OPERATIONAL AND ECONOMIC ANALYSIS OF MARKETPLACES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Zhen Lian

August 2022

OPERATIONAL AND ECONOMIC ANALYSIS OF MARKETPLACES

Zhen Lian, Ph.D.

Cornell University 2022

Technological advances are enabling more businesses to be organized as marketplaces, creating a global economic trend towards platform companies with abundant information and sophisticated technology. A distinct characteristic of marketplace operations is that, decisions are not made and enforced by a central planner, but are rather the outcome of decentralized agents' decisions, which can only be influenced by the marketplace but not directly controlled. This difference fundamentally changes the way firms must organize and manage their operations. It also changes the nature of competitive interactions among firms at an industry level. This thesis analyzes marketplaces by taking a combined view of both their economics and physical operations. Drawing on optimization, stochastic process, microeconomics, and industrial organization theories, it provides insights into both firm-level decisions and industry-level dynamics for modern marketplaces such as ride-hailing, retail and food delivery.

BIOGRAPHICAL SKETCH

Zhen Lian received her Bachelor's degree from Peking University in 2014 and Master's degree from Cornell University in 2016. Since then, she has pursued a Ph.D. in Operations Management at the S.C. Johnson Graduate School of Management, under the supervision of Prof. Garrett van Ryzin and Li Chen. Her research interest is in the economic impact of marketplaces, with a special focus on ride-hailing.

This document is dedicated to my family.

ACKNOWLEDGEMENTS

This thesis would not have been possible without my advisor Garrett van Ryzin. I met Garrett when I knew very little about the meaning of good research; I was also holding a quite naive opinion about academia, industry, and how they compare with each other. Under the supervision of Garrett, I was able to gain first-hand experiences in both, which shaped my taste in research and influenced my choice of career. Constantly amazed by Garrett's talent in solving complicated problems with simple and elegant ideas, I have learned so much from him – from developing a vision for important research, to presenting ideas to both academic and industry audiences, to being detail-oriented about things like properly capitalizing words on presentation slides. I'm sure I will continue to learn from Garrett; he will always be my role model as an advisor, researcher, and colleague.

I have also been extremely lucky to have Li Chen, Nagesh Gavirneni and Peter Frazier on my thesis committee. As my co-advisor, Li has also been giving me valuable advice on research and life; our research collaboration became an integral part of this thesis. Nagesh, who is the very first person I talked to about OM research, always cheers for my success and makes time for me when I need to talk. Peter generously agreed to serve on my committee and provide feedback for my thesis work at a later stage of my Ph.D., which I'm grateful for.

I also want to take this opportunity to thank my friends, who has made the Ph.D. journey really fun. Sébastien Martin and Arthur Delarue, who are also my super creative and talented coauthors, always motivate me to step out of the comfort zone. I can count on Sébastien to provide honest, constructive feedback whenever I need them, and am constantly impressed by Arthur's well-organized working style and great writing skills. I was fortunate to spend three years at

Cornell Tech, during which I shared the amazing view of Manhattan skyline with Amy, Alberto, Yichun, Angela, Mika, Xiaojie, Yuhang, Longqi, and Yin; in Ithaca, Hester, Cathy, Rihuan, Ruyu, Lin, Adeline, Dayoung, Jun, and Xiaoyan, my life there was a lot more enjoyable because of you.

More than half of the thesis was completed during the Covid-19 pandemic, during which millions of people lost their lives, jobs, or loved ones. I am grateful that I had access to good healthcare and funding to work uninterruptedly in this difficult time.

Last but not the least, I want to thank my family. Fan, thanks so much for being unconditionally supportive of my career; we are such a great team and I am so excited about the next chapter of our life together. Mom, dad and grandma, thanks for always trying to understand what I work on despite my poor explanation, and for tolerating my absence. I miss you so much. Cat Pipi, thanks for being so soft and fluffy. If there are typos in this document, it is because Pipi stepped on my keyboard.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Technological advances are enabling more businesses to be organized as marketplaces, creating a global economic trend towards platform companies with abundant information and sophisticated technology. These marketplace companies are highly disruptive. Consider the transportation sector; in major U.S. cities like New York and San Francisco, ride-hailing apps now serve more than five times the number of trips than the traditional taxi market (TNCs 2021 and wschneider 2021). In online retail, the 2020 profit of sellers on Amazon's third-party marketplace alone is estimated to be $25B, exceeding the total profits of airlines worldwide (Gilbert 2021 and IATA 2021). The revenue of the online food delivery marketplaces are tripled in the past five years; online advertising and cloud computing are trillion-dollar marketplaces (Sara Ashley O'Brien, CNN Business 2020, Factors 2021, and Anand 2017).

While marketplaces companies have traditional operational challenges, such as capacity, quality and inventory management, their operations are fundamentally different: decisions are not made and enforced by a central planner, but are rather the *outcome* of decentralized agents' decisions, which can only be *influenced* by the marketplace but not directly controlled. This difference fundamentally changes the way firms must organize and manage their operations. It also changes the nature of competitive interactions among firms at an industry level.

This thesis focuses on some of these challenges uniquely faced by marketplaces, taking a combined view of both their economics and physical operations. Drawing on optimization, stochastic process, microeconomics, and industrial

organization theories, it provides insights into both firm-level decisions and industry-level dynamics of marketplaces.

This thesis is based on the work over the course of the author's Ph.D. (Lian and Van Ryzin 2021; Lian and Ryzin 2020; Lian, Martin, and Ryzin 2021; Chen, Lian, and Yao 2021;) Chapter 2 analyzes the long-run optimal growth and subsidy policy for a two-sided platform such as ride-hailing, shedding light on whether the "race-to-growth" strategy is rational or just meeting the eye of the investors. Chapter 3 investigates the economic impacts of autonomous vehicles (AVs) and human drivers in a ride-hailing market, highlighting the critical roles of dispatch platform and market structure in achieving the potential economic benefits from AVs. Chapter 4 studies the wage equilibrium among gig economy firms who share a pool of independent contractor workers, showing that surprisingly, the equilibrium wage is increasing in the firms' demand rates, and smaller firms enjoy a labor cost advantage over larger firms. Chapter 5 explores the issue of counterfeiting when consumers knowingly purchase fake luxury products, a pressing issue in retail marketplaces like Amazon and Facebook marketplaces.

CHAPTER 2

OPTIMAL GROWTH IN TWO-SIDED MARKETS

2.1 Introduction

How to optimally grow a two-sided market is an important strategic problem in marketplace companies. In particular, all major ride-sharing companies (Uber, Lyft, Didi) have large internal organizations focused on how best to manage growth. This includes the acquisition of new passengers and drivers, encouraging more usage from existing passengers and drivers, passenger and driver pricing and how to balance growth spending between the demand and supply sides of the market. Similar growth problems are faced in two-sided markets in hospitality (Airbnb), food delivery (Door Dash), and casual labor (Task Rabbit, Handy). A fundamental problem in growth planning is managing the tradeoff between the cost of subsidizing growth (through prices or incentives) and the benefit of the scale economies that are achieved through growth. We analyze a stylized version of this problem.

In particular, our model assumes outputs (transactions) are "produced" by two primary inputs: the stock of supply and demand, which can be thought of as the number of adopters on each side of the market. This production is modeled using a homogeneous production function, which could have either increasing or decreasing returns to scale. We assume a monopolist platform controls the price of supply and demand. Supply and demand stock levels (number of adopters) in turn evolve according to a growth model in which the rate of growth is a function of both the surplus each side of the market receives as well as losses due to attrition (finite demand and supply "lifetimes"). The platform can apportion

this surplus between the two sides of the market by changing the price paid to sellers and the price charged to buyers. The platform can also subsidize the market by setting the price of demand lower than the price of supply. Using this model, we analyze how to optimally grow the market through prices and subsidies to maximize the platform's discounted total profits. Because ours is a monopoly model, there is no competitive benefit to growth (such as preemption); rather, the optimal growth policy is determined solely by whether the benefits of getting to scale quickly justify the costs of subsidizing growth.

Our analysis relies on a novel state space reduction in which we transform our two-state market model (with stock of supply and stock of demand as states) to a model with a single scalar "market size" state. This market size has units of dollars and can be interpreted as the level of investment in the market. Indeed, in our model the platform could liquidate the market by charging prices that reduce the stock of supply and demand to zero and generate a one-time profit equal to the market size.

Having a single market size greatly simplifies the analysis of market growth, since it provides a clear ordering of growth paths; in particular, a growth path which has a market size always greater than another growth path corresponds to "faster" growth. How the market size is apportioned between the supply and demand sides then reduces to a scalar "market balance" control variable. This too is conceptually appealing, since it allows us to cleanly distinguish between the total size of the market and how that market size is allocated between supply and demand. Optimal supply and demand balance itself is also a fundamental concern of marketplace growth organizations and our reformulation isolates and highlights this important choice variable.

While derived from our stylized model, this state space reduction is not unrealistic. Intuitively, it corresponds to a case where it is easier for a platform to shift surplus from the supply to the demand side (or vice versa) than it is to grow the total market. That is, the platform can raise the price significantly for buyers (demand) and pay suppliers (supply) significantly more by simply passing the revenues from the demand to the supply side. This will result in a loss of adoption on the demand side and an increase in adoption on the supply side, but requires no net subsidy. Hence, in the limit we can use such abrupt pricing changes to effectively choose the market balance we want at any point in time. However, such transfers of value between supply and demand do not change the total market value; they simply amount to “cashing out" of one side of the market to “buy up" the other side - leaving the total market value unchanged. True size growth requires increasing the total value of the market overall. Our size and balance decomposition concisely reflects this distinction.

We then investigate the behavior of optimal market growth, including the point at which the market becomes self-sustaining (the “critical size") and the long-run stable size of the market (the “saturation size"). These size thresholds provide important information on what size is necessary to achieve in order for a market to operate subsidy-free when markets have increasing returns to scale and the rational limits to market growth when markets have decreasing returns to scale.

We then characterize the optimal balance between supply and demand as a market grows. In particular, we show that optimal market balance at any point in time is uniquely determined by the market size at that point in time, for both increasing and decreasing returns. That is, regardless of the market size growth

path, the optimal market balance choice at each time is determined only by the market size at that time, the structural parameters governing the supply and demand growth process and the platform's discount rate. This balance choice reflects a fundamental trade-off between minimizing loss of size due to attrition and maximizing market output, which increases surplus and adoption. The path of optimal balance as a function of size, however, depends on the relative durability of supply and demand (which has the longer mean lifetime) and whether the market has increasing or decreasing returns to scale. Together, these results provide robust insights into the optimal choice of supply and demand balance as markets grow.

We then analyze how prices should be set to maximize discounted total profits. In general, we show that rapid subsidization is optimal in a range of cases – though not always. In particular, for both increasing and decreasing returns to scale where prices are unconstrained, an impulse of subsidy spending (a "subsidy shock") that takes the market instantaneously from its initial size to its final (steady-state or terminal) size is optimal. When there are constraints on the growth trajectory, we show that in the decreasing returns case, faster growth paths are always better than slower growth paths. For increasing returns, faster paths may not always dominate depending on whether the market potential is above or below a threshold size. However, if the market potential is large enough, a subsidy shock is still optimal.

The fact that faster growth is better closely mirrors the observed growth strategies in many marketplace companies, as noted in the popular business

press [1] [2] [3]. In ride-sharing in particular, heavy subsidization to incentivize rapid growth has been the norm in the industry; ride-sharing firms spent billions of dollars and multiple years subsidizing riders and drivers to build scale in the cities they serve. Despite the skepticism expressed by some in the popular business press about such growth strategies [4], our results suggest that they may indeed be rational.

The organization of the paper is as follows: Section 2.2 is a review of related literature; Section 2.3 is the model setup, consisting of four parts: (1) an output model about how supply and demand generate market outputs; (2) a growth model about how the growth rates of supply and demand react to wage and price at any time; (3) a market size-and-balance reformulation; (4) the optimal growth formulation. Section 2.4 shows our main results, consisting of three parts: (1) the optimal balance of supply and demand; (2) the stationary solution; (3) the optimal growth policies. Section 2.5 gives some numerical examples.

2.2 Related literature

There is an extensive literature in economics on two-sided markets in which how to price the two sides of the market is a central question. In particular Caillaud and Jullien 2003 study competition among intermediaries who perform matchmaking service between two groups of users. The authors characterize

[1] *Uber Aims to Maintain Heavy Spending to Keep Rivals at Bay.* (April 12, 2019). https://www.ft.com/content/8a28ba78-5d09-11e9-9dde-7aedca0a081a

[2] *How A Venture Capitalist Would Look At Uber's Value Today: It's All About The Marketplace*. (May 14, 2019). https://www.forbes.com/sites/mikeghaffary/2019/05/14/how-a-venture-capitalist-would-look-at-ubers-value-today-its-all-about-the-marketplace/

[3] *Blitzscaling*. (April 2016). https://hbr.org/2016/04/blitzscaling

[4] *Uber's Path of Destruction*, (Summer 2019, Vol. 3, No. 2), https://americanaffairsjournal.org/2019/05/ubers-path-of-destruction/

the intermediary's optimal pricing strategy as "divide-and-conquer", i.e., subsidizing the participation of one group of users while recovering the profit loss from the other group. Rochet and Tirole 2003 look at the a two-sided market that exhibits usage externalities (intensive margin) and the platform charges per-transaction fees on both sides of the market (e.g., the credit card industry). Armstrong 2006 analyzes a two-sided market that exhibits membership externalities (extensive margin) for a platform that charges a lump-sum membership fee on both sides of the market (e.g., night clubs and shopping malls). Both papers show the optimal price allocation for a monopoly platform follows a standard Lerner formula with the appropriate reinterpretation of marginal costs, implying that the price on one side of the market is inversely related to its elasticity of demand. Rochet and Tirole 2006 build a unifying model that combines the usage and membership externality. Weyl 2010 extends Rochet and Tirole 2006 by incorporating user heterogeneity on each side of the platform. Hagiu 2004 builds on Caillaud and Jullien 2003 and looks at the commitment issue in a two-sided market with one side of the market arriving earlier than the other side of the market.

However due to their static nature, these models do not consider the process of acquiring users; that is, once prices are set by the platform, users instantly join as long as they gain positive utility. While this formulation is appropriate for analyzing the steady state of two-sided platforms that are already established, our model specifically focuses on the growth process to get to the steady state. Moreover, although subsidy policies are widely discussed in these papers, the driving forces behind subsidies in these works are the asymmetry of price elasticities and cross externalities between the two sides of the market. In contrast, the main trade-offs we analyze involve the time value of subsidies and profits

as markets grow, in particular how to optimally balance the cost of subsidizing early growth and the benefits of scale economies achieved by growth.

Our model is also similar to those in the modern economic growth literature. For example, Solow 1956 explains long-run economic growth by representing the economy as a simple one-good economy and abstracting away from individual decisions (Acemoglu 2009). In the Solow model, the output of the economy is generated by labor and capital following a Cobb-Douglas production function with constant returns to scale, and the economy grows following the "law of motion" in the capital-labor ratio. The neoclassical growth literature builds on the Solow model by endogenizing consumers' consumption decisions by using the representative household's utility maximization problem. Seeking to explain the empirical evidence of balanced growth (Kaldor 1957), this stream of literature puts an emphasis on policies that lead to a constant capital-output ratio and constant output growth rates. In other words, their approach is to identify conditions under which empirically-observed balanced growth paths are optimal (Acemoglu 2009).

In contrast, our paper seeks to provide insights for the optimal growth of profit maximizing two-sided markets like ride-sharing, for which rapid growth and heavy subsidies are the empirical reality. Therefore, we do not constrain our analysis to only balanced growth paths. In contrast to the macro growth literature, we also consider production that exhibits economies of scale, which are essential features of many two-sided markets. Moreover, in the growth literature, the common assumption that the population is constant or grows at an exogenous rate and the fundamental policy objective is per-capita output. Therefore, their optimization problem is inherently a one-dimensional problem. However, in our

context total discounted profit (firm value) is the objective – not profit per unit of supply or demand. Since it is essential to consider the growth of both supply and demand simultaneously, this raises the added methodological challenge of characterizing the optimal policy for a problem with a two-dimensional state (the stock of supply and demand) and a two-dimensional control (the price of supply and demand), for which closed-form solutions are difficult to derive. In this sense, our paper adds to the neoclassical growth literature by providing a stylized and tractable formulation for two-dimensional growth.

Our work is also related to the matching literature. In the context of online labor markets like TaskRabbit and Upwork, the production of successful transactions from supply and demand can also be thought as a matching process between workers and jobs, which is closely connected to the literature on aggregate matching functions (see a survey paper by Petrongolo and Pissarides 2001). In particular, in our analysis we examine two special cases of our homogeneous production function, the Cobb-Douglas function (Section 2.4.1) and the instant matching function (Section 2.4.5). The Cobb-Douglas model is, to close approximation, the aggregate matching function resulting from a two-sided matching process that involves Poisson queueing together with linear search costs (see Stevens 2007). The instant matching function models a friction-less market where successful matching happens instantly.

In the operations management field, there is a burgeoning literature that studies pricing issues in the context of ride-sharing platforms and online two-sided markets (e.g. Cachon, Daniels, and Lobel 2017, Banerjee, Johari, and Riquelme 2015, Castillo, Knoepfle, and Weyl 2017, Gurvich, Lariviere, and Moreno 2019, Taylor 2018, Bai et al. 2018 Hu and Zhou 2017, Bimpikis, Candogan, and Saban

2019). While these papers consider a variety of important problems related to dynamic pricing in two sided markets, they focus on operational pricing in static markets and do not consider the problem of pricing to optimize long-run market growth.

Our paper is also related to the literature on market thickness. Early theoretical work in labor markets commonly assumed increasing or constant returns to scale (e.g. Diamond 1982). However, there is more recent work on returns to scale in the context of innovative marketplaces. Kabra, Belavina, and Girotra 2016 use data from a ride-hailing market and show evidence of increasing returns to scale. Cullen and Farronato 2014 use data from TaskRabbit, an online peer-to-peer freelance marketplace and finds no evidence of increasing returns to scale. Li and Netessine 2019 use data from an online holiday rental platform and find that increased market thickness actually leads to lower matching efficiency, which implies decreasing returns to scale. Nikzad 2017 studies how market thickness and competition influence the market equilibrium in the ride-hailing market.

Lastly, our paper is based on optimal control theory and hence is related to applications of optimal control theory in management science. For example, Vidale and Wolfe 1957 study a firm's optimal advertising expenditure problem when the effect of advertising carries over but diminishes over time. The rate of sales is affected by the advertising effort in two ways: new adoption from the unsold portion of the market due to advertising, and loss from the sold portion of the market over the time. Dhebar and Oren 1986 study a single supplier's pricing decision for a new product that exhibits demand side network externalities, subject to dynamics of market growth. In our model, we consider the optimal growth problem as a profit maximization problem with growth rates of supply

and demand driven by users' adoption and attrition from both sides of the market over time.

2.3 Model Setup

We begin by defining how supply and demand combine to produce output in a two-sided market. We then define how the supply and demand sides of the market grow as a function of the surplus each side receives from the market as well as attrition losses due to finite supply and demand lifetimes. Next, we show how this two-sided model can be reduced to a one-dimensional model in which a scalar market size is the only state variable and market balance (the ratio of supply size to demand size) becomes a control. The reformulation simplifies our analysis of growth and provides useful intuition. Lastly, we define the optimal growth problem for our model.

2.3.1 Supply, demand and production

Let $s(t)$ denote the stock of supply and $d(t)$ denote the stock of demand at time t. It is most natural to think of these stocks as being the number of suppliers and buyers who have joined the platform (adopters), but the measure of supply and demand could be more general. For example, in ride-sharing networks, supply is typically measured in driver-hours and demand is measured as the number of unique app sessions which could convert to a request for a ride.

To keep the terminology generic, we refer to a unit of market output as a *transaction* (e.g., a transaction is a ride in a ride-sharing network). We assume the

total number of transactions generated by a given stock of supply s and demand d follows a homogeneous production function, denoted as $g(s, d)$.

$$g(ks, kd) = k^{\alpha} g(s, d) \tag{2.1}$$

where $g(s, d)$ is homogeneous of degree α.

The factor α represents the *total returns to scale* of the market. If $\alpha > 1$, then transactions exhibit increasing returns to scale, meaning if we were to increase supply and demand by the same factor k, total transactions would increase by more than a factor k. If $\alpha < 1$, then transactions exhibit decreasing returns to scale, meaning if we were to increase the supply and demand by the same factor k, total transactions would increase by less than a factor k. However, total returns to scale can vary by type of market and stage of growth, in which case our model should be interpreted as describing the optimal growth problem within a given growth stage.

Increasing returns to scale corresponds to a market where as the supply and demand increase, transactions between buyers and sellers get more efficient. For example, in a ride-sharing market, if growth occurs due to more riders and drivers operating in the same geographical area, this leads to an increase in the density of available (open) drivers, which in turn leads to shorter pickup times. The decreased time to pick up each rider means drivers spend a higher proportion of their time on trip (higher on-trip utilization), enabling them to complete more trips per hour on the platform. Hence, in this case, increasing both supply and demand by a factor k will lead to an increase in rides by more than a factor k.

This scale economy effect was popularized in the famous “Uber napkin diagram" (Chen 2015) to explain a virtuous cycle in ride-sharing – greater volume

leading to higher density, leading to higher driver utilization, leading to lower costs, leading to lower prices, leading to even greater volume – that was used by Uber insiders to explain its early success. This virtuous growth cycle is typical of the early phases of ride-share growth in a city.

In contrast, decreasing returns to scale can occur when growth involves expanding the market to increasingly difficult-to-serve segments or harder-to-match buyers and sellers, such that transactions among sellers and buyers become less efficient. Consider ride-sharing again. Once growth in a city has saturated the city "core", additional growth tends to occur by expanding the geographical coverage of service beyond the city core to outskirts and surrounding suburban areas with lower population density. Such growth can result in a decrease in the average density of available drivers, which leads to increases in average pickup time and consequently reduced driver on-trip utilization. If driver density declines due to geographical growth, then an increase in supply and demand by a factor k will result in less than a factor k increase in rides. This type of growth by geographical expansion is typical of the more mature growth phase of ride-sharing service in a city.

Lastly, we note that one can combine our results from these two cases into a single model of growth in which we assume the market makes a transition from increasing to decreasing returns after it grows beyond a transition "market size". This is described in more detail below in Section 2.4.2.

2.3.2 Prices and supply/demand growth

Let p_s denote the price paid to the seller for a transaction and p_d denote the price charged to the buyer for a transaction. We assume the platform controls both the prices p_s and p_d and can set $p_d < p_s$ if it wants to subsidize the market. Both buyers and sellers have homogeneous utilities; buyers receive a gross utility v for each transaction and sellers incur a cost c for each transaction. Hence the net seller surplus per unit of supply is

$$w = \frac{(p_s - c)g(s,d)}{s} \tag{2.2}$$

and the net buyer surplus per unit of demand is

$$u = \frac{(v - p_d)g(s,d)}{d} \tag{2.3}$$

We then model the growth rate of adopters on the supply and demand side as a linear function of these net utilities

$$\dot{s}(t) = (-\beta_0^s + \beta_1^s w)s(t) \tag{2.4}$$

$$\dot{d}(t) = (-\beta_0^d + \beta_1^d u)d(t) \tag{2.5}$$

where here we assume all the coefficients above are positive. Note net utilities w and v can be negative in this formulation, which produces a negative growth effect.

The interpretation of the coefficients β_0^s and β_0^d are as attrition rates of sellers and buyers, respectively. Specifically, $1/\beta_0^s$ and $1/\beta_0^d$ can be interpreted as the average life time of adopters on the supply and demand side. Consider the example where supplier surplus is zero, $w = 0$, so that suppliers are indifferent to staying in the market or exiting. In this case we have $\dot{s}(t)/s(t) = -\beta_0^s$, so the

stock of supply adopters declines by a constant β_0^s percent per unit time, i.e., suppliers have a mean time to exit of $1/\beta_0^s$ and the stock of supply adopters decays at a rate β_0^s.

The coefficients β_1^s and β_1^d determine, respectively, how strongly supply and demand growth respond to the surplus generated for sellers and buyers. This can be seen by considering the case where buyers are persistent, so $\beta_0^d = 0$. In this case we have $\dot{d}(t)/d(t) = \beta_1^d u$, which means an increase of one dollar in per-buyer utility u leads to a β_1^d percent increase in demand side adopters per unit time. The value $1/\beta_1^d$ (resp. $1/\beta_1^s$) has units of dollars and can be interpreted as the *adoption cost*; that is, how much surplus must be provided to buyers (resp. sellers) to generate an incremental adopter on the demand (resp. supply) side.

In terms of micro-structure, one can think of this adoption growth as coming from word-of-mouth adoption of new users as in the "imitation effect" of the Bass model (Bass 1969) (see also Kalish 1985 and Oren and Schwartz 1988), more usage by the same population of users or some combination of both. In the first interpretation, the growth equations describe the number of adopters on each side of the market - keeping the usage frequency per adopter constant. Using the demand side as an example, in this interpretation buyers who successfully transact and pay a price p_d lower than their valuation v, have a positive experience that is either observed or communicated to potential new adopters. Hence, the adoption rate is positively related to the per-transaction surplus $(v - p_d)$ and the average number of transactions per current demand adopter $g(s, d)/d$. In the second interpretation, the growth functions describe the change in the usage frequency for a representative user, fixing the total number of users in the market. This aligns with the marketing literature in consumer loyalty; once

a user has a positive experience (receives positive surplus) from purchasing a product, they are more likely to purchase it again due to brand-specific user skills, reduced quality uncertainty and reduced search costs (Wernerfelt 1991). Under this loyalty interpretation of growth, previous successful (resp. negative) usage encourages (resp. discourages) future usage. Hence, the change in usage frequency is again positively related to the per-transaction surplus and the average number of transactions per user.

Substituting the values w and u above and simplifying, we can rewrite the supply and demand growth as

$$\dot{s}(t) = -\beta_0^s s(t) + \beta_1^s (p_s - c) g(s(t), d(t)) \tag{2.6}$$

$$\dot{d}(t) = -\beta_0^d d(t) + \beta_1^d (v - p_d) g(s(t), d(t)) \tag{2.7}$$

Note above that the growth of both supply and demand is increasing in total output $g(s(t), d(t))$.

2.3.3 Optimal growth formulation

We next formulate the platform's optimal growth problem. Note that at any time t, the platform's profit rate is the price difference between the demand side and the supply side, $p_d - p_s$, multiplied by the number of successful transactions g. We assume the platform has a discount rate $\rho > 0$. The platform's objective is then to maximize its discounted aggregated profit:

$$\max_{p_d, p_s} \int_0^\infty e^{-\rho t} (p_d - p_s) g(s, d) dt \tag{2.8}$$

subject to the growth equations (2.6), (2.7). (There may also be some additional growth constraints which we will discuss later in detail.)

2.3.4 Market size-and-balance state-space reduction

We next show that this two-sided market model with two states, $s(t)$ and $d(t)$, can be reduced to an equivalent model with a single “market size" state variable. How this market size is allocated to the supply and demand side is then determined by a new “market balance" control variable. This market size-and-balance representation is a state-space reduction, not merely a reformulation in the sense that market size is the only quantity necessary to determine both the rewards and the evolution of the system at any given time. It both simplifies and clarifies the analysis of optimal growth and balance. We next formally define the state-space reduction and then discuss the interpretation of the resulting market size and market balance variables.

State-space reduction

While the supply and demand stocks are separate state variables in our model, the following proposition establishes that these two states can be reduced to a single market size state provided prices are unconstrained:

Proposition 1 *If there are no constraints on prices p_s, p_d, then the state (s, d) can be reduced to a single-dimensional market size state:*

$$x \triangleq \frac{s}{\beta_1^s} + \frac{d}{\beta_1^d} \tag{2.9}$$

That is, for any given initial state (s_0, d_0) at time $t = 0$, let $x = \frac{s_0}{\beta_1^s} + \frac{d_0}{\beta_1^d}$ and $V(s_0, d_0)$ denote the optimal value of equation (2.8). Then for any (s_1, d_1) satisfying $\frac{s_1}{\beta_1^s} + \frac{d_1}{\beta_1^d} = x$, it holds that $V(s_1, d_1) = V(s_0, d_0)$.

Proof. Suppose we are at state (s_0, d_0) at time τ. Then we will show that without cost we can instantaneously shift to any other state (s_1, d_1) that satisfies

$$x \equiv \frac{s_0}{\beta_1^s} + \frac{d_0}{\beta_1^d} = \frac{s_1}{\beta_1^s} + \frac{d_1}{\beta_1^d} \tag{2.10}$$

To see why, consider the pair of impulse price paths (price shocks) $p_s(t;\tau) = \frac{s_1 - s_0}{\beta_1^s g(s_0,d_0)}\delta(t-\tau)$ and $p_d(t;\tau) = \frac{d_0 - d_1}{\beta_1^d g(s_0,d_0)}\delta(t-\tau)$ where $\delta(\cdot)$ is the Dirac function defined by

$$\delta(x) = \begin{cases} +\infty, & x = 0 \\ 0, & x \neq 0 \end{cases} \text{ and } \int_{-\infty}^{\infty} \delta(x)dx = 1$$

which implies $\int_{\tau^-}^{\tau^+} \delta(t-\tau)f(t)dt = f(\tau)$, the sifting property of the Dirac function (Weisstein 2020).

Define $s(\tau) = s(\tau^-) = s_0$. Then the stock of supply right after τ, i.e., $s(\tau^+)$ is given by

$$\begin{aligned} s(\tau^+) &= s(\tau^-) + \int_{\tau^-}^{\tau^+} \dot{s}(t)dt \\ &= s_0 + \int_{\tau^-}^{\tau^+} -\beta_0^s s(t) + \beta_1^s (p_s(t;\tau) - c) g(s(t), d(t))dt \\ &= s_0 + \int_{\tau^-}^{\tau^+} -\beta_0^s s(t) + \beta_1^s \left(\frac{s_1 - s_0}{\beta_1^s g(s_0, d_0)} \delta(t-\tau) - c \right) g(s(t), d(t))dt \end{aligned}$$

By the sifting property of the Dirac function,

$$\begin{aligned} &\int_{\tau^-}^{\tau^+} \beta_1^s \left(\frac{s_1 - s_0}{\beta_1^s g(s_0, d_0)} \delta(t-\tau) \right) g(s(t), d(t))dt \\ =&\beta_1^s \left(\frac{s_1 - s_0}{\beta_1^s g(s_0, d_0)} \right) g(s(\tau), d(\tau)) \\ =&\beta_1^s \left(\frac{s_1 - s_0}{\beta_1^s g(s_0, d_0)} \right) g(s_0, d_0) = s_1 - s_0 \end{aligned}$$

and for all other terms in the integral without the Dirac function,

$$\int_{\tau^-}^{\tau^+} -\beta_0^s s(t) - \beta_1^s c g(s(t), d(t))dt = 0$$

This is because $[\tau^-, \tau^+]$ is an infinitely small interval. Thus, we have

$$s(\tau^+) = s_0 + s_1 - s_0 = s_1$$

An identical argument proves that d_1 can be instantly achieved by setting $p_d(t) = \frac{d_0 - d_1}{\beta_1^d g(s_0, d_0)} \delta(t - \tau)$. Now by (2.10), $\frac{s_1 - s_0}{\beta_1^s} = \frac{d_0 - d_1}{\beta_1^d}$. Hence, the total cost to the platform of the two price shocks is

$$\int_{\tau^-}^{\tau^+} (p_s(t) - p_d(t)) g(s(t), d(t)) dt$$
$$= \int_{\tau^-}^{\tau^+} \left(\frac{s_1 - s_0}{\beta_1^s g(s_0, d_0)} - \frac{d_0 - d_1}{\beta_1^d g(s_0, d_0)} \right) \delta(t - \tau) g(s(t), d(t)) dt = 0$$

So the change from (s_0, d_0) and (s_1, d_1) is costless. □

Intuitively, the reason the state space collapses to a single market size variable is that the platform can use a cost-less price shock to arbitrarily shift the market balance at any point in time between supply and demand. These price shocks keep the total market size x fixed and are budget neutral, so they do not affect the profit of the platform. This means the platform's optimal growth problem (which we define formally below) reduces to one with a single state variable x together with a new *market balance* control variable

$$\gamma = \frac{s}{\beta_1^s x}$$

with $x\gamma = \frac{s}{\beta_1^s}$, $x(1 - \gamma) = \frac{d}{\beta_1^d}$. The balance control γ then represents the fraction of the market size allocated to the supply side; a high value of γ (close to one) corresponds to a supply-heavy market balance, while a low value of γ (close to zero) corresponds to a demand-heavy market balance.

Interpretation of market size and balance

In the definition of market size x in (2.9), $1/\beta_1^s$ and $1/\beta_1^d$ are interpreted, respectively, as the "cost" of a unit of supply and demand and have units of dollars per unit of supply/demand. The market size x therefore has units of dollars and can be interpreted as the "market value" or "level of investment". Indeed, one can show, using an argument similar to that in the proof of Proposition 1, that the platform can at any time choose to shock the market with extremely high demand prices and low supply prices such that it brings the market size from x down to 0 and in the process converts its user base to a cash value of x. In this sense, x can also also be thought of as the liquidation value of the market.

The market balance γ is the fraction of market value contained in the supply side while $1-\gamma$ is the fraction of market value contained in the demand side. In this sense, balance is a concept of relative value not relative quantity. For example, if suppliers are expensive to acquire and buyers are not (as measured by $1/\beta_1^s$ and $1/\beta_1^d$), then even if γ is close to one (a high concentration of total value in the supply side), this does not necessarily correspond to a high number of adopters on the supply side relative to adopters on the demand side.

Besides providing analytical simplification, this market size and balance formulation is extremely helpful in interpreting optimal growth. For one, having a scalar market size allows us to unambiguously identify one growth strategy as being "faster" than another if it produces a market size that is larger at all times (just as the concept of "distance" is fundamental to defining "speed" in physics). It also lets us identify important size thresholds. In particular, we show below that in the increasing returns case there is a critical market size such that growth must be subsidized if market size is below this critical size. And in the decreasing

returns case, there is a market size at which the market becomes saturated and it is not optimal to grow beyond this size.

Likewise, the concept of balance is also quite useful and enables us to show precisely how the optimal allocation of size between supply and demand changes during the evolution of market growth. Indeed, in practice understanding market balance is of first-order practical importance when managing growth in two-sided markets like ride-sharing and our formulation highlights this important choice variable.

Output and growth equations as a function of market size and balance

Under this size-and-balance reformulation, the market output g can be written as a function of γ and x,

$$g(s,d) = g(\beta_1^s \gamma x, \beta_1^d (1-\gamma)x) = x^\alpha g(\beta_1^s \gamma, \beta_1^d (1-\gamma))$$

The rightmost equality is due to the fact that $g(s,d)$ is homogeneous of degree α. Define $h(\gamma) = g(\beta_1^s \gamma, \beta_1^d (1-\gamma))$. Then the market output is given by

$$g(\gamma, x) = x^\alpha h(\gamma) \tag{2.11}$$

The market size x has initial condition: $x(0) = \frac{s(0)}{\beta_1^s} + \frac{d(0)}{\beta_1^d} > 0$, and evolves according to the differential growth equation:

$$\dot{x} = \frac{\dot{s}}{\beta_1^s} + \frac{\dot{d}}{\beta_1^d} = -(\beta_0^s \gamma + \beta_0^d (1-\gamma))x + (v - c - \pi) g(\gamma, x) \tag{2.12}$$

where $\pi = p_d - p_s$ is the platform's *profit margin*. Note this implies that under this reformulation price also reduces to a one-dimensional control.

Recovering the original state variables and prices

The two-dimensional state $s(t)$,$d(t)$ and control $p_s(t)$, $p_d(t)$ can be recovered from $x(t)$, $\gamma(t)$ (Proposition 30 and 31 in the Appendix). An important consequence of these propositions is that while the market size and balance decomposition assumes impulse price shocks can be used to choose the market balance at any point in time, if growth paths and balance paths are smooth, then so are prices. Moreover, we show below in Theorem 1 that the optimal balance itself is a continuous (and continuously differentiable) function of the market size $x(t)$. This means if a growth path $x(t)$ is continuous, then so is the optimal balance $\gamma(t)$, and therefore smooth growth paths imply smooth price paths.

2.3.5 Optimal growth reformulation

We next reformulate the optimal growth problem applying the size-and-balance reformulation:

$$\max_{\pi,\gamma} \int_0^\infty e^{-\rho t} \pi g(\gamma, x) dt \tag{2.13}$$

subject to the state equation

$$\dot{x} = -(\beta_0^s \gamma + \beta_0^d (1-\gamma))x + (v - c - \pi) g(\gamma, x)$$

and the constraint on the market balance:

$$0 \leq \gamma \leq 1$$

with the initial condition:

$$x(0) > 0$$

Moreover, the growth path may be subject to additional constraints of the form

$$x(t) \in \mathcal{X}$$

where $\mathcal{X}$ is the set of all feasible growth paths with respect to those constraints.

Again, we consider this growth problem in the space of the market size trajectory $x(t)$ rather than in the space of the price policy. As noted, we show in the next section that the optimal balance at any time t is a function of the market size $x(t)$, and by Propositions 30 and 31, optimal prices in turn are uniquely determined by the market size and balance decisions. This means we can analyze and compare growth strategies based solely on analyzing their respective market size trajectories.

2.3.6 Definitions

Before proceeding, we introduce some useful terminology.

Definition 1 *$x(t)$ is an* increasing growth path *if $x(t_1) \geq x(t_2), \forall t_1 > t_2$. It is* strictly increasing *if $x(t_1) > x(t_2), \forall t_1 > t_2$.*

That is, a growth path is increasing if the market size is non-decreasing over time and strictly increasing if market size is strictly increasing with time. We also need:

Definition 2 *$x_1(t)$ is a* faster growth path *from $\underline{x}$ to $\bar{x}$ over $[\underline{t}, \bar{t}]$ than $x_2(t)$ if $\underline{x} = x_1(\underline{t}) = x_2(\underline{t})$, $\bar{x} = x_1(\bar{t}) = x_2(\bar{t})$, and $x_1(t) \geq x_2(t)$ at all $\underline{t} \leq t \leq \bar{t}$. It is* strictly faster *if $x_1(t) > x_2(t)$ for all $t \in (\underline{t}, \bar{t})$.*

This definition is useful when comparing two growth paths; if the market size in path 1 is at least as large as that in path 2 at all times, then path 1 is a faster growth path.

We say the market is *subsidized at time t* if $\pi(t) < 0$. That is, the platform pays the supply side a higher price than the price it charges the demand side when the market is subsidized. Growth may require subsidies. In an extreme case the platform may want to inject an instantaneous subsidy into the market to achieve rapid growth. Specifically, consider an impulse of subsidy at time t that moves the market size from $x(t^-) = x_0$ to $x(t^+) = x_1 > x_0$ of the form

$$\pi(t) = -\frac{x_1 - x_0}{g(\gamma_0, x_0)}\delta(t)$$

We call $\pi(t)$ a *subsidy shock*. The cost of the subsidy shock is

$$\int_{t^-}^{t^+} \pi(t)g(\gamma_0, x_0)dt = \int_{t^-}^{t^+} -(x_1 - x_0)\delta(t)dt = -(x_1 - x_0)$$

We call $M = x_1 - x_0$ the *magnitude* of the subsidy shock.

Definition 3 *A market is* viable *if there exists a feasible increasing growth path that generates a strictly higher total discounted profit in (2.13) than the path* $x(t) = x_0, \forall t$.

In other words, a market is viable if growing it strictly improves the total discounted profit.

We also define the following function which is important in our analysis:

Definition 4 $G(\gamma, x) = (v - c)g(\gamma, x) - (\rho + \beta_0^s\gamma + \beta_0^d(1 - \gamma))x$

Intuitively, $G(\gamma, x)$ is the instantaneous rate of net social welfare generated by being in state x with balance choice γ. That is, it is the rate of welfare generated by current transactions (the term $(v - c)g(\gamma, x)$) minus the loss in value due to the decline in market size x (the term proportional to x). This market value loss is the sum of the time discount rate ρ and the attrition rates per unit of supply

and demand, $\beta_0^s\gamma + \beta_0^d(1-\gamma)$, times the market size. Hence, $G(\gamma, x)$ is the net social welfare rate.

2.3.7 Assumptions

We impose the following regularity conditions on $h(\gamma)$:

Assumption 1 *$h(\gamma)$ is twice-differentiable and $h(\gamma) > 0$ on $(0,1)$. Moreover, $h(0) \geq 0$, $h(1) \geq 0$, and both are finite.*

Remark 1 *$h(\gamma)$ is bounded for $\gamma \in [0,1]$.*

Assumption 2 *$h(\gamma)$ is strictly concave in γ on $(0,1)$, i.e., $h''(\gamma) < 0$.*

Assumption 3 *$h'(\gamma) = 0$ has a unique solution on [0,1].*

Assumption 4 *When $\alpha < 1$, $(1-\alpha)h'(\gamma)^2 < -\alpha h(\gamma)h''(\gamma)$ for $\gamma \in (0,1)$.*

We then have the following result on the constant elasticity of substitution (CES) function:

Lemma 1 *Suppose the output function $g(s,d)$ is the CES production function, i.e.,*

$$g(s,d) = (\theta s^m + (1-\theta)d^m)^{\frac{\alpha}{m}}, 0 \leq m < 1, 0 < \theta < 1$$

Then the production function satisfies Assumption 1-4 if $0 < \alpha < 1$ or $1 < \alpha < 1 + \epsilon$, where $\epsilon > 0$ is sufficiently small. In particular, if $m = 0$, $g(s,d)$ follows the Cobb-Douglas function, and it satisfies Assumption 1-4 as long as $\alpha < \min\{\frac{1}{\theta}, \frac{1}{1-\theta}\}$.

2.4 Main Results

Moreover, attrition losses in market size are minimized by concentrating the market balance in the "most durable" side of the market: the supply side ($\gamma = 1$) when supply attrition is lower than demand attrition ($\beta_0^s < \beta_0^d$) and the demand side ($\gamma = 0$) when demand attrition is lower than supply attrition ($\beta_0^d < \beta_0^s$). We denote this balance choice as $\hat{\gamma}$ and call it the *durability-maximizing balance*. Thus, $\hat{\gamma} = \mathbb{1}_{\beta_0^s < \beta_0^d}$.

Our main result on optimal market balance is then (proof in the Appendix):

Theorem 1 (Optimal Balance) *For any level of market size $x(t) = x$, the optimal market balance $\gamma^*(x)$ is given by*

$$\gamma^*(x) = \max\Big\{ \min\Big\{ (h')^{-1}\big(\frac{(\beta_0^s - \beta_0^d)x^{1-\alpha}}{v - c}\big), 1\Big\}, 0\Big\} \tag{2.14}$$

Moreover, as the market size grows from zero to infinity, when the market exhibits decreasing returns to scale, then $\gamma^*(x)$ goes from the output-maximizing balance γ^* to the durability-maximizing balance $\hat{\gamma}$; when the market exhibits increasing returns to scale, then $\gamma^*(x)$ goes from the durability-maximizing balance $\hat{\gamma}$ to the output-maximizing balance γ^*.

In summary, as x increase from zero to infinity, we have

Optimal Balance $\gamma^*(x)$ as Market Size x Increases[5]

[5] $\beta_0^s > \beta_0^d$: demand is more durable than supply, and vice versa; α is the total returns to scale; $0 \to \gamma^*$ means $\gamma(x)$ goes from 0 to the output-maximizing balance γ^* as x goes from 0 to ∞, and so on.

	$\alpha > 1$	$\alpha < 1$
$\beta_0^s > \beta_0^d$	$0 \to \gamma^*$	$\gamma^* \to 0$
$\beta_0^s < \beta_0^d$	$1 \to \gamma^*$	$\gamma^* \to 1$

To see the intuition behind these optimal balance results, note that there are two main components of supply and demand growth in our model (2.12): 1) loss in market size due to the supply and demand attrition coefficients β_0^s and β_0^d (the terms proportional to γ and $1-\gamma$ in (2.12)); and 2) growth in market size due to the rate of adoption (the term proportional to total market output $g(\gamma, x) = h(\gamma)x^\alpha$ in (2.12)). 1) is minimized at the durability-maximizing balance $\hat{\gamma}$, and 2) is maximized at the output-maximizing balance γ^*. There is a tension between these two forces. And their relative importance is a function of both market size and returns to scale.

With increasing returns to scale and low market size x, market output relative to size is very small and attrition losses are the dominant factor influencing growth. Hence, the optimal balance is to concentrate investment in the most durable size of the market. As the size grows, increasing returns imply output per unit of size increases and hence for large size, adoption growth dominates attrition losses. Therefore, the optimal balance shifts to the output-maximizing balance.

With decreasing returns to scale the effects are the opposite. At low market size, output relative to market size is large and hence adoption growth dominates attrition losses and the optimal balance choice is to maximize output. But as the market grows, decreasing returns imply that output relative to market size decreases and hence attrition loss becomes the dominant factor influencing

growth. Therefore, the optimal balance shifts to the durability-maximizing balance.

Importantly, this optimal balance result also implies that we can analyze growth policies by simply analyzing the market size path $x(t)$, since the optimal balance is determined by (2.14) once $x(t)$ is specified. Therefore, in the rest of the paper, we only consider policies that satisfy (2.14).

2.4.1 Stationary solution

We next define an important size threshold corresponding to the first-order conditions for our optimal control problem, namely:

Theorem 2 (Stationary Solution) *The stationary solution to the infinite-horizon problem (2.13) is*

$$x^* = \left\{ \frac{\beta_0^s \gamma^*(x^*) + \beta_0^d (1 - \gamma^*(x^*)) + \rho}{(v-c)\alpha h(\gamma^*(x^*))} \right\}^{\frac{1}{\alpha - 1}}$$

$$\gamma^*(x^*) = \max\left\{ \min\left\{ h'^{-1}\left(\frac{(\beta_0^s - \beta_0^d) x^{*1-\alpha}}{v-c}\right), 1 \right\}, 0 \right\}$$

For example, when using Cobb-Douglas function $g(s,d) = As^{\alpha_s} d^{\alpha_d}$, the stationary solution is in closed-form:

$$x^* = \left\{ \frac{A(v-c)(\beta_1^s)^{\alpha_s}(\beta_1^d)^{\alpha_d}\alpha_s^{\alpha_s}\alpha_d^{\alpha_d}}{(\rho+\beta_0^d)^{1-\alpha_s}(\rho+\beta_0^s)^{1-\alpha_d}} \right\}^{\frac{1}{1-\alpha}} \{\alpha_s(\rho+\beta_0^d) + \alpha_d(\rho+\beta_0^s)\}$$

$$\gamma^*(x^*) = \frac{\alpha_s(\rho+\beta_0^d)}{\alpha_s(\rho+\beta_0^d) + \alpha_d(\rho+\beta_0^s)}$$

When $\alpha < 1$ (decreasing returns), one can show the Hamiltonian (shown in the appendix) is concave in x and hence this stationary solution is the unique steady-state solution to the infinite horizon problem (2.13). In this case, the stationary

size x^* is a profit-maximizing market size, which we call the *saturation size*. When $\alpha > 1$ (increasing returns), the stationary solution above only characterizes a saddle point.

Therefore, a decreasing returns to scale market naturally stops growing once reaching x^*, while an increasing returns to scale market will not stop growing unless it reaches its size limit.

2.4.2 Optimal growth policy

We first show the optimal policy for an increasing returns to scale market with some growth constraints. We consider an upper bound on the market size, which we call the *market potential*, denoted as $\bar{x}$. We also consider an upper bound on the growth rate for a given market size, denoted as $f(x)$.

Theorem 3 *For an increasing returns to scale market, consider the infinite horizon problem (2.13) with the constraints*

$$\mathcal{X} = \{x(t)|x(0) = x_0, x(t) \leq \bar{x}, 0 \leq \dot{x} \leq f(x)\}^6$$

where $f(x) > 0$ for $x > 0$, and let $F(t)$ denote the solution to the differential equation $\dot{x} = f(x), x(0) = x_0$. For any given $x_0 < x^*$, let $\tilde{x} > x_0$ be uniquely defined by

$$\int_0^{F^{-1}(\tilde{x})} e^{-\rho t} G_x(\gamma^*(F(t)), F(t)) F'(t) dt = 0.$$

where x^* is the stationary market size defined in Theorem 2.

Then there are two cases:

[6]The assumption of $\dot{x} \geq 0$ is made because increasing growth paths are the focus of our analysis. Allowing for $\dot{x} < 0$ (decreasing growth) adds to the complexity of the analysis without providing additional insights.

1) If $x^* \leq x_0 < \bar{x}$ or $x_0 < x^*$, $\bar{x} > \tilde{x}$, then it is optimal to grow the market as fast as possible to $\bar{x}$, i.e.,

$$x^*(t) = \begin{cases} F(t), & 0 \leq t \leq F^{-1}(\bar{x}) \\ \bar{x}, & t > F^{-1}(\bar{x}) \end{cases} \tag{2.15}$$

2) If $x_0 < x^*$ and $x_0 < \bar{x} < \tilde{x}$, then it is optimal not to grow the market at all, i.e., $x^*(t) = x_0, t \geq 0$.

In other words, for an increasing returns market the optimal policy is either to grow the market from x_0 to $\bar{x}$ as fast as possible, or not to grow at all, depending on whether the market potential $\bar{x}$ is sufficiently large relative to x_0.

For a decreasing returns to scale market, the optimal policy is given in the following proposition:

Theorem 4 (Optimal Growth for Decreasing Returns) *Consider a market with decreasing returns to scale ($\alpha < 1$) and an initial market size, x_0, that is below the saturation size ($x_0 < x^*$). Then for the infinite horizon problem (2.13) with $\mathcal{X} = \{x(t) | x(0) = x_0, x(t) \leq \bar{x}, 0 \leq \dot{x} \leq f(x)\}$, faster growth dominates slower growth getting from x_0 to $\min\{x^*, \bar{x}\}$, where x^* is the stationary market size from Theorem 2.*

In the decreasing returns case, the output per unit size is high when the market size is initially low, but as the market size increases, the output per unit size declines due to the decreasing returns. At some point, the optimal growth rate reaches zero. This stopping point is characterized by the steady-state solution given by Theorem 2 where x^* is the long-term profit-maximizing size. Once $x(t) = x^*$ is achieved, it is optimal to stay at x^*. Hence, we call

x^* the market *saturation size* and the corresponding optimal balance $\gamma^*(x^*)$ the *saturation balance*.

Combining Theorem 3 and Theorem 4 together, we conclude that fast growth is generally optimal for both increasing and decreasing returns to scale markets:

Theorem 5 (Optimal Growth) *For the infinite-horizon problem (2.13), it is optimal to grow the market as fast as possible, given that the market is viable. Moreover, if the market has increasing returns to scale, the optimal growth path converges to the upper bound of the market size; if the market has decreasing returns to scale, the optimal growth path converges to the saturation size.*

Corresponding to the previous discussion on the stationary solution, the concavity/convexity of the objective function (2.13) in x leads to different long-run optimal sizes for decreasing/increasing returns to scale markets. Nevertheless, as long as there is sufficient potential in market growth, it is optimal to grow the market as fast as possible in both cases. In the next section, we discuss the specific optimal growth policies in some real-world settings, and introduce conditions that identify a viable market in those settings.

An unconstrained market

We start with the simplest case without bounds on market size or the growth rates.

Proposition 2 *When the market exhibits increasing returns to scale, if $\mathcal{X}$ is the set of all increasing growth paths from x_0, then the optimization problem (2.13) is unbounded.*

In short, when there is no limiting market size in an increasing returns to scale market, the optimal growth policy results in a market size that is unbounded and generates infinite profit. As a practical matter, this result implies that real-world two-sided markets cannot indefinitely exhibit increasing returns to scale as they grow, else there would be opportunities for unbounded profit. Moreover, due to the scale effect, the output per unit size also becomes unbounded, which is also unrealistic. This is simply the result of a production model that cannot make physical sense in the limit of ever increasing market size.

A more surprising result is the decreasing returns case. Specifically, while a decreasing-returns market can be self-sufficient at a low scale and could in fact grow to saturation size organically without subsidies, it is in fact still optimal to subsidize to achieve faster growth in this case. Indeed, the next result, which follows directly from Theorem 4, shows that when starting at an initial size $x_0 < x^*$, a strategy that produces faster growth from x_0 to x^* is always better.

Corollary 1 (Optimal Growth for Decreasing Returns) *If the market has decreasing returns to scale ($\alpha < 1$), the initial market size, x_0, is below the saturation size ($x_0 < x^*$) and prices are unconstrained, then the optimal growth strategy is to apply a subsidy shock of magnitude $M = x^* - x_0$ at time $t = 0$ to immediately grow the market to size x^*.*

If there is no constraint on the growth rate, the fastest growth path is simply a jump from x_0 to x^* at time 0, which can be achieved with a subsidy shock of magnitude $x^* - x_0$. As noted, this result is more surprising in the decreasing returns case because the market is profitable at low scale and therefore is able to grow from its initial size to the saturation size without subsidy. However, such

organic growth comes at the cost of delaying the time to reach the saturation size and hence delaying the profits that come from operating at the long-run maximum size. As a result, subsidizing more rapid growth is still beneficial, despite not being strictly necessary.

In both the increasing and decreasing returns case, growth leads to higher profit. We conclude with the condition for a viable market in this case:

Proposition 3 *When there is no constraint on market growth or size, a decreasing returns to scale market is viable as long as $x_0 < x^*$, and an increasing returns to scale market is always viable.*

A market with finite market size

We next consider a market with finite market potential $\bar{x}$, but unconstrained growth rates. Moreover, to avoid trivialities, we will assume the initial market size x_0 to be sufficiently small $(x_0 < x^*)$. Intuitively, a subsidy shock should still be optimal if the market is viable.

However, in contrast to the unconstrained case, an increasing returns to scale market is not always viable. Indeed, the market potential $\bar{x}$ needs to be sufficiently large so that the potential future reward can compensate for the cost of growth. In an increasing returns to scale market, growth is expensive because when the size is small, output relative to market size is low and hence the platform has to heavily incentivize growth through subsidies. If the market cannot recoup this subsidy cost by reaching a sufficient size, it may be unprofitable to grow. Specifically, define:

Definition 5 *The* critical size *is the smallest market size x such that $\dot{x} \geq 0$ and $\pi \geq 0$.*

The next proposition characterizes the critical size:

Proposition 4 *The critical size x_c is given by*

$$x_c = \left(\frac{\beta_0^s \gamma_c^\dagger + \beta_0^d (1 - \gamma_c^\dagger)}{(v - c) h(\gamma_c^\dagger)} \right)^{\frac{1}{\alpha - 1}} \tag{2.16}$$

$$\gamma_c^\dagger = \begin{cases} 0, & \frac{h'(0)}{h(0)} < \frac{\beta_0^s - \beta_0^d}{\beta_0^d} \\ \gamma_c, & \beta_0^s \frac{h'(1)}{h(1)} < \beta_0^s - \beta_0^d < \beta_0^d \frac{h'(0)}{h(0)} \\ 1, & \frac{h'(1)}{h(1)} > \frac{\beta_0^s - \beta_0^d}{\beta_0^s} \end{cases} \tag{2.17}$$

where γ_c uniquely defined by $(\beta_0^s \gamma_c + \beta_0^d (1 - \gamma_c)) \frac{h'(\gamma_c)}{h(\gamma_c)} = \beta_0^s - \beta_0^d$.

We next show that to grow to the critical size requires continuous subsidies:

Proposition 5 *When $\alpha > 1$, consider any market size $x < x_c$, then any increasing growth path from x to x_c must be subsidized at all times t along the path.*

This shows that when a platform's initial size is below the critical size, subsidies are the only growth option. Since no point on the growth path generates positive profit, one may expect that a market with a maximal market size below or at the critical size is not viable. The next corollary confirms this intuition:

Corollary 2 (Optimal Growth to the Critical Size) *For the infinite horizon problem (2.13) with $\alpha > 1$, x_0 sufficiently close to 0, and $x(t) \leq x_c$, the market is not viable, and it is optimal to keep $x(t) = x_0$.*

Growth to the critical size is a special case. Next, we analyze the optimal growth policy for a general upper bound $\bar{x}$:

Proposition 6 *Consider the infinite horizon problem (2.13) with $\alpha > 1$ and the constraint that $\mathcal{X} = \{x(t)|x(0) = x_0, x(t) < \bar{x}\}$. Define $\tilde{x}$ such that*

$$G(\gamma^*(x_0), x_0) = G(\gamma^*(\tilde{x}), \tilde{x}), \tilde{x} > x_0. \tag{2.18}$$

If the market is viable ($\bar{x} > \tilde{x}$), then it is optimal to apply a subsidy shock of magnitude $M = \bar{x} - x_0$ at time $t = 0$ to immediately grow the market to size $\bar{x}$. If the market is not viable ($\bar{x} < \tilde{x}$), then it is optimal to keep $x(t) = x_0$.

Note $\tilde{x}$ is the threshold size that determines whether an increasing returns to scale market is viable. The following lemma relates this minimal viable size to the critical size:

Lemma 2 *For an increasing returns to scale market, if the initial market size is sufficiently close to 0, then the critical market size $x_c < \tilde{x}$.*

From Proposition 6, the optimal policy to grow an increasing returns to scale market is either to grow it as fast as possible, or not to grow at all. From Proposition 6, whether the market is worth growing at all hinges on the market potential, $\bar{x}$, since $\bar{x}$ determines the long-term reward the platform can eventually collect once the growth phase completes.

When $\bar{x}$ is small, it is optimal to grow the market as slowly as possible, and the slowest growth path is to apply a subsidy shock at the terminal time. Since we are considering an infinite-horizon problem with free endpoints, the slowest path is effectively a path along which the market size never increases. The

intuition is that in an increasing returns to scale market, when the market size is small, output per unit size is very low and hence attrition losses dominate surplus-driven growth. In this regime, it is very costly to maintain market size and hence the optimal policy is to maintain as small a size as long as possible. Indeed, for a small $\bar{x}$, the increase in the long-term reward for reaching $\bar{x}$ does not compensate for the cost of growing the market from x_0 to $\bar{x}$; hence, it is optimal not to grow the market at all.

Nevertheless, when $\bar{x}$ is large, meaning there is a positive reward for achieving the maximal market size, a slower growth path delays the time that the platform can collect that reward. In this case, the benefit from collecting it sooner dominates the loss from a higher subsidy cost. Therefore, for a $\bar{x}$ that is large enough, the optimal growth policy switches to a subsidy shock at time zero; that is, the fastest growth path possible.

While stylized, this result reflects the intuition in the ride-sharing industry about the merits of rapid growth. Ride-sharing is simply not economically viable at low density and low scale due to the long pickup times, which significantly increase the driver labor time per ride. To sustain such an inefficient scale through subsidies – or even slowly grow the market at all while it is at inefficient scale – is uneconomical. A large injection of funding to bring the market rapidly up to an efficient scale is warranted. This is in fact the growth strategy that most ride-sharing companies have adopted.

As for the decreasing returns to scale market, since the market is efficient even at a small size and can grow organically, it is viable regardless of the value of the upper bound $\bar{x}$. Therefore, $\bar{x}$ only limits the long-run market size, but does not change the structure of the optimal policy.

Proposition 7 *Consider the infinite horizon problem (2.13) with $\alpha < 1$ and subject to the constraint that $x(t) < \bar{x}$. Then it is optimal to grow the market to $\min\{\bar{x}, x^*\}$ as fast as possible, i.e., by applying an impulse shock and bringing the market size from x_0 to $\min\{\bar{x}, x^*\}$ instantaneously.*

Returns to scale that change with market size

The optimal growth policies for finite and unconstrained market size can be pasted together to model a market that transitions from increasing to decreasing returns after growing beyond a given transition size $\bar{x}$ as follows: (We simply sketch out the idea since the details are straightforward.) By backward induction, first solve for the second-stage optimal growth policy by considering a decreasing returns optimal growth problem with initial condition $x(0) = \bar{x}$. The optimal value of this second-stage problem gives us the present value of the total reward for reaching the transition size $\bar{x}$. Next, solve an increasing returns to scale optimal growth problem with finite market size $\bar{x}$. The results of Proposition 6 then hold for this problem if we suitably modify the definition of viability to reflect that there is a fixed reward for reaching $\bar{x}$ equal to the present value of the optimized second-stage profit.

2.4.3 Finite relative growth rates

In the discussion thus, the growth rate itself was unconstrained. In reality, many factors could slow market growth rates. The first friction we consider is that there are typically decreasing marginal benefits to spending on subsidies that limit the relative growth rate one can achieve. We abstract this sort of friction as

a bound on the maximal relative growth rate of the form:

$$\dot{x}(t) \leq bx(t) \tag{2.19}$$

In other words, we can subsidize and achieve growth according to our growth model (2.12), but once the relative growth rate $\dot{x}(t)/x(t)$ reaches b, further subsidies have no marginal effect on growth. We call (2.19) the *relative growth rate constraint*.

When (2.19) is binding, the market grows exponentially, i.e., $x(t) = x_0 e^{bt}$. This is the fastest growth path in subject to this constraint. Along this fastest path, the profit rate is given by

$$\pi(t)g(\gamma(t), x(t)) = (v-c)g(\gamma(t), x(t)) - (\beta_0^s \gamma(t) + \beta_0^d(1-\gamma(t)) + b)x(t)$$

One can check that the optimal balance $\gamma^*(x(t))$ maximizes the profit rate, but does not affect the trajectory of the market size.

Again, we consider a case where the initial market size is small ($x_0 < x^*$). The optimal policy follows Theorem 3:

Proposition 8 *Consider the infinite horizon problem (2.13) with $\alpha > 1$ and subject to the constraints that $\mathcal{X} = \{x(t)|x(0) = x_0, \dot{x}/x \leq b, x(t) \leqslant \bar{x}\}$. If $\bar{x} > \tilde{x}_r$,where $\tilde{x}_r$ is uniquely defined by*

$$\int_0^{1/b \ln \frac{\tilde{x}_r}{x_0}} G_x(\gamma^*(x_0 e^{bt}), x_0 e^{bt}) x_0 e^{(b-\rho)t} dt = 0$$

then it is optimal to grow the market as fast as possible, and the optimal growth path is

$$x^*(t) = \begin{cases} x_0 e^{bt}, & t \leq \frac{1}{b} \ln \frac{\bar{x}}{x_0} \\ \bar{x}, & t > \frac{1}{b} \ln \frac{\bar{x}}{x_0} \end{cases}$$

Otherwise if $\bar{x} < \tilde{x}_r$, it is optimal to not grow the market at all, i.e., $x^*(t) = x_0, t \geq 0$.

For a decreasing returns to scale market, by Theorem 4, faster growth dominates slower growth from x_0 to $\min\{\bar{x}, x^*\}$. Therefore, the fastest growth path is still optimal. One interesting observation is that, in this case, the optimal policy may be profit-taking at a very early stage, then gradually shift to subsidizing, and eventually switch back to profit-taking at the end of market expansion. This is because the maximal growth rate $bx(t)$ is small at the beginning; the growth cannot further speed up even with a subsidy. After the growth rate gradually picks up, a subsidy is needed to expedite growth. In the long run, the market generates positive profit at the minimum of the saturation size x^* or the market potential $\bar{x}$.

The optimal policy follows Theorem 4:

Proposition 9 *Consider the infinite horizon problem (2.13) with $\alpha < 1$ and subject to the conditions that $\mathcal{X} = \{x(t)|x(0) = x_0, \dot{x}/x \leq b, x(t) \leqslant \bar{x}\}$. Then the optimal growth path is*

$$x^*(t) = \begin{cases} x_0 e^{bt}, & t \leq \frac{1}{b} \ln \frac{\min\{\bar{x}, x^*\}}{x_0} \\ \min\{\bar{x}, x^*\}, & t > \frac{1}{b} \ln \frac{\min\{\bar{x}, x^*\}}{x_0} \end{cases}$$

2.4.4 Fund raising limits

In the base model, we implicitly assumed that there is an abundant amount of subsidy budget available within a short period of time. In reality, subsidies must be funded and typically there are constraints on the ability to raise funds that are a function of the current market size. We analyze a stylized model of such a constraint next.

At time t, the profit rate is $\pi g(\gamma, x)$. If $\pi < 0$, it means the market is subsi-

dized, and the platform is losing money at the rate of $\pi g(\gamma, x)$. Consider the following constraint:

$$-\pi(t)g(\gamma(t), x(t)) \leq mx(t) \tag{2.20}$$

Constraint (2.20) requires that the loss from subsidizing the two-sided market at time t should be at most a fraction m of the market size at that time; the larger the market, the larger the subsidy budget. This reflects the fact that in reality, subsidy budgets come from rounds of investment with investors willing to provide money as firms prove that they can grow their target markets. Thus, a larger market size justifies more funding. Hence, we model this using a multiple m of the market size $x(t)$ as the upper bound for the amount of funding available at a given time t. We call (2.20) the *fund-raising budget* and m the *funding multiple*.

By the state equation (2.12), $-\pi g(x, \gamma) \leq mx$ is equivalent to $\dot{x} \leq (v - c)g(\gamma, x) - (\beta_0^s \gamma + \beta_0^d(1-\gamma) - m)x$. We require the funding multiple m not be too small (e.g., $m > max\{\beta_0^s, \beta_0^d\}$) such that growing to $\bar{x}$ from x_0 is feasible.

In contrast to the relative growth rate constraint, when the fund-raising budget (2.20) is binding, the market balance does not directly affect the profit rate $mx(t)$, but affects how fast the market grows. Thus, the optimal balance from Theorem 1 remains optimal here since it maximizes the growth rate for any given market size x. From this perspective, the fund-raising budget (2.20) can also be thought of as an upper bound on the growth rate.

We show that the optimal policy has a similar structure as in the previous case:

Proposition 10 *Consider the infinite horizon problem (2.13) with $\alpha > 1$ and subject to the conditions that $\mathcal{X} = \{x(t) | x(0) = x_0, -\pi g(x, \gamma) \leq mx, x(t) \leq \bar{x}\}$, where*

$m > \max\{\beta_0^s, \beta_0^d\}$. *Let* $F(t)$ *be the solution to the differential equation*

$$\dot{x} = (v-c)g(\gamma^*(x), x) - (\beta_0^s \gamma^*(x) + \beta_0^d(1-\gamma^*(x)) - m)x, x(0) = x_0 \quad (2.21)$$

If $\bar{x} > \tilde{x}_b$*, it is optimal to grow the market as fast as possible to* $\bar{x}$*, i.e.,*

$$x^*(t) = \begin{cases} F(t), & 0 \le t \le F^{-1}(\bar{x}) \\ \bar{x}, & t > F^{-1}(\bar{x}) \end{cases} \quad (2.22)$$

If $\bar{x} < \tilde{x}_b$*, it is optimal not to grow the market, i.e.,* $x^*(t) = x_0$*,* $t \ge 0$*, where* $\tilde{x}_b > x_0$ *is uniquely defined such that*

$$\int_0^{F^{-1}(\tilde{x}_b)} e^{-\rho t} G_x(\gamma^*(F(t)), F(t)) F'(t) dt = 0 \quad (2.23)$$

Note that as the funding multiple m goes to infinity, $\dot{x}$ goes to infinity, and $F^{-1}(\bar{x})$ approaches 0. Then the optimal path (2.22) in Proposition 10 gradually approaches the optimal path in Proposition 6, i.e., $x(t)$ jumps from x_0 to $\bar{x}$ at time 0 and stays at $\bar{x}$. Recall that without a bound on the growth rate, an increasing returns to scale market is viable if the market size is above $\tilde{x}$. The following lemma shows that, the threshold size $\tilde{x}$ without the growth rate constraint is lower than that with the fund-raising constraint.

Lemma 3 $\tilde{x}_b > \tilde{x}$*, where* $\tilde{x}_b$ *is defined by (2.23) and* $\tilde{x}$ *is defined by (2.18).*

This comparison reveals that when the growth rate is bounded, a larger market potential $\bar{x}$ is required for the market to be profitable. Combining with the results in Proposition 10, this implies that the funding multiple m and the market potential $\bar{x}$ complements each other; if sufficient funds cannot be raised fast enough, then even if the market has potential profitability (i.e., $\bar{x} > \tilde{x}$), there is not enough momentum to jump start the growth. On the flip side, if the market

limit $\bar{x}$ is small, even if the budget is abundant and allows impulse-like growth, it does not help because the potential profitability is too small. The intuition is that early growth is expensive in an increasing returns to scale market due to the low output-to-size ratio. Slow growth means that the market is stuck in an inefficient phase for even longer, and hence a large long-term reward is required to compensate for the high growth cost.

A decreasing returns to scale market, however, is self-sustaining even early on. Organic growth without subsidy is not only feasible but also generates a positive profit. Hence, the fund-raising budget (2.20) only affects how fast we can grow, but it does not change the decision about whether or not to grow:

Proposition 11 *Consider the infinite horizon problem (2.13) with $\alpha < 1$ and subject to the conditions that $\mathcal{X} = \{x(t)|x(0) = x_0, -\pi g(\gamma, x) \leq mx, x(t) \leq \bar{x}\}$, where $m > \max\{\beta_0^s, \beta_0^d\}$. $F(t)$ is the solution to the differential equation*

$$\dot{x} = (v - c)g(\gamma^*(x), x) - (\beta_0^s \gamma^*(x) + \beta_0^d(1 - \gamma^*(x)) - m)x, x(0) = x_0 \quad (2.24)$$

Then the optimal growth path is

$$x^*(t) = \begin{cases} F(t), & t \leq F^{-1}(\min\{x^*, \bar{x}\}) \\ \min\{x^*, \bar{x}\}, & t > F^{-1}(\min\{x^*, \bar{x}\}) \end{cases}$$

2.4.5 Matching function

In the previous discussion, we required the production function $g(s, d)$ to satisfy certain smoothness conditions (Assumption 1). And we focused our attention on strictly increasing and decreasing returns to scale markets. A natural question then is: what growth policy is optimal in a constant returns to scale market?

The perfect matching function, $g(s,d) = A\min\{s,d\}$, is constant returns to scale in (s,d). It characterizes a frictionless market in which supply and demand are matched instantaneously as long as both inputs are positive. Although this function does not satisfy the smoothness conditions in Assumption 1, its simplicity allows us to derive the optimal policy directly without Theorem 1 and 5. Hence, we use this function as an example to characterize optimal growth for a constant returns to scale market.

Under the market size-and-balance reformulation, the production function can be written as $g(\gamma,x) = A\min\{\beta_1^s\gamma, \beta_1^d(1-\gamma)\}x$. As before, we show the optimal growth policies in the space of (γ,x).

Proposition 12 (Optimal Balance for a Perfect Matching Market) *In a perfect matching market, if*

$$-(v-c)A\beta_1^d < \beta_0^s - \beta_0^d < (v-c)A\beta_1^s \tag{2.25}$$

then the optimal balance is the output-maximizing balance, i.e.,

$$\gamma^*(x) = \frac{\beta_1^d}{\beta_1^d + \beta_1^s} \tag{2.26}$$

Otherwise, the optimal balance is the durability maximizing balance, i.e.,

$$\gamma^*(x) = \mathbb{1}_{\beta_0^s < \beta_0^d} \tag{2.27}$$

The fact that the optimal balance is constant is not surprising. Recall that by Theorem 1, the optimal balance is a monotone function in market size, and the direction of the change depends on the interplay between attrition and adoption. For a constant returns to scale market like the perfect matching market, however, the relation between attrition and adoption does not change with market size.

Whichever force is stronger will always be stronger, and the optimal balance just maximizes that force. If condition (2.25) holds, adoption is stronger, and thus the optimal balance stays at the output-maximizing balance (2.26); if not, attrition is stronger, then the optimal balance stays at the durability maximizing balance (2.27).

Another way to understand (2.25) is from the perspective of redistributing the surpluses. Rewrite the condition as $A(v-c) > \max\left\{\frac{\beta_0^s-\beta_0^d}{\beta_1^s}, \frac{\beta_0^d-\beta_0^s}{\beta_1^d}\right\}$. $A(v-c)$ here is the maximal profit margin per transaction without external subsidy; $\frac{\beta_0^d-\beta_0^s}{\beta_1^d}$ ($\frac{\beta_0^s-\beta_0^d}{\beta_1^s}$) is the monetary value required to increase demand (supply) to meet supply (demand). If (2.25) does not hold, it means the market cannot maintain $s = d$ without subsidy. Such a market cannot be viable, because unlike an increasing returns to scale market, the output per unit size is fixed; growth will always require subsidy, and there is no future reward. Later we will show that (2.25) is indeed a necessary condition for a perfect matching market to be viable.

Next, we show that the optimal growth policy is similar to an increasing returns to scale market:

Proposition 13 (Optimal Growth Policy for a Perfect Matching Market) *In a perfect matching market, it is optimal to grow the market to its upper bound as fast as possible given that the market is viable. If not, it is optimal not to grow the market at all. Moreover, a market is viable if its parameters satisfy*

$$\rho < \frac{(v-c)A\beta_1^s\beta_1^d - (\beta_0^s\beta_1^d + \beta_0^d\beta_1^s)}{\beta_1^s + \beta_1^d} \tag{2.28}$$

Lemma 4 *Condition (2.28) implies condition (2.25).*

Hence, a viable market requires that adoption dominates attrition. Moreover, the discount factor ρ should be sufficiently small.

2.4.6 Summary of main results

Combining the result on the optimality of fast growth in an increasing, decreasing, and constant returns to scale market, our analysis suggests that subsidizing during all phases of market growth is warranted – not only during the initial phase of increasing returns in order to reach break-even size, but also later as markets mature and experience decreasing returns. Subsidizing growth in mature markets is desirable in order to reach the limiting saturation size more quickly. While we do not dismiss the fact that these findings might not be robust to changes to our particular specification of market output and supply and demand growth, the results nevertheless provide intriguing theoretical evidence that there may be more to the intense focus on subsidized growth in industries like ride-sharing than meets the eye. It suggests that highly subsidized growth is plausibly an economically optimal growth strategy and not just the result of "irrational enthusiasm" on the part of company founders and private-market investors.

2.5 Numerical example

To illustrate the results of our model, we next apply it to a stylized numerical example of a ride-sharing market with a Cobb-Douglas production function, $g(s,d) = As^{\alpha_s}d^{\alpha_d}$. Table 2.1 shows the parameters used in the numerical example.

Table 2.1: Parameters used in numerical examples

Parameter	Value	Notes
v	18.30 \$/trip	
c	3.75 \$/trip	
β_0^s	0.5 /year	2 year avg. life
β_0^d	0.2 /year	5 year avg. life
β_1^s	0.067 drivers/dollar	\$15 per driver adoption cost
β_1^d	0.100 riders/dollar	\$10 per rider adoption cost
ρ	0.15	15% per year
α_s	0.800	Incr. returns
α_d	0.400	
α	1.200	
A	0.05	
α_s	0.300	Decr. returns
α_d	0.286	
α	0.586	
A	1,000	

The unit of time is years, transactions are trips, and the stock of supply and demand is measured in terms of the number of riders and drivers on the platform. Riders have a value of \$18.30 per trip and drivers have a cost of \$3.75 per trip. Drivers have a mean lifetime on the platform of 2 years while riders have a mean lifetime of 5 years, so demand is more durable than supply. In terms of elasticity of growth to surplus, an increase of one dollar in the surplus for riders results in a 10% increase (per year) in the number of riders, while an increase in one dollar of surplus for drivers results in a increase of 6.7% increase (per year) in the number of drivers. This in turn implies an adoption cost of \$10 per rider and \$15 per driver.

2.5.1 Market size and balance trajectories

Figures 2.1 and 2.2 illustrate the growth paths and total profits resulting from alternative (non-optimal) subsidy policies subject to fund-raising budget constraints. Figure 2.1 shows the growth paths and total discounted profit for the increasing returns case, comparing a high-subsidy policy of $m = 1$, a low-subsidy policy of $m = 0.2$, and an unsubsidized policy of $m = 0$, for the case where the market potential is $\bar{x} = \$100\text{M}$. The initial market size is $x_0 = \$1\text{M}$. Note in this case the high-subsidy policy yields a total discounted profit of $\$26,404,464$ and the low-subsidy policy yields a total discounted profit of $-\$782,162$, while the total discounted profit from the zero-subsidy policy is zero. Note the zero-subsidy policy is not able to reach the critical market size, which is $x_c = \$8,334,002$, and thus also cannot reach the market potential $\bar{x}$. For the low-subsidy policy, although the market is able to reach the market potential, the total discounted profit goes negative due to the prolonged inefficient phase. For comparison, the optimal discounted profit of the subsidy shock policy is $\$1,000,000 - \$100,000,000 + \$23,028,440/0.15 = \$54,522,933$, where $\$23,028,440$ is the profit rate when the market size is at it potential, $\bar{x}$ and 0.15 is the discount rate.

Figure 2.2 show a similar comparison of policies for the decreasing returns case, comparing a high-subsidy policy of $m = 1$ to a zero-subsidy policy of $m = 0$. The initial market size is $x_0 = 100$. Here, the total discounted profit from the subsidy policy is $\$383,682,517$, while the total discounted profit from the zero-subsidy policy is $\$363,378,728$. Note in this case, both policies are able to reach the saturation size, which is $x^* = \$206,543,541$. However, the discounted total profit of the subsidized policy is greater. For comparison, the optimal discounted

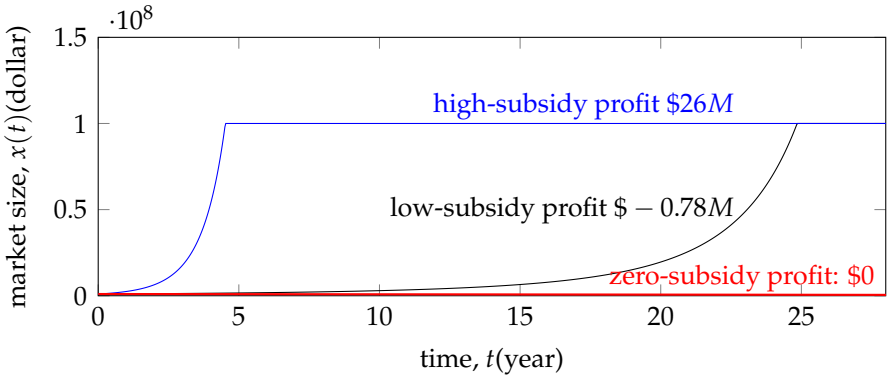


Figure 2.1: Policies and profits under fund-raising budgets (increasing returns to scale)
Note. High-subsidy policy: $m = 1$; low-subsidy policy: $m = 0.2$; zero-subsidy policy: $m = 0$; $\bar{x} = \$100M$; $x_0 = \$1M$

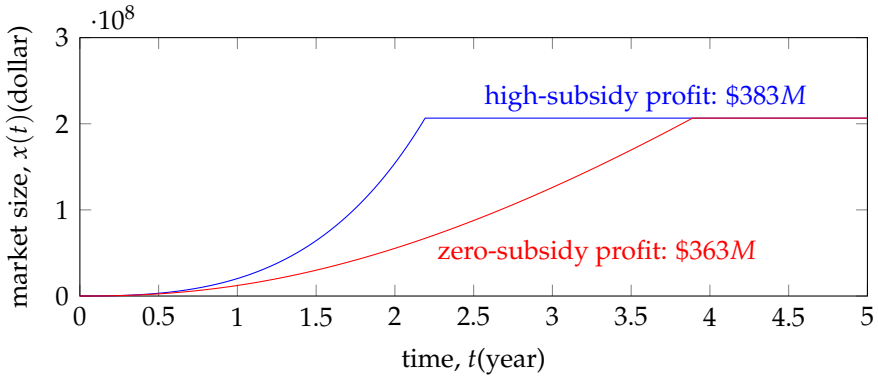


Figure 2.2: Policies and profits under fund-raising budgets (decreasing returns to scale)
Note. High-subsidy policy: $m = 1$; zero-subsidy policy: $m = 0$; $x_0 = 100$;

profit of the subsidy shock policy is $\$100 - \$206,543,541 + \$97,805,254/0.15 = \$445,491,589$, in which $\$97,805,254$ is the optimal profit rate and 0.15 is the discount rate.

Figures 2.3 and 2.4 show how optimal market balance changes as market size evolves in the numerical examples for increasing and decreasing returns to scale, respectively. Since demand is more durable than supply in this example, the durability-maximizing balance is $\gamma = 0$. The output-maximizing balance is

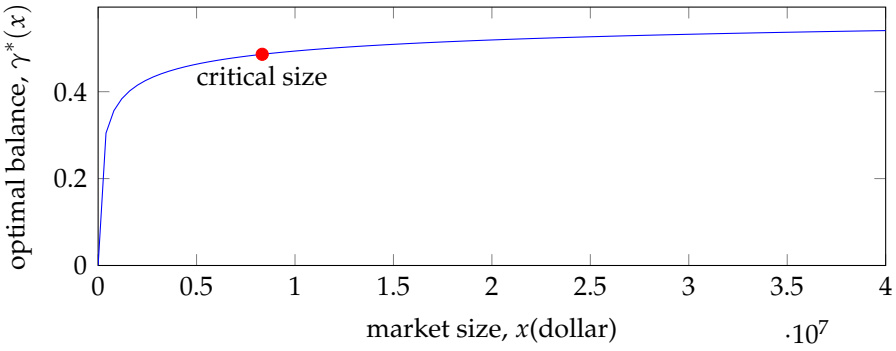


Figure 2.3: Evolution of optimal balance in the change of market size (increasing returns)

$\alpha_s/(\alpha_s + \alpha_d)$, which is equal to 0.667 in the increasing returns case and 0.512 in the decreasing returns case.

As is shown in Figure 2.3, in the increasing returns case the market balance starts out at zero (the durability-maximizing balance) and increases toward the output-maximizing balance as size increases, albeit slowly once the size approaches the critical size $x_c = \$8,334,002$. Market balance at the critical size is approximately $\gamma = 0.49$.

In contrast, as shown in Figure 2.4, market balance in the decreasing returns case starts out at the output-maximizing balance of $\gamma = 0.512$ when the market size is small and decreases toward zero (the durability-maximizing balance) as the market grows. Figure 2.4 shows the optimal balance is approximately $\gamma = 0.37$ when the market reaches the saturation size of $x^* = \$206,543,541$.

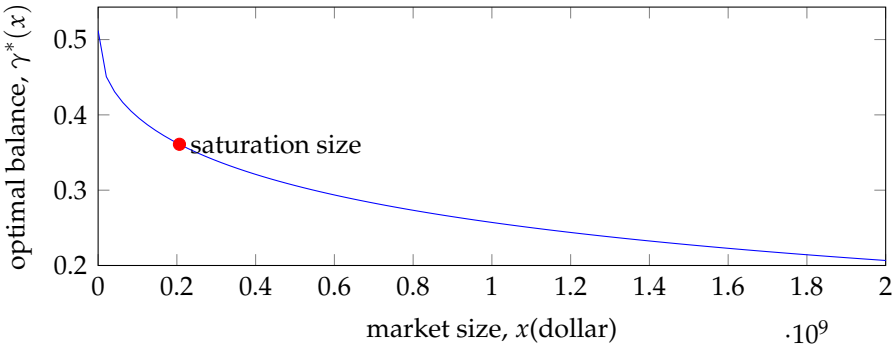


Figure 2.4: Evolution of optimal balance in the change of market size (decreasing returns)

2.5.2 Supply, demand and price trajectories

As we have briefly discussed in Section 2.3.4, once the trajectory of market size $x(t)$ and balance $\gamma(t)$ are determined, the original state variable $s(t)$, $d(t)$ and control $p_s(t)$, $p_d(t)$ can be recovered. Here we use Figure 2.5 and 2.6 as an illustration of the recovered trajectories. Both figures are calculated using the high-subsidy ($m = 1$) growth path in Figure 2.1 and its corresponding optimal balance trajectory.

Figure 2.5 shows the optimal trajectory of supply and demand as a function of time. Similar to the trajectory of $x(t)$, $s(t)$ and $d(t)$ gradually increase until reaching the market potential, and stay constant thereafter.

Figure 2.6 illustrates how the optimal pricing strategy changes over time. Before the market reaches its potential, the driver's wage per trip p_s is set to be higher than the price per trip p_d, reflecting the subsidy on each trip to expedite growth. Moreover, the difference between the wage and the price, $p_s - p_d$, is the largest at $t = 0$ and gradually decreases over time, showing that subsidies are more aggressive at the earlier stage. Once the market potential is reached, the

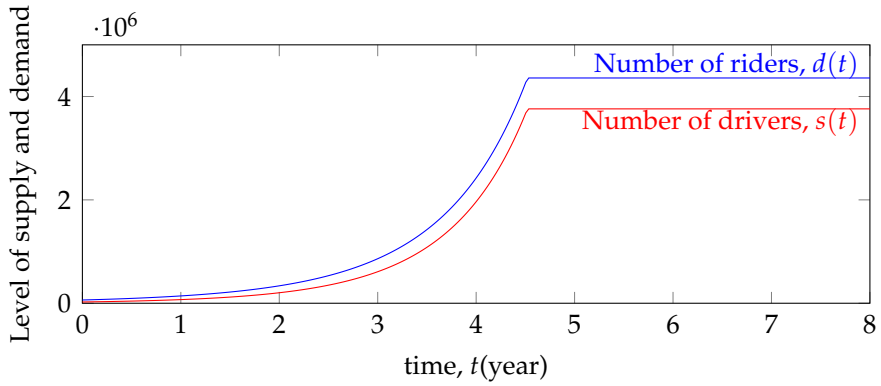


Figure 2.5: Trajectories of supply and demand (increasing returns to scale)
Note. Under the high-subsidy policy: $m = 1$; $\bar{x} = \$100M$; $x_0 = \$1M$

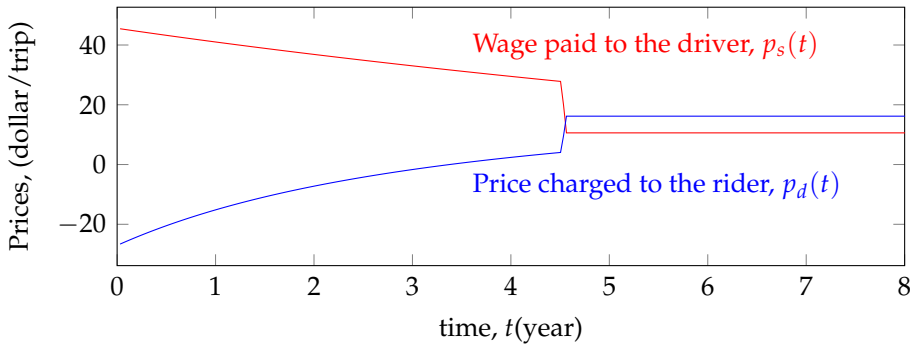


Figure 2.6: Trajectories of price policies (increasing returns to scale)
Note. Under the high-subsidy policy: $m = 1$; $\bar{x} = \$100M$; $x_0 = \$1M$

wage and the price switch to $p_s < p_d$ and stay constant from the time onward, meaning that the platform makes a positive profit thereafter.

2.6 Conclusion

Our model and results provide a theoretical framework to understand optimal growth in two-sided markets. The state-space reduction to a scalar market size together with a market balance control provide both conceptual clarity and

analytical tractability that should prove useful when analyzing other variations of the growth problem.

Our results on optimal balance provide useful insights into the factors that should be considered when deciding how to invest in the supply or demand sides of the market. In particular, it highlights the importance of the attrition (durability) of supply and demand, the adoption cost of generating supply and demand, and the elasticity of output produced by supply and demand – and how these three forces evolve in relative importance as markets grow.

We also characterize important size thresholds: the critical size at which a market becomes self-sustaining (increasing returns) and the saturation size at which optimal growth reaches its limit (decreasing returns). It would be interesting to map the predictions of our model empirically to determine what critical and saturation market sizes look like in real-world markets.

Lastly, we showed that in general faster growth is better and that subsidy shocks are optimal, both in early phases of growth when returns to scale are increasing and in later phases of growth when returns to scale are decreasing. The findings provide theoretical evidence of the value of bold subsidized growth strategies in two-sided markets – even when platforms have a monopoly and do not face competitive entry threats. It would be interesting to see if these extreme subsidy strategies hold up under other modeling assumptions and to what extent there is empirical evidence supporting their optimality.

Extending our theory of optimal growth to competing two-sided platforms is another worthwhile extension. It is important to understand to what extent the results here are robust to environments in which two or more platforms compete.

Toward that end, we have work in process on an oligopoly version of our growth model, which is forthcoming.

Acknowledgment.

The authors gratefully acknowledge the helpful feedback of several academic and industry colleagues on this work, especially Karan Girotra. Keith Chen also contributed to early ideas about this topic while he and the second author were colleagues together at Uber. We also acknowledge both Cornell Tech and Lyft for their support in making this research possible.

CHAPTER 3

CAPTURING THE BENEFITS OF AUTONOMOUS VEHICLES IN RIDE-HAILING

3.1 Introduction

Autonomous vehicles (AVs) hold the promise of significantly reducing the cost of ride-hailing services. But precisely how AVs will impact market outcomes is not well understood. Toward this end, we develop an economic model of autonomous vehicle (AV) ride-hailing markets in which uncertain aggregate demand is served with a combination of a fixed fleet of AVs and an unlimited, perfectly-elastic supply of human drivers (HVs). We analyze market outcomes under two dispatch platform structures (common platform vs. independent platforms) and two levels of AV competition (monopoly AV vs. competitive AV). Table 3.1 shows motivating examples from current AV firms for each market configuration.

Table 3.1: Motivating examples for different dispatch structures and levels of competition

	Common Platform	**Independent Platform**
Monopoly	Uber's own AV fleet	GM Cruise's AV ride-hailing network
Competitive	Lyft open platform	Tesla robotaxi

Note. *Note.* 1. Monopoly, common platform market: A ride-hailing company like Uber deploys its own AV fleet on the same dispatch platform as drivers. (Shetty 2020) 2. Monopoly, independent platform market: An auto maker like GM develops its own ride-hailing service supplied by AVs. (Bellan 2021) 3. Competitive, common platform market: A ride-hailing platform like Lyft collaborates with multiple AV suppliers (Waymo, Aptiv...) and shares the same dispatch network. (Blog 2021) 4. Competitive, independent platform market: Tesla announced plans to build its own ride-hailing network, where individual owners can deploy their Tesla to make profit. (McCracken 2021)

We look at a market that is subject to aggregate demand variability hour to

hour and faces a “loose" market for labor [1] – the typical setting for real-world ride-hailing markets. The incumbent HV ride-hailing market is assumed to be perfectly competitive: riders and drivers are both price takers and market prices for rides and labor are determined by a zero-profit equilibrium. This is a stylized representation of the fact that the current ride-hailing industry based on independent contractor drivers is highly competitive on both the demand and supply side. AVs in contrast are provided by suppliers who we assume make irreversible fleet capacity (quantity) choices. HVs are assumed to incur only variable per hour costs related to their variable vehicle costs and the reservation value of driver time. AVs incur both a fixed investment cost per calendar hour and a variable cost per operating hour – the sum of which is assumed to be lower than the HV's total variable cost.[2]

A salient feature of our model is the network density effect. Specifically, while operating, AVs and HVs are in one of three states: waiting for dispatch, traveling to pick up riders, or transporting riders. A higher density of supply and demand leads to a shorter average distance between riders and vehicles, leading to shorter pickup times and hence higher on-trip utilization. (See for example Castillo, Knoepfle, and Weyl 2017, Nikzad 2017). Since a vehicle only generates revenue when it is serving demand, the less time it spends waiting or traveling to pick up riders, the more revenue it can generate per unit time. This effect of density on revenue per unit time plays an important role in driving market outcomes in our model.

[1] A labor market is loose if the supply of potential drivers is abundant relative to demand. More precisely, the equilibrium price between supply and demand is in the elastic region of the aggregate market revenue function. See Asadpour, Lobel, and Ryzin 2019 for discussion of lose and tight labor markets in the context of ride-hailing.

[2] It is possible that eventually the adoption of AVs may achieve a point where there is a huge potential pool of AVs from individual owners that can stand by and provide ride-hailing service on the point, like the market for HVs nowadays. Nonetheless, this is not likely to happen before AVs become viable for commercial use, which is the case we focus on.

Our analysis provides a number of important insights. First, we show AVs will, in most plausible cases, not fully replace HVs even if AVs have lower total cost (the sum of their fixed and variable cost). The reason is the lower flexibility of AVs: AVs are expense assets that must be pre-committed to a market and incur fixed costs regardless of whether they are operating or not. This means AVs are cost-competitive only if their utilization is sufficiently high. Indeed, a supplier's decision on their AV capacity involves balancing the underage cost of missing rides in high demand scenarios against the overage cost of wasting capacity in low demand scenarios. With sufficient variability in demand scenarios, it is therefore not cost-effective to AVs to serve all demand. In contrast, HVs – which are operated by a large pool of independent contractors who typically use their vehicles for other purposes – have greater flexibility to vary their aggregate supply and do not incur the same degree of fixed costs. [3] Hence, HVs effectively have no capacity constraint and, in the case of perfectly-elastic supply, will fulfill all demand as long as the revenue per unit time exceeds the HV variable cost. The existence of this flexible, elastic market for HVs is critical determinant of market outcomes for AV service.

More surprisingly, we show that the lower cost of AVs does not necessarily translate into lower prices for ride-hailing service; rather, the price impact of introducing AVs to a market is ambiguous and depends critically on both the dispatch platform structure and the level of competition.

In the extreme case, we show that if AVs and HVs operate on independent dispatch platforms and there is a monopoly AV supplier, then prices are even

[3]The extent to which ride-hailing drivers commit to the fixed cost of a vehicle of course varies from driver to driver. That said, human contract drivers have many alternatives for using their vehicles for other contract work (e.g. delivery) and/or personal use that are quite distinct from an AV operating in a commercial fleet. This is why, in our stylized model, we assume HVs have purely variable costs.

higher than they are in a pure-HV market and consumers would be better off without AVs. These higher prices are driven by the interaction of market power and the density effect. If AVs and HVs operate on separate dispatch platforms, then adding AVs to the market reduces the average density on the HV platform. As a result, HVs will have a lower utilization rate and (until prices adjust) earn lower revenues per unit time. HVs will then respond by exiting the market until the market price is high enough to compensate for the lowered utilization. Through this density effect, reducing the market share for HVs raises the market price. A monopoly AV provider recognizes and exploits this price effect of reducing HV market share by increasing its supply, and hence the resulting market price in the mixed AV-HV regime is in fact higher than in the pure-HV market. This outcome, while bad for consumers, produces the highest total AV profit.

In contrast, with a common dispatch platform for both AVs and HVs, the average density remains the same as long as the total number of AVs and HVs does not change. In other words, while the share of demand served by HVs declines, there is no utilization loss from the entry of AVs. Even so, we find that the equilibrium price does not necessarily decrease. Whether the price is lower or not depends on the level of AV competition. If AVs are owned by a monopoly supplier, then the equilibrium price is the same as in a pure-HV market in every demand scenario. The reason is that a profit-maximizing monopoly AV supplier will seek to limit its supply to increase prices. But the market price can never be higher than the pure-HV equilibrium price else HVs will enter until the price returns to the HV equilibrium price. Hence, the monopoly AV supplier chooses supply quantities such that prices in all scenarios are identical to the pure-HV prices.

Indeed, the only market structure that leads to unambiguously lower prices (and expanded service) in all demand scenarios is when AVs and HVs operate on a common dispatch platform and the AV supply is competitive. In this case, for high demand scenarios where the entire AV fleet participates along with some HVs, the pure-HV equilibrium price is realized because, on a common platform, the density and utilization of vehicles is the same as in the pure-HV market. In scenarios where demand is too low to support HV participation, the market price is determined by a zero-variable-profit equilibrium among the competitive AV suppliers, who have variable costs strictly lower than HV variable costs. The result is a market price that is strictly lower than the pure-HV market price.

Our results illustrate the critical role market structure plays in realizing potential welfare gains from AVs. A common dispatch platform is necessary to ensure that economies of density are not adversely affected by introducing AVs in a market. At the same time competition among AV suppliers on this common platform is needed to ensure that market prices will in fact decline in scenarios where there is ample AV supply to serve all demand. The worst outcome for consumers is a monopoly AV supplier providing service on an independent platform; such a supplier can exploit a combination of market power and the price effects of reduced density in the HV market to drive prices even higher than in the pure-HV market. Unfortunately, this bad outcome for consumers is the preferred one for AV suppliers as it generates the highest AV profit. These findings reflect, in a stylized way, basic tensions emerging in the ride hailing industry between platform providers, like Lyft and Uber, who are promoting open access to multiple AV suppliers on a common platform co-mingled with existing human drivers; and new AV entrants, like GM/Cruise and (to a lesser extent) Waymo, who are pursuing a strategy of building their own stand-alone

AV-only services.

3.2 Related literature

Research on the economic impact of AVs to date is limited but growing. Ostrovsky and Schwarz 2019 study the relationship among autonomous transportation, carpooling, and road pricing. They show that AVs make carpooling and road pricing more attractive by eliminating the labor cost and rider's disutility from congestion. Their model focuses on a pure AV market, while ours looks at a mixed HV-AV market. Siddiq and Taylor 2022 study the implications of AVs on the competition between two ride-hailing platforms by examining three performance measures: the platform profit, drivers' welfare, and social welfare (defined as the sum of the former two terms). The focus of their analysis is the supply-side implications. In contrast, our key insight is centered around consumers on whether AVs will lead to a lower price. While we also examine optimal supplier decisions (i.e., fleet sizing), our primary focus is on the consumer welfare implications of various market structures not the profit consequence for competing platforms. Furthermore, their model does not consider the dispatch platform structure and density effects on utilization; thus, in their model, AVs will always make consumers better off (due to the increase in supply) and drivers worse off (due to the displacement of human labor). However, in our model, consumers are not better off with AVs under certain market structures.

Our paper is also related to the burgeoning literature on ride-hailing. Besbes, Castro, and Lobel 2021 study the spatial capacity planning for service platforms like ride-sharing. Banerjee, Kanoria, and Qian 2018 study the dynamic assign-

ment control of supply in a closed network where the number of supply units is finite, and a controller dynamically assign nodes when new demand arises. Bimpikis, Candogan, and Saban 2019 study the spatial pricing for networks such as ride-hailing, where riders are heterogeneous in destinations and willingness-to-pay and drivers relocate to maximize their expected earnings. In addition, Cachon, Daniels, and Lobel 2017, Bai et al. 2019, Hu and Zhou 2020, Gurvich, Lariviere, and Moreno 2019, Lobel, Martin, and Song 2021 and Dong and Ibrahim 2020 also study the pricing and optimal staffing in the ride-hailing market. This literature mainly considers independent contractor drivers as supply. Therefore, these models focus more on how to optimally dispatch, price and match the flexible workforce in complicated market settings, such as those with spatial and temporal demand uncertainties. In contrast, our work focuses on the market-level economic outcome when both drivers (HVs) and AVs are capable of supplying the ride-hailing market, and the two types of supply are distinct in the level of flexibility and cost structures. The key operational feature in our model lies in the HVs and AVs' service process; namely, that they must spend time waiting for requests and time picking up riders before completing a trip – an important characteristic of real-world ride-hailing service. Furthermore, among all these works, Lobel, Martin, and Song 2021 and Dong and Ibrahim 2020 are the only ones that investigate a hybrid model of independent contractors and permanent employees. While there are some similarities between the employee model and AVs in their cost structures, these papers focus on a platform's internal staffing decision; thus, they do not consider the structure of competitive in the wider ride-hailing market.

Because we consider multiple platforms, there are connections to the literature on platform competition. The seminal works of Rochet and Tirole 2003, Rochet

and Tirole 2006 analyze the competition among two-sided markets and discuss the optimal price allocation and user surplus under different governance structures. Furthermore, the literature on competition among ride-hailing platforms has been rapidly growing (for example, Bryan and Gans 2019, Ahmadinejad et al. 2019, Nikzad 2020, Loginova, Wang, and Liu 2019, Chen et al. 2020, Ahmadinejad et al. 2019, Tan and Zhou 2020, Bai and Tang 2018, Cohen and Zhang 2018, Bernstein, DeCroix, and Keskin 2021). Our paper makes an interesting contribution to the platform competition literature in that the two dispatch platform structures (common vs. independent) can be thought as different types of competition between the AV and HV supply. Under a common platform market, AVs and HVs are in coopetition (see also Cohen and Zhang 2018 for discussions on coopetition) where they compete over supply quantities but collaborate for scale economies; under an independent platform market, AVs and HVs are in competition for both supply quantities and scale. Our findings show that the competition between AVs and HVs (in the independent platform market) can lead to the highest price and worst outcome for consumers, due to the efficiency loss from separating the dispatch platforms.

Our paper is also related to work on market equilibria and driver utilization in ride-hailing. Specifically, Hall, Horton, and Knoepfle 2019 study the impact of fare increases on equilibrium driver earning. They show that driver utilization is decreasing in price, and thus increased fares do not lead to earnings increases due to the reduced utilization. Asadpour, Lobel, and Ryzin 2019 study how utilization-based minimum earning regulations affect the stability of marketplaces like ride-hailing. In our paper, different market structures (common platform vs. independent platform) of the mixed AV-HV market have a

significant impact on the equilibrium price due to their effect on HV utilization.[4]

The next two streams of literature are not directly related to ride-hailing markets or the AV technology, but they share similar high-level trade-offs or modeling features. In the electricity market, it is a common practice to have mixed power generation from different energy sources, such as hydro-electric power and natural gas turbines. These power generation technologies differ in their costs structure and flexibility. In the presence of uncertain and volatile demand for electricity, a mix of generating technologies is used to serve demand, with some technologies serving the base-load demand and others the peak-load demand, based on their cost structure. (See Harris 2006 for a comprehensive overview of the economics of electricity markets.) This is consistent with our findings that a combination of AVs and HVs achieve the best market outcome; AVs are more suited to serving base-load demand due to their low total cost when they can be highly utilized, while HVs are better suited to serving any residual peak demand due to their flexibility (no fixed cost, albeit higher variable cost).

Lastly, our characterization of the HV ride-hailing market has some similarity to the taxi market literature, e.g. Douglas 1972 and Arnott 1996. In this literature, the main driving forces are the effect that increasing both trips and taxis has on reducing waiting times, and the focus is on the best way to regulate and subsidize the market. Our model also includes this density effect. Moreover, our

[4]One difference between our definition of utilization and those in Hall, Horton, and Knoepfle 2019 and Asadpour, Lobel, and Ryzin 2019 is that, we explicitly consider the pickup time as part of the driver process; on average, a trip consists of three stages: waiting for requests, picking up, and completing the trip; our utilization is defined as the proportion of time drivers spend on completing the trip. In contrast, Hall, Horton, and Knoepfle 2019 and Asadpour, Lobel, and Ryzin 2019 only consider the two stages of waiting for requests and completing the trip, and thus tend to be higher than our utilization. Nonetheless, all the works illustrate the importance of utilization as a key metric for platform efficiency in real-world ride-hailing markets.

assumptions of a perfectly elastic HV supply with a fixed reservation earnings level and free entry also resemble those made in Douglas 1972. On the other hand, AVs have the unique characteristic of fixed capacity that must be pre-committed to a market with uncertain demand. Thus, our analysis generates new insights that have not been addressed by this literature.

3.3 Model

We next define our model in detail. We start with our economic and technical assumptions about HVs and AVs, then define our two dispatch platform structures (independent and common platform), and our two models of AV market competition (monopoly and competitive).

We start by summarizing the main assumptions of our model:

1. *AVs incur a fixed cost and require pre-commitment, while HVs incur only variable costs and can be utilized as needed. However, the sum of AVs' fixed and variable costs is strictly less than the HVs' variable cost.* The total cost advantage of AVs is the primary reason why AVs have the potential to reduce prices for customers and generate profits for ride-hailing companies.
2. *The HV supply is perfectly elastic and participation adjusts much faster than potential adjustments in AV supply.* In other words, HV supply adjusts quickly based on demand and market prices, while AVs are committed assets whose capacity cannot be changed easily. [5]
3. *The ride-hailing platforms make zero profit from HVs.* In practice ride-hailing

[5]The relaxation of the perfect elasticity assumption for HVs is discussed in Section 3.7.1.

firms face stiff price competition due to riders who multi-app and stiff wage competition from contract drivers who multi-app, leaving them little market power. Hence, we assume a zero-profit equilibrium for these firms. Indeed, low profits along with the high cost of human drivers are primary motivations behind ride-hailing platforms' investments in AV technology.[6]

4. *The supply of potential HV labor is plentiful.* We assume the labor market to be "loose" in the sense discussed in Asadpour, Lobel, and Ryzin 2019, which means that competitive market clearing prices are lower than the aggregate revenue maximizing price.[7]

3.3.1 AVs, HVs, and the market

Consider a ride hailing market (city or metro area) with two types of supply: HVs and AVs. HVs are operated by human contractor drivers who provide service using their own vehicle. HVs have a cost per hour on the platform w_0 (their reservation earnings level), which is the sum of variable vehicle cost per hour and the opportunity cost of the driver's time. All time spent on the platform (idle time, pickup time and trip time) is assumed to have the same cost per hour. We assume the supply market for HVs is perfectly elastic. That is, if drivers' realized earnings w exceeds their reservation earnings w_0, then an unlimited number of drivers is willing to enter; if the realized earnings w is less than w_0, no drivers are willing to participate. Hence, if in equilibrium there are a non-zero number of drivers participating in the market, then $w = w_0$. Let n_0 denote the number of HVs in the market.

[6] In Section 3.4, we provide more detailed discussion about this assumption.
[7] The implications of relaxing this assumption are discussed in Section 3.7.4

AVs are assumed to be fully capable of servicing all trips.[8] They are provided by fleet operators that make investment commitments to the number of AVs they will deploy in a market. Because AVs are capital-intensive assets that must be purchased in advance and committed to a single market, this fleet sizing decision is considered an irreversible investment. [9] AVs have a fixed cost per calendar hour of $c_f > 0$ and a variable cost per operating hour on the platform of $c_v > 0$. AVs are assumed to have a total cost advantage over HVs in the sense that the sum of the fixed and variable AV costs satisfy, $c_v + c_f < w_0$, where recall w_0 is the HV's reservation earnings. Were this not true, there would be no viable market for AVs.

Aggregate demand is stochastic. In particular, let M (a random variable) denote the "mass" of potential riders (market size) in a representative hour of operation. We assume M has a discrete distribution consisting of I demand scenarios, $M = \{m_i | i = 1, ...I\}$ with probability mass function $P(m_i), i = 1, \ldots, I$. The market price for a ride-hour of service is denoted by p. Demand (in ride-hours per calendar hour) in scenario i is given by $d_i(p) = m_i(1 - F(p))$, where the *rider value distribution* $F(p)$ is the fraction of riders whose monetary value for a ride hour v is less than p. For simplicity, we assume v is uniformly distributed on $[0, V]$. Thus, $F(p) = \min\{p/V, 1\}$. The inverse demand function is then given by $p_i(d) = V(1 - \frac{d}{m_i}), d \in [0, m_i]$. In different scenarios, the number of AVs deployed by the suppliers may vary, whereas the AV capacity remains the

[8] In reality, AVs will have significant restrictions on the trips they can serve, at least while the technology is maturing. These restrictions include: limits on travel speeds, night time driving, restricted pickup/drop-off locations, snow and ice restrictions, etc. For simplicity, we will ignore these restrictions. Alternatively, one can consider our model as applying to only the AV-addressable portion of the overall ride hailing market.

[9] Of course, it is possible to move AVs from one market to another at at cost. But this is not likely something that would make economic sense on a tactical (hourly) basis – though one could imagine that seasonal movements of fleets is certainly a possibility. Nevertheless, the core investment trade-off we analyze is effectively unchanged provided the supplier is committing the AV to a market for a sufficient length of time (one that spans multiple demand scenarios).

same across scenarios once chosen. We refer to the number of AVs deployed in a scenario as the *AV fleet size* in scenario i, to differentiate it from the AV capacity choice.

The product market for ride hailing service is assumed to be perfectly competitive. In other words, there is no quality difference between the service provided by AVs and HVs. HVs and AVs therefore compete on quantity and the market price is determined by their total supply, as in the Cournot model of competition.

Lastly, an AV or HV is assumed to be in one of three states: waiting for dispatch, enroute for pickup, or on-trip serving a ride. Let s denote the total number of open (idle) vehicles in the market and $t_1(s)$ denote the mean time to pick up a rider given s (the ETA function). From Kolesar 1975, this ETA function can be well approximated as a power (constant-elasticity) function of the number of vacant servers. That is,

$$t_1(s) = as^{-r}, \;\; 0 < r < 1$$

where a and r are parameters whose values depend on physical characteristics of the service region. Let t_2 denote the mean trip time. A complete trip involves the pickup time and the transport time, so the average total time a vehicle spends for a trip is $t_1(s) + t_2$, but the only revenue-generating part of the trip is the transport time t_2.

3.3.2 Dispatch platform structure

When introducing AVs to this existing HV market, we consider two possible dispatch structures: 1) AVs can be co-mingled with HVs on the same dispatch

platform or 2) HVs and AVs can serve rides on independent dispatch platforms. If AVs share the dispatch platform with HVs, we call this a *common platform market*. If AVs operate on their own platform, we call this an *independent platform market*. In both cases, we assume the platforms themselves operate at zero-profit and only serve to clear the market between the supply and demand sides by allowing the prices of supply and demand to adjust. [10]

Common platform market

In a common platform market, we assume symmetry in dispatch between HVs and AVs; that is, there is no differentiation made between AVs and HVs when assigning riders to vehicles.

As a result, both the total number of open cars and total demand served for each vehicle type are proportional to their fleet sizes. Specifically, when there are n_0 HVs and n_a AVs participating in the market, then by dispatch symmetry, HVs must be receiving a fraction $\alpha = n_0/(n_0 + n_a)$ of all dispatches and must constitute a fraction α of all s open cars. The dispatch symmetry assumption also implies that AVs and HVs have the same expected pickup time $t_1(s)$ and the same average trip time t_2. This means that the proportion of time that a vehicle spends serving trips – or *utilization* – is the same for AVs and HVs. Consequently, the demand served by AVs (or HVs) is also proportional to their fleet size.

[10] Alternatively, one can assume platform fees that are charged to either the supply or demand side are included as part of the variable costs of HVs and AVs and/or included in the definition of the demand function, so that market prices in our model are net of any platform fees. Admittedly, when the platform fee does not follow a fixed commission model (e.g., when the supply and demand side pricing are decoupled), our analysis may not carry through. Still, to our knowledge, the fixed-commission policy is a quite widely-adopted model. For example, Uber lists a fixed commission rate of 25% as its main payment policy (Uber 2022).

Independent platform market

In contrast, in an independent platform market, the expected pickup time for AVs and HVs are not necessarily the same. Consider introducing just one AV and operating it independently from an existing HV market with a fleet size of n_0 cars. Then the pickup time of the AV is going to be $t_1(1)/t_1(n_0) = n_0^r$ times longer than that of a HV. Clearly, the total demand that can be served by this single AV within a given time period (its capacity) will be much lower than that by an HV. Hence, the proportion of demand served by the AV is going to be less than $1/(n_0 + 1)$, implying that dispatch symmetry no longer holds.

3.3.3 AV competition

We also consider two levels of AV competition: 1) a monopoly AV market in which a single AV supplier provides all AV service, and 2) a competitive AV market in which each AV is operated by a separate supplier. These represent two stylized extremes of competition among AV suppliers.

Monopoly AV

In this setting, AVs are provided by a monopoly supplier that makes an irreversible decision on the AV capacity, denoted N, with the objective of maximizing its aggregate profit. The cost of this initial investment is $c_f N$ per hour. Then, in each scenario i with potential demand m_i, the supplier maximizes its variable profit by choosing the optimal number of AVs to operate subject to its capacity choice N.

Formally, the monopoly supplier's profit maximization problem is defined as

$$\max_{N} \sum_{1}^{I} P(m_i) \left\{ \max_{n \leq N} \pi_i^D(n) n \right\} - c_f N \tag{3.1}$$

where π_i^D denotes the variable profit per vehicle per hour an AV earns in scenario i, given an AV fleet size of n. $D = C, I$ represents the market structure, with I meaning the independent platform market and C meaning the common platform market.

Competitive AV

To analyze how competition affects market outcomes, we also consider the extreme case of a perfectly competitive AV market. In this case, one can think of each AV as being owned by an independent supplier, just as each HV is operated by an independent contractor. In contrast to a monopoly supplier that decides on the total number of AVs to invest, in the competitive case the decision is whether a potential owner (supplier) of a single AV invests or not. Conditional on deciding to invest in an AV, an AV supplier then decides whether to provide service in each scenario.

We assume a large market and an unlimited number of potential AV suppliers, so that in equilibrium each AV earns zero expected profit. This holds because if AVs generated positive expected profit, then potential AV suppliers would continue to enter the market until the expected profit dropped to zero. The AV capacity is thus determined by a zero-profit equilibrium.

More precisely, for each AV supplier, given an AV fleet size of n, the variable profit per AV is $\pi_i^D(n)$ per hour in scenario i, where $D = C, I$ represents the market structure as in Eq. (3.1). The AV supplier's decision is simple: if the

variable profit $\pi_i^D(n)$ is greater than zero, then the supplier will choose to provide service. If not, the AV supplier will be better off not providing service and earning zero variable profit in scenario i. Thus, aggregating all scenarios, the expected variable profit for each AV supplier is just $\sum_1^I P(m_i) \max\{\pi_i^D(n), 0\}$. Because potential AV suppliers can freely purchase AVs and enter the market, in equilibrium, the AV capacity is such that the expected variable profit is exactly the cost for purchasing an AV (i.e. c_f). Thus, the equilibrium capacity N is determined by

$$\sum_1^I P(m_i) \max\{\pi_i^D(N), 0\} = c_f \tag{3.2}$$

Note that the definition of π_i^D remains the same as in the monopoly case, but there is no longer a scenario choice about the number of vehicles to operate; all N vehicles will operate as long as variable profits are not negative.

3.4 Pure-HV market

Our baseline is the market equilibrium in a ride-hailing market served only by HVs. We then look at how this pure-HV market is impacted by the introduction of AVs. We assume that prices adjust instantaneously and that HVs enter and exit the market instantaneously according to the earnings they receive. Hence, equilibria over time are decoupled and it is sufficient to independently analyze the equilibrium within each given hour and demand scenario i.

Recall that it takes time $t_1(s)$ to pick up a rider when s cars are open, and the average transport time to serve a trip is t_2. Hence, for a given demand rate d, by Little's law the number of cars busy picking up or on trip is $d(t_1(s) + t_2)$. Given

a total supply of cars n, the number of open cars is then

$$s = n - d(t_1(s) + t_2)$$

Hence, the proportion of time that the HV fleet spends on completing trips is

$$u = \frac{dt_2}{s + d(t_1(s) + t_2)}$$

Thus, u is the HV fleet's on-trip utilization.

Moreover, riders only pay for the time HVs are on-trip, but not for the time HVs are waiting for rides or picking up riders. Thus, at demand rate d, an HV earns $p_i(d)u$ per hour on average. Because the supply of HVs are perfectly elastic, HVs will continue to enter the market until their hourly earnings are no longer above the reservation earnings w_0. Therefore, in equilibrium, we must have that $p_i(d)u = w_0$ so that the equilibrium utilization can be expressed as

$$u = \frac{w_0}{p_i(d)}$$

Substituting above, this implies in equilibrium we must have

$$\frac{w_0}{p_i(d)} = \frac{dt_2}{s + d(t_1(s) + t_2)} \tag{3.3}$$

Rewriting the equation gives

$$dp_i(d)t_2 = w_0(s + dt_1(s) + dt_2)$$

Note the left-hand side above is simply the total market revenue per hour, $r_i(d) \equiv dp_i(d)t_2$, while the right-hand side is the total driver cost per hour (individual driver cost per hour times the number of drivers open, en-route to pickup and on trip).

For any demand rate d, there is a (unique) minimal level of supply required to support this demand, defined by

$$\underline{n}(d) = \min_s \{s + d(t_1(s) + t_2)\}$$

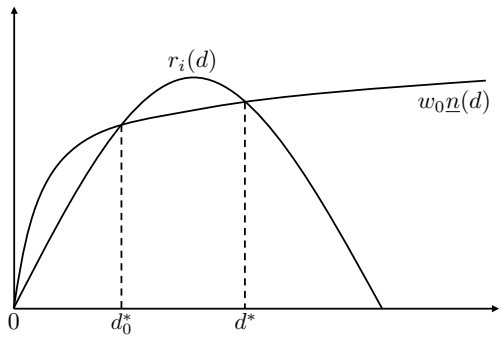


Figure 3.1: pure-HV equilibrium

This minimal level of supply must be the competitive equilibrium, otherwise total supply costs would be strictly higher which (since demand rate d is fixed) would lead to strictly higher rider prices. This in turn would trigger a competing platform to enter and serve the same demand d at lower prices.

For d to be an equilibrium demand, we must have that $r_i(d) = w_0\underline{n}(d)$. That is, market revenues must just be able to cover the minimum cost of supply necessary to support the demand rate d. This is illustrated in Fig. 3.1. If there is no positive d that satisfies the condition, then the only equilibrium is $d = 0$, meaning that the market cannot form. This could occur, for example, if the mass of riders M is not sufficient to support the cost of serving them at any positive level of demand and price. (The "thin market" case.) There can also be multiple equilibria with postive values of d; for example, in Fig. 3.1, both $d = d_0^*$ and $d = d^*$ satisfy the equilibrium condition. However, there is a key difference: the equilibrium with $d = d^*$ is a *stable* equilibrium, whereas the one with $d = d_0^*$ is not.

Formally, we define the stable equilibrium as follows:

Definition 6 (Stable equilibrium) *An equilibrium with demand rate d^* is not stable*

if $r_i(d^* - \epsilon) < w_0\underline{n}(d^* - \epsilon)$ *and* $r_i(d^* + \epsilon) > w_0\underline{n}(d^* + \epsilon)$ *for a small disturbance* $\epsilon > 0$*. If not, then the equilibrium is stable.*

In other words, for d^* to be a stable equilibrium demand rate, the revenue curve $r_i(d)$ must cross the cost curve $w_0\underline{n}(d)$ from above (like in the case of d^*). If not, as in the case of d_0^* in Fig. 3.1, with slightly more (fewer) drivers participating, the market revenue would go above (below) the cost, inducing even more (fewer) drivers to participate, which shifts the market *away* from d_0^*.

Throughout the paper, we restrict our interests in the stable equilibrium. Moreover, when there are multiple stable equilibria, we focus on the one with the highest demand rate, since it leads to the highest driver utilization and level of employment. We call such an equilibrium the *maximal stable equilibrium*.[11]

Formally, we summarize the pure-HV equilibrium in the following proposition:

Proposition 14 *In a pure-HV market, for each demand scenario* i*, there always exists a maximal stable equilibrium* $(d_i^*, p_i^*, n_i^*, u_i^*)$*. Moreover, in the equilibrium,*

1. *demand* $d_i^* = \max\{d : r_i(d) = w_0\underline{n}(d))\}$*, price* $p_i^* = p_i(d_i^*)$*, fleet size* $n_i^* = \underline{n}(d_i^*)$*, and utilization* $u_i^* = d_i^* t_2 / \underline{n}(d_i^*) = w_0 / p_i^*$*;*
2. d_i^**,* n_i^**, and* u_i^* *are increasing in* m_i*, and* p_i^* *is decreasing in* m_i*. Fixing a demand distribution,* d_i^**,* n_i^* *and* u_i^* *are increasing in* i*, and* p_i^* *is decreasing in* i*.*

Note the pure-HV equilibrium demand depends both on the potential demand m_i in scenario i and on the HV's reservation earnings w_0. To see how the

[11] As we show in Appendix B.1.3, there are at most two stable equilibria with one of them always being $d = 0$. Thus, the maximal stable equilibrium is either the non-zero stable equilibrium, or $d = 0$ which only happens when the market is not viable.

potential demand affects the equilibrium, consider a demand increase from low to high, e.g. from midnight to morning rush hours. The increase in demand will instantly raise the price $p_i(d)$, making $r_i(d) > w_0\underline{n}(d)$. This positive surplus for HVs will in turn induce more HVs to enter and prices to fall until the HV surplus goes down to zero again. Thus, both the equilibrium supply and demand is increasing in the potential demand.

The reservation earnings w_0 also affects the equilibrium demand in an intuitive way. Specifically, if labor becomes more expensive (w_0 increases), then under the previous supply level the revenue $r_i(d)$ will no longer cover the supply cost $w_0\underline{n}(d)$, leading to negative HV surplus. HVs will then start to exit the market and prices will rise until HV surplus increases back to zero. Hence, both the equilibrium supply and demand is decreasing in the reservation earnings for a fixed level of potential demand m_i.

In the rest of the paper, we only consider the maximal stable equilibrium. Thus, whenever we refer to an equilibrium, it is a maximal stable equilibrium.

Remark 2 *In our pure-HV market, the ride-hailing platform does not make profit and only mediates the transactions between supply and demand. This is to reflect the reality in ride-hailing, where platforms like Uber and Lyft engage in fierce competition on both the supply and demand side. When one platform lowers its pay (or increases its price), drivers (or riders) can always switch to the competing platform; the result is that all platforms have to give up their profit margins to remain in the market. Indeed, in practice, their difficulty in generating profit from the human driver market is one of the primary incentives for ride-hailing platforms to invest heavily in AV technologies (Teale 2021).*

3.4.1 Supply, demand and cost assumptions

We next make the following supply, demand and cost assumptions:

Assumption 5 (Loose market) *We assume a loose labor market for HVs, in which the revenue-maxizing price is above the human driver's equilibrium price, i.e.*

$$\arg\max_{p} r_i(d_i(p)) \geq p_i^*, \forall i$$

If a ride-hailing market doesn't satisfy Assumption 5, the HV's equilibrium price must be higher than the demand-side revenue-maximizing price. This would correspond to an extremely tight labor market in which revenue extraction from consumers is the limiting factor in expanding service – an unlikely case in most real-world ride-hailing markets as shown in Asadpour, Lobel, and Ryzin 2019. However, for completeness, in Section 3.7.4 we will briefly discuss what happens when Assumption 5 doesn't hold.

Assumption 6 (AVs have lower total cost) *Sum of the variable and fixed cost of AVs is still lower than HVs' reservation earnings, i.e.* $c_v + c_f < w_0$.

This assumption means AVs have strictly lower total cost than HVs when they utilized 100% of the time. Again, if this were violated, then there would not be a market for AVs.

3.5 Summary of main results and numerical example

Our analysis examines two dispatch platform structures (common platform vs. independent platform) under two levels of AV competition (monopoly AV vs.

competitive AV). By comparing these four cases with the pure-HV market, we explore how the introduction of AVs could affect market outcomes. We begin by summarizing our main results and then provide a numerical example to illustrate the results. The remainder of the paper then provides the detailed analysis behind each of the four cases.

Our first main finding is:

Proposition 15 *Even when AV technology has strictly lower total costs than HVs, HVs will still supply the market along with AVs if the demand distribution has a strict maximum-demand scenario (demand peak) with sufficiently low probability.*

The specific conditions can be found in Appendix B.4. Intuitively, this follows from the difference in cost structures for HVs and AVs combined with demand uncertainty. While HVs are more expensive per unit of working time, they are more flexible and do not require a large fixed cost commitment. An AV, in contrast, requires a significant capital investment which is only justified by the lower variable cost if the AV's utilization is sufficiently high. As a result, whether AVs are supplied by a monopoly supplier or through a competitive market, the equilibrium supply of AVs is typically less than what is required to serve demand in all scenarios. [12] The residual market (the demand in excess of the AV capacity) is then served by the HV market.

That a mixed market of AVs and HVs becomes the long-run state of ride-hailing is plausible. A useful analogy is electric power markets, which (like ride-hailing) are characterized by highly variable demand that is served with a

[12]There are extreme cases where an AV-only market emerges. For example, when there is no demand uncertainty, AV fixed costs are zero or the total cost of AVs are much lower than HVs.

Table 3.2: Market price comparisons under different dispatch structures and levels of competition

	Common Platform	**Independent Platform**
Monopoly	Identical prices in all scenarios	Higher prices in all scenarios
Competitive	Equal or lower prices in all scenarios	Higher prices in some scenarios

mix of production technologies for analogous cost-structure reasons.[13] "Base-load" technologies like hydroelectric and nuclear power require large capital investments but have low variable (and total) costs of generation. Gas turbines, in contrast, require a much smaller capital investment but have higher variable (and total) costs of generation. Yet despite the total cost advantage of base-load technologies, there is still a residual market for gas turbines to supply peak demand periods that cannot be supplied economically with base load technologies. In this sense, AVs in our analysis are akin to the base load technology of ride-hailing transportation while HVs play the role of peak-load suppliers.

Our next main result shows the importance of market structure for the price and consumer welfare effects of AVs:

Theorem 6 *With an AV technology that has lower total costs than HVs, the only market structure that leads to unambiguously lower (or non-increasing) prices in all demand scenarios is a perfectly competitive, common platform market. Specifically, the price outcomes in the four market cases are given by Table 3.2.*

In other words, whether the lower cost of AVs leads to lower prices (and thus

[13]For example, according to Electric Reliability Council of Texas, 99.5% of the energy in Texas came from four generation types: 43.7% natural gas, 28.8% coal, 15.1% wind, and 12% nuclear. (ERCOT 2016)

increased consumer welfare) depends critically on 1) having AVs integrated with HVs on a common dispatch platform and 2) ensuring AV operators compete. In the worst case market structure where AVs are supplied by a monopolist supplier and operate on an independent platform, market prices are in fact higher in all scenarios relative to a pure-HV market.

The intuition is that without a common dispatch platform, fragmenting the market into two independent services (AV and HV service) causes a loss of efficiency due to the spatial density (market thickness) effect. All else equal, this drives up the equilibrium price of both HV and AV service. This effect can further be exploited by a monopolist AV provider to drive the equilibrium price even higher than it would be in a pure-HV market. Competition in the AV market eliminates the monopoly effect, but cannot overcome the inefficiencies of spatial density lost by having two independent dispatch platforms.

A common dispatch platform preserves spatial density, but a monopoly AV supplier has no incentive to lower the price below the pure-HV market equilibrium price since it can price at (or epsilon below) that price and still serve the entire market. (This is true provided the HV equilibrium price is in the inelastic portion of the demand curve, which is the “loose" labor market regime of Assumption 5.) However, when there is AV’s competition in a common platform market, the equilibrium price will fall below the pure-HV equilibrium price in scenarios where the entire market can be served by AVs.

Hence, we need both the technical efficiency of a common platform market and competition among AV suppliers in low-demand scenarios to ensure that AV service preserves or lowers prices in all demand scenarios. We next provide a numerical example to illustrate our results.

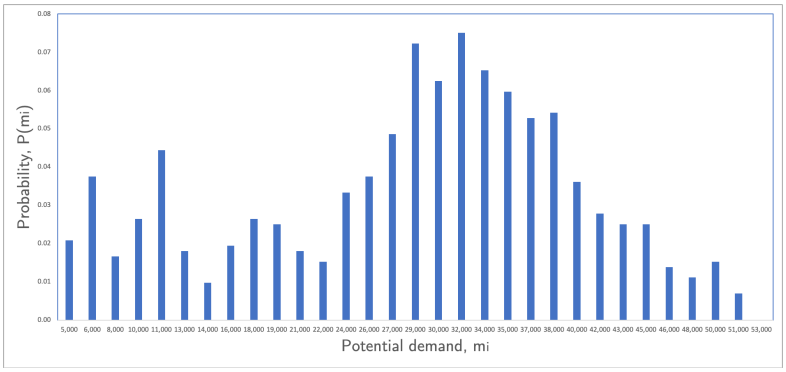


Figure 3.2: Probability density distribution of the hourly demand

3.5.1 Numerical examples

We applied our model to a numerical example of a ride-hailing market supplied by AVs and HVs. Table 3.3 shows the parameters used in the numerical example. HVs have reservation earnings of \$15/hour. AVs have an operating cost of \$3/hour, which is only incurred when AVs are utilized. AVs also have a fixed cost of \$6.85/hour, which is incurred regardless of the usage. The sum of the two types of AV cost is \$9.85/hour, which is lower than HVs' reservation earnings.

Table 3.3: Parameters used in numerical examples

	Meaning	Value
a	Mean pickup time for 1 open car	50 min
t_2	Mean trip time	25 min
r	ETA function parameter	0.4
w_0	The HV's reservation earnings	\$15/hour
c_v	The AV's operating cost	\$3/hour
c_f	The AV's fixed cost	\$6.85/hour
V	Price function parameter	\$200/ride-hour

The demand distribution is obtained from the FHV trip record by NYC Taxi & Limousine Commission (TLC) from June 1, 2019 to July 1, 2019, shown in Fig. 3.2. The number of trips is counted for each hour during this period and aggregated

into 30 demand scenarios, with the lowest scenario to be 4,720 requests/hour and the highest scenario to be 51,166 requests/hour.

We want emphasize that the purpose of the numerical example is to illustrate the high-level insights from the theoretical analysis, not to provide realistic empirical estimates. For example, the data used for generating the demand distribution is raw trip data, and thus may suffer from demand censoring. The choice of parameters, even though backed by news articles and industry reports, can benefit from a more rigorous empirical examination.[14] That said, the example provides some sense of the magnitude of the market structure effects.

We numerically solved the four cases: two dispatch platform structures (common platform vs. independent platform) combined with two levels of AV competition (monopoly AV vs. competitive AV). We then compared the market outcomes for these four cases.

3.5.2 Equilibrium prices

Fig. 3.3 is a numerical illustration of our main insights in Theorem 6. The equilibrium price is a function of the potential demand m_i and varies by market structure. Under an independent platform market, the equilibrium prices for a monopoly AV supplier are higher than a pure-HV market in all demand scenarios; when AVs are perfectly competitive, the prices are still higher in scenarios with high potential demand. Under a common platform market, the equilibrium prices for a monopoly AV supplier are identical to a pure-HV market in all demand scenarios. The only market structure that leads to unambiguously lower

[14] See Appendix B.6 for how we chose the parameters.

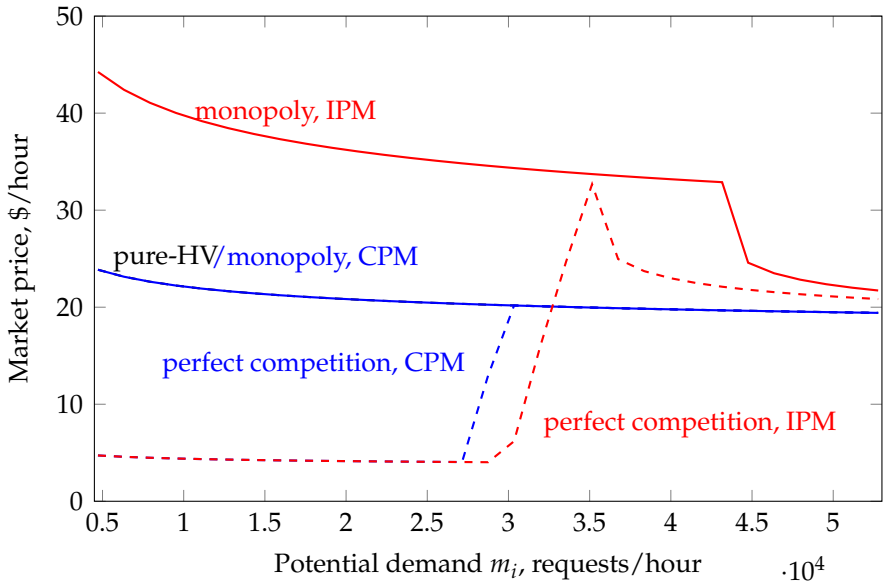


Figure 3.3: Price in each demand scenario at market equilibrium
Note. CPM: common platform market; IPM: independent platform market

prices is the perfectly competitive common platform market.

3.5.3 Revenues and social welfare

Table 3.4 shows a comparison of rider surplus, AV surplus and human driver employment among the four market structuers as well as a pure-HV market. AV surplus is the total profit generated by AVs in equilibrium. The calculation of rider surplus can be found in the appendix. The number of drivers is calculated as the average HV fleet size in equilibrium among all demand scenarios for the given market structure.

One can see that the competitive common platform market leads to the highest total surplus for riders and AVs, $\$603,310$, and a relatively high average number of HVs, $1,071$. In contrast, the monopoly independent platform market

Table 3.4: Surplus and HV employment impact

	Surplus			Driver
	Rider	AV	Total	Avg. #
Pure-HV	\$573,700	\$0	\$573,700	8,578
Monopoly, CPM	\$573,700	\$28,241	\$601,941	1,258
Competition, CPM	\$603,310	\$0	\$603,310	1,071
Monopoly, IPM	\$493,910	\$88,397	\$582,307	155
Competition, IPM	\$600,400	\$0	\$600,400	726

Note. CPM: common platform market; IPM: independent platform market

leads to the lowest total surplus, $582,307, but (not surprisingly) generates the highest AV profit of $88,397. Moreover, it leads to the lowest HV employment - with 155 HVs on average, much lower than the other structures.

Table 3.5 provides more details on the equilibrium market outcomes for each market structure. The average price for a monopoly independent platform market, $33.09, is significantly higher than the average price in all other structures. Moreover, in line with the AV surplus given in Table 3.4, the revenue under this market structure is also significantly higher than others. This illustrates how a monopoly AV supplier can take advantage of the independent platform market, setting high prices and making high profits at the expense of riders and HVs.

The numerical example again highlights the importance of market structure in unlocking the benefits of AV technology.

3.6 Analysis of main results

We next provide our detailed analysis of the common and independent platform markets.

Table 3.5: Equilibrium AV capacity, average demand fulfilled by AVs and HVs, and revenue for AVs

	AV capacity	Total demand	AV revenue	Average price
Monopoly, CPM	8,700	25,521/hr	$128,660/hr	$ 20.16/hr
Competition, CPM	9,050	26,155/hr	$101,130/hr	$ 15.47/hr
Monopoly, IPM	11,900	23,673/hr	$195,860/hr	$ 33.09/hr
Competition, IPM	9,835	26,072/hr	$102,800/hr	$ 15.77/hr

Note. "AV capacity" is the capacity determined by the monopoly supplier or the perfect competition equilibrium. "Total demand" is the combined number of requests supplied by AVs and HVs per hour on average. "AV revenue" is the total revenue generated by all AVs per hour. "Average price" is the average equilibrium price among all demand scenarios.

3.6.1 Common platform market

We begin by looking at a dispatch structure where AVs and HVs operate on the same platform, as described in Section 3.3.2. In this case, there is symmetry in dispatch and the total number of open cars and total demand served is proportional to the fleet sizes.

As before, we analyze the equilibrium in a single hour in a given demand scenario i. For scenario i, the AV supplier(s) chooses the AV fleet size n_a to deploy to maximize their variable profit, subject to capacity constraints. To see how the variable profit would change at different fleet size, we begin by analyzing the market with the AV fleet size n_a as exogenously given. After that, we analyze the AV fleet size and capacity choice for the monopoly and competitive AV supply.

Exogenous AV fleet size

Given that the AV fleet size n_a is fixed, the market equilibrium is determined by HVs' participation decisions. If HVs expect to make more than their reserva-

tion earnings w_0, then they will join and provide service. As the HV fleet size increases, more demand gets satisfied; the market price will decrease, lowering HVs' earnings. In equilibrium, HVs will make exactly w_0 per hour just like in the pure-HV market. Such an equilibrium can be reached when the AV fleet size n_a is not too high, so that AVs are only able to serve a small fraction of the potential demand and the market price is high enough to induce the participation of HVs.

On the flip side, it is also possible that HVs cannot make w_0 at any HV fleet size. This happens when the AV fleet size n_a is sufficiently high, such that a large fraction of the potential demand is satisfied by AVs alone. HVs thus cannot make enough to cover its reservation earnings w_0.

Formally, the results are characterized in Proposition 16:

Proposition 16 *Consider scenario i in a common platform market. Given an AV fleet size n_a, in equilibrium, there are only two possibilities:*

1. *AVs and HVs serve the market together. Moreover, the variable profit per AV is $(w_0 - c_v)$ per hour; the equilibrium total fleet size, price, demand rate, and utilization are identical to those in a pure-HV market. This case happens when $n_a < n_i^*$.*
2. *AVs can serve the entire market alone. Moreover, the variable profit per AV is no greater than $(w_0 - c_v)$ per calendar hour; the equilibrium price, demand rate and utilization solely depend on n_a. This case happens when $n_a \geq n_i^*$.*

Proposition 16 has important implications. It shows that, for the AV supplier(s), AVs' profitability hinges on HVs' reservation earnings. In any scenario, the hourly earnings per AV is guaranteed to be exactly w_0 when HVs participate,

and is bounded by w_0 when AVs serve the market alone. In other words, AVs are more profitable as HVs' opportunity costs increase. Indeed, high labor costs is one of the main incentives for ride-sharing companies like Uber and Lyft to invest heavily in the AV technology. From this perspective, the gap between HVs' and AVs' variable costs, $(w_0 - c_v)$, represents the maximum unit profit increase achievable by AV technology.

Nevertheless, the implications of AVs for consumers is more ambiguous. When AVs and HVs coexist, market-level characteristics such as the equilibrium price and the total supply and demand are the same as the pure-HV market. In other words, for consumers, it does not make any difference economically when AVs and HVs serve the market together.[15] When AVs serve the market alone, the market characteristics can be quite different from the pure-HV market depending on n_a; however, this is not necessarily good for the consumers either, because it is up to the AV supplier(s) to decide the level of service and market price that they find most profitable. As we will show in Section 3.6.1, when AVs are supplied by a monopoly, the supplier may not expand service or offer lower prices for consumers even if it has the ability to do so.

Appendix B.2 contains additional technical details about the common platform market, including the mathematical expression for the variable profit function π_i^C. Next, we proceed to analyze how the AV fleet sizes and capacity are determined under a monopoly AV supplier and perfectly competitive AVs.

[15] This is of course under the assumption that AVs and HVs provide identical service. If the service provided by AVs is inferior (or superior) to that by HVs, then the composition of supply will also make a difference to the consumers.

Monopoly

In this section, we analyze a common platform market for AVs and HVs, with AVs supplied by a monopoly. This is motivated by the real-world setting when a monopoly ride-hailing platform has its own AV fleet, or collaborates with a monopoly AV supplier (e.g., the collaboration between Lyft and Aptiv).

There are two stages of decisions for the monopoly AV supplier: first, the choice of the AV capacity N, and second the AV fleet size deployed in each scenario. In the first stage capacity choice, the AV supplier faces a trade-off between high fixed costs and lost sales. If the AV capacity is too low, then in high demand scenarios, the supplier's profit is constrained by the low capacity, missing the potential opportunity of higher revenue; if the AV capacity is too high, then in low demand scenarios, a large number of AVs may sit idle, wasting the cost for acquiring the capacity. Indeed, as we will soon show, the monopoly supplier's choice of AV capacity has the structure of a newsvendor model.

In the second stage, for the AV fleet size decision, the monopoly supplier also faces a trade-off: if too many AVs are deployed, then the market price will drop; if too few AVs are deployed, then the number of trips completed by AVs (quantity served) will be low. Moreover, the decision is constrained by the capacity chosen in the first stage.

Thus, we solve the problem backwards by first solving the optimal AV fleet size in each scenario, taking the capacity N as given; then we solve for the optimal capacity. Proposition 17 presents results for the AV fleet size:

Proposition 17 *Consider a monopoly AV supplier and a common platform market.*

Suppose the market is loose. Then in scenario i, given an AV capacity N, the monopoly supplier's decisions are the following:

1. *If $N < n_i^*$, then the optimal AV fleet size is its capacity N; Moreover, $(n_i^* - N)$ HVs join the market to supply the residual demand.*
2. *If $N \geq n_i^*$, then the optimal AV fleet size is n_i^*, which is the supply level that exactly keeps HVs out of the market.*

In both cases, the total fleet size and equilibrium market price are identical to the pure-HV market.

When N is low (or demand is high), AVs share the market with HVs; when N is high (or demand is low), AVs serve the market alone by just supplying enough units to keep HVs out. Such a decision reflects the monopoly power from the AV supplier. In a loose market (Assumption 5), it is not optimal for the AV supplier to expand the supply level beyond the pure-HV equilibrium level, because the marginal benefit from serving one more unit of demand does not make up for the reduction in price from the extended supply. The supplier's profit starts to decrease when the AV fleet size is above n_i^*. Thus, to maximize its profit, the supplier chooses to keep the fleet size at the same level as the pure-HV market.

An important takeaway from Proposition 17 is that, despite having lower costs, AVs lead to identical prices as the pure-HV market. Moreover, an AV's hourly earnings are constant and always the same as the HVs reservation earnings w_0 in any scenario. Hence, the AV supplier's net aggregated profit can be simplified as:

$$\Pi(N) = (w_0 - c_v) \sum_{1}^{I} P(m_i) \min\{N, n_i^*\} - c_f N \tag{3.4}$$

This simplified objective function has the structure of a newsvendor problem, with an underage cost being the variable profit per AV $(w_0 - c_v)$ and the overage cost being the fixed cost c_f. Hence, we can fully characterize the optimal capacity choice of the monopoly AV supplier:

Proposition 18 *Suppose the potential demand m_i is increasing in i, i.e. $m_1 < m_2 < ... < m_I$. In a loose market, the optimal AV capacity for a monopoly AV supplier in a common platform market is given by*

$$N^* = n_K^*$$

where K is the first scenario such that

$$\sum_{K+1}^{I} P(m_i) < \frac{c_f}{w_0 - c_v}$$

If such a K doesn't exist ($P(m_I) \geq \frac{c_f}{w_0 - c_v}$), then the profit-maximizing capacity is $N^* = n_I^*$.

The optimal choice of the AV capacity is determined by its fixed cost, the variable profit, and the demand distribution. In a world where AVs' fixed cost is low, the critical fractile $c_f/(w_0 - c_v)$ is low, and thus the optimal AV capacity is high. In an extreme case where the fixed cost c_f is close to 0, it is optimal to choose an AV capacity that can fulfill the demand in all scenarios ($N^* = n_I^*$). In this case, in any scenario $i < I$, there are a number of $(n_I^* - n_i^*)$ AVs being idle. HVs are then driven out of the market completely.

On the flip side, when AVs' fixed cost is high, the optimal capacity is low. In an extreme case where c_f is close to the variable profit $(w_0 - c_v)$, it is optimal to choose a capacity such that all AVs are operating 100% of the time ($N^* = n_1^*$). In

other words, if AVs are an expensive investment, the supplier cannot afford to let AVs be idle, and AVs and HVs will coexist and operate together for all scenarios except for the lowest one.

Another way to interpret the results in Proposition 18 is by looking at the demand distribution. Consider a simple distribution of only two potential scenarios: a high demand or a low demand scenario, with the probability of the high demand scenario being less than the critical fractile, i.e. m_1 and m_2, where $m_2 > m_1$ and $P(m_2) < \frac{c_f}{w_0 - c_v}$. Then by Proposition 18, the optimal AV capacity is just n_1^*. Then in the low demand scenario, AVs serve the market alone. This can be thought as non-peak hours when the demand is relatively low. The monopoly AV supplier can leverage its low cost to keep all HVs out of the market. On the other hand, the high demand scenario can be thought as peak hours that have high demand but occur with low probability. When peaks occur, HVs join the market to serve the residual demand that cannot be fulfilled by AVs.

Perfect competition

We consider a setup where multiple suppliers have access to the AV technology, and AVs share the dispatch platform with HVs. For example, Lyft is building a commercial self-driving network, in which AVs from multiple companies like Waymo and Aptiv can complete commercial rides on the same dispatch network with Lyft drivers. For tractability, in this section we analyze an extreme case where there is perfect competition among the AV suppliers. In the extension (Section 3.7.3), we discuss the case of oligopoly.

The key difference between competitive and monopoly AVs is that no supplier

has market power over the price in the market; instead, the equilibrium AV fleet size is an outcome of the joint decisions of all AV suppliers. Therefore, unlike the monopoly AV case, here the supply of AV is cannot be limited to maximize profit, and in equilibrium each AV supplier makes zero net profit. Therefore, the prices are expected to be lower than the monopoly case. This is indeed the case:

Proposition 19 *For a perfectly competitive common platform market, the equilibrium price is no higher than that in a pure-HV market in each demand scenario.*

Note this is the only market structure that leads to unambiguously lower prices in all scenarios. Similar to the monopoly case, when the AV capacity is below the pure-HV equilibrium fleet size, AVs and HVs serve the market together in all scenarios and the market price will be the same as in the pure-HV case. Each AV then earns a positive profit $(w_0 - c_v)$, and so all AVs will participate. But unlike the monopoly case, AVs will continue to participate at full capacity until the variable profit is zero. Hence, in these scenarios the AV supply is higher and the equilibrium price is lower than in the pure-AV regime.

3.6.2 Independent platform market

We next look at a market structure where AVs and HVs operate on two separate dispatch platforms. Since AVs do not share the dispatch network with HVs, they no longer share the same utilization and the scale economy with HVs. As before, we analyze the equilibrium in a given demand scenario i and start by analyzing the equilibrium with an exogenously given AV fleet size. We leave the detailed analysis in Appendix B.3 and only highlight results that are different from the

common platform market. We start by considering the fleet size and capacity choice for the two cases of monopoly and competitive AV suppliers.

Monopoly

Consider a monopoly AV supplier that makes an irreversible decision on the AV capacity N, the cost of which is $c_f N$ per hour. In each demand scenario m_i, the supplier maximizes its variable profit by choosing the optimal fleet size, subject to the capacity constraint N. Similar to the common platform market, there is a trade-off in the capacity selection; if the capacity is set too high, the supplier will incur high fixed costs in low demand scenarios in which AVs will be idle ("overage cost"), and if the capacity is set too low, there will be high opportunity costs in the high demand scenarios due to lost demand ("underage cost"). However, the difference is that, under an independent platform market, the trade-off may be more extreme. When the AV capacity is too low, not only does the supplier lose the opportunity to serve more demand, the AV fleet also becomes less efficient, making AVs less competitive than HVs. In certain demand scenarios where an HV can still make hourly earnings of w_0, an AV may not even be able to make c_v if its capacity is too low to achieve a break-even utilization. (See more on this in Appendix B.3.) With some demand distributions, this nonviable regime leads to surprising results, as shown in Fig. 3.4 where there exist cases where any AV capacity choice produces negative profit. In effect, the market may be "too thin" to support AV technology.

We summarize this observation in the following proposition:

Proposition 20 *In an independent platform market, even though AVs have lower total*

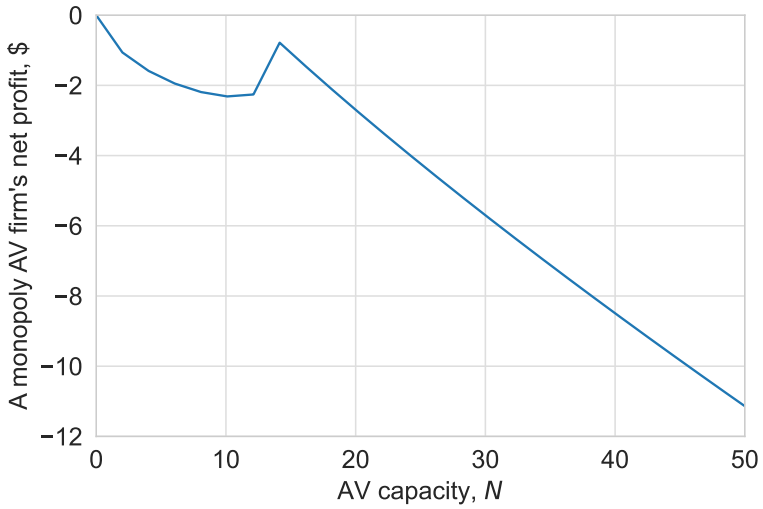


Figure 3.4: An example of an independent platform market in which a monopoly supplier only generates negative profits
Note. Parameters: Demand distribution: $M = \{m_1 = 15 \text{ requests/hour}, m_2 = 1,000,000 \text{ requests/hour}\}$, $P(m_1) = 0.2, P(m_2) = 0.8$; cost and price parameters: $w_0 = 5$, $c_v = 3.85$, $c_f = 1$, $V = 11$; trip parameters: $a = 0.1$, $t_2 = 2$, $r = 0.4$

cost, there exist demand distributions under which AVs cannot make positive aggregate profits even when they are supplied by a monopolist.

This is particularly surprising given that AVs have strictly lower costs than HVs and are managed by a profit-maximizing monopoly. It means that in an independent platform market, even a monopoly AV supplier with lower cost may not be able to break-even. The driving force behind this result is again the density effect. Separating the dispatch network and the rider pool for two platforms means AVs and HVs need to "compete" for density and scale. As AVs require pre-committed capacity investment but HVs do not, AVs can be disadvantaged in this competition.

However, when the demand distribution and cost parameters allow AVs to profit, the monopoly supplier can exploit the density effect. If AVs are profitable

at a reasonably large capacity, this leaves little residual demand for HVs, which lowers the HV utilization and drives up the equilibrium HV price. A monopoly AV supplier recognizes this price effect of crowding out HVs and chooses a large capacity in order to drive up prices. From Proposition 32 in Appendix B.3.1, we know that when AVs and HVs coexist in the market, the price is strictly higher than in a pure-HV market. This combined with the AV monopolist's incentives gives us the following proposition:

Proposition 21 *Conditional on AVs being viable, for a monopoly independent platform market, the equilibrium price in each scenario is at least as high as that in a pure-HV market.*

The rationale is that in scenarios in which AVs and HVs coexist, the equilibrium price is strictly higher than the pure-HV price due to the efficiency loss for HVs. In scenarios where AVs serve the market alone, a monopoly AV supplier does not benefit from extending supply above the HV equilibrium supply level; hence, prices are at least as high as in the pure-HV case.

To conclude, a monopoly supplier faces extreme situations in an independent platform market: it's either nonviable or is able to extract higher prices than a pure HV market. In either case, consumers do not benefit.

Perfect competition

When AVs are supplied by a large number of small competitive suppliers, in equilibrium the expected variable profit equals the fixed cost c_f. If this were not true, then there would be an incentive for additional suppliers to invest

in AVs. The equilibrium capacity under perfect competition can therefore be obtained by solving Eq. (3.2) using the profit function $\pi_i^I(n)$ given by Eq. (B.12) in Appendix B.3.1.

Since AV suppliers do not jointly optimize their profits under perfect competition, the result that AVs can be nonviable is directly implied by Proposition 20. That is, if a monopolist AV supplier cannot make a positive profit, then neither can competing AV suppliers:

Corollary 3 *In a perfectly competitive independent platform market, even though AVs have lower total cost, there exist demand distributions where AVs cannot make positive aggregate profits.*

Furthermore, the results in Proposition 32 from Appendix B.3.1 continue to apply here:

Proposition 22 *For a perfectly competitive independent platform market, there exists demand scenarios in which the equilibrium price is higher than that in a pure-HV market.*

In the regime of coexisting AVs and HVs, the price is higher than that in the pure-HV market, regardless of the level of competition. The rationale is that, the higher prices in the coexisting regime are driven by the efficiency loss from separating the two types of supply, not from the monopoly supplier's profit maximization; thus, it holds under all levels of competition. Nonetheless, when there is no HV participation and AVs operate alone (see Case 2 of Proposition 32, Appendix B.3), the level of AV competition would play a role; as a result of competition, the price can be lower than that in the pure-HV equilibrium.

To summarize, in a perfectly competitive, independent platform market, the equilibrium prices are still not uniformly lower than the pure-HV market.

3.7 Extensions

Next, we analyze some extensions to the main analysis.

3.7.1 HV supply elasticity

In this section, we examine the implications of HVs having finite elasticity. To avoid repetition, we only present the analysis for the common platform market; the analysis for the independent platform market can be found in Appendix B.3.2. Both analyses yield similar insights. The conclusion is that with finite supply elasticity, there will be additional welfare gains for consumers from AVs because the reduced demand for human labor will lower equilibrium prices for HV service.

All else being equal, suppose there is a supply curve for HVs, $w(n_0)$. That is, to get a supply of n_0 HVs, the hourly earnings per HV must be at least $w(n_0)$. Moreover, $w(n_0)$ is an increasing function of n_0. In other words, it is increasingly more expensive to acquire one more HV.

Fig. 3.5 illustrates how the equilibrium point shifts after introducing AVs in a common platform market, assuming HVs have a supply curve $w(n_0)$. The revenue curve, $r_i(d)$, has the same definition as in Section 3.4. The cost curve, "Total costs (pure-HV)", represents the total hourly earnings that HVs expect to

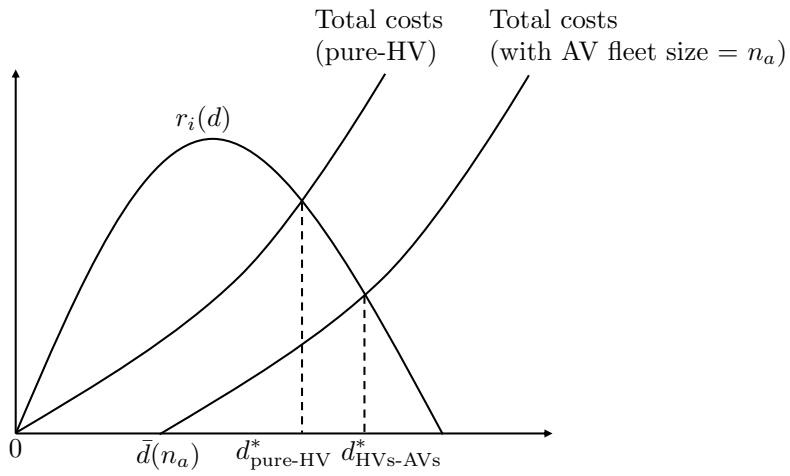


Figure 3.5: Equilibrium with finitely elastic HVs, common platform market
Note. Total cost (pure-HV) $= w(n_0)\underline{n}(d)$, with $n_0 = \underline{n}(d)$ which is the minimum fleet size to supply demand rate d; total costs (with AV fleet size = n_a) $= w(n_0)\underline{n}(d)$, with $n_0 = (\underline{n}(d) - n_a)^+$. (Without loss of generality, we let $w(0) = 0$ in this figure.) In both cases, the cost per vehicle is determined by HVs' reservation earnings $w(n_0)$; this is because under a common platform market, HVs and AVs always face the same price and utilization, and thus have the same hourly earnings.

receive when the market demand is at d in a pure-HV market. "Total costs (with AV fleet size $= n_a$)" represents the total hourly earnings that both AVs and HVs expect to receive when the market demand is at d and there are a fleet of AVs with size n_a. The market clears when the cost curve intersects with the revenue curve, same as that in Fig. 3.1.

When there are AVs with a fleet size of n_a, the cost curve shifts to the right. That is, to serve the same level of demand, the total costs (including for both HVs and AVs) strictly decrease after introducing AVs. The rationale is that, in the presence of the AV fleet, there is less demand for HVs. Because the supply curve is upward-sloping, HVs no longer require the same compensation as the pure-HV market. Thus, the market equilibrium is reached at a higher demand rate $d^*_{\text{HVs-AVs}}$ and a lower price, which increases consumer surplus. Such a shift

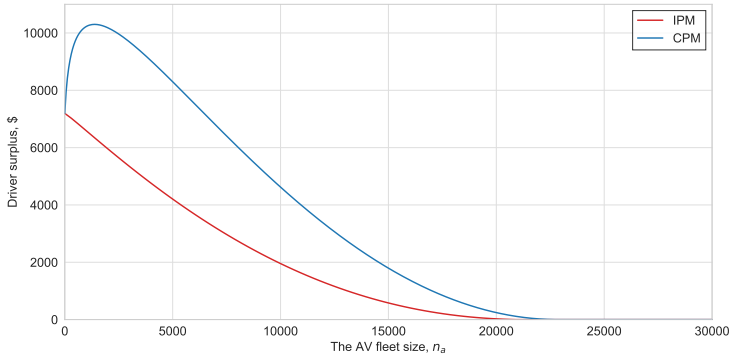


Figure 3.6: Driver surplus in the change of the AV fleet size in a single demand scenario
Note. In both CPM and IPM, the potential demand is $m_i = 50,000$ requests/hour.

of the cost curve will not happen when the HV supply is perfectly elastic – in that case, the reservation earnings remain constant and do not decrease even though the demand for HVs decreases after introducing AVs; the total costs curve will be identical to the pure-HV curve. Therefore, a finite labor supply elasticity does lead to increased consumer welfare from the introduction of AVs.

3.7.2 Driver welfare

In this section, we compare the driver welfare under various market structures with the pure-HV market. To do so, we continue to relax the perfect elasticity assumption for HVs (otherwise, HVs always work at their reservation earnings and have zero surplus) and consider finite elasticity. Our findings show that, in an independent platform market, introducing AVs always reduces drivers' welfare; in a common platform market, AVs can improve drivers' welfare, providing the HV supply is sufficiently inelastic.

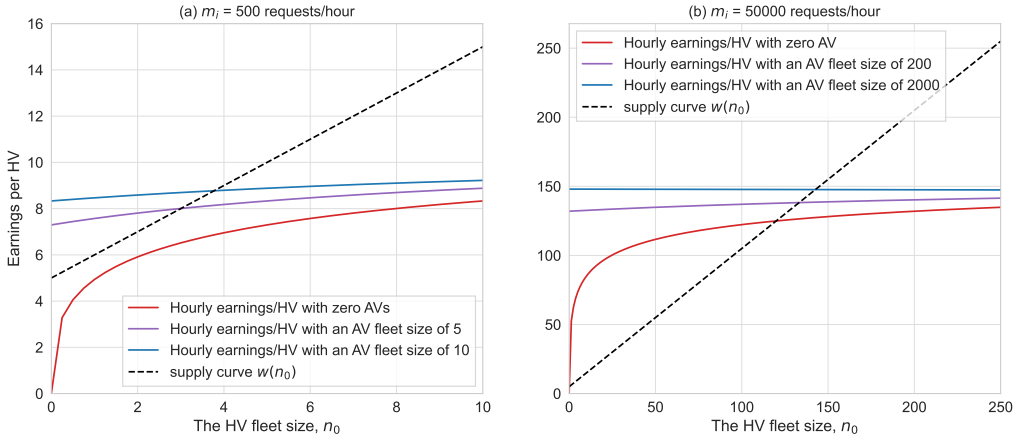


Figure 3.7: Hourly earnings per HV at different AV fleet size in a single demand scenario, CPM

Fig. 3.6 provides a numerical example of how driver surplus changes as a function of the AV fleet size in a particular demand scenario. As is illustrated by the blue curve in Fig. 3.6, in a common platform market the driver surplus is increasing in the AV fleet size when the AV fleet size is small. The intuition is that, because AVs and HVs are sharing the same platform, a larger AV fleet increases the utilization of HVs, allowing more HVs to participate, which increases their surplus. The formal definition of driver surplus can be found in Appendix B.5.

Fig. 3.7 takes a closer look at AVs impact on per HV hourly earnings and the equilibrium HV fleet size. The two sub-figures are examples of how AVs improve the earnings of HVs, with a relatively inelastic HV supply curve. Fig. 3.7(a) shows a situation where HVs cannot break-even in a market without AVs due to the low demand mass (the black dash line does not intersect with the red line). After introducing AVs, the earnings curve for HVs moves upward as a result of the increase in utilization by sharing the same platform with AVs. Fig. 3.7(b) is an example where the demand mass is sufficient to support HVs when they are

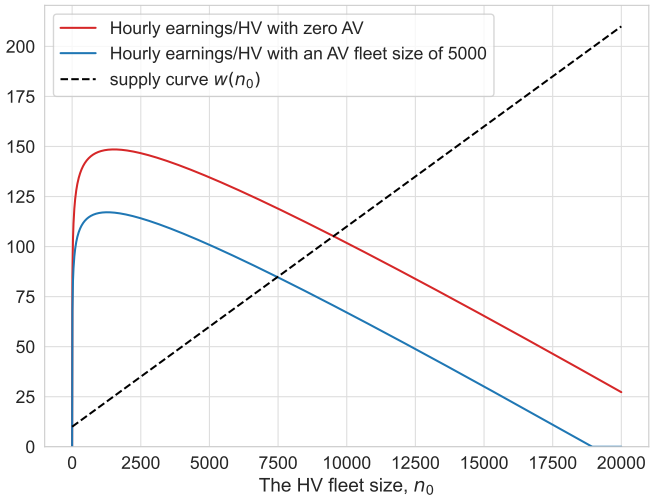


Figure 3.8: Earnings per HV at different AV fleet sizes in a single demand scenario, IPM

operating alone (the black dash line intersects with the redline). Still, HVs are better off because of the existence of AVs. Therefore, AVs may increase drivers' welfare both when HVs can and cannot sustain the market alone.

In an independent platform market, however, AVs always reduce the level of HV employment (i.e. the HV fleet size) and HVs' surplus in any demand scenario. Formally, we have

Proposition 23 (HV surplus) *In an independent platform market, HVs are always worse off because of the introduction of AVs.*

The formal proof of Proposition 23 is given in Appendix B.5. The reason is that, the HV fleet size depends on two main factors: the market price, and the HV utilization. With more AVs in the market, the supply is more abundant and the price is lower; in the meantime, HVs no longer benefit from AVs' scale and do

not have a higher utilization, because of the separate platforms. Fig. 3.8 shows the change in the hourly HV earnings in a demand scenario with $m_i = 50,000$ requests/hour. All else being equal, when there are $5,000$ AVs in the market, the HV earning curve moves downwards, and the equilibrium HV size decreases. Therefore, when AVs are deployed, they merely drive HVs out of the market and always reduce drivers' welfare.

3.7.3 Oligopoly AV suppliers

In this section, we discuss the setting of a finite number of AV suppliers. Because we consider a quantity competition among AV suppliers that provide identical service and have the same cost structure, the model follows the classical results of Cournot competition; that is, the equilibrium price should be inversely related to the number of suppliers in the market. (Shapiro 1989) Thus, we argue that the price outcomes for an oligopoly are bounded by the case of monopoly and perfect competition. Thus, in a common platform market, oligopoly AV suppliers will lead to prices lower than in the monopoly setting, which means that the prices can be lower than the pure-HV market. In an independent platform market, oligopoly AV suppliers will lead to prices higher than in the perfect competition setting, which means the prices can be higher than that in a pure-HV market. Therefore, our insights that prices are not unambiguously lower after introducing AVs still hold.

3.7.4 Tight labor market

In this section, we discuss the implications when the labor market is tight. That is, HVs' equilibrium price lies on the elastic portion of the demand curve, and a monopoly transportation service provider can improve its revenue by increasing the supply beyond the HV equilibrium.

We find that the insights from Theorem 6 largely hold when the labor supply is tight: AVs do not necessarily lead to lower prices, and whether they do depends on the structure of the marketplace. The main difference is that a common platform market will lead to unambiguously lower prices even with a monopoly AV supplier, because extending the supply above the HV equilibrium point improves the monopoly supplier's profit. Nonetheless, for an independent platform market, prices are still not uniformly lower than in a pure-HV market. Based our analysis in Appendix B.3, an independent platform market may be in one of the two regimes: coexisting AVs and HVs, or pure AV. The loose market condition will affect the decision of the monopoly AV supplier, providing an incentive to increase supply and lower prices. However, it does not change the way prices are determined in the coexisting regime, nor the existence of this regime. As a result, AVs still can lead to higher prices in an independent platform market.

3.7.5 Demand variability

As we have discussed in the main analysis, there is a trade-off between AVs and HVs, arising from their distinct cost structures. AVs are expense assets that require upfront investment, while HVs are flexible and incur costs only when

being utilized. In this section, we discuss how demand variability impacts the equilibrium outcome. Our conclusion is that, without demand variability, AVs serve the market alone in all scenarios and configurations; the HV participation is most significant when there is some variation in demand and AVs are relatively expensive.

We start by considering a market with a common platform shared by AVs and HVs and a monopoly AV supplier. For simplicity, consider a two-point distribution with two demand scenarios, high (H) and low (L), given by

$$M = \begin{cases} m_L, & \text{with probability } P_L \\ m_H, & \text{with probability } P_H = 1 - P_L \end{cases}$$

From the newsvendor structure in this configuration (see Proposition 18), it is easy to characterize the optimal AV capacity and the fleet sizes. Fig. 3.9 shows the average equilibrium AV and HV fleet sizes across the two scenarios, as a function of the probability mass of the high demand scenario P_H. The optimal AV capacity is just a piece-wise function, with $N^* = n_L^*$ when P_H is below the critical fractile and $N^* = n_H^*$ when P_H is at or above the critical fractile. The average fleet size including both AVs and HVs is just a weighted average of n_L^* and n_H^*, with the weight being P_L and P_H, respectively. When P_H is below the critical fractile, AVs serve the market alone in the low scenario and coexist with HVs in the high scenario (except when $P_H = 0$); furthermore, the difference between the average fleet size and the AV capacity equals to the average HV fleet size (the light gray area in Fig. 3.9), which we use as a measure for the HV participation. When P_H is at or above the critical fractile, AVs serve the market alone in all scenarios, and there is no HV participation; the difference between the AV capacity and the average fleet size equals to the average number of idle

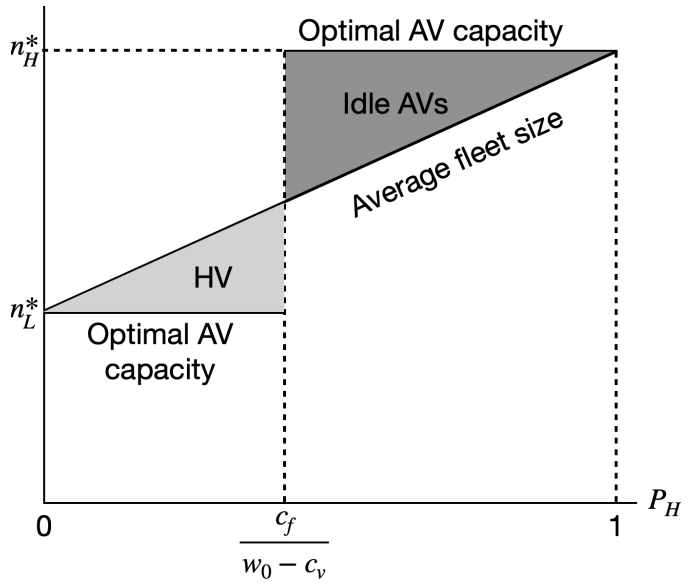


Figure 3.9: The optimal AV capacity and average fleet sizes, in the change of the potential demand distribution
Note. When $P_H < c_f/(w_0 - c_v)$, the optimal AV capacity $N^* = n_L^*$, which is the equilibrium HV fleet size for the low demand scenario; AVs are always 100% utilized, and HVs participate in the high demand scenario at a fleet size of $(n_H^* - n_L^*)$. When $P_H \geq c_f/(w_0 - c_v)$, $N^* = n_H^*$, AVs serve the market alone; in the low demand scenario, there is a number of $(n_H^* - n_L^*)$ idle AVs.

AVs.

One can observe that when there is no demand variability ($P_H = 0$ or $P_H = 1$), HVs do not participate; however, this does not mean that when there is increasing variability, the HV participation will always increase. Using the variance of M as a metric, the variability is maximized when $P_H = 0.5$. Consider the case of increasing variability as P_H goes from 0 to 0.5. Then when AVs are relatively more expensive ($c_f > 0.5(w_0 - c_v)$), the average HV fleet size is increasing in variability. The intuition is that, when the cost for AVs are relatively high, the monopoly supplier would rather lose the opportunity to make more profit in the high demand scenario, rather than letting AVs sit idle in the low demand scenario. Thus, as there is more variability, HVs serve an increasingly important

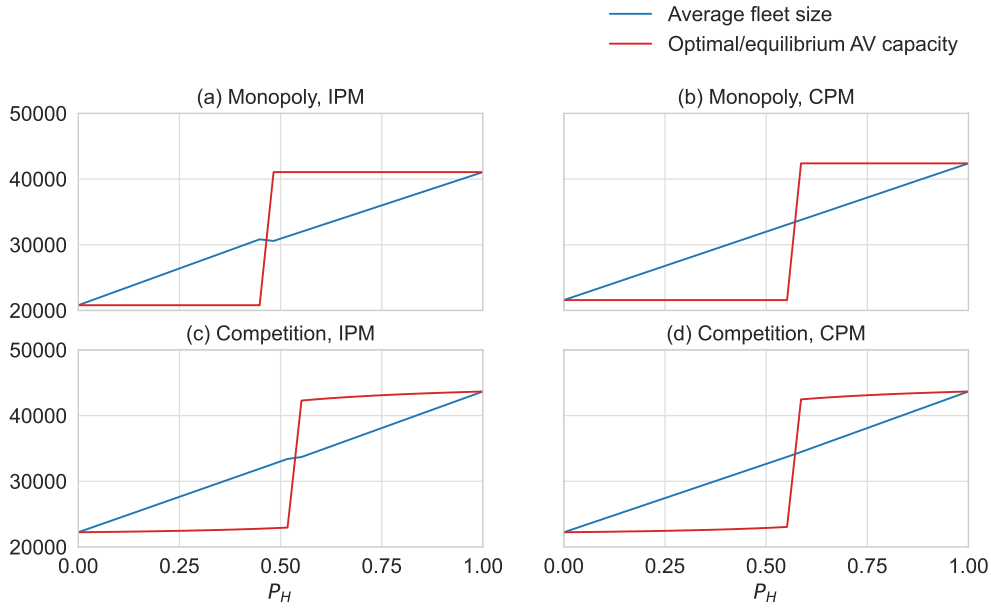


Figure 3.10: Numerical examples of the supply composition and optimal AV capacity, in the change of the potential demand distribution
Note. "Average fleet size": average number of vehicles (including both AVs and HVs) operating in the two scenarios; "Optimal/equilibrium AV capacity": the optimal AV capacity for a monopoly supplier, or the equilibrium AV capacity with perfectly competitive AVs. Parameters: In all four examples, $c_f/(w_0 - c_v) = 0.57$.

role of fulfilling the residual demand in the high demand scenario.

In contrast, when AVs are relatively less expensive ($c_f \leq 0.5(w_0 - c_v)$), the average HV fleet size can actually decrease in variability (for example, the average HV fleet size jumps down to zero as P_H exceeds the critical fractile). This is because, given the low AV costs, the supplier cares more about losing demand than having AVs idle. Thus, they can afford to hedge demand variability with a high AV capacity, which eliminates the HV participation.

Fig. 3.10 shows the same analysis with a numerical example for all four configurations, which exhibits a similar pattern. In all configurations, there is little HV participation when P_H is close to 0 and 1. As P_H goes from 0 to 0.5 and

demand variability increases, the HV participation can either increase (case b, c, d) or decrease (case a).

3.8 Conclusions

Our model provides a theory to understand the potential economic impact of AVs. By examining the market outcomes with AVs and HVs under various market settings, we find the lower cost of AVs do not necessarily lead to the expected outcomes of lower prices and expanded service. Rather, the impacts critically depend on the dispatch platform structure and level of AV competition.

Our findings also show that even when AV technology has strictly lower costs, there is still a market for HVs in cases where demand are sufficient peak scenarios. AVs are expense assets whose capacity needs to be pre-committed to the market, while HVs are provided by independent contractors and have more flexibility. With high demand variability, a mixed market of AVs serving the base demand and HVs serving the peak demand becomes the most cost-effective market state.

More surprisingly, we show that even with lower costs, AVs do not necessarily lead to lower prices. Whether AVs lead to lower prices and expanded services depend on the dispatch platform structure and the level of AV competition. Among all four market structures examined, the only structure that leads to unambiguously lower prices in all demand scenarios is the perfectly competitive common platform market. In contrast, an independent platform market with monopoly AVs indeed leads to higher prices in all demand scenarios. These results illustrate the critical role market structure plays in realizing the benefits

of AV technology.

There are many interesting extensions for this model. For example, we have assumed fully functional AVs that can serve the same set of trips as HVs. Although our results can carry through by only considering those areas that AVs are able to serve, it is worthwhile to consider how the cost structure of AVs will impact what segments of trips AVs will ultimately serve. An equilibrium that looks not only at total AV capacity but also the mix of trips that AVs serve can provide further insights into how AVs might affect market outcomes. Such an analysis is the subject of ongoing work by the authors.

CHAPTER 4

LABOR COST FREE-RIDING IN THE GIG ECONOMY

4.1 Introduction

The gig economy has grown rapidly over the past decade. Between 2010 and 2020, the share of gig workers rose to 15% of the total U.S. labor force (Iacurci 2020) and two-thirds of major U.S. companies now report using freelance contract workers (Arruda 2020). Gig workers provide labor to many industries – with ride-hailing, food delivery, casual labor and similar online platform businesses representing the fastest-growing sectors. Leading companies in this group – Uber, Lyft, Didi, Doordash, Airbnb – have valuations in the $10-100B range, exceeding the value of long-standing incumbents in transportation, food service and hospitality. In short, gig economy companies and gig work have become a major feature of the modern economy.

Yet gig economy work is fundamentally different than traditional employment. Many gig jobs do not require substantial training from the hiring firm[1] and do not involve long-term commitments between a firm and its contractors. Indeed, the unit of contracted work is usually quite small – often a single task or job (e.g., providing a ride or making a delivery).

Gig workers also have the flexibility to work as much or as little as they like, and they have the freedom to sign up and work for multiple firms to improve their earnings – often switching between different platforms in real time (called

[1]For example, see this article for a list of common gig jobs: *What kinds of work are done through gigs?*, Gig Economy Data Hub, `https://www.gigeconomydata.org/basics/what-kinds-work-are-done-through-gigs`

“multi-apping"). For example, it is common for ride-share drivers in the U.S. to accept rides from both Lyft and Uber, and to switch to food delivery apps during lunch and dinner time[2]. As estimated by Chen, Rossi, and Chevalier 2019, this flexibility is extremely valuable for gig workers, providing twice the surplus they would otherwise derive from less flexible work.

Indeed, multi-apping among gig workers is gaining in popularity. Working for multiple firms is perceived to have many benefits. In addition to reducing idle time and increasing utilization for workers, it also helps hedge the risk of major down time due to unexpected events like app outages or being banned from a certain platform. And the threat of losing the ability to multi-app has been cited as a key reason for the successful passing of Proposition 22 in California, a ballot measure jointly backed by Uber, Lyft, Postmates, DoorDash and Instacart to exempt app-based gig workers from classification as employees[3].

Multi-apping is often encouraged by firms too, especially new entrants to a market. For example, when Juno (who became at one point the second-largest ride-hailing company in New York City (NYC) by driver count) entered the NYC ride-hailing market, it paid Uber drivers $25/week to keep the Juno app open when they drove for Uber (Solomon 2016).

Not all gig workers are “multi-homing”, since signing up and learning to work for a new firm is costly. Nonetheless, using bank data that tracked workers’ earnings up to 2018, the JPMorgan Chase Institute (Farrell, Greig, and Hamoudi 2018) report a steady increase in the percentage of gig transportation workers that are multi-homing, from approximately 2.5% in 2012 to over 20% in 2018.

[2] *Don’t Put All Your Delivery Eggs in One Basket – Keep Your Earnings Options Open*. The Entre Courier. `https://entrecourier.com/2019/07/05/know-your-revenue-options/`

[3] See `https://yeson22.com/` for example.

So while not all gig workers are multi-homing, multi-homing is a fundamental benefit of gig work and in many ways defines what it means to be a “gig worker".

At a market level, when gig workers take jobs from multiple sources, they effectively form a common pool of labor that can be accessed on an on-demand basis by multiple firms. The structure of a shared labor pool raises interesting questions for both workers and firms. For example, when workers generate earnings from multiple platforms (such as drivers that both work for ride-hailing and food-delivery), how do they decide whether or not to work and which jobs to accept while working? When a firm sets a per-job pay rate for workers who are also taking jobs from other firms, how do their decisions differ from traditional wage-setting? Do all firms, in equilibrium, share equally in the cost of maintaining the worker pool or do some firms pay more than others? And how is such a labor pool formed in the first place? Under what conditions would the first firm enter such a market and when would subsequent firms follow? Our theory provides precise answers to these questions which yield important insights about the formation and structure of the gig economy.

We model a gig economy as a set of firms that utilize workers from a common pool. Firms receive stochastic arrivals of potential demand (jobs) and set the amount they pay workers from the pool to perform jobs. Firms are not able to serve a job if all workers in the pool are busy. This means firms benefit from some level of idleness in the worker pool as a reserve (buffer) against demand uncertainty. The revenue each firm receives for each job is fixed [4] and firms seek to maximize their profits, which are their total revenues minus their total labor

[4]The assumption of fixed revenue (equivalently fixed prices) is both for simplification and to focus our analysis on the labor market consequences of the common pool of workers. One can also think of this as arising from a product market that is perfectly competitive. But more general pricing and demand-side effects are beyond the scope of our analysis here.

cost for all jobs they are able to serve. Workers in the pool are perfectly flexible; jobs from different firms may provide different end services, but can be fulfilled by any worker in the pool – just as ride-hailing and food delivery are different services but may be served by the same workers.

Workers, in turn, have the freedom to choose whether to join the pool or not and if they do join, which jobs to accept. They seek to maximize their expected hourly earnings subject to earning at least a reservation wage rate (their reservation wage) for not participating. The total supply of workers participating in the pool is therefore determined by their aggregate earnings across all firms. Aware of the effects of their pay on both their own profit margin as well as the overall participation of workers in the common pool, firms select a pay rate that maximizes their expected profit. We analyze the resulting equilibrium in this gig economy and prove a number of interesting properties.

Our first finding is that the smaller [5] a firm is, the less it pays workers for jobs. This means when firms are sharing workers, there is a market force that gives smaller firms a labor cost advantage. Indeed, we show sufficiently small firms only need to pay the workers' reservation wage. For firms that pay strictly higher than the reservation wage, their per-job profit margin is inversely proportional to their job arrival rate. This means all such firms earn the same profit regardless of size, so there is a very strong dis-economy of scale due to the shared worker pool.

The intuition is that the pay rates of larger firms have a greater impact on worker participation in the shared worker pool. When a large firm pays too

[5]The size of the firm refers to the amount of potential earnings if all their job requests get fulfilled. For example, "smaller" firms may have a lower rate of job requests, or a lower expected revenue for completed jobs

little, it significantly reduces workers' average earnings due to the large share of jobs it provides. These lower earnings reduce the reserve of idle workers, which reduces the ability of all firms to complete jobs. In contrast, if a small firm reduces its pay for jobs, it does not significantly affect workers' overall earnings and participation in the pool, so small firms can lower their pay without affecting the reserve supply of workers – and hence the pool's ability to serve jobs – by much. In short, while the direct benefit of lower pay in terms of increasing the marginal profit on each job served is the same for both large and small firms, the indirect cost of lower pay on reduced worker participation in the pool is much higher for large firms than for small firms. Therefore, in equilibrium, larger firms end up carrying more of the burden of maintaining the worker pool.

In an extreme case where the size differences between large and small firms are significant, in equilibrium, small firms can get away with paying the minimal pay rate for jobs (i.e., the workers' reservation wage) without contributing any payment to workers' idle time, completely relying on the large firms to maintain the buffer of idle workers in the pool. This is in essence a free-rider problem as described in Olson 1965, Stigler 1974, Groves and Ledyard 1977, etc. From this perspective, idle gig workers waiting for jobs can be viewed as a shared public resource that is non-excludable (due to the flexible work arrangement) and partially non-rivalrous (due to the stochasticity in the job arrival), creating the possibility for small firms to access this collective resource without contributing to the aggregate compensation for workers' idle time.

The result that larger firms pay more than smaller firms has important implications. For one, it contradicts the conventional wisdom that gig economy companies enjoy "winner takes all" markets. For example, in the prospectus

for its initial public offering, Uber highlighted their "massive network" and "creating a liquidity network effect" as two of its most important competitive advantages; they also state that: "Generally, for a given geographic market, we believe that the operator with the larger network will have a higher margin than the operator with the smaller network." [6] Our results suggest quite the opposite.

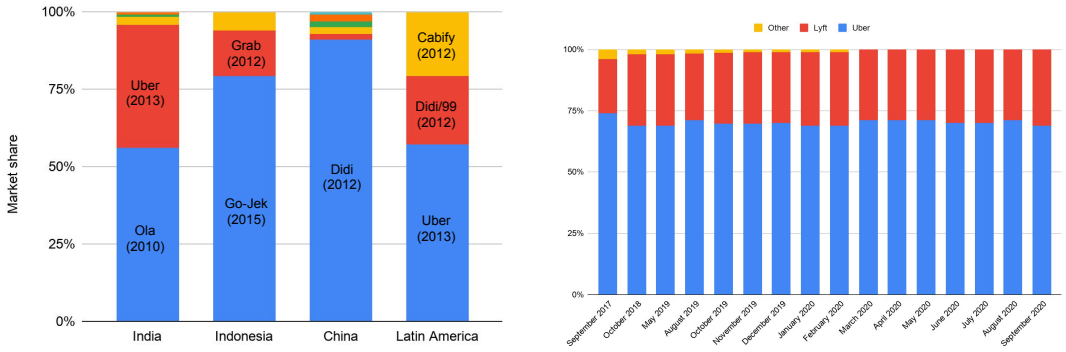
In fact, despite these claims that network effects favor larger firms in ride-hailing markets, empirical evidences suggest that dominance by a single firm is rare. For example, Fig. 4.1a shows the major ride-hailing companies, their market shares, and the year they entered four international markets in India, Indonesia, China, and Latin America, respectively. Among the four markets, China (a market with significant government intervention) is the only one in which a single firm dominates the market with more than 90% market share; for the other markets, oligopoly is clearly the more common outcome.

In the U.S. – the oldest ride-share market – we also do not observe larger firms gaining share over time, which would tend to be true if size conveyed a significant advantage. This is shown in Fig. 4.1b, which plots the market composition in the U.S. ride-hailing market from September 2017 to September 2020. (This is the entire post-IPO period for Uber and Lyft.) Fig. 4.1b shows that Uber and Lyft's market shares over these years are quite steady, suggesting a stable market structure with Lyft maintaining a viable position despite Uber's larger size.

Although these are limited empirical examples, they are consistent with the general hypothesis that large firms are not dominating in the gig economy. That

[6] Form S-1, Uber Technologies, Inc., https://www.sec.gov/Archives/edgar/data/1543151/000119312519103850/d647752ds1.htm.

is, despite other network and fixed-cost advantages that may benefit large firms [7], our work points to a compensating drag on large firms' costs due to sharing a pool of workers with competitors. As such, our results provide a new perspective on scale advantages (or lack thereof) in the gig economy.



(a) Market share of the leading ride-hailing companies in India, Indonesia, China, Latin America region and their years of market entry. (Time of data collection: India (December 2017), Indonesia (April 2018), China (2018), Latin America (September 2018). See sources in Appendix C.7.)

(b) Trend of the market share of the leading ride-hailing companies in the US from Sept. 2017 to Sept. 2020. (Source: *Uber vs. Lyft: Who's tops in the battle of U.S. rideshare companies*, Second Measure, `https://secondmeasure.com/datapoints/rideshare-industry-overview/`)

Figure 4.1: Market shares in various ride-hailing markets.

Despite the labor cost advantage that small firms enjoy, our results also show that small firms may need the existence of large firms to be profitable – and indeed may be unable to form a market on their own. If small firms were to form a gig economy on their own, their low demand rates may force them to pay a much higher rate to compensate for workers' idle time. Indeed, if their demand rates are too low, then the minimum pay to maintain the worker pool may be too high to break even. However, once a large firm creates a market, it lowers the barrier for small firms to subsequently enter. In other words, with a large incumbent that has already created a worker pool, smaller firms – who might not

[7] For example, Nikzad 2020 shows an economy of spatial density in ride-hailing can confer a scale advantage.

be viable on their own – can utilize the existing labor at a much lower cost. This is consistent with industry history, where we have seen large market leaders, like Uber in ride-hailing, spend massively to establish new markets, which then open the door for smaller players in ride-hailing and other gig businesses to prosper.

Indeed, our theory supports a more speculative conjecture that perhaps Uber – an unprecedentedly well-capitalized firm with an appetite for rapid growth – was perhaps the "Johnny Appleseed" of the gig economy, creating the conditions for worker pools to form in many regional markets across the globe. While such a conjecture clearly warrants proper empirical validation, it is roughly consistent with casual empiricism on how the gig economy has evolved. For example, according to the US Census Bureau, following the founding of Uber, the number of self-employed drivers in the U.S. more than tripled from 2013 to 2016. [8] The important role of large firms in forming a gig economy also has unconventional implications for anti-trust regulation of the sector.

The remainder of the paper is organized as follows: we discuss the related literature in Section 4.2. In Section 4.3, we then introduce our base model and assumptions. In Section 4.4, we provide our main results and illustrate them with numerical examples. Lastly, in Section 4.5, we provide a full characterization of a more general version of our model.

[8] See the trend of self-employed drivers from 1997 to 2016 here: *Detailed Look at Taxi, Limousine Services*, United States Census Bureau, https://www.census.gov/library/stories/2018/08/gig-economy.html This study from the Aspen Institute also cites Uber and Lyft as the main players driving the overall growth in the transportation sector based on the US Census Bureau data: *The Gig Economy: Research and Policy Implications of Regional, Economic, and Demographic Trends*, Page 4, Aspen Institute, https://assets.aspeninstitute.org/content/uploads/2017/02/Regional-and-Industry-Gig-Trends-2017.pdf

4.2 Related literature

Empirical evidence has suggested a positive relationship between wages and firm size (see Moore 1911, Brown and Medoff 1989, Idson and Oi 1999 for examples) and multiple theories have been proposed to explain this effect. For example, Idson and Oi 1999 shows that larger firms have higher capital-labor ratios and are early adopters of new technologies, and thus have a higher performance standard for workers. As a result, to properly incentivize workers and prevent skilled workers from taking other job offers, larger firms choose to compensate them with higher wages. Other explanations include that larger firms offer inferior work conditions, have a greater ability to pay higher wages, and are less able to monitor workers (Brown and Medoff 1989). Despite the similarity in outcomes, the driving forces in these explanations are qualitatively different from ours. In this literature, the notion of a shared labor pool does not exist; workers are full-time employees who only work for one firm at a time. Firms' wage decisions are thus only driven by their own characteristics, like hiring standards or working environment. In contrast, our setting is specific to the modern-day gig economy: workers are homogeneous, paid per-job and accessible to all firms in the gig economy. Since workers' participation decisions depend on their aggregate earnings across all gig economy firms, which are generated by all firms jointly, there is an inherent externality in firms' wage-setting decisions. These externalities in wage-setting are central to our theory, but not considered by the previous literature.

In addition, in gig economies workers may not have fixed schedules and jobs may arrive stochastically over time. As a result, gig workers are typically not necessarily fully utilized during their working hours and may have to spend

some time waiting without pay to receive job requests. For gig workers, the proportion of time they spend waiting can be interpreted as a form of very short-term "unemployment", which links our work to the literature on the natural rate of unemployment (see Mortensen et al. 1986 for a comprehensive review). Our work shares some similarities with this literature in the way we perceive workers' participation decisions. For example, Burdett et al. 1984 models an individual's labor market history as a Markov Chain with three states: employment, unemployment, and non-participation. Events that will affect the individual's utility in each state, such as receiving a job offer or getting fired, arrive in a Poisson process. The individual decides which state to occupy with the objective of maximizing their lifetime utility. In our model, workers have similar states: fulfilling jobs, waiting for jobs and not participating in the gig economy. Moreover, they are also utility maximizers who decide to transition states based on the expected wage. Despite these similarities, this prior literature mostly focuses on explaining the empirical evidences of aggregate unemployment rates and does not address the connection to firms' wage-setting, which is the primary focus of our work.

Our paper is also broadly related to the burgeoning literature on two-sided market competition, which studies the impact of multi-homing on a platform's profit and social welfare (e.g. the seminal work of Rochet and Tirole 2003, Rochet and Tirole 2006; works on competition among ride-hailing platforms by Bryan and Gans 2019, Ahmadinejad et al. 2019, Nikzad 2020, Loginova, Wang, and Liu 2019, etc.). In particular, Tan and Zhou 2020 shows when increasing the level of platform competition, consumer surplus can decline and platform profits can increase, which is contrary to the conventional intuition on the effects of competition for single-sided platforms. Ahmadinejad et al. 2019 also studies competition between two ride-hailing platforms. Interestingly, they also view

available workers as a shared resource and show that competition can lead to market failure in the form of a "tragedy of the commons" type outcome. However, since competition is the central topic of this stream of literature, it focuses primarily on how two-sided competition among platforms providing similar services affects surplus and prices. In contrast, neither product-market competition nor two-sided platforms are elements of our model; the only competition in our setting is over attracting workers. Hence, our results on cost advantages and firm size derive solely from the feature that gig economy firms share a common pool of workers.

4.3 A model of the gig economy

As noted, two fundamental characteristics of the gig economy are that workers have the flexibility to "multi-home" (i.e., work for several firms) and are paid for jobs not for their idle time between jobs. These workers form a common labor pool. A firm can change how much workers are paid for its jobs, but the total number of workers participating in the pool is determined by their aggregate earnings across all firms. Yet the size of the pool matters for firms because demand is stochastic and hence the larger the pool, the more likely it is that any given firm can find an available worker when it gets a job request. In this sense, the pool of available workers forms a shared resource that each firm both benefits from and contributes to through their choice of pay rates. How these worker pay choices are determined in equilibrium is our main focus.

Firms and jobs There are N firms that provide jobs to gig-workers. Each firm i ($1 \leq i \leq N$) receives new jobs requests (demand) at a mean rate μ_i, which is assumed constant over time. Job arrivals are stochastic and modeled as independent Poisson processes with rate μ_i. We let $\mu = \sum_{i=1}^{N} \mu_i$ denote the total job arrival rate. While job arrival rates in practice vary because of seasonality effects, a stationary setting allows us to simplify our analysis and is sufficient to establish the main equilibrium effects.

Each job generates a fixed revenue of v per hour and takes an expected time of T hours for a worker to complete. The revenue v and job time T are considered to be the same for all jobs and firms, which are just for ease of exposition and are relaxed in Section 4.5. Each firm i chooses the hourly payment rate to workers, p_i, for its jobs.[9] The set of all such payment rates is denoted $\mathbf{p} = (p_1, p_2, ..., p_N)$. Thus, a completed job from firm i generates a net profit of $(v - p_i)$ per hour. Such a pay structure with pay and revenue being proportional to job duration is common for gig jobs. For example, in ride-hailing, driver pay and passenger prices are largely based on time and distance of the trip[10].

Pool of Workers Workers are flexible and able to accept jobs from any firm. They join the pool of available workers after they finish jobs, and leave the pool when they accept a job. Let λ denote the rate at which they join (and leave) the pool. We assume the arrival process of workers to the pool is Poisson with rate λ.

[9] In practice, for some platform companies like TaskRabbits, the pay is chosen by workers instead of directly controlled by the firms, and firms only charge a transaction fee. Nonetheless, our analysis will carry through by considering workers' pay as the pay set by themselves minus the transaction fee set by firms.

[10] *Regulated driver pay rates in New York City*. NYC Taxi & Limousine Commission. `https://www1.nyc.gov/site/tlc/about/driver-pay-rates.page`

When a worker gets a job, they choose whether or not to accept it [11]. This decision depends on the job pay rate p_i. For example, if a food-delivery company pays much less than its competitors or other gig firms using the pool, workers may simply not be willing to accept their jobs. If a worker accepts a job, they work for some expected time T, get an hourly pay p_i and firm i gets an expected profit $T(v - p_i)$. If they reject the job, they simply stay in the pool and wait for another job. If there are no workers available when a job request arrives, the firm with the job to serve loses that job. The assumption of the Poisson arrival of workers – and that workers only leave the pool when they accept a job – is again for simplicity, and can be relaxed as shown in Section 4.5.

This model is a simple representation of why firms value maintaining a large pool of workers. It is indeed preferable for gig firms to have λ as high as possible in order to minimize the risk of having no available workers when their jobs arrive. On the other hand, the value λ depends on the workers' decisions. For example, if λ increases, the queue of waiting workers becomes large, which lowers workers' utilization and, hence, their hourly earnings. Workers may then decide not to join the pool, which would decrease λ. In our model, λ is actually endogenous and depends on both workers' utilization and the firms' prices, as discussed next.

Workers Equilibrium Next consider worker participation in the pool. We assume that the supply of workers is perfectly elastic with reservation wage w_0; that is, we assume that any number of workers will be willing to join the gig economy if their expected hourly earnings are at least w_0 and that workers will refuse to join the gig economy if they earn less than w_0. This assumption

[11] We will discuss how jobs are assigned in later paragraphs.

of perfect elasticity is not fundamental to our analysis but does simplify it. Moreover, it has been observed empirically that gig labor markets like ride-hailing are indeed highly elastic. [12] To avoid trivialities we assume that $w_0 < v$, else firms' revenues for jobs do not exceed the reservation wage of workers and hence there is no opportunity for any firm to make a positive profit.

When workers arrive at the pool, they wait until they accept a job from some firm i and then get paid an hourly rate p_i for some expected time T. Let W be the expected wait time until they accept a job, and P their expected earnings from the job (which depends on which jobs they are offered and the hourly payment of the firms). Then, at any steady-state worker equilibrium, we must have that

$$P = w_0(W + T) \tag{4.1}$$

Note that we include the workers' waiting time W, as the workers have to be available when they wait for a job, and so we need to include this wait time to compute their average earnings. This is true because if $P > w_0(W + T)$, then the expected earnings per hour in the gig economy is $P/(W + T) > w_0$, and more workers would be willing to join the pool, which increases λ and consequently increases the wait time W (as more workers share the same number of jobs), until we reach $P = w_0(W + T)$. Similarly, if $P/(W + T) < w_0$, then workers would refuse to join the pool, which lowers λ and decreases the wait time for the other workers, until again $P = w_0(W + T)$ is restored. Alternatively, one can view w_0 as the opportunity cost of workers' time. Hence, the expected worker surplus from a job is $\Pi = P - w_0(W + T)$. Perfect supply elasticity implies that their expected surplus must be zero and (4.1) must hold.

[12] For example, Hall, Horton, and Knoepfle 2020 showed that, after an adjustment period of about two months, Uber drivers' earnings revert to their previous equilibrium value whenever driver pay is changed – an outcome consistent with perfectly-elastic supply.

This equilibrium is not yet fully specified since P and W depend on the workers' strategy for accepting or rejecting the jobs as well as the vector of payments offered by firms, which is an outcome of the firms' strategic interactions. We look at both next, starting with the workers' strategy.

Workers' strategy Here we provide an informal description of the workers decision process and optimal strategy for ease of exposition. Appendix C.2 in the appendix contains a detailed micro-foundations model formally proving the results.

We assume workers are rational and risk-neutral, so they choose the jobs they are willing to accept in order to maximize their expected surplus. Consider a worker's decision when faced with a job request from firm i. If the worker accepts the job, they will get the hourly pay p_i for the duration T of the job, incur an opportunity cost $w_0 T$ and then depart the system – earning their reservation wage w_0 thereafter. Note that even if a worker returns to the pool for another job, they would still earn their reservation wage because of the equilibrium condition (4.1). Therefore, a worker's expected future surplus from accepting job i is $p_i T - w_0 T$.

If the worker rejects job i, they rejoin the pool of workers and will wait for an expected time W_{next} until they finally accept a job for an expected pay P_{next}. Therefore, the choice to wait produces an expected surplus of $(P_{\text{next}} - w_0 T) - w_0 W_{\text{next}}$. Note that we do not necessarily have that $P = P_{\text{next}}$ or $W = W_{\text{next}}$ as the expected wait time and pay of a newly-arriving worker is not necessarily the same as that of a worker that has been waiting. (Equality strictly holds only when both the waiting time and job arrival process are memoryless.)

Putting these choices together, the worker should accept job i if and only if

$$p_i T - w_0 T \geq P_{\text{next}} - w_0(T + W_{\text{next}}) \iff p_i T \geq P_{\text{next}} - w_0 W_{\text{next}} \tag{4.2}$$

We assume that in the case of equality in (4.2), the workers choose to accept job i. Note the left-hand side above is the total pay for job i while the right-hand side is the expected pay of a future job minus the waiting cost to get the next job. This means that workers will accept jobs that pay less than their expected pay P_{next}, because this can be better than incurring the waiting cost for a higher-paying future job. This simple fact is central to our results and is the key characteristic of the gig economy that enables smaller firms to pay less.

That workers are willing to accept a job that does not pay the most is similar to finding from search theory (e.g. Stigler 1961, McCall 1970, Mortensen et al. 1986). Unemployed workers in a traditional labor market have to make sequential decisions about whether to accept a job offer, where there is a cost for searching. This results in an optimal-stopping problem, implying there is a reservation wage above which workers find it optimal to accept a job. As in our model, the optimal reservation wage is always below the highest-possible wage.

As noted, the values W_{next} and P_{next} in general depend on the process that matches requests to workers. However, a reasonable and greatly simplifying assumption is:

Assumption 7 (Memoryless Beliefs) *Workers believe that* $W_{\text{next}} = W$ *and* $P_{\text{next}} = P$

That is, a worker who rejects a request expects to wait a similar time and expects to have similar pay for their next job as a worker who just arrived. This

assumption can be justified on both physical and behavioral grounds. It is a good approximation in ride-hailing, where the ride requests process is approximately Poisson in space and time and requests are allocated, approximately, to the nearest driver (Yan et al. 2020). The model is less accurate in first-come-first-serve settings, where workers that have waited longer are prioritized. That said, one could still derive similar results for this case, albeit with a more complicated analysis. On behavioral grounds the assumption can be justified as a form of bounded rationality; if the job assignment process is obscure, workers have limited information about the state of the system and/or workers have difficulty estimating their expected residual waiting time, then assuming workers use the heuristic $W_{\text{next}} \approx W$ and $P_{\text{next}} \approx P$ is plausible.

Assumption 7 significantly simplifies (4.2). Using $W_{\text{next}} = W$, $P_{\text{next}} = P$, and (4.1), we get that $w_0(T + W_{\text{next}}) = P$. Therefore (4.2) becomes:

$$p_i \geq w_0 \tag{4.3}$$

In other words, workers will accept any job that pays at least w_0 and reject all other jobs. This simple strategy maximizes their expected earnings. However, this does not mean all firms can pay w_0; if they did, (4.1) could not hold as workers would, on average, make strictly less than w_0 per unit time accounting for their waiting time. Hence, no worker would join the pool, λ would be zero and all firms would make zero profit.

So we have a simple market outcome: firms that pay $p_i < w_0$ will have all their requests rejected and firms that pay $p_i \geq w_0$ get have their requests accepted, provided there is a worker available in the pool. We can then evaluate the wait time, W, as a function of the workers' arrival rate λ and the vector of firm payment rates $\boldsymbol{p}$. This result is confirmed and proved formally in Appendix C.2,

where we study the worker's decision problem as a Markov Decision Process.

Formally, in queuing theory terminology, our model reduces to a queue where the queue length is the number of available workers in the pool and the service time is the inter-arrival time of jobs from any firm that pays $p_i \geq w_0$. This is equivalent to a queue with a single server where customers arrive in a Poisson process and job service time follows an exponential distribution, i.e. an M/M/1 queue (Morse 2004). This implies a service rate of $\mu_{\geq w_0} = \sum_{i,\, p_i \geq w_0} \mu_i$ (the total job rate of firms that pay above w_0). Thus, for a given service rate $\mu_{\geq w_0}$, the expected waiting time is a function of the arrival rate λ is

$$W(\lambda; \mu_{\geq w_0}) = \frac{1}{\mu_{\geq w_0} - \lambda} \tag{4.4}$$

Equation (4.4) confirms that when there are more workers in the gig economy (higher λ) and fewer jobs available (higher $\mu_{\geq w_0}$), workers must wait longer to get a job.[13].

Note an accepted job has a probability $\mu_i / \mu_{\geq w_0}$ to belong to firm i if $p_i \geq w_0$. Importantly, all firms that pay at least w_0 have the same job completion rate $\lambda / \mu_{\geq w_0}$, which means that all firms benefit equally from maintaining a large pool of workers. Additionally, we can also evaluate the expected pay P from an accepted job, which is simply the weighted average

$$P = T \sum_{i,\, p_i \geq w_0} \frac{\mu_i}{\mu_{\geq w_0}} p_i \tag{4.5}$$

Thus, when workers' arrivals are in the steady state, i.e. when Eq. (4.1) hold,

[13]We will generalize our models to more complex settings in Section 4.5, where we will consider more general waiting functions $W(\lambda; \mu_{\geq w_0})$ that cover a larger class worker arrival and job assignment processes.

combining Eq. (4.4) and Eq. (4.5) implies that workers' equilibrium must satisfy

$$T \sum_{j,\, p_j \geq w_0} \frac{\mu_j}{\mu_{\geq w_0}} p_j = w_0 (T + W(\lambda; \mu_{\geq w_0})) \tag{4.6}$$

Eq. (4.6) reveals the relation between the worker's equilibrium arrival rate λ and the vector of firm payment rates $\boldsymbol{p}$. The higher the aggregate equilibrium pay, the higher the arrival rate for workers. In the case where the average pay is too low, the equilibrium in Eq. (4.6) may not be feasible (i.e. when the left-hand side of Eq. (4.6) is lower than $w_0(T + W(0; \mu_{\geq w_0}))$). This means the aggregate equilibrium pay is simply too low for workers to earn their reservation wage. Thus, there is no workers' participation, leading to $\lambda = 0$.

Combining the two possibilities and rearranging Eq. (4.6), the endogenous equilibrium arrival rate λ is given by the following function of $\boldsymbol{p}$:

$$\lambda(\boldsymbol{p}) = W^{-1}\Big(\max\big(T \sum_{i, p_i \geq w_0} \frac{\mu_i}{\mu_{\geq w_0}} \frac{p_i}{w_0} - T, W(0; \mu_{\geq w_0})\big); \mu_{\geq w_0} \Big) \tag{4.7}$$

$$= \left(\mu_{\geq w_0} - \frac{w_0}{T \sum_{j,\, p_j \geq w_0} \frac{\mu_j}{\mu_{\geq w_0}} p_j - T w_0} \right)^+ \tag{4.8}$$

where $W^{-1}(\cdot, \mu_{\geq w_0})$ is just the inverse of the waiting time $W(\lambda; \mu_{\geq w_0})$ in Eq. (4.4) and is thus parameterized by $\mu_{\geq w_0}$ as well.

Firms strategy We next examine firms' strategies. We assume Nash behavior of firms; that is, each firm i chooses their labor pay rate p_i to maximize their expected profit given the other firms' pay rates, which we denote as p_{-i}. Using the notation $\lambda(p_i; p_{-i})$ to denote the arrival rate $\lambda(\boldsymbol{p})$ when p_{-i} is fixed and p_i varies, the expected profit rate of firm i in the steady state is

$$\pi(p_i; p_{-i}) = \begin{cases} \frac{\mu_i}{\mu_{\geq w_0}} \lambda(p_i; p_{-i})(v - p_i) T & \text{if } p_i \geq w_0 \\ 0 & \text{otherwise} \end{cases}$$

In words, the profit of the firm paying at least w_0 is simply the expected profit per successful job $(v - p_i)T$, multiplied by the frequency of successful jobs $\frac{\mu_i}{\mu_{\geq w_0}}\lambda(p_i; p_{-i})$. A firm has zero profit if it pays less than w_0 since no worker is willing to accept their jobs.

Using this closed-form profit function, we can formulate firm i's decision problem: Given p_{-i}, either pay $p_i < w_0$ and get zero profit or pay $p_i \geq w_0$ and choose p_i by solving the following best-response problem:

$$\max_{p_i} \quad \frac{\mu_i}{\mu_{\geq w_0}}\lambda(p_i; p_{-i})(v - p_i)T \cdot \mathbb{1}_{\{p_i \geq w_0\}} \tag{4.9}$$

To achieve the optimal profit firm i faces a trade-off: its profits are proportional to the rate of jobs $\lambda(p_i; p_{-i})$ and to raise $\lambda(p_i; p_{-i})$, firm i must raise its pay p_i to increase the left-hand side of (4.6). However, raising its pay p_i also reduces its profit margin per job completed.

The best-response problem illustrates the complex interaction of firms: all firms benefit from a large pool of workers (and therefore high λ) and they each contribute to increasing λ through (4.7). But if other firms are paying enough to maintain a high λ, then each firm has an incentive to lower their own pay rate to increase their marginal profit per job served. To understand how these interactions play out, we next analyze the Nash equilibria of this economy.

4.4 Nash equilibria

In this section, we analyze the equilibrium pay rates and illustrate our results with numerical examples.

4.4.1 Larger firms pay more

We say that firms are "participating" in the gig economy if they set their pay rate above the reservation wage $p_i \geq w_0$ (so that the workers accept their job requests). Our first result shows that in a Nash equilibrium, if the worker pool is not empty, then all firms participate.

Proposition 24 (Firms participation) *Only two types of Nash equilibria are possible: either some workers participate and all firms are profitable ($p_i \in [w_0, v)$ for all i and $\lambda(\boldsymbol{p}) > 0$), or no worker participates and none of the firms make profit ($\lambda(\boldsymbol{p}) = 0$).*

(All proofs are in the Appendix.) Proposition 24 states that there is no equilibrium with a subset of firms making a positive profit while others do not participate. The intuition is as follows: whenever at least one firm is making profits, it means that there is a worker pool supported by this firm. Given the existence of a worker pool, additional firms can attract workers to accept their jobs as long as they pay at least w_0. Since firms make a positive profit on each such job assuming $v > w_0$, a firm always benefits by participating. [14] Thus, the only possibility for a firm to not participate is when none of the other firms are participating and making profits.

[14] This is because we are considering the simplified setting where all firms have the same hourly revenue v. In the general case considered later, some firms may not be able to enter if their revenue is not high enough to generate a profit.

We are most interested in the equilibria of a viable gig economy in which $\lambda > 0$ and, by Proposition 24, all the firms make positive profits. Our main theorem characterizes the equilibria in this viable case:

Theorem 7 *Consider a gig economy with N firms that have job arrival rates $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_N$. Then the following hold:*

1. *A Nash equilibrium with all firms participating exists if and only if the total job arrival rate μ is high enough to satisfy*

$$\mu > \frac{w_0}{T(v - w_0)} \tag{4.10}$$

2. *The equilibrium is unique.*
3. *At equilibrium, the larger the firm, the more it pays workers:*

$$p_1 \geq p_2 \geq ... \geq p_N$$

4. *All firms with $p_i > w_0$ must have the same hourly profit $\pi(p_i; p_{-i})$ at equilibrium. That is, their profit margin $v - p_i$ is inversely proportional to their job arrival rate μ_i.*
5. *It is possible for smaller firms to have a lower hourly profits than the firms that pay $p_i > w_0$, but these firms pay the minimal rate $p_i = w_0$ and therefore have the highest-possible profit margin.*

Results 1 and 2 above state that as soon as the combined demand of all firms is enough to sustain a gig economy, a unique profitable equilibrium exists. Note that Eq. (4.10) is equivalent to $Tv \geq w_0(T + 1/\mu)$, which simply states that the maximum hourly revenue Tv must exceed the workers' opportunity cost

for the minimum expected wait time workers must spend to complete a job, $T + 1/\mu$. (This is true because $1/\mu$ is the average wait time for a job to arrive.) Interestingly, the minimum job arrival rate required for a viable gig-economy Eq. (4.10) is independent of the number of firms N; only their combined job arrival rate matters.

Results 3, 4 and 5 describe the equilibrium. Surprisingly, each firm either pays the minimum rate w_0 or has the same hourly profit as the largest firm. Considering “large firms" as firms that pay more than w_0 and “small firms" as those that pay w_0, this means all large firms (regardless of size) have the same total profit while all small firms enjoy the highest-possible profit margin. This is a rather striking form of dis-economy of scale. Also, these results imply the hourly pay to workers is (weakly) increasing in firms' job arrival rates (size).

4.4.2 Intuition

An explanation of Theorem 7 is that while all firms suffer the same loss of profit margin from increasing pay, a large firm that increases its pay will increase the job completion rate λ more than a smaller firm. Thus, larger firms have more of an incentive to increase their pay. Conversely, when a small firm raises its pay it has little effect on the size of the pool of workers, which leads to them paying less. The striking feature is the degree of this effect; the increased pay of large firms is so high that they end up making the same total profit regardless of their size. Also, small firms below a certain size are able to “free ride" and get away with paying workers the minimal rate possible, w_0 – relying entirely on large firms to support the common worker pool.

To better understand the intuition behind Theorem 7, we can reformulate the problem in terms of profits rather than prices. That is, instead of the firms choosing the payments p_i, consider instead that they choose their normalized profit margin, q_i, defined by

$$q_i \triangleq \frac{\mu_i}{\mu_{\geq w_0} w_0}(v - p_i)T, \tag{4.11}$$

where q_i is the expected profit of firm i per unit of supply normalized by the reservation wage w_0. Thus, firm i's normalized profit π is equal to $q_i\lambda$ (where one unit of hourly profit is w_0). Whenever the worker arrival rate λ increases by 1 unit, a fraction $\mu_i/\mu_{\geq w_0}$ of this increased supply is rationed to firm i, and generates $(v - p_i)T$ net profit for the firm. Note this means that at the same payment level, firms with higher market share benefit more from increasing the pool size, and have a larger q_i than smaller firms.

Furthermore, define

$$k_i \triangleq \frac{\mu_i}{\mu_{\geq w_0} w_0}(v - w_0)T \tag{4.12}$$

k_i is similar to q_i except the p_i in (4.11) is replaced by w_0. In words, k_i is the total welfare per unit of supply firm i generates by participating in the gig economy – again normalized by w_0.

With q_i and k_i, we can now reformulate the decision problem of the firms in a much simpler way. The equilibrium condition Eq. (4.6) becomes:

$$\sum_{j, q_j \leq k_j}^{N} k_j = \sum_{j, q_j \leq k_j}^{N} q_j + W(\lambda; \mu_{\geq w_0}) \tag{4.13}$$

Therefore the arrival rate $\lambda(\boldsymbol{p})$ from Eq. (4.7) can also be written as a function of $\boldsymbol{q} = (q_1, ..., q_N)$:

$$\lambda(\boldsymbol{q}) = W^{-1}\Big(\max\big(\sum_{j, q_j \leq k_j}^{N} (k_j - q_j), W(0; \mu_{\geq w_0})\big); \mu_{\geq w_0}\Big) \tag{4.14}$$

and firm i's best response (4.9) can simply be written as:

$$\max_{q_i} \quad q_i \lambda(q_i; q_{-i}) \cdot \mathbb{1}_{\{q_i \leq k_i\}} \tag{4.15}$$

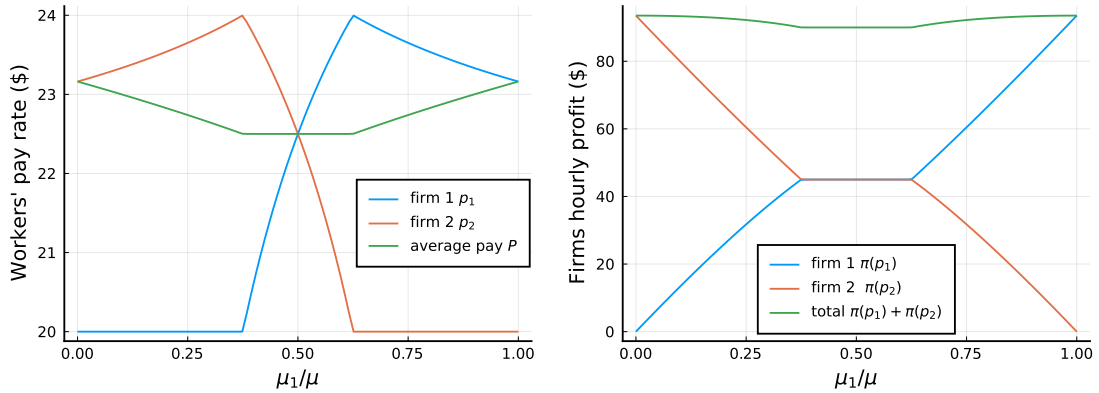
That is, firm i selects the optimal profit margin to maximize the hourly profit given by (4.15). The hourly profit depends on both the margin q_i and the size of the worker pool $\lambda(q_i; q_{-i})$. λ is governed by the workers' equilibrium condition (4.14).

The reformulation provides a clear and intuitive interpretation of the worker's equilibrium. From (4.13), the left-hand side, $\sum_{j, q_j \leq k_j}^{N} k_j$ can be interpreted as the aggregate welfare created by all firms. This aggregate welfare is either extracted by firms as their profits, represented by $\sum_{j, q_j \leq k_j}^{N} q_j$, or is consumed by compensating workers for waiting in the pool, represented by $W(\lambda; \mu_{\geq w_0})$. For example, if a firm chooses to pay the minimum rate $p_i = w_0$, then we have that $k_i = q_i$ and the left and right-hand-side terms for firm i in (4.13) cancel each other, so firm i does not contribute to maintaining idle workers in the pool. Equation (4.13) illustrates clearly that the size of the waiting worker pool is determined by the total welfare leftover in the economy after firms take their profits. Thus, for each firm, the more profit margin they take for themselves, the less they contribute to a common pool.

The reformulation also allows us to understand why, at equilibrium, the firms that do not free ride ($q_i < k_i$) all make the same profit. To see this, note that Eqs. (4.14) and (4.15) are symmetrical (locally) for all firms with $q_i < k_i$. This observation shows that the non free-riding firms must have the same equilibrium profit.

4.4.3 Examples

We next provide a few numerical examples to illustrate the main equilibrium results. We start with the simplest example of two firms and then show results for a four-firm economy.

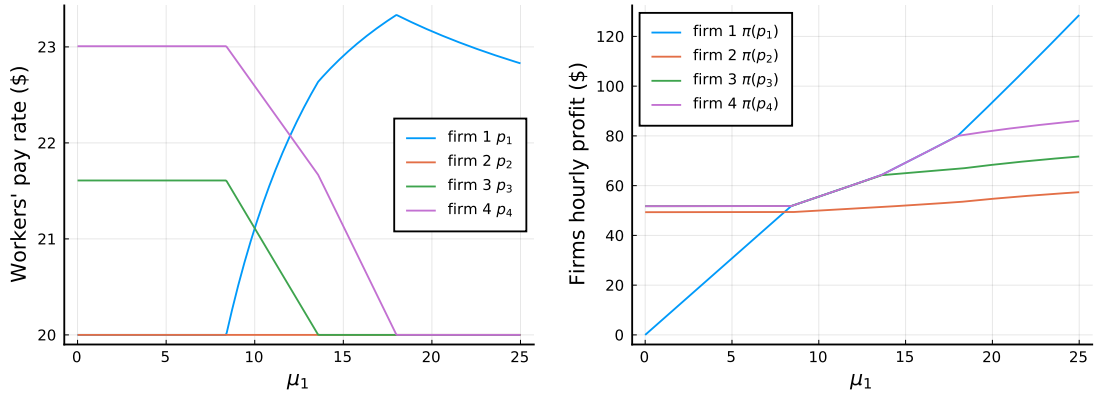


(a) Workers' pay rates p_1, p_2 and $P = (\mu_1 p_1 + \mu_2 p_2)/\mu$ at the Nash equilibrium.

(b) Firms' hourly profit $\pi(p_1)$, $\pi(p_2)$ and total profit $\pi(p_1) + \pi(p_2)$ at the Nash equilibrium.

Figure 4.2: Equilibria in a setting with two firms where the total job arrival frequency is constant.
Note: $\mu_1 + \mu_2 = 20$, $w_0 = 20\$/\text{hour}$ and $v = 30\$/\text{hour}$. We vary the relative size of the two firms, from $\mu_1/\mu = 0$ (firm 2 is a monopoly) to $\mu_1/\mu = 1$ (firm 1 is a monopoly)

Two firms Figure 4.2 provides data and outcomes for an example with two firms. Here, we keep μ constant, vary the relative size of each firm and then plot the equilibrium prices. Looking at the equilibrium prices in Figure 4.2a, we see that the smaller firm can have up to 38% of the total job arrival rate and still free ride (pay w_0). In the most extreme case, the smaller firms' profit margin can be \$10 per job while the largest firm is only \$6. The firms' hourly profits are represented in Figure 4.2b and illustrate results 4 and 5 of Theorem 7: if the smaller firm pays more than w_0, then the two firms make exactly the same hourly profit. It can also be seen from the total profit line of Figure 4.2b that



(a) Workers' pay rates p_1, p_2, p_3, p_4 at the Nash equilibrium.

(b) Firms' hourly profit $\pi(p_1), \pi(p_2), \pi(p_3), \pi(p_4)$ at the Nash equilibrium.

Figure 4.3: Equilibria in a setting with four firms when the first one is growing. We set $w_0 = 20\$/\text{hour}$ and $v = 30\$/\text{hour}$. We fix the job arrival rates $\mu_2 = 8, \mu_3 = 10, \mu_4 = 12$, and we vary μ_1 from 0 to 25.

the equilibrium has slightly lower total profit than a monopoly with the same total demand (the two endpoints of the graph) – so the smaller firm's free-riding makes the market slightly less profitable.

This duopoly example sheds light on the Uber/Lyft market share history presented in Figure 4.1b, since both companies roughly share the same demand and hence the assumption $\mu_1 + \mu_2 = \mu$ is reasonable. Our numerical example shows that as the smaller firm's (e.g., Lyft's) market share diminishes, its equilibrium labor costs declines. Such a market force helps counteract other positive effects of scale and hence serves to stabilize market share, consistent with the data in Figure 4.1b.

Four firms Figure 4.3 illustrates an example with four firms. Here we examine the impact of a growing firm (by varying μ_1) on the equilibrium prices and profit, keeping the other firms' constant ($\mu_2, \mu_3, \mu_4 = 8, 10, 12$). If we first look at Figure 4.3a, we see that when μ_1 is small, firms 1 and 2 pay w_0 while firms 3 and

4 have higher pay rates (and therefore the same hourly profit). As firm 1 grows, it starts to pay workers more, which allows firms 3 and 4 to pay them less, until they can also free ride and pay w_0.

Figure 4.3b shows the resulting profit outcomes. Note the profits of firms 2, 3 and 4 stay constant as long as firm 1 is small enough to pay w_0, but increase when firm 1 starts to pay above w_0 and begins contributing to increasing the size of the worker pool. This benefits all firms. (Every firm's profit is a non-decreasing function of μ_1.) When firm 1 grows and pays above w_0, it generates the same hourly profit as the firms that also pay above w_0. Interestingly, while firm 1's profit grows rapidly at first with its job rate μ_1, this growth is slowed down once it becomes big enough to allow the other firms to "free ride" and pay less. After a "phase transition", in which firm 1 is big enough to be the only firm that pays above the minimal rate and all other firms are free-riding, the growth of firm 1's profit recovers as the other firms cannot increase their marginal profit further.

4.4.4 A large firm is needed for market formation

Theorem 7 only describes the equilibria with $\lambda > 0$. We next study equilibria with $\lambda = 0$, i.e., where a stable gig economy cannot be created. This allows us to derive important insights on the formation of gig economy markets.

Theorem 8 *A Nash equilibrium with $\lambda = 0$ exists if and only if none of the firms have enough jobs to create a gig economy on their own:*

$$\mu_i \leq \frac{w_0}{T(v - w_0)} \qquad \forall 1 \leq i \leq N \tag{4.16}$$

Consequently, market formation (an equilibrium with $\lambda > 0$) is the unique stable

equilibrium if and only if the largest firm can sustain the market on its own, i.e., if its demand is above $\frac{w_0}{T(v-w_0)}$.

Theorem 8 has several implications. First, a sufficient condition for the creation of a gig economy is the existence of a large enough firm. To see this, compare (4.16) with (4.10): the right-hand sides correspond to the minimal job rate for a firm (or a group of firms) to be able to create a profitable gig economy. This explains the second part of Theorem 8: if one firm has enough jobs to create a profitable economy on its own, then there are no equilibria with $\lambda = 0$ and the only possible equilibrium is the one described in Theorem 7.

On the other hand, if none of the firm are big enough to "jump start" the gig economy, then (4.16) holds and there exists equilibria without market formation. Actually, there are many such equilibria. They correspond to the sets of pay rates $(p_i)_i$ such that $\lambda = 0$ at the worker equilibrium, and such that none of the firms can individually increase their pay rate so that $\lambda > 0$. Formally:

$$\frac{T}{\mu_{\geq w_0}}\left(\mu_i v + \sum_{j\neq i,\ p_j \geq w_0} \mu_j p_j\right) < w_0\left(T + \frac{1}{\mu_{\geq w_0}}\right) \qquad \forall 1 \leq i \leq N \qquad (4.17)$$

Eq. (4.17) implies that the only way for a market to be created in this situation is if the conditions of Theorem 7 apply (the combined jobs of all the firms are large enough to sustain a gig economy), and if enough firms coordinate to enter at the same time, e.g., they all enter the market simultaneously and settle on an equilibrium set of prices. Given that such coordinated-entry is unlikely, Theorems 7 and 8 indicate a potential hysteresis phenomenon: a gig economy needs to be created by a large firm that is willing to pay a high price to create a market, and once the market is created, many smaller firms will join it. However, once formed, the market can be sustained even if the larger firm exits provided

the total demand of the remaining small firms is large enough.

4.5 Generalization

The reformulation of the firms' profit maximization problem introduced in Section 4.4.2 allows us to generalize our model to a more realistic setting. By simply replacing v and T with v_i and T_i in (4.11) and (4.12), the model allows firms to have various revenue and job durations without changing the problem statement (4.15), (4.13). Furthermore, we find that as long as the expected waiting time $W(\lambda; \mu_{\geq w_0})$ satisfies certain regularity conditions, all of the main results in Section 4.4 continue to hold. The generalization is summarized in the following proposition:

Proposition 25 (sketch) *Consider a gig economy with N firms and firm-dependent revenue v_i and job duration t_i, $1 \leq i \leq N$. Suppose that the waiting function $\lambda \to W(\lambda; \mu_{\geq w_0})$ is strictly increasing, twice-differentiable and convex in λ on $[0, \mu_{\geq w_0})$; $W'(\lambda; \mu_{\geq w_0})$ is nonzero on $[0, b)$; $\lim_{\lambda \to \mu_{\geq w_0}} W(\lambda; \mu_{\geq w_0}) = +\infty$; and $W(0; \mu_{\geq w_0}) = 1/\mu_{\geq w_0}$.*

Then similar results to Proposition 24 and Theorems 7 and 8 hold. Theorem 10 of the Appendix present these results in detail.

Note that the regularity conditions on W are not very restrictive. The strict monotonicity simply means that we need more workers to have more accepted jobs, and the convexity implies that each new worker produced a smaller marginal increase in accepted jobs. $\lim_{\lambda \to \mu_{\geq w_0}} W(\lambda; \mu_{\geq w_0}) = +\infty$ means that we need an infinite number of workers to never miss a job. $W(0; \mu_{\geq w_0}) = 1/\mu_{\geq w_0}$

simply means that in the limit where we have almost no workers (light traffic limit), the average wait of a work is $1/\mu_{\geq w_0}$. This makes sense because the worker will almost certainly be alone and will take the first request that pays at least w_0, which corresponds to an average wait time $1/\mu_{\geq w_0}$. Many realistic worker arrival distributions and job assignment processes satisfy these conditions. For example, they can allow us to represent a setting where workers take several jobs in a row, that is, if they join the waiting queue immediately after finishing a job.

Proposition 25 confirms the robustness of our model. Indeed, at its core, the result that larger firms pay more does not rely on the specific assumption on job parameters or the precise waiting time function. The driving force is something more central to the gig economy: a pool of available workers waiting for jobs with firms jointly paying for the waiting time to maintain the pool. But since larger firms benefit more from the pool, and it is in their self interest to pay more to maintain it. This, in turn, creates an opportunity for smaller firms to free-ride and pay minimal rates to workers.

4.6 Conclusions

Our paper demonstrates the industrial organization consequences of a shared worker pool in a gig economy. We show that in equilibrium, firms' decisions critically depend on their size. Specifically, smaller firms can free ride on the existence of larger firms, relying on larger firms to pay the price of maintaining the shared labor pool. This results in small firms enjoying the lowest-possible labor costs while large firms end up earning the same profit regardless of size.

The resulting scale disadvantage is stark and helps explain why gig industries seem to be highly contestable and resistant to market concentration. At the same time, we show the existence of a large firm is necessary to form a gig economy in the first place; baring simultaneous entry by many small firms, a firm large enough to support a gig economy on its own is necessary for such an economy to form. This helps explain why it may have taken the enormous funding and persistence of early entrants like Uber to trigger the formation of gig economies throughout the world. It also suggests the regulatory efforts to limit market concentration in the gig economy must be approached with caution; such efforts could block the formation of gig economies.

Still, our analysis is limited. It only considers a case where firms do not compete in their product markets and firms' sizes do not change over time. An obvious and valuable extension would be to consider how product-market competition affects our results. Another worthwhile extension would be to understand the dynamics of competition if smaller firms can invest their profits to grow over time. Lastly, a rigorous empirical investigation of the many industry formation and structure predictions of our theory would be welcome.

CHAPTER 5

CONSPICUOUS CONSUMPTION IN THE PRESENCE OF NON-DECEPTIVE COUNTERFEITS

5.1 Introduction

> *In order to gain and to hold the esteem of men, it is not sufficient merely to possess wealth or power. The wealth or power must be put in evidence, for esteem is awarded only on evidence.* —Thorstein Veblen, *The Theory of the Leisure Class*, 1899

Consumers often have a desire to display their consumption taste and purchase power. American social economist Thorstein Veblen attributed such desire as a way to assure one's position in society. Since one's social status, such as their wealth, is private information that others cannot observe directly, wealthy consumers can signal their status to others through observable actions such as displaying their luxury purchases. This behavior is commonly known as *status signaling* (Ireland 1994).

The digital era has made status signaling unprecedentedly easy, creating immense opportunities for the luxury industry. Social media platforms like Instagram and TikTok have made it as simple as a touch/click for people to display their luxury purchases to hundreds, if not thousands or millions, of followers. Being able to signal one's status to a large audience thus significantly amplifies the utility that people can gain from such activities, stimulating the sales for brand-name status goods. In the United States, it has been reported that more than 90% of the consumer engagement with luxury brands is through

Instagram, whose primary function is to share photos among one's friend circle (Guan 2018). In China, a country that contributes to more than half of the global growth of luxury sales, e-commerce, and social media platforms have worked closely to make it easy for consumers to purchase and post (Li 2020).

However, the new era has also created a new challenge to the luxury (status product) industry: *non-deceptive* counterfeits are more popular than ever. Non-deceptive counterfeits refer to those purchased by consumers on purpose. Platforms such as Amazon and Alibaba have made it easy for consumers to search and purchase cheap knockoffs, thus enabling counterfeit sellers to reach numerous consumers at almost no cost (which would be otherwise unthinkable in the traditional retail channels). As a result, counterfeit products have become even more obtainable for consumers than before. When consumers knowingly purchase counterfeits, it poses a serious threat to the luxury industry; since consumers do not have the incentive to disclose the purchase, it is more challenging to detect counterfeits. Therefore, understanding the rationales behind consumers' decisions in purchasing counterfeits is a timely and pressing issue.

This chapter investigates how consumers signal their status through the purchase of status goods and their counterfeits. Furthurmore, we seek to understand the consequence on the profit and pricing decisions of the firms that produce authentic status goods. Thus, we are interested in the interactions between status signaling, counterfeits, and authentic status goods, at both the consumer and the market level. For example, what are the main driving factors behind the consumption decisions? How do counterfeits impact consumers' demand for the authentic status goods? How do counterfeits impact the sales and the demand for the authentic goods, and what strategies can be taken to mitigate the damage?

As an initial attempt to investigate these questions, we consider an incumbent firm (the firm) who produces status goods, and a counterfeiter (the counterfeiter) who is able to produce high-quality copycats of the status goods and attempts to enter the market. Consumer differ in their wealth levels (high or low), which are private information that they intend to signal through the consumption of the status goods and counterfeits.

A unique feature of our model is that the market demands for the firm and the counterfeiter are endogenously determined by the equilibrium of a consumer status signaling subgame. In particular, we introduce a *spectator*, whose opinion the consumer cares about. The spectator can observe a consumer's total consumption of the status and counterfeit goods; with a certain chance, they can differentiate between the two. The spectator thus makes an inference of the consumer's type through the consumption bundle. Consumers are better off when being considered as of the high type; they are worse off when the spectator finds out they have purchased counterfeits. Therefore, for the consumer, there is a trade-off in counterfeit consumption between the savings in signaling costs and the penalty of being exposed.

This idea of the spectator is built on the seminal work by Ireland 1994, who first proposes a signaling game model to explain the status product consumption and shows that a lower-type consumers have the incentive to distort their consumption upward to mimic higher-type consumers. In Ireland 1994, in equilibrium, consumers of all types always end up revealing their true wealth, which is a standard result in the signaling game. Introducing counterfeits to the consumer's choice set greatly complicates the problem – because the spectator cannot easily differentiate between the status and the counterfeit goods, consumers have

an additional lever to signal their wealth and do not always reveal their true status in the equilibrium. How consumers employ this fake status signal,and how does it impacts the consumer's dmemand for the status goods, are also one of the focus of this chapter.

Our contribution is two-fold. First, we enriched the econ literature on status signaling. In most standard two-type signaling models, high type senders always end up selecting the exact signal that separate themselves from low type senders. Since the signal has to be costly to be convincing, the result usually leads to a social welfare loss. In our work, the presence of counterfeits breaks the monotonic relation between types and signaling power. Due to the huge expected penalty costs that a high-type consumer incurs in purchasing counterfeits, under some conditions, low-type consumers may be the only type who has the privilege of using counterfeits to signal status. This non-monotonic relation leads to a pooling equilibrium that survives common refinements such as Intuitive Criterion, which is not often seen in most status signaling works. Furthurmore, in the pooling equilibrium, the high-type consumer no longer attempts to signal status through excessive consumption, which may indeed be beneficial from a social welfare standpoint.

Our work also builds a theoretical foundation for future studies on the operational issues raised by non-deceptive counterfeits. We built a consumer model for conspicuous consumption in the presence of counterfeits, which is to the best of our knowledge the first of such model. In our model, the status and counterfeit goods are not merely competing with each other; they are also competing with necessity products. In other words, the price changes of authentic and counterfeit goods influence not only the ratio of demands, but also the sum of the demands

for these two products. Therefore, this consumer model generates richer insights than those that only consider the interplay between the status and the counterfeit goods.

This chapter is organized as follows. Section 5.2 is a review of the relevant literature in the field of operations and economics. Section 5.3 contains our main analysis. Section 5.4 summaries the main findings.[1]

5.2 Literature Review

There is a growing literature in the field of operations and marketing on counterfeiting, most of which focus on deceptive counterfeits. Qian 2014 builds a theoretic model to analyze brand-protection strategies for counterfeiting mitigation in an economy with weak intellectual property rights. This paper mostly focuses on deceptive counterfeits, although the author briefly discusses a scenario in which buyers knowingly purchase counterfeits and fool others through counterfeit status signaling as a benchmark to the rest analysis. Qian, Gong, and Chen 2015 build a vertical differentiation model to analyze the market equilibria under the competition between brand-name companies and counterfeiters, showing that authentic brands can combat deceptive counterfeits by improving the product quality. Pun, Swaminathan, and Hou 2021 study how blockchain technology can be used for combating deceptive counterfeits. The central topic in these papers is the information asymmetry between consumers and counter-

[1]This chapter is based on an early draft of Chen, Lian, and Yao 2021. The analysis mainly focuses on the consumer (the signaling game and the resulting equilibrium demand). The firm's optimal price and profit are briefly discussed as a model extension in Section 5.3.5; for an extensive analysis of the firm's decision, please refer to the latest version of Chen, Lian, and Yao 2021.

feiters, while the information asymmetry in this chapter lies between consumers and the spectators around them.

In the literature concerning non-deceptive counterfeits, Pun and DeYong 2017 study the competition between a manufacturer and a counterfeiter over strategic consumers through a two-period game-theoretic model and shows that the timing of consumers' purchase plays an important role in determining the manufacturer's optimal strategy and social welfare. Cho, Fang, and Tayur 2015 analyze the anti-counterfeiting strategies for a brand-name company and shows that the effectiveness of the strategies critically depends on whether counterfeiters are deceptive or non-deceptive. Zhang, Hong, and Zhang 2012 derive and compare different strategies for brand-name products to fight non-deceptive counterfeits through a vertical differentiation model. Yi, Yu, and Cheung 2020 study the effect of non-deceptive counterfeits on the global supply chain and analyze the anti-counterfeit strategies. Although these works focus on non-deceptive counterfeits, they do not consider the status effect of the brand-name product or consumers' status signaling behavior, which are key to our analysis.

Gao, Lim, and Tang 2017 is perhaps the most closely related to this chapter in the literature, as it studies how brand-name companies combat non-deceptive counterfeits when consumers are status seeking. This chapter differs from Gao, Lim, and Tang 2017 in several significant ways. First, this chapter explicitly takes into account the counterfeiter market entry cost and wealth inequality. Our discussion is centered around how these two parameters would affect the status product firm's price and profit as well as social welfare in general. Second, we model consumer status-seeking behavior by a consumer status signaling game

in which the consumer status utility is derived from the spectator's status belief. As a result, different consumer types have different status valuations, depending on their willingness-to-pay. By contrast, the model considered by Gao, Lim, and Tang 2017 emphasizes more on the product side (such as the resemblance and quality of the counterfeit) and treats the consumer status utility as the average social status of a group of consumers who make the same purchase/no purchase decision. This modeling difference enables us to derive new insights about the impact of potential counterfeiter entry on the equilibrium market outcomes.

This chapter is also closely related to the literature on conspicuous consumption. Amaldoss and Jain 2005 study how the desire for exclusivity and conformity drives consumers purchasing decision for conspicuous products and propose a rational expectation mechanism for consumers' behavior. Heffetz 2011 empirically tests the relationship between the visibility of a product and the income elasticity and show that consumers purchase a product not just for the intrinsic value but also for the signaling value. Tereyağoğlu and Veeraraghavan 2012 analyze the production strategy for a firm that sells conspicuous products and derive conditions when scarcity strategies are optimal. These papers focus on conspicuous consumption through genuine status goods and do not consider counterfeits. This chapter complements this literature by further allowing consumers to purchase non-deceptive counterfeits to send a fake status signal.

Another related stream of literature is the study of status signaling in the field of economics. Frank 1985 first associates status signaling with the idea of positional goods and non-positional goods, arguing that interpersonal comparison among consumers can affect their purchasing patterns on positional goods more profoundly than non-positional goods. Ireland 1994 builds on this idea and

precisely model the interpersonal comparison with a game-theoretic approach. In particular, the model assumes that consumers have private information about their status level and are better off when others (spectators) view them to be wealthier than their true status type. Consumers and spectators are then modeled as two players of a signaling game. This chapter employs a similar consumer status signaling model when deriving the demand for the status product and the counterfeit, but goes a step further to embed this signaling game in a market entry deterrence game between the status product firm and the counterfeiter who attempts to enter the market.

Finally, in the field of operations management, the signaling game is also commonly employed for modeling firms' operational decisions under information asymmetry see, e.g., Lai and Xiao 2018; Zhao, Lai, and Xiao 2019; Chakraborty and Swinney 2021. This chapter contributes to this literature by introducing a market demand model that is endogenously determined by the equilibrium of a consumer status signaling game.

5.3 Analysis

5.3.1 Model Formulation

Products Suppose there are three goods, denoted X, Y and Z, which are consumed at level x, y and z. X refers to the authentic status good (the status good), the price of which is assumed to be p and $p > 1$. Y refers to a copycat of the status good X (the counterfeit), with a lower price $\delta p, 0 < \delta < 1$. Both X and Y are what we call the *visible* good, as their consumption can be easily observed

and recognized by others. Z refers to the necessity products (such as toilet paper, bread, pencil, etc.), the price of which is assumed to be 1. Since the necessity products are typically not for display or show-off, the consumption of Z is assumed to be invisible to spectators. Moreover, the total consumption level of X and Y, which is $x + y$, is visible, while the level of each type of consumption is not.

The spectator By observing the consumer's consumption level of X and Y, the spectator will draw a conclusion about the consumer's status. We define status as the relative wealth level, i.e. the difference between the consumer's wealth and the lowest wealth in the social group (Rao and Schaefer 2013). Since the consumer's wealth w is private information, the status perceived by the spectator is a function in x and y that reflects the spectator's belief in the consumer's true wealth. We denote this perceived status as $s(x, y)$.

The consumer We assume that consumers' utility function is a linear combination of their private utility and the perceived status, which is given by

$$f(x, y, z) + \lambda s(x, y)$$

Parameter λ thus reflects how important the spectator's conclusion is to the consumer, which we call the *status factor*. The private utility $f(x, y, z)$ can be considered as the utility when the status s is endogenously given, or in other words, the intrinsic utility from X, Y and Z.[2]

[2]We did not give an explicit expression for the spectator's utility. It can be assumed that for each observation, the spectator minimizes the distance between the perceived status and the conditional expected status of the consumer

We further assume function f to have the following quasi-linear form:

$$f(x, y, z) = x + y + u(z) - \mathbb{1}_{\{y>0\}}\varphi k,$$

where u is strictly concave and strictly increasing in z, φ is the probability of being caught for purchasing counterfeits, and k is the associated disutility to the consumer when the counterfeit purchase is exposed. We call φ the *exposure risk* and k the *exposure cost*.

The above formulation of private utility has several implications. First, f is assumed to be quasi-linear. This form has the property that for consumers under certain income level, it is optimal to purchase only the necessity good Z; as the income grows, all additional incomes are spent on status goods X and Y. This captures the real-life consumption behavior in the sense that the spending on necessities does not vary much among different income groups, while the spending on status goods is subject to income change. This reflects the fact that people only purchase luxury goods after sufficient necessities have been acquire, and the spending on necessities doesn't vary much for people of different income level.

Second, the utility from one unit of counterfeit Y is less than the utility from one unit of status good X by a constant, φk. As has been mentioned, this is the risk for purchasing a counterfeit. It can also be viewed as a reflection of the inferior quality of counterfeits compared with the status good.

In the consumer's decision-making process, a type i consumer solves the following utility maximization problem, subject to a budget constraint:

$$\max_{x,y,z} x + y + u(z) - \mathbb{1}_{\{y>0\}}\varphi k + \lambda(1 - \mathbb{1}_{\{y>0\}}\varphi)s(x, y)$$

$$\text{s.t. } px + \delta py + z \leq w_i,$$

Since it is less interesting when neither type purchases X, we make the following assumption:

Assumption 8 *Throughout the equilibrium analysis, we assume that both types will purchase some status goods, i.e.*

$$p < \frac{1}{u'(w_L)}$$

5.3.2 Game Description

There are two possible types for the consumer, a high type with wealth $w = w_L$ and a low type with $w = w_L$. The spectator has a prior belief that there is a probability of π for the consumer being high type and a probability of $1 - \pi$ for the consumer being low type.

The consumer moves first and chooses a consumption level in (x, y, z). Then the spectator observes (x, y), updates the belief in the consumer's true type and responds with $s(x, y)$. At the equilibrium, the consumer's overall utility is maximized and the spectator's belief aligns with the Bayesian rule. Note that we only consider pure strategy because of the strict concavity of the utility function.

5.3.3 Market with no counterfeit risk

We start with a benchmark case when there are only status goods in the market. When status goods are the only type of visible goods in the market, the spectator infers the consumer's status solely from her consumption level in status goods, x. For a consumer with type $i \in \{H, L\}$, the utility maximization problem is given

by

$$max_x x + u(w_i - px) + \lambda s(x)$$

Note that since the utility is strictly increasing in x and z, the budget constraint is always binding. Therefore, we replace z by $w_i - px$. When the consumer is considered as a high type, $s(x) = \Delta$; otherwise, $s(x) = 0$.

This follows the typical signaling game situation where it is more costly for the low type than for the high type in investing in status signal, i.e. the status goods. Hence, we look for a separating equilibrium that survives the Intuitive Criterion by Cho and Sobel 1990.

In the absence of the status effect

Proposition 26 *Without the status effect ($\lambda = 0$), there is a unique utility maximizer for a consumer with type $i \in \{H, L\}$, given by*

$$x_i^0 = \max\{0, \frac{w_i}{p} - \frac{1}{p} z^0(p)\} \tag{5.1}$$

where $z^0(p) = (u')^{-1}(\frac{1}{p})$. Moreover, x_i^0 is non-decreasing in the wealth level w_i.

Function $z^0(p)$ can be interpreted as the cross demand function of necessity goods over the price of status goods when there is no status effect. It can be verified that, $z^0(p)$ is increasing in p (by applying the inverse function theorem). This implies that when there is no status effect, the consumer's expenditure on luxury products is strictly decreasing in p. The reason is that status goods and necessities are substitutes; increasing status good price will shift consumers' demand towards necessities.

Proposition 26 implies that when there is no status effect, the high-type consumer always purchases more status goods than the low type does, except when both of them purchase zero status good; that is, $x_H^0 > x_L^0$ whenever $x_H > 0$. Proposition 26 provides a benchmark for measuring the consumer's demand distortion in the presence of the status effect, and functions x_L^0, x_H^0 will appear again in the following sections.

With the status effect

In the presence of the status effect, there is a utility boost for being considered as a high type. This provides an incentive for the low type to deviate from x_L^0, if this extra status utility can offset the disutility from over-consuming status goods. Aware of the possibility of being mixed up with a low-type consumer, the high-type consumer may have to raise her consumption to deter the low type. The spectator adjusts the belief on the consumer's type upon seeing different consumption level of X. There exists a unique perfect Bayesian equilibrium. We denote the equilibrium output for the high type as x_H^* and the low type as x_L^*. Denote the wealth difference, $w_H - w_L$, as Δ.

Proposition 27 *Given a status factor $\lambda > 0$, there always exists a unique separating equillibrium. Denote the equilibrium consumption level for the low-type and high-type consumer as x_L^* and x_H^*, respectively. Then there are two cases:*

1. *when the status factor λ or the price of the status good is below a threshold, a consumer of any type purchases the same quantity as when there is no status effect. That is, $x_L^* = x_L^0$, and $x_H^* = x_H^0$;*

Table 5.1: equilibrium consumption level functions in a market with no counterfeit risk

	$\lambda \leq \frac{1}{\Delta}\Big\{u(z^0(p)) - u(z^0(p)-\Delta)\Big\} - \frac{1}{p}$ or $z^0(p) < \Delta$	$\frac{1}{\Delta}\Big\{u(z^0(p)) - u(z^0(p)-\Delta)\Big\} - \frac{1}{p} < \lambda \leq \frac{1}{\Delta}\Big\{u(z^0(p)) - \frac{z^0(p)}{p}\Big\}$ and $z^0(p) \geq \Delta$	$\lambda > \frac{1}{\Delta}\Big\{u(z^0(p)) - \frac{z^0(p)}{p}\Big\}$
x_L^*	x_L^0	x_L^0	x_L^0
x_H^*	x_H^0	$\bar{x}_L$	$\frac{w_L}{p}$

Note. The function $\bar{x}_L$ is defined uniquely as the solution to the following equation:

$$\bar{x}_L + u(w_L - \bar{x}_L p) = \frac{w_L}{p} - \frac{1}{p}z^0(p) + u(z^0(p)) - \lambda\Delta$$

where $\bar{x}_L > \frac{w_L}{p} - \frac{1}{p}z^0(p)$. Moreover, both the function $\bar{x}_L$ and w_L/p is strictly above x_H^0. $\bar{x}_L$ represents the maximum quantity of the status good X that a low-type consumer is willing to purchase to be considered as a high-type consumer.

2. *when the status factor λ and the price of the status good are above the thresholds, a low-type consumer purchases the same quantity as when there is no status effect, while a high-type consumer purchases more than when there is no status effect. That is, $x_L^* = x_L^0$, and $x_H^* > x_H^0$.*

The expressions of the conditions and the equilibrium consumption level are given by Table 5.1.

The condition $z^0(p) < \Delta$ corresponds to the case that $x_H^0 > w_L/p$, which means even if the low type spends all her budget on the status good, she is not able to match the high type's optimal consumption. This happens when the wealth difference between the two types is so great that mimicking the high type is not an available option.

When the status factor λ is sufficiently large, the status utility for being considered as a high type is so high that the low type would like to spend all

her wealth w_L to purchase status goods, assuming she will be rewarded the high type's status. In this case, the high-type consumer can choose a consumption level that is slightly higher than w_L/p and separates from the low type. Being aware of this outcome, the low type will choose to stay at her private utility maximizer x_L^0, and the high-type consumer will consume w_L/p and enjoy a high status.

From proposition 27, in the presence of status effect, the low-type consumers stick to their private utility maximizing behavior . In other words, although they have an incentive to mimic the high type, they fail to do so and end up revealing their true type to the spectator. The reason is that high-type consumers are always able to choose a consumption level so high that the benefit from status seeking is offset by the utility loss for low-type consumers.

For high-type consumers, the optimal consumption level is driven by two factors, the income difference and the status effect. If the income difference Δ is sufficiently high, high-type consumers' private utility maximizer x_H^0 already makes it hard enough for low-type consumers to mimic. If the status effect λ is small, low-type consumers don't care much about the spectator's opinion; then again x_H^0 already induces enough utility loss for low type to give up mimicking. Therefore, in both cases, $x_H^* = x_H^0$. It is only when the income difference is moderate and consumers have a strong desire for status that low-type consumers become a real threat. In this case, high-type consumers have to distort their consumption to a level higher than x_H^0, otherwise low-type consumers will actually obtain more utility from mimicking. As is given in proposition 27, $\bar{x}_L$ is the consumption level with the least distortion that prevents low types from mimicking, and thus maximizes high-type consumers' overall utility. Moreover,

given p,

$\bar{x}_L$ is increasing as the status effect grows. Therefore, when the income difference is below a threshold and the status effect is above a threshold, status seeking boosts the demand for the status good.

5.3.4 Market with the counterfeit good

There exists a trade-off in purchasing the counterfeit good: the cost discount δ and the exposure cost k. When the k is high, even the low-type consumer won't purchase the counterfeit good, because the disutility from being exposed dominates utility boost the consumer obtains from mimicking the high type; when k is low, even the high-type consumer may purchase the counterfeit good, because the exposure cost is so low that the low price of counterfeits dominates the risk of getting caught. The most interesting case is when k is intermediate, in which case the high type will purchases the status product, and the low type will purchase the counterfeit product, of the same quantity. Such a purchasing structure has several important implications. For example, from the firm's perspective, by manipulating the exposure cost factor k, the consumer's purchase choice can vary significantly.

Lemma 5 helps understand how the consumption bundle looks like when both the status and counterfeit products are available:

Lemma 5 *If the consumer purchases some counterfeit product, then she won't purchase any status good.*

This implies that all mixed consumption bundles with $x > 0, y > 0$ are

suboptimal compared to only purchasing counterfeits. Therefore, we only need to compare the maximal utility from purchasing only X and that from purchasing only Y when looking for optimal consumption level.

In our benchmark case, we have discussed that equilibrium outputs are influenced by the income difference and status effects. In the presence of counterfeits, low-type consumers now have more ability for status seeking. The price discount δ amplifies low-type consumers' purchasing power.

Next, we present the analysis for two sets of consumption bundles: (1) both types of consumers purchase the status good X; in this case, we denote the equilibrium consumption level as $x_L^\dagger$ and $x_H^\dagger$ for the low-type and high-type consumer, respectively; (2) the high-type consumer purchases the status good X and the low-type consumer purchases the counterfeit good Y, in this case we denote the equilibrium consumption level as $y_L^\dagger$ and $x_H^\dagger$ for the low-type and high-type consumer, respectively. (3) both types of consumers purchase the counterfeit good Y. This case is similar to (1) except that the status good is replaced by the counterfeit good. Thus, we focus our discussion on (1) and (2).

Moreover, we introduce the following notation, which will provides insights into the consumer's equilibrium consumption. In Section 5.3.3, we have defined a threshold quantity $\bar{x}_L$ for the status good, above which a low-type consumer is not better off even if being considered as a high-type consumer. Here we extend the definition to include both the status and the counterfeit good, for both high-type and low-type consumers.

Define $\bar{v}_{i,j}$, $i \in \{L, H\}$, $j \in \{X, Y\}$. $\bar{v}_{i,j}$ represents the maximum quantity of product j that a type i consumer is willing to purchase to be considered as a

high-type consumer. More precisely, the formulation for the status good X is given by:

$$\bar{v}_{i,X} + u(w_i - \bar{v}_{i,X} \cdot p) = \frac{w_i}{p} - \frac{1}{p} z^0(p) + u(z^0(p)) - \lambda\Delta,\ i \in \{L, H\} \tag{5.2}$$

The formulation for the counterfeit good Y is given by:

$$\bar{v}_{i,Y} + u(w_i - \bar{v}_{i,Y} \cdot \delta p) = \frac{w_i}{\delta p} - \frac{1}{\delta p} z^0(\delta p) + u(z^0(\delta p)) - (1-\varphi)\lambda\Delta,\ i \in \{L, H\} \tag{5.3}$$

It is easy to verify that $\bar{v}_{L,X}$ is identical to $\bar{x}_L$.

Next, we proceed to present our analysis for the equilibrium.

Equilibrium: both types purchase the status good X

We start with the setting where both the low-type and the high-type consumer purchases the status good X. This setting happens when the exposure cost or risk are high, that even the low-type consumer does not prefer to purchase counterfeits. Proposition 28 summarizes the consumer's equilibrium consumption level:

Proposition 28 *Consider an equilibrium with both the high-type and low-type consumers purchasing the status good X. Then the equilibrium is identical to that in a market with no counterfeit risk. That is, $x_L^\dagger = x_L^*$, and $x_H^\dagger = x_H^*$.*

In other words, when neither type of consumers purchase counterfeits, the equilibrium is identical to that in Proposition 27.

Equilibrium: the low-type consumer purchases the counterfeit good Y and the high-type consumer purchases the status good X

Next, we analyze the setting where the low-type consumer switches to the counterfeit good. This happens when the exposure cost or risk is moderately high, such that the high-type consumer does not purchase any counterfeit out of the fear of losing status.

Furthurmore, even for the low-type consumer, an increase in the exposure risk φ will still discourage them from purchasing the counterfeit good. This is due to two reasons. The direct reason is that, the increased risk increases the chance of the consumption being exposed, which leads to a higher expected cost; the less obvious reason is that, it also reduces the consumer's opportunity of being considered as a high type. Thus, if we combine the status factor λ and the exposure risk φ together, it can be viewed as the low-type consumer's discounted sensitivity to status when using the counterfeit good to signal status.

In Proposition 29, we introduce the three types of equilibria that may happen when the low-type and high-type consumers are purchasing different kinds of visible goods for status signaling.

Proposition 29 (Sketch) *Given the status factor λ and the price of the status good p, there exists equilibrium where the low-type consumer purchases the counterfeit, and the high-type consumer purchases the status good. Moreover, the equilibrium consumption depends on the how cheap the counterfeit is in relative to the status good (i.e. the price discount factor δ):*

1. *when δ is low, there is a separating equilibirium, in which the consumption level of*

a low-type consumer is higher than that of a high-type consumer. That is, $x_H^\dagger < y_L^\dagger$.

2. *when* δ *is moderate, there is a pooling equilibrium, in which the consumption level of a low-type consumer is the the same as that of a high-type consumer. That is,* $x_H^\dagger = y_L^\dagger$.
3. *when* δ *is high, there is a separating equilibrium, in which the consumption level of a low-type consumer is lower than that of a high-type consumer. That is,* $x_H^\dagger > y_L^\dagger$.

The expressions of the conditions and the equilibrium consumption level can be found in Table 5.2.

Table 5.2: Conditions and equilibrium consumption level when the high-type consumer purchases X and the low-type consumer purchases Y

Condition	Equilibrium type	Low-type	High-type
Low δ	separating	y_L^0	x_H^0
Intermediate δ	pooling	x_H^0	x_H^0
High δ	separating	y_L^0	$\max\{\bar{v}_{L,Y}, x_H^0\}$

Proposition 29 reveals several interesting insights. Case 1 points out an equilibrium where the low-type consumer purchases a higher quantity than the high-type consumer, due to the low price of the counterfeit good. This result is quite striking in the sense that, when the counterfeit good is extremely cheap, a low-type consumer will intentionally separate from the high-type consumer and purchases a higher quantity, even though they could have settle for purchasing the same quantity as the high-type consumer and obtain the status effect from the pooling outcome. This setting will be discussed more extensively in the section "special case".

Case 2 is an equilibrium when the counterfeit price is moderate, and the low-type consumer settles for a pooling equilibrium with the high-type consumer.

Note that even though the spectator can not always tell the counterfeit from the status good from observation, she can work out the low type's utility and be aware when the low type purchases counterfeits. Therefore, upon observing the pooling consumption level, she knows that the products are either (1) the status good, purchased by a high-type consumer or (2) the counterfeit good that is not discovered with probability $1 - \varphi$, purchased by a low-type consumer. Therefore, the spectator will update her belief of the high-type consumer to be $P(H) = \frac{\pi}{\pi+(1-\pi)(1-\varphi)}$, upon observing the equilibrium consumption level.

Case 3 is an equilibrium when the counterfeit price is high, and the low-type consumer cannot mimick the high-type consumer, even with the price advantage from the counterfeit consumption. In this case, even when purchasing the counterfeit good, the low-type consumer is worse off by elevating her consumption to the high-type equilibrium consumption level. Thus, the low-type consumer gives up mimicking and just chooses the consumption level that maximizes her private utility. In contrast, the high-type consumer purchases exactly the quantity that can deter the low-type consumer from mimicking (unless that quantity is below the high-type consumer's private utility maximizer x_H^0, in which case the high-type consumer will just purchase x_H^0).

As is discussed, among the three cases, Case 1 is an unusual case where the low-type consumer is more capable of status signaling (i.e. can afford to purchase more visible goods than the high-type consumer). In the next a few paragraphs, we discuss a special case of such equilibrium.

Special case: a low counterfeit price and zero exposure risk (low δ, $\varphi = 0$) In an extreme case where the spectator cannot not differentiate between the status

and the counterfeit good at all (i.e. $\varphi = 0$), the spectator will make an inference on the consumer's type solely based on the consumption level. Interestingly, when the price discount factor δ is sufficiently low, the low-type consumer will be able to send a status signal even stronger than the high-type consumer, as is characterized by Case 1 in Proposition 29.

However, this does not mean that the low-type consumer can fool the spectator and achieve a "reverse" signaling equilibrium, in which the low-type consumer is mistakenly considered as high type and high type is mistakenly considered as low-type consumer.

To see why this is the case, we apply the *elimination of type-message pairs by dominance* (Cho and Kreps 1987), and show that such a reverse signaling equilibrium does not survive the intuitive criteria. For a message $v > w_H/p - z^0(p)/p$, the high type may be eliminated for this message if

$$U_H(w_H/p - z^0(p)/p, 0) > U_H(v, \Delta).$$

where $U_H(v, s) = v + u(w_H - pv) + \lambda s$. This condition says that, even if recognized as a high type, the high-type consumer is still worse off by consuming v of status goods, compared with consuming $w_H/p - z^0(p)/p)$ and recognized as a low type. By some mathematical manipulation, this condition is equivalent to $v > \gamma_H$, with

$$\gamma_H + u(w_H - p\gamma_H) + \lambda\Delta = \frac{w_H}{p} - \frac{1}{p}z^0(p) + u(z^0(p)), \gamma_H > \frac{w_H}{p} - \frac{1}{p}z^0(p)$$

Therefore, any message $v > \gamma_H$ can not be sent by a high-type consumer. It can be easily checked that any $v < w_H/p - z^0(p)/p)$ is dominated for a high-type consumer, too. Hence, if a low-type consumer want to obtain $s = \Delta$ from the spectator, she has to choose $v \in [w_H/p - z^0(p)/p), \gamma_H]$; arguably, with some

regularity conditions, v is not dominated by $w_L/p - z^0(p)/p$ for the low type, otherwise the low type would not be willing to purchase v. Consequently, from the spectator's point of view, the belief she holds upon observing v should be $\mu(H|v) < 1$, as both types are possible to consume this level of visible products.

In other words, with the signal power from consuming perfect the counterfeit good, the best a low-type consumer can do is to mimick the high-type consumer; there is no way she can be considered as a high type with probability 1, even if purchasing the counterfeit good will not be discovered.

5.3.5 Extension: The firm's profit maximization

Next, we explore the relationship between the status effect and the firm's pricing decision. To achieve this, we use a concrete function form that allows us to compute the consumer's equilibrium consumption level in closed-forms.

More precisely, let $u(z) = 2\sqrt{z}$. That is, the utility a consumer generates from the necessity good Z is a concave function in the necessity consumption level, and the consumption has a decreasing marginal value.[3]

By proposition 27, the equilibrium consumption level of a low-type consumer for the status good is given by:

$$x_L^* = \frac{w_L}{p} - p \tag{5.4}$$

The structure of the equilibrium consumption level of a high-type consumer for the status good depends on the price p.

[3]Consequently, $u'(z) = \frac{1}{z^{1/2}}$, $u''(z) = -\frac{1}{2z^{3/2}}$, and $z^0(p) = p^2$.

For $p < \sqrt{\Delta}$,

$$x_H^* = \frac{w_H}{p} - p \tag{5.5}$$

For $p \geq \sqrt{\Delta}$,

$$x_H^* = \begin{cases} \frac{w_H}{p} - p, & \text{if } \lambda \leq 2p - 2\sqrt{p^2 - \Delta} - \frac{\Delta}{p} \\ \frac{w_L}{p} - (\sqrt{p} - \sqrt{\lambda})^2, & \text{if } 2p - 2\sqrt{p^2 - \Delta} - \frac{\Delta}{p} < \lambda \leq p \text{ and } p \geq \sqrt{\Delta} \\ \frac{w_L}{p}, & \text{if } \lambda > p \text{ and } p \geq \sqrt{\Delta} \end{cases} \tag{5.6}$$

In the rest of the chapter, we focus on the setting where the production cost of the status good, is sufficiently low that the firm will choose a price point where both types of consumers purchase some status good when there is no counterfeit good. In other words, we focus on the case where $p,\ c \leq \sqrt{w_L}$.

Similar to the main model, we start by examining the setting with no status effect.

In the absence of the status effect

When there is no status effect ($\lambda = 0$), the demand function is just x_H^0 for the high-type consumer and x_L^0 for the low-type consumer. Therefore, the firm's profit function is given by

$$\Pi(p) = \pi x_H^0 + (1 - \pi) x_L^0 \tag{5.7}$$

$$= (\frac{\bar{w}}{p} - p)(p - c) \tag{5.8}$$

where $1 < p \leq \sqrt{w_L}$; $\bar{w}$ is the average wealth of the consumer, the expression of which is given by $\bar{w} = \pi w_H + (1 - \pi) w_L$; c is the unit cost of the status good.

It is easy to check that the profit function $\Pi(p)$ in Eq. (5.7) is strictly concave in p. Therefore, the optimal price p^* can be solved by setting the first-order derivative of $\Pi(p)$ zero. The expression of the optimal price is thus given by

$$p^{*2}(\frac{2p^*}{c} - 1) = w_L + \pi\Delta, \text{ if } \pi\Delta \geq 2w_L(\frac{\sqrt{w_L}}{c} - 1) \tag{5.9}$$

$$p^* = \sqrt{w_L}, \text{ if } \pi\Delta < 2w_L(\frac{\sqrt{w_L}}{c} - 1) \tag{5.10}$$

Next, we proceed to analyze the case when there is the status effect.

With the status effect

As one may observe, the optimal price function given by Eq. (5.9) is already not well-behaved even when $\lambda = 0$. Thus, for tractability, we impose the following conditions:

First, we assume that the status effect is sufficiently strong; that is,

$$\lambda > \max\left\{\frac{1}{c}, \frac{1}{(1-\pi)\Delta}, \frac{1}{\sqrt{\Delta}}\right\} \tag{5.11}$$

Moreover, we are interested in the case where the low-type consumer still has some affordability to the status product:

$$w_L > \lambda^2\Delta^2 \tag{5.12}$$

Under the assumptions, the low-type consumer's demand function remains the same as Proposition 27; the high-type consumer's demand function can be simplified as below:

$$x_H^*(p) = \begin{cases} 0, & \text{if } p > \left(\sqrt{\lambda\Delta} + \sqrt{2\sqrt{w_L} + \lambda\Delta}\right)^2 \\ \frac{w_L}{p} - \frac{p}{4} - \lambda\Delta + \sqrt{\lambda p\Delta}, & \text{if } 4\lambda\Delta < p \leq \left(\sqrt{\lambda\Delta} + \sqrt{2\sqrt{w_L} + \lambda\Delta}\right)^2 \\ \frac{w_L}{p}, & \text{if } 2\sqrt{\Delta} < p \leq 4\lambda\Delta \\ \frac{w_H}{p} - \frac{p}{4}, & \text{if } 1 < p \leq 2\sqrt{\Delta} \end{cases} \tag{5.13}$$

By definition, the company's profit function is given as:

$$\Pi(p) = (\pi x_H^*(p) + (1-\pi)x_L^*(p))(p-c)$$

Plugging in $x_H^*(p)$ and $x_L^*(p)$ gives

$$\Pi(p) = \begin{cases} 0, & \text{if } p > (\sqrt{\lambda\Delta} + \sqrt{2\sqrt{w_L} + \lambda\Delta})^2 \\ \pi\left(\frac{w_L}{p} - \frac{p}{4} - \lambda\Delta + \sqrt{\lambda p\Delta}\right)(p-c), & \text{if } p > 2\sqrt{w_L} \\ \left(\frac{w_L}{p} - \frac{p}{4} - \pi\lambda\Delta + \pi\sqrt{\lambda p\Delta}\right)(p-c), & \text{if } p > 4\lambda\Delta \\ \left(\frac{w_L}{p} - \frac{(1-\pi)p}{4}\right)(p-c), & \text{if } p > 2\sqrt{\Delta} \\ \left(\frac{\bar{w}}{p} - \frac{p}{4}\right)(p-c) & \text{if } p > 1 \end{cases} \tag{5.14}$$

The profit function in Eq. (5.14) is a piecewise function. When the status good price exceeds the threshold $\left(\sqrt{\lambda\Delta} + \sqrt{2\sqrt{w_L} + \lambda\Delta}\right)^2$, then neither type of the consumer purchases the status good. When the price is below the threshold, then at least the high-type consumer will purchase some status goods. It can be verified that, as the status factor becomes stronger, it is more likely for the high-type consumer to make purchase.

However, when the price becomes extremely low (e.g. when $p \leq 2\sqrt{\Delta}$) and approaches the necessity good price 1, the consumption is no longer influences by λ. In this case, the status good is so affordable that the consumer purchases

it without the consideration of status signaling. In other words, as the price of the status good decreases, status signaling plays a less important role in the consumer's consumption. This shows that, the high price tage of the status good is not just a way to collect more profit; more importantly, it plays the role of screening the consumers and maintaining the brand value, which resonates with our observation of the practice.

5.4 Summary

In this chapeter, we looked into how a status-seeking consumer choosing between the status good and its counterfeit. The consumer faces a tradeoff: the counterfeit offers the price discount, which is particularly valuable for a low-type consumer who wants to engage in status signaling; however, when being found out, the counterfeit consumption leads to a loss of status for both types of consumers. We build a signaling game model to investigate such complicated dynamics behind the consumer's status signaling decision.

We started with a benchmark case when the status good is the only tool for status signaling. We found that to signal their status and differentiate from the low-type consumer, the high-type consumer distorts their demand and over-purchases the status good. The firm recognizes and exploits this by raising the price in response to an increase in the status effect.

We then introduce the counterfeit good to the consumer's choice set. The price discount of the counterfeit good gives the low-type consumer an opportunity to mimick the consumption behavior of the high-type consumer without break the bank. Our findings show that, the wealth difference, the status factor,

the exposure risk and cost jointly influence the equilibrium outcome. Moreover, the presence of counterfeits impacts the firm in two ways: first, there is a substitution effect, under which the consumer prefers the counterfeit over the status good, and the counterfeiter directly steals sales from the firm; second, there is a counter status-seeking effect, under which the high-type consumer is discouraged from status signaling due to the easy access of the counterfeit by the low-type consumer. Both effects hurt the sales and profit of the firm.

An interesting extension of the model is to the entry-deterrence game between the counterfeit and the firm. By examining the equilibrium entry and pricing decision of the counterfeiter, serveral important insights may arise. For example, how does the exposure cost and risk impact the entry decision of the counterfeiter? As the incumbent, what is the best strategy for the firm? Does the counterfeit good increase the social welfare in any way, and if so, under what conditions? These questions are being investigated by the author and the focus of Chen, Lian, and Yao 2021.

CHAPTER 6

CONCLUSION AND FUTURE DIRECTIONS

This thesis investigates the decisions of marketplace companies and the resulting implications both internally and across the industry. Across the various settings, a coherent theme is to design a market where buyers and sellers are incentivized to participate at the right time and the right place.

Results from this thesis highlights the importance of an industry-level angle: because both buyers and sellers are free agents with the freedom to switch among platforms quickly, it is particularly meaningful to evaluate the decisions of a company based on the broader context. For example, the economic benefits of new technology like autonomous vehicles may depend on the way ride-hailing drivers respond to prices; the profit of a ride-hailing company may also rely on the pricing strategy of a food-delivery company given they share the same driver pool. Viewing the decisions locally leads to inefficient and undesirable outcomes.

There are immediate next steps that extend this thesis works in important ways. For example, a natural extension of Chapter 4 is to analyze the demand-side competition among gig economy companies when they share a pool of workers on the supply side. In Chapter 3, we studied AVs' welfare implications for riders and the same question can be asked for drivers. The common belief is that AVs will make drivers strictly worse off by reducing their employment rate. However, preliminary analysis suggests that AVs may actually help expand the market for drivers when carefully designed.

From a methodological perspective, the possibility of validate the theories with empirical analysis is particular exciting. For example, Chapter 4 shows that

it is the larger firms' best interest to pay a higher wage to maintain the worker pool in the gig economy. While casual empiricism exists that supports the result, it would be interesting to study this relationship rigorously.

Technology has been and will keep surprising us with more and more applications of the marketplace design that match the demand with supply in smarter ways. For example, several types of blockchain-based marketplaces have been widely adopted, e.g., in the area of decentralized finance (DeFi) and digital art trades (NFT). These marketplaces have new characteristics such as being permissionless and providing full transparency, posing new design challenges such as the need for carefully crafted transaction fee mechanisms. These are exciting new areas worthwhile for future investigation.

APPENDIX A

OPTIMAL GROWTH IN TWO-SIDED MARKETS

Proposition 30 *Consider a trajectory of size $x(t)$ and balance $\gamma(t)$ that are continuously differentiable. Then it can be uniquely determined that the stock of supply is*

$$s(t) = \beta_1^s \gamma(t) x(t) \tag{A.1}$$

the stock of demand is

$$d(t) = \beta_1^d (1 - \gamma(t)) x(t) \tag{A.2}$$

the price paid to the seller is

$$p_s(t) = c + \frac{\gamma'(t)x(t) + \gamma(t)x'(t) + \beta_0^s \gamma(t)x(t)}{g(\gamma(t), x(t))} \tag{A.3}$$

and the price charged to the buyer is

$$p_d(t) = v - \frac{-\gamma'(t)x(t) + (1-\gamma(t))x'(t) + \beta_0^d (1-\gamma(t))x(t)}{g(\gamma(t), x(t))} \tag{A.4}$$

Proof. First, note that (A.2) and (A.1) are directly given by the definition of $\gamma(t)$. Next, by (2.6),

$$p_s = c + \frac{s' + \beta_0^s s}{\beta_1^s g(s, d)}$$

Since $s(t) = \beta_1^s \gamma(t) x(t)$ and $\gamma(t), x(t)$ are both differentiable, by the chain rule we have

$$s' = \beta_1^s (\gamma'(t)x(t) + \gamma(t)x'(t)) \tag{A.5}$$

Substituting s' in the expression of p_s with (A.5) gives $p_s(t)$. The expression of $p_d(t)$ can be obtained using similar steps by (2.7) and (A.2). □

Proposition 31 *Suppose there are finite number of jumps in the trajectory of $\gamma(t), x(t)$. Then for t around such a discontinuous time point t_0, the price paid to the seller is*

$$p_s(t) = \frac{\gamma(t_0^+)x(t_0^+) - \gamma(t_0^-)x(t_0^-)}{g(\gamma(t_0^-), x(t_0^-))}\delta(t_0 - t)$$

and the price charged to the buyer is

$$p_d(t) = \frac{(1-\gamma(t_0^-))x(t_0^-) - (1-\gamma(t_0^+))x(t_0^+)}{g(\gamma(t_0^-), x(t_0^-))}\delta(t_0 - t)$$

Proof. By proposition 30, around time t_0,

$$s(t_0^-) = \beta_1^s \gamma(t_0^-)x(t_0^-), s(t_0^+) = \beta_1^s \gamma(t_0^+)x(t_0^+)$$

We have shown in the proof of Proposition 1 that a price shock on the supply side $p_s(t) = \frac{s_1 - s_0}{\beta_1^s g(s_0, d_0)}\delta(\tau - t)$ can instantly shift the stock of supply from s_0 to s_1 at time τ. As a result, a price shock given below will shift the stock of supply from $s(t_0^-)$ to $s(t_0^+)$:

$$\begin{aligned} p_s(t) &= \frac{\beta_1^s \gamma(t_0^+)x(t_0^+) - \beta_1^s \gamma(t_0^-)x(t_0^-)}{\beta_1^s g(\gamma(t_0^-), x(t_0^-))}\delta(t_0 - t) \\ &= \frac{\gamma(t_0^+)x(t_0^+) - \gamma(t_0^-)x(t_0^-)}{g(\gamma(t_0^-), x(t_0^-))}\delta(t_0 - t) \end{aligned}$$

The price shock for the demand side can be shown using identical analysis. □

Proof of Lemma 1

Proof. By the definition of h,

$$h(\gamma) = (\theta\gamma^m(\beta_1^s)^m + (1-\theta)(1-\gamma)^m(\beta_1^d)^m)^{\frac{\alpha}{m}}$$

The smoothness can be checked by taking the derivative of $h(\gamma)$. Moreover, given that $m \geq 0$, $h(\gamma)$ is defined on $\gamma = 0$ and $\gamma = 1$. Hence, Assumption 1 is confirmed.

For Assumption 4, Denote $l(\gamma) = \theta\gamma^m(\beta_1^s)^m + (1-\theta)(1-\gamma)^m(\beta_1^d)^m$. $l(\gamma) > 0$. Then

$$h'(\gamma) = \frac{\alpha}{m} l(\gamma)^{\frac{\alpha}{m}-1} l'(\gamma) \tag{A.6}$$

$$h''(\gamma) = \frac{\alpha}{m}\left(\frac{\alpha}{m}-1\right) l(\gamma)^{\frac{\alpha}{m}-2} l'(\gamma)^2 + \frac{\alpha}{m} l(\gamma)^{\frac{\alpha}{m}-1} l''(\gamma) \tag{A.7}$$

Then

$$\begin{aligned}
&(1-\alpha)h'(\gamma)^2 + \alpha h(\gamma)h''(\gamma) \\
&= \frac{\alpha^2}{m^2} l(\gamma)^{\frac{2\alpha}{m}-2} l'(\gamma)^2 + l(\gamma)^{\frac{2\alpha}{m}-2}\left(-\frac{\alpha^2}{m}\right) l'(\gamma)^2 + \frac{\alpha^2}{m} l(\gamma)^{\frac{2\alpha}{m}-1} l''(\gamma) \\
&= \frac{\alpha^2}{m} l(\gamma)^{\frac{2\alpha}{m}-2}\left\{\left(\frac{1}{m}-1\right) l'(\gamma)^2 + l(\gamma) l''(\gamma)\right\}
\end{aligned}$$

To check the sign of $\left(\frac{1}{m}-1\right) l'(\gamma)^2 + l(\gamma)l''(\gamma)$,

$$\begin{aligned}
l'(\gamma) &= \theta(\beta_1^s)^m m\gamma^{m-1} - (1-\theta)(\beta_1^d)^m m(1-\gamma)^{m-1} \\
l''(\gamma) &= \theta(\beta_1^s)^m m(m-1)\gamma^{m-2} + (1-\theta)(\beta_1^d)^m m(m-1)(1-\gamma)^{m-2}
\end{aligned}$$

By some algebraic manipulation,

$$\left(\frac{1}{m}-1\right) l'(\gamma)^2 + l(\gamma)l''(\gamma) = (m-1)m\theta(1-\theta)(\beta_1^s)^m(\beta_1^d)^m\gamma^{m-2}(1-\gamma)^{m-2}$$

Then

$$\begin{aligned}
&(1-\alpha)h'(\gamma)^2 + \alpha h(\gamma)h''(\gamma) \\
&\qquad = \alpha^2 l(\gamma)^{\frac{2\alpha}{m}-2}(m-1)\theta(1-\theta)(\beta_1^s)^m(\beta_1^d)^m\gamma^{m-2}(1-\gamma)^{m-2}
\end{aligned} \tag{A.8}$$

Since $m < 1$, $(1-\alpha)h'(\gamma)^2 + \alpha h(\gamma)h''(\gamma) < 0$. Hence Assumption 4 is confirmed.

For Assumption 3, setting $h'(\gamma) = 0$ gives

$$\gamma = \frac{1}{1 + \left(\frac{\theta}{1-\theta}\left(\frac{\beta_1^s}{\beta_1^d}\right)^m\right)^{\frac{1}{m-1}}}$$

Since $\left(\frac{\theta}{1-\theta}(\frac{\beta_1^s}{\beta_1^d})^m\right)^{\frac{1}{m-1}} > 0$, the solution of γ is always in the range of $(0,1)$.

For Assumption 2, we have shown in (A.7) that

$$h''(\gamma) = \frac{\alpha}{m} l(\gamma)^{\frac{\alpha}{m}-2} \left((\frac{\alpha}{m} - 1) l'(\gamma)^2 + l(\gamma) l''(\gamma) \right)$$

Since $l(\gamma)$ does not contain any term related to α, for $m > 0$, one can show that $(\frac{\alpha}{m} - 1) l'(\gamma)^2 + l(\gamma) l''(\gamma)$ is increasing and continuous in α. If $\alpha < 1$,

$$(\frac{\alpha}{m} - 1) l'(\gamma)^2 + l(\gamma) l''(\gamma) < (\frac{1}{m} - 1) l'(\gamma)^2 + l(\gamma) l''(\gamma) < 0$$

If $\alpha > 1$, by continuity, there always exists $\epsilon > 0$ that is sufficiently small, such that

$$(\frac{1}{m} - 1) l'(\gamma)^2 + l(\gamma) l''(\gamma) < (\frac{1+\epsilon}{m} - 1) l'(\gamma)^2 + l(\gamma) l''(\gamma) < 0$$

Therefore, if $1 < \alpha \leq 1 + \epsilon$, $h''(\gamma) < 0$ still holds.

For $m = 0$, $h(\gamma) = \gamma^{\alpha\theta}(1-\gamma)^{\alpha(1-\theta)}$, which is the Cobb-Douglas function. The proof for Assumption 1 to 4 still hold by taking the limit of $m \to 0$. In particular, for Assumption 2, it can be shown that $h''(\gamma) < 0$ for any $\alpha < \min\{1/\theta, 1/(1-\theta)\}$, given $m = 0$:

$$\begin{aligned}
&\lim_{m\to 0} \frac{\alpha}{m} l'(\gamma) = \lim_{m\to 0} (\frac{\alpha}{m} - 1) l'(\gamma) = \alpha(\theta\gamma^{-1} - (1-\theta)(1-\gamma)^{-1}) \\
&\lim_{m\to 0} \frac{\alpha}{m} l''(\gamma) = \alpha\{-\theta\gamma^{-1} - (1-\theta)(1-\gamma)^2\} \\
&\lim_{m\to 0} l(\gamma)^{\frac{\alpha}{m}-2} = \gamma^{\alpha\theta}(1-\gamma)^{\alpha(1-\theta)}
\end{aligned}$$

Hence,

$$\lim_{m\to 0} h''(\gamma) = \gamma^{\alpha\theta}(1-\gamma)^{\alpha(1-\theta)} \left((\frac{\alpha\theta}{\gamma} - \frac{\alpha(1-\theta)}{1-\gamma})^2 - \frac{\alpha\theta}{\gamma^2} - \frac{\alpha(1-\theta)}{(1-\gamma)^2} \right) \tag{A.9}$$

(A.10)

which can be furthur written as

$$\lim_{m\to 0} h''(\gamma)$$
$$= \gamma^{\alpha\theta}(1-\gamma)^{\alpha(1-\theta)}\left(\frac{\alpha\theta(\alpha\theta-1)}{\gamma^2}+\frac{\alpha(1-\theta)(\alpha(1-\theta)-1)}{(1-\gamma)^2}-\frac{2\theta(1-\theta)\alpha^2}{\gamma(1-\gamma)}\right) \tag{A.11}$$

Since $\alpha\theta < 1$, $\alpha(1-\theta) < 1$, (A.9) is negative. Hence, $h''(\gamma) < 0$. □

Lemma 6 *For any finite trajectory $x(t)$, the following equation holds:*

$$\int_0^T e^{-\rho t}\pi g(\gamma, x)dt = \int_0^T e^{-\rho t}G(\gamma, x)dt - e^{-\rho T}x(T) + x(0)$$

Proof. By Definition 4,

$$G(\gamma, x) = (v-c)g(\gamma, x) - (\rho + \beta_0^s\gamma + \beta_0^d(1-\gamma))x$$

By the formulation of $\dot{x}$,

$$\pi g(\gamma, x) = (v-c)g(\gamma, x) - (\beta_0^s\gamma + \beta_0^d(1-\gamma))x - \dot{x}$$

Then

$$\int_0^T e^{-\rho t}\pi g(\gamma, x)dt$$
$$= \int_0^T ((v-c)g(\gamma, x) - (\beta_0^s\gamma + \beta_0^d(1-\gamma))x)e^{-\rho t}dt - \int_0^T \dot{x}e^{-\rho t}dt \tag{A.12}$$

Integration by parts gives

$$\int_0^T \dot{x}e^{-\rho t}dt = \int_0^T x(t)\rho e^{-\rho t}dt + e^{-\rho T}x(T) - x(0)$$

Therefore,

$$\int_0^T e^{-\rho t}\pi g(\gamma, x)dt$$
$$= \int_0^T ((v-c)g(\gamma, x) - (\rho + \beta_0^s\gamma + \beta_0^d(1-\gamma))x)e^{-\rho t}dt - e^{-\rho T}x(T) + x(0)$$

□

Lemma 7 *Under Assumption 2 and 4, $G(\gamma, x)$ is strictly concave in γ. Moreover, if $\alpha < 1$, then $\max_\gamma G(\gamma, x)$ is strictly concave in x and reaches maximum at $x = x^*$; if $\alpha > 1$, then $\max_\gamma G(\gamma, x)$ is strictly convex in x and reaches global minimum at $x = x^*$.*

Proof. By the concavity of $h(\gamma)$ (Assumption 2),

$$\frac{\partial^2 G(\gamma, x)}{\partial \gamma^2} = (v - c)h''(\gamma)x^\alpha < 0$$

Then $\arg\max_\gamma G(\gamma, x)$ can be obtained by setting

$$\frac{\partial G(\gamma, x)}{\partial \gamma} = -(\beta_0^s - \beta_0^d)x + (v - c)h'(\gamma)x^\alpha = 0$$

Note that $h'(\gamma)$ is bounded by $h'(0)$ and $h'(1)$, due to $\gamma \in [0, 1]$. Thus, the optimal γ can be solved by

$$\gamma^*(x) = \max\left\{\min\{(h')^{-1}(\frac{(\beta_0^s - \beta_0^d)x^{1-\alpha}}{v - c}), 1\}, 0\right\} \tag{A.13}$$

Then for those x that $\gamma^*(x)$ has an interior solution,

$$\frac{d\gamma^*(x)}{dx} = \frac{(\beta_0^s - \beta_0^d)(1 - \alpha)}{(v - c)h''(\gamma)x^\alpha}$$

Since $h''(\gamma) < 0$,

$$\text{sgn}(\frac{d\gamma^*(x)}{dx}) = \text{sgn}((\alpha - 1)(\beta_0^s - \beta_0^d)) \tag{A.14}$$

By the envelope theorem,

$$\begin{aligned}\frac{\partial G(\gamma^*(x), x)}{\partial x} &= \frac{\partial G(\gamma, x)}{\partial x}\Big|_{\gamma=\gamma^*(x)} \\ &= -(\rho + \beta_0^s\gamma^*(x) + \beta_0^d(1 - \gamma^*(x))) + (v - c)\alpha h(\gamma^*(x))x^{\alpha-1}\end{aligned} \tag{A.15}$$

When (A.13) has an interior solution, plugging in (A.13)gives

$$\frac{\partial G(\gamma^*(x),x)}{\partial x} = \alpha(\beta_0^s - \beta_0^d)\frac{h(\gamma^*(x))}{h'(\gamma^*(x))} - (\beta_0^s\gamma^*(x) + \beta_0^d(1-\gamma^*(x)) + \rho)$$

Then

$$\frac{\partial G^2(\gamma^*(x),x)}{\partial x^2} = (\beta_0^s - \beta_0^d)\left(\frac{\alpha h'(\gamma)^2 - \alpha h(\gamma)h''(\gamma)}{h'(\gamma)^2} - 1\right)\frac{d\gamma^*(x)}{dx}$$

By (A.14), $\operatorname{sgn}(\frac{\partial G^2(\gamma^*(x),x)}{\partial x^2}) = \operatorname{sgn}((\alpha-1)((\alpha-1)h'(\gamma)^2 - \alpha h(\gamma)h''(\gamma)))$. When (A.13) does not have an interior solution, $\operatorname{sgn}(\frac{\partial G^2(\gamma^*(x),x)}{\partial x^2}) = \operatorname{sgn}(\alpha-1)$.

Therefore, if $\alpha > 1$, $G(\gamma^*(x),x)$ is strictly convex; if $\alpha < 1$, $G(\gamma^*(x),x)$ is strictly concave under Assumption 4.

Setting $\frac{\partial G(\gamma^*(x),x)}{\partial x} = 0$ gives

$$x = \left(\frac{\rho + \beta_0^s\gamma^*(x) + \beta_0^d(1-\gamma^*(x))}{(v-c)\alpha h(\gamma^*(x))}\right)^{\frac{1}{\alpha-1}}$$

The solution to the above equation is γ^*, and the corresponding market size x is x^*. To check the existence of x^*, it can be verified that $\lim_{x\to 0} G_x(\gamma^*(x),x))$ and $\lim_{x\to+\infty} G_x(\gamma^*(x),x)$ have opposite signs. Since $G_x(\gamma^*(x),x)$ is monotone and continuous, $G_x(\gamma^*(x),x) = 0$ must have a unique solution.

Therefore, x^* is the global minimum (maximum) for $\max_\gamma G(\gamma,x)$ under $\alpha > 1$ ($\alpha < 1$). □

Proof of Theorem 1

Proof. By Lemma 6, the infinite-horizon problem (2.13) can be written as

$$\int_0^\infty e^{-\rho t}G(\gamma,x)dt + x(0) \tag{A.16}$$

Since the selection of γ only affects the term $G(\gamma,x)$, by Lemma 7, it is optimal to set $\gamma = \arg\max G(\gamma,x)$. Steps for obtaining $\gamma^*(x)$ and its monotonicity can be found in the proof of Lemma 7.

To see the limit of γ^*, consider $\beta_0^s < \beta_0^d$ and $\alpha < 1$. $\lim_{x\to 0} h'^{-1}(\frac{(\beta_0^s-\beta_0^d)x^{1-\alpha}}{v-c}) = h'^{-1}(0) = \gamma^*$. By Assumption 3, $\min(1,\gamma^*) = \gamma^*$. For $x \to +\infty$, $\lim_{x\to+\infty} h'^{-1}(\frac{(\beta_0^s-\beta_0^d)x^{1-\alpha}}{v-c}) = h'^{-1}(-\infty)$. if $h'(1)$ is bounded, then $h'(1) > -\infty$, $h'^{-1}(h'(1)) < h'^{-1}(-\infty)$, and $\min(1, h'^{-1}(-\infty)) = 1$. If $h'(1)$ is unbounded, by Assumption 2 and 3, $h'(1) \to -\infty$. Then $h'^{-1}(-\infty)$ just equals to 1. The other three cases can be checked similarly. □

Proof of Theorem 2

Proof. The current value Hamiltonian to infinite-horizon problem (2.13) is

$$H(x,\pi,\gamma,\psi) = \pi g(\gamma,x) + \psi\left(-(\beta_0^s\gamma + \beta_0^d(1-\gamma))x + (v-c-\pi)g(\gamma,x)\right)$$

By the maximum principle, a stationary solution must satisfy

$$H_\pi = 0, H(x^*,\pi^*,\gamma^*,\psi) > H(x^*,\pi^*,\gamma,\psi), \psi' = \rho\psi - H_x = 0, x' = 0$$

By $H_\pi = 0$ and $g(\gamma,x) = h(\gamma)x^\alpha$,

$$(1-\psi)\pi h(\gamma)x^\alpha = 0$$

Since $x > 0, 0 < \gamma < 1, h(\gamma) \neq 0$,

$$\psi = 1$$

which satisfies the limiting transversality condition

$$\lim_{T\to\infty} e^{-\rho T}\psi(T) = 0$$

To maximize $H(x^*,\pi^*,\gamma,\psi)$, set $H_\gamma = 0$ assuming it has a solution on $[0,1]$:

$$(\pi - \psi\pi + v - c)h'(\gamma)x^\alpha - \psi(\beta_0^s - \beta_0^d)x = 0$$

Substitute by $\psi = 1$ and rearrange the terms,

$$(v-c)h'(\gamma) = (\beta_0^s - \beta_0^d)x^{1-\alpha} \tag{A.17}$$

When (A.17) does not have a solution on $[0,1]$, the optimal γ is on the boundary; whether $\gamma = 1$ or 0 depends on its gradient. One can check that the balance that maximizes H is given by

$$\gamma^*(x) = \max\Big\{ \min\Big\{ h'^{-1}\big(\frac{(\beta_0^s - \beta_0^d)x^{1-\alpha}}{v-c}\big), 1\Big\}, 0\Big\}$$

By $\psi' = \rho\psi - H_x$ and $\psi = 1$,

$$0 = \rho - H_x$$

which gives

$$\rho = -(\beta_0^s \gamma^*(x) + \beta_0^d(1-\gamma^*(x))) + (v-c)\alpha h(\gamma^*(x))x^{\alpha-1} \tag{A.18}$$

Both x and φ are constants, so the last two equations are satisfied. Combining (A.17) and (A.18) gives the stationary solution.

We also need to check whether H is jointly concave in (γ, x) for sufficiency.

$$\begin{aligned} &H(x,\pi,\gamma^*(x),\psi) \\ &= \{(1-\psi)\pi + \psi(v-c)\}h(\gamma^*(x))x^\alpha - \psi(\beta_0^s\gamma^*(x) + \beta_0^d(1-\gamma^*(x)))x, \end{aligned} \tag{A.19}$$

$$\begin{aligned} &H_x(x,\pi,\gamma^*(x),\psi) \\ &\quad = \{(1-\psi)\pi + \psi(v-c)\}h(\gamma^*(x))\alpha x^{\alpha-1} - \psi(\beta_0^s\gamma^*(x) + \beta_0^d(1-\gamma^*(x))) \end{aligned} \tag{A.20}$$

and

$$\begin{aligned} &H_{xx}(x,\pi,\gamma^*(x),\psi) = ((1-\psi)\pi + \\ &\psi(v-c))\alpha\left(h'(\gamma)\frac{d\gamma^*(x)}{dx} + h(\gamma)(\alpha-1)x^{\alpha-2}\right) - \psi(\beta_0^s - \beta_0^d)\frac{d\gamma^*(x)}{dx} \end{aligned} \tag{A.21}$$

When $\gamma^*(x)$ is a boundary solution, i.e. $\gamma^*(x) = 1$ or 0, H_{xx} can be simplified as

$$H_{xx}(x, \pi, \gamma^*(x), \psi) = ((1-\psi)\pi + \psi(v-c))\alpha h(\gamma^*(x))(\alpha-1)x^{\alpha-2}$$

When $\gamma^*(x)$ is an interior solution, by $H_\gamma = 0$,

$$((1-\psi)\pi + \psi(v-c))h'(\gamma)x^{\alpha-1} = \psi(\beta_0^s - \beta_0^d)$$

Moreover,

$$\frac{d\gamma^*(x)}{dx} = \frac{\psi(\beta_0^s - \beta_0^d)(1-\alpha)}{((1-\psi)\pi + \psi(v-c))h''(\gamma)x^\alpha} = \frac{h'(\gamma)(1-\alpha)}{h''(\gamma)x}$$

Hence,

$$H_{xx}(x, \pi, \gamma^*(x), \psi) = \frac{(\alpha-1)\psi^2(\beta_0^s - \beta_0^d)^2}{((1-\psi)\pi + \psi(v-c))h''(\gamma)x^\alpha}\left\{\frac{\alpha h(\gamma)h''(\gamma)}{h'(\gamma)^2} + 1 - \alpha\right\}$$

By the maximum principle, $H_\pi = 0$, which implies that $\psi(t) = 1$. Therefore, when $\gamma^*(x)$ is an interior solution,

$$H_{xx}(x, \pi, \gamma^*(x), \psi) = \frac{(\alpha-1)(\beta_0^s - \beta_0^d)^2}{(v-c)h''(\gamma)x^\alpha}\left\{\frac{\alpha h(\gamma)h''(\gamma)}{h'(\gamma)^2} + 1 - \alpha\right\}$$

When $\gamma^*(x)$ is a boundary solution,

$$H_{xx}(x, \pi, \gamma^*(x), \psi) = (v-c)\alpha h(\gamma^*(x))(\alpha-1)x^{\alpha-2}$$

If $\alpha < 1$, under Assumption 2 and 4, $h''(\gamma) < 0$, $\frac{\alpha h(\gamma)h''(\gamma)}{h'(\gamma)^2} + 1 - \alpha < 0$, which implies that $H_{xx} < 0$. Therefore, Hamiltonian is jointly concave in (γ, x). If $\alpha > 1$, $\frac{\alpha h(\gamma)h''(\gamma)}{h'(\gamma)^2} < 0$, $1 - \alpha < 0$, and again by Assumption 2, $h''(\gamma) < 0$. Therefore, $H_{xx} > 0$. This implies that the Hamiltonian $H(x, \pi, \gamma^*(x), \psi)$ attains its minimum at $x = x^*$. Thus, the stationary solution characterizes a saddle point. □

Theorem 9 (Fast vs. Slow) *Consider a set of increasing growth paths from $\underline{x}$ to $\bar{x}$ and the fixed endpoint problem*

$$\max \int_0^{t_0} e^{-\rho t}\pi g(\gamma^*(x), x)dt$$

$$s.t. \quad x(0) = \underline{x},\ x(t_0) = \bar{x}$$

Then in a decreasing returns to scale market, if $0 < \underline{x} < \bar{x} \leq x^*$*, faster growth dominates slower growth; and if* $x^* \leq \underline{x} < \bar{x}$*, slower growth dominates faster growth. Conversely, in an increasing returns to scale market, if* $0 < \underline{x} < \bar{x} \leq x^*$*, then slower growth dominates faster growth; if* $x^* \leq \underline{x} < \bar{x}$*, then faster growth dominates slower growth. To summarize,*

	$\alpha > 1$	$\alpha < 1$
$\underline{x} > x^*$	*faster is better*	*slower is better*
$\bar{x} < x^*$	*slower is better*	*faster is better*

This result shows that the stationary point in Theorem 2 defines a threshold between growth strategies over an interval – determining when it is optimal to grow fast and when it is optimal to grow slow. We next apply this result to analyze the optimal growth strategies overall.

Proof. By Lemma 6 in the Appendix,

$$\int_0^{t_0} e^{-\rho t} \pi g(\gamma^*(x), x) dt = \int_0^{t_0} G(\gamma^*(x), x) e^{-\rho t} dt - e^{-\rho t_0} \bar{x} + \underline{x}$$

Let $x_1(t)$ be a faster growth path from $\underline{x}$ to $\bar{x}$ over $[0, t_0]$ than $x_2(t)$. Let the corresponding pricing policies be $\pi_1(t)$ and $\pi_2(t)$. Then

$$\begin{aligned} &\int_0^{t_0} e^{-\rho t} \pi_1 g(\gamma^*(x_1), x_1) dt - \int_0^{t_0} e^{-\rho t} \pi_2 g(\gamma^*(x_2), x_2) dt \\ &\quad = \int_0^{t_0} (G(\gamma^*(x_1), x_1) - G(\gamma^*(x_2), x_2)) e^{-\rho t} dt \end{aligned} \tag{A.22}$$

Setting $\frac{dG(\gamma^*(x),x)}{dx} = 0$ gives

$$-(\rho + \beta_0^s \gamma^*(x) + \beta_0^d(1-\gamma^*(x))) + (v-c)\alpha h(\gamma^*(x))x^{\alpha-1} = 0$$

This is the same equation as (A.18). Therefore, $x = x^*$ is the solution to $\frac{dG(\gamma^*(x),x)}{dx} = 0$.

By Lemma 7 in the Appendix, when $\alpha > 1$, $G(\gamma^*(x), x)$ is convex in x. $x = x^*$ is the global minimum. $G(\gamma^*(x), x)$ is strictly decreasing for $x < x^*$ and strictly increasing for $x > x^*$.

If $\bar{x} < x^*$, then $x_2(t) < x_1(t) < x^*$, and

$$G(\gamma^*(x_1), x_1) < G(\gamma^*(x_2), x_2)$$

for all $0 \leqslant t \leqslant t_0$. As a result, (A.22) is negative. This implies that slower growth paths dominate faster growth paths.

Similarly, if $x^* < \underline{x}$, then $x^* < x_2(t) < x_1(t)$, and

$$G(\gamma^*(x_1), x_1) > G(\gamma^*(x_2), x_2)$$

for all $0 \leqslant t \leqslant t_0$. As a result, (A.22) is positive. This implies that faster growth paths dominate slower growth paths.

For $\alpha < 1$, under Assumption 4, $G(\gamma^*(x), x)$ is concave in x. So the monotonicity flips for $x > x^*$ and $x < x^*$. Using a similar analysis as for $\alpha > 1$ completes the proof. □

Lemma 8 *Any increasing growth path from x_0 to $\bar{x}$ is weakly dominated by*

$$x(t;t_i) = \begin{cases} x_0, & t \leq t_i \\ F(t-t_i), & t_i < t \leq t_i + F^{-1}(\bar{x}) \\ \bar{x}, & t_i + F^{-1}(\bar{x}) < t \leq T \end{cases} \tag{A.23}$$

where $0 \leq t_i \leq T - F^{-1}(\bar{x})$.

Proof. If $x_0 < \bar{x} < x^*$, by Theorem 9, it is optimal to grow the market as slow as possible. In this case, the slowest growth path is given by

$$x_s(t) = x(t; T - F^{-1}(\bar{x})) \tag{A.24}$$

To see why, suppose there is another increasing growth path $y(t)$ from x_0 to $\bar{x}$ over $[0, T]$ that is admissible and not faster than $x_s(t)$ given here. By definition, there exists a time point $t' \in [0, T]$ such that $y(t') < x_s(t')$. if $t' \leqslant T - F^{-1}(\bar{x})$, then $y(t') < x_s(t') = x_0$. This can't be true because $y(t)$ is an increasing growth path, and thus $y(t) \geqslant x_0$. If $t' > T - F^{-1}(\bar{x})$, then $y(t') < F(t' - T + F^{-1}(\bar{x}))$. Then $F^{-1}(y(t')) < t' - T + F^{-1}(\bar{x})$. The shortest time it takes to grow from $y(t')$ to $\bar{x}$ follows

$$F^{-1}(\bar{x}) - F^{-1}(y(t')) > T - t'$$

Therefore, $y(t)$ cannot reach $\bar{x}$ before or at $t = T$. Contradiction.

Similarly, if $x_0 > x^*$, again by Theorem 9, it is optimal to grow the market as fast as possible. The fastest growth path is given by

$$x_f(t) = x(t; 0) \tag{A.25}$$

The proof is similar to that for the slowest growth path (A.24).

For $x_0 < x^* < \bar{x}$, again consider an increasing growth path $y(t)$ from x_0 to $\bar{x}$ over $[0, T]$ that is admissible but doesn't satisfy (A.23). Since the growth rate is bounded, $y(t)$ is continuous, and thus must cross x^*. Denote the time $y(t) = x^*$ as $t_{y=x^*}$. We construct a the following growth path:

$$x_{s-f}(t) = x(t; t_{y=x^*}) \tag{A.26}$$

One can check that this growth path is the slowest from x_0 to x^* over $[0, t_{y=x^*}]$ and the fastest from x^* to $\bar{x}$ over $[t_{y=x^*}, T]$ using similar arguments for proving (A.24). □

Proof of Theorem 3

Proof. By Lemma 6, the infinite-horizon problem (2.13) can be written as

$$\int_0^\infty e^{-\rho t} G(\gamma, x) dt + x(0) \tag{A.27}$$

Since the selection of γ only affects the term $G(\gamma, x)$, by Lemma 7 and Theorem 1, it is optimal to set $\gamma = \gamma^*(x(t))$. Then (A.27) is just a function of $x(t)$. We show that the solution has the property of a *most rapid approach path* (see Spence and Starrett 1975). The next part is also similar to the steps taken in Spence and Starrett 1975:

(1) Any path from x_0 to $\bar{x} > x_0$ is feasible. This is true by having $f(x) > 0$ for all x.

(2) The optimal growth path either (a) stays at x_0 forever, or (b) goes to $\bar{x}$. This can be shown by contradiction. Suppose (a) and (b) are both not optimal; Then the optimal increasing growth path must grow to a market size y such that $x_0 < y < \bar{x}$. If $G(\gamma^*(y), y) \leq G(\gamma^*(x_0), x_0)$, by convexity, $G(\gamma^*(x), x) < G(\gamma^*(x_0), x_0)$ for any x that $x_0 < x < y$. Then any path from x_0 to y is clearly dominated by (a); If $G(\gamma^*(y), y) > G(\gamma^*(x_0), x_0)$, by Lemma 7, it must be true that $G(\gamma^*(x), x) > G(\gamma^*(y), y)$ for any x that $y < x \leq \bar{x}$. But this implies that a growth path from x_0 to y is strictly dominated by (b). Contradiction.

(3) By Lemma 8, a candidate for an optimal growth path from x_0 to $\bar{x}$ must have the following property:

$$x(t;t_i) = \begin{cases} x_0, & 0 \le t \le t_i \\ F(t-t_i), & t_i < t \le t_i + F^{-1}(\bar{x}) \\ \bar{x}, & t_i + F^{-1}(\bar{x}) < t \end{cases} \tag{A.28}$$

We show that the optimal t_i is either 0 or $+\infty$. (A.27) can be further expanded as

$$J = \int_{t_i}^{t_i+F^{-1}(\bar{x})} e^{-\rho t} G(\gamma^*(F(t-t_i)), F(t-t_i))dt + G(\gamma^*(x_0), x_0)\frac{1-e^{-\rho t_i}}{\rho} + G(\gamma^*(\bar{x}), \bar{x})\frac{e^{-\rho(t_i+F^{-1}(\bar{x}))}}{\rho} \tag{A.29}$$

and furthurmore,

$$J = e^{-\rho t_i}\int_0^{F^{-1}(\bar{x})} e^{-\rho t} G(\gamma^*(F(t)), F(t))dt + G(\gamma^*(x_0), x_0)\frac{1-e^{-\rho t_i}}{\rho} + G(\gamma^*(\bar{x}), \bar{x})\frac{e^{-\rho(t_i+F^{-1}(\bar{x}))}}{\rho} \tag{A.30}$$

The second equality is by the change of variable. Taking derivative of J over t_i gives:

$$\frac{\partial J}{\partial t_i} = -e^{-\rho t_i}(\rho\int_0^{F^{-1}(\bar{x})} e^{-\rho t} G(\gamma^*(F(t)), F(t))dt - G(\gamma^*(x_0), x_0) + e^{-\rho F^{-1}(\bar{x})} G(\gamma^*(\bar{x}), \bar{x})) \tag{A.31}$$

which can be furthur simplified as

$$\frac{\partial J}{\partial t_i} = -e^{-\rho t_i}\int_0^{F^{-1}(\bar{x})} e^{-\rho t} G_x(\gamma^*(F(t)), F(t))F'(t)dt \tag{A.32}$$

The second equality is obtained by integration by parts. Hence, the sign of $\frac{\partial J}{\partial t_i}$ does not change in t_i. In particular, if $\int_0^{F^{-1}(\bar{x})} e^{-\rho t} G_x(\gamma^*(F(t)), F(t))F'(t)dt > 0$, $\frac{\partial J}{\partial t_i} < 0$, the growth path $x(t;0)$ is optimal; otherwise, the growth path $x(t;+\infty)$ is optimal.

(4) It can be checked that $x(t;+\infty)$ generates the same profit (A.27) as $x(t) = x_0, \forall t$. So it is sufficient to compare $x(t;0)$ and $x(t;+\infty)$. Define

$$S(\bar{x}) = \int_0^{F^{-1}(\bar{x})} e^{-\rho t} G_x(\gamma^*(F(t)), F(t)) F'(t) dt \tag{A.33}$$

We show that for any $x_0 < x^*$, there exists an $\tilde{x}$ such that $S(\bar{x}) > 0$ for any $\bar{x} > \tilde{x}$, and $S(\bar{x}) < 0$ for any $\bar{x} < \tilde{x}$:

$$S'(\bar{x}) = \frac{dF^{-1}(\bar{x})}{d\bar{x}} e^{-\rho F^{-1}(\bar{x})} G_x(\gamma^*(\bar{x}), \bar{x}) F'(F^{-1}(\bar{x})) = e^{-\rho F^{-1}(\bar{x})} G_x(\gamma^*(\bar{x}), \bar{x})$$

The derivative is obtained following Leibniz integral rule. Then by Lemma 7, for $\bar{x} > x^*(\bar{x} < x^*)$, $S'(\bar{x}) > 0 (S'(\bar{x}) < 0)$. Moreover, $S(x_0) = 0$. Therefore, if $x_0 < x^*$, then $S(\bar{x}) = 0$ must have a unique positive root. Denote it as $\tilde{x}$. Then $S(\bar{x}) < 0$ for $\bar{x} < \tilde{x}$ and $S(\bar{x}) > 0$ for $\bar{x} > \tilde{x}$.

If $x_0 > x^*$, then $G_x(\gamma^*(F(t)), F(t)) > 0$ on $[x_0, \bar{x}]$. Hence, $S(\bar{x}) > 0$ as well.

Hence, given that $x_0 < x^*$, for $\bar{x} > \tilde{x}$, $S(\bar{x}) > 0$, $\frac{\partial J}{\partial t_i} < 0$, the optimal policy is to grow as fast as possible, and the optimal growth path is $x(t;0)$; for $\bar{x} < \tilde{x}$, $S(\bar{x}) < 0$, $\frac{\partial J}{\partial t_i} > 0$, the optimal policy is to grow as slow as possible, and the optimal growth path is $x(t) = x_0, \forall t$. Given that $x_0 > x^*$, it is optimal to grow as fast as possible, and the optimal growth path is $x(t;0)$. □

Proof of Theorem 4

Proof. By Theorem 2, in a decreasing returns to scale market, the long-run optimal size is the saturation size x^*. By Theorem 9, faster growth dominates slower growth below x^*. Hence, the optimal growth policy is to grow to the saturation size as fast as possible. □

Proof of Theorem 5

Proof. This is a combination of Theorem 3 and Theorem 4. □

Proof of Proposition 2

Proof. We will construct a feasible policy with unbounded value. Denote the initial market size as x_0. Fix $\gamma = 0.5$ and $\pi = 0.5(v - c)$. Apply an impulse to instantly increase the market size from $x(0)$ to $\tilde{x}$, where $\tilde{x} > (\frac{\beta_0^s + \beta_0^d}{(v-c)h(0.5)})^{\frac{1}{\alpha-1}}$. Then we can directly obtain the expression of $x(t)$ by integrating the differential equation:

$$\dot{x} = -0.5(\beta_0^s + \beta_0^d)x + 0.5(v - c)h(0.5)x^\alpha$$

which gives

$$x(t) = \left(\frac{0.5(\beta_0^s + \beta_0^d)}{0.5(v - c)h(0.5) - C_0 e^{0.5(\beta_0^s + \beta_0^d)(\alpha-1)t}} \right)^{\frac{1}{\alpha-1}}$$

$C_0 = 0.5(v - c)h(0.5) - 0.5(\beta_0^s + \beta_0^d)\tilde{x}^{1-\alpha}$. Then for this policy, x goes to infinity as t approaches $\frac{1}{0.5(\alpha-1)(\beta_0^s+\beta_0^d)} \ln(0.5(v - c)h(0.5)/C_0)$. Since the integrand becomes unbounded in a finite amount of time, the discounted objective value is unbounded. □

Proof of Proposition 3

Proof. For $\alpha > 1$, by Proposition 2, there exists a feasible growth path that leads to unbounded profit, while keeping $x(t) = x_0$ generates finite profit. Thus, an increasing returns to scale market is viable.

For $\alpha < 1$, if $x_0 < x^*$, then by Lemma 7, $G(\gamma^*(x), x) > G(\gamma^*(x_0), x_0)$ for all $x_0 < x \leq x^*$. Hence, by Lemma 6, any increasing growth path from x_0 to x^* generates a higher profit than $x(t) = x_0$. □

Proof of Proposition 4

Proof. Prove by contradiction. Suppose there exists some $\tilde{x} < x_c$ such that $\dot{x} \geq 0$ and $\pi \geq 0$. When the market size is $\tilde{x}$ and $\dot{x} \geq 0$, by (2.12),

$$-(\beta_0^s \gamma + \beta_0^d(1 - \gamma))\tilde{x} + (v - c - \pi)h(\gamma)\tilde{x}^\alpha \geq 0$$

Since $h(\gamma) > 0$,

$$\pi \leq v - c - \frac{\beta_0^s \gamma + \beta_0^d (1-\gamma)}{h(\gamma)\tilde{x}^{\alpha-1}}$$

Maximize the right-hand side over γ on $[0,1]$:

$$\left(v - c - \frac{\beta_0^s \gamma + \beta_0^d (1-\gamma)}{h(\gamma)\tilde{x}^{\alpha-1}} \right)_\gamma = \frac{-(\beta_0^s - \beta_0^d)h(\gamma) + (\beta_0^s \gamma + \beta_0^d (1-\gamma))h'(\gamma)}{h(\gamma)^2 \tilde{x}^{\alpha-1}} \tag{A.34}$$

Setting the numerator of (A.34) to 0 gives

$$\frac{h'(\gamma)}{h(\gamma)}(\beta_0^s \gamma + \beta_0^d (1-\gamma)) = \beta_0^s - \beta_0^d \tag{A.35}$$

When (2.17) has a solution on $[0,1]$, it is the expression for γ_c. Moreover, (A.34) is positive for $\gamma < \gamma_c$ and negative for $\gamma > \gamma_c$. To see why, the numerator in (A.34) is decreasing in γ. This can be seen by checking its first-order derivative:

$$\begin{aligned} &-(\beta_0^s - \beta_0^d)h'(\gamma) + (\beta_0^s - \beta_0^d)h'(\gamma) + (\beta_0^s \gamma + \beta_0^d (1-\gamma))h''(\gamma) \\ &\qquad = (\beta_0^s \gamma + \beta_0^d (1-\gamma))h''(\gamma) < 0 \end{aligned} \tag{A.36}$$

The last step is by Assumption 2. Therefore, $\gamma = \gamma_c$ is the global maximizer. When (2.17) does not have a solution on $[0,1]$, it means that the maximizer is a boundary solution. Extending the definition of γ_c to include the boundary solutions gives by

$$\gamma_c^\dagger = \begin{cases} 0, & \frac{h'(0)}{h(0)} < \frac{\beta_0^s - \beta_0^d}{\beta_0^d} \\ \gamma_c, & \beta_0^s \frac{h'(1)}{h(1)} < \beta_0^s - \beta_0^d < \beta_0^d \frac{h'(0)}{h(0)} \\ 1, & \frac{h'(1)}{h(1)} > \frac{\beta_0^s - \beta_0^d}{\beta_0^s} \end{cases}$$

This means

$$v - c - \frac{\beta_0^s \gamma + \beta_0^d (1-\gamma)}{h(\gamma)\tilde{x}^{\alpha-1}} \leq v - c - \frac{\beta_0^s \gamma_c^\dagger + \beta_0^d (1-\gamma_c^\dagger)}{h(\gamma_c^\dagger)\tilde{x}^{\alpha-1}}$$

Since $\alpha > 1$ and $\tilde{x} < x_c$,

$$v - c - \frac{\beta_0^s \gamma_c^\dagger + \beta_0^d (1-\gamma_c^\dagger)}{h(\gamma_c^\dagger)\tilde{x}^{\alpha-1}} < v - c - \frac{\beta_0^s \gamma_c^\dagger + \beta_0^d (1-\gamma_c^\dagger)}{h(\gamma_c^\dagger){x_c}^{\alpha-1}}$$

By (2.16), the right-hand side equals to 0. Then

$$\pi \leq v - c - \frac{\beta_0^s \gamma + \beta_0^d (1-\gamma)}{h(\gamma)\tilde{x}^{\alpha-1}} \leq v - c - \frac{\beta_0^s \gamma_c^\dagger + \beta_0^d (1-\gamma_c^\dagger)}{h(\gamma_c^\dagger)\tilde{x}^{\alpha-1}}$$
$$< v - c - \frac{\beta_0^s \gamma_c^\dagger + \beta_0^d (1-\gamma_c^\dagger)}{h(\gamma_c^\dagger){x_c}^{\alpha-1}} = 0 \quad \text{(A.37)}$$

Therefore, such an $\tilde{x}$ does not exist. □

Proof of Proposition 5

Proof. By the proof of proposition 4, for any $x < x_c$, $\dot{x} \geq 0$ implies that $\pi < 0$. □

Proof of Proposition 6

Proof. By Lemma 6, for the infinite horizon problem, the objective function is equivalent to

$$\int_0^\infty e^{-\rho t} G(\gamma, x) dt + x(0)$$

By Theorem 1, $\gamma(t) = \gamma^*(x(t))$. By the convexity of $G(\gamma^*(x), x)$ by Lemma 7, $\arg\max_{x \leqslant \bar{x}} G(\gamma^*(x), x)$ is either x_0 or $\bar{x}$. If $G(\gamma^*(\bar{x}), \bar{x}) > G(\gamma^*(x_0), x_0)$, then a jump from x_0 to $\bar{x}$ is optimal; if not, $x(t) = x_0$ is optimal.

By Lemma 7, $G(\gamma^*(x), x)$ is continuous. $\lim_{x\to 0} G(\gamma^*(x), x) \to 0$. $\lim_{x\to\infty} G(\gamma^*(x), x) \to \infty$. Then for $x_0 < x^*$, $\tilde{x}$ defined in Proposition 6 must always exist. Moreover, if $\bar{x} > \tilde{x}$, $G(\gamma^*(\bar{x}), \bar{x}) > G(\gamma^*(x_0), x_0)$; if $x_0 < \bar{x} < \tilde{x}$, $G(\gamma^*(\bar{x}), \bar{x}) < G(\gamma^*(x_0), x_0)$. □

Proof of Lemma 2

Proof. We first show that $G(\gamma^*(x_c), x_c) < 0$. By Theorem 1, $\gamma^*(x_c) = \gamma_c$, and

$$-(\beta_0^s \gamma_c + \beta_0^d (1-\gamma_c)) x_c + (v-c) h(\gamma_c) x_c^\alpha = 0$$

Therefore, by Lemma 6,

$$G(\gamma^*(x_c), x_c) = -\rho x_c < 0$$

Since $\alpha > 1$,

$$\lim_{x \to 0} G(\gamma^*(x), x) = 0$$

By Lemma 7, $\gamma^*(x)$ is continuous. By Assumption 1, $h(\gamma)$ is continuous. Hence, $G(\gamma^*(x), x)$ is continuous too. Then there always exists an $x_0 > 0$ sufficiently small such that

$$G(\gamma^*(x_c), x_c) < G(\gamma^*(x_0), x_0) = G(\tilde{x})$$

Then for such an x_0, it must be true that $x_c < \tilde{x}$), since for any $x > \tilde{x}$, $G(\gamma^*(x), x))$ is increasing in x by the definition of $\tilde{x}$. □

Proof of Proposition 7

Proof. This is a direct result of Theorem 4. □

Proof of Proposition 8, 9, 10, 11

Proof. Proposition 8 and 10 are direct results of Theorem 3; Proposition 9 and 11 are direct results of Theorem 4. □

Proof of Lemma 3

Proof. Integrating by parts and (2.18) gives

$$\begin{aligned}
&\int_0^{F^{-1}(\tilde{x})} e^{-\rho t} G_x(\gamma^*(F(t)), F(t)) F'(t) dt \\
&= \rho \int_0^{F^{-1}(\tilde{x})} e^{-\rho t} G(\gamma^*(F(t)), F(t)) - G(\gamma^*(x_0), x_0) + e^{-\rho F^{-1}(\tilde{x})} G(\gamma^*(\tilde{x}), \tilde{x}) \\
&= \int_0^{F^{-1}(\tilde{x})} \rho e^{-\rho t} \left(G(\gamma^*(F(t)), F(t)) - G(\gamma^*(x_0), x_0) \right) dt
\end{aligned}$$

Since $G(\gamma^*(x_0), x_0) = G(\gamma^*(\tilde{x}), \tilde{x})$, by convexity of G, $G(\gamma^*(x), x) < \gamma^*(x_0), x_0)$ for all $x \in [x_0, \tilde{x}]$. Hence,

$$\int_0^{F^{-1}(\tilde{x})} e^{-\rho t} G_x(\gamma^*(F(t)), F(t)) F'(t) dt < 0$$

Then it is implied that

$$\int_0^{F^{-1}(\tilde{x})} e^{-\rho t} G(\gamma^*(F(t)), F(t)) + \frac{e^{-\rho F^{-1}(\tilde{x})}}{\rho} G(\gamma^*(\tilde{x}), \tilde{x}) < \frac{1}{\rho} G(\gamma^*(x_0), x_0)$$

Hence, $\tilde{x} < \tilde{x}_b$. □

Proof of Proposition 12 and Proposition 13

Proof. First, Proposition 1 still holds because it only requires no constraint on p_s, p_d. Lemma 7 holds too because it only requires g to be integrable. Hence,

$$G(\gamma, x) = (v-c)A\min\{\beta_1^s\gamma, \beta_1^d(1-\gamma)\}x - (\rho + \beta_0^s\gamma + \beta_0^d(1-\gamma))x$$
$$= \begin{cases} (((v-c)A\beta_1^s - (\beta_0^s - \beta_0^d))\gamma - (\rho + \beta_0^d))x, & \gamma < \frac{\beta_1^d}{\beta_1^d+\beta_1^s} \\ ((-(v-c)A\beta_1^d - (\beta_0^s - \beta_0^d))\gamma + (v-c)A\beta_1^d - (\rho+\beta_0^d))x, & \gamma \geqslant \frac{\beta_1^d}{\beta_1^d+\beta_1^s} \end{cases}$$

There are three cases:

(i) $\beta_0^s - \beta_0^d \geqslant (v-c)A\beta_1^s$. For a given $x > 0$, $G(\gamma, x)$ is decreasing in γ. $G(\gamma, x) \leqslant G(0, x) = -(\rho + \beta_0^d)x$. The optimal balance is then $\gamma^*(x) = 0$, and the optimal policy is to keep $x(t) = x_0$. The market is not viable.

(ii) $\beta_0^s - \beta_0^d \leqslant -(v-c)A\beta_1^d$. For a given $x > 0$, $G(\gamma, x)$ is increasing in γ. $G(\gamma, x) \leqslant G(1, x) = -(\rho + \beta_0^s)x$. The optimal balance is then $\gamma^*(x) = 1$, and the optimal policy is again to keep $x(t) = x_0$. The market is not viable.

(iii) $-(v-c)A\beta_1^d < \beta_0^s - \beta_0^d < (v-c)A\beta_1^s$ For a given $x > 0$, $G(\gamma, x)$is increasing in γ in the first piece and decreasing in γ in the second piece. Hence, $\gamma^*(x) = \frac{\beta_1^d}{\beta_1^d+\beta_1^s}$. $G\left(\frac{\beta_1^d}{\beta_1^d+\beta_1^s}, x\right) = \frac{(v-c)A\beta_1^s\beta_1^d - \beta_1^s(\rho+\beta_0^d) - \beta_1^d(\rho+\beta_0^s)}{\beta_1^s+\beta_1^d}x$. $h(\gamma^*(x)) = \frac{(v-c)A\beta_1^s\beta_1^d - \beta_1^s(\rho+\beta_0^d) - \beta_1^d(\rho+\beta_0^s)}{\beta_1^s+\beta_1^d}$.

If (2.28) holds, $h(\gamma^*(x)) > 0$, $G^*(\gamma^*(x), x)$ is increasing in x. Hence, faster growth is better than slower growth. Otherwise, $h(\gamma^*(x)) \leq 0$, $G^*(\gamma^*(x), x)$ is decreasing in x. Hence, it is optimal to keep $x(t) = x_0$. The market is not viable

□

Proof of Lemma 4

Proof. Since $\rho > 0$, (2.28) implies that $(v-c)A\beta_1^s\beta_1^d > \beta_0^s\beta_1^d + \beta_0^d\beta_1^s$. Then

$$\begin{aligned}&(v-c)A\beta_1^s\beta_1^d \\ &> \beta_0^s\beta_1^d + \beta_0^d\beta_1^s > \max\{\beta_0^s\beta_1^d, \beta_0^d\beta_1^s\} > \max\{(\beta_0^s - \beta_0^d)\beta_1^d, (\beta_0^d - \beta_0^s)\beta_1^s\} \quad \text{(A.38)}\end{aligned}$$

Dividing each term by $\beta_1^s\beta_1^d$ and rearranging the terms give (2.25). □

An illustrative example of supply and demand surpluses In the analytical model, we look at optimal balance as a measure for the market value of the supply relative to the total market value. Here we take a similar approach and calculate the ratio of the supply side surplus relative to the total surplus, which we call the *balance of surplus*. The formula is given by

$$\frac{(p_s - c)g(s,d)}{(p_s - c)g(s,d) + (v - p_d)g(s,d)} = \frac{\Delta(\gamma x)/\Delta t + \beta_0^s\gamma x}{\Delta x/\Delta t + \beta_0^s\gamma x + \beta_0^d(1-\gamma)x}$$

The above expression can be obtained by Proposition 30. Hence, if the trajectory of $x(t)$ is fixed, γ is also fixed at $\gamma^*(x(t))$. The trajectory of the balance of surplus can then be computed as a function of $x(t)$. Here we provide two examples.

For the increasing returns to scale market, we use the optimal market size trajectory under the high-subsidy policy $m = 1$ (shown in Figure 2.1). The trajectory of the balance of surplus is shown in Figure A.1. The balance of surplus is increasing before the market reaches its potential at $t = 4.5$ years. Compared with Figure 2.3, it shows that the balance of surplus is increasing as the market grows, similar to the optimal balance.

For the decreasing returns to scale market, we use the optimal market size trajectory under the high-subsidy policy $m = 1$ (shown in Figure 2.2). The balance of surplus is decreasing before the market reaches the saturation size at

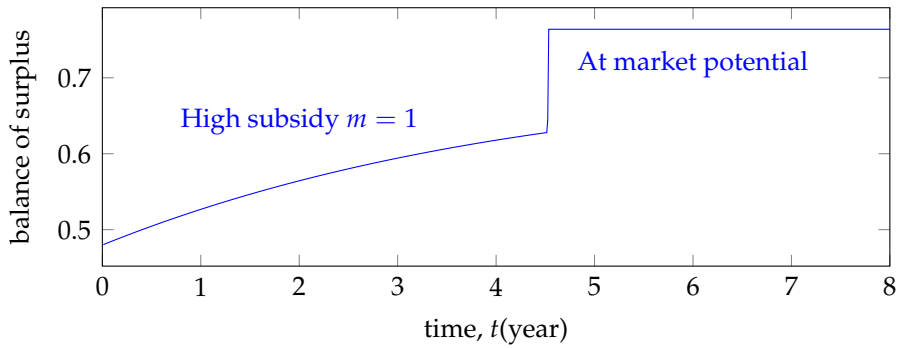


Figure A.1: Evolution of balance of surplus in the change of optimal market size trajectory (increasing returns)
Note. The market size trajectory $x(t)$ is the optimal policy under high subsidy $m = 1$

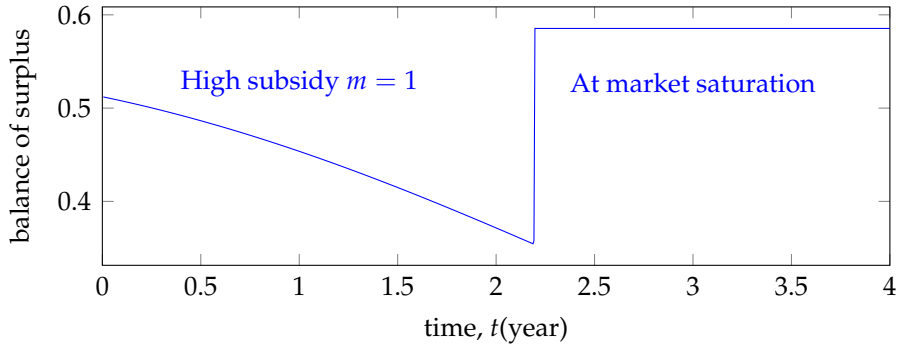


Figure A.2: Evolution of balance of surplus in the change of optimal market size trajectory (decreasing returns)

$t = 2.2$ years. It also shows a similar trend as as the optimal balance in Figure 2.4.

Hence, although it is intractable to analytically show the connection between the optimal balance and the balance of surplus, numerical examples suggest that a larger market balance implies that the total surplus of supply and demand is more concentrated on the supply side, and a smaller market balance implies that the total surplus of supply and demand is more concentrated on the demand side.

APPENDIX B

CAPTURING THE BENEFITS OF AUTONOMOUS VEHICLES IN RIDE-HAILING

The appendices can be divided into two parts: supplemental analysis and technical lemmas. In Appendix B.1, Appendix B.2 and Appendix B.3, we present additional analysis and illustrations for the pure-HV market, common platform market, and independent platform market, respectively. We also present proofs that are important and provide more general insights in these sections; for proofs that are more self-contained and with technical details, we present them in Appendix B.7. Appendix B.4 discusses conditions under which AVs and HVs will both operate. Appendix B.5 provides insights on the computation of surplus. Appendix B.6 talks about the choice of parameters for the numerical example in Fig. 3.3.

B.1 Pure-HV market

In this section, we provide the mathematical details for our benchmark case. The goal is to establish the stable equilibrium, and show comparative statistics around it.

B.1.1 Optimal level of open cars

We first introduce the formal definition of $\underline{n}(d)$, $\bar{d}(n)$ and $\bar{u}(d)$, which will be used in the rest of the appendix.

Lemma 9 *Given the ETA function $t_1(s) = as^{-r}$, for any demand level d, there exists a unique minimal supply level $\underline{n}(d)$, for which the following holds:*

1. *$\underline{n}(d)$ has a closed-form expression, given by*

$$\underline{n}(d) = \tilde{a}d^{\frac{1}{r+1}} + t_2 d$$

   where $\tilde{a} = r^{\frac{1}{r+1}}a^{\frac{1}{r+1}} + r^{-\frac{r}{r+1}}a^{\frac{1}{r+1}}$.

2. *$\underline{n}(d)$ has an inverse function, denoted as $\bar{d}(n)$. $\bar{d}(n)$ represents the highest demand rate that a fleet size of n can supply.*

3. *When using $\underline{n}(d)$ to serve the demand level d, the utilization of the HV fleet is maximized. Denote this utilization as $\bar{u}(d)$. Then it's expression is given by*

$$\bar{u}(d) = \frac{dt_2}{\underline{n}(d)} = \frac{dt_2}{\tilde{a}d^{1/(r+1)} + t_2 d}$$

B.1.2 Technical lemmas

We first introduce Lemma 10 that discuss the characteristics of utilization function $\bar{u}(d)$.

Lemma 10 (Utilization) *The maximum utilization $\bar{u}(d)$ defined in Lemma 9 is increasing and strictly concave in d. Moreover, $\bar{u}(0) = 0$, $\lim_{d\to\infty} \bar{u}(d) = 1$.*

In other words, with a higher level of demand rate, the fleet can have a higher utilization.

The next lemma is about the hourly earnings that a vehicle can expect to earn given the demand level and the market configuration. In particular, we define

a function $f(d;n) = p_i(d)\bar{u}(d - \bar{d}(n))$; when $n = 0$, $f(d;n)$ is the product of the market price $p_i(d)$ and the utilization $\bar{u}(d)$, which represents the per vehicle hourly earnings for a market with just one type of supply (HVs or AVs) to satisfy the demand rate d. It is also the per vehicle hourly earnings under a common platform market when AVs and HVs are supplying the demand together, because the utilization rate is shared among vehicles on the same dispatch platform.

When $n > 0$, it represents the per vehicle hourly earnings of one type of supply when AVs and HVs are operating on independent dispatch platforms. If n represents the AV fleet size, then $f(d;n)$ is the hourly earnings per HV when the demand rate is at d. In this case, as the AV fleet size n increases, the residual demand for HVs decreases, HVs thus have a lower utilization and a reduced hourly earnings.

Lemma 11 (Hourly earnings per vehicle) *Let n be a fleet size that satisfies $\bar{d}(n) < m_i$. For a given n, define function $f(d;n) = p_i(d)\bar{u}(d - \bar{d}(n))$, $\bar{d}(n) \leq d \leq m_i$, which represents the hourly earnings per vehicle at demand rate d when the residual demand for this type of vehicle is $(d - \bar{d}(n))$. Then $f(d;n)$ has the following properties:*

1. *Fixing n, $f(d;n)$ is strictly concave in d.*
2. *Fixing n, $f(d;n)$ is unimodal with a maximizer $d^\dagger$ that can be uniquely determined by the first-order condition $f'(d^\dagger;n) = 0$.*
3. *Define $g(n) = \max_{\bar{d}(n) \leq d \leq m_i} f(d;n)$, which represents the highest per vehicle hourly earnings for HVs (AVs) when n AVs (HVs) are operating on an independent dispatch platform. Moreover, $g(n)$ is continuous and decreasing in n; $g(0) > w_0$ and $g(\underline{n}(m_i)) = 0$.*

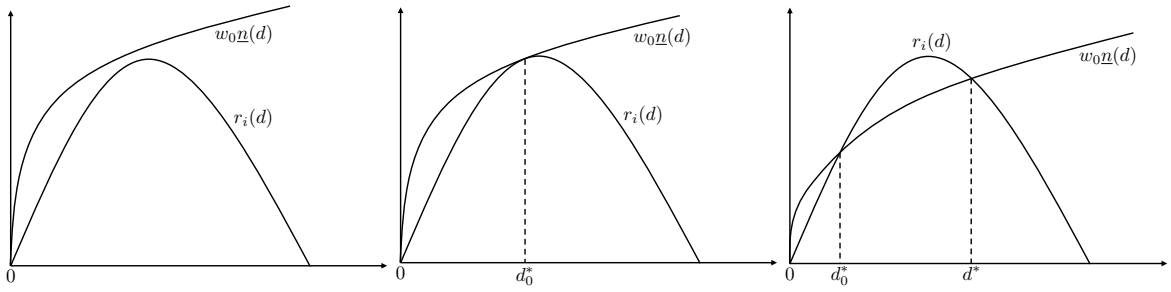


Figure B.1: Illustration of all possible equilibria

Note. Left to right: Case (a), (b), (c). In (a), the only equilibrium demand rate is $d = 0$. This is the case when the market is too "thin" to break even. It can be easily checked that $d = 0$ is a stable equilibrium because when the demand is slightly above 0, the cost is greater than the revenue, which causes the drivers to exit the market and moves the demand back to 0. In (b), there are two equilbrium point, $d = 0$ and $d = d_0^*$. One can check that both equilibria are stable. In (c), there are three equilibrium points. However, $d = d_0^*$ is not a stable equilibrium because the market will shift towards d^* when d slightly goes above d_0^* and towards 0 when d slightly goes below d_0^*.

B.1.3 Stable equilibrium (Proof of Proposition 14)

With the preparation of Appendix B.1.2, we prove results for the stable equilibrium in Proposition 14. We focus on proving the existence for the maximal stable equilibrium demand d^*; once d^* is determined, it is easy to derive the expressions for the rest variables.

Existence: As stated in Section 3.4, the equilibrium points are determined by

$$r_i(d) = w_0\underline{n}(d) \tag{B.1}$$

The possible relationships between $r_i(d)$ and $w_0\underline{n}(d)$ are illustrated by Fig. B.1. To see why, note that $d = 0$ is always a solution to Eq. (B.1) since $\underline{n}(0) = 0$. For $d > 0$, Eq. (B.1) is equivalent to

$$p_i(d)\bar{u}(d) = w_0$$

This is obtained by dividing $\underline{n}(d)$ on both sides. In other words, the equilibrium point is where the hourly earnings per HV exactly equals to w_0. By Lemma 11, we know that the function $p_i(d)\bar{u}(d)$ is unimodal and smooth; moreover, at both ends of $d = 0$ and $d = m_i$, $p_i(0)\bar{u}(0) = 0$, $p_i(m_i)\bar{u}(m_i) = 0$. Therefore, there are only three possibilities: (1) $\max_d p_i(d)\bar{u}(d) < w_0$, meaning HVs can never make w_0. This is case (a) in Fig. B.1; (2) $\max_d p_i(d)\bar{u}(d) = w_0$, meaning that there is only one demand rate at which HVs can make w_0; other than that, the cost exceeds the revenue. This is case (b) in Fig. B.1. (3)$\max_d p_i(d)\bar{u}(d) > w_0$, meaning that there are two demand rates at which HVs can make w_0; moreover, in the region between the two demand rates, the revenue is strictly higher than the cost. This is case (c) in Fig. B.1.

Combining all three cases, we conclude that there always exists at least one stable equilibrium. Moreover, the maximal stable equilibrium is given by

$$d_i^* = \max\{d : r_i(d) = w_0\underline{n}(d)\}$$

i.e., when $\max_d p_i\bar{u}(d) < w_0$, the maximal stable equilibrium is just $d^* = 0$; when $\max_d p_i\bar{u}(d) \geq w_0$, the maximal stable equilibrium is the largest d that satisfies the equilibrium condition Eq. (B.1).

The monotonicity for n_i^*, u_i^* and p_i^* can be verified by checking Lemma 9, Lemma 10, and the definition for $p_i(d)$, respectively.

Comparative statics: When m_i increases, the revenue curve $r_i(d)$ moves upward because at the same d, the market price $p_i(d)$ is now higher as a result of the increased potential demand. Thus, the maximal stable equilibrium demand d^* increases (as shown in Fig. B.2).

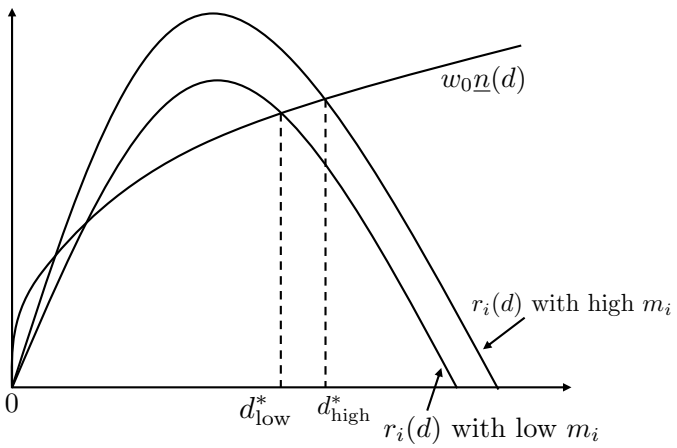


Figure B.2: Illustration of equilibrium demand rates at different potential demand m_i

B.1.4 Extension of the equilibrium concept

As a preparation for the analysis for markets with AVs, we extend the equilibrium concept in Definition 6 to include common and independent platforms between AVs and HVs. We use $R_i(d_0)$ to represent the total HV earnings (revenue) per hour, where d_0 is the demand supplied by HVs. We use $C(d_0)$ denote the total HV costs per hour. Then in equilibrium, by the perfect elasticity assumption, HVs just make enough earnings to cover their costs. More precisely, we have the following definition:

Definition 7 (Equilibrium) *A market is in equilibrium if $R_i(d_0) = C(d_0)$, where d_0 is the demand supplied HVs.*

Furthermore, the definition of a stable equilibrium is the following:

Definition 8 (Stable equilibrium with AVs) *A market is not stable if $R_i(d_0 - \epsilon) < C(d_0 - \epsilon$ and $R_i(d_0 + \epsilon) > C(d_0 + \epsilon)$ for a small disturbance $\epsilon > 0$. If not, then the market is in a stable equilibrium.*

The revenue and cost functions will be formally proved in Appendix B.2 and Appendix B.3, but for the completeness of the definition, we provide their expressions here. Consider a market where AVs serve a demand level of d_a and HVs serve a demand level of d_0. Under an independent platform market,

$$R_i(d_0) = p_i(d_0 + d_a)d_0 t_2,\ C(d_0) = w_0 \underline{n}(d_0)$$

Under a common platform market,

$$R_i(d_0) = p_i(d_0 + d_a)d_0 t_2,\ C(d_0) = \frac{d_0}{d_a + d_0} w_0 \underline{n}(d_a + d_0)$$

Similar to the pure-HV market, we call the stable equilibrium with the largest HV demand level d_0 as the *maximal stable equilibrium*. In the main body, for brevity, we use the term "equilibrium", "stable equilibrium", and "maximal stable equilibrium" interchangeably unless otherwise stated.

B.2 Common platform market

Since Proposition 15 and Theorem 6 are just summaries of Section 3.6.1 and Section 3.6.2, next we start by analyzing the common platform market. We look at the price, total fleet size and total demand, taking the AV fleet size as exogenously given.

B.2.1 Exogenous AV fleet size (Proof of Proposition 16)

There can only be two cases:

Case 1 (coexisting AV-HV): $n_a < n_i^*$. When the number of operating AVs is small, AVs and HVs jointly supply the market. In this case, there is an equilibrium with a positive number of HVs participating, i.e. $n_0 > 0$. By dispatch symmetry, HVs must be receiving a fraction $\alpha = n_0/(n_a + n_0)$ of all dispatches and must constitute a fraction α of all open cars. Thus, the equilibrium HV fleet size is determined by

$$\frac{w_0}{p_i(d)} = \frac{\alpha d t_2}{\alpha s + \alpha d(t_1(s) + t_2)} = \frac{d t_2}{s + d(t_1(s) + t_2)}$$

But this is the same equilibrium condition (3.3) as in the pure-HV case. Therefore, if there is participation in the market by HVs, we must have the same equilibrium total number of vehicles in the market and the same equilibrium price, demand rate, and utilization given by Proposition 14.

The AV variable profit per vehicle per hour is therefore

$$p_i(d_i^*)\bar{u}(d_i^*) - c_v = w_0 - c_v$$

Since this case can only happen when a positive number of HVs are operating, the AV fleet size need to be strictly less than the pure-HV equilibrium fleet size, n_i^*.

Case 2 (pure AV): $n_a \geq n_i^*$. When AVs operate at a fleet size higher than the pure-HV equilibrium size, the price is lower than p_i^* and can no longer support HVs' outside earnings w_0. The market outcomes thus solely depend on the number of AVs that are operating.

This suggests n_i^* is the highest fleet size that allows AVs to operate together with HVs in a common platform market. □

As a summary, in scenario i, for a common platform market, the variable profit per AV is given by

$$\pi_i^C(n_a) = \begin{cases} w_0 - c_v, & n_a \leq n_i^* \\ p_i(\bar{d}(n_a))\bar{u}(\bar{d}(n_a)) - c_v, & n_a > n_i^* \end{cases} \tag{B.2}$$

B.2.2 Optimal AV fleet size for a monopoly supplier (Proof of Proposition 17)

In this section, we discuss the AV fleet size n_a that maximizes the monopoly supplier's variable profit (i.e. $\arg\max_{n_a \leq N}\{\pi_i^C(n_a)\ n_a\}$). We divide the discussion by two cases:

Case 1: $N \leq n_i^*$**.** In scenario i, if the AV capacity N is at or below the equilibrium demand n_i^*, then at any fleet size, each AV makes a flat rate $(w_0 - c_v)$ (Case 1 in Appendix B.2.1); thus, it is optimal for the monopoly supplier to put as many AVs on road as possible; Because N does not exceed n_i^*, there will be leftover demand for HVs, and HVs will participate until the point where the total fleet size from HVs and AVs is exactly at n_i^*.

Case 2: $N > n_i^*$**.** If the AV capacity N is above n_i^*, then the monopoly supplier faces a trade-off: by extending the supply above n_i^*, it increases the demand and the fleet utilization, but reduces the market price.

We found that, in a loose market (i.e. the HV equilibrium price is below the revenue-maximizing price), the following lemma provides an answer to the trade-off:

Lemma 12 *In a loose market, in scenario i, a monopoly AV supplier cannot improve its profit by extending the supply beyond the pure-HV equilibrium demand d_i^*, i.e.*

$$d_i^* = \underset{d_i^* \le d \le \bar{d}(N)}{\arg\max} \{p_i(d)d - c_v \underline{n}(d)\}$$

For the monopoly supplier, the basic trade-off is whether to serve a larger market, or to limit the demand and keep a high price. Given that the pure-HV equilibrium is the outcome of perfectly competitive HVs, at d_i^*, the demand is high enough that the marginal value of further increasing demand does not offset the reduction in price (Assumption 5). Therefore, the monopoly supplier does not benefit from further extending the demand above d_i^* (or equivalently, the AV fleet size above n_i^*).

Thus, regardless of whether N is above or below n_i^*, it is always optimal for the supplier to deploy up to n_i^* of its AV capacity. Moreover, the equilibrium utilization, demand rate, and price is the same as u_i^* and d_i^*, p_i^*, respectively, given that the total fleet size and dispatch does not change after introducing AVs.

B.2.3 Optimal AV capacity for a monopoly supplier (Proof of Proposition 18)

In this section, we discuss the optimal choice of the capacity N.

We first show that the optimal fleet size must be one of the equilibrium fleet size, i.e. $N^* = n_i^*, i \in \{1, ..., I\}$: for any $n_1^* \le n \le n_I^*$, there must exist a K such that $1 \le K \le I-1$, $n^*(m_K, w_0) \le n \le n_{K+1}^*$. Then (3.4) can be written as

$$(w_0 - c_v) \sum_{1}^{K} P(m_i) n_i^* + (w_0 - c_v) \sum_{K+1}^{I} P(m_i) n - c_f n \tag{B.3}$$

Since (B.3) is linear in n, it is maximized at the endpoints of the interval, i.e. either n_K^* or n_{K+1}^*. Moreover, it can be easily verified that any $n < n_1^*$ or $n > n_I^*$ is strictly dominated. Hence, the optimal fleet size N^* must be one of the equilibrium size.

Then the optimization problem is equivalent to finding the optimal K. For n_K^* to be the optimal fleet size, K must be the first i such that

$$\Pi(n_{K+1}^*) - \Pi(n_K^*) < 0$$

where

$$\Pi(n_{K+1}^*) - \Pi(n_K^*) = \Big((w_0 - c_v)\sum_{K+1}^{I} P(m_i) - c_f\Big)\Big(n_{K+1}^* - n_K^*\Big) \tag{B.4}$$

Since n_i^* is increasing in m_i and thus increasing in I, $n_{K+1}^* > n_K^*$. Then the above inequality is equivalent to

$$\sum_{K+1}^{I} P(m_i) < \frac{c_f}{w_0 - c_v}$$

If the above condition has no solution, implying $\Pi(n_{K+1}^*) - \Pi(n_K^*) >= 0$ for all K, then the optimal K is just the largest index, I.

B.2.4 Equilibrium price for perfectly competitive AVs (Proof of Proposition 19)

When AVs are perfectly competitive with each other, they will end up all making zero profit. A potential AV supplier will then be indifferent about whether to purchase an AV or not. The equilibrium capacity under this scenario can be obtained by solving condition Eq. (3.2) using the profit function $\pi_i^C(n)$ given by Eq. (B.2).

In the perfect competition case, for scenarios where the AV capacity $N \leq n_i^*$, HVs enter the market and the equilibrium price is just identical to the pure HV equilibrium. For scenarios where $N > n_i^*$, suppose the equilibrium price remains to be p_i^*. This implies there are n_i^* AVs operating in the market and each AV earns a positive variable profit $(w_0 - c_v)$ while $(N - n_i^*)$ AVs are idle. This cannot be true under perfect competition, since the rest AVs will join the market until the variable profit is 0. Thus, the price will always be lower than p_i^*.

B.3 Independent platform market

In this section, we discuss the market configuration with AVs and HVs operating on independent platforms. Similar to Appendix B.2, we start by analyzing the market for an exogenously given AV fleet size.

B.3.1 Exogenous AV fleet size

Consider scenario i. Suppose HVs serve a demand level of d_0 and AVs serve a demand level of d_a. The total ride-hours supplied by AVs and HVs is just $(d_0 + d_a)$. The market price is then a function of d_a and d_0, given by

$$p = p_i(d_0 + d_a) \tag{B.5}$$

Thus, the total HV revenue is given by

$$R_i(d_0) = p_i(d_0 + d_a)d_0 t_2$$

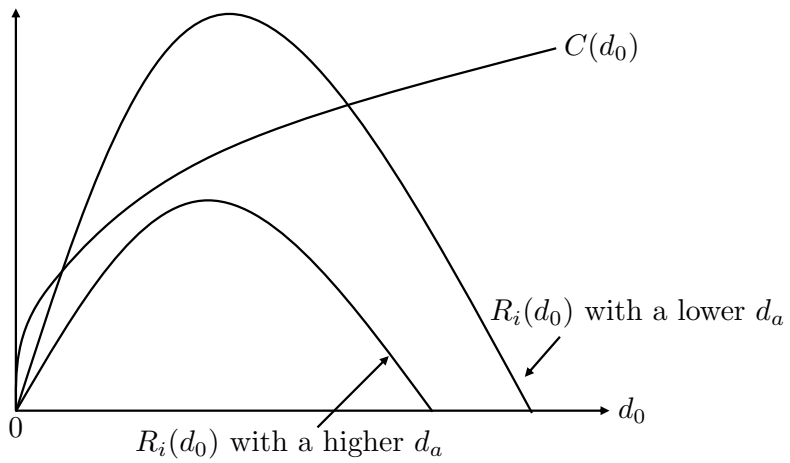


Figure B.3: Equilibrium in an independent platform market

Moreover, the number of HVs required to supply d_0 is just $\underline{n}(d_0)$. Thus, the total HV cost is given by

$$C(d_0) = w_0 \underline{n}(d_0)$$

By Definition 7, in equilibrium, it must hold that

$$R_i(d_0) = C(d_0) \Leftrightarrow p_i(d_0 + d_a)d_0 t_2 = w_0 \underline{n}(d_0) \tag{B.6}$$

Fig. B.3 illustrates the equilibria at different AV fleet sizes. All else being equal, the HV revenue $R_i(d_0)$ strictly decrease as the demand served by AVs d_a increases. When the AV fleet size is sufficiently large (as in the "$R_i(d_0)$ with a higher d_a" case), the residual demand for HVs is so low that the revenue no longer covers the cost. In this case, the only equilibrium is $d_0 = 0$, meaning that AVs serve the market alone.

Thus, we use define a threshold, $d_i^\dagger$, to denote the highest level of demand served by AVs that allows $R_i(d_0)$ to intersect with $C(d_0)$ at $d_0 > 0$. In other words, $d_i^\dagger$ is the AV demand level such that

$$\max_{d_0} \{R_i(d_0|d_i^\dagger) - C(d_0)\} = 0 \tag{B.7}$$

Denote the corresponding AV fleet size as $n_i^\dagger$, i.e.

$$n_i^\dagger = \underline{n}(d_i^\dagger) \tag{B.8}$$

Lemma 13 formally establishes how the threshold $n_i^\dagger$ impacts the HV participation and the monotonicity of $n_i^\dagger$ in m_i:

Lemma 13 *Consider $n_i^\dagger$ defined in equation (B.8). Then in an independent platform market, there is positive HV participation if and only if the AV fleet size n_a is no greater than $n_i^\dagger$; that is, Eq.* (B.6) *has a positive solution for d_0 only under the condition $n_a \leq n_i^\dagger$. Moreover, $n_i^\dagger$ is increasing in m_i.*

Next, the following proposition summarizes the market outcomes for an independent platform market:

Proposition 32 *Consider scenario i in an independent platform market. Given an AV fleet size n_a, in equilibrium, there are only two possibilities:*

1. *AVs and HVs serve the market together. The equilibrium price is given by*

$$p^* = \frac{w_0}{\bar{u}(d_0^*)}, \tag{B.9}$$

   where d_0^* is equilibrium HV demand rate determined by Eq. (B.6)*. Moreover, the equilibrium price p^* is strictly increasing in n_a. This case happens when $n_a \leq n_i^\dagger$.*

2. *AVs can serve the entire market alone. The equilibrium price is given by $p_i(\bar{d}(n_a))$. This happens when $n_a > n_i^\dagger$.*

The fact that the coexisting equilibrium price is increasing in the AV fleet size n_a is closely related to the density effect we have discussed in the common

platform market section. As AVs take away part of the demand from HVs, the average rider density for HVs goes down. Thus, HVs will have to spend more time picking up riders, and to make the same earnings, the equilibrium price must go up to compensate for the resulting utilization loss. This utilization loss for HVs will be further exacerbated as more AVs join, leading to an even higher equilibrium price.

When $n_a = 0$, the equilibrium price given by Proposition 32 converges to the equilibrium price in the pure HV market. Therefore, an immediate result regarding the comparison between the coexisting regime and the pure HV regime follows:

Corollary 4 *For an independent platform market, if* $0 < n_a < n_i^\dagger$*, the equilibrium price is strictly higher than that in a pure HV market.*

In other words, utilization loss for HVs is caused by introducing AVs in an independent platform market design. Hence, the presence of AVs means riders have to pay higher prices and face longer waiting times for pickup, both of which reduce rider welfare.

Furthermore, there are two things worth noticing in Proposition 32 that have interesting contrasts with the common platform market result in Proposition 16, which we discuss separately in the following paragraphs.

Nonviable AVs The first contrast is that, for an AV fleet size that falls under Case 1 of Proposition 32, even though it is true that HVs can break-even, there is no guarantee that AVs can break-even too. In other words, in later stages when the AV supplier(s) decide the fleet size, it is possible that they would rather not

operate in some scenarios. This can never happen in a common platform market. More precisely, we have the following lemma:

Lemma 14 *Under an independent platform market, for each scenario i, there exists a threshold n_i^l, such that when the AV fleet size $n_a < n_i^l$, the AV variable profit $\pi_i^I(n_a) < 0$.*

In other words, if the AV fleet size is lower than threshold n_i^l, the AV's utilization is so low that its variable profit is lower than c_v. The threshold n_i^l can be computed by jointly solving:

$$p_i(\bar{d}(n_i^l) + d_0)\bar{u}(\bar{d}(n_0)) = w_0 \tag{B.10}$$

$$p_i(\bar{d}(n_i^l) + d_0)\bar{u}(\bar{d}(n_i^l)) = c_v \tag{B.11}$$

That is, at threshold n_i^l, AVs and HVs operate together and each AV earns exactly c_v per operating-hour. The following lemma establishes the existence of n_i^l and how it changes monotonically with the potential demand m_i:

Lemma 15 *Let $d_x(n) = \max\{d | p_i(d)\bar{u}(d - \bar{d}(n)) = w_0\}$. $d_x(n)$ is the equilibrium demand at an AV fleet size n, in scenario i. Define $h(n) = \frac{\bar{u}(\bar{d}(n))}{\bar{u}(d_x(n) - \bar{d}(n))}$. $h(n)$ is the ratio of the equilibrium utilization between AVs and HVs at the AV fleet size n.*

1. *When $h(n_i^\dagger) \geq c_v/w_0$, n_i^l exists and is strictly increasing in m_i.*
2. *When $h(n_i^\dagger) < c_v/w_0$, n_i^l does not exist and AVs cannot make positive variable profits while coexisting with HVs.*

This implies that at a fixed fleet size, AVs may be able to break-even in a scenario with lower potential demand but becomes nonviable as the potential demand

gets higher. The intuition is that, under an independent platform market, a smaller scale means disadvantage in utilization. This may give HVs, which have unlimited capacity providing sufficient earnings, an advantage over AVs which are capacity constrained. When the AV is operating at a low fleet size, the resulting low utilization can make AVs less competitive than HVs, even though AVs have lower costs.

Regime boundary The second contrast is that, just like how $n_i^\dagger$ in Proposition 32 divides the two regimes of the coexisting AV-HV and pure-AV, in a common platform market, there is a similar threshold, which equals to n_i^*. Interestingly, we find that this threshold n_i^* is higher than that in an independent platform market, $n_i^\dagger$, for all scenarios. In other words, we have the following lemma:

Lemma 16 *In any scenario, given the same AV fleet size, if HVs cannot break-even under a common platform market, then they also cannot break-even under an independent platform market. However, the opposite does not hold true. That is,* $n_i^\dagger < n_i^*$ *for all* i.

This suggests that, it is more challenging for HVs to coexist with AVs in an independent platform market than in a common platform market (providing that AVs are operating) in any scenario. This confirms our intuition that an independent platform market tends to lead to more extreme outcomes; a monopoly AV can drive HVs out and become the sole supplier of the transportation service at a smaller capacity.

As a summary of the above analysis, in an independent platform market, when a fleet size of n AVs are providing service, the variable profit rate per AV is

given by:

$$\pi_i^I(n) = \begin{cases} p_i(\bar{d}(n) + d_0)\bar{u}(\bar{d}(n)) - c_v, & n \leq n_i^\dagger \\ p_i(\bar{d}(n))\bar{u}(\bar{d}(n)) - c_v, & n > n_i^\dagger \end{cases} \tag{B.12}$$

where d_0 is the equilibrium demand rate given by Eq. (B.6) and $n_i^\dagger$ is given by (B.8). When $n < n_i^l$, $\pi_i^I(n) < 0$, and hence the AV supplier(s) will never have the incentive to choose an AV fleet size lower than n_i^l in scenario i.

B.3.2 Equilibrium with finitely elastic HVs

For the independent platform market, we compare two settings: (1) The HV supply is perfectly elastic with reservation earnings w_0. This is the setting discussed in our main model (Section 3.6.2) (2) The HV supply is finitely elastic with supply curve $w(n_0)$, where n_0 is the HV fleet size.

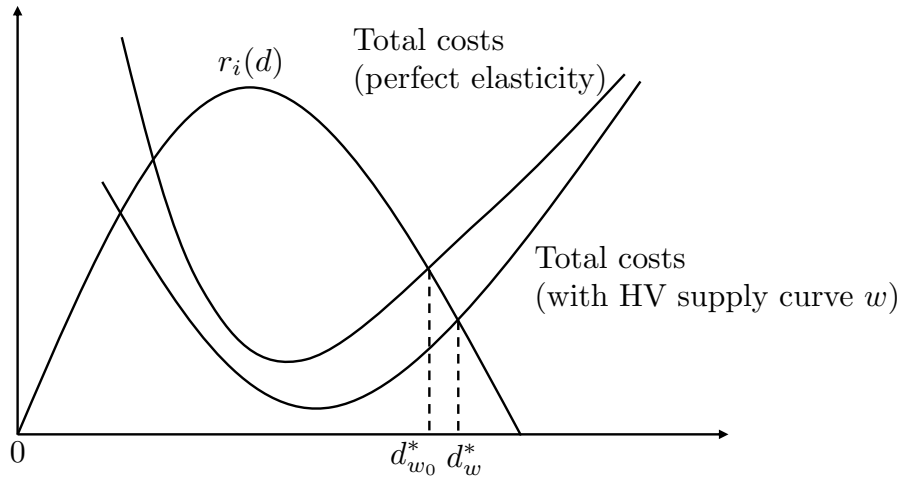


Figure B.4: Equilibrium with finitely elastic HVs, independent platform market

Fig. B.4 illustrates how the equilibrium point shifts after the elasticity changes from (1) to (2). For the sake of comparison, we consider supply curve $w(n_0)$ to

be equal to w_0 when n_0 is at the equilibrium HV fleet size in (1); in other words, $w(n_0) = w_0$ when $\bar{d}(n_0) + \bar{d}(n_a) = d^*_{w_0}$. The revenue curve $r_i(d)$ is the same as in Section 3.6.2. The cost curve, "Total costs (perfect elasticity)" represents the total hourly earnings that AVs and HVs expect to receive, when the HV supply is perfectly elastic at reservation earnings w_0, as described by setting (1); "Total costs (with HV supply curve w)" represents the total hourly earnings that AVs and HVs expect to receive, when HVs have a supply curve $w(n_0)$, as described by setting (2). For both curves, the AV fleet size is fixed at n_a.

The short conclusion is that, similar to the common platform market case in Section 3.7.1, relaxing the perfect elasticity assumption leads to additional savings brought by AVs, but this is driven by the fact that HVs lower their earnings standard. To see why, the equilibrium demand under supply curve $w(n_0)$, which is denoted as d^*_w, is higher than $d^*_{w_0}$, implying that the price is lower and consumers are better off. Moreover, at the equilibrium point d^*_w, HVs' hourly earnings are strictly lower than the hourly earnings w_0 at the equilibrium point $d^*_{w_0}$.

Below, we provide a proof sketch that shows the shape of the curves and explains why $d_{w_0} < d_w$. When the HV supply is perfectly elastic (Setting 1), given the AV fleet size n_a, the total costs can be represented by

$$w_0 n_0 + w_0 \frac{\bar{u}(\bar{d}(n_a))}{\bar{u}(\bar{d}(n_0))} n_a, \text{ where } n_0 = \underline{n}(d) - n_a \tag{B.13}$$

The first term is the cost for HVs and the second term is the cost for AVs. Note that similar to Section 3.7.1, the cost here is not the variable or fixed cost for AVs; it is the total earnings that AVs make under the market condition, which can be thought as a cost for the ride-hailing platform to pay the AVs. One can see that hourly earnings for an AV is the HV's reservation earnings w_0 discounted

by the ratio of their utilization. This is because they always face the same price when they are both operating, and whoever has a higher utilization would make a higher hourly earnings.

When the HV supply is finitely elastic (Setting 2), given the AV fleet size n_a, the total costs can be represented by

$$w(n_0)n_0 + w(n_0)\frac{\bar{u}(\bar{d}(n_a))}{\bar{u}(\bar{d}(n_0))}n_a, \text{ where } n_0 = \underline{n}(d) - n_a \tag{B.14}$$

Everything is the same as Eq. (B.13), except that the HV hourly earnings are $w(n_0)$.

By comparing the two expressions, it can be verified that Eq. (B.13) is strictly greater than Eq. (B.14) when d is at $d^*_{w_0}$. The two curves intersect at $d = \bar{d}(\underline{n}(d^*_{w_0}) + n_a)$; after that, as d further increases, Eq. (B.14) becomes greater than Eq. (B.13). Such a single crossing property verifies that $d^*_{w_0} < d^*_w$.

B.4 Conditions for AVs and HVs to coexist

Here we give sufficient conditions under which HVs will participate in at least one scenario for each market configuration. The condition is an upper bound on the probability mass of the highest demand scenario, i.e. $P(m_I)$. The intuition is that, when the demand distribution has spikes with high mass and low probability of occurrence, it is not economical for the AV supplier(s) to have an AV capacity that is so high that it can even supply the peak demand alone. What defines a low probability and a high demand mass in each configuration is shown below in Table B.1.

In the monopoly, common platform market, regardless of how high the

Table B.1: Conditions for HV to participate in at least one scenario

	Common Platform	**Independent Platform**
Monopoly	$P(m_I) < \frac{c_f}{w_0 - c_v}$	$P(m_I) < \frac{c_f(1-\max\{\tilde{n}_{I-1}, n^\dagger_{I-1}\}/n^\dagger_I)}{p_I(d^\dagger_I)\bar{u}(d^\dagger_I)-c_v}$ and $n^\dagger_I > \tilde{n}_I$
Competitive	$P(m_I) < \frac{c_f - \sum_1^{I-1} P(m_i)\left(p_i(d^*_I)\bar{u}(d^*_I)-c_v\right)^+}{w_0-c_v}$	$P(m_I) < \frac{c_f - \sum_{i=1}^{I-1} P(m_i)\left(p_i(d^\dagger_I).\bar{u}(d^\dagger_I)-c_v\right)^+}{p_I(d^\dagger_I)\bar{u}(d^\dagger_I)-c_v}$ and $n^\dagger_I > \tilde{n}_I$

Note. The condition for the monopoly, common platform market is sufficient and necessary; conditions for the rest three cases are sufficient. Definition for $\tilde{d}_i$: $\tilde{d}_i = \arg\max\{p_i(d)dt_2 - c_v\underline{n}(d)\}$, which represents the demand rate that maximizes the total variable profit rate $(p_i(d)dt_2 - c_v\underline{n}(d))$ in scenario i. $\tilde{n}_i = \underline{n}(\tilde{d}_i)$ which is the corresponding fleet size that can supply demand level $\tilde{d}_i$. Recall (d^*_i, n^*_i) represents the pure HV equilibrium demand and fleet size and are also the demand and AV fleet size above which AVs will serve the market alone in scenario i under a common platform market; similarly, $(d^\dagger_i, n^\dagger_i)$ are the threshold demand and AV fleet size that allows AVs to operate alone under an independent platform market.

potential demand mass is, as long as $P(m_I)$ is below the threshold, AVs will share the market with HVs in scenario i. In the rest three cases, the demand mass m_I also plays a role. Furthermore, under a common platform market, the threshold for the competitive case is lower than that for the monopoly case. In other words, all else being equal, AVs are more likely to serve the market alone when they are perfectly competitive in a common platform market; The difference between the two thresholds, $\sum_1^{I-1} P(m_i)\left(p_i(d^*_I)\bar{u}(d^*_I) - c_v\right)^+$ (normalized by $(w_0 - c_v)$), is the aggregate variable profit an AV extracts from all but the highest demand scenario under the competitive setting. In other words, AVs will participate and squeeze out HVs as long as they can still break-even; in contrast, the monopoly supplier will strategically coexist with HVs to maximize its total profit.

B.5 Surplus calculation

Rider surplus Rider surplus is determined by riders' demand curve $d(p)$, and the market price p, in the equilibrium. In each scenario i, for a given price p_i, the rider's surplus is given by

$$\int_{p_i}^{V} d(p)pdp$$

where V is the price that demand will fall to 0. In this model, we assume a linear demand curve; hence, the rider surplus can be further simplified as

$$\int_{p_i}^{V} d(p)pdp = \frac{1}{2}d(p_i)(V - p_i) = \frac{m_i}{2V}(V - p_i)^2 \tag{B.15}$$

The expected rider surplus is then

$$\sum_{i=1}^{I} \mathbb{P}(m_i)\frac{m_i}{2V}(V - p_i)^2 \tag{B.16}$$

Driver surplus Driver surplus is calculated under the assumption of the HV supply curve $w(n)$, where n represents the HV fleet size and $w(n)$ represents the hourly earnings required for n HVs to participate in ride-hailing. Then in a market with the HV equilibrium fleet size at n^*, driver surplus is given by

$$\int_{0}^{n^*} w(n^*) - w(n)\ dn$$

That is to say, driver surplus is the area between the supply curve and the HV equilibrium hourly earnings, for drivers who are willing to work at or below $w(n^*)$. Clearly, driver surplus is increasing in the equilibrium HV fleet size n^* because a larger n^* means both an increased number of drivers who get employed as well as higher earnings for every participating driver. Thus, when comparing the driver surplus in two settings, it is equivalent to comparing the equilibrium HV fleet size.

Next, we provide a formal proof to Proposition 23 on the implications of AVs on the HV fleet under different market configurations.

B.6 Choice of parameters in the numerical analysis

Table B.2 shows how parameters are chosen in Table 3.3.

Table B.2: Data and rationale behind calculations

	Data source and rationales behind calculations
a	$\approx 3(\text{area})^{1/2}$ (Kolesar 1975), area=300 sq ml (NYC, Population 2022)
t_2	25 min
r	Kolesar 1975
w_0	GOBankingRates 2022
c_v	Fuel and maintenance cost per hour
c_f	Straight-line depreciation of \$180k total cost over 3 years.
V	Average taxi trip price in the U.S. (Appendix B.6.1)

B.6.1 Riders' maximum valuation of a trip

The choice of parameter V in Table 3.3, Section 3.5.1 is supported by the taxi price from major U.S. cities. According to our calculation, the average price among the cities is around \$101.85/hour. We round the number to \$100/hour and use it as a proxy for the valuation of ride-hailing service for an average rider, which corresponds to $V/2$ in our model because we assume a uniform distribution in $[0, V]$ for riders' trip valuation. Thus, in Table 3.3, we set $V = \$200$/ride-hour.

Table B.3 provides the relevant data for calculating the taxi price. For each city, Column "Price" represents the average price per hour charged by taxis, which is

Table B.3: Average taxi price in major U.S. cities

City	Price ($/hour)	Initial charge ($)	Per mile charge ($)	Avg. trip distance (miles)	Avg. travel speed (miles/hour)
NYC	86.67	2.5	2.5	3.0	26.0
San Francisco	84.66	3.5	2.75	5.5	25.0
Boston	88.16	2.6	2.8	4.4	26.0
Chicago	82.39	3.25	2.25	5.5	29.0
Washington DC	59.64	3.25	2.16	5.9	22.0
Los Angeles	118.76	2.85	2.7	6.7	38.0
San Diego	166.06	2.8	3	7.2	49.0
Dallas	100.59	2.25	1.8	8.9	49.0
Phoenix	129.77	5	2.3	7.7	44.0
Average	101.85				

Note. Data source: Initial charge and per mile change – Taxi fares (2015), `https://www.taxifarefinder.com/rates.php`; avg. trip distance – Rideguru(2018), `https://ride.guru/lounge/p/what-is-the-average-trip-distance-for-an-uber-or-lyft-ride`; avg. travel speed – Geotab, `https://www.geotab.com/`.

computed using data from Column "Initial Charge" to Column "Average travel speed". The detailed calculation of the taxi price in Table B.3 is the following:

Avg. price per trip = Initial charge + per mile charge $\times$ avg. trip distance

Avg. duration per trip = avg. trip distance/avg. travel speed

Price = Avg. price per trip/Avg. duration per trip

Thus, we obtain an estimation of the average taxi price per hour in Table B.3, Column "Price". By taking an average over all cities, we obtain the number \$101.85/hour during the trip.

B.7 Proofs for technical lemmas and propositions

Proof of Lemma 9

Proof. We prove each item one by one.

Expression for $\underline{n}(d)$: For a given d, let $n(s) = s + d(t_1(s) + t_2)$. Then by definition, $\underline{n}(d) = \min_s n(s)$, where $t_1(s) = as^{-r}$. Taking the derivative of $n(s)$ gives $n'(s) = 1 - (rad)s^{-r-1}$, which implies $s^* = \arg\min_s n(s) = (ra)^{\frac{1}{r+1}} d^{\frac{1}{r+1}}$. Plugging in s^* to $n(s)$ gives the expression for $\underline{n}(d)$

Inverse function $\bar{d}(n)$: One can check that $\underline{n}(d)$ is strictly increasing in d. Thus, by the inverse function theorem, its inverse function $\bar{d}(n)$ exists and is strictly increasing in n.

Maximum utilization $\bar{u}(d)$: The maximum utilization is achieved when demand is satisfied by the minimum number of cars. Therefore, $\underline{u}(d)$ is equal to the ratio of dt_2 and $\underline{n}(d)$, which is the highest possible proportion of time spent on-trip for demand rate d. □

Proof of Lemma 10

Proof. By Lemma 9,

$$\bar{u}(d) = \frac{dt_2}{\tilde{a}d^{1/(r+1)} + t_2 d} = \frac{1}{\tilde{a}d^{-r/(r+1)} + 1},$$

where $\tilde{a} = r^{\frac{1}{r+1}} a^{\frac{1}{r+1}} + r^{-\frac{r}{r+1}} a^{\frac{1}{r+1}}$. The limit can be easily checked from the

expression. Then

$$\begin{aligned}\bar{u}'(d) &= -(\tilde{a}d^{-r/(r+1)}+1)^{-2}\tilde{a}(-\frac{r}{r+1})d^{-r/(r+1)-1} \\ &= \frac{\tilde{a}r}{r+1}\frac{1}{(\tilde{a}d^{-r/(r+1)}+1)(\tilde{a}d+d^{r/(r+1)+1})} > 0\end{aligned}$$

Then

$$\begin{aligned}\bar{u}''(d) = -\frac{\tilde{a}r}{r+1}\Big(&\frac{\tilde{a}(-r/(r+1))d^{-r/(r+1)-1}(\tilde{a}d+d^{r/(r+1)+1})}{((\tilde{a}d^{-r/(r+1)}+1)(\tilde{a}d+d^{r/(r+1)+1}))^2} \\ &+\frac{(\tilde{a}d^{-r/(r+1)}+1)(\tilde{a}+(r/(r+1)+1)d^{r/(r+1)})}{((\tilde{a}d^{-r/(r+1)}+1)(\tilde{a}d+d^{r/(r+1)+1}))^2}\Big)\end{aligned} \tag{B.17}$$

which can be further writen as

$$\bar{u}''(d) = -\frac{\tilde{a}r}{r+1}\frac{(\tilde{a}d^{-r/(r+1)}+1)(\tilde{a}/(r+1)+(r/(r+1)+1)d^{r/(r+1)})}{((\tilde{a}d^{-r/(r+1)}+1)(\tilde{a}d+d^{r/(r+1)+1}))^2} < 0$$

Thus, the concavity and monotonicity of $\bar{u}(d)$ is confirmed. □

Proof of Lemma 11

Proof. We prove each property one by one.

Property 1. The first-order derivative is given by

$$f'(d;n) = p_i'(d)\bar{u}(d-\bar{d}(n_a)) + p_i(d)\bar{u}'(d-\bar{d}(n_a))$$

The second-order derivative is given by

$$f''(d;n) = p_i''(d)\bar{u}(d-\bar{d}(n)) + 2p_i'(d)\bar{u}'(d-\bar{d}(n)) + p_i(d)u''(d-\bar{d}(n))$$

Moreover, since $p_i(d)$ is linear and decreasing in d,

$$p_i''(d) = 0, p_i'(d) < 0$$

By Lemma 10, $\bar{u}(\cdot)$ is increasing and strictly concave; hence,

$$\bar{u}'(d-\bar{d}(n)) > 0, u''(d-\bar{d}(n)) < 0$$

Therefore, $f''(d) < 0$ for $d \geq \bar{d}(n)$.

Property 2. When $d = \bar{d}(n)$, $f'(d) \to \infty$; when $d = m_i$, $f'(d) < 0$. Since $f'(d)$ is decreasing in d, and $[\bar{d}(n), m_i]$ is a compact set, by the intermediate value theorem, $f'(d) = 0$ has a unique solution on $[\bar{d}(n), m_i]$. Moreover, $f'(d) > 0$ for $d < d^\dagger$ and $f'(d) < 0$ for $d > d^\dagger$.

Property 3. The continuity of g is implied by the continuity of $\bar{u}$ and p_i. By the envelop theorem,

$$g'(n) = \frac{\partial f}{\partial n}|_{d=d^\dagger} = p_i(d^\dagger)\bar{u}'(d^\dagger - \bar{d}(n))(-\bar{d}'(n)) < 0$$

Moreover, since $g(0)$ is the highest per vehicle hourly earnings that HVs can make in a pure-HV market, then if HVs are viable (which is always the case in a loose market described in Assumption 5), it must be true that $g(0) = \max_{0 \le d \le m_i} p_i(d)\bar{u}(d) \ge p_i(d_i^*)\bar{u}(d_i^*) = w_0$, where d_i^* is the pure HV equilibrium demand; otherwise, HVs will not be able to make w_0 and no HVs will enter the market. In addition, one can easily check that $f(d; \underline{n}(m_i)) = 0$ for any d. Thus, $g(\underline{n}(m_i)) = 0$. □

Proof of Lemma 12

Proof. By Assumption 5, $p_i^* \le V/2$, which is equivalent to $d_i^* \ge m_i/2$. Thus, if we can show that the function $\pi(d) = (p_i(d)d - c_v\underline{n}(d))$ is strictly decreasing for $d \ge d_i^*$, then it proves that a monopoly firm does not have the incentive to extend the demand above d_i^*. Thus, we focus on showing $\pi(d)$ being strictly decreasing in the rest of the proof.

The first-order derivative of $\pi(d)$ is:

$$\pi'(d) = V(1 - \frac{2d}{m_i}) - c_v\Big(\frac{\tilde{a}}{r+1}d^{-\frac{r}{r+1}} + t_2\Big)$$

The second-order derivative of $\pi(d)$ is:

$$\pi''(d) = -\frac{2V}{m_i} + \frac{r\tilde{a}c_v}{(r+1)^2} d^{-\frac{r}{r+1}-1}$$

Then

$$\pi''(d) > 0 \Leftrightarrow d < \left(\frac{r\tilde{a}c_v m_i}{2V(r+1)^2}\right)^{\frac{1}{r/(r+1)+1}}$$

Hence,

$$\arg\max_d \pi'(d) = \left(\frac{r\tilde{a}c_v m_i}{2V(r+1)^2}\right)^{\frac{1}{r/(r+1)+1}}$$

If $\pi'\left(\left(\frac{r\tilde{a}c_v m_i}{2V(r+1)^2}\right)^{\frac{1}{r/(r+1)+1}}\right) \leq 0$, then $\pi'(d) \leq 0$, $\pi(d)$ is decreasing in d; then $\max \pi(d) = \pi(0) = 0$ and the maximizer is just 0.

If $\pi'\left(\left(\frac{r\tilde{a}c_v m_i}{2V(r+1)^2}\right)^{\frac{1}{r/(r+1)+1}}\right) > 0$, by the fact that $\pi'(0) \to -\infty$, $\pi'(-\infty) \to -\infty$ and $\pi'(d)$ is continuous, there must exist exactly two solutions, denoted as d_1, d_2 to the equation $\pi'(d) = 0$, with $d_1 < \left(\frac{r\tilde{a}c_v m_i}{2V(r+1)^2}\right)^{\frac{1}{r/(r+1)+1}}$, $d_2 > \left(\frac{r\tilde{a}c_v m_i}{2V(r+1)^2}\right)^{\frac{1}{r/(r+1)+1}}$. Moreover, $\pi'(d) < 0$ for $0 \leq d < d_1$, $\pi'(d) > 0$ for $d_1 < d < d_2$, and $\pi'(d) < 0$ for $d > d_2$.

As a result, $\pi(d)$ is either maximized at $d = 0$ or $d = d_2$. Since we consider a situation where HVs can make non-negative profit with a marginal cost $w_0 > c_v$, it implies that $\max_d \pi(d) > \pi(0) = 0$. Therefore, under our assumption, it must be true that $\arg\max \pi(d) = d_2$ and $\pi(d_2) > 0$.

Our last step is to show that $d_2 \in [0, m_i]$:

$$\pi'(d_2) = 0 \Leftrightarrow V(1 - \frac{2d_2}{m_i}) = c_v(\frac{\tilde{a}}{r+1} {d_2}^{-\frac{r}{r+1}} + t_2)$$

Then it must be true that $d_2 < \frac{m_i}{2}$, otherwise the left-hand side of the above equation would be negative while the right-hand side is positive. Therefore, the maximizer is just d_2 and $d_2 < m_i/2$. □

Proof of Lemma 16

Proof. Proof by contradiction. Suppose not. Then $p_i(\bar{d}(n_i^\dagger))\bar{u}(\bar{d}(n_i^\dagger)) < w_0$. Then for any $d > 0$, $p_i(\bar{d}(n_i^\dagger) + d)\bar{u}(\bar{d}(n_i^\dagger)) < p_i(\bar{d}(n_i^\dagger))\bar{u}(\bar{d}(n_i^\dagger)) < w_0$. This contradicts with the definition of $n_i^\dagger$. □

Proof of Lemma 15

Proof. It is equivalent to show that there exists a solution n_i^l such that $h(n_i^l) = c_v/w_0$. Under the condition that $h(n_i^\dagger) \geq c_v/w_0$, we prove the existence and uniqueness of the root to $h(n) = c_v/w_0$ using the intermediate value theorem.

1. $h(n)$ is continuous on $[0, n_i^\dagger]$. By the definition of $d_x(n)$, $\bar{d}(n) \leq d_x(n) \leq m_i$. Thus, $\bar{u}(d_x(n) - \bar{d}(n)) > 0$. By lemma 11, $d_x(n)$ is continuous on $n \in [0, n_i^\dagger]$. $\bar{u}(\bar{d}(n))$ is also continuous in n. Therefore, the quotient of $\bar{u}(\bar{d}(n))$ and $\bar{u}(d_x(n) - \bar{d}(n))$ is also continuous.
2. $h(n)$ is strictly increasing on $[0, n_i^\dagger]$. $\bar{d}(n)$ is increasing in n; $\bar{u}(\bar{d}(n))$ is increasing in n. $d_x(n)$ is decreasing in n. Thus, $h(n)$ is increasing in n.
3. $h(0) = 0$, $h(n_i^\dagger) \geq c_v/w_0$. By the intermediate value theorem and the monotonicity of h, there must exist a unique n on $[0, n_i^\dagger]$ such that $h(n) = c_v/w_0$.
4. n_i^l must be increasing in m_i. This can be seen from $\frac{\partial h}{\partial m_i} < 0$. For a given n, the only part in $h(n)$ that contains m_i is $d_x(n)$, and $\frac{\partial d_x(n)}{\partial m_i} > 0$. Thus, $\frac{\partial h}{\partial m_i} < 0$. We have shown that $\frac{\partial h}{\partial n} > 0$. Therefore, when holding $h(n) = c_v/w_0$ fixed, increasing m_i leads to an increasing root n.

Under the condition that $h(n_i^\dagger) < c_v/w_0$, we have $h(n) \leq h(n_i^\dagger) < c_v/w_0$, i.e. for $p_i(\bar{d}(n) + \bar{d}(n_0))\bar{u}(\bar{d}(n_0)) = w_0$, $p_i(\bar{d}(n) + \bar{d}(n_0))\bar{u}(\bar{d}(n_a) < c_v$ for all $n \leq n_i^\dagger$. In this case, HVs and AVs do not coexist. □

Proof of Proposition 32

Proof. The main arguments in Proposition 32 consist of three parts: 1) $n_i^\dagger$ is the threshold value that divides the two cases of coexisting AV-HV and pure-AV; this has been proved when we discuss the definition of $n_i^\dagger$ as well as Lemma 13 and its proof. 2) When AVs and HVs are in the coexisting regime, the price is given as Eq. (B.9) and is increasing in the AV fleet size n_a. 3) When AVs serve the market alone, the price is $p_i(\bar{d}(n_a)$; this is straightforward because $\bar{d}(n_a)$ is the highest demand rate that AVs can serve at a fleet size n_a, and the price now only depends on the participation of AVs. Thus, we focus on proving 2).

Price in the coexisting regime The equilibrium price can be obtained by solving

$$p\bar{d}(n_0)t_2 = w_0 n_0 \tag{B.18}$$

where n_0 is the positive root. Then the expression for the equilibrium price p^* can be obtained by dividing both sides by n_0. As for the relation between p^* and n_a, it is equivalent to show that the equilibrium demand served by HVs, d_0, is decreasing in n_a.

Proof by contradiction. Suppose not. Then there exists $n_1 < n_2 \leq n_i^\dagger$ such that $d_1 < d_2$, where $d_1 = \max\{d|p_i(d + \bar{d}(n_1))\bar{u}(d) = w_0\}$ and $d_2 = \max\{d|p_i(d + \bar{d}(n_2))\bar{u}(d) = w_0\}$. By Lemma 11, for any $d > d_1$, $p_i(d + \bar{d}(n_1))\bar{u}(d) < w_0$. Since $n_2 > n_1$ and $d_2 > d_1$, $p_i(d_2 + \bar{d}(n_2))\bar{u}(d_2) < p_i(d_2 + \bar{d}(n_1))\bar{u}(d_2) < p_i(d_1 + \bar{d}(n_1))\bar{u}(d_1) = w_0$. Contradiction. □

Proof of Lemma 13

Proof. For this proof, we leverage the Lemma 11 from the technical lemma section

(Appendix B.1.2). Thus, we first transform the definition of $n_i^\dagger$ in Eq. (B.7) and Eq. (B.8) in the form of hourly earnings *per HV*, like those used in Lemma 11.

The threshold $n_i^\dagger$ is defined based on the total net profit of HVs in scenario i for a market with an HV demand level d_0 and an AV demand level d_a, i.e.

$$R_i(d_0|d_a) - C(d_0)$$

Dividing the above term by the corresponding HV fleet size $\underline{n}(d_0)$ yields the net profit per HV, given by

$$p_i(d_0 + d_a)\bar{u}(d_0) - w_0$$

The threshold $n_i^\dagger$ is achieved when the above term is just at zero when the HV demand is at the profit-maximizing level. Let $d = d_0 + d_a$ and replace d_a with $\bar{d}(n)$, we have the above expression to be equivalent to

$$p_i(d)\bar{u}(d - \bar{d}(n)) - w_0$$

where the first term is just the hourly earnings per vehicle denoted as $f(d|n)$ in Lemma 13. Then it can be verified that the threshold $n_i^\dagger$ can be determined by solving $g(n_i^\dagger) = w_0$ from Lemma 11 in Appendix B.1.2. Thus, in the rest of the proof, we use the characteristics of g to provide insights for $n_i^\dagger$.

By Item 3 of Lemma 11, the function $g(n)$ is decreasing and continuous, $g(0) > w_0 > g(\underline{n}(m_i))$. By the mean value theorem, $g(n) = w_0$ must have a unique solution on $[0, \underline{n}(m_i)]$, which is just $n_i^\dagger$. This shows the existence and uniqueness of $n_i^\dagger$. Since $g(n)$ is monotone, its inverse function exists on $[0, \underline{n}(m_i)]$. Then $n_i^\dagger = g^{-1}(w_0)$.

Sufficient condition Next we show that for a given $n_a \leq n_i^\dagger$, there is a solution $d > \bar{d}(n)$ such that the hourly earnings per HV is w_0. Thus, it is equivalent to

show $f(d; n_a) = w_0$ has a solution such that $d > \bar{d}(n_a)$ when $n_a \leq n_i^\dagger$. This can be shown by the monotonicity of function g:

$$n_a \leq n_i^\dagger \Leftrightarrow n_a \leq g^{-1}(w_0) \Leftrightarrow w_0 \leq g(n_a) = \max_{\bar{d}(n_a) \leq d \leq m_i} f(d; n_a) = f(d^\dagger; n_a) \tag{B.19}$$

In other words, by Lemma 11, the maximum hourly earnings per HV $g(d; n)$ is decreasing in the AV fleet size n. Therefore, if $g(d; n)$ is just at zero when $n = n_i^\dagger$, then it must be true that it is positive when $n < n_i^\dagger$; moreover, by continuity, it shows that there must exists some intervals of d such that the hourly earnings per HV $f(d; n) > 0$, showing that this is a market that is viable for HVs.

Our next step is to verify if the root $f(d; n_a) = w_0$ is indeed positive. We prove this by continuity. Denote the maximizer of $f(d; n_a)$ as $d^\dagger$; its existence and uniqueness has been shown by Lemma 11. Thus, it must be true that we have $f(d^\dagger; n_a) = g(n_a) > w_0$. Moreover, it also must be true that $f(m_i; n_a) = 0 < w_0$. Then by the continuity of f, the equation $f(d; n_a) = w_0$ has a root on $d \in (d^\dagger, m_i)$.

Necessary condition Next we show if there exists a demand rate $d > \bar{d}(n_a)$ at which the hourly earnings per HV is w_0 and the AV fleet size is n_a, then it must hold that $n_a \leq n_i^\dagger$.

For an n_a, denote the root described above as d^*, where $d^* > \bar{d}(n_a)$. Then

$$p_i(d^*)\bar{u}(d^* - \bar{d}(n_a)) = w_0 \Leftrightarrow f(d^*|n_a) = w_0$$

Then by the definition of g,

$$w_0 = f(d^*|n_a) \leq g(n_a) \Leftrightarrow n_a \leq g^{-1}(w_0) = n_i^\dagger$$

which completes the proof.

Monotonicity of $n_i^\dagger$ Lastly, we show that $n_i^\dagger$ is increasing in m_i. Since $n_i^\dagger = g^{-1}(w_0)$, by the inverse function theorem,

$$\frac{\partial n_i^\dagger}{\partial m_i} = \frac{\partial g^{-1}(w_0)}{\partial m_i} = \frac{1}{\frac{\partial g(n_i^\dagger)}{\partial m_i}}$$

Again by the envelope theorem,

$$\frac{\partial g}{\partial m_i} = \frac{\partial f}{\partial m_i}|_{d^\dagger} > 0 \text{ for all } n < \underline{n}(m_i).$$

Hence,

$$\frac{\partial n_i^\dagger}{\partial m_i} > 0 \text{ for } n_i^\dagger < \underline{n}(m_i)$$

When $n_i^\dagger = \underline{n}(m_i)$, by Lemma 9, $n_i^\dagger$ is also increasing in m_i. □

Proof of Proposition 20

Proof. The numerical example in Figure 3.4 proves the existence of a market in which AVs cannot make a positive profit in a monopoly, independent platform market. A sketch for an analytical proof: for a two-point distribution where $m_1 << m_2$, the profit is maximized at either the optimal fleet size for potential demand m_1, or the optimal fleet size for potential demand m_2. Then show that the expected variable profit over the two scenarios is less than the fixed cost as in Figure 3.4. □

Proof of Proposition 21

Proof. Given any demand scenario, the market is either (a) a pure AV market, implying that an HV can make its reservation earnings w_0 but an AV cannot make its variable cost c_v; or (b) a coexisting AV-HV market, implying that an AV can make at least c_v and HVs can make w_0; or (c) a pure HV market, implying that an HV can make w_0 but an AV cannot make c_v.

Among all three cases, none of them can have price strictly lower than the pure HV equilibrium price. To see why, Proposition 32 and Corollary 4 implies

the price in (b) is strictly higher than that in a pure HV market. The price in (c) is just the pure HV equilibrium price. It remains to show that the price in (a) is at least as high as a pure HV market price.

Given capacity N and the condition that HVs cannot make w_0, a monopoly AV firm maximizes the total variable profit in scenario i:

$$\max_d p_i(d)dt_2 - c_v\underline{n}(d)$$

s.t.

$$d_i^\dagger < d \leq \bar{d}(N)$$

where $d_i^\dagger$ is defined by Eq. (B.7) and is the maximal demand rate satisfied by $n_i^\dagger$.

If the capacity $N < n_i^\dagger$, the pure AV regime does not exist under the given capacity. (The AV fleet size is not large enough to drive HVs out of the market.)

If $N \geq n_i^\dagger$, the monopoly supplier will have no incentive to extend its supply beyond $n_i^\dagger$. To see why, by Lemma 16, $n_i^\dagger < n_i^*$, which on the demand side implies that $d_i^\dagger < d_i^*$. By Lemma 12, the variable profit is decreasing for $d \geq d_i^*$. (The maximizer of the variable profit is no greater than d_i^*.) In the case where the maximizer is also no greater than $d_i^\dagger$, the variable profit is strictly decreasing on $[d_i^\dagger, \bar{d}(N)]$; the optimal demand rate is then above but infinitely close to $d_i^\dagger$, and the price is strictly higher than in the pure-HV market. In the case where the maximizer is greater than $d_i^\dagger$ (but still lower than d_i^*), the monopoly supplier just chooses the maximizer, which means the price is at least the same as the pure-HV market (strictly higher in scenarios where the maximizer is strictly lower than d_i^*).

Thus, in all cases, the price cannot be lower than the pure-HV equilibrium price. This completes the proof. □

Proof of Proposition 22

Proof. To prove this, we leverage the result from Proposition 32. By Proposition 32 and the related Corollary 4, which shows that, whenever in a scenario where AVs and HVs coexist, the price must be strictly higher than the pure-HV equilibrium price. Then, it is sufficient to show that, in a perfectly competitive independent platform market, there exists demand scenarios under which AVs and HVs coexist.

In Appendix B.4, we provide sufficient conditions under which HVs will supply the market along with AVs at least in the scenario with the highest potential demand. Then under this set of conditions, at least in the highest demand scenario, AVs and HVs coexist, and the price must be higher than the pure-HV equilibrium price; Numerical example Fig. 3.3 also exhibits a significant range of potential demand m_i under which AVs and HVs coexist under a perfectly competition common platform market (the region with a strictly decreasing market price of the dashed red curve). □

Proof for the condition of coexistence (Table B.1)

Proof. We prove for each market configuration.

(a) Common platform market, monopoly Proposition 18 shows the structure and the optimal choice of the AV capacity. It can be easily verified that, when $P(m_I) < \frac{c_f}{w_0 - c_v}$, the optimal capacity N^* will be lower than the total equilibrium fleet size n_I^* in scenario I, implying that HVs will participate at least in the highest demand scenario.

(b) Common platform market, competition **Intuition**: A sufficient condition for the HV participation is that, at any AV capacity that can drive HVs out of the market in all scenarios, AVs cannot break-even (i.e. the aggregate variable profit per vehicle is below c_f).

Recall that under a common platform market, AVs can supply the market alone in scenario i if and only if the AV fleet size is no less than n_i^*; thus, the condition is equivalent to:

$$\sum_{i=1}^{I} P(m_i)(p_i(d)\bar{u}(d) - c_v)^+ < c_f, \text{ for all } d \geq d_I^* \tag{B.20}$$

This condition can be simplified as

$$\sum_{i=1}^{I} P(m_i)(p_i(d_I^*)\bar{u}(d_I^*) - c_v)^+ < c_f \tag{B.21}$$

To see why, we show that the left-hand side of Eq. (B.20), which represents the variable profit per vehicle at demand rate d, is decreasing in d for $d \geq d_I^*$. Essentially, we show that in any scenario i, the variable profit per vehicle, $((p_i(d)\bar{u}(d) - c_v)^+$, is decreasing for $d \geq d_I^*$.

Consider two functions, $r_i(d)$ (defined in Section 3.4), and the variable revenue per vehicle $p_i(d)\bar{u}(d)$. Both functions are strictly concave and unimodal. ($r_i(d)$ is just a quadratic function of d; Lemma 11 shows the characteristics of $p_i(d)\bar{u}(d)$) By Assumption 5, it is true that $d_i^* \geq \arg\max_d r_i(d)$. By definition, $r_i(d) = p_i(d)dt_2$, which can further be written as

$$r_i(d) = (p_i(d)\bar{u}(d))\underline{n}(d)t_2$$

Then $\arg\max_d r_i(d) > \arg\max_d p_i(d)\bar{u}(d)$ must hold; to see why, taking the first order derivative of $r_i(d)$ gives the following:

$$r_i'(d) = (p_i(d)\bar{u}(d))'\underline{n}(d)t_2 + (p_i(d)\bar{u}(d))\underline{n}'(d)t_2$$

when the revenue per vehicle $p_i(d)\bar{u}(d)$ reaches its maximum (i.e. $(p_i(d)\bar{u}(d))' = 0$), $r_i'(d) > 0$, meaning that $r_i(d)$ is still increasing. Thus, we have

$$d_I^* \geq d_i^* \geq \arg\max_d r_i(d) > \arg\max_d p_i(d)\bar{u}(d)$$

for all $i = 1, ..., I$. The leftmost inequality is by the monotonicity of d_i^* in Proposition 14. Since $p_i(d)\bar{u}(d)$ is unimodal, it is strictly decreasing once d is over the peak. Therefore, $p_i(d_i^*)\bar{u}(d_i^*) > p_i(d)\bar{u}(d)$ for any $d > d_i^*$. Since this is true for every i, this completes the proof for the equivalence between Eq. (B.20) and Eq. (B.21).

Therefore, Eq. (B.21) is a sufficient condition for the HV participation. Next, we rewrite it to make it more readable. By definition of d_i^*, $p_i(d_i^*)\bar{u}(d_i^*) = w_0$. Thus, Eq. (B.21) can be rewritten as the following:

$$P(m_I)(w_0 - c_v) + \sum_{i=1}^{I-1} P(m_i)(p_i(d_I^*)\bar{u}(d_I^*) - c_v)^+ < c_f$$

By rearranging the terms, we have an inequality about $P(m_I)$ with the same expression as the expression in Table B.1 under "Competitive, Common Platform".

(c) Independent platform market, monopoly **Intuition:** A sufficient condition for the HV participation is that, at any AV capacity that can drive HVs out of the market in all scenarios, the monopoly AV supplier do not make as much total profit as when it ignores the highest demand scenario. In other words, the supplier makes more profit by only considering scenario 1 to $I - 1$, rather than elevating its capacity and trying to become the sole supplier in scenario I.

We start by computing the supplier's optimal profit with a requirement that the AV fleet size should always be high enough such that the demand level $d > d_i^\dagger$ and HVs cannot participate in all scenarios. In other words, This further

translate into a lower bound on capacity N:

$$N \geq n_I^\dagger$$

When the capacity N is below $n_I^\dagger$, AVs and HVs will operate together at least in the highest demand scenario. We show that, among all $N \geq n_I^\dagger$, capacity $N = n_I^\dagger$ leads to the highest total profit. Then, it is sufficient to only consider and compare $N = n_I^\dagger$ with other capacity levels below $n_I^\dagger$.

To show this, we first present the expression for the supplier's profit, given a capacity N:

$$\sum_1^I P(m_i) \max_{d_i^\dagger \leq d \leq \bar{d}(N)} (p_i(d)dt_2 - c_v \underline{n}(d)) - c_f N$$

which is the same as Eq. (3.1) by plugging into the variable profit Eq. (B.12) and imposing the condition that $d \geq d_i^\dagger$. We show that the total profit in scenario i, $(p_i(d)dt_2 - c_v \underline{n}(d))$, is decreasing for $d > d_I^\dagger$ for all $i = 1, ...I$. This leads to two subcases:

Case 1: $i = I$**.** In the special case of $d_I^\dagger > \tilde{d}_I$ that we are considering (recall that $\tilde{d}_I$ is the maximizer for the total profit $(p_I(d)dt_2 - c_v \underline{n}(d))$ in scenario I), it must be true that $\arg\max_{d_i^\dagger \leq d \leq \bar{d}(N)} (p_i(d)dt_2 - c_v \underline{n}(d))$ in scenario I is just $d_I^\dagger$. This is because the total profit in scenario I reaches its maximum at $\tilde{d}_I$ and is decreasing for $d \geq d_I^\dagger > \tilde{d}_I$.

Case 2: $i = 1, ...I - 1$**.** Because $d_I^\dagger > \tilde{d}_I$, it must also be true that $d_I^\dagger > \tilde{d}_i$ for all i that is below I ($\tilde{d}_i$ is increasing in the potential demand m_i). Then we have that $(p_i(d)dt_2 - c_v \underline{n}(d))$ is decreasing for $d > d_I^\dagger$ for all i. Therefore, the maximal aggregate total profit the supplier can earn while keeping AVs out in all scenarios is given by

$$\sum_1^I P(m_i) \max_{d_i^\dagger \leq d \leq d_I^\dagger} (p_i(d)dt_2 - c_v \underline{n}(d)) - c_f n_I^\dagger \tag{B.22}$$

Since we have shown that $d_I^\dagger > \tilde{d}_i$ and $d_I^\dagger \geq d_i^\dagger$ for all i, in all scenarios (except for $i = I$), the capacity constraint is not binding. Thus, the optimal demand at each scenario is $\tilde{d}_i$ when it is feasible, and $d_i^\dagger$ otherwise. Define $\hat{d}_i = \max\{\tilde{d}_i, d_i^\dagger\}$, which represents the optimal demand rate that can keep HVs out for scenario i. Let $\hat{n}_i = \underline{n}(\hat{d}_i)$. Then Eq. (B.22) can be further written as

$$P(m_I)(p_i(d_I^\dagger)d_I^\dagger t_2 - c_v \underline{n}(d_I^\dagger)) + \sum_1^{I-1} P(m_i)(p_i(\hat{d}_i)\hat{d}_i t_2 - c_v \underline{n}(\hat{d}_i)) - c_f n_I^\dagger \quad \text{(B.23)}$$

Next, we compute the profit when the supplier does not consider scenario I as part of its objective, and only care for the aggregate total profit from scenario 1 to I. To further simplify the problem, suppose the supplier decides to choose a capacity $N = \hat{n}_{I-1}$, which is strictly below $n_I^\dagger$, meaning that AVs and HVs will operate together in scenario I under this capacity; moreover, the supplier decides to supply the market alone in scenarios from 1 to $N - 1$. (We can allow the capacity here to be non-optimal because we are only characterizing a sufficient condition.) Then the supplier's profit will just be

$$\sum_1^{N-1} P(m_i)(p_i(\hat{d}_i)\hat{d}_i t_2 - c_v \underline{n}(\hat{d}_i)) - c_f \hat{n}_{I-1} \quad \text{(B.24)}$$

Let Eq. (B.23) $<$ Eq. (B.24). When this happens, a monopoly AV supplier will never selects a capacity at or above $n_I^\dagger$, and AVs will always coexist with HVs at least in the highest demand. Plugging in Eq. (B.23) and Eq. (B.24), the terms concerning $i < I$ cancels out, and the condition is equivalent to

$$-c_f \hat{n}_{I-1} > P(m_I)(p_i(d_I^\dagger)d_I^\dagger t_2 - c_v n_I^\dagger) - c_f n_I^\dagger$$

By dividing both sides by $n_I^\dagger$ and rearranging the terms, we obtain the condition in Table B.1 under "Monopoly, Independent platform".

(d) Independent platform market, competition The intuition is similar as (b). A sufficient condition would be AVs cannot break-even at any capacity that

allows the AV supply level to be high enough to drive HVs out in all scenarios. Then, we obtain a similar condition to Eq. (B.20):

$$\sum_{i=1}^{I} P(m_i)(p_i(d)\bar{u}(d) - c_v)^+ < c_f, \text{ for all } d \geq d_I^\dagger \tag{B.25}$$

The only difference is that, AVs serve the market alone in all scenarios if and only if $d \geq d_I^\dagger$ rather than d_I^*. (See the definition of $d_i^\dagger$ in and after Eq. (B.8).) Similar to Case (b), we start by showing that the left-hand side of Eq. (B.25) is decreasing for $d \geq d_I^\dagger$; then we derive a condition based on $d_I^\dagger$. Again, we prove the first part by showing that $(p_i(d)\bar{u}(d) - c_v)^+$ is decreasing in d for $d \geq d_I^\dagger$ for any i.

First, we have assumed that the market in consideration satisfies $d_I^\dagger > \tilde{d}_I$, where $\tilde{d}_I$ is the maximizer for the total variable profit $p_i(d)d - c_v\underline{n}(d)$ in scenario I. This implies that, $d_I^\dagger$ is also greater than $\arg\max_d p_I(d)\bar{u}(d)$; to see why, the total variable profit for scenario I can be rewritten as

$$(p_I(d)\bar{u}(d) - c_v)\underline{n}(d)$$

Taking the first-order derivative of the above term gives

$$(p_I(d)\bar{u}(d) - c_v)'\underline{n}(d) + (p_I(d)\bar{u}(d) - c_v)\underline{n}'(d)$$

At the maximizer of $p_I(d)\bar{u}(d)$, the first term is zero; the second term must also be greater than zero, otherwise $p_I(d)\bar{u}(d) < c_v$ at any d, which means an AV cannot even make its variable cost c_v at any demand level in a scenario with the highest demand and no competition from HVs and contradicts Assumption 5. Therefore, it must be true that $\arg\max_d p_I(d)\bar{u}(d) < \arg\max_d p_I(d)dt_2 - c_v\underline{n}(d) < d_I^\dagger$. Thus, in scenario I, $p_I(d)\bar{u}(d)$ is decreasing on $d \geq d_i^\dagger$.

Next, we show that for any $i \leq I-1$, $p_i(d)\bar{u}(d)$ is also decreasing on $d \geq d_I^\dagger$. Since we have shown that $d_I^\dagger > \arg\max_d p_I(d)\bar{u}(d)$, it is sufficient to show

$\arg\max_d p_i(d)\bar{u}(d)$ is increasing in i, i.e. the demand rate that maximizes the per vehicle revenue is increasing in the potential demand. To see why, take the first-order derivative over $p_i(d)\bar{u}(d)$, which givens the following

$$p_i'(d)\bar{u}(d) + p_i(d)\bar{u}'(d)$$

Setting the above term to zero and rearranging the terms gives

$$-\frac{p_i'(d)}{p_i(d)} = \frac{\bar{u}'(d)}{\bar{u}(d)}$$

The right-hand side of the equation is the same regardless of i. The left-hand side of the equation can be further simplified by plugging in the expression for $p_i(d)$:

$$-\frac{p_i'(d)}{p_i(d)} = -\frac{-V/m_i}{V(1-d/m_i)} = \frac{1}{m_i - d}$$

Then the first-order condition is equivalent to

$$\frac{1}{m_i - d} = \frac{\bar{u}'(d)}{\bar{u}(d)} \Leftrightarrow m_i = d + \frac{\bar{u}(d)}{\bar{u}'(d)}$$

By Lemma 10, $\bar{u}(d)$ is strictly increasing and concave, meaning that $\frac{\bar{u}(d)}{\bar{u}'(d)}$ is increasing in d. Therefore, as i and correspondingly, m_i, increase, the maximizer for $p_i(d)\bar{u}(d)$ also increases. This completes the proof for $d_I^\dagger > \arg\max_d p_i(d)\bar{u}(d)$ for all $i \leq I-1$.

From above, we have shown that $p_i(d)\bar{u}(d)$ is decreasing in d for $d \geq d_I^\dagger$ for all i. As a result, Eq. (B.25) is equivalent to

$$\sum_{i=1}^{I} P(m_i)(p_i(d_I^\dagger)\bar{u}(d_I^\dagger) - c_v)^+ < c_f \tag{B.26}$$

From Eq. (B.26) to the condition in Table B.1, the only missing step is to take $i = I$ out of the summation, remove the $()^+$ and rearrange the terms. Thus, we next show that $p_i(d_I^\dagger)\bar{u}(d_I^\dagger) > c_v$. By Lemma 16, the highest AV fleet size that allows AVs and HVs to coexist is higher in a common platform market than

an independent platform market, i.e. $d_I^\dagger < d_I^*$. Moreover, by definition, the equilibrium per vehicle revenue $p_I(d)\bar{u}(d)$ at $d = d_I^*$ is just w_0. Since we have shown that $p_I(d)\bar{u}(d)$ is decreasing for $d \geq d_I^\dagger$, it must be true that $p_I(d_I^\dagger)\bar{u}(d_I^\dagger) > w_0 > c_v$.

Therefore, we have Eq. (B.26) equivalent to

$$P(m_I)(p_I(d_I^\dagger)\bar{u}(d_I^\dagger) - c_v) + \sum_{i=1}^{I-1} P(m_i)(p_i(d_I^\dagger)\bar{u}(d_I^\dagger) - c_v)^+ < c_f$$

Rearranging the terms, we have the condition in Table B.1 under "Competitive, Independent Platform". □

Proof of Proposition 23

Proof.

We first focus on the case where HVs are viable both before and after introducing AVs. Consider any demand scenario where HVs are viable in a pure-HV market, i.e. there is a positive number of n_0^* HVs that can make reservation earnings $w(n_0^*)$. Denote the equilibrium HV fleet size when there is a fleet of n_a AVs as $n_0^\dagger$ and suppose that $n_0^\dagger > 0$. By definition, n_0^* should be the largest root such that

$$r_i(d) = w(n_0^*)n_0^*, \text{ where } d = \bar{d}(n_0^*)$$

When $n_0 > 0$, the above equation is equivalent to

$$p_i(d)\bar{u}(d) = w(n_0^*), \text{ where } d = \bar{d}(n_0^*) \tag{B.27}$$

On the other hand, $n_0^\dagger$ is defined as the largest root such that

$$p_i(d)\bar{u}(\bar{d}(n_0^\dagger)) = w(n_0^\dagger), \text{ where } d = \bar{d}(n_0^*) + \bar{d}(n_a)$$

We show that $n_0^* > n_0^\dagger$ for all $n_a > 0$ by contradiction. Suppose not. Then there exists an n_a such that $n_0^* \leq n_0^\dagger$. Then by the monotonicity of the supply curve w,

it must be true that

$$w(n_0^\dagger) \geq w(n_0^*)$$

Thus, by the definitions of n_0^* and $n_0^\dagger$, we have

$$p_i(\bar{d}(n_0^\dagger) + \bar{d}(n_a))\bar{u}(\bar{d}(n_0^\dagger))) > p_i(\bar{d}(n_0^*))\bar{u}(\bar{d}(n_0^*))$$

Moreover, given that $n_a > 0$, it must be true that

$$p_i(\bar{d}(n_0^\dagger) + \bar{d}(n_a))\bar{u}(\bar{d}(n_0^\dagger))) < p_i(\bar{d}(n_0^\dagger))\bar{u}(\bar{d}(n_0^\dagger)))$$

Therefore, we have

$$p_i(\bar{d}(n_0^\dagger))\bar{u}(\bar{d}(n_0^\dagger)) > p_i(\bar{d}(n_0^*))\bar{u}(\bar{d}(n_0^*))$$

But this violates the definition that n_0^* is the largest root of Eq. (B.27). Contradiction.

Now consider the first corner case where $n_0^\dagger > 0$ does not exist, but $n_0^* > 0$ exists. That is, the demand is large enough to support just HVs, but after introducing n_a AVs, HVs can no longer break even. Then it is clear that $n_0^* > n_0^\dagger$ and HVs are worse off.

The last corner case is when a positive n_0^* does not exist in the first place. That is, the demand cannot support HVs' operation in a pure-HV market. Then it implies that

$$p_i(d)\bar{u}(d) < w(n_0) \text{ for any } n_0 > 0, \text{ where } d = \bar{d}(n_0)$$

For any $n_a > 0$, it must be true that

$$p_i(\bar{d}(n_0) + \bar{d}(n_a))\bar{u}(\bar{d}(n_0)) < p_i(\bar{d}(n_0))\bar{u}(\bar{d}(n_0))$$

Thus, it must be true that

$$p_i(\bar{d}(n_0) + \bar{d}(n_a))\bar{u}(\bar{d}(n_0)) < w(n_0)$$

Thus, if HVs cannot break even in a pure-HV market, it will never break even in an independent platform market with any AV fleet size. In this case, HVs are also not better off because of AVs. □

APPENDIX C

LABOR COST FREE-RIDING IN THE GIG ECONOMY

C.1 Structure

The appendix is organized as follows: Appendix C.2 contains a comprehensive model that provides the micro-foundation for workers' decisions. In Appendix C.3, we state our most general results (Theorem 10), with firm-dependent revenue v_i, expected job duration t_i and a general waiting function $W(\lambda; \mu_{\geq w_0})$, of which a brief version appears in Proposition 25. In Appendix C.4, we give the proof of this theorem, and in Appendix C.5 we show that Proposition 24 and Theorems 7 and 8 from the main paper are simply special cases of Theorem 10 in Appendix C.3. Appendix C.6 provides proofs of several technical Lemmas used for the proof of Theorem 10. Appendix C.7 provides the data source for the empirical evidences in the introduction.

C.2 Micro-structure model of worker's decisions

We model the decision process of an individual worker as a Markov decision process as follows: Workers are in one of three states, denoted s, as shown in Table C.1:

s	Description
-1	out of the worker pool
0	waiting for work in the worker pool
1	working

Table C.1: Worker states

Workers earn their reservation wage w_0 for any time spent outside the worker pool. Jobs have the same average duration of $T = 1/\lambda_0$, where the job duration is exponentially distributed and λ_0 is the rate at which jobs finish. We assume workers rejoin the pool upon completion of a job. (They can immediately exit the pool once they rejoin it.) Firm j offers jobs to the worker at an independent Poisson rate λ_j, $j = 1, \ldots, N$, and job completion events are Poisson at rate λ_0. Without loss of generality, we assume the overall rate of events $\lambda_0 + \sum_{j=1}^{N} \lambda_j = 1$, which implies the mean interarrival time between events is one unit of time. Let the random variable ω denote the type of arrival event (i.e., $\omega = j$ if the arrival event is a job offer from firm j) and note that $P(\omega = j) = \lambda_j$, $j = 0, \ldots, N$.

We assume the worker makes decisions only when events occur, so their decision making process is embedded at these arrival time epochs, which we index by t. Formally, let a denote the action (decision) of the worker, $A(s, \omega)$ denote the set of feasible actions when a worker is in state s and faced with event ω. The possible actions are defined in Table C.2:

a	Description
-1	leave the worker pool
0	join (or stay in) the worker pool
j	accept a job from firm j, $j = 1, \ldots, N$

Table C.2: Worker actions

The feasible action sets are

$$A(-1, \omega) = \{-1, 0\}$$

$$A(0, \omega) = \{-1, 0, \omega\}$$

$$A(1, \omega) = \{1\}$$

Each state-action pair (s, a) generates an expected reward $r(s, a)$ defined by

$$r(s,a) = \begin{cases} w_0 & s = -1, 0;\ a = -1 \\ p_j/\lambda_0 & s = 0;\ a = j \\ 0 & s = -1, 0;\ a = 0 \\ 0 & s = 1 \end{cases}$$

This implies workers are paid for an accepted job at the time they decide to accept, where p_j is the pay rate of firm j and $1/\lambda_0$ is the average duration of a job. Similarly, the reservation wage rate w_0 is earned when a worked decides to leave the pool and remains outside the pool until the next decision epoch (event arrival), an expected duration of one unit of time. For an average cost expected value problem, accounting for these expected rewards at the decision epoch is without loss of generality.

The non-zero state transition probabilities, $q_{ss'}(a) = P(s_{t+1} = s' | s_t = s, a)$, are

$$\begin{aligned} q_{-1,-1}(-1) &= 1, & q_{-1,0}(0) &= 1 \\ q_{0,-1}(-1) &= 1, & q_{0,0}(0) &= 1, & q_{0,1}(j) &= 1,\ j > 0 \\ q_{1,0}(1) &= \lambda_0, & q_{1,1}(1) &= 1 - \lambda_0 \end{aligned}$$

That is, a decision to join or quit the pool changes the state deterministically accordingly, and a decision to accept a job deterministically puts the worker in the working state. The only probabilistic transitions are when a worker is working, in which case they finish work when a job completion event occurs and continue working otherwise.

The worker's objective is to maximize their expected average long-run reward. This maximization is achieved by solving the following average-cost dynamic programming optimality condition

$$\gamma + h(s) = E_\omega \left[\max_{a \in A(s,\omega)} \left\{ r(s,a) + \sum_{s'} q_{ss'}(a) h(s') \right\} \right], \quad s = -1, 0, 1. \qquad \text{(C.1)}$$

where γ is the long-run reward rate and $h(s)$ is the differential reward of state s. For the feasible action sets, rewards and state transition probabilities defined above, this reduces to solving:

$$\gamma + h(-1) = \max\{w_0 + h(-1),\ h(0)\} \tag{C.2}$$

$$\gamma + h(0) = E_\omega\left[\max\{w_0 + h(-1),\ h(0),\ p_\omega/\lambda_0 + h(1)\}\right] \tag{C.3}$$

$$\gamma + h(1) = \lambda_0 h(0) + (1-\lambda_0)h(1) \tag{C.4}$$

If the market is in equilibrium, it must be true that workers who are out of the pool are indifferent between staying out or joining. Likewise, workers who are waiting must be indifferent between exiting the pool and continuing waiting. This means both alternative actions in (C.2) of staying out or joining must have equal value, and similarly the two actions of continuing to wait or exiting in (C.3) must have equal value. These conditions reduce to satisfying the following equilibrium condition in addition to the optimality condition (C.1).

$$w_0 + h(-1) = h(0) \tag{C.5}$$

Lastly, we also require that the solution satisfy the equilibrium condition (4.1), which recall states that the opportunity cost of waiting for work must equal the excess earnings on that work.

The next proposition provides the optimal solution:

Proposition 33 *The unique optimal solution to both the equilibrium conditions (C.5) and (4.1), and the optimality condition (C.1) is* $\gamma^* = w_0$, $h^*(-1) = 0$, $h^*(0) = w_0$ *and* $h^*(1) = -w_0(1-\lambda_0)/\lambda_0$. *Moreover, the worker's optimal decision while in the worker pool is to accept a job offer from firm* j *iff* $p_j \geq w_0$.

Proof. It is easy to verify by direct substitution that this solution satisfies the equilibrium condition (C.5) and optimality conditions (C.2) and (C.4). This leaves only the optimality condition (C.3). To analyze it, note that for $h = h^*$, $w_0 + h(-1) = h(0) = w_0$, so the inner maximization in (C.3) reduces to

$$\max \{w_0, p_\omega/\lambda_0 + h^*(1)\}$$

Substituting $h^*(1) = -w_0(1-\lambda_0)/\lambda_0$ for the second term in the max (accepting the job ω), we see that

$$p_\omega/\lambda_0 + h^*(1) = \frac{1}{\lambda_0}(p_\omega - w_0(1-\lambda_0)),$$

where the right hand side above is greater than or equal to w_0 iff $p_\omega \geq w_0$. This proves the claim about the worker's optimal decision while in the worker pool.

Now, let $\lambda^* = \sum_{j:p_j \geq w_0} \lambda_j$ denote the rate of jobs from firms offering at least w_0 and define the average accepted price by

$$P^* = \left(\sum_{j:p_j \geq w_0} \lambda_j p_j \right) / \lambda^*$$

Then (C.3) becomes

$$\gamma^* + h^*(0) = \lambda^*(P^*/\lambda_0 + h^*(1) + (1-\lambda^*)w_0$$

Substituting the values of γ^* and h^* and simplifying, we have

$$w_0 \frac{1}{\lambda^*} = (P^* - w_0)\frac{1}{\lambda_0}.$$

But note that $1/\lambda^*$ is simply the worker's waiting time for a job they are willing to accept (under their optimal policy), so the left hand side is the worker's cost of waiting for work. And the right hand side is the total excess pay (pay above the

reservation wage) a worker receives from an accepted job. Hence, this condition simply says that the worker's waiting cost equals the excess pay generated from waiting, which is precisely the equilibrium condition (4.1). □

C.3 Statement of the general theorem

Theorem 10 *Consider a gig economy with N firms and firm-dependent revenue v_i and job duration t_i, $1 \leq i \leq N$. Suppose that for a given $\mu_{\geq w_0}$, the waiting function $\lambda \to W(\lambda; \mu_{\geq w_0})$ is strictly increasing, twice-differentiable and convex in λ on $[0, b)$; $W'(\lambda; \mu_{\geq w_0})$ is nonzero on $[0, b)$; $\lim_{\lambda \to b} W(\lambda; \mu_{\geq w_0}) = +\infty$; and $W(0; \mu_{\geq w_0}) = 1/\mu_{\geq w_0}$. Let $\mu_1(v_1 - w_0)t_1 \geq \mu_2(v_2 - w_0)t_2 \geq ... \geq \mu_N(v_N - w_0)t_N$. Then the following holds:*

1. *Only two types of equilibria are possible: either all firms participate and are profitable ($\lambda > 0$), or none of the firms participate ($\lambda = 0$).*
2. *A Nash equilibrium with all firms participating uniquely exists if and only if*

$$\sum_{1}^{N} \mu_i (v_i - w_0) t_i > w_0 \tag{C.6}$$

Moreover, the equilibrium $(q_1^*, q_2^*, ..., q_N^*)$ is given by

$$q_i^* = \min(\theta^*, k_i) \tag{C.7}$$

where θ^* can be uniquely obtained by solving

$$\theta^* = h(\sum_{1}^{N} \min(\theta^*, k_i) - \theta^*) \tag{C.8}$$

within the closed interval $[0, \bar{\theta}]$, where $\bar{\theta}$ is defined by $\sum_1^N \min(\bar{\theta}, k_i) = \sum_{i=1}^N k_i - W(0)$. The definition of function h can be found in Appendix C.6

3. *A Nash equilibrium with no firms participate exists if and only if none of the firms have enough jobs to attract workers and make positive profit on their own. That is, for each firm $i \leq N$, it holds that*

$$\mu_i \leq \frac{w_0}{t_i(v_i - w_0)}$$

Moreover, these equilibria correspond to any set of prices (p_i) such that for each firm i,

$$\mu_i(v_i - w_0)t_i + \sum_{j \neq i, p_j \geq w_0} \mu_j(p_j - w_0)t_j \leq w_0 \tag{C.9}$$

In other words, whenever $\sum_1^N \mu_i(v_i - w_0)t_i \leq w_0$ holds, there must exist a Nash equilibrium in which all firms participate in the gig economy and make positive profit, and such equilibrium is the only one with all firms participating and making positive profit; whenever $\mu_i \leq w_0/(t_i(v_i - w_0))$ holds, there always exist nash equilibria with no firms participate in the gig economy. Indeed, for the parameter space $(\mu_1, \mu_2, ..., \mu_N) \in M_1 \times M_2 \times ... \times M_N$, what Theorem 10 claims is that there are three possibilities:

1. $\sum_1^N \mu_i(v_i - w_0)t_i > w_0$ and $\exists i$, such that $\mu_i > w_0/(t_i(v_i - w_0))$. There exists one unique Nash equilibrium, and in this Nash equilibrium, all firms participate and make positive profit.
2. $\sum_1^N \mu_i(v_i - w_0)t_i > w_0$ and $\mu_i \leq w_0/(t_i(v_i - w_0)), \forall i$. There exists two types of equilibrium: one with all firms participating and making positive profit (there is only one such equilibrium), and the other with no firms participating and all make zero profit.
3. $\sum_1^N \mu_i(v_i - w_0)t_i \leq w_0$ and $\mu_i \leq w_0/(t_i(v_i - w_0)), \forall i$. There exists a unique type of Nash equilibrium, and in this type of Nash equilibrium, no firms participate and all make zero profit.

Nonetheless, essentially there are just two types of equilibria: one with a *viable* market (all firms participate and make profit) and one with an *unviable* market (none firms participate and all make zero profit). For ease of exposition, we categorize the equilibria based on whether the market is viable (Part 2, Part 3 in 10), respectively, instead of enumerating the three possibilities in the parameter space of μ_i.

C.4 Proof of the general theorem

The main structure of the proof is the following: we first show that there are two types of equilibria (Part 1 in Theorem 10); Then we characterize the condition for the equilibrium with all firms participating and profitable to hold, show the expression of the equilibrium, and prove the existence and uniqueness of the equilibrium (Part 2, 3 in Theorem 10).

C.4.1 Two types of equilibria

Proof. We show that there is no Nash equilibrium in which some firms participate and are profitable, while other firms do not participate. The intuition is that, whenever some firms are actively participating in the gig economy by setting payments at or above w_0, it means the hurdle of establishing a sizable worker pool has been overcome, and it is much easier for other firms to make profit.

Mathematically, we use a proof by contradiction, and suppose that there exists a Nash equilibrium $(p_1^*, p_2^*, ..., p_N^*)$ such that a subset of firms participate

and make profit, and the rest firms do not participate. Let $J = \{j : p_j^* \geq w_0\}$, i.e. J is the set of firms who participate in the gig economy and J is non-empty. We will prove the results by showing that if firms in J make profit, then any firm i who are not in J will be strictly better off by deviating to $p_i = w_0$ and participating, thus $(p_1^*, p_2^*, ..., p_N^*)$ cannot be a Nash equilibrium.

Define $\sum_{j\in J} \mu_j = \mu_J$, and let λ_J denote the equilibrium worker arrival rate in the given Nash equilibrium. As firms in J make profit, we have $\lambda_J > 0$ and by (4.13) we obtain:

$$\sum_{j\in J} \frac{\mu_j}{\mu_J} p_j^* t_j = w_0 (\sum_{j\in J} \frac{\mu_j}{\mu_J} t_j + W(\lambda_J; \mu_J)) \iff w_0 W(\lambda_J; \mu_J) = \sum_{j\in J} \frac{\mu_j}{\mu_J} (p_j^* - w_0) t_j$$

Suppose non-participating firm $i \notin J$ (that make zero profit) instead participates with $p_i = w_0$. Then its profit is given by:

$$\pi_i(w_0; p_{-i}) = \frac{\mu_i}{\mu_J + \mu_i} \lambda_{J+i} (v_i - w_0) t_i$$

where, λ_{J+i} verifies

$$\sum_{j\in J} \frac{\mu_j}{\mu_J + \mu_i} p_j^* t_j + \frac{\mu_i}{\mu_J + \mu_i} w_0 t_i = w_0 (\sum_{j\in J} \frac{\mu_j}{\mu_J + \mu_j} t_j + \frac{\mu_i}{\mu_J + \mu_i} t_i) + w_0 W(\lambda_{J+i}; \mu_J + \mu_i)$$

which can be simplified as

$$W(\lambda_{J+i}; \mu_J + \mu_i) = \frac{1}{w_0} \sum_{j\in J} \frac{\mu_j}{\mu_J + \mu_i} (p_j^* - w_0) t_j = \frac{\mu_J}{\mu_J + \mu_i} W(\lambda_J; \mu_J)$$

Recall that by definition, $W(0; \mu_{\geq w_0}) = 1/\mu_{\geq w_0}$. Thus, when i does not participate, $W(0; \mu_J) = \frac{1}{\mu_J}$. When i participates, $W(0; \mu_J + \mu_i) = \frac{1}{\mu_J + \mu_i}$. Therefore, we have the following: if $\lambda_J > 0$, by the assumption in Theorem 10 that $W(\lambda; \mu_{\geq w_0})$ being strictly increasing, it must be true that $W(\lambda_J; \mu_J) > W(0; \mu_J) = \frac{1}{\mu_J}$. Thus,

$$W(\lambda_{J+i}; \mu_J + \mu_i) = \frac{\mu_J}{\mu_J + \mu_i} W(\lambda_J; \mu_J) > \frac{\mu_J}{\mu_J + \mu_i} W(0; \mu_J) = \frac{\mu_J}{\mu_J + \mu_i} \frac{1}{\mu_J} = W(0; \mu_J + \mu_i)$$

which implies that $\lambda_{J+i} > 0$, $\pi_i(w_0; p_{-i})$ is strictly above 0, and firm i is strictly better off by participating. □

C.4.2 Existence, uniqueness and the closed-form expression for the Nash equilibrium with a viable market

Proof. In this section we prove Theorem 10 on the existence and uniqueness of the Nash equilibria with profitable firms and characteristics of this equilibrium. This proof is constructive and uses the symmetry of the best response decision problem (4.15) in order to reduce the search of the equilibrium to the search of the fixed point of a one-dimensional function.

Simplified best-response problem Note that the equilibrium we search for is one with all firms participating. This means that for firm i's best response, we only need to consider other firms' strategies being $0 < q_{-i} \leq k_{-i}, \forall j \neq i$. This greatly simplifies the best response decision problem Eq. (4.15). if firm i picks $q_i > k_i$, the profit $\pi(q_i; q_{-i})$ is just zero; if firm i picks $q_i \leq k_i$, then every firm is participating the economy, which means $\sum_{j,q_j \leq k_j}^{N}(k_j - q_j) = \sum_{1}^{N}(k_j - q_j)$ and $\mu_{\geq w_0} = \mu$.

Therefore, given all other firms strategy $q_j \leq k_j, j \neq i$, firm i's best response decision problem Eq. (4.15) can be simplified as

$$\max_{q_i} q_i W^{-1}\Big(\max\big(\sum_{j}^{N}(k_j - q_j), \frac{1}{\mu}\big); \mu_{\geq w_0}\Big) \cdot \mathbb{1}_{\{q_i \leq k_i\}} \tag{C.10}$$

Given that $\mu_{\geq w_0} = \mu$, which is a constant, k_j is just a simple constant term $\mu_j/(\mu w_0)(v_j - w_0)t_j$. Thus, the objective function in Eq. (C.10) only depends on the value of q_i and the sum of all other firms' profit $\sum_{j \neq i} q_j$. We approach the

problem by defining the following "unconstrained" problem:

$$\pi(q; q_s) = qW^{-1}(\sum_{i=1}^{N} k_i - q_s - q) \tag{C.11}$$

and the corresponding best response function

$$h(q_s) = \arg\max_q \pi(q; q_s)$$

In other words, $h(q_s)$ is the optimal profit when all other firms collectively take q_s profit if we ignore the market entry constraint that firm i's profit drops to 0 if q_i exceed k_i. Clearly, when ignoring the condition for the lowest pay that workers will take a job, the best response problem Eq. (C.11) is the same for all firms. Thus, for the best response decision problem Eq. (C.10), given all other firms' profit $q_s = \sum_{j\neq i} q_j$, if $h(\sum_{j\neq i} q_j)$ happens to be below k_i, then firm i's best response is just $h(\sum_{j\neq i} q_j)$; if $h(\sum_{j\neq i} q_j)$ is above k_i, it means firm i can potentially be better off taking a profit above k_i, but with the constraint that workers will not take a job if $q_i > k_i$, firm i can only choose the highest profit rate possible, which is just k_i. Thus, the best response for Eq. (C.10) is just equivalent to

$$\min\left(h(\sum_{j\neq i} q_j), k_i\right) \tag{C.12}$$

In Appendix C.6, we prove some technical lemmas about the unconstrained problem $\pi(q; q_s)$ and $h(q_s)$, which we will refer to in the following sections.

Proofs by steps Next we prove the theorem in the following steps:

1. Under the condition that $\sum_1^N \mu_i(v_i - w_0)t_i > w_0$,

   (a) Eq. (C.7) and Eq. (C.8) uniquely define a set of strategies with $0 < q_i^* \leq k_i$, $\forall i$ and $\lambda > 0$.

(b) Any Nash equilibrium with $0 < q_i^* \leq k_i$, $\forall i$ and $\lambda > 0$ must satisfy Eq. (C.7) and Eq. (C.8).

(c) The strategies defined by Eq. (C.7) and Eq. (C.8) is a Nash equilibrium.

2. Under the condition that $\sum_1^N \mu_i(v_i - w_0)t_i \leq w_0$, there does not exist a Nash equilibrium with $0 < q_i^* \leq k_i$, $\forall i$ and $\lambda > 0$ at the same time.

Step 1a: Define

$$g(\theta) = h(\sum_1^N \min(\theta, k_i) - \theta) - \theta, 0 \leq \theta \leq \bar{\theta} \tag{C.13}$$

Note that $g(\theta)$ is well defined on $[0, \bar{\theta}]$. The reason is that for any $0 \leq \theta \leq \bar{\theta}$, $\sum_1^N \min(\theta, k_i) - \theta \leq \sum_1^N \min(\bar{\theta}, k_i) - \theta = \sum_{i=1}^N k_i - W(0) - \theta \leq \sum_{i=1}^N k_i - W(0)$, which means $(\sum_1^N \min(\theta, k_i) - \theta)$ is in the domain of h.

Next, we prove 1a by showing that $g(\theta)$ is (1) continuous on $[0, \bar{\theta}]$, (2) strictly decreasing in θ, and (3) $g(0) > 0$, $g(\bar{\theta}) < 0$. Then by the Intermediate Value Theorem, there is a unique root on $(0, \bar{\theta})$ such that $g(\theta) = 0$, which is just the θ^* given in Eq. (C.8). Lastly, in (4), we verify that in the equilibrium given by Eq. (C.7) and Eq. (C.8), all firms participate and are profitable.

(1) For $\theta \in [0, \bar{\theta}]$, $0 \leq \sum_1^N \min(\theta, k_i) - \theta < \sum_{i=1}^N k_i - W(0)$ and is continuous in θ. Thus, the continuity of g is implied by the continuity of h (Lemma 19).

(2) To show $g(\theta)$ being decreasing on $\theta \in [0, \bar{\theta}]$, since $(-\theta)$ is strictly decreasing in θ, it is sufficient to show $h(\sum_1^N \min(\theta, k_i) - \theta)$ is non-increasing in θ. Moreover, by Lemma 19, h is strictly decreasing. Therefore, it is equivalent to show that $(\sum_1^N \min(\theta, k_i) - \theta)$ is non-decreasing in θ over $[0, \bar{\theta}]$. To see this, we

rewrite the expression as

$$\sum_1^N \min(\theta, k_i) - \theta = \sum_{\theta \le k_i} \theta + \sum_{\theta > k_i} k_i - \theta = \begin{cases} (N-1)\theta & \theta \le k_N \\ (N-2)\theta + k_N & \theta \le k_{N-1} \\ \dots & \\ \sum_2^N k_i & \theta \le k_1 \end{cases}$$

From the above expression, it is easy to see that $(\sum_1^N \min(\theta, k_i) - \theta)$ is non-decreasing in θ when $\theta \le k_1$, as the expression is continuous, piece-wise affine with non-negative slopes. Thus, we only need to show $\bar{\theta} \le k_1$ to prove $(\sum_1^N \min(\theta, k_i) - \theta)$ being non-decreasing on $[0, \bar{\theta}]$. To see this, note that $\sum_1^N \min(\bar{\theta}, k_i) = \sum_{i=1}^N k_i - W(0)$ (by definition) and $\sum_1^N \min(k_1, k_i) = \sum_{i=1}^N k_i$. Thus,

$$\sum_1^N \min(\bar{\theta}, k_i) < \sum_1^N \min(k_1, k_i)$$

Since the min function is non-decreasing, we have $\bar{\theta} \le k_1$.

(3) As $\sum_{i=1}^N k_i - W(0) > 0$, it follows from Lemma 19 that $h(0) > 0$. Moreover, by Lemma 19 point 1 , $h(q_s) + q_s < \sum_{i=1}^N k_i - W(0)$. Thus, let $q_s = \sum_1^N \min(\bar{\theta}, k_i) - \bar{\theta}$, and then we have

$$h(\sum_1^N \min(\bar{\theta}, k_i) - \bar{\theta}) + \sum_1^N \min(\bar{\theta}, k_i) - \bar{\theta} < \sum_{i=1}^N k_i - W(0)$$

Note that by definition, $\min(\bar{\theta}, k_i) = \sum_{i=1}^N k_i - W(0)$. Thus, we have

$$g(\bar{\theta}) = h(\sum_1^N \min(\bar{\theta}, k_i) - \bar{\theta}) - \bar{\theta} < 0$$

Thus, $g(0) = h(0) - 0 > 0$. $g(\bar{\theta}) < 0$.

(4) In this step, we verify that in the equilibrium given by Eq. (C.7) and Eq. (C.8), all firms participate and are profitable, i.e. $0 < q_i^* \le k_i$, $\forall i$ and $\lambda > 0$.

In (3), we have shown that $\theta^* > 0$. Since q_i^* is the min of θ^* and k_i, the first condition must be true. For the second condition, by the definition of λ in Eq. (4.13), $\lambda > 0$ if and only if $\sum_{i=1}^N k_i - \sum_1^N q_i \geq W(0)$. Then for the q_i^* given by Eq. (C.7), by $\theta^* \leq \bar{\theta}$,

$$\sum_{i=1}^N k_i - \sum_1^N q_i^* = \sum_{i=1}^N k_i - \sum_1^N \min(\theta^*, k_i) \geq \sum_{i=1}^N k_i - \sum_1^N \min(\bar{\theta}, k_i) = W(0)$$

The inequality is by the fact that min is a non-decreasing function; the rightmost equality is by the definition of $\bar{\theta}$. Therefore, $\lambda > 0$ is also verified.

Step 1b: By definition, if $(q_1^*, ..., q_N^*)$ is a Nash Equilibrium while $0 < q_i^* \leq k_i$ and $\lambda > 0$, each firms' marginal profit q_i^* must be a best response to the other firms' choices, and $(\sum_{i=1}^N k_i - \sum_1^N q_i^*)$ should be sufficiently high to support a positive λ. In other words, it must be true that:

$$\sum_{i=1}^N k_i - \sum_1^N q_i^* > W(0) \tag{C.14}$$

and

$$q_i^* = \arg\max_{q_i \leq k_i} q_i W^{-1}(\sum_{i=1}^N k_i - \sum_{j \neq i} q_j^* - q_i) = \arg\max_{q_i \leq k_i} \pi(q_i | \sum_{j \neq i} q_j^*), \forall i \tag{C.15}$$

By Lemma 18 and Lemma 19, $\pi(q_i | \sum_{j \neq i} q_j^*)$ is a strictly concave function, and as we assume $\sum_{i=1}^N k_i - W(0) > 0$, it is maximized at $h(\sum_{j \neq i} q_j^*)$. With the constraint that $q_i \leq k_i$, the maximizer either remains at the unconstrained maximizer $h(\sum_{j \neq i} q_j^*)$ (when $h(\sum_{j \neq i} q_j^*) < k_i$), or the constraint is binding (when $h(\sum_{j \neq i} q_j^*) \geq k_i$). Thus, (C.15) is equivalent to

$$q_i^* = \min(h(\sum_{j \neq i} q_j^*), k_i), \forall i \tag{C.16}$$

Thus, q_i^* is either k_i or strictly less than k_i, and when q_i^* is strictly less than k_i, $q_i^* = h(\sum_{j \neq i} q_j^*)$. We now show that all $q_i^* < k_i$ must be equal (there must

exists at least one i such that $q_i^* < k_i$, otherwise Eq. (C.14) cannot hold). To see it, note that $q_i^* = h(\sum_{j\neq i} q_j^*) = h(\sum_j q_j^* - q_i^*)$. Using Lemma 19, and knowing that $q_i^* > 0$, this is equivalent to :

$$q_i^* = \frac{W^{-1}(\sum_{i=1}^N k_i - \sum_{j\neq i} q_j^* - h(\sum_{j\neq i} q_j^*))}{(W^{-1})'(\sum_{i=1}^N k_i - \sum_{j\neq i} q_j^* - h(\sum_{j\neq i} q_j^*))} \tag{C.17}$$

$$= \frac{W^{-1}(\sum_{i=1}^N k_i - \sum_{j\neq i} q_j^* - q_i^*)}{(W^{-1})'(\sum_{i=1}^N k_i - \sum_{j\neq i} q_j^* - q_i^*)} \tag{C.18}$$

$$= \frac{W^{-1}(\sum_{i=1}^N k_i - \sum_1^N q_i^*)}{(W^{-1})'(\sum_{i=1}^N k_i - \sum_1^N q_i^*)} \tag{C.19}$$

Since the right-hand side is the same for all i, it must be true that there is a unique value for q_i^* when the constraint $q_i^* \leq k_i$ is not binding. We name the corresponding value as γ. Then it must be true that

$$q_i^* = \min(\gamma, k_i), \forall i \tag{C.20}$$

Moreover, as we have shown previously, any $q_i^* < k_i$ must satisfy $q_i^* = h(\sum_j q_j^* - q_i^*)$. Replacing q_i^* with γ and q_j^* with $\min(\gamma, k_j)$, it implies

$$\gamma = h(\sum_j \min(\gamma, k_j) - \gamma) \tag{C.21}$$

(C.20) and (C.21) are just identical to (C.7) and (C.8).

Step 1c: We show that for any firm i, given the strategies from other firms $q_j^* = \min(\theta^*, k_j), j \neq i$, the best response is $q_i^* = \min(\theta^*, k_i)$. Note that θ^* is given by (C.8).

There are two options for firm i: either selecting a high profit margin q_i and making the remaining surplus too low to support a positive λ (i.e. $q_i + \sum_{j\neq i} q_j^* \geq \sum_{i=1}^N k_i - W(0)$), or selecting a lower profit margin and maintaining a positive

λ (i.e. $q_i + \sum_{j\neq i} q_j^* < \sum_{i=1}^N k_i - W(0)$). In the first option, firm i earns zero profit. In the second option, given that $q_j^* = \min(\theta^*, k_j), j \neq i$, firm i can always make a strictly positive profit. In other words, there exists $0 < q_i \leq k_i$ such that $\lambda > 0$. This is true because we have shown in Item 1a that $\sum_{j\neq i} \min(\theta^*, k_j) < \sum_{i=1}^N k_i - W(0)$, thus making it feasible for firm i to maintain a positive λ and a positive profit margin q_i at the same time. Thus, given $q_j^* = \min(\theta^*, k_j), j \neq i$, firm i will always choose the second option.

Next, we show that in the second option, firm i's best response is $q_i^* = \min(\theta^*, k_i)$. Eq. (C.15) and (C.16) continue to apply here:

$$q_i^* = \arg\max_{q_i \leq k_i} q_i W^{-1}(\sum_{i=1}^N k_i - \sum_{j\neq i} q_j^* - q_i) = \min(h(\sum_{j\neq i} q_j^*), k_i)$$

Thus,

$$h(\sum_{j\neq i} q_j^*) = h(\sum_{j\neq i} \min(\theta^*, k_j)) = \begin{cases} h(\sum_1^N \min(\theta^*, k_i) - \theta^*), & k_i > \theta^* \\ h(\sum_1^N \min(\theta^*, k_i) - k_i), & k_i \leq \theta^* \end{cases}$$

Thus, for $k_i > \theta^*$, $h(\sum_{j\neq i} q_j^*) = h(\sum_1^N \min(\theta^*, k_i) - \theta^*) = \theta^*$, by the definition of θ^*. For $k_i \leq \theta^*$, we first show an intermediate result: $h(\sum_1^N \min(\theta^*, k_i) - k_i) \geq k_i$. The reason is the following: Let $x_1 = \sum_1^N \min(\theta^*, k_i) - k_i$ and $x_2 = \sum_1^N \min(\theta^*, k_i) - \theta^*$. Thus,

$$k_i \leq \theta^* \Leftrightarrow x_1 \geq x_2$$

By Lemma 19, $h(x_1) + x_1 \geq h(x_2) + x_2$, which gives

$$h(\sum_1^N \min(\theta^*, k_i) - k_i) - k_i \geq h(\sum_1^N \min(\theta^*, k_i) - \theta^*) - \theta^*$$

By the definition of θ^*,

$$h(\sum_1^N \min(\theta^*, k_i) - \theta^*) - \theta^* = 0 \Rightarrow h(\sum_1^N \min(\theta^*, k_i) - k_i) - k_i \geq 0$$

Thus, we have

$$h(\sum_{j\neq i}\min(\theta^*,k_j)) = \theta^*, k_i > \theta^*$$
$$h(\sum_{j\neq i}\min(\theta^*,k_j)) \geq k_i, k_i \leq \theta^*$$

This implies that for $k_i > \theta^*$,

$$\min(h(\sum_{j\neq i}\min(\theta^*,k_j)),k_i) = \min(\theta^*,k_i) = \theta^*$$

and for $k_i \leq \theta^*$,

$$\min(h(\sum_{j\neq i}\min(\theta^*,k_j)),k_i) = k_i$$

Therefore,

$$q_i^* = \min(h(\sum_{j\neq i} q_j^*),k_i) = \min(h(\sum_{j\neq i}\min(\theta^*,k_j)),k_i) = \min(\theta^*,k_i)$$

Proof of 2 When $\sum_{i=1}^N k_i \leq W(0)$, whenever there exists a $0 < q_i^* \leq k_i$, it must be true that

$$\sum_{i=1}^N k_i - \sum_1^N q_i^* < W(0)$$

which means $\lambda > 0$ cannot be true. Therefore, if $\sum_{i=1}^N k_i \leq W(0)$, $0 < q_i^* \leq k_i$ and $\lambda > 0$ cannot happen at the same time. □

C.4.3 Existence and expressions for the Nash equilibria with an unviable market

Proof. (1) First show the existence of a Nash equilibrium with $\lambda = 0$ when $\mu_i \leq \frac{w_0}{t_i(v_i - w_0)}$ for all i. We show this by proving the special case with all firms

setting payment strictly below w_0 ($p_i < w_0, \forall i$) is a Nash equilibrium under the given condition.

Given all $p_j < w_0, j \neq i$, if $p_i < w_0$, then the profit for firm i is just 0. If $p_i \geq w_0$, then

$$\sum_{p_k \geq w_0} \mu_k (v_k - w_0) t_k = \mu_i (v_i - w_0) t_i \leq w_0$$

Therefore, $\lambda = 0$, and the profit is also 0. Thus, for firm i, not participating is a best response.

(2) Then show that if there is an equilibrium with $\lambda = 0$, it must be true that $\mu_i \leq \frac{w_0}{t_i(v_i - w_0)}$ for all i. Prove by contradiction. Suppose not. Then there exists a Nash equilibrium with $\lambda = 0$ and there exists an M such that $\mu_M > \frac{w_0}{t_M(v_M - w_0)}$. We show that firm M can earn strictly above 0 by increasing its payment such that $\lambda > 0$. By definition, $\lambda = 0$ if and only if

$$\sum_{p_i \geq w_0} \mu_i (p_i - w_0) t_i = \mu_M (p_M - w_0)^+ t_M + \sum_{i \neq M, p_i \geq w_0} \mu_i (p_i - w_0) t_i \leq w_0 \quad \text{(C.22)}$$

But $\mu_M (v_M - w_0) t_M > w_0$. Therefore, it must be true that the p_M in (C.22) is less than v_M. Moreover, let $p_M = v_M - \epsilon$, where $\epsilon > 0$ is infinitely small. Then

$$\sum_{p_i \geq w_0} \mu_i (p_i - w_0) t_i = \mu_M (v_M - \epsilon - w_0) t_M + \sum_{i \neq M, p_i \geq w_0} \mu_i (p_i - w_0) t_i > w_0$$

which leads to $\lambda > 0$. Thus, by letting $p_M = v_M - \epsilon$, it will hold that $\lambda > 0$ and firm M is better off.

3. If (C.9) holds for all i, then for any firm, given the $p_j, j \neq i$ in (C.9), $\lambda > 0$ can only happen when $p_i > v_i$. Thus, to have $\lambda > 0$, firm i's profit $\pi = \frac{\mu}{\mu_{\geq w_0}} \lambda (v_i - p_i) t_i < 0$. Therefore, choosing p_i such that $\lambda = 0$ is a best response.

If (C.9) does not hold, i.e. for some $i = M$,

$$\mu_M(v_M - w_0)t_M + \sum_{j \neq M, p_j \geq w_0} \mu_j(p_j - w_0)t_j > w_0$$

Then firm M could earn strictly positive profit by setting price $p_M = v_M - \epsilon$ as shown in 2. Therefore, in a Nash equilibrium, (C.9) must hold. □

C.5 Theorems in Section 4.4: a special case of the general theorem

Proposition 24 is Part 1 in Appendix C.3. Theorem 7 and Theorem 8 are just Part 2 and Part 2 in Appendix C.3. To show the existence and uniqueness of the equilibria as well as the conditions for the two types of equilibria (Theorem 7, Theorem 8) to hold, it suffices to prove that the waiting function used in Section 4.4, $W = \frac{1}{\mu_{\geq w_0} - \lambda}$, satisfies the regularity conditions of the waiting function stated in Theorem 10. In addition, we illustrate how the equilibrium expressions in Part 2, Appendix C.3 translate to the properties in Theorem 7.

Since W in is a simple reciprocal function, it is straightforward to see that $\lim_{\lambda \to \mu_{\geq w_0}} W = +\infty$ and its differentiability on $[0, \mu_{\geq w_0})$. Moreover, its first-order and second-order derivatives in λ are given by:

$$W' = \frac{1}{(\mu_{\geq w_0} - \lambda)^2}, W'' = \frac{2}{(\mu_{\geq w_0} - \lambda)^3}$$

Thus, for $\lambda \in [0, \mu_{\geq w_0})$, $W' > 0$, $W'' > 0$, which confirms that W is strictly increasing, convex and W' is nonzero on $[0, \mu_{\geq w_0})$. At $\lambda = 0$, $W = \frac{1}{\mu_{\geq w_0}}$. Therefore, all conditions for the waiting function stated in Theorem 10 are satisfied.

Then we show that the Nash equilibrium characterized by Part 2 in the general theorem Theorem 10 can be translated into the properties in Theorem 7:

1. Existence and uniqueness: Condition (4.10) can be obtained by replacing t_i, v_i in condition (C.6) with T and v.

2. Larger firms pay more: Again let $v_1 = v_2 = ...v_N = v$ and $t_1 = t_2 = ... = t_N$ and suppose $\mu_1 \geq \mu_2... \geq \mu_N$. Then $\mu_1 \geq \mu_2... \geq \mu_N$ is equivalent to $k_1 \geq k_2... \geq k_N$. Moreover, by Eq. (C.7), $q_i^* = \min(\theta^*, k_i)$. Suppose $k_M \geq \theta^* > k_{M+1}$. Then for $i <= M$, $q_1^* = q_2^* = ... = q_M^* = \theta^* \leq k_M$, implying that $\mu_1(v - p_1^*) = \mu_2(v - p_2^*) = ... = \mu_M(v - p_N^*) \leq \mu_M(v - w_0)$, which further implies that $p_1^* \geq p_2^*... \geq p_M^* \geq w_0$. For $i > M$, by Eq. (C.7), $q_i^* = k_i$ which implies $p_i^* = w_0$. Thus, we have $p_1^* \geq p_2^*... \geq p_N^*$.

3. Same hourly profit for larger firms:. In equilibrium, for any firm with $p_i^* > w_0$, it must also be true that $q_i^* < k_i$, which is implied by the definition of q_i and k_i. By Eq. (C.7), $q_i^* = \min(\theta^*, k_i)$ and θ^* is the same for all i. Thus, for any firm with $q_i^* < k_i$, it must be true that $q_i^* = \theta^*$. Thus, all firms with $p_i^* < w_0$ has the same profit per supply θ^*. The hourly profit rate is just the product of θ^* and the supply level $\lambda(\mathbf{p}^*)$, which is also the same for all firms with $p_i^* < w_0$.

4. Smaller firms free ride: In equilibrium, any firm that does not pay $p_i^* > w_0$ will pay exactly w_0. the reason is that $p_i^* \leq w_0$ is equivalent to $q_i^* \geq k_i$. By Eq. (C.7), in equilibrium, it must be true that $q_i^* = k_i$, which implies $p_i^* = w_0$. This gives the firms the highest profit margin $(v - w_0)$.

C.6 Technical Lemmas

Our first Lemma describes W^{-1}, the inverse function of W that associates any given average wait time $W = t$ with a corresponding completed job rate $\lambda = W^{-1}(t; \mu_{\geq w_0})$, given the service rate $\mu_{\geq w_0}$. Note that for simplicity, we omit the parameter $\mu_{\geq w_0}$ in some parts of the proof, whenever there is no ambiguity.

Lemma 17 *For a given $\mu_{\geq w_0}$, assume that $W(\lambda; \mu_{\geq w_0})$ is strictly increasing, twice differentiable and strictly convex on $[0, b)$. Moreover, $\lim_{\lambda \to b} W(\lambda; \mu_{\geq w_0}) = +\infty$, and the derivative $W'(\lambda; \mu_{\geq w_0})$ is non-zero on $[0, b)$. Then we have the following:*

- *The inverse of W exists on $[W(0; \mu_{\geq w_0}), +\infty)$.*
- *Denote the inverse function as W^{-1}, i.e. for any $t = W(\lambda; \mu_{\geq w_0})$, $\lambda = W^{-1}(t; \mu_{\geq w_0})$*
- *$W^{-1}(t; \mu_{\geq w_0})$ is strictly increasing, twice-differentiable, strictly concave and non-negative for $t \in [W(0), +\infty)$.*
- *$(W^{-1})'(t; \mu_{\geq w_0}) > 0$ for $t \in [W(0; \mu_{\geq w_0}), +\infty)$.*

Proof. The existence, continuity, differentiability and monotonicity of W^{-1} are implied by the Inverse Function Theorem, and the strict concavity of W^{-1} is a consequence of W being strictly convex and strictly increasing. Since W^{-1} is strictly increasing, for any $t > W(0)$, $W^{-1}(t) > W^{-1}(W(0)) = 0$. Thus, $W^{-1}(t)$ is non negative in its domain $[W(0), +\infty)$. As $W' > 0$, by the Inverse Function Theorem, $(W^{-1})'(t) = \frac{1}{W'(\lambda)} > 0$ for $t \in [W(0), +\infty)$. □

We next turn to a firm's i profit maximization problem. In the profit maximization formulation of Section 4.5, if we fix $q_s = \sum_{j \neq i} q_i$, then the profit of firm

i is given by the formula $q_i\lambda = q_i W^{-1}(\sum_{j=1}^{n} k_j - q_s - q_i)$. Firm i will optimize q_i to maximize this quantity, and we need to understand well this optimization problem. This is done with the following technical Lemma.

Lemma 18 *Let*

$$\pi(q; q_s) = qW^{-1}\left(\sum_{i=1}^{N} k_i - q_s - q\right)$$

where $W^{-1}(\cdot)$ is the inverse function of W. Then for any finite number q_s, we have the following:

1. *$q \to \pi(q; q_s)$ is defined on $(-\infty, \sum_{i=1}^{N} k_i - q_s - W(0)]$.*
2. *$q \to \pi(q; q_s)$ is strictly concave.*
3. *If $q_s < \sum_{i=1}^{N} k_i - W(0)$, $\pi(q; q_s)$ has a unique maximizer on the open interval $(0, \sum_{i=1}^{N} k_i - q_s - W(0))$ and it can be obtained by solving $\frac{\partial \pi}{\partial q} = 0$.*

Proof. We prove each of these properties in turn.

Proof of 1 W^{-1} is defined on $[W(0), +\infty)$. Thus, $\sum_{i=1}^{N} k_i - q_s - q \geq W(0)$, which gives $q \leq \sum_{i=1}^{N} k_i - q_s - W(0)$.

Proof of 2 We prove the strict concavity by showing $\pi''(q; q_s) < 0$. Given q_s, let $t(q) = \sum_{i=1}^{N} k_i - q_s - q$. $\pi(q; q_s) = qW^{-1}(t(q))$. By the chain rule,

$$\pi'(q; q_s) = W^{-1}(t(q)) + q(W^{-1})'(t(q))t'(q) = W^{-1}(t(q)) - q(W^{-1})'(t(q))$$

$$\begin{aligned}\pi''(q; q_s) &= (W^{-1})'(t(q))t'(q) - (W^{-1})'(t(q)) - q(W^{-1})''(t(q))t'(q) \\ &= -2(W^{-1})'(t(q)) + q(W^{-1})''(t(q)) \quad \text{(C.23)}\end{aligned}$$

By Lemma 17, $(W^{-1})'(t(q)) > 0$ and $(W^{-1})''(t(q)) \leq 0$. Thus, $\pi''(q; q_s) < 0$.

Proof of 3 If $q < \sum_{i=1}^{N} k_i - W(0)$, the profit $\pi(q; q_s)$ is well defined for $q = 0$ and we have $\pi(0|q_s) = 0$. We also have $\pi(q; q_s) = 0$ when $q = \sum_{i=1}^{N} k_i - q_s - W(0)$, as we have:

$$W^{-1}(\sum_{i=1}^{N} k_i - q_s - q) = W^{-1}(\sum_{i=1}^{N} k_i - q_s - (\sum_{i=1}^{N} k_i - q_s - W(0))) = W^{-1}(W(0)) = 0$$

By the strict concavity of $\pi(q; q_s)$, for any $0 < q < \sum_{i=1}^{N} k_i - q_s - W(0)$, $\pi(q; q_s) > 0$, and by continuity of π there exists a maximum of π in this interval. Using the strict concavity of π, we know that this maximum is unique. □

We have established the technical details that will help us define the firms' best response to the other firms prices. The following technical Lemma shows how a firm's best response varies as a function of the other firms' choices.

Lemma 19 *Assume* $\sum_{i=1}^{N} k_i > W(0)$. *Let* $h(q_s) = \arg\max_q \pi(q; q_s)$, $q_s \in [0, \sum_{i=1}^{N} k_i - W(0)]$, *where* $\pi(q; q_s)$ *is defined by Lemma 18. Then we have the following:*

1. $h(q_s) = \frac{W^{-1}(\sum_{i=1}^{N} k_i - q_s - h(q_s))}{(W^{-1})'(\sum_{i=1}^{N} k_i - q_s - h(q_s))}$, *and* $0 < h(q_s) < \sum_{i=1}^{N} k_i - q_s - W(0)$.
2. $h(q_s)$ *is continuous.*
3. $h(q_s)$ *is strictly decreasing, and* $h(q_s) + q_s$ *is strictly increasing.*

Proof. We prove the points of the Lemma one by one.

Proof of Item 1 By Lemma 18, given $q_s \leq \sum_{i=1}^{N} k_i - W(0)$, $h(q_s)$ is just the q such that $\pi'(q; q_s) = 0$.

$$\pi'(h(q_s)|qs) = 0 \Leftrightarrow W^{-1}(\sum_{i=1}^{N} k_i - q_s - h(q_s)) - h(q_s)(W^{-1})'(\sum_{i=1}^{N} k_i - q_s - h(q_s)) = 0$$

Rearranging the terms gives the expression in Lemma 19. Note that $h(q_s)$ is well-defined as $(W^{-1})'(t)$ is nonzero by Lemma 17. Moreover, $0 < h(q_s) < \sum_{i=1}^{N} k_i - q_s - W(0)$ is a direct result of property 3 in Lemma 18.

Proof of Item 2 We first show the continuity of h on $q_s \in [0, \sum_{i=1}^{N} k_i - W(0)]$ by applying Berge's Maximum Theorem. To apply the Maximum Theorem, we set up the problem in the following way:

Let $f(q, q_s) = \pi(q; q_s)$. $f : Q \times Q_s \to \mathbb{R}$ and is continuous. $Q = Q_s = [0, \sum_{i=1}^{N} k_i - W(0)]$. Let $C(q_s) = [0, \sum_{i=1}^{N} k_i - q_s - W(0)]$. The goal is to show $C^*(q_s)$ is continuous in q_s, where

$$C^*(q_s) = \arg\max\{f(q, q_s) : q \in C(q_s)\}$$

By Berge's Maximum Theorem, $C^*(q_s)$ is continuous if a) $C(q_s) \neq \emptyset$ and $C(q_s)$ is compact for all $q_s \in Q_s$, b) f is a continuous function, and c) C is upper and lower hemicontinuous. a) and b) are easy to verify by Lemma 18 and c) is continuous by construction. Therefore, we have the continuity of h on $[0, \sum_{i=1}^{N} k_i - W(0)]$.

Proof of Item 3 We first use a proof by contradiction to show $h(q_s)$ is strictly decreasing. Let $0 \leq y < x \leq \sum_{i=1}^{N} k_i - W(0)]$, and suppose $h(x) \geq h(y)$. Thus, $x + h(x) > y + h(y)$. Moreover, $x + h(x)$ and $y + h(y)$ both lie in the open interval $(0, \sum_{i=1}^{N} k_i - W(0))$, as suggested by point 1, which makes $(\sum_{i=1}^{N} k_i - (x + h(x)))$ and $(\sum_{i=1}^{N} k_i - (y + h(y))$ within the domain of function W^{-1}. By Lemma 17,

W^{-1} is strictly increasing and positive, and $(W^{-1})'$ is strictly decreasing and positive. Thus,

$$W^{-1}(\sum_{i=1}^{N} k_i - (x + h(x))) < W^{-1}(\sum_{i=1}^{N} k_i - (y + h(y)))$$

$$(W^{-1})'(\sum_{i=1}^{N} k_i - (x + h(x))) > (W^{-1})'(\sum_{i=1}^{N} k_i - (y + h(y)))$$

Then by the expression of h given by Property 1 in Lemma 19, $h(x) < h(y)$, a contradiction.

Now we show $q_s + h(q_s)$ is strictly increasing, again using a proof by contradiction. We still let $0 \leq y < x \leq \sum_{i=1}^{N} k_i - W(0)]$, and now suppose $x + h(x) \leq h(y) + y$. Similarly, $(x + h(x))$ and $(h(y) + y)$ are still within the domain of W^{-1}. Then by the definition of h,

$$h(x) = \frac{W^{-1}(\sum_{i=1}^{N} k_i - (x + h(x)))}{(W^{-1})'(\sum_{i=1}^{N} k_i - (x + h(x)))} \geq \frac{W^{-1}(\sum_{i=1}^{N} k_i - (y + h(y)))}{(W^{-1})'(\sum_{i=1}^{N} k_i - (y + h(y)))} = h(y)$$

The inequality is again because W^{-1} is strictly increasing and positive, and $(W^{-1})'$ is strictly decreasing and positive. But this cannot be true, because we have already proved that h is strictly decreasing, and we assume $x > y$, i.e. $h(x) < h(y)$ must hold. □

C.7 Data sources for the market share and time of market entry in ride-hailing platforms (India, Indonesia and China)

- India: the market share data for leading companies are according to this Quarz article: *As Uber sputters, Ola is really stepping on the gas in India*, Quarz, `https://qz.com/india/1200878/with-uber-in-crisis-ola-`

zooms-ahead-in-indias-taxi-wars/. References for the market entry year for each company is listed in the "Source" column below.

Table C.3: Share of ride hailing market across India as of December 2017 and leading companies' year of market entry, by company.

	Market share	Market entry time	Source
Ola	56.2%	Dec. 2010	Crunchbase
Uber	39.6%	Aug. 2013	Uber blog
Jugnoo	2.5%	Nov. 2014	Crunchbase
Meru	0.8%	2014*	Economic Times
ixigo Cabs	0.9%	July 2015	startupleadership.com

Note: 2014 is the year Meru introduced app-based ride-hailing service; the company was founded much earlier in 2007 as a call-based taxi company.

- Indonesia: the market share data for leading companies are according to this article: *The market share of Gojek is almost 80 percent*, industrycoid, http://en.industry.co.id/read/6771/the-market-share-of-gojek-is-almost-80-percent. References for the market entry year for each company is listed in the "Source" column below.

Table C.4: Market share of the ride-hailing transportation industry in Indonesia and their year of market entry as of April 2018, by company

	Market share	Market entry time	Source
Go-Jek	79.2%	2015*	Reuters
Grab	14.69%	June 2012	Crunchbase
Uber*	6.11%	N/A	N/A

Note: 2015 is the year Go-jek introduced app-based ride-hailing services; the company was founded in 2009. The time when Uber entered Indonesia market is not available; It exit Indonesia market in April 2018.

- China: the market share data for leading companies are according to this report: *China internet report 2019*, page 58, south China Morning Post, https://multimedia.scmp.com/infographics/china-internet/pdf/china_internet_report_2019.pdf. References for the market entry year for each company is listed in the "Source" column below.

Table C.5: Market share of ride-sharing market in china in 4th quarter 2018 and the year of market entry, by company

	Market share	Market entry time	Source
Didi	91%	Sept. 2012	Crunchbase
Shouqi	2%	2015	Crunchbase
Meituan	2%	Feb. 2017	Financial Times
Shenzhou	2%	Jan. 2015	Owler
Caocao	2%	Nov. 2015	South China Morning Post
Yidao	1%	May 2010	Crunchbase

- Latin America: market shares are computed based on active user numbers for leading players by Techcrunch: Latin America is the next stage in the race for dominance in the ride-hailing market, Techcrunch, `https://techcrunch.com/2018/09/07/latin-america-is-the-next-stage-in-the-race-for-dominance-in-the-ride-hailing-market/`. References for the market entry year for each company is listed in the "Source" column below.

Table C.6: Market share of ride-sharing market in latin America 2018 and the year of market entry, by company

	Market share	Market entry time	Source
Uber	57.14%	2013	Techcrunch
Didi/99	22.22%	2012*	Crunchbase
Cabify	20.63%	2012	CoMotion News

Note: Didi acquired 99 in 2018. Year 2012 is the time that 99 entered the Latin America market

APPENDIX D

CONSPICUOUS CONSUMPTION IN THE PRESENCE OF NON-DECEPTIVE COUNTERFEITS

D.1 Process

The step of discovering counterfeits with probability φ can be considered as nature's move. After nature's move, the spectator updates her belief about the consumer, and then generates an inference w'. Note that here we didn't model the spectator's utility explicitly. The spectator does not strategically choose a w'; rather, she just passively generates inferences following the Bayes rule.

D.2 Proofs for the market with no counterfeit risk

Proposition 26

Proof. Consumer's utility function is

$$U(w, x) = x + u(w - px)$$

Take first and second order derivatives over x:

$$U' = 1 - pu'(w - px)$$

$$U'' = p^2 u''(w - px) - u'(w - px) < 0$$

Therefore, U is strictly concave in x. Setting U' to 0 gives the unique global maximizer. Since the consumption must be non-negative, if $\frac{w}{p} - \frac{1}{p} z^0(\frac{1}{p}) \geq 0$, then it is the maximizer for U; if not, 0 is the maximizer.

□

Proposition 27

Proof. We first prove the uniqueness and existence of a PBE. This is shown by proving the single-crossing property:

Suppose $s' > s, b' > b$ and $U(w_L, b', s') \geq U(w_L, b, s)$.Then

$$x' - x + u(w_L - px') - u(w_L - px) + \lambda(s' - s) \geq 0$$

Define function $g(w) = u(w - px') - u(w - px)$. Then $g'(w) = u'(w - px') - u'(w - px) > 0$ because u is strictly concave. Hence, $g(w_H) > g(w_L)$. Then

$$\begin{aligned} x' - x + u(w_H - px') - u(w_H - px) + \lambda(s' - s) \\ > x' - x + u(w_L - px') - u(w_L - px) + \lambda(s' - s) \geq 0 \end{aligned} \quad \text{(D.1)}$$

Rearranging the above inequality gives $U(w_H, x', s') > U(w_H, x, s)$.

Next, we show the expressions of the equilibrium consumption level x_L^* and x_H^*. Since we only have two types of consumers, the Riley outcome is the only equilibrium that survives Intuitive Criterion (Cho and Kreps 1987). The equilibrium outcome is constructed as follows: The low type chooses a consumption level x that maximizes $x + u(w_L - px)$, which is the optimal consumption level given that the spectator makes the correct speculation. As is shown in Proposition 26, the optimal choice for the low type is $x_L^* = x_L^0 = w_L/p - z^0(p)/p$.

Then the high type chooses the unique level of consumption that maximizes their utility, subject to the constraint that the low-type consumer is not better off from mimicking the high-type consumption level:

$$x + u(w_H - px) + \lambda\Delta \quad \text{(D.2)}$$

s.t.

$$x_L^* + u(w_L - px_L^*) \geq x + u(w_L - px) + \lambda\Delta \tag{D.3}$$

Case 1: The high-type consumer does not distort consumption Under the condition that

$$w_L/p + u(z^0(p)) \geq w_H/p + u(w_L - w_H + z^0(p)) + \lambda\Delta \text{ or } z^0(1/p) < \Delta, \tag{D.4}$$

the optimal solution x_H^* of Eq. (D.2) is given by

$$x_H^* = x_H^0 = w_H/p - z^0(p)/p,$$

Rearranging (D.4) gives

$$\lambda \leq \frac{1}{\Delta}\Big\{u(z^0(p)) - u(z^0(p) - \Delta)\Big\} - \frac{1}{p} \text{ or } z^0(1/p) < \Delta \tag{D.5}$$

which gives the condition for the first column in Table 5.1.

Case 2: The high-type consumer over-purchase to prevent mimicking When Eq. (D.5) does not hold, the high-type consumer cannot stop the low-type from mimicking, unless they distort their consumption upward. In that case, it is optimal to purchase exactly the quantity makes Eq. (D.3) binding. When that happens, the equilibrium high-type consumption level x_H^* is just the solution to the following equation:

$$x_H^* + u(w_L - x_H^* p) + \lambda\Delta = w_L/p - z^0(p)/p + u(z^0(p))$$

Our last step is to ensure the existence of x_H^* in the above equation. For the equation to be feasible, it must be true that $\min_x x + u(w_L - xp) + \lambda\Delta \leq$

$w_L/p - z^0(p)/p + u(z^0(p))$. By the concavity of the utility function, for $x \geq x_L^0$, $x + u(w_L - px)$ is decreasing in x. Therefore,

$$\min_x x + u(w_L - px) + \lambda\Delta = w_L/p + \lambda\Delta \leq w_L/p - z^0(p)/p + u(z^0(p))$$

which is equivalent to

$$\lambda \leq \frac{1}{\Delta}\Big\{u(z^0(p)) - \frac{z^0(p)}{p}\Big\} \tag{D.6}$$

This gives the condition in the second column of Table 5.1.

When Eq. (D.6) does not hold, it means the status effect is so high, that the low-type consumer is willing to spend all their wealth on status goods, if the status utility can be gained. Therefore, the optimal equilibrium consumption level for the high-type consumer is exactly the largest quantity the low-type consumer can afford, which is w_L/p. This gives the condition in the third column in Table 5.1. □

D.3 Proofs for the market with counterfeits

Lemma 5

Proof. Suppose the consumer chooses $x > 0, y > 0$, and the spectator infers her status as s. Her utility is:

$$U(w, x, y, s) = x + y + u(w - px - \delta py) + \lambda s$$

Given $y > 0$, the spectator's inference on status only depends on the total amount of visible goods. Therefore, purchasing $x + y$ counterfeits, 0 status goods and y counterfeits, x status goods receive the same status s. Hence,

$$U(w, 0, x + y, s) - U(w, x, y, s) = u(w - \delta p(x + y)) - u(w - px - \delta py) > 0$$

Therefore, for any consumer who purchases x and y units of genuine and counterfeit products,she can do strictly better by purchasing $(x + y)$ units of counterfeits instead. □

Proposition 28

Proof. Conditional on the fact that both types only purchase the status good, the game dynamics is identical to that in the market without the counterfeit good. Thus, the equilibrium outcome is the same as in Proposition 27. □

Proposition 29 (Proof sketch)

Proof. This is the setting where the high-type consumer purchases the status good, and the low-type consumer purchases counterfeit. Similar to the rationale in Proposition 27, the key factor here is whether there is a consumption level at which the low-type consumer loses the interest of mimicking the high-type consumer. If the answer is yes, then there exists a separating equilibrium with the low-type consumer purchases its "private" optimal consumption level (as if there's no status effect), and the high-type consumer keeps their consumption at the level that just discourages the low-type consumer (or, their private optimal level if it is sufficiently large to deter the low-type consumer). This is Case 1 in Proposition 29. When both consumers purchase the status good, this is the only equilibrium.

With the counterfeit, there is an additional possibility; since the counterfeit is cheaper, the low-type consumer has a higher ability to mimic; thus, it is possible that even when the high-type consumer distorts the consumption to the highest level, and the low-type consumer is still willing to match the same consumption quantity. In that case, a pooling equilibrium emerges. This is Case 2 in Proposition 29.

If the counterfeit is extremely cheap, the low-type consumer may have the incentive to purchase more than the high-type consumer, even if it means losing the status utility (imagine $\delta = 0$, then it is optimal for the low-type consumer to purchase an infinite quantity of counterfeits, even if this separates them from the high-type consumer.) In other words, it is a "reverse"-separating equilibrium that the low-type intentionally separates from the high-type consumer. This is Case 3 in Proposition 29.

□

BIBLIOGRAPHY

Moore, Henry Ludwell. 1911. *Laws of wages: An essay in statistical economics*. Macmillan.

Solow, Robert M. 1956. "A contribution to the theory of economic growth". *The quarterly journal of economics* 70 (1): 65–94.

Kaldor, Nicholas. 1957. "A model of economic growth". *The economic journal* 67 (268): 591–624.

Vidale, ML, and HB Wolfe. 1957. "An operations-research study of sales response to advertising". *Operations research* 5 (3): 370–381.

Stigler, George J. 1961. "The economics of information". *Journal of political economy* 69 (3): 213–225.

Olson, Mancur. 1965. "The Logic of Collective Action". *Contemporary Sociological Theory*: 124.

Bass, Frank M. 1969. "A new product growth for model consumer durables". *Management science* 15 (5): 215–227.

McCall, John Joseph. 1970. "Economics of information and job search". *The Quarterly Journal of Economics*: 113–126.

Douglas, George W. 1972. "Price regulation and optimal service standards: The taxicab industry". *Journal of Transport Economics and Policy*: 116–127.

Stigler, George J. 1974. "Free riders and collective action: An appendix to theories of economic regulation". *The Bell Journal of Economics and Management Science*: 359–365.

Kolesar, Peter. 1975. "A model for predicting average fire engine travel times". *Operations Research* 23 (4): 603–613.

Spence, Michael, and David Starrett. 1975. “Most rapid approach paths in accumulation problems”. *International Economic Review*: 388–403.

Groves, Theodore, and John Ledyard. 1977. “Optimal allocation of public goods: A solution to the" free rider" problem”. *Econometrica: Journal of the Econometric Society*: 783–809.

Diamond, Peter A. 1982. “Aggregate demand management in search equilibrium”. *Journal of political Economy* 90 (5): 881–894.

Burdett, Kenneth, et al. 1984. “Earnings, unemployment, and the allocation of time over time”. *The Review of Economic Studies* 51 (4): 559–578.

Frank, Robert H. 1985. “The demand for unobservable and other nonpositional goods”. *The American Economic Review* 75 (1): 101–116.

Kalish, Shlomo. 1985. “A new product adoption model with price, advertising, and uncertainty”. *Management science* 31 (12): 1569–1585.

Dhebar, Anirudh, and Shmuel S Oren. 1986. “Dynamic nonlinear pricing in networks with interdependent demand”. *Operations Research* 34 (3): 384–394.

Mortensen, Dale T, et al. 1986. “Job search and labor market analysis”. *Handbook of labor economics* 2 (15): 02005–02009.

Cho, In-Koo, and David M Kreps. 1987. “Signaling games and stable equilibria”. *The Quarterly Journal of Economics* 102 (2): 179–221.

Oren, Shmuel S, and Rick G Schwartz. 1988. “Diffusion of new products in risk-sensitive markets”. *Journal of Forecasting* 7 (4): 273–287.

Brown, Charles, and James Medoff. 1989. “The employer size-wage effect”. *Journal of political Economy* 97 (5): 1027–1059.

Shapiro, Carl. 1989. “Theories of oligopoly behavior”. *Handbook of industrial organization* 1:329–414.

Cho, In-Koo, and Joel Sobel. 1990. "Strategic stability and uniqueness in signaling games". *Journal of Economic Theory* 50 (2): 381–413.

Wernerfelt, Birger. 1991. "Brand loyalty and market equilibrium". *Marketing Science* 10 (3): 229–245.

Ireland, Norman J. 1994. "On limiting the market for status signals". *Journal of public Economics* 53 (1): 91–110.

Arnott, Richard. 1996. "Taxi travel should be subsidized". *Journal of Urban Economics* 40 (3): 316–333.

Idson, Todd L, and Walter Y Oi. 1999. "Workers are more productive in large firms". *American Economic Review* 89 (2): 104–108.

Petrongolo, Barbara, and Christopher A Pissarides. 2001. "Looking into the black box: A survey of the matching function". *Journal of Economic literature* 39 (2): 390–431.

Caillaud, Bernard, and Bruno Jullien. 2003. "Chicken & egg: Competition among intermediation service providers". *RAND journal of Economics*: 309–328.

Rochet, Jean-Charles, and Jean Tirole. 2003. "Platform competition in two-sided markets". *Journal of the european economic association* 1 (4): 990–1029.

Hagiu, Andrei. 2004. "Two-sided platforms: Pricing and social efficiency". *Available at SSRN 621461*.

Morse, Philip McCord. 2004. *Queues, inventories and maintenance: the analysis of operational systems with variable demand and supply*. Courier Corporation.

Amaldoss, Wilfred, and Sanjay Jain. 2005. "Conspicuous consumption and sophisticated thinking". *Management science* 51 (10): 1449–1466.

Armstrong, Mark. 2006. "Competition in two-sided markets". *The RAND Journal of Economics* 37 (3): 668–691.

Harris, Chris. 2006. *Electricity markets: pricing, structures and economics*. Vol. 328. John Wiley & Sons.

Rochet, Jean-Charles, and Jean Tirole. 2006. "Two-sided markets: a progress report". *The RAND journal of economics* 37 (3): 645–667.

Stevens, Margaret. 2007. "New Microfoundations for the Aggregate Matching Function". *International Economic Review* 48, no. 3 (): 847–868.

Acemoglu, Daron. 2009. *Introduction to Modern Economic Growth*. Princeton New Jersey: Princeton University Press.

Weyl, E Glen. 2010. "A price theory of multi-sided platforms". *American Economic Review* 100 (4): 1642–72.

Heffetz, Ori. 2011. "A test of conspicuous consumption: Visibility and income elasticities". *Review of Economics and Statistics* 93 (4): 1101–1117.

Tereyağoğlu, Necati, and Senthil Veeraraghavan. 2012. "Selling to conspicuous consumers: Pricing, production, and sourcing decisions". *Management Science* 58 (12): 2168–2189.

Zhang, Jie, L. Jeff Hong, and Rachel Q Zhang. 2012. "Fighting strategies in a market with counterfeits". *Annals of Operations Research* 192 (1): 49–66.

Rao, Raghunath S., and Richard Schaefer. 2013. "Conspicuous consumption and dynamic pricing". *Marketing Science* 32 (5): 786–804.

Cullen, Zoë, and Chiara Farronato. 2014. "Outsourcing tasks online: Matching supply and demand on peer-to-peer internet platforms". *Job Market Paper*.

Qian, Yi. 2014. "Brand management and strategies against counterfeits". *Journal of Economics & Management Strategy* 23 (2): 317–343.

Banerjee, Siddhartha, Ramesh Johari, and Carlos Riquelme. 2015. "Pricing in ride-sharing platforms: A queueing-theoretic approach". In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 639–639. ACM.

Chen, Andrew. 2015. https://andrewchen.co/ubers-virtuous-cycle-5-important-reads-about-uber/.

Cho, Soo-Haeng, Xin Fang, and Sridhar Tayur. 2015. "Combating strategic counterfeiters in licit and illicit supply chains". *Manufacturing & Service Operations Management* 17 (3): 273–289.

Qian, Yi, Qiang Gong, and Yuxin Chen. 2015. "Untangling searchable and experiential quality responses to counterfeits". *Marketing Science* 34 (4): 522–538.

ERCOT. 2016. *Inside the Promise: 2016 State of the Grid Report*. https://www.ercot.com/files/docs/2017/02/23/2016_StateoftheGridReport.pdf.

Kabra, Ashish, Elena Belavina, and Karan Girotra. 2016. "Designing promotions to scale marketplaces". PhD thesis, Working paper, INSEAD, Fontainebleau, France.

Solomon, Brian. 2016. "Stealth Startup Juno Will Take On Uber By Treating Drivers Better". *Forbes*. https://www.forbes.com/sites/briansolomon/2016/02/16/stealth-startup-juno-will-take-on-uber-by-treating-drivers-better.

Anand, Priya. 2017. *Uber Burned More Than $1 Million A Week On UberPool To Win San Francisco*. https://www.buzzfeednews.com/article/priya/uber-pool-burn-rate-frisco.

Cachon, Gerard P, Kaitlin M Daniels, and Ruben Lobel. 2017. "The role of surge pricing on a service platform with self-scheduling capacity". *Manufacturing & Service Operations Management* 19 (3): 368–384.

Castillo, Juan Camilo, Dan Knoepfle, and Glen Weyl. 2017. "Surge pricing solves the wild goose chase". In *Proceedings of the 2017 ACM Conference on Economics and Computation*, 241–242. ACM.

Gao, Sarah Y., Wei S. Lim, and Christopher S Tang. 2017. "Entry of copycats of luxury brands". *Marketing Science* 36 (2): 272–289.

Hu, Ming, and Yun Zhou. 2017. "Price, wage and fixed commission in on-demand matching".

Nikzad, Afshin. 2017. "Thickness and competition in ride-sharing markets". *Available at SSRN 3065672*.

Pun, Hubert, and Gregory D. DeYong. 2017. "Competing with copycats when customers are strategic". *Manufacturing & Service Operations Management* 19 (3): 403–418.

Bai, Jiaru, and Christopher S Tang. 2018. "Can two competing on-demand service platforms be profitable?" *Available at SSRN 3282395*.

Bai, Jiaru, et al. 2018. "Coordinating supply and demand on an on-demand service platform with impatient customers". *Manufacturing & Service Operations Management*.

Banerjee, Siddhartha, Yash Kanoria, and Pengyu Qian. 2018. "Dynamic assignment control of a closed queueing network under complete resource pooling". *arXiv preprint arXiv:1803.04959*.

Cohen, Maxime C, and Renyu Zhang. 2018. "Competition and coopetition for two-sided platforms". *Production and Operations Management*.

Farrell, Diana, Fiona Greig, and Amar Hamoudi. 2018. "The online platform economy in 2018: Drivers, workers, sellers, and lessors". *JPMorgan Chase Institute*.

Guan, Sophia. 2018. *93% of Consumer Engagement with Luxury Brands Happens on Instagram*. https://www.digimind.com/en/news/93-of-consumer-engagement-with-luxury-brands-happens-on-instagram. (Accessed on 06/14/2021).

Lai, Guoming, and Wenqiang Xiao. 2018. "Inventory decisions and signals of demand uncertainty to investors". *Manufacturing & Service Operations Management* 20 (1): 113–129.

Taylor, Terry A. 2018. "On-demand service platforms". *Manufacturing & Service Operations Management*.

Ahmadinejad, AmirMahdi, et al. 2019. "Competition in Ride-Hailing Markets". *Available at SSRN 3461119*.

Asadpour, Arash, Ilan Lobel, and Garrett van Ryzin. 2019. "Minimum Earnings Regulation and the Stability of Marketplaces". *Available at SSRN*.

Bai, Jiaru, et al. 2019. "Coordinating supply and demand on an on-demand service platform with impatient customers". *Manufacturing & Service Operations Management* 21 (3): 556–570.

Bimpikis, Kostas, Ozan Candogan, and Daniela Saban. 2019. "Spatial pricing in ride-sharing networks". *Operations Research* 67 (3): 744–769.

Bryan, Kevin A, and Joshua S Gans. 2019. "A theory of multihoming in rideshare competition". *Journal of Economics & Management Strategy* 28 (1): 89–96.

Chen, M. Keith, Peter E. Rossi, and Emily Chevalier Judith A.and Oehlsen. 2019. "The Value of Flexible Work: Evidence from Uber Drivers". *Journal of Political Economy* 127 (6).

Gurvich, Itai, Martin Lariviere, and Antonio Moreno. 2019. "Operations in the on-demand economy: Staffing services with self-scheduling capacity". In *Sharing economy*, 249–278. Springer.

Hall, Jonathan V, John J Horton, and Daniel T Knoepfle. 2019. "Pricing efficiently in designed markets: The case of ride-sharing". *New York University, New York*.

Li, Jun, and Serguei Netessine. 2019. "Higher Market Thickness Reduces Matching Rate in Online Platforms: Evidence from a Quasiexperiment". *Management Science*.

Loginova, Oksana, X Henry Wang, and Qihong Liu. 2019. *The Impact of Multi-Homing in a Ride-Sharing Market*. Tech. rep. University of Missouri Working Paper.

Ostrovsky, Michael, and Michael Schwarz. 2019. "Carpooling and the economics of self-driving cars". In *Proceedings of the 2019 ACM Conference on Economics and Computation*, 581–582. ACM.

Zhao, Xinyi, Guoming Lai, and Wenqiang Xiao. 2019. "Strategic financing and information revelation amid market competition". *NYU Stern School of Business*.

Arruda, William. 2020. "6 trends that will shape the gig economy in the 2020s". *Forbes*. `https://www.forbes.com/sites/williamarruda/2020/07/12/6-trends-that-will-shape-the-gig-economy-in-the-2020s`.

Chen, Li, et al. 2020. "Bonus competition in the gig economy". *Available at SSRN 3392700*.

Dong, Jing, and Rouba Ibrahim. 2020. "Managing Supply in the On-Demand Economy: Flexible Workers, Full-Time Employees, or Both?" *Operations Research* 68 (4): 1238–1264.

Hall, J., J. Horton, and Daniel T. Knoepfle. 2020. "Ride-Sharing Markets Re-Equilibrate".

Hu, Ming, and Yun Zhou. 2020. "Price, wage, and fixed commission in on-demand matching". *Available at SSRN 2949513*.

Iacurci, Greg. 2020. "The gig economy has balloned by 6 million people since 2010. Financial worries may follow". *CNBC news*. https://www.cnbc.com/2020/02/04/gig-economy-grows-15percent-over-past-decade-adp-report.html.

Li, Miro. 2020. *Xiaohongshu is turning a giant as a social media & an ecommerce platform*. https://daxueconsulting.com/latest-facts-and-insights-about-xiaohongshu/. (Accessed on 06/14/2021).

Lian, Zhen, and Garrett van Ryzin. 2020. "Autonomous vehicle market design". *under major revision at Management Science*.

Nikzad, Afshin. 2020. "Thickness and Competition in On-demand Service Platforms".

Sara Ashley O'Brien, CNN Business. 2020. *DoorDash soars 85% in Wall Street debut*. https://edition.cnn.com/2020/12/09/tech/doordash-ipo/index.html.

Shetty, Sameepa. 2020. *Uber's self-driving cars are a key to its path to profitability*. https://www.cnbc.com/2020/01/28/ubers-self-driving-cars-are-a-key-to-its-path-to-profitability.html.

Tan, Guofu, and Junjie Zhou. 2020. "The effects of competition and entry in multi-sided markets". *The Review of Economic Studies*.

Weisstein, Eric W. 2020. *Delta Function*. https://mathworld.wolfram.com/DeltaFunction.html.

Yan, Chiwei, et al. 2020. "Dynamic pricing and matching in ride-hailing platforms". *Naval Research Logistics (NRL)* 67 (8): 705–724.

Yi, Zelong, Man Yu, and Ki L. Cheung. 2020. "Impacts of Counterfeiting on a Global Supply Chain". *Manufacturing & Service Operations Management, fortcoming*.

Bellan, R. 2021. *Cruise launches driverless robotaxi service in San Francisco*. https://techcrunch.com/2021/11/03/cruise-launches-driverless-robotaxi-service-for-employees-in-san-francisco/.

Bernstein, Fernando, Gregory A DeCroix, and N Bora Keskin. 2021. "Competition between two-sided platforms under demand and supply congestion effects". *Manufacturing & Service Operations Management* 23 (5): 1043–1061.

Besbes, Omar, Francisco Castro, and Ilan Lobel. 2021. "Spatial capacity planning". *Operations Research*.

Blog, Lyft. 2021. *2021 The key to AV deployment: the Rideshare Network*. https://www.lyft.com/blog/posts/the-key-to-av-deployment-the-rideshare-network.

Chakraborty, Soudipta, and Robert Swinney. 2021. "Signaling to the crowd: Private quality information and rewards-based crowdfunding". *Manufacturing & Service Operations Management* 23 (1): 155–169.

Chen, Li, Zhen Lian, and Shiqing Yao. 2021. "Consumer Status Signaling, Wealth Inequality and Non-deceptive Counterfeits". *under review at Management Science*.

Factors, Facts. 2021. *Global Cloud Computing Market Size & Share Will Reach USD 1025.9 Billion by 2026: Facts & Factors*. https://www.globenewswire.com/news-release/2021/01/22/2162789/0/en/Global-Cloud-Computing-Market-Size-Share-Will-Reach-USD-1025-9-Billion-by-2026-Facts-Factors.html.

Gilbert, B. 2021. *Jeff Bezos is about to hand over the keys of Amazon to a new CEO. Read his final letter to shareholders right here.* https://www.businessinsider.com/amazon-jeff-bezos-final-letter-to-shareholders-as-ceo-2021-4?international=true&r=US&IR=T.

IATA. 2021. *Industry Statistics Fact Sheet*. https://www.iata.org/en/iata-repository/publications/economic-reports/airline-industry-economic-performance---april-2021---data-tables/.

Lian, Zhen, Sebastien Martin, and Garrett van Ryzin. 2021. "Labor cost free-riding in the gig economy". *under major revision at Management Science*.

Lian, Zhen, and Garrett Van Ryzin. 2021. "Optimal growth in two-sided markets". *forthcoming at Management Science*.

Lobel, Ilan, Sebastien Martin, and Haotian Song. 2021. "Employees, Contractors, or Hybrid: An Operational Perspective". *Available at SSRN 3878215*.

McCracken, Harry. 2021. *Elon Musk's failed Tesla robotaxi promise is the height of self-driving hype*. https://www.fastcompany.com/90677822/elon-musks-tesla-robotaxi-promise-typifies-self-driving-overexuberance#:%7E:text=At%20the%202019%20Autonomy%20Day,revenue%20from%20its%20App%20Store..

Pun, Hubert, Jayashankar M Swaminathan, and Pengwen Hou. 2021. "Blockchain adoption for combating deceptive counterfeits". Forthcoming, *Production and Operations Management*.

Teale, Chris. 2021. *Uber, Lyft pursue diverging paths in quest for profitability | Smart Cities Dive*. https://www.smartcitiesdive.com/news/uber-lyft-q4-earnings-quest-for-profitability-cities/594945/. (Accessed on 07/05/2021).

TNCs. 2021. *TNCs Today*. https://www.sfcta.org/projects/tncs-today.

wschneider, T. 2021. *Taxi and Ridehailing App Usage in New York City*. https://toddwschneider.com/dashboards/nyc-taxi-ridehailing-uber-lyft-data/.

GOBankingRates. 2022. *How Much Do Lyft Drivers Make?* https://www.gobankingrates.com/money/side-gigs/how-much-do-lyft-drivers-make/#:%7E:text=Hourly%20Earnings,-Though%20driver%20rates&text=On%20a%20typical%20day%2C%20a,taxes%20and%20not%20including%20expenses..

Population, World. 2022. *The 200 Largest Cities in the United States by Population 2022*. https://worldpopulationreview.com/us-cities.

Siddiq, Auyon, and Terry A Taylor. 2022. "Ride-hailing platforms: Competition and autonomous vehicles". *Manufacturing & Service Operations Management*.

Uber. 2022. *Tracking Your Earnings | Driver App*. https://www.uber.com/gh/en/drive/basics/tracking-your-earnings/.