



## Management Science

Publication details, including instructions for authors and subscription information:  
<http://pubsonline.informs.org>

### The Effects of Autoscaling in Cloud Computing

Amir Fazli, Amin Sayedi, Jeffrey D. Shulman

To cite this article:

Amir Fazli, Amin Sayedi, Jeffrey D. Shulman (2018) The Effects of Autoscaling in Cloud Computing. Management Science 64(11):5149-5163. <https://doi.org/10.1287/mnsc.2017.2891>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact [permissions@informs.org](mailto:permissions@informs.org).

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2018, INFORMS

Please scroll down for article—it is on subsequent pages






With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.

For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

# The Effects of Autoscaling in Cloud Computing

Amir Fazli,<sup>a</sup> Amin Sayedi,<sup>a</sup> Jeffrey D. Shulman<sup>a</sup>

<sup>a</sup>Foster School of Business, University of Washington, Seattle, Washington 98195

Contact: fazli@uw.edu,  <http://orcid.org/0000-0002-1506-271X> (AF); aminsa@uw.edu,  <http://orcid.org/0000-0003-2409-2387> (AS); jshulman@uw.edu,  <http://orcid.org/0000-0001-5288-3421> (JDS)

Received: September 21, 2016

Revised: March 13, 2017; June 5, 2017

Accepted: June 22, 2017

Published Online in Articles in Advance:  
January 19, 2018

<https://doi.org/10.1287/mnsc.2017.2891>

Copyright: © 2018 INFORMS

**Abstract.** Web-based firms often rely on cloud-based computational resources to serve customers, but the number of customers they will serve is rarely known at the time of product launch. A recent innovation in cloud computing, known as autoscaling, allows companies to automatically scale their computational load up or down to match customer demand. We build a game theory model to examine how autoscaling will affect firms' decisions to enter a new market and the resulting equilibrium prices, profitability, and consumer surplus. The model produces novel results depending on the likelihood of a firm's success in the new market, differentiation among potential entrants, and the cost of computational capacity. For instance, in contrast to previous capacity commitment models with demand certainty, we show that autoscaling can mitigate price competition if the likelihood of a firm's success in the market is moderate and the cost of capacity is sufficiently low. This is because without autoscaling, the firms' uncertainty about demand would lead to excessive computational capacities and thus aggressive price competition. We also find that when the likelihood of success is sufficiently high, autoscaling facilitates entry for one firm yet deters a second firm from simultaneously entering the market.

**History:** Accepted by Juanjuan Zhang, marketing.

**Supplemental Material:** The online appendix is available at <https://doi.org/10.1287/mnsc.2017.2891>.

**Keywords:** entrepreneurship marketing • competitive strategy • marketing • pricing • cloud computing

## 1. Introduction

Consider an entrepreneur or firm deciding whether to launch a web service in a new market. Getting started requires an investment of time and money into research, development, legal, and other starting expenses prior to knowing whether the new product will ever succeed. Furthermore, web- and mobile-based firms rely on computational capacity to serve customers. Every interaction a customer has with an application such as a page load, data transfer, and object viewing requires computational resources. This is particularly relevant to companies launching a software as a service (SaaS), an industry estimated at \$49 billion and expected to grow to \$67 billion by 2018 (Columbus 2015). For any company relying on computational resources, cloud computing allows the company to outsource the server setup and maintenance to a cloud provider.

While cloud computing technology has long allowed for the outsourcing of computational costs, firms who ran their tasks in the cloud often needed to decide, and precommit to, their computational capacity at the time of purchasing the cloud service. As such, because of the unpredictable nature of demand for firms entering a new market, the prepurchased capacity may be excessive or insufficient for the traffic. This issue is especially important in the case of entrepreneurs launching a web start-up. For example, BeFunky, an online photo editing start-up, was featured on a popular social media

site three weeks after launch and saw 30,000 visitors in three hours, crashing their servers (Nickelsburg 2016). Such events happen regularly enough that there is a term for it: *the slashdot effect*, which Klems et al. (2008) describe as occurring when a start-up is featured on a popular network, resulting in a significant increase in traffic load and causing the firm's servers to slow down or crash. Though capacity can later be increased, the missed demand can be costly. As Amazon CEO Jeff Bezos describes, start-ups face a serious challenge when choosing computational capacity:<sup>1</sup>

And you do face this issue [demand uncertainty] whenever you have a startup company. You want to be prepared for lightning to strike because if you're not that generates a big regret. If lightning strikes and you weren't ready for it, that's kind of hard to live with. At the same time, you don't want to prepare your physical infrastructure to hubris levels either in the case that lightning doesn't strike.

To address this challenge, cloud providers such as Amazon, Microsoft, and Google have begun to offer *autoscaling*, a feature that allows firms to scale their computational capacity up or down automatically in real time. Using autoscaling, firms launching a new web-based service can maintain application availability and scale their computational capacity for serving consumers without having to make capacity precommitments.

For companies, autoscaling means having just the right number of servers required for meeting the demand at any point in time, which can provide an attractive solution to handling uncertain demand in new markets. Autoscaling is offered with no additional fees and has been celebrated as one of the most beneficial features of cloud computing.

In this paper, we develop an analytical model to examine how the emergence of autoscaling in cloud computing will affect web-based firms' decisions regarding market entry and prices. In particular, despite popular belief, we identify conditions for when autoscaling negatively affects market entry. In other words, fewer firms will enter in certain markets because of the advent of autoscaling. Autoscaling has several properties that make it unique from some previously explored areas in marketing and operations. In particular, autoscaling

- removes a capacity decision that otherwise has to be made prior to pricing;
- converts computational capacity costs from fixed costs to variable costs at the time of pricing;
- allows capacity to be set after the uncertainties regarding consumers' level of interest and competitors' strategies are resolved; however, autoscaling still
- preserves the uncertainty that exists at the time of making the entry decision.

Given these properties, the effects of autoscaling on company strategies cannot be addressed by prior research on capacity choices and demand uncertainty. In fact, our model and predictions diverge from prior literature substantively.

We uniquely incorporate these properties of autoscaling into a game theory model in which two horizontally differentiated firms have the option to enter a market upon incurring an entry cost. We compare a model in which firms choose computational capacity to a model in which firms can choose to adopt autoscaling in cloud computing. The model is constructed to address the following research questions:

1. How does autoscaling affect pricing of market entrants?
2. How does autoscaling affect firms' profits in the new market?
3. How does autoscaling affect market entry decisions?
4. How does autoscaling affect consumer surplus?

We explore the roles of several important market factors in determining the answer to each of the research questions. First, we model the *ex ante* likelihood of a successful product launch. As Griffith (2014) suggests, the value that a firm brings to a new market is unknown before market entry. In some markets, consumer needs are well known and established, therefore yielding a higher likelihood of successfully creating a product that matches consumer needs. For other markets, the consumer needs are less understood and

there is a lower likelihood of a successful venture. We show that the likelihood of a successful venture plays a critical role in determining how autoscaling affects equilibrium strategies and profits.

Secondly, we model competition among market entrants. As Burke and Hussels (2013) suggest, competition in new markets is a key factor in determining the performance of new products. In fact, without accounting for competition, autoscaling has a strictly positive impact on the profit of the firm introducing a new product. However, a conventional study of autoscaling for a monopoly does not capture the strategic interactions inherent to autoscaling. By analyzing a competitive game theory model, we capture these strategic effects and find autoscaling can actually decrease the profits of competing firms.

We identify three general effects of autoscaling: First, the *downside risk reducing* effect of autoscaling prevents firms from investing in excess capacity in case their product is not successful. Second, the *demand satisfaction* effect of autoscaling allows firms to fully serve the demand without facing insufficient capacity in case their product is successful. Finally, autoscaling also has a *competition intensifying* effect, which can result in lower prices and profitability if multiple firms enter the market.

In addressing the first research question, we find that autoscaling may increase or decrease average prices charged by competing firms. Existing economic theory would suggest that removing the capacity decision prior to pricing would result in a shift from a Cournot game to a Bertrand game, thereby decreasing prices (e.g., Kreps and Scheinkman 1983). However, our model shows that demand uncertainty coupled with the capacity decision can result in autoscaling increasing the average prices set by each firm. Intuitively, autoscaling has two opposing effects on price. On the one hand, autoscaling can intensify price competition among firms by removing capacity constraints and allowing firms to aggressively lower their price to attract more customers. On the other hand, while capacity costs are sunk in the absence of autoscaling and do not affect the firms' pricing decisions, with autoscaling each firm's total cost of capacity depends on the number of customers purchasing from that firm, thus affecting the firm's price: With autoscaling, firms have less incentive to reduce price and attract more customers, since serving every additional customer requires an additional cost of capacity. When the probability of success is not high, the likelihood of head-to-head competition between two successful firms is low with or without autoscaling and thus the former price decreasing effect becomes dominated by the latter price increasing effect. Thus, our model shows that the probability of success plays a critical role in the effect of removing capacity commitments on pricing strategies.

In answering the second research question, we find that when the probability of success is sufficiently low, autoscaling increases the firms' expected profits. Intuitively, removing the capacity commitment through autoscaling can have a positive impact on firms' profits by allowing for demand satisfaction without extra capacity. For a low probability of success, there is little chance of both firms achieving success and competing head to head. Thus, the competition intensifying effect of autoscaling is dominated by the demand satisfaction and downside risk reducing effects. On the other hand, autoscaling may also create a prisoner's dilemma situation where firms choose autoscaling, but autoscaling lowers their equilibrium profits. In particular, when entry costs are sufficiently small and the probability of success is moderately high, firms adopt autoscaling in equilibrium; however, their equilibrium profits would be higher if autoscaling was not available. This result is driven by the competition intensifying effect outweighing the demand satisfaction and downside risk reducing effects of autoscaling.

With regard to the third research question, we find that autoscaling can actually decrease market entry. Though we confirm common intuition that the likelihood of a market being served by at least one firm is improved with autoscaling, we find that, under certain conditions, entry by multiple firms will not occur because of autoscaling. The counterintuitive result occurs when entry costs are moderately high and there is a high probability that entrants will have a successful venture. In this region, the two firms anticipate autoscaling will heighten price competition after entry, and one firm, therefore, avoids entering the market altogether.

Finally, in addressing the fourth research question, we show that autoscaling may increase or decrease expected consumer surplus depending on the likelihood of a successful venture and entry costs. Given the fact that autoscaling guarantees companies have the capacity to serve consumers in case of high demand, thereby resolving issues such as the slashdot effect, one might expect that consumers benefit from autoscaling. However, our model shows that when entry costs are low enough such that both firms enter the market, autoscaling decreases expected consumer surplus if and only if the probability of a successful venture is moderate. Intuitively, autoscaling decreases consumer surplus in the region where firms would not have set constraining capacities without autoscaling. In this region, the competition intensifying effect of autoscaling is weak and autoscaling increases firms' prices, resulting in lower surplus for consumers.

The findings of this study provide implications for various players in new markets, including start-ups, cloud providers, and consumers. Our analysis informs firms entering web-based markets on how autoscaling affects competitive dynamics in pricing and entry. The

results suggest that a firm should consider not only the positive direct effect of autoscaling in reducing costs, but also the negative strategic effect in altering the nature of competition. By evaluating the likelihood of success, the cost of computational capacity, and entry costs, managers can use the findings from this study to determine whether autoscaling increases or decreases the likelihood of monopoly power over the new market. Our findings also inform cloud providers about how autoscaling affects not only the number of firms using the cloud, but also the number of servers each of those firms purchases. Our model provides insights for consumers on how autoscaling affects the prices charged in the market, showing that for high probabilities of success, average prices decrease with autoscaling and for lower probabilities of success they increase with autoscaling. We also find conditions for which autoscaling will decrease or increase consumer surplus, which can be used for consumer surplus maximizing policy design. To the best of our knowledge, this paper is the first to study the marketing aspects of cloud computing, and how it can affect prices and market entry. With the growing trend of adopting the cloud by firms, cloud computing is expected to become a major part of any business and this provides the field of marketing with a variety of related new topics to explore.

In addition to contributions to practice, our work contributes to economic theory regarding capacity commitments. Our benchmark model uniquely solves a capacity choice game with demand uncertainty and horizontal differentiation between sellers. Contrary to the previous literature, where capacity commitments lead to higher prices, we show that under demand uncertainty, capacity commitments (relative to autoscaling) can intensify the competition and cause lower equilibrium prices. We also uniquely study the effects of removing capacity commitments made under demand uncertainty (via autoscaling) on firms' market entry decisions.

The rest of this paper is organized in the following order. In Section 2, we review the literature related to our research problem. In Section 3, we introduce the model. In Section 4, we present the analysis of the model and derive the results. Finally, the discussion of our findings is presented in Section 5.

## 2. Literature Review

Academic research on cloud computing is still relatively new and most of the work done on this topic focuses on technological issues of the cloud (e.g., Yang and Tate 2012). The few existing business and economics studies of cloud computing have mainly offered conceptual theories and evidence from surveys and specific cases (e.g., Leavitt 2009, Walker 2009, Armbrust et al. 2009, Marston et al. 2011, Sultan 2011, Gupta et al. 2013). Our



paper is the first to model autoscaling in the cloud and study its effects on market entry.

In addition to cloud computing, our research is related to a number of topics in the literature. Previous research shows uncertainty in demand plays an important role in capacity and production decisions (e.g., Desai et al. 2007, Ferguson and Koenigsberg 2007, Desai et al. 2010). A body of literature (e.g., Van Mieghem and Dada 1999, Anupindi and Jiang 2008, Anand and Girotra 2007, Goyal and Netessine 2007) has studied postponing production decisions until after more demand information is revealed. These previous studies on postponing production separate the capacity decision and the production decision: the capacity decision is assumed to occur before demand is revealed and it is the production decision that can be postponed. With cloud computing, there is zero production cost after the firm chooses its computational capacity and autoscaling uniquely allows both capacity and production to simultaneously match with demand, resulting in findings different from the production postponement literature. For instance, Anupindi and Jiang (2008) find production postponement increases capacity investment and profitability, whereas we find when autoscaling may decrease equilibrium computational expenditures and when it may decrease profitability.

Che et al. (2010) allow for eliminating the capacity decision under demand uncertainty by considering a firm's decision between adopting a make-to-stock system, a backorder system, or a combination of both. However, a firm's decision of using autoscaling is conceptually different from the decision between make-to-stock and backorder production. Backorder production delays the time at which customers are served compared to a make-to-stock production. With autoscaling, on the other hand, firms avoid capacity decision under demand uncertainty, while satisfying the demand at the exact same time as they would have with pre-purchased capacity. Che et al. (2010) find it is optimal for firms to use a combination of make-to-stock and backorder production. However, autoscaling and purchasing fixed capacity simultaneously is not an optimal strategy in our context.

Our examination of how cloud computing with autoscaling affects a firm's entry decision also relates to the literature on market entry. A body of literature looks at the timing of entry and how an incumbent can deter entry (e.g., Spence 1977, Joshi et al. 2009, Milgrom and Roberts 1982, Ofek and Turut 2013). Narasimhan and Zhang (2000) consider firms' decisions on order of entry into markets with demand uncertainty. In contrast, our model examines simultaneous entry decisions by firms, such that both firms face equal demand uncertainty when making entry decisions. Our model adds to the entry literature by jointly considering both entry and capacity decisions,

such that each firm's decision to enter depends on the expected future capacity of both firms and whether capacity autoscales to demand.

We compare autoscaling with cases where firms commit to their computational capacity before pricing and realizing demand. This relates to other papers examining capacity commitments. Kreps and Scheinkman (1983) find that a Bertrand pricing game becomes a Cournot game when capacity is chosen prior to pricing. Reynolds and Wilson (2000), Swinney et al. (2011), and Nasser and Turcic (2015) consider extensions of Kreps and Scheinkman's model and replicate their finding that capacity commitments dampen competition. Our research expands this literature by considering demand uncertainty in a model of capacity commitment. In contrast, we are the first to show that, depending on the level of demand uncertainty, capacity commitments may indeed intensify the competition.

Autoscaling in cloud computing also has the effect of converting up-front capacity costs to variable costs that change with demand. Prior research examining the effect of converting fixed to variable costs via outsourcing (e.g., Shy and Stenbacka 2003, Buehler and Haucap 2006, Chen and Wu 2013) have found that prices and profitability rise with this conversion. However, these models do not allow for demand uncertainty. Accounting for demand uncertainty and endogenous upfront costs in our model provides novel insights on fixed versus variable capacity costs. In contrast to prior outsourcing literature, we show average prices can fall when autoscaling is used even in conditions for which entry is unaffected.

In summary, our paper uniquely compares market entry, pricing, and profitability between computing resources requiring capacity precommitments and cloud computing with autoscaling. Though previous research has separately examined demand uncertainty, the timing of capacity and pricing decisions, or the conversion of fixed costs to variable costs, our paper is unique in its comprehensive examination of the effect of autoscaling.

### 3. Model

We consider two symmetric firms who could potentially enter a particular web or mobile application market. To enter the market, the firms would incur a fixed entry cost,  $F$ . This cost includes starting expenses such as legal, research and development, and human capital investments. To model postentry competition, we adopt a discrete horizontal differentiation model (e.g., Narasimhan 1988, Iyer et al. 2005, Zhang and Katona 2012, Zhou et al. 2015) with three consumer segments, each consumer demanding at most one unit of the product.<sup>2</sup> Upon entry, each firm  $i$  will find a segment of consumers, segment  $i$  with  $i \in \{1, 2\}$ , who will buy from firm  $i$  if and only if the price  $p_i$  is below their

reservation value  $v_i$  and who will derive zero value from the competitor's product. This captures the reality that consumers vary in their taste preferences regardless of firm entry and that firms have idiosyncratic differences that will allow them to serve these tastes differently from each other upon successful entry. The size of each segment  $i$  is given by  $\alpha < \frac{1}{2}$  for  $i \in \{1, 2\}$ . The remaining  $1 - 2\alpha$  consumers are in segment 3 and are indifferent between firms, preferring to buy from the firm with the lowest price. The parameter  $\alpha$  can be interpreted as the extent to which consumers vary in their taste preferences. Note that  $\alpha$  also represents the level of competition in the market; for  $\alpha = \frac{1}{2}$ , segment 3 disappears, each firm gets a local monopoly, and there is no competition between the firms. As  $\alpha$  becomes smaller, the segment of consumers for which both firms compete grows and competition intensifies.

In the absence of autoscaling, the timing of the game is as follows.

*Stage 1.* Firms simultaneously decide whether or not to enter the market and thereby incur the cost  $F$ . We allow for uncertainty in whether a firm will find the venture successful in terms of whether  $v_i$  is high or low. We assume the ex ante probability of a firm finding success in this market is  $\gamma$ , which is common knowledge. In other words, if firm  $i$  enters the market,  $v_i$  is an i.i.d. draw from a binary distribution in which  $v_i = 1$  with probability  $\gamma$  and  $v_i = 0$  with probability  $1 - \gamma$ . This assumption reflects the idea that the value provided to customers is unclear for potential entrants. As Lilien and Yoon (1990) argue, the fit between market requirements and the offering of the new entrant is highly unpredictable and is critical to the success of the entrant. In a survey of 101 start-ups, it was reported that the number one reason for the failure of a start-up is the lack of market need for the offered product (Griffith 2014), suggesting that the value created for customers in a new market is unknown to many firms before entry. In an extension presented in the online appendix, we allow the low value condition to be  $v_i = V_i > 0$  and verify our results are robust to the assumption.<sup>3</sup>

To remark on the structure of demand and uncertainty, notice that our model setup has several desirable properties. In particular, it allows a firm to be uncertain about the size of the potential market and the effect of its price on realized demand: the firm may find itself a monopolist, the firm may find itself with very low demand (normalized to zero), or the firm may find itself competing head-to-head. Moreover, a firm's price relative to its competitor's is not the only source of demand uncertainty. Though one can explore alternative model specifications to capture these same properties, the current specification allows for tractability while uncovering a novel mechanism.

*Stage 2.* Firms that enter simultaneously choose computational capacity  $k_i$  and incur a computational capacity cost  $ck_i$ .<sup>4</sup>

*Stage 3.* The reservation value for each firm  $i$ ,  $v_i$ , becomes common knowledge and each firm in the market simultaneously chooses  $p_i$  to maximize profit.

*Stage 4.* Demand is realized. In the case a firm experiences demand greater than its computational capacity, we assume an efficient rationing rule (see Tirole 1988, p. 213) in which demand from segments 1 and 2 is satisfied prior to demand from segment 3. Residual demand from segment 3 is allocated to the competing firm, provided it has available capacity. In the online appendix, we show that our results are robust to an alternative proportional rationing rule, where consumers of all segments arrive randomly and firm  $i$ 's capacity is allocated simultaneously to segment 3 and segment  $i$ .

When autoscaling is available, firms simultaneously decide whether to use autoscaling or to choose a computational capacity in Stage 2.<sup>5</sup> If a firm chooses autoscaling, it incurs the computational cost  $c$  only on each unit of realized demand. In practice, changing capacity decisions in the absence of autoscaling takes at least a few hours, and in some cases days, before coming into effect on the cloud servers. The Befunky example, the slashdot effect, and the "lightning strike" analogy by Amazon's CEO, discussed in the introduction, highlight the fact that demand often changes faster than what firms can respond to in terms of computational capacity. Our assumption that capacity decision is made before the demand is realized captures this reality. However, our main results are robust to this assumption. In particular, even if firms can choose to adopt autoscaling after the demand is realized, our results in Propositions 2, 3, 4, and 5 still hold.

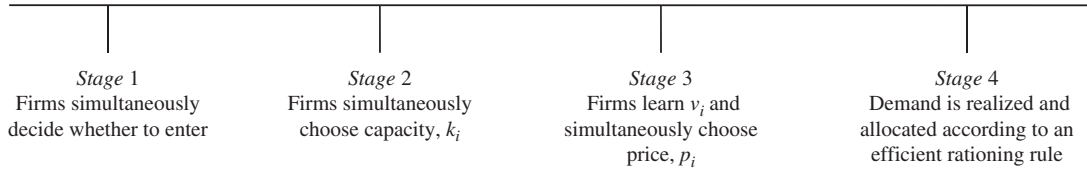
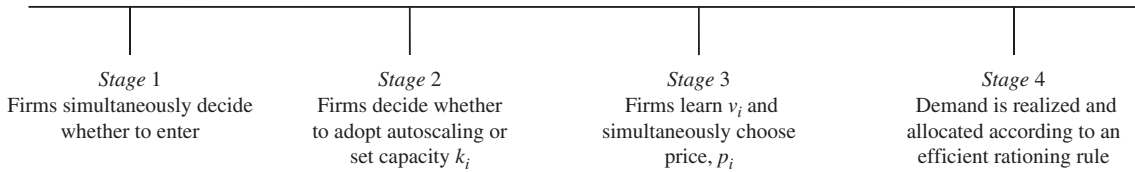
The timing of the game is depicted in Figures 1 and 2. A summary of notation is in Table 1.

## 4. Analysis

Our research objective is to identify how the advent of autoscaling affects equilibrium prices, profits, and market entry. To this end, we first examine equilibrium capacity and prices in the situation in which computational capacity must be determined prior to demand realization. We will subsequently characterize the equilibrium when autoscaling is available. We will conclude with a comparison across these possibilities.

### 4.1. Choice of Computational Capacity

In this section, we find the equilibrium choices of price and capacity and evaluate the effect of autoscaling on these choices. Throughout this section, we assume entry costs are such that both firms will have entered the market. The analysis of firms' choice of market entry is left for Section 4.3.

**Figure 1.** Sequence of Events Without Autoscaling**Figure 2.** Sequence of Events With Autoscaling

**4.1.1. Equilibrium Choices Without Autoscaling.** We start with solving the model in which there is no autoscaling via backward induction, beginning with the pricing subgame equilibrium. First suppose that  $k_1 + k_2 > 1$ . We denote this condition as *overlapping capacities*. We want to calculate equilibrium prices of this game. Without loss of generality, assume that  $k_2 \geq k_1$ . Also, it is easy to see that firms never set their capacity to  $k_i > 1 - \alpha$  or  $k_i < \alpha$ ; therefore, it is sufficient to consider the case where  $k_i \in [\alpha, 1 - \alpha]$  for  $i \in \{1, 2\}$ .

We start by showing that this game does not have a pure-strategy equilibrium. Assume for sake of contradiction that the firms use prices  $p_1$  and  $p_2$  in a pure-strategy equilibrium. If  $p_1 \neq p_2$ , then the firm with a lower price can benefit from deviating by increasing its price to  $(p_1 + p_2)/2$ . If  $p_1 = p_2$ , then firm 2 can benefit from deviating by decreasing its price to  $p_2 - \epsilon$ , for sufficiently small  $\epsilon$ , to acquire more consumers from segment 3. Therefore, a pure-strategy equilibrium cannot exist.

Next, we find a mixed-strategy equilibrium for this game. Mixed strategies can be interpreted as sales or promotions and are common in the marketing literature (e.g., Chen and Iyer 2002, Iyer et al. 2005, Zhang and Katona 2012). Provided  $k_1 \leq 1 - \alpha$ , firm 2 can choose to *attack* with a price that clears its capacity or *retreat* with a price equal to 1 that harvests the value from the  $1 - k_1$  consumers that firm 1 cannot serve

because of its capacity constraint. Let  $z$  be the price at which firm 2 is indifferent between attacking to sell to  $k_2$  consumers at price  $z$  and retreating to sell to  $1 - k_1$  consumers at price 1. We have  $z = (1 - k_1)/k_2$ . Figures 3 and 4 demonstrate the different appeals of these two pricing strategies. The choice between retreating and attacking for each firm depends on the choice of the other firm. If firm 1's price is high, it becomes easier for firm 2 to attract the consumer segment that is in both firms' reach resulting in firm 2 choosing to attack. On the other hand, if firm 1's price is low, firm 2 would prefer to retreat than to compete with firm 1 over the overlapping consumers. In the equilibrium that we find, both firms use a mixed strategy with prices ranging from  $z$  to 1. Suppose that  $F_i(\cdot)$  is the cumulative distribution function of price set by firm  $i$  and  $F_j(\cdot)$  is the cumulative distribution function of price set by competing firm  $j$ .<sup>6</sup> The profit of firm  $i$  earned by setting price  $x$ , excluding the sunk cost of capacity, is

$$\pi_i(x) = F_j(x)(1 - k_j)x + (1 - F_j(x))k_jx.$$

Using equilibrium conditions, we know that the derivative of this function must be zero for  $x \in (z, 1)$ . Therefore, we have

$$-x(k_i + k_j - 1)F'_j(x) - F_j(x)(k_i + k_j - 1) + k_i = 0.$$

The solution to this differential equation is

$$F_j(x) = \frac{k_i}{k_i + k_j - 1} + \frac{C_j}{x},$$

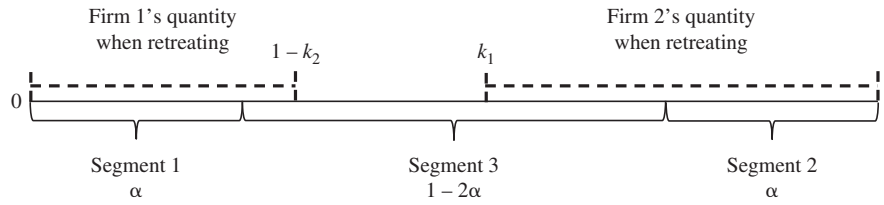
where constant  $C_j$  is determined by the boundary conditions. As for the boundary conditions, we use  $F_1(1) = 1$ . Therefore, we get

$$F_1(x) = \begin{cases} 0 & \text{if } x < z, \\ \frac{(k_1 - 1) + k_2 x}{x(k_1 + k_2 - 1)} & \text{if } z \leq x < 1, \\ 1 & \text{if } x \geq 1. \end{cases} \quad (1)$$

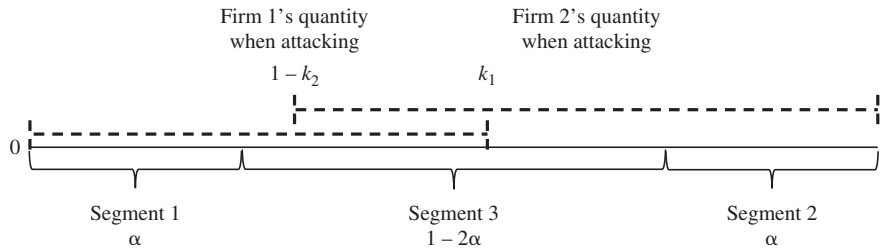
**Table 1.** Summary of Notation

Symbol	Description
$\alpha$	Size of each of segments 1 and 2
$k_i$	Capacity chosen by firm $i$
$v_i$	Reservation value consumers have for firm $i$
$\gamma$	Probability that $v_i = 1$
$c$	Cost per unit of computational capacity
$p_i$	Price chosen by firm $i$
$F$	Cost of entry

**Figure 3.** How  $k_1$  and  $k_2$  Affect Firm Incentives to Adopt a Retreating Price



**Figure 4.** How  $k_1$  and  $k_2$  Affect Firm Incentives to Adopt an Attacking Price



This implies that firm 1 mixes on prices between  $z$  and 1 such that firm 2 is indifferent between using any two prices in this range. Furthermore, given  $F_1(\cdot)$ , firm 2 strictly prefers any price in  $[z, 1]$  to any price outside this interval. To have an equilibrium, the strategy of firm 2 should be such that firm 1's strategy is not suboptimal. In other words, firm 1 should be indifferent between any two prices in  $[z, 1]$ , and should weakly prefer any price in  $[z, 1]$  to any price outside this interval. Therefore, we have to use the boundary condition  $F_2(z) = 0$  to make sure that (1) firm 1's indifference condition is satisfied in  $[z, 1]$ , and (2) firm 2 does not set the price to lower than  $z$ , as we already know from  $F_1(\cdot)$  that such prices are suboptimal for firm 2. As such, we get

$$F_2(x) = \begin{cases} 0 & \text{if } x < z, \\ \frac{k_1((k_1 - 1) + k_2 x)}{k_2 x(k_1 + k_2 - 1)} & \text{if } z \leq x < 1, \\ 1 & \text{if } x \geq 1. \end{cases} \quad (2)$$

Note that  $F_2(x)$  is discontinuous at  $x = 1$ , and jumps from  $k_1/k_2$  to 1. This implies that firm 2 uses price 1 with probability  $1 - k_1/k_2$ . In other words,  $f_2(1) = (1 - k_1/k_2)\delta(0)$ , where  $f_2(\cdot)$  is the probability density function for price of firm 2 and  $\delta(\cdot)$  is the Dirac delta function (see Hassani 2009, pp. 139–170).<sup>7</sup>

Given  $F_i(\cdot)$ , we can calculate the expected profit of each firm in this mixed-strategy equilibrium. Excluding the sunk cost of capacity, we have

$$\pi_1 = \frac{(1 - k_1)k_1}{k_2} \quad \text{and} \quad \pi_2 = 1 - k_1.$$

Note that the higher capacity firm, firm 2, earns a profit equal to the profit it would have made if it had

chosen a retreating strategy while pricing at 1; as such, the expected profit of firm 2 is independent of its capacity  $k_2$ . On the other hand, the lower capacity firm, firm 1, earns more than what it would have earned if retreating was chosen, since  $1 - k_2 < k_1 < k_2$  requires  $(1 - k_1)k_1/k_2 > 1 - k_2$ . As expected, after excluding sunk costs, the higher capacity firm makes a higher profit than the lower capacity firm.

Note that the mixed-strategy pricing equilibrium bears some resemblance to Chen and Iyer (2002) who find in a model of customized pricing the ratio of profits is equal to the ratio of consumer addressability. In our model, the profit ratio is equal to the ratio of capacities. However, the model in Chen and Iyer (2002) is conceptually very different from ours. In particular, the overlap between the customers of the two firms is always nonzero in Chen and Iyer (2002), whereas in our model the overlap is nonzero only if the sum of the capacities is larger than the market size, i.e.,  $k_1 + k_2 > 1$ . Furthermore, even though the ratio of profits is the same in both papers, the actual profit functions are very different. For example, as mentioned above, and in contrast to Chen and Iyer (2002), the profit of the firm with larger capacity does not depend on its own capacity in our model.

Now suppose  $k_1 + k_2 \leq 1$ . We denote this condition as *separated capacities*. This implies that each firm that enters the market can sell to its capacity without directly competing with the other firm for consumers in segment 3. As such, each firm that successfully enters the market can charge  $p_i = 1$  and sell  $k_i$  units for profit  $(1 - c)k_i$ . Increasing the price will result in zero sales and profit, decreasing the price will still sell  $k_i$  units but at lower revenue.

Next consider the capacity subgame equilibrium. The capacity decision is made in anticipation of the



possible combinations of values for  $v_1$  and  $v_2$ . If both firms find success (i.e.,  $v_1 = v_2 = 1$ ), then the profit depends on how  $k_i$  and  $k_j$  relate to each other and relate to  $\alpha$ . The expected profit for firm  $i$  depends on its capacity relative to the capacity of competing firm  $j$  and can be written as follows for  $k_i \in [\alpha, 1 - \alpha]$ :

$$E(\pi_i) = \begin{cases} \gamma k_i - ck_i & \text{if } k_i + k_j \leq 1, \\ \gamma(1-\gamma)k_i + \gamma^2 \frac{(1-k_i)k_i}{k_j} - ck_i & \text{if } 1-k_j < k_i < k_j, \\ \gamma(1-\gamma)k_i + \gamma^2(1-k_j) - ck_i & \text{if } 1-k_i < k_j \leq k_i, \end{cases}$$

where index  $j$  indicates the other firm. The equilibrium capacity choices are summarized in the following proposition.

**Proposition 1.** Suppose both firms enter the market initially. The equilibrium capacity choices depend on  $\gamma$  as follows:

- If there is a low probability of a successful venture (i.e.,  $\gamma < c$ ), then both firms choose  $k_i = 0$  and earn zero profit.
- If there is a moderate probability of a successful venture (i.e.,  $\gamma(1-\gamma) > c$ ), then capacities overlap such that one firm sets  $k = 1 - \alpha$  and the other firm sets  $k = k^*$ , where  $\alpha < k^* \leq 1 - \alpha$  is defined in the online appendix.
- If there is a high probability of a successful venture (i.e.,  $\gamma > c$  and  $\gamma(1-\gamma) < c$ ), then capacities do not overlap and the unique symmetric equilibrium is  $k_1 = k_2 = 1/2$ .

The results of Proposition 1 are depicted in Figure 5. Region 2 represents overlapping capacities such that  $k_1 + k_2 > 1$ . Region 3 represents separated capacities such that  $k_1 + k_2 = 1$  and in the symmetric equilibrium  $k_1 = k_2 = \frac{1}{2}$ .<sup>8</sup>

Proposition 1 highlights a nonmonotonic effect of  $\gamma$  on the equilibrium capacity choice. Intuitively, if there is a low probability of success then neither firm wishes to invest in computational capacity because there is a high probability of it going unused. Interestingly,

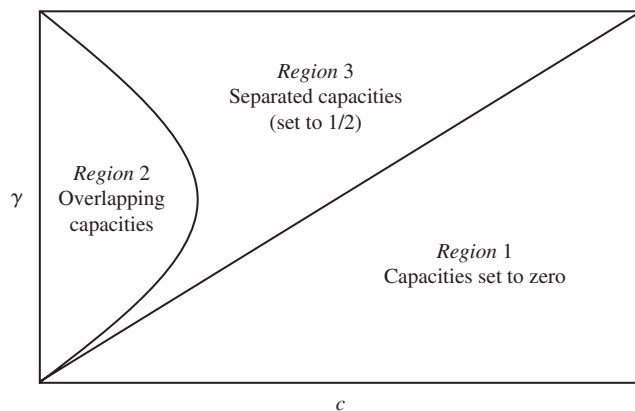
when there is a high probability of a successful venture, firms dampen competition by choosing a capacity that just covers the market. To understand this, consider the extreme case in which  $\gamma = 1$ . If firms choose separated capacities such that  $k_1 + k_2 = 1$ , both firms can charge their monopoly price for all of their consumers. The moment capacities overlap such that firms compete even for a single consumer, the firms are unable to avoid intense price competition for that consumer, thereby affecting revenues from all of their customers. Though an additional unit of capacity can result in an additional sale, the subsequent effect on price competition is severe enough such that firms refrain from competing directly.<sup>9</sup>

Another interesting facet of Proposition 1 is that a moderate probability of a successful venture leads to excessive capacity choices. Therefore, a reduction in the probability of success can actually cause an increase in capacity. To understand this result, consider two competing effects of decreasing  $\gamma$ . On the one hand, lower  $\gamma$  implies greater downside risk that the chosen capacity will go completely unused because of  $v_i = 0$  and the resulting failure in the market. This effect would suggest that capacity should decrease as  $\gamma$  decreases. On the other hand, a firm's chance at having monopoly power over all of segment 3 is maximized at moderate levels of  $\gamma$ . The latter monopoly harvesting effect dominates the former downside risk effect at moderate levels of  $\gamma$  resulting in overlapping capacities.

We show in the online appendix that when cost of capacity is low enough for overlapping capacities,  $c < \gamma(1 + \alpha(\gamma - 1))/(1 - \alpha) - 2\gamma^2$ , then both firms set maximum capacity (i.e.,  $k^* = 1 - \alpha$ ). Intuitively, when cost of capacity is negligible, even if capacity goes unused, firms do not incur a big loss. Thus both firms focus on fully benefiting from the high probability of being a monopolist,  $\gamma(1 - \gamma)$ , by setting maximum capacity, without being concerned about the *downside risk* effect.

To understand the role the assumptions play in the result, it is worth mentioning that firms can have excessive or insufficient capacity because of either demand uncertainty or randomized price competition or both. In other words, the mixed strategy in pricing decisions is not the only source of mismatch between capacity and demand. For instance, in the region for separated capacities, firm  $i$  realizes high value with probability  $\gamma$  and low value with probability  $1 - \gamma$ . Even if it does not face a competitor in the pricing game (i.e., firm  $j$  draws  $v_j = 0$ ), for high  $\gamma$  firm  $i$  will have insufficient capacity if  $v_i = 1$  and will have excessive capacity if  $v_i = 0$ . On the other hand, consider when both firms have high value. With probability  $\gamma^2$ , excessive capacity can exist because of firms setting overlapping capacities and competing over price. Thus, for a moderate probability of success (i.e.,  $\gamma(1 - \gamma) > c$ ), two high-value

**Figure 5.** Equilibrium Capacities as a Function of  $\gamma$  and  $c$



firms compete with mixed pricing strategies, resulting in excessive capacity for the firm with the higher price. Introducing autoscaling, by definition, eliminates the mismatch between capacity and demand. Next, we look at equilibrium strategies in a model of autoscaling.

**4.1.2. Equilibrium Choices With Autoscaling.** We now turn our attention to the equilibrium when autoscaling is available. Major cloud providers offer autoscaling with no additional fees.<sup>10</sup> We solve for the equilibrium pricing by entrants supposing both firms enter and choose autoscaling.<sup>11</sup>

With probability  $\gamma^2$ , we have  $v_1 = v_2 = 1$ , and it is straightforward to show there is no pure-strategy pricing equilibrium; instead the pricing subgame leads to a mixed-strategy equilibrium where the prices of both firms range between  $z'$  and 1. Similar to our analysis of mixed-strategy equilibrium without autoscaling,  $z'$  is the price for which each firm is indifferent between attacking, resulting in a profit of  $(z' - c)(1 - \alpha)$ , and retreating, resulting in a profit of  $\alpha(1 - c)$ . This results in  $z' = \alpha(1 - c)/(1 - \alpha) + c$ .

Supposing that  $G_i(\cdot)$  is the cumulative distribution function for the price of firm  $i$ , the profit of firm  $i$  when setting price  $x$ , is

$$\pi_i(x) = \alpha G_j(x)(x - c) + (1 - G_j(x))(1 - \alpha)(x - c).$$

Setting the derivative of this function equal to zero for  $x \in (z', 1)$  and using the boundary conditions  $G(z') = 0$  or  $G(1) = 1$ , we find

$$G_j(x) = \begin{cases} 0 & \text{if } x < z', \\ \frac{(1 - \alpha)(x - c) - \alpha(1 - c)}{(1 - 2\alpha)(x - c)} & \text{if } z' \leq x \leq 1, \\ 1 & \text{if } x > 1, \end{cases}$$

which results in the profit  $\alpha(1 - c)$  for each firm.

With probability  $\gamma(1 - \gamma)$ ,  $v_1 = 1$  and  $v_2 = 0$ , giving firm 1 monopoly power over all of segment 3 and profit of  $(1 - c)(1 - \alpha)$ . Thus, the expected profit of each firm when both use autoscaling is  $(1 - c)(\gamma^2\alpha + \gamma(1 - \gamma) \cdot (1 - \alpha))$ . Though we allow for firms to choose to set capacity (rather than adopt autoscaling) even when autoscaling is available, we show in the online appendix that there exists a  $\hat{c}$  such that both firms will choose autoscaling in equilibrium if  $c \geq \hat{c}$ . In the paper, we focus on both firms choosing autoscaling (i.e.,  $c \geq \hat{c}$ ), but show in the online appendix that our results also hold for  $c < \hat{c}$ , which results in only one firm choosing autoscaling.

Next, we study the effect of autoscaling on firms using the cloud and find how average prices change with the introduction of autoscaling.

## 4.2. Effect of Autoscaling on Firms' Prices

Given the firms' equilibrium strategies, we can examine how autoscaling will affect equilibrium prices in the event that entry costs are low enough such that both firms enter.

**Proposition 2.** Suppose entry costs are such that both firms enter the market with or without autoscaling. The effect of autoscaling on average prices depends on  $\gamma$  as follows:

- In the region for separated capacities (i.e.,  $\gamma > c$  and  $\gamma(1 - \gamma) < c$ ), autoscaling decreases the average price set by each firm.
- In the region for overlapping capacities (i.e.,  $\gamma(1 - \gamma) > c$ ), autoscaling increases the average price set by each firm for low enough cost of capacity (i.e.,  $c < \gamma(1 + \alpha(\gamma - 1))/(1 - \alpha) - 2\gamma^2$ ).

Proposition 2 shows demand uncertainty creates an important distinction from previous literature on capacity choice (e.g., Kreps and Scheinkman 1983), as it results in a capacity game that *decreases* average prices relative to the pricing game that arises with autoscaling. It may be expected that autoscaling should strictly reduce prices by allowing both firms to freely compete over all consumers without capacity constraints. However, Proposition 2 shows firms may actually increase their prices when they no longer have to commit to a fixed capacity under demand uncertainty. Therefore, the probability of a successful venture is critical in determining whether autoscaling increases or decreases prices; a result that is new to the literature.

In general, autoscaling has two opposing effects on price: First, autoscaling turns the cost of each server from a sunk cost to a cost that depends on the number of consumers served by the firm. Without autoscaling, firms do not consider the cost of servers in their pricing decision, as this cost is sunk. Thus, they incur no incremental cost from serving a larger portion of the market and are more flexible to do so by decreasing price. However, with autoscaling, each additional customer adds an additional cost, resulting in firms having less incentive to decrease their price to get more customers compared to when costs were sunk. This is the positive effect of autoscaling on average prices.

Second, autoscaling can intensify competition between two firms, in cases where capacity constraints without autoscaling stopped firms from competing head-to-head. This is the negative effect of autoscaling on average prices.

Proposition 2 shows the effect of autoscaling on the average prices charged by each firm for different regions of Figure 5. In the region for separated capacities, the second effect of autoscaling is the dominant effect. In this region firms choose not to attack without autoscaling, since they restrict their capacity to dampen competition. The equilibrium choice of capacity results in both firms charging the monopoly price.

Autoscaling removes this separation of targeted consumers and increases competition between the two firms, resulting in decreased average prices.

In the region for overlapping capacities, a low enough cost of capacity (i.e.,  $c < \gamma(1 + \alpha(\gamma - 1))/(1 - \alpha) - 2\gamma^2$ ) reduces the downside risk of excessive capacity and thus results in both firms setting maximum capacity;  $k_1 = k_2 = 1 - \alpha$ . This means there are no capacity constraints preventing the firms from competing over price without autoscaling. Thus, the second and negative effect of autoscaling on average prices diminishes, since autoscaling does not intensify competition in this region. Therefore, the first effect is dominant and autoscaling increases average prices.<sup>12</sup>

So far, we solved the pricing and capacity subgames conditional on both firms having had entered the market. Next, we consider firms' choice of entering the market and study how autoscaling affects this decision. Autoscaling allows the firms to avoid overspending or underspending on computational capacity. However, there is also a strategic effect of autoscaling that can lead to dampened or intensified competition. In the following section, we analyze how these two effects of autoscaling combine to influence entry decisions.

### 4.3. Effect of Autoscaling on Entry Decisions

We now turn our attention to the entry decision. We start by deriving the expected profits for each equilibrium strategy given the price and capacity choices described in Section 4.1.

First, suppose both firms enter the market. In the absence of autoscaling, the profits depend on equilibrium capacities. With overlapping capacities, the higher capacity firm earns expected profit equal to  $(1 - \alpha)((\gamma(1 - \gamma)) - c) + \gamma^2(1 - k^*)$  and the lower capacity firm earns expected profit equal to  $\gamma(1 - \gamma)k^* + \gamma^2(1 - k^*)k^*/(1 - \alpha) - ck^*$ , where  $k^*$  is the capacity chosen by the lower capacity firm and is defined in the online appendix. With separated capacities, each firm earns an expected profit of  $(\gamma - c)/2$ . As stated previously, when both firms use autoscaling, each firm's profit is  $(1 - c)(\gamma^2\alpha + \gamma(1 - \gamma)(1 - \alpha))$ .

Now consider the case when only one firm enters the market. Without autoscaling, the firm will be a monopolist, optimally choosing  $k = 1 - \alpha$  and earning expected profit  $(\gamma - c)(1 - \alpha)$ , if  $\gamma > c$ , and optimally choosing  $k = 0$  to earn zero profit otherwise. In the presence of autoscaling, a single entrant earns a profit of  $\gamma(1 - c)(1 - \alpha)$ .

Given the firms' equilibrium strategies and their expected profits, we can derive their entry decisions. We summarize the entry decisions with and without autoscaling in the following lemma.

**Lemma 1.** *Firms' entry decisions depending on the presence of autoscaling are as follows.*

- Without autoscaling: If  $F > (\gamma - c)(1 - \alpha)$ , then there is no entry. Otherwise, we have the following:

- In the region for overlapping capacities (i.e.,  $\gamma(1 - \gamma) - c > 0$ ), both firms enter if  $F < \gamma(1 - \gamma)k^* + \gamma^2(1 - k^*)k^*/(1 - \alpha) - ck^*$ . Otherwise only one firm enters.

- In the region for separated capacities (i.e.,  $\gamma > c$  and  $\gamma(1 - \gamma) - c < 0$ ), both firms enter if  $F < (\gamma - c)/2$ . Otherwise only one firm enters.

- With autoscaling: If  $F > \gamma(1 - c)(1 - \alpha)$ , then there is no entry. Both firms enter the market if  $F < \max[(1 - c) \cdot (\gamma^2\alpha + \gamma(1 - \gamma)(1 - \alpha)), ((c(\alpha(\gamma^2 - 1) + 1) + \gamma(\alpha(-\gamma) + \alpha - 1))^2)/(4(\alpha - 1)\gamma^2(c - 1))]$ . Otherwise, only one firm enters.

Lemma 1 is graphically depicted in Figures 6(a) and 6(b). Note that in Figure 6(a), there is a jump in the size of the region with double entry as  $\gamma$  grows. This is because when  $\gamma$  becomes sufficiently large, firms change their strategies from overlapping capacities to separated capacities.

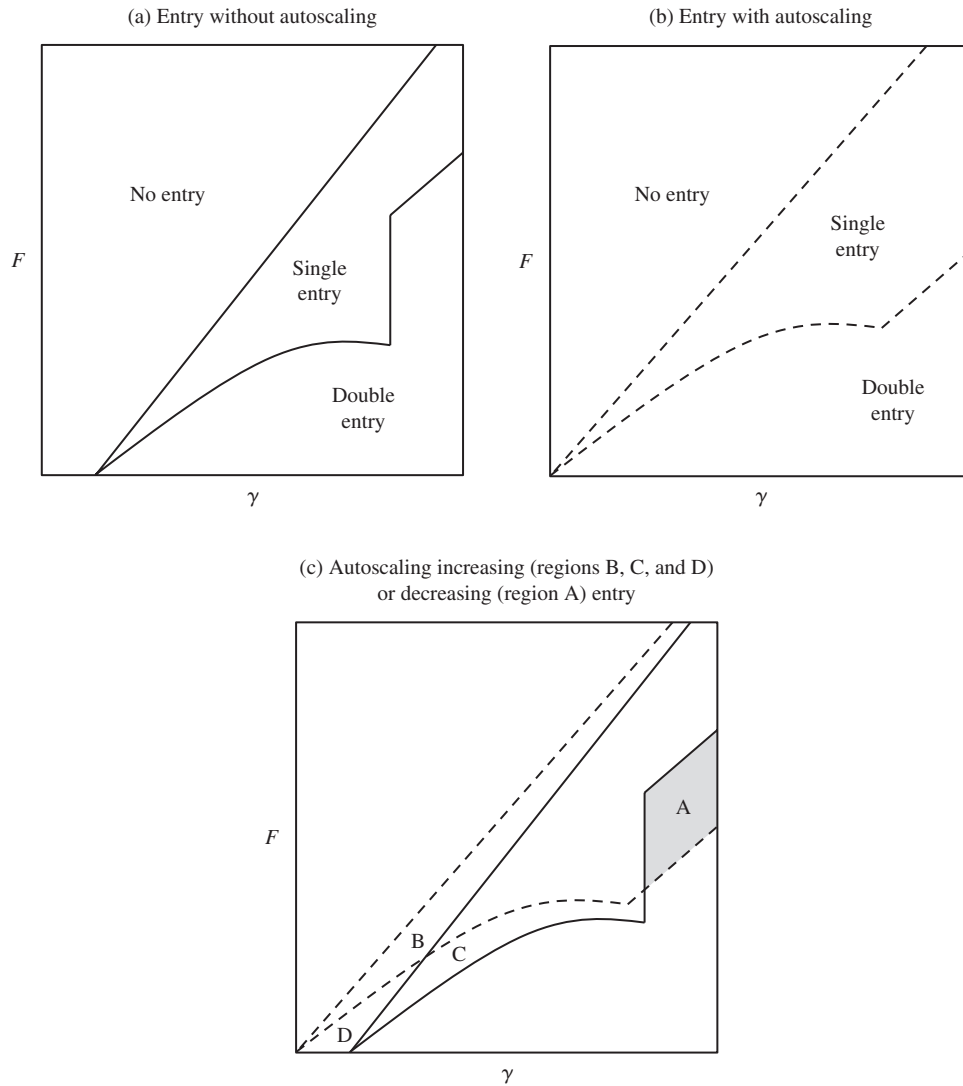
By comparing the results presented in Lemma 1, we can determine the effect of autoscaling on firm entry. We next examine how autoscaling affects whether multiple firms enter the market.

**Proposition 3.** *When probability of success,  $\gamma$ , is sufficiently large and cost of entry,  $F$ , is moderate (region A in Figure 6(c)), fewer firms enter the market in equilibrium when autoscaling is available than when it is not.*

Proposition 3 finds the counterintuitive result that autoscaling can decrease market entry. Though autoscaling has a downside risk reducing effect, it also has a competition intensifying effect. In other words, autoscaling makes it less costly for a firm to find out if it has a successful venture on its hands, but also allows a firm to fight aggressively for consumers in segment 3. To further explain these effects and when each is dominant, we consider the three potential outcomes if both firms enter.

When both firms enter, there is a  $\gamma(1 - \gamma)$  probability that a firm finds itself a monopolist, a  $\gamma^2$  probability that a firm finds itself competing head-to-head, and a  $1 - \gamma$  probability that a firm finds  $v_i = 0$ . In the former case, autoscaling weakly benefits firms because they are assured of having the computational capacity to satisfy the demand of all consumers in segment 3. Without autoscaling, firms acknowledging the downside risk choose  $k_i \leq 1 - \alpha$  and thus cannot satisfy all demand when given monopoly power over all consumers in segment 3. This is the problem that start-up Befunky experienced without autoscaling in the earlier example and represents the positive demand satisfaction effect of autoscaling. In the latter case, autoscaling weakly benefits firms because it prevents them from overpurchasing capacity. Without autoscaling,  $k_i \geq 0$  and thus firms have excess computational capacity

**Figure 6.** The Effect of Autoscaling on Entry



when  $v_i = 0$ . This is the positive downside risk reducing effect of autoscaling. However, autoscaling weakly disadvantages firms if they compete head-to-head. With autoscaling, firms are assured of having computational capacity to satisfy demand of all consumers in segment 3. As such, autoscaling intensifies competition. Without autoscaling, the fact that  $k_i \leq 1 - \alpha$  allows the firms to include a more profitable retreating price in the equilibrium mixed strategy. In this case, the demand satisfaction effect actually leads to the negative competition intensifying effect of autoscaling.

The downside risk reducing effect is most dominant when  $\gamma$  is low. The demand satisfaction effect is most dominant when  $\gamma(1 - \gamma)$  is high (i.e., moderate  $\gamma$ ) and the competition intensifying effect is most dominant when  $\gamma$  is high. A high  $\gamma$  increases the probability of competition and also makes it such that firms without autoscaling choose capacities such that this competition is avoided. Therefore, autoscaling decreases

market entry when the probability of a successful venture is sufficiently high.

We next consider the effect of autoscaling on participation in the market by any of the firms, i.e., how autoscaling affects whether at least one of the two firms enters.

**Proposition 4.** Autoscaling increases the range of entry costs,  $F$ , such that at least one firm enters the market.

Proposition 4 confirms common intuition that autoscaling can make entry more attractive for at least one firm. The reason is that it allows firms with uncertain likelihood of success to incur the cost of computational needs after demand is realized. This highlights the downside risk reducing effect of autoscaling. Without autoscaling, firms have to invest in the cost of entry  $F$  and the cost of computational capacity  $ck_i$  prior to realizing whether the venture will be successful. Autoscaling increases the range of entry costs for which at least



one firm enters by  $c(1 - \gamma)(1 - \alpha)$ . Thus, the higher the cost of capacity and the lower the probability of success, the more effective autoscaling will be in guaranteeing that the market will be served by at least one firm. Propositions 3 and 4 suggest that to find the effect of autoscaling on the number of new entrants, we must consider the probability of success as well as the cost of entry, which goes against the intuition that autoscaling always facilitates market entry.

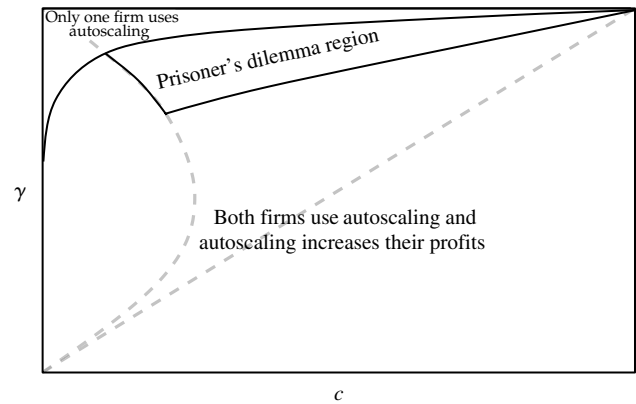
The combined results of Propositions 3 and 4 are graphically depicted in Figure 6(c). The solid (dotted) lines in the figure are the thresholds on  $F$  without (with) autoscaling, described in Lemma 1. As shown in this figure, there are four regions of interest. In region A, autoscaling decreases entry because of the competition intensifying effect. Autoscaling allows one firm to be a monopolist because the other firm cannot profitably enter given the anticipated level of competitive intensity. In regions B and D, the market will not be served by either firm unless there is autoscaling. In region C, a firm would have monopoly power because the downside risk of capacity prepurchase makes it unprofitable for a second entrant, but autoscaling alleviates this effect and results in competing firms entering the market.<sup>13</sup>

Next, we consider the effect of autoscaling on the expected profit of the two firms entering the market.

**Corollary 1.** Suppose entry costs are such that both firms enter the market with or without autoscaling. Autoscaling can create a prisoner's dilemma, such that both firms use autoscaling even though they earn greater expected profit in the absence of autoscaling.

As noted previously, the competition intensifying effect can outweigh the demand satisfaction effect and the downside risk reducing effect for sufficiently high  $\gamma$ . If  $F$  is sufficiently low, both firms will choose to enter with or without autoscaling. Furthermore, as shown in Corollary 1, they both choose autoscaling in equilibrium. Interestingly, this leads to a prisoner's dilemma situation where the firms' adoption of autoscaling results in diminished expected profitability of both firms. This result is depicted in Figure 7. The dashed lines in Figure 7 correspond to regions when autoscaling is not available (from Figure 5), and show how autoscaling affects firms' equilibrium profits in different regions. When the probability of success,  $\gamma$ , is very high, only one firm uses autoscaling while the competing firm can strategically limit its computational capacity to soften competition. Also, when  $\gamma$  is sufficiently low, both firms use autoscaling, but because of the downside risk reducing effect of autoscaling, both firms get higher profits with autoscaling. However, a moderately high  $\gamma$  creates a prisoner's dilemma situation where the competition intensifying effect of autoscaling dominates the downside risk reducing effect, but

**Figure 7.** When Existence of Autoscaling Can Lead to a Prisoner's Dilemma Effect for the Firms



the firms still use autoscaling. Therefore, both firms would be better off if autoscaling was not available in this region.

We now turn our attention to the impact of autoscaling on consumers.

#### 4.4. Effect of Autoscaling on Consumer Surplus

So far, we studied the effects of autoscaling on firms' strategies and their profit. Autoscaling can increase price competition between firms. It can also increase market entry. Both of these effects, intuitively, should lead to higher surplus for consumers. However, autoscaling also changes the capacity cost from sunk cost at the time of pricing to variable cost. Therefore, as shown in Proposition 2, autoscaling can lead to higher average prices, and thus lower surplus, for consumers. Furthermore, as shown in Proposition 3, autoscaling can also increase the likelihood of a monopoly market. In this section, we study the effect of these opposing forces on consumer surplus.

Expected consumer surplus can be derived from calculating the difference between expected social welfare and combined expected firm profit. Social welfare is equal to the combined value consumers get (i.e., the number of purchases times a value of 1) minus the cost to deliver that value (i.e.,  $c$  times the computational capacity). Therefore, the expected consumer surplus can be written as

$$E[CS] = E[\#purchases] - c \times E[\text{computational capacity}] - E[\pi_1] - E[\pi_2], \quad (3)$$

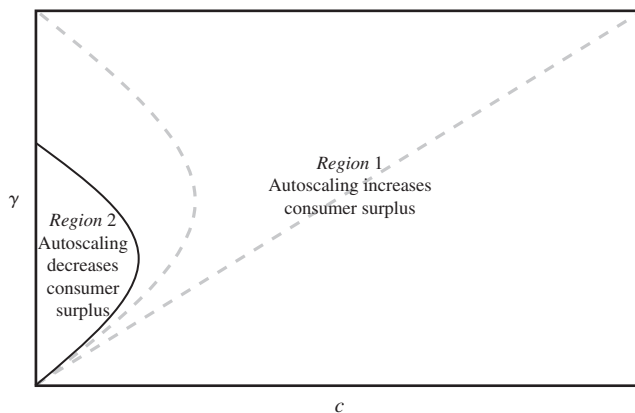
where  $\#purchases$  and  $\text{computational capacity}$  indicate the total number of consumers who purchase the product and the total computational capacity reserved by the firms, respectively. If there is only one firm in the market, in both cases with and without autoscaling, that firm sets the price to 1, resulting in zero consumer surplus. If both firms are in the market, consumer surplus

depends on whether firms use autoscaling or set capacity. We present the values of consumer surplus derived from Equation (3) in the online appendix. Comparing across conditions, we have the following result.

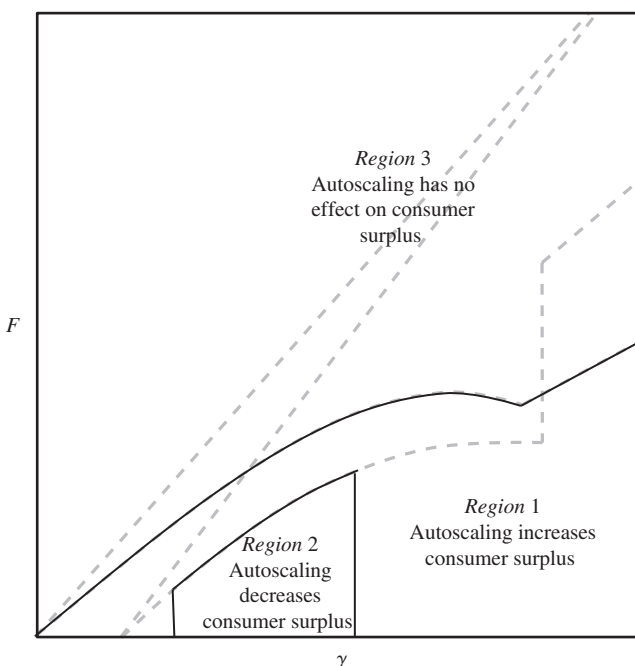
**Proposition 5.** *For sufficiently small  $F$ , sufficiently small  $c$ , and a moderate value of  $\gamma$ , autoscaling decreases expected consumer surplus. If  $F$  is sufficiently large, autoscaling has no effect on the expected consumer surplus. Otherwise, autoscaling increases the expected consumer surplus.*

Proposition 5 shows the effect of autoscaling on consumer surplus. The result is also depicted in Figures 8 and 9. Figure 8 is analogous to Figure 5 with the contours from Figure 5 marked with dashed gray lines. As we can see, the region where autoscaling decreases

**Figure 8.** Effect of Autoscaling on Consumer Surplus as a Function of  $c$  and  $\gamma$



**Figure 9.** Effect of Autoscaling on Consumer Surplus as a Function of  $F$  and  $\gamma$



consumer surplus is located in the region for overlapping capacities. Intuitively, since autoscaling increases average prices in this region, it decreases consumer surplus.

Figure 9 is analogous of Figure 6 with the contours from Figure 6 marked as dashed gray lines. The figure shows that autoscaling can only affect consumer surplus if the cost of entry is sufficiently low. In particular, unless both firms enter the market in at least one of the two cases (i.e., autoscaling or no autoscaling available), consumer surplus will be zero; therefore, if  $F$  is not low enough, autoscaling can have no effect on consumer surplus. Figure 9 also shows that when autoscaling increases market entry to two firms, it also increases expected consumer surplus.

## 5. Discussion

The emergence of cloud computing and its feature of autoscaling has the potential to influence the entry of web-based firms into new markets. Before autoscaling was available, firms needed to invest heavily in computational capacity before entering a new market so that they could serve highly unpredictable demand levels. The uncertainty in demand meant that the capacity chosen by firms could be more or less than the actual needs, resulting in extra costs or unfulfilled demand. However, autoscaling allows the new entrants a flexible capacity such that the cloud provider will designate the specific computational resources as the actual demand and required resources are realized. This feature can help new entrants to the market pay only for the capacity that is required to meet their demand, reducing the disadvantages of uncertainty in demand. In this paper, we find the effects of autoscaling on entry decisions of web-based firms and their prices, capacity decisions, and profits.

Our model shows how the demand satisfaction effect of autoscaling, in which firms can be assured of having capacity to satisfy demand, turns out to be a double-edged sword in that it frees competing firms to aggressively pursue customers. Our research identifies this competition intensifying effect of autoscaling and establishes the conditions under which this effect will outweigh the positive effects of autoscaling. This has several important implications for web-based firms and their product launch strategies.

The results can help guide pricing decisions by firms launching a new product with autoscaling. We find that autoscaling will increase the average launch prices if the probability of a successful venture is not too high. In this case, autoscaling decreases a firm's costs, but these savings are not passed on to consumers because autoscaling converts a sunk fixed cost into a variable cost that a firm incurs if and only if it attracts more customers. As a consequence, price competition is dampened by autoscaling in the case that entrants would otherwise choose excessive capacities. However,

if the probability of a successful venture is sufficiently high, firms would optimally limit their capacities to dampen competition in the absence of autoscaling. In this case, autoscaling gives the competing firms freedom to aggressively pursue customers with price and thus can decrease average prices.

The trade-off between the intensified competition and the reduced downside risk of failure affects firms' decisions on whether to enter the market. In particular, when the probability of success is sufficiently high, the increased competition effect of autoscaling dominates the decreased downside risk of failure. As a result, while autoscaling facilitates the entry for one firm, it deters a second firm from simultaneously entering the market. Our findings expand the market entry literature by looking at the topic from the new angle of capacity choice before versus after demand is realized.

Our research also guides managers of web-based firms with their response to the introduction of autoscaling and cloud computing. The research shows if the probability of success is moderately high, autoscaling leads to a prisoner's dilemma where both entrants adopt autoscaling, but they would have been better off if both had avoided it. Our findings suggest in industries with relatively low costs of entry, the introduction of the autoscaling feature in cloud computing can reduce the profit of new entrants to the market even though it decreases their downside risk of failure.

Autoscaling could also have positive or negative effects on consumer surplus. On the one hand, it can increase price competition between the firms and facilitate their entry, both of which lead to higher consumer surplus. On the other hand, autoscaling can lower consumer surplus by dampening price competition or by reducing the number of firms that enter the market. Our results show that when cost of entry is low and probability of success is moderate, autoscaling decreases expected consumer surplus.

Overall, our results highlight how web-based firms should use information on cost of entry, probability of success, cost of capacity, and level of differentiation to evaluate their entry decisions, capacity commitments, and pricing strategies in the presence of autoscaling. Our research is one of the first to consider the marketing aspects of cloud computing and autoscaling. With the rapid adoption of cloud computing by firms across different industries, marketing and economics research on the cloud can be a rich and important topic of study for future research.

### Acknowledgments

All authors contributed equally to this paper. The authors thank department editor Juanjuan Zhang, the associate editor, and anonymous reviewers for clear and constructive guidance during the review process. The authors also thank seminar participants at Stanford University, Texas A&M, and Washington University in St. Louis for their insightful comments.

### Endnotes

<sup>1</sup>"Amazon.com CEO Jeff Bezos on Animoto" by Jason Hsiao (April 21, 2008), <https://animoto.com/blog/news/company/amazon-com-ceo-jeff-bezos-on-animoto/>.

<sup>2</sup>We should note that the Hotelling model also leads to mixed-strategy equilibrium in the pricing subgame when there are capacity constraints. The reason that we use the discrete model in Narasimhan (1988) is that, unlike the Hotelling model, it gives us *ordinary* differential equations when we add capacity decisions to that model.

<sup>3</sup>In the online appendix, we show increasing the low state value strengthens the perverse effect of autoscaling on market entry. Intuitively, a high  $V_l$  means less difference between the good and bad outcomes when realizing value, which reduces the level of uncertainty faced by firms. Therefore, as  $V_l$  increases, the downside risk reducing effect of autoscaling is reduced. This in turn stops autoscaling from causing more entry into the market.

<sup>4</sup>The cost per unit of capacity,  $c$ , is assumed exogenous in the main model. We relax this assumption and allow  $c$  to be set endogenously by the cloud provider in the online appendix, where we show the general insights from the model remain the same.

<sup>5</sup>In practice, many companies such as Facebook and Netflix have revealed that they use autoscaling, and start-ups are highly recommended to do so. We should also note that assuming that firms can observe competitors' choices of autoscaling when setting their prices allows us to characterize the equilibria in the whole parameter space; however, we do not need this assumption for our main findings. Our main results come from regions in which using autoscaling is a weakly dominant strategy, and therefore, firms can rationally infer that their competitors are using autoscaling even if they cannot observe their decisions.

<sup>6</sup>We are implicitly assuming that  $F_i$  and  $F_j$  are piecewise differentiable. The game could have other mixed-strategy equilibria where cumulative distribution functions of prices are not differentiable. We cannot find those equilibria using this method.

<sup>7</sup>One might wonder if the probability density  $1 - k_1/k_2$  allocated to price 1 by firm 2 could be instead allocated to price  $z$ . The answer is that it cannot. While such strategy would still keep firm 1 indifferent between any two prices in  $[z, 1]$ , it would make price  $z - \epsilon$  (for sufficiently small  $\epsilon$ ) a strictly better strategy for firm 1, which violates equilibrium conditions.

<sup>8</sup>In Figures 5–9, we use parameters  $\alpha = \frac{1}{4}$ ,  $c = \frac{1}{2}$ , and  $F = 0$ , unless that parameter is being used as a variable in the figures.

<sup>9</sup>Note that  $\alpha$  does not affect the decision between separated and overlapping capacities. This is because regardless of the size of segment 3, the mere existence of this segment is what drives price competition when capacities overlap. Thus, as long as overlapping capacities fall in the region  $\alpha \leq k_i \leq 1 - \alpha$ , the mixed-strategy pricing chosen by each firm and therefore firms' profits do not depend on  $\alpha$ . However,  $\alpha$  does determine the equilibrium capacities chosen in the overlapping capacity region as detailed in the online appendix.

<sup>10</sup><https://aws.amazon.com/autoscaling/pricing/>, accessed May 2017.

<sup>11</sup>Theoretically, a firm could buy a fixed capacity  $k$  and also use autoscaling. We do not allow that in our model to simplify exposition, however, it is easy to see that doing so is always dominated by only using autoscaling (and not buying any fixed capacity). A formal proof is available upon request.

<sup>12</sup>Note that for  $c > \gamma(1 + \alpha(\gamma - 1))/(1 - \alpha) - 2\gamma^2$  in the region for overlapping capacities, the two firms have different average prices without autoscaling. Since with both firms using autoscaling they both set the same price, evaluating the effect of autoscaling on average charged prices is not as straight forward as for  $c < \gamma(1 + \alpha(\gamma - 1))/(1 - \alpha) - 2\gamma^2$ . Later in this section, we use consumer surplus as a proxy to average prices to study the effects of autoscaling in this region.



<sup>13</sup>Note that  $\alpha < \frac{1}{2}$  is a necessary condition in the proof of Proposition 3; when  $\alpha = \frac{1}{2}$ , region A in Figure 6(c) disappears. In other words, autoscaling decreases market entry only in the existence of competition. Without competition (i.e., when  $\alpha = \frac{1}{2}$  and size of segment 3 equals zero), there are no downsides to using autoscaling, and thus autoscaling always increases market entry.

## References

- Anand KS, Girotra K (2007) The strategic perils of delayed differentiation. *Management Sci.* 53(5):697–712.
- Anupindi R, Jiang L (2008) Capacity investment under postponement strategies, market competition, and demand uncertainty. *Management Sci.* 54(11):1876–1890.
- Armbrust M, Fox A, Griffith R, Joseph AD, Katz RH, Konwinski A, Lee G, et al. (2009) Above the clouds: A Berkeley view of cloud computing. Technical Report UCB/EECS-2009-28, Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley.
- Buehler S, Haucap J (2006) Strategic outsourcing revisited. *J. Econom. Behav. Organ.* 61(3):325–338.
- Burke A, Hussels S (2013) How competition strengthens start-ups. *Harvard Bus. Rev.* 91(3):24–25.
- Che H, Narasimhan C, Padmanabhan V (2010) Leveraging uncertainty through backorder. *Quant. Marketing Econom.* 8(3):365–392.
- Chen P, Wu S (2013) The impact and implications of on-demand services on market structure. *Inform. Systems Res.* 24(3):750–767.
- Chen Y, Iyer G (2002) Consumer addressability and customized pricing. *Marketing Sci.* 21(2):197–208.
- Columbus L (2015) Roundup of cloud computing forecasts and market estimates Q3 update, 2015. *Forbes* (September 27), <http://www.forbes.com/sites/louiscolombus/2015/09/27/roundup-of-cloud-computing-forecasts-and-market-estimates-q3-update-2015/>.
- Desai P, Koenigsberg O, Purohit D (2007) The role of production lead time and demand uncertainty in marketing durable goods. *Management Sci.* 53(1):150–158.
- Desai P, Koenigsberg O, Purohit D (2010) Forward buying by retailers. *J. Marketing Res.* 47(1):90–102.
- Ferguson M, Koenigsberg O (2007) How should a firm manage deteriorating inventory. *Production Oper. Management* 16(3):306–321.
- Goyal M, Netessine S (2007) Strategic technology choice and capacity investment under demand uncertainty. *Management Sci.* 53(2):192–207.
- Griffith E (2014) Why startups fail, according to their founders. *Fortune* (September 25), <http://fortune.com/2014/09/25/why-startups-fail-according-to-their-founders/>.
- Gupta P, Seetharaman A, Raj JR (2013) The usage and adoption of cloud computing by small and medium businesses. *Internat. J. Inform. Management* 33(5):861–874.
- Hassani S (2009) *Mathematical Methods for Students of Physics and Related Fields*, 2nd ed., Chap. 5 (Springer, New York).
- Iyer G, Soberman D, Villas-Boas J (2005) The targeting of advertising. *Marketing Sci.* 24(3):461–476.
- Joshi Y, Reibstein D, Zhang J (2009) Optimal entry timing in markets with social influence. *Management Sci.* 55(6):926–939.
- Klems M, Nimis J, Tai S (2008) Do clouds compute? A framework for estimating the value of cloud computing. Weinhardt C, Luckner S, Stößer J, eds. *Workshop on E-Business: Designing E-Business Systems—Markets, Services, and Networks*, Vol. 22 (Springer, Berlin), 110–123.
- Kreps D, Scheinkman J (1983) Quantity precommitments and Bertrand competition yield Cournot outcomes. *Bell J. Econom.* 14(2):326–337.
- Leavitt N (2009) Is cloud computing really ready for prime time? *Computer* (1):15–20.
- Lilien GL, Yoon E (1990) The timing of competitive market entry: An exploratory study of new industrial products. *Management Sci.* 36(5):568–585.
- Marston S, Li Z, Bandyopadhyay S, Zhang J, Ghalsasi A (2011) Cloud computing: The business perspective. *Decision Support Systems* 51(1):176–189.
- Milgrom P, Roberts J (1982) Limit pricing and entry under incomplete information: An equilibrium analysis. *Econometrica* 50(2):443–459.
- Narasimhan C (1988) Competitive promotional strategies. *J. Bus.* 61(4):427–449.
- Narasimhan C, Zhang J (2000) Market entry strategy under firm heterogeneity and asymmetric payoffs. *Marketing Sci.* 19(4):313–327.
- Nasser S, Turcic D (2015) To commit or not to commit: Revisiting quantity vs. price competition in a differentiated industry. *Management Sci.* 62(6):1719–1733.
- Nickelsburg M (2016) Startup spotlight: BeFunky makes online photo editing and graphic design simple. *GeekWire* (January 28), <http://www.geekwire.com/2016/befunky/>.
- Ofek E, Turut O (2013) Vaporware, suddenware, and trueware: New product preannouncements under market uncertainty. *Marketing Sci.* 32(2):342–355.
- Reynolds S, Wilson B (2000) Bertrand-Edgeworth competition, demand uncertainty, and asymmetric outcomes. *J. Econom. Theory* 92(1):122–141.
- Shy O, Stenbacka R (2003) Strategic outsourcing. *J. Econom. Behav. Organ.* 50(2):203–224.
- Spence M (1977) Entry, capacity, investment and oligopolistic pricing. *Bell J. Econom.* 8(2):534–544.
- Sultan NA (2011) Reaching for the cloud: How SMEs can manage. *Internat. J. Inform. Management* 31(3):272–278.
- Swinney R, Cachon G, Netessine S (2011) Capacity investment timing by start-ups and established firms in new markets. *Management Sci.* 57(4):763–777.
- Tirole J (1988) *The Theory of Industrial Organization* (MIT Press, Cambridge, MA).
- Van Mieghem J, Dada M (1999) Price versus production postponement: Capacity and competition. *Management Sci.* 45(12):1631–1649.
- Walker E (2009) The real cost of a CPU hour. *Computer* 42(4):35–41.
- Yang H, Tate M (2012) A descriptive literature review and classification of cloud computing research. *Comm. Assoc. Inform. Systems* 31(2):35–60.
- Zhang K, Katona Z (2012) Contextual advertising. *Marketing Sci.* 31(6):980–994.
- Zhou B, Mela C, Amaldoss W (2015) Do firms endowed with greater strategic capability earn higher profits? *J. Marketing Res.* 52(3):325–336.