



Competition in Digital Markets: Role of Data and Network

Citation

Valavi, Ehsan. 2022. Competition in Digital Markets: Role of Data and Network. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences.

Permanent link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37373699>

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

HARVARD UNIVERSITY
Graduate School of Arts and Sciences



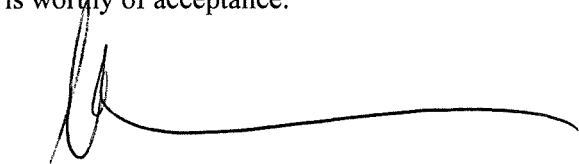
DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the Committee for the
PhD in Business Administration have examined a dissertation
entitled


Competition in Digital Markets: Role of Data and Network

Presented by **Ehsan Valavi**


candidate for the degree of Doctor of Philosophy and hereby
certify that it is worthy of acceptance.

Signature 

Marco Iansiti, Chair

Signature 

Karim Lakhani

Signature 

Feng Zhu

Date: May 27, 2022

Competition in Digital Markets: Role of Data and Network

A DISSERTATION PRESENTED

BY

EHSAN VALAVI

TO

HARVARD BUSINESS SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN THE SUBJECT OF

BUSINESS ADMINISTRATION

HARVARD UNIVERSITY

CAMBRIDGE, MASSACHUSETTS

MAY 2022

©2022 – EHSAN VALAVI
ALL RIGHTS RESERVED.

Competition in Digital Markets: Role of Data and Network

ABSTRACT

This dissertation studies how resources like proprietary data or market factors such as network interconnectivity influence the barrier to entry into digital markets, and analyzes their impact on the competitiveness and growth of digital firms.

Network interconnectivity measures the degree to which consumers in one market purchase products and services from the providers in a different market. In chapter 2, in the context of digital platforms, I examine how such network interconnectivity affects interactions between an incumbent firm serving in multiple markets and an entrant seeking to enter one of these markets.

Chapter 3 investigates the role of data, as a critical resource for digital firms, in creating barriers to entry of competitors. In this chapter, I mainly study how data perishability, which measures the loss in the value of data over time in dynamically changing environments, influences the barrier to entry and, thereby, the competitiveness of an incumbent firm.

Finally, chapter 4 proposes a framework to measure the data perishability rate in various business contexts. The proposed method uses user-generated text data from Reddit.com and estimates the loss in the value of text data to the algorithmic prediction of conversations over time. In this chapter, I argue that the measurement is correlated with the speed of change (and how fast data loses its value) in various business areas.

Contents

TITLE	i
COPYRIGHT PAGE	ii
ABSTRACT	iii
TABLE OF CONTENTS	iv
LIST OF FIGURES	vi
LIST OF TABLES	vii
DEDICATION PAGE	viii
ACKNOWLEDGMENTS	ix
1 INTRODUCTION	1
2 NETWORK INTER-CONNECTIVITY AND ENTRY INTO PLATFORM MARKETS	7
2.1 The Model	11
2.2 Equilibrium Analysis	17
2.3 Extensions	25
2.4 Discussion and Conclusion	33
3 TIME AND THE VALUE OF DATA	39
3.1 Background and Framework	46
3.2 Effectiveness Curve and Value Depreciation	53
3.3 Datasets Collected Over Time	60
3.4 Experimental Design	67
3.5 Conclusion	77
4 TIME-DEPENDENCY, DATA FLOW, AND COMPETITIVE ADVANTAGE	82
4.1 Perishability Measurement Method	84
4.2 Perishability Curves Track Real-World Changes	85
4.3 Characterizing the Perishability Trends	86
4.4 Pairwise Comparison of Macro Trends	90
4.5 Implications	91
APPENDIX A APPENDIX FOR CHAPTER 2	95

APPENDIX B	APPENDIX FOR CHAPTER 3	112
B.1	Appendix B-1	112
B.2	Appendix B-2	123
APPENDIX C	APPENDIX FOR CHAPTER 4	125
REFERENCES		141

List of Figures

- 2.1 Sequence of the game between an incumbent and the entrant 14
- 2.2 Firm’s profit vs. advertising intensity θ 21
- 2.3 The entrant’s optimal advertising intensity and firms’ profits under different values of interconnectivity coefficient r 23
- 2.4 Examples of platform markets with different degrees of network interconnectivity 34

- 3.1 Power-law learning curve. 53
- 3.2 Substitution curves for different dataset sizes 60
- 3.3 Size of datasets processed for each month (To be used in the experiment) 69
- 3.4 Measured learning curves for models that have been trained at different times 73
- 3.5 Cross entropy loss value when we use a model that has been trained on year z (each curve) and is tested on data from year x (x-axis) 74
- 3.6 Equivalent sizes over time (x-axis) when we used 100MB of data in the training phase 76
- 3.7 Effectiveness curve 78

- 4.1 Conceptual illustration of the loss in data value due to time-dependency 85
- 4.2 Effective dataset sizes for multiple topics 86
- 4.3 Effective dataset sizes for the politics subreddit (Blue) and entire Reddit data (Yellow) 87
- 4.4 Half-life-time measured for a few most visited subreddits 89
- 4.5 The p-values for estimated β (The difference between perishability rates) 92

- B.1 Effectiveness graphs for various training sizes 124

- C.1 Power-law (Left plot) and exponential (Right plot) curve fitting. We expect to see a linear representation in either graph 128

List of Tables

2.1	List of notation in the duopoly game model between the incumbent and the entrant	13
3.1	Examples of functional forms for famous ML models.	49
4.1	Perishability rate measurements for several topics.	88

DEDICATED TO MY PARENTS

Acknowledgments

I AM TRULY HONORED AND FORTUNATE TO HAVE BEEN ADVISED by Professors Marco Iansiti, Karim Lakhani, and Feng Zhu. They supported me unconditionally and trusted me with challenging assignments and research projects. They always encourage me to challenge myself and make a difference. I am forever grateful for their trust and devotion. I am also thankful for the advice and guidance I received from Professors Kris Ferreira and Joel Goh. I started at HBS working with Kris and Joel and learned so much from them. I am also forever grateful for the mentorship I received from Professors Ananth Raman, Nathan Craig, Ryan Buell, Amy Edmondson, Gary Pisano, and Iav Bojinov.

I have benefited immensely from all of you as patient teachers, brilliant researchers, and true friends. You have been highly influential and supportive in all aspects of my life.

Special thanks are extended to other TOM unit faculty Amitabh Chandra, Raj Choudhury, Chiara Farronato, Frances Frei, Hise Gibson, Shane Greenstein, Janice Hammond, Robert Huckman, Hima Lakkaraju, Alan MacCormack, Rory McDonald, Edward McFowland, Allison Mnookin, Toni Moreno, Kyle Myers, Seth Neel, Michael Parzen, Willy Shih, Ariel Stern, Stefan Thomke, Mike Toffel, Christina Wing, and RC strategy teaching group, Ashish Nanda, Debora Spar, Jan Rivkin, David Fubini, Frank Nagle, Rem Koning, Jorge Tamayo, Dan Gross, and Andy Wu for their constant advice and open doors.

I am also forever grateful for the unwavering support I received from the doctoral program's office, particularly Jen Mucciarone, Angela Valvis, Marais Young, Kathy Randel, Keith Foster, Tina Christodouleas, Darlene Le, and Maryna Macdonald, and our exceptional doctoral students Moonsoo Choi, Ashley Palmarozzo, Kala Viswanathan, Ryan Allen, Maya Balakrishnan, Meitong Li, Hayley Blunden, Tommy Pan Fang, Chris Fulton, Cheng Gao, Grace Gu, Raha Imanirad, Olivia Jung, Do Yoon Kim, Ohchan Kwon, Michael Anne Kyle, James Sappenfield, Peter Scoblic, Lumumba Seegars, Michelle Shell, Lauren Taylor, Mike Teodorescu, Daniel Yue, Justine Boudou, Natalie Epstein, Jeff Fossett, Caleb Kwon, Nataliya Langburd Wright, and Hashim Zaman.

My research would not have been possible without the support of several coauthors and engaged research partners. I want to thank Xinxin Li from the University of Connecticut, Joel Hestness from Cerebras Systems, and Newsha Ardalani from Meta AI (Formerly known as Facebook AI research -FAIR). Xinxin (With Professors Marco Iansiti and Feng Zhu) is my coauthor in the pa-

per presented in chapter 2. Joel and Newsha are my coauthors in the papers presented in chapter 3 (With Marco Iansiti) and chapter 4 (With Marco Iansiti, Feng Zhu, and Karim Lakhani).

1

Introduction

Extensive literature exists on how firms compete and ultimately thrive in a market. These studies offer a variety of theories on how various factors, like the resources firms own, their capabilities, and the characteristics of the market they tap into, influence their competitiveness and survival (16,99,102,62,25,69,115). These factors contribute to creating competitive advantage differently, and their impact varies across markets, contexts, and business areas. For example, producing cheaper and at a larger scale is often what managers at traditional firms (Firms that usually produce physi-

cal goods and have people at the core of their operations) are after to create competitive advantage. These cost and scale desires are mostly met by finding and securing cheaper supplies, developing technologies and innovations to increase productivity and efficiency, and leveraging the cost advantages gained from scaling the size and the variety of offered goods, known as economies of scale and scope (^{16,62,99,25}). Market characteristics related to the networks these firms connect to are also important* (⁶⁹). However, their impacts are weaker for firms that aren't sufficiently large due to physical, technological, and organizational constraints that limit the scale. For example, expanding a physical shopping mall to include more shops increases the variety of services a shopper can get. Despite that, beyond a physical limit, shoppers find the expansion overwhelming, and the quality of their shopping experience deteriorates. University is another example where offering more variety of courses to a larger audience is a plus. Yet, physical and organizational challenges prevent it from such expansion.

For firms offering digital goods and services, in contrast to traditional firms, competitive advantage is increasingly defined by controlling resources like data and their ability to shape digital networks (⁶³). In other words, in making a firm more competitive in digital markets, the contribution of the market's network characteristics and securing resources like data is more significant. Because, first, certain costs associated with economic activities fall substantially with digitization, making cost reduction a second priority. For example, operations, storage, search, replication, transportation, tracking, and verifications costs are considerably lower in digital markets (⁵⁰). Second, digital technologies are not bound by physical constraints and often present substantial scalability. For example, Coursera Inc and EdX offer far more variety of courses to a significantly larger user base than any traditional university. The Stanford Machine Learning course that Andrew Ng offered

*There are other market characteristics like heterogeneity in participants' preferences where firms can compete by crafting different business models. For example, firms offering similar, even identical, products and services may have different revenue formulas (Subscription vs. pay-as-you-go) and compete to gain different market segments. However, in this dissertation, I am only concerned about the role of networks these firms tap into.

at Coursera in 2011 had more than 2 million enrolled students, which in terms of enrollee number surpasses any class in the history of the traditional system. Another example is the Amazon marketplace, which, compared to brick-and-mortar chains, offers a significantly greater variety of goods from distant locations because physical constraints do not bind it. Such scalability amplifies the role of market characteristics like the network in creating and sustaining competitive advantages for digital firms. Therefore, cost reduction as the second priority, together with substantial scalability of digital solutions, puts securing resources and shaping and controlling digital networks at the forefront of managers' priorities in digital firms.

This dissertation investigates how resources like proprietary data and market factors such as network interconnectivity influence the barrier to entry into digital markets and analyzes their impact on the competitiveness and growth of digital firms.

Chapter 2 investigates the role of network interconnectivity in creating barrier to entry in the context of digital platform markets. In certain platform networks, buyers in one market purchase products and services from providers in many different markets, whereas in others, buyers primarily purchase from providers within the same market. Accordingly, network interconnectivity — which measures the degree to which consumers in one market purchase from providers in a different market — varies across different industries. This chapter examines how network interconnectivity affects interactions between an incumbent platform serving multiple markets and an entrant platform seeking to enter one of these markets. The model yields several interesting results. First, even if the entrant can advertise at no cost, it still may not want to make every user in a local market aware of its service, as doing so may trigger a competitive response from the incumbent. Second, having more mobile buyers, which increases interconnectivity between markets, can reduce the incumbent's incentive to fight and, thus, increase the entrant's incentive to expand. Third, more robust interconnectivity between markets may or may not make the incumbent more defensible: when advertising is not costly and mobile buyers consume in both their local markets and the markets they visit, a

large number of mobile buyers will increase the entrant's profitability, thereby making it difficult for the incumbent to deter entry. However, when advertising is costly or mobile buyers only consume in the markets they travel to, a large number of mobile buyers will help the incumbent deter entry. When advertising cost is at an intermediate level, the entrant prefers a market with moderate interconnectivity between markets. Fourth, it finds that even if advanced targeting technologies can enable the entrant to also advertise to mobile buyers, the entrant may choose not to do so in order to avoid triggering the incumbent's competitive response. Finally, we find that the presence of network effects is likely to decrease the entrant's profit. The results offer managerial implications for platform firms and help understand their performance heterogeneity.

In chapter 3, this thesis investigates how heterogeneity in data sourcing of an incumbent firm influences the barrier to entry of a new competitor. For simplicity and without loss of generality, I focus on the heterogeneity across time and study the time value of data in a dynamically changing environment. This chapter challenges managers' belief that collecting more data will continually improve the accuracy of machine learning models. This belief is engrained by the statistical theories on how the accuracy of machine learning models scales with the dataset size and the managerial theories on how the economic value created by a firm scales with the resources (2,6,21,23).

We argue that when data lose relevance over time, it may be optimal to collect a limited amount of recent data instead of keeping around an infinite supply of older (less relevant) data. In addition, we argue that increasing the stock of data by including older datasets may, in fact, damage the model's accuracy. Expectedly, the model's accuracy improves by increasing the flow of data (defined as data collection rate); however, it requires other tradeoffs in terms of refreshing or retraining machine learning models more frequently.

We use these results to investigate how the business value created by machine learning models scales with data and when the stock of data establishes a sustainable competitive advantage. We argue that data's time-dependency weakens the barrier to entry that the stock of data creates. As a

result, a competing firm equipped with a limited (yet sufficient) amount of recent data can develop more accurate models. This result, coupled with the fact that older datasets may deteriorate models' accuracy, suggests that created business value doesn't scale with the stock of available data unless the firm offloads less relevant data from its data repository. Consequently, a firm's growth policy should incorporate a balance between the stock of historical data and the flow of new data.

This research complements its theoretical results with an experiment. In the experiment, it uses the simple and widely used machine learning task known as next-word prediction. We empirically measure the loss in the accuracy of a next-word prediction model trained on datasets from various time periods. Our empirical measurements confirm the economic significance of the value decline over time known as data perishability. For example, 100MB of text data, after seven years, becomes as valuable as 50MB of current data for the next-word prediction task.

Chapter 4 aims to measure the data perishability rate for various business areas. What is often less appreciated in management literature is that the time value of data for digital firms ranges widely with the business area they are operating in. This variance call for new strategies and management practices and has significant implications for policymakers and regulators (35,76,85,110,111,45,98) now faced with designing policy to guard against bias, enhance user privacy, increase consumer welfare and safeguard competition. In this chapter, I use user-generated text data from Reddit.com and compare the time-dependency across various Reddit topics (Subreddits). I make this comparison by measuring the rate at which the user-generated text data loses its relevance to the algorithmic prediction of conversations. I show that different subreddits have different rates of relevance decline over time. The decay rate on slow-varying subreddit like "history" is very low, as data maintains its value indefinitely. In contrast, subreddits like "world news" have a significantly higher decay rate and lose their value relatively quickly. Relating the text topics to various business areas of interest, I argue that competing in a business area in which data value decays rapidly alters strategies to acquire a competitive advantage. When data value decays rapidly, access to a continuous flow of data will

be more valuable than access to a fixed stock of data. In this kind of setting, improving user engagement and increasing the user-base help create and maintain a competitive advantage (⁶⁹).

2

Network Inter-connectivity and Entry into Platform Markets

Digitalization has led to the emergence of numerous platforms in our economy today (¹⁰⁸, ⁶⁴, ⁸⁸). Examples of popular platforms include Uber in the transportation industry, Airbnb in the accommodation industry, Craigslist in the classifieds market, and Groupon in the local daily deals market. A growing body of literature on information systems attempts to understand the optimal strategies

for digital platforms to scale and compete. Scholars have examined a variety of issues, including optimal pricing (⁸⁸), interactions between competing platforms (^{71, 86, 114}), optimal business models (^{26, 87, 105}), strategies to motivate third-party providers (^{60, 73}), matching efficiency between buyers and sellers (^{59, 113}), platforms' investment decisions (⁹), managing multigenerational platforms (⁵⁵), and contractual relationships or tensions between platform owners and third-party providers (^{61, 56, 79}).

This study adds to this literature by examining how network characteristics affect the strategies and performance of competing platforms. All platforms exhibit two-sidedness in that they facilitate matching and transactions between consumers and service providers in their markets, but the interconnectivity of their businesses—which measures the degree to which consumers in one market purchase services from service providers in a different market—varies considerably across industries. For example, the network structure of Upwork, an online marketplace that connects millions of businesses with freelancers around the globe, exhibits high interconnectivity among different markets. In contrast, Uber's network consists of local network clusters with some interconnectivity: riders transact with drivers in their city, and, except for frequent travelers, they care most about the local availability of Uber drivers. We observe similar local network clusters with some interconnectivity in group buying platforms such as Groupon, classifieds sites such as Craigslist, food delivery platforms such as Grubhub, and restaurant-reservation platforms such as OpenTable, and marketplaces that match freelance labor with local demand such as TaskRabbit, Instacart, and Rover.

The network interconnectivity of a platform market has important implications for the profitability and defensibility of incumbent platforms. When the network is strongly interconnected, it is difficult for a new entrant to compete, particularly when consumers in one local market mostly purchase services from other markets. A platform that enters one local market, for example, would waste a significant amount of marketing resources to build awareness among local consumers and service providers without generating a large number of transactions. Therefore, for a new plat-

form, entry into highly interconnected markets is costly. In contrast, when consumers and service providers mostly transact within their local clusters, it is relatively easy for a new platform to enter, as it can specialize in one local cluster and build awareness from there. In the ridesharing industry, many entrants have challenged market leaders in local markets. Fasten entered the Boston market in 2015 to compete with Uber and Lyft with a much smaller budget. In New York City, Juno and Via have been competing with Uber and Lyft for years, and Myle was launched recently. Uber also faced a wave of rivals in London, including Estonia's Bolt, France's Kapten, Israel's Gett, and India's Ola. Didi, the largest ridesharing company in China, constantly faced new entrants in multiple cities.

This study adopts a game-theoretical approach to examine how network interconnectivity affects competitive interactions between an incumbent platform and an entrant platform. The incumbent platform has an installed base of buyers and service providers in multiple local markets; the entrant is interested in entering one of these markets. To capture interconnectivity between local markets, the assumption is that some buyers are mobile: they travel between markets, purchasing services in each. In the first stage, the entrant invests money to build brand awareness in one of these markets. In the second stage, the incumbent and the entrant set prices for buyers and wages for service providers in that market. Finally, in the third stage, buyers and service providers in that market choose one platform to conduct transactions.

The model yields several interesting results. First, even if the entrant can advertise at no cost, it still may not want to make every user in a local market aware of its service, as doing so may trigger a competitive response from the incumbent. Second, having more mobile buyers, which increases interconnectivity between markets, can reduce the incumbent's incentive to fight and, thus, increase the entrant's incentive to expand. Third, stronger interconnectivity across markets may or may not make the incumbent more defensible. When advertising is not costly, and mobile buyers consume in both their local markets and the markets they visit, a large number of mobile buyers (i.e., great interconnectivity) will increase the entrant's profitability, thereby making it difficult for

the incumbent to deter entry. This result is somewhat surprising: great interconnectivity is supposed to provide the entrant with a disadvantage because it increases the size of the incumbent's potential market while retaining the size of the entrant's potential market. When advertising cost is at an intermediate level, the entrant prefers a market with moderate interconnectivity between markets. When advertising is costly, or mobile buyers only consume in the markets they travel to, a large number of mobile buyers will help the incumbent deter entry. Fourth, we find that even if advanced targeting technologies can enable the entrant to also advertise to mobile buyers, the entrant may choose not to do so to avoid triggering the incumbent's competitive response. Finally, we find evidence that the presence of network effects is likely to decrease the entrant's profit.

In the literature on platform strategies, our approach to the problem is closely related to the literature examining entry into platform markets. Studies have identified a number of factors that influence the success or failure of entrants in platform markets, such as the strength of network effects (^{115, 86}), platform quality (^{80, 103}), multi-homing (^{24, 71, 8}), and exclusivity (³²). All these studies assume a strongly interconnected network. As indicated by Afuah et. al (⁴), this assumption does not reflect the actual networks in most industries. This study extends this literature by examining how network interconnectivity affects the strategies and performance of incumbents and entrants.

Broadly, this research is related to competitive interactions between incumbents and entrants. Theoretical models in the literature focus on incumbent strategies such as capacity investment to deter or accommodate entry (^{43, 106}). Empirical studies often find that incumbent reactions to entrants are selective (⁴⁷): while some incumbents choose to react to entrants aggressively, others do not appear to respond to entry. Studies have also shown that this variation in responses often depends on entrant characteristics, such as scale (^{36, 68}). Our model finds support for these empirical results. Chen and Guo (²⁷) document a similar result regarding an entrant refraining from over-advertising to avoid competitive response from a competitor but in a very different setting. In their setting, a third-party seller sells the same product as a retail platform, and the seller needs to decide

how much to advertise through other channels, such as search engines and social media. This research differs from these studies by examining how network interconnectivity changes the incumbent's incentives to react, the entrant's incentives to advertise, and the entrant's profit. We show that buyers' consumption behavior matters: when mobile buyers consume in local markets, greater network interconnectivity sometimes increases entrant profits; however when they do not, greater network interconnectivity reduces entrant profits.

This research is also related to studies that examine how network structures affect product diffusion (¹, ¹⁰⁰, ¹⁰¹, ¹⁰⁹). These studies typically focus on social networks, like instant messaging platforms, and examine questions related to issues such as seeding within these networks (⁴⁶, ⁸¹), pricing policies to facilitate product diffusion (²⁰, ⁷⁷), network formation processes and how local network clustering leads to local bias (⁷⁸), prediction accuracy (⁹¹), and market segmentation (¹⁵). These networks have more complicated connections because they depend on individuals' own social networks, and, consequently, these studies rely on simulations or descriptive results. We adopt a different perspective to focus on how interconnectivity between local markets affects market entry and derive closed-form solutions.

The remainder of the chapter is organized as follows. In Section 2.1, we introduce the model and analyze the competitive interactions between an incumbent and an entrant. In Section 2.2, we examine extensions to our main models. In Section 2.3, we conclude by discussing the implications of our results and potential future research.

2.1 THE MODEL

2.1.1 MODEL SETUP

Assume that there are multiple local markets each with N buyers who are currently using the incumbent's platform (denoted as I) for transactions. A fraction of buyers in each market are mobile

— r percent of them travel between markets. Assume the movement is random, so that in equilibrium, in each market, rN buyers visit other markets and rN additional buyers come from other markets to make purchases. Hence, r measures the interconnectivity between these markets. Each mobile buyer places one order for the service in his local market and another order when he travels. For example, riders use ride-sharing services in their local markets; when they travel, they use ride-sharing services in other markets. (We consider the scenario in which mobile buyers do not consume in their local markets in an extension.) Each service provider fulfills one order at most. To accommodate these mobile buyers, each market has $(1 + r)N$ service providers. Table 2.1 provides a summary of the notations used in the main model.

Table 2.1: Notation in the main model

Variable	Meaning
N	Total number of local buyers
r	Proportion of buyers who are mobile and travel between markets, $r \in [0, 1]$
n	Number of buyers and service providers (users) who are exposed to the entrant's advertising
$L(n)$	Advertising cost for the entrant for reaching n buyers and service providers
k	Advertising cost parameter
θ	Proportion of users that the entrant targets for advertisement, $\theta \in [0, 1]$
θ^*	Optimal proportion of potential users that the entrant targets for advertisement, $\theta^* \in [0, 1]$
v	Buyer's willingness to pay for the service
m	Maximum switching cost
a_i	Switching cost for buyer i to adopt the entrant's platform, $a_i \sim Uni(0, m)$
c_j	Switching cost for service-provider j to adopt the entrant's platform, $c_j \sim Uni(0, m)$
a^*	The threshold at which buyers with lower switching cost will adopt the entrant's platform.
c^*	The threshold at which service providers with lower switching cost adopt the entrant's platform.
N_l^B	Number of buyers who use platform $l \in \{I, E\}$
N_l^S	Number of service providers who use platform $l \in \{I, E\}$
$U_{i_l}^B$	Utility of buyer i who uses platform $l \in \{I, E\}$
$U_{j_l}^S$	Utility of service provider j who uses platform $l \in \{I, E\}$
p_l	Price for buyers that is set by platform $l \in \{I, E\}$
w_l	Wage for service providers that is set by platform $l \in \{I, E\}$
π_l	Profit for platform $l \in \{I, E\}$

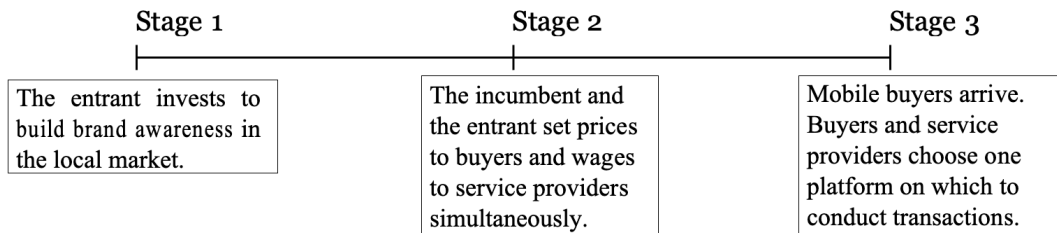


Figure 2.1: Sequence of the game

Before an entrant (denoted as E) enters one of these markets, the incumbent serves the market as a monopoly and all the users (i.e., both service providers and buyers) are aware of the incumbent. (This assumption is relaxed in an extension of the model in which not all buyers and sellers are aware of the incumbent.) Neither the buyers nor the service providers are aware of the entrant, but the entrant can advertise to build awareness.

The game proceeds as follows, as depicted in Figure 2.1. In the first stage, the entrant invests to build brand awareness among users in the local market. Advertising is costly, and it costs the entrant $L(n)$ to reach n potential users. The entrant decides on θ , a fraction of the potential users reached through advertising.* Because we have N buyers and $(1+r)N$ service providers, $n = \theta(2N + rN)$. Following the literature (¹⁰⁴, ¹⁰⁶, ³⁹, ⁶⁶), we assume the advertising cost is a (weakly) increasing and convex function of n : $L'(n) \geq 0$ and $L''(n) \geq 0$. Note that even with digital technologies, it remains costly to build awareness. While certain platforms may be able to attract their first tranche of customers relatively inexpensively, through word-of-mouth or other low-cost strategies, the cost typically begins escalating when the platform begins to look for new and somewhat different customers through search advertising, referral fees, and other marketing strategies.† Consequently,

*Our results continue to hold qualitatively if the performance of the entrant’s advertising level is uncertain (i.e., when the entrant decides θ the fraction of potential users reached through advertising becomes $\theta + \varepsilon$ where ε is a random variable).

†See, for example, “Unsustainable customer acquisition costs make much of ecommerce profit proof,” Steve Dennis, Forbes, August 31, 2017.

many platforms exit the market after burning too much money on customer acquisition. In our model, we allow advertising cost to vary and examine its implications on platform strategies and performance. In the main model, we also assume that the entrant is not able to advertise to mobile buyers. We relax this assumption in an extension.

In the second stage, the incumbent sets the price to each buyer, denoted as p_I , and the wage to each service provider, denoted as w_I , in the local market. The entrant also sets the price for the service buyers, denoted as p_E , and the wage for the service providers, denoted as w_E . Here, the subscript denotes the platform (I for incumbent and E for entrant). For example, Instacart, decides the prices for users and the wages for shoppers. Uber decides the rates for riders and the commissions it takes before passing on the revenue from riders to drivers, which effectively determines the wages for drivers. Consistent with the practice, we allow firms to set different prices and wages in different markets, but they do not price discriminate based on whether a buyer is local or mobile within a market. We denote each buyer's willingness to pay for the service as v . Further, we normalize the value of outside options to zero and the service providers' marginal cost to zero.[‡] Hence, without the entrant, as a monopoly, the incumbent will choose $p_I = v$ and $w_I = 0$.

In the third stage, the rN mobile buyers from other markets arrive. Buyers and service providers choose one platform on which to conduct transactions. Mobile buyers are not exposed to the entrant's advertisements and are, therefore, only aware of the incumbent. Hence, the entrant and the incumbent compete for buyers and service providers from the local market, but the mobile buyers will only use the incumbent platform.

The $(1 - \theta)$ portion of users in the local market is only aware of the incumbent and will buy or provide the service on the incumbent platform as long as they receive a non-negative utility from the incumbent. Specifically, a buyer will buy the service as long as $p_I \leq v$, and a service provider will

[‡]If we allow the marginal cost to be a positive constant, then the equilibrium service prices will increase by this constant.

provide the service as long as $w_I \geq 0$. Because $p_I \leq v$ and $w_I \geq 0$ always hold, these users will always use the incumbent's platform.

The θ portion of users in the local market becomes aware of both the incumbent and the entrant and will remain with the incumbent's platform unless the entrant provides a higher utility. If a user elects to switch to the entrant's platform, there is a switching cost that varies across users. We denote this cost for a service provider, i , as c_i and for a buyer, j , as a_j . Similar to Ruiz-Aliseda (⁹⁵), we assume that both c_i and a_j follow a uniform distribution between zero and m , where m captures the difficulty in switching to a new service in the market. To be consistent with real world scenarios, we assume that m is sufficiently large (i.e., there are some users whose switching cost is sufficiently large) so that, in equilibrium, the entrant will not take away the entire segment of users who are aware of both platforms. (Mathematically, this assumption requires that the distribution of the switching cost be sufficiently sparse—that is, $m > \frac{2(1+r)v}{16(2+r)}$.)

Among the θ portion of service providers, a service provider, i , will choose the entrant if the utility from using the entrant's platform ($U_{E_i}^S = w_E - c_i$) is greater than the utility from using the incumbent's platform ($U_{I_i}^S = w_I$). Here, the subscript again denotes the platform (I for incumbent and E for entrant) and the superscript denotes the user (S for service provider and B for buyer). The solution to the equation $U_{E_i}^S = U_{I_i}^S$ is $c^* = w_E - w_I$, describing the switching cost of the indifferent service provider. Thus, service providers with $c_i < c^*$ will choose the entrant and those with $c_i \geq c^*$ will choose the incumbent. Let N_I^S denote the number of service providers selecting the incumbent and N_E^S denote the number of service providers selecting the entrant. Then, we have the following two equations:

$$N_I^S = \left(1 - \frac{c^*}{m}\theta\right)(1+r)N \quad (2.1)$$

$$N_E^S = \frac{c^*}{m}\theta(1+r)N. \quad (2.2)$$

Similarly, a buyer, j , will choose the entrant if the utility from using the entrant's platform ($U_{Ej}^B = v - p_E - a_j$) is greater than the utility from using the incumbent's platform ($U_{Ij}^B = v - p_I$). The solution to the equation $U_{Ej}^B = U_{Ij}^B$ is $a^* = p_I - p_E$. Thus, buyers with $a_j < a^*$ will choose the entrant and those with $a_j \geq a^*$ will choose the incumbent. Let N_I^B denote the number of service buyers selecting the incumbent and N_E^B denote the number of service buyers selecting the entrant. We obtain the following two equations:

$$N_I^B = \left(1 - \frac{a^*}{m}\theta + r\right)N. \quad (2.3)$$

$$N_E^B = \frac{a^*}{m}\theta N. \quad (2.4)$$

We can then derive the incumbent's profit, π_I , and the entrant's profit, π_E , from the local market as follows:

$$\pi_I = \min(N_I^S, N_I^B)(p_I - w_I) \quad (2.5)$$

$$\pi_E = \min(N_E^S, N_E^B)(p_E - w_E) - L(\theta(2N + rN)) \quad (2.6)$$

It is possible that under some prices and wages of the two platforms, the number of buyers is not the same as the number of service providers. In such cases, either some buyers' orders are not fulfilled, or some service providers will not serve any buyers and hence earn no income.

2.2 EQUILIBRIUM ANALYSIS

We use backward induction to derive the equilibrium. Specifically, we first derive each platform's optimal price and profit given the entrant's advertising decision and then solve for the entrant's

optimal advertising decision in the first stage.

To derive each platform's optimal price given the entrant's advertising decision, we recognize that it is often difficult to derive closed-form equilibrium solutions when we allow two competing platforms to set prices on both sides, especially when the platforms are heterogeneous. Prior studies have often had to make simplifying assumptions, such as a fixed price (or royalty rate) on one-side of the market, symmetric pricing, or one platform being an open source platform and, thus, free (108,38,22,3). In this study, we take advantage of a market clearing condition to derive the optimal prices and wages for the two competing platforms. Specifically, we prove that, in equilibrium, the incumbent and the entrant will always choose their prices and wages so that the number of service providers using a platform equals the number of buyers using the same platform: $N_I^S = N_I^B$ and $N_E^S = N_E^B$. Lemma 2.2.1 states this result (proofs of all lemmas and propositions for the main model are provided in the appendix).

Lemma 2.2.1. *The incumbent and the entrant will set their prices and wages so that the number of service providers using a platform equals the number of buyers using the same platform.*

The intuition for Lemma 2.2.1 is that if the numbers on the two sides are not balanced, a firm can adjust its price or wage to get rid of excess supply or demand to increase its profitability. The lemma suggests that $a^* = (1 + r)c^*$. Hence, $(p_I - p_E) = (1 + r)(w_E - w_I)$. Thus, we can rewrite the profit functions as follows:

$$\pi_I = \left(1 - \frac{p_I - p_E}{m}\theta + r\right) N \left(p_I + \frac{p_I - p_E}{1 + r} - w_E\right). \quad (2.7)$$

$$\pi_E = \frac{p_I - p_E}{m}\theta N \left(p_E - \frac{p_I - p_E}{1 + r} - w_I\right) - L(\theta(2N + rN)). \quad (2.8)$$

We can then derive each platform's optimal price and profit, given the entrant's advertising deci-

sion, as shown in proposition 2.2.1.

Proposition 2.2.1. *Given the entrant's choice of advertising intensity θ , the optimal prices, number of buyers and service providers, and platform profits can be determined as follows:*

$$\begin{aligned}
 i \text{ If } 0 \leq \theta \leq \min\left(\frac{2m(2+r)}{3v}, 1\right), \text{ then } p_I^* &= v, w_I^* = 0, p_E^* = \frac{(3+r)v}{2(2+r)}, w_E^* = \frac{v}{2(2+r)}, N_I^{B^*} = \\
 N_I^{S^*} &= \frac{N(1+r)}{2} \left(2 - \theta \frac{v}{m(2+r)}\right), N_E^{B^*} = N_E^{S^*} = \frac{N(1+r)\theta v}{2m(2+r)}, \pi_I^*(\theta) = \frac{N(1+r)v}{2} \left(2 - \frac{\theta v}{m(2+r)}\right), \\
 \text{and } \pi_E^*(\theta) &= \frac{N(1+r)\theta v^2}{4m(2+r)} - L(\theta(2N + rN)).
 \end{aligned}$$

$$\begin{aligned}
 ii \text{ If } \min\left(\frac{2m(2+r)}{3v}, 1\right) < \theta \leq 1, \text{ then } p_I^* &= \frac{(2(2+r)m)}{3\theta}, w_I^* = 0, p_E^* = \frac{(3+r)m}{3\theta}, w_E^* = \frac{m}{3\theta}, \\
 N_I^{B^*} = N_I^{S^*} &= \frac{2N(1+r)}{3}, N_E^{B^*} = N_E^{S^*} = \frac{N(1+r)}{3}, \pi_I^*(\theta) = \frac{4Nm(1+r)(2+r)}{9\theta}, \text{ and } \pi_E^*(\theta) = \\
 \frac{Nm(1+r)(2+r)}{9\theta} &- L(\theta(2N + rN)).
 \end{aligned}$$

When θ is smaller than a certain threshold $\left(\min\left(\frac{2m(2+r)}{3v}, 1\right)\right)$, we find that the incumbent platform chooses not to respond to the entrant. It continues to charge the monopoly price, v , and offer the monopoly wage, zero, although its profit does decrease as θ increases because it loses market share to the entrant. The entrant platform incentivizes some buyers and service providers to switch by charging a lower price and offering a higher wage.

The threshold for θ (weakly) increases with r because mobile buyers are only aware of the incumbent platform (i.e., the incumbent platform has monopoly power over them) and their existence reduces the incumbent's incentive to respond to the entrant. It is thus not surprising that the entrant can take advantage of this lack of incentive and increase its advertising intensity. The number of transactions hosted on the incumbent platform increases with r because of mobile buyers from other markets, even though the incumbent loses more transactions from local buyers to the entrant when r increases. The incumbent platform's profit increases with r because of the increase in transactions at the same monopoly price it charges. The number of transactions the entrant serves also increases with r because it can advertise more aggressively without triggering a competitive response from the incumbent. The entrant's profit increases with r without taking the advertising cost into

account. If advertising cost increases significantly with r , the entrant's profit may decrease with r , a scenario which will be examined later.

When θ is larger than the threshold, however, the entrant platform has the potential to steal a large market share from the incumbent. The incumbent platform chooses to respond by lowering its price to buyers. The entrant platform thus lowers its price to buyers as well. Note that the wages offered by the entrant in this case decrease with θ . This is because even though advertising reaches many service providers, there is no demand for all the service providers due to the competitive response from the incumbent on the buyer side, thereby allowing the entrant to offer lower wages.

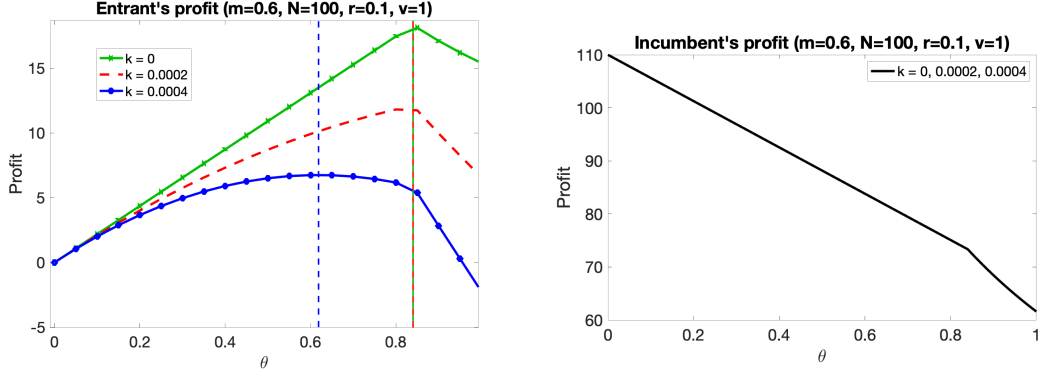
We again find that because mobile buyers reduce the incumbent's incentive to fight, both the incumbent and the entrant can charge (weakly) higher prices to buyers while maintaining the same wages as r increases. They both have more transactions when r increases. The incumbent's profit increases with r , while the entrant's profit increases with r when its advertising cost does not increase too much with r .

Note that when θ is larger than the threshold, as θ increases, the profits of both platforms decrease due to intense competition even without considering advertising cost. Thus, we expect the entrant's optimal choice of θ to be no more than the threshold $\left(\min\left(\frac{2m(2+r)}{3v}, 1\right)\right)$. That is, it is in the best interest of the entrant not to trigger the incumbent's competitive response.

Corollary 2.2.1. *The entrant's optimal choice of advertising intensity θ always satisfies $\theta \leq \min\left(\frac{2m(2+r)}{3v}, 1\right)$.*

The exact optimal level of θ for the entrant depends on the cost of advertising, $L(n) = L(\theta(2N + rN))$. Following the literature, Thompson and Teng (¹⁰⁴), Tirole (¹⁰⁶), Esteves and Resende (³⁹), Jiang and Srinivasan (⁶⁶), we assume a quadratic cost function, $L(n) = kn^2$, where $k \geq 0$. A large k suggests that advertising is costly, while a small k suggests that it is inexpensive.[§]

[§]If there is no cost for the entrant to reach its fans, we can modify the cost function to be of the form $L(n) = k(\max(n - z, 0))^2$, where z is the total number of fans. Our results hold qualitatively.



(a) Entrant's profit. The vertical lines indicate the optimal θ (b) Incumbent's profit for each scenario.

Figure 2.2: Firm's profit vs. advertising intensity θ .

Figure 2.2 illustrates how the entrant's profit changes with the choice of θ for different values of k . We notice that for a given level of k , the entrant's profit increases and then decreases with θ . Even if advertising has no cost (i.e., $k = 0$), there is an optimal advertising level for the entrant. As k increases (i.e., advertising becomes more expensive), the optimal advertising intensity, θ^* , decreases. However, the incumbent's profit always decreases with θ and is independent of k . The following proposition formalizes the relationship between the optimal advertising intensity, θ^* , and the value of k .

Proposition 2.2.2. *The optimal advertising intensity, θ^* , depends on the value of k .*

- i If $k \geq \max\left(\frac{3(1+r)v^3}{16m^2N(2+r)^4}, \frac{(1+r)v^2}{8mN(2+r)^3}\right)$, then $\theta^* = \frac{1+r}{8(2+r)^3kNm}$, which decreases with r . The entrant's profit is $\frac{(1+r)^2v^4}{64km^2(2+r)^4}$ and the incumbent's profit is $\frac{N(1+r)v}{2} \left(2 - \frac{(1+r)v^3}{8kNm^2(2+r)^4}\right)$.
- ii If $0 \leq k < \max\left(\frac{3(1+r)v^3}{16m^2N(2+r)^4}, \frac{(1+r)v^2}{8mN(2+r)^3}\right)$, then $\theta^* = \min\left(\frac{2m(2+r)}{3v}, 1\right)$, which weakly increases r . When $\frac{2m(2+r)}{3v} < 1$, the entrant's profit is $\frac{N(1+r)v}{6} - \frac{4kN^2m^2(2+r)^4}{9v^2}$ and the incumbent's profit is $\frac{2Nv(1+r)}{3}$. When $\frac{2m(2+r)}{3v} \geq 1$, the entrant's profit is $\frac{N(1+r)v^2 - 4kN^2m(2+r)^3}{4m(2+r)}$ and the incumbent's profit is $\frac{N(1+r)v}{2} \left(2 - \frac{v}{m(2+r)}\right)$.

We have two cases. When k is large, advertising is costly. In this case, the optimal advertising intensity $\theta^* \leq \min\left(\frac{2m(2+r)}{3v}, 1\right)$. The entrant and the incumbent have no strategic interactions with each other and the entrant's optimal advertising intensity is determined by the marginal benefits and marginal cost from reaching another user. Consequently, the entrant's equilibrium profit is independent of the market size, N . This result also highlights the impact of network interconnectivity, independent of the market size.

When k is small, advertising is inexpensive, and the entrant platform, thus, has an incentive to increase advertising intensity θ . The entrant's profit increases with θ until $\theta = \min\left(\frac{2m(2+r)}{3v}, 1\right)$. When $\frac{2m(2+r)}{3v} \geq 1$, the entrant will advertise to everyone in the market. Otherwise, the entrant's profit first increases as θ increases up to $\frac{2m(2+r)}{3v}$ and, then—because of the competitive response from the incumbent discussed in Corollary 2.2.1—decreases with θ afterwards. Thus, the entrant will choose $\theta^* = \frac{2m(2+r)}{3v}$. Note that the entrant's optimal choice of θ is independent of market size N but increases with r , which again highlights that market size and network interconnectivity affect equilibrium outcomes differently.

The discussion above leads to the following corollary:

Corollary 2.2.2. *Even if the advertising cost is zero (i.e., $L(n) = 0$ or $k = 0$), the entrant will not necessarily advertise to the entire market but instead choose the optimal advertising intensity $\theta^* = \frac{2m(2+r)}{3v}$ when $\frac{2m(2+r)}{3v} < 1$.*

We then examine how the fraction of mobile buyers, r , affects the optimal θ and the platforms' profits in the two cases in Proposition 2.2.2. Figure 2.3 illustrates the relationships under different values of k . When k is large (in Proposition 2.2.2i), advertising is costly. As r increases, the number of service providers, $(1 + r)N$, increases in the market, but the number of buyers accessible to the entrant remains the same. Thus, the likelihood that advertising is wasted on some service providers without matched buyers also increases. With a large k , it is optimal for the entrant to reduce θ^* to

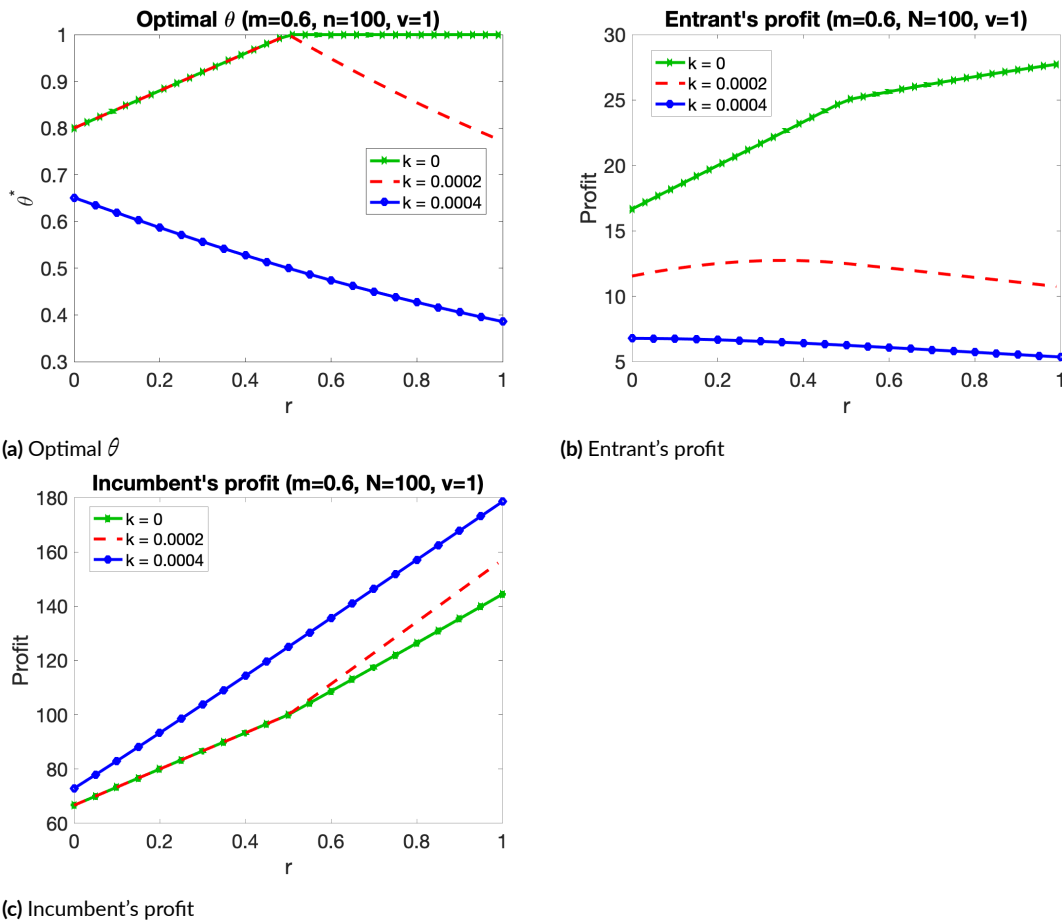


Figure 2.3: The entrant's optimal advertising intensity and firms' profits under different values of interconnectivity coefficient r

reduce its advertising cost, $L(\theta^*(2N + rN))$, even if a large r reduces the incumbent's incentive to respond. This explains the declining curve in Figure 2.3.a for a large k (e.g., $k = 0.0004$). Because the entrant advertises to fewer buyers, the entrant's profit also decreases with r , as depicted in Figure 2.3.b for $k = 0.0004$.

In contrast, when k is small (in Proposition 2.2.2ii), θ^* (weakly) increases with r . This is because when advertising is inexpensive, the advertising wasted on unmatched service providers becomes a less significant issue and the entrant wants to take advantage of the incumbent's disincentive to

respond instead. Thus, we observe an increasing curve of θ^* in Figure 2.3.a for a small k (e.g., $k = 0$). The impact of r on the entrant's profit is also positive as long as k is sufficiently small. ¶ This result is consistent with Proposition 2.2.1, where we have shown that if the advertising cost is small for the entrant, the entrant's profit will increase with r regardless of θ . When we use k to capture the cost of advertising, as long as k is sufficiently small (e.g., $k = 0$ in Figure 2.3.b), the entrant's profit increases with r . The result shows that when the incumbent has more captive buyers and, therefore, less incentive to fight, the entrant could be more profitable when advertising is not costly.

Note that the threshold of k , $(\max(\frac{3v^3}{32Nm^2(2+r)^3}, \frac{v^2}{8mN(2+r)^3}))$, below which the entrant's profit increases with r , is a decreasing function of r . Therefore, an intermediate value of k may begin from below the threshold when r is small, but then exceed the threshold as r increases. This implies that the equilibrium may switch between the two cases (where k falls above or below the threshold) as r changes. This explains why we may observe a non-monotonic relationship between the optimal advertising intensity (θ^*) and r and the same for the relationship between the entrant's profit and r , as shown by the case of $k = 0.0002$ in Figures 2.3.a and 2.3.b.

Regardless of the value of k , the incumbent's profit always increases with r , because it has more captive buyers when r is larger (as illustrated in Figure 2.3.c).

Below we summarize the relationship between the entrant's advertising intensity and the fraction of mobile buyers in Corollary 2.2.3 and the relationship between the platform profit and the fraction of mobile buyers in Proposition 2.2.3:

Corollary 2.2.3. *When k is small, as the fraction of mobile buyers, r , increases, the entrant has incentive to advertise more (higher θ^*) until it reaches the entire market. Conversely, when k is large, as r increases, the entrant has incentive to reduce advertising (lower θ^*). For intermediate values of k , as r increases from zero, the optimal θ^* increases with r first and then decreases with r .*

¶When $\frac{2m(2+r)}{3v} < 1$, $\theta^* = \frac{2m(2+r)}{3v}$ and the entrant's profit increases with r , as long as $k < \frac{3v^3}{32Nm^2(2+r)^3}$.
When $\frac{2m(2+r)}{3v} \geq 1$, $\theta^* = 1$ and the entrant's profit increases with r , as long as $k < \frac{v^2}{8mN(2+r)^3}$.

Proposition 2.2.3. *The incumbent's profit always increases with the fraction of mobile buyers, r . How the fraction of mobile buyers, r , affects the entrant's profit depends on the value of k . When k is small, the entrant's profit increases with r , and conversely, when k is large, the entrant's profit decreases with r . For intermediate values of k , as r increases from zero, the entrant's profit increases with r first and then decreases with r .*

The results suggest that under certain circumstances (e.g., when advertising is cheap), network interconnectivity, which does not influence the size of the entrant's potential demand in a single market but is supposed to benefit the incumbent that operates in multiple interconnected markets, may, in fact, encourage the entrant to enter the market and increase entrant profit. In this case, a higher network interconnectivity will make it even more difficult for an incumbent to deter entry. These results have important managerial implications for platform owners to understand their competitiveness in markets with a given level of market interconnectivity for resource planning and marketing strategy design, and for policymakers to take into account the network interconnectivity when considering the anti-competitive issues in platform markets. These results also have important implications for the entrant regarding the choice of market to enter when the fraction of incoming mobile buyers varies across markets; we will examine this further in the next section.

2.3 EXTENSIONS

||

||In the extensions, we make a similar assumption as in the main model that m is sufficiently large so that, in equilibrium, the entrant will not take away the entire segment of users who are aware of both platforms. Because the condition changes in different extensions, for consistency, we use the most restrictive condition, $m > \frac{3v}{5}$, in all extensions. This approach does not affect the key insights drawn from different extensions.

2.3.1 HETEROGENEOUS MARKETS

In our main analysis, we assume that all markets are homogenous. Consequently, the entrant could begin by entering any one of these markets. If these markets have different fractions of mobile buyers visiting from other markets, assuming the entrant only has the budget to enter one market only, which market should the entrant choose to enter?

Suppose there are H markets. Let r_b be the fraction of mobile buyers coming into market b ($b = 1, 2, \dots, H$). Hence, market b has N local buyers, $(1+r_b)N$ service providers, and $r_b N$ incoming mobile buyers. We obtain the following proposition: **

Proposition 2.3.1. *When k is small, the entrant should choose the market with the highest fraction of mobile buyers from other markets, r , to enter; when k is large, the entrant should choose the market with the lowest fraction of mobile buyers from other markets, r , to enter. For an intermediate value of k , the entrant may choose a market where r is also intermediate***.*

The result echoes Proposition 2.2.3, where we find that the entrant's profit increases with r when k is small, decreases with r when k is large, and has a non-monotonic relationship with r for an intermediate value of k . While Proposition 2.2.3 focuses on how the entrant's profit changes when r in a local market increases, this proposition extends our finding to how the entrant should choose a market among the markets that vary in the fraction of incoming mobile buyers. The proposition suggests that when the fraction of visitors is high, the incumbent is less likely to fight the entrant. At the same time, however, a high fraction of visitors means that a large fraction of the entrant's advertising expenditure will be wasted on unmatched service providers. Hence, the entrant will find such

**Proofs of all lemmas and propositions for the extensions are provided in the online appendix.

***The thresholds for k are provided in the online appendix. We also explored heterogeneous market sizes. As indicated by Proposition 2.2.2, when k is sufficiently large, market sizes would not affect the entrant's profit. When k is small, if market size is sufficiently large for a certain market, the fraction of mobile buyers (r) will have a negligible effect and, thus, the entrant should simply choose the largest market to enter; otherwise, the entrant's choice will depend on both r and N .

a market attractive when advertising is not costly. For example, if Google wants to offer ride-sharing services because it already has a larger number of users from its current services and can build awareness at a low cost (k is small), Google should start offering these services in large cities with a large fraction of visitors. However, a new startup, for which advertising is rather costly, should target small cities with a small fraction of visitors in order to improve advertising efficiency.

2.3.2 THE INCUMBENT DOES NOT OWN THE ENTIRE MARKET

In our main model, we also assume that the incumbent owns the entire market (i.e., all potential buyers and service providers are aware of the incumbent) before the entrant emerges. In reality, it is possible that not every user in the local market is aware of the incumbent. Thus, it is possible for the entrant to attract users who are not aware of the incumbent. We consider this possibility in this extension. Assume the incumbent's market share before the entrant arrives is s , where $0 < s < 1$. We then have the following proposition:

Proposition 2.3.2. *The results from our main model are qualitatively the same when $s \geq \frac{m(2+r)}{2m+mr+v+rv}$. If $s < \frac{m(2+r)}{2m+mr+v+rv}$, both platforms charge buyers $p_I^* = p_E^* = v$ and offer service providers $w_I^* = w_E^* = 0$.*

The results from the main model remain qualitatively the same as long as s is sufficiently large. But when s is below a certain threshold, the results differ from our main results. When the incumbent has a small share of the market, the entrant and the incumbent can effectively avoid direct competition by targeting different segments of that market. Hence, both will charge monopoly prices and offer monopoly wages, and no buyers and service providers will switch from the incumbent to the entrant.

2.3.3 MOBILE BUYERS ONLY CONSUME WHEN THEY TRAVEL

In our main model, mobile buyers purchase services in both their local markets and the markets they visit. This assumption fits with markets such as those in the ride-sharing industry, where riders hail cars in their own markets and also in other markets when they travel, or daily local deal markets, where consumers buy deals in their own markets and also in other markets when they travel. In this case, interconnectivity affects the size of the potential market for the incumbent but does not affect the size of the potential market for the entrant. Our main model thus enables us to examine the impact of market interconnectivity on the entrant independent of the market size effect. Interestingly, although the size of the potential market for the entrant is unchanged, its profit may increase with the interconnectivity between markets. This possible profit enhancement for the entrant, independent of the market size effect, is the most interesting result of our model and provides novel insights regarding the role of market interconnectivity in influencing platform competition.

Although the assumption that mobile buyers purchase services in both their local markets and the markets they visit is consistent with the practice for many platforms, in this extension, we examine to what extent our results are affected by this assumption by looking at the scenario in which mobile buyers do not consume in their local markets. We obtain the following result under this assumption:

Proposition 2.3.3. *The results from the main model are qualitatively the same when mobile buyers do not consume in their local markets, except that the entrant's profit under the optimal θ always decreases with r .*

Unlike the main model, in this case, by assuming away local consumption, we keep the size of the potential market for the incumbent fixed. The size of the potential market for the entrant, however, changes with interconnectivity: a larger fraction of mobile buyers will have fewer potential buyers for the entrant. The results suggest that the incumbent's profit always increases with r , irrespective

of whether or not local consumption occurs. However, although the entrant can continue to take advantage of the incumbent's disincentive to fight and advertise more aggressively, its demand decreases (i.e., the market size effect and interconnectivity effect take place jointly). Consequently, its profit decreases with r . In the case of Airbnb, for example, travelers typically do not care about the number of hosts in their home cities; they care more about the number of hosts in the cities they wish to visit.

This result explains why it is more difficult to challenge an incumbent platform like Airbnb for which local consumption occurs less frequently compared to one like Uber.

2.3.4 THE ENTRANT CAN TARGET MOBILE BUYERS

In the main model, we assume that the entrant is not able to advertise to mobile buyers. We make this assumption because mobile buyers often stay in the market they visit briefly. Even if the entrant is continuously advertising in that market, without sufficient exposure to its advertisement, a mobile buyer may not consider the entrant's product. We now relax this assumption and assume that advanced targeting technologies can help the entrant identify mobile buyers and can advertise to them effectively. Let θ and θ_i be the fractions of local and mobile users, respectively, that become aware of the entrant's platform after the entrant's advertising. The demand on the service provider side remains the same as in Equations (2.1) and (2.2). The demand on the buyer side becomes

$$N_I^B = \left(1 + r - \frac{a^*}{m}(\theta + \theta_i r)\right) N \quad (2.9)$$

$$N_E^B = \frac{a^*}{m}(\theta + \theta_i r) N \quad (2.10)$$

Note that local advertising and advertising to mobile buyers are different in that when advertising to local, the entrant advertises both to buyers and service providers, which helps balance demand and supply; however, mobile users only include buyers and, thus, advertising targeted mobile buyers can target buyers only. Consequently, advertising to the two groups of users has different effects on the pricing strategies of the entrant and incumbent. In this case, we find that the entrant may not want to advertise to the mobile buyers even if there is no cost of advertising, as summarized by the following proposition.

Proposition 2.3.4. *Even if the cost of advertising is zero (i.e., $L(n) = 0$), the entrant will not necessarily choose to advertise to mobile buyers even if it is able to, that is, $\theta_t^* = 0$ if $\frac{2m(2+r)}{3v} \leq 1$.*

As we show in the main analysis, mobile buyers help deter the incumbent from fighting with the entrant. Hence, the entrant may not want to steal the mobile buyers from the incumbent even when it can target them and advertise to them at zero cost. The entrant is more likely to avoid advertising to the mobile segment when the value of these buyers to the incumbent is high (large v), the mobile segment is not large so the entrant does not lose a huge number of potential buyers (small r), and a small amount of advertising can steal a large number of mobile buyers away from the incumbent and thus trigger its response (small m). When the advertising cost for mobile buyers is higher than the cost for local users, the entrant will be even less likely to advertise to mobile buyers. The only situation in which the entrant will advertise to mobile buyers is when the entrant has advertised to all local buyers and has not triggered competitive responses from the incumbent, which is rarely observed in practice. Thus, Proposition 2.3.4 helps justify the assumption in our main model that mobile buyers are only aware of the incumbent.

2.3.5 THE PRESENCE OF NETWORK EFFECTS

In our main model, we focus on matching between the buyers and service providers. Similar to other matching models (¹¹⁴), we do not model network effects. This approach enables us to separate the network-interconnectivity effect from the network effects, but network effects may have an impact on matching quality or speed. For example, in the case of ride-sharing services, a large number of drivers on a platform can reduce the wait time for riders. Similarly, a large number of riders reduces the idle time for drivers. In the accommodation market, a large number of hosts and travelers on a platform increase the likelihood that each traveler and each host is matched with a party close to his or her personal preference. To capture such benefits, we add a utility component to capture the network effects in the buyers' and service providers' utility functions and allow this utility component to increase with the number of users on the other side of the same platform:

$$U_I^B = eN_I^S + v - p_I \quad (2.11)$$

$$U_E^B = eN_E^S + v - p_E - a_i \quad (2.12)$$

$$U_I^S = eN_I^B + w_I \quad (2.13)$$

$$U_E^S = eN_E^B + w_E - c_i \quad (2.14)$$

Here, we use parameter e ($e \geq 0$) to capture the strength of network effects. We first consider the case where e is small and both platform firms can co-exist. To avoid multiple equilibria due to network effects, we assume e to be much smaller than the value of the transaction it-

self. (Mathematically, we require $e < \min\left(\frac{v}{2N}, \frac{m}{4N}\right)$). This assumption is reasonable because in such markets most benefits to buyers or service providers come from the transaction itself. We find our main results to be qualitatively unchanged, as summarized in the following proposition:

Proposition 2.3.5. *The results from the main model are qualitatively the same in the presence of network effects when the strength of network effects is small.*

We also examine how the strength of network effects affects the profits of both the entrant and incumbent. Given the computational complexity, we explore this effect as the strength of network effects, e , approaches zero. We find that as long as m is sufficiently large (e.g., $m > v$), because the incumbent has a larger market share, network effects make the incumbent more attractive to users, thereby reducing users' tendencies to switch to the entrant. Hence, as the network effects become stronger, the entrant's profit decreases and the incumbent's profit increases.

When e is sufficiently large, we find that the equilibrium in which the entrant has positive demand cannot be sustained and the incumbent becomes the monopoly. This result is expected because when network effects dominate pricing effects, if an entrant enters the market, the incumbent always has the incentive and is able to take advantage of its installed base advantage to drive the entrant out of the market.

Proposition 2.3.6. *When network effects become sufficiently large, the incumbent can deter the entrant from entering the market and thereby monopolize the market.*

2.3.6 HETEROGENEOUS SWITCHING COSTS

In our model, we assume that buyers and service providers face the same switching costs. In practice, however, their switching costs may differ. For example, in the ride-sharing industry, riders only need to download a new app to switch to a different platform, while drivers may have to undergo background checks and verification processes to switch to a different platform. In order to investigate

how heterogeneity in switching costs affects the platforms, we allow buyers' switching cost to be uniformly distributed between zero and m_b and service providers' switching costs to be uniformly distributed between zero and m_s . Then, we obtain the following proposition:

Proposition 2.3.7. *When we allow buyers' switching cost to be uniformly distributed between 0 and m_b and service providers' switching costs to be uniformly distributed between 0 and m_s , we have $\frac{\partial \pi_E^*}{\partial m_b} < \frac{\partial \pi_I^*}{\partial m_s} < 0$ and $\frac{\partial \pi_I^*}{\partial m_b} \geq \frac{\partial \pi_I^*}{\partial m_s} \geq 0$.*

Proposition 2.3.7 suggests that an increase in switching costs on the buyer side harms the entrant or benefits the incumbent more than the same increase on the service provider side. The intuition is that because of the existence of mobile buyers, we have more service providers than local buyers. Hence, the total number of transactions that the entrant platform serves depends largely on the number of buyers the entrant can incentivize to switch to the entrant platform. Thus, buyers' switching cost affects firm profits more than that of service providers. Note that when the optimal θ, θ^* , reaches the threshold that is just high enough to not trigger the incumbent's response, the incumbent's profit is independent of m_b and m_s . This explains why sometimes $\frac{\partial \pi_I^*}{\partial m_b} = \frac{\partial \pi_I^*}{\partial m_s} = 0$.

2.4 DISCUSSION AND CONCLUSION

Extant studies in the platform strategy literature typically assume that each participant on one side of a market is (potentially) connected to every participant on the other side of the market. Our research departs from this assumption to explore the impact of network interconnectivity on the defensibility of an incumbent with presence in multiple markets against an entrant that seeks to enter one of these markets.

As depicted in Figure 2.4, our model captures heterogeneous network interconnectivity across different industries, ranging from isolated network clusters ($r = 0$) to a fully connected network ($r = 1$). Examples of isolated local clusters (i.e., no mobile buyers) include Handy, a marketplace

tions where markets are heterogenous, the entrant is able to target mobile buyers, buyers and service providers have different switching costs, and network effects are present. Overall, these results help explain barriers to entry in platform markets and the resulting performance heterogeneity among platform firms in different markets.

These results corroborate empirical observations of many platform markets. For example, we show that it is optimal for an entrant not to trigger incumbent responses. The founders of Fasten, an entrant into the ride-sharing market in Boston, were very clear from the beginning that they did not want to trigger Uber's response by strategically minimizing their advertising activities. Fasten also chose not to target visitors in Boston: it did not advertise in Boston's Logan Airport or in its South Station Bus Terminals. Indeed, although Fasten grew rapidly in Boston during the period 2015–2017, Uber and Lyft did not change their prices or wages to compete. As a counterexample, when Meituan—a major player in China's online-to-offline services such as food delivery, movie ticketing, and travel bookings—entered the ride-sharing business, it was able to build awareness of its service at almost no cost through its existing app, which had an extensive user base. Meituan's entry into the Shanghai ride-sharing market triggered strong responses from the incumbent, Didi, thereby leading to a subsidy war between the two companies. Meituan subsequently decided to halt ride-sharing expansion in China.

The results also suggest that Airbnb's and Booking.com's business models are more defensible than those of Uber because most of their customers are travelers and do not use the service in their local markets as often, while Uber's consumers primarily use its services in their local markets. The difference in defensibility is a key aspect for why Airbnb and Booking.com were able to achieve profitability, while Uber has been hemorrhaging money. ^{††}

^{††}See, for example, <https://techcrunch.com/2019/01/15/ahead-of-ipo-airbnb-achieves-profitability-for-second-year-in-a-row/>, <https://www.macrotrends.net/stocks/charts/BKNG/booking-holdings/gross-profit> and <https://www.reuters.com/article/us-uber-ipo/uber-unveils-ipo-with-warning-it-may-never-make-a-profit-idUSKCN1RN2SK>, accessed October 2019.

The study offers important managerial implications for platform owners. We find that an incumbent's profit increases with interconnectivity regardless of whether or not mobile buyers consume in local markets; thus, incumbent platforms should seek to build strong interconnectivity in their networks. In our model, the level of interconnectivity is given exogenously; however, in practice, how firms design their platforms can influence interconnectivity. For example, while Craigslist is a local classifieds service, its housing and job services attract users from other markets. Our research suggests that such services are important sources of Craigslist's sustainability and, thus, Craigslist should strategically devote more resources to grow these services. As another example, many social networking platforms such as Facebook and WeChat allow companies or influencers to create public accounts that any user can connect with. Such moves increase interconnectivity among their local network clusters.

This research suggests that an entrant needs to conduct a thorough network analysis to understand the interconnectivity among different markets, the strength of network effects, the capability of its targeting technologies, and whether or not mobile users consume in their local markets. These factors, together with the cost of reaching users, can help inform the entrant's location choice and how aggressively it should build awareness in a new market. The entrant needs to realize that even if advertising incurs little cost, it is not always optimal for it to advertise to every user. The entrant should advertise to the extent that it does not trigger competitive responses from the incumbent. Equally important, it is not always the case that an entrant should choose a market with low interconnectivity. When advertising is inexpensive and mobile buyers consume in local markets, it could be more profitable to enter a market with high interconnectivity.

This research also offers important implications for policymakers. With the growing popularity of digital platforms, policymakers around the world are increasingly concerned about the market power of these platforms. Our research suggests that regulators should pay close attention to the network structures of these platform markets to improve their understanding of market competi-

tiveness and entry barriers.

As one of the first research that explicitly models network interconnectivity of platform markets, our research opens a new direction for future research on platform strategies. For example, our model focuses on an entrant's entry strategy and only allows the incumbent to react through pricing. Future research could consider the incumbent's perspective and examine other strategies for entry deterrence.

This research focuses on examining an entrant with limited resources (to overcome entry cost in each market) and an incumbent that already exists in many markets. Even if we allow the entrant to enter more than one market, as long as the number of markets the entrant can realistically enter is small compared to the total number of available markets available (which is true in most cases), our results would not change qualitatively because network interconnectivity (or awareness spillover) plays a rather minor role for the entrant relative to the incumbent. Take the Uber and Fasten cases as examples. Even if Fasten enters a second market, the number of Fasten users from that market to Boston is rather small compared to the number of Uber users from the hundreds of cities outside Boston to Boston. This also implies that Fasten's advertising in the second market has little impact, relative to Uber, on the first market that Fasten entered. Future research can extend our analysis to examine cases involving a resourceful entrant that can enter many markets at once, such as in the case of Uber vs. Grab in Southeast Asia.

In our model, one buyer and one service provider are matched during each transaction. In other words, at a given time, a buyer cannot buy from multiple service providers (regardless of whether they are on the same platform or different ones) and a service provider cannot serve multiple buyers (regardless of whether these buyers are on the same or different platforms). This assumption matches with the rides-sharing industry in that the same rider or the same driver does not show up in multiple cars at a time. If we allow multiple transactions for each user, we may observe multi-homing in that a rider may be matched to Uber drivers for certain transactions and Lyft drivers

for other transactions. Future research can extend our model to incorporate multiple transactions for each user and allow buyers and service providers to multi-home. In this case, the entrant needs to decide on the entry and advertising strategy based on how many transactions it expects to serve. Buyers and service providers will decide whether to adopt the entrant platform based on their switching costs and expected benefits from future transactions. While the game will be more complicated, we believe that our key insights would continue to hold. For example, in equilibrium, only the buyers and service providers with low switching cost will adopt the entrant platform to multi-home. Incoming mobile buyers will continue to disincentivize the incumbent to respond in each period, which ultimately drives the impact of network interconnectivity on the entrant's advertising strategy and profitability, as illustrated in our model.

Furthermore, to focus on the impact of network interconnectivity, we abstract away many factors that could influence competitive interactions between incumbents and entrants. For example, in the ride-sharing industry, riders may not care much about vehicle features. However, in the accommodation industry, travelers are likely to care about the features of properties, thereby making it easier for an entrant into the accommodation industry to differentiate itself from an incumbent and reducing the competitive intensity. In addition, because of tractability, we could not examine all possible parameter values after incorporating network effects into our main model. Future research could further explore how these factors affect competitive interactions.

3

Time and the Value of Data

We witness a dramatic acceleration of digitization in firms' infrastructure, products, and services. Artificial Intelligence (AI) enabled solutions are on the rise, and more than ever, data appears to be a critical strategic asset (², ⁶, ²¹, ³⁴). As a result, in almost all industries and economic sectors, firms amass substantial volumes of user data to improve their current and future services, anticipating that it also gives them an advantage over their competitors. In regulatory debates, this accumulation of data by firms is considered to be a critical source of competitive advantage that could lead to a con-

centration in digital markets (^{21, 35, 44, 107}). In addition, from users perspective, there are privacy concerns (¹⁸) on when and how firms use the accumulated data and if in any way it can adversely harm users. Because of this aggressive data accumulation by firms, it is crucial to understand how and when the value created by AI-enabled services scales with the size of available data. Particularly, since the data accumulation process often happens over time, it is of great interest to research how the created business value changes over time, especially when the dataset is sampled from a dynamically changing environment.

Current literature on how the increase in the stock of available data scales the business value has mixed results. Managers often believe that collecting more data continually improves the accuracy of machine learning models. This belief is engrained by the statistical theories on how the accuracy of machine learning models scales with the dataset size (^{23, 57}) and the managerial theories on how the economic value created by a firm scales with the resources (^{2, 6, 21}). All these theories attest that more data is always better, and securing a vast amount of such resource leads to the firm's success in the long run. In addition, recent research hypothesizes a feedback loop (^{52, 65}) between the size of available data and the quality of AI-based solutions. (⁵²) theorize and compare this data externality to network effects, where the value of a service or product increases in user-base size. In this "data network effect" (^{52, 54, 90}), more data leads to a higher accuracy of algorithms, which means better services (⁵⁷). Better service then leads to a higher user engagement or a larger user-base, which creates even more data. Despite these theories, empirical research (^{11, 29}) finds limited or no economic significance in accumulating large datasets. For example, (²⁹) investigates the effect of historical search data on search results' quality. They found little empirical evidence on the effectiveness of old data in the quality of search engine results. (¹¹) also raise a similar question on the data's economies of scale for specific problems. They suggest a diminishing return to scale value model for data and argue that increasing data volume in advertisement applications does not improve the service quality. We believe that diminishing return to scale theory doesn't explain the aggressive data

accumulation by firms already equipped with massive datasets.

This research investigates how the business value changes over time when the dataset is sampled from a dynamically changing environment and how this change explains the mixed results seen in the literature. A dataset sampled from a dynamically changing environment loses relevance over time, making the created business value time-dependent. This time-dependency is referred to as concept drift in the machine learning literature. Concept drift is known to cause a deterioration in the algorithm's performance. It manifests itself as a decrease in the algorithm's accuracy score or an increase in its loss value or error. However, the extent to which it affects accuracy or loss values and how time-dependency affects a firm's data strategy is still unknown.

Our approach to the problem has similarities and differences with machine learning and AI literature. Similar to machine learning literature, we model the change in environment with a shift in data-generating probability distributions. In contrast with the literature, we fix a model/task and vary the data generating distribution to study the effect of time-dependency on business value. In machine learning research, given a dataset, researchers alter the models to improve the accuracy score or the loss value. In addition to this difference, we also define new metrics such as "equivalent size" and "effectiveness" to compare the value of datasets sampled from different times. Unlike machine learning literature which reports the effect of time-dependency in accuracy scores or loss values, our measures report the effect in dataset sizes. In doing so, for any machine learning task, we first define an oracle dataset. Subsequently, we train our model on the given dataset (referring to as baseline dataset) and then measure the model's accuracy score by testing it on the oracle dataset. We then ask what size of the oracle dataset leads to a similar accuracy score if used for training the model. We call it the equivalent oracle size. Thus, we can quickly compare various datasets by comparing their equivalent sizes. Another benefit of measuring the value this way is that we can borrow terminology from economics and management research, making our findings more relatable to a broader audience. For example, if a dataset's equivalent oracle size declines over time, we call the dataset per-

ishable.

In our theoretical setting for this research, we investigate the effect of time-dependency for the task of learning the probability distribution and use the maximum likelihood estimation method to accomplish this task. Learning the data-generating probability distribution is a fundamental problem in the statistical learning theory. It is because we can evaluate any statistics (like expectation or variance of any quantity of interest) from the distribution. Hence, we believe that theorems and propositions we prove for this task can be, with slight modification, used for a wide variety of other tasks. We use the Maximum Likelihood Estimation (MLE) for our analysis since consistency and efficiency are essential for our mission in this research. Consistency is critical since we model dynamically changing environments using probability distributions. Hence, for any given time, it is crucial to learn the distribution consistently. Efficiency is essential to achieving the lowest estimation variance with the smallest dataset size. Intuitively, efficiency makes MLE the most scalable method (in gaining a better accuracy score) for a fixed dataset size and hence, a wise choice by firms.

We derive several managerial and economic intuitions by comparing the value of datasets sampled from different times. We argue that due to shifts in the data generating distribution, it may be optimal for a firm to collect a more limited amount of recent data instead of keeping around an infinite supply of older data. This is a direct result of our first proposition in this research. This proposition shows that even a perfect model trained on an infinite supply of time-dependent data may have lower accuracy than the same model trained on a recent (perfectly relevant) dataset of limited size. In other words, a less relevant dataset of infinite size has a finite (bounded) equivalent oracle size (defined as the perfectly relevant dataset in this case). Hence, a competing firm with an oracle dataset of sufficient size can easily attain a better accuracy score or loss value. This proposition has several other economic and managerial implications that we discuss later in this research. In pursuit of comparing the value of datasets from different times, we define a substitution function that measures how much oracle size a firm gains/loses if it substitutes its baseline dataset with another dataset of

the same size from a different time. We prove that substitution gain is a function of the baseline dataset size as well as time. It becomes sharper with the increase in the size of the baseline dataset, meaning that the gain/loss percentage increases as the baseline dataset size increases. As we discuss later in this chapter, it has immediate implications for firms regarding training frequency, i.e., how often firms should retrain their models.

We use the machinery we developed for comparing values of datasets sampled from different times to compare values of datasets curated over a period of time. We do so by defining the “equivalent time” in Proposition 3.3.1. This proposition states that, for any baseline dataset curated over a period of time, there exists an “equivalent time” such that, fixing the size, a model trained on a dataset from the equivalent time produces a similar accuracy score as the model trained on the baseline dataset. As a result, we can compare datasets curated over various time periods by first calculating their equivalent times and then by comparing the oracle sizes each equivalent time produces. A direct result of this method is the introduction of offloading algorithm. Offloading algorithm removes less relevant data hoping that gain in relevance counterbalances the loss in size. We then use the offloading algorithm to argue that increasing the stock of data by including older datasets may, in fact, damage the model’s accuracy, putting a firm in a disadvantageous position. Together with the increase in sharpness of substitution gain as a function of the flow of data (Number of data points in a given time or the rate of acquiring new data points), these results build the case for defining the optimal scaling and growth path for a firm. When the firm is small, the optimal growth path focuses on the stock of available data curated over time. As the flow of data increases (A firm acquires more users or user-engagement increases, for example), the firm offloads older data and focuses on the flow of data as the primary value driver.

To confirm the economic significance of our findings, we empirically measure the decline in the value of data for the next-word-prediction task. It is a widely used machine learning task with applications in auto-completion software in cellphones and the search recommendations in search

engines. In our experiment, we use a user-generated text dataset from Reddit.com (⁴⁰). We divide this dataset into smaller datasets based on data points' sampling time (Month-year format). We then train a variation of GPT-2 from OpenAI (⁵¹) on each of these smaller datasets and measure their equivalent sizes over time. Our measurements confirm the economic significance of time-dependency as we show that in roughly seven years, 100MB of text data becomes as valuable as 50MB of current data for the next-word-prediction task.

Our findings can explain the mixed the result in the literature. We acknowledge that increasing the dataset size improves the accuracy of machine learning models. Accordingly, we find the feedback loop logic compelling. However, we show that the stock of available data produced by the feedback loop has a limited oracle size because of time dependency. Hence, despite the accelerated growth in the size of the data repository, we shouldn't expect a significant increase in created business value. In other words, the feedback loop stalls in dynamically changing environments unless the firm offloads its less relevant data and focuses on the flow of data as the primary value driver. This finding supports the reported results in (²⁹) and (¹¹) since both search engine (²⁹) and advertisement (¹¹) businesses use time-sensitive data and hence, face significant time-dependency.

Our work also adds to machine learning and economics literature in several ways. This research adds to machine learning (particularly Natural Language Processing) literature by providing a different view of the domain/concept shift problem. Our method and analysis make machine learning researchers and practitioners better explain the tradeoffs and challenges that variation in training data has for their models. For example, our method allows them to realize how often they should retrain their models. As a result, they can formulate a better scaling/growth plan by adequately crafting their data management and resource prioritization strategy. It is worth noting that in this research, we only measure the predictive value of data. Hence, we don't talk about the value of data for inference. There is a subtle distinction between the two in the statistics literature.

We also contribute to the economics literature by providing a better understanding of data's time

dependency and its implication on modeling data economy and the growth of digital firms. In the economics literature, the impact of data on economy and AI's widespread applications have been examined by several authors (e.g., ^{5, 7, 14, 19, 31, 33, 67, 72, 82, 93, 107, 111}). In this literature, AI is a general-purpose technology that brings down the cost of prediction. Accuracy of the prediction increases with the size of training datasets, making a case for arguments on the importance of data in the growth of digital firms. Such arguments motivate research on how data influences firm dynamics (^{41, 42}) and how it disproportionately benefits large firms (¹⁷), which then stimulates debates on the implication of AI and, more precisely, data on competition (^{34, 35, 76, 85, 89, 94}). Of course, the degree to which data impacts business value and hence the competition varies with the design parameters like the degree of personalization (^{58, 96}) or the externalities between recommendation clusters (¹³). Nevertheless, studying the effect of data's time dependency on competition remains crucial. Several lines of research (^{21, 67, 76, 94, 110}) recruit the resource base view as a framework to explore how a firm can exploit data to create a sustainable competitive advantage. These researches mostly focus on the non-rivalry, exclusivity, or imitability of data. In our research, we argue that time-dependency plays a major role as well. We show that the business value doesn't solely scale with the size of available data, and even a dataset of infinite size may have a finite equivalent oracle size. As we discuss more in section 3, an immediate impact of this result is that we can't use regular discounting functions to model decay in the value of data. An adequate discounting model is a function of time and the size of the dataset. Our research also distinguishes between the flow of data and stock of historical data in creating business value. Such distinction is also noted by (³⁰) in their experiment in the context of online news. Our research argues that small firms should focus on the stock of available data and gradually shift their focus to the flow of data as they grow over time.

In this chapter, in the framework section, we explain our approach to the problem and clarify why we made particular choices. Then, in section 3, we introduce the effectiveness curve and explain value depreciation over time. We show the bounded size equivalence in this section. Section 4 inves-

tigates the effectiveness of datasets curated over time. These datasets are a combination of datasets samples from various times. We explain sequential offloading and suggest that old data may even put a firm in a disadvantaged position in businesses with high time dependency. Section 5 empirically measures the value depreciation for the next-word-prediction task. Finally, in the conclusion section, we wrap this chapter with a discussion.

3.1 BACKGROUND AND FRAMEWORK

In this section, we introduce our approach to the problem and explain the particular modeling choices we made. We start by describing the relevance loss of a dataset as a shift in the underlying data-generating distribution over time and dig into its possible cause. We argue that the relevance loss mostly stems from exceptional reasons that often cause a monotonic decrease in the value of data over time. We then provide a brief introduction to machine learning models and explain our focus on the data-generating probability distribution's maximum likelihood estimation. We introduce a decomposition of the MLE's objective function to lay the groundwork for the next section's propositions.

3.1.1 CHANGE IN DISTRIBUTION

Time-dependency of data occurs for many reasons. For example, it can happen if consumers taste or behavior changes over time. If we look at best seller music albums from the 80s and compare them with the best sellers in 2020, we can see the difference in taste. Another example is the continuous innovation in products and services space. Telegram is a perfect example of widely used innovation that is ancient nowadays due to the invention of other communication devices like hardline telephones. Soon, hardline telephones will be ancient too because of the introduction of voice over IP phones (VOIP) or cellphones. Given the rapid development of new technologies in this sector, if we

tried to use consumer behavior data from the 1960s to predict how consumers will use the newest iPhone, such a task would be impossible and absurd. This fundamental shift in consumers behavior caused by innovation makes time-dependency a severe problem.

In the machine learning context, time-dependency is usually referred to as concept drift or non-stationarity. As an example, suppose we use letters written a hundred years ago to train a text auto-completion model for smartphones today. In that case, the users will be disappointed since the way we write and communicate has fundamentally changed over time which means that the model will suggest words or phrases that we no longer use. When this occurs, ML researchers try to use tools known as transfer learning to deal with this non-stationarity and combat time-dependency. These techniques usually describe the change in the data over time as a change in the data-generating probability distribution. Consequently, they either use the dataset's histogram to learn this change over time, or they assume a time-model for the change and accordingly adjust the ML models. In either case, dealing with the data generating distributions has its own problems. To name one, these solutions mostly approximate the change over time and hence, they still incur penalties for not being perfect. But most importantly, they assume that the set of elements Ω in the probability space (Ω, σ, P) is known in advance, which is a fairly big assumption and one that doesn't account for continuous innovation over time. Going back to our text auto-completion example, the data generating distribution changes over time in two ways. First, as time passes there is a lower probability for historical words and phrases to be used. Second, the environment (language in our example) allows for the birth of new elements (new phrases and words), which is equivalent to an increase in their frequency of use. An example of such words is "covfefe" which President Trump used in a tweet and became viral. This word was never recorded in any dataset before the president coined it. This birth and death process of probability space elements, if not accounted for, is among the very reasons we see the depreciation in data value over time. Unfortunately, it is hard to predict and adjust for such changes in data-generating processes and despite the best effort of ML researchers, these

transfer learning models are not perfect. In our research, we assume that the prediction model is fixed, and it may account for transfer learning. Nevertheless, due to imperfection, achieved model accuracy changes depending on the time the training data was collected.

Similar to machine learning literature on transfer learning and non-stationary, we also observe time-dependency as an outcome of change in data-generating probability distributions. To compare distributions at different points in time, we create a universal set of elements. It is the union of all element sets across all time periods. For example, the word iPhone is created in the 2000s. In the language dataset from 1900, this element does not exist, and hence, it is not measurable. We create the universal element set by including this word. Then, for the dataset from the 1900, we should designate zero probability for its appearance. Still, instead of assigning this word a zero probability, we give it an infinitesimal value. This infinitesimal probability helps us use different functional forms like the log function without being worried about issues with functional domains.

Formalizing the assumptions we made so far, we assume that the prediction is for time 0 with a model trained on data from the past. The training data from the past is from t periods prior to time 0 where $t \in \{0\} \cup \mathbb{R}^+$. For the sake of simplicity, we will say this historical data was sampled at time t . The element set at time t is Ω_t for the probability space $(\Omega_t, \sigma_t, \tilde{P}_t)$, where σ_t is the sigma-field over Ω_t and \tilde{P}_t is the probability over the sigma-field. The universal probability space is then (Ω, σ, P_t) , where $\Omega = \bigcup \Omega_t$ and $\delta = \sum_{\omega \in \Omega - \Omega_t} \delta_\omega$ and $\delta_\omega > 0$. As explained earlier, we prefer δ to be zero, but due to some regulatory conditions in the MLE's loss function, we assume δ is infinitesimal.

$$P_t(\omega) = \begin{cases} (1 - \delta)\tilde{P}_t & \omega \in \Omega_t \\ \delta_\omega & \omega \in \Omega - \Omega_t \end{cases} \quad (3.1)$$

With this change in measure, it is possible to compare datasets and define the shift in distribution. A change in distribution between the time i and j means $\exists \omega \in \Omega$ s.t. $P_i(\omega) \neq P_j(\omega)$.

3.1.2 LEARNING DATA DISTRIBUTIONS

Machine learning aims to find the relation between inputs and outputs of an unknown system. The unknown system is usually seen as a black box with little or no information about its function. The goal is then to observe examples of this unknown system's function and adjust a model's parameters accordingly so that the model can replicate the system's function as similarly as possible. These observed examples of systems' function are referred to as data.

Putting this into mathematical semantics, given a dataset $D_{n,t} = \{(x_i, y_i)_t\}_{i=1}^n$, which is composed of n input-output samples $d_i = (x_i, y_i)_t \in \Omega$ collected at time t , we want to find a model $m(d, \theta) \in \mathcal{M}$ that describes the relationship between the input x and the output y . Here, θ represents the model's parameters and \mathcal{M} is the set of all candidate models distinguished by the parameters θ . Linear, logistic, and deep neural network compositional functions are examples of $m(d, \theta)$. Table 3.1 provides the functional forms for these three examples. In most machine learning cases, the goal is to make $m(d, \theta)$ as close as possible to the system's function by learning the parameters θ . The notion of closeness depends on the problem formulation and the objective of the learning task.

Table 3.1: Examples of functional forms for famous ML models.

Case	Functional form $m(d, \theta)$ where $d = (x, y)$
Linear functional	$y = \theta x$
Logistic functional	$y = \frac{e^{\theta x}}{1 + e^{\theta x}}$
Simple Deep Learner with L layers and non-linear functions γ	$y = \theta_L \gamma(\theta_{L-1} \gamma_{L-1}(\dots(\theta_2 \gamma_2(\theta_1 x)) \dots))$

In this research, we restrict our theoretical analysis to the problem of learning the data-generating probability distribution. We do this because identifying this probability distribution is the fundamental problem in statistical learning theory. Once we know the probability distribution that characterizes the data-generation process, we are able to calculate any statistics of interest about the

data, such as its expected value or variance. In general, all statistical models are a function of data distributions. Consequently, under specific regulatory conditions like continuity of models in the probability space, a sequence of distributions converging to the data-generating probability distribution also defines a converging sequence of the model to its converging value. This argument attests that learning the underlying distribution is the fundamental problem in machine learning.

Finally, we choose the maximum likelihood estimator for learning the data-generating probability distribution since it is an unbiased and efficient estimator. Due to its efficiency, it is rational to prefer it over other unbiased estimators. Note that in this research, we are not concerned about time-complexity or other computational issues. Our goal is to get the most from a limited number of data points, and hence, we care about efficiency.

3.1.3 MAXIMUM LIKELIHOOD ESTIMATION AND LEARNING THE PROBABILITY DISTRIBUTION

In the problem of learning a probability distribution, the unknown system is the distribution's functional form. The unknown distribution is defined over the set Ω . The goal is to introduce an estimator $m(d, \theta)$ that converges to $P(\omega)$ for all $\omega \in \Omega$ as dataset size approaches infinity ($n \rightarrow \infty$). The MLE's objective function for estimating the probability distribution, using the model $m(d, \theta)$ and the dataset $D_n = \{d_i\}_{i=1}^n$, has following form

$$\theta_n = \arg \max_{\theta} \sum_{i=1}^n \log(m(d_i, \theta))$$

By dividing the sum by the number of samples and multiplying it by -1, we reach the following equivalent minimization problem. The objective function denotes a loss function called empirical cross-entropy.

$$\theta_n = \arg \min_{\theta} -\frac{1}{n} \sum_{i=1}^n \log(m(d_i, \theta))$$

Convergence to a local minimum happens as the size of the dataset, that is sampled independently and from an identical distribution, grows. For the sake of simplicity and for not dealing with issues of local optimums, we assume our optimization reaches the global optimum and $\lim_{n \rightarrow \infty} \theta_n = \theta^*$ where $m(d, \theta^*) = P(\omega) \forall \omega \in \Omega$. Of course, this is true with the assumption that $P \in \mathcal{M}$ (The solution exists in the search domain). As explained earlier, in this research we assume that algorithms are wisely chosen, and our goal is to see how they perform when the training and testing data are from different distributions. From the Central Limit Theorem, we can see the following approximation for the loss function's value.

Theorem 3.1.1. *Assuming $E(\log m(x, \theta^*))^2 < \infty$, for a sufficiently large number of data points ($n \gg 1$), the loss function can be approximated with*

$$\frac{-1}{n} \sum_{i=1}^n \log(m(x_i, \theta)) = H(P) + D(P||m(x, \theta)) + O\left(\frac{C_1}{\sqrt{n}}\right)\mathcal{N}(0, 1)$$

Where C_1 is a constant, and, is a function of $\text{var}(\log m(d, \theta^*))$. $H(P)$ is the Shannon entropy defined as $H(P) = -\sum_{\omega \in \Omega} P(\omega) \log(P(\omega))$ (⁹⁷), and the summation is over the element set Ω .

$KL(P||m(d, \theta_n)) = \sum_{\omega \in \Omega} P(\omega) \log\left(\frac{P(\omega)}{m(d, \theta_n)}\right)$ is the Kullback-Leibler (KL) divergence (⁷⁴) between the actual distribution $P(\omega)$ and the estimator/model $m(d, \theta)$.

As the size of the dataset approaches infinity, the error term is getting smaller. Immediately from theorem 3.1.1, we can see that KL-divergence is the only component of the loss function that is a function of θ (Model). Hence, minimizing the loss function is equivalent to minimizing $KL(P||m(d, \theta))$. The property of KL-divergence is that it is always non-negative. Besides, it is equal to zero if and only if $P(d) = m(d, \theta)$ almost anywhere. With KL-divergence equal to zero, the only term remaining in the loss function is $H(P)$, which describes the system's entropy. The convergence speed of loss function to $H(P)$ as the function of dataset size is called the learning curve.

3.1.4 LEARNING CURVE

Learning curves represent the expected value of the objective function (loss function in our case) versus the size of the dataset. The expected value is taken with respect to randomness in sampling or algorithm's initialization. In other words, fixing the size, if we sample the dataset and train the model infinite times and then take the average loss value, we reach the expected value of the loss function. Putting this in mathematical semantics, the learning curve is a function $G_t(n) : \mathbb{R}^+ \rightarrow \mathbb{R}$ that takes the size of a dataset as input and outputs the value we should expect for the loss function. For the problem of learning the data-generating probability distribution, this function shows how KL-divergence $KL(P||m(d, \theta))$ changes with the dataset's size.

From theorem 3.1.1, with infinite sample size, we can see the loss function's convergences to the entropy of the data-generating distribution. Since the data-generating distribution changes over time, its entropy changes as well and hence, we added the subscript t to $G_t(n)$ to capture this time-dependency. This function is monotonically decreasing and hence, invertible. Due to its asymptotic convergence to a bounded value $H(P_t)$, it has a convex form for large dataset sizes. We further assume that it is continuous and differentiable, meaning that $(\frac{\partial G_t(n)}{\partial n} < 0)$.

In practice, the learning curve is shown to be predictable for deep learning algorithms (⁵⁷) and is composed of small data, power-law, and irreducible loss regions (Shown in Figure 3.1). In the small data region, the model is not scaling significantly with dataset size. The power-law region is where model performance scales with dataset size. In this region, for deep learners, the function $G_t(n)$ is believed (⁵⁷) to have a power-law functional form. Lastly, in the irreducible loss region, the model's generalization loss value does not improve significantly. Between the regions, the power-law region is the one that we can see significant improvement in performance as we increase the dataset size.

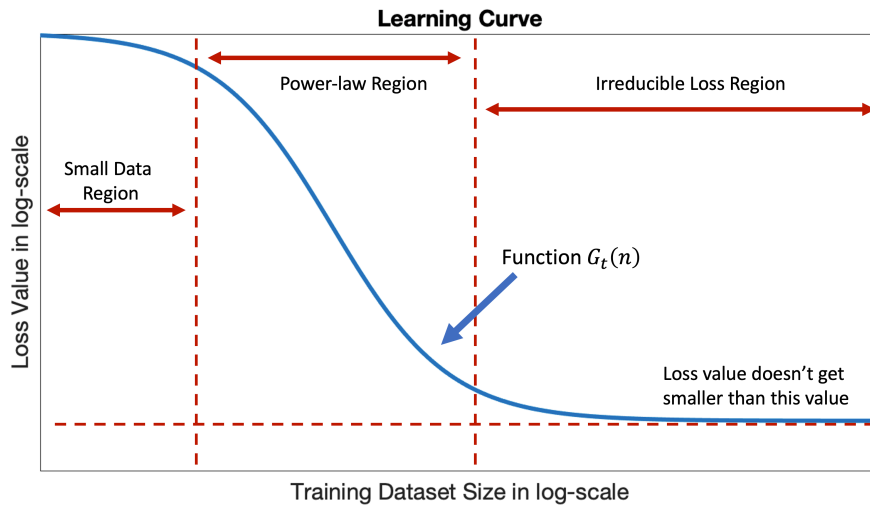


Figure 3.1: Power-law learning curve.

3.2 EFFECTIVENESS CURVE AND VALUE DEPRECIATION

To achieve our goal of measuring the depreciation of datasets' value over time, we need a mechanism to find the value of any given dataset at any given time. Unfortunately, value is subjective and dependent on the context, problem definition, and implementation. Therefore, we define a novel measure of datasets value that reports it as a function of the size. To explain this measure, we first define an oracle dataset that is sampled at time o from P_0 . We use this oracle dataset as a base of comparison and say that the loss value that a model trained on this dataset achieves is our reference loss value and the best loss value we can achieve. We then compare the model's performance trained on the baseline dataset to the model's performance trained on oracle. We finally ask what amount of data from the oracle dataset allows our model to achieve the equivalent loss value that the model achieves after being trained on the baseline. The "equivalent size" of the baseline dataset is the amount of data from the oracle dataset needed to train a model achieving the same loss values.

Before formally defining the notion of equivalent size, a good starting point would be to compare

a baseline dataset of infinite size and see how well-performing a model trained on this dataset is in predicting P_0 (the oracle). This is particularly important since we expect the infinite sized baseline dataset to be as valuable as possible in reaching the ultimate algorithm's performance. In a way, this gives us clues on whether the infinite baseline dataset can scale created value. Proposition 3.2.1 investigates this ultimate scaling behavior.

Proposition 3.2.1. *Assuming $P_t(\omega) \neq P_0(\omega)$, a model trained on a dataset of infinite size from the wrong distribution $P_t(\omega)$ has limited predicting power at time t , and in probability, a model trained on a dataset of bounded size from the right distribution $P_0(\omega)$ can reach the same loss value.*

The argument in proposition 3.2.1 is that in the training phase, due to the change in the probability distribution, $m(d, \theta)$ converges to the wrong distribution $P_t(\omega)$. Therefore, MLE's loss value has an additional term $KL(P_0 || P_t)$ besides the Shannon entropy $H(P_0)$. It is as if we used a dataset of bounded size from $P_0(\omega)$, and due to its limited size, we did not reach the minimum loss value possible. The minimum loss value possible is reached when MLE's loss function is equal to $H(P_0)$.

Intuitively, this proposition tells us that created value in a machine learning-based product or service scales differently with the size of datasets compared with the way tangible assets scale the value. For example, suppose we have a company that produces apple juice boxes. Apples that are stored for a little while longer in inventory will dry out and produce less juice compared to fresh apples. Nevertheless, if we scale the number of older apples in the juice production business, the number of potential juice boxes scales with it. In contrast, when it comes to time-dependent data, increasing the number of older data points doesn't necessarily drive an increase in the model's accuracy and, thereby, the created business value. This is because an oracle dataset of bounded size can produce a loss value equal to that of an infinite amount of older (less relevant) baseline data.

This proposition has implications in economic modeling, academic antitrust debates, and data

management strategy for practitioners. Proposition 3.2.1 states that curating massive datasets over time does not create a significant barrier to entry of a competitor if the data-generating distribution changes. In our interviews with practitioners, we always found them hopeful that increasing dataset size can compensate for the shortcomings in scaling. On the regulatory side, also there are debates on the role of super large datasets and if they create a barrier to entry advantage for big tech companies. This proposition states that scaling of value is different from what we previously knew and hence, the role of time-dependency should be accounted for. This proposition has also implications for economic modeling. The immediate implication is the way we should model the decline in the value of data. Specially, when modelers want to treat data as an asset, they should be aware of the way they account for the value decline. The usual approach to account for the value decline is to use a time-dependent discounting function like an exponential decay $e^{-\alpha t}$ to be multiplied by the accumulated capital at time t . This proposition states that such discounting functions should be a function of accumulated capital as well as the sampling time since there is not a multiplicative discounting function that is multiplied in infinity and results in a finite value. Hence, the discounting function should be a function of both size and the time.

As explained earlier, something lacking from proposition 3.2.1 is that it talks about loss value, which is not very informative in making comparisons. It is not informative because we do not know how to interpret the excess loss value term $KL(P_0 || P_t)$. We just know that it is positive, and therefore, the loss value should be bigger than the one for the oracle dataset. To solve this issue, we use the learning curve inverse function to translate the loss function back into an “equivalent” dataset size. Dataset sizes are easy to understand and compare.

Recall that learning curve at time zero $G_0(n)$ is a monotone function and therefore has an inverse. Using the inverse of the learning curve $G_0^{-1}(\cdot)$, we can find the expected size of a dataset from time zero (the oracle dataset) with an equivalent MLE loss value. Briefly, what we do to form the equivalent size is to first train a model on data sampled from $P_t(\omega)$. Then, we use the trained model

to find the loss value on the data that has been sampled from $P_0(\omega)$. Finally, we use the function $G_0^{-1}(\cdot)$ to see what size of the data from $P_0(\omega)$ can generate a similar loss values. This is the basis for our definition of equivalent size.

Definition 3.2.1. *Dataset $D_{n,t}$ has the equivalent size $\bar{n}_{D_{n,t}}$ at time o :*

$$\bar{n}_{D_{n,t}} = E_{\theta_{n,t}} \left(G_0^{-1} \left(-E_{P_0} \left(\log m(d, \theta_{n,t}) \right) \right) \right)$$

Where $\theta_{n,t}$ is the solution to:

$$\theta_{n,t} = \operatorname{argmin}_{\theta} - \frac{1}{|D_{n,t}|} \sum_{d \in D_{n,t}} \log(m(d, \theta))$$

In this definition, there exist two expectations. The expectation inside of $G_0^{-1}(\cdot)$ measures the expected model's loss over the test set. The second one is the outer expectation and calculates the expectation with respect to randomness in the algorithm's initialization and steps. In practice, we can approximate the outer expectation by solving for $\theta_{n,t}$ multiple times. Using averaging limit, we can calculate the equivalence empirically in the following way

$$\lim_{k \rightarrow \infty} \frac{1}{k} \sum_{j=1}^k \left(G_0^{-1} \left(\lim_{l \rightarrow \infty} -\frac{1}{l} \sum_{i=1}^l \log \left(m \left(d_i, \theta_{n,t}^{(j)} \right) \right) \right) \right)$$

Where $d_i \sim P_0(\omega)$ and the outer sum is over multiple runs of the algorithm. For a fairly large number of testing data points, the inner expectation converges. Using theorem 3.1.1 to simplify the definition further, we have

$$\bar{n}_{D_{n,t}} = E \left(G_0^{-1} \left(H(P_0) + KL(P_0 || m(d, \theta_{n,t})) \right) \right)$$

Letting $n \rightarrow \infty$ eliminates algorithms' initialization issues as well as other types of randomness and

hence, $m(d, \theta_{n,t}) \rightarrow P_t(\omega)$. Therefore, in the limit

$$\bar{n}_{D_{\infty,t}} = G_0^{-1}(H(P_0) + KL(P_0||P_t))$$

It is in agreement with proposition 3.2.1 where it argues that $\bar{n}_{D_{\infty,t}} < \infty$ if $P_0(\omega) \neq P_t(\omega)$.

Because $G_0^{-1}(H(P_0) + KL(P_0||P_t)) < G_0^{-1}(H(P_0)) = \infty$.

Notice that the equivalent size is a function of the algorithm as well as the dataset itself. Dependence on the algorithm is recognized through the inverse function $G_0^{-1}(\cdot)$. It means that the algorithm's power in scaling with dataset size shapes the effectiveness of a dataset. The following example makes it clear. Suppose we have a very large dataset, but we do not use it to train a model. In that case, the sampling time is not essential and, regardless of time, the dataset is as effective as not having it in the first place ($n = 0$). On the other hand, if the algorithm scales in a faster pace (in the number of data points), a small dataset from $P_0(\omega)$ can reach $H(P_0) + KL(P_0||P_t)$ with a smaller number of data points which means $\bar{n}_{D_{\infty,t}}$ is indeed small.

Definition 3.2.2. *Effectiveness of dataset $D_{n,t}$ is defined as $E_{D_{n,t}} = \frac{\bar{n}_{D_{n,t}}}{n}$.*

Intuitively the effectiveness should be always between zero and one, i.e., $E_{D_{n,t}} \in [0, 1]$. $E_{D_{n,t}} = 1$ means that the given dataset's value is equal to the value of the oracle dataset. Note that $E_{D_{n,t}}$, by definition, can't be more than 1. $E_{D_{n,t}} = 0$ means that data is worthless compared to the oracle dataset and the prediction power of a model trained on this dataset is equivalent to a uniformly random guessing of output values. The value of $E_{D_{n,t}} \in [0, 1)$ signals that the equivalent size is less than the actual size of a dataset. It is as if dataset perishes over time. The more perishable the data, the less its effectiveness over time. For example, if the effectiveness is equal to 0.8, we say that the dataset lost 20% of its effective size.

Proposition 3.2.1 argues that effectiveness $E_{D_{\infty,t}} = 0$ if $P_0(\omega) \neq P_t(\omega)$. It is because $\bar{n}_{D_{n,t}}$ remains bounded and therefore, $\lim_{n \rightarrow \infty} \frac{\bar{n}_{D_{n,t}}}{n} = 0$.

Definition 3.2.3. *Substitution curve is a function $f_n(t_1, t_2) : \mathbb{R}^2 \rightarrow \mathbb{R}$ and is defined as*

$$f_n(t_1, t_2) = \frac{\bar{n}_{D_n, t_1}}{\bar{n}_{D_n, t_2}}$$

It shows how well we will be off in terms of effectiveness if we substitute a dataset of size n from time t_2 with a dataset of the same size that has been sampled at time t_1 . Note that choosing $t_2 = 0$ brings us back to the definition of effectiveness. Using theorem 3.1.1, the substitution curve has following formulation

$$f_n(t_1, t_2) = \frac{\bar{n}_{D_n, t_1}}{\bar{n}_{D_n, t_2}} = \frac{E(G_0^{-1}(H(P_0) + KL(P_0 || m(d, \theta_{n, t_1}))))}{E(G_0^{-1}(H(P_0) + KL(P_0 || m(d, \theta_{n, t_2}))))}$$

Theorem 3.2.1. *Substitution curve has the following properties.*

- a) *It is non-negative and bounded.*
- b) *It is a monotonic function of n .*
- c) *It is converging to a substitution frontier*

$$\lim_{n \rightarrow \infty} f_n(t_1, t_2) = \frac{\bar{n}_{D_\infty, t_1}}{\bar{n}_{D_\infty, t_2}} = \frac{G_0^{-1}(H(P_0) + KL(P_0 || P_{t_1}))}{G_0^{-1}(H(P_0) + KL(P_0 || P_{t_2}))}$$

Nonnegativity and boundedness are immediate. It is nonnegative because function G_0^{-1} is non-negative by definition. Boundedness is also immediate from proposition 3.2.1, because for $i \in \{1, 2\}$ and $t_i \neq 0$, $0 < \bar{n}_{D_n, t_i} < \infty$ for all n .

The substitution curve is an important definition in this research. It is a building block for the argument we make in the next section on the effectiveness of datasets gathered over a long period of time. The concept will be used in the sequential offloading algorithm defined in the next session.

Figure 3.2 depicts examples of substitution curves $f_n(t, 1)$ assuming a monotonic decline in the value of data over time. Each curve represents the substitution gain over time t when the substi-

tution time is fixed at $(t_1, t_2) = (t, 1)$. Blue curve is the frontier trajectory $f_\infty(t, 1)$ described in Theorem 3.2.1c. This figure pictorially shows the substitution function's monotonicity on n and its convergence to the frontier. As apparent in this Figure, we do not gain much in substituting data from different times for very small dataset sizes. It is because small datasets do not provide significant scaling in performance, and hence, it does not matter when they were sampled. This behavior is mostly seen in the small data region of the learning curve. For medium dataset sizes, when we are in the learning curve's power-law region, we gradually see significant gains in substituting datasets from different times. Increasing dataset size in the power-law region brings us to the medium-high dataset size regime. This region will be used in our experiments (In later sections) to measure perishability. Finally, the infinite dataset size speaks of the irreducible loss region and the highest sensitivity to substitution.

In Appendix B.2, building further on our observation, we empirically measure the substitution curve for our experiment in this research and show that gain/loss increases in n for $t_1 > t_2$ and decreases for $t_1 < t_2$. In other words, the substitution curves become sharper and the gains/loss in substituting data from various time increase. Proposition 3.2.2 formalizes this idea. To prove this proposition, we have two additional assumption. First, we assume that the monotonicity result proved in theorem 3.2.1 (b) is valid for all dataset sizes meaning that $f_n(t_1, t_2)$ is monotonic for all n . Second, we assume that $f_1(t_1, t_2) = 1$ for all $t_1, t_2 \in \mathbb{R}^+ \cup 0$. It is intuitive since, in our model, all elements have non-zero probability and hence, one data point carries in expectation same amount of information regardless of when it was sampled.

Proposition 3.2.2. *The substitution gain function becomes sharper with the increase in the size of the baseline dataset (n), meaning that the gain/loss increases as the baseline dataset size increases. Mathe-*

matically, for all time $t_1, t_2 \in \mathbb{R}^+ \cup \{0\}$ and sizes $n_1, n_2 \in \mathbb{N}$ with $n_1 < n_2$:

$$\begin{cases} f_{n_1}(t_1, t_2) \leq f_{n_2}(t_1, t_2) & \text{when } f_{n_1}(t_1, t_2) \geq 1 \\ f_{n_1}(t_1, t_2) > f_{n_2}(t_1, t_2) & \text{when } f_{n_1}(t_1, t_2) < 1 \end{cases}$$

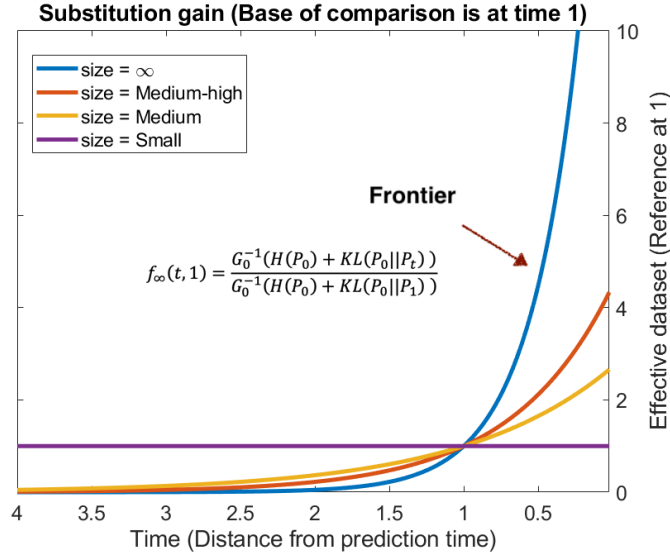


Figure 3.2: Substitution curves for different sizes of datasets. The frontier is marked as blue. It shows the maximum depreciation in substituting datasets of time 1 with a dataset of any other time.

This proposition is particularly crucial for our discussion in section 4. We use it to prove that flow of data becomes the main value driver when a firm grows, for example, in the user base. Hence, firms should retrain their models more frequently when the amount of data they collect in a given time (n in this case) grows.

3.3 DATASETS COLLECTED OVER TIME

So far, we have studied the effectiveness of datasets that have been sampled at a given time t . Nevertheless, most datasets are collected over time, and there is a need to study their effectiveness and

compare their values. This is particularly important since we want to study the historical value of data. In that pursuit, we compare two datasets that have been sampled over time; A baseline dataset that spans over a longer period and a subset of that dataset which has a smaller size and contains only recent data (most relevant data in case of semi-monotonic decline in value of data over time). When the value of smaller dataset is more than the value of larger dataset, we conclude that using historical data in the training set increases the loss value and may put a company in a disadvantageous position. In section 4.1, we use this idea and introduce the sequential offloading algorithm. This algorithm exploits the tradeoff between the size and the relevance of datasets gathered over time. It removes less relevant data from the training dataset, meaning we lose size, and hopes that increasing the relevance counterbalances the loss in size.

We use the machinery we developed so far, with slight modification, to compare the values of datasets sampled over time. In doing so, we first show that the accuracy score (loss value) of models trained on such datasets are equivalent to the accuracy score (loss value) of models trained on a dataset of similar size that has been sampled at a time t^* , i.e., there exist a time t^* such that a model trained on a dataset from this time has an equivalent loss value to the same model that is trained on the dataset gathered over time. t^* is called the equivalent time. Subsequently, we calculate the equivalent time for these datasets (that have been sampled over time) and use the substitution curve to compare their values.

Notation wise, we show datasets of size n that are collected over a period $[t_1, t_2]$ with $D_{n,[t_1,t_2],\lambda_t}$. λ_t is called the sampling density function and shows the proportion of samples that have been collected at time t . For the sake of simplicity, we assume $t_1 = 0$ and focus on datasets of the form $D_{n,[0,t],\lambda_t}$. Mainly because it is easier to turn a bigger period of time into smaller periods with the sampling function λ_t . For example, the dataset $D_{n,[t_1,t_2],\hat{\lambda}_t}$ is equivalent to dataset $D_{n,[0,t_2],\lambda_t}$ with λ_t

equal to

$$\lambda_t = \begin{cases} 0 & t < t_1 \\ \hat{\lambda}_t & t_1 \leq t \leq t_2 \end{cases}$$

Although the dataset $D_{n,[t_1,t_2],\lambda_t}$ is sampled from various time with different generating probability distribution functions, it still has a “Net Distribution” that can be used to measure its effectiveness. Lemma 3.3.1 presents the net distribution for $D_{n,[0,t],\lambda_t}$. It states that the net distribution is a convex combination of all distribution from time 0 to t with weights $\lambda_t \in [0, 1]$ & $\int_0^t \lambda_s ds = 1$.

Lemma 3.3.1. *Net distribution of dataset $D_{n,[0,t],\lambda_t}$ is equal to*

$$P_{[0,t],\lambda_t}(\omega) = \int_0^t P_s(\omega) \lambda_s ds$$

As this lemma states, the net distribution is not necessarily equal to P_0 . Therefore, using proposition 3.2.1, we can still argue that datasets curated over a period of time have limited relevance.

Corollary 3.3.1. *Datasets that are collected over time from dynamically changing environments have finite effectiveness and value. Hence, the growth between the accuracy of machine learning models and the stock of available data (Known as the data network effect) stalls.*

Data network effect (The growth cycle between the stock of data and the accuracy of models) stalls when we have time-dependency. As stated in Corollary 3.3.1, the accuracy of machine learning models and the value they create doesn't simply scale with the stock of available data. As a result, a firm should either incorporate the time dimension into their models or alternate datasets to improve its relevance. Incorporating time in the models is not easy and often impossible, mainly when dealing with innovation over time. For example, we can't tell what kind of medical innovations we should expect in the next couple of years or what news we should read in the papers next week. Besides, we already assumed that firms already choose best models, and hence, we expected them to

incorporate time if possible. Therefore, as the next step in this chapter, we focus on methods to alternate the dataset's composition and improve their relevance. Alternating datasets should be in the direction of improving their relevance, and hence we first need a method to compare the value of dataset pre and post alteration. Proposition 3.3.1, using the net distribution of datasets, maps the accuracy score of the dataset gathered over time to the accuracy score of a dataset that has been sampled in a given time. Hence, we can compare two datasets using the substitution curve method we developed in section 3.

Proposition 3.3.1. *There exists an equivalent time $t^* \in [0, t]$ such that the dataset $D_{n,[0,t],\lambda_t}$ provides an equivalent loss value to the dataset D_{n,t^*} i.e. $\bar{n}_{D_{n,[0,t],\lambda_t}} = \bar{n}_{D_{n,t^*}}$. The solution is unique when decline in value of data is monotonic.*

Proposition 3.3.1 is the key to understanding the next subsection on sequential offloading. As much as it is important to understand what it says, it is also important to realize what it does not. It does not say that the “Net distribution” is equal to P_{t^*} . Net distribution is a combination of many distributions, including P_{t^*} , and therefore, it is not necessarily equal to P_{t^*} . Instead, Proposition 3.3.1 suggests that $P_{[0,t],\lambda_t}$ and P_{t^*} are in a way that they make equal KL divergences with P_0 , i.e. $KL(P_0||P_{[0,t],\lambda_t}) = KL(P_0||P_{t^*})$. Consequently, they produce equivalent MLE loss value, which means $\bar{n}_{D_{n,[0,t],\lambda_t}} = \bar{n}_{D_{n,t^*}}$.

Note that having t^* between zero and t is important in this proposition. The emphasis is on the fact that the period $[0, t]$ starts from time 0. Even if the dataset has been sampled from $[t_1, t]$, still, $t^* \in [0, t]$. It is because, for the dataset $(D_{n,[t_1,t],\lambda_t})$ where $0 < t_1 < t$, there might exist a sampling density λ_t such that it makes the Net distribution $P_{[0,t],\lambda_t}(\omega) = P_0(\omega)$ for all $\omega \in \Omega$, i.e. $t^* = 0 \notin [t_1, t]$.

The most exciting thing about this theorem is that $t^* < t$. If we deliberately delete the portion $[t_1, t]$ from the dataset where $t_1 < t^*$, despite losing size, the remaining dataset $(D_{n_1,[0,t_1],\lambda_t})$ will

have a new equivalent time t^{**} which is $t^{**} < t^*$. In other words, datasets gain relevance. Altering a dataset composition by deleting the portion $[t_1, t]$ is what we investigate next as sequential offloading algorithm.

3.3.1 SEQUENTIAL OFFLOADING

The idea of sequential offloading is founded in increasing the value of a dataset by reducing its size. It looks to be counter-intuitive, but in a time-dependent context, data perish quickly, and it may be beneficial to discard useless information. Clearly, deleting old data means loss of dataset size. Nevertheless, gaining relevance may offset the loss of dataset size, and deletion likely improves the overall effectiveness.

The idea is centered around Proposition 3.3.1 and theorem 3.2.1. Proposition 3.3.1 states that for a dataset $D_{n,[0,t],\lambda_t}$ there exist a time $t^* \in [0, t]$ such that $\bar{n}_{D_{n,[0,t],\lambda_t}} = \bar{n}_{D_{n,t^*}}$. By deleting data $[t^*, t]$ from the dataset, we end up with a smaller size n_0 , but the equivalent time shifts from t^* to $t^{**} \in [0, t^*]$, which is more relevant. If the substitution gain is higher than the lost size due to deletion, it means we gained from deletion i.e.

$$f_{n-n_0}(t^{**}, t^*) > \frac{n}{n-n_0} \Rightarrow \bar{n}_{D_{n,[0,t],\lambda_t}} < \bar{n}_{D_{n-n_0,[0,t^*],\lambda_t}}$$

Where n_0 is the size that has been deleted from the dataset. Algorithm 1 Formalizes the sequential offloading. This algorithm stops when there is no gain in deleting old data. It also opens a philosophical question on what a successful iteration means for the data. A successful iteration means $\bar{n}_{D_{n,[0,t]}} < \bar{n}_{D_{n-n_0,[0,t^*]}}$ and hence, there is positive improvement upon losing a portion of data. Therefore, as the following corollary states, including less relevant data (older data in our case) actually did put us in a disadvantageous position.

Corollary 3.3.2. *Including older (less relevant) data in the training set may put a firm in a disad-*

vantageous competitive position.

Algorithm 1 Sequential offloading algorithm

Require: Dataset $D_{n,[0,t],\lambda_t}$, substitution gain function $f_n(t_1, t_2)$

$i \leftarrow 1$

$t^{(0)} \leftarrow t$

$n^{(0)} \leftarrow n$

while (Gain is possible) **do**

 Find t^* as explained in theorem 3.2.1 and call $t^{(i)}$

$n^{(i)} \leftarrow n^{(i-1)} \times \int_{t^{(i)}}^{t^{(i-1)}} \lambda_t dt$

 Delete sampled data $[t^{(i)}, t^{(i-1)}]$ from $D_{n^{(i-1)},[0,t^{(i-1)}],\lambda_t}$ and call it $D_{n^{(i)},[0,t^{(i)}],\lambda_t}$

if $\left(f_{n^{(i)}-n^{(i-1)}}(t^{(i)}, t^{(i-1)}) > \frac{n^{(i-1)}}{n^{(i)}-n^{(i-1)}} \right)$ **then**

 Gain is possible

$i \leftarrow i + 1$

else

 Gain is not possible

end if

end while

3.3.2 DATA AS A DRIVER OF GROWTH

A successful offloading iteration means that older (less relevant) data is lowering the modeling accuracy score. Hence, it is optimal for a firm to retrain its models more frequently with relevant datasets. The retraining frequency is determined by the sharpness of the substitution function.

From Proposition 3.2.2, we know that increasing the flow of data which is the rate of accumulating new information increases the gain/loss value of the substitution function and hence, changes the retraining frequency. Since the increase in flow is inevitable due to growth in the user-base or engagement, it is of paramount importance to understand the optimal growth strategy for a firm that derives value from data. In Proposition 3.3.2, we show that a sharper substitution gain function (which is the result of increase in flow) brings the equivalent time closer to the prediction time.

Proposition 3.3.2. *Increases in the gain/loss value of the substitution function, as a result of multiplying the data flow rate, brings the equivalent time closer to the prediction time.*

We now cite Proposition 3.2.2 and argue that an multiplying the flow of data, with a constant $\alpha > 1$, makes the offloading more likely, and hence, as stated in Proposition 3.3.2, it brings the equivalent time closer to the prediction time. Bringing the equivalent time closer to prediction time means that we rely on the flow of data to create value. Consequently, we conclude that when a firm grows, which leads to inevitable growth in the flow of data, it should shift its focus from the stock of available data to the flow of data as the primary value driver.

3.3.3 WEIGHT-ADJUSTED DATASETS

In previous subsections, we compared the values of two datasets, a baseline dataset that is collected over a period of time and a subset of this dataset that only contains recent data. In this comparison, we argued that there are cases where the subset of the baseline dataset has a higher value for a business application since the model trained on this subset has a higher accuracy score. Consequently, we argued that including older (less relevant) data in the training set might put a firm in a disadvantageous competitive position. We made this comparison to contest a widely known idea that having more data is always better, and as a result, companies who were collecting data earlier than others are in a better competitive position solely because of the data they own.

Still, despite feasibility and ease of implementation, deleting older data from the dataset is a sub-optimal action from the implementation perspective. A wiser choice is to use a weight-adjusted version of the baseline dataset. The sequential offloading algorithm is also weight-adjusted since it puts zero weight on older data and full weight on recently collected data. However, this zero-one weighting is not necessarily optimal despite being more advantageous than the baseline dataset.

The main challenge is to find the optimal weights. It is challenging because of two reasons. First,

as proposition 3.3.2 suggests, the value of data sampled from various times changes with the size of the dataset or the flow of data. Therefore, companies should continuously reevaluate and adjust weights over time since their dataset size or user-base changes. The second challenge is that companies should use the whole dataset to evaluate the weights, which is time-consuming and expensive to implement. Because, as proposition 3.2.2 suggests, the importance or value of recent data increases with the size of the subset, meaning that for larger subsets, the optimal weights of recent data will be more significant. Therefore, using a subset of the baseline dataset to evaluate the weights (Which is a standard solution to calibrate parameters in practice) miscalculates the weights and undervalues the importance of recent data. Because of these two points, i.e., continuously evaluating the optimal weights using the entire dataset a firm owns, we argue that it might be tractable to discard older data than repeating complicated repetition of the learning process.

There is a question on the transferability of wights between firms. In other words, since the cause of perishability mentioned in this research are at the market level, it makes sense that one firm calculates and sells the optimal weights to other firms. In that case, it is essential to know that weights depend on the company size, and the optimal weight that is useful for a big firm is not optimal for a smaller firm.

3.4 EXPERIMENTAL DESIGN

Our goal in this section is to measure effectiveness and thereby perishability of datasets empirically. In other words, after training an algorithm with data that has been sampled on one stationary period, we measure its equivalent oracle size at other periods. As it is shown later in this section, we observe a semi-monotonic decline in the value of a dataset. We expected the decline to be monotonic since datasets from the past lose relevance monotonically due to continuous innovation over time. However, the equivalent size has slight periodicity in the measurements due to seasonality in

certain topics like fashion or other periodic data generating sources. Note that the overall decline still looks monotonic.

We make the measurements in the language modeling contexts. Mainly because language modeling datasets tend to be the largest and most easily collected. Datasets are easily collected since language modeling tasks are most often unsupervised; For example, in the next-word-prediction task, the model predicts the next word in a given sentence, and hence, any book, magazine or text source could be a potential data source. Furthermore, the language modeling is currently used as a common pre-training objective for many other language tasks (¹¹²) making our measurements relevant to a wide variety of applications. In this section, we first explain data and how we process it for the task. Then, we explain the algorithm and model architecture, and lastly, we present the measurements.

3.4.1 DATA COLLECTION AND PROCESSING

Our challenge is to find a large enough dataset that has been collected over a long period of time. It is because text processing algorithms require large training set sizes to have significant improvement in quality. In addition, we need this dataset to be sampled over a long period of time to let us make an observable perishability measurement. From a technical standpoint, the dataset must be large enough to reliably measure the power-law portion of the learning curves associated with each time period. Thus, the dataset must span roughly two orders of magnitude in size larger than the smallest dataset in the power-law region. Prior results (⁵⁷) show that, for language modeling, the smallest such dataset is at most 1 million words. Consequently, the dataset should contain roughly 10-100 million words per time period.

We choose the Reddit post dataset as it fits to our needs. This data was collected and used in (⁴⁰). It is a collection of posts and comments from the years 2006 to 2018 and was scraped from Reddit between September 2016 and July 2018. We preprocessed the dataset to create flat text files with the

following format:

Title (6): What was the biggest scandal in your school?

Text:

Comment (4): Vampires. This was almost 6 years ago now at my high school, but vampires. Do a quick...

Comment (3): Not sure if I'd call it a "scandal", but when I was in college...

Comment (2): Freshman year a friend of mine found a paper bag at the bus stop full of money - and it...

'Title' is the title that the author specified when posting the submission, and 'Text' is an optional field of body text associated with the post. After the post, each line is a comment from other users designated by 'Comment'. Comments only contain text. The values in parenthesis are submission or comment scores based on upvotes or downvotes given to each by users. We filtered out comments and posts with scores less than 2.

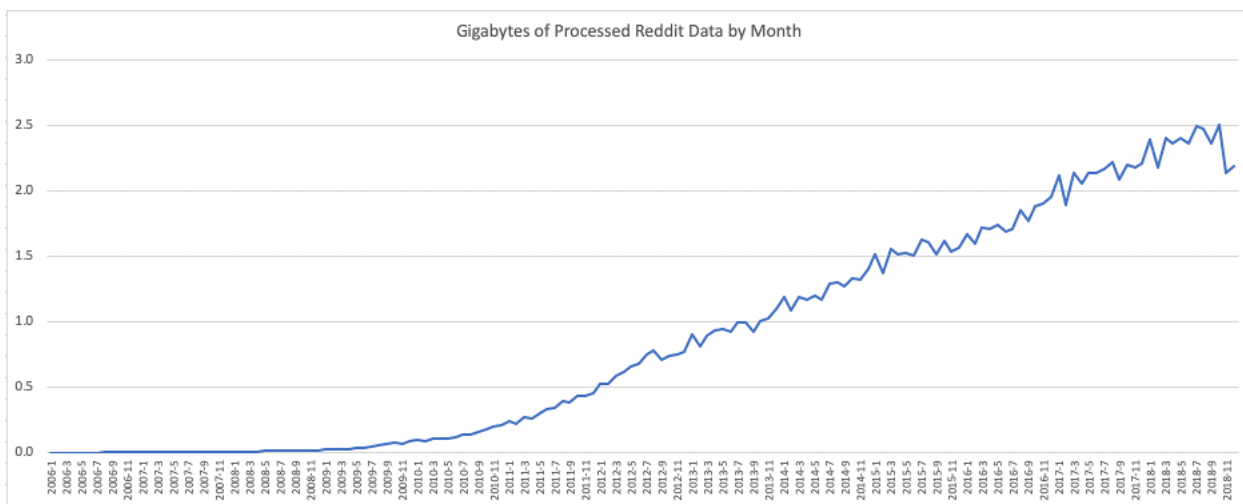


Figure 3.3: Size of datasets processed for each month. For example, for July 2013, 1 Gigabyte of text data is processed. This is not a cumulative dataset size. The growth in the size shows the growth in the number of topics discussed, the number of users as well as their engagement.

To evaluate how data distributions and value shifts over time, we split the dataset into chunks based on the time stamp of the submissions and comments. We aim for 100 million words per time period, so we group data until each split is at least that large. Specially, we group posts and comments into the following periods: the years 2006-2009, January-June 2010, July-December 2010, January-March 2011, April-June 2011, July-September 2011, October-December 2011, and then monthly for the years 2012-2018. Earlier years of Reddit dataset have less data because the platform was becoming established and growing, so we had to group more extended periods together. Figure 3.3 shows the amount of data we processed each month.

Finally, we subdivide the data from each time period to form a standard machine learning training and testing setup for collecting learning curves. First, we randomly sample and split the posts (and their comments) into training, development/validation, and test/evaluation subsets. The development and test sets are at least 2 million words each. The development set is used to validate that the model is learning to generalize during training and to early-stop training when the model performs the best on the development set. The test set is used after training to evaluate how well the training is done. We use these test sets to cross-evaluate models trained on data from other periods. The model never trains on these subsets.

After splitting out the development and test sets, we randomly shuffled the remaining data as the full training set for the time period. We subdivide this training set into chunks of exponentially increasing size by factors of 2. Empirically, we find that datasets of size 1.25 million words are large enough to be in the power-law portion of learning curves, so we break the training set into successively overlapping subsets of size 40 million, 20 million, 10 million, 5 million, 2.5 million, 1.25 million words by taking the first half of the prior subset. We train separate models on each training subset to collect how models generalize as they are allowed to train with increasing dataset size. The resulting data size-generalizability curves are learning curves for the time period.

3.4.2 MODEL ARCHITECTURE AND TRAINING PROCESS

We chose to train current state-of-the-art language models on the data to collect their generalization error and learning curves. Specifically, we train GPT-2, the Generative Pre-Training transformer-based model from OpenAI (^{51, 92}). Collecting learning curves can be costly due to the training time required to train large models on each of the training subsets. We chose to train a small variant of GPT-2 that was expected to be large enough (i.e., sufficient parameters) to overfit all of the training set splits and yet small enough to train in a reasonable amount of time – at most about 32 hours per training subset on a single GPU. We configure our GPT-2 model variant as follows: Vocabulary size 50257 sub-word tokens, maximum sequence length 512 tokens, depth 6 transformer blocks each with 8 self-attention heads and hidden dimension 512. The model has 44.9 million parameters total – a rule of thumb in language modeling is to use a model with as many parameters as words in the largest dataset.

We train the models using the Adam optimizer with a static learning rate of $2e-4$ and with batch sizes 12 and 24. The training objective is the cross-entropy loss of the model’s prediction of the probability of the target next token in the input sentence. We empirically find that changing the batch size marginally changes the final loss ($< 0.3\%$ change in cross-entropy), so we do not further explore optimization hyperparameters to mitigate total training time. Finally, we validate the models using the development dataset every 50-200 training steps, depending on the size of the dataset—smaller datasets require fewer training steps for the model to converge. We early-stop training when the development set loss stops improving for more than 15 validation runs.

3.4.3 EVALUATION PROCESS AND EFFECTIVE DATASET SIZE

Our objective is to measure how much the data distribution has changed over time. In that cause, we evaluate how well a dataset that has been sampled from one time period, can predict values for

each other time period’s data. To do so, we train a model and evaluate its test error for each time period over multiple time periods. Furthermore, we characterize the learning curves so that we can translate measured test errors back to equivalent dataset sizes. Finally, we present the effectiveness curve.

In the training phase, we first find the finest model for each time period and each dataset size. The finest model is the one that achieves a smaller development set loss. Its selection process mimics the way models are chosen for deployment in AI-enabled products. To find the finest model, at each training run, we validate the models on the given time period’s development set and choose the model weights that achieve smaller development set loss. When we test with multiple different batch sizes, the finest model is the one that achieves superior performance in separate training runs for the given time period and training set size.

We collect the finest model for each training set size ranging from 1.25 to 40 million words. Doing so allows us to construct learning curves across different time periods. We cross-evaluate all finest model – one for each time period and training set size – by evaluating them on the test sets for all other time periods. We use these results to curve fit learning curves and indirectly calculate its inverse: Given finest models for the time period t_1 , and their evaluation scores for the time period, t_0 (t_0 can be equal to t_1), these scores will be used to show how increasing the training set size from period t_1 might improve prediction accuracy for the time period t_0 . We curve fit learning curves with power-laws.

Figure 3.4 shows examples of learning curves for models trained at different times. Each curve shows a model that has been trained on a specific time-period. The learning curves are different from each other and form parallel curves. The offset is due to change in the entropy $H(P)$, which is different at different times. Earlier models like those that have been trained in 2010 have lower values than the model of 2018. To answer why this is happening, we should look at Figure 3.3. As apparent from Figure 3.3, the dataset size per month is growing, which is a clear sign of the in-

crease in the contribution and growth of the user base. This growth adds to the diversity in topics as well as language styles. The more diverse the dataset, the higher its entropy. It is also apparent from this graph that the learning curve is a decreasing function, and hence, more data causes lower cross-entropy value.

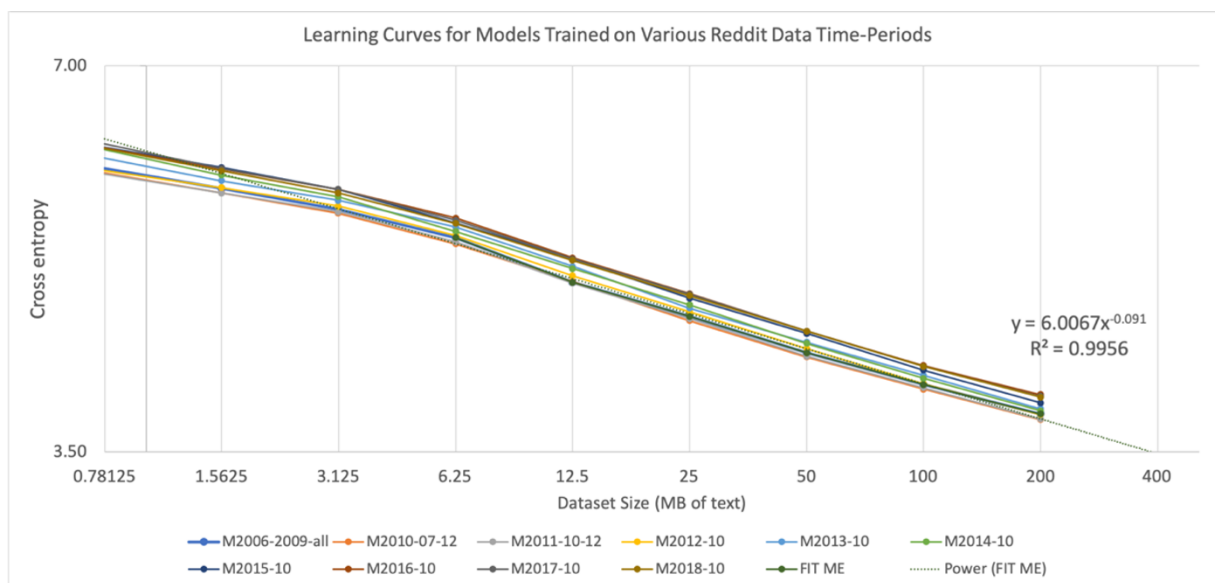


Figure 3.4: Measured learning curves for models that have been trained at different times. The x-axis is in the log-scale and shows the dataset size. Y-axis is the cross-entropy value. The legend describes the time we used to train these models. For example, the yellow curve shows a model that has been trained on data from October 2012.

Figure 3.5 shows test evolution results for models trained on different time periods. Training size is fixed, and the algorithm is trained on data from a few time-periods. Periods are shown in the legend section of this Figure. Each point in this graph is the evaluation result of a training and test pair and curves are made by joining pairs with similar training time. For example, the blue curve shows the finest model’s test results that have been trained 2006-2009 and tested on every other time.

The first observation is that the best model for prediction in t_0 is the one trained on data from t_0 . As an example, before January 2010, the model that has been trained on data from 2006-2009

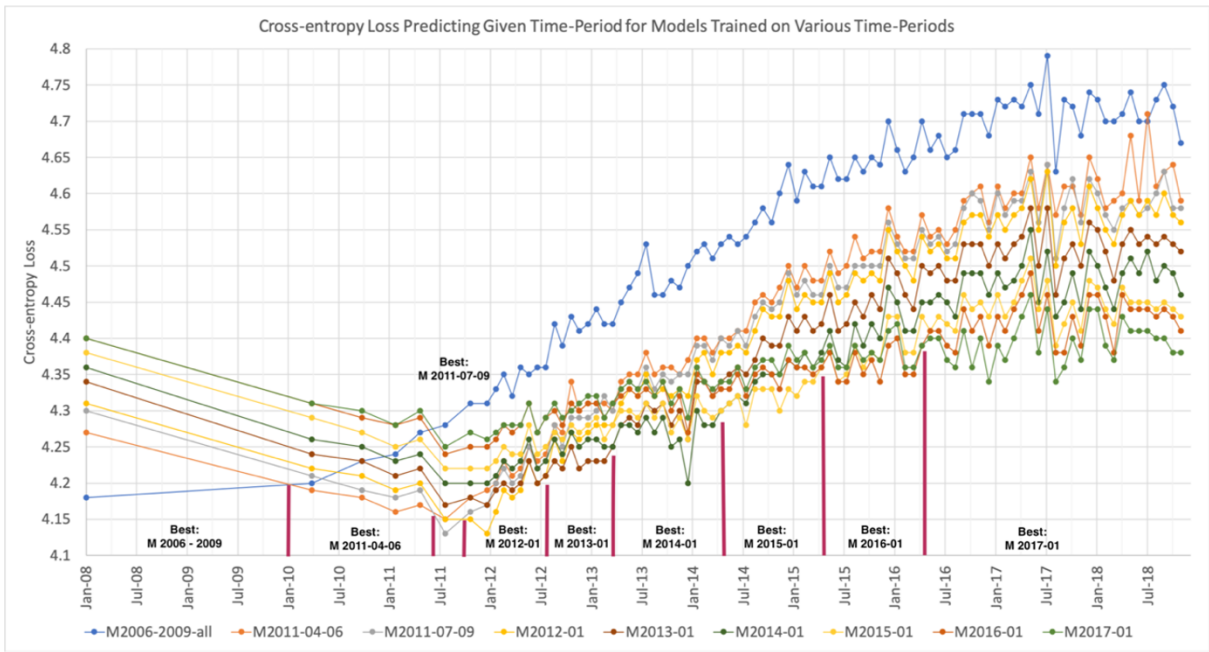


Figure 3.5: Cross entropy loss value when we use a model that has been trained on year z (each curve) and is tested on data from year x (x-axis). Y-axis is the cross-entropy loss. The legend describes the time we used to train these models. For example, the green curve shows a model that has been trained on data from January 2014. The best cross-entropy loss in each time period is mentioned in this graph as well.

($m_{2006-2009}$) has the lowest cross-entropy and hence, has the best predicting power compared to other curves. In contrast, from January 2010 to June 2011, the April 2010's model ($m_{2010-04-06}$) is the best performer replacing the blue curve. It immediately shows perishability. It is because the best performing model at one period loses its power as we move away from its sampling time. Despite apparent perishability, as time goes by, we see an increase in the cross-entropy values across all models. It is again due to the increase in the diversity of topics in Reddit data over time. In other words, the entropy function $H(P)$ is increasing.

Finally, we invert these learning curves to estimate the equivalent dataset size from time period t_1 when predicting for the time period t_0 . Start with the best model, $m_{t_1, 50M}$, for time period t_1 trained on 50 million words, for example. Evaluate $m_{t_1, 50M}$ to collect cross-entropy loss for time period t_0 . Now use the learning curve for models trained and tested on time period t_0 to estimate how much training data from time period t_0 is required to achieve that cross-entropy loss. Suppose the inverted learning curve yields 40 million words required in time period t_0 , then the equivalent dataset size from time period t_1 is 40 million words at time t_0 , or it is effectively 80% of its time t_1 size.

Figure 3.6 shows the equivalent dataset sizes for models trained on the 100MB of data sampled from different times. We chose 100MB for this graph to make it easier for readers to convert values to percentages. As seen in this Figure, for periods after sampling time, the equivalent sizes are monotonically decreasing. Despite overall monotonicity, we need to answer two questions about this graph:

1. Why do we observe higher equivalence variability on curves with higher equivalence (Closer to 100MB) sizes?
2. Why do we, on some occasions, observe a sudden increase in all equivalence curves?

For the first question, we believe it happens due to numerical errors in the inversion of learning curves. As we see in Figure 3.4, learning curves have power-law functional forms. Hence, in dif-

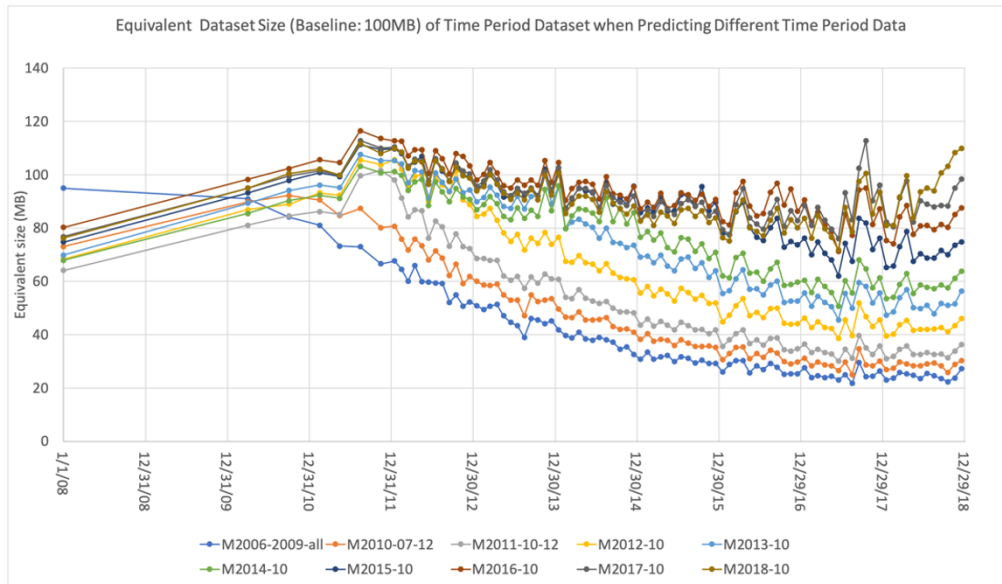


Figure 3.6: Equivalent sizes over time (x-axis) when we used 100MB of data in the training phase. Each curve is the trained model. The legend describes the time we used to train these models. For example, the yellow curve shows a model that has been trained on data from October 2012.

ferent regions of the learning curve, small change in measured cross-entropy translates to different magnitudes of change in equivalent sizes. For example, in Figure 3.4, if the training size is 100MB with measured cross-entropy of 5, the equivalent size is roughly 25MB. A small change of 0.1 in the measured cross-entropy translates to an equivalent size of roughly 20MB, which is 5MB different from the previous measurement. However, a similar small change, when the cross-entropy is 4, makes the difference of roughly 50MB. Therefore, the closer the equivalent size is to the training size, a smaller error causes a higher variability. This also explains the overshoots of later models (2017 and 2018) in equivalent sizes in August 2017.

For the second question, aside from the test set’s sampling issues, model errors, and numerical error in fitting the learning curve’s functional, we believe it is natural for events on those occasions to be slightly more predictable by all models. For example, for August 2017, if we look at the predictive power of $m_{2006-2009}$, we cannot find a considerable change, and sudden increase looks normal.

However, due to the magnification of error and variability in later models (Models with sampling time closer to 2017), we see considerable changes in their equivalence values that sometimes lead to overshoots above 100MB.

At last, Figure 3.7 shows the effectiveness curves. To deal with issues of the sudden increase in equivalent sizes, we made a slight alteration on the way we calculate the effectiveness curve. In this way, since theoretically $\bar{n}_{D_{n,0}} = n$, we calculate $E_{n,t} = \frac{\bar{n}_{D_{n,t}}}{n} = \frac{\bar{n}_{D_{n,t}}}{\bar{n}_{D_{n,0}}}$. In other words, instead of dividing the equivalent size of time t to 100MB, we divide it by the measured equivalent size of test time. It is as if we divide the measured value by the value of the best model predicting the test time. Doing this process over models from a few time periods creates Figure 3.7.

In this Figure, we can confirm a monotonic decrease of the effectiveness curve. It is interesting to see that the effectiveness curves of models from different times are all lined up. As this graph shows, roughly around 7 to 8 years, the value of data for the algorithm and the next-word-prediction task drops 50%. Furthermore, we can see small periodic behavior in the measurements. For example, looking at the values of days 365, 730, and 1095 and comparing them with the values of days 181, 550, and 915, we can see small ripples in the overall form of effectiveness functional. It suggests small periodicity in the data.

3.5 CONCLUSION

An increase in the size of a dataset improves the generalizability and the accuracy of machine learning models. This argument and managerial theories on how the business output scales with the availability of resources convinced economists and data scientists that having more data always improves the quality of AI-based products and services. For long, such statements triggered debates on whether stock of available data owned by big tech firms creates a barrier to the entry of competitors and if, much like the network effect, the data network effect creates a winner take all situation. Sev-

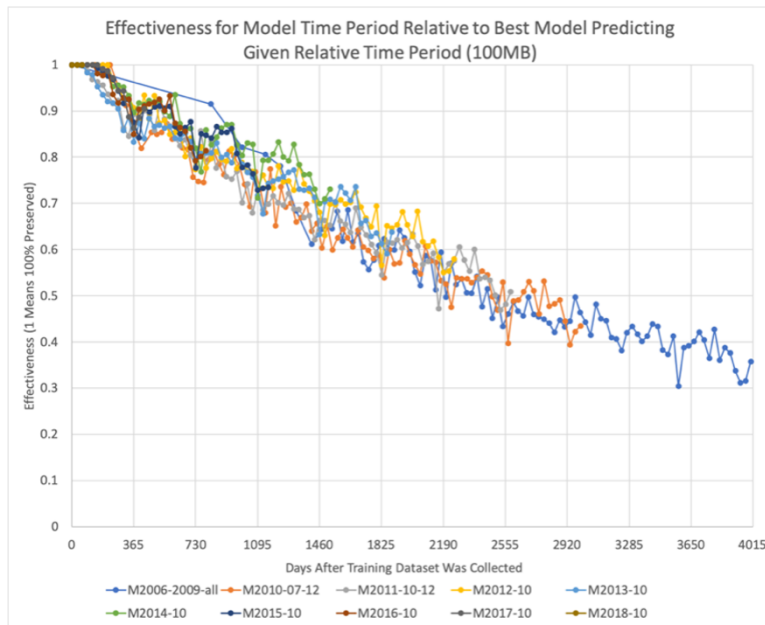


Figure 3.7: Effectiveness curve. The X-axis shows the number of days after the training dataset was collected. The Y-axis shows the effectiveness of the trained model. 1 means that 100% of the dataset's value has persevered. The legend describes the time we used to train these models. For example, the yellow curve shows a model that has been trained on data from October 2012.

eral empirical studies contested this view and argued that such economies of scale are hard to achieve in AI-based businesses for a variety of reasons. These empirical studies often cited data's diminishing return to scale and argued that the marginal value of new data points decreases as datasets size grows.

In our research, we argued that time-dependency, despite having a significant effect on scaling the business value, is neglected in these debates. Time-dependency refers to the change in business value over time because of training models on datasets sampled from a dynamically changing environment. We cite the innovation in products and services space as well as the change in consumers' tastes and behavior as contributing factors to the change in environments over time. We further argue that due to innovation, with certainty, older datasets lose relevance over time. Mainly because the data-generating probability distributions in the future are different in that there is no combination of distributions from the past that can add up the future probability distributions.

We theoretically show that even a perfect model trained on an infinite supply of time-dependent data may have lower accuracy than the same model trained on a recent (perfectly relevant) dataset of finite size. For dynamically changing business environments, this theorem immediately dismisses the role of stock of historical data in creating a barrier to entry of a competitor mainly because a competing firm equipped with a finite (yet sufficient) amount of recent data can attain a similar accuracy score and enter the market. In addition, this theorem has economic modeling implications. It states that a simple discounting function often used in the form of exponential decay over time is not suitable for modeling the growth in the data economy. Instead, an adequate discounting model should be a function of the size in addition to being a function of time.

We further introduced metrics like the "equivalent size" that report the impact of time-dependency in dataset sizes. To evaluate the equivalent size, we first defined an oracle dataset. We then measured the size of the oracle dataset that, if used for training the machine learning model, leads to a similar accuracy score as the model trained on our dataset. In other words, the oracle dataset size creates a

base of comparison in comparing various datasets. We then introduced the substitution function that measures the gain/loss in equivalent size when substituting datasets from various times. The equivalent size and the substitution function are the building blocks of our machinery to measure the value of data over time. Lastly, we proved the existence of “equivalent time” to extend our machinery’s capability to compare values of dataset curated over time. In our extended machinery, a dataset curated over time is represented by another dataset of similar size sampled from the equivalent time. Using the extended machinery, we formulated offloading algorithm. The outcome of this algorithm suggests that a business may remove old (less relevant) data from its repository and, despite losing size, end up in a better competitive position. In other words, the gain in relevance counterbalances the loss in size. On the other hand, a successful iteration of this algorithm suggests that increasing the dataset size by including an older dataset may put a firm in a disadvantageous position.

The offloading algorithm builds the case for our argument that, in a rapidly changing environment, a firm should focus on the flow of data (defined as data collection rate) as the primary value driver instead of the stock of data. To formalize this idea, we first showed that when the flow of data increases, a firm gains/loses more if it substitutes data from various times. Then, we proved that the increase in the gain/loss (in the substitution function) makes the firm offload more often to bring the equivalent time closer to the prediction time. Such action makes the firm focus on the most relevant (recently collected) data. Therefore, when a firm grows (which often leads to an increase in the flow of data), it focuses on the most relevant (recent) data as the primary value driver. Consequently, we argue that the flow of data is the main value driver for big tech companies. Therefore, the marginal value of new data for such firms is always positive and economically significant, forcing them to collect new data aggressively.

Our findings can explain the mixed the result in the literature. First, we acknowledge the feedback loop logic in (52). However, Proposition 3.2.1 shows that the stock of available data produced

by the feedback loop has a finite oracle size because of time dependency. Hence, despite the accelerated growth in the size of data repository, we shouldn't expect a significant increase in created business value. In other words, the feedback loop stalls in dynamically changing environments unless the firm offloads its less relevant data and focuses on the flow of data as the primary value driver. This finding supports the reported results in (29) and (11) since both search engine (29) and advertisement (11) businesses use time-sensitive data and hence, face significant time dependency.

We can extend our arguments and results to study other data characteristics as long as we can model them by a change in the data-generating probability distribution. It is because all our definitions, theorems, and propositions are a function of variation in the distributions. For example, we may extend this result to measure the value loss in the user dimension. In other words, we may model the heterogeneity in preferences across users by variation in their preference distributions. Then, we can measure the value of a user's data on predicting other user's preferences. In this research, we chose to center our arguments around the time dimension since it is easier to visualize value decline over time. Mainly because experimenting over time dimension has the benefit of having a possibly semi-monotonic decline in the value of data.

4

Time-Dependency, Data Flow, and Competitive Advantage

Virtually, for every industry, data-driven externalities (⁶⁵) create forces that shape the way businesses compete. Notably, data availability can create a growth cycle between data volume and algorithmic quality (^{54,90}): more data leads to a better quality of products and services, which in turn increases demand; the increase in demand leads to an even higher volume of data and thereby completes the

cycle.

From previous chapter we know that the magnitude of these competitive forces is subject to change and depends on data characteristics, such as time-dependency. Recall that time-dependency refers to the attribute that data's relevance and merit in making accurate predictions decline over time. In other words, data often is a non-durable asset, and its value perishes over time. For example, the advertisement data that is valuable now in predicting a person's purchasing preferences may be much less valuable tomorrow, next week, or next year, as the person's preferences will change. This kind of time-dependency can dramatically alter the balance of competitive advantage and transform data's influence in creating "moats," barriers to the entry of a new competitor.

This chapter, using natural language processing models and naturally occurring consumers' text data from Reddit.com (⁴⁰), shows unequal time-dependency and speed of change among different text topics representing various interest areas. We measure the change in data value for different subreddits and show they perish with different rates. For example, the value of data in the "relationship" subreddit, perishes much slowly than the value of data from "world news" and "politics" subreddits.

Reddit is a social news aggregation platform founded in 2005. As of January 2021, according to Alexa internet, Reddit is the 18th most visited website worldwide and 7th in the United States. 49% of the traffic is from the U.S. following by 10% and 5% from the United Kingdom and Canada. It has around 330 million monthly active users. On Reddit, users share their opinions on many different issues and contribute to multiple discussions.

Similar to what we have done in previous chapter, we train a small variant of GPT-2, the Generative Pre-Training transformer-based model from OpenAI (^{51,92}), for the next-word-prediction task. The next-word-prediction algorithm predicts the next word in a sentence given a sequence of words. We use cross-entropy as our loss function and choose the dataset size of 100MB that allows us to stay in the learning curve's power-law region (⁵⁷). We measure the effectiveness curve and fit an

exponential function to the measurements.

We train the algorithm with large dataset sizes to assure it is adequately tuned to linguistic models. Consequently, at this level of training, we believe errors in predicting the next words mostly stem from the changes in different topics. For example, in the computer operating system topic space, after the sequence, "download windows," we may expect 'XP' in 2002 and '11' in 2022. Similarly, in science, as time goes by, researchers propose better experimentation methods and may find altered results. For example, if they claimed in 2002 that "Coffee drinking is good for heart disease" and then change the claim in 2020 to "Coffee drinking is not good for heart disease," the next-word-prediction algorithm picks up this development. As a result, much like the vast literature on online word-of-mouth and its economic implications, (75,84,28,49,53,70) we believe that our findings in this research are informative about the speed of change and innovation in various business areas.

4.1 PERISHABILITY MEASUREMENT METHOD

This section briefly reminds us of the definitions we introduced in the previous chapter. Recall that data perishability studies change in the value of data over time. We defined a metric called data effectiveness to capture data's relative effectiveness in making predictions at every point in time. The perishability is then to see how the data effectiveness changes over time.

We elaborate on how to measure data effectiveness using an example shown in Figure 4.1. As depicted in Figure 4.1(i), we train a model on the dataset (\mathcal{A}) of size $|\mathcal{A}|$ sampled from time 0. We then evaluate the model's performance (Loss value) on a testing dataset sampled from time T . Let's say the model produces the loss value L . We then use a training dataset sampled at time T to see what training set size from this sample (if tested on a dataset from time T) would result in a similar loss value. Let's say size $|B|$ from time T reaches L . Size $|B|$ is expected to be less than size $|\mathcal{A}|$ since \mathcal{A} has lost its predictive relevance over time. We define the ratio $\frac{|B|}{|\mathcal{A}|} \in [0, 1]$ as dataset \mathcal{A} 's effectiveness

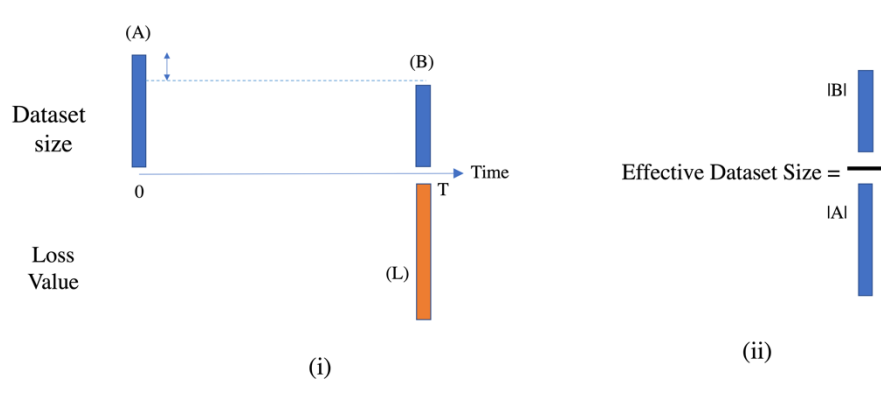


Figure 4.1: These figures conceptually illustrate the loss in data value due to time-dependency. The bars below the time axis represent the loss value, and the bars above the time axis are dataset sizes. In figure 4.1 (i), (A) represents a training set with size $|A|$ that is sampled at time 0. (L) is the loss value measured for the model trained on the dataset (A) and tested at time T . (B) represent a training dataset with size $|B|$ from time T that if we train the model using (B) and test it at T , the loss value would be (L). Figure 4.1 (ii) shows the dataset's effectiveness value. The perishability curve is then to measure effective datasets sizes for multiple T and study the evolution of effectiveness.

at time T .

4.2 PERISHABILITY CURVES TRACK REAL-WORLD CHANGES

To reassure that our method tracks real-world changes, we look into a few subreddits' perishability curves. For example, in Figure 4.2, we measure the perishability of datasets from October 2012 and see if the measured changes correspond with real-world events. We can see periodicity in perishability of sports datasets like "hockey" (This is expected since we have seasonality in sport) or a flat perishability curve in "history" (This is expected since commentators in such forums usually discuss events far in the past). Yet, the most interesting behavior arises in the "politics".

Figure 4.3 presents the perishability curves of the "politics" subreddit. We observe that the value of 2012's data declined mostly in mid 2015. The observation indicates that political discussions in 2015 and 2016 (Before 2016's presidential election) are not predictable from 2012's data, and we suspect a drastic shift in the political landscape and a change in political discussions on those years.

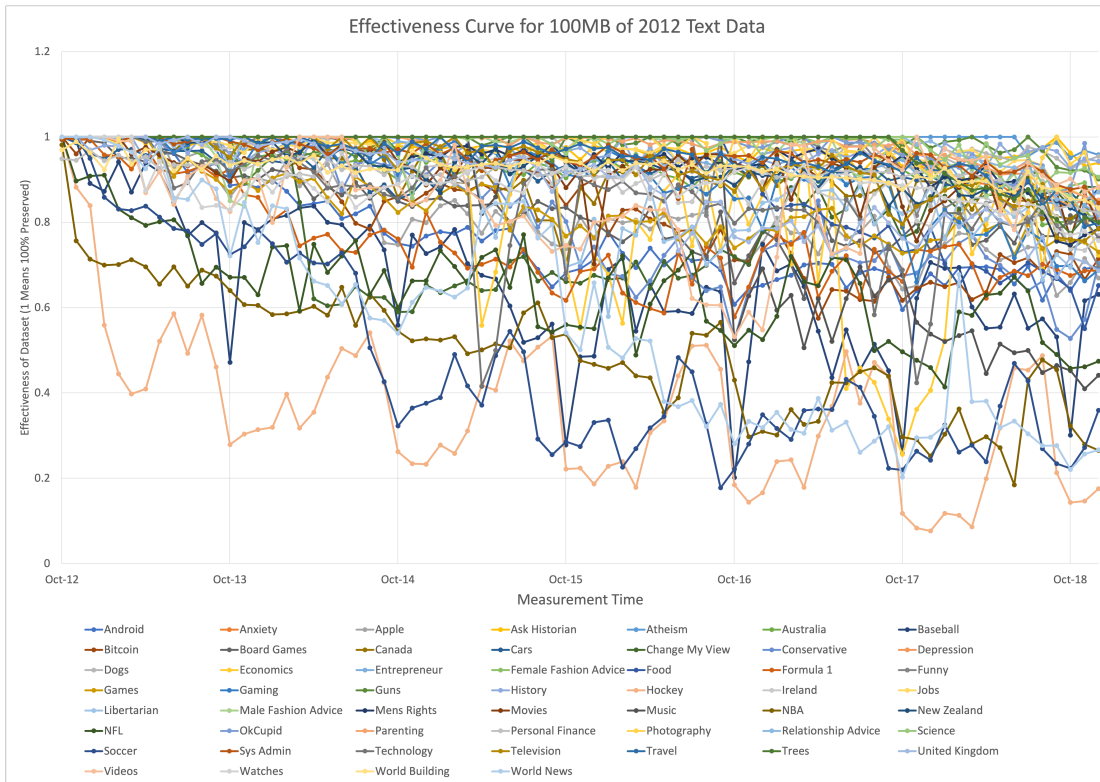


Figure 4.2: Effective dataset sizes for multiple topics. This graph visually shows the difference in data perishability between the topics. It is worth noting that for most subreddits, the dataset’s effectiveness is not constantly diminishing. We see ripples or sudden drops in value. The general trend, though, shows a decrease in the overall value of data over time.

These observation highlights our method’s functionality in tracking changes over time.

4.3 CHARACTERIZING THE PERISHABILITY TRENDS

The perishability curves in Figure 4.2 don’t lend themselves to a unique functional form. We can observe a macro trend for each curve, showing an overall decline in the value of data and a micro-trend that often manifests itself with periodicity. The micro-trend is particularly visible in the “hockey” subreddit dataset in Figure 4.2. For the rest of this section, we focus on the macro trends and characterize the overall decline rate in the data value for the entire Reddit dataset and a few sub-

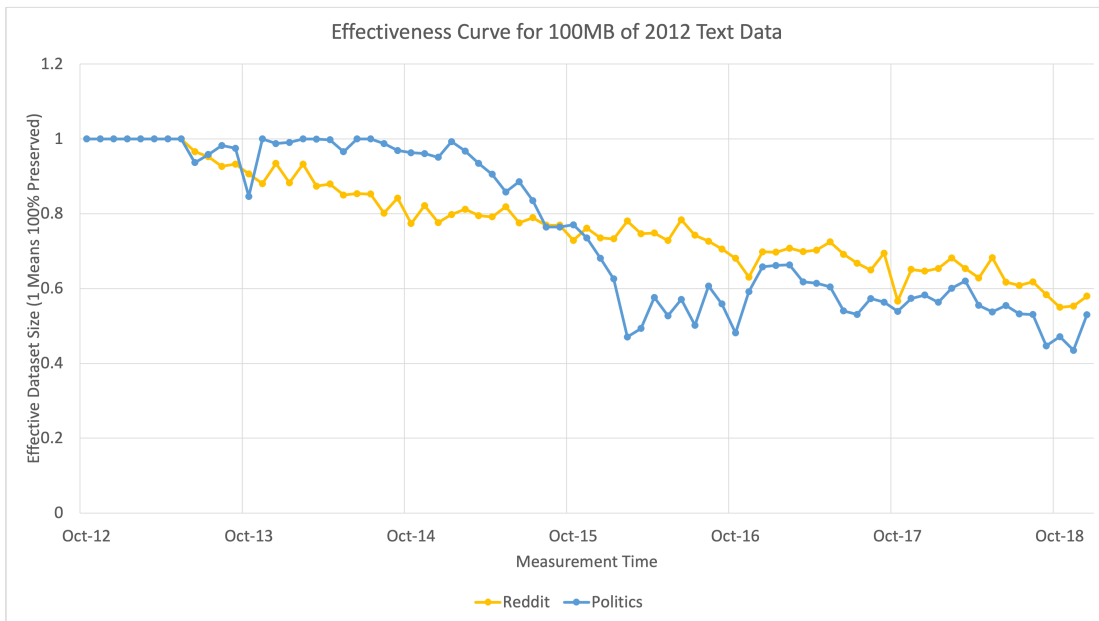


Figure 4.3: Effective dataset sizes for the politics subreddit (Blue) and entire Reddit data (Yellow). The dataset from all Reddit topics has a monotonic decrease in value. In contrast, the dataset sampled in October 2012 from the politics subreddit perishes mainly at the beginning of 2015 and reaches the lowest point in February 2016 (The months leading to the 2016 United States presidential election)

reddits. We find (Explained in supplementary information section) that exponential function fits the decay trend best. Table 4.1 shows the estimated exponential decay rates μ from the functional form $e^{-\mu t}$ for different subreddits. It is estimated using

$$\log(y) = \mu t + u \tag{4.1}$$

where y is the measured effectiveness, t is time (in years), and u is normally distributed fitting noise. Since decay rates might be hard to interpret, we also provided the dataset half-life-time. Half-life-time is the period that it takes for a dataset to loses half of its predictive substance. , i.e., the time t where $e^{-\mu t} = \frac{1}{2}$. Figure 4.4 provides the half-life-time for multiple subreddit datasets.

Table 4.1: Perishability rate measurements for several topics.

Topics (Subreddit)	Estimate ($-\mu \sim \frac{1}{\text{year}}$)	Half-Life-time (Years)	Standard Error (All estimates are significant at 10^{-3})
History	-0.004	100>(168.9)	6.56E-04
Relationship	-0.010	66.69	4.71E-04
Movies	-0.026	26.53	0.001
Food	-0.048	14.39	0.002
Technology	-0.054	12.78	0.003
Apple (the company)	-0.059	11.76	0.001
Entire Reddit	-0.084	8.22	0.001
NFL	-0.100	6.92	0.003
Music	-0.108	6.43	0.004
Baseball	-0.122	5.67	0.005
Politics	-0.151	4.58	0.004
NBA	-0.189	3.67	0.004
Soccer	-0.230	3.00	0.006
World News	-0.233	2.97	0.005
Hockey	-0.245	2.83	0.009

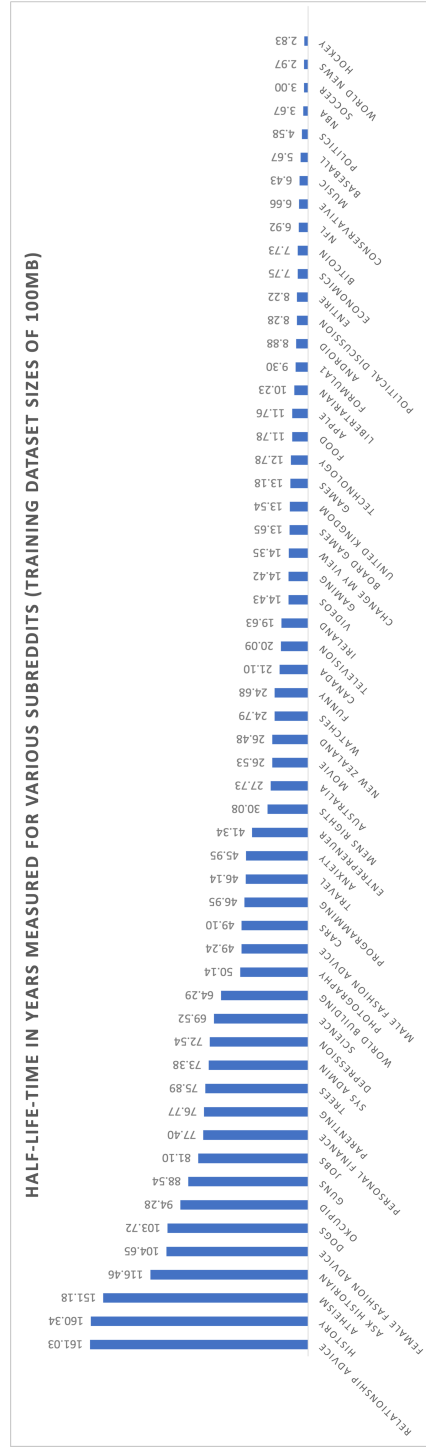


Figure 4.4: Half-life-time measured for a few most visited subreddits. We see sports, politics and economics on the right side (Higher perishability) followed by technology, society and leisure in the middle. On the far left, we see topics mostly related to personal development.

It is interesting to see that subreddits like “relationship” appearing to be durable. In other words, the way people talk about relationships is not changing quickly. In contrast, it takes 4-5 years for a dataset of 100MB in “politics” to lose half of its predictive substance.

Two other interesting topics arise from United States professional sports setting by comparing the National Basketball Association (NBA) and the National Football League (NFL). They have significantly different half-life-time. We believe it is due to their structural differences. To name one, the number of games each team plays each season in the NBA is 82, whereas it is 16 games in the NFL. In total, the NFL has 256 games per season, which is much smaller than the NBA. Therefore, the opportunity of new events in the NBA is naturally higher, which accounts for 3.67 years of half-life time comparing to 6.92 years in NFL. Similarly, National Hockey League schedules 82 games per team, and therefore the subreddit “hockey” has perishability rate similar to NBA’s.

Other factors like players’ longevity and movements are essential in the predictability of events. Basketball, football, and hockey are physically demanding sports and have roughly similar players’ longevity. Hence, the number of games is a good proxy for measuring the number of new events. In contrast, the baseball subreddit, despite MLB’s 162 games per season, has lower perishability. It means that the rate of new events per season is lower for MLB comparing with NFL, NBA, or NHL. A good explanation for this is players’ longevity, which is higher in baseball.

4.4 PAIRWISE COMPARISON OF MACRO TRENDS

In the previous section, we measured the decay rates for a few subreddits and observed varying degrees of perishability over different topics. This section wants to certify that the decay rates (macro trends) are different between subreddits, i.e., to see if two datasets have indistinguishable macro trends if they have similar perishability rates. To answer this question, we formulate a new test that

captures the difference between every subreddit pairs as follows:

$$\log(y_i) - \log(y_j) = -\beta t + \varepsilon \quad (4.2)$$

Where y_i is the measured effectiveness for topic $i \in \{relationship, history, \}$, t is time, and ε is the normally distributed noise. The coefficient β should be zero if two topics have identical decay rate. Therefore, in comparing different topics, the null hypothesis would be $\beta = 0$.

Figure 4.5 shows the p-value for subreddit pair's β estimates. We expect a higher p-value (shown in darker blue) if two datasets have relatively indistinguishable macro trends. Statistically speaking, darker blue means that a significant difference in perishability rates is not evident from our data using the exponential decay model.

In Figure 4.5, we see about ten dark blue clusters of subreddit pairs. Though we can't establish a causal relationship between the subreddits in each cluster, it is still worth noting that a casual relation would lead pairs of subreddits to belong to the same cluster. For example, we expect subreddits like "gaming", "board games," and "games" to belong to the same cluster. We don't want to draw a causal conclusion, but it is interesting to see that the pairs (economics, conservative) and (economics, bitcoin) have dark blue in Figure 4.5. Yet, the pair (conservative, bitcoin) has a light blue color.

4.5 IMPLICATIONS

Data perishability has strategic implications for businesses that provide data-driven products and services. High perishability undermines the importance of data volume or historical data in creating a competitive advantage. On the other hand, as proved in the supplementary information section and also suggested by Claussen et al.⁽³⁰⁾, increasing the flow of data can compensate for the volume's value loss due to perishability.

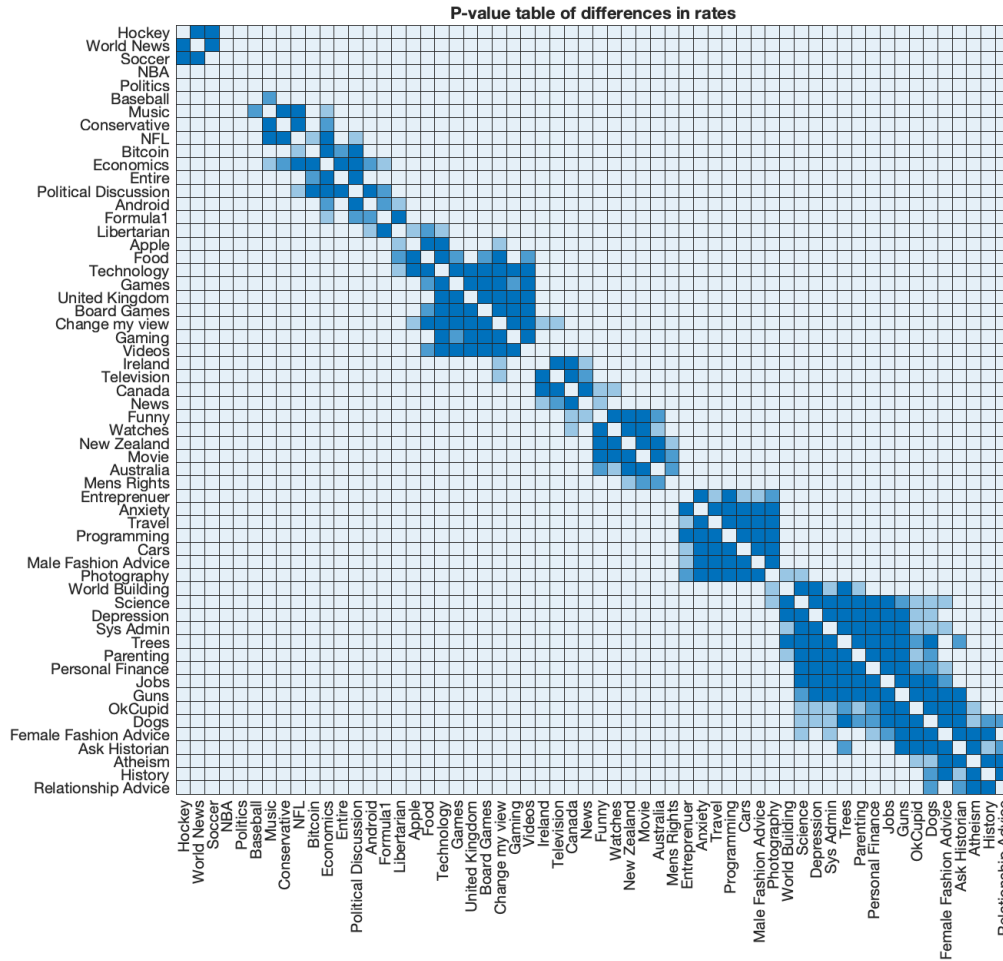


Figure 4.5: The p-values for estimated β (The difference between perishability rates). There are four colors in this graph. Very light blue means that the subreddit pair's estimated β is different from zero at 10^{-3} significance level. Light blue and blue colors show significance at 0.01 and 0.05 levels, respectively. The darkest blue color indicates that the p-value is higher than 0.05, meaning that for the exponential decay model, a significant difference in perishability rates is not evident from our data.

Our findings directly influence practices benefiting from user-generated text data (^{12,37,48,83,10}). For example, services and products like search engines, recommendation systems, and AI-enabled personal assistants and translators can adjust the algorithms and training data repository to account for the importance of data flow in different contexts.

Although our perishability analysis has been limited to products and businesses driven by the language models, we believe the conclusion that the data perishability rate is a function of dataset can be extended to other businesses. In other words, our method indirectly illustrates how different business areas might differ with respect to the rate of decay in data value and hence, the importance of data flow in their operations. Subsequently, business areas facing higher perishability offload historical data more frequently which not only reduces the dataset size and the complexity of operation, but it increases the effectiveness of the datasets. Besides, removing unnecessary data in the offloading process answers some of the users' privacy concerns.

To increase the data flow, businesses can, for example, increase user engagement or expand the user-base when there exist user-wide externalities. The importance of growing user-base highlights the possibility of market dominance and concentration (¹¹⁵), making our findings relevant to antitrust debates. Therefore, from the regulatory perspective, authorities should be aware of the profound difference in the value creation dynamics across business areas. They should craft policies considering economic models that include the flow dynamics for business areas facing higher perishability than models solely based on data stock.

Our method in measuring a dataset's effectiveness can be used in other research areas such as linguistic, economics, and social sciences. For example, one can study the socio-economic impact of innovation, policy introduction, or stimuli on the user's behavior by observing user-generated text data's predictability. While in our research, we used the next-word-prediction task to track behavior changes, other types of prediction tasks can be used depending on the research domain and question.

The supplementary information section contains proofs, details of implementation, and a broader discussion on the choices we made in this chapter.

A

Appendix for Chapter 2

PROOF OF LEMMA 2.2.1

Given the entrant's prices, p_E and w_E , the incumbent's best response is always to choose p_I and w_I , such that the number of service buyers equals the number of service providers on the incumbent's platform. Otherwise, the incumbent can always increase its profit by increasing p_I or decreasing w_I so that the profit margin ($p_I - w_I$) goes up without affecting the matched demand (i.e.,

$\min(N_I^S, N_I^B)$). Similarly, given the incumbent's prices, p_I and w_I , the entrant's best response is always to choose p_E and w_E , such that the number of service buyers equals the number of service providers on the entrant's platform. Q.E.D.

PROOF OF PROPOSITION 2.2.1

We first solve for the optimal prices for the interior equilibrium, where $0 < a_i^* < m$. We first confirm that the second-order derivatives are both negative: $\frac{\partial^2 \pi_I}{\partial p_I^2} = \frac{\partial^2 \pi_E}{\partial p_E^2} = -\frac{2N(1+\frac{1}{(1+r)})^\theta}{m} < 0$.

We then derive the first-order conditions:

$$\frac{\partial \pi_I}{\partial p_I} = N \left(2 + r + \frac{(p_E(3+r) - 2p_I(2+r) + (1+r)w_E)\theta}{m(1+r)} \right) = 0 \quad (\text{A.1})$$

$$\frac{\partial \pi_E}{\partial p_E} = \frac{N(p_I(3+r) - 2p_E(2+r) + (1+r)w_I)\theta}{m(1+r)} = 0 \quad (\text{A.2})$$

And we further obtain the following:

$$p_I = \frac{1}{2} \left(\frac{p_E(3+r) + (1+r)w_E}{2+r} + \frac{m(1+r)}{\theta} \right) \quad (\text{A.3})$$

$$p_E = \frac{(3+r)p_I + (1+r)w_I}{4+2r} \quad (\text{A.4})$$

Solving (A3) and (A4) along with $(p_I - p_E) = (1+r)(w_E - w_I)$ (according to Lemma 2.2.1), we obtain $p_I^* = \frac{2m(2+r)}{3\theta} + w_I$, $p_E^* = \frac{m(3+r)}{3\theta} + w_I$, and $w_E^* = \frac{m}{3\theta} + w_I$. The number of buyers and service providers using each platform under the optimal prices are $N_I^{B^*} = N_I^{S^*} = \frac{2N(1+r)}{3}$ and $N_E^{B^*} = N_E^{S^*} = \frac{N(1+r)}{3}$. The profits under the optimal prices are $\pi_I^*(\theta) = \frac{4mN(1+r)(2+r)}{9\theta}$ and $\pi_E^*(\theta) = \frac{mN(1+r)(2+r)}{9\theta} - L(\theta(2N+rN))$.

For this interior equilibrium to hold, we need to ensure that the incumbent has no incentive to deviate from this equilibrium by charging such a high price $p_I = v$ that no one from the overlapped market will transact on its platform but that it gets the most profit from the users who are not aware of the entrant.*

The highest possible deviation profit the incumbent gets in this case is $(1 - \theta + r)Nv$, whereas the incumbent's equilibrium profit is $\frac{4mN(1+r)(2+r)}{9\theta}$. To guarantee that the latter is higher (i.e., $\frac{4mN(1+r)(2+r)}{9\theta} > (1 - \theta + r)Nv$) for all values of θ , we assume that $m > \frac{9(1+r)v}{16(2+r)}$. This condition also ensures that the incumbent will never completely give up the overlapped market—that is, the entrant will never have the entire overlapped market. This is quite realistic because, in practice, there are always users who have a sufficiently high switching cost such that would rather remain with their current platform.

For this interior equilibrium to hold, we must also have $p_I^* = \frac{2m(2+r)}{3\theta} + w_I < v$. Because the choice of w_I does not affect either platform's profit and any border solution is inferior to the interior solution, the incumbent has incentive to ensure that $p_I^* < v$ holds as much as possible by setting $w_I^* = 0$. Consequently, $p_I^* = \frac{2m(2+r)}{3\theta}$, $p_E^* = \frac{m(3+r)}{3\theta}$, and $w_E^* = \frac{m}{3\theta}$, and the condition $p_I^* < v$ requires $\theta > \frac{2m(2+r)}{3v}$.

When $\theta \leq \min\left(\frac{2m(2+r)}{3v}, 1\right)$, even with $w_I^* = 0$, the optimal p_I^* cannot satisfy $p_I^* < v$. Thus, the incumbent's price is bounded at $p_I^* = v$ to the buyers. Then, based on (A4), $w_I^* = 0$ and $p_I - p_E = (1 + r)(w_E - w_I)$, we obtain $p_E^* = \frac{(3+r)v}{2(2+r)}$ and $w_E^* = \frac{v}{2(2+r)}$. The number of buyers and service providers using each platform under the optimal prices are $N_I^{B*} = N_I^{S*} = \frac{N(1+r)}{2} \left(2 - \frac{v\theta}{m(2+r)}\right)$ and $N_E^{B*} = N_E^{S*} = \frac{N(1+r)v\theta}{2m(2+r)}$. The profits under the optimal prices are $\pi_I^*(\theta) = \frac{N(1+r)v}{2} \left(2 - \frac{v\theta}{m(2+r)}\right)$ and $\pi_E^*(\theta) = \frac{N(1+r)v^2\theta}{4m(2+r)} - L(\theta(2N + rN))$.

*This deviation is not adequately captured by the optimization process above because its calculation automatically assigns a negative profit to the overlapped market if $p_I - p_E > m$

PROOF OF COROLLARY 2.2.1

When $\theta > \frac{2m(2+r)}{3v}$, $\pi_E^*(\theta) = \frac{mN(1+r)(2+r)}{9\theta} - L(\theta(2N+rN))$, which decreases with θ . Thus, the entrant never has the incentive to increase θ once $\theta > \frac{2m(2+r)}{3v}$. Therefore, the optimal choice of θ always satisfies $\theta \leq \min\left(\frac{2m(2+r)}{3v}, 1\right)$.

PROOF OF PROPOSITION 2.2.2

According to Corollary 2.2.1, the entrant always chooses $\theta \leq \min\left(\frac{2m(2+r)}{3v}, 1\right)$; then, according to Proposition 2.2.1, $\pi_I^*(\theta) = \frac{N(1+r)v}{2} \left(2 - \frac{v\theta}{m(2+r)}\right)$ and $\pi_E^*(\theta) = \frac{N(1+r)v^2\theta}{4m(2+r)} - L(\theta(2N+rN))$. Given $L(n) = kn^2$, $\pi_E^*(\theta) = \frac{N(1+r)v^2\theta}{4m(2+r)} - kN^2(2+r)^2\theta^2$. We first confirm that the second-order derivative is negative: $\pi_E^{*''}(\theta) = -2kN^2(2+r)^2 < 0$. Then, we derive the first-order condition:

$$\pi_E^{*'}(\theta) = \frac{N(1+r)^2v^2 - 8kN^2m(2+r)^3\theta}{4m(2+r)} = 0 \quad (\text{A.5})$$

This yields

$$\theta^* = \frac{(1+r)v^2}{8kNm(2+r)^3} \quad (\text{A.6})$$

Because the optimal choice of θ is bounded by $\theta \leq \frac{2m(2+r)}{3v}$ and $\theta \leq 1$, we compare $\frac{(1+r)v^2}{8kNm(2+r)^3}$ with the two bounds and obtain $\frac{(1+r)v^2}{8kNm(2+r)^3} \leq \frac{2m(2+r)}{3v}$ if $k \geq \frac{3(1+r)v^3}{16m^2N(2+r)^4}$ and $\frac{(1+r)v^2}{8kNm(2+r)^3} \leq 1$ if $k \geq \frac{(1+r)v^2}{8mN(2+r)^3}$. Thus, we derive the following two cases:

- i When $k \geq \max\left(\frac{3(1+r)v^3}{16m^2N(2+r)^4}, \frac{(1+r)v^2}{8mN(2+r)^3}\right)$, both $\frac{(1+r)v^2}{8kNm(2+r)^3} \leq \frac{2m(2+r)}{3v}$ and $\frac{(1+r)v^2}{8kNm(2+r)^3} \leq 1$ hold. Then $\theta^* = \frac{(1+r)v^2}{8kNm(2+r)^3}$, and it is easy to verify that $\frac{\partial \theta^*}{\partial r} < 0$. By replacing θ with $\frac{(1+r)v^2}{8kNm(2+r)^3}$ in $\pi_I^*(\theta) = \frac{N(1+r)v}{2} \left(2 - \frac{v\theta}{m(2+r)}\right)$ and $\pi_E^*(\theta) = \frac{N(1+r)v^2\theta}{4m(2+r)} - L(2N(\theta + \theta_2^r))$, we obtain $\pi_I^* = \frac{N(1+r)v}{2} \frac{2 - ((1+r)v^3}{8kNm^2(2+r)^4}$ and $\pi_E^* = \frac{(1+r)^2v^4}{64km^2(2+r)^4}$.
- ii When $k < \max\left(\frac{3(1+r)v^3}{16m^2N(2+r)^4}, \frac{(1+r)v^2}{8mN(2+r)^3}\right)$, either $\frac{(1+r)v^2}{8kNm(2+r)^3} \leq \frac{2m(2+r)}{3v}$ or $\frac{(1+r)v^2}{8kNm(2+r)^3} \leq$

1 does not hold. Then $\theta^* = \min\left(\frac{2m(2+r)}{3v}, 1\right)$, and it is easy to verify that $\frac{\partial\theta^*}{\partial r} \geq 0$.
 When $\frac{2m(2+r)}{3v} < 1$, by replacing θ with $\frac{2m(2+r)}{3v}$ in $\pi_I^*(\theta) = \frac{N(1+r)v}{2}\left(2 - \frac{v\theta}{m(2+r)}\right)$ and $\pi_E^*(\theta) = \frac{N(1+r)v^2\theta}{4m(2+r)} - L(2N(\theta + \theta_2^r))$, we obtain $\pi_I^* = \frac{2Nv(1+r)}{3}$ and $\pi_E^* = \frac{N(1+r)v}{6} - \frac{4kN^2m^2(2+r)^4}{9v^2}$. When $\frac{2m(2+r)}{3v} \geq 1$, by replacing θ with 1 in $\pi_I^*(\theta) = \frac{N(1+r)v}{2}\left(2 - \frac{v\theta}{m(2+r)}\right)$ and $\pi_E^*(\theta) = \frac{N(1+r)v^2\theta}{4m(2+r)} - L(2N(\theta + \theta_2^r))$, we obtain $\pi_I^* = \frac{N(1+r)v}{2}\left(2 - \frac{v}{m(2+r)}\right)$ and $\pi_E^* = \frac{N(1+r)v^2 - 4kN^2m(2+r)^3}{4m(2+r)}$.

PROOF OF COROLLARY 2.2.2

If $k = 0$, according to Proposition 2.2.2 (ii), $\theta^* = \frac{2m(2+r)}{3v}$ when $\frac{2m(2+r)}{3v} < 1$.

PROOF OF COROLLARY 2.2.3

In Proposition 2.2.2, $\frac{\partial\theta^*}{\partial r} < 0$ in (i) and $\frac{\partial\theta^*}{\partial r} \geq 0$ in (ii). Since $\max\left(\frac{3(1+r)v^3}{16m^2N(2+r)^4}, \frac{(1+r)v^2}{8mN(2+r)^3}\right)$ is a decreasing function of r , for intermediate values of k , as r increases, the region can shift from (ii) to (i). So when $k < \max\left(\frac{3(1+r_{\max})v^3}{16m^2N(2+r_{\max})^4}, \frac{(1+r_{\max})v^2}{8mN(2+r_{\max})^3}\right)$, $\frac{\partial\theta^*}{\partial r} \geq 0$. When $k \geq \max\left(\frac{3(1+r_{\min})v^3}{16m^2N(2+r_{\min})^4}, \frac{(1+r_{\min})v^2}{8mN(2+r_{\min})^3}\right)$, $\frac{\partial\theta^*}{\partial r} < 0$. And when $\max\left(\frac{3(1+r_{\max})v^3}{16m^2N(2+r_{\max})^4}, \frac{(1+r_{\max})v^2}{8mN(2+r_{\max})^3}\right) \leq k < \max\left(\frac{3(1+r_{\min})v^3}{16m^2N(2+r_{\min})^4}, \frac{(1+r_{\min})v^2}{8mN(2+r_{\min})^3}\right)$, as r increases from zero, the optimal θ^* increases with r first and then decreases with r .

PROOF OF PROPOSITION 2.2.3

From Proposition 2.2.2, it is easy to verify that $\frac{\partial\pi_I^*}{\partial r} > 0$ in both (i) and (ii) and $\frac{\partial\pi_E^*}{\partial r} < 0$ in (i). To examine $\frac{\partial\pi_E^*}{\partial r}$ in (ii), we consider the following two cases:

- i When $\frac{2m(2+r)}{3v} < 1$, $\max\left(\frac{3(1+r)v^3}{16m^2N(2+r)^4}, \frac{(1+r)v^2}{8mN(2+r)^3}\right) = \frac{3(1+r)v^3}{16m^2N(2+r)^4}$. In this case, $\frac{\partial\pi_E^*}{\partial r} > 0$ if $k < \frac{3v^3}{32m^2N(2+r)^3}$. Since $\frac{3(1+r)v^3}{16m^2N(2+r)^4}$ decreases with r , for intermediate values of k , as r increases, the region can shift from (ii) to (i). So when $k < \frac{3v^3}{32m^2N(2+r_{\max})^3}$, $\frac{\partial\theta^*}{\partial r} > 0$. When

$k \geq \frac{(3v^3)}{32m^2N(2+r_{min})^3}$, $\frac{\partial \theta^*}{\partial r} < 0$. And when $\frac{3v^3}{32m^2N(2+r_{max})^3} \leq k < \frac{3v^3}{32m^2N(2+r_{min})^3}$, as r increases from zero, π_E^* increases with r first and then decreases with r .

ii When $\frac{2m(2+r)}{3v} \geq 1$, $\max\left(\frac{3(1+r)v^3}{16m^2N(2+r)^4}, \frac{(1+r)v^2}{8mN(2+r)^3}\right) = \frac{(1+r)v^2}{8mN(2+r)^3}$. In this case, $\frac{\partial \pi_E^*}{\partial r} > 0$ if $k < \frac{v^2}{8mN(2+r)^3}$. Since $\frac{(1+r)v^2}{8mN(2+r)^3}$ decreases with r , for intermediate values of k , as r increases, the region can shift from (ii) to (i). So when $k < \frac{v^2}{8mN(2+r_{max})^3}$, $\frac{\partial \theta^*}{\partial r} > 0$. When $k \geq \frac{v^2}{8mN(2+r_{min})^3}$, $\frac{\partial \theta^*}{\partial r} < 0$. And when $\frac{v^2}{8mN(2+r_{max})^3} \leq k < \frac{v^2}{8mN(2+r_{min})^3}$, as r increases from zero, π_E^* increases with r first and then decreases with r .

PROOF OF PROPOSITION 2.3.1

If the entrant enters market b , we can similarly obtain the demand function in market b as follows:

$$N_{Ib}^S = \left(1 - \frac{(c_b^*)}{m} \theta_b\right) (1 + r_b)N, \quad (\text{A.7})$$

$$N_{Eb}^S = \frac{c_b^*}{m} \theta_b (1 + r_b)N, \quad (\text{A.8})$$

$$N_{Ib}^B = \left(1 + r_b - \frac{a_b^*}{m} \theta_b\right) N, \quad (\text{A.9})$$

$$N_{Eb}^B = \frac{a_b^*}{m} \theta_b N, \quad (\text{A.10})$$

where $c_b^* = w_{Eb} - w_{Ib}$ and $a_b^* = p_{Ib} - p_{Eb}$.

Following the same procedure as that in the main analysis, we can obtain the following two main propositions in market b with a similar assumption of switching cost ($m > \frac{3v}{5}$). Without loss of generality, let us focus on the case where $\frac{2m(2+r_{max})}{3v} \leq 1$, where r_{max} is the maximum r_b across all

markets.

Proposition 2.2.1A: Given the entrant's choice of θ_b , the optimal prices, number of buyers and service providers, and profits are as follows:

- i If $0 \leq \theta \leq \frac{2m(2+r_b)}{3v}$, $p_{Ib}^* = v$, $p_{Eb}^* = \frac{(3+r_b)v}{2(2+r_b)}$, $w_{Eb}^* = \frac{v}{2(2+r_b)}$, $N_{Ib}^{B^*} = N_{Ib}^{S^*} = \frac{N(1+r_b)}{2} \left(2 - \frac{\theta_b v}{m(2+r_b)}\right)$, $N_{Eb}^{B^*} = N_{Eb}^{S^*} = \frac{N(1+r_b)\theta_b v}{2m(2+r_b)}$, $\pi_{Ib}^*(\theta) = \frac{N(1+r_b)v}{2} \left(2 - \frac{\theta_b v}{m(2+r_b)}\right)$, and $\pi_{Eb}^*(\theta) = \frac{N(1+r_b)\theta_b v^2}{4m(2+r_b)} - L(\theta_b(2N + r_b N))$.
- ii If $\frac{2m(2+r_b)}{3v} < \theta \leq 1$, then $p_{Ib}^* = \frac{(2(2+r_b)m)}{(3\theta_b)}$, $w_{Ib}^* = 0$, $p_{Eb}^* = \frac{(3+r_b)m}{3\theta_b}$, $w_{Eb}^* = \frac{m}{3\theta_b}$, $N_I^{B^*} = N_I^{S^*} = \frac{2N(1+r_b)}{3}$, $N_E^{B^*} = N_E^{S^*} = \frac{N(1+r_b)}{3}$, $\pi_I^*(\theta) = \frac{4Nm(1+r_b)(2+r_b)}{9\theta_b}$, and $\pi_E^*(\theta) = \frac{Nm(1+r_b)(2+r_b)}{9\theta_b} - L(\theta_b(2N + r_b N))$.

We again confirm the entrant's optimal choice of θ_b to be no more than $\frac{2m(2+r_b)}{3v}$. That is, regardless of the market the entrant is in, it is in the best interest of the entrant not to trigger the incumbent's competitive response. Endogenizing θ_b , we obtain the following proposition.

Proposition 2.2.2A: The optimal θ_b^* depends on the value of k :

- i If $k \geq \frac{3(1+r_b)v^3}{(16m^2N(2+r_b)^4)}$, then $\theta_b^* = \frac{(1+r_b)v^2}{8(2+r_b)^3 k N m}$. The entrant's profit is $\frac{(1+r_b)^2 v^4}{64km^2(2+r_b)^4}$ which decreases with r_b .
- ii If $0 \leq k < \frac{3(1+r_b)v^3}{16m^2N(2+r_b)^4}$, then $\theta_b^* = \frac{2m(2+r_b)}{3v}$. The entrant's profit is $\frac{N(1+r_b)v}{6} - \frac{4kN^2m^2(2+r_b)^4}{9v^2}$, which increases with r_b if $k < \frac{3v^3}{(32m^2N(2+r_b)^3)}$.

Thus, the entrant will choose the market that yields the highest profit as follows:

- i If $0 \leq k \leq \frac{3v^3}{(32m^2N(2+r_{max})^3)}$, the entrant will choose the market with the largest fraction of incoming mobile users to enter.
- ii If $\frac{3v^3}{32m^2N(2+r_{max})^3} < k < \frac{3v^3}{32m^2N(2+r_{min})^3}$, the entrant will choose the market with an intermediate fraction of incoming mobile users to enter.

- iii If $k \geq \frac{3v^3}{32m^2N(2+r_{min})^3}$, $\pi_E^* = \frac{(1+r)^2v^4}{64km^2(2+r)^4}$ the entrant will choose the market with the smallest fraction of incoming mobile users to enter.

PROOF OF PROPOSITION 2.3.2

We can similarly obtain the demand function as follows:

$$N_I^S = \left(1 - \frac{c^*}{m}\theta\right)(1+r)sN \quad (\text{A.11})$$

$$N_E^S = \frac{c^*}{m}\theta(1+r)sN + \theta(1-s)N \quad (\text{A.12})$$

$$N_I^B = \left(1 - \frac{a^*}{m}\theta + r\right)sN \quad (\text{A.13})$$

$$N_E^B = \frac{a^*}{m}\theta sN + \theta(1-s)N. \quad (\text{A.14})$$

Following the same procedure as in the main analysis, we can obtain the following two main propositions given $m > \frac{3v}{5}$:

Proposition 2.2.1B: Given the entrant's choice of θ , the optimal prices, transaction quantity, and profits for the two platforms are as follows:

i If $s \geq \frac{m(2+r)}{2m+mr+v+rv}$

- (a) If $0 \leq \theta \leq \min\left(\frac{2m(1+r)(2+r)s}{3(1+r)sv-m(2+r)(1-s)}, 1\right)$, then $p_I^* = v$, $p_E^* = \frac{1}{2}\left(m\left(\frac{1}{s}-1\right) + \frac{(3+r)v}{(2+r)}\right)$, $w_E^* = \frac{1}{2}\left(\frac{v}{2+r} - \frac{m(1-s)}{(1+r)s}\right)$, $w_I^* = 0$, $N_I^{B*} = N_I^{S*} = \frac{N}{2}\left(\theta + s\left(2 + 2r - \theta - \frac{(1+r)v\theta}{m(2+r)}\right)\right)$, $N_E^{B*} = N_E^{S*} = \frac{N}{2}\left(1 - s\left(1 - \frac{(1+r)v}{m(2+r)}\right)\right)\theta$, $\pi_I^*(\theta) = \frac{Nv(m(2+r)(s(2+2r-\theta)+\theta)-(1+r)sv\theta)}{2m(2+r)}$, and $\pi_E^*(\theta) = \frac{N(m(2+r)(1-s)+(1+r)sv)^2\theta}{4m(1+r)(2+r)s} - L(\theta(2N+rN))$.

(b) If $\min\left(\frac{2m(1+r)(2+r)s}{3(1+r)sv-m(2+r)(1-s)}, 1\right) < \theta \leq 1$, then $p_I^* = \frac{m(2+r)(s(2+2r-\theta)+\theta)}{3(1+r)s\theta}$, $p_E^* = \frac{m}{3}\left(\frac{(3+2r)(1-s)}{(1+r)s} + \frac{(3+r)}{\theta}\right)$, $w_E^* = \frac{m}{3}\left(\frac{1}{\theta} - \frac{1-s}{(1+r)s}\right)$, $w_I^* = 0$, $N_I^{B^*} = N_I^{S^*} = N^{\frac{(s(2(1+r)-\theta)+\theta)}{3}}$, $N_E^{B^*} = N_E^{S^*} = \frac{N(s(1+r-2\theta)+2\theta)}{3}$, $\pi_I^*(\theta) = \frac{mN(2+r)(s(2+2r-\theta)+\theta)^2}{9(1+r)s\theta}$, and $\pi_E^*(\theta) = \frac{mN(2+r)(s(1+r-2\theta)+2\theta)^2}{9(1+r)s\theta} - L(\theta(2N+rN))$.

ii If $s < \frac{m(2+r)}{2m+mr+v+rv}$

(a) $p_I^* = v$, $p_E^* = v$, $w_E^* = 0$, $w_I^* = 0$, $N_I^{B^*} = N_I^{S^*} = (1+r)sN$, $N_E^{B^*} = N_E^{S^*} = (1-s)N\theta$, $\pi_I^*(\theta) = (1+r)sNv$, and $\pi_E^*(\theta) = (1-s)Nv\theta - L(\theta(2N+rN))$.

Endogenizing θ , we obtain the following proposition.

Proposition 2.2.2B: The optimal θ^* depends on the value of k and the value of s .

i If $s \geq \frac{m(2+r)}{2m+mr+v+rv}$

(a) If $k \geq \max\left(\frac{(m(2+r)(1-s)+(1+r)sv)^2(3(1+r)sv-m(2+r)(1-s))}{16m^2N(1+r)^2(2+r)^4s^2}, \frac{(m(2+r)(1-s)+(1+r)sv)^2}{8mN(1+r)(2+r)^3s}\right)$,

then $\theta^* = \frac{(m(2+r)(1-s)+(1+r)sv)^2}{8kmN(1+r)(2+r)^3s}$, which decreases with r and s . The entrant's profit is

$\frac{(m(2+r)(1-s)+(1+r)sv)^4}{64km^2(1+r)^2(2+r)^4s^2}$ and the incumbent's profit is

$$\frac{v(m^3(2+r)^3(1-s)^3 - m(1+r)^2(2+r)(1-s)s^2v^2 - (1+r)^3s^3v^3 + m^2(1+r)(2+r)^2s(16kN(1+r)(2+r)^2s + (1-s)^2v))}{16km^2(1+r)(2+r)^4s}$$

(b) If $0 \leq k < \max\left(\frac{(m(2+r)(1-s)+(1+r)sv)^2(3(1+r)sv-m(2+r)(1-s))}{16m^2N(1+r)^2(2+r)^4s^2}, \frac{(m(2+r)(1-s)+(1+r)sv)^2}{(8mN(1+r)(2+r)^3s)}\right)$,

then $\theta^* = \min\left(\frac{2m(1+r)(2+r)s}{(3(1+r)sv-m(2+r)(1-s))}, 1\right)$, which weakly increases with r and weakly

decreases with s . When $\frac{2m(1+r)(2+r)s}{3(1+r)sv-m(2+r)(1-s)} < 1$, the entrant's profit is

$$\frac{N((m(2+r)(1-s)+(1+r)sv)^2(3(1+r)sv-m(2+r)(1-s))-8km^2N(1+r)^2(2+r)^4s^2)}{2(3(1+r)sv-m(2+r)(1-s))^2}$$
 and the incum-

bent's profit is $\frac{2N(1+r)^2s^2v^2}{3(1+r)sv-m(2+r)(1-s)}$. When $\frac{2m(1+r)(2+r)s}{3(1+r)sv-m(2+r)(1-s)} \geq 1$, the entrant's

profit is $\frac{N(m(2+r)(1-s)+(1+r)sv)^2}{(4m(2+r)(1+r)s)} - k((2+r)N)^2$ and the incumbent's profit is

$$\frac{Nv(m(2+r)(1+s+2rs)-(1+r)sv)}{2m(2+r)}$$

ii If $s < \frac{m(2+r)}{2m+mr+v+rv}$,

- (a) If $k \geq \frac{(1-s)v}{2N(2+r)^2}$, then $\theta^* = \frac{(1-s)v}{2kN(2+r)^2}$, which decreases with s and r . The entrant's profit is $\frac{(1-s)^2v^2}{4k(2+r)^2}$ and the incumbent's profit is $(1+r)sNv$.
- (b) If $0 \leq k < \frac{(1-s)v}{2N(2+r)^2}$, then $\theta^* = 1$, the entrant's profit is $N((1-s)v - kN(2+r)^2)$ and the incumbent's profit is $(1+r)sNv$.

PROOF OF PROPOSITION 2.3.3

In this case, we only need N service providers to match N orders in each local market. We can obtain the demand functions as:

$$N_I^S = (1 - \frac{c^*}{m}\theta)N \quad (\text{A.15})$$

$$N_E^S = \frac{c^*}{m}\theta N \quad (\text{A.16})$$

$$N_I^B = (1 - \frac{a^*}{m}\theta)(1-r)N + rN \quad (\text{A.17})$$

$$N_E^B = \frac{a^*}{m}\theta(1-r)N \quad (\text{A.18})$$

We then follow the same procedure as in our main analysis to derive the following two main propositions given $m > \frac{3v}{5}$:

Proposition 2.2.1C: Given the entrant's choice of θ , the optimal prices, number of buyers and service providers, and profits are as follows:

- i If $0 \leq \theta \leq \min\left(\frac{2(2-r)m}{3(1-r)v}, 1\right)$, then $p_I^* = v, p_E^* = \frac{(3-2r)v}{2(2-r)}, w_E^* = \frac{(1-r)v}{2(2-r)}, w_I^* = 0, N_I^{B^*} = N_I^{S^*} = \frac{N(2(2-r)m - (1-r)\theta v)}{2(2-r)m}, N_E^{B^*} = N_E^{S^*} = \frac{(1-r)N\theta v}{2(2-r)m}, \pi_I^* = \frac{Nv(2(2-r)m - (1-r)\theta v)}{2(2-r)m}$, and $\pi_E^* = \frac{(1-r)N\theta v^2}{4(2-r)m} - L(2\theta N)$.

- ii If $\min\left(\frac{2(2-r)m}{3(1-r)v}, 1\right) < \theta \leq 1$, then $p_I^* = \frac{2(2-r)m}{3(1-r)\theta}$, $p_E^* = \frac{(3-2r)m}{3(1-r)\theta}$, $w_E^* = \frac{m}{3\theta}$, $w_I^* = 0$, $N_I^{B^*} = N_I^{S^*} = \frac{2N}{3}$, $N_E^{B^*} = N_E^{S^*} = \frac{N}{3}$, $\pi_I^*(\theta) = \frac{4N(2-r)m}{9(1-r)\theta}$, and $\pi_E^*(\theta) = \frac{N(2-r)m}{9(1-r)\theta} - L(2\theta N)$.

We again confirm the entrant's optimal choice of θ to be no more than $\min\left(\frac{2(2-r)m}{3(1-r)v}, 1\right)$. That is, it is in the best interest of the entrant not to trigger the incumbent's competitive response. Endogenizing θ , we obtain the following proposition.

Proposition 2.2.2C: The optimal θ^* depends on the value of k :

- i. If $k \geq \max\left(\frac{3(1-r)^2v^3}{64m^2N(2-r)^2}, \frac{(1-r)v^2}{32mN(2-r)}\right)$, then $\theta^* = \frac{(1-r)v^2}{32kNm(2-r)}$, which decreases with r . The entrant's profit is $\frac{(1-r)^2v^4}{256km^2(2-r)^2}$ and the incumbent's profit is $Nv - \frac{(1-r)^2v^4}{64km^2(2-r)^2}$.
2. If $0 \leq k < \max\left(\frac{3(1-r)^2v^3}{64m^2N(2-r)^2}, \frac{(1-r)v^2}{32mN(2-r)}\right)$, then $\theta^* = \min\left(\frac{2(2-r)m}{3(1-r)v}, 1\right)$, which weakly increases with r . When $\frac{2(2-r)m}{3(1-r)v} < 1$, the entrant's profit is $\frac{Nv}{6} - \frac{16kN^2m^2(2-r)^2}{9(1-r)^2v^2}$ and the incumbent's profit is $\frac{2Nv}{3}$. When $\frac{4m}{3v} \geq 1$, the entrant's profit is $\frac{N(1-r)v^2}{4m(2-r)} - 4kN^2$ and the incumbent's profit is $\frac{Nv}{2} \left(2 - \frac{(1-r)v}{m(2-r)}\right)$.

Proposition 2.2.2C is qualitatively similar to Proposition 2.2.2. The only difference is that in this case $\pi_E^*(\theta)$ always decreases with r . Since Corollary 2.2.2, 2.2.3 and 2.2.4 directly follow Proposition 2.2.2, these corollaries also remain qualitatively the same, except that in this case, the entrant's profit always decreases with r . Specifically, if $k = 0$, according to Proposition 2.2.2C (ii), $\theta^* = \frac{2(2-r)m}{3(1-r)v}$ when $\frac{2(2-r)m}{3(1-r)v} < 1$, so Corollary 2.2.2 holds qualitatively. It is easy to verify that in Proposition 2.2.2C, $\frac{\partial \theta^*}{\partial r} < 0$ in (i) and $\frac{\partial \theta^*}{\partial r} > 0$ in (ii). Also since $\max\left(\frac{3(1-r)^2v^3}{64m^2N(2-r)^2}, \frac{(1-r)v^2}{32mN(2-r)}\right)$ is a decreasing function of r , for intermediate values of k , as r increases, the region can shift from (ii) to (i), so Corollary 2.2.3 also holds qualitatively. Lastly, from Proposition 2.2.2C, it is easy to verify that in both (i) and (ii), $\frac{\partial \pi_I^*(\theta)}{\partial r} > 0$ and $\frac{\partial \pi_E^*(\theta)}{\partial r} < 0$, so Corollaries 2.2.4 remains the same for the incumbent's profit, but the entrant's profit always decreases with r .

PROOF OF PROPOSITION 2.3.4

Given the new demand functions, we can follow the same procedure as that in our main analysis to derive the following two main propositions, given $m > \frac{3v}{5}$:

Proposition 2.2.1D: Given the entrant's choice of θ and θ_t , the optimal prices, number of buyers and service providers, and profits are as follows:

1. If $0 \leq \frac{\theta(r\theta_t + \theta)}{2\theta + r(\theta_t + \theta)} \leq \min(\frac{2m}{3v}, \frac{1}{2})$, $p_I^* = v$, $p_E^* = \frac{2r\theta_t v + (3+r)v\theta}{4\theta + 2r(\theta_t + \theta)}$, $w_E^* = \frac{v(r\theta_t + \theta)}{4\theta + 2r(\theta_t + \theta)}$, $w_I^* = 0$, $N_I^{B^*} = N_I^{S^*} = \frac{N(1+r)(4m\theta + 2mr(\theta_t + \theta) - v\theta(r\theta_t + \theta))}{2m(2\theta + r(\theta_t + \theta))}$, $N_E^{B^*} = N_E^{S^*} = \frac{N(1+r)v\theta(r\theta_t + \theta)}{2m(2\theta + r(\theta_t + \theta))}$, $\pi_I^*(\theta, \theta_t) = \frac{N(1+r)v(4m\theta + 2mr(\theta_t + \theta) - v\theta(r\theta_t + \theta))}{2m(2\theta + r(\theta_t + \theta))}$, and $\pi_E^*(\theta, \theta_t) = \frac{N(1+r)v^2\theta(r\theta_t + \theta)}{4m(2\theta + r(\theta_t + \theta))} - L(\theta(2N + rN) + \theta_t rN)$.
2. If $\min(\frac{2m}{3v}, \frac{1}{2}) < \frac{\theta(r\theta_t + \theta)}{2\theta + r(\theta_t + \theta)}$, $p_I^* = \frac{2m(2\theta + r(\theta_t + \theta))}{3\theta(r\theta_t + \theta)}$, $p_E^* = \frac{m}{3}(\frac{2}{\theta} + \frac{(1+r)}{r\theta_t + \theta})$, $w_E^* = \frac{m}{3\theta}$, $w_I^* = 0$, $N_I^{B^*} = N_I^{S^*} = \frac{2N(1+r)}{3}$, $N_E^{B^*} = N_E^{S^*} = \frac{N(1+r)}{3}$, $\pi_I^*(\theta, \theta_t) = \frac{4mN(1+r)(2\theta + r(\theta_t + \theta))}{9\theta(r\theta_t + \theta)}$, and $\pi_E^*(\theta, \theta_t) = \frac{mN(1+r)(2\theta + r(\theta_t + \theta))}{9\theta(r\theta_t + \theta)} - L(\theta(2N + rN) + \theta_t rN)$.

We again confirm that the entrant's optimal choices of θ and θ_t satisfy the condition $\frac{\theta(r\theta_t + \theta)}{2\theta + r(\theta_t + \theta)} \leq \min(\frac{2m}{3v}, \frac{1}{2})$. That is, it is in the best interest of the entrant not to trigger the incumbent's competitive response. The threshold $\frac{\theta(r\theta_t + \theta)}{2\theta + r(\theta_t + \theta)}$ is an increasing function of θ and θ_t and it increases faster with θ than with θ_t . When this condition is satisfied, the entrant's profit is also an increasing function of θ and θ_t and it also increases faster with θ than with θ_t . Thus, endogenizing θ and θ_t , we obtain the following proposition for the case of $k = 0$.

Proposition 2.2.2D: When $k = 0$, the optimal θ^* and θ_t^* are

1. If $\frac{2m(2+r)}{3v} \leq 1$, then $\theta^* = \frac{2m(2+r)}{3v}$, $\theta_t^* = 0$.
2. If $\frac{2m(2+r)}{3v} > 1$, $\theta^* = 1$, and $\theta_t^* = \frac{2m(2+r) - 3v}{r(3v - 2m)}$ when $3v > 4m$ and $\theta_t^* = 1$ when $3v \leq 4m$.

Note that in case (ii), the entrant obtains all local users and has not triggered a response from the incumbent. Therefore, it proceeds to advertise to mobile buyers.

PROOF OF PROPOSITION 2.3.5

Assume that a^* is the switching cost of the indifferent user and c^* is the switching cost of the indifferent service provider. Then, Equations (1)–(4) define the number of buyers and service providers that select the entrant and the incumbent, respectively. Then, given the utility functions in Equations (9)–(12), we derive a^* and c^* by solving the following two equations simultaneously:

$$e\left(1 - \frac{c^*}{m}\theta\right)(1+r)N + v - p_I = e\frac{(c^*(1+r))}{m}\theta N + v - p_E - a^*, \quad (\text{A.19})$$

$$e\left(1 - \frac{a^*}{m}\theta + r\right)N + w_I = e\frac{a^*}{m}\theta N + w_E - c^*. \quad (\text{A.20})$$

Thus, we obtain $a^* = \frac{m(p_I - p_E - eN(1+r)) + 2eN(1+r)\theta(w_E - w_I - eN(1+r))}{m^2 - 4e^2N^2(1+r)\theta^2}$ and $c^* = \frac{m(m(w_E - w_I - eN(1+r)) + 2eN\theta(p_I - p_E - eN(1+r)))}{m^2 - 4e^2N^2(1+r)\theta^2}$. We can then prove that Lemma 2.2.1 holds in this extension, that is, the incumbent and the entrant will always choose their prices and wages so that $a^* = (1+r)c^*$, as long as $e < \frac{m}{4N}$.

We then follow the same procedure as that in our main analysis to derive the two main propositions and find that our key results hold qualitatively under conditions $m > \frac{3v}{5}$ and $e < \frac{v}{2N}$.

Proposition 2.2.1E: Given the entrant's choice of θ , the optimal prices, number of buyers and service providers, and profits are as follows:

$$\begin{aligned} \text{i If } 0 \leq \theta \leq \min\left(\frac{4m(2+r)}{16eN(1+r)+3v+\sqrt{(16e^2N^2(1+r)^2+9v^2)}}, 1\right), \text{ then} \\ p_I^* = \frac{(2m(2+r)(eN(1+r)+v) - eN(1+r)(6eN(1+r)+8v)\theta)}{(2m(2+r)-7eN(1+r)\theta)}, p_E^* = \frac{m(eN(1+r)^2+(3+r)v) - eN(1+r)(eN(1+r)+6v)\theta}{2m(2+r)-7eN(1+r)\theta}, \\ w_E^* = \frac{(m(eN(1+r)(3+2r)+v) - eN(1+r)(5eN(1+r)+2v)\theta)}{(2m(2+r)-7eN(1+r)\theta)}, w_I^* = 0, N_I^{B^*} = N_I^{S^*} = \frac{(N(1+r)(2m(2+r) - (6eN(1+r)+v)\theta))}{(2m(2+r)-7eN(1+r)\theta)}, \\ N_E^{B^*} = N_E^{S^*} = \frac{N(1+r)(v - eN(1+r))\theta}{2m(2+r)-7eN(1+r)\theta}, \\ \pi_I^* = \frac{2N(1+r)(2m(2+r) - (6eN(1+r)+v)\theta)(m(2+r)(eN(1+r)+v) - eN(1+r)(3eN(1+r)+4v)\theta)}{(2m(2+r)-7eN(1+r)\theta)^2}, \text{ and} \end{aligned}$$

$$\pi_E^* = \frac{(N(1+r)(v-eN(1+r))^2\theta(m(2+r)-4eN(1+r)\theta))}{(2m(2+r)-7eN(1+r)\theta)^2} - L(\theta(2N+rN)).$$

ii If $\min\left(\frac{4m(2+r)}{16eN(1+r)+3v+\sqrt{16e^2N^2(1+r)^2+9v^2}}, 1\right) < \theta \leq 1$, then $p_I^* = (2m(2+r))/3\theta - 2eN(1+r)$, $p_E^* = \frac{m^2(2+r)(3+r)-3meN(1+r)(8+3r)\theta+12eN(1+r)(2eN(1+r))\theta^2}{3\theta(m(2+r)-4eN(1+r)\theta)}$,

$$w_E^* = \frac{(m^2(2+r)-meN(4-r)(1+r)\theta)}{3\theta(m(2+r)-4eN(1+r)\theta)}, w_I^* = 0,$$

$$N_I^B = N_I^S = \frac{2N(1+r)(m(2+r)-3eN(1+r)\theta)}{3(m(2+r)-4eN(1+r)\theta)}, N_E^B = N_E^S = \frac{N(1+r)(m(2+r)-6eN(1+r)\theta)}{3(m(2+r)-4eN(1+r)\theta)},$$

$$\pi_I^*(\theta) = \frac{4N(1+r)(m(2+r)-3eN(1+r)\theta)^2}{9\theta(m(2+r)-4eN(1+r)\theta)}, \text{ and}$$

$$\pi_E^*(\theta) = \frac{N(1+r)(m(2+r)-6eN(1+r)\theta)^2}{9\theta(m(2+r)-4eN(1+r)\theta)} - L(\theta(2N+rN)).$$

We again confirm the entrant's optimal choice of θ to be no more than

$$\min\left(\frac{4m(2+r)}{16eN(1+r)+3v+\sqrt{16e^2N^2(1+r)^2+9v^2}}, 1\right).$$

That is, it is in the best interest of the entrant not to trigger the incumbent's competitive response. Therefore, the entrant will select θ to maximize $\pi_E^*(\theta) = \frac{(N(1+r)(v-eN(1+r))^2\theta(m(2+r)-4eN(1+r)\theta))}{(2m(2+r)-7eN(1+r)\theta)^2} - k(\theta(2N+rN))^2$ under the constraint that $\theta \leq \min\left(\frac{4m(2+r)}{16eN(1+r)+3v+\sqrt{16e^2N^2(1+r)^2+9v^2}}, 1\right)$. Because the first term of $\pi_E^*(\theta)$ increases with θ and the second term of $\pi_E^*(\theta)$ decreases with θ , we can conclude that there exists a k^* so that the two scenarios in Proposition 2.2.2 hold qualitatively. That is,

1) when $k \geq k^*$, the optimal θ^* is the solution to $\frac{\partial \pi_E^*(\theta)}{\partial \theta} = 0$,

and $\theta^* \leq \min\left(\frac{4m(2+r)}{16eN(1+r)+3v+\sqrt{16e^2N^2(1+r)^2+9v^2}}, 1\right)$;

and

2) when $k < k^*$, $\theta^* = \min\left(\frac{(4m(2+r))}{16eN(1+r)+3v+\sqrt{16e^2N^2(1+r)^2+9v^2}}, 1\right)$.

PROOF OF PROPOSITION 2.3.6

Following the same procedure as that in the proof of Proposition 2.3.5, we can prove that if there exists an equilibrium where both the entrant and the incumbent have a positive demand, Lemma

2.2.1 holds in this extension when e is very large—that is, the incumbent and the entrant will always choose their prices and wages so that $a^* = (1 + r)c^*$ when $e > \frac{m}{2N\theta}$. Then, we show that such an equilibrium cannot be sustained because the incumbent will always deviate to drive the entrant out of the market. This is because when e is very large, the highest possible profit that an incumbent can obtain if giving up xN ($x \leq \theta$ and $xN \geq 1$) local buyers and xN matched service providers to the entrant is lower than the profit an incumbent can obtain by deviating to prevent these local buyers and service providers from switching to the entrant.

Specifically, if the incumbent gives up xN local buyers and xN matched service providers to the entrant, a buyer's utility from using the incumbent platform is $e(1 - x + r)N + v - p_I$ and a service provider's utility from using the incumbent platform is $e(1 - x + r)N + w_I$. Thus, the highest possible profit the incumbent can obtain is $(2e(1 - x + r)N + v)(1 - x + r)N$, by charging $p_I = e(1 - x + r)N + v$ and $w_I = -e(1 - x + r)N$. Its actual profit can be lower because the incumbent may have to lower its margin to compete with the entrant, so that its $(1 - x + r)N$ buyers and $(1 - x + r)N$ service providers will not switch to the entrant.

In contrast, if the incumbent deviates to charge $p_I = e(1 - x + r)N + w_E$ and $w_I = w_E - e(1 - x + r)N$, it can prevent these xN local buyers and xN service providers from switching to the entrant. This can be shown as follows. The highest possible utility of a buyer switching to the entrant platform, assuming xN service providers will switch together, is $exN + v - p_E$, which is not higher than the utility of remaining with the incumbent, which is $e(1 + r)N + v - p_I$, given that $p_I = e(1 - x + r)N + w_E e(1 - x + r)N + p_E$. Similarly, the highest possible utility of a buyer switching to the entrant platform, assuming xN buyers will switch together, is $exN + w_E$, which is not higher than the utility of remaining with the incumbent, which is $e(1 + r)N + w_I$, given that $w_I = w_E - e(1 - x + r)N$. In this case, the profit the incumbent can obtain is $2e(1 - x + r)N(1 + r)N$, which is higher than the highest possible profit that the incumbent can obtain once it gives up these local buyers and service providers to the entrant (which is $(2e(1 - x + r)N + v)(1 - x + r)N$) when

$$e > \frac{v}{2N_x}.$$

Thus, we find that the condition under which the equilibrium in which the entrant has positive demand does not exist and the incumbent will take the entire market: $e > \max\left(\frac{m}{2N\theta}, \frac{v}{2N_x}\right)$. As $xN \geq 1$ and $x \leq \theta$, the sufficient condition is $e > \max\left(\frac{m}{2}, \frac{v}{2}\right)$.

PROOF OF PROPOSITION 2.3.7

The demand functions are now changed to

$$N_I^S = \left(1 - \frac{c^*}{m_s}\theta + r\right)N, \quad (\text{A.21})$$

$$N_E^S = \frac{c^*}{m_s}\theta N, \quad (\text{A.22})$$

$$N_I^B = \left(1 - \frac{a^*}{m_b}\theta + r\right)N, \quad (\text{A.23})$$

$$N_E^B = \frac{a^*}{m_b}\theta N. \quad (\text{A.24})$$

Then, we follow the same procedure as that in our main analysis to derive the following two main propositions given $m > \frac{3v}{5}$:

Proposition 2.2.1F: Given that the entrant's choice of θ , the optimal prices, number of buyers and service providers, and profits are as follows:

$$\begin{aligned} \text{i. If } 0 \leq \theta \leq \min\left(\frac{2(m_b+m_s+m_br)}{3v}, 1\right), p_I^* = v, p_E^* = \frac{(m_b+2m_s+m_br)v}{2(m_b+m_s+m_br)}, w_E^* = \frac{m_s v}{2(m_b+m_s+m_br)}, N_I^{B*} = \\ N_I^{S*} = \frac{N(1+r)}{2} \left(2 - \frac{\theta v}{(m_b+m_s+m_br)}\right), N_E^{B*} = N_E^{S*} = \frac{N(1+r)\theta v}{2(m_b+m_s+m_br)}, \pi_I^*(\theta) = N(1+r) \frac{v}{2} \left(2 - \theta \frac{v}{(m_b+m_s+m_br)}\right), \text{ and } \pi_E^*(\theta) = \frac{N(1+r)\theta v^2}{4(m_b+m_s+m_br)} - L(\theta(2N + rN)). \end{aligned}$$

2. If $\min\left(\frac{2(m_b+m_s+m_br)}{3v}, 1\right) < \theta \leq 1$, then $p_I^* = \frac{2(m_b+m_s+m_br)}{3\theta}$, $w_I^* = 0$, $p_E^* = \frac{m_b+2m_s+m_br}{3\theta}$, $w_E^* = \frac{m_s}{3\theta}$, $N_I^{B^*} = N_I^{\delta^*} = \frac{2N(1+r)}{3}$, $N_E^{B^*} = N_E^{\delta^*} = \frac{N(1+r)}{3}$, $\pi_I^*(\theta) = 4N(1+r)\frac{m_b+m_s+m_br}{9\theta}$, and $\pi_E^*(\theta) = N(1+r)\frac{m_b+m_s+m_br}{9\theta} - L(\theta(2N+rN))$.

We again confirm the entrant's optimal choice of θ to be no more than $\min\left(\frac{2(m_b+m_s+m_br)}{3v}, 1\right)$.

That is, it is in the best interest of the entrant not to trigger the incumbent's competitive response.

Endogenizing θ , we obtain the following proposition.

Proposition 2.2.2F: The optimal θ^* depends on the value of k :

- i If $k \geq \max\left(\frac{3(1+r)v^3}{16N(2+r)^2(m_b+m_s+m_br)^2}, \frac{(1+r)v^2}{8N(2+r)^2(m_b+m_s+m_br)}\right)$, then $\theta^* = \frac{(1+r)v^2}{8kN(2+r)^2(m_b+m_s+m_br)}$, which decreases with m_b , m_s , and r . The entrant's profit is $\frac{(1+r)^2v^4}{64k(2+r)^2(m_b+m_s+m_br)^2}$ and the incumbent's profit is $\frac{N(1+r)v}{2}\left(2 - \frac{(1+r)v^3}{8kN(2+r)^2(m_b+m_s+m_br)^2}\right)$.

- ii If $0 \leq k < \max\left(\frac{3(1+r)v^3}{16N(2+r)^2(m_b+m_s+m_br)^2}, \frac{(1+r)v^2}{8N(2+r)^2(m_b+m_s+m_br)}\right)$, then $\theta^* = \min\left(\frac{2(m_b+m_s+m_br)}{3v}, 1\right)$, which weakly increases with m_b , m_s , and r . When $\frac{2(m_b+m_s+m_br)}{3v} < 1$, the entrant's profit is $\frac{N(1+r)v}{6} - \frac{4kN^2(2+r)^2(m_b+m_s+m_br)^2}{9v^2}$ and the incumbent's profit is $\frac{2Nv(1+r)}{3}$. When $\frac{2(m_b+m_s+m_br)}{3v} \geq 1$, the entrant's profit is $\frac{N(1+r)v^2 - 4kN^2(2+r)^2(m_b+m_s+m_br)}{4(m_b+m_s+m_br)}$ and the incumbent's profit is $\frac{N(1+r)v}{2}\left(2 - \frac{v}{m_b+m_s+m_br}\right)$.

Comparative statics suggest that $\frac{\partial \pi_E^*}{\partial m_b} < \frac{\partial \pi_E^*}{\partial m_s} < 0$ and $\frac{\partial \pi_I^*}{\partial m_b} \geq \frac{\partial \pi_I^*}{\partial m_s} \geq 0$.

B

Appendix for Chapter 3

B.1 APPENDIX B-1

PROOF OF THEOREM 3.1.1)

Define $v = -\log(m(d, \theta))$. For a given θ and IID $d_i \sim P(\omega)$, v_i becomes IID samples of random variable v . If $Ev_i^2 < \infty$, for a large number of data points we can use central limit theorem and

hence,

$$\frac{1}{n} \sum_{i=1}^n v_i = E_P(v) + o\left(\frac{C_1}{\sqrt{n}}\right) \mathcal{N}(0,1)$$

Where C_1 is a function of $\text{var}(v)$. Note that

$$\begin{aligned} E_P(v) &= -E_P(\log(m(d, \theta))) = -E_P(\log(P)) + E_P(\log(P)) - E_P(\log(m(d, \theta))) = \\ &= -E_P(\log(P(d))) + E_P \log\left(\frac{P(d)}{m(d, \theta)}\right) = H(P) + KL(P||m(d, \theta)). \end{aligned}$$

Therefore,

$$-\frac{1}{n} \sum_{i=1}^n \log(m(d_i, \theta)) = H(P) + KL(P||m(d, \theta)) + O\left(\frac{C_1}{\sqrt{n}}\right) \mathcal{N}(0,1)$$

Q.E.D.

PROOF OF PROPOSITION 3.2.1)

From our assumptions in the paper and the asymptotic efficiency of MLE [23], we know that

$\lim_{n \rightarrow \infty} m(d, \theta_n) = P(d)$ where $\theta_n = \arg \max_{\theta} \sum_{i=1}^n \log(m(d_i, \theta))$. Hence, for $E|\log(m(d_i, \theta_n))| < \infty$

and using the strong law of large number we have

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \sum_{i=1}^n \log(m(d_i, \theta_n)) = H(P) + KL(P||m(d, \theta_{\infty})) = H(P) + KL(P||P) = H(P)$$

Therefore, a model that has been trained on $D_{\infty,0}$ should reach the loss value $H(P_0)$. Assume

$d^{(0)} \sim P_0(d)$ and $d^{(t)} \sim P_t(d)$. Consider a model that has been trained on a dataset from time t

$(D_{\infty,t})$ and been tested on a dataset from time 0, $D_{\infty,0}$. In this case, $\lim_{n \rightarrow \infty} m(d^{(t)}, \theta_n) = P_t(d)$

where $\theta_{n,t} = \arg \max_{\theta} \sum_{i=1}^n \log(m(d_i^{(t)}, \theta))$

The test loss value for this model is

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \sum_{i=1}^n \log \left(m \left(d_i^{(0)}, \theta_{\infty, t} \right) \right) = H(P(d)) + KL(P(d) || m(d, \theta_{\infty, t})) = H(P_0) + KL(P_0 || P_t)$$

Since both $H(P_0)$ and $KL(P_0 || P_t)$ are non-negative functions of distributions [74, 97], we conclude that the loss value is higher than $H(P_0)$. Therefore, a bounded size dataset should reach the loss value $H(P_0) + KL(P_0 || P_t)$.

Formalizing this argument, we define a neighborhood around $H(P_0)$ with the size $\delta > 0$ and prove that with probability $(1 - \varepsilon)$, any dataset of bounded size reaches a value in the neighborhood.

Mathematically, for large dataset samples $n \gg 1$ and $\delta > 0$, using theorem 3.1.1 we have

$$\begin{aligned} & P \left(\left| -\frac{1}{n} \sum_{i=1}^n \log \left(m \left(d_i^{(0)}, \theta_{n,0} \right) \right) - H(P_0) \right| > \delta \right) = \\ & P \left(\left| KL(P_0 || P_t) + O \left(\frac{1}{\sqrt{n}} \right) \mathcal{N}(0, 1) \right| > \delta \right) = P \left(\left| \mathcal{N} \left(KL(P_0 || P_t), o \left(\frac{1}{\sqrt{n}} \right) \right) \right| > \delta \right) = \\ & = P \left(\mathcal{N} \left(KL(P_0 || P_t) - \delta, o \left(\frac{1}{\sqrt{n}} \right) \right) > 0 \right) + P \left(\mathcal{N} \left(KL(P_0 || P_t) + \delta, o \left(\frac{1}{\sqrt{n}} \right) \right) < 0 \right) \\ & = \underbrace{\Phi \left(\frac{\delta - KL(P_0 || P_t)}{o \left(\frac{1}{\sqrt{n}} \right)} \right)}_{(i)} + \underbrace{\Phi \left(\frac{-\delta - KL(P_0 || P_t)}{o \left(\frac{1}{\sqrt{n}} \right)} \right)}_{(ii)} \end{aligned}$$

Where $\Phi(\cdot)$ is the cumulative distribution function of standard Normal. In above equation, since $\delta > 0$, (i) is bigger than (ii) which means

$$P \left(\left| -\frac{1}{n} \sum_{i=1}^n \log \left(m \left(d_i^{(0)}, \theta_{n,0} \right) \right) - H(P_0) \right| > \delta \right) < 2\Phi \left(\frac{\delta - KL(P_0 || P_t)}{o \left(\frac{1}{\sqrt{n}} \right)} \right)$$

Since for $\delta < D(P_0||P_t)$ the numerator is negative,

$$\lim_{n \rightarrow \infty} \Phi \left(\frac{\delta - KL(P_0||P_t)}{o\left(\frac{1}{\sqrt{n}}\right)} \right) = \Phi(-\infty) = 0$$

Therefore, For any $\varepsilon, \delta > 0, \exists n_0 < \infty$ s.t. $\forall n > n_0$

$$P \left(\left| -\frac{1}{n} \sum_{i=1}^n \log \left(m \left(d_i^{(0)}, \theta_{n,0} \right) \right) - H(P_0) \right| > \delta \right) < 2\Phi \left(\frac{\delta - KL(P_0||P_t)}{o\left(\frac{1}{\sqrt{n}}\right)} \right) < \varepsilon$$

Meaning that a dataset size of $n > n_0$ with probability $1 - \varepsilon$ surpass the performance of infinite dataset size from time t .

Q.E.D.

PROOF OF PROPOSITION 3.2.2)

As explained in the main text, the substitution function $f_n(t_1, t_2)$ measures the gain in substituting a dataset of size n from time t_2 with a dataset of same size from time t_1 . We also proved in theorem 3.2.1 (a) that the substitution function is non-negative. In proving the claims, we have two additional assumptions. First, we assume that the monotonicity result proved in theorem 3.2.1 (b) is valid for all dataset sizes meaning that $f_n(t_1, t_2)$ is monotonic for all n . Second, we assume that $f_1(t_1, t_2) = 1$ for all $t_1, t_2 \in \mathbb{R}^+ \cup 0$. It is intuitive since, in our model, all elements have non-zero probability and hence, one data point carries in expectation same amount of information regardless of when it was sampled.

Now, fixing t_2 , for all $t \in \mathbb{R}^+ \cup 0$ and all $n \in \mathbb{N}$ we have either

- $f_n(t, t_2) < f_\infty(t, t_2)$ which due to monotonicity means that it is increasing, and the claim is proved.

- $f_n(t, t_2) > f_\infty(t, t_2)$ which means that for $t = 0$ and $n < \infty, n > f_n(0, t_2) > f_\infty(0, t_2) \implies 1 > f_\infty(0, t_2) = \infty$ which is a contradiction.
- Or $f_n(t, t_2)$ intersects with $f_\infty(t, t_2)$ in a few points. Suppose t_0 is an intersection point. Since $f_n(t_0, t_2)$ is monotone in n and $f_n(t_0, t_2) = f_\infty(t_0, t_2) = c$, for all $n, f_n(t_0, t_2) = c$. If $c \neq 1$, then $f_1(t_0, t_2) \neq 1$ which is a contradiction and hence $c = 1$. Between the intersection points if $f_\infty(t, t_2) > 1, f_n(t, t_2)$ is increasing since it is monotonic between $f_1(t, t_2) = 1$ and $f_\infty(t, t_2)$ and if $f_\infty(t, t_2) < 1, f_n(t, t_2)$ is decreasing since it is monotonic between $f_1(t, t_2) = 1$ and $f_\infty(t, t_2)$.

In conclusion, $f_n(t_1, t_2)$ is increasing in n if $f_n(t_1, t_2) > 1$ and it is decreasing in n if $f_n(t_1, t_2) < 1$. Hence, the substitution gain/loss increases with n . Because of Proposition 3.3.1 which states that a dataset curates over time has a loss value equal to a dataset of same size (n) that is sampled from the equivalent time, we argue that n is equivalent to the flow of data. Hence, the substitution gain/loss increases with the flow of data.

Q.E.D.

PROOF OF THEOREM 3.2.1)

- This is a direct result of theorem 3.1.1 and proposition 3.2.1.
- Due to monotonic decline of effectiveness over time, $KL(P_0||P_{t_1}) < KL(P_0||P_{t_2})$ for $t_2 > t_1$ and $KL(P_0||P_{t_1}) > KL(P_0||P_{t_2})$ for $t_2 < t_1$.

For sufficiently large number of datapoints, the model $m(d, \theta)$ almost converged to $P(d)$. Therefore, due to continuity and differentiability of the learning curve, we can use Taylor expansion of learning curve's inverse in the neighborhood of $P(d)$.

$$\bar{n}_{D_{n,t}} = EG_0^{-1}(H(P_0) + KL(P_0||m(d, \theta_{n,t}))) \sim$$

$$\begin{aligned}
& G_0^{-1} (H(P_0) + KL(P_0||P_t)) \\
& + E \left[(KL(P_0||m(d, \theta_{n,t})) - KL(P_0||P_t)) \frac{\partial G_0^{-1}(q)}{\partial q} \Big|_{H(P_0)+KL(P_0||P_t)} \right] \\
& = G_0^{-1} (H(P_0) + KL(P_0||P_t)) - E \left[\left(E_{P_0} \log \frac{m(d, \theta_{n,t})}{P_t(d)} \right) \frac{\partial G_0^{-1}(q)}{\partial q} \Big|_{H(P_0)+KL(P_0||P_t)} \right]
\end{aligned}$$

Using Taylor expansion $\log(1+x) \sim x - \frac{x^2}{2} + \frac{x^3}{3} + o(x^4)$, in the neighborhood of $x = 0$.

We do this because we expect $m(d, \theta_{n,t}) \rightarrow P_t(d)$. Using Taylor expansion, we have

$$\begin{aligned}
n_{D_{n,t}} & = G_0^{-1} (H(P_0) + KL(P_0||P_t)) - \\
& E_{P_0} \left(\frac{m(d, \theta_{n,t}) - P_t(d)}{P_t(d)} - \frac{1}{2} \left(\frac{m(d, \theta_{n,t}) - P_t(d)}{P_t(d)} \right)^2 + \frac{1}{3} \left(\frac{m(d, \theta_{n,t}) - P_t(d)}{P_t(d)} \right)^3 \right. \\
& \left. + o \left(\left(\frac{m(d, \theta_{n,t}) - P_t(d)}{P_t(d)} \right)^4 \right) \right) \times \frac{\partial G_0^{-1}(q)}{\partial q} \Big|_{H(P_0)+KL(P_0||P_t)}
\end{aligned}$$

Assuming $m(d, \theta)$ to be a continuous function of θ , we can use theorem 10.1.12 in ²³ (Asymptotic efficiency of MLE) and approximate $m(d, \theta_{n,t})$ with respect to randomization in algorithms and choice of dataset in the training phase. Therefore,

$$m(d, \theta_{n,t}) \sim P_t(d) + \frac{1}{\sqrt{n}} \mathcal{N}(0, v(\theta))$$

Where $v(\theta)$ is the Cramer-Rao lower bound.

$$\begin{aligned}
\bar{n}_{D_{n,t}} & = G_0^{-1} (H(P_0) + KL(P_0||P_t)) \\
& - E \left[E_{P_0} \left(\frac{1}{\sqrt{n}} \mathcal{N} \left(0, \frac{v(\theta)}{P_t(d)} \right) - \frac{1}{2n} \left(\mathcal{N} \left(0, \frac{v(\theta)}{P_t(d)} \right) \right)^2 + \frac{1}{3n\sqrt{n}} \left(\mathcal{N} \left(0, \frac{v(\theta)}{P_t(d)} \right) \right)^3 + o \left(\frac{1}{n^2} \right) \right) \right. \\
& \left. \times \frac{\partial G_0^{-1}(q)}{\partial q} \Big|_{H(P_0)+KL(P_0||P_t)} \right]
\end{aligned}$$

$$\begin{aligned}
&= G_0^{-1} (H(P_0) + KL(P_0||P_t)) + \frac{1}{2n} \left[EE_{P_0} \left(\mathcal{N} \left(0, \frac{v(\theta)}{P_t(d)} \right) \right)^2 + o\left(\frac{1}{n^2}\right) \right] \\
&\quad \times \frac{\partial G_0^{-1}(q)}{\partial q} \Big|_{H(P_0)+KL(P_0||P_t)}
\end{aligned}$$

Since the first and third moment of centered Gaussian distribution is equal to 0.

As a side note, the argument inside the brackets is positive. Since

$$\frac{\partial G_0^{-1}(q)}{\partial q} < 0$$

we conclude

$$\frac{1}{2n} \left[EE_{P_0} \left(\mathcal{N} \left(0, \frac{v(\theta)}{P_t(d)} \right) \right)^2 \right] \frac{\partial G_0^{-1}(q)}{\partial q} \Big|_{H(P_0)+KL(P_0||P_t)} < 0$$

Hence, $\bar{n}_{D_{n,t}}$ is an increasing function in n for sufficiently large n .

Back to the prove, we now take the derivative of $f_n(t_1, t_2)$ with respect to n . For large n we use the following approximation

$$\begin{aligned}
\hat{f}_n(t_1, t_2) &= \frac{G_0^{-1} (H(P_0) + KL(P_0||P_{t_1})) + \frac{1}{2n} \left[EE_{P_0} \left(\mathcal{N} \left(0, \frac{v(\theta)}{P_{t_1}(d)} \right) \right)^2 \right] \frac{\partial G_0^{-1}(q)}{\partial q} \Big|_{H(P_0)+KL(P_0||P_{t_1})}}{G_0^{-1} (H(P_0) + KL(P_0||P_{t_2})) + \frac{1}{2n} \left[EE_{P_0} \left(\mathcal{N} \left(0, \frac{v(\theta)}{P_{t_2}(d)} \right) \right)^2 \right] \frac{\partial G_0^{-1}(q)}{\partial q} \Big|_{H(P_0)+KL(P_0||P_{t_2})}} \\
&= \frac{\bar{n}_{D_{\infty,t_1}} + \frac{1}{2n} \left[EE_{P_0} \left(\mathcal{N} \left(0, \frac{v(\theta)}{P_{t_1}(d)} \right) \right)^2 \right] \frac{\partial G_0^{-1}(q)}{\partial q} \Big|_{H(P_0)+KL(P_0||P_{t_1})}}{\bar{n}_{D_{\infty,t_2}} + \frac{1}{2n} \left[EE_{P_0} \left(\mathcal{N} \left(0, \frac{v(\theta)}{P_{t_2}(d)} \right) \right)^2 \right] \frac{\partial G_0^{-1}(q)}{\partial q} \Big|_{H(P_0)+KL(P_0||P_{t_2})}}
\end{aligned}$$

To show the derivative sign, we focus on the for large n (Omitting $o(\frac{1}{n^3})$)

$$\begin{aligned} \implies num \left(\frac{\partial \hat{f}_n(t_1, t_2)}{\partial n} \right) &\sim \frac{1}{2n^2} \left(\bar{n}_{D_{\infty}, t_1} \left[EE_{P_0} \left(\mathcal{N} \left(0, \frac{v(\theta)}{P_{t_2}(d)} \right) \right) \right]^2 \frac{\partial G_0^{-1}(q)}{\partial q} \Big|_{H(P_0)+D(P_0|P_{t_2})} \right. \\ &\quad \left. - \bar{n}_{D_{\infty}, t_2} \left[EE_{P_0} \left(\mathcal{N} \left(0, \frac{v(\theta)}{P_{t_1}(d)} \right) \right) \right]^2 \frac{\partial G_0^{-1}(q)}{\partial q} \Big|_{H(P_0)+D(P_0|P_{t_1})} \right) \end{aligned}$$

Since the argument in the brackets are not a function of n , we can conclude that for large n , the substitution function $f_n(t_1, t_2)$ is monotonic in n .

Q.E.D.

PROOF OF LEMMA 3.3.1)

Assume dataset $D_{n,t}$ is sampled over time with the density function $\lambda_{t=t_0} = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(t_i = t_0)$.

Considering each sample a random variable, number of times event “a” happens i.e. $\mathbf{1}(d \in a) = 1$ in the dataset is equal to $\sum_{i=1}^n \mathbf{1}_{t_i}(d \in a)$. Therefore, the expected frequency of the event $\{d \in a\}$ is equal to

$$P_{D_{n,t}}(d \in a) = E \left(\frac{\sum_{i=1}^n \mathbf{1}_{t_i}(d \in a)}{n} \right) \underset{\text{Fubini theorem}}{=} \frac{1}{n} \sum_{i=1}^n E(\mathbf{1}_{t_i}(d \in a)) = \frac{1}{n} \sum_{i=1}^n P_{t_i}(d \in a)$$

Integrating the density function λ_t into formulation

$$\begin{aligned} P_{n,[0,t],\lambda_t}(d \in a) &= \frac{1}{n} \sum_{i=1}^n P_{t_i}(d \in a) = \frac{1}{n} \int_0^t \sum_{i=1}^n P_s(d \in a) \mathbf{1}(t_i = s) ds \\ &= \int_0^t P_s(d \in a) \frac{1}{n} \sum_{i=1}^n \mathbf{1}(t_i = s) ds = \int_0^t P_s(d \in a) \lambda_s ds \end{aligned}$$

Q.E.D.

PROOF OF PROPOSITION 3.3.1)

Using lemma 3.3.1, we know that dataset's net distribution is

$$P_{[0,t],\lambda_t}(d \in a) = \int_0^t P_s(d \in a) \lambda_s ds$$

Therefore, training on the dataset of infinite size and test it at time 0, the error will be equal to

$$H(P_0) + KL(P_0 || P_{[0,t],\lambda_t}) = H(P_0) + KL\left(P_0 || \int_0^t P_s(d \in a) \lambda_s ds\right)$$

Since KL-divergence is a convex function [31], we use Jensen inequality to derive an upper bound

$$\begin{aligned} KL\left(P_0 || \int_0^t P_s(d \in a) \lambda_s ds\right) &= KL\left(\int_0^t P_0 \lambda_s ds || \int_0^t P_s(d \in a) \lambda_s ds\right) \\ &= \int_0^t \lambda_s KL(P_0 || P_s) ds \leq \max_{s \in [0,t]} KL(P_0 || P_s) \end{aligned}$$

Besides, we know that KL-divergence is non negative which means

$$KL(P_0 || P_0) = 0 \leq KL\left(P_0 || \int_0^t P_s(d \in a) \lambda_s ds\right) \leq \max_{s \in [0,t]} KL(P_0 || P_s)$$

Since we assumed in this paper that the function $b(t) = D(P_0 || P_t)$ is continuous over time (The change in distribution is gradual and hence, $b(t)$ is continuous) There exist a time $t^* \in [0, t]$ such that

$$KL(P_0 || P_{t^*}) = KL\left(P_0 || \int_0^t P_s(d \in a) \lambda_s ds\right)$$

Therefore

$$H(P_0) + KL(P_0 || P_{t^*}) = H(P_0) + KL\left(P_0 || \int_0^t P_s(d \in a) \lambda_s ds\right)$$

This means that P_{t^*} generate the same loss value as $P_{[0,t],\lambda_t}$.

Q.E.D.

PROOF OF PROPOSITION 3.3.2)

Without loss of generality we focus our attention to the case of monotonic decline in the value of data. Increase in the substitution gain or loss means that the substitution gain becomes sharper as the dataset size increase due to an increase in the flow of data i.e. for all $t > t_2 > t_1 > 0$, $\alpha \geq 1$, and for the flow rates $\psi_H(t) = \alpha\psi_L(t) > 0$

$$f_{n_H}(t_1, t_2) > f_{n_L}(t_1, t_2)$$

where

$$n_H = \int_0^t \psi_H(t) dt$$

$$n_L = \int_0^t \psi_L(t) dt$$

$$\lambda_H(t) = \frac{\psi_H(t)}{n_H} = \frac{\alpha\psi_L(t)}{\alpha n_L} = \frac{\psi_L(t)}{n_L} = \lambda_L(t)$$

Since $\lambda_L(t) = \lambda_H(t)$, according to Lemma 3.3.1, the net distribution for datasets created from both flow rates ψ_H and ψ_L are identical. Therefore, according to Proposition 3.3.1, both distributions have identical equivalent times. Lets call the equivalent time for these datasets $t_2 > 0$.

Remember the condition for a successful off-loading iteration from the equivalent time t^* to t^{**} is

$$f_{n-n_0}(t^{**}, t^*) > \frac{n}{n-n_0}$$

In that case for all $t_1, t_2 > 0$ such that offloading for n_L is feasible, i.e.

$$f_{n_L}(t_1, t_2) \geq \frac{n_L}{n_L - \int_{t_2}^t \psi_L(t) dt}$$

we have

$$f_{n_H}(t_1, t_2) > f_{n_L}(t_1, t_2) \geq \frac{n_L}{n_L - \int_{t_2}^t \psi_L(t) dt} = \frac{\alpha}{\alpha} \times \frac{n_L}{n_L - \int_{t_2}^t \psi_L(t) dt} = \frac{n_H}{n_H - \int_{t_2}^t \psi_H(t) dt}$$

$$\Rightarrow f_{n_H}(t_1, t_2) > \frac{n_H}{n_H - \int_{t_2}^t \psi_H(t) dt}$$

meaning that for all $t_1, t_2 > 0$ such that offloading is possible for the low flow rate $\psi_L(t)$, such offloading is also possible for high flow rate $\psi_H(t)$ and hence, the equivalent time for the high flow rate is weakly closer to the prediction time 0 compared to the equivalent time for low flow rate. And that completes the proof.

Q.E.D.

B.2 APPENDIX B-2

We ran four experiments with different dataset sizes over Reddit data. The up-left Figure shows the effectiveness curve when we trained the model over 25MB of data. Up-right, down-left, and down-right show the curves for 50, 100, 200 MBs, respectively. As can be seen in these graphs, the effectiveness curve is becoming steeper as expected. Meaning that substitution gain will be monotonically increasing in the number of samples. For example, looking at the effectiveness value for day 2920, we can see the effectiveness values of roughly 0.55, 0.5, 0.45, and 0.4 in the 25, 50, 100, and 200 MBs graphs, respectively.

$$f_{25MB}(0, 2920) \sim 1.81$$

$$f_{50MB}(0, 2920) \sim 2.00$$

$$f_{100MB}(0, 2920) \sim 2.22$$

$$f_{200MB}(0, 2920) \sim 2.50$$

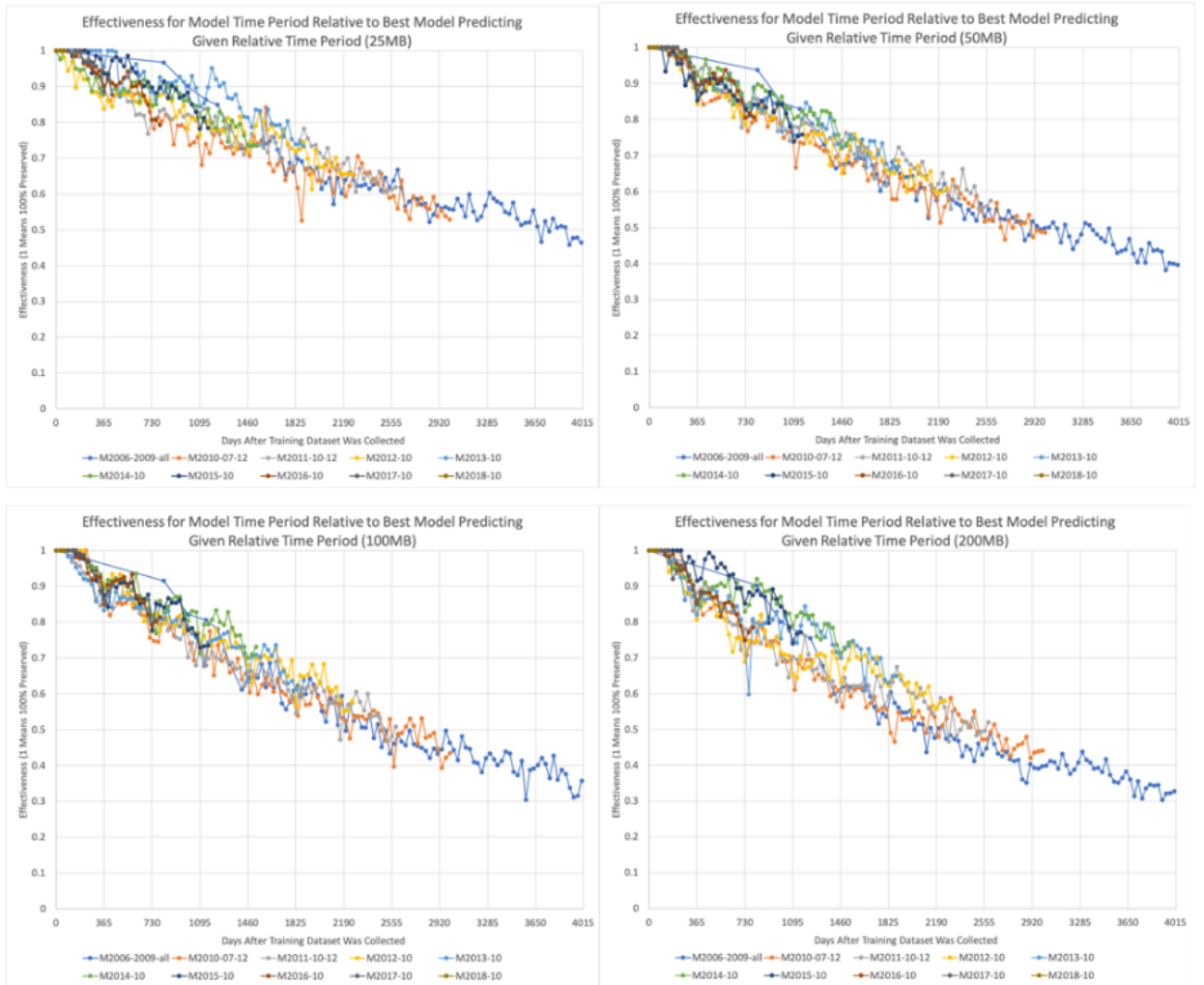


Figure B.1: Effectiveness graphs for various training sizes



Appendix for Chapter 4

WHY TIME-DEPENDENCY?

Data is time-dependent for many reasons, among which we can mention the change in consumers' taste or behavior and, more importantly, innovation in products and services. Innovation plays a key role in time-dependency. It is because it creates new needs that are hard to predict from old data. Besides, due to innovation, older solutions and technologies are gradually being eliminated, which

means that data becomes irrelevant in some cases.

As an example, consider data on kerosene's price and consumption. Before electricity, kerosene was used on lamps to light homes and offices. With the invention of electric bulbs, kerosene lost its lighting purpose, and rare is a time to see it being used for lighting. Consequently, consumers have different price elasticity, which means using kerosene's price data to study macroeconomic questions may not be as relevant as it was many years ago. Besides, the invention of electricity created new consumer needs like refrigerators, air conditioning systems, the Internet, and social media, which changes consumers' behavior.

In kerosene's example, it took decades to witness the change, and the speed of decline for data appeared to be slow. In contrast, in electronics and particularly the cellphone business, change happens at a faster pace. Less than two decades ago, smartphones were introduced, and with them came many innovations in communication methods and devices. Because of these changes, earlier cellphones are becoming less usable and, in some cases, not even compatible with telecom infrastructure.

Similar to our argument on kerosene, using data on old cellphones is not as relevant as using recent cellphone's data. It means the speed of change in the cellphone business is even faster than the speed of change in the energy sector. These observations are market-specific and are affecting every firm within a market. For example, if consumers' taste changes fast, all firms should follow the change quickly or lose the business. Because of it, the speed of change has market level consequences and may change modes of competition.

ADVANTAGES OVER ALTERNATIVE MEASUREMENT METHODS

One way to compare different markets/industries with respect to their speed of change is to create a pool of companies from different industries and study how relevant are the old data to their cur-

rent problems. For example, we can take companies like Uber in the rideshare business and New York Times from news and media to study how consumer's data lose value in time for those specific companies and generalize it to their industry, respectively.

This method has several challenges. To name a few, we can mention selection biases, algorithmic differences, and availability. About selection biases, not everyone is doing business with Uber, and similarly, New York Times readers have a particular taste. Therefore, there are biases in how data is generated, making the comparison difficult, and not generalizable to other companies in the same industry.

As of algorithmic difference, we can immediately tell that New York Times and Uber are in different businesses, and therefore, they need data for different purposes. Besides, perishability changes with the learning curve and scalability of algorithms. Since NYT and Uber use user's data for different tasks and use different algorithms with distinct scaling behavior, perishability measures, as defined in chapter 3 are not directly comparable.

Finally, even if we solve selection biases and algorithmic differences issues, it is not easy, if not impossible, to get users data from companies to do the analysis.

EXPONENTIAL OR POWER-LAW DECAY FUNCTIONS?

As shown in figure 4.2, there is no unique functional form describing the value loss. Sometimes, the wind of change blows strongly, and other times the entire world stops. An example of this can be seen in the politics topic in the years leading to the 2016 U.S. presidential election (Figure 4.3).

Therefore, we consider a functional that takes the difference between training and testing time as an input, and outputs the effective size. In other words, this functional is independent of the testing and the training times and only takes the difference as input. This is a valid assumption because of the measurements provided in chapter 3 on the value loss over entire Reddit data. In that chapter,

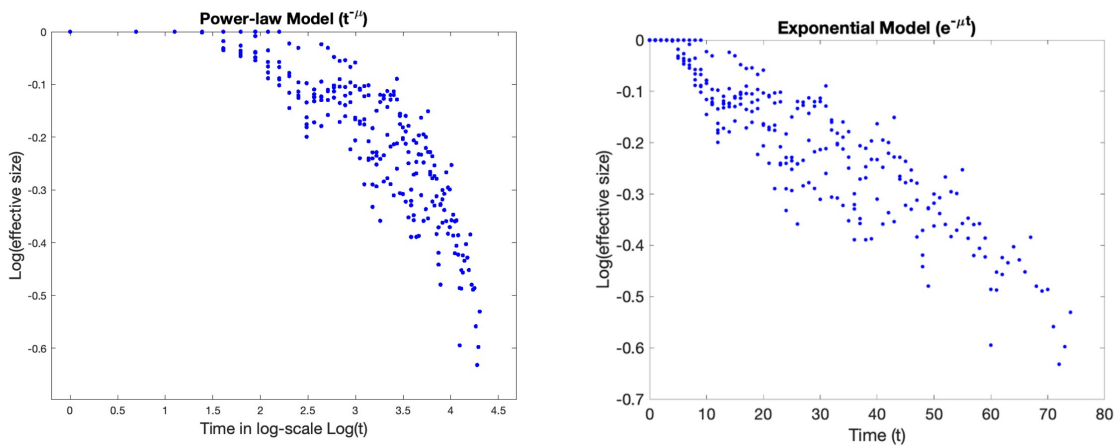


Figure C.1: Power-law (Left plot) and exponential (Right plot) curve fitting. We expect to see a linear representation in either graph

the equivalence graphs showed that text data's value decline is independent of sampling and testing times. In these graphs, equivalence appears as a function of the difference in testing and sampling times.

Besides, for this chapter's purpose, which compares the speed of change between different topics, we need the function to have only one scalar parameter measuring the decay. This function is decreasing and bounded from below by zero, which means that a convex function could be a suitable candidate.

Consequently, we decide to test the exponential and the power-law functional forms since they both satisfy the mentioned conditions. The visual inspection, provided in Figure C.1, suggests that the exponential function describes the value decay better between these functionals.

In Figure C.1 left, we took the log from the time dimension and expected to see a linear functional form if the power-law decay model describes the measurements. As apparent from the graph, it is not linear. In contrast, taking the log from the effectiveness, it is easy to observe a linear relation meaning that the exponential decay model describes the measurements best.

PERISHABILITY AND DATA FLOW:

As previously shown in chapter 3, for highly time-dependent businesses, the data volume does not significantly contribute to the scalability of data-driven AI solutions. To such an extent that there is a finite upper-bound on the effective size of perishable datasets. This bound shows that the data network effect cycle saturates and does not scale beyond the upper-bound even with a dataset of infinite sizes (That has shifted distribution).

In this paper, we show that data flow, to some extent, mitigates the scaling limitations. Expanding the user-base, in the presence of user-wide externalities, or increasing user engagement create the required data flow. Either way, the increase in data flow leads to an increase in the equivalent size.

The proof requires several theorems. First, we mention in chapter 3, that net distribution, in the dataset gathered over a period of time, is solely a function of the underlying data distribution and the sampling density. Hence, multiplying the dataset size by a constant that shows the increase in the flow does not change the net distribution, yet it improves the dataset's equivalent size. Second, we show that in comparison with non-perishable data, it is likelier for the perishable data to offload (refer to chapter 3) the dataset and move the equivalent time (refer to definitions in chapter 3) closer to the prediction time. Getting the equivalent time closer to the prediction time increases the upper-bound limit and, thereby, enhances the effectiveness of data flow in improving quality.

Putting these theorems together, we conclude that data flow derives the scale in time-dependent business. In a nutshell, the following two forces make the information flow the primary driver of value creation for a perishable dataset:

- Pushing the equivalent time closer to prediction time through off-loading, which makes historical data less critical.
- Increasing the upper-bound of equivalent size for equivalent times closer to the prediction

time, which makes the flow the main scalability driver.

TERMINOLOGY AND DEFINITIONS

Definitions 1-4: For a dataset $(D_{n,[0,t],\lambda_t})$ of size n that has been sampled since t periods prior to prediction time with the density function λ_t (Where for $\forall s \in [0, t], \lambda_s \in [0, 1]$ and $\int_0^t \lambda_s ds = 1$) we define following:

1. Equivalent time $t^* \in [0, t]$ is the time that an IID dataset of size n from the time t^* (D_{n,t^*}) produces the loss value equal to the loss value of $D_{n,[0,t],\lambda_t}$.
2. Equivalent size $\bar{n}_{D_{n,[0,t],\lambda_t}}$ is the size (E) in figure 4.1. If the dataset is sampled only at time t ($D_{n,t}$) we use the notation $\bar{n}_{D_{n,t}}$.
3. Effectiveness $E_{D_{n,[0,t],\lambda_t}} = \frac{\bar{n}_{D_{n,[0,t],\lambda_t}}}{n}$ or alternatively $E_{D_{n,t^*}} = \frac{\bar{n}_{D_{n,t^*}}}{n}$ is the ratio of $\frac{\text{size}(E)}{\text{size}(A)}$ in figure 4.1.
4. Substitution function $f_n(t_1, t_2) = \frac{\bar{n}_{D_{n,t_1}}}{\bar{n}_{D_{n,t_2}}}$ shows the gain in equivalent size when we substitute data from time t_2 with a dataset from time t_1 .

Assumption 1: The equivalent size $\bar{n}_{D_{n,t}}$ is monotonically decreasing over time.

Definition 5) Datasets from topic/business H is more perishable comparing to datasets from topic/business L if, fixing the sampling function λ_t and the size for both datasets, we have

$$\bar{n}_{H_n,t_1} - \bar{n}_{H_n,t_2} > \bar{n}_{L_n,t_1} - \bar{n}_{L_n,t_2}$$

For all $t_1, t_2 \in [0, t]$ and $t_1 < t_2$.

Theorem C.o.1. *Highly time-dependent datasets have equivalent time closer to the prediction time than less perishable datasets.*

Proof: We focus on the off-loading mechanism and how it is reasonable for a highly perishable dataset to off-load more often than a less perishable dataset. To prove the theorem, we first prove that highly perishable data has a sharper substitution curve meaning $f_n^H(t_1, t_2) > f_n^L(t_1, t_2)$ where H,L means high and low perishability. Then, we use this inequality to prove that every off-loading iteration for a low perishable dataset is also an off-loading iteration for a high perishable dataset. Therefore, we conclude that high perishable data has equivalent time closer to prediction time.

Definition 5 states that H is more perishable than L if

$$\bar{n}_{H_n, t_1} - \bar{n}_{H_n, t_2} > \bar{n}_{L_n, t_1} - \bar{n}_{L_n, t_2} \quad (\text{C.1})$$

Since $\bar{n}_{L_n, 0} = \bar{n}_{H_n, 0} = n$, we alternatively have $\bar{n}_{L_n, t} > \bar{n}_{H_n, t}$ or $E_{L_n, t} > E_{H_n, t}$. In other words, data from low perishable datasets remain effective for a longer time. Therefore, we have

$$\frac{1}{\bar{n}_{H_n, t_2}} > \frac{1}{\bar{n}_{L_n, t_2}}$$

Multiplying above inequality to inequality (C.1) we have

$$\frac{\bar{n}_{H_n, t_1} - \bar{n}_{H_n, t_2}}{\bar{n}_{H_n, t_2}} > \frac{\bar{n}_{L_n, t_1} - \bar{n}_{L_n, t_2}}{\bar{n}_{L_n, t_2}} \Leftrightarrow \frac{\bar{n}_{H_n, t_1}}{\bar{n}_{H_n, t_2}} - 1 > \frac{\bar{n}_{L_n, t_1}}{\bar{n}_{L_n, t_2}} - 1 \Leftrightarrow f_n^H(t_1, t_2) > f_n^L(t_1, t_2)$$

Which proves the first step. For the second step, consider a highly perishable dataset $H_{n, [0, t], \lambda_t}$ with identical sampling density function and equal size to a less perishable dataset $L_{n, [0, t], \lambda_t}$. The condition for a successful off-loading iteration from the equivalent time t^* to t^{**} is

$$f_{n-n_0}(t^{**}, t^*) > \frac{n}{n - n_0}$$

Since we assumed both $H_{n, [0, t], \lambda_t}$ and $L_{n, [0, t], \lambda_t}$ have identical sizes and sampling density functions,

deleting identical period's data from both datasets still makes them have equal sizes. Consequently, for $t_1 < t_2$, any off-loading iteration that changes $L_{n,[0,t_2],\lambda_t}$ to $L_{n-n_0,[0,t_1],\lambda_t}$ is also an iteration for $H_{n,[0,t_2],\lambda_t}$ to $H_{n-n_0,[0,t_1],\lambda_t}$ since:

$$f_n^H(t_1, t_2) > f_n^L(t_1, t_2) \geq \frac{n}{n - n_0}$$

And that completes the proof. Q.E.D.

Theorem C.o.2. *The upper-bound on the equivalent size decreases in t . The closer the equivalent time to the prediction time, the larger the upper-bound on equivalent size.*

Proof: The upper-bound is equal to $\bar{n}_{D_{\infty,t}}$ since $n_{D_{n,t}}$ is increasing in the number of data points. Assuming a monotonic decline in the equivalent size means that this upper bound is indeed decreasing in time. Q.E.D.

Bibliography

- [1] Abrahamson, E. & Rosenkopf, L. (1997). Social network effects on the extent of innovation diffusion: A computer simulation. *Organization science*, 8(3), 289–309.
- [2] Abrardi, L., Cambini, C., & Rondi, L. (2019). The economics of artificial intelligence: A survey. *Robert Schuman Centre for Advanced Studies Research Paper No. RSCAS*, 58.
- [3] Adner, R., Chen, J., & Zhu, F. (2020). Frenemies in platform markets: Heterogeneous profit foci as drivers of compatibility decisions. *Management Science*, 66(6), 2432–2451.
- [4] Afuah, A. (2013). Are network effects really all about size? the role of structure and conduct. *Strategic Management Journal*, 34(3), 257–273.
- [5] Aghion, P., Jones, B. F., & Jones, C. I. (2019). *9. Artificial Intelligence and Economic Growth*. University of Chicago Press.
- [6] Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press.
- [7] Agrawal, A., Gans, J., & Goldfarb, A. (2019). Economic policy for artificial intelligence. *Innovation Policy and the Economy*, 19(1), 139–159.
- [8] Anderson, S. P., Foros, Ø., & Kind, H. J. (2019). The importance of consumer multihoming (joint purchases) for market performance: Mergers and entry in media markets. *Journal of Economics & Management Strategy*, 28(1), 125–137.
- [9] Anderson Jr, E. G., Parker, G. G., & Tan, B. (2014). Platform performance investment in the presence of network externalities. *Information Systems Research*, 25(1), 152–172.
- [10] Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management science*, 57(8), 1485–1509.
- [11] Arnold, R., Marcus, J. S., Petropoulos, G., & Schneider, A. (2018). *Is data the new oil? Diminishing returns to scale*. Calgary: International Telecommunications Society (ITS).

- [12] Babić Rosario, A., Sotgiu, F., De Valck, K., & Bijmolt, T. H. (2016). The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *Journal of Marketing Research*, 53(3), 297–318.
- [13] Bajari, P., Chernozhukov, V., Hortaçsu, A., & Suzuki, J. (2018). *The impact of big data on firm performance: An empirical investigation*. Technical report, National Bureau of Economic Research.
- [14] Baldwin, R. (2019). *The globotics upheaval: Globalization, robotics, and the future of work*. Oxford University Press.
- [15] Banerji, A. & Dutta, B. (2005). *Local network externalities and market segmentation*. Technical report.
- [16] Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of management*, 17(1), 99–120.
- [17] Begenau, J., Farboodi, M., & Veldkamp, L. (2018). Big data in finance and the growth of large firms. *Journal of Monetary Economics*, 97, 71–87.
- [18] Bergemann, D., Bonatti, A., & Gan, T. (2020). *The economics of social data*. Cowles Foundation discussion paper.
- [19] Brynjolfsson, E., Mitchell, T., & Rock, D. (2018). What can machines learn, and what does it mean for occupations and the economy? In *AEA Papers and Proceedings*, volume 108 (pp. 43–47).
- [20] Campbell, A. (2013). Word-of-mouth communication and percolation in social networks. *American Economic Review*, 103(6), 2466–98.
- [21] Carriere-Swallow, M. Y. & Haksar, M. V. (2019). *The economics and implications of data: an integrated perspective*. International Monetary Fund.
- [22] Casadesus-Masanell, R. & Hałaburda, H. (2014). When does a platform create value by limiting choice? *Journal of Economics & Management Strategy*, 23(2), 259–293.
- [23] Casella, G. & Berger, R. L. (2021). *Statistical inference*. Cengage Learning.
- [24] Cennamo, C. & Santalo, J. (2013). Platform competition: Strategic trade-offs in platform markets. *Strategic management journal*, 34(11), 1331–1350.
- [25] Chandler, A. D. & Hikino, T. (1990). *Scale and scope: The dynamics of industrial capitalism*. Harvard University Press.
- [26] Chen, J., Fan, M., & Li, M. (2016). Advertising versus brokerage model for online trading platforms. *Mis Quarterly*, 40(3), 575–596.

- [27] Chen, J. & Guo, Z. (2022). New-media advertising and retail platform openness. *MIS Quarterly*, 46(1), 431.
- [28] Chen, Y. & Xie, J. (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management science*, 54(3), 477–491.
- [29] Chiou, L. & Tucker, C. (2017). *Search engines and data retention: Implications for privacy and antitrust*. Technical report, National Bureau of Economic Research.
- [30] Claussen, J., Peukert, C., & Sen, A. (2021). The editor and the algorithm: Returns to data and externalities in online news. *Available at SSRN 3479854*.
- [31] Cockburn, I. M., Henderson, R., & Stern, S. (2019). 4. *The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis*. University of Chicago Press.
- [32] Corts, K. S. & Lederman, M. (2009). Software exclusivity and the scope of indirect network effects in the us home video game market. *international Journal of industrial Organization*, 27(2), 121–136.
- [33] Cowgill, B. & Tucker, C. E. (2020). Algorithmic fairness and economics. *Columbia Business School Research Paper*.
- [34] Crémer, J., de Montjoye, Y.-A., & Schweitzer, H. (2019). Competition policy for the digital era. *Report for the European Commission*.
- [35] De Corniere, A. & Taylor, G. (2020). *Data and competition: a general framework with applications to mergers, market structure, and privacy policy*. CEPR Discussion Paper No. DP14446.
- [36] Debruyne, M., Moenaertb, R., Griffinc, A., Hartd, S., Hultinke, E. J., & Robben, H. (2002). The impact of new product launch strategies on competitive reaction in industrial markets. *Journal of Product Innovation Management: An International Publication of The Product Development & Management Association*, 19(2), 159–170.
- [37] Dhar, V. & Chang, E. A. (2009). Does chatter matter? the impact of user-generated content on music sales. *Journal of Interactive marketing*, 23(4), 300–307.
- [38] Economides, N. & Katsamakas, E. (2006). Two-sided competition of proprietary vs. open source technology platforms and the implications for the software industry. *Management science*, 52(7), 1057–1071.
- [39] Esteves, R. B. & Resende, J. (2013). Competitive targeted advertising with price discrimination.
- [40] Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., & Auli, M. (2019). Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*.

- [41] Farboodi, M., Mihet, R., Philippon, T., & Veldkamp, L. (2019). Big data and firm dynamics. In *AEA papers and proceedings*, volume 109 (pp. 38–42).
- [42] Farboodi, M. & Veldkamp, L. (2021). *A growth model of the data economy*. Technical report, National Bureau of Economic Research.
- [43] Fudenberg, D. & Tirole, J. (1984). The fat-cat effect, the puppy-dog ploy, and the lean and hungry look. *The American Economic Review*, 74(2), 361–366.
- [44] Furman, J., Coyle, D., Fletcher, A., McAuley, D., & Marsden, P. (2019). Unlocking digital competition: Report of the digital competition expert panel. *UK government publication, HM Treasury*.
- [45] Furman, J. & Seamans, R. (2019). Ai and the economy. *Innovation policy and the economy*, 19(1), 161–191.
- [46] Galeotti, A. & Goyal, S. (2009). Influencing the influencers: a theory of strategic diffusion. *The RAND Journal of Economics*, 40(3), 509–532.
- [47] Geroski, P. A. (1995). What do we know about entry? *International journal of industrial organization*, 13(4), 421–440.
- [48] Ghani, R., Probst, K., Liu, Y., Krema, M., & Fano, A. (2006). Text mining for product attribute extraction. *ACM SIGKDD Explorations Newsletter*, 8(1), 41–48.
- [49] Godes, D. & Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing science*, 23(4), 545–560.
- [50] Goldfarb, A. & Tucker, C. (2019). Digital economics. *Journal of Economic Literature*, 57(1), 3–43.
- [51] GPT-2 (2018-2020). Gpt-2 source code: <https://github.com/openai/gpt-2>. OpenAI.
- [52] Gregory, R. W., Henfridsson, O., Kaganer, E., & Kyriakou, H. (2021). Data network effects: Key conditions, shared data, and the data value duality. *Academy of Management Review*.
- [53] Gruhl, D., Guha, R., Kumar, R., Novak, J., & Tomkins, A. (2005). The predictive power of online chatter. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 78–87).
- [54] Hagi, A. & Wright, J. (2020). Data-enabled learning, network effects and competitive advantage. *working paper*.
- [55] Hann, I.-H., Koh, B., & Niculescu, M. F. (2016). The double-edged sword of backward compatibility: The adoption of multigenerational platforms in the presence of intergenerational services. *Information Systems Research*, 27(1), 112–130.

- [56] Hao, L., Guo, H., & Easley, R. F. (2017). A mobile platform's in-app advertising contract under agency pricing for app sales. *Production and Operations Management*, 26(2), 189–202.
- [57] Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M., Ali, M., Yang, Y., & Zhou, Y. (2017). Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*.
- [58] Holtz, D., Carterette, B., Chandar, P., Nazari, Z., Cramer, H., & Aral, S. (2020). The engagement-diversity connection: Evidence from a field experiment on spotify. In *Proceedings of the 21st ACM Conference on Economics and Computation* (pp. 75–76).
- [59] Hong, Y. & Pavlou, P. A. (2017). On buyer selection of service providers in online outsourcing platforms for it services. *Information Systems Research*, 28(3), 547–562.
- [60] Huang, N., Burtch, G., Gu, B., Hong, Y., Liang, C., Wang, K., Fu, D., & Yang, B. (2019). Motivating user-generated content with performance feedback: Evidence from randomized field experiments. *Management Science*, 65(1), 327–345.
- [61] Huang, P., Ceccagnoli, M., Forman, C., & Wu, D. (2013). Appropriability mechanisms and the platform partnership decision: Evidence from enterprise software. *Management Science*, 59(1), 102–121.
- [62] Hunt, S. D. (1999). *A general theory of competition: Resources, competences, productivity, economic growth*. Sage publications.
- [63] Iansiti, M. & Lakhani, K. R. (2020). *Competing in the age of AI: strategy and leadership when algorithms and networks run the world*. Harvard Business Press.
- [64] Iansiti, M. & Levien, R. (2004). *The keystone advantage: what the new dynamics of business ecosystems mean for strategy, innovation, and sustainability*. Harvard Business Press.
- [65] Ichihashi, S. (2021). The economics of data externalities. *Journal of Economic Theory*, 196, 105316.
- [66] Jiang, B. & Srinivasan, K. (2016). Pricing and persuasive advertising in a differentiated market. *Marketing Letters*, 27(3), 579–588.
- [67] Jones, C. I. & Tonetti, C. (2020). Nonrivalry and the economics of data. *American Economic Review*, 110(9), 2819–58.
- [68] Karakaya, F. & Yannopoulos, P. (2011). Impact of market entrant characteristics on incumbent reactions to market entry. *Journal of Strategic Marketing*, 19(02), 171–185.
- [69] Katz, M. L. & Shapiro, C. (1985). Network externalities, competition, and compatibility. *The American economic review*, 75(3), 424–440.

- [70] King, R. A., Racherla, P., & Bush, V. D. (2014). What we know and don't know about online word-of-mouth: A review and synthesis of the literature. *Journal of interactive marketing*, 28(3), 167–183.
- [71] Koh, T. K. & Fichman, M. (2014). Multihoming users' preferences for two-sided exchange networks. *Mis Quarterly*, 38(4), 977–996.
- [72] Korinek, A. & Stiglitz, J. E. (2019). *14. Artificial Intelligence and Its Implications for Income Distribution and Unemployment*. University of Chicago Press.
- [73] Kuang, L., Huang, N., Hong, Y., & Yan, Z. (2019). Spillover effects of financial incentives on non-incentivized user engagement: Evidence from an online knowledge exchange platform. *Journal of Management Information Systems*, 36(1), 289–320.
- [74] Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1), 79–86.
- [75] Kwark, Y., Chen, J., & Raghunathan, S. (2018). User-generated content and competing firms' product design. *Management Science*, 64(10), 4608–4628.
- [76] Lambrecht, A. & Tucker, C. E. (2015). Can big data protect a firm from competition? Available at SSRN 2705530.
- [77] Leduc, M. V., Jackson, M. O., & Johari, R. (2017). Pricing and referrals in diffusion on networks. *Games and Economic Behavior*, 104, 568–594.
- [78] Lee, G. M., Qiu, L., & Whinston, A. B. (2016). A friend like me: Modeling network formation in a location-based social network. *Journal of Management Information Systems*, 33(4), 1008–1033.
- [79] Li, Z. & Agarwal, A. (2017). Platform integration and demand spillovers in complementary markets: Evidence from facebook's integration of instagram. *Management Science*, 63(10), 3438–3458.
- [80] Liebowitz, S. (2002). Rethinking the networked economy: The true forces driving the digital marketplace. *AMACOM Div. American Management Association, Dallas*.
- [81] Manshadi, V., Misra, S., & Rodilitz, S. (2020). Diffusion in random networks: Impact of degree distribution. *Operations Research*, 68(6), 1722–1741.
- [82] Milgrom, P. R. & Tadelis, S. (2019). *How Artificial Intelligence and Machine Learning Can Impact Market Design*. University of Chicago Press.
- [83] Mishne, G., Glance, N. S., et al. (2006). Predicting movie sales from blogger sentiment. In *AAAI spring symposium: computational approaches to analyzing weblogs* (pp. 155–158).

- [84] Monske, S. (2018). *The Impact of Electronic Word-of-Mouth on Consumers and Firms*. Westfaelische Wilhelms-Universitaet Muenster (Germany).
- [85] Newman, N. (2014). Search, antitrust, and the economics of the control of user data. *Yale J. on Reg.*, 31, 401.
- [86] Niculescu, M. F., Wu, D., & Xu, L. (2018). Strategic intellectual property sharing: Competition on an open technology platform under network effects. *Information Systems Research*, 29(2), 498–519.
- [87] Parker, G., Van Alstyne, M. W., & Jiang, X. (2016). Platform ecosystems: How developers invert the firm. *Boston University Questrom School of Business Research Paper*, (2861574).
- [88] Parker, G. G. & Van Alstyne, M. W. (2005). Two-sided network effects: A theory of information product design. *Management science*, 51(10), 1494–1504.
- [89] Petit, N. (2017). Antitrust and Artificial Intelligence: A Research Agenda. *Journal of European Competition Law & Practice*, 8(6), 361–362.
- [90] Prufer, J. & Schottmuller, C. (2017). Competing with big data. *TILEC Discussion Paper*.
- [91] Qiu, L., Rui, H., & Whinston, A. B. (2014). The impact of social network structures on prediction market accuracy in the presence of insider information. *Journal of Management Information Systems*, 31(1), 145–172.
- [92] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.
- [93] Reimers, I. & Shiller, B. (2018). Welfare implications of proprietary data collection: an application to telematics in auto insurance. *Available at SSRN 3125049*.
- [94] Rubinfeld, D. L. & Gal, M. S. (2017). Access barriers to big data. *Ariz. L. Rev.*, 59, 339.
- [95] Ruiz-Aliseda, F. (2016). When do switching costs make markets more or less competitive? *International journal of industrial organization*, 47, 121–151.
- [96] Schaefer, M., Sapi, G., & Lorincz, S. (2018). The effect of big data on recommendation quality: The example of internet search. *DIW Berlin Discussion Paper*.
- [97] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3), 379–423.
- [98] Sokol, D. D. & Comerford, R. (2015). Antitrust and regulating big data. *Geo. Mason L. Rev.*, 23, 1129.
- [99] Stiglitz, J. E. (1986). Theory of competition, incentives and risk. In *New Developments in the Analysis of Market Structure* (pp. 399–449). Springer.

- [100] Suarez, F. F. (2005). Network effects revisited: The role of strong ties in technology selection. *Academy of Management Journal*, 48(4), 710–720.
- [101] Sundararajan, A. (2006). Local network effects and complex network structure. *Available at SSRN 650501*.
- [102] Teece, D. & Pisano, G. (2003). The dynamic capabilities of firms. In *Handbook on knowledge management* (pp. 195–213). Springer.
- [103] Tellis, G. J., Yin, E., & Niraj, R. (2009). Does quality win? network effects versus quality in high-tech markets. *Journal of Marketing Research*, 46(2), 135–149.
- [104] Thompson, G. L. & Teng, J.-T. (1984). Optimal pricing and advertising policies for new product oligopoly models. *Marketing Science*, 3(2), 148–168.
- [105] Tian, L., Vakharia, A. J., Tan, Y., & Xu, Y. (2018). Marketplace, reseller, or hybrid: Strategic analysis of an emerging e-commerce model. *Production and Operations Management*, 27(8), 1595–1610.
- [106] Tirole, J. (1988). *The theory of industrial organization*. MIT press.
- [107] Tirole, J. (2020). Competition and the industrial challenge for the digital age. *paper for IFS Deaton Review on Inequalities in the Twenty-First Century*.
- [108] Tirole, J. & Rochet, J.-C. (2003). Platform competition in two-sided markets. *Journal of the european economic association*, 1(4), 990–1029.
- [109] Tucker, C. (2008). Identifying formal and informal influence in technology adoption with network externalities. *Management Science*, 54(12), 2024–2038.
- [110] Van Til, H., Van Gorp, N., & Price, K. (2017). Big data and competition. *Ecorys Study for the Dutch Ministry of Economic Affairs, Ecorys, Rotterdam*. <https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2017/06/13/big-data-and-competition/big-data-andcompetition.pdf>.
- [111] Varian, H. (2019). *16. Artificial Intelligence, Economics, and Industrial Organization*. University of Chicago Press.
- [112] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.
- [113] Xu, H. (2018). Is more information better? an economic analysis of group-buying platforms. *Journal of the Association for Information Systems*, 19(11), 1.

- [114] Zhang Chenglong, Jianqing Chen, S. R. (2018). Platform competition in ride sharing economy. *Working paper*.
- [115] Zhu, F. & Iansiti, M. (2012). Entry into platform-based markets. *Strategic Management Journal*, 33(1), 88–106.