Alma Mater Studiorum – Università di Bologna
in cotutela con Toulouse School of Economics

DOTTORATO DI RICERCA IN

Economia

Ciclo XXVII

Settore Concorsuale di afferenza: 13/A4

Settore Scientifico disciplinare: SECS-P/06

Essays in Industrial Organization and Information Technology

Presentata da: **Wing Man Wynne Lam**

Coordinatore Dottorato

Matteo Cervellati

Relatore

Giacomo Calzolari

Relatore

Jacques Crémer

Esame finale anno 2014

Essays in Industrial Organization and Information Technology

Wing Man Wynne Lam

Toulouse School of Economics, and University of Bologna

E-mail: wingmanwynne.lam2@unibo.it

August, 2014

Contents

Executive Summary

My dissertation studies issues of competition and investment in Internet markets, and it is divided into three parts: the first chapter provides a general analysis of platform competition, which can be applied to Internet markets such as mobile applications, online advertisements and search, as well as more traditional markets such as credit cards, video games and shopping malls; the second and third chapters present two models to help understand the markets for cloud computing and cybersecurity.

Chapter 1 studies how consumers' switching costs affect the pricing and profits of firms competing in two-sided markets such as Apple and Google in the smartphone market. When two-sided markets are dynamic—rather than merely static—I show that switching costs lower the first-period price if network externalities are strong, which is in contrast to what has been found in one-sided markets. By contrast, switching costs soften price competition in the initial period if network externalities are weak and consumers are more patient than the platforms. Moreover, an increase in switching costs on one side decreases the first-period price on the other side.

Chapter 2 examines firms' incentives to invest in local and flexible resources when demand is uncertain and correlated. Before demand is realized, two firms decide to invest in their local capacity. Provider(s) of flexible resource observe these decisions and invest in their capacity. After demand is realized, firms buy flexible resource if demand exceeds their local capacity. I find that market power of the monopolist providing flexible resources distorts investment incentives, while competition mitigates them. The extent of improvement depends critically on demand correlation and the cost of capacity: under social optimum and monopoly, if the flexible resource is cheap, the relationship between investment and correlation is positive, and if it is costly, the relationship becomes negative; under duopoly, the relationship is positive. The analysis also sheds light on some policy discussions in markets such as cloud computing.

Chapter 3 develops a theory of sequential investments in cybersecurity in which the software vendor can invest ex ante and ex post. The regulator can use safety standards and liability rules as means of increasing security. I show that the joint use of an optimal standard and a full liability rule leads to underinvestment ex ante and overinvestment ex post because the software vendor does not suffer the full costs of the society in case of security failure. Instead, switching to a partial liability rule can correct the inefficiencies. This suggests that to improve security, the regulator should encourage not only the firms, but also the enterprises to invest in security. I also discuss the effect of network externality and explain why firms engage in "vaporware".

Chapter I

Switching Costs in Two-sided Markets

This paper studies a dynamic two-sided market in which consumers face switching costs between competing products. I first show that, in a symmetric equilibrium, switching costs lower the first-period price if network externalities are strong. By contrast, switching costs soften price competition in the initial period if network externalities are weak and consumers are more patient than the platforms. Second, an increase in switching costs on one side decreases the first-period price on the other side. Finally, consumer heterogeneity such as the presence of more loyal and naive customers on one side intensifies first-period competition on this side but softens first-period competition on the other side.

Keywords: switching costs, two-sided markets, network externality, naivety, loyalty

JEL Classification: D4, L1

> *"High price [and] lack of consumption apps... doomed the Surface. They could have broken through by pricing the Surface aggressively to drive sales volume that created a pull on app developers. But they didn't. Consumers stayed away."*
>
> Hal Berenson, President of True Mountain Group, LLC.[1]

1 Introduction

In many markets, there are switching costs and network effects. Previous work points out that large switching costs cause firms to charge a higher price to their locked-in customers, and

I thank Marc Bourreau, Giacomo Calzolari, Jacques Crémer, Vincenzo Denicolò, Lapo Filistrucchi, Bruno Jullien, Alireza Naghavi, Andrew Rhodes, Paul Seabright, and participants in numerous conferences and seminars for helpful comments. Any opinions expressed are those of the author only.

[1] Quoted from "Will Microsoft get the new Surface(s) right? Part 1," hal2020.com, May 8, 2014.

large network externalities cause platforms to charge a lower price, yet little is known about the interaction between the two concepts. This paper studies how switching costs affect price competition when network externality is present; I find that an increase in switching costs of one group intensifies price competition for the other group in the introductory period.

A good example is the smartphone operating system market. Apple, Google and Windows are key players in the market. Each of them faces two groups of consumers, application users and application developers. While it is easy for consumers to migrate data from an older version of Windows Phone to a newer version, a consumer who switches from Android to Windows Phone incurs the cost of migrating—if not re-purchasing—a set of apps, media files, as well as contacts, calendars, emails and messages. As suggested by Hal Berenson, one of the problems faced by Windows Phone is its weak app library. Suppose now that Windows improves its library by introducing more Android apps. This not only raises the utility of users through network externality but also lowers their switching cost in terms of data migration. For instance, making some Android movie or music streaming apps available also for Windows Phone allows users to migrate their media files across devices more easily without the hassle of moving the data manually, which results in lower switching costs.[2] Such change may seem to be welfare-improving because the extent to which platforms can exploit their locked-in customers is smaller. However, in markets with cross-group externalities, where participation of one group increases the value of participating for the other group, I show that a decrease in switching costs of the user leads to an increase in the price for developers. Since developers value the participation of the user and a decrease in switching costs of the user makes attracting users easier, the platform can price higher to extract rents from developers. As a consequence, lower switching costs may not improve consumer welfare. It is important that regulators can evaluate the outcome of these cross-group effects properly. The analysis also provides insight into other two-sided markets with switching costs, such as media, credit cards, video games, and search engines.

I consider a simple Hotelling model of duopoly with horizontal differentiation, where platforms 0 and 1 sell their product to consumers whose relative preference for the two platforms are indexed by their position along a unit interval. Consumers have unitary demand, so that in each period, each consumer purchases one good from either platform (single-homing). The penultimate section will extend the analysis to cover the multi-homing case. I assume that there are both switching costs and network externalities. Moreover, consumers are heterogeneous in terms of loyalty and naivety. Loyal consumers are attached to one platform and never switch.[3] Naive consumers are short-sighted and care only about today. This model is flexible enough that it can collapse to either a pure switching-cost model or to a pure two-sided market model for extreme parameter

[2] Klemperer (1995) gives many examples of different kinds of switching costs, for instance, learning costs, psychological costs, transactions costs, etc. The UK Office of Fair Trading documented some useful case studies.

[3] A survey published by Consumer Intelligence Research Partners (CIRP) reveals that almost half of smartphone buyers stay loyal to their previous brand, with Apple having the highest loyalty rate. This survey was taken from data surveying 500 subjects in the US who had purchased a new mobile phone in the previous 90 days over the last four quarters, between July 2012 and June 2013.

values. When both effects are at work, I show that conventional results may change. I focus on symmetric equilibrium in which platforms charge the same price to each side. I also show that such equilibrium exists even when parameters on the two sides are not symmetric.

This paper's contribution is twofold. First, it studies switching costs together with network externalities, whereas the existing literature has tended to focus on either of them. Discussing the two together is important—I show that switching costs work differently in a two-sided market and this result has important implications for consumer protection. In a one-sided market, switching costs may intensify or soften first-period price competition depending upon how patient consumers are relative to platforms; but in a two-sided market, under strong externalities, higher switching costs always make the first-period more competitive. I also find that there is a cross-effect: higher switching costs on one side unambiguously reduce the price on the other side. The second contribution relates to the investigation of consumer heterogeneity that has been neglected in the two-sided market literature. In particular, this model provides a general framework for examining how switching costs affect the pricing strategy of platforms depending on consumers' characteristics, such as sophistication and loyalty, which traditional arguments cannot deal with.

The main results can be summarized as follows. When cross-group externalities are weak, whether higher switching costs make the market more competitive in the first period depends on two forces. On the one hand, more patient consumers are less tempted by a temporary price cut because they understand that the price cut will be followed by a price rise in later periods. Their demand is therefore less elastic, and platforms will respond by charging higher prices. On the other hand, more patient platforms put more weight on future profits, and thus both compete aggressively for market share. Switching costs make markets more competitive if platforms are relatively more patient than the consumers. By contrast, when externalities become sufficiently strong, platforms' incentive to lock consumers in becomes stronger because by capturing one group of consumers, it helps to convince the other group to join. Consequently, higher switching costs cause the platform to charge a lower price in the first period. Additionally, there is a cross-group effect: an increase in switching costs on one side unambiguously decreases the price on the other side. The reason is that platforms can build market share either directly through one side or indirectly through the other side. When switching costs on one side are large, an easier way to build market share is to focus on the indirect channel; consequently first-period competition is increased on the other side (Proposition 5).

Considering consumer heterogeneity, I show that platforms offer lower prices to one side if there are many naive and loyal consumers. The intuitive reason is that after consumers make their purchase in the first period, consumers who are loyal know that they will patronize the same platform for an indefinite period of time, and feel that they deserve a bigger carrot in the first period. The presence of naive consumers, who care only about immediate cost and reward, gives even more incentive to platforms to compete aggressively. Platforms charge higher prices to one side if on the other side there are more naive consumers. This is because higher price elasticity on the side with more naive consumers reduces the opportunity cost of recruiting consumers on the other side. Therefore, it leads to less competitive behavior on the other side (Proposition 7).

These results yield clear policy recommendations. First, since asymmetric price structures are common in two-sided markets, attractive introductory offers do not necessarily call for consumer protection as in one-sided markets. Second, if disloyal consumers do not know their preferences in the first period, platforms may provide imprecise information about their tastes, so that these consumers are less loyal, and they will switch more, which platforms can exploit later. Therefore, there is room for government intervention, particularly in achieving a greater transparency of information. Disloyal consumers would benefit from more information, so that they are able to make choices that are best aligned to their tastes. As a result, they can build loyalty more easily and save considerable switching costs.

1.1 Related Literature

There is a sizeable literature on switching cost, which broadly speaking, can be categorized into two main groups.[4] One group of papers assumes that firms cannot discriminate between old and new consumers. Firms knowing that they can exercise market power in the second period over those consumers who are locked-in, they are willing to charge a lower price in the first period in order to acquire these valuable customers. This "bargains-then-ripoff" pattern is the main result of the first-generation switching-cost models (see for instance Klemperer (1987a, b)). A second group of works allows for price discrimination, so firms can charge a price to its old customers and a different price to new ones. Chen (1997) analyzes a two-period duopoly with homogeneous goods. Under duopoly, consumers who leave their current supplier have only one firm to switch to. Since there is no competition for switchers, this allows the duopolist to earn positive profits in equilibrium. Taylor (2003) extends Chen's model to many periods and many firms. With three or more firms, there are at least two firms vying for switchers, and if products are undifferentiated, these firms will compete away all their future profits. More recent contributions include Biglaiser, Crémer and Dobos (2013), which studies the consequence of heterogeneity of switching costs in an infinite horizon model with free entry. They show that even low switching cost customers are valuable for the incumbent.

The design of pricing strategies to induce agents on both sides to participate has occupied a central place in the research on two-sided markets.[5] The pioneering work is Caillaud and Jullien (2003), who analyze a model of imperfect price competition between undifferentiated intermediaries. In the case where all agents must single-home, the only equilibrium involves one platform attracting all agents and the platform making zero profit. In contrast, when agents can multi-home, the pricing strategy is of a "divide-and-conquer" nature: the single-homing side is subsidized (divide), while the multi-homing side has all its surplus extracted (conquer). Armstrong (2006)

[4] Farrell and Klemperer (2007), and Klemperer (1995) provide excellent overviews on the literature of consumer switching costs.

[5] See Rysman (2009) for a survey of the literature on two-sided markets.

advances the analysis by putting forward a model of competition between differentiated platforms by using the Hotelling specification. He finds that the equilibrium price is determined by the magnitude of cross-group externalities and whether agents single-home or multi-home. His approach is the closest to mine. However, he focuses on a static model of two-sided market without switching costs, while here with switching costs and different degrees of sophistication the problem becomes a dynamic one. Another closely related paper is Rochet and Tirole (2006), who combine usage and membership externalities (as opposed to the pure-usage-externality model of Rochet and Tirole (2003), and the pure-membership-externality model of Armstrong (2006)), and derive the optimal pricing formula. But they focus on the analysis of a monopoly platform.

Substantial studies have been separately conducted in the dual areas of switching costs and two-sided markets, but analysis is rarely approached from a unified perspective. This paper seeks to fill the gap. Besides this study, there is little literature that studies the interaction between switching-costs and network externalities. Su and Zeng (2008) analyze a two-period model of two-sided competing platforms. Their focus is on the optimal pricing strategy when only one group of agents has switching costs and their preferences are independent, while this paper studies a richer setting in which both sides bear switching costs, and consumers are heterogeneous in terms of loyalty and naivety. Therefore, one can view Su and Zeng (2008) as a special case of my model. Biglaiser and Crémer (2014) study the effect of switching costs and network externalities on competition, but they do not address the issue in a two-sided context.

2 Model

Consider a two-sided market with two periods. There are two groups of consumers, denoted A and B, such as smartphone users and application developers. Assume that for some exogenous reasons in each period consumers choose to single-home. Section 5.1 will extend the analysis to cover the multi-homing case. Both sides of consumers have switching costs: side i (A or B) consumers have to incur switching cost $s_i \geq 0$ if they switch platform in the second period. On each side, consumers are heterogeneous in two dimensions. First, consumers can be naive or rational. Naive consumers, who are a fraction α_i of the population on side i, make decisions based on their first-period utility; while rational consumers, who form a fraction $1 - \alpha_i$ of side i's population, make decisions based on their lifetime utility. Therefore, on each side, naive consumers have $\delta_i = 0$, while rational consumers have $\delta_i > 0$.[6] Moreover, I distinguish the firm's discount factor, denoted δ_F, from the consumer's discount factor δ_i. Second, consumers learn whether they are loyal or not after their purchase in the first period. With probability μ_i consumers' preferences do not change and they never switch ("loyal"), and with probability $1 - \mu_i$ their preferences are

[6] This is different from Klemperer (1987b) because he does not consider the possibility of having a mixture of naive and rational consumers. Consumers are either all naive or all rational.

re-distributed on the unit interval in the second period (independent preferences).[7] Independent preferences are needed for technical reason because it smooths the demand function. Since not all consumers have changing preferences in practice, I assume that there are some loyal consumers. There are two competing platforms, denoted 0 and 1, which enable the two groups to interact. Consider a simple Hotelling model, where consumers on each side are assumed to be uniformly located along a unit interval with the two platforms located at the two endpoints. Both α_i and μ_i are known by the platforms. Throughout the paper, we assume that platforms cannot price discriminate among his previous customers and customers who have bought the rival's product in the previous period.

The utility of a consumer on side i is

$$v_i + e_i n_{k,t}^j - |x-k| - p_{k,t}^i,$$

where $i, j \in \{A, B\}, i \neq j$ since the two sides are symmetric. v_i is the intrinsic value of consumers on side i for using either platform. Assume that v_i is sufficiently large such that the market is fully covered. e_i is the benefit that consumer from side i enjoys from interacting with each agent on the other side (for simplicity, I ignore the possibility that consumers also care about the number of people in the same group who joins the platform). Suppose that each side is of mass 1, so that $n_{k,t}^i$ is the number of agents from side i (A or B) who are attached to platform k (0 or 1) in period t (1 or 2), while the number of agents from the same side in the same time period who are attached to the other platform is denoted $1 - n_{k,t}^i$. Thus, $e_i n_{k,t}^j$ is the total external benefit from interacting with the other group. The location of the consumer is denoted x. To keep things simple, I assume unit transport cost. Thus, $|x-k|$ is the transport cost when the consumer purchases from platform k. Platform charges are levied on a lump-sum basis: each agent from side i incurs a cost of $p_{k,t}^i$ when he joins platform k at time t.

Platform k's profit at time t is given by

$$\pi_{k,t} = p_{k,t}^A n_{k,t}^A + p_{k,t}^B n_{k,t}^B, \tag{1}$$

which is the sum of revenues from side-A and side-B. I make three assumptions. First, assume that the marginal cost of production is equal to zero for simplicity. Second, assume that $s_i \in [0, 1)$, where one is the unit transport cost, so that at least some consumers will switch. Third, assume $e_i \in [0, 1)$ in order to ensure that the profit function is well-defined, and the demand is decreasing

[7] Loyalty in this model can be interpreted in two ways: First, it can be interpreted as exogenous. Loyal consumers are not able to switch because they have large switching costs. Second, loyalty can be interpreted as endogenous. Suppose that switching cost is drawn from a two-point distribution: s is small with probability $1-\mu$, and s is big with probability μ. In this case, the concept of loyalty is endogenized because it is determined by switching costs. Both interpretations lead to the same calculations, but for simplicity I adopt the first interpretation for the rest of the analysis.

Klemperer (1987b) makes a similar assumption, but he assumes that those consumers, who have fixed tastes, respond to prices in both periods.

in a platform's own price and increasing in its rival's price.[8]

The timing of the game is as follows.

- In the first period, consumers are unattached. They learn their preferences. Platforms set the first-period price. Consumers choose which platform to join.
- In the second period, consumers learn their switching cost and whether they are loyal or not.[9] Platforms set the second-period price. Consumers decide to switch or not.

The solution concept for the game is subgame perfect equilibrium (SPE).

2.1 Second Period: the mature market

I work backward from the second period, where each platform has already established a customer base. Given the first-period market shares $n^A_{0,1}$ and $n^B_{0,1}$, a consumer on side i, located at θ^i_0 on the unit interval, purchased from platform 0 in the first period is indifferent between continuing to buy from platform 0 and switching to platform 1 if

$$v_i + e_i n^j_{0,2} - \theta^i_0 - p^i_{0,2} = v_i + e_i(1 - n^j_{0,2}) - (1 - \theta^i_0) - p^i_{1,2} - s_i.$$

The indifferent consumer is given by

$$\theta^i_0 = \frac{1}{2} + \frac{1}{2}[e_i(2n^j_{0,2} - 1) + p^i_{1,2} - p^i_{0,2} + s_i].$$

Another consumer on side i, positioned at θ^i_1, previously purchased from platform 1 is indifferent between switching to platform 0 and continuing to purchase from platform 1 if

$$v_i + e_i n^j_{0,2} - \theta^i_1 - p^i_{0,2} - s_i = v_i + e_i(1 - n^j_{0,2}) - (1 - \theta^i_1) - p^i_{1,2}.$$

The indifferent consumer is given by

$$\theta^i_1 = \frac{1}{2} + \frac{1}{2}[e_i(2n^j_{0,2} - 1) + p^i_{1,2} - p^i_{0,2} - s_i].$$

We then substitute θ^i_0 and θ^i_1 into the following.

$$n^i_{0,2} = \mu_i n^i_{0,1} + (1 - \mu_i) n^i_{0,1}\theta^i_0 + (1 - \mu_i)(1 - n^i_{0,1})\theta^i_1. \qquad (2)$$

Consumers of platform 0 consists of three types, and similarly for platform 1. The first type is loyal customers, who buy from platform 0 in both periods. The second type is switchers (whose preferences are unrelated in the two periods), who did not switch away from platform 0. The third type is also switchers, but they switched away from platform 1 to platform 0.

Then, we solve for the market shares, plug them into the profit functions, and solve for the equilibrium prices. The details are shown in Appendix A.

[8] More specifically, one represents the unit transport cost. Assuming $e_i < 1$ ensures that in the symmetric equilibrium, both platforms serve some consumers.

[9] The analysis is the same even if consumers learn their switching cost in the first period. However, if they know whether they are loyal or not in the first period, the calculation changes slightly, but qualitative results should hold.

Effect of Switching Costs on Second-period Pricing

Proposition 1. *Given first-period market share, on each side, the platform with a larger market share increases the second-period price as switching costs increase; whereas the other platform with a smaller market share decreases the second-period price as switching costs increase.*

Proof. See Appendix A.1. □

The literature calls this price a "ripoff" because the second-period price paid by consumers in equilibrium is higher in a market with switching costs than in a market without switching costs.[10] However, the extent of the ripoff depends on market share. There are two possible strategies: On the one hand, the platform might want to exploit its existing customers with a high price because switching costs give platform market power over the consumers who are locked-in. On the other hand, the platform might want to poach its rival's customers with a low price. Proposition 1 shows that the platform with a larger market share charges a higher second-period price as switching costs increase because it focuses more on exploiting old customers than on poaching new customers; whereas the platform with a smaller market share charges a lower second-period price in order to win back some customers.

Notice that if the market share is equal between platforms, then switching cost has no effect on the second-period price, which is indeed the case when we solve the full equilibrium. The reason is that when platforms have an equal share of the market, their incentives to exploit old customers offset their incentives to attract new customers.

Proposition 2. *Given first-period market share, the second-period price paid by consumers on side i is increasing in switching costs of consumers on side j if*

(i) Consumers on side j are more valuable ($e_i > e_j$), and platform 0 has a larger market share on side j ($n^j_{0,1} > 1/2$), or

(ii) Consumers on side i are more valuable ($e_i < e_j$), and platform 0 has a smaller market share on side j ($n^j_{0,1} < 1/2$).

Proof. See Appendix A.2. □

The intuition behind Proposition 2 runs as follows. Part (i) shows that consumers on side j are more valuable to the platform because they exert stronger externalities on consumers on side i compared to externalities of side i on side j. If the platform has a larger market share of the more valuable side, it can charge higher second-period prices to both sides compared to the case without switching costs. That is, $\partial p^j_{0,2}/\partial s_j > 0$ from Proposition 1, and $\partial p^i_{0,2}/\partial s_j > 0$ from (i) of Proposition 2.

[10] As will be seen later, the second-period price in my model is $p^i_{0,2} = \frac{1-e_j(1-\mu_i)}{1-\mu_i}$, which is larger than the price in a two-sided market model without switching costs, $p^i = 1 - e_j$.

By contrast, part (ii) shows that if the platform has a smaller market share of side j, according to Proposition 1 it will focus more on poaching side j with a low price than exploiting them with a high price, that is, $\partial p_{0,2}^j/\partial s_j < 0$. It will then charge a higher second-period price to side i because decreasing the price on side j reduces the "opportunity cost" of recruiting consumers on side i: the platform loses less revenue on side j by recruiting one less consumer on side i.[11] Both platforms thus compete less aggressively for them. Consequently, higher switching costs on side j cause the platform to charge a higher price on side i, that is, $\partial p_{0,2}^i/\partial s_j > 0$. Note that what platform 1 will do is just the opposite of platform 0 because of the asymmetric market shares.

In a one-sided market with switching costs, a platform's market share is an important determinant of its pricing strategy because it affects the platform's future profitability (see Klemperer (1995)); in a multi-sided market it is crucial to also take into consideration network externalities. Relying on a one-sided logic may overestimate potential anti-competitive effects: according to Proposition 1 the second-period price tends to increase with switching cost on the side that the platform has a larger market share; but this does not necessarily imply anti-competitive motives in two-sided markets, since according to Propositions 2 larger margin on one side could be translated into smaller or even negative margin on the other side depending on the magnitude of externalities.

Effect of Switching Costs on Second-period Profit

Consider the case, where (i) the platform's first-period market shares of the two sides are not too small, and (ii) cross-group externalities are not too different from each other.

Proposition 3. *Platform's second-period profits are increasing in switching costs on one side if it has a larger market share on this side than the other platform, and decreasing in switching costs if it has a smaller market share.*

Proof. See Appendix A.3. □

In the literature, switching costs typically raise platforms' profits in the second period of a market with switching costs as compared to a market without switching costs because platforms charge a higher price to repeat buyers. However, Proposition 3 shows that whether second-period profits increase or decrease with switching costs depends on market share and cross-group externalities.

[11] Rochet and Tirole (2003, 2006) explain that the difference between a one-side market and a two-sided market lies in the change in this opportunity cost. In particular, the standard Lerner formula becomes

$$\frac{p^i - (c - p^j)}{p^i} = \frac{1}{\eta^i}$$

in a two-sided market, where c is the marginal cost and η is the price elasticity.

2.2 First Period: the new market

I now turn to the first-period equilibrium outcomes when consumers are unattached. All consumers have discount factor δ_i. However, on side i, a proportion α_i of consumers are naive (N) with $\delta_i = 0$. They make decisions based on their first-period utility only. A proportion $1 - \alpha_i$ of side i's population is rational (R) with $\delta_i > 0$. They make decisions based on their lifetime utility.

A naive consumer on side i located at θ_N^i is indifferent between buying from platform 0 and platform 1 if

$$v_i + e_i n_{0,1}^j - \theta_N^i - p_{0,1}^i = v_i + e_i(1 - n_{0,1}^j) - (1 - \theta_N^i) - p_{1,1}^i,$$

which can be simplified to

$$\theta_N^i = \frac{1}{2} + \frac{1}{2}[e_i(2n_{0,1}^j - 1) + p_{1,1}^i - p_{0,1}^i].$$

As for sophisticated consumers, they also take into consideration their second-period utility. If a sophisticated consumer on side i located at θ_R^i joins platform 0 in the first period, his expected second-period utility is given by

$$U_{0,2}^i = \mu_i(v_i + e_i n_{0,2}^j - \theta_R^i - p_{0,2}^i) + (1 - \mu_i)\int_0^{\theta_0^i}(v_i + e_i n_{0,2}^j - \theta_R^i - p_{0,2}^i)dx$$
$$+ (1 - \mu_i)\int_{\theta_0^i}^1(v_i + e_i(1 - n_{0,2}^j) - (1 - \theta_R^i) - p_{1,2}^i - s_i)dx.$$

$U_{0,2}^i$ is the sum of three terms. With probability μ_i the consumer is loyal and chooses to join platform 0 in both periods; with probability $(1 - \mu_i)\theta_0^i$ he has independent preferences but still chooses to stay with platform 0; and with probability $(1-\mu)(1-\theta_0^i)$ he has independent preferences and he switches to platform 1.

Similarly, if he joins platform 1 in the first period, his expected second-period utility is given by

$$U_{1,2}^i = \mu_i(v_i + e_i(1 - n_{0,2}^j) - (1 - \theta_R^i) - p_{1,2}^i)$$
$$+ (1 - \mu_i)\int_{\theta_1^i}^1(v_i + e_i(1 - n_{0,2}^j) - (1 - \theta_R^i) - p_{1,2}^i)dx$$
$$+ (1 - \mu_i)\int_0^{\theta_1^i}(v_i + e_i n_{0,2}^j - \theta_R^i - p_{0,2}^i - s_i)dx.$$

A sophisticated consumer on side i is indifferent between purchasing from platform 0 and platform 1 if

$$v_i + e_i n_{0,1}^j - \theta_R^i - p_{0,1}^i + \delta_i U_{0,2}^i = v_i + e_i(1 - n_{0,1}^j) - (1 - \theta_R^i) - p_{1,1}^i + \delta_i U_{1,2}^i.$$

After some rearrangement, this gives

$$\theta_R^i = \frac{1}{2} + \frac{1}{2}[e_i(2n_{0,1}^j - 1) + p_{1,1}^i - p_{0,1}^i + \delta_i(U_{0,2}^i - U_{1,2}^i)].$$

The first-period market share of side i is

$$n^i_{0,1} = \alpha_i \theta^i_N + (1 - \alpha_i)\theta^i_R. \tag{3}$$

Then, we can derive the profit functions, and solve for the equilibrium prices. Calculations are rather involved and interested readers can refer to Appendix B.

I focus on the platform-symmetric equilibrium: both platforms charge the same price to each side, that is, $p^A_{0,1} = p^A_{1,1}$ and $p^B_{0,1} = p^B_{1,1}$.

Proposition 4. *The single-homing model has a symmetric equilibrium.*

Proof. See Appendix B. □

Although I focus on a symmetric equilibrium, the existence of it does not require all parameters on the two sides to be symmetric. I show the existence condition, Equations (B.1) and (B.2), in Appendix B. In the next section, I will discuss the comparative statics of the price.

3 Discussion

The analysis of the effect of switching costs on first-period prices is complicated as several effects are at play. An easier way to interpret the result is to start the discussion from pure switching-cost model (à la Klemperer) and pure two-sided market model (à la Armstrong), and then turn to the main model of the paper: a two-sided market model with switching costs. In addition, I will study other interesting ingredients such as loyalty and naivety.

3.1 Pure Switching-cost Model

In a one-sided market with switching costs, all consumers are rational; network externalities and consumers' loyalty do not matter. Assuming that $\alpha_i, \mu_i, e_i = 0$, $i \in \{A, B\}$, the first-period equilibrium price becomes

$$p^i_{0,1} = 1 + \frac{2}{3}(\underbrace{\delta_i s_i^2}_{consumer's\ anticipation} - \underbrace{\delta_F s_i}_{firm's\ anticipation}),$$

which is equivalent to Equation (18) in Klemperer (1987b).

Since the level of the first-period price is lower in a market with switching costs than without them, the literature calls it a "bargain". This pattern of attractive introductory offers followed by higher prices to exploit locked-in consumers (see Proposition 1)—the "bargains-then-ripoffs" pricing—is well-known in the switching-cost literature.

However, the extent of the bargain depends on switching costs. More specifically, the first-period price is U-shape in switching costs. There are two effects at work: On the one hand, rational consumers anticipate that if they are locked-in in the second period, the platform will raise its price. Thus, consumers are less responsive to a first-period price cut. This explains why consumers' sophistication increases the first-period price through δ_i. On the other hand, forward-looking platforms have strong incentive to invest in market share because they anticipate the benefit of having a larger customer base in the future. Platforms thus compete more aggressively to capture market share, and platforms' sophistication decreases the first-period price through δ_F. While the platform's anticipation effect is first-order in switching costs, the consumer's anticipation effect is only second-order. Therefore, the platform's anticipation effect dominates initially, the first-period price decreases with switching costs; and later the consumer's anticipation effect becomes more powerful, and thus the first-period price increases with switching costs.[12] Consequently, we get the U-shape relationship.

3.2 Pure Two-sided Market Model

In a simple model of two-sided markets, there is only one period, so that $\delta_F, \delta_i, \alpha_i = 0$; and loyalty and switching costs are irrelevant, so that $s_i, \mu_i = 0$, $i \in \{A, B\}$.

The first-period equilibrium price is simplified to

$$p^i = 1 - e_j,$$

which is the same as in Proposition 2 of Armstrong (2006). This equation shows that platforms compete fiercely for the more valuable group, whose external benefit exerted on the other group of consumers is larger.

3.3 Switching Costs in Two-sided Markets

More generally, in a two-sided market with switching costs, I find that the "bargain" can be increasing in switching costs when externalities are strong, which is different from Klemperer's result. This model is a good representation of markets such as smartphone and video games. Smartphone: switching from Apple's iOS to Google's Android system, application developers need to re-code their programs for different interfaces, as well as to create additional support and maintenance; whereas application users need to migrate and re-purchase their applications. Video games: switching from Sony's PlayStation to Windows' Xbox, gamers need to re-learn how to use the controller and lose the progress of their games, whereas developers have to buy a separate development kit to create games for different consoles.

[12] Different papers use different terminologies, for example, Somaini and Einav (2013) use "anticipation effect" and "investment incentive", while Rhodes (2013) uses "consumer elasticity effect" and "investment effect". I simply call them consumer's and firm's anticipation effect because the mechanism goes through the discount factor. My paper is quite different from Somaini and Einav (2013) and Rhodes (2013): they examine the effect of switching costs in a dynamic setting without network externalities, while I discuss a model with both switching costs and network externalities.

Proposition 5. *In the single-homing model, with all consumers and both platforms equally rational, $\delta_i = \delta_F = \delta > 0$ and $\alpha_i = 0$; independent preferences, $\mu_i = 0$; and symmetric externalities, $e_i = e > 0$, $i \in \{A, B\}$,*

i. If externalities are weak, on each side the first-period price $p^i_{0,1}$ is U-shape in switching costs s_i.

ii. If externalities are strong, on each side the first-period price $p^i_{0,1}$ is decreasing in switching costs s_i.

iii. The first-period price charged to side i, $p^i_{0,1}$, is decreasing in switching costs on side j, s_j.

Proof. See Appendix C. □

As in Klemperer (1987b), the first-period price is lower with switching costs than without, which represents a bargain. This paper, however, finds that the extent of the bargain depends not only on switching costs on one side, but also on externalities and switching costs on the other side.

More specifically, part (i) shows that when externalities are weak, we get the result of Klemperer: the bargain is inverted U-shape in switching costs. For small switching costs, rational consumers understand that they can easily switch in the second period, and are therefore more responsive to price cut in the first period. Platforms have strong incentive to compete for market share. Consequently, switching costs are pro-competitive when they are small. By contrast, when switching costs are very large, rational consumers recognize that they will be exploited in the second period, and are therefore less tempted by a price cut. Their demand becomes less elastic, and platforms will respond by charging higher prices. This explains why switching costs are anti-competitive when they are large.

Interestingly, part (ii) shows that strong externalities overturn the U-shape result: in this case the bargain is increasing in switching costs, and the positive relationship between the first-period price and switching costs does not arise. The intuition is that externalities provide an additional downward push on the first-period price because recruiting one side helps to get the other side on board. This strengthens the incentives of platforms to invest in market share, which dominates the incentive of rational consumers to avoid being locked-in. Consequently, switching costs always make the market more competitive when externalities are strong.

Part (iii) shows that an increase in switching costs on one side unambiguously decreases the first-period price charged to the other side. The reason is that platforms can build market share on side j via two channels: directly through side j, and indirectly through side i. When switching costs on side j are large, rational consumers are less responsive to price cuts because they expect a price rise to follow in the second period. An easier way to build market share on side j is then to focus on the indirect channel, i.e. attracting side i. As a result, first-period competition is increased on side i.

Proposition 5 also provides new insights into the two-sided market literature. While Armstrong (2006) shows that prices are decreasing in externalities, I focus on the effect of the interaction between network externalities and switching costs on prices.

3.4 Naive Consumers

A straightforward interpretation of naive consumers is that these consumers only care about utility in the current period. Or this could also be interpreted as the case in which consumers are different in every period.[13]

Proposition 6. *In the single-homing model, when all consumers are naive, $\delta_i = 0$ and $\alpha_i = 1$; and have independent preferences, $\mu_i = 0$, $i \in \{A, B\}$, the first-period price $p^i_{0,1}$ is decreasing in switching costs s_i regardless of the level of externalities.*

Proof. See Appendix D □

The intuition underlying this proposition is as follows. When consumers are naive, they do not anticipate that a first-period price cut will lead to a second-period price rise, and will therefore react more responsively to price cut in the first period. This increases the incentives of platforms to reduce the first-period price in order to gain more market share. Since naive consumers have no incentive to avoid being locked-in, the platform's incentive to compete for market share dominates. This explains the fierce price competition for naive consumers.

Strictly speaking, expectation about whether the others will switch play no role here because μ_i and α_i are known. In a broader sense, however, Proposition 6 can be interpreted as in line with earlier work by von Weizsäcker (1984) and Borenstein, MacKie-Mason and Netz (2000). They show that if consumers expect that a firm's price cut is more permanent than their tastes, which can be interpreted as consumers being naive, then switching costs tend to lower prices.

3.5 Heterogeneous Consumers

I now turn to discuss, rather than having all consumers being rational or naive, the consequence of having heterogeneous consumers. On each side, a fraction α_i of consumers are naive, while $1 - \alpha_i$ of them are rational; and a proportion μ_i consumers are loyal, while the remaining ones have independent preferences.[14]

Proposition 7. *In the single-homing model,*

[13] For example, a company buys some software for their workers in the first period. Some workers leave the company in the second period, and purchase their own software. These workers have a switching cost of learning some new software that are different from that purchased by their company, but the company will not take into consideration this switching cost when buying in the first period.

[14] Gabszewicz, Pepall and Thisse (1992) also discuss heterogeneous consumers in terms of brand loyalty, but they consider the pricing strategy of a monopoly incumbent, who anticipates the entry of a rival in the subsequent period, and focus on the effect of loyalty on entry.

i. On each side, the first-period price $p^i_{0,1}$ is decreasing in the proportion of naive consumers α_i, if the proportion of loyal consumers, μ_i, is high.

ii. The first-period price on side i, $p^i_{0,1}$, is increasing in the proportion of naive consumers on side j, α_j.

iii. The first-period price $p^i_{0,1}$ is decreasing in the discount factor of the platform δ_F.

Proof. See Appendix E. □

The intuition behind this proposition is as follows. Part (i) shows that on each side, if there are many loyal consumers, the first-period price is lower with naive consumers than without.[15] The reason is that after consumers make their purchase in the first period, consumers who are loyal know that they will patronize the same platform for an indefinite period of time, and feel that they deserve a bigger carrot in the first period. Naive consumers, who care only about today, are more attracted by a price cut. Therefore, increasing the proportion of consumers who are loyal and naive makes the market more competitive in the first period.

Part (ii) shows that an increase in the proportion of consumers who are naive on one side will soften price competition on the other side. Intuitively, the demand of naive consumers on side j is more elastic, and platforms will react by charging lower prices. This, in turn, reduces the opportunity cost of recruiting consumers on side i. Platforms thus compete less aggressively for market share on side i. Consequently, consumers' naivety on one side mitigates the ferocity of first-period competition for market share on the other side.

Part (iii) shows that first-period prices are lower when platforms are more patient. Platforms compete harder on prices because they foresee the advantage of having a large customer base in the future.

More generally, Propositions 5 and 7 say that the strategy of lowering price is not simply due to network externalities in a two-sided market, a view that is central to the work of Rochet and Tirole (2003), and Armstrong (2006). But in my model whether the platform will act more aggressively also depends on the characteristics of consumers and their switching costs. This has important implications on regulations that alter switching costs and loyalty rate in real circumstances, which will be explored more fully in Section 4.

3.6 A Special Case: asymmetric sides

The model also covers the case of asymmetric sides, where consumers on one side, say side-B, do not incur any switching costs in the second period ($s_B = 0$). Examples of such a market include browsers, search engines, and shopping malls. Browsers: Internet users can switch relatively more easily between Internet Explorer, Chrome, and Firefox than content providers because when

[15] If consumers' tastes change ($\mu < 1$), it may nullify the competitive effect of naivety.

content providers switch, they need to rewrite the codes so that they are compatible with the new browser. Search engines: customers can switch easily between Google, Bing and Yahoo in as little as one click, but there are switching costs for top-listed publishers, who want their website to appear on the top list of another search engine. Shopping malls: shoppers are free to go to any shopping malls, but there are high transaction costs for shops in terminating the old contract and initiating a new one.

For simplicity, assume that consumer preferences are independent, $\mu_i = 0$; all consumers are rational, $\alpha_i = 0$; and they have the same discount factor as the firm, $\delta_i = \delta_F = \delta$, $i \in \{A, B\}$.

Corollary 1. *If only one side of consumers has switching costs, then switching costs only affect the price on this side but not the other side.*

Proof. Under the assumptions above,

$$p^B_{0,1} = 1 - e_A.$$

□

The intuition is that since preferences of side-B consumers in the two periods are unrelated and they do not have switching costs, every period's choice is independent. This means that the first-period price is not affected by the second-period price. Consequently, although side-A consumers' switching costs affect side-B's second-period price through externalities, it does not affect side-B's first-period price.

3.7 Effect of Switching Costs on First-period Profit

In a platform-symmetric equilibrium, the two platforms share consumers on each side equally, that is $n^A_{0,1} = n^B_{0,1} = 1/2$. Therefore, the expected profit of platform 0 is

$$\pi_0 = \frac{1}{2}p^A_{0,1} + \frac{1}{2}p^B_{0,1} + \delta\pi_{0,2},$$

where $\pi_{0,2}$ is the second-period profit.

Differentiating π_0 with respect to s_i, we obtain

$$\frac{\partial \pi_0}{\partial s_i} = \frac{1}{2}\frac{\partial p^i_{0,1}}{\partial s_i} + \frac{1}{2}\frac{\partial p^j_{0,1}}{\partial s_i}$$

because the profit in the last period, $\pi_{0,2}$, is not affected by s_i in equilibrium.

As is well-known from the switching-cost literature, switching costs raise platforms' profits in the second period compared to the case of no switching costs as second-period prices are usually higher. However, the presence of market power over locked-in consumers intensifies competition in the first period, and this may result in a decrease in overall profit.[16]

[16] See for instance Klemperer (1987a).

More interestingly, I identify an additional channel through which switching costs can reduce overall profit, namely, when network externalities are strong. The reason is that strong externalities increase the incentives of platforms to vie for market share, and therefore switching costs on side i intensify price competition on side i (see (ii) of Proposition 5). Higher switching costs on side i also lead to more competitive behavior on side j because capturing more consumers on side j is a cheaper way to build market share on side i. Side i consumers are harder to attract as they have strong incentives to avoid being locked-in and thus paying large switching costs in the second period (see (iii) of Proposition 5). Higher switching costs lower prices on both sides, and thereby reducing overall profit.

4 Welfare and Policy Implications

The first-period welfare is constant in switching costs because all consumers buy one unit of good, the size of the two groups is fixed, and the whole market is served. It ignores the possible demand-expansion and demand-reduction effects of switching costs as the total demand is fixed. However, the second-period welfare is decreasing in switching costs. The welfare loss is the sum of two deadweight losses:

$$2(1-\mu_i)[\quad \underbrace{(\frac{1-s_i}{2})s_i}_{DWL\ from\ switchers} \quad + \quad \underbrace{\frac{s_i^2}{4}}_{DWL\ from\ non-switchers} \quad].$$

Consider consumers who have independent preferences. Since their tastes will change in the second period, for those who have previously bought from platform 0, consumers whose tastes change a lot will switch to platform 1 with probability $(1-s_i)/2$ and each pays s_i; consumers whose tastes change a little will continue to buy from platform 0 even though they prefer platform 1. This happens with probability $s_i/2$ and each suffers an average loss of mismatch with an inferior product $s_i/2$. Similarly, consider consumers who have previously bought from platform 1. Consumers on both sides suffer this loss. As for loyal consumers, there is no loss for them because first, they do not switch; second, their preferences do not change, and hence there is no deadweight loss associated with mismatch.[17]

Although switching costs lower social welfare, from the consumer welfare point of view, consumers may still benefit from switching costs if the equilibrium price is lower. I therefore suggest the following policy implications. In one-sided markets, attractive introductory offers that induce early adoption may call for consumer protection in later periods, for example, through compatibility or standardization policies that lower switching costs. In two-sided markets, asymmetric price structures are common because they help to increase the participation of different groups

[17] Naivety does not affect welfare. The only thing that matters for welfare is whether consumers' preferences change or not. When consumers' preferences do not change, they make the right product choice and do not switch. When consumers' preferences change, switchers have to incur the switching costs, and some of the non-switchers are forced into buying an inferior product that does not match their tastes.

of consumers. For example, Proposition 2 shows that s_j may have a positive or negative impact on $p^i_{0,2}$, and Proposition 5 shows that the relationship between $p^i_{0,1}$ and s_i depends on e, and $p^i_{0,1}$ decreases with s_j. Therefore, when policy-makers alter switching costs of one group, it may have broader repercussions on the other group; sticking to a one-sided logic may lead to inefficient policies.[18]

In this model, I assume that all consumers know their preferences in the current period, but tastes of some consumers may change. One could alternatively interpret a fraction μ_i of consumers know their preferences, while $1 - \mu_i$ of them do not know theirs. Disloyal consumers receive a signal about their tastes in the first period, and after buying from the platform, they know their tastes in the second period. This would not change the result as long as the signal is uniformly distributed. This allows us to evaluate the effect of information transparency policy. For example, Proposition 7 shows that loyalty makes it more likely that naivety will hurt the platform. Thus, platforms may lack incentive to enhance consumers' understanding of their own preferences. They might try to provide imprecise information about consumers' tastes, so that consumers are less loyal, and they will switch more, which platforms can exploit later. Therefore, there is room for government intervention. In particular, increasing transparency of information would enable disloyal consumers to make choices that are best aligned to their tastes, build loyalty and save switching costs.

5 Extensions

The analysis so far is based on a single-homing model, but this is not the only market configuration in reality. There are various ways to extend the model, for instance, one may consider the case where one group single-homes while the other group join both (commonly termed as "competitive bottlenecks"). It might also be interesting to consider asymmetric platforms. I will sketch these extensions in turn.

5.1 Competitive Bottlenecks

Suppose that side A continues to single-home, while side B may multi-home. Competitive bottleneck framework is typical in markets such as computer operating systems, and online air ticket and hotel bookings. Operating systems: users use a single OS, Windows OS, Apple's Mac OSX platform or Linux-based OS, while engineers develop software for different OS. Travel bookings: consumers use one comparison site such as skyscanner.com, lastminute.com or booking.com, but airlines and hotels join multiple platforms in order to gain access to each comparison site's customers. Since side B can join both platforms, switching costs and loyalty on this side are

[18] Wright (2004) also shows that analyzing a two-sided market as if it was a one-sided market may lead to some policy errors. Different from him, however, this paper identifies some new issues raised by switching costs in two-sided markets that have not been discussed previously.

not relevant, so that $s_B, \mu_B = 0$.[19] The main difference from the single-homing model lies in the market share of side-B consumers, which can be described as follows. In period t, $t \in \{1, 2\}$, a side-B consumer located at $\theta^B_{0,t}$ is indifferent between buying and not buying from platform 0 if

$$v_B + e_B n^A_{0,t} - \theta^B_{0,t} - p^B_{0,t} = 0,$$

which can be simplified to

$$\theta^B_{0,t} = v_B + e_B n^A_{0,t} - p^B_{0,t}.$$

Similarly, a side-B consumer located at $\theta^B_{1,t}$ is indifferent between buying and not buying from platform 1 if

$$v_B + e_B(1 - n^A_{0,t}) - (1 - \theta^B_{1,t}) - p^B_{1,t} = 0,$$

which can be simplified to

$$\theta^B_{1,t} = v_B + e_B(1 - n^A_{0,t}) - p^B_{1,t}.$$

We solve the game by backward induction as before. Consider the symmetric equilibrium. Appendix F proves the existence of it. We can then derive the equilibrium prices.

Proposition 8. *In the multi-homing model, with all consumers and both platforms equally rational, $\delta_i = \delta_F = \delta > 0$ and $\alpha_i = 0$; independent preferences, $\mu_i = 0$; and symmetric externalities, $e_i = e > 0$, $i \in \{A, B\}$,*

i. For the single-homing consumers, if externalities are weak, the first-period price $p^A_{0,1}$ is U-shape in switching costs s_A. If externalities are strong, the first-period price $p^A_{0,1}$ is decreasing in s_A.

ii. If the market is fully covered, then first-period prices tend to be higher on the multi-homing side and lower on the single-homing side with respect to the single-homing model in Section 3.6.

Proof. See Appendix F. □

[19] Note that the concept of multi-homing is not compatible with switching costs in the current framework. I use two examples to illustrate. First, think of the smartphone market. If the option to multi-home means consumers are able to use both iPhone and Android systems, then it is not reasonable to impose an additional learning cost on them if they switch platform. Another example is the media market. If multi-homing means that advertisers are free to put ads on either or both platforms, then it does not make sense to impose an additional switching cost on these advertisers. One may argue that we can distinguish between learning switching costs (incurred only at a switch to a new supplier) and transactional switching costs (incurred at every switch), as in Nilssen (1992), but switching costs are not relevant on the multi-homing side because learning costs and transaction costs are equivalent in a two-period model. This also explains why it is not useful to consider the case in which both sides multi-home.

Part (i) implies that for single-homing consumers stronger externalities make it more likely that first-period equilibrium prices decrease with switching costs, which is consistent with Proposition 5 in the single-homing model. As for multi-homing consumers, both switching costs and the degree of sophistication do not affect the price paid by them because each period's choice is independent. This case and the previous case of asymmetric sides have similar intuition because $s_B, \mu_B = 0$. Part (ii) is different from results in the single-homing model. Since side B multi-homes, there is no competition between the two platforms to attract this group. Compared with the case of asymmetric sides, the higher first-period price faced by the multi-homing side is a consequence of each platform having monopoly power over this side, and the large revenue is used in the form of lower first-period price to convince the single-homing side to join the platform.

Before, in the single-homing model, switching costs do not affect the first-period welfare, but lowers the second-period welfare. However, in the multi-homing model switching costs affect first-period welfare through participation, which is, in turn, determined by the price. In the second period, switching cost has no effect on price because platforms have an equal share of the market, and their incentives to exploit old customers offset their incentives to poach new customers. If switching costs reduce first-period price (see (i) of Proposition 8, especially when externalities are strong), then switching costs may increase welfare.[20] This is because lower price induces more consumers to multi-home, and more multi-homing consumers increases the utility of single-homing consumers.

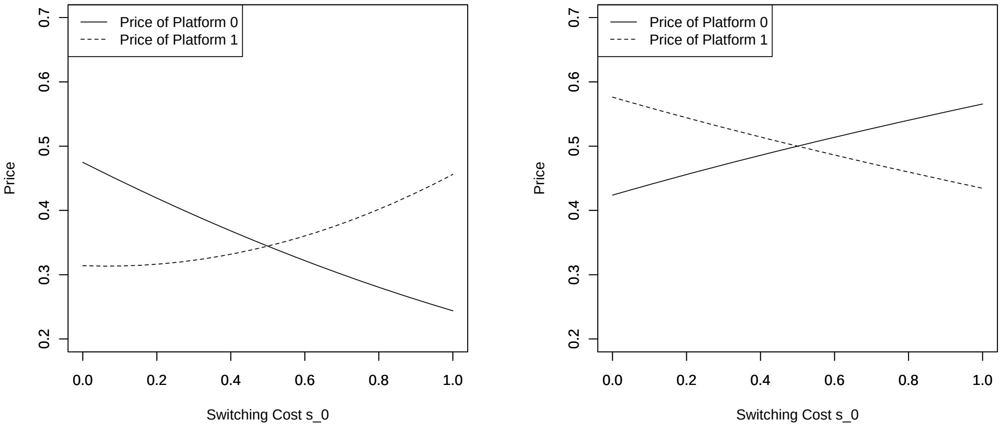
5.2 Asymmetric Platforms

Let us now consider asymmetric platforms. The cost of switching from platform 0 to 1, denoted s_0, is different from the cost of switching from platform 1 to 0, denoted s_1. As an example, some say "iPhones are more expensive than most Samsung smartphones."[21] Can we attribute the difference in the pricing of devices between Apple and Samsung to the fact that Apple has successfully built an ecosystem that makes users hard to switch? To address this question, consider two groups of consumers who are asymmetric in the sense that only consumers on side A incur switching costs in the second period. For simplicity, assume that all consumers single-home. Consider a numerical example where $\delta_A = \delta_B = \delta_F = 0.8$, $\mu_A = \mu_B = 0$, $e_A = e_B = 0.5$, $s_1 = 0.5$, and $s_0 \in [0, 1]$.

The results are illustrated in Figure 1. Panel (a) presents the first-period pricing, and panel (b) shows the second-period pricing as functions of switching costs s_0. Pricing of platform 0 is shown with a solid line, and that of platform 1 is drawn as a dotted line. It is shown that if $s_0 < s_1$, platform 1 charges a lower price than platform 0 in the first period, but a higher price in the second period. The intuitive reason is that since platform 1 is relatively more expensive to switch away from in the second period, it is willing to charge a lower price in the first period in

[20] If there is quality choice as in Anderson et al. (2013), then welfare effects are less clear-cut: platform's investment in quality may change depending on whether multi-homing is allowed.

[21] NBC News, "Apple is biggest US phone seller for first time," 1 February 2013, by Peter Svensson. http://www.nbcnews.com/technology/apple-biggest-us-phone-seller-first-time-1B8210244



(a) First-period Prices

(b) Second-period Prices

Figure 1: Equilibrium Pricing with Asymmetric Platforms.

order to acquire more customers whom it can exploit later. On the contrary, if $s_0 > s_1$, platform 1, knowing that consumers will easily switch away tomorrow, will raise its price today. This result holds as long as externalities are not too strong.[22]

6 Conclusion

This paper has characterized the equilibrium pricing strategy of platforms competing in two-sided markets with switching costs. The main contribution is that it has provided a useful model for generalizing arguments already used in the switching-cost and the two-sided market literature, and for extending beyond traditional results. In line with earlier research, there are some conditions under which switching costs reduce first-period prices but increase second-period prices (à la Klemperer); and prices tend to be lower on the side that exerts a stronger externalities (à la Armstrong). However, this model develops the idea further by proving that in a dynamic two-sided market—as opposed to a merely static one—under weak externalities, switching costs soften price competition in the first period if consumers are significantly more patient than the platforms; under strong externalities, switching costs always make the market more competitive. In terms of consumer heterogeneity, the presence of more loyal and naive consumers on one side intensifies price competition in the first period on this side.

The analysis could be extended in a number of different directions. First, this paper has taken switching costs as an exogenous feature of the market. Future research could consider endogenous switching costs. Second, this paper has focused on a two-period model, and it would be useful to understand the extent to which the results carry over to a multi-period model. Finally, this paper has explored heterogeneity such as loyalty and naivety, but one can think of other

[22] For large externalities ($e \to 1$), symmetric equilibrium does not exist because there is coordination problem. Given that externalities are so strong, all consumers might want to join one platform only.

forms of heterogeneity across consumers. For example, within-group switching costs may be different between the technologically advanced customers and the less advanced ones. Within-group externalities may also be different: youngsters use applications more heavily, and therefore care more about network externalities than their older counterparts, many of whom only use their smartphones for phone calls and text messages. However, including these forms of heterogeneity will complicate the analysis considerably. The current model captures a lot of ingredients in reality, yet is sufficiently tractable to allow for a complete characterization of the equilibrium. This seems to be a reasonable first step to extend a literature that has not fully explored the implications of consumer heterogeneity.

A Second Period Equilibrium

Solving for $n^A_{0,2}$ and $n^B_{0,2}$ in Equation (2) simultaneously, we obtain the second-period market shares as follows:

$$n^i_{0,2} = \frac{\gamma + \beta_i + (1-\mu_i)(p^i_{1,2} - p^i_{0,2}) + e_i(1-\mu_i)(1-\mu_j)(p^j_{1,2} - p^j_{0,2})}{2\gamma},$$

where

$$\begin{aligned}\gamma =& 1 - (1-\mu_A)(1-\mu_B)e_A e_B,\\ \beta_i =& (2n^i_{0,1} - 1)(\mu_i + (1-\mu_i)s_i) + (2n^j_{0,1} - 1)(1-\mu_i)e_i(\mu_j + (1-\mu_j)s_j).\end{aligned}$$

Because $e_i < 1$, we have $\gamma > 0$.

Substituting the market shares into the profit function in Equation (1), and differentiating it with respect to the prices, we obtain the following equations.

$$\begin{aligned}\frac{\partial \pi_{0,2}}{\partial p^i_{0,2}} &= n^i_{0,2} - \frac{p^i_{0,2}}{2\gamma}(1-\mu_i) - \frac{p^j_{0,2}}{2\gamma}e_j(1-\mu_i)(1-\mu_j),\\ \frac{\partial \pi_{1,2}}{\partial p^i_{1,2}} &= 1 - n^i_{0,2} - \frac{p^i_{1,2}}{2\gamma}(1-\mu_i) - \frac{p^j_{1,2}}{2\gamma}e_j(1-\mu_i)(1-\mu_j).\end{aligned}$$

Solving the system of first-order conditions, one finds the following second-period equilibrium prices.

$$\begin{aligned}p^i_{0,2} &= \frac{1 - e_j(1-\mu_i)}{1-\mu_i} + \frac{\eta_i\lambda_i + \epsilon_i\lambda_j}{(1-\mu_i)\Delta},\\ p^i_{1,2} &= \frac{1 - e_j(1-\mu_i)}{1-\mu_i} - \frac{\eta_i\lambda_i + \epsilon_i\lambda_j}{(1-\mu_i)\Delta}.\end{aligned} \tag{A.1}$$

where

$$\begin{aligned}
&\Delta = 9 - (1-\mu_A)(1-\mu_B)(e_A + 2e_B)(e_B + 2e_A) > 0,\\
&\lambda_i = (2n^i_{0,1} - 1)(\mu_i + (1-\mu_i)s_i),\\
&\eta_i = 3 - e_j(e_j + 2e_i)(1-\mu_i)(1-\mu_j) > 0,\\
&\epsilon_i = (1-\mu_i)(e_i - e_j).
\end{aligned}$$

A.1 Proof of Proposition 1

Differentiate Equation (A.1) with respect to s_i, we have

$$\begin{aligned}
\text{sign}\,\frac{\partial p^i_{0,2}}{\partial s_i} =&\,\text{sign}(n^i_{0,1} - \frac{1}{2}),\\
\frac{\partial p^i_{0,2}}{\partial s_i} =& -\frac{\partial p^i_{1,2}}{\partial s_i}.
\end{aligned}$$

A.2 Proof of Proposition 2

Differentiate Equation (A.1) with respect to s_j, we have

$$\begin{aligned}
\text{sign}\,\frac{\partial p^i_{0,2}}{\partial s_j} =&\,\text{sign}(e_i - e_j)(n^j_{0,1} - \frac{1}{2}),\\
\frac{\partial p^i_{0,2}}{\partial s_j} =& -\frac{\partial p^i_{1,2}}{\partial s_j}.
\end{aligned}$$

A.3 Proof of Proposition 3

The second-period profit of platform 0 is

$$\begin{aligned}
\pi_{0,2} &= p^A_{0,2}n^A_{0,2} + p^B_{0,2}n^B_{0,2}\\
&= \left[\frac{1-e_B(1-\mu_A)}{1-\mu_A} + \frac{\eta_A\lambda_A + \epsilon_A\lambda_B}{(1-\mu_A)\Delta}\right]\left[\frac{1}{2} + \frac{3\lambda_A + (1-\mu_A)(e_A+2e_B)\lambda_B}{2\Delta}\right]\\
&\quad + \left[\frac{1-e_A(1-\mu_B)}{1-\mu_B} + \frac{\eta_B\lambda_B + \epsilon_B\lambda_A}{(1-\mu_B)\Delta}\right]\left[\frac{1}{2} + \frac{3\lambda_B + (1-\mu_B)(e_B+2e_A)\lambda_A}{2\Delta}\right].
\end{aligned}$$

The first-order conditions with respect to s_A and s_B are

$$\frac{\partial \pi_{0,2}}{\partial s_i} = \frac{\partial \pi_{0,2}}{\partial \lambda_i}(2n^i_{0,1} - 1)(1-\mu_i),$$

where

$$\begin{aligned}\frac{\partial \pi_{0,2}}{\partial \lambda_i} =& \frac{\eta_i}{(1-\mu_i)\Delta}\left[\frac{1}{2} + \frac{3\lambda_i + (1-\mu_i)(e_i+2e_j)\lambda_j}{2\Delta}\right] \\ &+ \frac{3}{2\Delta}\left[\frac{1-e_j(1-\mu_i)}{1-\mu_i} + \frac{\eta_i\lambda_i + \epsilon_i\lambda_j}{(1-\mu_i)\Delta}\right] \\ &+ \frac{\epsilon_j}{(1-\mu_j)\Delta}\left[\frac{1}{2} + \frac{3\lambda_j + (1-\mu_j)(e_j+2e_i)\lambda_i}{2\Delta}\right] \\ &+ \frac{(1-\mu_j)(e_j+2e_i)}{2\Delta}\left[\frac{1-e_i(1-\mu_j)}{1-\mu_j} + \frac{\eta_j\lambda_j + \epsilon_j\lambda_i}{(1-\mu_j)\Delta}\right].\end{aligned}$$

Therefore,

$$\operatorname{sign}\frac{\partial \pi_{0,2}}{\partial s_i} = \operatorname{sign}(n^i_{0,1} - \frac{1}{2})$$

if

$$\frac{\partial \pi_{0,2}}{\partial \lambda_i} > 0.$$

For $\partial \pi_{0,2}/\partial \lambda_i > 0$, we need $n^A_{0,1}$ and $n^B_{0,1}$ not too close to zero, as well as e_A and e_B are not too different.

B First Period Equilibrium

The indifferent rational consumer is given by

$$\theta^i_R = \frac{1}{2} + \frac{e_i(2n^j_{0,1}-1) + p^i_{1,1} - p^i_{0,1} + \delta_i(\mu_i + (1-\mu_i)s_i)\frac{[(1-\mu_i)(e_i+2e_j)\lambda_j + (3-\Delta)\lambda_i]}{(1-\mu_i)\Delta}}{2(1+\delta_i\mu_i)}.$$

Substitute θ^i_N and θ^i_R into Equation (3), and solve simultaneously for $n^A_{0,1}$ and $n^B_{0,1}$:

$$n^i_{0,1} = \frac{1}{2} + \frac{e_i(1-\kappa_j)(p^i_{1,1} - p^i_{0,1}) + \tau_j(e_i\tau_i + \sigma_i)(p^j_{1,1} - p^j_{0,1})}{2[(1-\kappa_i)(1-\kappa_j) - (e_i\tau_i + \sigma_i)(e_j\tau_j + \sigma_j)]},$$

where

$$\begin{aligned}\tau_i =& \alpha_i + \frac{1-\alpha_i}{1+\delta_i\mu_i}, \\ \kappa_i =& \frac{\delta_i(\mu_i + (1-\mu_i)s_i)(3-\Delta)(1-\alpha_i)(\mu_i + (1-\mu_i)s_i)}{(1-\mu_i)\Delta(1+\delta_i\mu_i)}, \\ \sigma_i =& \frac{\delta_i(\mu_i + (1-\mu_i)s_i)(e_i+2e_j)(1-\alpha_i)(\mu_j + (1-\mu_j)s_j)}{\Delta(1+\delta_i\mu_i)}.\end{aligned}$$

The expected profit of platform 0 is

$$\pi_0 = p^A_{0,1}n^A_{0,1} + p^B_{0,1}n^B_{0,1} + \delta_F\pi_{0,2}.$$

The first-order conditions for maximizing π_0 with respect to $p_{0,1}^A$ and $p_{0,1}^B$ are given as follows.

$$\frac{\partial \pi_0}{\partial p_{0,1}^i} = n_{0,1}^i - p_{0,1}^i \frac{\tau_i(1-\kappa_j)}{2\varphi} - p_{0,1}^j \frac{\tau_i(e_j\tau_j + \sigma_j)}{2\varphi} + \delta_F \left[\frac{\partial \pi_{0,2}}{\partial n_{0,1}^i}\frac{\partial n_{0,1}^i}{\partial p_{0,1}^i} + \frac{\partial \pi_{0,2}}{\partial n_{0,1}^j}\frac{\partial n_{0,1}^j}{\partial p_{0,1}^i}\right]$$

where

$$\begin{aligned}\varphi =& (1-\kappa_i)(1-\kappa_j) - (e_i\tau_i + \sigma_i)(e_j\tau_j + \sigma_j),\\ \frac{\partial \pi_{0,2}}{\partial n_{0,1}^i} =& \left[\frac{6}{(1-\mu_i)\Delta} + \frac{(e_i - e_j) - (e_i + e_j)(e_j + 2e_i)(1-\mu_j)}{\Delta}\right](\mu_i + (1-\mu_i)s_i) \overset{\text{def}}{=} \xi_i.\end{aligned}$$

Similarly, there are two first-order conditions for platform 1.

I focus on the platform-symmetric equilibrium, where $p_{0,1}^A = p_{1,1}^A = p^A$ and $p_{0,1}^B = p_{1,1}^B = p^B$. I derive the sufficient condition for the existence of such symmetric equilibrium, which requires that platform k's profit is concave in its prices. The concavity condition is as follows.

$$1 - \kappa_A > e_A\tau_A + \sigma_A > 0; \quad 1 - \kappa_B > e_B\tau_B + \sigma_B > 0. \tag{B.1}$$

In addition to Equation (B.1), to ensure that the platform does not deviate from the equilibrium price, we need the following condition:

$$v_i + \frac{1}{2}e_i - \frac{1}{2} > \frac{1}{1-\mu_i} - e_i > (v_i + \frac{1}{2}e_i - \frac{1}{2})\mu_i, \quad i \in \{A, B\}. \tag{B.2}$$

The first inequality means that we need v_i to be big enough such that the market is covered. The second inequality means that we need μ_i to be small enough and v_i to be big, but not too big, in order to guarantee that the platform does not deviate to serve only loyal consumers in the second period. For example, Equations (B.1) and (B.2) are satisfied when α_i is big and/or $\mu_i = 0$ is small.[23]

Under symmetric equilibrium, the first-period equilibrium prices for side A and side B are given respectively by

$$p_{0,1}^A = \frac{1-\kappa_A}{\tau_A} - \frac{\sigma_B}{\tau_B} - e_B - \delta_F\xi_A; \quad p_{0,1}^B = \frac{1-\kappa_B}{\tau_B} - \frac{\sigma_A}{\tau_A} - e_A - \delta_F\xi_B, \tag{B.3}$$

and the second-period equilibrium prices are given by

$$p_{0,2}^A = \frac{1-e_B(1-\mu_A)}{1-\mu_A}; \quad p_{0,2}^B = \frac{1-e_A(1-\mu_B)}{1-\mu_B}.$$

C Proof of Proposition 5

If $\delta_A = \delta_B = \delta_F = \delta > 0$, $\alpha_A = \alpha_B = 0$, $\mu_A = \mu_B = 0$, and $e_A = e_B = e > 0$, Equation (B.3) becomes

$$p_{0,1}^i = 1 - e + \frac{\delta}{3(1-e^2)}\left[(2-3e^2)s_i^2 - 2(1-e^2)s_i - es_is_j\right].$$

[23] When $\alpha_i = 1$, we obtain the same existence condition for a symmetric equilibrium as in Armstrong (2006). I show that the equilibrium exists for a wider range of parameters.

Differentiating $p^i_{0,1}$ with respect to s_i, we obtain

$$\begin{aligned}
\frac{\partial p^i_{0,1}}{\partial s_i} &= \frac{\delta}{3(1-e^2)}\left[2(2-3e^2)s_i - 2(1-e^2) - es_j\right], \\
\frac{\partial^2 p^i_{0,1}}{\partial s_i^2} &= \frac{2\delta(2-3e^2)}{3(1-e^2)}\begin{cases} > 0 & \text{if } e < \sqrt{2/3}, \\ < 0 & \text{if } e \geq \sqrt{2/3}, \end{cases} \\
\frac{\partial p^i_{0,1}}{\partial s_i}|_{s_i=0} &= \frac{\delta}{3(1-e^2)}\left[-2(1-e^2) - es_j\right] < 0.
\end{aligned}$$

Therefore, $p^i_{0,1}$ is U-shape in s_i if $e < \sqrt{2/3}$, and decreasing in s_i if $e \geq \sqrt{2/3}$.

Differentiating $p^i_{0,1}$ with respect to s_j, we get

$$\frac{\partial p^i_{0,1}}{\partial s_j} = -\frac{\delta e s_i}{3(1-e^2)} < 0.$$

Therefore, $p^i_{0,1}$ is decreasing in s_j.

D Proof of Proposition 6

If $\delta_A = \delta_B = 0$, $\alpha_A = \alpha_B = 1$, and $\mu_A = \mu_B = 0$, Equation (B.3) becomes

$$p^i_{0,1} = 1 - \delta_F\left[\frac{6 + e_i - e_j - (e_i + e_j)(e_j + 2e_i)}{\Delta}\right] s_i - e_j.$$

Differentiating it with respect to s_i, we obtain

$$\frac{\partial p^i_{0,1}}{\partial s_i} < 0.$$

E Proof of Proposition 7

Differentiating Equation (B.3) with respect to α_A, α_B and δ_F, we obtain the following:

$$\frac{\partial p^i_{0,1}}{\partial \alpha_i}\begin{cases} \leq 0, & \text{if } \mu_i \to 1 \text{ or } e_i, e_j \to 0, \\ > 0, & \text{if } \mu_i \to 0 \text{ and } e_i, e_j \to 1, \end{cases}$$

since

$$\frac{\partial p^i_{0,1}}{\partial \alpha_i} > 0 \text{ if } \frac{\mu_i + 2\mu_i(1-\mu_i)s_i + (1-\mu_i)^2 s_i^2}{\mu_i^2 + 3\mu_i(1-\mu_i)s_i + (1-\mu_i)^2 s_i^2} > \frac{\Delta}{3}.$$

$$\begin{aligned}
\frac{\partial p^i_{0,1}}{\partial \alpha_j} &\geq 0. \\
\frac{\partial p^i_{0,1}}{\partial \delta_F} &= -\xi_i \leq 0.
\end{aligned}$$

F Proof of Proposition 8

The first-order conditions of π_k, $k \in \{0,1\}$, with respect to $p^A_{0,1}$ and $p^B_{0,1}$ are, respectively,

$$n^A_{k,1} - \frac{1}{2\omega}p^A_{k,1} - \frac{e}{2\omega}p^B_{k,1} - \frac{\delta}{2\omega}\frac{\partial \pi_{k,2}}{\partial n^A_{0,1}} = 0,$$
$$n^B_{k,1} - (1+\frac{e^2}{2\omega})p^B_{k,1} - \frac{e}{2\omega}p^A_{k,1} - \frac{\delta e}{2\omega}\frac{\partial \pi_{k,2}}{\partial n^A_{0,1}} = 0,$$

where

$$\omega = 1 - e^2 - \frac{\delta s_A^2(e^2-2\gamma)}{3\gamma}.$$

Using similar proof as in the single-homing model, the symmetric equilibrium exists in the multi-homing model. The existence conditions are as follows. First, platform k's profit is concave in its prices if $\omega \geq 0$, which means that δ, s_A and e are not too big.

Second, we need to ensure that the platform does not deviate to sell only to loyal consumers on side A.

$$v_A + e(\frac{v_B}{2}+\frac{e}{2}) - \frac{1}{2} > 1 - (1-\mu_A)e^2 - \frac{ev_B}{2} > \left[v_A + e(\frac{v_B}{2}+\frac{e}{2}) - \frac{1}{2}\right]\mu_A,$$

or equivalently, μ_A is small, and v_B is big, but not too big.

The first-period equilibrium prices are as follows.

$$p^A_{0,1} = 1 - e^2 - \frac{\delta(3e^2-2)s_A^2}{3(1-e^2)} - \frac{2\delta s_A}{3} - \frac{v_B e}{2},$$
$$p^B_{0,1} = \frac{v_B}{2}.$$

For part (i), differentiate $p^A_{0,1}$ with respect to s_A.

$$\frac{\partial p^A_{0,1}}{\partial s_A} = -\frac{2\delta}{3} - \frac{2\delta(3e^2-2)s_A}{3(1-e^2)},$$
$$\frac{\partial^2 p^A_{0,1}}{\partial s_A^2} = -\frac{2\delta(3e^2-2)}{3(1-e^2)}\begin{cases} >0 & \text{if } e < \sqrt{2/3}, \\ <0 & \text{if } e \geq \sqrt{2/3},\end{cases}$$
$$\frac{\partial p^A_{0,1}}{\partial s_A}|_{s_A=0} = -\frac{2\delta}{3} < 0.$$

Therefore, $p^A_{0,1}$ is U-shape in s_A if $e < \sqrt{2/3}$, and decreasing in s_A if $e \geq \sqrt{2/3}$.

For part (ii), we compare the first-period prices paid by consumers who bear switching costs (side-A) and those who do not (side-B) in the multi-homing model (denoted mh) with that in the single-homing model in Section 3.6 (denoted sh).

For side-A,

$$p^A_{mh} < p^A_{sh} \text{ if } e + \tfrac{v_B}{2} > 1.$$

For side-B,

$$p^B_{mh} > p^B_{sh} \text{ if } e + \tfrac{v_B}{2} > 1.$$

References

[1] Simon Anderson, Øystein Foros, and Hans Jarle Kind. Competition for Advertisers in Media Markets. In *Fourteenth CEPR/JIE Conference on Applied Industrial Organization*, 2013.

[2] Mark Armstrong. Competition in two-sided markets. *RAND Journal of Economics*, 37(3):668–691, 2006.

[3] Gary Biglaiser and Jacques Crémer. Switching Costs and Network Effects in Competition Policy. *Recent advances in the analysis of competition policy and regulation*, 2014, forthcoming.

[4] Gary Biglaiser, Jacques Crémer, and Gergely Dobos. The value of switching costs. *Journal of Economic Theory*, 148(3):935–952, 2013.

[5] Severin Borenstein, Jeffrey MacKie-Mason, and Janet Netz. Exercising Market Power in Proprietary Aftermarkets. *Journal of Economics and Management Strategy*, 9(2):157–188, 2000.

[6] Bernard Caillaud and Bruno Jullien. Chicken & egg: competition among intermediation service providers. *RAND Journal of Economics*, 34(2):309–328, 2003.

[7] Yongmin Chen. Paying Customers to Switch. *Journal of Economics & Management Strategy*, 6(4):877–897, 1997.

[8] Liran Einav and Paulo Somaini. A Model of Market Power in Customer Markets. *Journal of Industrial Economics*, 2013, forthcoming.

[9] Joseph Farrell and Paul Klemperer. Coordination and Lock-in: Competition with Switching Costs and Network Effects. In Mark Armstrong and Rob Porter, editors, *Handbook of Industrial Organization*, volume 3, chapter 31, pages 1967–2072. North-Holland, 2007.

[10] Jean Gabszewicz, Lynne Pepall, and Jacques-François Thisse. Sequential Entry with Brand Loyalty Caused by Consumer Learning-by-Using. *Journal of Industrial Economics*, 40(4):397–416, 1992.

[11] Paul Klemperer. Markets with Consumer Switching Costs. *Quarterly Journal of Economics*, 102(2):375–394, 1987a.

[12] Paul Klemperer. The competitiveness of markets with switching costs. *RAND Journal of Economics*, 18(1):138–150, 1987b.

[13] Paul Klemperer. Competition when Consumers have Switching Costs: An Overview with Applications to Industrial Organization, Macroeconomics, and International Trade. *Review of Economic Studies*, 62:515–539, 1995.

[14] Tore Nilssen. Two Kinds of Consumer Switching Costs. *RAND Journal of Economics*, 23(4):579–589, 1992.

[15] Office of Fair Trading. Switching costs. Report of the UK government Office of Fair Trading, 2003.

[16] Andrew Rhodes. Re-examining the Effects of Switching Costs. Toulouse School of Economics Working Paper, 2013.

[17] Jean-Charles Rochet and Jean Tirole. Platform Competition in Two-Sided Markets. *Journal of the European Economic Association*, 1(4):990–1029, 2003.

[18] Jean-Charles Rochet and Jean Tirole. Two-sided markets: a progress report. *RAND Journal of Economics*, 37(3):645–667, 2006.

[19] Marc Rysman. The Economics of Two-sided Markets. *Journal of Economic Perspective*, 23(3):125–143, 2009.

[20] Su Su and Na Zeng. The Analysis of Two Period Equilibrium of Two-sided Competing Platforms. In *Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference on Communication, Networking & Broadcasting; Computing & Processing (Hardware/Software)*, 2008.

[21] Curtis Taylor. Supplier surfing: competition and consumer behavior in subscription markets. *RAND Journal of Economics*, 34(2):223–246, 2003.

[22] Christian von Weizsäcker. The Costs of Substitution. *Econometrica*, 52(5):1085–1116, 1984.

[23] Julian Wright. One-sided Logic in Two-sided Markets. *Review of Network Economics*, 3(1):44–64, 2004.

Chapter II

Competition in the Market for Flexible Resources: an application to cloud computing

This paper considers firms' incentives to invest in local and flexible resources when demand is uncertain and correlated. Before demand is realized, two firms decide to invest in their local capacity. Provider(s) of flexible resource observe these decisions and invest in their capacity. After demand is realized, firms buy flexible resource if demand exceeds their local capacity. I find that market power of the monopolist providing flexible resources distorts investment incentives, while competition mitigates them. The extent of improvement depends critically on demand correlation and the cost of capacity: under social optimum and monopoly, if the flexible resource is cheap, the relationship between investment and correlation is positive, and if it is costly, the relationship becomes negative; under duopoly, the relationship is positive. The analysis also sheds light on some policy discussions in markets such as cloud computing.

Keywords: capacity investment, cloud computing, competition, demand correlation

JEL Classification: D4, L8

1 Introduction

For firms in various industries, capacity investment decision involves investing early in their own capacity before demand for their products is realized, and such investment is difficult to

I thank Giacomo Calzolari, Jacques Crémer, Vincenzo Denicolò, Florian Englmaier, Federico Etro, Neil Gandal, Michael Katz, Thomas-Olivier Léautier, Fabio Manenti, Paul Seabright, Tommaso Valletti, and participants in numerous conferences for helpful comments. A previous version of this paper was circulated under the title "Cloud Computing: Investment, Competition and Demand Correlation." Any opinions expressed are those of the author only.

reverse. After the demand is realized, firms have the option to undertake a second investment in a flexible resource to accommodate the excess demand, for instance by outsourcing. In the IT sector, cloud computing provides such an opportunity for outsourcing. Cloud computing is fundamentally the leasing of computer services, including computing power and storage, but on an unprecedented scale. While local computing capacity can support the average demand of the firm, cloud computing is able to scale services on demand and accommodate the workload that exceeds what the local capacity can handle.[1] Accordingly, firms can use cloud computing as a flexible resource for business continuity and disaster recovery plans.[2]

Moreover, in the cloud computing market, computing demand is uncertain as demand varies daily; and correlated at a global level. For example, a U.S. cloud provider such as Amazon, Google and Microsoft could have customers from Europe as well as Australia. Correlation is therefore driven to some extent by geography: computing demands from countries that are close to each other are positively correlated; demands from countries that are located in different time zones are negatively correlated. Moreover, as argued by Harms and Yamartino (2010), even the largest cloud provider will not be able to eliminate uncertainty and correlation.[3]

This paper focuses on the problem of capacity investment in two resources when demand is uncertain and correlated. In the cloud computing example, capacity is a key part of competition in this industry. In the introductory phase, it is common that cloud providers build far more capacity than needed, and one does not expect capacity to be an issue in this growing phase. However, as cloud computing enters a more mature phase, capacity may become constrained as demand grows quickly.[45] For example, on August 25, 2013, Amazon seems to struggle to keep up with the growing computing demand, and an IT problem at one of its datacenters has caused many users of major web services such as Instagram, Netflix, Vine and Airbnb to experience lengthy delays and reduced data transfer speeds for several hours.[6] Amazon's web stores, Microsoft's outlook.com, Google's Gmail email service and the YouTube video site have also faced similar glitches from time to time. This raises a number of interesting questions: what is the profit-maximizing investment

[1] The U.S. National Institute of Standards and Technology provides five defining characteristics of cloud computing: on-demand service, broad network access, resource pooling, rapid elasticity and measured service. This paper focuses on the definition of on-demand service and rapid elasticity.

[2] Business continuity and disaster recovery plans minimize any disruption of business operation due to insufficient local capacity or failure of critical systems.

[3] In the cloud computing market, retailers increase computing demand during the holiday season; and businesses need more computing power during the tax season. However, this type of correlation is not correlation across firms, and is therefore not the focus of this paper.

[4] International Data Corporation (IDC) estimates that worldwide spending on public cloud services is expected to reach $47.4 billion in 2013 and $107 billion in 2017, which represents a growth rate five times that of the IT industry as a whole.

[5] Capacity can be interpreted in two ways: number of physical servers or service quality. In the former case, there is a maximum traffic that each server can handle. In the latter case, even if the capacity does not hit the limit, high demands can put a costly strain on servers, which results in poor quality of service.

[6] BBC news, "Instagram, Vine and Netflix hit by Amazon glitch," available at `http://www.bbc.co.uk/news/technology-23839901`, August 26, 2013 (accessed on August 27, 2013).

strategy in flexible resource such that the problem of quality degradation can be avoided? How should we promote efficient investment from a public policy perspective?

The contribution of this paper is twofold: first, I consider investment in two resources: firms first invest in their local capacity, and later can use flexible resources as an alternative sourcing option to cover temporary shortage of local resources; Second, I focus on uncertain and correlated demand; whereas the existing literature either assumes one type of resources or ignores demand correlation. An interesting finding is that investment can increase with correlation, which is in contrast to the common belief that only negative correlations are valuable because the provider can aggregate demand and reduces the risk.[7] The reason why providers invest more as correlation increases is that when capacity is cheap, providers can benefit more from high demand realizations without worrying about the risk of low demand realizations.

Two firms, whose demand is uncertain and correlated, make their investment decision in local resource under demand uncertainty. Observing firms' local investment, providers of flexible resource (e.g. Amazon, Google and Microsoft) decide how much to invest in capacity, and set the price for their flexible resource (e.g. Amazon Web Services (AWS), Google Compute Engine, Microsoft Azure). After demand is realized, firms can buy flexible resources if demand exceeds their local capacity.

I consider both cases of monopoly and duopoly in providing the flexible resource. As should be expected, investment is suboptimal in the monopoly market. Particularly, the provider of the flexible resource tends to underinvest in its capacity with respect to the socially optimal level, whereas firms tend to overinvest in their local capacity. Such inefficiency comes from market power of the monopolist. Firms invest in local capacity to avoid being exploited by the monopolist, which in turn reduces investment incentive of the monopolist.

Competition always mitigates the underinvestment problem, but more interestingly, the extent of improvement depends crucially on demand correlation and the cost of capacity. Both socially optimal and monopoly investment in flexible resource increases with correlation if the investment cost of flexible resource is small enough, and decreases with correlation if the flexible resource is costly. The reason is that as correlation increases, firms either "win big" when demand realization is high for both firms or "lose big" when demand realizations is low for both firms. If the flexible resource is cheap, the planner or the provider need not worry about "losing". Rather, they will focus on reaping benefits from the "winning" outcome, and therefore they invest more as correlation increases. On the contrary, if the flexible resource is expensive, then "losing" is costly, and thus they invest less as correlation increases.

Under duopoly, I show that investment in flexible resource is increasing in correlation for high or low correlations with a numerical example. The reason for not observing the negative relationship between investment and correlation in this case, as opposed to the social optimum and the monopoly case, is that firms rely more on the flexible resource as competition between providers lowers the price of flexible resources. Firms' incentive to capture the windfall from the

[7] See, for instance, p. 218 of Bayrak et al. (2011).

"winning" high demand realizations increasingly outweighs their incentive to avoid the risk of "losing" as correlation increases. Knowing this, each provider is willing to build a bigger capacity of flexible resources. These results suggest that information on the cost condition and the degree of demand correlation have important consequences for investment. They also explain the need for far more data on costs and demand in order to underpin the appropriate degree of competition in an industry. I will discuss in more detail the implications of competition on investment in the cloud computing industry in the penultimate section.

1.1 Literature

This paper is closely related to the literature on capacity and resource flexibility in operational management. However, unlike this paper, this literature either studies monopolistic models that cannot explain the effect of competition or studies a competitive setting without demand correlation. For example, Lee (2009) studies the optimal capacity investment of a computing service provider in a single resource in the absence of correlated demands. Niyato, Chaisiri and Lee (2009) study the optimal choice of private and public computing service in the monopoly and oligopoly market, but again in a context without correlated demands. Both Van Miegham (1998), and Bish and Wang (2004) study the optimal investment strategy in flexible resources when a monopolist faces uncertain demands for its two products, which corresponds to the social optimum in this model. However, they did not identify the problem of suboptimal investment, and more importantly, how to correct the problem. There are few papers that study firms' choice of technology in a competitive setting. See, for instance, Goyal and Netessine (2007) and Anupindi and Jiang (2008). However, these papers focus on the production stage, without taking into account the incentives to provide flexible resource.

This paper is also related to the literature on Real Options (RO) in finance, which focuses on the role of RO in providing flexibility to management decisions. However, unlike financial assets, IT investments are not tradable, and therefore cannot be priced at the value of risk; rather they are priced by a third party, which is the service provider in this case. See, for instance, Angelou and Economides (2005), Benaroch and Kauffman (1999) and Kauffman et. al. (2002) for details on the limitation of RO's applicability in IT investments. Moreover, the RO literature usually assumes that the value of investment projects is uncorrelated, whereas demand correlation plays an important role here.

2 The Model

Consider two firms, 1 and 2, that need to build capacity in order to serve their customers. To do this, they can either invest in their own local resource L or they can buy flexible resources K from the market. The difference lies in that investments in local resources are irreversible and these resources are for the exclusive use of the investing firm, while flexible resources can be bought from the market instantly when needed and released when not needed. An example of flexible resources is cloud computing as cloud computing power is provisioned as an on-demand service. The firm gets a profit π for each consumer served.

Investment technology. The unit cost of local resource and flexible resource are denoted by c_L and c_K respectively. I assume that local resource is supplied competitively, so that firms can buy L at a price c_L. The flexible resource market can be either a monopoly or a duopoly.

Demand. The demand for the final services of the two firms is uncertain and correlated. More specifically, demands for firm 1 and 2, denoted by x and y respectively, are drawn from a joint distribution $h(x, y)$, with support $[0, \infty) \times [0, \infty)$. The demand of firm 1, x, is given by the marginal distribution $f(x) = \int_0^\infty h(x, y)dy$. Similarly, the demand of firm 2, y, is given by $g(y)$. In the following analysis, I focus on the case where demands (x, y) follow an exponential distribution with $\lambda = 1$,[8] but in Appendix E I show that the main results carry through in the linear case. More particularly, the exponential distribution can be described as follows. The marginal distributions $F(x)$ and $G(y)$ and marginal densities $f(x)$ and $g(y)$ are respectively

$$\begin{aligned} F(x) =&1 - e^{-x}, \\ G(y) =&1 - e^{-y}, \\ f(x) =&e^{-x}, \\ g(y) =&e^{-y}. \end{aligned}$$

The joint distribution function $H(x, y)$ and joint density function $h(x, y)$ follow Gumbel (1960):

$$\begin{aligned} H(x, y) =&(1 - e^{-x})(1 - e^{-y})(1 + \alpha e^{-x-y}), \\ h(x, y) =&e^{-x-y}[1 + \alpha(2e^{-x} - 1)(2e^{-y} - 1)], \end{aligned}$$

where $-1 < \alpha < 1$ is a measure of correlation.[9]

We consider the following game:[10][11]

- Stage 1: firm 1 and 2 invest in their own local capacity L_1 and L_2 simultaneously;
- Stage 2: the provider(s) invest(s) in capacity of flexible resources K;
- Stage 3: the provider(s) set(s) a per unit price of flexible resource p;
- Stage 4: demands (x, y) are realized and firms decides whether and how much to buy the flexible resource.

[8] A distribution is exponential when $F(\lambda, x) = 1 - \lambda e^{-\lambda x}$ is satisfied.

[9] Strictly speaking, $\rho = \frac{cov(x,y)}{\sqrt{var(x)var(y)}}$ is the coefficient of correlation, but since α and ρ move in the same direction (more precisely, $\rho = \frac{\alpha}{4}$, see Equation (3.10) on p. 706 of Gumbel (1960)), there is no loss of generality in saying that α is a measure of correlation.

[10] I do not model entry here, but I expect the same qualitative result with entry. Although entry will lower the price, the underinvestment problem still exists as long as $p > c_K$.

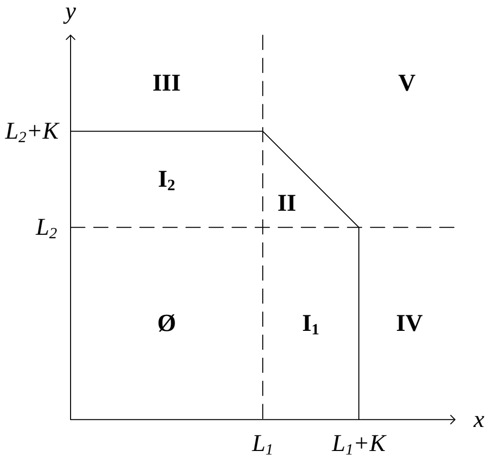
[11] Section 6.1 considers alternative timing.

It is clear that in Stage 4, if a firm's demand spikes above its local capacity, it will purchase flexible resources as long as the price is less than π. In other words, a firm's demand for flexible resources is price-inelastic.[12]

For simplicity, I make the following assumptions. First, $\pi > c_L$, so there is incentive to purchase local resources. Second, I focus on the more interesting case where $c_K < c_L$. For example, it is common in practice that cloud computing exhibits significant economies of scale. To facilitate our analysis, I focus on the specification with $\pi = 1$, $c_L = 0.5$ and $c_K \in [0, 0.5]$.[13] Third, when users are indifferent between buying and not buying the flexible resource, it will always buy for some exogenous reasons such as reputation: if its customer's demand is not served, the customer will never purchase from that firm again. The solution concept adopted here is subgame perfect equilibrium (SPE).

3 Social Optimum

The benevolent planner chooses L_1, L_2, K so as to maximize social welfare. Figure 1 illustrates the basic structure of the demand for flexible resources.



Area	Demand for Cloud
Ø	0
I_1	$x-L_1$
I_2	$y-L_2$
II	$x+y-(L_1+L_2)$
III	K
IV	K
V	K

Figure 1: Demand for Flexible Resources.

In *Area* $\emptyset$, both firms have sufficient local capacity to serve their customers, and therefore there is no demand for cloud. *Area* I_1 captures the situation where firm 2's local capacity is enough

[12] Qualitative results for the monopoly case would be similar if we consider elastic demand. As for the duopoly case, however, if we consider elastic demand, we can no longer follow the approach of de Frutos and Fabra (2011), who study a sequential capacity-price game under demand uncertainty with price-inelastic demands. Interested reader can see Reynolds and Wilson (2000) for a discussion of the two-stage game under the assumption of downward-sloping and uncertain demand.

[13] These assumptions are innocuous for two reasons. First, setting $\pi = 1$ is only a normalization, and it will not affect the qualitative conclusion. Second, the main results hold more generally as long as the flexible resource is more efficient, i.e. $c_K < c_L$.

to cover its demand, but firm 1's demand exceeds its local capacity and will therefore purchase flexible resources. *Area I_2* illustrates the reverse situation where only firm 2 buys cloud. In *Area II*, both firms buy cloud. In all the cases above, all demands are served. *Area III* represents the situation where firm 1 has enough local capacity, while firm 2 has too much demand such that the flexible resource provider is capacity constrained. *Area IV* shows the reverse situation: firm 1 has too much demand, while firm 2's local capacity is sufficient. *Area V* captures the situation where the demands of both firms are extremely high such that it exhausts the capacity of the flexible resource provider. Thus the social welfare is given by

$$\begin{aligned}\max_{L_1,L_2,K} S = &\int_{\emptyset+I_1+I_2+II}(x+y)h(x,y)dydx + \int_{III}(x+L_2+K)h(x,y)dydx \\ &+ \int_{IV}(L_1+K+y)h(x,y)dydx + (L_1+L_2+K)\int_V h(x,y)dydx \\ &- c_K K - c_L(L_1+L_2). \qquad (1)\end{aligned}$$

Let $\Omega(L_1, L_2, K)$ denote the probability of (x, y) falling in areas $\{III\} + \{IV\} + \{V\}$. The social planner only invests in flexible resources, and the socially optimal investment is given by

$$\Omega(0,0,K) = 1 - \int_0^K \int_0^{K-x} h(x,y)dydx = c_K.$$

The optimal capacity is such that the social marginal benefit equals the marginal cost.

Proposition 1. *The social planner only invests in the flexible resource, and the socially optimal investment in flexible resource increases with demand correlation if c_K is small, but decreases with demand correlation if c_K approaches c_L.*

Proof. See Appendix A. □

The intuition behind Proposition 1 runs as follows. The social planner only invests in flexible resources because $c_K < c_L$. As the demand correlation increases, so does the probability of getting either high demand realizations or low demand realizations from both firms: the firms either "win big" or "lose big." The impact of an increase in demand correlation therefore depends on the cost of the flexible resource. If the investment cost is sufficiently low, then "losing" is cheap and the planner would focus on reaping the benefits of high demand realizations. Therefore, investment increases with correlation for low cost. On the contrary, if investment cost is large enough, the planner aims at minimizing the risk of "losing," so investment decreases with correlation.

4 Monopoly

Suppose now that there is a monopoly provider for the flexible resource that chooses p and K to maximize its expected profit. Proceeding by backward induction, given L_1, L_2, K and monopoly price p^m, the demand for cloud is the same as in Figure 1 as long as $p^m \leq \pi$. As the monopolist can extract all the value of its cloud service, it is obvious that

$$p^m = \pi \tag{2}$$

in Stage 3.

The investment of the provider is determined by

$$\Omega(L_1, L_2, K) = 1 - \int_0^L \int_0^{L+K} h(x,y)dydx + \int_L^{L+K} \int_0^{2L+K-x} h(x,y)dydx = c_K, \tag{3}$$

In Stage 1, expecting that $p^m = \pi$, firm 1 chooses its local capacity L_1 so as to maximize its profit:

$$\max_{L_1} \int_0^{L_1} xf(x)dx + \int_{L_1}^{\infty} L_1 f(x)dx - c_L L_1.^{14}$$

The first two terms show that the whole demand is served when demand is below local capacity, whereas capacity is saturated when demand exceeds local capacity. The last term represents the total spending in local capacity.

Then, the first-order condition determines the equilibrium investment of L_1:

$$1 - F(L_1) \leq c_L. \tag{4}$$

The second-order condition is also satisfied.

Analogously, for firm 2, the equilibrium investment of L_2 is determined by

$$1 - G(L_2) \leq c_L. \tag{5}$$

The market equilibrium is characterized by Equations (2), (3), (4) and (5). It is clear that, unlike the social optimum, firms invest in a positive amount of local capacities; and unlike the duopoly case, firms' investments are independent of the provider's investment strategy.

Proposition 2. *In the market with a monopolistic flexible resource provider, the provider underinvests in the flexible resource relative to the social optimum, while the firms overinvest in their local capacity.*

Proof. See Appendix B. □

The intuition behind Proposition 2 is as follows. The monopolist sells the flexible resource at a monopoly price, which extracts all consumer surplus. Anticipating this, the firm will invest in L, even if L is a less efficient technology compared with K, in order to gain part of the consumer surplus. As a consequence, the benefit of investing in the flexible resource is lower for the monopolist than for the social planner, and hence the monopolist underinvests.[15]

[14] The firm only gets positive profit from its local capacity because the surplus of the consumers, who are served by utilizing the flexible resource, are extracted entirely.

[15] Notice that Proposition 2 holds more generally for any rationing rule. The reason is that users pay the monopoly price, and hence the rationing rule will not affect local investment.

To solve the problem, the regulator may ban local investments of the firms. However, this is a rather heavy-handed approach. Firms may prefer local resources for a variety of legitimate reasons. For instance, flexible resources are valuable for the firm as they offer the flexibility to modify a prior investment strategy as more information becomes available over time. More particularly, in case of "good news" the firm can scale up their services, and in case of "bad news" it can scale down. Therefore, firms are willing to pay extra to buy the flexible resource even though it is more expensive ($p^m > c_L$). Indeed, statistics shows that cloud computing is appealing to industries that have high variability in data traffic such as medical research and drug discovery in the healthcare sector.[16]

Therefore, I consider a lighter form of intervention. Since surplus appropriation originates from market power, it seems reasonable to investigate whether introducing more competition in the market—thereby forcing down the price—would incentivize the provider and the firms to behave optimally. As we will see later, the extent to which competition improves investment incentives is subtler than it appears as it varies with demand correlation and investment cost.

Let us now turn to the impact of correlation.

Proposition 3. *In the decentralized case with a single provider, there is positive local investment; and the monopolist's investment in flexible resource increases with demand correlation if c_K is small, but decreases with demand correlation if c_K approaches c_L.*

Proof. See Appendix C. □

The impact of an increase in demand correlation on both socially optimal and equilibrium investment depends on whether the flexible resource is significantly more efficient than the local resource. The intuition of Proposition 3 is in the same spirit as Proposition 1. However, the monopolist's investment is more likely to be decreasing in demand correlation as shown in the following corollary:

Corollary 1. *The smallest c_K under which investment in flexible resource decreases with demand correlation is larger at the social optimum than under monopoly.*

Proof. See Appendix D. □

The intuition behind Corollary 1 is that local investment is zero at the social optimum and positive in the monopoly case. Thus, the planner will not run into the risk of not being able to sell the flexible resource to firms that receive low demand and buy local resources only. As a consequence, the planner can better enjoy the possible windfall from high demand realizations than the monopolist.

[16] World Economic Forum (2010) identifies the healthcare industry as one of the major sectors which can benefit from cloud computing.

5 Duopoly

Let us now consider the case of competing providers. They play the game as before.[17] I solve the problem proceeding backwards. In the capacity-price stage, I apply some results in de Frutos and Fabra (2011), henceforth FF, which can be summarized as follows. In their paper, two firms make sequential capacity-price decision under demand uncertainty in markets with price-inelastic demands. They show that

- Proposition 7 of FF. The only equilibrium in the pricing stage is a mixed-strategy equilibrium.
- Proposition 8 of FF. Capacity choices are asymmetric.
- Proposition 9 of FF. If the density function of demand is non-decreasing, then the equilibrium is unique.

For a given L_1 and L_2, there is a stochastic demand function for the flexible resource that is price-inelastic. Thus, we can apply FF's results in the continuation game, where the aggregate capacity is defined by $K(L_1, L_2)$, the capacity chosen by the smaller provider $k^-(L_1, L_2)$, the capacity chosen by the larger provider $k^+(L_1, L_2)$, and the equilibrium expected profits of the two providers $\pi^-(L_1, L_2)$ and $\pi^+(L_1, L_2)$.

The main difference between this paper and FF is that the first stage in this paper is absent in FF. FF assume that demand is exogenously given, while here the demand for the flexible resource is endogenously determined by investments in local capacity and the strength of demand correlation. Therefore, unlike the monopoly case, firms' investments are no longer independent of the provider's strategy. This poses several difficulties in the analysis.

First, the endogenously determined demand function for the flexible resource is not necessarily non-decreasing, which means that the equilibrium in the continuation game may not be unique. If this is the case, we focus on the most symmetric case, where the difference between the big firm and the smaller firm is minimized, meaning that the degree of competitiveness is maximized.

Second, this introduces strategic interaction between the two firms: each firm's investment changes the demand for the flexible resource, which affects providers' investments and in turn affects the rival firm's investment. To simplify the analysis, I assume that L_1 and L_2 are chosen cooperatively such that $L_1 = L_2 = L$. The two firms maximize the following joint profit:[18][19]

$$\max_L \left[S(L) - \pi^+(L) - \pi^-(L)\right] - 2c_L L,$$

[17] Since there is demand uncertainty, this exercise requires more than just applying the classical result of Kreps and Scheinkman (1983), which proves outcome equivalence between the capacity-price game and the Cournot game. As pointed out by de Frutos and Fabra (2011), the introduction of demand uncertainty rules out the existence of symmetric equilibria due to a difference in marginal revenue between the large firm and the small firm even if the two firms are symmetric ex ante.

[18] Under this assumption, rationing rule does not affect investments in local and flexible resources.

[19] Even though I assume cooperative investment, the two firms act differently from the case with a single firm because the two firms cannot share their local capacity.

where $S(L)$ is the social surplus given by Equation (1). The surplus is shared between the firms and the providers (but not the consumers). This is because demand is inelastic, so firms can extract all consumer surplus.

Solving the above problem yields the equilibrium investment in local capacity L_1^d and L_2^d, where d denotes duopoly. Then, we can also determine the equilibrium investment in flexible resource $K^d(L_1^d, L_2^d)$.

As should be expected, competition always increases social welfare as compared to the monopoly case because it mitigates the underinvestment problem in flexible resources and the overinvestment problem in local resources. A formal proof is provided in Appendix F. More interesting is that the extent of improvement depends crucially on the cost of capacity and the degree of correlation, which is shown in the following numerical example.[20]

Figure 2 plots, for a given c_K, flexible resource investment against demand correlation. Social optimum is shown with a solid line, the duopoly case is drawn as a dotted line, and the monopoly case is illustrated by a long-dashed line.

The main observations in Figure 2 are summarized in the following remark.

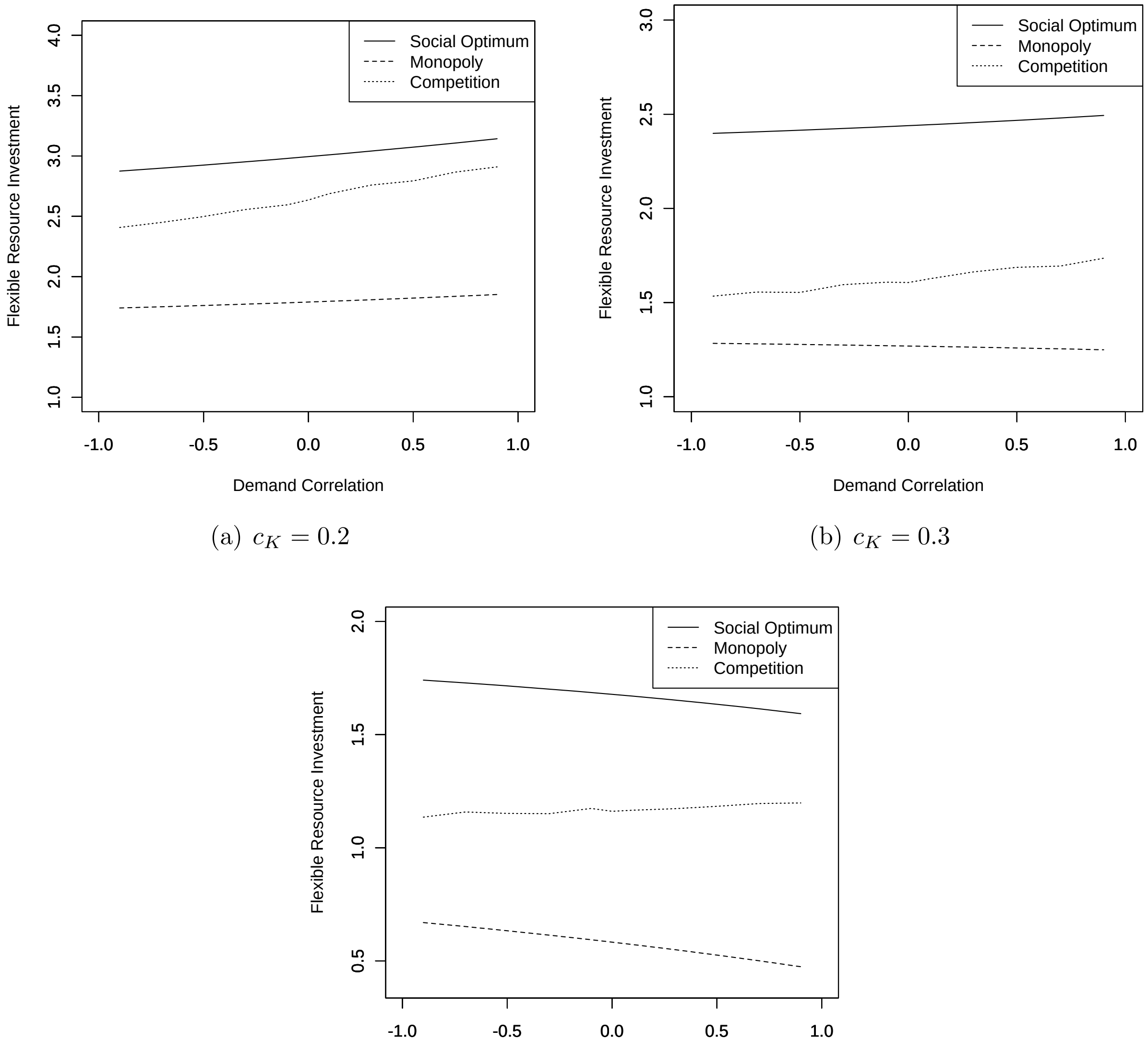
Remark 1. *Comparing the socially optimal, monopoly and duopoly solutions,*

(i) When c_K is sufficiently small, both the planner and the monopolist's investments in flexible resources increase with correlation. As c_K approaches c_L, both of these investments decrease with correlation. The threshold level such that the impact of correlation changes is larger at the social optimum than it is under monopoly.

(ii) Under duopoly, it can be shown that for high or low correlations, the investment in flexible resource is increasing in correlation.

Part (i) is already shown in Propositions 1 and 3, and Corollary 1. As for part (ii), the intuitive reason for not observing a negative relationship between investment and correlation under duopoly, unlike the socially optimal and monopoly regimes, is as follows. Under the socially optimal and monopoly regimes, local investment does not vary with correlation: at the social optimum local investment is zero; in the monopoly case firms pay the monopoly price, and thus their local investment is not affected by correlation. Unlike these regimes, in the duopoly case firms pay less than the monopoly price and are therefore more willing to switch to buying the flexible resource in order to capture the possible windfall of high demand realizations. As a consequence, firms invest less in local capacities, and hence providers invest more in flexible resources as correlation increases.

[20] The main difficulty in solving for an explicit solution in the duopoly case stems from the fact that the demand for the flexible resource is endogenously determined by L and α, and this, in turn, affects the mixed strategy in prices of the provider. Consequently, it is difficult to characterize the profit function of the firm without using a numerical method.



(a) $c_K = 0.2$

(b) $c_K = 0.3$

(c) $c_K = 0.5$

Figure 2: Flexible Resource Investment and Demand Correlation for different values of c_K.

6 Discussion

6.1 Alternative Timing

My analysis focuses on the timing where firms invest first. It fits the scenario where some flexible resources such as cloud computing offers more flexibility in managing demand uncertainty than local resources. However, one could alternatively consider the case where firms observe the provider's investment in flexible resources before deciding their own local investment. In this setting, firms still overinvest in L, and providers still underinvest in K, provided price is chosen after the capacity decision because the monopoly price will emerge as long as demand is inelastic. Another alternative is to consider the case where p is chosen prior to L, but the underinvestment problem will still occur because the provider will never charge $p = c_K$ as its profit will become zero and it will not have any incentive to invest. Moreover, it is difficult to think of a situation in practice that fits the scenario of choosing price prior to capacity.

6.2 Remedies

Although it is always more efficient for firms to use the flexible resource, there are two reasons that prevent everyone from using the flexible resource only: first, the stochastic nature of demand prevents the provider from contracting over the amount of investment ex ante; second, the provider of the flexible resource cannot commit to marginal-cost pricing. As a consequence, firms rely more on local capacity and the provider underinvests.

Throughout the paper, I focus on non-contingent and linear pricing.[21] One can think of other pricing structures such as non-linear tariffs and contingent pricing. First, considering non-linear tariffs, it is common for cloud providers such as Amazon, Dropbox and Google to use non-linear pricing for their storage service: they provide basic service for free, and then offer additional storage capacity for a fee. However, we can easily see that non-linear pricing does not solve the underinvestment problem because the provider will underinvest as long as $p > c_K$.

Second, considering contingent pricing, such practice is not very popular in the market for cloud computing: with the exception of AWS, which uses both contingent and non-contingent pricing, other large cloud providers such as Azure, Google and IBM rarely use spot pricing. On the contrary, in the electricity wholesale market, electricity is bought and sold at spot prices.[22] Yet, there is only one kind of capacity: firms typically buy energy from electricity companies, but do not generate their own electricity (although some firms may have their own emergency electricity generator, they are not for regular use). As argued by Carr (2005) and Jeff Bezos in

[21] Non-contingent pricing means that prices are determined before demand is realized, whereas contingent pricing are state-dependent.

[22] The electricity literature (see, for instance, Borenstein and Holland (2005), Murphy and Smeers (2005), Joskow and Tirole (2007), and Léautier (2011)) mostly considers a two-stage game, in which firms choose their capacity first, and then they bid prices for each state of the world in a spot market. See also Crew, Fernando and Kleindorfer (1995) for a survey of the literature on peak-load pricing.

Stone (2013), they both envisioned today's IT supply would transform from companies' private capacity into a centralized utility service, just like how electricity became a utility a century ago. It is therefore interesting to think about how spot pricing can change investment incentives in an environment with both flexible and local resources, as in the case of cloud computing, where firms buy flexible resource for its instant scalability and own local resource for data security and privacy reasons. A formal model of contingent pricing would entail a trade-off as follows: the provider tends to price high during peak periods, which induces firms to invest more in local capacity; but it tends to price low during off-peak periods, which induces firms to rely more on flexible resource. Consequently, the extent to which investment is distorted depends on the relative strength of these two effects. If, for instance, the second effect dominates, then contingent pricing can potentially remedy the problem of underinvestment in flexible resources. Despite this additional trade-off created by contingent pricing, investment decision still depends fundamentally on the degree of correlation and the cost of capacity, and therefore all the main qualitative results of this paper should remain valid.

Finally, it may be worthwhile to consider a subsidy. Suppose the regulator introduce a subsidy s for investment in flexible resource. The cost of flexible resource becomes $c_K - s$, so the provider will be more willing to offer a lower price. At the same time, it also has more incentives to undertake investment in flexible resource, which could potentially mitigate the underinvestment problem.

6.3 Policy Implications

Cloud computing has emerged as a new business model for computing and storage resource management for firms, and a new source of entertainment and communication services for consumers. As the cloud market is still in its infancy, many classic economic issues such as pricing, investment strategies, the appropriate market structure, competition policy, privacy and security concerns are still unclear.[23] We take the first step to understand the impact of competition on

[23] Recently, there has been a flurry of research on the opportunities and obstacles for the adoption of cloud services; see, for example, Armbrust et al. (2009), Harms and Yamartino (2010), and Marston et al. (2011). They mainly focus on three layers of the cloud architecture: infrastructure, platform, and application. However, as argued in Bayrak et al. (2011), such categorization are useful only in defining technological differences, but not so much in analyzing their economic impact. Indeed the existing literature on cloud computing are mostly descriptive, and only rarely is the problem approached from a theoretical perspective. Fershtman and Gandal (2012) raise important economic issues of cloud computing such as changes in the strength of network effects, compatibility among software applications, the development of standards, and the market structure that should emerge. However, most of these topics have already been well-documented in a separate literature; in order to work on theoretical advancement, one needs to clearly delineate the unique features of the cloud computing market.

Recent efforts to expand the theoretical study of cloud computing include Wang (2014), who studies the adoption of cloud services within a moral hazard framework, and this paper. However, they differ in two respects. First, this paper is about capacity investment, while Wang focuses on the problem of migration, which means that there is no investment on the provider's side. Second, this paper studies the effect of competition, but such effect is absent investment in this industry.

Although there are a number of competitors in the cloud computing market such as AWS, Azure, Google and IBM/SoftLayer, market power exists. For instance, large cloud providers build hyperscale datacenters that exhibit significant increasing returns to scale, which could come from the centralization of computing resources or from volume discount on the components that providers use to build their datacenter.[24] As a result, smaller firms may not be able to compete with these incumbents. Moreover, many consumers prefer to buy service from well-known brands because they expect higher quality. This raises concerns about the degree of competitiveness of this market.

This model predicts that the impact of competition on investment depends crucially on the investment cost. It is often argued that cloud computing reduces the cost of investing in computing power significantly. While the marginal cost of producing an extra unit of storage or computing power is close to zero, the costs of electricity for powering up the machines, cooling the systems, as well as management, maintenance and implementation of the software and hardware in a large server farm is far from negligible.[25] Therefore, information on the cost structure in the cloud computing industry should have been gathered and analyzed as it has important consequences for investment.

7 Conclusion

This paper has analyzed firms' incentives to invest in local and flexible resources. I find that market power of the monopolist providing flexible resources distorts investment, and competition always improves social welfare. The extent of improvement depends on demand correlation and investment cost. If investment cost is small, investment under social optimum, monopoly and competition is increasing in correlation; if cost is large, investment under competition is still increasing in correlation, whereas that under social optimum and monopoly goes in opposite direction. I have also examined the potential merits of policies such as spot pricing and a subsidy for investment in flexible resource to remedy the underinvestment problem.

These results have implications for investment decision in outsourcing, particularly in the market for cloud computing. Admittedly, the cloud computing market is growing unpredictably, and there is no clear indication or consensus on how it will develop. For now, this paper shows that even if the cloud computing market follows the footsteps of the electricity market and providers eventually adopt spot pricing, a similar trade-off that we derived here will arise. Therefore, analyzing data on cost and demand represents a useful first step towards a fuller understanding of the nascent industry.

in Wang.

[24] See Harms and Yamartino (2010) for more examples of how firms benefit from economies of scale.

[25] In September 2012, the New York Times reported that "the digital warehouses use about 30 billion watts of electricity, roughly equivalent to the output of 30 nuclear power plants."

I list some important topics that lie beyond the scope of this paper, but would be appropriate for further work. The first is to consider product differentiation. For example, assuming that cloud computing services (such as Dropbox storage services) and local storage services are differentiated—how, then, would the investment strategy change? Second, it would be interesting to study the consequences of vertical integration. For instance, what will happen if upstream cloud computing firms such as Microsoft and Google also enter the downstream market of software applications?

A Proof of Proposition 1

The social optimum is obtained by differentiating Equation (1) with respect to L_1, L_2 and K. The F.O.C. with respect to L_1 is given by

$$\{IV\} + \{V\} \leq c_L.$$

Similarly, the F.O.C. with respect to L_2 is

$$\{III\} + \{V\} \leq c_L.$$

Finally, the F.O.C. with respect to K is:

$$\{III\} + \{IV\} + \{V\} \leq c_K.$$

As $\{III\} + \{IV\} + \{V\} > \{IV\} + \{V\}$ or $\{III\} + \{V\}$, the marginal benefit of investing in the flexible resource is always higher than that of local capacity. Furthermore, the marginal cost of investing in the flexible resource is lower ($c_K < c_L$). Then we must have $L_1^* = L_2^* = 0$, where asterisk denotes the socially optimal level of investment. Since $c_K < c_L < \pi$, all F.O.C. are satisfied with equality.

The socially optimal investment in the flexible resource is determined by the F.O.C. with respect to K, which can be rewritten as

$$F(K, \alpha, c_K) = \int_0^K \int_0^{K-x} h(x,y)dydx - 1 + c_K = 0.$$

By implicit function theorem,

$$\frac{\partial K}{\partial \alpha} = -\frac{\frac{\partial F}{\partial \alpha}}{\frac{\partial F}{\partial K}}.$$

We can show that

$$\frac{\partial F}{\partial K} = \int_0^K e^{-K}[1 + \alpha(2e^{-x} - 1)(2e^{x-K} - 1)]dx$$

is positive. Moreover, we have

$$\begin{aligned}\frac{\partial F}{\partial \alpha} &= \int_0^K \int_0^{K-x} e^{-x-y}(2e^{-x} - 1)(2e^{-y} - 1)dydx \\ &= -e^{-K}[K + 3e^{-K} + 2Ke^{-K} - 3].\end{aligned}$$

It can be shown that there exists a $\bar{K}^*$ such that $\frac{\partial F}{\partial \alpha} < 0$ when $K > \bar{K}^*$, and $\frac{\partial F}{\partial \alpha} > 0$ when $K < \bar{K}^*$. In addition, it is obvious that K decreases with c_K. Therefore, if c_K is small such that $K > \bar{K}^*$, then $\frac{\partial K}{\partial \alpha} > 0$. On the contrary, if c_K is large, K is small such that $K < \bar{K}^*$, then $\frac{\partial K}{\partial \alpha} < 0$.

B Proof of Proposition 2

For firm 1, its equilibrium investment is determined by

$$1 - F(L_1) = c_L,$$

As $1 - F(L_1) > \{IV\} + \{V\}$, we must have $L_1^m > L_1^* = 0$, and hence there is overinvestment. The same happens for firm 2.

For the flexible resource provider, its equilibrium investment K^m is determined by

$$\begin{aligned}
\max_K \Pi = & \int_0^{L_1} \int_{L_2}^{L_2+K} (y - L_2) h(x,y) dy dx + \int_{L_1}^{L_1+K} \int_0^{L_2} (x - L_1) h(x,y) dy dx \\
& + \int_{L_1}^{L_1+K} \int_{L_2}^{L_1+L_2+K-x} (x + y - L_1 - L_2) h(x,y) dy dx \\
& + K \left[\int_0^{L_1} \int_{L_2+K}^{\infty} h(x,y) dy dx + \int_{L_1+K}^{\infty} \int_0^{L_2} h(x,y) dy dx \right. \\
& \left. + \int_{L_1}^{\infty} \int_{L_2}^{\infty} h(x,y) dy dx - \int_{L_1}^{L_1+K} \int_{L_2}^{L_1+L_2+K-x} h(x,y) dy dx \right] - c_K K.
\end{aligned}$$

which gives us

$$\Omega(L_1^m, L_2^m, K^m) = c_K = \Omega(0, 0, K^*).$$

Suppose that the flexible resource provider invests K such that $L^m + K = K^*$, Since $L^m > 0$, it must be $\Omega(L_1^m, L_2^m, K) < \Omega(0, 0, K^*)$, which means such K cannot be the equilibrium. Therefore, the flexible resource provider must invest K^m such that $L^m + K^m < K^*$, which means that $K^m < K^*$ (underinvestment).

C Proof of Proposition 3

The monopolist's investment is determined by

$$F(K, \alpha, c_K) = \int_0^L \int_0^{L+K} h(x,y) dy dx + \int_L^{L+K} \int_0^{2L+K-x} h(x,y) dy dx - 1 + c_K = 0.$$

By implicit function theorem,

$$\frac{\partial K}{\partial \alpha} = - \frac{\frac{\partial F}{\partial \alpha}}{\frac{\partial F}{\partial K}}.$$

It is straightforward to show that

$$\begin{aligned}\frac{\partial F}{\partial K} =& \int_0^L e^{-x-L-K}[1+\alpha(2e^{-x}-1)(2e^{-L-K}-1)]dx \\ &+ \int_0^L e^{-y-L-K}[1+\alpha(2e^{-y}-1)(2e^{-L-K}-1)]dy \\ &+ \int_L^{L+K} e^{-2L-K}[1+\alpha(2e^{-x}-1)(2e^{-x-2L-K}-1)]dy > 0, \\ \frac{\partial F}{\partial \alpha} =& \int_0^L \int_0^{L+K} e^{-x-y}(2e^{-x}-1)(2e^{-y}-1)dydx \\ &+ \int_L^{L+K} \int_0^{2L+K-x} e^{-x-y}(2e^{-x}-1)(2e^{-y}-1)dydx.\end{aligned}$$

Similar to the proof in Appendix A, there exists $\bar{K}^m$ such that $\frac{\partial F}{\partial \alpha} < 0$ when $K > \bar{K}^m$; and $\frac{\partial F}{\partial \alpha} > 0$ when $K < \bar{K}^m$. Moreover, as K^m is decreasing in c_K, then if c_K is small such that $K > \bar{K}^m$, then $\frac{\partial K}{\partial \alpha} > 0$. On the contrary, if c_K is such that $K < \bar{K}^m$, then $\frac{\partial K}{\partial \alpha} < 0$.

D Proof of Corollary 1

From the proof in Appendices A and C, it suffices to show $\frac{\partial F}{\partial \alpha}^*(K^*) < \frac{\partial F}{\partial \alpha}^m(L^m, K^m)$, where both terms integrate the same function over the respective area as shown in Figure 3. The difference between $\frac{\partial F}{\partial \alpha}^*(K^*)$ and $\frac{\partial F}{\partial \alpha}^m(L^m, K^m)$ lies in the shaded area. Comparing integrations over the triangles and the trapezium, we can conclude that the above condition is satisfied because the triangles have higher values of x or y.

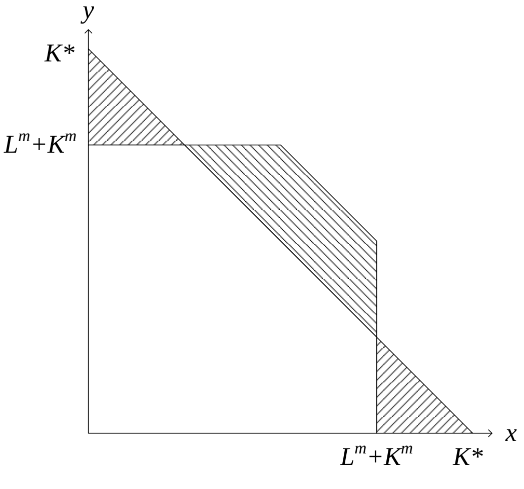


Figure 3: Investment under Social Optimum and Monopoly.

We therefore have

- If $\frac{\partial F}{\partial \alpha}^m < 0$, then $\frac{\partial F}{\partial \alpha}^* < 0$. Both $\frac{\partial K}{\partial \alpha}^*, \frac{\partial K}{\partial \alpha}^m > 0$, which is true for small c_K.

- If $\frac{\partial F}{\partial \alpha}^* > 0$, then $\frac{\partial F}{\partial \alpha}^m > 0$. Both $\frac{\partial K}{\partial \alpha}^*, \frac{\partial K}{\partial \alpha}^m < 0$, which is true for large c_K.
- For medium c_K, $\frac{\partial F}{\partial \alpha}^* < 0$ and $\frac{\partial F}{\partial \alpha}^m > 0$. Then, $\frac{\partial K}{\partial \alpha}^* > 0$ and $\frac{\partial K}{\partial \alpha}^m < 0$.

Thus, under social optimum there is a larger range of c_K under which investment increases with correlation as compared to the monopoly case.

E Linear Example

E.1 Social Optimum

The relationship between investment in flexible resource and demand correlation at the social optimum is slightly different when demands are uniformly distributed. To see this, consider a joint distribution $h(x, y)$ as follows:

- *Positive correlation.* With probability ρ, only pairs of demands on the $x = y$ line are possible (perfect positive correlation). With probability $1 - \rho$, demands are uniformly distributed on a unit square $[0, 1] \times [0, 1]$ (independent demands). We can use ρ as a measure of positive correlation.
- *Negative correlation.* With probability ρ, only pairs of demands on the $x + y = 1$ line are possible (perfect negative correlation). With probability $1-\rho$, demands are uniformly spread over a unit square $[0, 1] \times [0, 1]$ (independent demands). We can use $-\rho$ as a measure of negative correlation.

Since $c_K < c_L < \pi$, all the F.O.C. are satisfied with equality. In the case of positive correlation, the optimal capacity is chosen such that the marginal benefit is equal to the marginal cost:

$$\rho(1 - \frac{K}{2}) + (1 - \rho)\frac{1}{2}(2 - K)^2 = c_K.$$

Note that $K \geq 1$ because $c_K \leq 0.5$. Differentiating K with respect to ρ, we find that K^* increases with ρ.

In the case of negative correlation, we have

$$K^* = \max\left\{1, 2 - \sqrt{\frac{2c_K}{1 - \rho}}\right\}.$$

Note that $K \geq 1$. The reason is that if demands are perfectly negatively correlated and investment is less than 1, then marginal benefit always exceeds cost. When $K > 1$, the optimal investment is determined by

$$(1 - \rho)\frac{1}{2}(2 - K)^2 = c_K.$$

It is easy to see that K^* increases with $-\rho$.

We therefore have

Result 1. *In the case of uniformly distributed demands, the social planner only invests in the flexible resource, and the socially optimal investment always increases with demand correlation.*

The reason is that, for uniformly distributed demands, the marginal benefit of expanding capacity always increases as correlation increases.

E.2 Monopoly Case

In the monopoly case, the result in the linear example is the same as Proposition 3 in the main text. To keep things simple, further assume that $c_K \in [0.25, 0.5]$ such that $L_1 + K$ and $L_2 + K$ are smaller than 1. In the case of positive correlation, the monopolist chooses K such that

$$\rho(\frac{1}{2} - \frac{K}{2}) + (1-\rho)(\frac{3}{4} - K - \frac{1}{2}K^2) = c_K.$$

In the case of negative correlation, the monopolist choice of K solves

$$\rho(1-2K) + (1-\rho)(\frac{3}{4} - K - \frac{1}{2}K^2) = c_K.$$

We therefore have

Result 2. *In the case of uniformly distributed demands, there is positive local investment; and the monopolist's investment in flexible resource increases with demand correlation if c_K is small, but decreases with demand correlation if c_K approaches c_L.*

F Competition Improves Social Welfare

Competition always increases social welfare because it mitigates the underinvestment and overinvestment problem.

- $K^d \geq K^m$: The F.O.C. of K in the monopoly case is

  $$\{III\} + \{IV\} + \{V\} = c_K.$$

  As for the duopoly case, we refer to Equation (12) in FF: the F.O.C. of K is

  $$1 - D(K) = c_K,$$

  where $D(K)$ is the demand for the flexible resource. Since firms only buy the flexible resource when demand is above their local capacity, this condition can be rewritten as

  $$\frac{\{III\} + \{IV\} + \{V\}}{1 - \int_0^{L_1}\int_0^{L_2} h(x,y)dydx} = c_K.$$

  Therefore, $K^d \geq K^m$ because $1 - \int_0^{L_1}\int_0^{L_2} h(x,y)dydx < 1$. Note that $K^d = K^m$ only when $L_1, L_2 = 0$.

- $p^d \leq p^m$: Under duopoly, providers of the flexible resource randomize over price with the upper bound of π (see Proposition 7 of FF).

- $L^d \leq L^m$: Firms invest less in local resource under duopoly because the price of it is lower.

References

[1] Georgios Angelou and Anastasios Economides. Flexible ICT investment analysis using Real Options. *International Journal of Technology, Policy and Management*, 5(2):146–166, 2005.

[2] Ravi Anupindi and Li Jiang. Capacity Investment under Postponement Strategies, Market Competition, and Demand Uncertainty. *Management Science*, 54(11):1876–1890, 2008.

[3] Michael Armbrust, Armando Fox, Rean Griffith, Anthony Joseph, Randy Katz, Andrew Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia. Above the clouds: A Berkeley view of cloud computing. Technical Report No. UCB/EECS-2009-28, 2009. available at `http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html`.

[4] Ergin Bayrak, John Conley, and Simon Wilkie. The Economics of Cloud Computing. *The Korean Economic Review*, 27(2):203–230, 2011.

[5] Michel Benaroch and Robert Kauffman. A Case for Using Real Options Pricing Analysis to Evaluate Information Technology Project Investments. *Information Systems Research*, 10(1):70–86, 1999.

[6] Ebru Bish and Qiong Wang. Optimal Investment Strategies for Flexible Resources, Considering Pricing and Correlated Demands. *Operations Research*, 52(6):954–964, 2004.

[7] Severin Borenstein and Stephen Holland. On the Efficiency of Competitive Electricity Markets with Time-invariant Retail Prices. *RAND Journal of Economics*, 36(3):469–493, 2005.

[8] Nicholas Carr. The End of Corporate Computing. *MIT Sloan Management Review*, pages 67–73, Spring 2005.

[9] Michael Crew, Chitru Fernando, and Paul Kleindorfer. The Theory of Peak-load Pricing: A Survey. *Journal of Regulatory Economics*, 8:215–248, 1995.

[10] María-Ángeles de Frutos and Natalia Fabra. Endogenous Capacities and Price Competition: The role of demand uncertainty. *International Journal of Industrial Organization*, 29:399–411, 2011.

[11] Chaim Fershtman and Neil Gandal. Migration to the Cloud Ecosystem: Ushering in a New Generation of Platform Competition. *Communications & Strategies Digiworld Economic Journal*, 85(1):109–123, 2012.

[12] Manu Goyal and Serguei Netessine. Strategic Technology Choice and Capacity Investment under Demand Uncertainty. *Management Science*, 53(2):192–207, 2007.

[13] Emil Julius Gumbel. Bivariate Exponential Distributions. *Journal of the American Statistical Association*, 55(292):698–707, 1960.

[14] Rolf Harms and Michael Yamartino. The Economics of the Cloud. Microsoft White Paper, 2010.

[15] Paul Joskow and Jean Tirole. Reliability and Competitive Electricity Market. *RAND Journal of Economics*, 38(1):60–84, 2007.

[16] Robert Kauffman, Henry Lucas, Paul Tallon, Andrew Whinston, and Kevin Zhu. Using Real Options Analysis for Evaluating Uncertain Investments in Information Technology: Insights from the ICIS 2001 Debate. *Communications of the Association for Information Systems*, 9:136–167, 2002.

[17] Thomas-Olivier Léautier. The Invisible Hand: ensuring optimal investment in electric power generation. Working Paper, 2011.

[18] In Lee. A Model for Determining the Optimal Capacity Investment for Utility Computing. In Dirk Neumann, Mark Baker, Jörn Altmann, and Omer Rana, editors, *Economic Models and Algorithms for Distributed Systems*. Birkhäuser Verlag, Basel, Switzerland, 2009.

[19] Sean Marston, Zhi Li, Subhajyoti Bandyopadhyay, Juheng Zhang, and Anand Ghalsasi. Cloud Computing - The Business Perspective. *Decision Support Systems*, 51:176–189, 2011.

[20] Frederic Murphy and Yves Smeers. Generation Capacity Expansion in Imperfectly Competitive Restructured Electricity Markets. *Operations Research*, 53(4):646–661, 2005.

[21] Dusit Niyato, Sivadon Chaisiri, and Bu-Sung Lee. Economic Analysis of Resource Market in Cloud Computing Environment. Paper presented to the 2009 IEEE Asia-Pacific Services Computing Conference, Singapore, 7-11 December 2009.

[22] Stanley Reynolds and Bart Wilson. Bertrand-Edgeworth Competition, Demand Uncertainty, and Asymmetric Outcomes. *Journal of Economic Theory*, 92:122–141, 2000.

[23] Brad Stone. *The Everything Store: Jeff Bezos and the Age of Amazon.* Little, Brown and Company, United States, 2013.

[24] Tong Wang. Migration to the Cloud. Working Paper, 2014. available at https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=IIOC2014&paper_id=162.

[25] Jan Van Miegham. Investment Strategies for Flexible Resources. *Management Science*, 44(8):1071–1078, 1998.

[26] World Economic Forum. Exploring the Future of Cloud Computing: Riding the Next Wave of Technology-Driven Transformation, 2010. available at http://members.weforum.org/pdf/ip/ittc/Exploring-the-future-of-cloud-computing.pdf.

Chapter III

Ex Ante and Ex Post Investments in Cybersecurity

This paper develops a theory of sequential investments in cybersecurity in which the software vendor can invest *ex ante* and *ex post.* The regulator can use safety standards and liability rules as means of increasing security. A standard is a minimum level of safety, and a liability rule states the amount of damage each party is liable for. I show that the joint use of an optimal standard and a full liability rule leads to underinvestment *ex ante* and overinvestment *ex post* because the software vendor does not suffer the full costs of the society in case of security failure. Instead, switching to a partial liability rule can correct the inefficiencies. This suggests that to improve security, the regulator should encourage not only the firms, but also the enterprises to invest in security. I also discuss the effect of network externality and explain why firms engage in "vaporware".

Keywords: cybersecurity, sequential investment, standards, liability

JEL Classification: L1, L8

1 Introduction

New security concerns are constantly arising as privacy breaches proliferate and cyber attacks escalate. For example, a recent data breach on an unprecedented scale saw more than 1.2 billion credentials stolen by a Russian criminal group.[1] Moreover, we continue to see the rise of "ransomware" (a malicious program that encrypts files on the victim's computer and demands a fee

I thank Giacomo Calzolari and Jacques Crémer for their helpful comments. Any opinions expressed are those of the author only.

[1] See "Russia gang hacks 1.2 billion usernames and passwords," *BBC News*, August 6 2014, available at `http://www.bbc.com/news/technology-28654613`.

before unlocking those files), the discovery of security flaws on smartphones, and the emergence of new security risks from the "Internet of Things" (such as hackers stealing sensitive data from owners of Internet-connected objects—from locks, lights, thermostats, televisions, refrigerators, washing machines, to cars). A critical gap has thus emerged between firms' investment in cybersecurity and today's rapidly evolving technological advances, which warrants further research. More particularly, good security depends on more than just the technology. It requires a deeper understanding of the incentives of the agents who sell as well as those who use the technology. In the software industry, the incentives of those who are responsible for security and those who suffer from a security problem are often misaligned: while software vendors are motivated to minimize their own private costs, the social planner's goal is to minimize society's costs. Firms' incentives to invest are therefore suboptimal.[2]

The purpose of this paper is to understand how to use legislation such as safety standards and liability rules to provide incentives for software firms to make their product more secure. A standard is a minimum level of safety set by the regulator, and a liability rule states the amount of damage each party is liable for. In practice, there are different types of security standards, such as encryption standards, security breach notification standards, IT continuity standards, set by the National Institute of Standards and Technology (NIST) and Center for Internet Security (CIS) in the U.S., and more widely by the International Organization for Standards (ISO) and Internet Engineering Task Force (IETF). As for negligence liability, consumers continue to file lawsuits against firms for security breaches, data leakage, and infringement of privacy, and in this regard, these firms might be held accountable for consumer damages. This raises a number of interesting questions: Which of the interventions, standards or liability rules, would better incentivize firms and consumers to behave optimally? Should standards and liability rules be used separately or jointly? Is it socially optimal to shift some of the cost of investing in security from firms to consumers? To address these questions, I develop a model to study the investment incentives of a software firm when its software is subject to security problems and when consumers bear some precaution costs.

This paper makes two contributions. First, it studies a new type of inefficiency in the cybersecurity market, which is due to software vendors failing to take into account of consumers' cost of investing in security. Taking precautions is in general less costly for ordinary consumers as they only need to reboot their machines and the process of updating security is mostly automatic nowadays. However, the cost of precautions is significant for enterprise users, especially when they adopt sophisticated firewalls, cryptographic protocols, virus detection techniques, intrusion detection systems, data-loss prevention features, among others. Top-notch security tools are expensive and require a large number of man-hours to maintain and manage them. They are especially important for financial services, telecommunication sectors and government departments. Second, I introduce two types of investment the firms can undertake: *ex ante* care and *ex post* maintenance.

[2] See Anderson, Clayton and Moore (2009), and Anderson and Moore (2009) for surveys of the economics of network security.

In the software industry, as software is always evolving and adding new functionalities, they are never free of bugs. There are usually multiple rounds of debugging. Therefore, it is common for the software industry to have sequential investments. I further show that such possibility of sequential investments may lead to "vaporware" practice even in the absence of preemptive motives and reputation concerns: because *ex ante* and *ex post* investments are substitutes, allowing firms to identify security problem *ex post* increases the likelihood of releasing a less secure software product *ex ante*—a new perspective in the vaporware literature. In Sections 3.1 and 4, I also explore the consequences of public policies such as subsidizing the training of computer experts, synchronizing patch release and adoption cycles, and implementing vulnerability management by a third party.

To be more specific, I consider a model in which a firm sells software that is subject to potential security problems. The firm can invest *ex ante* to increase the security level and *ex post* to find the security problem before the hacker. If the firm discovers the bug, it can choose whether to disclose it or hide it. If the firm discloses the bug information, consumers can choose whether to take precaution or not. Consumers differ in their costs of taking precaution: actions are more costly for the laymen than for the computer experts.

I find that since the firm does not suffer the full costs of the society in case of security failure, its incentives to invest are suboptimal: it underinvests *ex ante* and overinvests *ex post.* I also show that there are inefficiencies associated with the joint use of a full liability rule and an optimal standard to increase security. Interestingly, a partial liability rule, which shifts some liability to the consumers, can correct the inefficiencies. This suggests that policies that encourage consumers and firms to share the costs of security could improve security. For example, since applying patches and malware-removal tools are costly for enterprise customers, the government could try to encourage them to put more effort in finding, testing and installing these tools as soon as the vendor makes them available. These results continue to hold in the presence of network externality.

I also show that if the firm has limited liability, increasing the number of computer experts mitigates suboptimal investment incentives. The reason is that the difference between the private and social incentives to invest arises from two effects. First, the firm does not pay fully for the damage, and the total amount of damage is decreasing in the number of experts. Second, the firm ignores the precautionary costs of the consumers when it makes its investment decision, and the total cost of precaution is increasing in the number of experts. When the firm has limited liability, the first effect dominates. This implies that to alleviate the inefficiency, the government can either impose limited liability on the firm and increase the number of computer experts, or simply allocate more liability to the firm. More particularly, under limited liability, the government can provide a subsidy for training in the area of cybersecurity so that enterprises become more competent in managing security threats. In contrast, if the firm bears substantial liability for consumers' damage, then the government needs to be careful about increasing the number of experts because the objectives of the planner and the firm will become more divergent.

1.1 Literature

This paper is primarily related to recent works on the economics of security investment. Gordon and Loeb (2002) and Kunreuther and Heal (2003) study the optimal security investment. Kunreuther and Heal (2003) consider the presence of network externality, but Gordon and Loeb (2002) do not. Both of them consider simultaneous investment, while I focus on sequential investment. Varian (2004) examines full liability in a model in which efforts of multiple parties are needed to increase security. He finds that liability should be assigned entirely to the party who can best manage the risk. Different from his analysis, I also consider partial liability, and the joint effect of partial liability and standards.

This paper also relates to the economics and legal literature on tort laws, but it departs from this literature by considering the possibility of consumers taking actions and sequential investments. More specifically, Shavell (1984) and Kolstad et al. (1990) compare standards with liability rules. However, Shavell's analysis is based on the inefficiencies associated with the potential bankruptcy of the firm and the uncertainty of lawsuit by the consumers, while the inefficiencies studied by Kolstad et al. are due to the uncertainty over the legal standard to which the firm will be held liable. Differently, inefficiencies here are caused by the firm failing to take into account of consumers' costs of investing in security. Moreover, the literature on torts has tended to focus on either *ex ante* investment, as in Daughety and Reinganum (1995, 2006), or *ex post* investment, as in Polinsky and Shavell (2010);[3][4] whereas this paper deals with both.

Finally, this paper shares with the literature on disclosure laws (see, for example, Granick (2005) and Choi et al. (2010)) the focus on the tradeoff that arises from disclosing software vulnerabilities: while secrecy prevents attackers from taking advantage of publicized security flaws, it interferes with scientific advancement in security, which is largely based on information sharing and cooperation. Choi et al. also examine the effect of a mandatory disclosure policy and a "bug bounty" program on welfare. However, they take security investments as given, and do not discuss optimal investment. Daughety and Reinganum (2005) study the effect of confidential settlement on product safety, but their focus is not on investment. This paper extends this literature by analyzing the optimal investment in security, and such investment is of two kinds: *ex ante* care and *ex post* maintenance.

2 The Model

Monopoly software vendor. Consider a firm that produces a software product which contains potential bugs. For simplicity, I assume away prices, so that the problem is simplified to choosing a level of security that minimizes the sum of the costs. The assumption is reasonable for con-

[3] See Daughety and Reinganum (2013) for a survey of the literature on torts.

[4] Polinsky and Shavell analyze information acquisition about product risks when product quality is uncertain. Therefore, their problem concerns *ex post*, rather than *ex ante*, investment.

sumers who have already bought the software and are therefore not concerned about the prices. Moreover, if the firm generates profit from channels other than selling the software product such as advertisement, then the objective is simply to minimize the costs.

Heterogeneous consumers. There is a unit mass of consumers. Consumers have different precaution costs: a proportion α of them are "computer experts" and have precaution cost γ drawn from a distribution $F(\gamma) \sim [0, +\infty)$, while the others are "laymen" with $\gamma = \infty$. Experts are security professionals who can take security precautions such as monitoring the system for attacks and patching the system if the firm discloses the presence of a security problem, while laymen without such professional knowledge will never take precautions.[5] In the main text, all experts have the same γ and there are two types of consumers, but in Appendix A I show that the results are robust to the introduction of a continuum of consumer types. Assume that consumers always have positive utility in using the software.

Timing of the game. (i) The firm invests s in security at a cost $c(s)$. This is *ex ante* care. Such investment could take the form of improvement in infiltration detection or authentication technologies. (ii) By investing $m(b)$ in *ex post* maintenance, the firm will find a bug before the hacker does with probability b. Let $p(s)$ be the probability that the hacker will attack. I assume away strategic attacks.[6] (iii) If the firm discovers a bug, it can choose whether or not to disclose the security problem. Assume that there is no cost in disclosing the bug. For example, the firm can simply post the information on its website. However, disclosure increases the probability of attack by a small ϵ.[7] (iv) If the firm discloses a bug, the experts can choose whether or not to take precaution.

Assumption 1. *$c'(0) = 0, c'(s) > 0, c''(s) > 0, c'''(s) > 0, m'(0) = 0, m'(b) > 0, m''(b) > 0, m'''(b) > 0, p'(s) < 0,$ and $p''(s) > 0$.*

Under Assumption 1, investment costs $c(s)$ and $m(b)$ are thrice differentiable, convex, and increasing in s and b respectively;[8] and that probability of attack $p(s)$ is convex and decreasing in s.

Damage. For the firm, the damage incurred from an attack is $\overline{\eta}$ in case the hacker discovers the bug before the firm does, and $\underline{\eta}$ in case the firm identifies the bug first. Assume that $\overline{\eta} > \underline{\eta}$. This could be the financial loss caused by stolen information of the firm becoming available to the hacker. Such loss is smaller if the firm finds the bug first as it can then try to fix the problem.

[5] I assume that consumers take precaution after the firm has disclosed the information about the bug. One could alternatively think of consumers taking precaution *ex ante*. However, the qualitative result will not change as long as the costs associated with these precautions are not borne by the firm.

[6] Strategic attacks are modeled in, for instance, Acemoglu et al. (2013). They show that strategic targeting provides additional incentives for overinvestment in security because larger investment shifts attacks from one agent to another.

[7] Arora, Nandkumar and Telang (2006) show empirically that in some cases vulnerability disclosure increases the frequency of attacks.

[8] The third derivatives ensure that the profit function is well-behaved.

However, the firm may face substantial loss if the hacker exploits a bug that has not been previously identified—a phenomenon known as "zero-day attacks". For the consumers, the damage from an attack is $\overline{\mu}$ if they do not take precaution and $\underline{\mu}$ if they do. This could be monetary loss due to fraudulent use of their personal information. Assume that $\overline{\mu} > \underline{\mu}$, meaning once informed, consumers can take actions to mitigate the risk of being attacked. Let $\lambda \in [0,1]$ denote the part of consumers' damages for which the firm is liable. I focus on three liability regimes:

- Full liability, under which the firm is liable for all damages faced by the consumers, i.e. $\lambda = 1$;
- Partial liability, under which the firm only compensates consumers partially, i.e. $\lambda \in (0,1)$;
- No liability, under which consumers will not receive any compensation from the firm, i.e. $\lambda = 0$.

Thus, the total loss for the firm is $\eta + \lambda\mu$, where $\eta \in [\overline{\eta}, \underline{\eta}]$ and $\mu \in [\overline{\mu}, \underline{\mu}]$.

3 Optimal Investment

I now work backward from the last stage. When the firm discloses a bug, the expected damage for a consumer who does not take precaution is $p(s)\overline{\mu}$, and that for a consumer who takes precaution is $p(s)\underline{\mu} + \gamma$. Therefore, the consumer will take precaution if

$$\gamma < p(s)(\overline{\mu} - \underline{\mu}). \tag{1}$$

In the disclosure stage, the firm can choose its disclosure policy in case it discovers a bug. If it does not disclose the security problem, its expected cost is $p(s)(\underline{\eta} + \lambda\overline{\mu})$. If it chooses to disclose, there are two cases. If consumers take precaution, the firm incurs a cost of $p(s)[\underline{\eta}+\lambda(\alpha\underline{\mu}+(1-\alpha)\overline{\mu})]$. However, if consumers do not take precaution, the cost becomes $p(s)(\underline{\eta} + \lambda\overline{\mu})$.[9] Therefore, the firm will only disclose if this leads consumers to take precaution, that is, if Equation (1) holds.

In the investment stage, the firm chooses s and b to minimize its expected loss, which is denoted by $\mathcal{L}^f$.

$$\begin{aligned} \min_{b,s} \mathcal{L}^f &= (1-b)p(s)(\overline{\eta} + \lambda\overline{\mu}) \\ &\quad + b\left\{\int_0^{p(s)(\overline{\mu}-\underline{\mu})} p(s)[\underline{\eta} + \lambda(\alpha\underline{\mu} + (1-\alpha)\overline{\mu})]dF(\gamma) + \int_{p(s)(\overline{\mu}-\underline{\mu})}^{\infty} p(s)(\underline{\eta} + \lambda\overline{\mu})dF(\gamma)\right\} \\ &\quad + m(b) + c(s). \end{aligned} \tag{2}$$

Let $b^m(s)$ denote the firm's optimal *ex post* investment strategy given *ex ante* security s, and let s^* and $b^* \equiv b^m(s^*)$ denote the solutions of Equation (2).

[9] When consumers do not take precaution, the firm is indifferent between disclosing and not disclosing. However, by assuming that disclosure would increase the probability of attack by ϵ, the firm will strictly prefer not to disclose.

The first term in Equation (2) is the expected cost of the firm when the hacker discovers the bug first, and in which case both the firm and the consumers suffer a large damage. When the firm finds the bug before the hacker, either it discloses the bug if consumers' cost is small, which is captured by the second term, or it does not disclose if consumers' cost is large, which is captured by the third term. In this case, the firm suffers a small damage from attack because it identifies the bug sooner than the hacker, while the extent of damages suffered by the consumers depends on whether precautionary measures are taken. The last two terms represent *ex ante* and *ex post* investment costs.

The social planner's incentive to disclose is the same as the firm, that is, the planner will disclose as long as γ is small enough. However, different from the firm, if the planner chooses to disclose, its expected cost is $p(s)(\underline{\eta} + \alpha\underline{\mu} + (1-\alpha)\overline{\mu}) + \alpha\gamma$, which is higher than that of the firm. This is because the planner also takes into account consumers' cost of taking precautions, and internalizes all the costs, so there is no liability issue. In case of non-disclosure, the expected cost is $p(s)(\underline{\eta} + \overline{\mu})$.

The social planner chooses s and b to minimize the expected loss of the society, which is denoted by $\mathcal{L}^{SP}$.

$$\begin{aligned}
\min_{b,s} \mathcal{L}^{SP} =& (1-b)p(s)(\overline{\eta}+\overline{\mu}) + b\left\{ \int_0^{p(s)(\overline{\mu}-\underline{\mu})} [p(s)(\underline{\eta} + \alpha\underline{\mu} + (1-\alpha)\overline{\mu}) + \alpha\gamma] dF(\gamma) \right. \\
&\left. + \int_{p(s)(\overline{\mu}-\underline{\mu})}^{\infty} p(s)(\underline{\eta}+\overline{\mu}) dF(\gamma) \right\} + m(b) + c(s) \\
=& \mathcal{L}^f|_{\lambda=1} + b\alpha \int_0^{p(s)(\overline{\mu}-\underline{\mu})} \gamma dF(\gamma). \qquad (3)
\end{aligned}$$

Let $b^{SP}(s)$ denote the social planner's optimal *ex post* investment strategy given *ex ante* security s, and let s^o and $b^o \equiv b^{SP}(s^o)$ denote the solutions of Equation (3).

The difference between $\mathcal{L}^f$ and $\mathcal{L}^{SP}$ is that the firm minimizes its own private costs, while the social planner minimizes the sum of firm's and consumers' costs.

Lemma 1. *Under full liability ($\lambda = 1$), $b^m(s)$ and $b^{SP}(s)$ decrease with s.*

Proof. See Appendix B. □

Lemma 1 shows that the firm has less incentive to find bugs *ex post* given a high security level *ex ante*, meaning that *ex ante* and *ex post* investments are substitutes.

Lemma 2. *Under full liability ($\lambda = 1$), $b^m(s) > b^{SP}(s)$ for all s. In particular, if the standard is set at the socially optimal level, $s^* = s^o$, the firm will overinvest in ex post maintenance, $b^m(s^o) > b^{SP}(s^o)$.*

Proof. See Appendix C. □

One might expect that under full liability and an optimal standard the firm will invest optimally, but it turns out differently when consumers also bear some costs in protecting their computers. The intuition runs as follows. If a bug is not found, both the firm and the society suffer the same loss. If a bug is discovered, the firm can reduce the loss more than the planner because it does not bear the costs of the consumers. Since the firm has more to gain in finding the bug, it will overinvest.

I assume that full liability is defined for "net" damages to the consumers. One can alternatively define it for "total" damages, which includes also consumers' precaution cost. In this case, full liability alone is enough to restore the first-best. I model the liability regime the way I did because in practice, firms are typically liable for financial damages to the consumers caused by, for example, a data breach. Liability sometimes also covers for litigation costs, but very rarely for investment costs in precaution. One difficulty lies in estimating the amount of time and effort consumers spent on managing, maintaining and patching a system.

Proposition 1. *(Full Liability). Under full liability* ($\lambda = 1$)*, the firm underinvests in ex ante care,* $s^* < s^o$*, and overinvests in ex post maintenance,* $b^* > b^o$*.*

Proof. See Appendix D. □

Proposition 1 shows that full liability alone does not achieve the first-best solution. The reason is that, as shown in Lemma 2, *ex post* the firm has more to gain in finding the bug than the planner, and hence it invests too much in *ex post* maintenance. The firm invests too little in *ex ante* care because it expects to overinvest *ex post*, as was shown in Lemma 1.

Proposition 2. *(Partial Liability). The socially optimal level of investment,* s^o *and* b^o*, can be achieved with the joint use of an optimal standard* s^o *and a partial liability rule* $\lambda \in (0, 1)$*.*

Proof. See Appendix E. □

When security standards are set at the socially optimal level, it is inefficient to implement full liability because the firm will overinvest *ex post*; it is also inefficient to set firm's liability to zero because it will then underinvest *ex post.* As a consequence, the optimal liability rule is a partial one. Note that in Appendix F I show that if liability regime is the only instrument of public policies, it is not enough to provide the right incentives for two investments.

3.1 Network Externality

In this subsection, I consider direct and indirect network effects. In practice, users whose computers are infected may create negative externalites on the other users in that attackers can use these computers to host phishing sites, distribute spam e-mails or other unlawful content. Kunreuther and Heal (2003), August and Tunca (2006), Acemoglu et al. (2013), and Riordan (2014), for instance, examine agents' incentive to invest in security under the presence of network externalities. While they focus on one type of security investment, this paper deals with two types.[10]

Let us first examine the situation with indirect network effects in which the firm's investment strategy is affected by the proportion of consumers taking precaution.

Corollary 1. *(Indirect network effects). When λ is large, increasing the proportion of computer experts, α, exacerbates the ex ante underinvestment and ex post overinvestment problems. When λ is small, increasing α mitigates the investment problem.*

Proof. See Appendix G. □

The intuition behind Corollary 1 runs as follows. Comparing Equations (2) with (3), the difference between the private and social incentives to invest that is related to α arises from the following.

$$p(s) \quad \underbrace{(1-\lambda)(\alpha\underline{\mu}+(1-\alpha)\overline{\mu})}_{distortion\ from\ liability\ assignment} \quad + \quad \underbrace{\alpha\gamma}_{distortion\ from\ consumers'\ costs} \quad .$$

Investment incentives are therefore distorted by two forces: first, the firm does not pay fully for the damage; second, the firm ignores the precautionary costs of the consumers when it makes its investment decision. If the firm is held liable for a large proportion of damage (i.e. λ is large), then reducing the proportion of experts (α) mitigates suboptimal investment incentives. The reason is that an increase in firm's liability reduces the first type of distortion, whereas a decrease in the proportion of experts reduces the second type of distortion. Taking the effects together, the objectives of the planner and the firm become more aligned, and thus this reduces the extent that the firm is investing suboptimally. If, on the other hand, the firm is held liable for a smaller proportion of damage, then increasing the proportion of experts will reduce the inefficiency. This is because the extent of the first type of distortion depends on the total amount of damage, and is decreasing in α, whereas the extent of the second type of distortion depends on the total cost of precaution of the consumers, and is increasing in α. When the firm has limited liability, the first type of distortion dominates.

This implies that to alleviate the inefficiency, the government can either impose limited liability on the firm and increase the number of computer experts, or simply allocate more liability to the firm. More particularly, under limited liability, the government can provide a subsidy for training in the area of cybersecurity so that enterprises become more competent in managing security threats. For example, many security breaches involve attackers trying to compromise users' accounts, and users are sometimes unaware of such attack. Even if they are aware of the attack, they sometimes lack the skills needed to resolve the security problem. Therefore, increasing training that aims to enhance the technical skills of these enterprise users appears to be appropriate provided that

[10] More particularly, August and Tunca (2006) focus on the problem of patch management, and therefore consider *ex post* investment only. Security investments are strategic complements in Kunreuther and Heal (2003), strategic substitutes in Acemoglu et al. (2013), and can be strategic complements or strategic substitutes in Riordan (2014) depending on whether the attacks are direct or indirect, but agents can only invest once in these models.

the cost of implementing this subsidy is not too large. In contrast, if the firm bears substantial liability for consumers' damage, then the government needs to be careful about increasing the number of experts because the objectives of the planner and the firm would further diverge. That being said, this does not mean that offering cybersecurity training is undesirable (e.g. it could potentially generate cost savings for firms through detecting, defending against and recovering from cyber-attacks), but that the potential adverse effects on incentives should not be ignored.

Previously, I have assumed that there are no direct network effects, but my qualitative results would not change even if we add this. Re-interpreting *ex post* investment as a patch release and consumers' action as the choice of patch installation, direct network effects between consumers could arise when consumers who do not patch increase the security risks on other consumers, and consumers who patch reduce the probability of others being attacked. In this case, increasing the proportion of experts α will lower the damage to all experts, $\underline{\mu}$, and that to all laymen, $\overline{\mu}$, meaning only magnitude changes. However, the main qualitative result of liability-sharing between the firm and the consumers remains valid, provided consumers have to take precautionary actions.

4 Discussion

Vaporware.—"Vaporware" refers to the software industry practice of announcing new products well in advance of their actual release on the market.[11] The previous literature, for instance, Bayus et al. (2001) and Haan (2003), studies how such product pre-announcements can be used as a means of entry deterrence in a signaling model. Choi et al. (2010) examine how reputation concerns may induce firms to make honest announcements in a repeated cheap-talk game. Although vaporware practice typically means the release dates of the products are much later than the original announced dates, we could alternatively view the announced product as a product characteristics (a security feature, for instance) instead of the physical product. Vaporware could then be interpreted as delivering a lower-quality product than promised, which is consistent with the current development in the industry: software products, mobile applications, and smart-home appliances are often launched prematurely while they are still in development and are therefore susceptible to security risks. The result of *ex ante* underinvestment in security in this model captures the essence of this situation. Moreover, I show that underinvestment may occur even in the absence of preemptive motives and reputation concerns. This is therefore different from the vaporware literature, where firms engage in vaporware only to prevent entry or when reputational concern is not so important. The new insight here is that the possibility of sequential investments, which allows the firm also to invest *ex post* in fixing the security problem, provides an alternative explanation at least in part for vaporware practice in the software market.

Policy Implications.—I have examined the investment incentive of a software vendor, both *ex*

[11] Vaporware may also mean the announced products never reach the market, but this is not the focus of this paper because the firm always introduces the product in this model.

ante and *ex post*, when consumers bear some costs of taking precaution. I find that security can be improved with the joint use of an optimal standard and a partial liability rule. This implies that the regulator can enforce some minimum standards for encryption and security breach notification. Sanctions can be imposed if these requirements are violated. Another policy we can consider is liability regime. Interestingly, I find that, given an optimal standard, shifting some liability to the consumers is welfare improving. This means that the regulator should not impose a one hundred percent liability on the software vendor because this will distort its investment incentives. Instead, an effective policy is to ask both the software vendor and its customers to share the costs of security.[12]

Despite the fact that users dislike or feel concerned about security problems, some of them ignore notifications from the vendor and do not take up any of the proposed solutions. For example, more than 90% of ChoicePoint customers whose personal information had been stolen did not take up the mitigating solutions offered by the firm such as free credit monitoring service and insurance after the data breach.[13] This may be due to the fact that consumers have other competing demands on their time, and paying attention to data breach notifications appears to be low on their priority list.

On enterprise level, installing patches could be costly especially for large companies because the plethora of security updates can often overwhelm software engineers, who have to keep track of all relevant bugs and patches, and match the version of all those updates to the version of software their company is using. Once a problem is identified, they need to figure out which updates get priority, and look for solutions to deal with it.[14] In addition, if the installation requires rebooting an enterprise's critical system, downtime can be expensive. As a consequence, this could easily lead to the missing of some major security updates.

This suggests that a desirable policy should try to eliminate the delay in applying the solutions to security problems. First, the government could persuade or mandate the users to react more quickly (for example, within a predetermined window of time) as soon as the vendor makes the solutions available and notifies them in a reasonable way. Second, third parties can be introduced to help enterprises to find, select and deploy the solutions that are relevant to their systems. An example of third-party vulnerability management that helps businesses to adhere to compliance and security standards in the IT and financial sectors is Qualys, Inc.

[12] Although this discussion interprets costs of security as a form of liability, they are different from the costs explained by γ in that consumers ignoring or not noticing security alerts is not an investment, but rather it shows a systematic lack of security consciousness. This raises the question of who should be responsible for the damages that arise from such negligence.

[13] See Jon Brodkin, "Victims of ChoicePoint Data Breach Didn't Take Advantage of Free Offers," *Network World*, April 10, 2007, `http://www.networkworld.com/news/2007/041007-choicepoint-victim-offers.html?page=1`.

[14] Practitioners have commonly considered patch management as a time- and resource-consuming activity. See, for instance, Symantec, "Automating Patch Management," February 8, 2005, `http://www.symantec.com/articles/article.jsp?aid=automating_patch_management`.

More General Applications.—The analysis also provides insight into other industries in which sequential investments are important, such as automobiles and pharmaceuticals. We can then reinterpret the seller as a firm that produces a product with some safety features. There are again two types of investments the firm can undertake: *ex ante* investment in pre-sale product design, and *ex post* investment in post-sale product testing. For example, *ex ante* investment could lead to the development of a new technology in cars that is subject to potential safety defect, or a new drug that has previously unknown side effects. The firm can invest *ex post* to remedy these safety problems. We can then use the previous analysis to study investment incentives of the firm, in particular whether there are incorrect incentives to provide safety *ex ante* and *ex post* and how to improve them.

5 Conclusion

To increase security, the key is not so much about holding the software vendor solely liable for the loss, but balancing the investment incentives between different players. This discussion represents a useful first step towards understanding sequential security investments. In future work, it might be interesting to relax the single-firm assumption and consider dynamic issues and contagion issues in a network of multiple firms.[15]

A Continuum of Consumers

With a slight abuse of the notation, suppose that there is a continuum of consumers whose precaution cost γ is drawn from a distribution $F(\gamma) \sim [0, +\infty)$. As before, consumers will take precaution if $\gamma < p(s)(\overline{\mu} - \underline{\mu})$, and the marginal consumer, who is indifferent between taking and not taking precaution, is given by $\gamma(s) \equiv p(s)(\overline{\mu} - \underline{\mu})$.

If the firm does not disclose the bug, its expected cost is $p(s)(\underline{\eta} + \lambda\overline{\mu})$; if it discloses the bug, it expected cost is $p(s)[\underline{\eta} + \lambda(F(\gamma(s))\underline{\mu} + (1 - F(\gamma(s)))\overline{\mu})]$. Since the latter is smaller than the former, the firm will always disclose. Therefore, the firm chooses s and b to minimize

$$\min_{b,s} \mathcal{L}^f = (1-b)p(s)(\overline{\eta} + \lambda\overline{\mu}) + bp(s)[\underline{\eta} + \lambda(F(\gamma(s))\underline{\mu} + (1 - F(\gamma(s)))\overline{\mu})] + m(b) + c(s). \quad \text{(A.1)}$$

As for the planner, the cost for non-disclosure is $p(s)(\underline{\eta} + \overline{\mu})$, whereas the cost for disclosure is $p(s)[\underline{\eta} + F(\gamma(s))\underline{\mu} + (1 - F(\gamma(s)))\overline{\mu}] + \int_0^{\gamma(s)} \gamma dF(\gamma)$. Since the latter is smaller than the former,

[15] See, for instance, Morris (2000), Acemoglu et al. (2013), and Goyal et al. (2014) for treatment of contagion in networks.

the planner will always disclose. The planner therefore solves

$$\begin{aligned}\min_{b,s}\mathcal{L}^{SP} &= (1-b)p(s)(\overline{\eta}+\overline{\mu})\\ &\quad + b\left\{p(s)[\underline{\eta}+F(\gamma(s))\underline{\mu}+(1-F(\gamma(s)))\overline{\mu}]+\int_0^{\gamma(s)}\gamma dF(\gamma)\right\}+m(b)+c(s). \quad \text{(A.2)}\end{aligned}$$

It is easy to see that since $\int_0^{\gamma(s)}\gamma dF(\gamma)>0$, $\mathcal{L}^{SP}>\mathcal{L}^f$ for any λ. Thus, the main results of *ex ante* underinvestment and *ex post* overinvestment carry through.

B Proof of Lemma 1

Since $\lambda=1$, the first-order conditions with respect to b are given by

$$\begin{aligned}\frac{\partial\mathcal{L}^{SP}}{\partial b} &= 0,\\ \Leftrightarrow m'(b) &= p(s)(\overline{\eta}+\overline{\mu})-\underbrace{\int_0^{p(s)(\overline{\mu}-\underline{\mu})}[p(s)(\underline{\eta}+\alpha\underline{\mu}+(1-\alpha)\overline{\mu})+\alpha\gamma]dF(\gamma)}_{G^{SP}(s)}\\ &\quad -\int_{p(s)(\overline{\mu}-\underline{\mu})}^{\infty}p(s)(\underline{\eta}+\overline{\mu})dF(\gamma), \quad \text{(B.1)}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial\mathcal{L}^{f}}{\partial b} &= 0,\\ \Leftrightarrow m'(b) &= p(s)(\overline{\eta}+\overline{\mu})-\underbrace{\int_0^{p(s)(\overline{\mu}-\underline{\mu})}p(s)(\underline{\eta}+\alpha\underline{\mu}+(1-\alpha)\overline{\mu})dF(\gamma)}_{G^{f}(s)}\\ &\quad -\int_{p(s)(\overline{\mu}-\underline{\mu})}^{\infty}p(s)(\underline{\eta}+\overline{\mu})dF(\gamma). \quad \text{(B.2)}\end{aligned}$$

The right hand sides of Equations (B.1) and (B.2) are decreasing in s.

C Proof of Lemma 2

We can see from Equations (B.1) and (B.2) that if $s^*=s^o$, then $G^f(s^o)<G^{SP}(s^o)$. Thus, $b^m(s^o)>b^{SP}(s^o)$.

D Proof of Proposition 1

Since $\lambda = 1$, the first-order conditions with respect to s are given by

$$\begin{aligned}\frac{\partial \mathcal{L}^{SP}}{\partial s} &= 0, \\ \Leftrightarrow -\frac{c'(s)}{p'(s)} &= (1-b)(\overline{\eta}+\overline{\mu}) + b\left[\int_0^{p(s)(\overline{\mu}-\underline{\mu})} (\underline{\eta}+\alpha\underline{\mu}+(1-\alpha)\overline{\mu})dF(\gamma)\right. \\ &\quad \left. + \int_{p(s)(\overline{\mu}-\underline{\mu})}^{\infty} (\underline{\eta}+\overline{\mu})dF(\gamma)\right], \qquad \text{(D.1)}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \mathcal{L}^{f}}{\partial s} &= 0, \\ \Leftrightarrow -\frac{c'(s)}{p'(s)} &= (1-b)(\overline{\eta}+\overline{\mu}) + b\left[\int_0^{p(s)(\overline{\mu}-\underline{\mu})} (\underline{\eta}+\alpha\underline{\mu}+(1-\alpha)\overline{\mu})dF(\gamma)\right. \\ &\quad \left. + \int_{p(s)(\overline{\mu}-\underline{\mu})}^{\infty} (\underline{\eta}+\overline{\mu})dF(\gamma) - \alpha p(s)(\overline{\mu}-\underline{\mu})^2 f(p(s)(\overline{\mu}-\underline{\mu}))\right]. \qquad \text{(D.2)}\end{aligned}$$

Define the right hand side of Equation (D.1) as $H^{SP}(b)$, and that of Equation (D.2) as $H^f(b)$. Clearly, the left hand sides of Equations (D.1) and (D.2) are equal. However, $H^{SP}(b^{SP}(s)) > H^f(b^{SP}(s)) > H^f(b^m(s))$. The first inequality follows from $H^{SP}(b) > H^f(b)$ for any b, whereas the second inequality is due to the fact that $H^f(b)$ is decreasing in b.

Since $c'''(s) > 0$ and $p'''(s) > 0$, it is easy to see that $-c'(s)/p'(s)$ is convex and increasing in s, and it has the limits $\lim_{s\to 0} -c'(s)/p'(s) = 0$ and $\lim_{s\to\infty} -c'(s)/p'(s) = \infty$. As for the right hand sides, the limits of both $H^{SP}(b)$ and $H^f(b)$ are bounded away from ∞ as s tends to ∞. Moreover, $H^{SP}(0) > 0$, and if $H^f(0) > 0$, the solution to both equations exists, and we denote them by s^* and s^o respectively. In addition, if the solution is unique, we must have $s^* < s^o$ due to the fact that $H^{SP}(b^{SP}(s)) > H^f(b^m(s))$.[16]

Using Lemma 1, if $s^* < s^o$, then $b^* > b^o$.

E Proof of Proposition 2

Suppose $s^* = s^o$. If $\lambda = 1$, Lemma 2 implies $b^m(s^o) > b^{SP}(s^o)$. If $\lambda = 0$, Equation (B.2) becomes

$$m'(b) = p(s)(\overline{\eta} - \underline{\eta}).$$

Comparing with Equation (B.1), $b^m(s^o) < b^{SP}(s^o)$. Therefore, there exists $\lambda \in (0,1)$ such that $b^m(s^o) = b^{SP}(s^o)$.

[16] For example, there exists a unique equilibrium investment when both $F(p(s))$ and $p(s)f(p(s))$ are convex, and $m(b)$ is quadratic.

F Liability regime as the only instrument

Suppose that there exists $\lambda \in [0, 1]$ such that $b^* = b^o$ and $s^* = s^o$. This implies that $\partial\mathcal{L}^f/\partial b = \partial\mathcal{L}^{SP}/\partial b$ and $\partial\mathcal{L}^f/\partial s = \partial\mathcal{L}^{SP}/\partial s$. However, we can easily verify that these two conditions cannot be satisfied at the same time.

G Proof of Corollary 1

The difference between Equations (B.1) and (B.2) is

$$m'(b^*) - m'(b^o) = \alpha \int_0^{p(s)(\overline{\mu}-\underline{\mu})} \gamma dF(\gamma),$$

which is positive and increasing in α, meaning that a larger α worsens the *ex post* overinvestment problem.

Similarly, the difference between Equations (D.1) and (D.2) is

$$(b^* - b^o)\left[\int_0^{p(s)(\overline{\mu}-\underline{\mu})} (\underline{\eta} + \alpha\underline{\mu} + (1-\alpha)\overline{\mu})dF(\gamma) + \int_{p(s)(\overline{\mu}-\underline{\mu})}^{\infty} (\underline{\eta} + \overline{\mu})dF(\gamma) - (\overline{\eta} + \overline{\mu})\right] \\ - \alpha b^* p(s)(\overline{\mu} - \underline{\mu})^2 f(p(s)(\overline{\mu} - \underline{\mu})).$$

The first term $(b^* - b^o)$ is positive and increasing in α, and the term in the square bracket is negative and decreasing in α. The product of these two terms is thus negative and decreasing α. Since the final term $-\alpha b^* p(s)(\overline{\mu} - \underline{\mu})^2 f(p(s)(\overline{\mu} - \underline{\mu}))$ is also negative and decreasing in α, taken together the difference between Equations (D.1) and (D.2) is negative and decreasing in α, meaning that the *ex ante* underinvestment problem is more severe as α increases.

This proof remains valid as long as λ is large enough.

References

[1] Daron Acemoglu, Azarakhsh Malekian, and Asu Ozdaglar. Network Security and Contagion. MIT Working Paper, 2013.

[2] Ross Anderson, Richard Clayton, and Tyler Moore. The Economics of Online Crime. *Journal of Economic Perspectives*, 23(3):3–20, 2009.

[3] Ross Anderson and Tyler Moore. Information Security: Where Computer Science, Economics and Psychology Meet. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 367(1898):2717–2727, 2009.

[4] Ashish Arora, Anand Nandkumar, and Rahul Telang. Does information security attack frequency increase with vulnerability disclosure? An empirical analysis. *Information Systems Frontiers*, 8(5):350–362, 2006.

[5] Terrence August and Tunay Tunca. Network Software Security and User Incentives. *Management Science*, 52(11):1703–1720, 2006.

[6] Barry Bayus, Sanjay Jain, and Ambar Rao. Truth or Consequences: An Analysis of Vaporware and New Product Announcements. *Journal of Marketing Research*, 38(1):3–13, 2001.

[7] Jay Pil Choi, Chaim Fershtman, and Neil Gandal. Network Security: Vulnerabilities and Disclosure Policy. *Journal of Industrial Economics*, 58(4):868–894, 2010.

[8] Jay Pil Choi, Eirik Kristiansen, and Jae Nahm. Vaporware. *International Economic Review*, 51(3):653–669, 2010.

[9] Andrew Daughety and Jennifer Reinganum. Product Safety: Liability, R&D and Signaling. *American Economic Review*, 85(5):1187–1206, 1995.

[10] Andrew Daughety and Jennifer Reinganum. Secrecy and Safety. *American Economic Review*, 95(4):1074–1091, 2005.

[11] Andrew Daughety and Jennifer Reinganum. Markets, Torts and Social Inefficiency. *RAND Journal of Economics*, 37(2):300–323, 2006.

[12] Andrew Daughety and Jennifer Reinganum. Economic Analysis of Products Liability: Theory. In Jennifer Arlen, editor, *Research Handbook on the Economics of Torts*, chapter 3, pages 69–96. Edward Elgar Publishing Ltd., 2013.

[13] Lawrence Gordon and Martin Loeb. The Economics of Information Security Investment. *ACM Transactions on Information and System Security*, 5(4):438–457, 2002.

[14] Sanjeev Goyal, Hoda Hiedari, and Michael Kearns. Competitive Contagion in Networks. *Games and Economic Behavior*, 2014, forthcoming.

[15] Jennifer Granick. The Price of Restricting Vulnerability Publications. *International Journal of Communications Law & Policy*, 9:1–35, 2005.

[16] Marco Haan. Vaporware as a Means of Entry Deterrence. *Journal of Industrial Economics*, 51(3):345–358, 2003.

[17] Charles Kolstad, Thomas Ulen, and Gary Johnson. Ex Post Liability for Harm vs. Ex Ante Safety Regulation: Substitutes or Complements? *American Economic Review*, 80(4):888–901, 1990.

[18] Howard Kunreuther and Geoffrey Heal. Interdependent Security. *Journal of Risk and Uncertainty*, 26(2-3):231–249, 2003.

[19] Stephen Morris. Contagion. *Review of Economic Studies*, 67(1):57–78, 2000.

[20] A. Mitchell Polinsky and Steven Shavell. Mandatory Versus Voluntary Disclosure of Product Risks. *Journal of Law, Economics, & Organization*, 28(2):360–379, 2010.

[21] Michael Riordan. Economic Incentives for Security. Powerpoint Slides presented at Cybercriminality Seminar at Toulouse School of Economics on 4 June, 2014.

[22] Steven Shavell. A Model of the Optimal Use of Liability and Safety Regulation. *RAND Journal of Economics*, 15(2):271–280, 1984.

[23] Hal Varian. System Reliability and Free Riding, 2004. Available at `http://people.ischool.berkeley.edu/~hal/Papers/2004/reliability` (accessed 1 December, 2013).