The Impact of Patent Life on Total Welfare: A Policy Experiment*

An-Hsiang Liu† Ralph Siebert‡ Christine Zulehner§

September, 2012

Abstract

This paper evaluates how different lengths of patent lives impact market structure and market performance. We formulate a dynamic oligopoly model in the tradition of Ericson and Pakes (1995) and allow entry costs to vary over time. Firms produce multiple products and decide on the optimal timing to enter a market, followed by production and exit decisions. Using quarterly firm-level data on the static random access memory industry from 1974 to 2003, we find that entry costs decline by more than 90% within the first ten quarters. Our policy experiments provide evidence that the life of product patents has a negative impact on consumer surplus. We also find that product patents increase total surplus if the patent life is either sufficiently short or sufficiently long. If patent life is short, the increase in monopolist's profits and entry cost saving dominate the reduction in consumer welfare, which affects total welfare positively. If patent life is long, dynamic efficiency gains, i.e., the delay of subsequent entry and savings on entry costs impact total welfare positively.

JEL: C1, L1, L6, O3.

Keywords: Adoption of Technologies, Development Costs, Entry Costs, Dynamic Oligopoly, Innovation, Intellectual Property Rights, Market Structure, Patent Life, Policy Experiment, Semiconductor Industry.

*We thank seminar participants at Purdue University and the University of Mannheim for their useful comments. We especially thank Victor Aguirregabiria, Jack Barron, Lanier Benkard, Gautam Gowrisankaran, David Hummels, Stephen Martin, Stephen Ryan, Philipp Schmidt-Dengler, Anson Soderbery, Yuya Takahashi, Justin Tobias, and Chong Xiang for their helpful discussions and valuable suggestions on this or an earlier draft. We gratefully acknowledge financial support from the Jubiläumsfonds of Oesterreichische Nationalbank (grant 11038). All errors are our own.

†Purdue University, Department of Economics, Krannert School of Management, Address: 403 West State Street, West Lafayette, IN 47907-2056, United States, Email: liu166@purdue.edu.

‡Purdue University, Department of Economics, Krannert School of Management, Address: 403 West State Street, West Lafayette, IN 47907-2056, United States, Email: rsiebert@purdue.edu.

§Johannes Kepler University Linz and Austrian Institute of Economic Research (Wifo), Address: Altenberger Strasse 69, A-4040 Linz, Austria, Email: christine.zulehner@jku.at.

1 Introduction

As argued by policy makers as well as economists one of the main components that stimulates economic growth is technological progress, see e.g. Solow (1957). Prominent studies focus on examining how different market structures relate to technological progress. Schumpeter (1942) argues that monopolistic markets are the major force behind technological advance. A monopoly position allows innovators to earn high profits and to recoup their research and development (R&D) investments in a short time period. Consequently, firms' incentives to invest in R&D increase, which promotes technological advance. Arrow (1962), on the other hand, claims that rather competitive pressure stimulates technological advance. Other empirical studies confirm an inverted U-shaped relationship between research intensity and market structure across industries, see e.g., Aghion et al. (2005), Levin, Cohen and Mowery (1985) and Scherer (1967).[1]

Most countries adopt the Schumpeter hypotheses and assign intellectual property rights to innovators by granting patents to them. Patents provide Intellectual Property Rights to firms and avoid the problem that competitors imitate their innovation. Patents grant a monopoly position to innovators for a certain time period, which differ across countries.[2] In the last twenty years, the U.S. Congress adopted several changes regarding patent laws. In the 1980's significant efforts have been undertaken in designing standards on U.S. Intellectual Property Rights with the purpose of unifying patent rights. In terms of patent life, the original patent term under the 1790 Patent Act was fourteen years. The 1836 Patent Act provided an extension of seven years from and after the expiration of the first term. In 1861 the seven year extension was eliminated and the term changed to seventeen years. In 1995, a major patent law change was enacted in the U.S., which extended the patent life from 17 from the date of issue to 20 years from the earliest filing date.

Several debates address the merits of patent protection. Research-based firms argue that patent laws also allow firms to earn monopoly profits and to recoup R&D investments in a shorter time period. Hence, firms' incentives to engage in research increases, which fastens the pace of technological advance. Moreover, patents might also be socially beneficial as they avoid excessive entry into markets. However, patents also have the following shortcomings: they confer

[1] In the automobile industry, however, there is a positive correlation between price-cost margins and the innovation intensity at the firm level, see e.g. Hashmi and van Biesebrock (2010). Other prominent studies in this area are Griliches (1984), Mansfield (1963), Saloner and Shepard (1995), Rose and Joskow (1990), and Hannan and McDowell (1984).

[2] For more information on patent rights in different countries, see Moser (2005).

monopoly power to firms. An extended patent life implies that monopoly prices will be charged for a longer time period which reduces consumer welfare. Gilbert and Shapiro (1990) contend that strong patent laws raise the number of innovations. The negative impacts that monopolies have on welfare make limitations on the legal protection conferred by intellectual property rights necessary. The optimal patent design considers the length and breadth as the two adjustable dimensions of a patent. Nordhaus (1969), for example, argues that 17 years of patent duration is close to optimal for many industries.

Up to date, only few empirical studies examine the optimal patent length. Hence, very little is known how patent protection affects consumer and producer surplus.[3] The objective of our study is to evaluate how different lengths of patent lives impact total surplus in the semiconductor industry. Evaluating the impact of the length of patent life on total welfare is a relevant but nontrivial task. Firms account for the fact that different lengths of patent lives determine market structure, market prices and firms' discounted payoffs. Hence, the length of patent lives crucially determines firms' incentives to enter. For example, a longer patent life protects the first entrant's monopoly position and allows the first entrant to earn profits for a longer time period due to lower competition. Firms account for the benefit of entering early and earning higher profits (see e.g., Fudenberg and Tirole, 1985, Riordan, 1992, and Klette and Kortum, 2004), which might be beneficial to technological advance. A longer time of patent protection delays the entry time of successive entrants and lowers their profits. Therefore, longer patent lives might not leave sufficient profits for later entrants and do not provide sufficient incentives for them to enter at all.

We frequently observe that entry or development costs decline over time.[4] This especially applies to the semiconductor industry, as early entry requires a higher pace of innovation and higher investments in R&D. Against the background of declining entry costs, determining the optimal timing to enter a market becomes a challenging task.[5] Firms face a trade off in deter-

[3]Some simulation studies (e.g., Fink (2000); Maskus and Konan (1994); Subramanian (1995); and Wattal (2000)) evaluate the welfare losses implied by patent protection. To the best of our knowledge, Chaudhuri, Goldberg and Jia (2006) is the only study that empirically estimates the impact of patent laws on welfare. They find that the Indian pharmaceuticals market would suffer from enormous losses in consumer welfare if the Indian pharmaceutical market was protected by patent laws similar to the U.S.

[4]We call the development costs from now onwards entry costs that need to be incurred to develop a new product and to enter a market.

[5]For studies on the optimal timing of entering new markets or adopting new technologies see e.g., Ackerberg and Gowrisankaran (2006), Einav (2010), Gowrisankaran and Stavins (2004), Schmidt-Dengler (2006), and Sweeting (2006 and 2009). Our study also relates to the literature on entry. The pioneering studies on entry by Bresnahan and Reiss (1990, 1991) show that the first 2-3 entrants have the largest impact on market price, and later entrants do not significantly reduce market price any further.

mining the optimal time to enter a market. On the one hand, the firm would like to enter early in order to ensure higher profits with fewer competitors. On the other hand, the firm has an incentive to delay entry because the new technology becomes cheaper over time, i.e., declining entry costs. The optimal entry time is determined at a point where the marginal profits from entering early equals the marginal cost from entering sooner, see also Kamien and Schwarz (1982).[6] Up to date, only very little is known on how entry costs evolve over time. This is mostly due to the fact that entry is associated with product specific R&D investments, which are rarely observed. This missing data problem represents a challenging task for identifying time-varying entry costs.

However, as is shown by theoretical studies free entry can lead to excessive entry causing inefficiencies, see e.g., Mankiw and Whinston (1986).[7] Using data on the U.S. commercial radio broadcasting industry, Berry and Waldfogel (1999) estimate that the loss on total welfare due to excessive entry relative to the social optimum is of about 45% of revenues. In a more recent study, Hsieh and Moretti (2003) show that a socially inefficient large number of real estate agents enter markets (cities) characterized by high housing prices.

To the best of our knowledge our model is the first study that incorporates a fully dynamic model in which firms make entry, exit and production decisions. We also allow entry costs to vary over time. To evaluate how different lengths of patent lives impact total surplus, we formulate a dynamic oligopoly model in the tradition of Ericson and Pakes (1995). Firms are allowed to produce multiple products and decide on the optimal timing to enter a market, followed by production and exit decisions. Our estimation procedure uses quarterly firm-level data on the static random access memory (SRAM) industry from 1974 to 2003. A static random access memory chip is a type of semiconductor memory classified into generations according to its information storage capacity.[8] The industry provides a natural setting for our policy experiment (to evaluate the impact of different patent lengths on total welfare) for a number of reasons. First, the SRAM industry is characterized by free market entry, as firms are limited to patent new processes rather than new products. Hence, firms are not able to patent the information capacity stored on a memory chip. In contrast, firms in the pharmaceutical market file for patents related to product innovations. Second, the SRAM industry is a high tech industry, which is characterized by rapidly declining entry costs.

[6]In other words, the optimal time of entering is determined at a point where the present discounted value of profits exceeds the present value of R&D costs by the largest amount.

[7]Note that free market entry corresponds to a patent life of zero periods in our study.

[8]A detailed industry description is provided in section 2.

Our estimation results show a considerable decline in entry costs of more than 90% within the first 10 quarters. The policy experiments show that different lengths of patent lives always reduce consumer welfare compared to the free entry case. Most interestingly is the fact that total welfare increases if the length of patent lives is sufficiently short or long. If patent life is very short, the savings of entry cost can cover the losses in consumer surplus and industry profits that makes total welfare to improve. If the length of patent life is in the intermediate range, the increase in producer surplus and the savings on entry costs from prohibiting excessive entry is not large enough to compensate for the reduction in consumer surplus. If the patent life is sufficiently long, excessive entry is limited which saves on entry costs. Moreover, the monopolist's profits significantly increase due to own learning effect. In this case, the entry cost savings and the monopolist's gain in profit are so large that they dominate the losses related to prohibiting other firms to enter the market.

The remainder of the paper is organized as follows. In the next section, we describe the patent system as well as the semiconductor industry emphasizing the development of new product generations and new process technologies. We also describe the industry and present summary statistics. Section 3 introduces our dynamic oligopoly model and Section 4 describes the econometric model. In Section 5 we discuss the estimation results and the results for our counterfactual experiments. We conclude in Section 6.

2 Institutional and Industry Background

In this section, we briefly describe the U.S. patent system and discuss it in the context of the semiconductor industry of which the SRAM industry is a part. Then, we describe the SRAM industry in more detail and also provide a description of the data and the summary statistics.

2.1 Patents

The Congress shall have power ... to promote progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries. U.S. Constitution, Article 1, Section 8. (See also Carlton and Perloff, 2005).

Under the Agreement on Trade-Related Intellectual Property Rights, the World Trade Organization members are required to enforce patents for certain industries. In order to receive a patent, the inventor must prove that the invention is useful, novel and not obvious. The most common type of patents are utility patents. Utility patents cover different types of inventions, such as mechanical devices, chemical compositions and processes, manufacturing methods, computer software, biotechnology and business methods.

The types of inventions which are patented vary across industries. In some industries, firms patent new products, whereas in other industries they patent new processes. For example, in the pharmaceutical industry, firms file patents for chemical compounds to ensure a monopoly position for a specific drug. Since it is difficult for other firms to patent around a chemical compound, a firm's monopoly position for a specific drug is highly protected.

On the contrary, in the semiconductor industry, firms mostly file patents for new process technologies rather than new products, i.e., firms are not able to claim a patent on the capacity of information storage on memory chips. Since semiconductor firms patent around other firms' process technologies, the products are not as highly protected in the semiconductor industry compared to other industries such as the pharmaceutical industry.[9]

A patent system has the following positive implications on social welfare: it overcomes the appropriability problem, i.e., inventions and information are not protected and spill over to opponents at no cost. Hughes, Moore and Snyder (2002) examine the long run impact of a policy change in the U.S. pharmaceutical industry. They find that that prices and future innovations go down if the government abruptly ended patent protection. However, patents prevent other firms from entering markets for a certain time, and reduces the pace of the diffusion process of new technologies.

Since the first U.S. Patent Act passed Congress in 1790, more than 5 million patents have been granted by the U.S. Patent and Trademark Office. The Office grants patents for initially 3.5 years, which can be extended to up to 20 years.

2.2 The SRAM Industry

In the following, we describe the innovation and manufacturing process for SRAMs. SRAMs belong to the family of semiconductors, which are primarily used as an input in computers, electronic devices like digital cameras or cell phones, automotive products and household appliances, among many others. The semiconductor industry is a high tech industry and has a

[9] Levin, Klevorick, Nelson and Winter (1987) report that R&D managers rated patents as the least effective mechanism of protecting an invention and appropriating the returns. Moreover, patents are considered to be more effective for protecting new products, rather than new processes.

significant impact on productivity growth. Most firms in this industry invest more than 30% of their revenues in R&D.

Semiconductors consist of memory chips, microprocessors, and application-specific integrated circuits. SRAMs are memory chips designed for the purpose of storing and retrieving information. The chips are classified into generations according to their storage capacity of information. The technological progress of memory chips is frequently described by Moore's Law. According to Moore's Law, the number of transistors per chip doubles every two years. A consequence of Moore's Law is that the number of transistors results in a fourfold increase in memory capacity per chip. The increase in memory capacity per chip is therefore determined by a constant technological relationship.

A higher capacity of SRAM chips requires advanced technological improvements along different dimensions, i.e., different architectures of transistor cells which allow for smaller transistor and cell sizes, an increase in the die area, lower temperature for performing operations, lower energy consumption, advanced masks and lenses to improve etching processes. To improve the scaling of SRAMs tremendous R&D investments are necessary. For example, in 2004, Samsung, Hynix, Toshiba, Intel and Micron announced a capital spending between 2 and 4.5 billion U.S. dollars. Note that these numbers are rarely reported and also represent averages only.

The manufacturing process of SRAMs is highly complex. Advanced photolithographic and chemical processes to etch electrical circuits onto the wafers surface are necessary to improve a chip's performance. Memory chips are cut from silicon wafers, which determines silicon as the base material entering the manufacturing process.

The production process of every new chip generation is characterized by significant learning effects throughout the life cycle. The production yield rate, which is defined as the percentage of wafers that successfully pass all production stages, can increase from around 20% to more than 80% over the life cycle. The increasing yield rate is a result of repetition of production processes, which increases the efficiency of operations and reduces the amount of material waste and other manufacturing errors. Learning effects determine firms' production decisions as more production experience generates higher cost savings. Beyond learning from own experience, firms also learn from their competitors' accumulated experience spilling over through labor turnover and information transfers. Finally, learning also results from information transmitted between generations. Spence (1981) as well as Fudenberg and Tirole (1983) demonstrate that such learning effects can have a sizable impact on costs, strategic decisions, and market performance. Ghemawat and Spence (1985) illustrate the impact of learning affects on competition and market performance.

Learning also impacts firms' optimal timing decisions to enter a new chip generation. If learning from own experience within a generation is strong, firms will enter earlier, see, e.g., Spence (1981), Fudenberg and Tirole (1983) and Dasgupta and Stiglitz (1988). If intra-generational learning spills over from other firms, barriers to entry will decline (Ghemawat and Spence, 1985). Experience can also be transferred between generations through intergenerational spillovers, see Irwin and Klenow (1994). Those spillovers provide incentives for firms to enter earlier.

2.3 Data description and summary statistics

Next, we introduce the data sources and provide some descriptive statistics. Our study uses firm level and industry level information on prices and quantities for different SRAM generations which are compiled by Gartner Inc. The data set encompasses six product generations, namely the 4Kb, 16Kb, 64Kb, 256Kb, 1Mb and 4Mb generations. It entails quarterly data from January 1974 until December 2003. The data set covers firm and industry units shipped, the average selling price, and the number of firms in the market.

Figure 2 shows the average prices of the 4Kb to 4Mb SRAM generations over time. Prices are rapidly declining throughout the product life cycle. An important phenomenon that explains this price decline for SRAMs are learning effects, see the discussion above. Figure 1 shows how industry shipments of the 4Kb to 4Mb SRAM generations evolve over time. Shipments of each generation are characterized by product life cycles, which take on similar shapes across different generations. At a certain point of time, multiple SRAM generations are offered on the market.

Figure 3 shows the number of firms present in the 4Kb to 4Mb SRAM generations over time. The number of firms follows an inverse U-shape within every generation. Interestingly, firms continue to enter the market many years after the product generation has been launched. To illustrate market structure in more detail, Figure B shows the entry process for the 64Kb generation. The figure illustrates that firms enter the 64Kb generation even 20 years after the generation has been launched. We observe that about half of the firms enter in the first 5 years. Most firms enter during the 8^{th} to the 33^{rd} quarter while less than 10% of the firms enter in the first two years.

The fact that firms simultaneously produce multiple product generations indicates that firms are able to benefit from learning effects across generations. In light of intergenerational learning effects, multiproduct firms might have an incentive to produce more output in previous generations to gain more experience and lower their production costs in the new generation, which increases profits and allows them to enter markets earlier. It is important to mention that intergenerational learning is one major reason why firms successively enter into different generations - rather than skipping one generation - in order to achieve an early headstart for production in subsequent generations. This fact is important in our policy experiment to make sure that firms will not skip producing a generation once patent rights become sufficiently long. Moreover, multiproduct firms also account for substitution patterns originated from the demand side. If chips are substitutes, the production of different chip generations will cannibalize the sales of other generations. Consequently, multiproduct firms will account for these substitution patterns when setting their production schedules for different generations.

3 Dynamic oligopoly model

In this section we construct a discrete-time infinite horizon model with time indexed by $t = 0, 1, \ldots, \infty$. Firms, denoted by $i = 1, 2, \ldots, N$, maximize the sum of profits over all periods for a specific generation k.[10] The model is formulated as a state space game in which firms use Markov strategies. Firms' actions in a given period determine not only their own and rival firms' current profits, but also their own and rival firms' future states. We build on the fact that firms are rational and forward-looking, i.e. they derive their discounted profit streams given the evolution of the state vector and their actions. At the beginning of each period, firms simultaneously decide whether to enter generation k and how much to produce in generation k. Firms who entered in generation k in the past, also decide whether to exit generation k or not.[11] We focus on pure entry strategies. Entering a new product generation incurs an entry

[10] We simplify our entry model to one generations and believe that this is an appropriate assumption when estimating time-variant entry costs for the following reasons. First, different product generations are usually produced at different fabs. Therefore, firms' output decisions are usually made at the fab level. Second, firms that consider to enter generation k in period t only know about the existence of generation k-1, but do not know of generation k+1 yet. Moreover, firms who actively produce one generation, introduce subsequent generations because of intergenerational learning effects. Therefore, firms skipping one generation to achieve an early headstart for the next generation is not a concern in the semiconductor industry. When considering the entry of generations k, we treat the states from the previous generation as predetermined and given from the data. The assumption might be restrictive, however, for other industries and should be reconsidered.

[11] Exit does not involve any cost or scrap value.

cost of C_t^k, where the subscript t refers to the time period when entry occurrs. The entry cost is deterministic and larger than zero. Since the entry cost is primarily a function of technological knowledge and inputs, we assume the function to be industry and generation specific, rather than firm-specific.[12]

The observable state variables to every firm at period t are denoted by a vector $s_{it} = (n_t^k, x_{it}^k, x_{-it}^k, C_t^k, z_t^k, X_{it}^{k-1})$, which contains all of the payoff relevant state variables for generation k. The variables n_t^k refer to the number of firms in period t for generations k and are supposed to capture the degree of competitiveness in the markets for generation k. We also account for firms' learning capabilities. Firms are supposed to learn from past experience lowering marginal costs. We distinguish between learning experience accumulated within a generation (intragenerational learning) and learning experience accumulated in the previous generation (intergenerational learning). With regard to intragenerational learning effects, we allow firms to learn from their own experience (x_{it}^k) as well as from other firms' experience via spillovers (x_{-it}^k). z_t^k includes all the demand and cost shifters which are known to the industry. $X_{it}^{k-1} = (n_t^{k-1}, x_{it}^{k-1}, x_{-it}^{k-1})$ are state variables from the previous generation. These variables are assumed to be exogenous and given from the data. The entry of generation k may be determined by the relative competitiveness in the markets of generations $k(n_t^k)$ and $k-1$ (n_t^{k-1}). We allow firms to learn from their own production experience in the previous generation (x_{it}^{k-1}). We also account for firms' intergenerational learning effects via spillovers resulting from other firms' production experience in the previous generation (x_{-it}^{k-1}). Finally, the time-variant entry cost C_t^k enter the state vector.

3.1 Evolution of States

Number of firms

We define the transition of the number of firms n_t^k within generation k as follows:

$$n_t^k = n_{t-1}^k + ne_t^k - nx_t^k, \tag{1}$$

where ne_t^k is the number of entering firms and nx_t^k the number of exiting firms.

[12] Note that firms might have different entry times in generation k due to the fact that we allow for heterogeneous firm sizes in generation $k-1$.

Learning

Firm i's marginal cost of production is characterized by several learning components. The first component describes a firm's intragenerational learning from own experience. Firm's own learning within generation k is given as follows:[13]

$$x_{it}^k = x_{it-1}^k + q_{it-1}^k, \tag{2}$$

where q_{it-1}^k is firm i's output in generation k at time $t-1$ and $x_{i0}^k = 0$, assuming that there is no own experience in the beginning of the product cycle of each product generation. The other components describe the extent to which firm i learns from other firms' experience via spillovers within and between generations. Intragenerational spillovers are specified as follows:

$$x_{-it}^k = x_{-it-1}^k + \sum_{j \neq i} q_{jt-1}^k, \tag{3}$$

with $x_{-i0}^k = 0$, assuming that there are no spillovers in the beginning of the product cycle of each product generation. In general, potential entrants in a specific generation may benefit from own experience in the previous generation as well as intragenerational and intergenerational spillovers.

Productivity shocks

In each period, each incumbent in generation k receives private productivity shock ν_{it}^k. The shocks are drawn from a known normal distribution G^k with zero means and constant variances σ_k^2 and are independently and identically distributed across firms and across periods.

[13] Note that the laws of motion for the state variables are deterministic.

Marginal costs

The marginal cost function depends on the set of payoff-relevant state variables and follows the specification in Irwin and Klenow(1994) to account for own learning and spillover learning. Firm i's marginal cost mc_{it}^k is specified as follows:

$$\ln(mc_{it}^k) = \rho_0^k + \rho_i + \rho_1 \ln(P_t^{sil}) + \rho_2 \ln\left(x_{it}^k + \rho_3 x_{-it}^k + \rho_4 \left[x_{it}^{k-1} + \rho_3 x_{-it}^{k-1}\right]\right) + \nu_{it}^k, \tag{4}$$

Marginal costs are described by generation-specific intercepts (ρ_0^k), firm fixed effects (ρ_i), the price of silicon P_t^{sil} and several learning components, i.e., intragenerational own learning (x_{it}^k) and learning from others via spillovers (x_{-it}^k), intergenerational own learning (x_{it}^{k-1}) and learning through spillovers x_{-it}^{k-1}. The parameters ρ_2 and ρ_4 measure the impact of the own intra- and intergenerational learning effects on the marginal costs. The parameter ρ_3 allows for a separate impact of own learning and learning via spillovers on marginal costs.

Strategies

Firm i's strategy is given by $\sigma_i^k(s, \nu_i)$, where σ_i is the set of firm i's strategies that describe the production, entry and exit decisions. The set of state vectors (s_{it}) are summarized in s and ν_i includes the generation-specific firm-level productivities. Rival firms' strategies are denoted by σ_{-i}.

3.2 Profit

Each firm maximizes its future discounted profits from generation k. If a firm is active in generation k, its per period profit is given by:

$$\pi_{it}^k(\sigma_i^k, \sigma_{-i}^k, s_t, \nu_{it}^k) = q_{it}^k[p_t^k - mc_{it}^k(\nu_{it}^k)],$$

where the variable p_t^k represents the price for generation k. The price is a function of the industry output $q_t^k = \sum_i q_{it}^k$ and observable demand shifters d_t. Hence, firm i's per period profits for generation k at time t depend on the actions of all firms, the state vector, and firm i's private shock. A firm earns zero profit if it is inactive in generation k.

If behavior is given by a Markov perfect equilibrium strategy profile $\sigma^k = (\sigma_i^k, \sigma_{-i}^k)$, firm i's value function can be written recursively:

$$V_i(s,\sigma^k;\theta) = E_\nu\left[\pi_i^k(\sigma^k,s,\nu_i^k) - C^k(\sigma^k) + \beta\int V_i(s';\sigma^k)dP(s'|\sigma^k,s)\middle|s\right], \tag{5}$$

where θ denotes a vector of parameters and P describes the evolution from today's state s to tomorrow's state s'.

3.3 Timing

The timing of our model can be summarized as follows. In each period t, events occur in the following order:

1. Firm i considers the following state $s_{it} = (n_t^k, x_{it}^k, x_{-it}^k, C_t^k, z_t^k, X_{it}^{k-1})$.
2. Given the state vector, potential entrants make their entry decisions for generation k and incumbents make their exit decisions for generation k .
3. Incumbent firms in generation k observe their productivity shocks ν_{it}^k, choose outputs q_{it}^k and realize profits.
4. The state moves to $(n_{t+1}^k, x_{it+1}^k, x_{-it+1}^k, C_{t+1}^k, z_{t+1}^k, X_{it+1}^{k-1})$.

4 Econometric model

In this section, we describe our econometric model. We apply the two-step estimation method developed by Bajari, Benkard and Levin (2007) to estimate the cost of entering the $64K$ SRAM generation. The first stage includes the estimation of the demand, the policy functions and the marginal cost functions. All policies are functions of the observed state, s, including the entry cost C_t^k and the productivity shocks ν_{it}^k which are unobserved to the econometrician. We include a time trend in all of the policy functions to proxy for the declining entry costs over time. In the second stage, we apply forward simulations to calculate the continuation values. Finally, we recover the entry cost over time that rationalizes firms' policies.

4.1 First Stage

Demand

We assume that every generation is homogeneous in itself, but different generations represent differentiated goods. The demand function is specified log-linearly as follows:

$$\ln(q_t^k) = \alpha_0^k + \alpha_1^k \ln(P_t^k) + \alpha_2 \ln(P_t^S) + \alpha_3 \ln(GGDP_t) + \alpha_4 \ln P_t^{DRAM} + \alpha_5 Time_t + d_t^k, \quad (6)$$

where we denote the vector of coefficients with α. We account for generation-specific demand shifters (α_0^k). P_t^k is the price for generation k. Note that we allow for generation-specific price elasticities (α_1^k). P_t^S is the price for the substitutable generation.[14] $GGDP_t$ is the growth rate of GDP, P_t^{DRAM} is the price of DRAMs and $Time$ refers to a time trend. We estimate equation (6) for generation $k = 64K$ and include the $16K$ and $256K$ product generations as substitutes.

Marginal Costs

In line with our theoretical model, we retrieve firms' marginal costs for $k = 64K$ generations. We recover firm i's marginal costs for the two generations (mc_{it}^{64K}) using the estimated price elasticities $\hat{\alpha}_1^{64K}$ of our demand equation (6) and the observed market shares, ms_{it}^{64K}, see also Siebert and Zulehner (2011):

$$mc_{it}^{64K} = P_t^{64K}\left(1 + \frac{ms_{it}^{64K}}{\hat{\alpha}_1^{64K}}\right). \quad (7)$$

Output Policy

The output policy depends on the set of observed state variables, the productivity shock and a time trend. Hence, the output policy accounts for factor prices, demand shifters, market structure, intra- and intergenerational own learning and learning via spillovers. The output function for the $64K$ generation is specified as:

$$\ln(q_{it}^{64K}) = \gamma_i + \gamma_1 \ln(P_t^{sil}) + \gamma_2 \ln P_t^{DRAM} + \gamma_3 GDP_t + \gamma_4 \ln(n_t^{16K}) + \gamma_5 \ln(n_t^{64K}) + \gamma_6 \ln(x_{it}^{16K})$$
$$+ \gamma_7 \ln(x_{it}^{64K}) + \gamma_8 \ln(x_{-it}^{16K}) + \gamma_9 \ln(x_{-it}^{64K}) + \gamma_{10} Time_t + \gamma_{11} \hat{\nu}_{it}^{64K} + \epsilon_{it}^{64K}, \quad (8)$$

where γ_i is a firm level fixed effect. We assume that the error term ϵ_{it}^{64K} is i.i.d normally distributed. $\hat{\nu}_{it}^{64K}$ is the residual from the marginal cost estimation and is interpreted as the productivity shock. A positive shock rises the marginal cost of production and hence reduces a

[14] As substitute we use the average SRAM price of the adjacent generations.

firm's output. Thus, we should expect a negative sign for γ_{11}.

Entry and Exit Policies

We estimate the following probit model for entering the $64K$ generation:[15]

$$Pr(e_{it}^{64K} = 1; s) = \Phi\Big(\lambda_i + \lambda_1 \ln(n_t^{64K}) + \lambda_2 \ln(x_{it}^{16K}) + \lambda_3 \ln(x_{-it}^{64K}) + \lambda_4 Time_t\Big), \tag{9}$$

where $\Phi(.)$ is the cumulative distribution function of the standard normal. The entry policy depends on the number of firms active in the $64K$ generation (n_t^{64K}), firm i's experience from the previous $16K$ generation (x_{it}^{16K}), the spillover from other firms in the $64K$ generation (x_{-it}^{64K}) and a time trend.

The exit policies for the $64K$ generation is formulated as:

$$Pr(ex_{it}^{64K} = 1; s) = \Phi\Big(\psi_0 + \psi_1 \ln(n_t^{64K}) + \psi_2 \ln(x_{it}^{64K}) + \psi_3 \ln(x_{-it}^{64K}) + \psi_4 Time_t\Big), \tag{10}$$

where the probability of exit is a function of the number of firms, the own experience and other firms' experience in the current generation.

4.2 Second Stage

Using the estimated policy functions, we take draws for the productivity shocks for each incumbent in each period and forward simulate the sequences of optimal outputs for each firm. In each simulation l, we record t_{il}^* (firm $i's$ optimal timing of entering the $64K$ generation in simulation l). Using $N \times L$ observations, where N is the number of firms and L is the number of simulations, we generate further potential equilibrium paths. Hence, based on the optimal timing (t_{il}^*), the demand equation, the marginal cost function, the drawn productivity shocks and the optimal policies, we compute a simulated sequence of outputs. Next, we compute the discounted profits before subtracting the entry costs, i.e., $\widetilde{V}_{il}(t_{il}^*)$. To summarize, we compute the value functions applying the following steps for each simulation:

1. We draw one productivity shock for each incumbents at each period i.e., ν_{it}^{64K}.

[15] Note, since our theoretical model explicitly formulates entry in generation k, we only have to estimate the entry policy for the $64K$ generation. This is also sufficient for our policy experiments to evaluate the lengths of patent lives in the $64K$ generation.

2. We use the optimal production, entry and exit policies to update firms' production choices $\hat{q}_{it}^{64K}$ and state variables $\hat{s}_{it}$. Note that the draws of productivity shocks enter $\hat{q}_{it}^{64K}$ and $\hat{s}_{it}$(see equation 8).

3. For every period, we update prices $\hat{p}^{64K}(\hat{q}_{it}^{64K}, \hat{q}_{-it}^{64K})$ using the estimated demand equation.

4. We compute the vector of marginal costs $\hat{mc}_i^{64K}$ using the estimated marginal cost function, $\hat{q}_{it}^{64K}$, $\hat{s}_{it}$ and ν_{it}^{64K}.

5. We calculate the value functions: $\widetilde{V}_{il}(t_{il}^*, t_{-il}^*) = \sum_{t=0}^{\infty} \beta^t(\hat{\pi}_{it}^{64K})$.

In order to identify the evolution of entry costs over time, we compare the changes in the discounted values at the optimal entry times with the values at the distorted and therefore suboptimal entry times to derive the upper and lower bounds for the entry costs.

Regarding the upper bound of the entry costs, we compare the discounted values (net of entry costs) at the optimal entry time $\widetilde{V}_{il}(t_{il}^*) - C(t_{il}^*)$ with the discounted value (net of entry costs) from entering one period later, i.e., $\widetilde{V}_{il}(t_{il}^* + 1) - C(t_{il}^* + 1)$. Based on firms' rational expectations, we can write:

$$\widetilde{V}_{il}(t_{il}^*) - C(t_{il}^*) \geq \widetilde{V}_{il}(t_{il}^* + 1) - C(t_{il}^* + 1)$$

Rearranging this inequality yields:

$$\Delta C^{max} = \widetilde{V}_{il}(t_{il}^*) - \widetilde{V}_{il}(t_{il}^* + 1) \geq C(t_{il}^*) - C(t_{il}^* + 1), \tag{11}$$

where ΔC^{max} defines the maximum change in entry costs between period t and period $t+1$ and represents the upper bound of the slope of the entry costs, see Figure 5.

Regarding the lower bound of the entry costs, we compare the discounted values (net of entry costs) when the optimal entry time is $(t_{il} + 1)^*$ with the discounted value (net of entry costs) evaluated at an earlier suboptimal entry time (t_{il}):

$$\widetilde{V}_{il}(t_{il} + 1)^* - C(t_{il} + 1)^* \geq \widetilde{V}_{il}(t_{il}) - C(t_{il})$$

Again, rearranging the inequality gives:

$$\Delta C^{min} = \widetilde{V}_{il}(t_{il}) - \widetilde{V}_{il}(t_{il}+1)^* \leq C(t_{il}) - C(t_{il}+1)^*, \tag{12}$$

where ΔC^{min} represents the lower bound of the slope of entry cost and defines the minimum entry costs between period t and period $t+1$, see Figure 6.

Next, we search over the entry costs that best rationalize the two conditions - inequalities (11) and (12) - using a minimum distance approach. We therefore define a penalty, when inequalities (11) and (12) are violated.[16] Then, ΔC can be recovered by searching over c, that minimizes the weighted sum of penalties. Denoting $A_t \equiv \{\,(i,l) \mid t^*_{il} = t\,\}$ as the set of observations where firms optimally enter at period t, and $A_{t+1} \equiv \{\,(i,l) \mid (t_{il}+1)^* = t+1\,\}$ as the set of observations where firms optimally enter at period $t+1$:

$$\begin{aligned}\widehat{\Delta C(t)} = \arg\min_c \Bigg\{ \frac{1}{|A_t|}\sum_{A_t} \min\left[0, \widetilde{V}_{il}(t^*_{il}) - (\widetilde{V}_{il}(t^*_{il}+1) + c)\right] \\ + \frac{1}{|A_{t+1}|}\sum_{A_{t+1}} \min\left[0, \widetilde{V}_{il}(t_{il}+1)^* - (\widetilde{V}_{il}(t_{il}) - c)\right] \Bigg\}\end{aligned} \tag{13}$$

where $|\cdot|$ refers to the number of elements in the set.

To summarize, we perform the following steps in our estimation algorithm:

1. We estimate the demand equation (6) to retrieve the elasticity of demand.

2. We use the elasticity of demand in order to calculate the marginal costs using the Lerner Index, see equation (7).

3. We estimate the marginal cost function to retrieve the cost parameters entering in equation (4).

4. We estimate the production policies (equation 8), the entry policy (equation 9) and the exit policy (equation 10).

5. We use the cost parameters and the optimal policies to apply forward simulations and to calculate the value functions (before entry costs), i.e., $\widetilde{V}_{il}(t^*_{il})$ and $\widetilde{V}_{il}(t_{il}+1)^*$.

[16] A penalty is defined when the simulated discounted value (net of entry costs) is higher at a suboptimal entry time than at the optimal entry time.

6. We distort the optimal time to enter and use the policies to calculate value functions (before entry costs) i.e., $\widetilde{V}_{il}(t^*_{il}+1)$ and $\widetilde{V}_{il}(t_{il})$.
7. Finally, we use the penalty conditions and apply a minimum distance procedure to recover the slope of the entry costs (equation 13).

5 Estimation results and policy experiment

In this section, we describe our estimation results. We first discuss the results for the demand and the marginal cost function. Afterwards, we discuss the estimation results for the firms' production, entry and exit policies. Finally, we describe the results for the entry costs and discuss the results of our policy experiment.

5.1 First stage results

Demand

To estimate industry demand, we pool the data from all product generations and estimate equation (6) using ordinary least squares and instrumental variable estimation techniques. We instrument for potential correlation of the price with the error term using supply-side variables: price of silicon (P_t^{sil}) and cumulative industry output ($\sum_{t'=1}^{t-1} q_{t'}^k$) of the current generation. The estimation results are shown in Table 1. We estimate two specifications: the first specification incorporates a constant price elasticity across product generations (Columns 1 and 2). The second specification incorporates generation-specific price elasticities (Columns 3 and 4). We employ a Chow test to test for identical price elasticities across product generations. The F-statistics returns a value of 6.01 which leads us to reject the null hypothesis that prices elasticities are equal to each other.

Comparing the results in columns 3 and 4, the instrumental variable estimation returns more elastic price elasticities of demand than the ordinary least squares estimation. This result indicates an upward bias of the ordinary least squares estimates and gives rise to the fact that prices are endogenous. To test for endogeneity, we conduct a Durbin-Wu-Hausman test. The test statistics is $F_{(1,300)} = 619.99$ for the first specification and it is $F_{(4,293)} = 6.40$ for the second specification. We therefore reject the exogeneity of prices in both specifications.

In the following, we discuss the results shown in column 4. The estimated price elasticities range from −2.75 to −3.19, confirming a highly elastic market demand in the SRAM industry. In comparing the price elasticities throughout different generations, we find that the $64K$ generation is characterized by the least elastic demand, which explains the fact that the average prices in this generation stay at higher levels. Demand becomes more elastic in the subsequent $256K$ and $1M$ generations. Finally, our estimations return positive cross-price elasticities, confirming the fact that SRAM chips are substitutes across generations.

Marginal cost

The marginal cost function, equation (4), is estimated using a non-linear least square procedure. We assume that the error in the marginal cost function is normally distributed. The results are shown in Table 2. Column 1 reports the estimation results without a time trend. All coefficients are significantly different from zero, carry the expected sign and reveal economically sensible magnitudes. In addition, the model fits remarkably well with an adjusted R^2 of about 0.90. We find a positive effect of input prices on output prices. The results further return a learning parameter of −0.34, which corresponds to a learning rate of 21%.[17] This result confirms that the SRAM industry is characterized by strong intragenerational learning effects. Moreover, we find that there are spillovers within one generation and across generations. The spillover effect within one generation is of about 30%, i.e., firms are able to benefit from about a third of other firms' output. Compared to this effect, the intergenerational spillover effect is small. About 10% from the output produced in the earlier generation contributes to learning.

Column 2 reports the estimation results from a second specification that includes a linear time trend. The negative significant time trend reflects the fact that marginal costs decline over time. The inclusion of the time trend slightly alleviates the magnitude of intra- and intergenerational learning. The learning effects, however, still remain economically and statistically significant. Especially, the intergenerational spillover effect reduces its effect from about 1% to about 0.1%.

Since the residual of the marginal cost estimation is interpreted as private productivity shocks, we use the standard deviation of the residuals, 0.3, for the distribution of productivity shocks. Therefore, we take draws of productivity shocks from a normal distribution with zero mean and 0.3 standard deviation in the forward simulation.

[17] The learning rate is defined as $1 - 2^{\rho_2}$, i.e. the percentage reduction of marginal cost when (past) experience doubles, see also Irwin and Klenow (1994).

Output policy

We estimate the output policies specified in equation (8). Table 3 shows the estimation results for the $64K$ generation using an ordinary least square estimator.[18] An increase in the price of silicon increases marginal costs and hence reduces the output. The coefficients of the demand shifters carry the expected signs. Firms produce more when the size of the market is larger(GDP) and when the price of a substitute is higher(P^{DRAM}). The coefficients of the number of firms within a generation is negative and significant. A more competitive market, i.e. a larger number of firms in the market, reduces firms' output. Interestingly, however, firms produce more in the $64K$ generation when the $16K$ generation becomes more competitive.

Our estimation results also reconfirm the significant intragenerational learning effects. The coefficients of the own cumulative outputs in the current generation is positive and highly significant. This illustrates the fact that firms raise production in order to gain more production experience which reduces own marginal costs via learning effects. In terms of intragenerational spillovers, other firms' cumulative output within the $64K$ generation (x_{-i}^{64K}) has an impact which is not significantly different from zero. This result seems to be surprising as Table 2 shows that intragenerational spillovers are prevalent. The estimation results of the output policy can be explained by the existence of two opposing effects: a direct effect having an impact on a firm's production via spillovers, and an indirect effect, as a firm produces less when other firms' marginal costs decline. From our estimation results, both effects cancel each other out.

The intergenerational own learning effects (x_i^{16K}) and spillover (x_{-i}^{16K}) are insignificant. This result is consistent with the marginal cost estimates which show insignificant intergenerational learning effects. Finally, the estimate of the coefficient of $\hat{\nu}_i$ is negative and significant as expected. A positive shock which rises marginal cost will reduce output.

Entry and exit policy We estimate the probabilities of entry, equation (9), and exit, equation (10), using probit models. The results are shown in Table 4. The first column reports the results of the entry policy. The results show that firms are more likely to enter when more firms have already entered the $64K$ generation in the past. The significant and positive estimate

[18] We also estimated the output equations using an instrumental variable estimator to correct for a potential endogeneity of the lagged own cumulative output. As instruments we use the twice and three times lagged variables of own cumulative output. The differences between the estimated coefficients from the instrumental variable and the ordinary least squares estimations are negligible. In addition, the Hausman test fails to reject the consistency of the ordinary least squares estimates. Moreover, the Hausman test statistics of 97.76 reject the random effects specifications.

of cumulative output in the $16K$ generation emphasizes the importance of experience from earlier generations through intergenerational own learning effects. The estimated positive and significant time trend indicates that at identical industry states but different time periods, firms choose to enter in a later period in order to save on entry costs. This result provides a first insight that entry costs are decreasing over time.

Column 2 in Table 4 show the results for the exit policy.[19] Firms are more likely to exit the $64K$ generation if they missed out on gaining sufficient production experience, and if many other firms entered the market. Moreover, firms are less likely to exit if other firms' cumulative outputs are larger.

5.2 Second stage results

Using our estimates, we apply forward simulations to calculate the discounted values (before entry costs) at different states given the optimal policy functions. We compare these discounted values to those generated from distorting the entry policies as described in Section 4.2. We retrieve the slope of the entry cost for the $64K$ generation using a simulated minimum distance estimator as shown in equation (13). The resulting entry costs are illustrated in Figure 7.[20] Our results indicate that entry costs decline by 342 million U.S. dollars during the first 30 periods. More than 90% of the reduction of entry cost occurs during the first 10 periods. In the first 10 periods, entry costs rapidly decline with an average of 31.8 million U.S. dollars per period. The entry costs decline less steeply after the 11^{th} period, which is the time when excessive entry begins (see Figure B). Between the 11^{th} and the 30^{th} period, entry costs decline by about 1.2 million U.S. dollars per period on average. To summarize, our entry cost estimates nicely support the fact that early entrants incur significantly higher entry costs.

5.3 Counterfactual experiment

Based on our structural estimates, we conduct a policy experiment that evaluates the impact of different lengths of "patent lives" in the $64K$ generation on market structure and quantifies the impact on total welfare. We formulate the patent life regulation as follows: The regulator grants a patent to the first firm entering the $64K$ SRAM chip generation. The patent expires after

[19] We do not include firm fixed effect in the exit policy since not all firms have exited the $64K$ generation at the end of data period

[20] We normalized the entry costs presented in the figure by setting the entry cost of the 30^{th} period to 0.

a certain number of periods, which is specified by the regulator.[21] The patent holder receives the exclusive right to sell $64K$ chips for the time the patent is being granted by the regulator. Once the patent life expires, potential entrants decide if they enter the chip generation, pay the corresponding entry costs and decide how much to produce.

Evaluating the impact of patent life on total welfare is not a straightforward exercise, as a longer patent life delays entry and also reduces the entry costs for firms. A lower entry cost has a significant impact on firms' discounted values, their entry decisions, as well as their quantity choices, prices and market structure. Firms are assumed to choose the same equilibrium strategies. We simulate output, prices, entry and exit depending on the different patent lengths granted by the regulator. We also account for learning effects to evaluate the entire impact on consumer and producer welfare.

The results are illustrated in Figures 8 and 9. Note that all changes in total welfare are evaluated relative to the free entry case. Figure 8 shows the impact of the regulation on industry consumer and producer surplus. A longer patent life increases market power for the first entrant, which causes prices to increase, output to fall, and consumer surplus to decline. The change in producer surplus is always positive comparing to the case of free entry and increases as patent life increases. The conclusion of welfare improving by patent protection depends on the comparison of the increases in producer surplus and decreases in consumer surplus. For 3 and 4 periods of patent life, the total surpluses are larger than that under free entry. For 5 to 14 periods of patent life, the changes in total surplus are negative and continue to fall. This results reflect that the decreases in consumer surplus outweigh the increase in producer surplus. For 14 to 18 and 23 to 30 periods of patent life, the changes in total surplus rise due to the strong increase in the changes in producer surplus. While the losses in consumer surplus fall steadily and level off for longer patent lives, the gains in producer surplus increase more rapidly for longer patent lives. As a result, the changes in total surplus start to rise for longer patent lives and become positive for more than 27 periods of patent lives.

To better understand the change in producer surplus, we decompose the producer surplus into the monopolist's profits, other firms' profits and entry costs, see Figure 9. As patent lives increase, the protected monopolist's profits drastically increase relative to the free entry case for several reasons: First, granting the first entrant a patent right for multiple periods preserves its

[21] We assume that the original first entrant is being granted the patent. Hence, firms do not anticipate the regulation and other firms have no chance to outrun or leapfrog the original first entrant.

monopoly position and enables the protected entrant to earn monopoly profits during this time period. Under free entry other firms instantly enter the market. Hence, the change in market structure lowers the first entrant's profits. Since the patent life regime is compared to the free entry case, the monopolist's profits increase over time. Second, the monopolist produces the entire industry output and is able to further benefit from intragenerational own learning effects, which further lowers costs and increases profits. Comparing the profits gained by the monopolist to the profits lost by other firms, the monopolist's gain is greater than other firms' losses for shorter patent lives (less than 15 periods) but less than other firms' losses for longer patent lives(more than 15 periods). With sufficiently long patent lives, the protected monopolist can accumulate sufficient experience in early periods and accelerate the learning paces which enables other firms to benefit from spillovers. Consequently, the industry profits become positive relative to the free entry case for more than 15 periods of patent lives.

The figure also show that entry cost savings increase with the number of monopoly periods, which is due to the fact that longer patent lives delay successive entry. As patent lives increase, the increasing rate of entry cost savings depends on the decreasing rate of entry cost and the distribution of entry under free entry. When the entry cost declines faster and the entries of more firms have been delayed, the savings of entry cost are larger. From Figure 9, the savings of entry cost rise more rapidly for less than 10 periods of patent life because the entry cost decline drastically during the first 10 periods(see Figure 7). For more than 10 periods of patent life, the savings of entry cost increases relatively slowly in general because of much lower decreasing rates of entry cost(see Figure 7). However, for 14 to 17 and 25 to 30 periods of patent lives, the savings of entry cost rise more rapidly since the entries of a larger number of firms have been delayed by patent protection(see Figure B). The entry cost savings are always larger than the change in industry profits (the combination of the monopolist's profits and other firms' profits) which makes the producer surplus to be positive relative to free entry case. For sufficiently long patent lives(more than 25 periods), both the change in the monopolist's profit and savings of entry cost increase drastically and consequently lead to more rapid increases in producer surplus.

To summarize the main results of our policy experiment: for shorter patent lives, there are losses in total welfare because of the losses in industry profit and small savings of entry costs. When the length of patent life lasts for sufficiently long periods, the large savings in entry costs plus the gain in monopolist profits due to learning effect rise producer surplus. Consequently, the gains in producer surplus outweigh the losses in consumer surplus and total welfare improves. Hence, free entry drives too many firms entering too early and therefore leads to excessive entry.

6 Summary and concluding remarks

The purpose of this paper is to empirically investigate the impact of different patent lengths on consumer and producer surplus. We estimate a dynamic structural model in which firms choose the optimal time to enter a market, and make production and exit decisions. Our estimation results show that entry costs drastically decline over time. In fact, we find that entry costs decline by more than 90% in the first 10 quarters. Our estimation results also provide evidence that firms account for learning effects when determining their optimal production. Early entrants incur higher entry costs, but are able to take advantage of learning effects early in the life cycle, which lowers their future costs and returns higher profits. Late entrants especially benefit from decreasing entry costs over time as well as from other firms' experience via spillovers. Our policy experiments show that declining entry costs over time play a dominant role in evaluating the length of patent lives on total welfare. We find that consumer surplus is always declining for different lengths of patent lives due to monopoly pricing. On the other hand, patents increase producer surplus compared to free entry. Our results also show that different lengths of patent lives increase total surplus in the industry if the patent duration is sufficiently short or long. If patent life is short, total welfare decreases relative to the free entry case as the savings in entry costs is not large enough to compensate for the decline in consumer welfare and the loss in industry profit. If patent life is sufficiently long, total surplus increases as the entry cost savings that result from avoiding excessive entry in the industry and the gain in the monopolist's profit compensate the losses in consumer welfare. Free entry results in excessive entry since firms under free entry only care about their own profits and not about the overall industry profits. Hence, entry regulation through patent protection diminishes the private incentives of entry, and optimizes social incentives by avoiding excessive entry and firms paying a socially undesirable amount on entry costs.

On a final note, it is worth emphasizing that declining entry costs over time are significant also in many other markets, such as pharmaceuticals, the automotive and electronics industry, and many others. Therefore, it would be interesting to explore in future research if the main finding, i.e., sufficiently long patent lives positively impact total welfare through avoiding excessive entry rates, also applies to other markets.

References

Ackerberg, D. and G. Gowrisankaran, 2006, "Quantifying Equilibrium Network Externalities in the ACH Banking Industry," Rand Journal of Economics, 37, pp. 738-761.

Aghion, P., N. Bloom, R. Blundell, R. Griffith and P. Howitt, 2005, "Competition and Innovation: An Inverted-U Relationship," The Quarterly Journal of Economics, pp. 701-728.

Arrow, K., 1962, "Economic Welfare and the Allocation of Resources for Invention" in National Bureau of Economic Research Conference Volume," The Rate and Direction of Inventive Activity," (Princeton, NJ: Princeton University Press, 1962), pp. 609-625.

Bajari, P., L. Benkard and J. Levin, 2007, "Estimating Dynamic Models of Imperfect Competition," Econometrica, Vol. 75, No. 5, pp. 1331-1370.

Berry, Steven T. and Joel Waldfogel (1999): Free Entry and Social Inefficiency in Radio Broadcasting, The RAND Journal of Economics, Vol. 30, No. 3 (Autumn, 1999), pp. 397-420.

Berry, Steven T. and Joel Waldfogel (2001): Do mergers increase product variety: Evidence from Radio Broadcasting, Quarterly Journal of Economics, 116(3), pp. 969-1007.

Bresnahan, T.F. and P.C. Reiss, 1990, "Entry in Monopoly Markets," Review of Economic Studies, Vol. 57(4), pp. 531-553.

Bresnahan, T.F. and P.C. Reiss, 1991, "Entry and Competition in Concentrated Markets," Journal of Political Economy, Vol. 99(5), pp. 977-1009.

Carlton, D.W. and J.M. Perloff, 2005, "Modern Industrial Organization," Pearson, Addison Wesley.

Chaudhuri, S., P.K. Goldberg and P. Jia, 2006, "Estimating the Effects of Global Patent Protection in Pharmaceuticals: A Case Study ofQuinolones in India, " The American Economic Review, Vol. 96, No. 5 (Dec., 2006), pp. 1477-1514.

Cohen, W.M., R.C. Levin and D.C. Mowery, 1987, "Firm Size and R&D Intensity: A Re-Examination," Journal of Industrial Economics, 35, p. 543-565.

Dasgupta, P. and J.E. Stiglitz, 1988, "Learning-by-Doing, Market Structure and Industrial and Trade Policies," Oxford Economic Papers, Oxford University Press, Vol. 40(2), pp. 246-268.

Einav, L., 2010, "Not All Rivals Look Alike: Estimating an Equilibrium Model of The Release Date Timing Game," Economic Inquiry, 48(2), pp. 369-390.

Ericson, R. and A. Pakes, 1995, "Markov-Perfect Industry Dynamics: A Framework for Empirical Work," Review of Economic Studies, 62, pp. 53-82.

Fink, C., 2000, "How Stronger Patent Protection in India Might Affect the Behavior of Transnational Pharmaceutical Industries," The World Bank, Policy Research Working Paper 2352.

Fudenberg, D. and J. Tirole, 1983, "Learning by Doing and Market Performance," Bell Journal of Economics, Vol. 14, pp. 522-530.

Fudenberg, D. and J. Tirole, 1985, "Preemption and Rent Equilization in the Adoption of New Technology," Review of Economic Studies, Vol. 52(3), pp. 383-401.

Galbraith, J.K., 1956, "American Capitalism," (Boston: Houghon Mifflin).

Ghemawat, P. and M.A. Spence, 1985, "Learning Curve Spillovers and Market Performance," The Quarterly Journal of Economics, Vol. 100(5), pp. 839-852.

Gilbert, R.J. and D.M.G. Newberry, 1982, "Preemptive Patenting and the Persistence of Monopoly," American Economic Review, 72, pp. 514-526.

Gilbert, R.J. and C. Shapiro, 1990, "Optimal Patent Length and Breath," Rand Journal of Economics, 21, pp. 106-112.

Gowrisankaran, G. and J. Stavins, 2004, "Network Externalities and Technology Adoption: Lessons from Electronic Payments," Rand Journal of Economics, 35, pp. 260-276.

Griliches, Z., 1984, "R&D, Patents, and Productivity," Chicago: University of Chicago Press.

Hannan, T. and J. McDowell, 1984, "The Determinants of Technology Adoption: The Case of the Banking Firm," Rand Journal of Economics, 15, pp. 328-335.

Hashmi, A.R. and J. Van Biesebroeck, 2010, “Market Structure and Innovation: A Dynamic Analysis of the Global Automobile Industry,” National Bureau of Economic Research Working Paper No. w15959 .

Helpman, E., 1993, “Innovation, Imitation, and Intellectual Property Rights,” Econometrica, 61(6),pp. 1247-80.

Hsieh, Chang-Tai and Enrico Moretti (2003): ”Can Free Entry Be Inefficient? Fixed Commissions And Social Waste In The Real Estate Industry,” Journal of Political Economy, v111(5,Oct), 1076-1122.

Hughes, J.W., M.J. Moore and E.A. Snyder, 2002, “Napsterizing Pharmaceuticals: Access, Innovation, and Consumer Welfare,” National Bureau of Economic Research Working Paper 9229.

Irwin, D.A. and P.J. Klenow, 1994, “Learning-by-Doing Spillovers in the Semiconductor Industry,” Journal of Political Economy, Vol. 102(6), pp. 1200-1227.

Kamien, M.I. and N.L. Schwartz, 1982, “Market Structure and Innovation,” Cambridge: Cambridge University Press.

Klette, T.J. and S. Kortum, 2004, “Innovating Firms and Aggregate Innovation,” Journal of Political Economy, Vol. 112(5), pp. 986-1018.

Lee, T. and L.L. Wilde, 1980, “Market Structure and Innovation,” Quarterly Journal of Economics, 94, pp. 429-436.

Levin, R.C., W.M. Cohen and D.C. Mowery, 1985, “R&D, Appropriability, and Market Structure: New Evidence on some Schumpeterian Hypotheses,” American Economic Review, 75, p. 20-24.

Levin, R.C., A.K. Klevorick, R.R. Nelson and S.G. Winter, 1987, “Appropriating the Returns from Industrial Research and Development,” Brookings Papers on Economic Activity, 3, pp. 783-820.

Loury, G.C., 1979, “Market Structure and Innovation,” Quarterly Journal of Economics, 93, pp. 395-410.

Mankiw, N.G. and Whinston, M.D. (1986): "Free Entry and Social Inefficiency." RAND Journal of Economics, Vol. 17 (1986), pp. 48-58.

Mansfield, E., 1963, "Size of Firm, Market Structure and Innovation," Journal of Political Economy, 71, pp. 556-576.

Maskus, K.E. and D.E. Konan, 1994, "Trade-Related Intellectual Property Rights: Issues and Exploratory Results," In Analytical and Negotiating Issues in the Global Trading System, ed. Alan V. Deardorff and Robert M. Stern, 401-46. Ann Arbor, MI: University of Michigan Press.

Mazzoleni, R. and R.R. Nelson, 1998, "The Benefits and Costs of Strong Patent Protection:A Contributiont o the Current Debate" Research Policy, 27(3), pp. 273-84.

Moser, P., 2005, "How Do Patent Laws Influence Innovation? Evidence from Ninetheenth-Century World Fairs," The American Economic Review, Vol. 95 (4), pp. 1215-1236.

Nelson, R.R., S.G. Winter, 1978, "Forces Generating and Limiting Concentration Under Schumpeterian Competition," Bell Journal of Economics, Vol 9, pp. 524-548.

Nelson, R.R., S.G. Winter, 1982, "The Schumpeterian Tradeoff Revisited," American Economic Review, Vol 72, pp. 114-132.

Nordhaus, W.D., 1969, "Invention, Growth, and Welfare," (Cambridge: MA: MIT Press.)

Reinganum, J., 1989, "The Timing of Innovation: Research, Development and Diffusion," (in: Richard Schmalensee and Robert D. Willig eds. 'The Handbook of Industrial Organization', Amsterdam: North-Holland).

Riordan, M.H., 1992, "Regulation and Preemptive Technology Adoption," Rand Journal of Economics, Vol. 23, pp. 334-349.

Rose, N.L. and P.L. Joskow, 1990, "The Diffusion of New Technologies: Evidence from the Electric Utility Industry," Rand Journal of Economics, Vol. 21(3), pp. 354-373.

Saloner, G. and A. Shepard, 1995, "Adoption of Technologies with Network Effects: An Empirical Examination of Automated Teller Machines," Rand Journal of Economics, 26, pp.479-501.

Scherer, F., 1967, "Market Structure and the Employment of Scientists and Engineers," American Economic Review, LVII, pp. 524531.

Schumpeter, J.A., 1942, "Capitalism, Socialism and Democracy," (New York: Harper and Brothers Publishers).

Schmidt-Dengler, 2006, "The Timing of New Technology Adoption: The Case of MRI," manuscript.

Siebert, R.B. and C. Zulehner, 2011, "The impact of innovation and market demand on market structure," manuscript.

Solow, R.M., 1957, "Technical Change and the Aggregate Production Function," Review of Economics and Statistics, 39, pp. 312-319.

Spence, A.M., 1981, "The Learning Curve and Competition," Bell Journal of Economics, Vol. 12(1), pp. 49-70.

Sweeting, A., 2006, "Coordination, Differentiation and the Timing of Radio Commercials," Journal of Economics and Management Strategy, 15(4).

Sweeting, A., 2009, "The Strategic Timing of Radio Commercials: An Empirical Analysis Using Multiple Equilibria," Rand Journal of Economics, 40(4).

A Appendix: Tables

Table 1: Demand Estimation Results

Variable	OLS	IV	OLS	IV
Constant	16.706***	18.728***	16.903***	17.406***
	(19.88)	(21.77)	(20.77)	(24.04)
$\ln(P)$	-2.745***	-3.187***		
	(-25.77)	(-22.15)		
$\ln(P^{16K})$			-2.717***	-3.067***
			(-13.47)	(-18.24)
$\ln(P^{64K})$			-2.579***	-2.751***
			(-28.95)	(-29.31)
$\ln(P^{256K})$			-2.804***	-2.927***
			(-20.57)	(-23.62)
$\ln(P^{1M})$			-3.006***	-3.185***
			(-21.53)	(-22.71)
Dummy 64K	2.041***	2.194***	1.846***	1.661***
	(16.20)	(17.32)	(8.45)	(8.44)
Dummy 256K	3.768***	4.149***	3.951***	3.742***
	(20.15)	(22.52)	(14.65)	(15.44)
Dummy 1M	5.201***	5.942***	5.889***	5.803***
	(16.15)	(18.37)	(13.43)	(13.80)
$\ln(P^S)$	0.345*	0.336**	0.185	0.436**
	(1.74)	(2.02)	(1.01)	(2.43)
$\ln(GGDP)$	10.095	7.984	11.546	8.274
	(0.79)	(0.60)	(0.95)	(0.71)
$\ln(P^{Dram})$	1.132***	1.223***	1.267***	1.217***
	(10.48)	(10.86)	(10.31)	(9.92)
Time	0.013*	-0.00003	0.023**	0.014
	(1.74)	(-0.00)	(2.54)	(1.53)
Number of observations	314	310	314	310
Adjusted R-squared	0.858	0.817	0.868	0.844

The dependent variable is the logarithm of industry output ($\ln(q_t^k)$. As instruments, we use the price of silicon, cumulative total output in the current generation, and cumulative total output from the previous generation. The t-statistics are shown in parentheses below the parameter estimates, and *** (**, *) denotes a 99% (95%, 90%) level of significance.

Table 2: Marginal Cost Estimation Results

Variable	(1)	(2)
$\ln(P^{sil})$ (ρ_1)	1.008^{***}	0.729^{***}
	(67.83)	(36.28)
Intragenerational own learning (ρ_2)	-0.344^{***}	-0.260^{***}
	(-61.00)	(-51.76)
Spillover learning (ρ_3)	0.317^{***}	0.308^{***}
	(6.14)	(5.19)
Intergenerational own learning (ρ_4)	0.010^{***}	0.0009^{**}
	(4.63)	(2.46)
Dummy 64K	0.239^{***}	0.293^{***}
	(21.02)	(25.96)
Dummy 256K	0.639^{***}	0.731^{***}
	(52.10)	(57.62)
Dummy 1M	1.008^{***}	1.204^{***}
	(67.23)	(69.54)
Time		-0.013^{***}
		(-20.46)
Firm fixed effect	Yes	Yes
Number of observations	5,835	5,835
Adjusted R-squared	0.907	0.913

Table 2 shows nonlinear least square results of the marginal cost function, as shown in equation (4). The dependent variable is the logarithm of the marginal costs. The t-statistics are shown in parentheses below the parameter estimates, and *** (**, *) denotes a 99% (95%, 90%) level of significance.

Table 3: Output Policy Estimation Results

Variable	$64K$
Constant	-5.971
	(-1.04)
GDP	0.00000038***
	(3.25)
$\ln(P^{DRAM})$	0.200**
	(1.98)
$\ln(P^{sil})$	-0.391
	(-1.48)
$\ln(n^{16K})$	0.430**
	(2.12)
$\ln(n^{64K})$	-0.296
	(-1.30)
$\ln(x_i^{16K})$	-0.044
	(-1.17)
$\ln(x_i^{64K})$	0.492***
	(11.34)
$\ln(x_{-i}^{16K})$	0.643
	(1.14)
$\ln(x_{-i}^{64K})$	0.013
	(0.07)
$\hat{\nu}_i$	-0.319**
	(-2.57)
Time	-0.094***
	(-6.64)
Firm fixed effect	Yes
Number of observations	1,693
Adjusted R-squared	0.782

The dependent variable is the logarithms of firm level output in the $64K$ generation. The t-statistics are shown in parentheses, and *** (**, *) denotes a 99% (95%, 90%) level of significance.

Table 4: Entry and Exit Policy Estimation Results

Variable	Entry $64K$	Exit $64K$
Constant	-4.680***	-2.658***
	(-5.58)	(-3.62)
$\ln(n^{64K})$	0.170***	1.104*
	(2.71)	(1.91)
$\ln(x_i^{16K})$	0.046**	
	(2.31)	
$\ln(x_i^{64K})$		-0.139***
		(-4.43)
$\ln(x_{-i}^{64K})$	0.051	-0.395*
	(1.63)	(-1.71)
Time	0.054***	0.051***
	(7.10)	(2.70)
Firm fixed effect	Yes	No
Number of observations	3,195	1,701
Pseudo R-squared	0.592	0.058

Table 4 shows the results for the entry and exit policies, as shown in equations (9) and (10). The entry and exit policies are estimated using probit models. The dependent variable in the entry model takes on a value of one when a firm chose to enter the $64K$ generation and zeros before entry occurred. In the exit model, the dependent variable takes on a value of one if a firm exits the generation and zero between the firm's entry and exit decisions. Firm fixed effects are not included in the exit models since not all firms have exited in the data. The z-statistics are shown in parentheses, and *** (**, *) denotes a 99% (95%, 90%) level of significance.

B Appendix: Figures

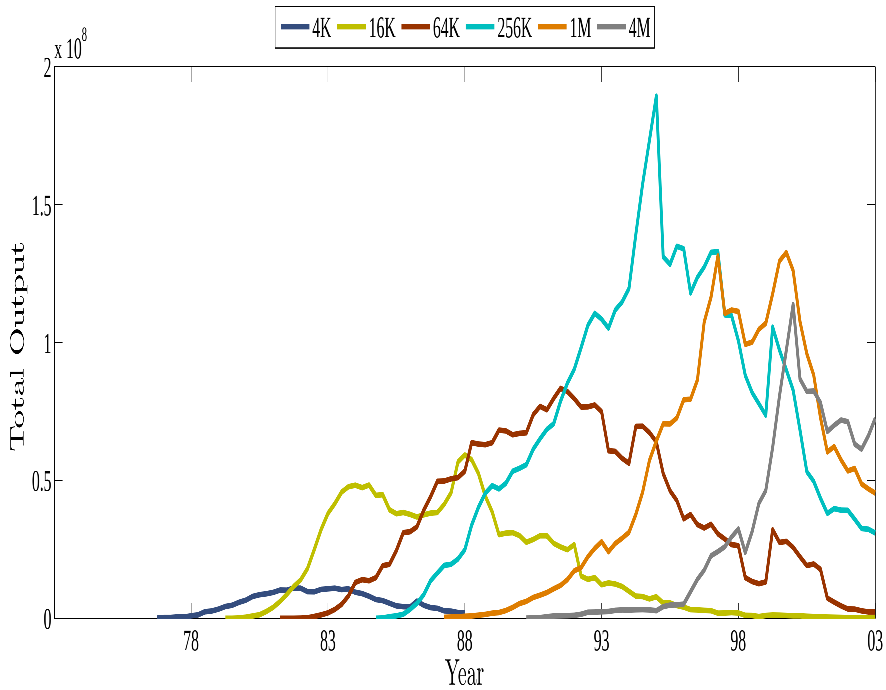

Figure 1: Industry units shipped, 1974-2003
Source: Gartner Inc.

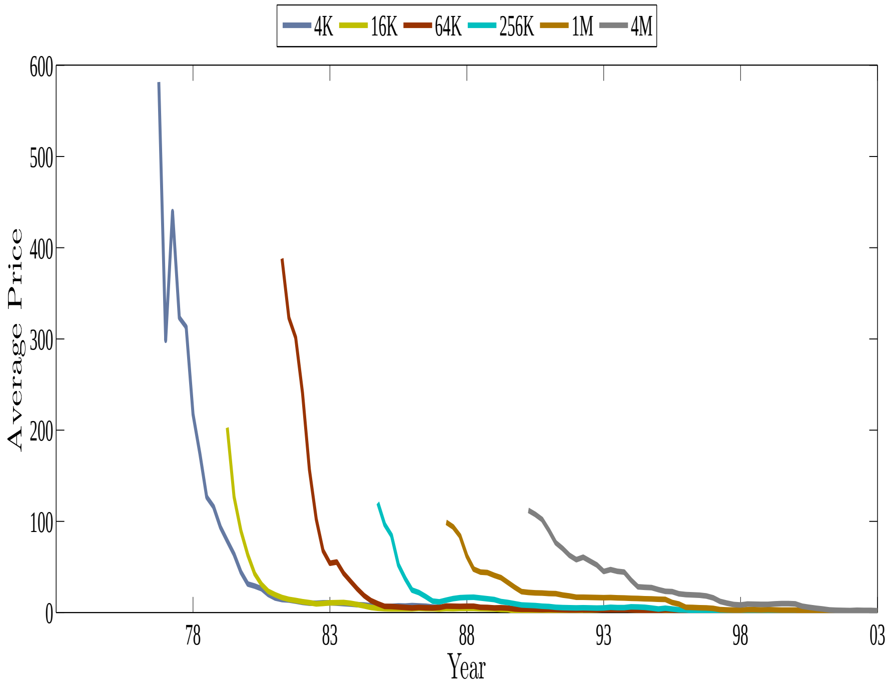

Figure 2: Average SRAM prices, 1974-2003
Source: Gartner Inc.

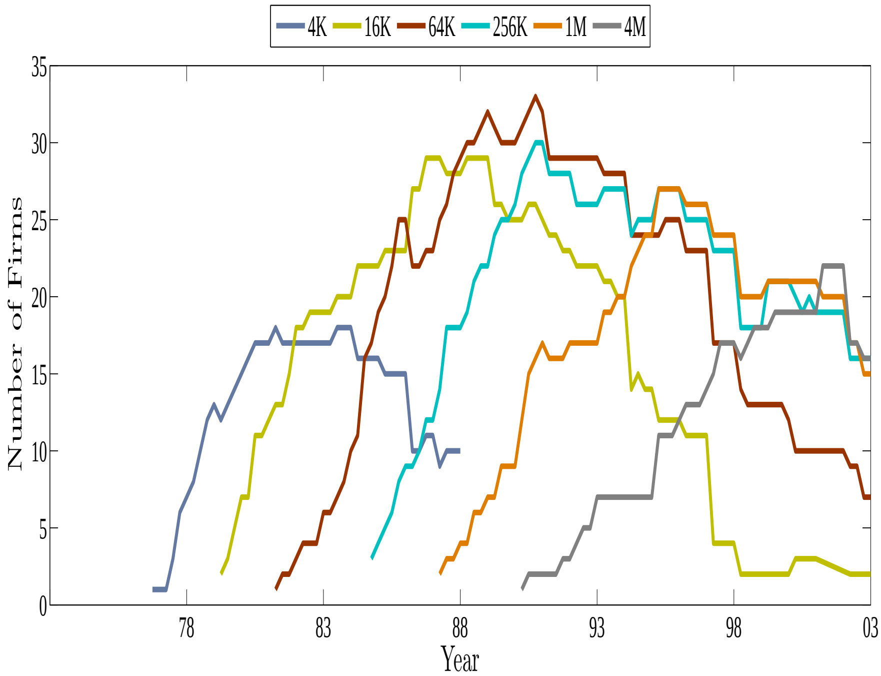

Figure 3: Number of Firms
Source: Gartner Inc.

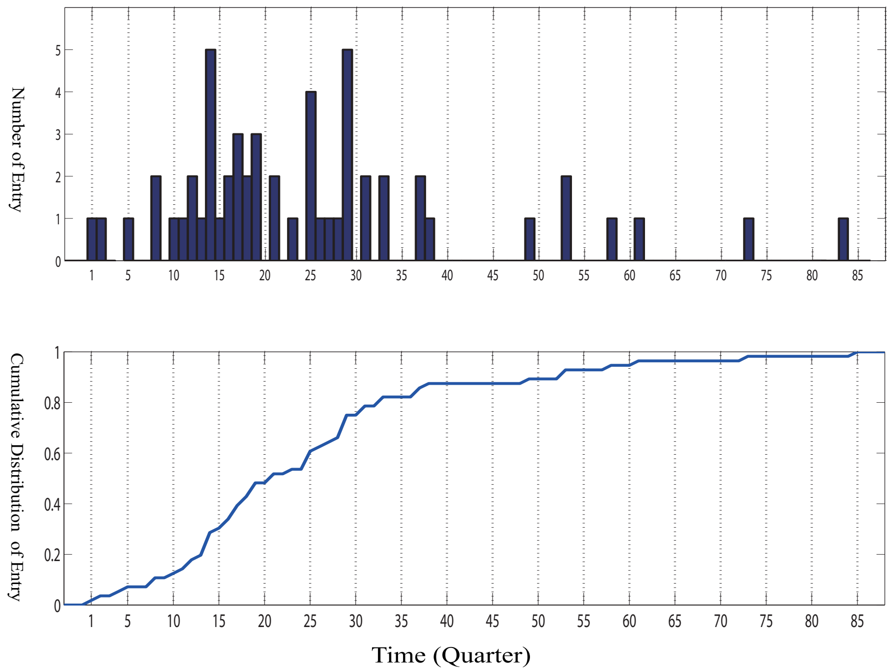

Figure 4: Distribution of Entry 64K
Source: Gartner Inc. The first entry occurs in the first quarter of 1982.

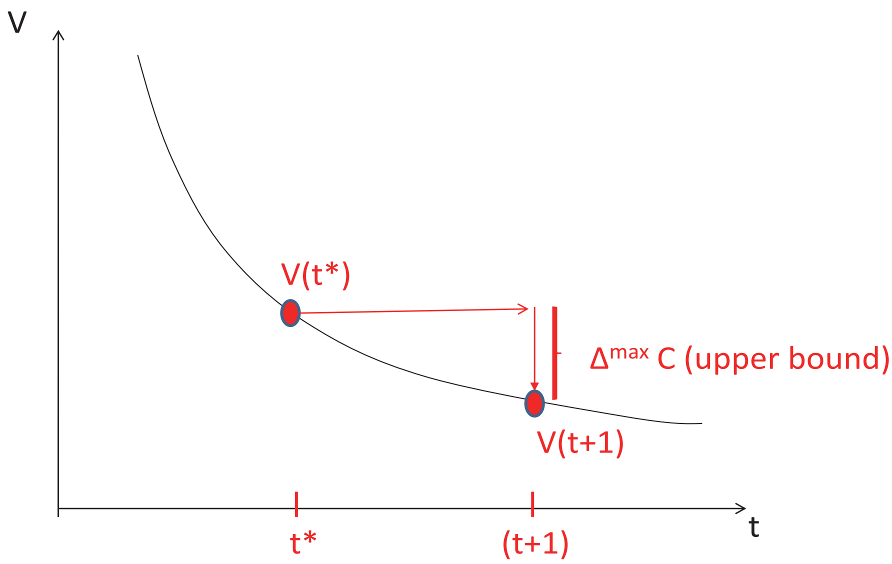

Figure 5: Identification of Entry Cost (Upper Bound)

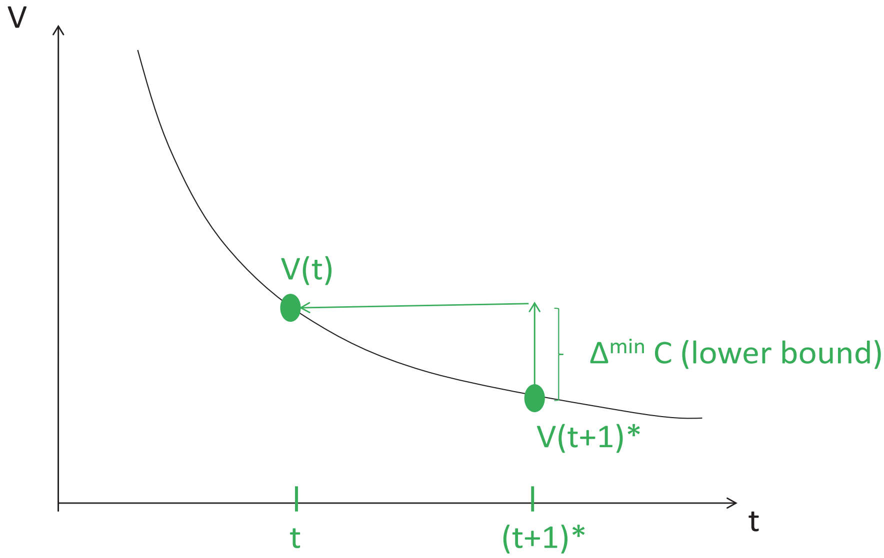

Figure 6: Identification of Entry Cost (Lower Bound)

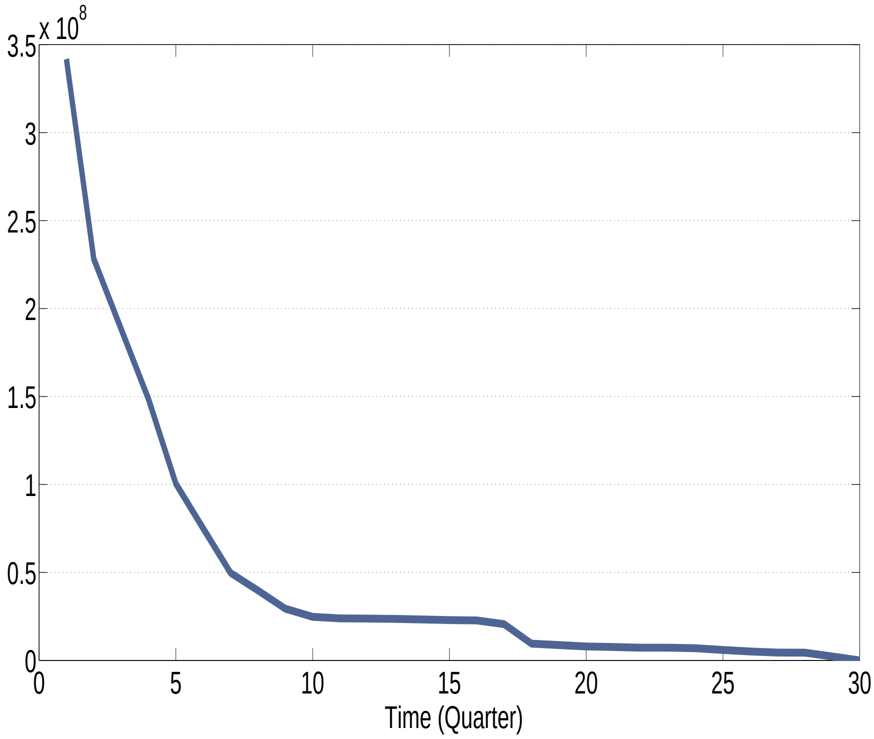

Figure 7: Slope of Entry Costs for the 64K generation

All values are the discounted values at the timing of the first entry in US dollars.

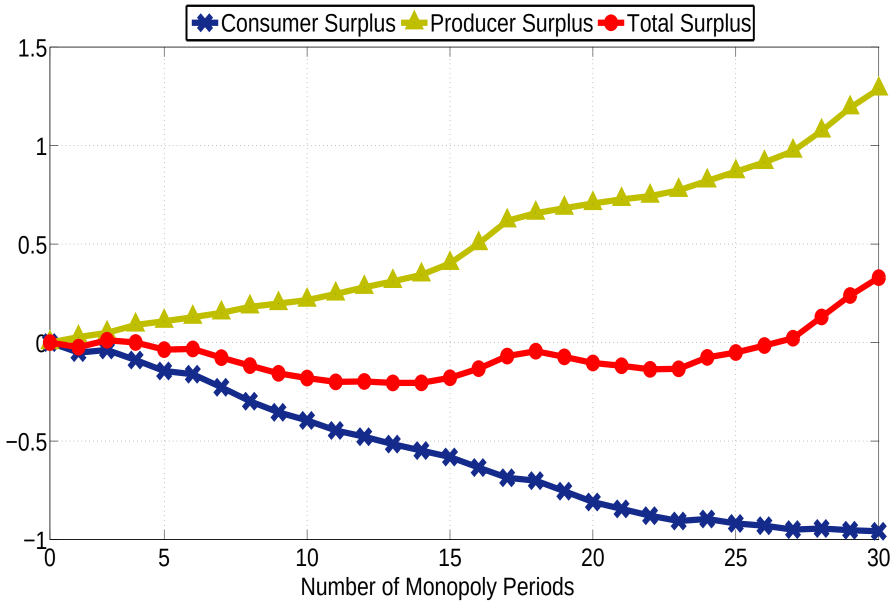

Figure 8: Change in Welfare

All values are the discounted values at the timing of the first entry in US dollars.

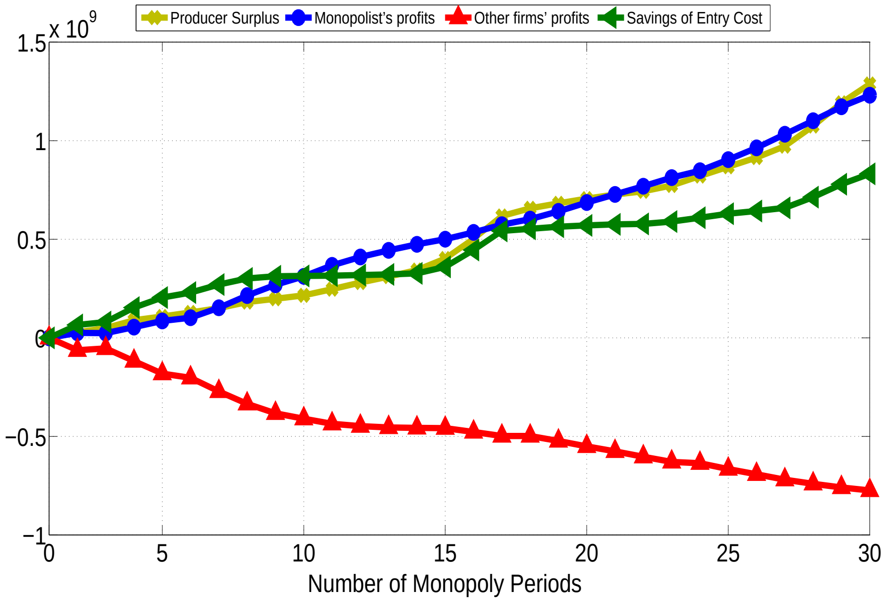

Figure 9: Change in Producer Surplus

All values are the discounted values at the timing of the first entry in US dollars.