UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

Graduate School - Master's Thesis in Economics

Price Dispersion and the Value of Information in Online Retail Markets

Hendrik Jahns
920110-3935

Abstract

Compared to conventional markets, online markets offer many informational advantages to consumers. It is seemingly easy to compare prices online and still, economists have found temporal price dispersions in markets for homogeneous goods. Existing research uses data from the turn of the millennial, yet it's findings may not apply to today's markets, as online markets have developed rapidly and are becoming increasingly important to consumers and firms.

In this thesis, I measure the price dispersion of homogeneous goods in German online markets. I find significant levels of price dispersion, comparable to those in previous research. I furthermore examine how price dispersion relates to market characteristics: My analysis indicates that price dispersion shows a negative relationship to average price levels and no significant relationship to the number of sellers in the market. Lastly, I simulate consumer search behavior to obtain estimates of the value of information to consumers. I find that consumers can be broadly categorized into two groups, namely of low and high search intensity. For the analyses, I collected price data of 207 homogeneous electronics products from a German price comparison website.

Supervisor: *Anna Bindler*
Keywords: *price dispersion, value of information, online markets*

2017 - 05 - 22

Acknowledgements

I would like to thank my supervisor, Anna Bindler, for her continuous support, dedication and helpful comments she has given me during my thesis work. I would also like to express my gratitude to my parents who have always supported and encouraged me in my studies throughout the years. Special thanks is also due to guenstiger.de GmbH, who have allowed me to collect data from their website.

Contents

List of Figures

List of Tables

1 Introduction

With the introduction of the Internet and the emergence of e-commerce, economists predicted that competition would be strengthened and price differences reduced (Levin 2011). By offering consumers comprehensive and instantaneous possibilities to compare prices between sellers, the Internet would significantly reduce search costs. Popular press even declared a new era of frictionless commerce with perfect information and competition (Ellison & Ellison 2004). Yet, economists have found persistent price dispersions in many examined markets, even among homogeneous goods (Levin 2011).

The phenomenon of price dispersion can be best described as the variation of prices of an underlying good with the same characteristics across sellers (Pan et al. 2004). For instance, Brynjolfsson & Smith (2000) find that the relative range of prices for homogeneous books and CDs lies at up to 47% percent in their sample. While there has been evidence for lower price levels in online retail compared to traditional brick-and-mortar stores, the relative price dispersion differed only little (Pan et al. 2004). In economic theory, price dispersion can be explained by search models: The intuition behind this strand of models is that consumers have a cost of searching and do not obtain all price quotes for comparison. Some sellers can then charge higher prices to relatively uninformed consumers, which results in price dispersion Baye et al. (2006).

Much of the existing research uses data from the dawn of e-commerce era; the impact of e-commerce on retail industries has, however, continued to develop rapidly and the effects measured previously may not be representative of today's market. In the United States for instance, the share of e-commerce to total retail sales volume has increased from approximately 0.9% in 2000 to 8.1% in 2016 (United States Census Bureau 2016). In Sweden (Germany), the percentage of individuals purchasing goods online in the last 12 months increased from 55% (49%) in 2006 to 76% (74%) in 2016 (Eurostat 2016). This gives rise to an important question: How large is price dispersion in online retail today and what factors relate to it?

The authors of several papers have noted that the measured price dispersion may be a result of the immaturity of online markets (Pan et al. 2004), and reexamining these markets may yield different results. In this study I collect new data to provide an insight into the current degree of price dispersion and the value of information to consumers. This contributes to existing research by reviewing its findings and illustrating the mechanisms in today's online retail markets. My descriptive analysis includes measuring price dispersions through the coefficient of variation as well as providing detailed information on other market characteristics. In the examined markets, I find significant levels of price dispersion comparable to those in previous research.

Previous literature has examined possible relating factors to these dispersions, such as the role of price levels and the number of sellers in a market (Ratchford et al. (2003), Baye et al. (2004*a*), Pan et al. (2004)). In today's highly dynamic online market however, these findings may not apply anymore as retailers may have adopted new strategies and consumers may have matured in using the Internet's informational resources. Based on a sequential search model by Carlson & McAfee (1983), I develop hypotheses and test the relationship between the above mentioned factors in several regression models. My analysis reveals a negative semi-elastic relationship of price dispersion and price levels, yet no significant relationship of price dispersion and the number of sellers in a market.

Moreover, previous research has aimed to quantify the value of information to consumers in online markets (Baye et al. (2003), Pan et al. (2004)). Using a simulation program, I extend this analysis by differentiating between different search intensities of consumers. For my simulation, I use a weighted sampling approach and utilize several assumptions from search theoretical models to obtain estimates of the value of information to consumers. To better model consumer search behavior, I use website traffic estimates to assign weights to the sellers in the examined markets. The simulation results suggest that consumers can be broadly categorized into two groups of high intensity and low intensity search.

The majority of papers focuses on American markets and while the United States surely has had the most decisive and innovative role in the digitization of markets, findings may not necessarily be applicable to other countries. I collected my data from German markets, which thus also adds to the scarcity of research outside of the United States. Analogue to previous research, the data was collected from an online price comparison website. The data is comprised of 207 consumer electronics products and was collected on a daily basis over the course of 28 days using an automated web scraping tool. The product sample is made up of televisions, hard drives and printers.

Overall, price dispersion measures are an important indicator of competition and information efficiency within a market (Pan et al. 2004). To economists it is of interest as to why online markets have shown persistent inefficiencies even though the Internet offers many informational advantages to consumers. This line of research is also relevant to consumers, as it shows to what degree intensifying search is effective, and to sellers, as it gives insights into the competitiveness of prices in online markets. Next to providing extensive information on the structure of online markets, my findings also provides a better understanding of the search behavior and informational value of search for consumers.

The remainder of this thesis is structured as follows: Section 2 will give an overview of existing literature; in Section 3, I will discuss the theoretical background of this research field and formulate hypotheses to summarize my research intentions; Section 4 will provide a description of the data I collect; in Section 5, I will specify my intended empirical approach for the regression analysis, and present and discuss the results; in Section 6, I will describe my simulation methodology, and discuss the results.

2 Literature Review

Generally, research has found that the introduction of the Internet has lowered prices. Although theory suggests that there has been a significant decrease in consumer search costs, price dispersions however remained persistent (Levin 2011). A prominent example of early research in this field is the paper of Brynjolfsson & Smith (2000). Brynjolfsson & Smith (2000) examine the price difference of homogeneous books and CDs between offline and online retailers, as well as only among online retailers. From their sample, they find that books (CDs) are 15.5% (16.1%) cheaper on the Internet. When investigating the level of price dispersion, they find that the average price range for books (CDs) is 25% (33%) in online retail. They conclude that price dispersion in conventional retail is only slightly larger at most.

Pan et al. (2003) examine a repeated cross sectional dataset from November 2001 to February 2003. Their data is collected from a web comparison website and consists of overall 2176 products. They find the price dispersion measured by the coefficient of variation to lie between 9.78% and 11.72%.[1] When examining the relationships between price dispersion and the average price level and the number of sellers in the respective market, they find a significant negative relationship of price dispersions and price levels, yet mixed results for the number of sellers. They further expand their analysis by estimating the value of information to consumers in these markets. The relative price difference of a consumer only searching for one price versus a consumer who has all information lies between 13.1% and 14.5% in their sample.[2]

Ratchford et al. (2003) collect two datasets in November 2000 and November 2001 from a price comparison website with overall 1407 products. For the different product categories, they find levels of price dispersion in the range of 6.51% to 16.63% measured by the coefficient of variation. Regarding the relationship of price dispersion and price levels, they find a significant negative relationship. Furthermore, they find a significant quadratic relationship for price dispersion and the number of sellers in the market, described by a downward facing parabola.

Baye et al. (2003) focus on the value of information in online markets. Similarly to Pan et al. (2003) they measure the value of information as the difference of the average market price to the lowest price in the market. Using a comprehensive panel dataset of 4 million price quotes ranging from August 2000 to March 2001, they find the average relative value of information to be 15.89%. The price dispersion measured by the relative price range is 35.52% for their average market. Baye et al. (2004*a*) examine the relationship of market size, i.e. the number of sellers, and price dispersion. They use an extensive dataset with

[1] The coefficient of variation is the standard deviation divided by the average of a variable.

[2] The uniformed consumer which only samples one price pays an expected price equal to the average market price, while the fully informed consumer pays the lowest price in the market.

over 200,000 market observations collected from an American online price comparison site between August 2000 to March 2001. They report an average price dispersion measured by the coefficient of variation of 9.10% across all observations. When controlling for product popularity and other market characteristics they find that price dispersion varies strongly with the number of sellers in the market. Their data suggests that price dispersion measured by the the percentage gap between the two lowest prices is larger for smaller markets and smaller for larger markets. A summary of the discussed research and further papers is shown in Table 1.

Table 1: Selection of Research Papers Examining Price Dispersion in Online Retail

Article	Country	Observations	Time	Measure	Price Dispersion
Brynjolfsson & Smith (2000)	USA	20 books, 20 CDs	02/1998-05/1999	Price range by avg. price	Books: 33%, CDs 25%
Clay et al. (2001)	USA	399 books	08/1999 - 01/2000	Coefficient of Variation	12.9%-27.7%
Clay et al. (2002)	USA	107 books	04/1999	Price range by avg. price	27%-73%
Pan et al. (2003)	USA	2176 mixed products (rep. cross section)	11/2000, 11/2001, 03/2003	Coefficient of Variation	9.78% - 11.72%
Ratchford et al. (2003)	USA	1407 mixed products (rep. cross section)	11/2000, 11/2001	Coefficient of Variation	6.51% - 16.63%
Baye et al. (2004*a*)	USA	214,337 observations (rep. cross section)	08/2000-03/2001	Coefficient of Variation	9.10%
Baye et al. (2004*b*)	USA	36 Products	11/1999-05/2001	Coefficient of Variation	12.5%

The majority of research refers to more than fifteen year old data from the turn of the millennial and possibly does not reflect the current state of the markets. Yet this topic remains highly relevant, especially in light of the ever growing e-commerce markets and the impact the digitization of markets has had on the economy. As mentioned in the introduction, the share of individuals shopping online and the share of e-commerce to overall retail has dramatically increased over the last years. I expand on previous literature by using a new and unique data set to examine the current degree of price dispersion for homogeneous goods. Similar to previous research, I further explore possible related factors to price dispersion, namely the number of the sellers and average price level in the market. Lastly, I expand on the analysis of the value of information by simulating consumer search behavior to provide insights into the optimization process of search consumers face when purchasing goods online.

3 Sequential Search Model and Hypotheses

When considering the classic Bertrand model, price dispersion should not exist in markets. This is owing to its core assumptions of perfectly-informed consumers and product homogeneity. This model however does not reflect the empirical findings in real markets (Brynjolfsson & Smith 2000). Alternative theoretical models have been put forward to help explain the existence of price dispersion.

The baseline approach behind these models is to relax the assumption of perfectly-informed consumers. Baye et al. (2006) categorize these models into two broad groups, namely search theoretic models and clearing house models. In search theoretic models, consumers do not have perfect information as they have to engage in costly search to obtain price quotes. These models can be further divided into two subcategories: Sequential search models and fixed sample search models, also called non-sequential search models. The key difference between these models is how consumers optimize their search behavior and how the number of searches is determined. In fixed sample search models, each individual search comes with a marginal cost to consumers. In this case, consumers determine a fixed number of searches according to the marginal cost and benefit of an additional search. In sequential search models on the other hand, the number of searches is a random variable. Consumers form expectations about their number of searches given the price distribution and their individual reservation price, and then search until they are satisfied with a price quote. Both models yield price dispersion due to the fact that firms can exploit the search costs of consumers and charge different prices.

Clearing house models assume that some consumers have access to all prices in the market, e.g. through a price-comparison website. These consumers are able to always observe the lowest price in the market, while other consumers can only randomly obtain price quotes. The outcome is price dispersion, as firms are able to charge higher prices to the group of uninformed consumers (Baye et al. 2006).

To derive my hypotheses, I build on a sequential search model by Carlson & McAfee (1983). In this model price dispersion arises due to the fact that firms differ in marginal costs and consumers exhibit heterogeneous search costs. This results in firms being able to set different prices and keep their profitable market share as not all consumers can find the lowest price. The advantages of this model include the assumption of heterogeneous marginal costs for firms. This seems to be realistic for online consumer goods markets, as there may be differences in economies of scale among online retailers. Furthermore, the model assumes there to be a continuum of search costs. This allows there to be a large number of different consumers, as opposed to clearing house models, where there usually only exist two groups of informed and uninformed consumers.

The supply side of the market consists of n firms with heterogeneous marginal costs that sell a homogeneous good to consumers. The firm's prices are ranked with sub-index $k = 1, ..., n$, such that $k = 1$ is the lowest and $k = n$ is the highest price. Consumers differ in their search costs and search sequentially. Each consumer's search costs (c) is drawn from a continuous distribution $G(c)$ with $c \in [0, \infty]$ (Equation (1)). Assume $G(c)$ to be a uniform distribution with T being the range of search costs, T/s the total amount of buyers and $1/s$ the density in every point.

$$\begin{aligned} G(c) &= c/s, \quad 0 \leq c \leq T \\ G(c) &= T/s, \quad T < c \\ G'(c) &= g(c) = 1/s \end{aligned} \tag{1}$$

Consumers are assumed to know the set of prices (p), yet they are only able to sample prices randomly. Their perceived price distribution is then given by Equation (2).

$$\begin{aligned} f(p) &= 1/n, \quad p = p_1, ..., p_n \\ &= 0 \qquad otherwise \end{aligned} \tag{2}$$

The consumer's sequential search can be described as follows: Consumers randomly draw prices from the distribution of prices. As soon as they find a price lower or equal to their reservation price, they stop searching and buy the good. Consumers optimize their expected gain from searching according to their perceived price distribution and their individual cost of search. Let x_k denote the consumers expected benefit of searching for a lower price than p_k (Equation (3)).

$$\begin{aligned} x_k &= \sum_{i=1}^{k-1} (p_k - p_i) f(p_i) \quad , \quad k = 2, ..., n \\ &= \Big(p_k - \sum_{i=1}^{k-1} \frac{p_i}{k-1}\Big) \frac{k-1}{n} \quad , \quad k = 2, ..., n \end{aligned} \tag{3}$$

When searching for a price lower than p_k, the expected benefit for the consumer will be the difference of p_k to the average of all prices lower than p_k, multiplied by the probability of finding a price lower than p_k. The index i refers to the $k - 1$ prices lower than p_k. This implies that consumers will search as long as the expected benefit of searching is higher than their search cost. More formally, consumers will only buy the good for a price below p_{k+1} iff $x_k \leq c < x_{k+1}$, meaning that their search cost has to be greater or equal to the expected benefit of searching for a price lower than p_k and be strictly smaller than the search for a price lower than p_{k+1}. With this condition, all consumers can be placed into groups with different effective reservation prices. Consumers with higher search costs will on average conduct less search, as they terminate their search earlier. The term in Equation (3) becomes equal to zero for the case that $k = 1$, as there is no lower price in the distribution. Next, the demand function for firm j with expected quantity q_j is derived (Equation (4)).

$$q_j = \sum_{k=j}^{n} \frac{1}{k}[G(x_{k+1}) - G(x_k)] \tag{4}$$

The highest priced firm with p_n will equally share all those consumers that buy at the first sampled store with all other firms.[3] The second highest priced firm additionally obtains a $1/(n-1)$th share of all those consumers who would buy at price p_{n-1} and so on. Using the assumption of uniformly distributed search costs, the demand function can be rewritten accordingly. This step requires a substantial amount of algebra and is described in detail in Carlson & McAfee (1982). In short, Carlson & McAfee (1983) substitute in the consumers benefits of search (Equation (3)) and the search cost distribution (Equation (1)). The final demand function for firm j is then given by Equation (5)[4].

$$\begin{aligned} q_j &= \frac{1}{sn}\Big(T - \frac{n-1}{n}p_j + \sum_{i \neq j} \frac{p_i}{n}\Big) \\ &= \frac{1}{sn}[T - (p_j - \bar{p})], \quad with \quad \bar{p} = \sum_{j=1}^{n} \frac{p_j}{n} \end{aligned} \tag{5}$$

From this equation we can make a few simple observations. Firm j's demand increases with an increasing average price in the market ($\bar{p}$), an increasing range of search costs (T), and an increase in the search cost density ($1/s$). The demand decreases with an increase in price j (p_j), and an increasing number of firms in the market (n). A company's demand thus primarily depends on the difference of their price to the average price in the market.

Given the consumers behavior, the price setting is determined by the firms profit maximization. The firm's profit function and first order condition are given by Equation (6), with firms maximizing their profit according to an optimal price.

$$\begin{aligned} \pi_j &= p_j q_j - c_j(q_j) \\ \frac{\partial \pi_j}{\partial p_j} &= q_j + (p_j - c_j'(q_j))\frac{\partial q_j}{\partial p_j} = 0 \end{aligned} \tag{6}$$

Carlson & McAfee (1983) use a cost function with increasing marginal costs in their model

[3] Note that similarly to k, the index j is also ranked, with $j = 1$ being the lowest priced firm, and $j = n$ the highest priced firm.

[4] Remember that the index i refers to the $k - 1$ prices lower than p_k.

(Equation (7)).

$$c_j(q_j) = \alpha_j q_j + \beta {q_j}^2 \tag{7}$$

In e-commerce markets it may however be more sensible to assume constant marginal costs, as e-commerce firms possibly do not see large cost surges when increasing scale. Next to setting up and maintaining a website, and paying for storage capacities for the goods, an increase in scale within a certain capacity would typically only lead to incremental costs in form of constant variable costs, e.g. material costs and shipping. I will discuss the implications of a constant marginal cost function later on in this section. To obtain firm j's price, we need to set up the explicit profit function and profit maximization function. Substituting in the first line from Equation (5) we obtain the following:

$$\begin{aligned} \pi_j &= (p_j - \alpha_j - \beta q_j) q_j \\ \frac{\partial \pi_j}{\partial p_j} &= q_j + p_j - \alpha_j - 2\beta q_j) \frac{\partial q_j}{\partial p_j} = 0 \\ &= \frac{1}{sn}\Big(T - \frac{n-1}{n} p_j + \sum_{i \neq j} \frac{p_i}{n}\Big) + (p_j - \alpha_j - 2\beta q_j)\Big(- \frac{n-1}{sn^2}\Big) = 0 \end{aligned} \tag{8}$$

After further rearrangements we obtain Equation (9).

$$\begin{aligned} p_j = \alpha_j + \frac{(1+\gamma)n}{n-1}\Big(T + \frac{n-1}{2n-1+\gamma n}(\bar{\alpha} - \alpha_j)\Big) \\ \mathit{with} \quad \gamma \equiv 2\beta \frac{(n-1)}{sn^2} \end{aligned} \tag{9}$$

It states that the price levels of firms vary in equilibrium, primarily depending on the firm's individual costs (α_j). If we now assume constant marginal costs, i.e. $\beta = 0$, this then leads to $\gamma = 0$. This however does not change the relationship of varying prices, as the firm's price depends on it's underlying cost parameter (α_j). From this equation I derive my first hypothesis:

Hypothesis 1: With heterogeneous search costs and firm costs, there exists price dispersion in equilibrium.

If we now assume that each homogeneous good constitutes a separate market, we can make predictions about how markets differ depending on the exogenous parameters. First, I will discuss the effect of a change in the number of sellers in the market on the price dispersion, followed by an analysis of the effect of a change in the price level of the good.

In their model, Carlson & McAfee (1983) demonstrate how the level of price dispersion depends on the number of firms in the market. The price dispersion is given by the variance of p. When summing up all prices in Equation (9) and dividing by n, the last term cancels out and we obtain the average market price (Equation (10)).

$$\bar{p} = \bar{\alpha} + \frac{(1+\gamma)nT}{n-1} \tag{10}$$

We can now derive the variance of p by averaging the squared difference of $\bar{p}$ from p_j for all j (Equation (11)).

$$\sigma_p{}^2 = \frac{1}{n}\sum_{j=1}^{n}(p_j - \bar{p})^2 \tag{11}$$

This gives us the final equation describing the price dispersion in the market (Equation (12)).

$$\begin{gathered} \sigma_p{}^2 = \Big(\frac{1}{2 + 1/(n-1) + (2\beta/sn)}\Big)^2 \sigma_\alpha{}^2 \\ \frac{\partial \sigma_p{}^2}{\partial n} > 0 \end{gathered} \tag{12}$$

When differentiating $\sigma_p{}^2$ by n, we can see that there is a positive relationship, suggesting that an increase in the number of sellers in the market leads to a higher dispersion.[5] This is only true under the assumption that the variance in the cost parameter α ($\sigma_\alpha{}^2$) remains constant. Note that this relationship also remains when assuming constant marginal costs, i.e. $\beta = 0$. A possible explanation to this is that with an increasing number of sellers, the market becomes more obfuscated to consumers. With unchanged search costs but more sellers, it would be easier for sellers to charge differentiated prices, as the probability of being sampled by a less informed consumer is higher. From this, I derive my second hypothesis:

Hypothesis 2: Price dispersion shows a positive relationship to the number of sellers.

There is reason to believe that price dispersion may vary with the price level of the underlying good under the assumption that each homogeneous good represents a separate market. A consumer's search may depend on the expected expenditure made on the good. For instance, hard drives tend to be more inexpensive than TVs, and thus make up a smaller share of expenditures. Assuming that search costs do not differ for these products, consumers

[5] From an empirical standpoint, the number of sellers may however not be exogenous to price dispersion. This relationship is discussed later on in Section 5.

should expect to gain more from search with higher overall price levels. This hypothesis was first introduced by Stigler (1961), yet is missing a formal derivation in his paper.

In the model of Carlson & McAfee (1983), it is very difficult to demonstrate the effect of higher price levels on price dispersion. However, I argue that the same effect can be shown when decreasing overall consumer search costs ceteris paribus. The expensiveness of a good in this context can be described as the ratio of price level to search cost. When assuming that search costs are constant for all goods, expensive goods will exhibit a higher ratio than less expensive goods. To simplify the analysis one could thus keep the price level unchanged and rather decrease search costs of expensive goods to obtain this ratio. Changing consumer's search costs in a market would then effectively show the differences of price dispersion by price levels between goods.

In the extension notes to their research paper (Carlson & McAfee (1982)), Carlson & McAfee (1983) show precisely this. Assume there to be a right shift in the distribution of search costs, i.e. higher overall search costs, such that:

$$\begin{aligned} G(c) &= 0, & 0 \leq c < w \\ G(c) &= (c-w)/s, & w \leq c \leq T + w \\ G(c) &= T/s, & T + w < c \end{aligned} \tag{13}$$

Carlson & McAfee (1982) demonstrate this for the specific case that only one firm charges the lowest price, as the demand specification varies with the number of firms charging the lowest price at the same time. Using the same steps as above, the equilibrium prices for firm j and firm 1 then become:

$$\begin{aligned} p_j &= \alpha_j + \frac{(1+\gamma)n}{n-1}\Big(T + \frac{n-1}{2n-1+\gamma n}(\bar{\alpha} - \alpha_j + w)\Big) \quad for \quad j = 2, ..., n \\ p_1 &= \alpha_1 + \frac{(1+\gamma)n}{n-1}\Big(T + \frac{n-1}{2n-1+\gamma n}(\bar{\alpha} - \alpha_1 - (n-1)w)\Big) \end{aligned} \tag{14}$$

We can show that the average market price remains the same as in Equation (10), as the terms with w cancel out when calculating the average price over all n firms. The lowest price (p_1) is lower and all other prices (p_j) rise when increasing search costs due to the included w term.[6] This increase in price range also leads to an increase in the variance of prices, while preserving the equilibrium mean. In reverse, this implies that lower search costs relative to the price level leads to a decrease in price dispersion. Extending this to my above reasoning, markets with relatively more expensive goods should show a lower price dispersion.

[6] Again, this relationship also remains when assuming constant marginal costs, i.e. $\beta = 0$.

Hypothesis 3: Price dispersion shows a negative relationship to price levels.

This theoretical model however also has some limitations. First, the model's predictions are based on the assumption of the search costs being uniformly distributed, which may not reflect the true composition of search costs. Next, consumers are assumed to know the set of prices, but can only randomly sample them. In reality however, consumers may have preferences in sampling certain retailers due to for instance brand trust or experience. Moreover, while consumers may form expectations about the set of prices in the market, it may be unrealistic to assume that consumers know the set of available prices in a given market.

4 Data Description

I collected my data from the German price comparison website *Guenstiger.de.* Price comparison websites, often also referred to as *shopbot* in literature, list product prices from different sellers as a service to consumers to be able to compare prices. Most shopbots restrict the systematic collection of the provided data from there website in their terms of use. It was therefore necessary to explicitly ask for permission before collecting my data. I contacted several German and Swedish web comparison websites in November 2016. *Guenstiger.de* is one of the leading shopbots in the German market and they granted me permission to collect data from their website. This shopbot lists prices from over 3000 online shops, with a focus mainly on consumer electronics.

In the course of November and December 2016 I examined different product categories and products to determine if they were suitable for my purposes. Most importantly, the products needed to be perfectly homogeneous for the approach to be valid according to the assumptions of the theoretical model and to avoid a bias. The product categories computer hard drives, printers and televisions proved to be robust to this requirement. I obtained a selection of products by extracting the 400 most popular products on the website in each of the respective categories before starting to collect the data.[7] I then randomly selected my products from this sample. The randomization was processed in Excel by assigning each product a random number between 0 and 1, and then selecting the 70 products with the 70 highest random numbers in each category. The final set of products consists of 207 electronics products, with 68 computer hard drives, 70 printers, and 69 televisions.

To collect my data I used the web scraping service *Parsehub*. This company provides *web*

[7] Consumer electronics may have relatively short life spans and older products may not be relevant to consumers and suppliers. To ensure that my product sample is composed of relevant products, I sorted the products in each respective category by popularity on the website. The popularity of products for this shopbot is determined by the amount of users visiting the product specific URL.

crawlers, which are able to visit web pages and collect specified information in an automated fashion. I configured the web crawlers to extract certain information from the HTML code of the web pages. Due to the fact that every product page has the same underlying web design, the web crawlers can then iterate through a specified list of product-URLs and extract the information according to the configuration.

The data was collected from the 17th of January to the 14th of February.[8] The data collection was scheduled to be collected every day between 13:00 and 14:00 CET. On each web page the web crawlers extracted the following information: the product URL, the product prices, the seller IDs, the shipping costs, the shipping time information, and the seller individual product description text. The shipping costs listed on this shopbot also include transactions fees of Paypal. In German e-commerce it is common for online shops to charge extra fees depending on the form of payment. Paypal is a commonly used payment method in Germany and these adjustments guarantee that the prices are comparable. More details on the data collection and data cleaning process are provided in Appendix A. The final data consists of 132,711 price quotes, over 28 days and 207 products. This results in a panel of 5,788 price dispersion observations.[9]

Common measures of price dispersion in previous literature include the range of prices, the standard deviation and variance, the difference between the average and lowest price, and the difference between lowest and second lowest price (Pan et al. 2004). All measures can be normalized by dividing them by the average price in the market. For this thesis, the price dispersion is measured using the coefficient of variation. One advantage of this measure is that it is a relative measure as opposed to for instance the standard deviation, thus allowing for comparisons across markets. Moreover, the coefficient of variation captures the full variation of prices within the market, in contrast to for instance the relative range of prices. Each of the 5788 price dispersion observations and average prices are calculated as shown by Equation (15).

$$\frac{\sigma_{jt}}{\bar{p}_{jt}} = \frac{1}{\bar{p}_{jt}} \sqrt{\sum_{i=1}^{N_{jt}} \left(p_{ijt} - \bar{p}_{jt}\right)^2} \quad with \quad \bar{p}_{jt} = \frac{1}{N_{jt}} \sum_{i=1}^{N_{jt}} p_{ijt} \tag{15}$$

The price dispersion for a product market for a specific day constitutes a single observation, where σ_{jt} describes the standard deviation, $\bar{p}_{jt}$ the average of prices and N_{jt} the number of sellers for the j^{th} product at day t. p_{ijt} describes the price of seller i for product j at day t. Table 2 shows the descriptive statistics for the examined products.

The upper part of Table 2 refers to prices excluding shipping and transaction fees, while

[8] On the 25th of January Parsehub experienced issues with there servers. Due to these complications it was not possible to collect data on this day.

[9] For eight observations the market only consisted of one listing. These observations were excluded as it is not possible to calculate the standard deviation.

Table 2: Summary Statistics of Data

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Products			**Coefficient of Variation**				**Price Average in Market**			
without Shipping	∅ Seller Number	Observations	Mean	Std. Dev.	Min.	Max.	Mean	Std. Dev.	Min.	Max.
All	22.924	5,788	.0969	.0506	.0018	.4149	508.67	687.07	43.079	4,142.27
70 Printers	29.265	1,959	.1076	.0425	.0018	.2884	297.12	408.04	43.079	2,912.28
68 Hard Drives	25.713	1,904	.0976	.0595	.0323	.4149	171.22	199.96	47.72	1,532.09
69 TVs	13.713	1,925	.0854	.0445	.0062	.2615	1,057.71	866.62	203.2	4142.27
with Shipping										
All	22.924	5,788	.0957	.0499	.0088	.4155	515.55	691.29	47.47	4,169.79
70 Printers	29.265	1,959	.1065	.0436	.0193	.2875	301.35	410.12	47.47	2,931.60
68 Hard Drives	25.713	1,904	.0958	.0575	.0225	.4155	175.46	200.20	52.15	1,537.41
69 TVs	13.713	1,925	.0847	.0453	.0088	.2629	1,069.92	870.56	206.38	4,169.79

Notes: *Descriptive statistics of examined markets, where each market observation includes the product market's seller number, coefficient of variation and price average. The calculations are shown in Equation (15). The upper half of the table is based on prices without additional fees; The lower half of the table is based on prices with additional fees.*
Source: *Own calculations of collected data from www.guenstiger.de*

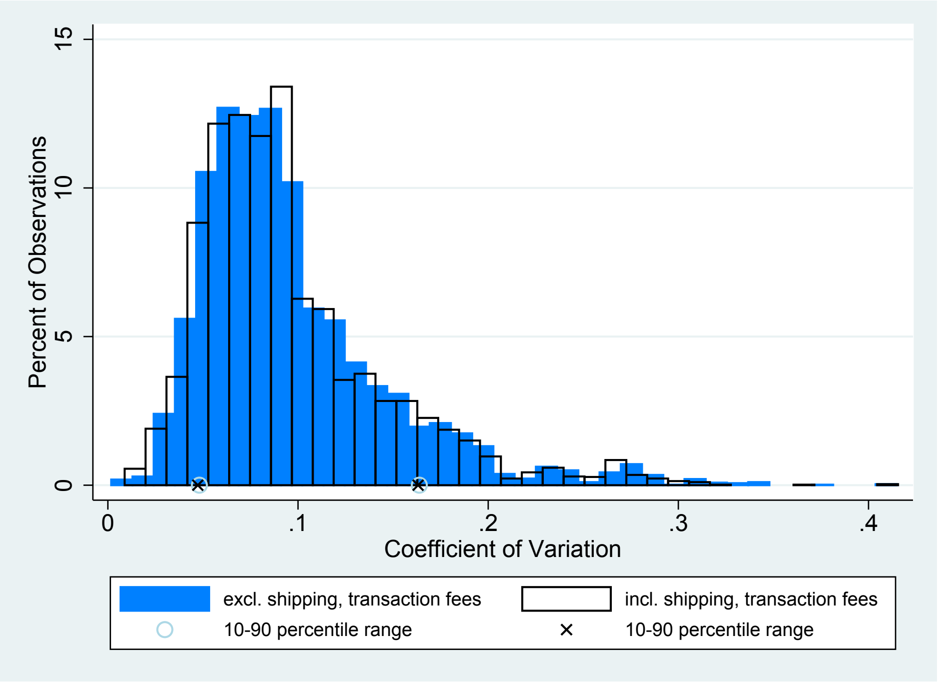
the lower half displays price statistics that include these fees. The average number of sellers (Column 1) for all markets is 22.9, with variation between product categories (on average 29.3 sellers for printers, 25.7 for hard drives and 13.7 for televisions). The same can be said for the mean of the average price levels of the product categories (Column 8), with hard drives being considerably cheaper on average than television for instance. The mean price dispersion (Column 4) measured with the coefficient of variation however does not vary as strongly across categories. Yet according to the relatively high standard deviations (Column 5) and range (Columns 6, 7) within the categories, the level of price dispersion across products and days does seem to vary strongly. Similarly, the same measures for the average price levels (Columns 9, 10, 11) indicate high variation of price levels between products. These descriptive statistics validate **Hypothesis 1** of there existing price dispersion in the examined markets. The overall average measured price dispersion including additional fees is 9.57% for my data. This result is similar to those of previous research papers who use the coefficient of variation as a dispersion measure (Table 1).[10]

Figure 1 shows histograms of price dispersion of all market observation including and excluding shipping and transaction fees. The histogram suggests that the percentage distributions of price dispersion only differ little when excluding or including shipping and transaction costs, yet the distribution of prices excluding additional fees is shifted slightly to the right. The markings on the x-axis show the 10 and 90 percentiles of price dispersion observations, which lie at 4.8% and 16.32% for prices including additional fees, and at 4.7%

[10] Pan et al. (2003), Ratchford et al. (2003) and Baye et al. (2004*a*) find comparable results. Clay et al. (2001) however find the dispersion to lie considerably higher in their sample.

and 16.31% for prices excluding additional fees. For the subsequent analysis in this thesis, I will use the prices including all fees, as they constitute the prices consumers actually pay when purchasing goods online. As the prices without additional fees may be a method of sellers to obfuscate consumers, I will carry out robustness checks in later sections.

Figure 1: Histogram of Price Dispersion



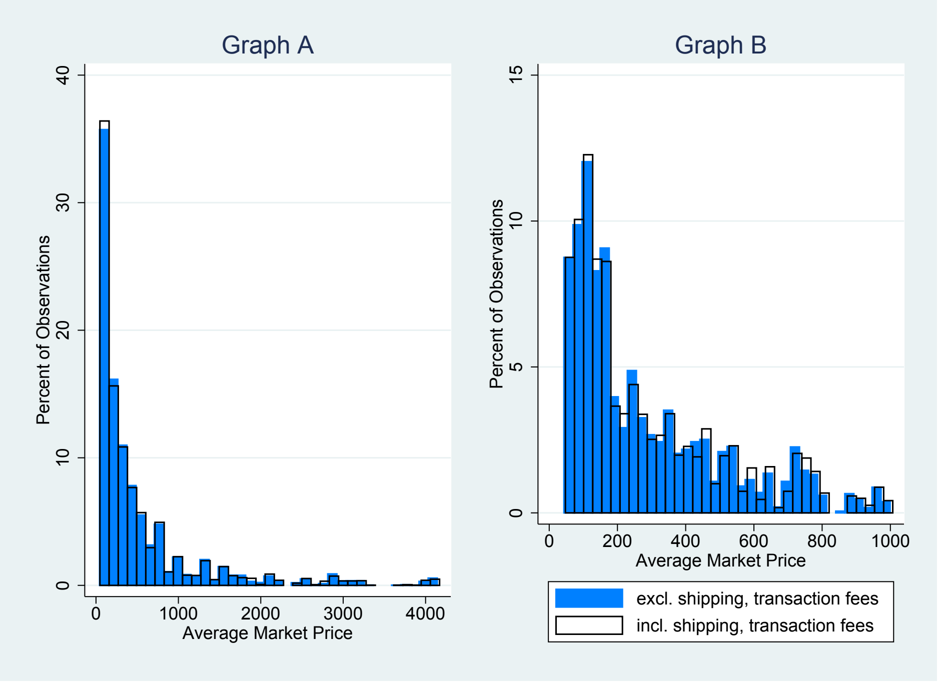
Source: *Collected data from www.guenstiger.de*

Figure 2 displays the histograms of average prices in markets for all products and across all days. Graph A shows the total price range of products, while Graph B only shows the sub sample of products with average prices below 1000 Euro, as these products make up the largest share. The comparison of the percentage distributions of prices including and excluding shipping and transaction costs suggest that they only differ little, apart from the fact that prices including shipping fees are naturally larger on average.

Figure 3 shows a histogram of the number of sellers in all product markets across all days. The graph indicates that there is strong variation in the number of sellers across markets, meaning that the market sizes for the examined products differ considerably.

Figure 4 shows the development of the average values of the main variables over time. The data suggests that there is a slight decrease of price dispersion over time, both when including and excluding additional fees. The price dispersion was slightly higher on the first day of measuring, yet when examining the data closer, it does not seem as this would come from a specific product or seller. There is also an overall decrease in the average price levels and average number of sellers, yet with temporary increases as well. This is in line with

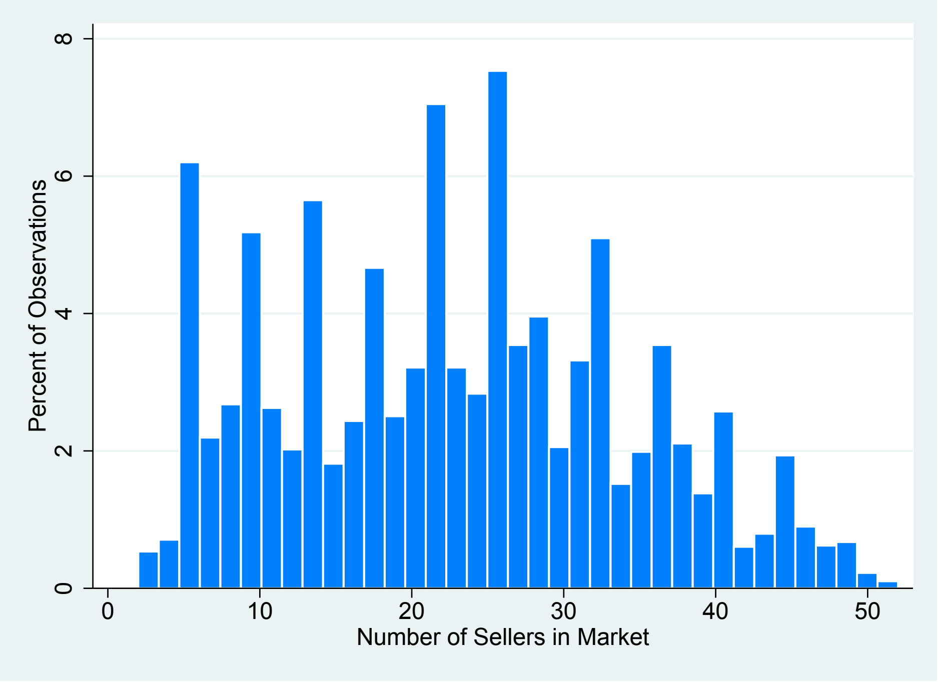
Figure 2: Histogram of Price Levels



Notes: *Graph A shows the distribution of average market prices for all observations. Graph B shows the distribution of average market prices for average market prices below 1000 Euro.*

Source: *Collected data from www.guenstiger.de*

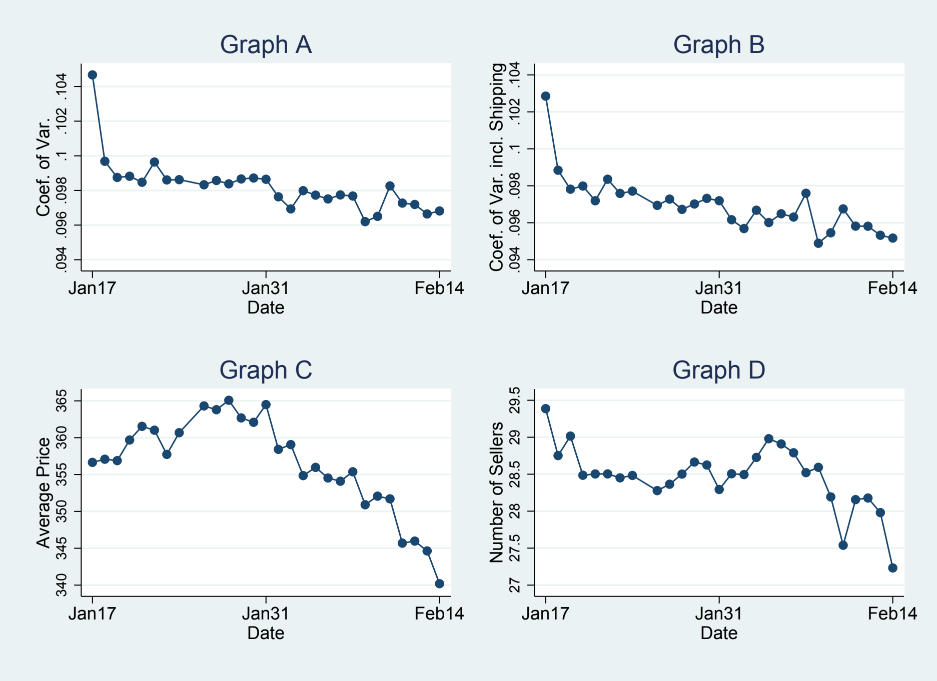
Figure 3: Histogram of the Number of Sellers in the Markets



Source: *Collected data from www.guenstiger.de*

other literature: Baye et al. (2004*a*) for instance also do not find a discernible time pattern for their half year data, noting that price dispersion is not a temporal inequilibrium.

Figure 4: Plots of Averages of Main Variables over Time



Notes: *Graph A shows the average coefficient of variation across days using prices excluding additional fees. Graph B shows the average coefficient of variation across days using prices including additional fees. Graph C shows the average market price and Graph D the average seller number across days.*
Source: *Collected data from www.guenstiger.de*

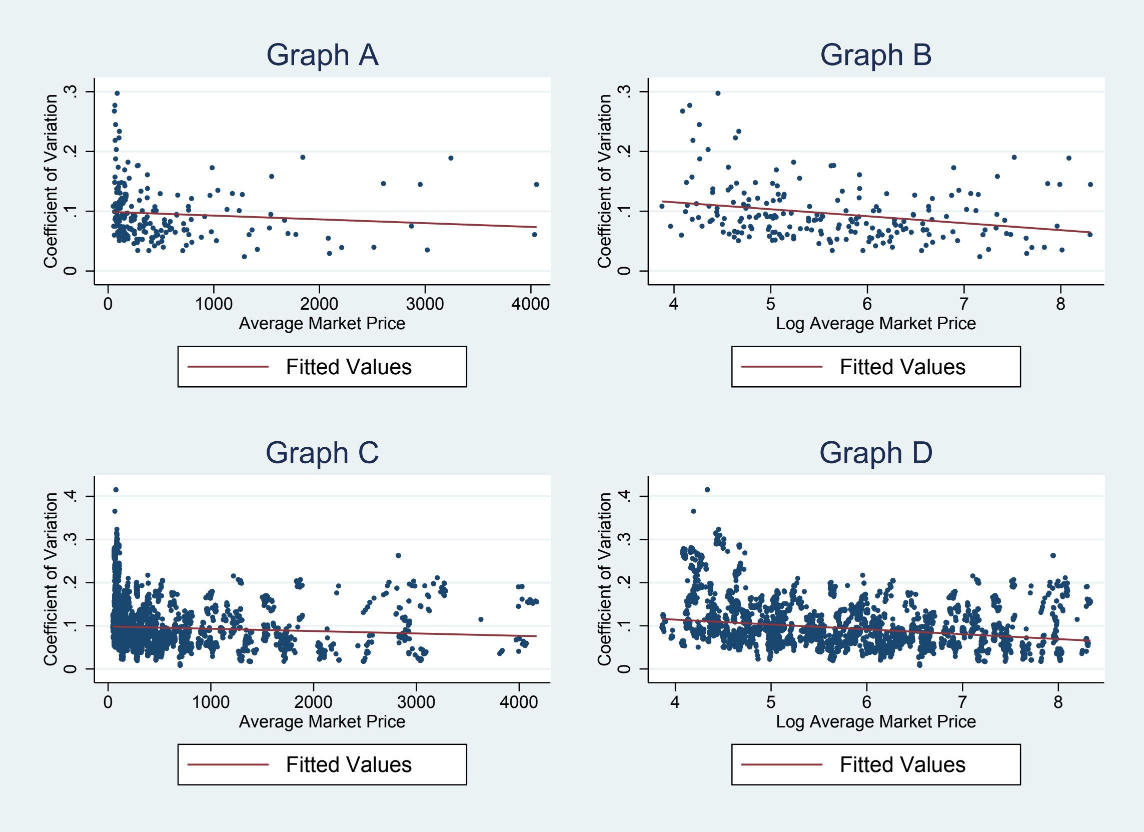
5 Regression Analysis

5.1 Methodology

To examine the relationship of price dispersion with the number of sellers in the market, as well as with the price level of products, I use a regression framework. A number different approaches have been used in previous research regarding the functional form and the treatment of the explanatory variables.

First, I will discuss the relationship of price dispersion to the price level in the market. Previous research papers have used different specifications: while Pan et al. (2003) use the linear average price in the market, Ratchford et al. (2003) find that the log of the average price provides a better fit. Figure 5 shows scatter plots of my data. Graphs A and B display all product markets averaged over days, while the Graphs C and D show all product markets across all days individually. The scatter plots of the log average price and the price dispersion in Graphs B and D seem to fare much better than a linear approach in Graphs A and C and therefore would best suit my specification. In a simple regression, the log price approach shows an R-squared of 6.77 % (5.51%) for the pooled (time-averaged) data, while the linear regression reports an R-squared of 0.56% (0.87%).

Figure 5: Scatterplots of the Average Market Price vs. the Price Dispersion

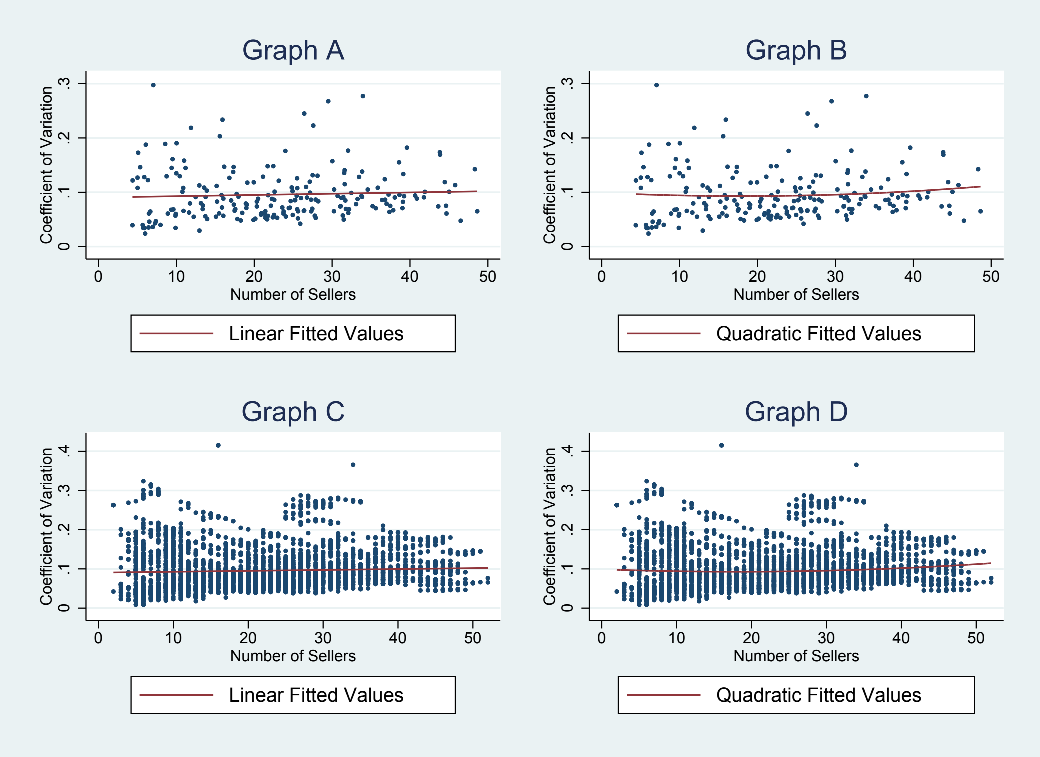


Notes: *Graph A shows a scatter plot of the coefficient of variation and the average market price with time averaged data. Graph B shows a scatter plot of the coefficient of variation and the log average market price with time averaged data. Graph C shows a scatter plot of the coefficient of variation and the average market price with all observations. Graph D shows a scatter plot of the coefficient of variation and the log average market price with all observations.*
Source: *Collected data from www.guenstiger.de*

Moreover, a semi-elastic relationship seems sensible when considering consumer behavior. From my theoretical model, I argued that higher price levels lead to more search and thus less price dispersion. As this stems from the relative relationship of search costs to price levels, it would be reasonable to also model this relative relation in the regression specification. With a lin-log specification, a 1% increase in price levels would imply an absolute increase in price dispersion by $\beta/100$, with β being the estimated coefficient. With this, two lower priced products with an absolute difference of X Euro in average price level will show a larger relative difference in price dispersion than two higher priced products with the same absolute difference in price level, i.e. a convex relation. For instance, the difference in price dispersion between two products priced 1000 and 1100 Euro will be lower than for two products priced 100 and 200 Euro, even though the absolute difference is identical.

Next, I discuss the specification of the number of sellers in the model. Baye et al. (2004*a*) point out that the number of sellers may have a non-linear relationship to price dispersion, based on their simulations with several different theoretical models. Similarly, Baye et al. (2003) find evidence of a non-linear relationship of price dispersion and the number of sellers in a market. Baye et al. (2004*a*), Ratchford et al. (2003), and Pan et al. (2003) all include squared seller terms in their specifications. When plotting these variables in my data, the inclusion of a squared seller term does not seem to better fit the data (Figure 6).

Figure 6: Scatterplots of the Number of Sellers vs. the Price Dispersion



Notes: *Graph A shows a scatter plot of the coefficient of variation and the number of sellers with time averaged data using a linear fitted line. Graph B shows a scatter plot of the coefficient of variation and the number of sellers with time averaged data using a quadratic fitted line. Graph C shows a scatter plot of the coefficient of variation and the number of sellers across all observations using a linear fitted line. Graph D shows a scatter plot of the coefficient of variation and the number of sellers with all observations using a quadratic fitted line.*

Source: *Collected data from www.guenstiger.de*

Again, Graphs A and B display all product markets averaged over days, while the Graphs C and D show all product markets across all days individually. Graphs A and C include a linear fitted line; Graphs B and D include a quadratic fitted line. The R-squared is very close to zero when running simple regressions and the adjusted R-squared is even negative for the regressions including both linear and quadratic terms. I choose to model a linear relationship in my specifications, yet will test for a non-linear relationship in my robustness analysis.

Depending on the nature of the underlying data, the specifications in previous research also differed with regard to addressing fixed effects. Ratchford et al. (2003) and Pan et al. (2003) use repeated cross sectional data with two and three time instances respectively, with time fixed effects to model the differences. Baye et al. (2004*a*) use both a linear time trend variable and time fixed effects in their different regression specifications for their panel data. To address the unobserved heterogeneity across products, Baye et al. (2004*a*) use a product fixed effects specification next to their pooled regression. Ratchford et al. (2003) on the other hand use fixed effects for the respective product categories and Pan et al. (2003) utilize a random effects specification for their product categories.

When examining my data closer, there seems to be substantial variation across products for my variables (Table 2), yet only very little variation over time, owing to the fact that the data was only collected over the course of four weeks. As there is only very little within variation for products, a products fixed effects approach would capture almost all variation and would not be a efficient method to use. I therefore use fixed effects for the product categories. My proposed baseline specification is shown in Equation (16),

$$\frac{\sigma_{it}}{\overline{p}_{it}} = \beta_0^1 + \beta_1^1 * \ln \overline{p}_{it} + \beta_2^1 * sellers_{it} + \beta_3^1 * HD_i + \beta_4^1 * TV_i + \epsilon_{it}^1 \tag{16}$$

where $\sigma_{it}/\overline{p}_{it}$ is the coefficient of variation of prices, $\ln \overline{p}_{it}$ is the log of the average price level, $sellers_{it}$ are the number of sellers, and HD_i and TV_i represent the product category fixed effects for the i$^{\text{th}}$ product at day t. The product category printers serves as the benchmark to the category dummy variables.

As there still seems to be a weak linear time trend for my dependent variable (Figure 4), I include one specification with a linear time variable (Equation (17)),

$$\frac{\sigma_{it}}{\overline{p}_{it}} = \beta_0^2 + \beta_1^2 * \ln \overline{p}_{it} + \beta_2^2 * sellers_{it} + \beta_3^2 * HD_i + \beta_4^2 * TV_i + t + \epsilon_{it}^2 \tag{17}$$

where the term t in represents the linear time trend variable. The fact that there is only very little time variation however also implies that the variables are nearly perfectly correlated over time within products, leading to nearly perfectly correlated residuals if using a pooled approach. Indeed, when examining the correlation of the residuals and lagged residuals, the correlation is very close to one. I thus include a specification where the data is averaged over time as an approach of eliminating the problem of autocorrelation within product clusters, while preserving the cross-sectional variance (Equation 18).

$$\frac{\sigma_i}{\overline{p}_i} = \beta_0^3 + \beta_1^3 * \ln \overline{p}_i + \beta_2^3 * sellers_i + \beta_3^2 * HD_i + \beta_4^2 * TV_i + \epsilon_i^3 \tag{18}$$

As suggested by Cameron & Miller (2015), I cluster the standards errors on product level for the first two pooled specifications to allow for autocorrelation and to avoid an underestimation of the models standard errors and subsequent over-prediction of my estimations. Standard OLS is used as estimator for all specifications.

It has to be noted that the estimated relationships cannot necessarily be interpreted as causal, as there are endogeneity concerns. Price dispersion in markets may not simply be a result of market sizes or market prices, as reverse causality may exist. For instance, a market with a large number of sellers may feature a large price dispersion, as consumers have to engage in more search. However, an already large price dispersion may attract further sellers, when there is the possibility to charge prices over marginal cost. The number of sellers and price levels can therefore not be seen as exogenous determinants, but endogenous to price dispersion.

Moreover, dispersing prices may not only be a consequence of search costs and imperfect information, but also incorporate premiums for seller heterogeneity. As already noted by Stigler (1961), products are never fully homogeneous, because they may be sold in a heterogeneous context. For instance, one seller may differentiate themselves by offering superior customer service, which allows for a higher price and no direct comparison possibility to other sellers. Previous research has however not managed to explain existing price dispersion by controlling for various seller characteristics (Baye et al. (2006), Pan et al. (2004)). Ratchford et al. (2003) for instance find that differences in e-tailer services explain only a very little portion of price dispersion, when controlling for factors such as the *ease of ordering*, *product selection*, *customer support*, or *shipping and handling.* Clay et al. (2002) compile a comprehensive list of store attributes, including informational aspects and services such as reviews and recommendations, to model the heterogeneity of their seller sample, but do not find strong correlations between these and price levels. In their paper, Brynjolfsson & Smith (2000) point out that many distinguishing factors between retailers are merely of informational value, and are not strictly bound to the product. For instance, customers may utilize the superior amount of product information or customer reviews at retailer A, but still purchase at retailer B.

Lastly, due to the time limitations of a master thesis, my data covers a relatively short time span. Other research papers have used more extensive datasets, that may provide more generalizable results.

5.2 Results and Discussion

Table 3 shows the regression results. In the baseline specification (Model (1)), a 10% higher priced market would feature a 0.00155 lower price dispersion. In a market with the mean price dispersion level of 0.0957, a 10% increase of market price level would imply a price dispersion of 0.0942, i.e. a relative decrease of 1.62%. This is in line with **Hypothesis 3**, suggesting that higher price levels are related to a lower dispersion in prices. As discussed in my theoretical model, this may be driven by the relation of price levels and consumers' search costs, where there is a greater incentive for consumers to search more for higher priced goods, as the potential benefits are higher in relation to their constant search costs. With higher search, the theoretical model predicted a decrease in price dispersion. The empirical results are also consistent with those of Ratchford et al. (2003) and Pan et al. (2003), who also find a significant negative relationship. In Ratchford et al. (2003) sample, the average electronics market features a price dispersion of 0.0965. They measure a relationship of -0.000112 of the log average price to the level of price dispersion, meaning that with a 10% increase of the average price level, the average electronics market would see an decrease of price dispersion of 0.00112 to 0.0954, i.e. a 1.16% relative decrease.[11]

Next, the number of sellers does not show a significant relationship to the price dispersion in any of the specifications, with the exception of Model (4), where unclustered standard errors are used. Based on the theoretical model, I hypothesized that price dispersion increases with more sellers in the market, as consumers may be more obfuscated when there are more sellers in the market. In a larger market, consumers have a lower probability of finding lower prices, making it easier for firms to differentiate prices. However, as there seems to be no significant relationship, there is no supporting evidence for **Hypothesis 2** in my analysis. In related literature, different results have been found. Baye et al. (2004*a*) and Baye et al. (2003) find that the price dispersion is negatively related to sellers. Pan et al. (2003) find a positive relationship of sellers and price dispersion in their specifications. In their paper, the relationship is however not significant when using the coefficient of variation as dispersion measure, only when using the relative range of prices. As previously mentioned, many research papers model a non-linear relationship in their specifications. In the robustness section I will test and discuss this possible non-linear relationship.

The product category dummies are partially significant. The results suggest that hard drives exhibit a significantly lower price dispersion than televisions and printers. The television dummy however does not show significance, indicating that televisions do not show significantly different levels of price dispersions than printers.

[11] Pan et al. (2003) find a linear relationship of -0.0003 of the log average price to the level of price dispersion. The specification is however linear and therefore not comparable. The average market has a price dispersion of 0.1172 in their sample, meaning that a $10 increase of the price level would come with a relative decrease of the price dispersion of -0,026% in this market.

Table 3: Regression Results

	(1) Price Dispersion	(2) Price Dispersion	(3) Price Dispersion	(4) Price Dispersion
Mean Price Dispersion	0.0957	0.0957	0.0957	0.0957
ln average price	-0.0155**	-0.0155**	-0.0162***	-0.0155***
	(0.001)	(0.001)	(0.001)	(0.000)
# of Sellers	-0.000514	-0.000522	-0.000569	-0.000514***
	(0.183)	(0.177)	(0.156)	(0.000)
TV Dummy	-0.00863	-0.00873	-0.00912	-0.00863***
	(0.356)	(0.351)	(0.338)	(0.000)
HD Dummy	-0.0192**	-0.0193**	-0.0199**	-0.0192***
	(0.009)	(0.009)	(0.008)	(0.000)
Day		-0.000207*		
		(0.015)		
Intercept	0.204***	0.208***	0.210***	0.204***
	(0.000)	(0.000)	(0.000)	(0.000)
Standard Errors	Clustered	Clustered	Robust	Robust
N	5788	5788	207	5788
R^2	0.082	0.083	0.101	0.082
adj. R^2	0.081	0.083	0.084	0.081
F	4.965	4.830	5.283	114.5
Prob > F	0.0008	0.0003	0.0005	0.0000

Notes: *p-values in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; Model (1) is baseline specification; Model (2) includes linear time variable; Model (3) uses time averaged data; Model (4) is baseline specification with unclustered standard errors; Clustered standard errors on product level for Models (1) and (2), 207 clusters; Robust standard errors for Model (3) and Model (4).*

Source: *Collected data from www.guenstiger.de*

In Model (2), the included linear time trend shows a small but significant negative trend. When considering the scale of this variable (28 days) and the negligible differences in intercepts between Model (1) and Model (2), this variable does not seem to add much explanatory value to the specifications. In Model (3) time averaged data is used, yet the results are very similar to those of Model (1) and (2). This indicates that the time variation is indeed very low, and a product fixed effects approach such as in Baye et al. (2004*a*) would not be a suitable specification. Lastly for comparison, Column (4) shows the results for the baseline model (Model (1)) when only using robust standard errors and not clustered standard errors. In this model all estimations are highly significant when not addressing the autocorrelation in the data, meaning that the standard errors are too low in this specification. All models show joint significance and an explanatory value of roughly 8% according to the adjusted R-squared.

Overall, the results are very similar across all specifications regarding sign, magnitude and significance, with the exception of Model (4) using robust standard errors. My data and analysis show that there still exists significant price dispersion in online retail markets. In line with previous research and against the initial conjecture that the Internet would decimate inefficiencies and informational asymmetries, the examined markets show that sellers manage to differentiate prices for homogeneous goods to a significant degree. From a search theoretical view point, this is best explained by consumers not engaging in exhaustive search to sample all sellers prices in the market. As noted by Levin (2011), consumer search costs in online markets may still be non-trivial, even though from a economic theory point of view, the markets may seem to show near perfect information. In several older research papers, it has been pointed out that the measured price dispersion may be a result of the immaturity of online markets (Pan et al. 2004). My results however show, that price dispersion in online markets has been consistent and constitutes a pricing equilibrium.

In my regression models, the log price level shows a significant negative relationship to price dispersion. This relationship suggests that consumers engage in varying levels of search depending on the expensiveness of a good. Furthermore, the non-existence of a relationship between the number of sellers and price dispersion suggests that the degree to which sellers can differentiate prices and discriminate between consumers does not relate to the market concentration. These findings stand in contrast to previous results, yet these also have not been consistent throughout research papers.

My data is composed of electronics products, meaning that results cannot readily be generalized to other product categories. Moreover, due to the time limitations of a master thesis, my sample size is relatively small compared to the studies of for instance Pan et al. (2003), Ratchford et al. (2003), or Baye et al. (2004*a*). I however argue that my results can be seen as representative of electronics products, as I extracted a sufficiently sized random sample of relevant products.

5.3 Robustness

To check the robustness of my empirical approach, I remodel my specifications in several different ways. First, a number of previous research papers found theoretical and empirical evidence of a non-linear relationship between the number of sellers and price dispersion.

When including a squared term to the baseline model, the new specification (Table 4, Model (2)) does not increase in significance or fit. Similar to the baseline model (Table 4, Model (1)), both *seller* terms remain insignificant in the non-linear specification. This suggests, that there is neither a linear nor non-linear relationship of price dispersions and sellers in my data. There is however a slight increase in the adjusted R-squared.

Table 4: Robustness Regression Results

	(1) Price Dispersion	(2) Price Dispersion	(3) Price Dispersion	(4) Price Dispersion
Mean Price Dispersion	0.0957	0.0957	0.0684	0.0969
ln average price	-0.0155**	-0.0159***	-0.0124***	-0.0157**
	(0.001)	(0.001)	(0.000)	(0.001)
# of Sellers	-0.000514	-0.00195	0.000219	-0.000522
	(0.183)	(0.149)	(0.491)	(0.191)
# of Sellers^2		0.0000290		
		(0.206)		
TV Dummy	-0.00863	-0.00866	-0.0105	-0.00881
	(0.356)	(0.355)	(0.092)	(0.356)
HD Dummy	-0.0192**	-0.0178*	-0.0244***	-0.0190*
	(0.009)	(0.019)	(0.000)	(0.012)
Intercept	0.204***	0.220***	0.146***	0.206***
	(0.000)	(0.000)	(0.000)	(0.000)
N	5788	5788	5726	5788
R^2	0.082	0.088	0.185	0.084
adj. R^2	0.081	0.087	0.184	0.083
F	4.965	4.575	11.69	5.139
Prob > F	0.0008	0.0006	0.0000	0.0006

Notes: *p-values in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; Model (1) is baseline specification; Model (2) includes squared seller term; Model (3) uses data excluding outer range of prices; Model (4) uses prices that exclude additional shipping and transaction fees; Clustered standard errors used on product level for all Models, 207 clusters.*

Source: *Collected data from www.guenstiger.de*

In a further approach, I included dummies for different groups of seller numbers, i.e. a dummy variable for 1-10 sellers, a dummy variable for 11-20 sellers, etc. This model only showed highly insignificant estimates for the coefficients of these dummies. The results are not reported in this paper.

Next, I test if my specifications are driven by outliers. When examining my data, some markets showed prices that were far from the market average. Abnormally low prices may be caused by certain sellers offering their product at a sales price, while abnormally high prices may be caused by sellers not updating their prices to market movements. To control for this, I exclude the highest and lowest price in every market and recalculate the price dispersion, seller number and average market price.

The price dispersion is reduced from 0.0957% to 0.0684% and the average price is slightly lower with 507.71 Euro compared to 515.55 Euro previously. This shows that the price dispersion changes drastically when excluding extreme values and that the upper bounds were on average relatively further away from the market mean than the lower bounds. Naturally, the seller number is reduced by two on average. This dataset includes only 5726 market observations across days and products, as some markets only consisted of no sellers or one seller after the exclusion of the price quotes.

The results are displayed in Table 4, Model (3). A 10% increase of average price level for the average market with a price dispersion of 0.0684 would lead to an absolute decrease of price dispersion of 0.00124, i.e. a relative decrease of 1.81%. When plotting the predictions of the two models next to each other, the difference can be can be best described by an almost parallel downward shift of the fitted line of the robust model. This means that the differences between models are mainly driven by the differences in intercepts and suggests that the price dispersion was approximately uniformly decreased over all price levels. Therefore, estimated relationship with the log average market price is stable relative to the newly calculated values of price dispersion. The other estimates do not vary strongly, with a still insignificant relationship of the number of sellers, and a significant negative sign of the hard drives dummy. Moreover, the fit of the model has increased greatly when considering the adjusted R-squared of 18.4%.

As a last check, I rerun my baseline specification using the price quotes that exclude shipping and transaction fees (Model (4)). The descriptive statistics suggested that the differences in price dispersion and average price levels were only very small, yet Ellison & Ellison (2009) suggest that online sellers may engage in obfuscation strategies. For instance, some sellers may obfuscate consumers by displaying lower list prices but charge high shipping and transaction fees. If this occurs systematically for certain markets, the estimated relationships may change. The results however are very similar to those of the baseline model (Model (1)), which suggests that shipping and transaction fees are neither a driver of price dispersion, nor of the relationships of price dispersion with price levels, and with the number of sellers.

6 Monte Carlo Simulations

6.1 Methodology

Using the price quotes in the examined markets, I next aim to estimate the value of information to consumers. As there exists price dispersion in the examined online markets, consumers must engage in costly search. They face the decision of increasing search with a possible price deduction, yet must invest more time and effort. Previous research papers such as Baye et al. (2003) or Pan et al. (2003) have provided estimates of the value of information by comparing the price paid by an uninformed consumer only obtaining one price quote to a fully informed consumer who pays the lowest price in the market. These values were calculated by subtracting the market minimum price, the fully informed consumer's price, from the average market price, the uninformed consumer's price. I differentiate this analysis by quantifying the payoff from incremental searches, i.e. examine the marginal value of information from additional search in online consumer markets.

Assume for the purpose of these simulations that consumers have heterogeneous search costs and engage in fixed sample search rather than sequential search. In this case, consumers decide upon a fixed sample size of price quotes prior to searching according to their marginal search cost and expected marginal benefit. Moreover, assume that consumers obtain a fixed sample of k price quotes from the N prices within a product market, and purchase the good at the lowest obtained price quote. Based on the model by Moraga-González et al. (2016), the sampling decision of a consumer is given by:

$$v - E(min\{p_1, ..., p_k\}) - kc > 0 \tag{19}$$

$$E(min\{p_1, ..., p_{k-1}\}) - E(min\{p_1, ..., p_k\}) > c \tag{20}$$

$$E(min\{p_1, ..., p_k\}) - E(min\{p_1, ..., p_{k+1}\}) < c \tag{21}$$

Consumers sample k price quotes with cost c per search with $k \in \{1, .., N\}$ (Inequalities (19), (20) and (21)). Inequality (19) states that the valuation v of the product subtracted by the expected minimum price paid when sampling k price quotes and the cost of searching for k price quotes must be greater than zero. The optimal k is reached when the sampling of one price quote less will have greater expected gains than costs of search (Inequation 20) and the sampling of one additional price quote will induce greater cost of search than expected gains (Inequality 21). This results in there existing N groups of consumers obtaining $k \in \{1, ..., N\}$ price quotes.

To obtain estimates of the value of information, I run simulations according to these consumer behavior assumptions. For each of the N groups of consumers I simulate the average sampling behavior by repeatedly sampling k price quotes in the examined markets.

Using this, the average paid premium for the k consumer groups can be obtained by calculating the average difference between the minimal sample price and the minimal market price p_{min}, with S being the number of simulation repetitions. Equations (22) and (23) show the calculations for the absolute and relative premium respectively.

$$\frac{1}{S}\sum_{i=s}^{S}\big(min\{p_{1,s},...,p_{k,s}\} - p_{min}\big) = "abs.Premium_k" \quad \forall \quad k = 1,...,N \tag{22}$$

$$\frac{1}{S}\sum_{i=s}^{S}\big(\frac{min\{p_{1,s},...,p_{k,s}\} - p_{min}}{p_{min}}\big) = "rel.Premium_k" \quad \forall \quad k = 1,...,N \tag{23}$$

With this, the marginal benefits, i.e. the value of additional information, between consumer groups sampling k and $k+1$ price quotes can be obtained. As an example, let consumers who search for five price quotes on average pay 200 Euro for product i, while consumers who search for four price quotes on average pay 210 Euro. The marginal value of information can be expressed as the average gain from one additional search, in this case 10 Euro or 5%.

First, it is however important to consider that consumers are more likely to sample certain online shops than others. To model this, I use website traffic estimates from *similarweb.com*[12] as a proxy for the likelihood of consumers observing certain price quotes, as online shops with a high amount of traffic are more likely to be sampled by consumers. Using this data, each seller is assigned a sample probability weight according to their traffic estimate. A more detailed description of the data is provided in Appendix B.1.

In the simulation a consumer observes k prices according to the underlying probability distribution in the market. As the sampling is weighted and conducted without replacement, it is processed consecutively by the simulation program, i.e. the first draw has an underlying probability distribution including all price quotes, while the j^{th} draw is only subject to the probability distribution of the remaining unsampled $N - (j - 1)$ prices in the market. Following this, the probability of price n to be drawn at draw j is given by Equation (24), with w_n being the probability weight of price n, j being the indicator of the draw, and i the index for the unsampled price quotes in the market.

$$\rho_{n,j} = \frac{\omega_n}{\sum_{i=1}^{N-(j-1)}\omega_i} \quad with \quad j \in \{1,...,k\} \tag{24}$$

For the first draw ($j = 1$) the equation simply becomes:

[12] SimilarWeb is a marketing intelligence company providing data on Internet user behavior. Amongst others, the data sources include a panel of browsing data from anonymous Internet users, currently the largest panel in the industry.

$$\rho_{n,1} = \frac{\omega_n}{\sum_{i=1}^{N} \omega_i} \tag{25}$$

Due to computational restrictions, I randomly chose 10 printers, 10 hard drives and 10 televisions from the sample of 207 products. I furthermore randomly sampled 10 of the 28 available days. The simulations were run 200 times for each product for each day, resulting in a total of 20,000 simulation runs for each product category. From the data description it became apparent that there was only little variation across time, yet significant variation across products. The results for each product are therefore reported separately, yet averaged over days. The simulation code is shown in Appendix B.2.

As already noted in the regression methodology section, a setback to this approach is the fact that it does not consider unobserved seller heterogeneity (see Section 5.1). Furthermore, only online offers are considered in this simulation, and consumers may well also sample prices at competing brick-and-mortar stores.

6.2 Results and Discussion

Table 5 displays the comparative statistics of the randomized simulation subsample and the total sample. With the exception of the average price of printers and the price dispersion for hard drives, the variable means are reasonably similar. The t-tests I used to compare means indicated that the sample means were not significantly different from each other. Yet as the sample sizes for each product category are fairly small, the t-tests lacked power and did not yield reliable results.

For each product category, I display the simulation results in graphs with the relative and absolute price premium in relation to the number of searches. Figure 7 shows the results for the printers, Figure 8 the results for the hard drives and Figure 9 the results for the televisions. The upper graphs show each of the 10 products individually, while the lower graphs show the averages. The complete list of randomly selected products and days can be found in Tables 7 and 8, Appendix B.3. The averaged numerical simulation results are provided in Appendix B.4. As the products exhibit different market sizes, the maximum number of searches varies for each line. The lines approach the x-axis with an increasing number of searches, as the average premium is reduced with an increasing number of sampled prices.

First, there seem to be significant differences of the relative and absolute premium levels for consumers only engaging in one search. Televisions are the most expensive product category, with an average price of 1150.33 Euro within this simulation sample. Their absolute premium for consumers only engaging in one search is also highest with an average value of 64.65 Euro paid, yet low in relative terms with a relative premium of 6.6%. The price averages for hard drives and printers in this sample are roughly the same (195.52 Euro and 204.66 Euro). Yet, the initial relative and absolute premiums differ as well. While consumers only engaging in one search on average pay 11.62 Euro (9.2%) more than the minimum market price for hard drives, the premium for printers is 24.53 Euro (11.9%). As already discussed in section 4, there are differences in market size between categories. While printers and hard drives typically show larger markets, with an average of 29.57 and 24.5 sellers respectively, television markets are relatively small with an average of 13.7 sellers in this simulation sample (Appendix B.3).

Table 5: Comparison of Products in Simulation Subsample to Complete Sample

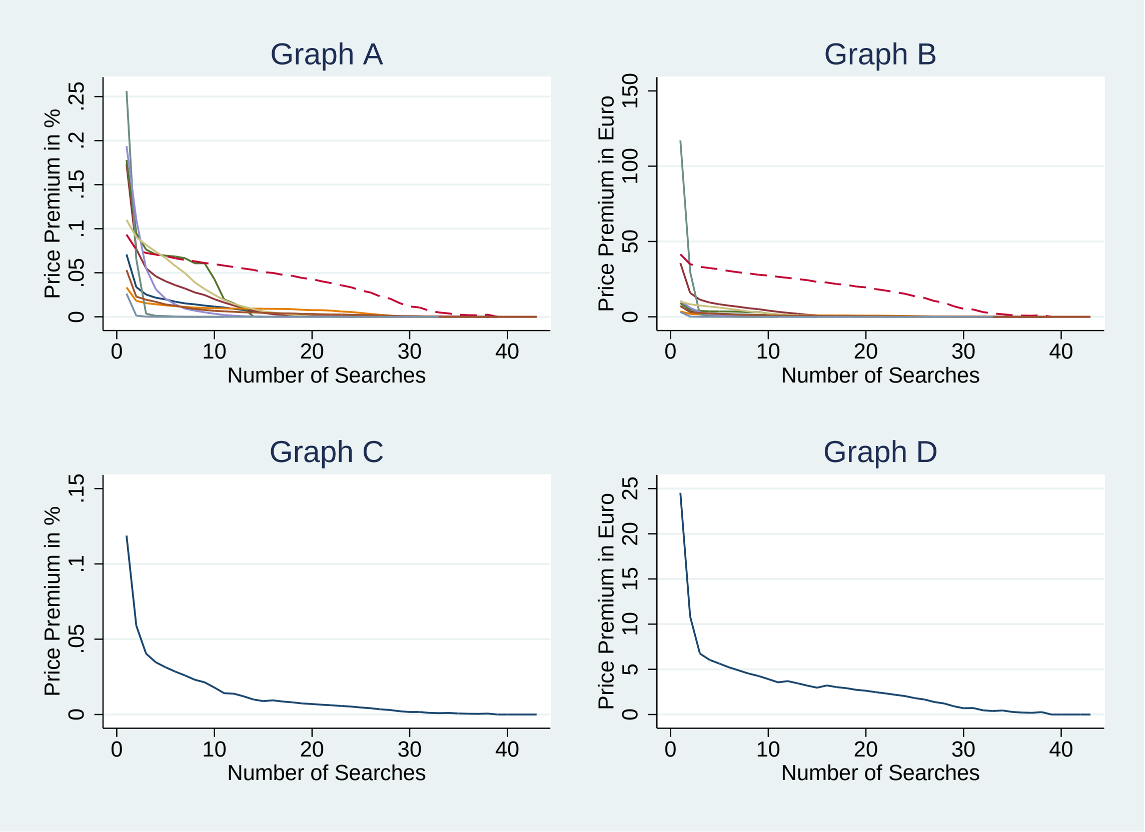
	N	**# Sellers**	**∅ -Price**	**Coef. of Var.**
Televisions Subsample	10	13.71	1,150.33	.0854
Televisions Total Sample	69	13.71	1,069.92	.0847
Hard Drives Subsample	10	24.49	195.52	.0788
Hard Drives Total Sample	68	25.71	175.46	.0958
Printers Subsample	10	29.57	204.66	.1161
Printers Total Sample	70	29.27	301.35	.1065

Notes: *This table provides summary statistics for the examined products in the simulation sample and for all products in the total sample. For the simulation sample, the measures reflect the mean values for the 10 sampled products and 10 sampled days in each product category. For the total sample, the measures are averages across all products and days in each product category. Displayed are the number of products (N), the number of sellers (# Sellers), the average market price (∅ -Price) and the coefficient of variation of prices.*
Source: *Own calculations of collected data from www.guenstiger.de*

The average premium graphs for all categories can be best described by a convex function. It can be noted that the reduction in the premium for the first few searches is substantial for the average values. By sampling approximately two to five price quotes, consumers will reduce the average premium greatly. The graphs of the separate product lines however also show that this is not necessarily true for each product individually. As an example, the dotted line in Figure 7 representing the printer *HP M553dn* shows no large decrease for the first searches and a relatively flat progression. On average however, this shows that consumers, who engage in very little search may already have large benefits. Especially for smaller markets, for instance televisions, a small amount of search may already be sufficient to have relatively large benefits.

For a moderate amount of searches, the average line rapidly flattens out. Across all product categories, the average premiums are not greatly reduced after approximately 10 searches. The individual product trends here differ greatly however. For some products, the premium is already zero or approaching it, while there is still a significant premium on other products. This is especially true for printers and televisions when examining Graphs

Figure 7: Simulation Results Printers



Notes: *Graph A shows the relative price premium in percent versus the number of searches; Graph B shows the absolute price premium in percent versus the number of searches; Graph C shows the mean relative price premium in percent versus the number of searches; Graph D shows the mean absolute price premium in percent versus the number of searches.*

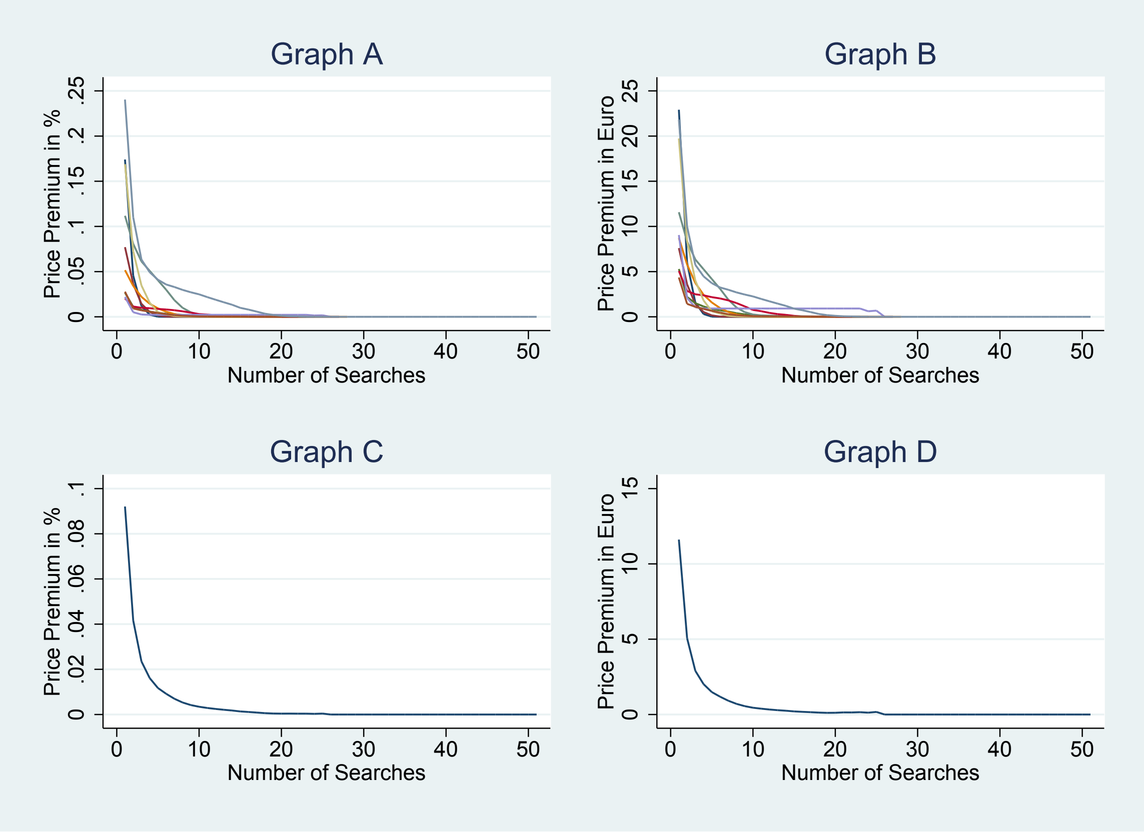
Source: *Simulations on collected data from www.guenstiger.de and www.similarweb.com*

A and B in Figures 7 and 9. With few exceptions, the premium typically approaches zero at no later than 20 searches across all product categories. Naturally, this occurs earlier for products with smaller market sizes.

Overall, we can say that there are typically large gains for the first number of searches, yet varying benefit developments after these initial searches. When assuming that consumers are sophisticated and can form their expectations based on previous search processes, it would be most beneficial to either engage in little search or exhaustive search. This depends on the individual search costs and is caused by the uncertainty in gains for moderate search.

This may provide a search theoretic explanation to the existence of price dispersion in online markets. While the Internet has provided many search cost reducing mechanisms for consumers, it may still be more beneficial for consumers with relatively high search costs to

Figure 8: Simulation Results Hard Drives



Notes: *Graph A shows the relative price premium in percent versus the number of searches; Graph B shows the absolute price premium in percent versus the number of searches; Graph C shows the mean relative price premium in percent versus the number of searches; Graph D shows the mean absolute price premium in percent versus the number of searches.*

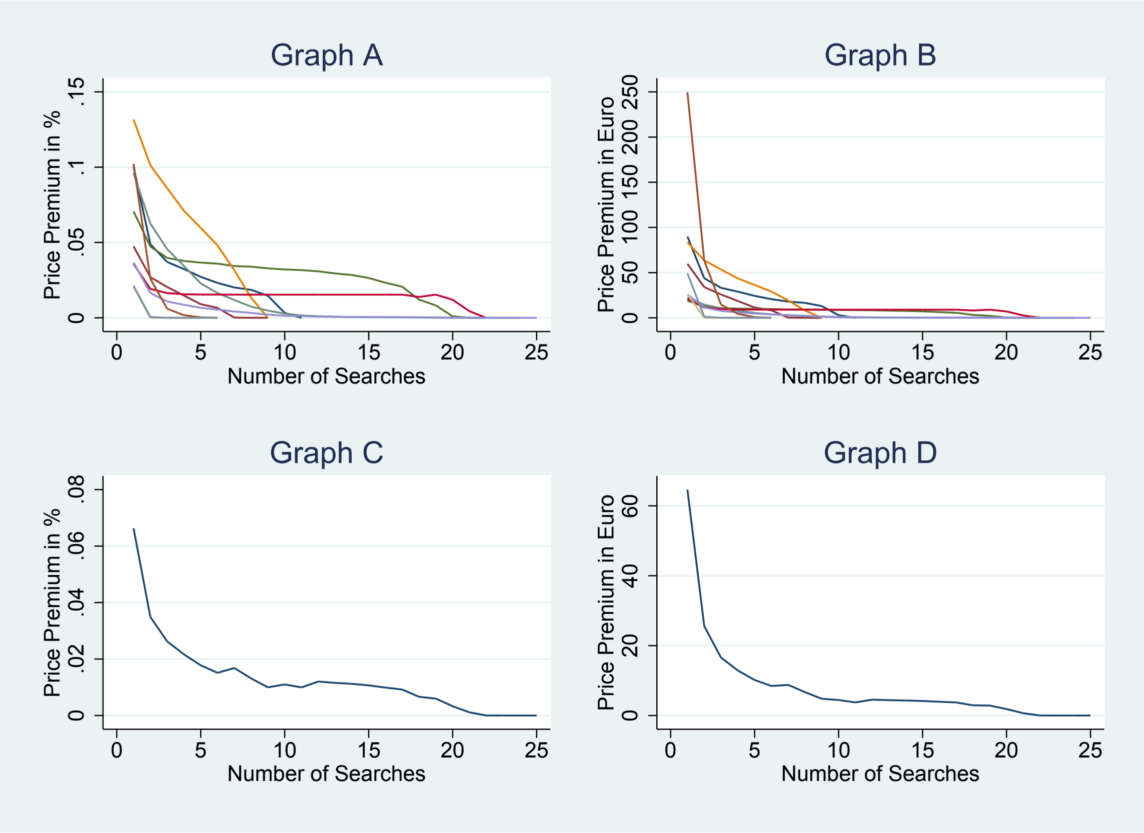
Source: *Simulations on collected data from www.guenstiger.de and www.similarweb.com*

only engage in little search, as moderate search does not necessarily provide significant gains with the underlying price distributions. Consumers with low search costs on the other hand also form their expectations accordingly to the fact that the payoff with moderate search is uncertain and approaching the maximum with high search. With this it is most beneficial for them to engage in high intensity search.

For comparison to previous papers, I also provide the value of information measured by the average market price in relation to the minimal market price, as used by Baye et al. (2003) and Pan et al. (2003). As these are simple computations, I calculated these values over all observations. The results are shown in Table 6, both for the average for all products as well as for separate product categories.

The left hand side of the table shows the absolute and relative differences between

Figure 9: Simulation Results Televisions



Notes: *Graph A shows the relative price premium in percent versus the number of searches; Graph B shows the absolute price premium in percent versus the number of searches; Graph C shows the mean relative price premium in percent versus the number of searches; Graph D shows the mean absolute price premium in percent versus the number of searches.*

Source: *Simulations on collected data from www.guenstiger.de and www.similarweb.com*

average market price and minimal market price averaged across products and days, while the right side of the table shows the absolute and relative differences between the weighted average price and minimal price. The traffic estimates were used as weights for the latter to model the relative size and importance of the sellers offering in the respective markets. The relative (absolute) value of information across all observations is 14.7% (53.10 Euro). This value lies very close to the results of Pan et al. (2003), who measure a value of information in the range of 13.1% to 14.5%, and Baye et al. (2003), who find the value of information to be 15.9% in their sample. When weighting the prices with the traffic estimates the value of information is lower with a relative value of 9.3% and absolute value of 42.41 Euro. A possible explanation to this is that more sampled sellers have lower prices, possibly resulting from consumers learning which sellers tend to offer lower prices. Another explanation is that firms with lower costs have more market power and market presence as they are able

Table 6: Results Value of Information

	N	unweighted rel. Value of Inf.	unweighted abs. Value of Inf.	weighted rel. Value of Inf.	weighted abs Value of Inf.
All Products	5796	14.7%	53.10 Euro	9.3%	42.41 Euro
Televisions	1932	10.6%	100.50 Euro	8.0%	88.82 Euro
Hard Drives	1904	15.5%	21.31 Euro	8.2%	9.16 Euro
Printers	1960	17.9%	37.26 Euro	11.8%	28.96 Euro

Notes: *This table displays the unweighted value of information as described by the relative and absolute difference between the average and lowest market price. The weighted value of information is calculated by computing . The table shows the averages across observations.*
Source: *Own calculations on collected data from www.guenstiger.de and www.similarweb.com*

to offer lower prices.

Table 5 shows that the randomized subsample the simulations are fairly similar to those of the total sample. The weighted value of information in Table 6 is calculated in the same way as the price premium for a consumer only searching for one price in the simulations. In the simulations, the printers show the same value of 11.8%, while the televisions have a slighter lower value of 6.6% in simulations compared to 8% in the table, and hard drives have a higher value of 9.2% in the simulations compared to 8.2% in the table. These values are also reasonably close to the simulation results, meaning that the averaged results displayed in the graphs can be seen as representative of the whole sample. Yet as already discussed in Section 5.2, the results of these simulations can possibly not be generalized to markets of other product categories, as only electronics products were examined.

7 Conclusion

In this thesis, I measure the degree of price dispersion in online markets, and examine relating factors and the value of information to consumers. For this, I collect online retail price data of 207 homogeneous electronics products over the course of four weeks from the German price comparison website www.guenstiger.de. I first measure the degree of price dispersion in the individual markets, which serves as a measure for market inefficiency. I find a distinct dispersion of prices showing that online markets are still characterized by informational inefficiencies.

Moreover, I run regressions to examine the relationship of price dispersion with the number of sellers, and with the average price level in the market. Contrary to previous findings, I find no evidence for there existing a relationship between the seller number and price dispersion. The average market price however shows a significant negative relationship with price dispersion, which is consistent with previous research. A possible theoretical explanation is that consumers individual search costs are constant across price segments of different products, meaning that they engage in more search for higher priced products. This in turn reduces the ability of firms to discriminate with prices, deflating the price dispersion.

Lastly, I run simulations to obtain the value of information to consumers depending on the number of searches. To simulate the behavior of consumers, I utilize website traffic estimate data to model the relative importance of online shops in the underlying markets and use several search theoretic assumptions to model consumer behavior. Assuming that consumers have sophisticated expectations of market prices, the results suggests that they can roughly be grouped into two search intensities. One group of consumers engages in low intensity search, and on average pays a premium on the product, while the other group engages in high intensity search and on average pays the lowest price in the market. I also compute the value of information with the same method as in previous research and find comparable results.

In conclusion, this thesis illustrates that in 2017, online markets show informational inefficiencies comparable to those measured during the dawn of the e-commerce era. Search costs to consumers seem to be significant even in a market setting with superior information provision when compared to conventional markets. The common conjecture in previous research that these inefficiencies may be resolved as the markets mature can therefore not be supported by my analysis.

Limitations to my approach include the possible unobserved heterogeneity of online retailers, that may have a significant explanatory power to price dispersion. The measured price dispersion and effects may therefore be slightly overstated. Previous literature has however not found measurable characteristics that manage to explain price dispersion. Furthermore, only online prices are included, although consumers may also consider the prices at conventional stores in their search behavior. A possible way to overcome theses problem could be to utilize data on consumer preferences in online shopping to better understand purchasing decisions. The examined data is composed of consumer electronics product, which means that these results may not be generally representative of all online retail markets. A larger dataset with more product categories would be a solution to this issue.

References

Baye, M. R., Morgan, J. & Scholten, P. (2003), 'The Value of Information in an Online Consumer Electronics Market', *Journal of Public Policy & Marketing* **22**(1), 17–25.

Baye, M. R., Morgan, J. & Scholten, P. (2004*a*), 'Price Dispersion in the Small and in the Large: Evidence from an Internet Price Comparison Site', *The Journal of Industrial Economics* **52**(4), 463–496.

Baye, M. R., Morgan, J. & Scholten, P. (2004*b*), 'Temporal price dispersion: Evidence from an Online Consumer Electronics Market', *Journal of Interactive Marketing* **18**(4), 101–115.

Baye, M. R., Morgan, J., Scholten, P. et al. (2006), 'Information, Search, and Price Dispersion', *Handbook on Economics and Information Systems* **1**, 323–77.

Brynjolfsson, E. & Smith, M. D. (2000), 'Frictionless Commerce? A Comparison of Internet and Conventional Retailers', *Management Science* **46**(4), 563–585.

Cameron, A. C. & Miller, D. L. (2015), 'A Practitioner's Guide to Cluster-Robust Inference', *Journal of Human Resources* **50**(2), 317–372.

Carlson, J. A. & McAfee, R. P. (1982), 'Discrete Equilibrium Price Dispersion: Extensions and Technical Details'.

Carlson, J. A. & McAfee, R. P. (1983), 'Discrete Equilibrium Price Dispersion', *Journal of Political Economy* **91**(3), 480–493.

Clay, K., Krishnan, R. & Wolff, E. (2001), 'Prices and Price Dispersion on the Web: Evidence from the Online Book Industry', *The Journal of Industrial Economics* **49**(4), 521–539.

Clay, K., Krishnan, R., Wolff, E. & Fernandes, D. (2002), 'Retail Strategies on the Web: Price and Non-Price Competition in the Online Book Industry', *The Journal of Industrial Economics* **50**(3), 351–367.

Ellison, G. & Ellison, S. (2004), 'Industrial Organization: Lessons from the Internet', *Citeseer* .

Ellison, G. & Ellison, S. F. (2009), 'Search, Obfuscation, and Price Elasticities on the Internet', *Econometrica* **77**(2), 427–452.

Eurostat (2016), 'ICT Usage in Households and by Individuals, Individuals who ordered Goods or Services over the Internet for Private Use (isoc_r_blt12_i)'. Accessed: 2017-04-17.

URL: *http://ec.europa.eu/eurostat/web/digital-economy-and-society/data/ database*

Levin, J. D. (2011), 'The Economics of Internet Markets', *NBER Working Paper Series* .
URL: *http://www.nber.org/papers/w16852*

Moraga-González, J. L., Sándor, Z. & Wildenbeest, M. R. (2016), 'Nonsequential Search Equilibrium with Search Cost Heterogeneity', *International Journal of Industrial Organization* .

Pan, X., Ratchford, B. T. & Shankar, V. (2003), The Evolution of Price Dispersion in Internet Retail Markets, *in* 'Organizing the New Industrial Economy', Emerald Group Publishing Limited, pp. 85–105.

Pan, X., Ratchford, B. T. & Shankar, V. (2004), 'Price Dispersion on the Internet: A Review and Directions for Future Research', *Journal of Interactive Marketing* **18**(4), 116–135.

Ratchford, B. T., Pan, X. & Shankar, V. (2003), 'On the Efficiency of Internet Markets for Consumer Goods', *Journal of Public Policy & Marketing* **22**(1), 4–16.

Stigler, G. J. (1961), 'The Economics of Information', *Journal of Political Economy* **69**(3), 213–225.

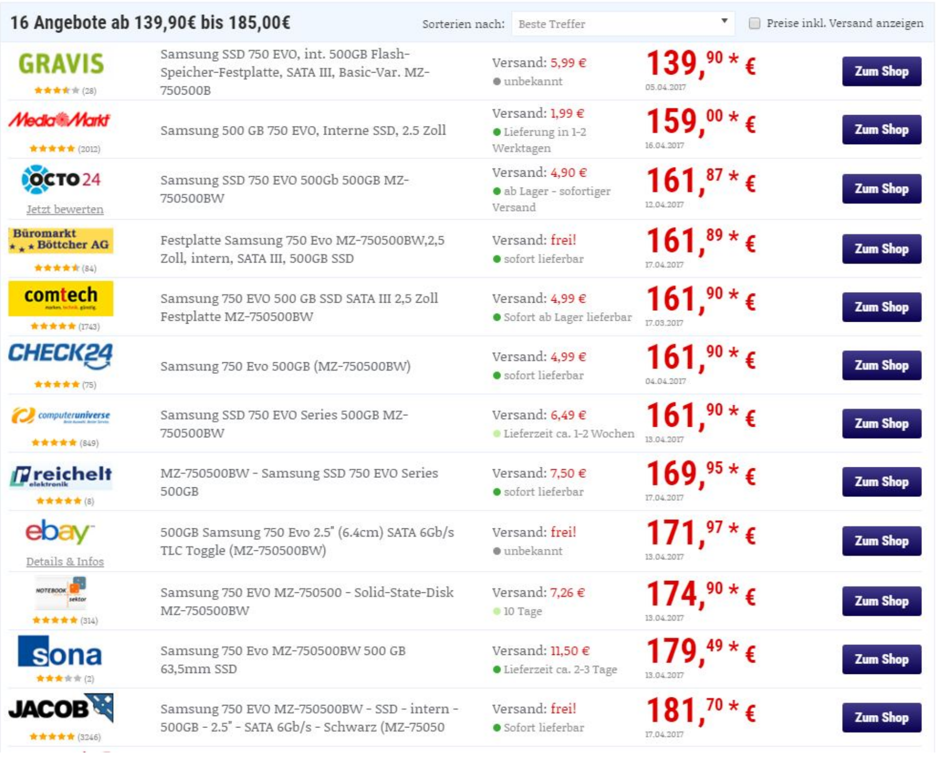
United States Census Bureau (2016), 'Latest quarterly e-commerce report'. Accessed: 2017-04-17.
URL: *https://www.census.gov/retail/index.html#ecommerce*

Appendices

A Appendix - Data Collection and Cleaning

This appendix provides additional information on the data collection and data cleaning process. Figure 10 shows an example product web page from the shop bot I collected my data from.

Figure 10: Screenshot from Price Comparison Website www.Guenstiger.de



16 Angebote ab 139,90€ bis 185,00€ Sorterien nach: Beste Treffer ▾ □ Preise inkl. Versand anzeigen

GRAVIS ★★★★☆ (28)	Samsung SSD 750 EVO, int. 500GB Flash-Speicher-Festplatte, SATA III, Basic-Var. MZ-750500B	Versand: 5,99 € ● unbekannt	$139,^{90}$ * € 05.04.2017	Zum Shop
MediaMarkt ★★★★★ (2012)	Samsung 500 GB 750 EVO, Interne SSD, 2.5 Zoll	Versand: 1,99 € ● Lieferung in 1-2 Werktagen	$159,^{00}$ * € 16.04.2017	Zum Shop
OCTO24 Jetzt bewerten	Samsung SSD 750 EVO 500Gb 500GB MZ-750500BW	Versand: 4,90 € ● ab Lager - sofortiger Versand	$161,^{87}$ * € 12.04.2017	Zum Shop
Büromarkt Böttcher AG ★★★★★ (84)	Festplatte Samsung 750 Evo MZ-750500BW,2,5 Zoll, intern, SATA III, 500GB SSD	Versand: frei! ● sofort lieferbar	$161,^{89}$ * € 17.04.2017	Zum Shop
comtech ★★★★★ (1743)	Samsung 750 EVO 500 GB SSD SATA III 2,5 Zoll Festplatte MZ-750500BW	Versand: 4,99 € ● Sofort ab Lager lieferbar	$161,^{90}$ * € 17.03.2017	Zum Shop
CHECK24 ★★★★★ (75)	Samsung 750 Evo 500GB (MZ-750500BW)	Versand: 4,99 € ● sofort lieferbar	$161,^{90}$ * € 04.04.2017	Zum Shop
computeruniverse ★★★★★ (849)	Samsung SSD 750 EVO Series 500GB MZ-750500BW	Versand: 6,49 € ● Lieferzeit ca. 1-2 Wochen	$161,^{90}$ * € 13.04.2017	Zum Shop
reichelt elektronik ★★★★★ (8)	MZ-750500BW - Samsung SSD 750 EVO Series 500GB	Versand: 7,50 € ● sofort lieferbar	$169,^{95}$ * € 17.04.2017	Zum Shop
ebay Details & Infos	500GB Samsung 750 Evo 2.5" (6.4cm) SATA 6Gb/s TLC Toggle (MZ-750500BW)	Versand: frei! ● unbekannt	$171,^{97}$ * € 13.04.2017	Zum Shop
NOTEBOOK sektor ★★★★★ (314)	Samsung 750 EVO MZ-750500 - Solid-State-Disk MZ-750500BW	Versand: 7,26 € ● 10 Tage	$174,^{90}$ * € 13.04.2017	Zum Shop
sona ★★★☆☆ (2)	Samsung 750 Evo MZ-750500BW 500 GB 63,5mm SSD	Versand: 11,50 € ● Lieferzeit ca. 2-3 Tage	$179,^{49}$ * € 13.04.2017	Zum Shop
JACOB ★★★★★ (3246)	Samsung 750 EVO MZ-750500BW - SSD - intern - 500GB - 2.5" - SATA 6Gb/s - Schwarz (MZ-75050	Versand: frei! ● Sofort lieferbar	$181,^{70}$ * € 17.04.2017	Zum Shop

To ensure that my approach was valid, I needed to make sure the sample consisted of perfectly homogeneous products. Usually shopbots have a unique web page for each product, yet this is not necessarily the case for certain products. For instance, smartphones and tablets may come in different colors and these product versions appeared to be mixed together in many cases. Another source of heterogeneity is bundling: an example are cameras, where it is common that SLR bodies were offered as a bundle with a lens. Furthermore, products where special editions or second generations are included in the listings are problematic. In test runs previous to the actual data collection I examined whether the chosen products suffered from any of the mentioned problems. In the case of there being a problem, I replaced the respective product with another randomly drawn product. Three products were excluded from the dataset ex post, as they showed flaws in the data.

To collect the data, Parsehub's web crawlers iterate through a specified list of web pages.

Each web page of the price comparison web page displays a maximum of 30 seller prices ranked by price, meaning that products with more than 30 seller listings were spread out over multiple URLs. In practice this means that consumers have to navigate to a second or even third page to see more higher priced seller listings. As the number of listed sellers varied from day to day, it was necessary to for me adjust the URL list for the web crawlers on a daily basis.

To automate the data cleaning process, I created a tool in Excel and Visual Basic. With this, I performed several steps on a daily basis. First, the tool deleted all price quotes from the auction platform *Ebay.de*, as Ebay offers are mostly not factory new goods and not necessarily sold by commercial sellers. Next, the tool identified duplicate listings and unwanted price quotes, usually caused by multiple listings of a company for one product. I then manually compared all duplicate listings on a daily basis and subsequently deleted faulty price listings. The duplicate listings were mostly caused by wrong products being listed, reduced prices for faulty items or bundled products. If products were identical and offered twice in the webshop, the lower price was kept. Furthermore, I deleted all offers from eight sellers that started listing prices on the price comparison website after I started collected the data, as they would have caused an exogenous change in the number of sellers that does not stem from the relationship to the price dispersion.

B Appendix - Simulations

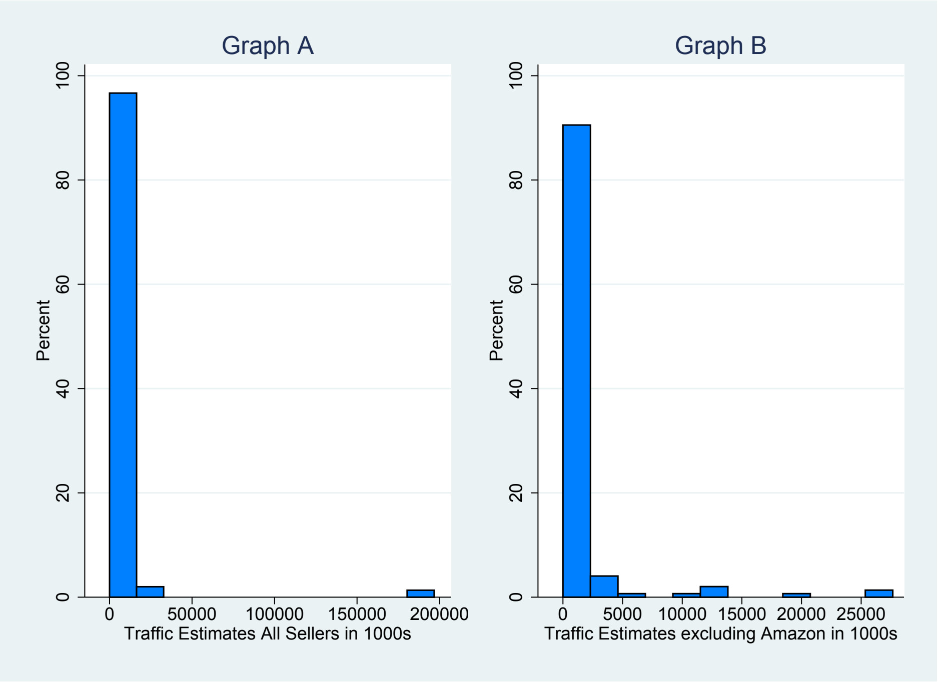
B.1 Data Description SimilarWeb

This appendix give additional information about the data used from SimilarWeb.

For 8 of the 150 sellers, no data was available on SimilarWeb. These sellers were then excluded from the simulations. For international sellers such as Amazon or Mediamarkt, the traffic for the German website domain was used. The seller Apple only operates an international *.com*-Domain with a traffic estimate for all users globally. I approximated German users to make up 5 percent of traffic, based on the fact that French and British users make up 4.93% and 5.55% of Apple.com traffic on SimilarWeb respectively. Furthermore, my data lists price quotes by Amazon and the lowest offer in the Amazon Marketplace as two separate sellers. The Amazon Marketplace is a third party seller platform within the Amazon online shop. The traffic estimates for Amazon.de were split evenly between these two sellers.

Figure 11 displays the traffic estimates in thousands for the sellers in the market. The large outlier in Graph A is Amazon.de with the by far largest traffic estimate. Graph B excludes this seller and shows that the majority of the 150 sellers only have relatively small traffic numbers, i.e. the market is characterized by a small number of large sellers and a large share of small sellers.

Figure 11: Histograms of Seller Traffic Estimates



Notes: *Graph A shows the distribution of traffic estimates for all sellers in the sample. Graph B shows the distribution of traffic estimates for all sellers excluding Amazon.de.*
Source: *Collected data from www.similarweb.com*

B.2 Monte Carlo Simulation Code

I wrote the following code sequence to simulate my in section 6.1 described approach. The program used is Stata. The percentage symbols indicate where information, such as the used file paths, has to be specified.

```
ssc install SAMPLEPPS, replace

global category "%% PRODUCT CATEGORY ACRONYM %%"

clear
set obs 1
gen searches=.
gen product="."
```

```
save "alldata $category.dta”, replace


global no_products 10
global no_days 10

clear
set more off
global datapath "%% FILEPATH %% / %% FILE NAME %% .dta”
global dataoutput "%% FILEPATH %%”
use "$datapath”, replace


//sample products
preserve

        collapse price_ship, by(product)
        sample $no_products, count
        drop price_ship
        save "products $category.dta”, replace


        levelsof product, local(productsample)
        local j=1

        foreach i of local productsample{

        global product‘j’ "‘i’”

        local ++j
        }

restore

//sample dates
preserve

        collapse price_ship, by(date)
        sample $no_days, count
        drop price_ship
        save "dates $category.dta”, replace

        levelsof date, local(datesample)
        local j=1

        foreach i of local datesample{
```

```
        global date‘j’ ‘i’

        local ++j
        }

restore

//define monte carlo program
capture program drop mce_search
program define mce_search, rclass

use ”$datapath”, replace

keep if product==”$productname” & date==$day

sum price_ship, meanonly
return scalar min_tot=r(min)

forvalues i=1/$marketsize {

        gen allpicks‘i’=0

        preserve

        forvalues j=1/‘i’ {

        samplepps pick‘j’, size(weights) n(1)

        replace allpicks‘i’=allpicks‘i’ + pick‘j’

        replace weights=0 if pick‘j’==1
        drop pick‘j’
        }

        gen allprices‘i’=allpicks‘i’*price_ship

        sum allprices‘i’ if allprices‘i’>0
        return scalar minprice‘i’ = r(min)

        restore
}

clear

end
```

```
//run loop
forvalues i=1/$no_products {

        global productname ”${product‘i’}”

        clear
        set obs 1
        gen searches=.
        gen product=”.”

        save ”$dataoutput\\$productname.dta”, replace


        forvalues l=1/$no_days {

        use ”$datapath”, replace
        global day ${date‘l’}

        keep if product==”$productname” & date==$day

        drop if weights==0
        qui sum sellers , meanonly
        global marketsize = r(N) − 1
        global marketmax = r(N)
        local difflist1 ”min1=r(minprice1)”
        local difflist2 ”diffabs1 diffrel1”

        forvalues k=2/$marketsize {

                local difflist1 ///
        ”‘difflist1’ min‘k’=r(minprice‘k’)”
                local difflist2 ///
        ”‘difflist2’ diffabs‘k’ diffrel‘k’”
                }
                local difflist2 ”‘difflist2’ ///
        diffabs$marketmax diffrel$marketmax”


                simulate market_min=r(min_tot) ‘difflist1’, ///
                reps(200) nodots: mce_search

        forvalues j=1/$marketsize {

                gen diffabs‘j’= min‘j’ − market_min
                gen diffrel‘j’= (min‘j’ − market_min)/market_min
                }
                gen diffrel$marketmax = 0
                gen diffabs$marketmax = 0

                collapse ‘difflist2’

                xpose, clear varname

                gen searches=0
                gen product=”$productname”
                gen diffabs=0
                gen diffrel=0

        forvalues k=1/$marketmax {

                replace searches=‘k’ if _varname==”diffrel‘k’”
                replace searches=‘k’ if _varname==”diffabs‘k’”
                replace diffabs=v1 if _varname==”diffabs‘k’”
                replace diffrel=v1 if _varname==”diffrel‘k’”

        }

                collapse (max) diffabs diffrel, by(searches product)

                append using ”$dataoutput\\$productname.dta”
                save ”$dataoutput\\$productname.dta”, replace
        }

        use ”$dataoutput\\$productname.dta”, replace

        collapse diffabs diffrel, by(searches product)


        append using ”alldata $category.dta”
        save ”alldata $category.dta”, replace

}
```

B.3 Descriptive Statistics of Simulation Sample

Table 7: Overview of Products used in Simulation

Televisions	**# Sellers**	**∅ -Price**	**σ - Price**	**Coef. of Var.**
LG_Electronics_60UH605V	10.9	978.24	66.55	.0680
LG_Electronics_65UH625V	8.7	1,333.57	79.18	.0593
Panasonic_TX_24DSW504	21.3	295.03	15.41	.0524
Philips_43PUS6501	9.6	723.23	46.60	.0649
Samsung_UE32J4570	24.2	277.46	47.71	.1715
Samsung_UE49K6379	21.7	647.62	60.40	.0932
Samsung_UE55K5589	24.9	762.49	76.47	.1002
Samsung_UE55KS7000_2	6.0	1,288.49	33.13	.0257
Samsung_UE65KS9000	5.6	2,516.67	101.31	.0401
Sony_KDL_75W855C	4.2	2,680.45	487.30	.7834
Total	13.71	1,150.33	101.41	.0854
Hard Drives	**# Sellers**	**∅ -Price**	**σ - Price**	**Coef. of Var.**
Canon_Connect_Station_CS100	17.5	183.81	18.10	.0985
Intel_SSD_600p_256GB_SSDPEKKW256G7X1	17.6	105.99	5.44	.0514
Samsung_Portable_SSD_T3_500GB_MU_PT500B	24.0	216.14	15.94	.0737
Samsung_SSD_850_EVO_500GB_MZ_M5E500BW	25.1	187.49	10.85	.0579
Samsung_SSD_850_PRO_128GB_MZ_7KE128BW	22.2	125.68	18.78	.1491
Samsung_SSD_960_EVO_500GB_MZ_V6E500BW	21.6	272.97	13.09	.0479
Seagate_SkyHawk_10TB_ST10000VX0004	26.2	438.62	18.22	.0415
Verbatim_Store_n_Save_4TB	16.8	152.48	22.71	.1489
Western_Digital_2_WD_Blue_SSD_M_2_500GB	24.5	166.76	9.24	.0554
Western_Digital_2_WD_Red_2TB_WD20EFRX	49.4	105.27	6.70	.0636
Total	24.49	195.52	13.91	.0788
Printers	**# Sellers**	**∅ -Price**	**σ - Price**	**Coef. of Var.**
Brother_DCP_J562DW	37.3	119.31	10.85	.0910
Brother_DCP_L2560DW	40.6	238.01	22.04	.0925
Canon_Pixma_MG2950	11.9	66.013	13.94	.2110
Epson_C11CF50403	30.9	120.52	8.86	.0734
Flashforge_Finder_3D	8.3	496.58	72.45	.1434
HP_Color_LaserJet_Enterprise_M553dn	36.5	533.71	47.04	.0882
HP_DeskJet_3630 d	29.7	65.74	9.99	.1519
HP_OfficeJet_5740	28.6	111.74	14.61	.1308
HP_OfficeJet_Pro_6970	40.6	152.05	13.32	.0875
Samsung_Xpress_C430	31.3	142.96	13.02	.0911
Total	29.57	204.66	22.61	.1161

Notes: *This table provides summary statistics for the examined products in the simulation sample in Section 6. The measures reflect the mean values for the 10 sampled products and 10 sampled days in each product category. Displayed are the number of sellers (# Sellers), the average market price (∅ -Price), the standard deviation of the prices (σ -Price) and the coefficient of variation of prices in the respective product markets.*
Source: *Own calculations on collected data from www.guenstiger.de*

Table 8: Overview of Dates used in Simulation

Televisions	**Hard Drives**	**Printers**
19.01.2017	19.01.2017	17.01.2017
21.01.2017	30.01.2017	21.01.2017
24.01.2017	21.01.2017	22.01.2017
27.01.2017	24.01.2017	26.01.2017
31.01.2017	27.01.2017	27.01.2017
01.02.2017	30.01.2017	31.01.2017
04.02.2017	01.02.2017	01.02.2017
05.02.2017	04.02.2017	02.02.2017
08.02.2017	06.02.2017	08.02.2017
12.02.2017	08.02.2017	13.02.2017

Notes: *This table shows the days of collected data used for the simulations in Section 6.*
Source: *Collected data from www.guenstiger.de*

B.4 Numerical Simulation Results

Table 9: Numerical Simulation Results Televisions

Searches	**Absolute Premium in Euro**	**Relative Premium**
1	64.65	.06633
2	25.60	.03493
3	16.57	.02624
4	12.89	.02165
5	10.17	.01779
6	8.438	.01510
7	8.76	.01682
8	6.69	.01313
9	4.79	.00997
10	4.44	.01099
11	3.75	.00996
12	4.54	.01202
13	4.40	.01156
14	4.30	.01122
15	4.16	.01069
16	3.94	.00990
17	3.75	.00920
18	2.92	.00663
19	2.86	.00599
20	1.86	.00330
21	.66	.00112
22	.0072	.00001
23	.0014	.000002
24	0	0
25	0	0

Notes: *Averaged simulation results of randomized subsample of televisions across all sampled products and days*

Source: *Simulations on collected data from www.guenstiger.de and www.similarweb.com*

Table 10: Numerical Simulation Results Hard Drives

Searches	**Absolute Premium in Euro**	**Relative Premium**
1	11.62	.09207
2	5.07	.04164
3	2.90	.02352
4	2.02	.01612
5	1.49	.01173
6	1.18	.00918
7	.92	.00698
8	.71	.00532
9	.56	.00418
10	.45	.00343
11	.39	.00290
12	.33	.00248
13	.29	.00210
14	.25	.00178
15	.21	.00136
16	.18	.00112
17	.15	.00087
18	.13	.00061
19	.113	.00045
20	.116	.00040
21	.139	.00041
22	.137	.00038
23	.155	.00039
24	.124	.00031
25	.171	.00042
26	0	0
...	...	...
51	0	0

Notes: *Averaged simulation results of randomized subsample of hard drives across all sampled products and days*

Source: *Simulations on collected data from www.guenstiger.de and www.similarweb.com*

Table 11: Numerical Simulation Results Printers

Searches	Absolute Premium in Euro	Relative Premium
1	24.53	.11883
2	10.85	.05902
3	6.74	.04046
4	6.04	.03466
5	5.63	.03136
6	5.22	.02847
7	4.87	.02590
8	4.53	.02308
9	4.27	.02139
10	3.92	.01791
11	3.56	.01416
12	3.70	.01383
13	3.45	.01207
14	3.20	.00999
15	2.97	.00891
16	3.22	.00939
17	3.03	.00862
18	2.92	.00807
19	2.73	.00736
20	2.63	.00695
21	2.47	.00651
22	2.34	.00612
23	2.18	.00567
24	2.04	.00527
25	1.82	.00465
26	1.66	.00418
27	1.38	.00344
28	1.23	.00299
29	.91	.00217
30	.70	.00166
31	.72	.00168
32	.47	.00109
33	.39	.00091
34	.45	.00105

Continued on next page

Table 11 – *Continued from previous page*

Searches	Absolute Premium in Euro	Relative Premium
35	.29	.00068
36	.22	.00052
37	.19	.00045
38	.27	.00061
39	.0023	.00001
40	.0005	.000002
41	0	0
42	0	0
43	0	0

Notes: *Averaged simulation results of randomized subsample of printers across all sampled products and days*

Source: *Simulations on collected data from www.guenstiger.de and www.similarweb.com*