decline in the average number of startup M&As in the EEA region in the post-enrollment period, which is entirely absent in the other two regions.

[13] Aside from Jia et al. (2021), literature that has equally employed difference-in-difference analyses to investigate the impact of the GDPR is, for instance, Peukert et al. (2022)

[14] As previous literature has noted (e.g., Johnson (2022)), this control group is likely not perfect. It is possible that, due to the "Brussels effect" and due to further privacy regulations being implemented in jurisdictions outside the EU (such as the CCPA in California, or the LGPD in Brazil; see https://gdpr.eu/gdpr-vs-lgpd/, accessed 06/04/2023), companies outside the EU might effectively have been treated as well. Nevertheless, I start with this control group, and will later try to rule out alternative explanations. In future research, I intend to explore the likely effect of the GDPR by using companies in industries that do not heavily rely on personal data as a control group, possibly employing a triple difference-in-differences framework.

[15] Including EU_g, $gdpr\_enact_t$, or $gdpr\_enroll_t$ individually (i.e., not interacted with each other) is redundant, as these are captured by time and state fixed effects, α_t and α_g.

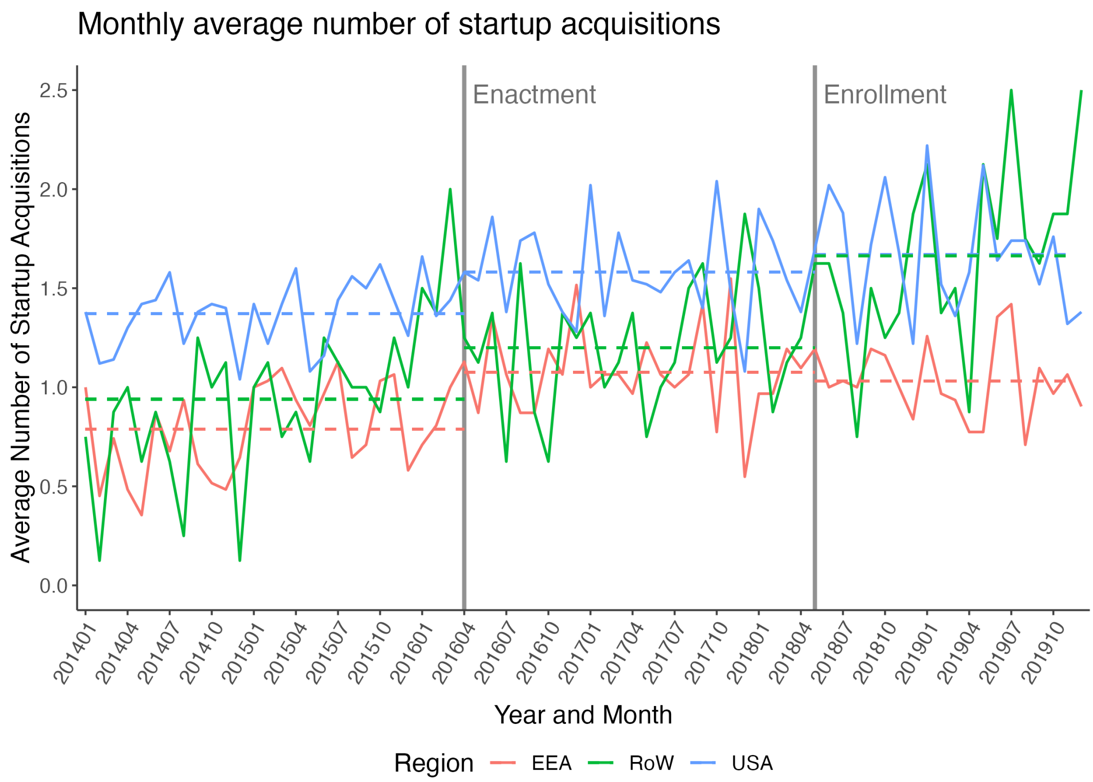


Figure 3.2: Monthly average number of startup acquisitions in the period under study in the difference-in-difference framework, i.e., 2014-2019. Averages (dashed horizontal lines) are computed across states (or countries, respectively) of the three regions.

4.1 Main results

Table 3.3: Linear model. Columns (1) and (2) use the full sample. Columns (3) and (4) focus on those EEA countries which recorded at least 6 acquisitions in 2014-2019. In particular, it excludes: Malta; Slovakia; Slovenia; Croatia; Lithuania; Estonia; Iceland; Latvia; and Bulgaria.

	Dependent variable: Number of Acquisitions			
	Full sample		Excluding EEA states with ≤5 acq	
	EEA vs. USA	EEA vs. RoW	EEA vs. USA	EEA vs. RoW
	(1)	(2)	(3)	(4)
european:post_gdpr_enact	0.077	0.027	0.183	0.133
	(0.131)	(0.162)	(0.166)	(0.191)
european:post_gdpr_enroll	−0.132**	−0.506*	−0.132	−0.506*
	(0.064)	(0.276)	(0.080)	(0.280)
Month FE	✓	✓	✓	✓
Country / State FE	✓	✓	✓	✓
Adjusted R^2	0.901	0.717	0.9	0.699
Observations	5,832	2,808	5,184	2,160
Standard errors clustered at state (respectively country) level.			*p<0.1; **p<0.05; ***p<0.01	

Table 3.3 displays results based on the difference-in-difference specification. I use the full sample in columns (1) and (2), using the USA and the RoW, respectively, as control groups. It shows a decline in acquisitions in the European Union that took place after GDPR enrollment (i.e., from May 2018 on-

wards). However, the results are significant at the 5% level when comparing the countries of the EEA to the USA, and significant only at the 10% level for the RoW. In columns (3) and (4), I use a restricted sample that does not include countries that recorded less than 6 acquiisitons in 2014 to 2019. Whereas the sizes of all coefficients remain very similar, standard errors increase, and a comparison to the USA becomes a insignificant.

A concern with the dependent variable is that it is highly right skewed. I therefore next employ (1) the logged number of acquisitions (plus 1) as a dependent variable, and (2) a Poisson model. Table 3.4 displays the analog to Table 3.3 using the logged dependent variable. The coefficients on the GDPR enrollment variable remain negative, but, except for column (2), lose all statistical significance. Table 3.5 displays the results when employing a Poisson model. Here, the coefficients on enrollment are significantly negative except for column (3). Interestingly, the coefficients on enactment become significantly positive when using US states as a control group.

Table 3.4: Linear model in logs. Columns (1) and (2) use the full sample. Columns (3) and (4) focus on those EEA countries which recorded at least 6 acquisitions in 2014-2019. In particular, it excludes: Malta; Slovakia; Slovenia; Croatia; Lithuania; Estonia; Iceland; Latvia; and Bulgaria.

	Dependent variable:			
	Log(Number of Acquisitions+1)			
	Full sample		Excluding EEA states with $\leq$5 acq	
	EEA vs. USA	EEA vs. RoW	EEA vs. USA	EEA vs. RoW
	(1)	(2)	(3)	(4)
european:post_gdpr_enact	0.022	−0.025	0.048	0.002
	(0.032)	(0.056)	(0.040)	(0.060)
european:post_gdpr_enroll	−0.027	−0.110*	−0.015	−0.098
	(0.026)	(0.057)	(0.033)	(0.060)
Month FE	✓	✓	✓	✓
Country / State FE	✓	✓	✓	✓
Adjusted R^2	0.733	0.684	0.723	0.651
Observations	5,832	2,808	5,184	2,160
Standard errors clustered at state (respectively country) level.			*p<0.1; **p<0.05; ***p<0.01	

Jia et al. (2021) find evidence for a decline in the number of VC deals after GDPR rollout (as opposed to enactment). Similarly to those authors, startup M&A activity seems to be affected only after enrollment, too.

4.2 Robustness

I first replicate the results employing a state-quarter panel: a possible advantage of using quarterly data may be that this dataset contains fewer observations with no acquisitions. This can be seen in Table 3.6, which details variation in the number of acquisitions across state-quarters for each of the three regions. A disadvantage for statistical power is that a quarterly panel contains fewer observations than a monthly one. Tables 3.7, 3.8, and 3.9 show the results of the difference-in-difference analysis. The results largely reflect what we have seen in previous tables, although (possibly driven by the low number of observations) the level of significance declines even more.

I next go back to the month-stage panel, but using a winsorized dependent variable (number of

Table 3.5: Poisson model. Columns (1) and (2) use the full sample. Columns (3) and (4) focus on those EEA countries which recorded at least 6 acquisitions in 2014-2019. In particular, it excludes: Malta; Slovakia; Slovenia; Croatia; Lithuania; Estonia; Iceland; Latvia; and Bulgaria.

	Dependent variable:			
	Number of Acquisitions			
	Full sample		Excluding EEA states with ≤5 acq	
	EEA vs. USA	EEA vs. RoW	EEA vs. USA	EEA vs. RoW
	(1)	(2)	(3)	(4)
european:post_gdpr_enact	0.168*	0.067	0.163*	0.062
	(0.087)	(0.102)	(0.087)	(0.104)
european:post_gdpr_enroll	-0.096*	-0.368***	-0.084	-0.355***
	(0.051)	(0.094)	(0.052)	(0.095)
Month FE	✓	✓	✓	✓
Country / State FE	✓	✓	✓	✓
Pseudo R^2	0.65635	0.53221	0.64210	0.47678
BIC	11,739.0	5,697.5	11,438.3	5,394.6
Observations	5,832	2,808	5,184	2,160

Standard errors clustered at state (respectively country) level. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 3.6: Distribution of startup acquisitions, for startups located in a given region, 2014-2019, **quarterly data**.

Region	Min	25^{th} perc	Median	Mean	75^{th} perc	95^{th} perc	Max	SD
EEA	0	0	1	2.87	3	17	33	5.60
USA	0	0	1	4.58	4	16	110	13.12
RoW	0	0	2	3.69	5	15	27	5.33

acquisitions) at the 95^{th} percentile, equivalent to six acquisitions. This approach effectively eliminates some of the outliers. The results can be seen in Appendix 3.C, Tables 3.14, 3.15, and 3.16. The sign of the coefficients remain the same. The coefficients estimated with using US states as a control group however tend to become insignificant, as the standard error increases and the magnitude of the coefficient (as expected) declines.

When using a linear probability model, i.e., employing as a dependent variable an indicator variable equal to one if any acquisition occurred in a given month and quarter, the results are insignificant and noisy (not reported).

Table 3.7: Linear model, **quarterly data**. Columns (1) and (2) use the full sample. Columns (3) and (4) focus on those EEA countries which recorded at least 6 acquisitions in 2014-2019. In particular, it excludes: Malta; Slovakia; Slovenia; Croatia; Lithuania; Estonia; Iceland; Latvia; and Bulgaria.

	Dependent variable:			
	Number of Acquisitions			
	Full sample		Excluding EEA states with $\leq$5 acq	
	EEA vs. USA	EEA vs. RoW	EEA vs. USA	EEA vs. RoW
	(1)	(2)	(3)	(4)
european:post_gdpr_enact	0.204	0.086	0.522	0.404
	(0.391)	(0.483)	(0.493)	(0.569)
european:post_gdpr_enroll	-0.318^{*}	-1.455^{*}	-0.320	-1.456^{*}
	(0.188)	(0.760)	(0.241)	(0.776)
Quarter FE	✓	✓	✓	✓
Country / State FE	✓	✓	✓	✓
Adjusted R^2	0.957	0.86	0.957	0.85
Magnitudes (percent), enrollment	-24.33	-143.67	-21.77	-111.58
Observations	1,944	936	1,728	720

Standard errors clustered at state (respectively country) level. *p<0.1; **p<0.05; ***p<0.01

Table 3.8: Model in logs, **quarterly data**. Columns (1) and (2) use the full sample. Columns (3) and (4) focus on those EEA countries which recorded at least 6 acquisitions in 2014-2019. In particular, it excludes: Malta; Slovakia; Slovenia; Croatia; Lithuania; Estonia; Iceland; Latvia; and Bulgaria.

	Dependent variable:			
	Log(Number of Acquisitions + 1)			
	Full sample		Excluding EEA states with $\leq$5 acq	
	EEA vs. USA	EEA vs. RoW	EEA vs. USA	EEA vs. RoW
	(1)	(2)	(3)	(4)
european:post_gdpr_enact	-0.001	-0.033	0.033	0.001
	(0.056)	(0.092)	(0.067)	(0.099)
european:post_gdpr_enroll	-0.020	-0.163^{*}	0.012	-0.132
	(0.054)	(0.092)	(0.067)	(0.101)
Quarter FE	✓	✓	✓	✓
Country / State FE	✓	✓	✓	✓
Adjusted R^2	0.831	0.815	0.822	0.784
Observations	1,944	936	1,728	720

Standard errors clustered at state (respectively country) level. *p<0.1; **p<0.05; ***p<0.01

Table 3.9: Poisson model, **quarterly data**. Columns (1) and (2) use the full sample. Columns (3) and (4) focus on those EEA countries which recorded at least 6 acquisitions in 2014-2019. In particular, it excludes: Malta; Slovakia; Slovenia; Croatia; Lithuania; Estonia; Iceland; Latvia; and Bulgaria.

	Dependent variable: Number of Acquisitions			
	Full sample		Excluding EEA states with ≤5 acq	
	EEA vs. USA	EEA vs. RoW	EEA vs. USA	EEA vs. RoW
	(1)	(2)	(3)	(4)
european:post_gdpr_enact	0.1620**	0.0675	0.1579*	0.0633
	(0.0810)	(0.0945)	(0.0816)	(0.0957)
european:post_gdpr_enroll	-0.0786	-0.3537***	-0.0682	-0.3433***
	(0.0517)	(0.0836)	(0.0523)	(0.0847)
Quarter FE	✓	✓	✓	✓
Country / State FE	✓	✓	✓	✓
Pseudo R^2	0.78742	0.68504	0.77826	0.63919
BIC	6,140.3	2,919.2	5,908.1	2,689.9
Observations	1,944	936	1,728	720

Standard errors clustered at state (respectively country) level. *p<0.1; **p<0.05; ***p<0.01

5 Impact on cross-border M&As

Anecotal evidence suggests that the GDPR has made data flows outside the EEA more difficult[16]. The seemingly high degree of legal uncertainty regarding cross-border data flows is also illustrated by the invalidation of the EU-US Privacy Shield in 2020[17], and the subsequent issueing of Standard Contractual Clauses in 2021 that govern data transfers between the EEA and third countries[18].

Moreover, I conducted an interview with a senior security engineer of a Germany based enterprise software startup in 2020, who confirmed that data transfer to North America in the light of the "EU-US Privacy Shield" was a main legal concern for them.

These anecdotes go along with findings by Jia et al. (2020) suggesting that VC investments have become more local with the roll-out of the GDPR. In the light of this evidence, I study the effect of the GDPR on cross-border mergers.

I define "domestic" acquisitions as acquisitions conducted by an acquirer active in the same state or country, respectively. I define "same union" acquisitions as acquisitions conducted by an acquirer active in the same union (i.e., an EEA-located firm acquiring an EEA-located startup). Table 3.10 shows that the average share of domestic or same union acquirers respectively varies greatly between the US, and the RoW or EEA regions: 83% of acquisitions of US-based startups are conducted by acquirers located in the same state as the startup, whereas only somewhat more than a third of acquisitions of EEA- or RoW-based startups are conducted by acquirers located in the same country.

As a first step, I again plot the average number of acquisitions conducted by a domestic acquirer in the period of study, see Figure 3.3. In the US as well as the RoW, we see a steady increase in domestic acquisitions over time. In the EEA, however, the number of domestic acquisitions declines after GDPR

16 For a blog article discussing international transfers of data under the GDPR, see https://tinyurl.com/yecxy7md (both accessed 06/04/2023).

17 See https://tinyurl.com/2kxkz6jv for the decision of the European Court of Justice, and https://tinyurl.com/28ahsb7w for a blog entry discussing implications for firms (both accessed 06/04/2023).

18 See https://tinyurl.com/4mxdmx93 for an explanation of Standard Contractual Clauses (accessed 06/04/2023).

Table 3.10: Average shares of domestic or same union acquirers out of all acquisitions (conditional on any acquisition occurring in a given state-quarter).

Region	Average share of domestic acquirers	Average share of same union acquirers
EEA	0.35	0.64
RoW	0.39	0.89
USA	0.83	0.90

enrollment.

After this initial visual evidence of the raw data, I carry out the specification of Equation (3.1) again that employs fixed effects. This time, I use as dependent variable either the number of acquisitions conducted by domestic acquirers, or alternatively the number of acquisitions conducted by acquirers in the "same union", i.e., outside the EU for startups based in the EU, and outside the USA for startups based in the USA (and in analogue for startups based in the RoW). The preliminary results are displayed in Table 3.11. Surprisingly, the decline in M&As in Europe with respect to the USA (or with respect to the RoW) seem to be driven by *domestic* acquirers, as opposed to foreign acquirers. Table 3.12 confirms this finding when employing logs. This finding stands in contrasts to findings by Jia et al. (2020), and in contrast to the anecdotal findings.

In further work, I would like to control for startup age or measures of maturity (such as the number of funding rounds) in these regressions, as cross-border mergers likely are targeting more mature startups.

Table 3.11: Linear model, GDPR and number of within-border acquisitions.

	Dependent variable:			
	Number of Acquisitions, domestic acquirer		Number of Acquisitions, acquirer from same union	
	EEA vs. USA	EEA vs. RoW	EEA vs. USA	EEA vs. RoW
	(1)	(2)	(3)	(4)
european:post_gdpr_enact	0.009	−0.012	0.062	−0.021
	(0.095)	(0.111)	(0.120)	(0.165)
european:post_gdpr	−0.207***	−0.181*	−0.150***	−0.453**
	(0.066)	(0.098)	(0.051)	(0.229)
Month FE	✓	✓	✓	✓
Country / State FE	✓	✓	✓	✓
Adjusted R^2	0.89	0.526	0.893	0.632
Observations	5,832	2,808	5,832	2,808

Standard errors clustered at state (respectively country) level. *p<0.1; **p<0.05; ***p<0.01

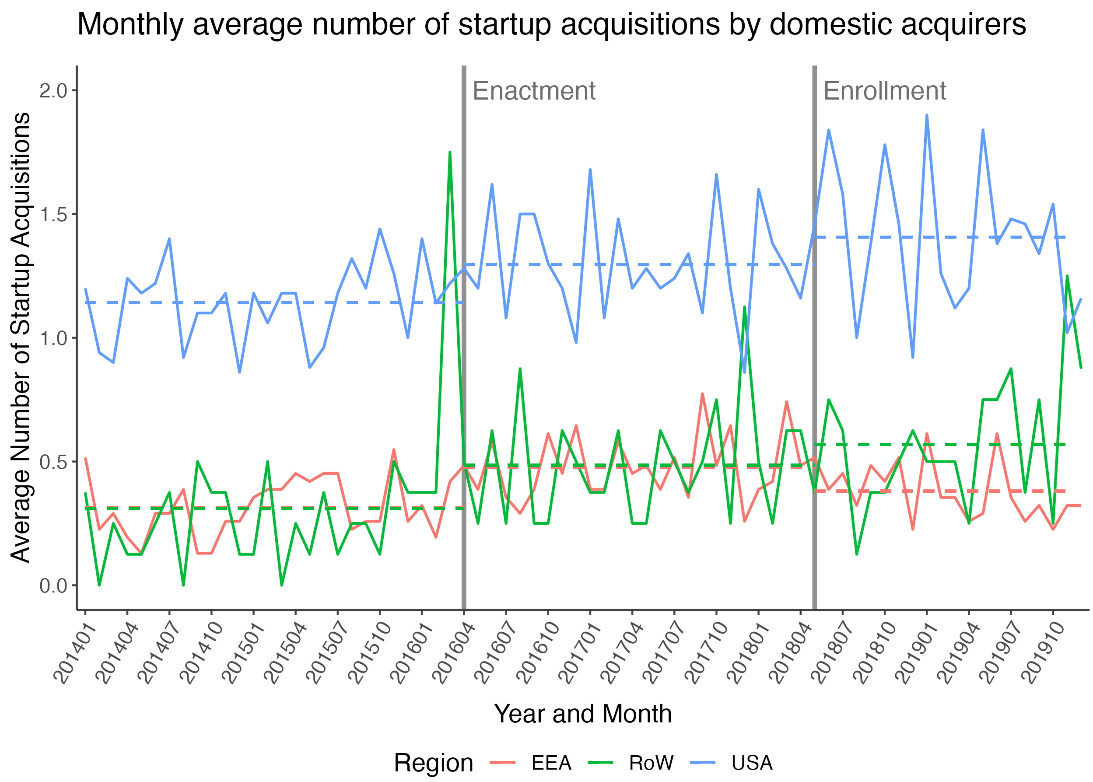


Figure 3.3: Monthly average number of startup acquisitions by domestic acquirers in the period under study in the difference-in-difference framework, i.e., 2014-2019. Averages (dashed horizontal lines) are computed across states (or countries, respectively) of the three regions. The outlier for the RoW in March 2016 is driven by a particularly strong month in Canada.

Table 3.12: Model in logs, GDPR and number of within-border acquisitions..

	Dependent variable:			
	log(Number of Acquisitions + 1), domestic acquirer		log(Number of Acquisitions+1), acquirer from same union	
	EEA vs. USA	EEA vs. RoW	EEA vs. USA	EEA vs. RoW
	(1)	(2)	(3)	(4)
european:post_gdpr_enact	0.003	−0.049	0.030	−0.027
	(0.031)	(0.054)	(0.035)	(0.056)
european:post_gdpr	−0.055***	−0.056*	−0.044*	−0.111**
	(0.021)	(0.031)	(0.023)	(0.051)
Month FE	✓	✓	✓	✓
Country / State FE	✓	✓	✓	✓
Adjusted R^2	0.712	0.547	0.714	0.637
Observations	5,832	2,808	5,832	2,808

Standard errors clustered at state (respectively country) level. *p<0.1; **p<0.05; ***p<0.01

6 Possible extensions

Heterogeneity between industries. An important next step would be to investigate how the above effects vary depending on the startup's industry. This requires, first, to designate industries that have been more, or less, impacted by privacy regulation. I propose to find out about heterogeneity in the extent to which firms were affected by the GDPR by looking at public firms' 10K reports. If companies in a given mention terms such as "GDPR", "privacy", or "CCPA" more heavily than companies in another industry, this should imply that firms in the former industry might be more influenced. If I find that startups in the former industries also display a larger drop in acquisitions, this could corroborate that my finding is driven by the GDPR, and not by other factors that happened around the same time in the EU.

Heterogeneity between countries. European countries are known to have different levels of GDPR enforcement. A further idea is thus to employ the measures of regulatory strictness developed in Johnson et al. (n.d.) and to explore whether countries have been affected to different degrees. If such an investigation reveals that this is indeed the case, it could again lend greater credibility to the initial findings.

Heterogeneity between firms, and types of mergers that were affected. Next, I would turn to the firm-level, and explore to what extent firms with different characteristics are differentially affected. First, is it early-, or rather later-stage startups that are mostly affected by the GDPR? Second, it would be interesting and economically important to find out whether it was mostly horizontal, or rather vertical, mergers that have been prevented by the GDPR: if startup M&As of vertical nature are becoming less likely after the GDPR, the types of innovations we observe would be different. Possibly there would be fewer "complementary" innovation that is useful for potential acquirers, and more "stand-alone" innovative products that can be sold to downstream firms (thus, possibly less incremental and more disruptive innovation). In contrast, if first and foremost M&As of horizontal nature are prevented, this would have consequences for market dynamics and industry concentration.

Impact on other aspects of M&A deals and larger sample of firms. I am currently starting to analyze data from Refinitiv. This data, first, allows to extend the analysis to non-VC funded firms, and to look at separate dates during the acquisition process: e.g., on the lag between the time of announcement and time of closure of the deal. Refinitiv includes a classification specifying whether or not an acquirer's core business can be viewed as "high-tech". If observed in this additional data, interesting follow-up questions to ask would be whether there is a change in the prevalence of breakup fees or stock lock-ups, and whether the GDPR affected the period between announcement and closing of mergers.

7 Conclusion

This paper studies the effect of privacy regulation on startup M&As using a difference-in-differences framework. The initial evidence shows that the GDPR enacted in the EU in 2016 and implemented in 2018 is associated with a decrease in acquisitions of European based startups compared to startups based in the US and other countries. Second, I find that this drop in M&As seems to be driven by domestic acquirers who reduce their M&A activity, which goes against anecdotal evidence. I suggest further follow-up questions that could be asked. These could both provide further support to my initial findings, as well as allow us to deduce conclusions for the economic significance and meaning of these findings.

A limitation of the findings that is shared with previous literature examining the effects of privacy regulation is that no perfect control group which EU-based startups could be compared with seems to exist: presumably, many businesses around the world have European citizens as their customers, or do business with European firms, and might thus have been directly or indirectly affected by the GDPR. I plan to strengthen my results by conducting between-industry analyses, which could increase the plausibility that the reduction in M&As is indeed driven by the GDPR, as opposed to other factors. Moreover, there are multiple robustness checks that still need to be conducted in further work.

Ultimately, the paper relates, first and foremost, to the unintended consequences of privacy regulation (and complex regulations more generally), and to the determinants of firms' M&A decisions. Secondly, a future version of this paper may also study how privacy protection affects firms' data acquisition strategies, and thus provide new insights into data-driven economic transactions. This could improve our understanding of how firms' profits depend on acquiring new data, and to what extent the value of data is affected by privacy regulation. A related follow-on question is to what extent any observed change in merger activity led to a change in the competitive landscape of different industries. The answers to these questions would have significant implications for antitrust policy and for the regulation of digital industries.

3.A Anecdotal Evidence

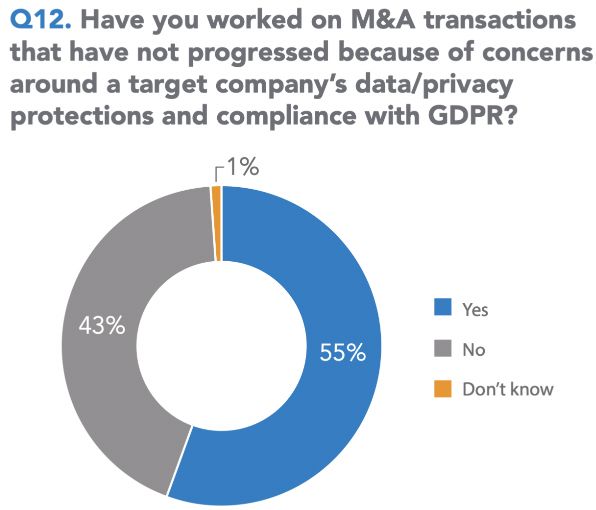


Figure 3.4: Source: https://www.datasite.com/content/dam/merrillcorpwebsite/en/documents/reports/Merrill-DatasiteOne-DueDiligence2022.pdf (accessed 06/04/2023).

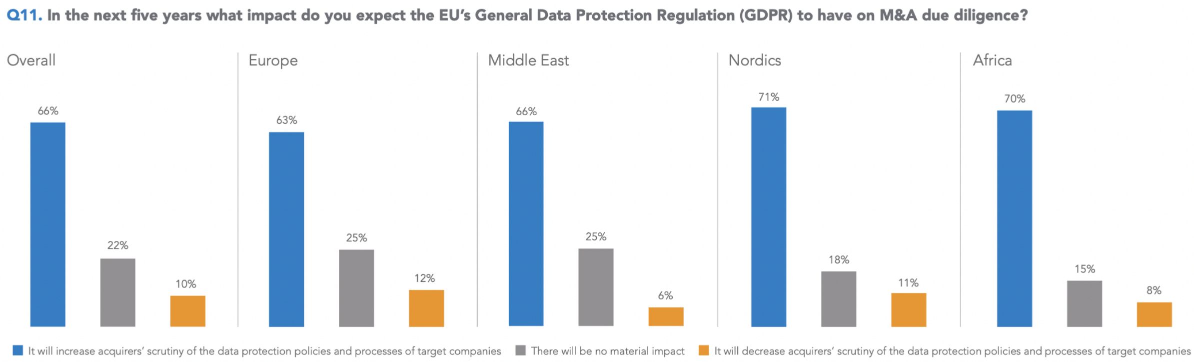


Figure 3.5: Source: https://www.datasite.com/content/dam/merrillcorpwebsite/en/documents/reports/Merrill-DatasiteOne-DueDiligence2022.pdf (accessed 06/04/2023).

3.B Crunchbase Data

3.B.1 Data construction

I obtained access to data on *Crunchbase*, a data portal that tracks financial information on over a million public and private companies, in particular VC-funded firms. *Crunchbase* records companies' founding dates, funding rounds, acquisitions, investments into other companies, initial public offerings (IPOs), and closures. Unlike other financial databases, having received a VC investment is not a pre-condition for being present on this database. The database is well-established in the empirical finance literature, and is believed to capture early-stage funding rounds and acquisitions of small sizes especially well compared to other data sources (Jin, 2019, Yu, 2020).

Information found on *Crunchbase* are sourced using Machine Learning, an in-house data team, a venture program, and via crowdsourcing. People who have worked for the VC industry mentioned to me that *Crunchbase*'s coverage may be most accurate for firms located in North America and Europe.

Defining "startups" and "acquisitions". As *Crunchbase* contains both, venture capital and other types of investments (such as private equity), I use *Crunchbase*'s "Glossary of Funding Types" (Crunchbase, 2022), industry reports and prior literature as guidance to know which types of investments to classify as venture capital.[19] I then define "startups" as companies that have received at least one such VC-type investment.

Crunchbase itself defines acquisitions as majority takeovers. I only consider "first exits", i.e. acquisitions of VC-backed firms that are still private.

The acquisition date recorded by *Crunchbase* is the date of announcement.

Sample. The dataset thus consists of first exits of VC-backed firms. I only include countries that are either part of the EEA plus Switzerland, *or* that have recorded at least 6 acquisitions and are part of the OECD in 2014-2019. See Table 3.13 for an overview of the countries contained in the sample.

[19] I define investments of the following types as being VC investments: *Angel*, *Pre-Seed*, *Seed*, *Series A* to *Series J*, *Series Unknown*, *Corporate Round*, *Undisclosed* and *Convertible Note*. I consider VC investments as financial investments into very early-stage, high-risk companies. The listed investment types' descriptions in *Crunchbase*'s *Glossary of Funding Types* match these characteristics (Crunchbase, 2022). Thus, investment types such as, for instance, *Post-IPO Debt*, *Grant* or *Product Crowdfunding* are not considered as typical VC investments.

Table 3.13: Number of startup acquisitions, for startups located in a given country, 2014-2019.

Country	number of acquisitions	percent
USA	5498	65.92
GBR	565	6.77
CAN	371	4.45
DEU	370	4.44
FRA	341	4.09
ISR	145	1.74
ESP	111	1.33
NLD	110	1.32
SWE	96	1.15
CHE	72	0.86
IRL	68	0.82
ITA	65	0.78
JPN	62	0.74
FIN	60	0.72
AUS	60	0.72
DNK	57	0.68
BEL	43	0.52
NOR	35	0.42
TUR	27	0.32
KOR	24	0.29
AUT	24	0.29
POL	22	0.26
PRT	17	0.20
HUN	14	0.17
MEX	14	0.17
CZE	10	0.12
GRC	8	0.10
ROM	8	0.10
LUX	7	0.08
CHL	6	0.07
CYP	6	0.07
BGR	5	0.06
LVA	4	0.05
ISL	4	0.05
EST	4	0.05
LTU	3	0.04
HRV	2	0.02
SVN	1	0.01
SVK	1	0.01
MLT	1	0.01

3.C Further robustness checks

Table 3.14: Linear model, using a winsorized dependent variable at the 95th percentile (equal to six acquisitions). Columns (1) and (2) use the full sample. Columns (3) and (4) focus on those EEA countries which recorded at least 6 acquisitions in 2014-2019. In particular, it excludes: Malta; Slovakia; Slovenia; Croatia; Lithuania; Estonia; Iceland; Latvia; and Bulgaria.

	Dependent variable:			
	Number of Acquisitions			
	Full sample		Excluding EEA states with $\leq$5 acq	
	EEA vs. USA	EEA vs. RoW	EEA vs. USA	EEA vs. RoW
	(1)	(2)	(3)	(4)
european:post_gdpr_enact	0.067	-0.055	0.133	0.012
	(0.077)	(0.130)	(0.098)	(0.144)
european:post_gdpr_enroll	-0.055	-0.341^{**}	-0.045	-0.331^{**}
	(0.057)	(0.156)	(0.072)	(0.163)
Month FE	✓	✓	✓	✓
Country / State FE	✓	✓	✓	✓
Adjusted R^2	0.749	0.739	0.741	0.718
Observations	5,832	2,808	5,184	2,160

Standard errors clustered at state (respectively country) level. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 3.15: Linear model in logs, using a winsorized dependent variable at the 95th percentile (equal to six acquisitions). Columns (1) and (2) use the full sample. Columns (3) and (4) focus on those EEA countries which recorded at least 6 acquisitions in 2014-2019. In particular, it excludes: Malta; Slovakia; Slovenia; Croatia; Lithuania; Estonia; Iceland; Latvia; and Bulgaria.

	Dependent variable:			
	Log(Number of Acquisitions+1)			
	Full sample		Excluding EEA states with $\leq$5 acq	
	EEA vs. USA	EEA vs. RoW	EEA vs. USA	EEA vs. RoW
	(1)	(2)	(3)	(4)
european:post_gdpr_enact	0.017	-0.033	0.040	-0.011
	(0.030)	(0.054)	(0.037)	(0.058)
european:post_gdpr	-0.020	-0.091^{*}	-0.007	-0.078
	(0.026)	(0.048)	(0.033)	(0.052)
Month FE	✓	✓	✓	✓
Country / State FE	✓	✓	✓	✓
Adjusted R^2	0.665	0.67	0.648	0.633
Observations	5,832	2,808	5,184	2,160

Standard errors clustered at state (respectively country) level. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 3.16: Poisson model, using a winsorized dependent variable at the 95th percentile (equal to six acquisitions). Columns (1) and (2) use the full sample. Columns (3) and (4) focus on those EEA countries which recorded at least 6 acquisitions in 2014-2019. In particular, it excludes: Malta; Slovakia; Slovenia; Croatia; Lithuania; Estonia; Iceland; Latvia; and Bulgaria.

	Dependent variable:			
	Number of Acquisitions			
	Full sample		Excluding EEA states with $\leq$5 acq	
	EEA vs. USA	EEA vs. RoW	EEA vs. USA	EEA vs. RoW
	(1)	(2)	(3)	(4)
european:post_gdpr_enact	0.0996	-0.0071	0.0930	-0.0137
	(0.0854)	(0.1049)	(0.0857)	(0.1059)
european:post_gdpr_enroll	-0.0566	-0.2675***	-0.0422	-0.2532***
	(0.0609)	(0.0839)	(0.0613)	(0.0849)
Month FE	✓	✓	✓	✓
Country / State FE	✓	✓	✓	✓
Pseudo R^2	0.46548	0.48391	0.43831	0.42097
BIC	11,251.0	5,529.5	10,951.8	5,226.6
Observations	5,832	2,808	5,184	2,160

Standard errors clustered at state (respectively country) level. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

References

Canayaz, M., Kantorovitch, I., & Mihet, R. (2022). Consumer Privacy and Value of Consumer Data. *Swiss Finance Institute Research Paper Series N ° 17-23*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3986562

Chen, Z., Choe, C., Cong, J., & Matsushima, N. (2022). Data-driven mergers and personalization. *The RAND Journal of Economics*, *53*(1), 3–31. Retrieved from https://doi.org/10.1111/1756-2171.12398 doi: https://doi.org/10.1111/1756-2171.12398

Condorelli, D., & Padilla, J. (2022). Data-driven Envelopment with Privacy-Policy Tying. *Working Paper*, 1–21. Retrieved from https://www.condorelli.science/PEPPT.pdf

Crunchbase. (2022). *Glossary of Funding Types.* Retrieved 2022-03-12, from https://support.crunchbase.com/hc/en-us/articles/115010458467-Glossary-of-Funding-Types

De Cornière, A., & Taylor, G. (2021). Data and Competition: A Simple Framework with Applications to Mergers and Market Structure. *Working Paper*. Retrieved from https://drive.google.com/file/d/1WFVUqYecWTkestv_RHJDbNPH8ygBPQ4M/view

Eisfeld, L. (2023). Entry and Acquisitions in Software Markets. *Working Paper*. Retrieved from https://luiseeisfeld.github.io/assets/docs/JMP_Eisfeld_TSE.pdf

Jia, J., Jin, G. Z., & Wagman, L. (2020). GDPR and the Localness of Venture Investment. *Working Paper*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3436535

Jia, J., Jin, G. Z., & Wagman, L. (2021). The short-run effects of the general data protection regulation on technology venture investment. *Marketing Science*, *40*(4), 661–684.

Jin, Z. (2019). How do Large Companies Affect Entrepreneurship: Evidence From Amazon's HQ2 Search. *Working Paper*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3280715

Johnson, G. (2022). Economic Research on Privacy Regulation: Lessons from the GDPR and Beyond. *NBER Working Paper No. 30705*, 1–37. doi: 10.2139/ssrn.4293618

Johnson, G., Shriver, S., & Goldberg, S. (n.d.). Privacy and Market Concentration: Intended and Unintended Consequences of the GDPR. *Management Science*.

Kerber, W., & Zolna, K. K. (2022). The German Facebook Case: The Law and Economics of the Relationship between Competition and Data Protection Law. *Working Paper*.

Lambrecht, A. (2017). E-Privacy Provisions and Venture Capital Investments in the EU.

Working Paper. Retrieved from https://www.ceps.eu/wp-content/uploads/2017/10/E-PrivacyProvisionsandVentureCapitalInvestmentsintheEU.PDF

Martin, N., Matt, C., Niebel, C., & Blind, K. (2019). How Data Protection Regulation Affects Startup Innovation. *Information Systems Frontiers*, *21*(6), 1307–1324. doi: 10.1007/s10796-019-09974-2

Peukert, C., Bechtold, S., Batikas, M., & Kretschmer, T. (2022). Regulatory Spillovers and Data Governance: Evidence from the GDPR. *Marketing Science*, *41*(4), 318–340. doi: 10.1287/mksc.2021.1339

Yu, S. (2020). How do accelerators impact the performance of high-technology ventures? *Management Science*, *66*(2), 530–552. doi: 10.1287/mnsc.2018.3256