Essays on Cloud Pricing and Causal Inference

Çınar Kılcıoğlu

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

Essays on Cloud Pricing and Causal Inference

Çınar Kılcıoğlu

In this thesis, we study economics and operations of cloud computing, and we propose new matching methods in observational studies that enable us to estimate the effect of green building practices on market rents.

In the first part, we study a stylized revenue maximization problem for a provider of cloud computing services, where the service provider (SP) operates an infinite capacity system in a market with heterogeneous customers with respect to their valuation and congestion sensitivity. The SP offers two service options: one with guaranteed service availability, and one where users bid for resource availability and only the "winning" bids at any point in time get access to the service. We show that even though capacity is unlimited, in several settings, depending on the relation between valuation and congestion sensitivity, the revenue maximizing service provider will choose to make the spot service option stochastically unavailable. This form of intentional service degradation is optimal in settings where user valuation per unit time increases sub-linearly with respect to their congestion sensitivity (i.e., their disutility per unit time when the service is unavailable) – this is a form of "damaged goods." We provide some data evidence based on the analysis of price traces from the biggest cloud service provider, Amazon Web Services.

In the second part, we study the competition on price and quality in cloud computing. The public "infrastructure as a service" cloud market possesses unique features that make it difficult to predict long-run economic behavior. On the one hand, major providers buy their hardware from the same manufacturers, operate in similar locations and offer a similar menu of products. On the other hand, the competitors use different proprietary "fabric" to manage virtualization, resource allocation and data transfer. The menus offered by each provider involve a discrete number of

choices (virtual machine sizes) and allow providers to locate in different parts of the price-quality space. We document this differentiation empirically by running benchmarking tests. This allows us to calibrate a model of firm technology. Firm technology is an input into our theoretical model of price-quality competition. The monopoly case highlights the importance of competition in blocking "bad equilibrium" where performance is intentionally slowed down or options are unduly limited. In duopoly, price competition is fierce, but prices do not converge to the same level because of price-quality differentiation. The model helps explain market trends, such the healthy operating profit margin recently reported by Amazon Web Services. Our empirically calibrated model helps not only explain price cutting behavior but also how providers can manage a profit despite predictions that the market "should be" totally commoditized.

The backbone of cloud computing is datacenters, whose energy consumption is enormous. In the past years, there has been an extensive effort on making the datacenters more energy efficient. Similarly, buildings are in the process going "green" as they have a major impact on the environment through excessive use of resources. In the last part of this thesis, we revisit a previous study about the economics of environmentally sustainable buildings and estimate the effect of green building practices on market rents. For this, we use new matching methods that take advantage of the clustered structure of the buildings data. We propose a general framework for matching in observational studies and specific matching methods within this framework that simultaneously achieve three goals: (i) maximize the information content of a matched sample (and, in some cases, also minimize the variance of a difference-in-means effect estimator); (ii) form the matches using a flexible matching structure (such as a one-to-many/many-to-one structure); and (iii) directly attain covariate balance as specified —before matching— by the investigator. To our knowledge, existing matching methods are only able to achieve, at most, two of these goals simultaneously. Also, unlike most matching methods, the proposed methods do not require estimation of the propensity score or other dimensionality reduction techniques, although with the proposed methods these can be used as additional balancing covariates in the context of (iii). Using these matching methods, we find that green buildings have 3.3% higher rental rates per square foot than otherwise similar buildings without green ratings —a moderately larger effect than the one previously found.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First, I would like to thank my advisors, Professors Costis Maglaras and José R. Zubizarreta, for their support and guidance throughout my studies. They have been great advisors all around. Apart from their input on the academic work, their friend-like approach and guidance to whatever issue I had were unique. I have learned so much from them.

I am very grateful to my committee members Professors Garrett van Ryzin, Omar Besbes, and Jay Sethuraman for their input on my thesis.

I owe a great deal to the other faculty members of the Decision, Risk, and Operations Division, especially Professor Nelson Fraiman. It was a privilege to be his teaching assistant. I have always had his support unconditionally. I would also like to thank the staff and students of the Decision, Risk, and Operations Division for the company.

I feel very fortunate to have worked with Preston R. McAfee and Justin M. Rao at Microsoft Research. I would like to thank Preston R. McAfee for having me in his group and giving me opportunity to work with him. He has the greatest vision I have ever seen. The work in Chapter 3 is completed with my amazing mentor, Justin M. Rao. Thanks to his deep understanding of the cloud market, the work got closer to the real world.

I have a long list of friends to thank for being there when I needed. I hope they all know themselves because there is no friends list here. I would also like to thank everyone who kept asking me the most annoying question "When are you graduating?" for pushing me to work more. Without them, I wouldn't have finished my thesis without enough stress!

Finally, I would like to thank my family, Gülten, İbrahim, Çağlar, and Özge Kılcıoğlu for their continuous support and love. They have always been there for me from thousands miles away. The biggest thanks goes to them.

To my family

Chapter 1

Introduction

Data made available in social networks, media and entertainment, electronic commerce, and mobile is exploding. Firms across industries are increasingly focusing on the use of data analytics to generate insightful and actionable insights to improve their profitability and growth, improve customer experience, design new and better products and services. Together these trends have led to a significant increase in IT storage and computing requirements across industries, and apart from significant infrastructure investments in computing and data storage clusters, they have led to increased support, management and maintenance costs. The operating loads of these large corporate storage and computing clusters exhibit significant intraday and seasonal variability, and additionally firms want flexibility for rapid growth in resource requirements as their needs evolve and mature. In this environment, cloud computing –a form of outsourcing of the aforementioned physical IT infrastructure resources– has become a cost effective alternative for these firms.

In broad terms, cloud computing refers to the provision of computing resources, such as storage, data management, and processing, over a network of remote servers hosted on a remote data center location and accessible via the internet, which is broadly available and at abundant speeds. Cloud computing uses two key pieces of technology. The first is virtualization, the ability to create a simulated environment that can run software just like a physical computer. Virtualization is governed by the "cloud fabric," which functions as the hypervisor, scheduler and manages fault tolerance. The second piece is network communication protocol, both within the datacenter and

between different datacenters. While both technologies have been around for decades, there have been many proprietary advances and thus the quality of the service offered can vary, even when datacenters use the same underlying physical hardware. An analogy is the operating system of a single computer—firms invest in operating system technology to improve performance given the expected capabilities of the underlying hardware.

Currently, Amazon, Google, and Microsoft are the leading providers of cloud computing services to a variety of customers, ranging from individuals and small firms, to global media companies and government agencies. These customers differ with respect to their resource needs, duration, valuation and sensitivity to service level. For instance, while a researcher who does not have a strict time constraint and has a limited budget may prefer to procure computing power anytime within a week and pay little. On the other hand, an online retailer that hosts its web servers in the cloud is very sensitive to service availability and the quality (speed) of the rendered service. This heterogeneity with respect to price and congestion sensitivity allows service providers to offer a menu of product options to segment and better serve this market, essentially offering hosted computing resources at different price levels depending on their anticipated service availability (e.g., as measured by the % of time that the resource will be available).

Chapters 2 and 3 of this thesis focus on economics and operations of cloud computing. Specifically, Chapter 2 studies a stylized revenue maximization problem for a monopolistic provider of cloud computing services, where the service provider (SP) operates an infinite capacity system in a market with heterogeneous customers with respect to their valuation and congestion sensitivity. The SP offers two service options: one with guaranteed service availability, and one where users bid for resource availability and only the "winning" bids at any point in time get access to the service. In this part, we focus on only one "product" with two service levels. The work in this chapter is done in collaboration with Costis Maglaras. In Chapter 3, we study multiple product, single service level (guaranteed service) case, first under monopoly, then under duopoly. In this chapter, we allow the SP to offer a menu of products, which involves a discrete number of choices (virtual machine sizes) with respective price points. The quality of a product is determined by the virtual machine size and firm technology. This work is done jointly with Justin M. Rao.

The backbone of cloud computing is datacenters. Through cloud, each datacenter hosts thousands of tenants at the same time which essentially provides a more efficient energy use compared to the case where each tenant owns an in-house server that requires energy for running the service, cooling the server room etc. However, the energy consumed by datacenters is massive. The National Resources Defense Council (NRDC) states:

> In 2013, U.S. data centers consumed an estimated 91 billion kilowatt-hours of electricity. This is the equivalent annual output of 34 large (500-megawatt) coal-fired power plants, enough electricity to power all the households in New York City twice over. Data center electricity consumption is projected to increase to roughly 140 billion kilowatt-hours annually by 2020, the equivalent annual output of 50 power plants, costing American businesses $13 billion per year in electricity bills and causing the emission of nearly 150 million metric tons of carbon pollution annually. (Whitney and Delforge, 2014)

In the past years, there has been an extensive effort on increasing the energy efficiency of datacenters, both to save on costs and to be environmentally green. Similar to datacenters, buildings are also in the process of going green as they have a major impact on the environment through excessive use of resources, such as energy and water, and large carbon dioxide emissions. Chapter 4 revisits a previous study about the economics of environmentally sustainable buildings. To be able to estimate the effect of green building practices on market rents, new matching methods, which achieve three critical goals simultaneously that current matching methods cannot provide all at once, are proposed; and the economic performance of green buildings is studied using one of this proposed matching methods under statistical causal inference setting. The research in this chapter is done in collaboration with José R. Zubizarreta.

The rest of this chapter introduces the following chapters in depth by positing the research questions and objectives along with their connection to the literature.

1.1 Revenue Maximization for Cloud Computing Services

In Chapter 2, we study a problem of market segmentation for a revenue maximizing (monopolist) service provider (SP) of cloud computing resources that offers two classes of service: guaranteed (on-demand instances) and best effort (spot instances). The latter is procured via a second price auction. This problem is motivated by the service menu offered by Amazon Web Services (AWS), the largest SP in the market currently. Insights extracted from asymptotic analysis of large scale multi-server systems suggest that the observed variation in spot prices is not consistent with the natural stochastic fluctuations between a two-class priority service system. Moreover, it is typically believed that these SP's are not capacity constrained in this stage, but rather experiencing a rapid phase of infrastructure investment aiming to capture market share. Motivated by these observations, we study a SP that operates a system with infinite capacity, and note that under that assumption there is no competition for scarce resources between the two service classes or amongst the users bidding for spot service; specifically all users bidding higher than the SP's reserve price get access to uninterrupted service. The quality of a product is defined as the fraction of time the product is available to customers. While guaranteed service offers 100% availability with a fixed price, the quality level and the payment depend on customers' bids in best effort service. Each customer gains some positive utility from the service proportional to the time that the service is received and suffers a negative utility proportional to the length of time that the service is unavailable. The market is heterogeneous, and, specifically, users differ with respect to two parameters: valuation per hour of service and disutility per hour of service disruption (sensitivity to congestion). The former is how much customers are willing to pay for one hour service, and the latter is how much disutility one hour of service interruption creates. For example, for an online retailer hosting its web servers, valuation is the customer's willingness to pay to have the web server running for one hour, and sensitivity to congestion is the cost of not having the server running, which may include the lost revenue or profit margin as well as lost goodwill from affected customers. Both valuation and sensitivity to congestion are private information and thus unknown to the SP. All users are assumed to have infinite duration service requirements.

We formulate and solve the deterministic SP's revenue maximization problem. We treat two cases separately. First, we study the case where valuation per unit time grows sub-linearly as a function of the disutility per unit time of service disruption, i.e., where `(valuation/time)/(cong. sens./time)` $\downarrow$ as `(cong. sens./time)` $\uparrow$. In other words, for users that are congestion sensitive, disutility due to service disruption grows faster than the user's valuation. This case is the main focus of this chapter and we extensively analyze the solution to the revenue maximization problem under this regime. In practice, user's valuation from a rented resource is bounded above by the valuation attained from a resource owned; while disutility due to service disruption is not bounded and can be multiple times of the valuation depending on use cases. This observation motivates the reason of focusing on this sub-linear increase case. In the first part of our analysis, we assume that the user's valuation per unit time of service is an affine, increasing, function of her congestion cost per unit time of service disruption. In this case, user types are one-dimensional, and we assume that user congestion costs per unit time are independent identically distributed (i.i.d.) draws from a continuous distribution, with a strictly positive density function and bounded support. We model the prevailing spot price as a discrete process (e.g., in $.01 increments per hour) and focus on the associated steady state distribution. We assume that the SP can select the steady state distribution, i.e., the long run average fraction of time for which the spot price spends at each price level; if the SP's reserve price is constant through time, then the steady state distribution will reduce to a point mass at that respective level. (We discuss the validity of the assumption of using steady state distribution in the last section of the chapter.) Users (have infinite service time requirements) observe the steady state distribution of the spot price path, and choose between guaranteed and spot, and, if they select the spot service option, they also determine how much to bid. We prove that i) the SP should set the price levels of the spot service option such that the lowest spot price level will be below the lowest valuation across all users in the market (that is, nobody is priced out); ii) it is optimal to use two distinct price levels in spot service for positive fractions of time, respectively, and offering more than two price levels does not generate more revenue for the SP; and iii) the fraction of time that the spot service price is "high" depends on the coefficients of the affine relation between congestion cost rate and valuation rate, but not the distribution of types itself.

These "high" periods make the spot service option stochastically unavailable and create intentional service degradation —a form of "damaged goods." In the last part of our analysis, we show that it may be optimal for the SP to offer multiple (> 2) price levels for the spot service option, if the valuation rate grows sub-linearly with respect to the congestion cost rate but the respective relation is general (not affine).

For completeness, the second case we study is one where the valuation rate increases faster than the user's congestion cost rate. In this case, we prove that it is never optimal to offer spot service. Intuitively, in this setting congestion sensitive users are willing to pay increasingly high amounts, and the SP is not willing to sacrifice any revenue from these high types by offering an incentive compatible lower priced spot price option. Therefore, if more congestion sensitive customers have comparatively higher valuations, then it is optimal to serve only the high-valuation market segment by offering the high quality service. This case is analyzed in the Appendix.

Next, we analyze the price traces of over 1,000 products that the dominant provider in the market offers for a six-month period, and present descriptive statistics that sheds light on the dynamics of the spot price. This work, by no means, claims that the dominant provider in the market sets the prices as described here. It provides an alternative explanation to the observed spot prices which is found to be not consistent with the asymptotic analysis of large scale multi-server systems. Calibrating our model on the observed data, we offer some insight on the dynamics of spot price valuations, and characterize the relative magnitude of valuation rate to the congestion rate; the latter may be as much as 10 times larger than the former.

Lastly, we verify our state independent, stationary model using data and show that utilities under the state dependent transient system converges to the utilities under the steady state when users have service times in the order of days.

Our work is related to the literature on "economics of queues," which goes back to the work of Naor (1969) that introduced the study of strategic customer behavior in a queueing setting. Mendelson (1985) and Mendelson and Whang (1990) studied (primarily) social welfare optimization in an $M/M/1$ system serving a market of heterogeneous, utility maximizing customers. Afèche (2013) studied the revenue maximization problem for a SP operating in a market with two segments

that differ with respect to their delay sensitivity, and importantly showed that the SP may use the notion of "strategic delay" to optimally segment the market and optimize the system's revenue rate. Strategic delay amounts to (artificially) increasing the realized waiting time of some customers beyond the waiting time that they would experience due to the system's congestion effects. This is akin to the idea of "damaged goods" introduced earlier on in economics and marketing, e.g., Deneckere and McAfee (1996) and McAfee (2007) that showed that profit maximizing firms may intentionally "crimp" their products to price discriminate, and Pareto improve performance; these papers provide striking examples from high-tech industry; see also, Anderson and Dana (2009).

Our model does not involve any congestion phenomena that arise due to the dynamics of a finite capacity physical system, and as such resembles in its nature the marketing and economics references on damaged goods. In terms of model formulation and motivation, however, it is closer to several papers from the economics of queues literature that we highlight below. Afèche and Pavlin (2015) studied a model with one-dimensional types, where the valuation is a linear function of the delay cost parameter. For this model they characterized for a SP that operates an $M/M/1$ system. We will consider the same model in §2.2.3 and study the SP's revenue maximizing solution in that case. Our model differs from the one above in its utility function: specifically, users extract value from the service and pay only when service is available, and incur disutility but stop paying when service is interrupted. Our result that shows that the use of "damaged goods" may be optimal is similar to theirs. The affine model is an example of a model where valuation grows sub-linearly as a function of the congestion sensitivity. §2.3.3 shows that when the monotonicity result holds but the relation between the two parameters is general, then the optimal solution may involve again the use of damaged goods but the structure of the optimal policy is more complex. In the Appendix, we look at the case where the valuation rate grows super-linearly as a function of the congestion cost parameter, which is akin to the model studied in Katta and Sethuraman (2005). Our utility function is again different and the details of the analysis are not the same, but one of the key findings that the use of damaged goods is not needed is consistent with their results (considered when capacity grows large and the system becomes uncongested). Nazerzadeh and Randhawa (2015) look at a similar model as the one studied in Katta and Sethuraman (2005)

and among other things show that in the unconstrained capacity setting, offering one service level performs "well," which is consistent with our findings.

Our work is also related to the stream of work that studies economic optimization problems in a queue in the context of large scale systems. Maglaras and Zeevi (2003) showed that in a single type market where demand is elastic, the revenue maximizing operating regime in an $M/M/C$ system where the system size C and the market potential grow large is the so-called heavy-traffic regime. Maglaras *et al.* (2015) extended this analysis to multiple types of customers, establishing again, under some conditions, the phenomenon of strategic delay mentioned above. Finally, our model operates as a two class priority system. The asymptotic behavior of such a system in a multi-server setting was studied in Maglaras and Zeevi (2005).

Abhishek *et al.* (2012) ask a question similar to ours and analyze the problem of the SP using tools from mechanism design to show that offering only high a quality (guaranteed service) product with a fixed price generates more revenue than offering both high and low quality products at the same time. This result is in contrast to our findings in §2.3.3, as well as those in Afèche and Pavlin (2015). In our model of §2.3.3, users with valuation v_i have congestion cost parameter κ_i (deterministic), whereas in Abhishek *et al.* (2012) such users may have a random congestion cost parameter with distribution F_i. If we approximate our model in their setting by letting the capacity grow large, and, more importantly, restrict the support of their congestion rate parameter to a narrow support (centered around κ_i), then one of the key conditions needed for their main finding no longer holds, therefore removing the apparent inconsistency. Afèche and Mendelson (2004) studied the revenue maximization problem in a queue with priority auctions and generalized delay cost structure. They show that in some cases, revenue maximizing uniform pricing provides no or only little surplus loss. Moreover, using priority auctions instead of uniform pricing yields lower prices and higher utilization in the system.

In a recent study, Mitra and Wang (2015) consider a monopoly broadband access internet service provider that offers a guaranteed service with a usage fee, and a best effort service free of charge. In profit maximization setting, they show why best effort service "harvests" possibly new guaranteed service clients; at its core lies a stylized model for the dynamics of adoption of new

users (applications) that start as best effort users (subsidized) and then some of these transition to successful applications that switch to guaranteed service quality.

Armbrust *et al.* (2010) provide an overview of cloud computing from different perspectives including cloud computing economics. Xu and Li (2013) show that throttling the resource generates more revenue than uniform usage pricing and performance guarantees can be provided with an extra fee. In their model customers differ only with respect to their valuation per unit time and each customer is allowed to choose different number of resources. Borgs *et al.* (2014) study a multiperiod pricing problem where the SP offers a service with varying capacity in a market that customers are strategic and heterogeneous in their valuations, arrival and departure periods. They used the cloud computing market as an example of such a setting, and provided an efficient algorithm to find a dynamic pricing mechanism that satisfies service guarantees. Savin *et al.* (2005) look at the problem of capacity allocation of rental equipment used by two customer types, with stochastic rental period requirements. They formulate the problem as a queueing control problem and provide a heuristic control based on a fluid model approximation. Baron (2003) considers a system (similar to cloud computing) that the SP shares her computing resources. He presents token-bucket admission control and pricing schemes. In this work customers compete for the shared resource.

Our work provides descriptive statistics and some analysis on a rich data set from a leading SP. Similar datasets have been analyzed in different works to find possible explanations for the observed price fluctuations. Agmon Ben-Yehuda *et al.* (2011) draw the conclusion that the SP varies her reserve price over time. They empirically show that the spot prices seem to follow trends that show significant regularity when views under an appropriate prism, and could be the result of the SP's control of the reserve price.

1.2 Competition on Price and Quality in Cloud Computing

While the importance of the technologies that cloud computing uses is widely researched in the systems community (see *e.g.*, Nurmi *et al.*, 2009; Rimal *et al.*, 2009), the "infrastructure as a service" public cloud marketplace is often described as "commoditized" from an economic competition per-

spective (Marston *et al.*, 2011). The reasoning is putatively straightforward. Since cloud providers use similar, if not identical, physical hardware they cannot meaningfully differentiate their products and thus profit margins should converge to zero. In Chapter 3, we begin our analysis by empirically assessing this claim by running a series of benchmarking workloads across two major provider's various service levels ("virtual machine" (VM) size), similar to the approach used in Li *et al.* (2010). We find different run-times for similarly described offerings, such as "2 virtual cores, 4GB memory." While run-time decreases for both providers as one moves to larger VMs, the price-performance trade-offs are different, which means there are different feasible price-quality combinations. We formalize this insight with a two-parameter model of the firm's production technology and the calibrated model achieves good fit to our data.

The fitted parameters are used in our theoretical model as one source of differentiation across firms. We view these technologies as fixed for our analysis, imagining they are the result of countless engineering decisions made over the years. Endowed with a technology, firms then choose *performance menus*, which provides a second source of potential differentiation. A performance menu is a set of VMs with different CPU, memory and disk configurations. For example, Amazon Web Services (AWS) offers about 20 different VM configurations, ranging from low performance "micro" to high performance "extra large." We model customers as having heterogeneous types with varying sensitivity to job completion time, but with a common job completion valuation and workload requirement. Customers choose optimally from the price-quality menus provided by firms.

We start with the monopoly case. There are a number of reasons this is a useful starting point even though most large regional markets are not currently characterized by monopoly. First, SEC Filings reveal that AWS is currently many times larger than the next closest competitor, indicating that one provider "pulling away" from the competition is certainly not implausible. Second, smaller countries often have only a single major provider with a datacenter within national boundaries. Finally, a customer that has used a given provider for some time could face large switching costs, leading to potential monopolistic dynamics targeted at "locked in" customers.

For the monopoly case, we characterize the optimal base price, quality level and associated customer demand functions. Interestingly, under some conditions, offering an additional quality

level does not generate more revenue. We provide sufficient conditions for when a firm should offer multiple quality levels. The conditions show that when the quality level is increasing almost linearly in price and there are some customer types in the system that are highly sensitive to delay, offering an additional higher quality products, up to a point, generates more revenue.

The results also reveal an interesting dynamic with respect to customer valuations and quality. When valuations increase, the optimal strategy for the service provider is to intentionally degrade the quality level of lower tier offerings as opposed to increasing the unit price. While this might sound counter-intuitive at first, it is readily understood by recognizing that customers are paying per time-unit. A higher quality product is not only more expensive, but offers faster runtime—the faster runtime reduces the net payment on the margin. As valuations increase, there is an incentive to make the low quality options less attractive to "high types." By damaging the product, it is effectively more expensive *and* less attractive due to increased delay. This "double dividend" for damaging the good has previously been observed in the computing hardware and shipping/transport industries (Deneckere and McAfee, 1996). Overall, the results for the monopoly case highlight the nuanced role of competition in this marketplace.

We next move on to the duopoly case. We start by characterizing the Nash equilibrium when each service provider is restricted to offer only one quality level. In this case the higher quality provider attracts high-type customers (the ones that are more sensitive runtime delays) at a higher price. In other words, there is stable differentiation on the quality dimension. When providers are allowed to offer multiple quality levels, we no longer have a closed form solution. We thus simulate the game under different market settings where providers compete in base price level. Interestingly, prices do not converge and instead display Edgeworth cycles (as in Maskin and Tirole, 1988). The intuition for these cycles is the standard one, with a a bit of tweak. Despite the quality differentiation, the goods are relatively good substitutes for each other and thus Bertrand-like price competition leads to successive undercutting of price, albeit at different price levels (the tweak). That is, prices move in parallel down to a point of very low returns for the firm. At this point, a war of attrition ensues and one firm "leads" the pair back up to a higher price point and the cycle repeats.

Past research has shown that these types of cycles, though commonly predicted, are empirically quite rare. Exceptions occur in markets where prices change flexibly and there are other sources of price volatility (e.g. due to cost shifters such as oil prices in the retail gasoline market Noel, 2007). Perhaps unsurprisingly, then, we do not observe classical Edgeworth cycles in cloud computing. It turns out, however, that once we consider important market features the observed price patterns share qualitatively similar features with classic cycles.

The most important dynamic is the relatively rapid reductions in the cost per compute cycle due to technological advances, which are commonly attributed to Moore's law. In reality the situation is more complex, with Moore's law slowly giving way and other advances breaking through (Chien and Karamcheti, 2013). Nonetheless, these advances provide both a real decline in costs for the provider and a strong consumer perception that prices should fall, not rise. In practice, cloud providers tend to replace physical hardware approximately every three years. The release of new hardware enables new, superior "generations" of VMs. But the old generations can nonetheless be virtualized on the new hardware, just with less physical resources required than before and thus at a lower cost. This means constant prices for older generations are effective increases relative to costs. We examine historical prices and observe that the largest provider, AWS, tends to offer newer generations at lower prices and keep older generation prices relatively high. Indeed we document that older, inferior generations are often priced *higher* than the comparable VMs in the new generations. So while the model predicts varying intensity of price competition over time, in practice we observe this variance across products by release date. In other words, some "regions" of the product space have vigorous competition—we view this as substantively similar to the cycling prediction.

Further, we highlight that our model predicts that price differences can be maintained in equilibrium and the market will not totally commoditized. Interestingly, in the Summer of 2015 one major provider dropped prices rather substantially and the other two major providers did not follow suit. Our model gives a rigorous explanation as to why.

To the best of our knowledge, this is the first study that models the cloud computing products from price-quality perspective under competition. The analysis draws on three main streams of

literature. The first is from economics and marketing literature on price-quality competition. Most papers here focus on the case when players are symmetric and each player chooses one quality and one price under competition or one player chooses two distinct quality levels under monopoly (Maskin and Tirole, 1988; Moorthy, 1988; Shaked and Sutton, 1982). Here we have two asymmetric players each choosing multiple quality levels and prices, and both quality levels and prices are interdependent, which is why we have to rely on simulations at times.

The second stream of literature is on cloud pricing. Abhishek *et al.* (2012) and Xu and Li (2013) look at the problem from a higher level and try to find the best pricing strategy by offering the same product in different pricing mechanisms. In this work, we aim to find a revenue maximizing price-quality menu with fixed prices. There are papers on competition in an oligopoly market with multiple providers. Feng *et al.* (2014) studies non-cooperative competition model in a cloud market and computes an equilibrium price. However, each player has single product type in this study. Anselmi *et al.* (2014) studies the price competition in cloud computing by considering all three layers of cloud. Our focus in this study is only the IaaS market.

The third stream is the analysis reports prepared by private cloud companies (CloudHarmony[1], Cloud Spectator[2], ProfitBricks[3]). They investigate the performance of different cloud providers from different angles. Although their methodology contains extensive performance analysis, it does not have a solid economic framework, and performance values and units prices are not incorporated into the analysis in a transparent manner.

[1] https://cloudharmony.com

[2] http://cloudspectator.com

[3] https://www.profitbricks.com

1.3 Maximizing the Information Content of a Balanced Matched Sample in a Study of the Economic Performance of Green Buildings

Buildings have a major impact on the environment through greenhouse gas emissions and excessive use of natural resources. For example, the United States Environmental Protection Agency (EPA) reported that in 2013 nearly 39% of total U.S. carbon dioxide emissions were due to residential and commercial buildings.[4] For the same year, the U.S. Energy Information Administration reported that about 40% of total U.S. energy consumption was from these types of buildings.[5] At the same time, there is growing scientific consensus that current levels of carbon dioxide and related greenhouse gas emissions greatly increase the risks of climate change, and that excessive use of resources can lead to resource depletion and habitat degradation. For these reasons, the construction and operation of buildings can have a substantial impact on the earth's environment.

In an interesting and relevant study, Eichholtz *et al.* (2010) analyzed the effect of environmentally sustainable building practices on their rents and selling prices. This is an important study subject for the reasons already stated and also because there is not much empirical evidence for the development of environmentally sustainable or green buildings. Among the available evidence, there are the results of a study by the U.S. General Service Administration Public Buildings Service that analyzed the performance of 22 green buildings and found that, compared to national averages, green buildings have 36% fewer carbon dioxide emissions and 25% less energy use, in addition to 19% lower aggregate operational costs and 27% higher occupant satisfaction.[6] Given the environmental and social benefits of green buildings, one important question is how much these benefits affect the rent of green commercial buildings. This is important to investors, developers and property owners in order to invest in green buildings.

In their study, Eichholtz *et al.* (2010) analyzed a large sample of commercial green- and non-

[4] http://www.epa.gov/climatechange/Downloads/ghgemissions/US-GHG-Inventory-2015-Main-Text.pdf, Table ES-7.

[5] http://www.eia.gov/totalenergy/data/monthly/pdf/mer.pdf, Table 2.1.

[6] http://www.gsa.gov/graphics/pbs/Green_Building_Performance.pdf.

green-rated buildings in the United States. Using linear regression and propensity score methods, they found that buildings with green ratings have 2.8% higher rental rates per square foot compared to similar buildings without green ratings. In Chapter 4, we revisit this important question using new matching methods that adjust more precisely for covariates and better exploit the structure of the buildings data.

In the United States, green buildings are certified as energy-efficient or sustainable by different agencies. The EPA gives the "Energy Star" certification to commercial buildings if their amount of energy used meets certain criteria.[7] The Green Building Council (USGBC) labels a building as LEED (Leadership in Energy and Environmental Design) based on its performance in different categories such as indoor environmental quality, site sustainability and water conservation. Following Eichholtz *et al.* (2010), we consider a building to be green if it is certified as Energy Star or LEED and focus our analysis on commercial buildings.

To estimate the effect of energy efficiency and sustainability on the economic returns of buildings, we compare green-rated buildings to similar non-green-rated buildings in the same market. For this, we use multivariate matching methods and find matches of green and non-green buildings that are nearby and similar along a number of covariates, including age, amenities, number of stories, quality and whether the building was recently renovated. However, standard matching methods do not have the flexibility to exploit the particular structure of the buildings data and will typically result in imbalanced or inefficient analyses. In particular, the data consists of 694 green buildings and 7,411 non-green buildings, organized in 694 geographic clusters. In each of these clusters, there is one green building and one or more non-green buildings not further apart than one quarter mile from the green building. While some clusters have only one non-green building, others have as many as 83 non-green buildings. As a result of this structure, pair matching (or matching with a $1:1$ ratio) would result in many non-green buildings not being used in the analysis, and matching with a fixed $1:\kappa$ ratio (where κ is an integer greater than 1) would result in some clusters not being used at all. Naturally, for our analyses we would like to use a flexible matching ratio in

[7] Specifically, the EPA can give the "Energy Star" certification to buildings in the top quarter of energy efficiency compared to similar buildings nationwide. The energy efficiency calculation is done by the EPA using a scoring algorithm that takes into account the characteristics of the building, such as size, location, number of occupants.

order to match as many buildings as possible, while precisely balancing covariates. However, to our knowledge existing matching methods are not able to achieve all of these goals simultaneously. To analyze the effect of energy efficiency and sustainability on the economic returns of buildings, in this work we build on the method of cardinality matching and propose a general matching framework to maximize the information content of a balance matched sample. Within this framework, we present new matching methods that simultaneously achieve three goals: (i) to maximize the information content of a matched sample and, in some cases, minimize the variance of a widely used effect estimator; (ii) to form the matched groups of the matched sample using a flexible matching structure (such as a one-to-many/many-to-one or, in a sense, a full matching structure; Rosenbaum, 1989, Hansen, 2004); and (iii) to directly attain covariate balance as specified —before matching— by the investigator. On the one hand, standard matching methods are not designed to achieve goals (i) and (iii), but on the other hand, cardinality matching does not allow flexible matching structures beyond a one-to-many fixed matching ratio. Achieving these three goals simultaneously poses a number of difficulties. First, maximizing the size of matched sample with a flexible matching ratio requires a different notion of sample size than the one used in cardinality matching, since, for instance, two one-to-one treated and control matches should not count the same as one one-to-two match. This requires defining the information content of the matched sample. Second, the differential weighting of the different matched groups needs to be taken into account when assessing covariate balance and in the analyses, but this poses a number of challenges in building a mathematical program and in computing its optimal solutions. Third, a sound implementation of this method needs to take advantage of modern advancements in parallel computing.

Chapter 2

Revenue Maximization for Cloud Computing Services

In this chapter, we study a stylized revenue maximization problem for a cloud computing provider. Section 2.1 offers a short introduction to the services and pricing that we encounter in today's cloud computing SPs. This section provides a basis for Chapter 3 as well. Section 2.2 describes our model, which we analyze in Section 2.3. Section 2.4 offers a more detailed look into the pricing data from the currently largest provider of cloud computing, Amazon Web Services, and briefly discuss some of its implications. Finally, Section 2.5 builds a state-dependent model and shows how it converges to our main model after users stay in the system for few days.

2.1 Glimpse of Cloud Computing Market and Pricing Mechanisms

In this section, we describe the market and practice, then motivate our analysis. In Section 2.4 we will return and take a closer look at the data.

The two key participants in the market for cloud computing are users and providers. The users can be individuals or companies requiring temporary (short-term) or permanent (long-term) computing resources that can be reached over the internet. The providers are the operators of the cloud computing services. Currently there are many small and large SPs in the market, with

Amazon, Google, and Microsoft being the leading providers. There are three main services that cloud providers offer: software-as-a-service (SaaS), platform-as-a-service (PaaS), and infrastructure-as-a-service (IaaS). In this work we are focusing on IaaS service, where the product is defined as the bundle of a machine type, an operating system, and a location.

Each provider offers its products under one or multiple price models. The dominant provider in the market is Amazon and it offers the richest pricing options. Currently, Amazon rents out its computing resources under three different pricing models: pay-as-you-go (on-demand instances), pay-as-you go under contract (reserved instances), and second price auction (spot instances). On-demand and reserved instances offer guaranteed service and in the sequel, we will focus on a model with only 2 service options: guaranteed and best effort, which we refer to "on-demand" and "spot." Although there are multiple providers in the market and they compete, as of May 2015, Amazon's IaaS cloud is ten times larger than the next 14 competitors combined (Leong *et al.*, 2015).

We are focusing on two pricing models: on-demand and spot. Each product has a fixed hourly price in the on-demand market and users continue paying this fixed rate as long as they use the service. Amazon has no control of ending a running service, while customers can end their service at any point in time with no penalty. The spot market has a more complicated pricing structure. For each product, Amazon sets a reserve price, possibly time-varying, and customers bid their maximum willingness-to-pay per hour for that product. The spot price at any time point is defined as the minimum bid accepted at that time, which in some cases may be the reserve. The spot price fluctuates over time in response to variations to the available capacity not utilized by the "guaranteed" instances rented by Amazon, and to the number of active spot customers and their corresponding bids. If the bid of a particular customer falls below the spot price, this customer is temporarily out of access to the cloud (priced out) until the spot price falls again at or below her bid.

The data on hand shows the the spot price exhibits significant fluctuations over time. They may be around one tenth of the corresponding on-demand prices; and, can and do fluctuate to up to five or ten times of the corresponding on-demand prices; interrupting service for many spot instance customers, resulting in some form of disutility. If customers in the spot market bid sufficiently high

and continuously pay the prevailing spot price (even when it is above the price of on-demand), in the long run they will receive uninterrupted service. The corresponding time-average spot price is cheaper than the corresponding on-demand price for some of the products, but certainly not all. We present further descriptive statistics in Section 2.4.

Another choice customers make is whether to use cloud or in-house resources for their computing needs, and there are multiple size of products that can be chosen under the cloud option. Table 2.1 shows the configuration and prices for a product family (*m4* machine types) with Linux operating system residing in *us-west-2* (Oregon) region. The table offers a glimpse on the magnitude of these per hour costs that later on are traded off against disutility from service disruption per hour. Among these products, we analyze *m4.xlarge* machine more closely. Hourly on-demand price for this product is \$0.254/hr, while it can go up to \$0.374 in other regions. This product was available both in spot and on-demand markets approximately in the last 80 days of our time window. Usage in this period cost \$486 in on-demand market, while it was between \$113 to \$207 (depending on the subregion selected) in spot market. One year of continuous usage of this product costs \$2,208 in on-demand market. As a comparison, a similar in-house server (HP ProLiant DL380 Gen9 - Xeon E5-2620V3 2.4 GHz - 16 GB, which has 6 cores) costs around the same to purchase without any IT, rack space, or peripheral costs for mounting, networking, etc.. However, if one wants to rent a product for long-term continuous usage, reserved instances offers much cheaper options. For instance, the same product can be rented by paying \$1,271 upfront for one year of usage (see Armbrust *et al.*, 2010 for a more detailed cost analysis). We study the optimal price-product size menu selection in Chapter 3.

Table 2.1: Prices in on-demand and reserved markets for a group of products and their configurations

Machine name	# cores	# RAM	price/hr	on-demand price/year	reserved price/year
m4.large	2	8	\$0.126	\$1,103.76	\$635
m4.xlarge	4	16	\$0.252	\$2,207.52	\$1,271
m4.2xlarge	8	32	\$0.504	\$4,415.04	\$2,541
m4.4xlarge	16	64	\$1.008	\$8,830.08	\$5,082

2.2 Model Formulation

2.2.1 Detour: Asymptotic Behavior of Large Scale Multi-Server Systems

In the sequel we will motivate why we will adopt a system model with infinite capacity. To do that, we will review known results from multi-server priority systems with large capacity (like the ones operated by cloud SPs).

We briefly discuss a system where the SP has a finite processing capacity C and offers two nonsubstitutable service classes: guaranteed-rate (G) service and best-effort (BE) service. In the former, customers receive a constant service rate as long as there is capacity and are blocked otherwise; in the latter, service rate is dependent on the number of customers in the whole system. G service has priority over BE. BE users get one unit of capacity, if this is available, or share the available capacity (not used by G users) equally if there are more BE users connected than the available number of servers, thus experiencing congestion. Customers arrive to the system according to independent Poisson processes and the service requirements are exponentially distributed. Maglaras and Zeevi (2005) studied this system and showed that when the system size grows large, the G class occupies $\alpha C + B(t)\sqrt{C}$ servers, where $0 < \alpha < 1$ and $B(t)$ is standard Brownian motion, and the remaining capacity, $(1-\alpha)C - B(t)\sqrt{C}$, is available to BE service. A similar analysis could be carried through under the auction model for BE service. The important observation is that the variation in the available capacity for BE users will be second order, and this would result in fluctuations of the prevailing spot price that would also be second order (i.e., small). This prediction does not agree with what we observe in the data. This suggests that perhaps a different mechanism gives rise to the fluctuations to the spot price that may be exogenous to the capacity dynamics of the BE class, as defined crudely by their supply-demand imbalance.

The above discussion has three important caveats that are worth noting. First, the model assumes the same (or reasonably similar) service durations for both services. Second, the model assumes each customer has unit demand. If users may demand a random number of servers and this follows a heavy-tailed distribution, it may be possible to observe big price spikes. The observed frequency and duration of price spikes would require frequent, random arrival of users with unusually

large capacity needs that are short-lived (which may be implausible). Last, the model assumes that in equilibrium, the fraction of the overall system capacity consumed by each of the two service classes are comparable (and first order). If BE service used a very small fraction of the total capacity and the overall system was heavily utilized, then significant spot price fluctuations could emerge; e.g., the BE usage is of order $\sqrt{C}$, which is the same as the order of magnitude of the G service class, thus resulting in fluctuations of BE available capacity that are of the same order as the overall capacity used by BE. Nevertheless, in this case the revenue generated from BE service would be insignificant, rendering the parameter regime less interesting. Cloudyn, a cloud management platform, estimated in October 2013 that the spot instances only consume 3%–5% of all instances with 40% monthly increase. Since then, Amazon invested heavily on its spot instances, developed new features and alternative product groups, and acquired a few companies working on spot market optimization. Therefore, we believe that the size of the spot market has reached to a significant level. It also seems unlikely that the data centers of large-scale cloud computing SPs are operating at full utilization at this point in time of rapid expansion and effort to capture market share.

2.2.2 The Infinite Capacity Model

Motivated from the above we will model the market as follows. The SP has infinite capacity and operates multiple resources and offers a service (or a product) from two distinct channels: guaranteed (G, on-demand instances) and best effort (BE, spot instances). Customers differ with respect to their willingness-to-pay for a unit-time service, v, and their congestion sensitivity parameter, κ. We assume that each customer has unit demand and infinite service time. Customers are individual utility maximizers and they make two decisions: i) which service to choose, and ii) if BE is chosen, how much to bid. A customer's decision is independent of the size of her demand. Although the decision of which service to choose might be dependent on the service time and the current state of the system, we show later in Section 2.5 that these effects disappear if the service time is at least a couple of days. Moreover, since customer decisions become state independent, coupled with infinite capacity, the arrival distribution (stationary or non-stationary) plays no role in the system dynamics.

Let i denote the service class such that $i = 1$ for G service and $i = 2$ for BE service. G service is offered with a fixed price p_G per unit time and each customer paying this price gets a dedicated resource. The price for BE service, $p_{BE}(t)$, is a RCLL (right-continuous with left limits) discrete-level stochastic process in the interval $[\underline{p}, \overline{p}]$ with N price levels ($\underline{p} = p_N \leq p_{N-1} \leq \ldots \leq p_1 = \overline{p}$). We will not characterize the dynamics at this stage, but assume that users with infinite service level requirements decide based on the steady state probability mass function associated with $\{p_{BE}(t), t \geq 0\}$ which is denoted by $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$, and assumed to exist, has support $\mathcal{P} = [\underline{p}, \overline{p}] \in (0, \infty)$. In this option, customers place their bids and the SP offers service to each BE user whose bid is larger than or equal to the prevailing spot price $p_{BE}(t)$, and interrupts service to all bidders below $p_{BE}(t)$. That is, a BE user that bids b is *active* $\forall t$ s.t. $p_{BE}(t) \leq b$ and *interrupted* $\forall t$ s.t. $p_{BE}(t) > b$. We assume interruptions have no cost to the SP and an interrupted job resumes without any additional setup cost (if service disruptions are infrequent and service times are long—infinite in our model—this modeling idealization may be reasonable). It is worth noting that in our infinite capacity model the price dynamics are controlled by the SP as opposed to stochastic supply-demand imbalance effects.

Users are heterogeneous and characterized by their idiosyncratic (monetary) valuation per unit time of receiving service v and disutility (congestion sensitivity) parameter κ, which measures the monetary loss per unit of time where the service is unavailable. Consider a user with valuation v and congestion sensitivity parameter κ and bid $\$b$ for BE service. For a user that selects BE service and bids $\$b$, we will define $\alpha(b)$ to be the fraction of time her service is active, and $p(b)$ to be the payment per unit time:

$$\alpha(b) = \sum_{i:p_i \leq b} \pi_i \qquad \text{and} \qquad p(b) = \sum_{i:p_i \leq b} \pi_i p_i.$$

The net utilities for the two service options are:

$$U_1(v, \kappa) = v - p_G \quad \text{and} \quad U_2(v, \kappa, b) = \alpha(b)v - \kappa(1 - \alpha(b)) - p(b).$$

That is, customers extract the per unit value $\$v$ when their service is active; when their service is

interrupted they forgo this value and incur a cost of $\$\kappa$ per unit time. They only pay while their service is active, captured by p_G and $p(b)$ for each option, respectively.

The optimal BE bid for a user with parameters v and κ is

$$b(v,\kappa) = \underset{\underline{p}\leq b\leq \overline{p}}{\operatorname{argmax}} U_2(v,\kappa,b).$$

Lemma 1. *Without loss of generality,* $b(v,\kappa) \in \{p_1, p_2, \ldots, p_N\}$.

Moreover, if there are multiple bids that achieve the maximum, we assume that the lowest maximizing bid is selected. We let $U_2(v,\kappa) := U_2(v,\kappa,b(v,\kappa))$. A user with parameters v and κ chooses service-$i^*(v,\kappa)$, where

$$i^*(v,\kappa) = \underset{i=1,2}{\operatorname{argmax}} \{U_i(v,\kappa) : U_i(v,\kappa) \geq 0\} \text{ and set } i^*(v,\kappa) = 0 \text{ if } U_i(v,\kappa) < 0 \text{ for } i = 1,2,$$

where $i = 0$ represents the no-buy option.

In the next part, we will formulate our revenue maximization problem for the case where there is an affine relation between v and κ.

2.2.3 Revenue Maximization Problem

We study a market where the valuation rate v grows more slowly than her corresponding congestion sensitivity parameter κ; i.e., users with increasing congestion sensitivity may indeed value the service more, but their valuation does not grow as fast as the corresponding disutility from service interruption.

This market regime has two assumptions: i) valuations increase with congestion sensitivity; ii) valuation over congestion sensitivity ratio is increasing in congestion sensitivity. Both assumptions are aligned with the real world. First, as users value the service more, their disutility from service disruption hurts them more. Second, user valuations are capped by the valuation gained from a server that is owned rather than rented. On the other hand, congestion sensitivity does not have a clear upper bound and this value may be very high for businesses doing critical work that requires

100% uptime.

We consider a continuum of user types indexed by η. A type η user has a positive willingness-to-pay $v := A + \eta$ per unit time of service and a positive congestion sensitivity parameter $\kappa := B\eta$, where A, B are positive constants common across all consumers. User types are assumed to be independent and identically distributed (i.i.d.) draws from a continuous distribution F with density f, which is assumed strictly positive and continuously differentiable on the interval $\mathcal{N} = [\underline{\eta}, \overline{\eta}] \subseteq [0, \infty)$. Let $\bar{F} = 1 - F$. Hence, v and κ are linearly dependent and user heterogeneity is one dimensional. Note that in this setting, both the valuation rate ($v = A + \eta$) and the congestion rate ($\kappa = B\eta$) are increasing function of the user type η, and that relative rate of growth of $v/\kappa = \frac{A+\eta}{B\eta}$ is decreasing in their type. We summarize this model for ease of reference below:

$$\text{Model 1: } \; v = A + \eta, \; \kappa = B\eta, \; A, B > 0, \; \eta \sim F. \tag{2.1}$$

The SP offers the BE service on an N-price grid given by $p_1 \geq p_2 \geq \ldots \geq p_N \geq 0$ and let $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$, where π_i is the fraction of time the prevailing BE service price is p_i ($\mathbf{1}^T\boldsymbol{\pi} = 1$ and $\boldsymbol{\pi} \geq 0$). Guaranteed service price is p_G. Define $\bar{\pi}_k = \sum_{j=k}^{N} \pi_j$ and $\bar{p}_k = \sum_{j=k}^{N} \pi_j p_j$, and redefine the net utility function for BE as

$$U_2(\eta, p_k) = \bar{\pi}_k(A + \eta) - B(1 - \bar{\pi}_k) - \bar{p}_k, \quad k = 2, \ldots, N$$

for a type η customer bidding p_k. Note that $U_1(\eta)$ takes the form of $U_1(v, \kappa)$ and $U_2(\eta, b)$ takes the form of $U_2(v, \kappa, b)$ with one dimensional user types.

Let $\mathcal{S}_G$ denote the interval for customer types that choose G service, and sets $\mathcal{S}^i_{BE}$ the interval for customer types that choose BE service and bid p_i ($i = 2, 3, \ldots, N$):

$$\mathcal{S}_G = \{\eta | U_1(\eta) \geq U_2(\eta, p_i), \; i = 2, 3, \ldots, N \text{ and } U_1(\eta) \geq 0\} \text{ and}$$
$$\mathcal{S}^i_{BE} = \{\eta | U_2(\eta, p_i) \geq U_2(\eta, p_k), \; k \neq i; U_2(\eta, p_i) > U_1(\eta), \text{ and } U_2(\eta, p_i) \geq 0\}, \; i = 2, 3, \ldots, N.$$

Note that bidding p_1 (or higher) means receiving uninterrupted BE service, equivalent to G

service. If $\bar{p}_1 < p_G$, then $\mathcal{S}_G$ becomes empty, which makes G service obsolete. However, bidding p_1 is essentially receiving G service. Therefore we need to include $\bar{p}_1 \geq p_G$ as a constraint in our model so that the uninterrupted service is labeled as G.

Then the SP's revenue maximization problem is:

$$\underset{p_G,p_1,p_2,\ldots,p_N,\boldsymbol{\pi}}{\text{maximize}} \quad p_G \cdot \int_{\eta\in\mathcal{S}_G} f(\eta)d\eta + \sum_{i=2}^{N} \bar{p}_i \cdot \int_{\eta\in\mathcal{S}^i_{BE}} f(\eta)d\eta \tag{2.2}$$

$$\text{subject to} \quad \bar{\pi}_1 \geq p_G \tag{2.3}$$

$$\mathbf{1}^T\boldsymbol{\pi} = 1 \text{ and } \boldsymbol{\pi} \geq 0 \tag{2.4}$$

$$p_1 \geq p_2 \geq \ldots \geq p_N \geq 0 \text{ and } p_G \geq 0. \tag{2.5}$$

This formulation can be simplified further by removing (2.3) and setting p_1 high enough in the solution, since p_1 does not have an effect in (2.2).

As mentioned in the introduction, the monotonicity of v/κ as κ grows plays a crucial role. In the following section, we will analyze this infinite capacity stylized model, and, specifically consider the SP's revenue maximization problem under the setting of sub-linear increase in valuation with congestion sensitivity. We will consider two models of customer heterogeneity (v, κ). In Section 2.3.1–2.3.2, we will consider that types are continuous; in Section 2.3.3, we will consider a discrete model. Super-linear increase in valuation with congestion sensitivity case is treated briefly in the Appendix.

2.3 Main Results

2.3.1 BE randomizes between 2 price levels (high/low).

The SP will offer the BE service at two price levels $\$p_H$, $\$p_L$ with $p_H \geq p_L$, and choose π, the fraction of time the BE service is priced at $\$p_L$. A customer that bids $\$p_L$ for BE will enjoy the service for π fraction of time, and if she bids $\$p_H$, she enjoys the service without interruption, and pays $\pi p_L + (1-\pi)p_H$. From Lemma 1, customers do not bid any other amount. Guaranteed service

is priced at $\$p_G$. Without loss of generality we will assume that $\pi p_L + (1-\pi)p_H > p_G$, that is, if a user wants guaranteed service, then she will choose the G service option at $\$p_G$. We will add this as a constraint to our downstream revenue optimization formulation.

In this special case, the utilities for two services can be written as a function of η as

$$U_1(\eta) = (A+\eta) - p_G \text{ and } U_2(\eta) = \pi(A+\eta) - B\eta(1-\pi) - \pi p_L = \pi(A+\eta-p_L) - B\eta(1-\pi).$$

Note that we do not need the bid value in U_2 function as it is p_L for all customers choosing the BE service.

We will first assume that the utility gained from the BE service is non-decreasing in η for any η, i.e.,

$$U_2'(\eta) \geq 0 \iff \pi - B + B\pi \geq 0 \iff \pi \geq \frac{B}{1+B}, \tag{2.6}$$

which implies a constraint on the choice of π to the SP.

Later on we will formulate and solve the problem for the case $\pi < \frac{B}{1+B}$ and show that the respective solution is sub-optimal. We redefine $\mathcal{S}_G$ and $\mathcal{S}_{BE}$ and let

$$\mathcal{S}_G = \{\eta | U_1(\eta) \geq U_2(\eta), \text{ and } U_1(\eta) \geq 0\} \text{ and } \mathcal{S}_{BE} = \{\eta | U_2(\eta) > U_1(\eta), \text{ and } U_2(\eta) \geq 0\},$$

denote the sets of customer types that choose G and BE service, respectively. From (2.6) and the fact that $U_1'(\eta) \geq U_2'(\eta)$ for any $B > 0$ and $0 \leq \pi \leq 1$, we get that

$$\mathcal{S}_G = \{\eta | \eta \geq \eta_H \text{ and } \eta \geq p_G - A\} \text{ and } \mathcal{S}_{BE} = \{\eta | \eta < \eta_H \text{ and } \eta \geq \eta_L\},$$

where η_H and η_L satisfy

$$(1+B)(1-\pi)\eta_H = p_G - \pi p_L - (1-\pi)A \text{ and } (\pi - B(1-\pi))\eta_L = \pi(p_L - A). \tag{2.7}$$

That is, customer type η chooses G if $\eta \geq \eta_H$ and $\eta \geq p_G - A$, chooses BE if $\eta_L \leq \eta < \eta_H$, and

does not join the system if $\eta < \eta_L$. The marginal types η_L, η_H are controlled by the SP through p_G, p_L, p_H, and π. Here we are restricting our analysis to the case that $\eta_H \geq \eta_L$. If $\eta_L > \eta_H$, then the BE service becomes unattractive, and the SP is offering only G service (this is also the case when $\eta_L = \eta_H$); Based on this observation, we can disregard from consideration the case where $\eta_L > \eta_H$.

We will first assume that $\eta_H \geq p_G - A$ and formulate and solve SP's revenue maximization problem. Then we will show that any solution with $\eta_H < p_G - A$ is sub-optimal and verify the assumption is satisfied under the optimal solution.

Assuming $\eta_H \geq p_G - A$, the revenue function of the SP is

$$\begin{aligned} R_1 &= p_G \bar{F}(\eta_H) + \pi p_L (F(\eta_H) - F(\eta_L)) \\ &= (p_G - \pi p_L)\bar{F}(\eta_H) + \pi p_L \bar{F}(\eta_L) \\ &= \left[\eta_H(1+B)(1-\pi) + (1-\pi)A\right]\bar{F}(\eta_H) + \left[\pi(A+\eta_L) - B\eta_L(1-\pi)\right]\bar{F}(\eta_L) := R(\eta_H, \eta_L, \pi). \end{aligned}$$

The SP's revenue maximization problem is:

$$\underset{\eta_H, \eta_L, \pi}{\text{maximize}} \quad R(\eta_H, \eta_L, \pi) \tag{2.8}$$

$$\text{subject to} \quad \eta_L \leq \eta_H, \ \pi \geq \frac{B}{1+B}, \ \pi \leq 1. \tag{2.9}$$

In contrast, if $\eta_H < p_G - A$, the revenue function reduces to

$$R_2 = p_G \bar{F}(p_G - A) + \pi p_L (F(\eta_H) - F(\eta_L)) \leq p_G \bar{F}(\eta_H) + \pi p_L (F(\eta_H) - F(\eta_L)) = R_1,$$

and the corresponding constraint set is smaller than in (2.9). It follows that any solution with $\eta_H < p_G - A$ is sub-optimal.

Next we solve (2.8)–(2.9) in terms of η_H, η_L, and π. These three parameters uniquely determine p_G and p_L from (2.7), and we show that the optimal solution satisfies $p_G \geq p_L$.

Proposition 1. *Consider the model specified by (2.1) and let $(\eta_H^*, \eta_L^*, \pi^*)$ be an optimal solution*

to (2.8)–(2.9), and p_G^* and p_L^* be the optimal prices corresponding to the solution triple. Then,

1. $\pi^* = \dfrac{B}{1+B}$,
2. $\eta_L^* = \underline{\eta}$ *with* $p_L^* = A$,
3. $\eta_H^* = p_G^* - A$.

(All proofs are given in the Appendix.) We can simplify the revenue maximization problem to

$$\underset{\eta_H}{\text{maximize}} \quad \left(\eta_H + \frac{A}{1+B}\right)\bar{F}(\eta_H) + \frac{B}{1+B}A. \tag{2.10}$$

Proposition 2. *Under the model specified by (2.1) with two price levels for the BE service, it is optimal to offer G and BE services if and only if*

$$f(\underline{\eta}) < \left(\underline{\eta} + \frac{A}{1+B}\right)^{-1}. \tag{2.11}$$

Once η_H^* and π^* are identified, the optimal price pair (p_G^*, p_L^*) can be chosen so as to satisfy Proposition 1. We mentioned earlier that without loss of generality we will restrict attention to prices such that

$$(1-\pi^*)p_H^* + \pi^* p_L^* > p_G^*, \tag{2.12}$$

i.e., it is sub-optimal for users that want uninterrupted service to choose BE but submit a high bid ($\$p_H$). Any choice of p_H that satisfies (2.12) will suffice.

To establish that Proposition 1 indeed characterizes the globally optimal solution, we need to rule out any solution where $\pi < \dfrac{B}{1+B}$ and, as a consequence, $U_2(\eta)$ is decreasing in η. If $U_2(\underline{\eta}) \leq 0$, there is no BE service, i.e., reducing to a one-service solution. Assuming $U_2(\underline{\eta}) > 0$, $\mathcal{S}_G$ and $\mathcal{S}_{BE}$ can be written as

$$\mathcal{S}_G = \{\eta | \eta \geq \eta_H \text{ and } \eta \geq p_G - A\} \text{ and } \mathcal{S}_{BE} = \{\eta | \eta < \eta_H \text{ and } \eta \leq \eta_L\},$$

where η_H and η_L satisfy (2.7). Then the SP's revenue maximization problem is:

$$\underset{\eta_H,\eta_L,p_G,\pi}{\text{maximize}} \quad p_G \cdot \int_{\eta\in\mathcal{S}_G} f(\eta)d\eta + \pi p_L \cdot \int_{\eta\in\mathcal{S}_{BE}} f(\eta)d\eta \quad \text{subject to} \quad 0 \le \pi < \frac{B}{1+B}. \tag{2.13}$$

Proposition 3. *Consider the model specified by (2.1) with two price levels for the BE service. The optimized revenue rate for (2.13) is bounded above by the optimized objective in (2.10). Therefore,* $\pi < \dfrac{B}{1+B}$ *is sub-optimal.*

2.3.2 Can the SP do better by offering BE with N >2 price levels?

In this section we are allowing the SP to choose more than two price levels as in Section 2.2.3.

Similar to the two-price-level case, we assume that the utility gained from the BE service is non-decreasing in η for any η and $\bar{\pi}_i$ values, i.e.,

$$\begin{aligned} U_2'(\eta, p_i) \ge 0,\ i = 2,3,\ldots,N &\iff \bar{\pi}_i \ge \frac{B}{1+B},\ i = 2,3,\ldots,N \\ &\iff \pi_N \ge \frac{B}{1+B}. \end{aligned}$$

This assumption ensures that as users value more and become more congestion sensitive, the utility they receive from the BE service does not decrease.

With the addition of this assumption, the problem (2.2)–(2.5) becomes

$$\underset{p_G,p_2,p_3,\ldots,p_N,\boldsymbol{\pi}}{\text{maximize}} \quad p_G \cdot \int_{\eta\in\mathcal{S}_G} f(\eta)d\eta + \sum_{i=2}^{N} \bar{p}_i \cdot \int_{\eta\in\mathcal{S}^i_{BE}} f(\eta)d\eta \tag{2.14}$$

$$\text{subject to} \quad \pi_N \ge \frac{B}{1+B},\ 1^T\boldsymbol{\pi} = 1,\ \boldsymbol{\pi} \ge 0 \tag{2.15}$$

$$p_2 \ge p_3 \ge \ldots \ge p_N \ge 0 \text{ and } p_G \ge 0. \tag{2.16}$$

Proposition 4. *Consider the model specified by (2.1) and let k^* be the number of distinct price levels offered in BE service at the optimal solution of (2.14)–(2.16). Then, an optimal solution is to use $k^* = 2$ with the structure specified in Proposition 1.*

Once again, for the model in (2.1) with the affine relation between (v, κ), it is optimal to offer G service and BE service with two-price-level if and only if (2.11) is satisfied.

2.3.3 General Dependence Between Valuation and Congestion Sensitivity

So far we have restricted attention to the affine dependence between (v, κ) that allowed us to solve the resulting revenue maximization problem in closed form. In this subsection we briefly consider a market where the (v, κ) dependence is general, yet still v/κ grows sub-linearly with respect to κ, and primarily show that in such a setting the SP may wish to offer more than 2 price levels for BE service.

Suppose there are n customer types and N price levels in BE service ($N > n$). Let $\kappa_1 > \kappa_2 > \ldots > \kappa_n > 0$ with $v_1 \geq v_2 \geq \ldots \geq v_n > 0$ such that $\frac{v_1}{\kappa_1} < \frac{v_2}{\kappa_2} < \ldots < \frac{v_n}{\kappa_n}$. The fraction of users that are of type i is λ_i. Let $p_1 \geq p_2 \geq \ldots \geq p_N \geq 0$ be the price levels with $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ such that π_j is the fraction of time the system is in price level p_j ($\mathbf{1}^T\boldsymbol{\pi} = 1$ and $\boldsymbol{\pi} \geq 0$).

$$\text{Model 2:} \quad v_1 \geq v_2 \geq \ldots \geq v_n > 0, \ \kappa_1 > \kappa_2 > \ldots > \kappa_n > 0, \ \frac{v_1}{\kappa_1} < \frac{v_2}{\kappa_2} < \ldots < \frac{v_n}{\kappa_n}. \tag{2.17}$$

The objective of the SP is to maximize its revenue rate by offering a price vector $(p_G, \boldsymbol{p})$ and an availability vector $\boldsymbol{\pi}$. As previously, one can restrict attention to customer bids in $\{p_1, p_2, \ldots, p_N\}$.

Proposition 5. *Consider the model specified by (2.17) and let p^* be the optimal price when there is only G service. Then, it is optimal to offer G and BE services together if and only if $p^* > v_n$.*

Proposition 5 shows that if some customer types choose not to buy under the optimal single service level (only G) solution, then the SP can extract more revenue by offering the second service level (BE). We have shown above that when there is linear dependence between v and κ, it is enough to offer BE service with two price levels. The example given below shows that when (v, κ) have a general dependence structure and still v grows sub-linearly with respect to κ, i.e., the model in (2.17), it may be optimal to use $k > 2$ price levels. In particular, the following example shows that the property proven in Proposition 4 no longer holds (the relation between (v, κ) is quadratic):

Example: Three customer types with $\boldsymbol{\lambda} = (1, 1, 1)$, $\boldsymbol{v} = (4, 2, 1)$, $\boldsymbol{\kappa} = (16, 4, 1)$. The optimal solution is $p_G = 4$, $\boldsymbol{p} = (p_1, 6, 2/3)$, $\boldsymbol{\pi} = (1/7, 3/28, 3/4)$ where $p_1 > 20$.

2.4 Data

We first offer a description of price data from AWS (a cloud computing platform offered by Amazon), and then offer a brief calibration and discussion of our model on AWS data. This section links the observed data with our model and provides useful insights about customer characteristics and price points set by Amazon.

2.4.1 Descriptive Statistics

Amazon is the biggest cloud computing SP. They offer over 1,000 products to the IaaS market in 9 regions globally. For each product, the price trace of the last 90 days is made publicly available by Amazon. We have obtained price traces from August 2013 onwards for the spot instances using an automated script that we programmed, which runs everyday and downloads and stores the price traces of the last 24 hours, for all products. This script has enabled us to have a longer time frame for the price history. Amazon does not disclose any information other than the price traces.

We have analyzed the data traces from March 1, 2015 to August 31, 2015 for 1,122 products. The products are categorized under five different classes by AWS: "compute optimized," "general purpose," "GPU instances," "memory optimized," and "storage optimized." In each of these classes there are multiple machine sizes. Moreover, prices differ with respect to the location of the product and the operating system the product has. To facilitate reporting statistics on pools of different products with different on-demand prices, we normalize the spot and on-demand prices of each product by the respective on-demand price. In this manner, a normalized spot price is unit-less and expressed and understood as a multiple of the underlying on-demand price; all products have a normalized on-demand price equal to 1. To get a better sense of pricing dynamics, first we look at the descriptive statistics per product. For each product, we calculate: the average normalized spot price; the normalized spot price range; the average uptick and downtick inter-arrival times;

the average magnitudes of the corresponding spot price jumps; and, the fraction of time the spot price is greater than on-demand price. Table 2.2 shows that the mean of the average normalized spot prices across products is about half of the on-demand price. More than 92% of the products have a time-average spot price less than 1, which means that for more than 92% of the products, procuring spot instances, with sufficiently high bids so as never to be shut off, would cost less than on-demand instances for the whole 6-month period. We discuss this result more in Section 2.4.2. Further, summary statistics shows that the range of spot price fluctuations is wide, more than three times of the corresponding on-demand price on average. The average inter-arrival time of an uptick (downtick) price change is in the order of hours, and the average magnitude of an uptick (downtick) is about one third of the on-demand prices. Lastly, for most products the spot price is below the corresponding on-demand price for more than 90% of the time. Figure 2.1 shows the distribution of each of these categories (with a few outliers discarded in each plot).

Table 2.2: Summary of descriptive statistics per product

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Average normalized spot price	0.037	0.213	0.366	0.515	0.780	3.756
Normalized spot price range	0.000	0.829	1.511	3.216	4.772	39.050
Avg. uptick inter-arrival time (hrs)	0.000	3.071	7.115	35.200	27.930	1428.000
Avg. downtick inter-arrival time (hrs)	0.000	3.088	6.924	33.950	27.200	1111.000
Average uptick magnitude	0.000	0.144	0.284	0.449	0.537	9.740
Average downtick magnitude	0.000	0.143	0.290	0.459	0.537	12.150
Fraction of time spot>on-demand	0.000	0.000	0.008	0.072	0.066	1.000

Next, we assume that a user selects spot service and bids sufficiently high so that she is never outbid and would enjoy uninterrupted service. For each different possible time of arrival, we record the average price she would pay per hour if she stayed in the system for 1 hour, 1 day, 1 week, or 1 month. For each of the 4 usage durations, we average across time of arrival. The results are reported in Figure 2.2. These plots show that most Windows products have higher spot prices compared to Linux/UNIX products, i.e., the potential gain from the spot market is less for Windows products. The four panels in Figure 2.2 are similar, suggesting that the usage duration does not play an important role on the selection between spot (at maximum bid) versus on-demand.

To illustrate the fluctuations in the prevailing spot prices over time, we focused on the running

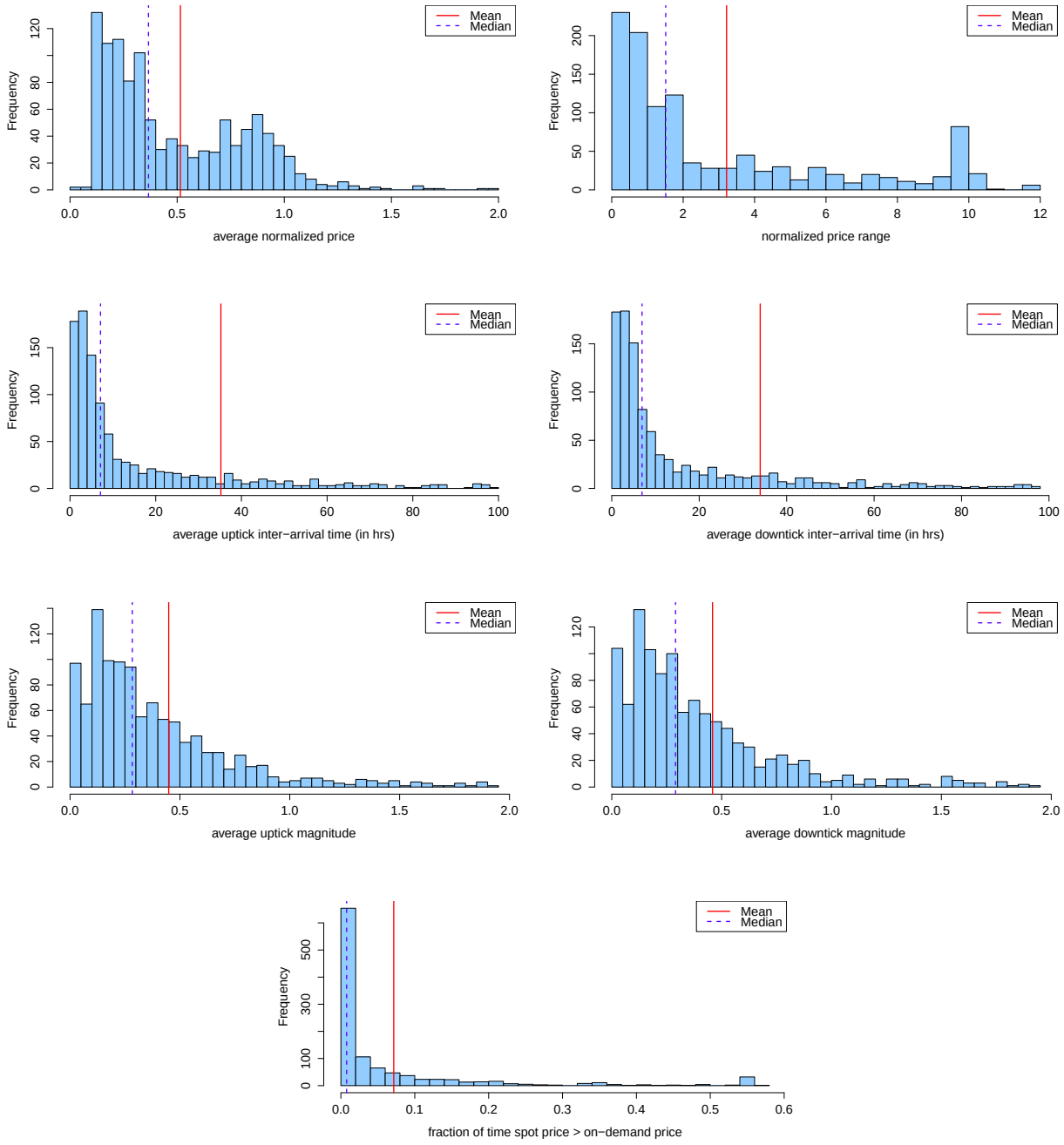


Figure 2.1: Histogram of descriptive statistics per product

averages of the prevailing spot prices for daily, weekly, and monthly usage updated daily for every class of product. Figure 2.3 summarizes how the average spot price fluctuates over time under daily,

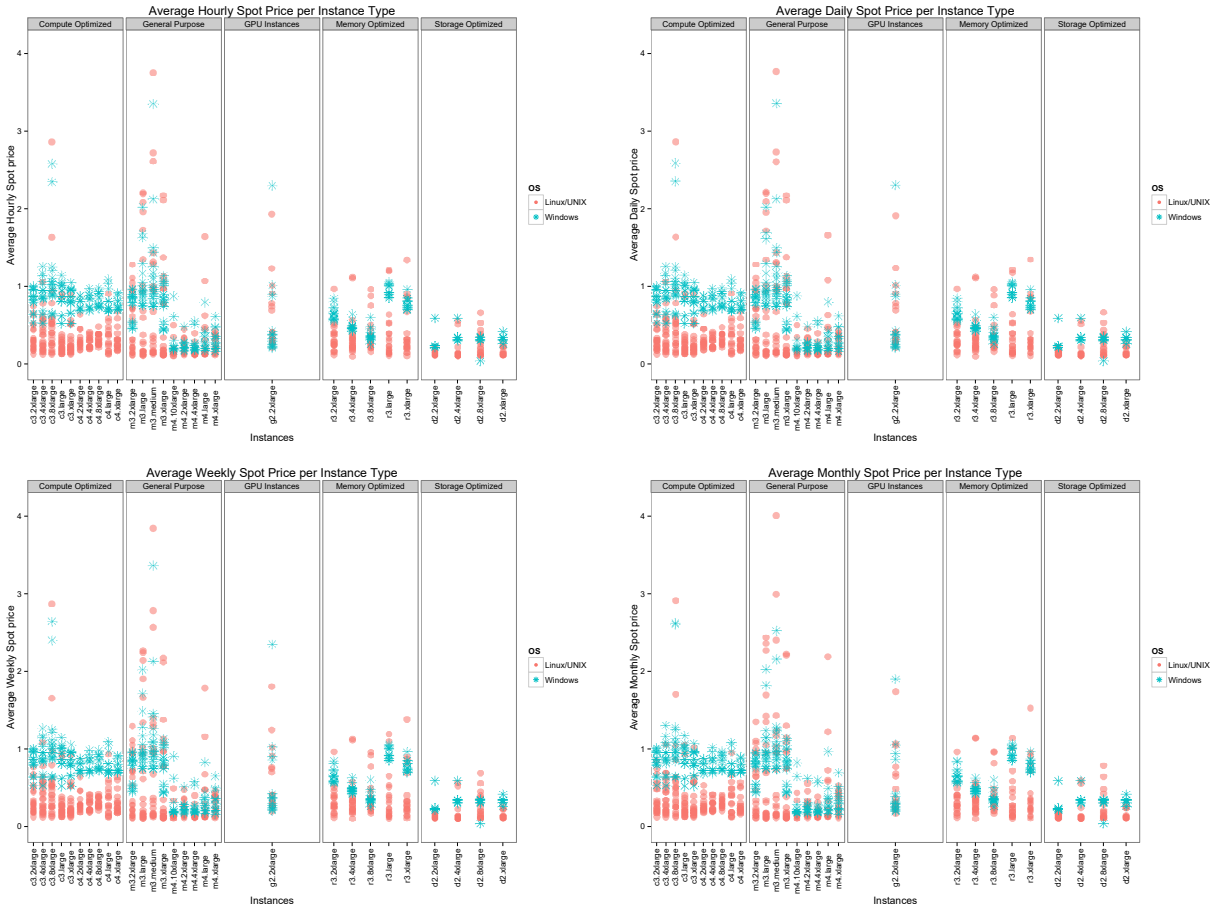


Figure 2.2: Average prices for different duration of usage

weekly, and monthly usage of "GPU Instances" for Linux/UNIX and Windows machines; there are 18 such products for each operating system in total. While the solid line represents the average price of all products in this class, the blue (shaded) area denotes ± one standard deviation band from the average price (computed across the respective 18 data points in each time point). For this product class, prices for both operating systems follow similar patterns whereat the spot market is cheaper than the on-demand market until the beginning of August. The standard deviation also increased during that period, implying also increased variation across different "GPU Instances" products during this peak period. We observe different patterns in other product classes.

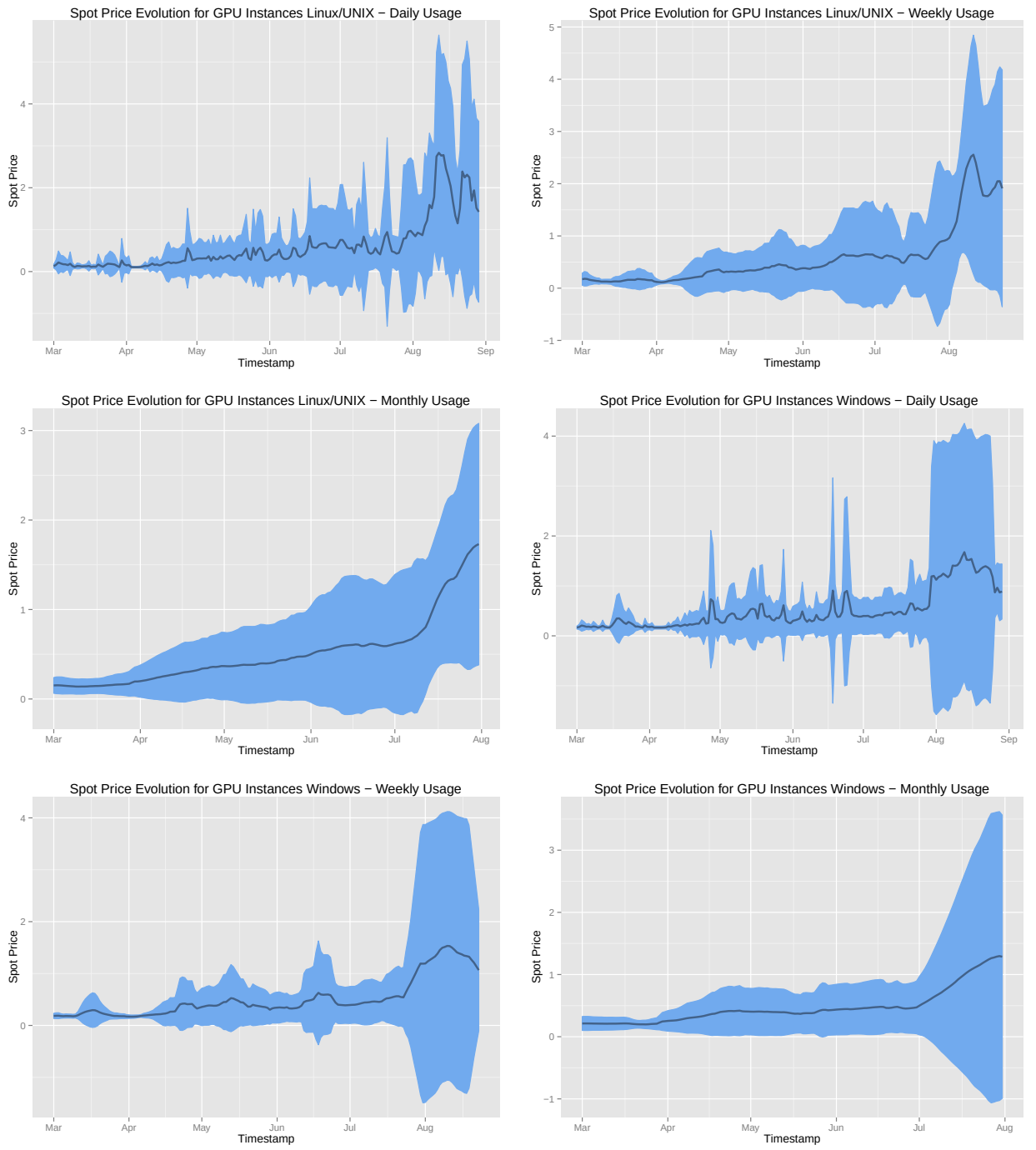


Figure 2.3: Average price change over time with different usage times

2.4.2 Data Evidence

We can use the AWS data to calibrate our model primitives. To repeat, the model of Section 2.2.3 assumes a linear dependence between valuation and congestion rates, which implies that valuation grows sub-linearly relative to the congestion rate. The results of Section 2.3.2 suggest that it suffices for the SP to offer the spot price service with just two price levels (high/low). We proceed as follows. We assume the model specified in Section 2.2.3 is in force and that the SP follows the optimal policy derived in Section 2.3.1. We compute the empirical spot price occupancy distribution, and approximate it with a two-level (p_H, p_L, π) distribution. We then derive the implied user valuation and congestion model parameters. We will approximate the empirical distribution by the triple (p_H, p_L, π) that is closest in the sense of the Kantorovich metric, where p_H is the high price, p_L is the low price, and π is the fraction of time the price is equal to p_L. (The Kantorovich metric between two random variables X and Y in $\mathbb{R}$ is defined as $K(X,Y) = \int_{\mathbb{R}} |F_X(x) - F_Y(x)|\, dx$, where F_X and F_Y are the cumulative distribution function of X and Y, respectively.)

Assuming that user types are distributed uniformly on $(0, \eta_{max})$, the parameters $A, B, \eta_{\max}$ can be calculated using the results from Section 2.3.1. Specifically, for given (p_H, p_L, π),

$$B = \frac{\pi}{1-\pi},\ A = p_L,\ \eta_{max} = 2(p_G - A) + \frac{A}{1+B}$$

where $p_G = 1$ since the price path is normalized based on the G service price. Using these parameters we can get the implied valuation and congestion sensitivity parameter for each user type η: her valuation rate is equal to $A + \eta$ and her congestion sensitivity parameter is equal to $B\eta$. Lastly, we analyze if the implied parameters satisfy the conditions on p_H and p_L, i.e. whether $p_G < (1-\pi)p_H + \pi p_L$, $p_H > 1$, $p_L < 1$ hold. Our analysis containing the data for the period Mar. 1, 2015 – Aug. 31, 2015 has shown that out of 1,122 products, only 63 of them satisfy all these conditions. The summary statistics of the normalized price path of these 63 products is given in Table 2.3. Calibrating our model on the observed data we get the parameters shown in Table 2.4.

The normalized estimated parameters suggest the following:

- Valuation per unit time: $(A + \eta) \sim U(0.6, 1.5)$.

Table 2.3: Summary statistics of price paths

Avg. Price	Price Range	Reserve Price	Price>1
1.446	6.438	0.608	29%

Table 2.4: Estimated parameters on average

A	B	η_{max}	η_H	p_L	p_H	π
0.638	14.992	0.835	0.362	0.638	5.291	0.784

- Congestion cost per unit of downtime: $B\eta \sim U(0, 12.5)$. Specifically, we note that congestion costs when the system is down, due to lost revenue and possibly lost goodwill/reputation, can be up to 4x-10x of the valuation per unit time.
- Fraction of downtime: $1 - \pi = 0.216$.
- Congestion cost per unit time (due to service interruption): $\sim U(0, 2.7)$.
- Lowest valuation per unit time choosing G: $A + \eta_H = 1$.

These parameter estimates suggest that for high customer types, the disutility from service disruption in spot service is of the same order of magnitude (or higher) as the valuation itself, and as a result, only the least congestion-sensitive users will choose that option. In our data, this seems to be the lower 40% of the distribution that wants G service.

Finally, as noted earlier, for more than 92% of the products, a user would be better off selecting the spot option and bid sufficiently high so as to receive continuous uninterrupted service for the whole 6-month period. Based on our model, this would suggest insufficient degradation of the spot service option by the SP so as to incentivize congestion sensitive customers to choose the on-demand service option. Assuming the estimated parameters on Table 2.4 also hold for all offered products and the demand for each of these products is the same, our back-of-the-envelope calculation shows that Amazon could almost double the revenue extracted from these products by further optimizing the pricing of the spot option. Of course, this calculation disregards other (unmodeled) economical and technological considerations that may affect such tactical pricing decisions, and for which we lack transparent data.

2.5 Discussion: State Dependent Bidding

In this section, we are going to create a state dependent model where users observe the state, current spot price, and choose their bid accordingly. The system is modeled as a Discrete Time Markov Chain (DTMC) and transitions between the states happen based on this DTMC (with no absorbing state). We assume that user arrivals happen at the beginning of a period and users are not allowed to change their bid over time.

Let $R(b, m, p_i)$ be the amount a user pays by bidding b for a job of length m periods when the state of the system is p_i at the time of arrival. Let $T(b, m, p_i)$ be the number of periods the system is down to complete m period job by bidding b when the initial state is p_i. Finally, let P_{ij} be the transition probability from state i to j, where $\sum_j P_{ij} = 1 \; \forall i$ and $P_{ij} \geq 0 \; \forall i, j$.

We will analyze the system where there are two states low price L and high price H.

Under the two-state DTMC model, if the user bids p_l when the current price level is p_l:

$$\mathbb{E}[R(p_L, m, p_L)] = mp_L \text{ and } \mathbb{E}[T(p_L, m, p_L)] = \frac{(m-1)(1-P_{LL})}{(1-P_{HH})}.$$

If the user bids p_L when the current price level is p_H:

$$\mathbb{E}[R(p_L, m, p_H)] = mp_L \text{ and } \mathbb{E}[T(p_L, m, p_H)] = \frac{1+(m-1)(1-P_{LL})}{(1-P_{HH})}.$$

Before analyzing the cases when she bids p_H, let's simplify the notation and define

$$R_L(m) := \mathbb{E}[R(p_H, m, p_L)] \text{ and } R_H(m) := \mathbb{E}[R(p_H, m, p_H)]$$

If she bids p_H when the current price level is p_L:

$$R_L(m) = p_L + (1-P_{LL})R_H(m-1) + P_{LL}R_L(m-1) \text{ and } \mathbb{E}[T(p_H, m, p_L)] = 0,$$

and if she bids p_H when the current price level is p_H:

$$R_H(m) = p_H + P_{HH}R_H(m-1) + (1-P_{HH})R_L(m-1) \text{ and } \mathbb{E}[T(p_H, m, p_L)] = 0,$$

with $R_L(0) = R_H(0) = 0$

If the system is in steady state, that is, the payment and processing time are independent of the initial state, and the user bids p_L:

$$\mathbb{E}[R(p_L, m)] = mp_L \text{ and } \mathbb{E}[T(p_L, m)] = \frac{m(1-\pi)}{\pi},$$

where $\boldsymbol{\pi} = (\pi, 1-\pi)$ is the steady state distribution for the states (p_L, p_H), respectively, and using $\boldsymbol{\pi}P = \boldsymbol{\pi}$,

$$\pi = \frac{1-P_{HH}}{2-P_{LL}-P_{HH}}.$$

If the system is in steady state and the user bids p_H:

$$\mathbb{E}[R(p_H, m)] = m(\pi p_L + (1-\pi)p_H) \text{ and } \mathbb{E}[T(p_H, m)] = 0.$$

Note that we drop the third parameter in the functions R and T in steady state formulations.

We want to see the effect of state dependency on customer choice. More precisely, we want to see how customer utility (as a function of payment and total downtime) changes based on the initial state of the system, and compare it with the utility under the steady state.

If the user bids p_L, the payment function becomes independent of the initial state of the system; however, the downtime depends on the initial state. If the user bids p_H, then the payment function depends on the initial state, not the downtime. Therefore, for the case that the user bids p_L, we are going to compare the downtime with the downtime under steady state; for the case that the user bids p_H, the comparison will be on the payment.

When the user bids p_L, the ratio of expected downtime over the whole processing time under

steady state is

$$\frac{\mathbb{E}[T(p_L, m)]}{m + \mathbb{E}[T(p_L, m)]} = \frac{\frac{m(1-\pi)}{\pi}}{m + \frac{m(1-\pi)}{\pi}} = 1 - \pi$$

It is easy to see that

$$\frac{\mathbb{E}[T(p_L, m, p_L)]}{m + \mathbb{E}[T(p_L, m, p_L)]} \xrightarrow[m\to\infty]{} 1 - \pi \quad \text{and} \quad \frac{\mathbb{E}[T(p_L, m, p_H)]}{m + \mathbb{E}[T(p_L, m, p_H)]} \xrightarrow[m\to\infty]{} 1 - \pi.$$

We pick the duration of one period as 1 hour. The data shows that the transitions from low price to high price or vice versa in every few hours. Based on that observation, we set $P_{LL} = 0.8, P_{HH} = 0.5$, which means that a low to high jump occurs in every 5 hours on average, while a high to low jump occurs in every 2 hours on average. Lastly, using Table 2.4, we pick p_H to be 8.8 times p_L. Figure 2.4 shows if the user stays in the system at least for a few days, then the observed downtime is almost equal to the steady state case.

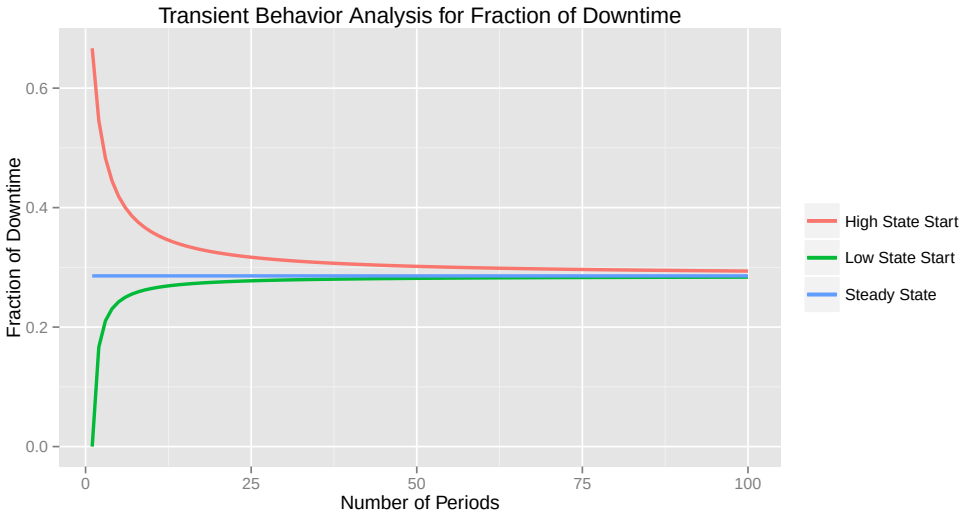


Figure 2.4: Transient behavior of fraction of downtime

Next we will compare the payment functions when the user bids p_H. The payment function

under steady state is

$$\begin{aligned}\mathbb{E}[R(p_H, m)] &= m(\pi p_L + (1-\pi)p_H) \\ &= m\left(\frac{(1-P_{HH})p_L}{2-P_{LL}-P_{HH}} + \frac{(1-P_{LL})p_H}{2-P_{LL}-P_{HH}}\right).\end{aligned}$$

We can plot $\frac{\mathbb{E}[R(p_H,m)]}{m}$, $\frac{\mathbb{E}[R(p_H,m,p_L)]}{m}$, and $\frac{\mathbb{E}[R(p_H,m,p_H)]}{m}$ for given p_L, p_H, P_{LL}, and P_{HH}; and see how the normalized payment functions per unit time change as the required processing time increases. ($\frac{\mathbb{E}[R(p_H,m,p_L)]}{m}$, and $\frac{\mathbb{E}[R(p_H,m,p_H)]}{m}$ are normalized using $\frac{\mathbb{E}[R(p_H,m)]}{m}$.) Using the same parameters as before, similar to our conclusion on downtime, Figure 2.5 shows that the expected payments given the initial state converge to the expected payment under steady state after a few days.

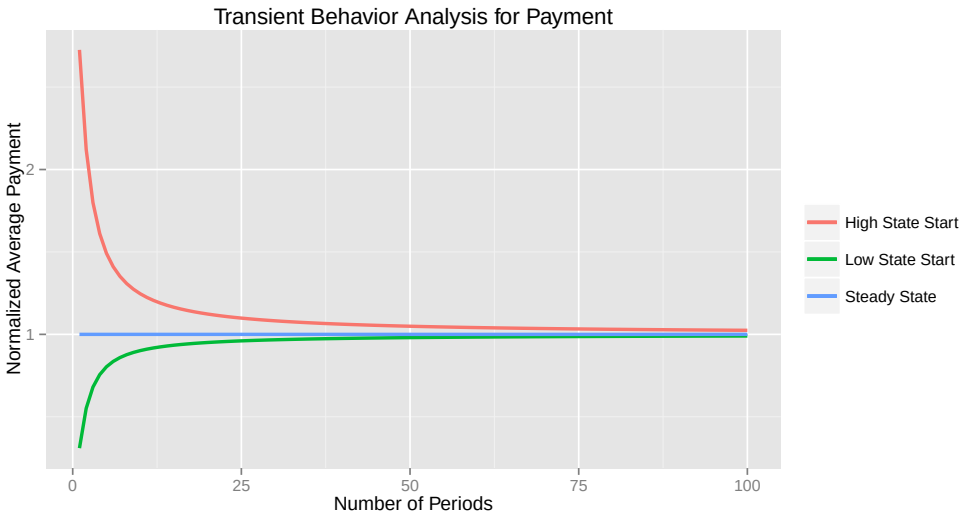


Figure 2.5: Transient behavior of payment

Chapter 3

Competition on Price and Quality in Cloud Computing

In this chapter, we study the optimal price-quality menu to offer for cloud computing providers. In Section 3.1, we provide the basics of our model and validate it. In Section 3.2, we study the revenue maximization problem under monopoly. Section 3.3 discusses the problem under duopoly. Section 3.4 highlights some possible model extensions. In Section 3.5, we reconcile our model predictions and real-world behavior. Finally, we discuss our findings and conclude our work in Section 3.6.

3.1 Model

On the customer side, there are n customer types indexed by i, where customer type i has a valuation (v_i), delay sensitivity (c_i), both per unit time of workload[1] under nominal quality level, and arrival rate (λ_i). We assume there is only one type of workload which can be parallelizable up to a certain extent, and all customer types need to run the same workload. We relax this assumption and discuss the results in §3.4.

On the provider side, there are m different service providers indexed by j, where service provider j chooses a base quality level q_{j1} ($0 < q_{j1} < \bar{q}_j$), where $\bar{q}_j$ is the maximum base quality level that can

[1] We use *workload* and *job* interchangeably throughout the chapter.

be offered, price per unit time for the base quality level p_{j1}, and number of quality levels to offer L_j. Each service provider has an inherent performance scaling factor α_j determined by the structure and technology used (which will later estimate, $0.5 < \alpha_j < 1$), and each offers a price-quality menu (p_{jk}, q_{jk}), where $p_{jk} = 2^{k-1}p_{j1}$, $q_{jk} = 2^{k-1}\alpha_j^{k-1}q_{j1}$ for $k = 1, 2, \ldots, L_j$.

The size of a workload is defined as the time it takes to complete the job using a baseline quality product. We are assuming job completion time function $W(w, q) := \frac{w}{q}$ where w is the completion time of a job under baseline quality and q is the quality level.

The utility of customer type i with workload w, choosing quality level k of service provider j is

$$\begin{aligned} U_{ijk} &= v_i w - c_i W(w, q_{jk}) - p_{jk} W(w, q_{jk}) \\ &= w\left(v_i - \frac{c_i + 2^{k-1}p_{j1}}{2^{k-1}\alpha_j^{k-1}q_{j1}}\right), \end{aligned}$$

with $U_{ij0} = 0$ representing the no-buy option.

Then, customer type i chooses quality level k^* of service provider j^*, where

$$j^* = \underset{j\in\{1,2,\ldots,m\}}{\operatorname{argmax}} \{\max_{k\in\{0,1,2,\ldots,L_j\}} U_{ijk}\} \text{ and } k^* = \underset{k\in 0,1,2,\ldots,L_{j^*}}{\operatorname{argmax}} U_{ij^*k}.$$

Service providers are revenue maximizers.[2] Assuming each customer type has workload w, the revenue function for service provider j is

$$\Pi_j(p_{j1}, q_{j1}) = w\left[\sum_{i\in\mathcal{S}_{j1}} \lambda_i \frac{p_{j1}}{q_{j1}} + \sum_{i\in\mathcal{S}_{j2}} \lambda_i \frac{p_{j1}}{\alpha_j q_{j1}} + \ldots + \sum_{i\in\mathcal{S}_{jL_j}} \lambda_i \frac{p_{j1}}{\alpha_j^{L_j-1} q_{j1}}\right],$$

where S_{jk} is the set of customer types that choose quality level k of service provider j ($k = 1, 2, \ldots, L_j$).

Model Validity. All big cloud providers offer different product families to their customers, and

[2] We later discuss how to incorporate costs in the analysis.

each product family is customized for special kind of workloads. Amazon has *t2*, *m4*, *c4*[3]; Google has *standard*, *high-mem*[4]; and Microsoft has *A*, *D*, *G*[5], to name a few. In most of these product families companies offer 4 different product sizes with different prices; however, what they actually pick is a base level product configuration and a price for this base level. Once the base level is picked, second product is configured as the double the size of the base product with twice the price, third product is configured as the double of the second product, and finally fourth is configured as the double of the third product. Price - Configuration menu for Microsoft's D product family with Linux Machine for Central US region is given in Table 3.1 as an example of this structure.

Table 3.1: Azure price – configuration menu

Product	Cores	Ram	Disk Sizes	Unit Price
*D*1	1	3.5 GB	50 GB	$0.077/hr
*D*2	2	7 GB	100 GB	$0.154/hr
*D*3	4	14 GB	200 GB	$0.308/hr
*D*4	8	28 GB	400 GB	$0.616/hr

To validate our price-quality model, we have picked two service providers (a and b) with one product family for each. Therefore, we have products a_i and b_i, lower i indicating smaller size product, with unit prices $2^{i-1}0.100$ and $2^{i-1}0.126$ ($i = 1, 2, 3, 4$) for providers a and b, respectively.[6] The workload we have chosen for this experiment is *DaCapo*[7] (Blackburn *et al.*, 2006). DaCapo is a benchmark suite that runs different Java workloads with non-trivial memory loads. We have run the workload once a day for one week at the same time for both providers with different product sizes in similar regions. Average running times and cost values are summarized in Table 3.2.[8] Contrary to the previous literature (Ou *et al.*, 2012; Schad *et al.*, 2010; Wang and Ng, 2010), our experiment with one type of workload has shown that the job completion time does not vary too much over time for the same product (the average standard deviation in completion time is less

[3] https://aws.amazon.com/ec2

[4] https://cloud.google.com/compute

[5] https://azure.microsoft.com/en-us/services/virtual-machines

[6] For anonymity, names are filtered and unit prices are transformed.

[7] http://www.dacapobench.org

[8] In total cost calculations, it is assumed that cost is incurred per second basis.

than 5% of the mean completion time per product), unless the product is a burstable type product, or has a shared CPU (*t2* product family in AWS, *f1-micro* in Google).

Table 3.2: Price-quality comparison

Product	Unit Price	Avg. Comp. Time	Total Cost
a_1	\$0.100/hr	738.14 sec	\$0.021
a_2	\$0.200/hr	490.47 sec	\$0.027
a_3	\$0.400/hr	383.90 sec	\$0.043
a_4	\$0.800/hr	360.57 sec	\$0.080
b_1	\$0.126/hr	719.71 sec	\$0.025
b_2	\$0.252/hr	468.00 sec	\$0.033
b_3	\$0.504/hr	360.71 sec	\$0.051
b_4	\$1.008/hr	308.71 sec	\$0.086

Figure 3.1 shows how products are located in time/cost space for this specific workload. User utility increases as we move towards the origin, as it signals faster performance and lower cost. Interestingly, all product offerings are Pareto efficient, that is, there is no product that is both cheaper and faster than any other products. Therefore, each product can be chosen by a rational customer based on her time/cost trade-off. Since users differ with respect to time sensitivity, they will choose different performance level.

Assuming $w = 1000$ for the workload we are experimenting with, we try to estimate the scaling factor and base quality level for both products. We find that $(\alpha_1, q_1) = (0.693, 1.355)$, and $(\alpha_2, q_2) = (0.616, 1.733)$ for service providers a and b. respectively. The mean percentage absolute error of our fit is 8% for both providers.[9] Hence, we can conclude that our model with quality level function $2^{k-1}\alpha_j^{k-1}q_{j1}$ is fairly realistic. Note that in reality, α value is not only provider dependent, but also workload dependent. No matter how good infrastructure one provider has, if the workload to be run is not parallelizable, α value would end up being low. We are doing our analysis for a specific type of workload which is fairly parallelizable, and we discuss possible extensions to this in §3.4. Reader may refer to Amdahl (1967) and Gustafson (1988) for detailed analysis on maximum achievable performance gain with parallelization formulations based on workload type.

[9]The accuracy of our fit is not dependent on $w = 1000$ assumption. Any w would yield the same accuracy.

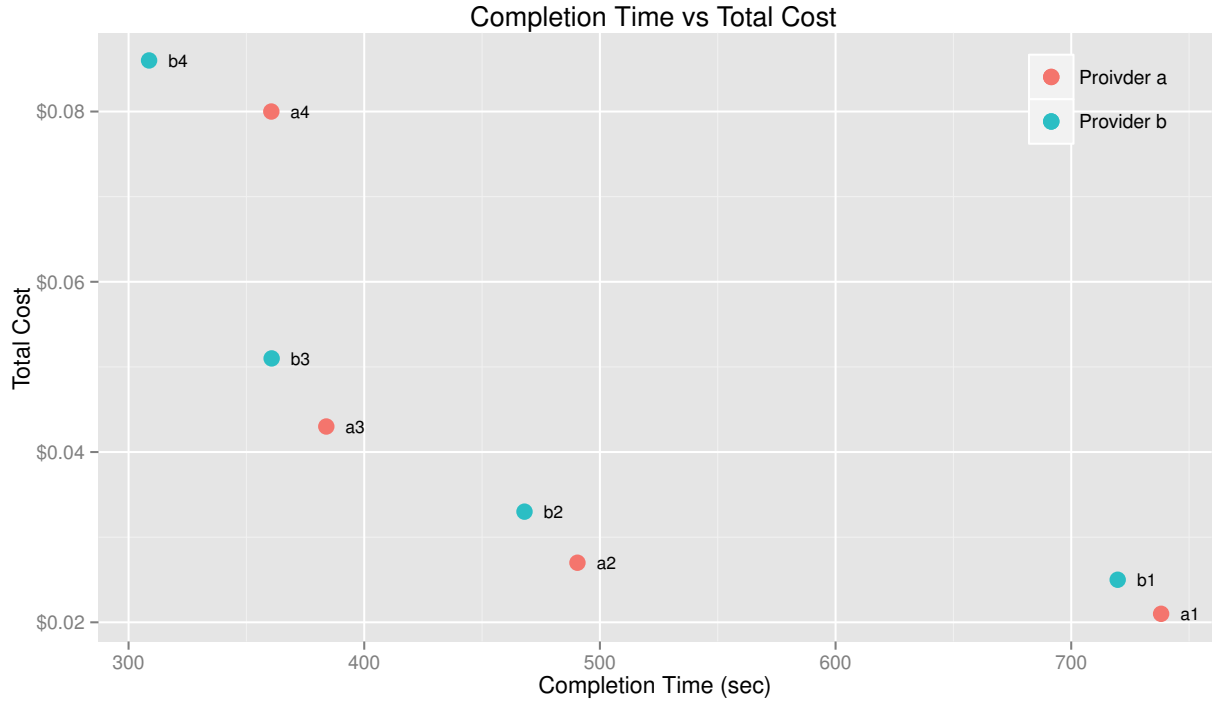


Figure 3.1: Completion time vs total cost

3.2 Revenue Maximization under Monopoly

After describing our model and validate it, we start our analysis with a monopolistic, revenue maximizing service provider, and therefore, drop the subscript j. In the first part of this section, we allow the provider to set both base price and quality level, and in the second part we maximize the provider's revenue for a given base quality level and we provide some numerical examples.

3.2.1 Optimal Price-Quality Menu

The service provider chooses base quality level q_1, base price level p_1 and number of quality levels to offer L; scaling factor α is endogenous.

The revenue of the monopolistic service provider when she offers only one quality level (q_1):

$$\Pi_1(p_1, q_1) = \sum_{i \in \mathcal{I}(p_1, q_1)} \lambda_i p_1 \frac{w}{q_1},$$

where $\mathcal{I}(p_1, q_1)$ is the set of customer types that choose to buy the product when the price is p_1 and the quality level is q_1.

When the service provider offers two quality levels $(q_1, 2\alpha q_1)$, the revenue becomes

$$\Pi_2(p_1, q_1) = \sum_{i \in \mathcal{I}_1(p_1,q_1)} \lambda_i p_1 \frac{w}{q_1} + \sum_{i \in \mathcal{I}_2(p_1,q_1)} \lambda_i p_1 \frac{w}{\alpha q_1},$$

where $\mathcal{I}_1(p_1, q_1)$ is the set of customer types that choose to buy the low quality product and $\mathcal{I}_2(p_1, q_1)$ is the set of customer types that choose to buy the high quality product when the price-quality menu is $\{(p_1, q_1), (2p_1, 2\alpha q_1)\}$.

Lemma 2. *$\mathcal{I}_1(p_1, q_1) \cup \mathcal{I}_2(p_1, q_1) \supseteq \mathcal{I}(p_1, q_1)$. for any (p_1, q_1).*

Proposition 6. *Offering an additional quality level generates at least as much revenue as offering fewer number of quality levels.*

Proposition 6 shows that offering a higher quality level does not cannibalize the service provider's revenue. The next step is to formulate customer preferences on different quality levels.

Proposition 7. *If $c_i \in \left[0, \frac{2p_1(1-\alpha)}{2\alpha - 1}\right)$, then customer type i chooses the first quality level given that her utility is nonnegative. Similarly, if $c_i \in \left[\frac{2^{k-1}p_1(1-\alpha)}{2\alpha-1}, \frac{2^k p_1(1-\alpha)}{2\alpha-1}\right)$, then customer type i chooses quality level k given that her utility is nonnegative.*

So far we have assumed that all customer types have the same valuation v and workload requirement w. From this point on, we make an additional assumption that there is a continuum of customer types that differ with respect to the delay sensitivity parameter, c, where $c \sim U(0, \bar{c})$. Moreover, we assume that the number of quality levels can be at most 4, which is aligned with what we observe in the market.[10]

[10] There are a few product families with 5 or 6 quality levels, such as Google's *n1-standard* and *n1-highmem* families. We only choose 4 quality levels here for simplification.

We start with the revenue maximization problem with one quality level:

$$\begin{aligned} \underset{p_1,q_1}{\text{maximize}} \quad & \Pi_1(p_1,q_1) = \frac{1}{\bar{c}}\frac{p_1}{q_1}\left(\min\{vq_1 - p_1, \bar{c}\}\right) \\ \text{subject to} \quad & 0 < q_1 \leq \bar{q_1},\ p_1 \geq 0. \end{aligned} \tag{3.1}$$

Let (p_1^*, q_1^*) be the optimal base price and quality level and Π_1^* be the optimal revenue for (3.1). Then, using first order conditions, it can easily be shown that

$$p_1^* = \begin{cases} \dfrac{v\bar{q}}{2}, & \text{if } \bar{q} \leq \dfrac{2\bar{c}}{v} \\ v\bar{q} - \bar{c}, & \text{if } \bar{q} > \dfrac{2\bar{c}}{v} \end{cases}, \text{ and } \Pi_1^* = \begin{cases} \dfrac{v^2\bar{q}}{4\bar{c}}, & \text{if } \bar{q} \leq \dfrac{2\bar{c}}{v} \\ v - \dfrac{\bar{c}}{\bar{q}}, & \text{if } \bar{q} > \dfrac{2\bar{c}}{v} \end{cases}.$$

As the number of quality levels offered increases, the revenue maximization problem gets more complicated, and the closed form solutions have multiple cases. Therefore, we only present the results for the case where there are *exactly* four quality levels assuming $\bar{q}$ is high enough that it is not binding in the problem.[11] The revenue function can be written as

$$\begin{aligned} \Pi_4(p_1,q_1) = \frac{1}{\bar{c}}\frac{p_1}{q_1}\Bigg\{ & \frac{1}{\alpha^3}\left[\min\left\{8\alpha^3 vq_1 - 8p_1, \bar{c}\right\} - \min\left\{\frac{8p_1(1-\alpha)}{2\alpha-1}, \bar{c}\right\}\right] \\ & + \frac{1}{\alpha^2}\left[\min\left\{\frac{8p_1(1-\alpha)}{2\alpha-1}, \bar{c}\right\} - \min\left\{\frac{4p_1(1-\alpha)}{2\alpha-1}, \bar{c}\right\}\right] \\ & + \frac{1}{\alpha}\left[\min\left\{\frac{4p_1(1-\alpha)}{2\alpha-1}, \bar{c}\right\} - \min\left\{\frac{2p_1(1-\alpha)}{2\alpha-1}, \bar{c}\right\}\right] \\ & + \min\left\{\frac{2p_1(1-\alpha)}{2\alpha-1}, \bar{c}\right\}\Bigg\}. \end{aligned} \tag{3.2}$$

Proposition 8. *Let the optimal price and base quality level in (3.2) be (p_1^*, q_1^*). Then*

$$p_1^* = \frac{\bar{c}}{8}\left[\frac{\sqrt{\alpha^6 + 2\alpha^5 - 3\alpha^4 - 8\alpha^2 + 8\alpha}}{\alpha^3 + \alpha^2 - 6\alpha + 4} - 1\right] \quad \textit{and } q_1^* = \frac{\bar{c} + 8p_1^*}{8\alpha^3 v}. \tag{3.3}$$

Proposition 8 shows that as $\bar{c}$ increases, both p_1^* and q_1^* increase. Moreover, as v increases p_1^* does not change while q_1^* decreases. It means that as customers are willing to pay more for the

[11] The formulation presented assumes the third quality level is chosen by at least some customer types.

service, instead of increasing the unit price, the provider would deliberately degrade the quality level and sell it with the same unit price, which increases the revenue in return since the processing time becomes longer. The intuition for this result is that increasing the base price makes some customer types choose lower quality levels. Since higher quality products always generate more revenue to the provider, this shift lowers the impact of revenue increase coming from the price increase. On the other hand, decreasing the base quality level does not make any changes on customer preferences and all customer types pays more for the service completion.

Next, we provide sufficient conditions on the optimal number of quality levels to offer under monopoly.

Proposition 9. *Sufficient conditions for offering multiple quality levels:*

1. *If* $v\bar{q}\left(2\alpha - \dfrac{2\alpha-1}{\alpha}\right) \leq \bar{c}$ *and* $\dfrac{2}{3} < \alpha < 1$, *offering two quality levels generate more revenue than offering only one quality level.*
2. *If* $2\alpha v\bar{q}\left(2\alpha - \dfrac{2\alpha-1}{\alpha}\right) \leq \bar{c}$ *and* $\dfrac{2}{3} < \alpha < 1$, *offering three quality levels generate more revenue than offering two quality levels.*
3. *If* $4\alpha^2 v\bar{q}\left(2\alpha - \dfrac{2\alpha-1}{\alpha}\right) \leq \bar{c}$ *and* $\dfrac{2}{3} < \alpha < 1$, *offering four quality levels generate more revenue than offering three quality levels.*

3.2.2 Optimal Price Menu under Fixed Quality Levels

While controlling both price and quality levels at the same time potentially generates more revenue to the service provider, another interesting question is to find the optimal prices given quality levels. When the service provider offers only one quality level p_1, the optimal price is similar to what we presented in the previous section:

$$p_1^* = \begin{cases} \dfrac{vq_1}{2}, & \text{if } q_1 \leq \dfrac{2\bar{c}}{v} \\ vq_1 - \bar{c}, & \text{if } q_1 > \dfrac{2\bar{c}}{v} \end{cases}. \tag{3.4}$$

When there are two quality levels, $(q_1, 2\alpha q_1)$, the optimal price menu is $(p_1^*, 2p_1^*)$, where

$$p_1^* = \max\left\{\frac{2v\alpha q - \bar{c}}{2}, \frac{vq(2\alpha - 1)}{2\alpha}\right\}, \tag{3.5}$$

only if $p_1^* \leq \frac{\bar{c}(2\alpha - 1)}{2(1-\alpha)}$; otherwise offering one quality level is preferred to offering two.

When there are more than two quality levels, the optimal price depends on multiple conditions and it is beyond the scope of this exercise. Instead, we provide some numerical examples.

Numerical Examples. In this part, we are going to illustrate cases on how many quality levels the monopolistic service provider offers in the optimal solution given its base quality level, scaling factor, and customer characteristics.[12]

1. If service provider a from the previous section with $(\alpha, q_1) = (0.693, 1.355)$ is the only provider in the market with $v = 0.488$ and $\bar{c} = 0.961$, then the optimal price is indeed \$0.100 and offering 4 quality levels is the revenue maximizing strategy. In other words, if service provider a has $(\alpha, q_1) = (0.693, 1.355)$ and offers 4 quality levels with base price level \$0.100, then, we can infer the market conditions as $v = 0.488$ and $\bar{c} = 0.961$ (using Proposition 8).

2. If service provider b from the previous section with $(\alpha, q_1) = (0.616, 1.733)$ is the only provider in the market with $v = 0.488$ and $\bar{c} = 0.961$ (as above), then the optimal price is \$0.423 and only the base quality level product is being chosen by some customers and the rest choose the *no-buy* option. Setting a price of \$0.126 in this market generates less revenue although all four quality levels are chosen by some customer types and there is no customer type that chooses the *no-buy* option.

These examples show that given market conditions and selected product quality, the monopolistic service provider may choose to offer multiple products (as in Example 1 above) or choose to offer only one product with a price level that may be too high for low customer types (as in Example 2). This behavior is intuitive when the service provider has a relatively high base quality

[12]The optimal prices found here are searched on a grid with \$0.001 increments. Therefore, the sensitivity of the optimal prices is \$0.001.

level and a low scaling factor as higher product types do not provide much higher quality than the base quality, which is already high for the market.

3.3 Revenue Maximization under Duopoly

In this section we extend our previous analysis to duopoly case where providers have their own base quality levels and scaling factor set and announced, and they compete with the base price. We still assume that each provider can offer at most 4 quality levels and customers have common valuation v and workload w, and different delay sensitivities $c \sim U(0, \bar{c})$.

We start with a simple model where each provider offers only one quality level. Let (p_1, q_1) and (p_2, q_2) be the price and quality for the first and second providers, respectively. Without loss of generality, assume $q_1 < q_2$. Then, customers with lower type (lower delay sensitivity) choose the first provider, while high types choose the second. Customer type $\hat{c}$ is indifferent between the first and second provider, where

$$\hat{c} = \frac{p_2 q_1 - p_1 q_2}{q_2 - q_1},$$

assuming $\hat{c} \geq 0$.[13] Given p_2, the objective function of the first provider is

$$R_1(p_1) = \frac{1}{\bar{c}} \frac{p_1}{q_1} \hat{c} = \frac{1}{\bar{c}} \frac{p_1}{q_1} \frac{p_2 q_1 - p_1 q_2}{q_2 - q_1}$$

and given p_1, the objective function of the second provider is

$$R_2(p_2) = \frac{1}{\bar{c}} \frac{p_2}{q_2} \left[\max \left\{ \min \left\{ v q_2 - p_2, \bar{c} \right\} - \frac{p_2 q_1 - p_1 q_2}{q_2 - q_1}, 0 \right\} \right].$$

Proposition 10. *Let p_1^e and p_2^e be the equilibrium prices for the first and second provider. Then*

[13] In Nash equilibrium, $\hat{c}$ is indeed nonnegative, which could be derived using Proposition 10.

the Nash equilibrium satisfies

$$p_1^e = \frac{p_2^e q_1}{2q_2} \text{ and } p_2^e = \underset{p_2 \in \{0, p_2^x, p_2^y\}}{\text{argmax}} R(p_2),$$

where $R(p_2)$ *is evaluated for* $p_1 = p_1^e$, *and*

$$p_2^x = \max\left\{vq_2 - \bar{c}, \frac{2vq_2(q_2 - q_1)}{4q_2 - q_1}\right\}, \quad p_2^y = \min\left\{vq_2 - \bar{c}, \frac{2\bar{c}(q_2 - q_1)}{3q_1}\right\}.$$

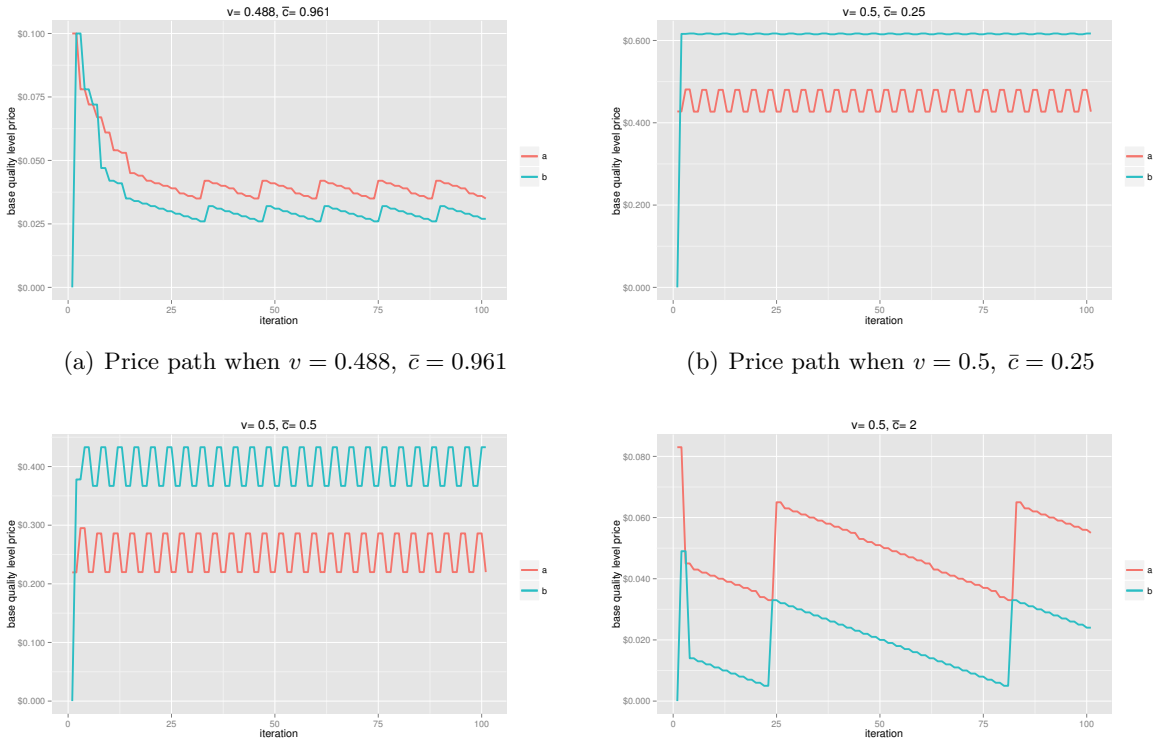
As we point out in the previous section, when we allow the service provider to have more than one quality level, the solution depends on v, $\bar{c}$, and the base quality level in a more complicated way. Therefore, it is not straightforward to find closed-form solutions for duopoly case. Instead we simulate the market with different parameters.[14] In our simulation model, first, service provider a from the previous section sets its monopoly price. Second, given a's price, service provider b finds its best response. Then, service provider a finds its best response given b's price, so on and so forth. We iterate this game for 100 times to see if the game reaches a Nash equilibrium that neither of the players would want to change their prices. We analyze four different cases below. In none of the cases we reach a Nash equilibrium. Each case has a different Edgeworth cycle with different price ranges and periodicity. These case are depicted in Figure 3.2 and described below.

1. $v = 0.488$, $\bar{c} = 0.961$: We have shown that the optimal price for a in this market is \$0.100 when there is monopoly, while it is \$0.423 for b; and we have concluded that if b is the monopoly, there is no point of offering more than one quality level. However, when there is competition, offering more than one quality level becomes preferable to offering only one level for b.

   The price competition makes a decrease its monopoly prices by more than 50%. The price for a varies between \$0.035 and \$0.042 in the cycle, while it is \$0.026 and \$0.032 for b (Figure 3.2(a)).

2. $v = 0.5$, $\bar{c} = 0.25$: a only uses the base quality level under monopoly, where the optimal

[14] As before, we use a price grid with \$0.001 increments.



(a) Price path when $v = 0.488,\ \bar{c} = 0.961$

(b) Price path when $v = 0.5,\ \bar{c} = 0.25$

(c) Price path when $v = 0.5,\ \bar{c} = 0.5$

(d) Price path when $v = 0.5,\ \bar{c} = 2$

Figure 3.2: Price paths under duopoly with different parameter settings

price is \$0.4275, which is found using (3.4). Under duopoly, we have found that only the first quality level is used in both providers in the Edgeworth cycle, and the prices range from \$0.427 to \$0.481 for a, \$0.615 to \$0.617 for b (Figure 3.2(b)).

Since higher quality levels are not selected in either of the providers, we can assume that each provider offers only one quality level and try to find the equilibrium prices that could potentially be aligned with Figure 3.2(b). Under this assumption, the equilibrium prices can be calculated by using Proposition 10 as $p_1^e = \$0.018$ and $p_2^e = \$0.045$.

However, when we relax this assumption and let both providers to offer four quality levels, these prices are no longer equilibrium prices because they are so low that high type customers prefer higher quality levels; and therefore, the equilibrium is no longer sustained.

3. $v = 0.5,\ \bar{c} = 0.5$: two quality levels are used in a under monopoly with the optimal price of

$0.220, which is found using (3.5). Under duopoly, first two quality levels are used in a and only one quality level is used in b in the cycle. The price ranges in the Edgeworth cycle are from $0.22 to $0.286 and from $0.367 to $0.433 for a and b, respectively (Figure 3.2(c)).

4. $v = 0.5$, $\bar{c} = 2$: all four quality levels are used in a under monopoly. Under duopoly, all quality levels in both providers are used as well. In this setting, the price varies more for both providers in the Edgeworth cycle (Figure 3.2(d)).

While the price of a is higher than the price of b in Cases 1 & 4, it is reversed in Cases 2 & 3. It is important to note that the price cycle ranges depend on the initial price level we start the iterative pricing procedure. For instance, if we start Case 2 with a lower price level for provider a, we reach a price cycle with ranges from $0.005 to $0.006 and from $0.003 to $0.004 for providers a and b, respectively. In this solution, both providers generate lower revenue although all four quality levels are selected by some customer types, which in turn, pushes the prices for provider a to be higher than provider b in the price cycle, contrary to the one quality level case.

3.4 Model Extensions

There are many avenues to explore by using our price-quality model as a building block. In this work, we have assumed there is one common workload for all customer types, which implies that the scaling factor, α, only depends on the provider in our model. In reality customers have different workloads and the scaling factor is a combination of the type of workload and the scaling performance of the provider. One potential way to modify the model would be to write the scaling factor as $\alpha = \beta\gamma$, where $\beta \in [0.5, 1]$ is a workload dependent parameter that denotes how parallelizable the workload is, and γ is the scaling factor of the provider. Assuming that our DaCapo workload has $\beta = 0.8$, since it is moderately parallelizable, γ values become 0.866 and 0.770 for providers a and b, respectively. We have simulated scenarios where β is uniformly distributed between 0.5 and 1 and reached similar results with Edgeworth cycles.

Another extension is to solve profit maximization problem instead of revenue maximization. However, this would add an additional layer of complication on the cost side. At the simplest level,

unit cost of a product depends on the configuration that the provider uses, rather than the quality level, and the scale of the provider, since economies of scale plays an important role. With enough information on cost, the model can be modified for profit maximization.

The cloud computing market is a fast growing market and in such market dynamics, sometimes players aim to maximize their market share in the short run before revenue or profit maximization, which could potentially generate higher profits to a player in the long run once the it has its own customer base. In market share maximization case, the duopoly prices are determined based on how much providers can handle profit loss in the short run. In the extreme case, both set prices equal to zero. On the other hand, in zero profit case prices would be set based on costs which may give rise to interesting results as the quality levels and the scaling factors play an important role.

One potential work would be to extend our duopoly game to a two-stage game in which providers first compete in quality and then compete in price.

3.5 Reconciling Model Predictions and Real-World Behavior

As mentioned in the introduction, price cycles in cloud computing are not observed in practice. In the computing, technological advances mean costs are constantly falling. This provides both a market perception that prices should not rise and means that constant prices can be effectively viewed as price increases relative to costs. In Figure 3.3 we show AWS prices for the "general compute" (m series), large size, with the number indicating the generation. Later generations can only run on newer, higher performance hardware. This new hardware can also run the older generations more cost effectively than before. While this is only one product family, the trends are representative.

A few interesting observations can be derived from the figure. First, in the most recent time period, the best VM sells at the lowest price, whereas the worst sells at the highest price. Second, during the price war period of April 2014, the then-newest generation saw a larger price decrease than the older generation. Finally, the oldest generation is still offered and sold in the marketplace. Relative to the falling prices of new generations, this constant price can be conceptualized as a price

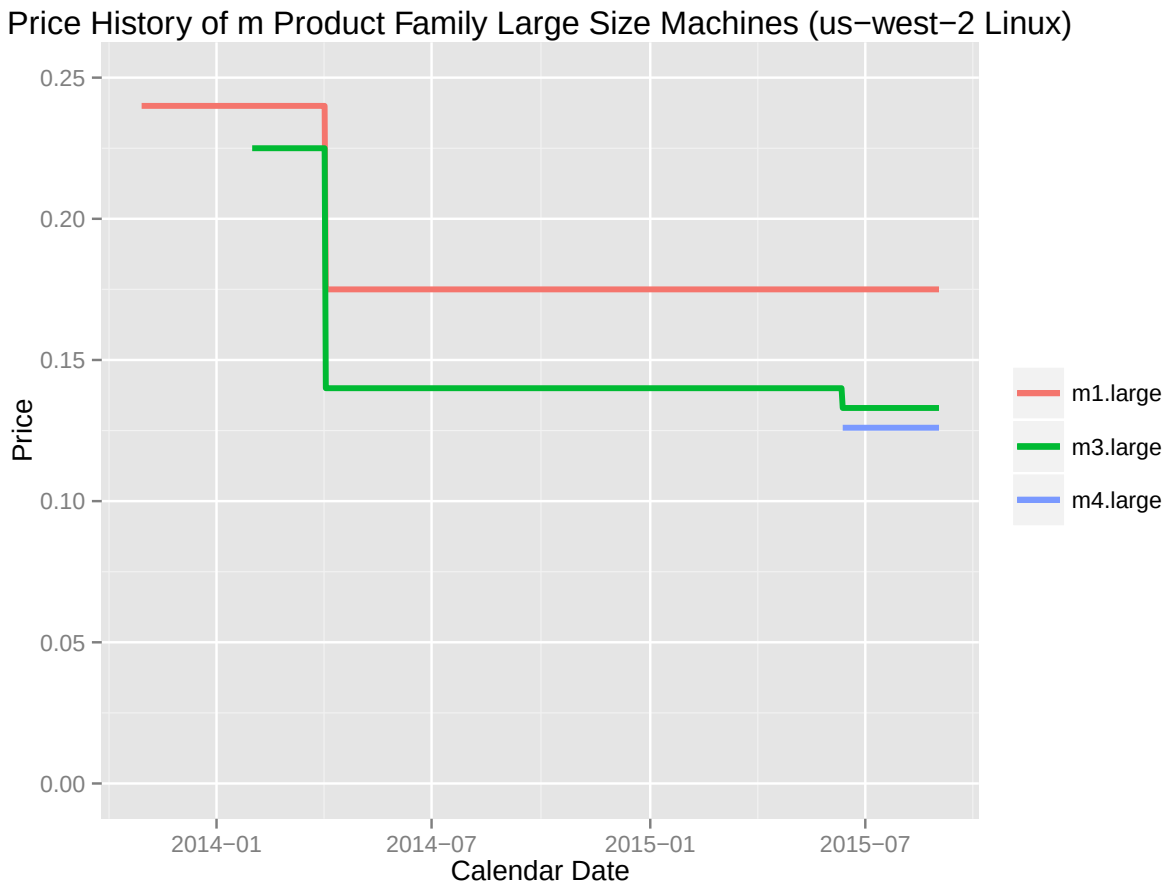


Figure 3.3: m-large machine price history

increase. While these patterns are certainly not equivalent to Edgeworth cycles, they do evidence "price wars" in one segment of the product space (new generations) and relatively high prices in other parts.

Finally we note some caveats to the realism of our model. We calibrated the model using certain benchmark workloads, but in practice customers will have heterogenous needs and we, by no means, captured all of them. Further, providers may innovate to serve a particular niche, such as genomics, with customized offerings. These "menu choices" could be incorporated into our model but at present we do not address this layer of detail. Finally, customers vary in terms of sophistication and how "active" they are in their choice processes. Out of simplicity we ignore these complexities, but concede they could play a significant role in market dynamics and thus are a

fruitful area for future research.

3.6 Discussion and Conclusion

The public cloud "infrastructure as a service" market possesses interesting features that make it hard to predict important facets of competition, such as market shares and provider margins, in the long-run. On the one hand, major providers buy their hardware from the same manufacturers (who in turn generally use the same chipsets and so forth), operate in similar locations and offer seemingly similar products (e.g. VMs specified by number of virtual cores, RAM and disk). On the other hand, the competitors use different proprietary "fabric" to manage virtualization, resource allocation and data transfer. Just as a laptop would tend to run applications differently depending on the operating system, this opens up the space for performance differentiation in the cloud. Further, the menus offered by each provider involve a discrete number of choices and allow providers to locate in different parts of the price-quality space. Our empirical work documents such differentiation.

Our theoretical model gives a long-run view on competition. First, the monopoly case highlights how additional competitors can block "bad equilibrium" where performance is intentionally slowed down or options are unduly limited. In duopoly, price competition is fierce, but prices do not converge to the same low level because of price-quality differentiation. The model also predicts Edgeworth cycles and we have discussed institutional factors that help explain why these are not observed. Once these factors are taken into consideration, the observed patterns can be viewed as being qualitatively similar to the model's predictions: periods of constant prices punctuated by price wars that do not necessarily end with providers having the same prices, and older generations having substantially less vigorous competition than the newest offerings. Further, in Q2 2015 Amazon itemized AWS earnings for the first time and revealed the service has a health operating profit. Our empirically calibrated model helps not only explain price cutting behavior but also how providers can manage a profit despite predictions that the market "should be" totally commoditized.

Chapter 4

Maximizing the Information Content of a Balanced Matched Sample in a Study of the Economic Performance of Green Buildings

In this chapter, we investigate the effect of green building practices on market rents using new matching methods in observational studies. The chapter is organized as follows. In Section 4.1, we give an overview of matching in observational studies. In Section 4.2, we review cardinality matching, discuss different matching structures, and finally present a definition of the information content of a matched sample for a simple difference-in-means effect estimator. In Section 4.3, we first introduce a general framework for matching to maximize the information content of a balanced matched sample, then show that cardinality matching is a particular case of this framework, and present a formulation for matching with a variable one-to-many ratio (in two other appendices, we present formulations for matching to minimize the variance of the difference-in-means effect estimator and matching with a flexible one-to-many/many-to-one or full matching structure). In Section 4.4, we evaluate the building matches in terms of covariate balance and effective sample sizes,

and also describe the details of the computational implementation. In Section 4.5, we investigate the economic effects of green buildings. In Section 4.6, we discuss the new matching methods proposed. In Section 4.7, we close with a summary and remarks.

4.1 Overview of Matching in Observational Studies

In observational studies of causal effects, matching methods are often used in an attempt to compare like with like; i.e., units that are the same ideally in every respect except in their assignment to a treatment (Cochran and Rubin, 1973). In our study, these units are buildings similar in terms of age, amenities, number of stories, etc., except in their green building practices. Of course, this comparison can be assessed in terms of observed covariates only, and with matching methods (the same as with other regression or weighting methods of adjustment for observed covariates) the question about the influence of unobserved covariates in effect estimates remains open (for instance, see Chapter 4 of Rosenbaum, 2002 for a formal discussion). With standard matching methods, other devices such as differential effects, evidence factors, multiple control groups and sensitivity analyses can be used to limit and assess the influence of such unobserved covariates (see Rosenbaum, 2015 for a review of these devices).

The appeal of matching as a method of adjustment lies in part in its conceptual simplicity (comparing like with like while keeping the unit of analysis intact; Rosenbaum and Silber, 2001), that its adjustments are an interpolation instead of an extrapolation based on a parametric model (Rosenbaum, 1987; Imbens, 2015), and in the fact that it is conducted without using outcomes, thus preventing exploratory expeditions in the data to choose the form of adjustments that better suits the hypotheses of the investigation (Rubin, 2008). It is for this last reason that matching is considered to be part of the design as opposed to the analysis of an observational study (Rosenbaum, 2010). However, some matching methods are cumbersome in practice.

The main goal of matching is to find matched groups with similar or balanced observed covariate distributions (Stuart, 2010). Ideally, these groups would be formed by units identical in every way (by "clones" of treated and control units), but usually this is not feasible in practice. There is a

curse of dimensionality in exact matching: as the number of observed covariates increases, there is a combinatorial explosion in the resulting types of units. In fact, with two binary covariates there are 2^2 or four types of units, but with twenty binary covariates there are 2^{20} or over a million types of units. Thus, for an observational study of the typical size (like our building study with a few thousand observations), there will not be enough units to match each treated unit to one control exactly. It is for this reason, and also because randomization does not produce exact matches but balance in expectation, that weaker, aggregate forms of covariate balance than exact matching tend to be pursued in practice, leaving exact matching for a few covariates of overriding prognostic importance (see Sections 3.3 and 9.3 of Rosenbaum, 2010 for a detailed exposition of this argument). The propensity score (Rosenbaum and Rubin, 1983) is an important tool used to achieve aggregate covariate balance.

The propensity score is the probability of treatment assignment given the observed covariates. It constitutes a dimensionality reduction technique in which a P-dimensional observed covariate is summarized into a single scalar with important theoretical properties. Informally, theorems 1 and 3 in Rosenbaum and Rubin (1983) state that matching on the propensity score tends to balance the P observed covariates used to estimate the score, and that for balancing the P covariates it suffices to balance the one-dimensional propensity score. However, these are stochastic properties that hold over repeated realizations of the data-generation mechanism, and for a given realization (this is, for a given data set), even if the true treatment assignment is known, it is not certain that the propensity score will balance the observed covariates (especially if the covariates have many categories or are sparse; see Zubizarreta *et al.*, 2011 and Yang *et al.*, 2012 for related discussions). Also, in practice the true assignment mechanism is unknown, and this makes the task of balancing the observed covariates even more difficult due to misspecification of the propensity score model. Furthermore, while matching on the propensity score is typically used for balancing means, in some settings it is desirable to balance other features of the distribution of the P observed covariates, such as the marginal distributions (Rosenbaum *et al.*, 2007), and this can be very difficult by matching on the propensity score (for a related argument in the context of weighting see, for instance, Zubizarreta, 2015). It is for these reasons that matching on the propensity score involves a considerable amount

of guesswork in practice.

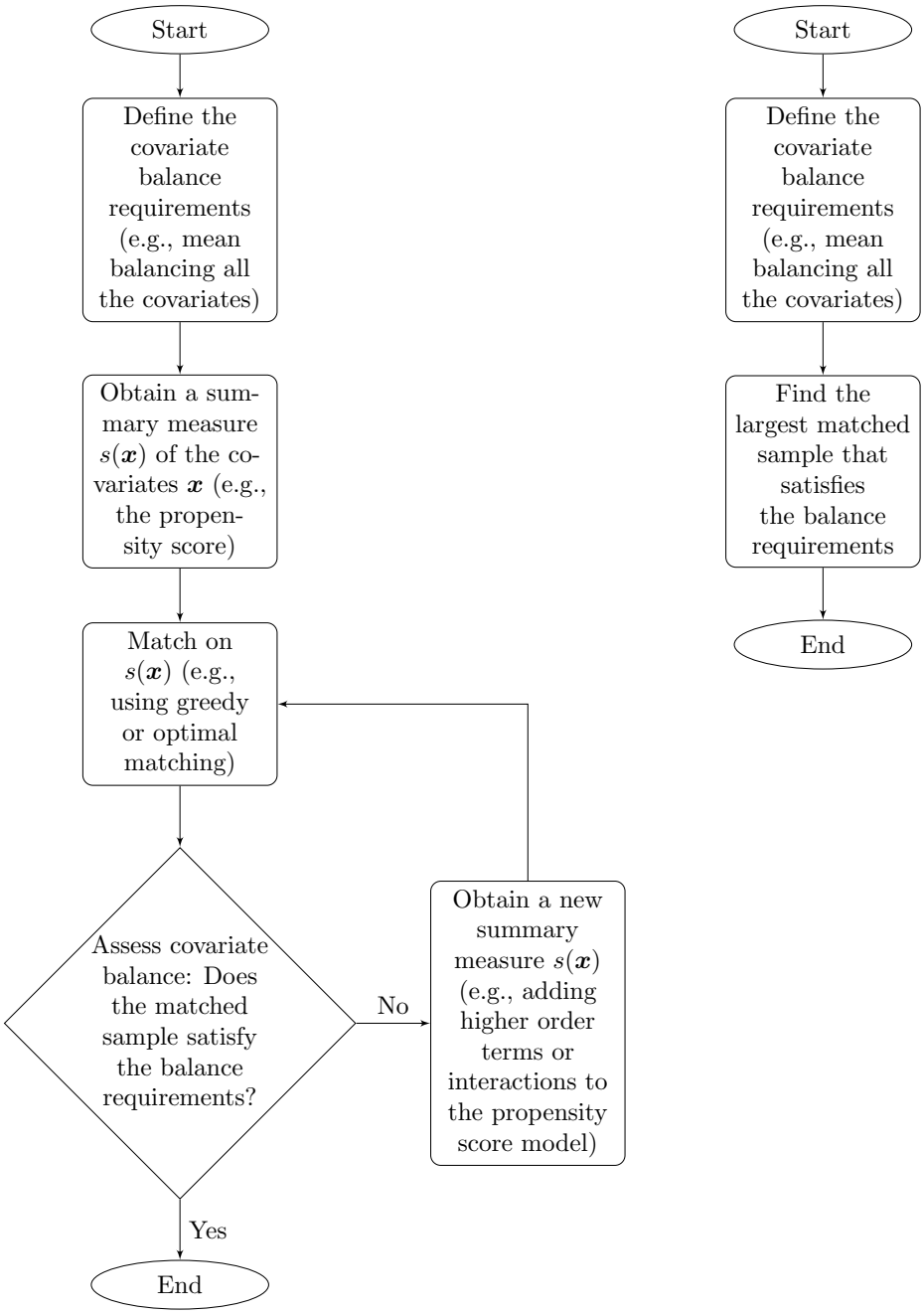
A recent method that addresses these limitations is optimal cardinality matching, or cardinality matching for short (Zubizarreta *et al.*, 2014). Cardinality matching solves an integer programming problem to maximize the cardinality or size of a matched sample subject to constraints on covariate balance. These constraints allow the investigator to balance the covariates directly and in a very precise manner. In their weakest form, these constraints can require the means to be balanced (see Zubizarreta, 2012 for details), but they can also require other forms of distributional balance such as fine balance (Rosenbaum *et al.*, 2007) and strength-k matching (Hsu *et al.*, 2015).[1] In this way cardinality matching directly balances covariates.

Other interesting matching methods that aim at covariate balance include coarsened exact matching (Iacus *et al.*, 2012), balance optimization subset selection (Nikolaev *et al.*, 2013), genetic matching (Diamond and Sekhon, 2013), and refined covariate balance via network flows (Pimentel *et al.*, 2015). Other related weighting methods include inverse probability tilting (Graham *et al.*, 2012), entropy balancing (Hainmueller, 2012), stable balancing weights (Zubizarreta, 2015), calibration weighting (Chan *et al.*, 2016), and the overlap weights (Li *et al.*, 2016).

The flowcharts in Figure 4.1 compare the basic steps involved in cardinality matching and in standard matching methods based on the propensity score or other summary measures of the observed covariates (such as the Mahalanobis distance). While standard matching methods can entail many iterations to meet the covariate balance requirements by fine-tuning the summary measure, cardinality matching directly finds the largest matched sample that meets these requirements. In a sense, with cardinality matching subject matter knowledge of the scientific question at hand comes naturally into the matching problem through the balancing constraints, finding the largest matched data set that satisfies the investigator's specifications for covariate balance or comparability between treated and control units. For simplicity, in Figure 4.1(a) we omit the decisions involved in propensity score matching about overlap, but typically additional steps would be present (for example, see

[1]Fine balance forces the marginal distributions of a nominal variable to be identical, but without constraining units to be matched within each of the categories of a nominal variable (see Chapter 10 of Rosenbaum, 2010 for details); whereas strength-k matching is a stronger form of balance in which low dimensional joints are forced to be identical: out of K nominal covariates, each of the $\binom{K}{k}$ possible interactions of covariates is finely balanced, so the joint distributions of each of the $\binom{K}{k}$ combinations of covariates is perfectly balanced.

Chapter 15 of Imbens and Rubin, 2015 for an extensive discussion). In contrast, with cardinality matching the possibility of covariate distributions exhibiting limited overlap is addressed in terms of the original covariates, finding the largest match that meets the investigator's specifications for covariate balance.



(a) Common matching methods

(b) Cardinality matching

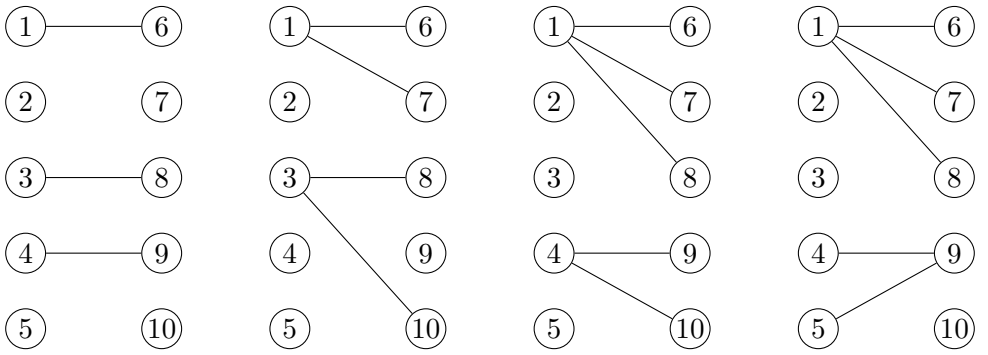
Figure 4.1: Flowcharts of common matching methods and cardinality matching

4.2 Review: Cardinality Matching; Matching Structures; Information Content

4.2.1 Cardinality Matching

As described above, most matching methods target covariate balance indirectly, by matching treated and control units (green and non-green buildings) that are close on a summary measure of the covariates such as the propensity score. Unlike these matching methods, cardinality matching uses the original covariates to match units and directly balance their covariate distributions (Zubizarreta *et al.*, 2014). Specifically, cardinality matching finds the *largest* matched sample that satisfies the investigator's specifications for covariate balance. Following Zubizarreta (2012), these specifications for covariate balance may not only require mean balance, but perhaps also other forms of distributional balance such as fine balance (Rosenbaum *et al.*, 2007), x-fine balance (Zubizarreta *et al.*, 2011), strength-k matching (Hsu *et al.*, 2015), and exact matching (Rosenbaum, 2010, Section 9.3), all this on several covariates simultaneously. For example, cardinality matching will find the largest matched sample in which all the marginal distributions of the covariates are balanced. In this manner, cardinality matching focuses on covariate balance in aggregate, allowing the investigator to re-match the treated and control units in the balanced matched sample to emphasize covariates that are strongly correlated with the outcome. As illustrated in Zubizarreta *et al.* (2014), this has the effect of reducing the heterogeneity of matched-group differences in outcomes and, in turn, also reducing sensitivity to biases due to unmeasured confounders (see Rosenbaum, 2005 for a detailed exposition of this argument and Baiocchi, 2011 for an original alternative approach).

From a computational standpoint, cardinality matching requires solving a linear integer programming problem, and while it has not been found a polynomial time algorithm to solve the cardinality matching problem, there is considerable structure in this problem and many instances of it can be solved in time that from a user perspective is comparable to that of common matching methods (see Appendix C.1). At the present, cardinality matching is solved with the optimization solvers CPLEX, GLPK, Gurobi and Symphony via the statistical package `designmatch` for `R` (Zubizarreta, 2012; Zubizarreta and Kilcioglu, 2016).

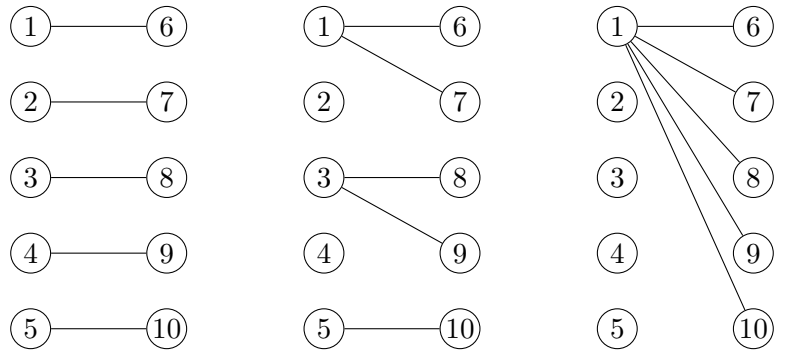


(a) 1 : 1 or pair matching (b) 1 : κ fixed (c) 1 : κ variable (d) 1 : $\kappa_{\mathcal{C}}/\kappa_{\mathcal{T}}$: 1 or full matching

Figure 4.2: Different matching structures

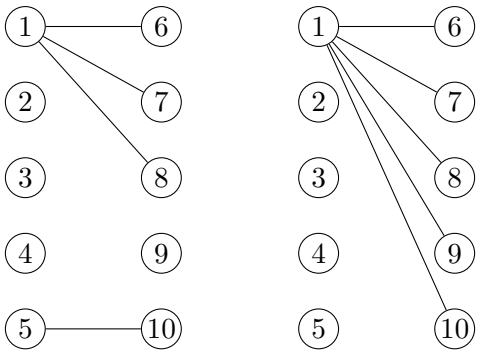
4.2.2 Matching Structures

In its simplest form, a matched sample is assembled by pairs of treated and control units selected from larger reservoirs of both types of units. As in our buildings study, the reservoir of controls is often much larger than the one of the treated units, and it is feasible to match more than one control to each treated unit. One possible way of doing this is by matching with a fixed 1 : κ ratio and either matching each treated unit to κ controls or not matching it at all. A more flexible structure is a variable 1 : κ ratio, in which each treated unit is matched at most to κ controls (if matched at all). The most flexible structure is matching with a one-to-many/many-to-one structure, or, loosely speaking, full matching (Rosenbaum, 1989; Hansen, 2004). (In rigor, the term full match refers not only to a one-to-many/many-to-one structure but to an optimal design for an observational study in which all the treated units are matched to controls forming groups as similar as possible in terms of a summary of the covariates, $s(\boldsymbol{x})$; see Section 10.3.6 of Rosenbaum, 2002. In this sense, a one-to-many/many-to-one matching structure always dominates a many-to-many structure Rosenbaum, 1991. Also, by matching without replacement it is straightforward to conduct inference with existing methods Rosenbaum, 1993, 2001.) We denote the one-to-many/many-to-one structure as 1 : $\kappa_{\mathcal{C}}/\kappa_{\mathcal{T}}$: 1, where $\kappa_{\mathcal{C}}$ is the maximum number of control units matched to each treated unit, and $\kappa_{\mathcal{T}}$ is the maximum number of treated units matched to each control. These different matching structures are illustrated in Figure 4.2 below.



(a) 5 matches with 10 units (b) 5 matches with 8 units (c) 5 matches with 6 units

Figure 4.3: Different matching structures with the same number of matches



(a) 6 units in two matches (b) 6 units in one match

Figure 4.4: Different matching patterns with the same number of matched units

It is desirable to extend cardinality matching to matching with a variable one-to-many or a one-to-many/many-to-one structure, but a question that arises is how to define the size of the matched sample with these flexible matching structures. Naturally, five 1 : 1 matches of green and non-green buildings (exemplified in Figure 4.3(a)) should count more than than two 1 : 2 matches plus one 1 : 1 match (Figure 4.3(b)), and this, in turn, should count more than one 1 : 5 match (Figure 4.3(c)). Although the first and second matchings have the same number of different controls, in the second matching there are only two different treated units; so, subject to the same constraints on covariate balance, the first matching should be preferable. Intuitively, there is more information in the first match. In the following section we formalize this notion using the concept of information content of a matched sample for a difference-in-means effect estimator.

4.2.3 Information Content of a Matched Sample

Let $i \in \mathcal{I} = \{1, 2, ..., I\}$ index the set of matched groups and $j \in \mathcal{J}_i = \{1, 2, ..., J_i\}$ index the set of units (in our study, buildings) within each of these matched groups. Using this notation, for example in Figure 4.2(a), $J_i = 2$ for each $i \in \mathcal{I}$ and the matched groups constitute pairs, and in Figure 4.2(c), $J_1 = 4$ and $J_2 = 3$ and so the groups form quadruples and triples, respectively. To accommodate the more general one-to-many/many-to-one or full matching structure, we adopt the convention that the first unit in each group is either a treated unit and all the other units are controls, or that the first unit is a control and all the other units are treated.

Following Haviland *et al.* (2007), we pose a simple treatment effect model

$$Y_{ij} = \alpha_i + \beta Z_{ij} + \varepsilon_{ij} \tag{4.1}$$

where Y_{ij} is the observed outcome of unit j in matched group i, α_i is a group effect for all the units in group i (this indicates there is dependence between units in each group, but that it may be eliminated by taking differences within groups), Z_{ij} is the treatment assignment indicator, and ε_{ij} is a residual term with $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Here, we assume the outcome variance is constant across units. Consider the matched group difference in outcomes

$$D_i = Z_{i1}\left(Y_{i1} - \frac{\sum_{j \neq 1} Y_{ij}}{\kappa_i}\right) + (1 - Z_{i1})\left(-Y_{i1} + \frac{\sum_{j \neq 1} Y_{ij}}{\kappa_i}\right) \tag{4.2}$$

where κ_i is the number of controls units in matched group i. We can calculate the variance of this difference and find that

$$\text{Var}(D_i) = \sigma^2\left(1 + \frac{1}{\kappa_i}\right) \propto \left(\frac{2}{\frac{1}{1} + \frac{1}{\kappa_i}}\right)^{-1}. \tag{4.3}$$

In other words, the variance of the difference is inversely proportional to the harmonic mean of the number of treated and control units in each matched group (Kalton, 1968; see also Hansen and Bowers, 2008). We denote $h^{(\kappa)}$ as the harmonic mean of the number of units in a matched group

with a $1:\kappa$ (or $\kappa:1$) matching ratio

$$h^{(\kappa)} = \frac{2}{\frac{1}{1}+\frac{1}{\kappa}} = \frac{2\kappa}{1+\kappa}. \tag{4.4}$$

In this manner, in a $1:1$ match or pair match, $h^{(1)} = 1$; in a $1:2$ match, $h^{(2)} = 4/3$; in a $1:3$ match, $h^{(3)} = 3/2$; and so on.

We call the information content of a matched sample the sum of the harmonic means of the number of treated and control units in each matched group, $\sum_{i\in\mathcal{I}} h^{(\kappa_i)}$; that is, the sum of the Fisher information of the matched groups. In this way, for example, the information content of two $1:1$ matches will be 50% larger than the information of one $1:2$ match ($1+1=2$ instead of $4/3$), and the information of three $1:1$ matches will be the same as the information of two $1:3$ matches ($1+1+1=3/2+3/2$).

Another way of defining the information content in a matched sample about the parameter β is the reciprocal of the variance of an effect estimator, for example of the average of the group differences

$$\hat{\delta} = \frac{1}{I}\sum_{i\in\mathcal{I}}\left(Z_{i1}\left(Y_{i1} - \frac{\sum_{j\neq 1} Y_{ij}}{\kappa_i}\right) + (1-Z_{i1})\left(-Y_{i1} + \frac{\sum_{j\neq 1} Y_{ij}}{\kappa_i}\right)\right). \tag{4.5}$$

However, we find that this particular definition is somewhat restrictive, as other estimators may be preferable in practice such as regressing the group differences in outcomes on group differences in covariates as in Rubin (1979), or using the weighted M-statistics in Rosenbaum (2014). Also, this definition is less intuitive and more difficult to implement in practice (see Appendix C.2), and has a weaker connection with cardinality matching. Clearly, if the matching ratio given by κ_i is constant, then maximizing the information content is equivalent to cardinality matching with a fixed $1:\kappa$ ratio as in Zubizarreta *et al.* (2014), so this provides a more general framework and a richer interpretation for cardinality matching.

For these reasons we consider maximizing the sum of the harmonic means of the number of treated and control units in each matched group; in other words, maximizing the sum of the Fisher

information of the matched groups. Building upon this notion of information content, in the next section we present a general matching framework and specific matching formulations that maximize the information content of a matched sample subject to covariate balance and matching structure constraints.

4.3 Maximizing the Information of a Balanced Matched Sample

4.3.1 A General Matching Framework

Let $t \in \mathcal{T} = \{1, \ldots, T\}$ index the set of treated units (in our study, green buildings) and $c \in \mathcal{C} = \{1, \ldots, C\}$ index the set of controls (non-green buildings), with $T \leq C$. Define $p \in \mathcal{P} = \{1, \ldots, P\}$ as the label of the P observed covariates. Each treated unit $t \in \mathcal{T}$ has a vector of observed covariates $\boldsymbol{x}_t = \{x_{t,p_1}, \ldots, x_{t,p_P}\}$, and each control $c \in \mathcal{C}$ has a similar vector $\boldsymbol{x}_c = \{x_{c,p_1}, \ldots, x_{c,p_P}\}$. We introduce the decision variable m_{tc}, which is 1 if treated unit t is matched with control c, and 0 otherwise.

In the abstract, we want to solve

$$\max_{\boldsymbol{m}} \{\mathbb{I}(\boldsymbol{m}) : \boldsymbol{m} \in \mathcal{M} \cap \mathcal{B}\} \tag{4.6}$$

where $\mathbb{I}(\boldsymbol{m})$ is the information content of the matched sample, and $\mathcal{M}$ and $\mathcal{B}$ are matching and balancing constraints, respectively. This general formulation pursues the goal of finding the largest matched sample —or, in general, the matched sample with the largest information content— that satisfies certain requirements for matching structure $\mathcal{M}$ and covariate balance $\mathcal{B}$. Generally, the requirements for covariate balance are guided by scientific knowledge of the research question at hand (in our study, what drives buildings' rent). Ideally one would match with a flexible matching structure, but as we discuss below this imposes computational restraints. We now discuss the specific forms of $\mathbb{I}$, $\mathcal{M}$ and $\mathcal{B}$ when matching with a $1 : \kappa$ fixed ratio, a $1 : \kappa_{\mathcal{C}}$ variable ratio, and, due to space considerations, we relegate the case of matching with a flexible $1 : \kappa_{\mathcal{C}}/\kappa_{\mathcal{T}} : 1$ matching ratio to Appendix C.3.

4.3.2 Matching with a Fixed $1:\kappa$ Ratio

Matching with a fixed $1:\kappa$ ratio is equivalent to cardinality matching. In (4.6), $\mathbb{I}$, $\mathcal{M}$ and $\mathcal{B}$ take the forms

$$\mathbb{I}(\boldsymbol{m}) = \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} m_{tc}, \tag{4.7}$$

$$\mathcal{M} = \left\{ \sum_{c\in\mathcal{C}} m_{tc} = \kappa,\ t\in\mathcal{T} \text{ if } \kappa > 1 \text{ and } \sum_{c\in\mathcal{C}} m_{tc} \leq \kappa,\ t\in\mathcal{T} \text{ if } \kappa = 1; \right. \\ \left. \sum_{t\in\mathcal{T}} m_{tc} \leq 1,\ c\in\mathcal{C}; m_{tc}\in\{0,1\},\ t\in\mathcal{T}, c\in\mathcal{C} \right\}, \tag{4.8}$$

$$\mathcal{B} = \left\{ -\varepsilon_p \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} m_{tc} \leq \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} m_{tc}(f(x_{t,p}) - f(x_{c,p})) \leq \varepsilon_p \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} m_{tc}, \right. \\ \left. m_{tc}\in\{0,1\},\ t\in\mathcal{T}, c\in\mathcal{C}; p\in\mathcal{P} \right\}, \tag{4.9}$$

where $\varepsilon_p \geq 0$ is a given constant, and $f(\cdot)$ is a suitable transformation of the covariates. For example, if $f(x_{\cdot,p}) = x_{\cdot,p}$, then (4.9) constrains the matched samples to have means that differ at most by ε_p for covariate p. Also, if $f(\cdot)$ is a binary indicator for the categories of a nominal covariate p and $\varepsilon_p = 0$, then (4.9) requires the matched samples to have the same number of treated and control units within each category, but without constraining which units are matched together.[2] Similar ideas can be used to balance the interactions of several nominal covariates. See Zubizarreta (2012) and Zubizarreta *et al.* (2014) for more balancing examples.

4.3.3 Matching with a Variable $1:\kappa_{\mathcal{C}}$ ratio

To generalize cardinality matching for maximizing the information content of the matched sample with a variable $1:\kappa_{\mathcal{C}}$ matching ratio, we introduce a new decision variable n_t, the number of control

[2]This technique is called fine balance (Rosenbaum *et al.*, 2007) and it has the effect of exactly balancing the mean of every linear combination of the categories of the covariates finely balanced.

units that treated unit t is matched to, which is bounded above by $\kappa_{\mathcal{C}}$. Then problem (4.6) becomes

$$\mathbb{I}(\boldsymbol{m}, \boldsymbol{n}) = \sum_{t \in \mathcal{T}} h^{(n_t)}, \tag{4.10}$$

$$\begin{aligned} \mathcal{M} = \Bigg\{ & \sum_{c \in \mathcal{C}} m_{tc} = n_t,\ t \in \mathcal{T}; n_t \leq \kappa_{\mathcal{C}},\ t \in \mathcal{T}; \sum_{t \in \mathcal{T}} m_{tc} \leq 1,\ c \in \mathcal{C}; \\ & m_{tc} \in \{0,1\},\ t \in \mathcal{T}, c \in \mathcal{C}; n_t \geq 0,\ t \in \mathcal{T} \Bigg\}, \end{aligned} \tag{4.11}$$

$$\begin{aligned} \mathcal{B} = \Bigg\{ & -\varepsilon_p \sum_{t \in \mathcal{T}} h^{(n_t)} \leq \sum_{t \in \mathcal{T}} h^{(n_t)} x_{t,p} - \sum_{c \in \mathcal{C}} \left(\sum_{t \in \mathcal{T}} m_{tc} \frac{h^{(n_t)}}{n_t} \right) x_{c,p} \leq \varepsilon_p \sum_{t \in \mathcal{T}} h^{(n_t)}, \\ & p \in \mathcal{P}, m_{tc} \in \{0,1\},\ t \in \mathcal{T}, c \in \mathcal{C}; n_t \geq 0,\ t \in \mathcal{T} \Bigg\}. \end{aligned} \tag{4.12}$$

Here, we let $f(x) = x$ for mean balance. Note that by using transformations of the covariates, it is possible to balance statistics other than means (e.g., by mean balancing indicators for the quantiles of x in the treated units it is possible to approximately balance its marginal distribution; see Zubizarreta, 2012 for details). Also, note that $h^{(\kappa)}$ is an increasing, convex transformation of κ; that is, $h^{(\kappa)}$ increases as κ increases at a decreasing rate. However, this optimization problem has the expressions $h^{(n_t)}$ and $m_{tc}\frac{h^{(n_t)}}{n_t}$ which are not linear in m_{tc} and n_t. To linearize $h^{(n_t)}$, we define a new decision variable $m_t^{(r)}$, which is 1 if treated unit t is matched with at least r controls, and 0 otherwise ($t \in \mathcal{T}, r \in \{1, \ldots, \kappa_{\mathcal{C}-1}\}$). This new decision variable can be written using linear constraints as

$$m_t^{(r)} \leq n_t - \sum_{s=1}^{r-1} m_t^{(s)},\ t \in \mathcal{T}, r \in \{1, \ldots, \kappa_{\mathcal{C}-1}\} \tag{4.13}$$

$$\kappa_{\mathcal{C}} m_t^{(r)} \geq n_t - \sum_{s=1}^{r-1} m_t^{(s)},\ t \in \mathcal{T}, r \in \{1, \ldots, \kappa_{\mathcal{C}-1}\}. \tag{4.14}$$

Here we do not need to define the decision variable $m_t^{(\kappa_{\mathcal{C}})}$ since $m_t^{(\kappa_{\mathcal{C}})} = n_t - \sum_{s=1}^{\kappa_{\mathcal{C}}-1} m_t^{(s)}$;

therefore, it is not a decision variable. Using the $m_t^{(r)}$'s, we can rewrite $h^{(n_t)}$ as

$$\begin{aligned} w_t^{(1)} &:= h^{(n_t)} \\ &= \sum_{s=1}^{\kappa_{\mathcal{C}}-1} \left(h^{(s)} - h^{(s-1)}\right) m_t^{(s)} + \left(h^{(\kappa_{\mathcal{C}})} - h^{(\kappa_{\mathcal{C}}-1)}\right) \left(n_t - \sum_{s=1}^{\kappa_{\mathcal{C}}-1} m_t^{(s)}\right). \end{aligned} \tag{4.15}$$

Hence, we can write the objective function in the linear form: $\sum_{t\in\mathcal{T}} w_t^{(1)}$.

The next step is to write $m_{tc}\frac{h^{(n_t)}}{n_t}$ in linear form. Define

$$\begin{aligned} w_t^{(2)} &:= \frac{h^{(n_t)}}{n_t} \\ &= \sum_{s=1}^{\kappa_{\mathcal{C}}-1} \left(\frac{h^{(s)}}{s} - \frac{h^{(s-1)}}{s-1}\right) m_t^{(s)} + \left(\frac{h^{(\kappa_{\mathcal{C}})}}{\kappa_{\mathcal{C}}} - \frac{h^{(\kappa_{\mathcal{C}}-1)}}{\kappa_{\mathcal{C}}-1}\right) \left(n_t - \sum_{s=1}^{\kappa_{\mathcal{C}}-1} m_t^{(s)}\right), \end{aligned} \tag{4.16}$$

where $\frac{h^{(0)}}{0}$ is set to 0. The expression of interest becomes $m_{tc}w_t^{(2)}$ which is still not linear. Therefore, we define the decision variable $q_{tc} = m_{tc}w_t^{(2)}$, which is equal to $w_t^{(2)}$ if $m_{tc} = 1$, 0 otherwise. It can be written using linear constraints as

$$q_{tc} \leq m_{tc},\ t \in \mathcal{T}, c \in \mathcal{C} \tag{4.17}$$

$$q_{tc} \leq w_t^{(2)},\ t \in \mathcal{T}, c \in \mathcal{C} \tag{4.18}$$

$$q_{tc} \geq w_t^{(2)} - (1 - m_{tc}),\ t \in \mathcal{T}, c \in \mathcal{C}. \tag{4.19}$$

Lastly, we define $w_c = \sum_{t\in\mathcal{T}} q_{tc},\ c \in \mathcal{C}$, and rewrite mean balancing constraints

$$-\varepsilon_p \sum_{t\in\mathcal{T}} w_t^{(1)} \leq \sum_{t\in\mathcal{T}} w_t^{(1)} x_{t,p} - \sum_{c\in\mathcal{C}} w_c x_{c,p} \leq \varepsilon_p \sum_{t\in\mathcal{T}} w_t^{(1)}, p \in \mathcal{P} \tag{4.20}$$

This program is no longer a pure integer programming (IP) problem, as cardinality matching; it is a mixed integer programming (MIP) problem with considerably less structure than the MIP problem solved by Zubizarreta (2012). In fact, the constraints (4.13)-(4.20) make the program quite complicated to solve in general.

4.3.4 Matching with a Flexible $1 : \kappa_{\mathcal{C}}/\kappa_{\mathcal{T}} : 1$ Ratio

One step further is to formulate (4.6) to match with a flexible $1 : \kappa_{\mathcal{C}}/\kappa_{\mathcal{T}} : 1$ matching ratio or full matching. Due to space constraints, this is discussed in Appendix C.3.

4.4 Description of the Matches

In our study, we find the matched sample of green and non-green buildings with largest information content (4.10) that satisfies the matching structure (4.11) and that balances the original covariates in the sense of (4.12). In particular, we match with a variable $1 : \kappa_{\mathcal{C}}$ matching ratio because each geographic cluster has only one green building and a variable number of non-green buildings. We choose $\kappa_{\mathcal{C}} = 4$ because the gains from matching with a higher $1 : 5$ or a $1 : 6$ ratio are not very marked assuming the same number of treated units are matched (see Table 2 of Haviland *et al.*, 2007) and because increasing the maximum matching ratio by one adds $2T$ constraints and T binary variables to the mathematical program making it more difficult to solve (see Section 4.4.4 below).

4.4.1 Covariate Balance

Table 4.1 shows the absolute standardized differences in means of the observed covariates before and after matching with a variable $1 : 4$ ratio. In the table, before matching there are a number of substantial differences, most notably in the building classes, age (>40 years) and amenities, whereas after matching all these differences are smaller than 0.1. Within the framework of (4.6), we designed the matched sample to be balanced in this way.

4.4.2 Information of the Matched Samples

Table 4.2 below shows the information content or, loosely speaking, the effective samples sizes of the samples matched with fixed $1 : 1$, $1 : 2$, $1 : 3$ and $1 : 4$ ratios, and with a variable $1 : 4$ ratio. With a $1 : 1$ ratio or pair matching, the resulting information content is 666, meaning that 666 buildings were paired. With fixed $1 : 2$, $1 : 3$ and $1 : 4$ ratios, the information content is equivalent to 757, 708, and 642 pairs, whereas with a variable $1 : 4$ ratio it is 941. In other words, matching

Table 4.1: Standardized differences in means before and after matching

Covariate	Standardized difference in means	
	Before matching	After matching
Building size	0.362	0.076
Building class A	1.005	0.096
Building class B	-0.650	0.053
Building class C	-0.557	-0.068
Net contract	0.127	0.020
Employment growth	0.043	0.000
Employment growth missing	-0.010	0.000
Age $\leq$10 years	0.323	0.049
Age 11-20 years	0.400	0.034
Age 21-30 years	0.392	0.018
Age 31-40 years	-0.066	-0.044
Age >40 years	-0.974	-0.050
Age missing	-0.150	-0.007
Renovated	-0.389	0.033
Stories low	-0.145	-0.066
Stories intermediate	0.032	0.046
Stories high	0.141	0.031
Stories missing	-0.061	-0.014
Amenities	0.474	0.079

with a variable $1:4$ ratio produces an effective sample size 47% larger than matching with a fixed $1:4$ ratio. This shows the gains from matching with a variable ratio.

Table 4.2: Effective sample sizes as measured by $\mathbb{I}$ in (4.10)

Matching structure	Information or effective sample size
$1:1$ fixed	666
$1:2$ fixed	757.3
$1:3$ fixed	708
$1:4$ fixed	641.6
$1:4$ variable	940.6

4.4.3 Comparison to Optimal Matching

Following the suggestion of a reviewer, we compare our method to optimal matching as implemented in `optmatch` (Hansen, 2007). In optimal matching, we calculate the Mahalanobis distance with propensity score calipers as suggested in Rosenbaum and Rubin (1985). For a strict comparison,

in both methods we use a variable 1 : 4 matching ratio. As a result, with optimal matching the effective sample size is somewhat smaller than with our method (730 versus 940.6) and there are substantial imbalances in several covariates (more than half of the covariates exhibit differences in means larger than 0.1 standard deviations). Arguably, covariate balance could be improved by recalculating the covariate distances, but this would involve iteration in order to achieve covariate balance (as described in Figure 1(a) above). With the proposed method, the differences in means are constrained to be at most 0.1 standard deviations by design. However, `optmatch` is optimal in another important sense — it minimizes the total sum of covariate distances between matched units — and it runs in polynomial time, so relatively large data sets can be handled quickly (Hansen and Klopfer, 2006). As we discuss in the following section, computation is an important aspect to consider in the implementation of our method.

4.4.4 Computation and Details of the Implementation

Matching with a variable $1 : \kappa_{\mathcal{C}}$ ratio, (4.10)-(4.20), as in our study, and also matching with a flexible $1 : \kappa_{\mathcal{C}}/\kappa_{\mathcal{T}} : 1$ ratio, (C.16)-(C.42), as in Appendix C.3, have more complicated structure than cardinality matching, mainly due to the harmonic means used in the objective function and mean balancing constraints. Specifically, while cardinality matching with a $1 : 1$ ratio and mean balancing has $T \times C$ binary decision variables and $T + C + 2 \times P$ constraints, matching with a variable $1 : \kappa_{\mathcal{C}}$ ratio with harmonic means has additional $T \times (\kappa_{\mathcal{C}} + C)$ continuous decision variables and $T \times (2 \times \kappa_{\mathcal{C}} + 3 \times C - 1)$ constraints, after some simplifications.

Although these two matching problems are considerably larger than cardinality matching, by using optimization solvers such as CPLEX and Gurobi it is still possible to reach solutions with a small optimality gap in a reasonable amount of time depending on the problem size (see Appendix C.4 for a simulation study using the buildings data). Nemhauser (2013) reports that algorithmic speed in solvers such as CPLEX and Gurobi has increased 256000 times between 1991 and 2013. This, combined with a modest computer speedup of 1000 times, translates into the ability to solve problems that took nearly seven years in the early 1990's to one second today (Nemhauser, 2013). These major improvements have been made possible by a combination of advancements in

preprocessing and heuristics for finding good feasible solutions quickly, branch-and-bound methods to reduce the feasible set, linear programming implementations as the basic tool for solving IP and MIP problems, and parallel computing (Bixby and Rothberg, 2007, Linderoth and Lodi, 2010, Nemhauser, 2013; see also Bertsimas, 2014 for a related discussion and applications of MIP to statistical and machine learning).

In addition to these optimization techniques, we used exact matching constraints on the location covariate (see Appendix C.5), and divided the problem into 10 subproblems to solve each of them in parallel. Using the `R` packages `doParallel` and `foreach` (Weston and Calaway, 2014), we solved the 10 subproblems independently and simultaneously using 10 processors with 15-minute time limit. Among these subproblems, one gives the optimal solution within the time limit, and the others give solutions with about 2% optimality gap at the end of the specified time. This computational implementation method enables us to solve this problem under 20 minutes. It would take more than 2 hours to reach the same solution if no parallel computing methods were used. At the present time, the code that we used for the analyses is available upon request, but soon it will be available within the package `designmatch` for `R`.

4.5 Economic Performance of Green Buildings

From our balanced matched sample, we find that green buildings have 3.3% higher rental rates per square foot than otherwise similar non-green buildings. The 95% confidence interval associated to this estimate is $[1.3\%, 5.5\%]$ (obtained using the inferential procedures in Hansen *et al.*, 2014). For comparison, this estimate is moderately larger than the one of Eichholtz *et al.* (2010), who reported that green buildings have rental rates 2.8% higher per square foot than similar non-green buildings (with 95% confidence interval of $[1\%, 4.6\%]$).

In principle, our estimand is not the same as the one of Eichholtz *et al.* (2010), since our approach restricts the analysis to the sample with largest information that is balanced, usually discarding some treated units (in our study, these are 19 out of the 694 green buildings available before matching). To get a better understanding of our matched sample, in Table C.2 of Appendix

C.6 we provide a description of the samples of green buildings before matching, after matching, and of those green buildings that were unmatched and left out from the analyses. Overall, this sample closely resembles that of all the available green buildings before matching, so in principle these results can be generalized to a population of buildings of similar characteristics.

Next, when conducting a sensitivity analysis to hidden biases, we find that for an unobserved covariate to explain away the estimated effect of 3.3% it would need to simultaneously increase the odds of a building having green ratings and of a positive difference in rent both by a factor of 1.9, so the results are only moderately insensitive to hidden biases (see Rosenbaum and Silber, 2009 and Hansen *et al.*, 2014 for details of this analysis).

To interpret these results, let us remember that about 30% of building operating costs are driven by energy consumption and that green buildings typically have 25% less energy use and in aggregate 19% lower operating costs. Therefore, in rough terms, savings from operating costs overcome the extra amount paid for a green building rent if the rent to operating costs ratio is 5.75 (= 0.19/0.033) or more. Thus, it is an economically sound decision for some companies to prefer green buildings and pay more rent. Moreover, as Eichholtz *et al.* (2010) mention, even a small improvement on the energy use of existing buildings has a big impact not only on the economy but also on the environment. In this way, companies are also willing to pay more to "go green" for a sustainable environment.

4.6 Discussion of the Proposed Matching Methods

The main objective of matching in observational studies is to balance observed covariates and thereby remove biases due to systematic differences in their distributions (Cochran, 1965, Section 2.2). As discussed in Section 8.7 of Rosenbaum (2010), efficiency is a secondary concern in observational studies. The explanation for this is that if there is a bias that does not decrease as the sample size increases, then it tends to dominate the mean squared error in large samples, resulting in a very precise estimate of the wrong quantity (Haviland *et al.*, 2007). For these reasons, in view of the bias-variance —or, stated differently, the balance-precision— tradeoff involved in matching,

we give priority to balance over precision, and, subject to removing systematic biases by balancing covariates, we maximize precision, or more specifically, the information content of the matched sample.

The framework we proposed in Section 4.3.1 encompasses these objectives in a general way. Within this framework, cardinality matching is a special case when matching with a fixed $1 : \kappa$ ratio. Also, the formulations presented in Section 4.3.3, and in Appendices C.2 and C.3, are different methods for maximizing the information content of a balanced matched sample. Ideally, if the outcome model follows (4.1) and if the outcome analyses use the effect estimator (4.5), then one would solve the matching problem in Appendix C.2, but as discussed this is a very complicated optimization problem because the number of matched pairs I is also a decision variable. Interestingly, if the solution to the cardinality matching problem uses all the available treated units, then this solution also minimizes the variance of the effect estimator (4.5). With other estimators or non-constant variances across units, the formulations in Section 4.3.3 and Appendix C.3 for matching with a $1 : \kappa_{\mathcal{C}}$ variable ratio and the more flexible $1 : \kappa_{\mathcal{C}}/\kappa_{\mathcal{T}} : 1$ matching ratio, respectively, may be more appropriate.[3] As discussed in Section 4.2.3, these formulations are not only easier to implement but also more intuitive as they maximize the sum of the Fisher informations of the matched groups.

Building on cardinality matching, the proposed methods do not require estimation of the propensity score as they directly balance the original covariates. Nonetheless, the propensity score may be used as an additional covariate in the balancing constraints $\mathcal{B}$. In this work we mainly discussed mean balancing constraints, but other constraints can be implemented for distributional balance such as fine balance (Rosenbaum *et al.*, 2007) and strength-k matching (Hsu *et al.*, 2015); for a related discussion, see Zubizarreta (2012).

[3] In model (4.1) we assumed that the variance is constant across units. One way to relax this assumption is to suppose instead that the variance in the treated group is f times bigger than the variance in the control group. Then $h^{(\kappa)}$ becomes the sum of the harmonic means of 1 (treated unit) and κ_i/f ("control" units) for each matched group (as opposed to 1 and κ_i, as before). As another example, suppose that the variance in one category of a binary covariate is f times bigger than in the other category. Then the weighting becomes $h^{(\kappa_i)}/f$ for the matched group with greater variance and emphasizing to match f times as many groups from the strata with smaller variance. Extending this example, there may be important strata and one could estimate the variance in those strata and plug in the estimates, but this would require using the outcomes for matching. In general, if the variances vary arbitrarily, then the weights become intractable.

Assessing overlap or lack of common support in covariate distributions is a widespread practice undertaken in observational studies in order to avoid extrapolating or fabricating results from regression models that assume a particular functional form (Rosenbaum, 2010, Section 18.2; Imbens and Rubin, 2015, Chapter 14). This is typically done in two steps: first, by trimming the sample on the propensity score, and second, by checking balance. For instance, Imbens (2015) suggests dropping units with extreme values of the estimated propensity score (Crump *et al.*, 2009) and then checking balance in normalized differences in average covariates. As in cardinality matching, the methods proposed in this work directly "trim" the sample to satisfy the requirements for covariate balance of the original covariates. To the extent that these requirements balance the covariates adequately, these methods will avoid extrapolation by restricting the analysis to the matched treated and control samples that overlap the most (again, in the sense of information and the balance requirements).

Of course, restricting the analysis to the samples of treated and control units that overlap will typically change the estimand. In the case that treated units are matched to a subset of the controls, the estimand will cease to be the average treatment effect on the treated and it will become a more local estimand, the average treatment effect on the matched treated units. In view of this limitation of the data, one way to proceed without further modeling assumptions is by describing both the matched and unmatched samples as in Appendix C.6. This provides a basic understanding of the population to which, in principle, the results of the matched analysis can be generalized (Hill, 2008; see also Traskin and Small, 2011 and Fogarty *et al.*, 2015). Another way to proceed is by weighting the matched samples to a target population of greater policy interest perhaps by using the method in Zubizarreta (2015).

In cardinality matching, finding the largest balanced matched sample is followed by re-matching the pairs or groups that constitute the matched sample to minimize their total sum of covariate distances. If these covariates are predictive of the outcome, this re-matching will reduce heterogeneity within matched groups and therefore sensitivity to biases due to unobserved covariates (Rosenbaum, 2005). A possible direction for future research would be to extend the proposed methods along these lines. Also, the proposed methods can be used for adjustment in observational

studies with a time-dependent treatment and time-dependent covariates via risk set matching (Li *et al.*, 2001, Lu, 2005). Under weaker identification assumptions than those of "no unmeasured confounders," the proposed methods can also be used for treatment effect estimation with an instrumental variable (Baiocchi *et al.*, 2010, Zubizarreta *et al.*, 2013) or a discontinuity design (Keele *et al.*, 2015).

4.7 Summary

In this chapter, we revisited the study of Eichholtz *et al.* (2010) about the market performance of green buildings. To analyze the effect of energy efficiency and sustainability on the economic returns of buildings, we used new matching methods that take more advantage of the clustered structure of the buildings data than standard matching methods. We proposed a general framework for matching in observational studies and specific matching methods within this framework that simultaneously achieve three goals: (i) maximize the information content of a matched sample (and, in some cases, also minimize the variance of a widely used effect estimator); (ii) form the matches using a flexible matching structure (such as a one-to-many/many-to-one structure); and (iii) directly attain covariate balance as specified —before matching— by the investigator. To our knowledge, existing matching methods are only able to achieve, at most, two of these goals simultaneously. Using these methods, we obtained a larger effective sample size and found that green buildings have 3.3% higher rental rates per square foot than otherwise similar buildings without green ratings (a moderately larger effect than the one previously found by Eichholtz *et al.*, 2010). Thus, besides being environmentally responsible it is also an economically sound decision to pursuit environmentally sustainable building practices.

Bibliography

V. Abhishek, I. A. Kash, and P. Key. Fixed and market pricing for cloud services. *arXiv preprint arXiv:1201.5621*, 2012.

P. Afèche and H. Mendelson. Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Management Science*, 50(7):869–882, 2004.

P. Afèche and M. Pavlin. Optimal price/lead-time menus for queues with customer choice: segmentation, pooling, strategic delay. *Management Science (forthcoming)*, 2015.

P. Afèche. Incentive-compatible revenue management in queueing systems: optimal strategic delay. *Manufacturing & Service Operations Management*, 15(3):423–443, 2013.

O. Agmon Ben-Yehuda, M. Ben-Yehuda, A. Schuster, and D. Tsafrir. Deconstructing amazon ec2 spot instance pricing. In *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on*, pages 304–311. IEEE, 2011.

G. M. Amdahl. Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April 18-20, 1967, spring joint computer conference*, pages 483–485. ACM, 1967.

E. T. Anderson and J. D. Dana, Jr. When is price discrimination profitable? *Management Science*, 55(6):980–989, 2009.

J. Anselmi, D. Ardagna, J. Lui, A. Wierman, Y. Xu, and Z. Yang. The economics of the cloud: price competition and congestion. *ACM SIGecom Exchanges*, 13(1):58–63, 2014.

M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, et al. A view of cloud computing. *Communications of the ACM*, 53(4):50–58, 2010.

M. Baiocchi, D. S. Small, S. Lorch, and P. R. Rosenbaum. Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association*, 105(492):1285–1296, 2010.

M. Baiocchi. Designing robust studies using propensity score and prognostic score matching. *Chapter 3 in Methodologies for Observational Studies of Health Care Policy, Dissertation, Department of Statistics, The Wharton School, University of Pennsylvania*, 2011.

O. Baron. *Pricing and admission control for shared computer services using the token bucket mechanism.* PhD thesis, Massachusetts Institute of Technology, 2003.

D. P. Bertsekas. A new algorithm for the assignment problem. *Mathematical Programming*, 21:152–171, 1981.

D. Bertsimas. Statistics and machine learning via a modern optimization lens. *The 2014-2015 Philip McCord Morse Lecture*, 2014.

R. E. Bixby and E. Rothberg. Progress in computational mixed integer programming—a look back from the other side of the tipping point. *Annals of Operations Research*, 149:37–41, 2007.

S. M. Blackburn, R. Garner, C. Hoffman, A. M. Khan, K. S. McKinley, R. Bentzur, A. Diwan, D. Feinberg, D. Frampton, S. Z. Guyer, M. Hirzel, A. Hosking, M. Jump, H. Lee, J. E. B. Moss, A. Phansalkar, D. Stefanović, T. VanDrunen, D. von Dincklage, and B. Wiedermann. The DaCapo benchmarks: Java benchmarking development and analysis. In *OOPSLA '06: Proceedings of the 21st annual ACM SIGPLAN conference on Object-Oriented Programing, Systems, Languages, and Applications*, pages 169–190, New York, NY, USA, October 2006. ACM Press.

C. Borgs, O. Candogan, J. Chayes, I. Lobel, and H. Nazerzadeh. Optimal multiperiod pricing with service guarantees. *Management Science*, 60(7):1792–1811, 2014.

K. C. G. Chan, S. C. P. Yam, and Z. Zhang. Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society*, page forthcoming, 2016.

A. A. Chien and V. Karamcheti. Moore's law: The first ending and a new beginning. *Computer*, (12):48–53, 2013.

W. G. Cochran and D. B. Rubin. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446, 1973.

W. G. Cochran. The planning of observational studies of human populations. *Journal of the Royal Statistical Society*, 128:234–266, 1965.

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

R. J. Deneckere and R. P. McAfee. Damaged goods. *Journal of Economics & Management Strategy*, 5(2):149–174, 1996.

A. Diamond and J. S. Sekhon. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *The Review of Economics and Statistics*, 95(3):932–945, 2013.

P. Eichholtz, N. Kok, and J. M. Quigley. Doing well by doing good? green office buildings. *American Economic Review*, pages 2492–2509, 2010.

Y. Feng, B. Li, and B. Li. Price competition in an oligopoly market with multiple iaas cloud providers. *Computers, IEEE Transactions on*, 63(1):59–73, 2014.

C. Fogarty, M. Mikkelsen, D. Gaieski, and D. Small. Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *Journal of the American Statistical Association*, page forthcoming, 2015.

B. S. Graham, C. Campos de Xavier Pinto, and D. Egel. Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies*, 79:1053–1079, 2012.

J. L. Gustafson. Reevaluating amdahl's law. *Communications of the ACM*, 31(5):532–533, 1988.

J. Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

B. B. Hansen and J. Bowers. Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 23(2):219–236, 2008.

B. B. Hansen and S. O. Klopfer. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627, 2006.

B. B. Hansen, P. R. Rosenbaum, and D. S. Small. Clustered treatment assignments and sensitivity to unmeasured biases in observational studies. *Journal of the American Statistical Association*, 109(505):133–144, 2014.

B. B. Hansen. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, 99(467):609–618, September 2004.

B. B. Hansen. Flexible, optimal matching for observational studies. *R News*, 7:18–24, 2007.

A. Haviland, D. S. Nagin, and P. R. Rosenbaum. Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods*, 12(3):247, 2007.

J. L. Hill. Discussion of research using propensity-score matching: Comments on ï£¡a critical appraisal of propensity-score matching in the medical literature between 1996 and 2003ï£¡ by peter austin, statistics in medicine. *Statistics in Medicine*, 27(12):2055–2061, 2008.

J. Y. Hsu, J. R. Zubizarreta, D. S. Small, and P. R. Rosenbaum. Strong control of the family-wise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika*, page forthcoming, 2015.

S. M. Iacus, G. K. King, and G. Porro. Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24, 2012.

G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

G. W. Imbens. Matching methods in practice: Three examples. *Journal of Human Resources*, 50(2):373–419, 2015.

G. Kalton. Standardization: A Technique to Control for Extraneous Variables. *Applied Statistics*, 17(2):118–136, 1968.

A. Katta and J. Sethuraman. Pricing strategies and service differentiation in queues – a profit maximization perspective. Technical report, Computational Optimization Research Center, Columbia University. TR-2005-04, 2005.

L. Keele, R. Titiunik, and J. R. Zubizarreta. Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *Journal of the Royal Statistical Society: Series A*, 178:223–239, 2015.

H. W. Kuhn. The Hungarian Method for the Assignment Problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

L. Leong, D. Toombs, Gill. B., G. Petri, and Haynes T. Magic quadrant for cloud infrastructure as a service, worldwide report. Technical report, Gartner, 2015.

Y. P. Li, J. K. Propert, and P. R. Rosenbaum. Balanced risk set matching. *Journal of the American Statistical Association*, 96(455):pp. 870–882, 2001.

A. Li, X. Yang, S. Kandula, and M. Zhang. Cloudcmp: shopping for a cloud made easy. In *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, pages 5–5. USENIX Association, 2010.

F. Li, K. L. Morgan, and A. M. Zaslavsky. Balancing covariates via propensity score weighting. *Working Paper*, 2016.

J. T. Linderoth and A. Lodi. Milp software. In James J. Cochran, Louis A. Cox, Pinar Keskinocak, Jeffrey P. Kharoufeh, and J. Cole Smith, editors, *Wiley Encyclopedia of Operations Research and Management Science*. Wiley, 2010.

B. Lu. Propensity score matching with time-dependent covariates. *Biometrics*, 61(3):721–728, 2005.

C. Maglaras and A. Zeevi. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Science*, 49(8):1018–1038, 2003.

C. Maglaras and A. Zeevi. Pricing and design of differentiated services: Approximate analysis and structural insights. *Operations Research*, 53(2):242–262, 2005.

C. Maglaras, J. Yao, and A. Zeevi. Optimal price and delay differentiation in queueing systems. *Management Science (forthcoming)*, 2015.

S. Marston, Z. Li, S. Bandyopadhyay, J. Zhang, and A. Ghalsasi. Cloud computing—the business perspective. *Decision Support Systems*, 51(1):176–189, 2011.

E. Maskin and J. Tirole. A theory of dynamic oligopoly, ii: Price competition, kinked demand curves, and edgeworth cycles. *Econometrica: Journal of the Econometric Society*, pages 571–599, 1988.

R. P. McAfee. Pricing damaged goods. Economics Discussion Paper 2007-2, Kiel Institute for the World Economy, 2007.

H. Mendelson and S. Whang. Optimal incentive-compatible priority pricing for the m/m/1 queue. *Operations Research*, 38(5):870–883, 1990.

H. Mendelson. Pricing computer services: queueing effects. *Communications of the ACM*, 28(3):312–321, 1985.

D. Mitra and Q. Wang. Preservation of best-effort service on the internet in the presence of managed services and usage-generated applications. *Available at SSRN 2587828*, 2015.

K. S. Moorthy. Product and price competition in a duopoly. *Marketing Science*, 7(2):141–168, 1988.

P. Naor. The regulation of queue size by levying tolls. *Econometrica*, 37(1):15–24, 1969.

H. Nazerzadeh and R. S. Randhawa. Near-optimality of coarse service grades for customer differentiation in queueing systems. *Available at SSRN 2438300*, 2015.

G. L. Nemhauser. Integer programming: Global impact. *EURO INFORMS July 2013*, 2013.

A. G. Nikolaev, S. H. Jacobson, W. K. T. Cho, J. J. Sauppe, and E. C. Sewell. Balance optimization subset selection (boss): An alternative approach for causal inference with observational data. *Operations Research*, 61(2):398–412, 2013.

M. D. Noel. Edgeworth price cycles: Evidence from the toronto retail gasoline market*. *The Journal of Industrial Economics*, 55(1):69–92, 2007.

D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov. The eucalyptus open-source cloud-computing system. In *Cluster Computing and the Grid, 2009. CCGRID'09. 9th IEEE/ACM International Symposium on*, pages 124–131. IEEE, 2009.

Z. Ou, H. Zhuang, J. Nurminen, A. Ylä-Jääski, and P. Hui. Exploiting hardware heterogeneity within the same instance type of amazon ec2. In *4th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud)*, 2012.

C. H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Prentice-Hall, 1982.

C. Papadimitriou. *Computational Complexity*. Addison-Wesley, Reading (Mass.), 1994.

S. D. Pimentel, R. R. Kelz, J. H. Silber, and P. R. Rosenbaum. Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *Journal of the American Statistical Association*, 110(510):515–527, 2015.

B. P. Rimal, E. Choi, and I. Lumb. A taxonomy and survey of cloud computing systems. In *INC, IMS and IDC, 2009. NCM'09. Fifth International Joint Conference on*, pages 44–51. Ieee, 2009.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

P. R. Rosenbaum and D. B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

P. R. Rosenbaum and J. Silber. Matching and thick description in an observational study of mortality after surgery. *Biostatistics*, 2:217–232, 2001.

P. R. Rosenbaum and J. H. Silber. Amplification of sensitivity analysis in observational studies. *Journal of the American Statistical Association*, 104(488):1398–1405, 2009.

P. R. Rosenbaum, R. N. Ross, and J. H. Silber. Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association*, 102(477):75–83, 2007.

P. R. Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394, 1987.

P. R. Rosenbaum. Optimal matching for observational studies. *Journal of the American Statistical Association*, 84:1024–1032, 1989.

P. R. Rosenbaum. Discussing hidden bias in observational studies. *Archives of Internal Medicine*, 115:901– 905, 1991.

P. R. Rosenbaum. Hodges-lehmann point estimates of treatment effect in observational studies. *Journal of the American Statistical Association*, 88(424):1250–1253, 1993.

P. R. Rosenbaum. Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika*, 88(1):219–231, 2001.

P. R. Rosenbaum. *Observational Studies*. Springer, 2002.

P. R. Rosenbaum. Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *The American Statistician*, 59(2):147–152, April 2005.

P. R. Rosenbaum. *Design of Observational Studies*. Springer, 2010.

P. R. Rosenbaum. Weighted m-statistics with superior design sensitivity in matched observational studies with multiple controls. *Journal of the American Statistical Association*, 109(507):1145–1158, 2014.

P. R. Rosenbaum. How to see more in observational studies: Some new quasi-experimental devices. *Annual Review of Statistics and Its Application*, 2(1):null, 2015.

D. B. Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74:318–328, 1979.

D. B. Rubin. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840, 2008.

S. V. Savin, M. A. Cohen, N. Gans, and Z. Katalan. Capacity management in rental businesses with two customer bases. *Operations Research*, 53(4):617–631, 2005.

J. Schad, J. Dittrich, and J. Quiané-Ruiz. Runtime measurements in the cloud: observing, analyzing, and reducing variance. *Proceedings of the VLDB Endowment*, 3(1-2):460–471, 2010.

A. Shaked and J. Sutton. Relaxing price competition through product differentiation. *The review of economic studies*, pages 3–13, 1982.

E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21, 2010.

M. Traskin and D. Small. Defining the study population for an observational study to ensure sufficient overlap: a tree approach. *Statistics in Biosciences*, 3:94–118, 2011.

G. Wang and T. S. E. Ng. The impact of virtualization on network performance of amazon ec2 data center. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9. IEEE, 2010.

S. Weston and R. Calaway. *Getting Started with `doParallel` and `foreach`*, 2014.

J. Whitney and P. Delforge. Data center efficiency assessment. *Natural Resources Defense Council, New York City, New York*, 2014.

H. Xu and B. Li. A study of pricing for cloud resources. *ACM SIGMETRICS Performance Evaluation Review*, 40(4):3–12, 2013.

D. Yang, D. Small, J. H. Silber, and P. R. Rosenbaum. Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. *Biometrics*, 68(2):628–636, 2012.

F. Yoon. Entire matching and its application in an observational study of treatments for melanoma. *Dissertation, Department of Statistics, The Wharton School, University of Pennsylvania*, 2009.

J. R. Zubizarreta and C. Kilcioglu. *designmatch: Construction of Optimally Matched Samples for Randomized Experiments and Observational Studies that are Balanced by Design*, 2016. R package version 0.1.1.

J. R. Zubizarreta, C. E. Reinke, R. R. Kelz, J. H. Silber, and P. R. Rosenbaum. Matching for several sparse nominal variables in a case-control study of readmission following surgery. *The American Statistician*, 65(4):229–238, 2011.

J. R. Zubizarreta, D. S. Small, N. K. Goyal, S. A. Lorch, and P. R. Rosenbaum. Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *Annals of Applied Statistics*, 7:25–50, 2013.

J. R. Zubizarreta, R. D. Paredes, and P. R. Rosenbaum. Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile. *Annals of Applied Statistics*, 8(1):204–231, 2014.

J. R. Zubizarreta. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500):1360–1371, 2012.

J. R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

Appendix A

Appendix to Chapter 2

A.1 Super-Linear Increase in Valuation with Congestion Sensitivity

This section uses the discrete type (v, κ) model of Section 2.3.3 but assumes that $\frac{v_1}{\kappa_1} \geq \frac{v_2}{\kappa_2} \geq \ldots \geq \frac{v_n}{\kappa_n}$, i.e., that the valuation rate grows super-linearly with respect to the corresponding congestion sensitivity rate.

$$\text{Model 3:} \quad v_1 \geq v_2 \geq \ldots \geq v_n > 0, \ \kappa_1 > \kappa_2 > \ldots > \kappa_n > 0, \ \frac{v_1}{\kappa_1} \geq \frac{v_2}{\kappa_2} \geq \ldots \geq \frac{v_n}{\kappa_n}. \tag{A.1}$$

The objective of the SP is to maximize its revenue by offering price vector $\boldsymbol{p}$ and availability vector $\boldsymbol{\pi}$. Similar to Lemma 1, in this setting users need only consider bids that are equal to one of the offered price points. Here we do not introduce a separate G service at first. User types that bid equal to the highest price level, p_1, receive uninterrupted service (i.e., G service) and pay $\overline{p}_1$. Hence, we first consider only BE service first, find the optimal pricing mechanism in this case, and then introduce a separate G service with price $\overline{p}_1$. The next proposition characterizes the structure of the optimal solution when the SP maximizes its revenue over the price grid and associated π's.

Proposition 11. *Consider the model specified by (A.1). Let k^* be the number of distinct price levels offered in BE service. Then, $k^* \leq 2$.*

Proposition 11 shows that it is optimal to offer BE service with at most two price levels. Let these price levels be p_H and p_L with $p_H \geq p_L$, and π is the fraction of time BE service is priced at p_L. If $p_H = p_L$, then the solution has only one service level, which is uninterrupted service. If $p_H > p_L$, then customers that bid p_H enjoys uninterrupted service by paying $\pi p_L + (1-\pi)p_H$. In this case, in addition to BE service we can offer G service with price $p_G := \pi p_L + (1-\pi)p_H$. Customer types bidding p_H previously are now indifferent between bidding p_H for BE service or paying p_G for G service. To ensure that these customer types choose G service over BE service with bid p_H, we can increase p_H without changing p_G. Now, these customers are no longer indifferent between the two options. Note that this change would not affect the choice of customer types that choose to bid p_L.

Next, we compare the optimal revenue that the SP makes with at most two service levels, i.e., two price levels, with the optimal revenue under one service level. Proposition 12 shows that the revenue under the former case is bounded above by the latter, or equivalently, offering one price level is optimal.

Proposition 12. *Consider the model specified by (A.1). Then,* $k^* = 1$.

Offering BE service with one price level means that the price is constant over time and customer types that bid this constant price get uninterrupted service, which is equivalent to G service, and the rest of the customer types do not get any service. Thus, we conclude that offering only G service is optimal if (v, κ) follows (A.1).

The result above is consistent with the policy identified in Katta and Sethuraman (2005), whereat the authors showed that for the model considered in this section that optimal policy for a SP operating a system with congestion effects (arising through the operation of an $M/M/1$ system) is optimal not to inject any strategic delay. In a large scale system, the service level that arises due to stochastic congestion effects becomes small, and in an infinite capacity system altogether disappears; i.e., if the SP can avoid congestion effects, she will indeed select to do so. This is what we see in our model as well. (Similarly to what we mentioned earlier the structure of the user utility function is different in our model, so a direct application of these earlier findings is not possible.)

A.2 Proofs

Proof. Proof of Proposition 1

1. Suppose $\pi^* = \dfrac{B}{1+B} + \varepsilon$, $\varepsilon > 0$ and $\bar{\pi} = \dfrac{B}{1+B}$, where $\bar{\pi}$ is not one of the optimal values for π. Then,

$$\begin{aligned} R(\eta_H^*, \eta_L^*, \bar{\pi}) - R(\eta_H^*, \eta_L^*, \pi^*) = \varepsilon \left\{ [(1+B)\eta_H^* + A]\, \bar{F}(\eta_H^*) - [(1+B)\eta_L^* + A]\, \bar{F}(\eta_L^*) \right\} < 0 \\ \iff [(1+B)\eta_H^* + A]\, \bar{F}(\eta_H^*) < [(1+B)\eta_L^* + A]\, \bar{F}(\eta_L^*) \\ \implies \eta_H^* = \eta_L^* + \gamma \ (\gamma > 0) \end{aligned}$$

   However, decreasing η_H^* by γ increases R. Therefore, $(\eta_H^*, \eta_L^*, \pi^*)$ is not the optimal solution. Contradiction.

2. If $\pi^* = \frac{B}{1+B}$, then the willingness to pay for BE service is $A\frac{B}{1+B}$, which is independent of the customer types. Therefore, η_L^* is either $\underline{\eta}$ or $\overline{\eta}$. If $\eta_L^* = \overline{\eta}$, then $\eta_H^* \geq \overline{\eta}$ since $\eta_L^* \leq \eta_H^*$. However, $R(\overline{\eta}, \overline{\eta}, \frac{B}{1+B}) = 0$, and $R(\eta_H, \eta_L, \frac{B}{1+B})$ is nonnegative for all $\underline{\eta} \leq \eta_L \leq \eta_H \leq \overline{\eta}$. Therefore, $\eta_L^* = \underline{\eta}$. Finally, if $\eta_L^* = \underline{\eta}$, then $p_L^* = A$.

3. $\eta_H^* = \eta_H^*(1+B)(1-\pi^*) = p_G^* - A + \pi^*(A - p_L^*) \iff \eta_H^* - \pi^*(A - p_L^*) = \eta_H^* = p_G^* - A$.

□

Proof. Proof of Proposition 2 The first order condition for the objective function is

$$\begin{aligned} &\frac{d}{d\eta_H} \left[\left(\eta_H + \frac{A}{1+B} \right) \bar{F}(\eta_H) \right] \Big|_{\eta_H = \eta_H^*} = 0 \\ \Rightarrow &\bar{F}(\eta_H^*) - \left(\eta_H^* + \frac{A}{1+B} \right) f(\eta_H^*) = 0 \\ \Rightarrow &f(\eta_H^*) = \frac{\bar{F}(\eta_H^*)}{\eta_H^* + \dfrac{A}{1+B}} \end{aligned}$$

$\Rightarrow$: If there are two services offered, then $\eta_H > \underline{\eta}$, which implies

$$\frac{d}{d\eta_H}\left[\left(\eta_H + \frac{A}{1+B}\right)\bar{F}(\eta_H)\right]\Big|_{\eta_H=\underline{\eta}} > 0 \quad \Rightarrow \; f(\underline{\eta}) < \frac{1}{\underline{\eta} + \frac{A}{1+B}}$$

$\Leftarrow$: Suppose (2.11) holds and the SP offers only G. This implies $\eta_H^* = \underline{\eta}$. However, if (2.11) holds

$$\frac{d}{d\eta_H}\left[\left(\eta_H + \frac{A}{1+B}\right)\bar{F}(\eta_H)\right]\Big|_{\eta_H=\eta_H^*} > 0,$$

which contradicts the optimality condition of η_H. □

Proof. Proof of Proposition 3 We allow $0 \leq \pi \leq \frac{B}{1+B}$ instead of strict inequality and show the proposition holds for this larger feasible region. We analyze four collectively exhaustive alternatives for the bounds of S_G and S_{BE} below.

- *Case 1: $\eta_L \geq \eta_H$ and $\eta_H \geq p_G - A$:* The revenue function becomes

$$\begin{aligned}\bar{R}_1 &= \left[\eta_H(1+B)(1-\pi) + \pi\eta_L - B(1-\pi)\eta_L + A\right]\bar{F}(\eta_H) + \left[\pi\eta_L - B(1-\pi)\eta_L + A\pi\right]F(\eta_H) \\ &= \left[\eta_H + (\eta_L - \eta_H)(\pi - B(1-\pi)) + A\right]\bar{F}(\eta_H) + \left[\pi\eta_L - B(1-\pi)\eta_L + A\pi\right]F(\eta_H)\end{aligned}$$

  $\bar{R}_1$ is non-decreasing in π. Hence $\eta = \frac{B}{1+B}$ is the optimal solution. In the optimal π,

$$\bar{R}_1 = (\eta_H + A)\bar{F}(\eta_H) + \frac{AB}{B+1}F(\eta_H) \leq \left(\eta_H + \frac{A}{1+B}\right)\bar{F}(\eta_H) + \frac{AB}{B+1}.$$

  Hence, the solution is sub-optimal.

- *Case 2: $\eta_L \geq \eta_H$ and $\eta_H < p_G - A$:* If $\eta_L \geq \eta_H$, then $U_1(\eta_H) = U_2(\eta_H) \geq U_2(\eta_L) = 0$ since $U_2(\eta)$ is decreasing in η. Then, $U_1(\eta_H) \geq U_1(p_G - A) = 0$. Since $U_1(\eta)$ is increasing in η, $p_G - A \leq \eta_H$. Therefore, this case is not possible.

- *Case 3: $\eta_L \leq \eta_H$ and $\eta_H > p_G - A$:* If $\eta_L \leq \eta_H$, then $U_1(\eta_H) = U_2(\eta_H) \leq U_2(\eta_L) = 0$ since

$U_2(\eta)$ is decreasing in η. Then, $U_1(p_G - A) = 0 \geq U_1(\eta_H)$. Since $U_1(\eta)$ is increasing in η, $p_G - A \geq \eta_H$. Therefore, this case is not possible.

- *Case 4:* $\eta_L \leq \eta_H$ *and* $\eta_H \leq p_G - A$*:* The revenue function for this case is

$$\begin{aligned}\bar{R}_4 = &\left[\eta_H + (\eta_H - \eta_L)(-\pi + B - B\pi) + A\right] \bar{F}(\eta_H + (\eta_H - \eta_L)(-\pi + B - B\pi)) \\ &+ \left[\pi\eta_L - B(1-\pi)\eta_L + A\pi\right] F(\eta_L).\end{aligned}$$

Let $\bar{R}_4^{\eta_L}$ be the revenue function for a fixed η_L value and $(\eta_H^{\eta_L}, \pi^{\eta_L})$ be the optimal solution to the problem

$$\underset{\eta_H, \pi}{\text{maximize}} \quad R_4^{\eta_L}(\eta_H, \pi) \tag{A.2}$$

$$\text{subject to} \quad \eta_H \geq \eta_L,\ \pi \leq \frac{B}{1+B},\ \pi \geq 0. \tag{A.3}$$

We can easily show that $\pi^{\eta_L} = \frac{B}{1+B}$. Suppose $\pi^{\eta_L} < \frac{B}{1+B}$ with $\eta_H^{\eta_L} = \eta_L$. Then, $\bar{R}_4^{\eta_L}(\eta_L, \pi^{\eta_L}) = (\eta_L + A)\,\bar{F}(\eta_L) + [(\pi^{\eta_L} - B + B\pi^{\eta_L})\eta_L + A\pi^{\eta_L}]\,F(\eta_L)$. This function is increasing in π. Hence η_L is not the optimal solution, contradiction. Similarly, suppose $\pi^{\eta_L} < \frac{B}{1+B}$ with $\eta_H^{\eta_L} > \eta_L$. However, the solution can be improved by decreasing η_H and increasing π simultaneously, contradiction. Hence, $\pi^{\eta_L} = \frac{B}{1+B}$. Then

$$\begin{aligned}R_4^{\eta_L} &= max\left\{\max_{\eta_H} (\eta_H + A)\,\bar{F}(\eta_H) + \frac{AB}{1+B}F(\eta_L) \text{ s.t. } \eta_H \geq \eta_L, \left(\eta_L + \frac{A}{1+B}\right)\bar{F}(\eta_L) + \frac{AB}{1+B}\right\} \\ &\leq \max_{\eta_H}\left(\eta_H + \frac{A}{1+B}\right)\bar{F}(\eta_H) + \frac{AB}{1+B} \text{ s.t. } \eta_H \geq \eta_L.\end{aligned}$$

The solution of (A.2)–(A.3) is bounded above by the problem

$$\underset{\eta_H}{\text{maximize}}\left(\eta_H + \frac{A}{1+B}\right)\bar{F}(\eta_H) + \frac{AB}{1+B} \text{ s.t. } \eta_H \geq \eta_L.$$

i.e., $\bar{R}_4 = \underset{\eta_L}{max}\bar{R}_4^{\eta_L}$ is bounded above by R_1. Hence the solution under $\eta_L \leq \eta_H$ and $\eta_H \leq p_G - A$ is sub-optimal.

Therefore, $\pi \leq \dfrac{B}{1+B}$ is sub-optimal. □

Proof. Proof of Proposition 4 We start with the following lemma.

Lemma 3. *For any given* $(p_G, p_2, \ldots, p_N, \boldsymbol{\pi})$ *that satisfies* $p_2 \geq p_3 \geq \ldots \geq p_N \geq 0$, $1^T\boldsymbol{\pi} = 1$, $\pi_N \geq \dfrac{B}{1+B}$, *and* $\boldsymbol{\pi} \geq 0$, *there exist* $\eta_1 \geq \eta_2 \geq \ldots \geq \eta_N$ *such that*

$$\mathcal{S}_G = \{\eta | \eta \geq \eta_1\} \text{ and } \mathcal{S}_{BE}^i = \{\eta | \eta_i \leq \eta \leq \eta_{i-1}\},\ i = 2, 3, \ldots, N,$$

and similarly, for any given $(\eta_1, \eta_2, \ldots, \eta_N \geq 0, \boldsymbol{\pi})$ *that satisfies* $\eta_1 \geq \eta_2 \geq \ldots \geq \eta_N$, $1^T\boldsymbol{\pi} = 1$, $\pi_N \geq \dfrac{B}{1+B}$, *and* $\boldsymbol{\pi} \geq 0$, *there exists a set of price levels* p_G *and* $p_2 \geq p_3 \geq \ldots \geq p_N \geq 0$.

Assuming $\eta_1 \geq \ldots \geq \eta_N$, the revenue function of the SP is

$$\begin{aligned} R &= p_G\bar{F}(\eta_1) + \bar{p}_2(F(\eta_1) - F(\eta_2)) + \ldots + \bar{p}_N(F(\eta_{N-1}) - F(\eta_N)) \\ &= (p_G - \bar{p}_2)\bar{F}(\eta_1) + (\bar{p}_2 - \bar{p}_3)\bar{F}(\eta_2) \ldots + (\bar{p}_{N-1} - \bar{p}_N)\bar{F}(\eta_{N-1}) + \bar{p}_N\bar{F}(\eta_N) \\ &= \sum_{i=1}^{N-1} \left[\pi_i(A + \eta_i) + B\eta_i\pi_i\right] \bar{F}(\eta_i) + \left[\pi_N(A + \eta_N) - B\eta_N(1 - \pi_N)\right] \bar{F}(\eta_N). \end{aligned}$$

Therefore, the revenue maximization problem becomes

$$\underset{\boldsymbol{\eta}, \boldsymbol{\pi}}{\text{maximize}} \quad \sum_{i=1}^{N-1} \pi_i \left[(A + \eta_i) + B\eta_i\right] \bar{F}(\eta_i) + \left[\pi_N(A + \eta_N) - B\eta_N(1 - \pi_N)\right] \bar{F}(\eta_N) \tag{A.4}$$

$$\text{subject to} \quad \eta_1 \geq \eta_2 \geq \ldots \geq \eta_N,\ \pi_N \geq \frac{B}{1+B},\ 1^T\boldsymbol{\pi} = 1,\ \boldsymbol{\pi} \geq 0. \tag{A.5}$$

Lemma 4. *There exists an optimal solution to the problem above such that at most one of the optimal* $(\pi_1, \pi_2, \ldots, \pi_{N-1})$ *values is nonnegative.*

From Lemma 4, the problem can be simplified to

$$\begin{aligned} \underset{\eta_H, \eta_L, \pi}{\text{maximize}} \quad & (1 - \pi)\left[(A + \eta_H) + B\eta_H\right] \bar{F}(\eta_H) + \left[\pi(A + \eta_L) - B\eta_L(1 - \pi)\right] \bar{F}(\eta_L) \\ \text{subject to} \quad & \eta_H \geq \eta_L,\ \pi \geq \frac{B}{1+B},\ \pi \leq 1. \end{aligned}$$

which is equivalent to the two-price-level problem. □

Proof. Proof of Proposition 5 ⇒: Assume there are G and BE services in the optimal solution and $p^* \leq v_n$. If $p^* \leq v_n$, then $p^* = v_n$ by the optimality of p^*, which implies there is no customer type that chooses no-buy option. If offering G and BE services together generates more revenue, then there is at least one customer type that chooses BE over G.

When there is only G service, the optimal revenue is $R_1 = \sum_{i=1}^{n} \lambda_i v_n$. When there are two services, the optimal revenue is $R_2 = \sum_{i \in S_1} \lambda_i p_G + \sum_{i \in S_2} \lambda_i \bar{p}_{[i]}$, where S_1 is the set of customer types that chooses G and S_2 is the set of customer types that chooses BE ($S_1 \cap S_2 = \emptyset$ and $S_1 \cup S_2 = \{1, 2, \ldots, n\}$). p_G is the optimal price for G service, which implies $p_G = v_M$ where $M = \max\{i | i \in S_1\}$, and $\bar{p}_{[i]} = \sum_{j=[i]}^{N} \pi_j p_j$ which is the payment of customer type i with her optimal bid value $p_{[i]}$.

Now we will show that $\bar{p}_{[i]} < p_G = v_M \ \forall i \in S_2$. Suppose $\bar{p}_{[k]} \geq p_G = v_M,\ k \in S_2$. Since $k \in S_2$, $\bar{\pi}_{[k]} v_k - (1 - \bar{\pi}_{[k]})\kappa_k - \bar{p}_{[k]} \geq v_k - v_M$, where $\bar{\pi}_{[k]} = \sum_{j=[k]}^{N} \pi_j$. If $\bar{p}_{[k]} \geq v_M$, then $\bar{\pi}_{[k]} v_k - (1 - \bar{\pi}_{[k]})\kappa_k \geq v_k$, which is not possible since $\bar{\pi}_{[k]} < 1$. Therefore, $\bar{p}_{[i]} < p_G = v_M \ \forall i \in S_2$.

$$R_2 = \sum_{i \in S_1} \lambda_i p_G + \sum_{i \in S_2} \lambda_i \bar{p}_{[i]} < \sum_{i=1}^{n} \lambda_i p_G \leq \sum_{i=1}^{n} \lambda_i p^*,$$

where the second inequality comes from the optimality of p^* in one product case. Therefore, offering G and BE services together does not generate more revenue than offering only G service. Contradiction.

⇐: Let $H = \underset{1 \leq i \leq n}{\operatorname{argmin}}\{p^* \geq v_i\}$. From the optimality of p^*, $v_H = p^*$, and the set of customer types $\{1, 2, \ldots, H\}$ choose G. If we offer a BE service such that no customer types from the set $\{1, 2, \ldots, H\}$ prefer BE and at least one customer type from the set $\{H+1, H+2, \ldots, n\}$ chooses BE, then the revenue generated by G and BE services together becomes higher than that of G service only.

Set $\pi = \dfrac{\kappa_H - \kappa_{H+1}}{v_H - v_{H+1} + \kappa_H - \kappa_{H+1}}$, $p_2 = \dfrac{\kappa_H v_{H+1} - \kappa_{H+1} v_H}{\kappa_H - \kappa_{H+1}}$ and $p_1 = \infty$. Then,

$$v_i - p^* \geq \pi v_i - (1-\pi)\kappa_i - \pi p_2 \text{ for } i \leq H \text{ and } \pi v_{H+1} - (1-\pi)\kappa_{H+1} - \pi p_2 \geq 0,$$

which implies all customer types $i \leq H$ choose G and customer type $H+1$ chooses BE service. □

Proof. Proof of Proposition 11 Let $p_{[i]}$ be the optimal bid for customer type i. Therefore, customer type i either makes a bid of $p_{[i]}$ or leaves the system with no purchase. Clearly, customers with high valuations prefer bidding higher, that is, $p_{[i]}$ is non-increasing in i. Let $s \in \{1, 2, \ldots, n\}$ be the highest customer index that makes a bid, which is determined by $\boldsymbol{p}$ and $\boldsymbol{\pi}$. Therefore, s is not a decision variable. However, an alternative way to solve the problem is to find the optimal $\boldsymbol{p}$ and $\boldsymbol{\pi}$ for any possible s value, and then choose the s that generates the maximum revenue. Now we characterize and solve the revenue maximization problem for a given s value.

For any $k \in \{1, 2, \ldots, n - [i]\}$, type i customer prefers bidding $p_{[i]}$ over $p_{[i]+k}$ if

$$\begin{aligned} & U_2(v_i, \kappa_i, p_{[i]}) \geq U_2(v_i, \kappa_i, p_{[i]+k}) \\ \bar{\pi}_{[i]} v_i - \kappa_i(1 - \bar{\pi}_{[i]}) - \bar{p}_{[i]} & \geq \bar{\pi}_{[i]+k} v_i - \kappa_i(1 - \bar{\pi}_{[i]+k}) - \bar{p}_{[i]+k} \\ (\bar{\pi}_{[i]} - \bar{\pi}_{[i]+k})(v_i + \kappa_i) & \geq \bar{p}_{[i]} - \bar{p}_{[i]+k}, \qquad i = 1, \ldots, s \end{aligned} \tag{A.6}$$

and prefers bidding $p_{[i]}$ over $p_{[i]-k}$ if

$$\begin{aligned} & U_2(v_i, \kappa_i, p_{[i]-k}) \leq U_2(v_i, \kappa_i, p_{[i]}) \\ \Leftrightarrow \ & \bar{\pi}_{[i]-k} v_i - \kappa_i(1 - \bar{\pi}_{[i]-k}) - \bar{p}_{[i]-k} \leq \bar{\pi}_{[i]} v_i - \kappa_i(1 - \bar{\pi}_{[i]}) - \bar{p}_{[i]} \\ \Leftrightarrow \ & (\bar{\pi}_{[i]-k} - \bar{\pi}_{[i]})(v_i + \kappa_i) \leq \bar{p}_{[i]-k} - \bar{p}_{[i]}. \qquad i = 1, \ldots, s \end{aligned} \tag{A.7}$$

Lastly, type i customer prefers bidding $p_{[i]}$ over no bidding (i.e., leaving the system with no purchase) if

$$\bar{\pi}_{[i]} v_i - \kappa_i(1 - \bar{\pi}_{[i]}) - \bar{p}_{[i]} \geq 0. \qquad i = 1, \ldots, s \tag{A.8}$$

Lemma 5. *$\bar{\pi}_{[i]}v_i - \kappa_i(1-\bar{\pi}_{[i]})$ is decreasing in i ($i \le s$).*

Lemma 6. *(A.8) can be simplified to*

$$\bar{\pi}_{[s]}v_s - \kappa_s(1-\bar{\pi}_{[s]}) - \bar{p}_{[s]} \ge 0. \tag{A.9}$$

For a given s, the objective function for the SP is

$$R_s = \sum_{i=1}^{s} \lambda_i \bar{p}_{[i]} = \sum_{j=1}^{s-1}\sum_{i=1}^{j} \lambda_i(\bar{p}_{[j]} - \bar{p}_{[j+1]}) + \sum_{i=1}^{s} \lambda_i \bar{p}_{[s]}.$$

Therefore, the revenue maximization problem can be written as

$$\underset{\boldsymbol{\pi},\boldsymbol{p}}{\text{maximize}} \quad R_s = \sum_{j=1}^{s-1}\sum_{i=1}^{j} \lambda_i(\bar{p}_{[j]} - \bar{p}_{[j+1]}) + \sum_{i=1}^{s} \lambda_i \bar{p}_{[s]} \tag{A.10}$$

$$\text{subject to} \quad (\bar{\pi}_{[i]} - \bar{\pi}_{[i]+k})(v_i+\kappa_i) \ge \bar{p}_{[i]} - \bar{p}_{[i]+k} \qquad i=1,\dots,s;\ \ k=1,2,\dots,N-[i] \tag{A.11}$$

$$(\bar{\pi}_{[i]-k} - \bar{\pi}_{[i]})(v_i+\kappa_i) \le \bar{p}_{[i]-k} - \bar{p}_{[i]} \qquad i=1,\dots,s;\ \ k=1,2,\dots,[i]-1 \tag{A.12}$$

$$\bar{\pi}_{[s]}v_s - \kappa_s(1-\bar{\pi}_{[s]}) - \bar{p}_{[s]} \ge 0 \tag{A.13}$$

$$p_1 \ge p_2 \ge \dots \ge p_N \ge 0 \tag{A.14}$$

$$\mathbf{1}^T\boldsymbol{\pi} = 1,\ \boldsymbol{\pi} \ge 0. \tag{A.15}$$

Lemma 7. *Let $(\boldsymbol{\pi}^*, \boldsymbol{p}^*)$ denote an optimal solution to the problem above. Then,*

$$\begin{aligned}
p^*_{[1]} = p^*_{[1]+1} = \dots = p^*_{[2]-1} &= v_1 + \kappa_1,\\
p^*_{[2]} = p^*_{[2]+1} = \dots = p^*_{[3]-1} &= v_2 + \kappa_2,\\
&\vdots\\
p^*_{[s-1]} = p^*_{[s-1]+1} = \dots = p^*_{[s]-1} &= v_{s-1} + \kappa_{s-1},\\
p^*_{[s]} = p^*_{[s]+1} = \dots = p^*_N &= v_s + \kappa_s - \frac{\kappa_s}{\bar{\pi}^*_{[s]}}.
\end{aligned} \tag{A.16}$$

Lemma 7 simplifies the revenue maximization problem to

$$\begin{aligned} \underset{\boldsymbol{\pi}}{\text{maximize}} \quad & R_s = \sum_{j=1}^{s-1}\sum_{i=1}^{j}\lambda_i(\bar{\pi}_{[j]}-\bar{\pi}_{[j+1]})(v_j+\kappa_j)+\sum_{i=1}^{s}\lambda_i\left[\bar{\pi}_{[s]}(v_s+\kappa_s)-\kappa_s\right] \\ \text{subject to} \quad & \bar{\pi}_{[s]} \geq \frac{\kappa_s}{v_s+\kappa_s},\ \mathbf{1}^T\boldsymbol{\pi}=1,\ \boldsymbol{\pi}\geq 0. \end{aligned}$$

This problem has N decision variables. It can be simplified further using the following change of variables

$$\begin{aligned} &\alpha_0 = 1-\bar{\pi}_{[1]} \\ &\alpha_i = \bar{\pi}_{[i]}-\bar{\pi}_{[i+1]} && i=1,2,\ldots,s-1 \\ &\alpha_s = \bar{\pi}_{[s]} \end{aligned}$$

where α_i can be interpreted as the time that the price level is equal to $v_i+\kappa_i$ for $i=1,2,\ldots,s-1$, α_0 as the time that the price level is above $v_1+\kappa_1$ and α_s as the time that the price level is at its minimum.

Therefore, the problem becomes

$$\underset{\alpha_0,\alpha_1,\ldots,\alpha_s}{\text{maximize}} \quad R_s = \sum_{j=1}^{s}\sum_{i=1}^{j}\lambda_i\alpha_j(v_j+\kappa_j)-\sum_{i=1}^{s}\lambda_i\kappa_s \tag{A.17}$$

$$\text{subject to} \quad \alpha_s \geq \frac{\kappa_s}{v_s+\kappa_s},\ \alpha_0+\alpha_1+\ldots+\alpha_s=1,\ \alpha_0,\alpha_1,\ldots,\alpha_s\geq 0. \tag{A.18}$$

Lemma 8. *Let $(\alpha_0^*,\alpha_1^*,\ldots,\alpha_s^*)$ denote the optimal solution to (A.17)–(A.18) and $k=\underset{j\in\{1,\ldots,s\}}{argmax}\{(v_j+\kappa_j)\sum_{i=1}^{j}\lambda_i\}$. If $k=s$, $\alpha_0^*=\alpha_1^*=\ldots=\alpha_{s-1}^*=0,\alpha_s^*=1$, else $\alpha_0^*=\ldots=\alpha_{k-1}^*=\alpha_{k+1}^*=\ldots=\alpha_{s-1}^*=0,\alpha_k^*=\frac{v_s}{v_s+\kappa_s},\alpha_s^*=\frac{\kappa_s}{v_s+\kappa_s}$.*

Lemma 8 shows that the SP offers at most two price levels, and there is no price level higher than the bid of the highest value customer. Since this result holds for any s, it also holds for the optimal s. Therefore, the optimal solution has at most two price levels. □

Proof. Proof of Proposition 12 The proposition is equivalent to the following statement.

Let Π_2 be optimal revenue that the SP achieves by offering at most two price levels:

$$\Pi_2 = \max_{1 \le s \le n} R_s,$$

and Π_1 be the optimal revenue by offering only one price level:

$$\Pi_1 = v_{k^*} \sum_{i=1}^{k^*} \lambda_i$$

where $k^* = \underset{j \in \{1,2,\ldots,n\}}{\operatorname{argmax}} \left\{ v_i \sum_{i=1}^{j} \lambda_i \right\}$. Then, $\Pi_2 = \Pi_1$.

If $k^* = n$, i.e., all customers are served with the price v_n, degrading the service for some customers would not increase the revenue, therefore, $k^* < n$ is the first condition to offer two price levels. We need to find two indexes, $\overline{k}$ and $\underline{k}$, the high-level threshold and low-level threshold, respectively, such that customer types $\{1, 2, \ldots, \overline{k}\}$ bid high price level, $\{\overline{k}+1, \overline{k}+2, \ldots, \underline{k}\}$ bid low price level, and $\{\underline{k}+1, \underline{k}+2, \ldots, n\}$ leave the system ($\overline{k} \le \underline{k} \le n$). Thus, the optimal solution for the two-price case can be written as

$$\Pi_2 = (1 - \alpha_{\underline{k}})(v_{\overline{k}} + \kappa_{\overline{k}}) \sum_{i=1}^{\overline{k}} \lambda_i + \alpha_{\underline{k}}(v_{\underline{k}} + \kappa_{\underline{k}}) \sum_{i=1}^{\underline{k}} \lambda_i - \kappa_{\underline{k}} \sum_{i=1}^{\underline{k}} \lambda_i,$$

Since $\alpha_{\underline{k}} = \frac{\kappa_{\underline{k}}}{v_{\underline{k}} + \kappa_{\underline{k}}}$ from Lemma 8, the optimal revenue for two-price level case becomes

$$\Pi_2 = \frac{v_{\underline{k}}}{v_{\underline{k}} + \kappa_{\underline{k}}} \left(v_{\overline{k}} + \kappa_{\overline{k}}\right) \sum_{i=1}^{\overline{k}} \lambda_i.$$

Next we need to find $\overline{k}$ and $\underline{k}$ such that $\Pi_2 > \Pi_1$. Note that Π_2 is decreasing in $\underline{k}$. Therefore, $\underline{k} = \overline{k}$. However, this means that no customer type bids low price level, which is equivalent to one price level solution. Therefore, $\Pi_2 = \Pi_1$. □

A.3 Additional Proofs

Proof. Proof of Lemma 1 Bidding a value between two price levels is the same as bidding the

lower price level amount in terms of availability of the product and payment amount, therefore sub-optimal. □

Proof. Proof of Lemma 3 $\mathcal{S}^i_{BE}$ contains all η values that satisfy the following constraints

$$U_2(\eta, p_i) \geq U_2(\eta, p_{i+j}), j = 1, \ldots, N-i \tag{A.19}$$

$$U_2(\eta, p_i) \geq 0, \tag{A.20}$$

$$U_2(\eta, p_i) \geq U_2(\eta, p_{i-j}), j = 1, \ldots, i-2 \tag{A.21}$$

$$U_2(\eta, p_i) \geq U_1(\eta). \tag{A.22}$$

There exists an $\underline{\eta}_i$ such that any $\eta \geq \underline{\eta}_i$ satisfies (A.19) and (A.20). Similarly, there exists an $\overline{\eta}_i$ such that any $\eta \leq \overline{\eta}_i$ satisfies (A.21) and (A.22). Hence, $\mathcal{S}^i_{BE} = \{\eta | \underline{\eta}_i \leq \eta \leq \overline{\eta}_i\}$. For the rest of the analysis, we assume $\underline{\eta}_i \leq \overline{\eta}_i$ for $i = 2, 3, \ldots, N$.

Next, we will show that $\overline{\eta}_{i+1} \leq \underline{\eta}_i$ for $i = 2, 3, \ldots, N-1$. Suppose there exists an i such that $\overline{\eta}_{i+1} > \underline{\eta}_i$. Then, since $\bar{\pi}_i \geq \bar{\pi}_{i+1} \geq 0$, $U_2(\overline{\eta}_{i+1}, p_i) > U_2(\overline{\eta}_{i+1}, p_{i+1})$. However, from A.21, $U_2(\overline{\eta}_{i+1}, p_{i+1}) \geq U_2(\overline{\eta}_{i+1}, p_i)$. Contradiction. Therefore, $\overline{\eta}_{i+1} \leq \underline{\eta}_i$ for $i = 2, 3, \ldots, N-1$.

Now, we will find conditions for $\underline{\eta}_i, \overline{\eta}_i$ for $i = 2, 3, \ldots, N$. First, we show that $U_2(\underline{\eta}_i, p_i) = U_2(\underline{\eta}_i, p_{i+1})$ for $i = 2, 3, \ldots, N-1$. Suppose it is not true, which implies $U_2(\underline{\eta}_i, p_i) = U_2(\underline{\eta}_i, p_{i+k}) > U_2(\underline{\eta}_i, p_{i+1})$ for some $k > 1$. Moreover, $U_2(\overline{\eta}_{i+1}, p_{i+1}) \geq U_2(\overline{\eta}_{i+1}, p_{i+k})$. Since $\underline{\eta}_i \geq \overline{\eta}_{i+1}$ and $U_2'(\eta, p_{i+1}) \geq U_2'(\eta, p_{i+k})$, $U_2(\underline{\eta}_i, p_{i+1}) \geq U_2(\underline{\eta}_i, p_{i+k})$. Contradiction. Therefore, $U_2(\underline{\eta}_i, p_i) = U_2(\underline{\eta}_i, p_{i+1})$. For $i = N$, $U_2(\underline{\eta}_N, p_N) = 0$. Second we show that $U_2(\overline{\eta}_i, p_i) = U_2(\overline{\eta}_i, p_{i-1})$ for $i = 3, 4, \ldots, N$. Suppose not true, which implies $U_2(\overline{\eta}_i, p_i) = U_2(\overline{\eta}_i, p_{i-k}) > U_2(\overline{\eta}_i, p_{i-1})$ for some $k > 1$. Moreover, $U_2(\underline{\eta}_{i-1}, p_{i-1}) \geq U_2(\underline{\eta}_{i-1}, p_{i-k})$. Since $\underline{\eta}_{i-1} \geq \overline{\eta}_i$ and $U_2'(\eta, p_{i-1}) \geq U_2'(\eta, p_{i-k})$, $U_2(\overline{\eta}_i, p_{i-1}) \geq U_2(\overline{\eta}_i, p_{i-k})$. Contradiction. Therefore, $U_2(\overline{\eta}_i, p_i) = U_2(\overline{\eta}_i, p_{i-1})$. For $i = 2$, $U_2(\overline{\eta}_2, p_2) = U_1(\overline{\eta}_2)$. From the two conditions on $\underline{\eta}_i$ and $\overline{\eta}_i$, we reach $\underline{\eta}_{i-1} = \overline{\eta}_i$ for $i = 3, 4, \ldots, N$.

Next step is to rename the boundaries. Let $\eta_i = \underline{\eta}_i$ for $i = 2, 3, \ldots, n$ and $\eta_1 = \overline{\eta}_2$. This concludes the first part of the proposition.

For any given $(\eta_1, \eta_2, \ldots, \eta_N, \boldsymbol{\pi})$ that satisfies $\eta_1 \geq \eta_2 \geq \ldots \geq \eta_N$, $1^T\boldsymbol{\pi} = 1$, $\pi_N \geq \dfrac{B}{1+B}$, and $\boldsymbol{\pi} \geq 0$, the following prices satisfy $p_2 \geq p_3 \geq \ldots \geq p_n \geq 0$ and they are aligned with $\mathcal{S}_G$ and $\mathcal{S}^i_{BE}$,

$i = 2, 3, \ldots, N$:

$$\begin{aligned} p_G &= A + \eta_1, \\ p_i &= A + \eta_i + B\eta_i, && i = 2, 3, \ldots, N-1, \\ p_N &= A + \eta_N + B\eta_N - \frac{B\eta_N}{\pi_N}. \end{aligned}$$

□

Proof. Proof of Lemma 4 Let $(\boldsymbol{\eta}^*, \boldsymbol{\pi}^*)$ be an optimal solution. Assume that there are two π_i^* $(i = 1, 2, \ldots, N-1)$ values such that $\pi_j^* > 0$, $\pi_k^* > 0$, and all others are equal to 0. Without loss of generality, $j < k$ which implies $\eta_j^* \geq \eta_k^*$. If $\eta_j^* = \eta_k^*$, then another optimal solution would be $\pi_j^* := \pi_j^* + \pi_k^*$ and $\pi_k^* := 0$, which has only one nonnegative π_i value for $i = 1, \ldots, N-1$. If $\eta_j^* > \eta_k^*$, then there are three possible cases:

Case 1: $[(A + \eta_j^*) + B\eta_j^*] = [(A + \eta_k^*) + B\eta_k^*]$: $\pi_j^* := \pi_j^* + \pi_k^*$ and $\pi_k^* := 0$ is another optimal solution where at most one π value is nonnegative.

Case 2: $[(A + \eta_j^*) + B\eta_j^*] > [(A + \eta_k^*) + B\eta_k^*]$: $\pi_j^* := \pi_j^* + \pi_k^*$ and $\pi_k^* := 0$ give a better solution, contradiction.

Case 3: $[(A + \eta_j^*) + B\eta_j^*] < [(A + \eta_k^*) + B\eta_k^*]$: $\pi_j^* := 0$ and $\pi_k^* := \pi_j^* + \pi_k^*$ give a better solution, contradiction.

Therefore, there cannot be two nonnegative π_i values $(i = 1, 2, \ldots, N-1)$. Using the same idea, we can generalize the result to more than two nonnegative value case. □

Proof. Proof of Lemma 5 From (A.8) and prices being nonnegative,

$$\bar{\pi}_{[i]} \geq \frac{\kappa_i}{v_i + \kappa_i}, \qquad i = 1, \ldots, s$$

and since $\frac{v_i}{\kappa_i}$ is decreasing in i,

$$\frac{\kappa_{i+1}}{v_{i+1}+\kappa_{i+1}} \geq \frac{\kappa_i}{v_i+\kappa_i}. \qquad i = 1, 2, \ldots, s-1.$$

Using these two inequalities and $\bar{\pi}_{[i]}$ being decreasing in i, for any $i = 1, 2, \ldots, s-1$,

$$\begin{aligned}
\bar{\pi}_{[i]}v_i - \kappa_i(1-\bar{\pi}_{[i]}) &= \bar{\pi}_{[i]}(v_i+\kappa_i) - \kappa_i \\
&= \bar{\pi}_{[i+1]}(v_i+\kappa_i) + (\bar{\pi}_{[i]} - \bar{\pi}_{[i+1]})(v_i+\kappa_i) - \kappa_i \\
&= \bar{\pi}_{[i+1]}\left[v_i+\kappa_i - (v_{i+1}+\kappa_{i+1})\right] + (\bar{\pi}_{[i]} - \bar{\pi}_{[i+1]})(v_i+\kappa_i) - \kappa_i + \bar{\pi}_{[i+1]}(v_{i+1}+\kappa_{i+1}) \\
&\geq \frac{\kappa_{i+1}}{v_{i+1}+\kappa_{i+1}}(v_i+\kappa_i) - \kappa_{i+1} + (\bar{\pi}_{[i]} - \bar{\pi}_{[i+1]})(v_i+\kappa_i) - \kappa_i + \bar{\pi}_{[i+1]}(v_{i+1}+\kappa_{i+1}) \\
&\geq \frac{\kappa_i}{v_i+\kappa_i}(v_i+\kappa_i) - \kappa_{i+1} + (\bar{\pi}_{[i]} - \bar{\pi}_{[i+1]})(v_i+\kappa_i) - \kappa_i + \bar{\pi}_{[i+1]}(v_{i+1}+\kappa_{i+1}) \\
&= \bar{\pi}_{[i+1]}(v_{i+1}+\kappa_{i+1}) - \kappa_{i+1} + (\bar{\pi}_{[i]} - \bar{\pi}_{[i+1]})(v_i+\kappa_i) \\
&\geq \bar{\pi}_{[i+1]}(v_{i+1}+\kappa_{i+1}) - \kappa_{i+1} \\
&= \bar{\pi}_{[i+1]}v_{i+1} - \kappa_{i+1}(1-\bar{\pi}_{[i+1]})
\end{aligned}$$

□

Proof. Proof of Lemma 6 Using Lemma 5, it can easily be shown that

$$\bar{\pi}_{[i+1]}(v_i+\kappa_i) - \kappa_i \geq \bar{\pi}_{[i+1]}(v_{i+1}+\kappa_{i+1}) - \kappa_{i+1}. \qquad i = 1, 2, \ldots, s-1$$

Then, using (A.6), for any $i = 1, 2, \ldots, s-1$,

$$\begin{aligned}
\bar{\pi}_{[i]}v_i - \kappa_i(1-\bar{\pi}_{[i]}) - \bar{p}_{[i]} &= \bar{\pi}_{[i]}(v_i+\kappa_i) - \kappa_i - \bar{p}_{[i+1]} - (\bar{p}_{[i]} - \bar{p}_{[i+1]}) \\
&\geq \bar{\pi}_{[i]}(v_i+\kappa_i) - \kappa_i - \bar{p}_{[i+1]} - (\bar{\pi}_{[i]} - \bar{\pi}_{[i+1]})(v_i+\kappa_i) \\
&= \bar{\pi}_{[i+1]}(v_i+\kappa_i) - \kappa_i - \bar{p}_{[i+1]} \\
&\geq \bar{\pi}_{[i+1]}(v_{i+1}+\kappa_{i+1}) - \kappa_{i+1} - \bar{p}_{[i+1]}.
\end{aligned}$$

Therefore,

$$\bar{\pi}_{[1]}v_1 - \kappa_1(1-\bar{\pi}_{[1]}) - \bar{p}_{[1]} \geq \bar{\pi}_{[2]}v_2 - \kappa_2(1-\bar{\pi}_{[2]}) - \bar{p}_{[2]} \geq \ldots \geq \bar{\pi}_{[s]}v_s - \kappa_s(1-\bar{\pi}_{[s]}) - \bar{p}_{[s]} \geq 0.$$

□

Proof. Proof of Lemma 7 First, we need to show that (A.16) is a feasible solution. Since $v_i + \kappa_i$ is decreasing in i, and (A.13) implies $\bar{\pi}^*_{[s]} \geq \frac{\kappa_s}{v_s+\kappa_s} > 0$, the solution satisfies (A.14). (A.15) is trivially satisfied since (A.16) does not impose anything on $\boldsymbol{\pi}$ and uses the optimal $\boldsymbol{\pi}^*$, which is also feasible. (A.13) holds with equality. Trivially (A.11) and (A.12) are also satisfied. Therefore, (A.16) is a feasible solution. Now we need to show that this solution is optimal. The objective function is equivalent to

$$R_s = \sum_{i=1}^{s} \lambda_i \bar{p}_{[i]},$$

where all $\bar{p}_{[i]}$ variables have nonnegative coefficients. Moreover, (A.16) provides a solution where all p variables are equal to their upper bounds. Therefore, the solution is an optimal solution. □

Proof. Proof of Lemma 8 Suppose $k = s$ and $\alpha^*_s = 1 - \varepsilon$ with $\alpha^*_j = \varepsilon$ $(0 \leq j < s, \varepsilon > 0)$. Since $(v_s + \kappa_s)\sum_{i=1}^{s}\lambda_i > (v_j + \kappa_j)\sum_{i=1}^{j}\lambda_i$, $\alpha_s = 1, \alpha_j = 0$ gives a higher objective function value. Contradiction. Similarly, if $k < s$, α^*_s has to be equal to its lower bound in the optimal solution. If $\underset{j\in\{1,\ldots,s\}}{argmax}\{(v_j + \kappa_j)\sum_{i=1}^{j}\lambda_i\}$ is not unique, there are alternative optimal solutions which has a solution given above. This proves that the SP offers at most two price levels. Moreover, since $k > 0$, $\alpha^*_s = 0$, which means the fraction of time the price level is above $v_1 + \kappa_1$ is equal to zero. Therefore, the price never goes beyond the bid of the highest value customer, which is equal to $v_1 + \kappa_1$. □

Appendix B

Appendix to Chapter 3

B.1 Proofs

Proof of Lemma 2. Suppose $i \in \mathcal{I}(p_1, q_1)$ and $i \notin \mathcal{I}_1(p_1, q_1) \cup \mathcal{I}_2(p_1, q_1)$. If $i \in \mathcal{I}(p_1, q_1)$, then $v_i \geq \frac{c_i + p_1}{q_1}$. If $i \notin \mathcal{I}_1(p_1, q_1) \cup \mathcal{I}_2(p_1, q_1)$ then $v_i < \frac{c_i + p_1}{q_1}$ and $v_i < \frac{c_i + 2p_1}{2\alpha q_1}$. Contradiction. □

Proof of Proposition 6. It is enough to show that $\Pi_{k+1}(p_1, q_1) \geq \Pi_k(p_1, q_1)$ for any (p_1, q_1) and $k \geq 1$.

We can easily show the inequality holds for $k = 1$ case, i.e., $\Pi_2(p_1, q_1) \geq \Pi_1(p_1, q_1)$ for any (p_1, q_1).

$$\begin{aligned}
\Pi_2(p_1, q_1) &= \sum_{i \in \mathcal{I}_1(p_1, q_1)} \lambda_i p_1 \frac{w_i}{q_1} + \sum_{i \in \mathcal{I}_2(p_1, q_1)} \lambda_i p_1 \frac{w_i}{\alpha q_1} \\
&\geq \sum_{i \in \mathcal{I}_1(p_1, q_1)} \lambda_i p_1 \frac{w_i}{q_1} + \sum_{i \in \mathcal{I}_2(p_1, q_1)} \lambda_i p_1 \frac{w_i}{q_1} \\
&\geq \sum_{i \in \mathcal{I}(p_1, q_1)} \lambda_i p_1 \frac{w_i}{q_1} = \Pi_1(p_1, q_1).
\end{aligned}$$

The same procedure follows for any $k > 1$. □

Proof of Proposition 7. Let $f_k(c) = \frac{c + 2^{k-1} p_1}{2^{k-1} \alpha^{k-1}}$. $f_k(c)$ is linearly increasing in c for any nonnegative integer k. The slope of $f_k(c)$ is $\frac{1}{2^{k-1}\alpha^{k-1}}$ which is decreasing in k. Hence, for any k, if $\frac{\bar{c} + 2^k p_1}{2^k \alpha^k q_1} \leq$

$\frac{\bar{c}+2^{k-1}p_1}{2^{k-1}\alpha^{k-1}q_1}$ for a given $\bar{c}$, then the inequality holds $\forall\ c \geq \bar{c}$.

Let c_k be the level such that $f_{k+1}(c_k) = f_k(c_k)$ (which implies $f_{k+1}(c) \leq f_k(c)\ \forall\ c \geq c_k$). If $c_i \in [0, c_1)$, then customer type i chooses the first quality level, and if $c_i \in [c_{k-1}, c_k)$. then customer type i chooses quality level k.

Finally, if $f_{k+1}(c_k) = \frac{c_k + 2^k p_1}{2^k \alpha^k} = \frac{c_k + 2^{k-1}p_1}{2^{k-1}\alpha^{k-1}} = f_k(c_k)$, then $c_k = \frac{2^k p_1(1-\alpha)}{2\alpha - 1}$ for $0.5 < \alpha_j < 1$. □

Proof of Proposition 8. The revenue function can be rewritten as

$$\Pi_4(p_1, q_1) = \frac{1}{\bar{c}}\frac{p_1}{q_1}\left\{\frac{1}{\alpha^3}\min\left\{8\alpha^3 v q_1 - 8p_1, \bar{c}\right\} - \left(\frac{1}{\alpha^3} - \frac{1}{\alpha^2}\right)\min\left\{\frac{8p_1(1-\alpha)}{2\alpha-1}, \bar{c}\right\}\right.$$
$$\left. - \left(\frac{1}{\alpha^2} - \frac{1}{\alpha}\right)\min\left\{\frac{4p_1(1-\alpha)}{2\alpha-1}, \bar{c}\right\} - \left(\frac{1}{\alpha} - 1\right)\min\left\{\frac{2p_1(1-\alpha)}{2\alpha-1}, \bar{c}\right\}\right\}.$$

First, in the optimal solution, there will be some customer types that choose the highest quality product (by assumption), which implies

$$8\alpha^3 v q_1 - 8p_1 \geq \bar{c} > \frac{8p_1(1-\alpha)}{2\alpha-1}. \tag{B.1}$$

Therefore, the revenue function can be written as

$$\Pi_4(p_1, q_1) = \frac{1}{\bar{c}}\frac{p_1}{q_1}\left\{\frac{1}{\alpha^3}\min\left\{8\alpha^3 v q_1 - 8p_1, \bar{c}\right\} - \left(\frac{1}{\alpha^3} - \frac{1}{\alpha^2}\right)\frac{8p_1(1-\alpha)}{2\alpha-1}\right.$$
$$\left. - \left(\frac{1}{\alpha^2} - \frac{1}{\alpha}\right)\frac{4p_1(1-\alpha)}{2\alpha-1} - \left(\frac{1}{\alpha} - 1\right)\frac{2p_1(1-\alpha)}{2\alpha-1}\right\}.$$

Assume $(\bar{p}_1, \bar{q}_1)$ is a global maximizer of the function above with $8\alpha^3 v\bar{q}_1 - 8\bar{p}_1 > \bar{c}$. Since $\Pi_4(\bar{p}_1, \bar{q}_1) \geq \Pi_4(0, \bar{q}_1) \geq 0$, it can easily be shown that $\Pi_4(\bar{p}_1, \bar{q}_1 - \varepsilon) \geq \Pi_4(\bar{p}_1, \bar{q}_1)$ for a positive ε. Hence, $8\alpha^3 v q_1^* - 8p_1^* = \bar{c}$ where (p_1^*, q_1^*) is the optimal base price, base quality level couple. Using this equality, the revenue function can be rewritten as

$$\Pi_4(p_1) = \frac{1}{\bar{c}}\left\{\frac{8\alpha^3 v p_1}{\bar{c} + 8p_1}\left[\frac{1}{\alpha^3}\bar{c} - \frac{8p_1(1-\alpha)^2}{(2\alpha-1)\alpha^3} - \frac{4p_1(1-\alpha)^2}{(2\alpha-1)\alpha^2} - \frac{2p_1(1-\alpha)^2}{(2\alpha-1)\alpha}\right]\right\},$$

where $p_1 \in \left[0, \dfrac{\bar{c}(2\alpha-1)}{8(1-\alpha)}\right)$. The upper bound is found by using (B.1). Then,

$$p_1^* = \underset{0 \le p_1 < \frac{\bar{c}(2\alpha-1)}{8(1-\alpha)}}{\operatorname{argmax}} \left\{ \frac{8\alpha^3 v p_1}{\bar{c}+8p_1} \left[\frac{1}{\alpha^3}\bar{c} - \frac{8p_1(1-\alpha)^2}{(2\alpha-1)\alpha^3} - \frac{4p_1(1-\alpha)^2}{(2\alpha-1)\alpha^2} - \frac{2p_1(1-\alpha)^2}{(2\alpha-1)\alpha} \right] \right\},$$

$\Pi_4(p_1)$ has a global maximum in $\left[0, \frac{\bar{c}(2\alpha-1)}{8(1-\alpha)}\right)$ and the first order conditions give

$$p_1^* = \{p_1 > 0 \mid \bar{c}^2(2\alpha-1) = 4p_1(\bar{c}+4p_1)(1-\alpha)(4-2\alpha-\alpha^2)\}.$$

The positive root of p_1^* is

$$p_1^* = \frac{-\alpha^3\bar{c} - \alpha^2\bar{c} + 6\alpha\bar{c} - 4\bar{c}}{8(\alpha^3+\alpha^2-6\alpha+4)} + \frac{\sqrt{\alpha^6\bar{c}^2 + 2\alpha^5\bar{c}^2 - 3\alpha^4\bar{c}^2 - 8\alpha^2\bar{c}^2 + 8\alpha\bar{c}^2}}{8(\alpha^3+\alpha^2-6\alpha+4)},$$

which is equivalent to (3.3). □

Proof of Proposition 9.

1. $v\bar{q}\left(2\alpha - \dfrac{2\alpha-1}{\alpha}\right) \le \bar{c}$ implies $\bar{q} \le \dfrac{2\bar{c}}{v}$. Hence, the revenue of the one-quality-level case is $\dfrac{v^2\bar{q}}{4\bar{c}}$. Under the given conditions, the revenue of the two-quality-level case is $\dfrac{v^2\bar{q}(2\alpha-1)}{2\alpha\bar{c}}$. For positive v and $\bar{c}$,

$$\frac{v^2\bar{q}(2\alpha-1)}{2\alpha\bar{c}} > \frac{v^2\bar{q}}{4\bar{c}}$$

when $\dfrac{2}{3} < \alpha < 1$.

2. Let Π_2^* be the optimal revenue when two quality levels are offered, with the optimal base quality level $\bar{q}$ (optimal base quality level has to be equal to $\bar{q}$ when the condition in the first point is satisfied), and the optimal base price p_2^*. Let $\bar{\Pi}_2$ be the revenue when the service provider offers only the second quality level with price $2p_2^*$. Then $\bar{\Pi}_2 > \Pi_2^*$, since the second quality level still gives nonnegative utility to the customer types that chose the first quality level when the first quality level is present, and they pay more than before. Now, assume

that the new highest base quality level is $2\alpha\bar{q}$, and follow the first item of this proposition.

3. The same idea follows.

□

Proof of Proposition 10. First, we write down the first order conditions for p_1:

$$p_1^e = \frac{p_2^e q_1}{2q_2}.$$

Then, we separate the second provider's problem into two cases, $(P1)$ and $(P2)$, and solve both.

$$\begin{aligned} (P1): \underset{p_2}{\text{maximize}} \quad & \frac{1}{\bar{c}}\frac{p_2}{q_2}\left[vq_2 - p_2 - \frac{p_2 q_1 - p_1 q_2}{q_2 - q_1}\right] \\ \text{subject to} \quad & p_2 \geq vq_2 - \bar{c}. \end{aligned}$$

By taking the derivative of the revenue function and then plugging p_1^c for p_1, we reach

$$p_2^x = \max\left\{vq_2 - \bar{c}, \frac{2vq_2(q_2 - q_1)}{4q_2 - q_1}\right\}.$$

$$\begin{aligned} (P2): \underset{p_2}{\text{maximize}} \quad & \frac{1}{\bar{c}}\frac{p_2}{q_2}\left[\bar{c} - \frac{p_2 q_1 - p_1 q_2}{q_2 - q_1}\right] \\ \text{subject to} \quad & p_2 \leq vq_2 - \bar{c}. \end{aligned}$$

By taking the derivative of the revenue function and then plugging p_1^c for p_1, we reach

$$p_2^y = \min\left\{vq_2 - \bar{c}, \frac{2\bar{c}(q_2 - q_1)}{3q_1}\right\}.$$

If both $(P1)$ and $(P2)$ give a negative revenue in the optimal solution, then, $p_2^e = 0$ which generates zero revenue; otherwise, $p_2^e = \underset{p_2 \in \{p_2^x, p_2^y\}}{\text{argmax}}\, R(p_2)$ □

Appendix C

Appendix to Chapter 4

C.1 Computational Complexity of Cardinality Matching

In complexity theory, computational problems are categorized in terms of their inherent difficulty, usually in connection with the time it takes to find a solution (Papadimitriou, 1994). While some problems can be solved quickly, with algorithms that run in polynomial time, other problems cannot. Problems that can be solved with polynomial time algorithms are considered tractable in the sense that the number of arithmetic steps it takes to solve a problem instance increases as a polynomial function of the size of the problem. For instance, the assignment problem of matching treated and controls units to minimize the total sum of covariate distances between matched units (as in Rosenbaum, 1989) is considered tractable because it has a worst-case time bound of $O(U^3)$ where U is the number of units available before matching (Kuhn, 1955; Bertsekas, 1981; Papadimitriou and Steiglitz, 1982). General IP and MIP problems are NP-hard in the sense that no polynomial time algorithm has been found to solve any problem in their general class so far.

Cardinality matching (4.7)-(4.9) is an IP problem and, although a polynomial time algorithm has not been found to solve this specific problem, from a user standpoint the time it takes in practice to solve a typical instance of this problem is comparable to the time it takes to solve the assignment problem. In the cardinality matching problem the constraint matrix defined by (4.8)-(4.9) is not totally unimodular (meaning that the feasible region it defines is not an integral polyhedron, so

the problem cannot be solved by relaxing the original problem and solving a linear program as in the assignment problem), however there is much structure in the constraints (4.8)-(4.9) so it can be solved in reasonable time with modern optimization solvers.

C.2 Matching to Minimize the Variance of a Difference-in-Means Effect Estimator

Consider the effect estimator (4.5) and calculate its variance

$$\sum_{i\in\mathcal{I}} \text{Var}(\hat{\delta}) = \frac{\sigma^2}{I^2} \sum_{i\in\mathcal{I}} \left(1 + \frac{1}{\kappa_i}\right). \tag{C.1}$$

Ideally, within the matching framework of (4.6), we would define the information content $\mathbb{I}$ as the inverse of this variance; however, since the number of matched pairs I is also a decision variable, the resulting optimization problem is very complicated. A simplification is to fix I by matching all the treated units with a variable $1 : \kappa_{\mathcal{C}}$ ratio. For fixed I, the variance of the effect estimator is proportional to

$$\sum_{i\in\mathcal{I}} \text{Var}(\hat{\delta}) \propto \sum_{i\in\mathcal{I}} \frac{1}{\kappa_i}. \tag{C.2}$$

Put $\ell^{(\kappa_i)} = \frac{1}{\kappa_i}$. Since maximizing the inverse of the variance is equivalent to minimizing the variance, the problem we want to solve can be written as

$$\min_{\boldsymbol{m},\boldsymbol{n}} \{\mathbb{V}(\boldsymbol{m},\boldsymbol{n}) : (\boldsymbol{m},\boldsymbol{n}) \in \mathcal{M} \cap \mathcal{B}\} \tag{C.3}$$

where

$$\mathbb{V}(\boldsymbol{m},\boldsymbol{n}) = \sum_{t\in\mathcal{T}} \ell^{(n_t)}, \tag{C.4}$$

$$\begin{aligned} \mathcal{M} = \Bigg\{ &\sum_{c\in\mathcal{C}} m_{tc} = n_t,\ t \in \mathcal{T}; n_t \le \kappa_{\mathcal{C}},\ t \in \mathcal{T}; \sum_{t\in\mathcal{T}} m_{tc} \le 1,\ c \in \mathcal{C}; \\ &m_{tc} \in \{0,1\},\ t \in \mathcal{T}, c \in \mathcal{C}; n_t \ge 1,\ t \in \mathcal{T} \Bigg\}, \end{aligned} \tag{C.5}$$

$$\mathcal{B} = \left\{ -\varepsilon_p T \leq \sum_{t\in\mathcal{T}} x_{t,p} - \sum_{c\in\mathcal{C}} \left(\sum_{t\in\mathcal{T}} m_{tc}\ell^{(n_t)} \right) x_{c,p} \leq \varepsilon_p T, \right. \\ \left. p \in \mathcal{P}, m_{tc} \in \{0,1\},\ t \in \mathcal{T}, c \in \mathcal{C}; n_t \geq 1,\ t \in \mathcal{T} \right\}. \tag{C.6}$$

Note that the set $\mathcal{B}$ above is only written for mean balancing constraints. In a similar way to the model described in Chapter 4, $\ell^{(n_t)}$ and $m_{tc}\ell^{(n_t)}$ have to be linearized using $m_t^{(r)}$s ($t \in \mathcal{T}, r \in \{2,3,\ldots,\kappa_{\mathcal{C}}-1\}$). The only difference is we do not need $m_t^{(1)}$, it is set to 1 since all treated units are forced to be matched with at least one control unit in this formulation. Therefore,

$$m_t^{(r)} \leq n_t - \sum_{s=2}^{r-1} m_t^{(s)} - 1,\ t \in \mathcal{T}, r \in \{2,\ldots,\kappa_{\mathcal{C}-1}\} \tag{C.7}$$

$$\kappa_{\mathcal{C}} m_t^{(r)} \geq n_t - \sum_{s=2}^{r-1} m_t^{(s)} - 1,\ t \in \mathcal{T}, r \in \{2,\ldots,\kappa_{\mathcal{C}-1}\}, \tag{C.8}$$

and

$$\begin{aligned} w_t &:= \ell^{(n_t)} \\ &= 1 + \sum_{s=2}^{\kappa_{\mathcal{C}}-1} \left(\frac{1}{s} - \frac{1}{s-1} \right) m_t^{(s)} + \left(\frac{1}{\kappa_{\mathcal{C}}} - \frac{1}{\kappa_{\mathcal{C}}-1} \right) \left(n_t - \sum_{s=2}^{\kappa_{\mathcal{C}}-1} m_t^{(s)} - 1 \right). \end{aligned} \tag{C.9}$$

To linearize $m_{tc}\ell^{(n_t)}$, define $q_{tc} = m_{tc}\ell^{(n_t)}$ which can be formulated as

$$q_{tc} \leq m_{tc},\ t \in \mathcal{T}, c \in \mathcal{C} \tag{C.10}$$

$$q_{tc} \leq w_t,\ t \in \mathcal{T}, c \in \mathcal{C} \tag{C.11}$$

$$q_{tc} \geq w_t - (1 - m_{tc}),\ t \in \mathcal{T}, c \in \mathcal{C}. \tag{C.12}$$

Lastly, we define $w_c = \sum_{t\in\mathcal{T}} q_{tc}, c \in \mathcal{C}$, and rewrite mean balancing constraints

$$-\varepsilon_p T \leq \sum_{t\in\mathcal{T}} x_{t,p} - \sum_{c\in\mathcal{C}} w_c x_{c,p} \leq \varepsilon_p T, p \in \mathcal{P}. \tag{C.13}$$

C.3 Matching with a Flexible $1 : \kappa_{\mathcal{C}}/\kappa_{\mathcal{T}} : 1$ Ratio

First, let us define g_t and g_c

$$g_t = \begin{cases} h^{(n_t)} & \text{if } n_t \geq 2 \\ \sum_{c \in \mathcal{C}} m_{tc} \dfrac{h^{(n_c)}}{n_c} & \text{if } n_t \leq 1 \end{cases} \qquad t \in \mathcal{T}, \tag{C.14}$$

$$g_c = \begin{cases} h^{(n_c)} & \text{if } n_c \geq 2 \\ \sum_{t \in \mathcal{T}} m_{tc} \dfrac{h^{(n_t)}}{n_t} & \text{if } n_c \leq 1 \end{cases} \qquad c \in \mathcal{C}. \tag{C.15}$$

Then, the problem can be formulated as

$$\mathbb{I}(\boldsymbol{m}, \boldsymbol{n}) = \sum_{t \in \mathcal{T}} g_t, \tag{C.16}$$

$$\mathcal{M} = \mathcal{M}_1 \cap \mathcal{M}_2 \cap \mathcal{M}_3 \tag{C.17}$$

where

$$\mathcal{M}_1 = \left\{ \sum_{c \in \mathcal{C}} m_{tc} = n_t,\ t \in \mathcal{T}; n_t \leq \kappa_{\mathcal{C}},\ t \in \mathcal{T}; m_{tc} \in \{0,1\}, n_t \geq 0,\ t \in \mathcal{T} \right\}, \tag{C.18}$$

$$\mathcal{M}_2 = \left\{ \sum_{t \in \mathcal{T}} m_{tc} = n_c,\ c \in \mathcal{C}; n_c \leq \kappa_{\mathcal{T}},\ c \in \mathcal{C}; m_{tc} \in \{0,1\}, n_c \geq 0,\ c \in \mathcal{C} \right\}, \tag{C.19}$$

$$\begin{aligned} \mathcal{M}_3 = \{ (n_t - 1)(n_c - 1) m_{tc} = 0,\ t \in \mathcal{T}, c \in \mathcal{C}; m_{tc} \in \{0,1\},\ t \in \mathcal{T}, c \in \mathcal{C}; \\ n_t \geq 0,\ t \in \mathcal{T}; n_c \geq 0,\ c \in \mathcal{C} \}, \end{aligned} \tag{C.20}$$

and

$$\begin{aligned} \mathcal{B} = \Bigg\{ -\varepsilon_p \sum_{t \in \mathcal{T}} g_t \leq \sum_{t \in \mathcal{T}} g_t x_{t,p} - \sum_{c \in \mathcal{C}} g_c x_{c,p} \leq \varepsilon_p \sum_{t \in \mathcal{T}} g_t,\ p \in \mathcal{P}; \\ m_{tc} \in \{0,1\},\ t \in \mathcal{T}, c \in \mathcal{C}; n_t \geq 0,\ t \in \mathcal{T}; n_c \geq 0,\ c \in \mathcal{C} \Bigg\}. \end{aligned} \tag{C.21}$$

Note that g_t, g_c and the constraint set $\mathcal{M}_3$ have to be written in linear form. Let us define $m_c^{(r)}$, $w_c^{(1)}$ and $w_c^{(2)}$ analogous to $m_t^{(r)}$, $w_t^{(1)}$, and $w_t^{(2)}$. The decision variable $m_c^{(r)}$ is equal to 1 if control unit c is matched with at least r number of treated units, and 0 otherwise ($c \in \mathcal{C}, r \in \{1, \ldots, \kappa_{\mathcal{T}} - 1\}$). With linear constraints

$$m_c^{(r)} \leq n_c - \sum_{s=1}^{\kappa_{\mathcal{T}}-1} m_c^{(s)}, \qquad c \in \mathcal{C}, r \in \{1, \ldots, \kappa_{\mathcal{T}} - 1\} \tag{C.22}$$

$$\kappa_{\mathcal{T}} m_c^{(r)} \geq n_c - \sum_{s=1}^{\kappa_{\mathcal{T}}-1} m_c^{(s)}. \qquad c \in \mathcal{C}, r \in \{1, \ldots, \kappa_{\mathcal{T}} - 1\} \tag{C.23}$$

Using $m_c^{(r)}$,

$$w_c^{(1)} := \sum_{s=1}^{\kappa_{\mathcal{T}}-1} \left(h^{(s)} - h^{(s-1)}\right) m_c^{(s)} + \left(h^{(\kappa_{\mathcal{T}})} - h^{(\kappa_{\mathcal{T}}-1)}\right) \left(n_c - \sum_{s=1}^{\kappa_{\mathcal{T}}-1} m_c^{(s)}\right), \tag{C.24}$$

$$w_c^{(2)} := \sum_{s=1}^{\kappa_{\mathcal{T}}-1} \left(\frac{h^{(s)}}{s} - \frac{h^{(s-1)}}{s-1}\right) m_c^{(s)} + \left(\frac{h^{(\kappa_{\mathcal{T}})}}{\kappa_{\mathcal{T}}} - \frac{h^{(\kappa_{\mathcal{T}}-1)}}{\kappa_{\mathcal{T}}-1}\right) \left(n_c - \sum_{s=1}^{\kappa_{\mathcal{T}}-1} m_c^{(s)}\right), \tag{C.25}$$

where $\frac{h^{(0)}}{0}$ is set to 0.

Now, we can rewrite g_t and g_c as

$$g_t = \begin{cases} w_t^{(1)}, & \text{if } m_t^{(2)} = 1 \\ \sum_{c \in \mathcal{C}} m_{tc} w_c^{(2)}, & \text{if } m_t^{(2)} = 0 \end{cases}, \qquad t \in \mathcal{T} \tag{C.26}$$

$$g_c = \begin{cases} w_c^{(1)}, & \text{if } m_c^{(2)} = 1 \\ \sum_{t \in \mathcal{T}} m_{tc} w_t^{(2)}, & \text{if } m_c^{(2)} = 0 \end{cases}. \qquad c \in \mathcal{C} \tag{C.27}$$

The expressions $m_{tc} w_c^{(2)}$ and $m_{tc} w_t^{(2)}$ are still not in linear form; therefore, we define two new sets of decision variables $u_{tc} := m_{tc} w_c^{(2)}$ and $v_{tc} := m_{tc} w_t^{(2)}$, and formulate them in the following

way:

$$u_{tc} \leq m_{tc}, \qquad t \in \mathcal{T}, c \in \mathcal{C} \tag{C.28}$$

$$u_{tc} \leq w_c^{(2)}, \qquad t \in \mathcal{T}, c \in \mathcal{C} \tag{C.29}$$

$$u_{tc} \geq w_c^{(2)} - (1 - m_{tc}), \qquad t \in \mathcal{T}, c \in \mathcal{C} \tag{C.30}$$

$$v_{tc} \leq m_{tc}, \qquad t \in \mathcal{T}, c \in \mathcal{C} \tag{C.31}$$

$$v_{tc} \leq w_t^{(2)}, \qquad t \in \mathcal{T}, c \in \mathcal{C} \tag{C.32}$$

$$v_{tc} \geq w_t^{(2)} - (1 - m_{tc}). \qquad t \in \mathcal{T}, c \in \mathcal{C} \tag{C.33}$$

As the last step, we write g_t and g_c using conditional constraints

$$g_t \leq h^{(\kappa_{\mathcal{C}})} m_t^{(2)} + \sum_{c \in \mathcal{C}} u_{tc}, \qquad t \in \mathcal{T} \tag{C.34}$$

$$g_t \leq 1 - m_t^{(2)} + w_t^{(1)}, \qquad t \in \mathcal{T} \tag{C.35}$$

$$g_t \geq \sum_{c \in \mathcal{C}} u_{tc} - \kappa_{\mathcal{C}} m_t^{(2)}, \qquad t \in \mathcal{T} \tag{C.36}$$

$$g_t \geq w_t^{(1)} - \left(1 - m_t^{(2)}\right) h^{(\kappa_{\mathcal{C}})}, \qquad t \in \mathcal{T} \tag{C.37}$$

$$g_c \leq h^{(\kappa_{\mathcal{T}})} m_c^{(2)} + \sum_{t \in \mathcal{T}} v_{tc}, \qquad c \in \mathcal{C} \tag{C.38}$$

$$g_c \leq 1 - m_c^{(2)} + w_c^{(1)}, \qquad c \in \mathcal{C} \tag{C.39}$$

$$g_c \geq \sum_{t \in \mathcal{T}} v_{tc} - \kappa_{\mathcal{T}} m_c^{(2)}, \qquad c \in \mathcal{C} \tag{C.40}$$

$$g_c \geq w_c^{(1)} - \left(1 - m_c^{(2)}\right) h^{(\kappa_{\mathcal{T}})}. \qquad c \in \mathcal{C} \tag{C.41}$$

Finally, constraint set $\mathcal{M}_3$ can be written in linear form as

$$m_t^{(2)} + m_c^{(2)} \leq 2 - m_{tc}. \qquad t \in \mathcal{T}, c \in \mathcal{C} \tag{C.42}$$

C.4 Running Times

Here we present the results of a small simulation study to provide a sense of the running times of the proposed methods. In the original data, we have 694 clusters with one treated (green) building and multiple control buildings. In the simulation study, we randomly selected 100, 500 and 2000 clusters with replacement, and for each of these number of clusters we tested our method with different number of covariates: 5, 10, 20, and 50. For covariate sizes 5 and 10 we randomly selected the covariates from our covariate set, and for covariate sizes 20 and 50 (since there are not that many covariates to begin) we included interactions of covariates that are fairly independent of each other. As in the actual study, we divided each of the optimization problem into 10 subproblems using exact matching constraints (as explained in Appendix C.5) and solved each of them in parallel. We gave a time limit of 60 minutes to each of the problems. The results are presented in the following tables.

Table C.1(a) presents the running times and optimality gaps for the method used in the actual study; this is, matching with a variable $1:4$ ratio with the weighted balancing constraints (4.12). In each cell of the table, the first row shows the running time of the optimization problem, which is the maximum running time of the 10 subproblems. (Since these problems are run in parallel, the total running time of the optimization portion is the maximum running time of all the subproblems. A running time greater than 60 minutes indicates that an optimal solution would be found after the reported duration.) The second row shows the optimality gaps in terms of the maximum effective sample size reached in the given time limit and the tightest upper bound found by the solver after the branch and bound procedure also within the time limit. One can evaluate how close the provided solution is to the theoretical solution from these numbers. In the table, we observe that one can obtain relatively small optimality gaps within the time limit for samples of size $(n_t, n_c) \approx (500, 5000)$, and for larger sample sizes if the number of covariates is smaller than 10.

As described in Appendix C.5, one way to decrease the complexity of the problem, and therefore to reduce computing times is by omitting the weights in the balancing constraints. Table C.1(b)

presents the results for this approach with a time limit of 15 minutes for each of the optimization problems. Within this time limit, it is possible to find solutions with a small optimality gap for all the instances in the table.

Table C.1: Running times and optimality gaps for matching for different combinations of sample sizes and number of covariates. The running times are reported in minutes and the optimality gaps appear in terms of two numbers: the best solution found within the given time limit and the bounding (perhaps infeasible) solution found also within the time limit.

(a) *Matching with a variable* $1:4$ *ratio with the weighted balancing constraints* (4.12)

Number of units	Number of covariates			
(n_t, n_c)	5	10	20	50
(100, 1228)	0.1 130.3–130.3	60.0 123.1–123.8	60.0 77.2–77.5	0.1 62.2–62.2
(500, 5237)	60.0 673.4–690.6	60.0 655.2–678.4	60.0 617.8–646.8	60.0 535.7–574.9
(1000, 10806)	60.0 1355.5–1405.8	60.0 1332.0–1391.4	60.0 1274.4–1365.1	60.0 1165.8–1286.8
(2000, 21190)	60.0 2700.3–2876.7	60.0 2630.0–2869.2	60.0 2468.7–2864.6	60.0 2244.4–2750.7

(b) *Matching with a variable* $1:4$ *ratio with the unweighted balancing constraints* (4.9)

Number of units	Number of covariates			
(n_t, n_c)	5	10	20	50
(100, 1228)	0.1 132.4–132.4	0.1 127.0–127.0	0.1 78.9–78.9	0.1 66.2–66.2
(500, 5237)	0.1 700.5–700.5	0.2 692.1–692.1	15.0 658.4–658.4	15.0 585.8–592.5
(1000, 10806)	1.7 1405.3–1405.3	0.3 1397.2–1397.2	15.0 1366.0–1367.9	15.0 1288.6–1293.1
(2000, 21190)	0.5 2841.1–2841.1	15.0 2827.8–2827.9	15.0 2781.0–2782.7	15.0 2677.0–2683.5

C.5 Devices for Speed

One tactic for more quickly solving the previous matching problems is exact matching for nominal covariates of prognostic relevance or which are to be used for subgroup analyses. Let $x_{\cdot,p}$ be a nominal covariate taking integer values $\tilde{n} \in \mathcal{N} \subset \mathbb{N}$. To match exactly for $x_{\cdot,p}$, one possibility is to

include the constraint

$$\sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} m_{tc}\left|\mathbb{1}_{x_{t,p}=\tilde{n}} - \mathbb{1}_{x_{c,p}=\tilde{n}}\right| = 0,\ \forall\tilde{n}\in\mathcal{N}, \tag{C.43}$$

where $\mathbb{1}$ is the indicator function. Exact matching constraints reduce the feasible region considerably and therefore the optimal solution is found faster. Another possibility to match exactly for $x_{\cdot,p}$ is to divide the dataset into smaller, mutually exclusive and collectively exhaustive pieces based on the categories of $\mathcal{N}$ and solve a matching problem for each piece in parallel. If the problem is run on a machine with multiple processors and/or multiple cores, each subproblem can be assigned to be solved independently by a processing unit. The default settings in `R` do not use all the cores available in the machine running the code; however, there are some packages available to create a parallel backend so that independent subproblems can be solved simultaneously on different processing units (see, for instance, Weston and Calaway, 2014 for the `R` packages `doParallel` and `foreach`).

Other tactics that can be used to attain computational speedups include simplifying the matching problem by eliminating the harmonic mean weights from the balancing constraints (but not the objective function) or using Yoon (2009)'s entire number to determine the matching ratio for each unit before matching (Zubizarreta, 2012). However, we do not recommend the first of these approaches because it results in an inconsistency between the balance criteria used to assess the quality of the match and the balance criteria needed for unbiased estimation with an estimator that uses the harmonic mean weights. Also, we are not enthusiastic about the second approach because it requires that one estimate the propensity score in order to calculate the entire number.

C.6 Description of the Matched Sample

Table C.2 below describes the samples of green buildings before matching, after matching and of those green buildings that were left out from the matched analyses due to lack of good controls. We observe that the sample of matched green buildings is very similar to that of all the green buildings (after all, only 19 green buildings were unmatched and left out from the analyses). Among others, the unmatched buildings are larger on average, have better quality (are all in class A and have a higher proportion of amenities), are not very old, less of them are renovated, and have high stories.

Table C.2: Means and sizes of the samples of green buildings before matching ("All"), after matching ("Matched") and of those green buildings that were left out from the analyses due to lack of good controls ("Unmatched").

Covariate	Sample		
	All	Matched	Unmatched
Building size	0.324	0.327	0.520
Building class A	0.794	0.780	1.000
Building class B	0.195	0.207	0.000
Building class C	0.012	0.012	0.000
Net contract	0.058	0.059	0.053
Employment growth	0.035	0.037	-0.028
Employment growth missing	0.009	0.009	0.000
Age $\leq$10 years	0.143	0.140	0.158
Age 11-20 years	0.241	0.234	0.316
Age 21-30 years	0.434	0.425	0.526
Age 31-40 years	0.111	0.120	0.000
Age >40 years	0.059	0.066	0.000
Age missing	0.013	0.014	0.000
Renovated	0.210	0.213	0.158
Stories low	0.463	0.455	0.211
Stories intermediate	0.267	0.264	0.263
Stories high	0.271	0.281	0.526
Stories missing	0.000	0.000	0.000
Amenities	0.718	0.711	0.895
Sample size	694	675	19