Some Aspects of Internet Portal Market Competition

A Dissertation Presented

by

Volha Chuvakin

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Economics

Stony Brook University

December 2007

Stony Brook University

The Graduate School

VOLHA CHUVAKIN

We, the dissertation committee for the above candidate for the

Ph.D. degree, hereby recommend acceptance of this dissertation.

Hugo Benitez-Silva - Dissertation Supervisor

Associate Professor, Department of Economics

Mark R. Montgomery - Chairperson of Defense

Professor, Department of Economics

Wei Tan

Assistant Professor, Department of Economics

Thomas Sexton

Professor, College of Business

This dissertation is accepted by the Graduate School

Lawrence Martin

Dean of the Graduate School

Abstract of the Dissertation

Some Aspects of Internet Portal Market Competition

by

Volha Chuvakin

Doctor of Philosophy

in

Economics

Stony Brook University

2007

This dissertation investigates consumer behavior using the web portal industry as an empirical setting. Specifically, it explores the connection between features proposed by the portal and its success on the market and models the relationship between consumer characteristics and his online choices.

Two important questions concerning the online market competition are explored. First, the behavior of users on the Internet market is studied by examining the market shares of Internet portals, and establishing the connections between different portal characteristics and their attractiveness for people. Later, users' switching decisions are analyzed as a function of their own demographic characteristics and portal attributes using the survival analysis methods.

It was demonstrated that individual portal features such as Portal age, Mail and Search quality, are very important in explaining the overall market share, but less powerful in explaining the market shares of separate services. Although Mail and Search can be treated as major determinants of market shares: increase in Mail and Search quality can lead to an increase of market share for 5.6% and 4.4% respectively; the existence of Greetings, News service, Messenger and Weather service plays positive

role in forming the consumer preferences towards the portal, adding up to 5% to the number of existing customers. In addition, separate market shares for the most popular portal services are estimated and interconnections between them are analyzed. The results of the estimations point that market shares of search and services associated with virtual communities are determined not only by the overall quality of portal attributes, but also by the demographic characteristics of users, namely, Age and Education. It was also discovered that it is not the number of services, but the presence of high quality services improves the portal market share.

Survival analysis for the portal switching was introduced to further investigate the patterns of consumers' behavior on the online portal market. Logit probability estimation together with duration models in three different specifications is utilized to understand what factors lead to potential users' drop off.

Again, it was confirmed that the main factors contributing to the survival probability are existence of high quality portal services. Portal can raise the probability of survival by a factor of 1.22 by offering the high quality Mail service; existence of such portal features as Shopping, Finance, News, which increase the probability of survival by factors of 1.93, 1.64 and 1.21 respectively. Among the demographic characteristics, User age and Household size increase the probability of survival by 0.7% and 0.4%, and higher levels of user education reduces the probability of survival by 3.6%.

This dissertation is among the first to explore online consumer heterogeneity. The portal users' behavior is explored by dividing them into two groups based on the level of activity online. It is assumed that more active users, who surf the Internet intensively while switching constantly from site to site, will demonstrate different rates of portal drop-offs than regular users. This result is used to support the assumption that Internet users' population is heterogeneous and the behavior of different groups of users should be modeled separately. Kaplan-Meir model suggests that active users are attracted by the higher quality of services and have 25-35% higher proportion of switching. A question whether users from multiple member households produce higher switching rates is also introduced. Due to the nature of information spillover, it is hypothesized that increased number of switches will result from household information sharing. The work suggests the existence of such extra switching phenomenon.

The approach developed and applied in this dissertation can be effectively used for new, more detailed click stream data from today's portals. More importantly, the results can be used for the modern market of mobile portals, which is currently undergoing the stage equivalent to one explored in this research.

Table of contents

List of Figures viii
List of Tables ix
Acknowledgements xii
Chapter 1. Introduction. 1
Chapter 2. The web portal industry 4
Chapter 3. Literature review. 11
3.1. Information availability, switching costs and consumer behavior: theoretical background. 11
3.2. Online consumer behavior. 13
3.3. Portal competition. 15
3.4. What have we learned? 16
Chapter 4. Summary of contributions and policy implications. 19
4.1. Data and methodology. 19
4.2. Big multi-purpose portals vs. specialization. 21
4.3. Consumer heterogeneity. 21
Chapter 5. Data sources and description. 23
Chapter 6. Portal competition and the determinants of market shares. 30
6.1. Introduction. 30
6.2. Modeling the portal market share 31
6.2.1. Defining the market share for online portal 31
6.2.2. Data. 33
6.2.3. Conceptual and econometric model. 33
6.3. Empirical results. 36
6.3.1. General panel estimation. 36
6.3.2. Aggregated model estimation. 37
6.3.3. Estimation with lagged dependent variable. 38
6.3.4. Seemingly unrelated regression estimation. 40
6.3.5. Heckman selection estimation. 43
6.4. Conclusions. 45

Chapter 7. Analysis of switching behavior. 59
7.1. Introduction. 59
7.2. Duration model of switching. 62
7.2.1. Definition of hazard. 62
7.2.2. Data. 63
7.2.3. Conceptual and econometric model. 64
7.2.3.1. Logit model of probability of switching. 65
7.2.3.2. Cox's proportional hazard model. 66
7.2.3.3. Parametric estimations of hazard model. 67
7.3. Empirical Results. 68
7.3.1. Logit estimation of the probability of switching. 68
7.3.2. Cox's proportional hazard estimation. 69
7.3.3. Lognormal regression. 70
7.3.4. Loglogistic estimation. 71
7.3.5. Weibull hazard rate model. 72
7.4. Consumer heterogeneity. 73
7.4.1. Identification of different user types. 74
7.4.2. Heterogeneity test. 76
7.4.3. Kaplan-Meier estimation. 77
7.5. Knowledge spillover within the household. 78
7.6. Conclusions. 80
Chapter 8. Conclusions and policy implications 108
References 111
Appendix 117

List of Figures

Figure 2.1. Users' activity online. 7
Figure 2.2. Portal market shares for week ending May 13, 2006 7
Figure 7.1. Generic hazard rate. 81
Figure 7.2. Cox proportional hazard. 82
Figure 7.3. Lognormal estimated survival rates. 83
Figure 7.4. Loglogistic estimated survival rates. 84
Figure 7.5. Weibull estimated survival rates. 85
Figure 7.6. Indicators of user activity online. 86
Figure 7.7. Kaplan-Meier survival estimate 87
Figure 7.8. Kaplan-Meier survival estimates by user activity, Model I. 88
Figure 7.9. Kaplan-Meier survival estimates for multiple spells, Model I. 88
Figure 7.10. Kaplan-Meier survival estimates for single- and multiple-member households, Model I. 89

List of Tables

Table 2.1. Top 12 portals, share of market exceeds 1 %. 8
Table 2.2. Portal ranking and market share by service for week ending May 13, 2006.... 9
Table 2.3. Weekly market share of visits among all US web sites for week ending July 8, 2006..... 10
Table 2.4. Response time for second mover introducing various portal services..... 10
Table 5.1. Clickstream raw data sample. 25
Table 5.2. Summary statistics on number of visits and demographic characteristics of users. 26
Table 5.3. Summary of online visits per category for December 27, 1999 – March 31, 2000..... 27
Table 5.4. Summary of visits to portals for December 27, 1999 – March 31, 2000..... 29
Table 6.1. Correlation coefficients for the portal attributes. 47
Table 6.2 Definition of variables introduced into the model. 48
Table 6.3. Naïve estimation of market shares. 49
Table 6.4. General estimation coefficients. 50
Table 6.5. Estimation coefficients for grouped factors..... 51
Table 6.6. Estimation coefficients for grouped factors with lagged dependent variable. 51
Table 6.7. Estimation coefficients for seemingly unrelated regression..... 52
Table 6.8. Correlation matrix of residuals for SUR. 53
Table 6.9. Estimation coefficients for seemingly unrelated regression without Auction and Sport features 54
Table 6.10. Correlation matrix of residuals for SUR without Auction and Sport features. 55
Table 6.11. Heckman selection estimation. Selection into mail service..... 56
Table 6.12. Heckman selection estimation. Selection into virtual community 57
Table 6.13. Heckman selection estimation. Selection into search sevice..... 58
Table 7.1. The results of logit estimation, Model I..... 90
Table 7.2.Logit odds ratios, Model I..... 91

Table 7.3. The results of logit estimation, Model II. 92
Table 7.4.Logit odds ratios, Model II. 93
Table 7.5. The results of Cox proportional hazard estimation, Model I 94
Table 7.6. The results of Cox proportional hazard estimation, Model II..... 95
Table 7.7. Fit parametric survival models using lognormal underlying distribution, Model I..... 96
Table 7.8. Fit parametric survival models using lognormal underlying distribution, Model II 97
Table 7.9. Fit parametric survival models using loglogistic underlying distribution, Model I..... 98
Table 7.10. Fit parametric survival models using loglogistic underlying distribution, Model II 99
Table 7.11. Fit parametric survival models using Weibull underlying distribution, Model I 100
Table 7.12. Fit parametric survival models using Weibull underlying distribution, Model II..... 101
Table 7.13. Regression of activity indicator on household characteristics..... 102
Table 7.14. Log-rank test for equality of survivor functions by user activity, Model I. 103
Table 7.15. Wilcoxon (Breslow) test for equality of survivor functions by user activity, Model I..... 103
Table 7.16. Kaplan-Meier survival function comparison by user activity, Model I..... 103
Table 7.17. Log-rank test for equality of survivor functions by user activity, Model II. 104
Table 7.18. Wilcoxon (Breslow) test for equality of survivor functions by user activity, Model II. 104
Table 7.19. Kaplan-Meier survival function comparison by user activity, Model II. ... 104
Table 7.20. Log-rank test for equality of survivor functions of multiple spells estimation, Model I..... 105
Table 7.21. Wilcoxon (Breslow) test for equality of survivor functions of multiple spells estimation, Model I. 105
Table 7.22. Kaplan-Meier survival function comparison of multiple spells estimation,

Model I. 105
Table 7.23. Log-rank test for equality of survivor functions of multiple spells estimation, Model II. 106
Table 7.24. Wilcoxon (Breslow) test for equality of survivor functions of multiple spells estimation, Model II. 106
Table 7.25. Kaplan-Meier survival function comparison of multiple spells estimation, Model II. 106
Table 7.26. Log-rank test for equality of survivor functions of estimation for single- and multiple-member households, Model I. 107
Table 7.27. Wilcoxon (Breslow) test for equality of survivor functions of estimation for single- and multiple-member households, Model I. 107
Table 7.28. Kaplan-Meier survival function comparison of estimation for single- and multiple-member households, Model I. 107
Table A1. Collinearity diagnostics for portal attributes 117
Table A2. Eigenvalues and condition index for portal attributes. 117
Table A3. Collinearity diagnostics for demographic characteristics of users. 118
Table A4. Eigenvalues and condition index for demographic characteristics of users. 118

Acknowledgements

First, I would like to thank my advisor, Hugo Benitez-Silva, for the constant support through the evolution of this research, for his encouragement, support and advice, without which this work would not be completed. I wish to thank my committee members Mark Montgomery, Wei Tan and Thomas Sexton for their helpful comments and suggestions. Their aid in the development of this work is greatly appreciated.

Thanks are due to Jan-Dieter Spalink, Jeremy Stanley and Adrian Giles for their help in obtaining the data and in responding to data-related questions. I would also like to acknowledge Sangin Park and John Hause for their help at early stages of this work.

Finally, I am grateful to my husband, Anton, and all friends of mine, for their invaluable help and support throughout the years of my studies.

Chapter 1. Introduction.

One of the important characteristics of consumer behavior is the persistence of consumer choices. In the markets where customers are loyal, a firm's current position is an important determinant of its future profitability.

The Internet revolutionized how businesses and consumers interact between each other. It enabled new ways of gathering information about goods and changed the way these goods are sold. People use the Internet in the same way they make all other choices: determine needs, gather information about products, evaluate product choices, and assess their satisfaction level afterwards. With an increasing amount of Internet usage, it should come as no surprise that its impact on consumers has been heavy and will only continue at an increasing pace and thus approaching this problem with scientific methods will only grow more challenging.

In this dissertation, I would like to investigate the consumer behavior using the web portal industry as an empirical setting. The Internet market makes it hard to master the trust and loyalty, with millions of sites just a mouse click away, little keeps consumers from jumping from Yahoo! to Google if they are not fully satisfied. Brynjolfsson and Smith (2000) call the Internet "The Great Equalizer" because the technological capabilities of the medium reduce buyer search and switching costs and eliminate competitive advantages that retailers would enjoy in a physical marketplace. In this environment, attracting consumers and maintaining customer base remains one of the most important tasks for the portal management.

The number of players on a web portal market has grown from a lone portal in 1993 to several hundred at its peak in 2000 (Nie and Erbring (2000)) and then slowly declined. The competition among the top portals has been fierce during the dot-com boom and continues to be intense as of now.

Engaging into online competition, managers face two important questions: how to acquire new customers and how to retain the existing ones. They must decide under dynamic conditions which technological features must be included into a portal, how to improve the user experience on the website and what can be the additional ways to keep customers locked-in. Increasing competition on the online markets and rising users' expectations make this decision process even more complicated.

This study was conducted to address the above issues, to define the connection between features proposed by the portal and its success on the market, to model the relationship between consumer characteristics and his online choices.

Using data of online user behavior I will explore two important questions concerning the online market competition. First I use a panel dataset to analyze the market shares of the portals. The importance of market share is often stressed in the marketing literature (Fogg (1974), Buzell and Wiersema (1981), Cook (1985), Bridges, Yim and Briesch (1995)) but not well investigated in the context of the online industry (Gallaugher and Downing (2000)). Due to the specific revenue model of the web portal industry the customer base becomes the vital asset of the online firm. In this paper I analyze the general relationship between market shares of the online portals and their online features as well as variables representing the demographic characteristics of the consumers. I extend this analysis allowing the portals to compete on different online features separately and define the determinants of market leadership together with the set of most important online features.

In order to further investigate the consumer behavior on the Internet portal market and add dynamics into the model, I use survival analysis to follow user's switching between different portals. Switching is not uncommon on online markets (Brynjolfsson and Smith (2000), Chen and Hitt (2001), Bucklin and Sismeiro (2001)). Indeed, a portal is experience good, the user does not know in advance the utility level he will obtain from the visit to a given portal and he may try different alternative before choosing the best for him. I estimate the hazard model using different specifications to see how consumers respond to certain portal characteristics, what portal features attract them and lead to the switching. Several factors ranging from the quality of portal services to switching costs contribute to the declining hazard of switching. Finally, allowing for

consumer heterogeneity I explore the differences between hazard rates among different user groups.

The remainder of the dissertation is organized as follows. Chapter 2 presents the short overview of the portal industry and its unique characteristics. Chapter 3 reviews the literature on the problem of consumer behavior and highlights several approaches in the empirical research. Chapter 4 summarizes my contribution and policy implications. Chapter 5 discusses the data used in this dissertation. In Chapter 6 I present the market share model for the online portal industry. The impact of portal features and individual consumer characteristics on consumer switching behavior is analyzed in Chapter 7. The discussion of the results and conclusions are presented in Chapter 8.

Chapter 2. The web portal industry.

Portal is a term, generally synonymous with gateway, for a World Wide Web site that either proposes to be a major starting site for users when they get connected to the Web, or that users tend to visit as an anchor site.

The first web directory was created in 1993. This can be considered as a starting point of the web portal industry. At that time it was called "Jerry's Guide to the World Wide Web" but in 1994 it got a new name: Yahoo! Along with Yahoo, other Internet search engines and directories, like Altavista, Excite, Open Text, Magellan, Infoseek, and Lycos also became popular. All of them started as search engines or directories, but when they began experiencing page views numbering in the millions each day, most realized they could use their popularity by offering more features that would keep people at their sites once the users finished their initial search.

In the year 2000, which will be the center of our analysis, the Internet had millions of pages and several hundreds of portals. Only 10 to 15 of them could be considered as main portals, others were fringe. Portals offered a wide range of customization options and functionality including: Internet search and navigation; email; homestead[1]; customized news, weather, sports, and horoscopes; planners, calendars, and contact managers; bookmark managers; real-time chat and gaming; message boards; shopping; small business services; and much more.

Portals took the leading position on online markets. The most visited services at the time included e-mail service and different types of search services (Figure 2.1). The unbeatable leader among portals was Yahoo! with more than 30% of all visits followed by MSN, Excite and Netscape (Table 2.1).

Today there is one more contender for the portal market leadership. Google entered the portal space being a leader on the search engine market and its aggressive efforts led

[1] Homestead is a complete web site hosting, free or fee based.

him to its current third position, but were not sufficient to dislodge Yahoo! and MSN from their top spots (Table 2.2, Figure 2.2). According to the Forrester analyst Charlene Li, "Yahoo, AOL and MSN had millions of users before Google came along. Getting them to surrender their accounts, passwords and familiarity with the Web services they have long used will take more than just offering more e-mail storage space. Compelling services will win out in the end, however." (Reagan (2006)).

At the same time advances of the mobile connectivity led to a development of the separate mobile portal market currently shared by four players: Yahoo!, Google, AOL and MSN. The integration of the Internet, mobility and communications at the device creates a new set of business opportunities for portal companies. Portal market will continue to grow and rigorous scientific study of this growth is needed due to an ever-changing nature of the Internet phenomenon and its overall impact on the entire society.

An important feature of the web portal market is that any innovation introduced by a market agent can be easily implemented by competitors. Average time of implementation, according to Gallaugher and Downing (2000) is less than 2.5 months (Table 2.4).

Most of the portal services are offered to consumers free of charge. For portal companies it means that their competition is not in prices, but in qualities.

In addition, the Internet portal market has the property of reducing search and switching costs. All the groups of switching costs[2] (Klemperer (1987)) are significantly lowered in online markets:

- monetary transaction cost are low or non-existing for portal customers, they can easily switch between different portals without significant money loss;

[2] There may be transaction costs of switching between two almost identical services. Different banks may offer completely identical checking accounts, but there are monetary costs involved into closing the account with one bank and opening it with another. Similarly, it may be costly to switch between different cell phone providers, since customers are required to pay activation fees.
The learning required to use one brand may not be fully transferable to other brands of the similar products with the identical functionality. When consumer started to use the product of the particular brand he has the strong incentive to continue using this. For example when choosing a cake mix, it is easier for consumer to buy the brand he used before.
These two types of switching costs reflect real social cost of switching, although their magnitudes can be influenced by the firms. The third type - artificial or contractual costs arise entirely at firm's discretion. A good example is airline frequent flyer program rewarding the repeated travels with the same (or partner) company.

- learning cost is the most important source of costs online since learning is required to use portal features and may not be transferable to another web portal (Bucklin and Sismeiro (2001));
- artificial or contractual costs, researchers associate these type of costs with virtual communities of the particular portal (Gallaugher and Downing (2000)); by switching portals, a user may potentially lose contact with people who is unaware of his new address, in addition, he may lose friends made in this community, chat and game partners.

Low search and switching costs online allow users to combine services from different providers at no additional cost, the ability they cannot get on a conventional market.

These distinct features described above together with portal specific revenue model differentiate web portal market from all others existing in the economy.

The revenue model in the portal industry includes the following:

- advertising revenue – portal charges money for delivering audience to the advertiser:

  banner advertising -- x amount per 1,000 banner views;

  fees from advertisers or partner retailers who are "featured" on the main start page;

  keyword-based advertising on search engines;
- traffic revenue – charging “linked” content for each transaction driven through the portal;
- service revenue – charging customers for the access to premium portal services.

For the timeframe that will be covered in this dissertation the most important source of the portal revenue was the advertising revenue, i.e. portal earned most of its money when customers just visited the pages and were exposed to the ads featured on them.

Portals do not charge their customers directly for the basic services provided but all of the above sources of revenue are directly linked to the size of the customer base. A large number of customers have always been seen to be the goal of the “portal wars.” And the main question remains how to acquire customers when technology and business model can be easily imitated.

Figure 2.1. Users' activity online. (Source: Nie and Erbring (2000))

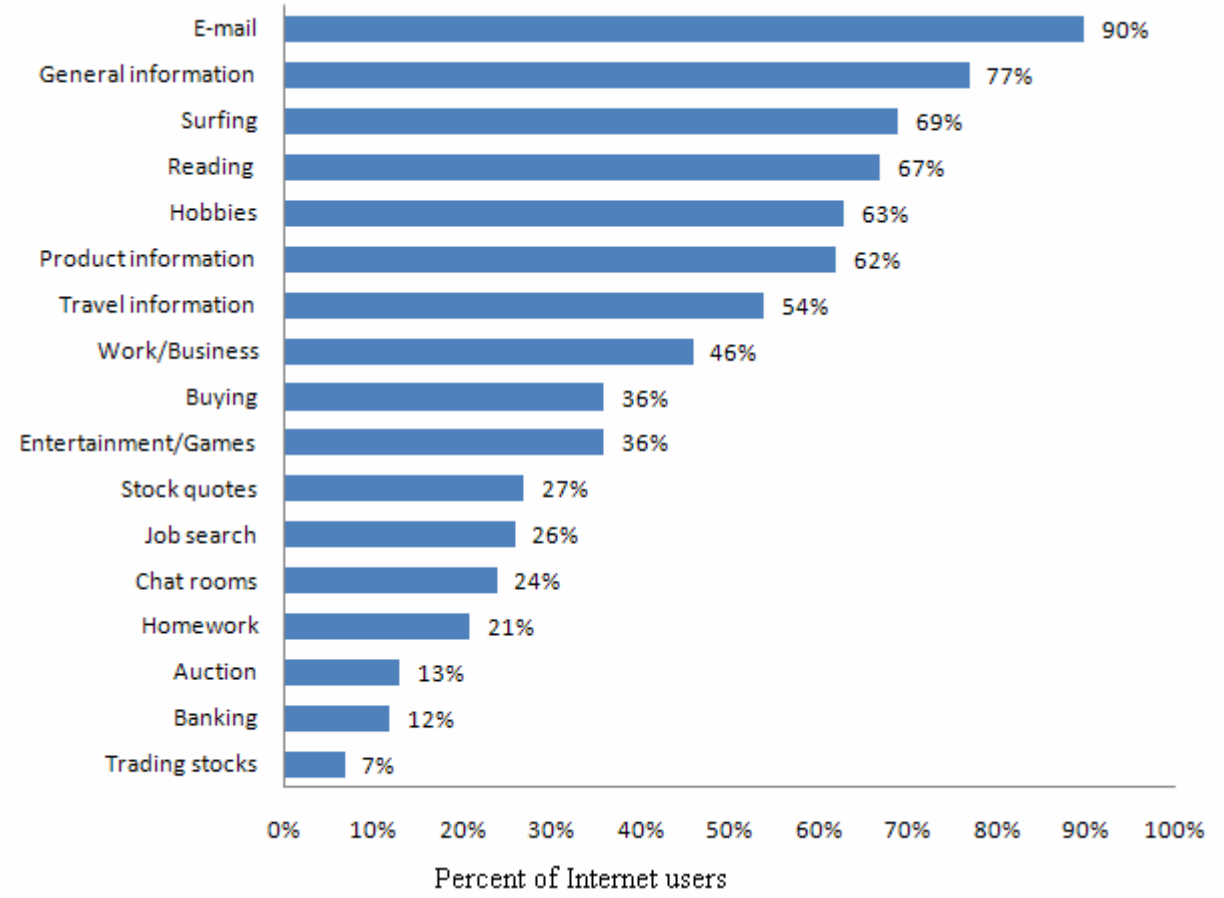
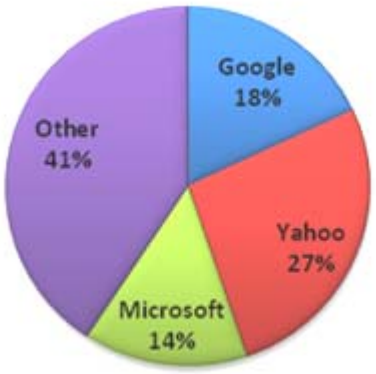


Figure 2.2. Portal market shares3 for week ending May 13, 2006 (Source: Hitwise).



[3] Percentages of total number of users

Table 2.1. Top 12 portals, share of market exceeds 1 %.

site_id	site_type	total visits	visits, %	total time	time, %
com.yahoo	Portal	378781	34.495	41237760	32.543
com.msn	Portal	207923	18.935	26014508	20.529
com.netscape	Portal	96310	8.771	10943132	8.636
com.aol	Portal	77650	7.071	7444956	5.875
com.go	Portal	75183	6.847	10808148	8.529
com.excite	Portal	71216	6.485	7594896	5.994
com.lycos	Portal	45091	4.106	5426206	4.282
com.altavista	Portal	36247	3.301	3950609	3.118
com.iwon	Portal	22646	2.062	3442158	2.716
com.myway	Portal	18381	1.674	2813050	2.219
com.hotbot	Portal	16084	1.465	1452725	1.146
com.snap	Portal	15210	1.385	1392590	1.099
Total			96.599		96.688

Table 2.2. Portal ranking and market share by service for week ending May 13, 2006 (Source: Hitwise).

Rank	Name	Market share in category (%)
Search engines (2322 sites, 7.3 of all Internet visits)		
1	Google	47.4
2	Yahoo! Search	16.0
3	MSN Search	11.5
E-mail services (1089 sites, 9.3 of all Internet visits)		
1	Yahoo! Mail	42.4
2	MSN Hotmail	22.9
3	MySpace – Mail	19.5
4	Gmail	2.54
News and Media (6080 sites, 3.4 of all Internet visits)		
1	Yahoo! News	6.3
2	Google News	1.9
* MSN News results appear within search.msn.com domain		
Business and Finance (1030 sites, 0.57 of all Internet visits)		
1	Yahoo! Finance	34.9
2	MSN Money Central	13.4
# 40	Google Finance	0.29
Travel – Maps (164 sites, 0.47 of all Internet visits)		
1	Mapquest	56.3
2	Yahoo! Maps	20.5
3	Google Maps	7.5
4	MSN Virtual Earth	4.3
5	Google Earth	2.0

Table 2.3. Weekly market share of visits among all US web sites for week ending July 8, 2006 (Source: Hitwise).

Rank	Name	Domain	Market share
1	MySpace	www.myspace.com	4.46
2	Yahoo Mail	mail.yahoo.com	4.42
3	Yahoo	www.yahoo.com	4.25
4	Google	www.google.com	3.89
5	MySpace – Mail	mail.myspace.com	2.85
6	MSN Hotmail	www.hotmail.com	2.39
7	MSN	www.msn.com	1.92
8	eBay	www.ebay.com	1.59
9	Yahoo Search	search.yahoo.com	1.36
10	MSN Search	search.msn.com	0.93

Table 2.4. Response time for second mover introducing various portal services. (Source: Gallaugher and Downing (2000)).

Service	Response time (months)
Auctions	3
Calendar	4
Chat	4
Classifieds	1
E-mail	3
Games	2
Shop	less than 1
Homestead	2

Chapter 3. Literature review.

This chapter summarizes the existing literature on consumer behavior on high-technological markets where information about goods and services is widely available and search and switching costs are low. First, I consider theoretical research related to the consumer behavior on such markets and discuss the prognosis about consumer choices. Then I review the empirical studies of the online consumer behavior. Finally, I discuss the limitations of the utilized approaches and methodologies used to explore online markets, to justify the approach and methods used in this study.

3.1. Information availability, switching costs and consumer behavior: theoretical background.

The question of consumer choice under different market conditions has received much attention in economic literature. The explosive growth of high-tech industry led the economic thought in a new direction. Based on the findings of game theorists (Nelson (1970), Beggs (1989), Wernerfelt(1991), Beggs and Klemperer (1992), Klemperer (1995)) a new set of literature has developed. All the above research suggests that presence of switching or search costs and lack of the information availability,

present certain companies with a source of market power. Group of researchers started to investigate what will be the market outcomes for newly developed “turbulent markets” where hundreds of goods are available, information is widespread and switching and search costs may be present. Nelson (1970) argues that company has market power on the market for experience goods and competition on such markets is reduced. Erdem and Keane (1996) use dynamic models of product choice behavior with uncertainty about the product attributes. Their results suggested that choices are based on the experience with certain brands as well as signals received from the outside; brand loyalty occurs from the low riskiness of familiar products, although reduces level of competition at equilibrium. Bakos (1997) explored how reducing search costs influence the electronic marketplaces. In contrast, his theory suggests that reduced search costs result in increased competition among sellers, increasing efficiency of electronic marketplaces, potentially leading to a substantial grows in the economic efficiency. Moshkin and Shachar (2000) examined this problem more deeply separating the effects of search costs and switching costs on the market outcome. The study shows that lack of information and high search costs can be the reason for brand persistence even without the presence of switching costs, and it is important to distinguish different sources of brand loyalty in order to improve market outcomes; information availability and lowering of search costs will lead to increased search and better equilibrium. Anand and Shachar (2000) later showed that consumer choices may be driven by other products of the same brand, and the influence of brand image is lower for more informed individuals. Bolton (1998) again studies the market for experience goods and discovered high level of dependence between customer satisfaction of the individual and duration of his stay with the same brand or service.

3.2. Online consumer behavior.

Unique qualities of online markets provided an excellent ground for different applications of the theoretical models. Thanks to the new rich source of data collected by Internet service providers, we can now examine consumer behavior, search and switching behavior and, more generally, main determinants of consumer movements online.

One of the first attempts to understand trends on online market were made by IT professionals in the late 1990s. In 1999, International Data Corporation and RelevantKnowledge conducted the research and found out that data suggests that users are not loyal to any one web portal. This research suggested that for rapidly changing markets it is hard to master true loyalty among customers. Future investigations of online markets disproved this statement. Later, Gallaugher and Downing (2000) focused on four major determinants of market leadership on the Internet market, that include first-mover advantage, brand effect, stickiness associated with switching costs and virtual communities. They first identified the specific features of online markets that distinguish them from physical marketplace and influence the market outcomes, such as first mover disadvantage, switching costs that are not associated with monetary costs; they also found first evidence of brand effect or brand loyalty on online market. These two papers did not propose the systematic approach to the analysis of Internet consumer behavior, but raised several important questions.

In the next few years, many researchers concentrated on different aspects of online consumer behavior and made several interesting observations. Even though search costs on online markets are low, the amount of search is less than predicted (Johnson et al (2001)) and several studies supported the hypothesis that brand loyalty may be observed on the online markets. Bucklin and Sismeiro (2001) applied Markov Chain Monte Carlo method to estimate a generalized tobit model and identified learning over multiple

sessions and within-site lock-in, supporting the existence of learning effect and brand loyalty effect for the Internet stores (Amazon and CDNow). Park and Fader (2002) used a multivariate timing model to explain browsing patterns for music (CDNow and Musicboulivard) and books (Amazon and Barnes & Noble). They found that consumers are generally brand loyal and their browsing pattern depends positively on the last site visited. Moe and Fader (2002) introduced nonstationarity in the searching behavior; they concluded that some form of evolution and learning occurs, supporting the hypothesis of persistence on the Internet market.

Each of these studies above restricts their analysis to only two websites at a time. It means that industry dynamics is missing from the analysis. In rapidly developing Internet environment (data used covers periods from March to October 1998 when Internet development skyrocketed), it may not be sufficient to concentrate on two online stores to analyze the searching behavior. Some surfers may use certain store less frequently as they are gaining experience from other places, and their behavior is influenced by quantity and quality of available online stores.

Miller (2007) reports the results of a recent study conducted by researchers at Penn State's College of Information Sciences and Technology. Taking the results from four different search engines: Google, MSN Live Search, Yahoo! and in-house non-branded engine, they compared customers' response. Despite the fact that the results were identical, the study participants picked Yahoo! as the most relevant, with Google and MSN trailing behind. This study again stresses the importance of "emotional branding" online.

Brynjolfsson and Smith (2000) empirically analyzed consumer behavior while using Internet shopbots and discovered that consumers usually use brand as a proxy of retailer's credibility and in general branded retailers are able to charge higher price than their rivals. At the same time, Lynch and Ariely (2000) found that lowering the search costs for quality information on online vineyards reduces the price sensitivity and produces welfare gains for consumers and sellers. Later, Waldfogel and Chen (2003) found evidence that information transparency increases competition and undermines brand attachment for Internet stores.

Chen and Hitt (2001) studied the online brokerage industry in order to measure and

describe switching costs. They found that system usage measure and quality are associated with decreased switching, and customer demographic characteristics have little influence on the switching process. Modeling the online media market Goldfarb (2002b) also finds that switching costs are present there and generate market power; publishers can earn extra profit because of the locked-in users.

Lack of consensus of existing empirical findings as well as numerous contradictions to the theoretical outcomes demonstrate that online markets are more complex than any theory could have suggested; and it will take serious expertise to understand the details of their operation.

3.3. Portal competition.

The literature on the competition among web portals is limited. A formal model was first presented by Gallaugher and Downing (2000). “Portal combat” was modeled as follows: market penetration or market reach was chosen as dependent variable, a linear function of portal age, brand, features that establish virtual communities and features that create switching costs. Analysis included only four big portals: Yahoo!, Lycos, Excite and Infoseek. Combining factors in different ways, Gallaugher and Downing conclude that not all featured variables have influence on market reach. Age and make effect was proven to be strong, games and chat suggest positive impact of virtual communities and switching costs on users’ stickiness. Other variables did not produce any significant influence.

The strong part of the paper is that most factors that may potentially influence portal

competition were highlighted in this study. Some limitations to this research come from the highly aggregated data; model also does not count for consumer preferences.

Goldfarb (2002a) utilizes the household-level data to estimate true state dependence or loyalty on Internet portals. Nested logit framework allows portals to be close substitutes, separating fringe portals and destination websites; the outside good is not using the portal. Using generalized extreme value distribution author estimates household-level regressions, controlling for heterogeneity across households. He finds the loyalty coefficient to be significant and robust to different model specifications, meaning that households tend to exhibit loyalty in their online choices, which suggests important implications for web portals. This work covers the entire portal market and uses demographic characteristics of the consumers. However, some limitations must be addressed. For the data collected at household level, we cannot distinguish between different household members and therefore, some conclusions about search activity may be misleading. Also, in the study, the objective portal characteristics have been ignored, that could lead to overstating the importance of brand effects.

3.4. What have we learned?

Theory provided some behavioral hypotheses that can be verified using the available data of online consumer activity. However, there is no clear way of approaching the problem of online market modeling; in fact, even market itself is not well defined.

Contemporary studies of online markets are concentrated at marketing companies (Media Metrix / comScore, Nielsen NetRatings, Hitwise), who own most of the data and

often publish online market ratings and brief analyses of the current market situation (Sullivan (2003), Tancer (2006a), Tancer (2006b), Regan (2006), Burns (2006), Reardon (2007), Ray (2007)). Not surprisingly, the publications lack the deep economic analysis of the described market changes.

Most of the previous research focused on certain aspects of consumer choice and specific features that may lead to change in consumer preferences. The limited nature of their findings, which does not allow to extrapolate results for other online markets, as well as lack of dynamics in some of them forms a gap that can be fulfilled. To provide a thorough analysis of online market, it is necessary to define the market properly, to identify the basic trends on this market, as well as to investigate its distinguished features from different standpoints. A portal is a complex good, offering more than one service at once, and users make their decisions based on more than one factor. At the same time, portal is an experience good, and users update their behavior based on previous visits to portals. Therefore, it will be beneficial to create a comprehensive model that allows to analyze all services featured by the portals, to identify the main factors that affect consumers' behavior and influence their decision to visit this portal again. In addition, this allows analyzing the interferences of different portal features.

Main findings that appear in almost all studies are the existence of brand loyalty and switching costs online. In presence of these factors, maintaining the strong position and keeping customers give a lot of gain. Keeping that in mind, I will model the market share of online portal for all portals available for the analysis; I will include both portal characteristics (Gallaugher and Downing (2000), Goldfarb (2002a)) and consumer characteristics (Goldfarb (2002a), Chen and Hitt (2001)) into the analysis in order to create a comprehensive model of competition in one of the biggest online markets – the web portal market.

Again, since portals are experienced goods, it is necessary to study them under dynamic conditions. This was stressed by Nelson (1970), Bolton (1998), Park and Fader (2002), and Moe and Fader (2002). I will apply hazard/survival analysis to examine the determinants of online consumers switching behavior and better understand how their choices are made.

The literature directly related to definitions and methodology used for estimating

market shares and hazard functions will be briefly reviewed prior to the estimations in chapters 6 and 7 respectively.

Chapter 4. Summary of contributions and policy implications.

This chapter outlines the main contributions of this dissertation. In addition, it discusses the hypotheses that are addressed in this study and policy implications that may be drawn based on the results.

4.1. Data and methodology.

It is well established in the literature that building and maintaining the large customer base is the most important goal of portal managers. Two sources of easy market share gain are discussed in the research:

- true brand loyalty;
- switching costs.

Although both of the above factors may impose barriers to switching, they produce completely different implications in terms of gaining and maintaining the customer base. In the presence of true brand loyalty on the online markets, portal's main goal transforms into simply attracting new loyal customers. At the same time, switching costs

can be created artificially for the existing customers preventing them from leaving for another portal.

Learning costs represent a part of the online switching costs. In previous studies, learning effect has been identified, but the connection between online learning and demographic characteristics of users was never established.

Using the appropriate dataset that includes not only the information about customer movements online and length of stays, but also detailed information about portal attributes and set of the consumer's demographic characteristics, this dissertation re-evaluates the previous findings related to brand loyalty and switching costs effects. I assume that by introducing detailed information about portal attributes I will be able to explain most of the customer retention.

Including the demographic characteristics will help to identify different sources of switching costs, equipping portal managers with important information about means for customer retention.

In my dissertation, I will study the portal market from different perspective. First, I analyze the behavior of users on the Internet market by exploring the shares of Internet portals, establishing connection between different portal characteristic and its attractiveness for people. Then, I study the users' decisions to drop from a portal as a function of their own demographic characteristics and portal attributes using the survival analysis methods. Such comprehensive analysis provides better insights for understanding the reasons that lead customers to and from certain portals.

4.2. Big multi-purpose portals vs. specialization.

Big universal portals top the list, but at the same time small portals, specializing in games or greetings exist on the same market, maintaining the stable market shares. Specialization often leads to better quality of supplied services and helps niche and regional portals to stay in good shape. This is why it is imperative to understand which of the following is more important: the individual contribution of portal services or the relationships between them; and which is better: to improve the existing services or generate new ones.

I will estimate separate market shares for the most popular portal services and analyze the interconnections between them. The results of such estimations will provide a better understanding of the importance of different portal attributes, which may help to improve the overall position on the market.

4.3. Consumer heterogeneity.

It is implicitly assumed in previous research that Internet users are homogenous. I want to extend the portal users' behavior by dividing them into two groups based on the level of activity online. I assume that more active users, who surf the Internet

intensively, switching constantly from site to site, will demonstrate different rates of portal drop-offs than regular users. If so, the different methods and techniques must be used for attracting and retaining different types of users.

Chapter 5. Data sources and description.

The data for this research was originally collected by Plurimus Corporation and is currently owned by Hitwise. This is a click stream household level data that keeps track of all customers' movements online (the example of raw dataset is presented in Table 5.1). The dataset covers 2654 households, for the period December 27, 1999 - March 31, 2000; for a total of 3,228,595 observations. This data allows us to analyze a relatively early stage of the portal market development since most of the portals came into being between 1996 and 1998.

The data collected was anonymous and does not allow to trace the records to any specific person, this allows to assume that data does not significantly suffer from the behavioral bias. Moreover, general trends in the data are similar to the results presented in US Internet usage surveys (Nie and Erbring (2000), Sullivan (2003)).

The data set provides detailed information about Internet usage: it contains the time of arrival at and departure from the destination website (with down to the second accuracy), website type and name, the number of pages viewed at the website, the number of bytes downloaded from and uploaded to the given website; it also indicates whether the portal page is user's starting page. Data also contains the demographic characteristics of users: gender, age, education, income, marital status, family size, geographic region (summarized in Table 5.2).

Variables reflecting different portal attributes, such as mailbox size, search quality, virtual community indicators and size, existence of weather, finance, news and other portal services were obtained from SearchEngineWatch.com, PCWorld.com and archive.org.

The data set, however, has some limitations. First, the geographic distributions of the samples are not representative. New York, Chicago and Los Angeles are under-

represented. This limitation cannot be considered to be a major problem since web portals are a national product and we can extend the results to different geographic regions. The second limitation of this data is that it does not contain information about users at work. Online behavior at work is very likely to be different from that at home. However, Nie and Erbring (2000) report that only 16.8% of Internet users use it primarily at work, 18.9% use it equally at home and at work and 64.4% use Internet mostly or only at home. This means that second limitation should not create major problem.

Using the full data set, we can summarize user habits in Internet usage. Summary of websites visits per category is presented in Table 5.2. All pages are divided into 88 different categories. Portal visits are leading in both number of visits (31.4% of all visits) and time spent on portals (about 26 % of all time spent online), followed by several news, communication and entertainment services.

Dataset includes 45 different portals visited by households in the sample. Yahoo! tops the portal list with almost 35% of all visits followed by MSN (18.9%), Netscape (8.7%) and Excite (6.4%) (Table 5.3). The number of leading portals with more than 1% of all visits as well as the amount of time spent is equal to 12 and these twelve portals receive 95.59% of all portal visits among the customers in the dataset (Table 2.1).

In this dissertation, I study the web portal market consisting of all 45 portals available in this dataset. For the purposes of this study, the data will be rearranged in order to enable me to complete the estimation of the determinants of portal market share and conduct a hazard/survival analysis.

Although data is comparatively old, the approach and techniques developed in this work may be applied for any dataset and provide useful insights for the analyst. More importantly, my results can be used for the modern market of mobile portals, which is currently undergoing the stage equivalent to one explored in this research.

Table 5.1. Clickstream raw data sample.

id	sdate	bfrom	Bto	session	endview	npgs	week	age	educ	host	dnldtime	viewtime
100	29-Jan-00	91349	7568	91	29-Jan-00	5	5	68.9	12.4	com.lifeminders	308	565
100	31-Jan-00	44068	3210	97	31-Jan-00	3	6	68.9	12.4	com.lifeminders	121	121
100	31-Jan-00	6175	1522	97	31-Jan-00	1	6	68.9	12.4	com.lifeminders	138	138
26	2-Feb-00	273132	17115	104	2-Feb-00	29	6	68.9	12.4	com.aol	461	466
23	2-Feb-00	2990	958	104	2-Feb-00	3	6	68.9	12.4	com.newaol	27	27
26	2-Feb-00	34613	5030	104	2-Feb-00	5	6	68.9	12.4	com.aol	149	165
26	2-Feb-00	8114	83	104	2-Feb-00	1	6	68.9	12.4	com.aol	9	9
23	3-Feb-00	2542	611	107	3-Feb-00	1	6	68.9	12.4	com.fiber-net	48	106
26	3-Feb-00	19270	2275	107	3-Feb-00	5	6	68.9	12.4	com.looksmart	58	345
26	3-Feb-00	141769	22761	108	3-Feb-00	8	6	68.9	12.4	com.ask	160	160
26	10-Feb-00	143722	24310	137	10-Feb-00	8	7	68.9	12.4	com.ask	435	435
24	14-Feb-00	169508	28033	142	14-Feb-00	13	8	68.9	12.4	com.bigfoot	222	222
26	14-Feb-00	58627	1854	142	14-Feb-00	3	8	68.9	12.4	com.hotsheet	139	203
26	15-Feb-00	60212	1317	147	15-Feb-00	3	8	68.9	12.4	com.hotsheet	100	100
26	15-Feb-00	184795	30712	147	15-Feb-00	8	8	68.9	12.4	com.ask	473	473
3	15-Feb-00	436	360	147	15-Feb-00	1	8	68.9	12.4	uk.co.ndirect	4	4
26	15-Feb-00	6568	3646	147	15-Feb-00	3	8	68.9	12.4	com.ask	19	19
23	15-Feb-00	38316	3409	147	15-Feb-00	7	8	68.9	12.4	com.dnai	96	199
26	15-Feb-00	3113	1803	147	15-Feb-00	2	8	68.9	12.4	com.ask	10	180

Table 5.2. Summary statistics on number of visits and demographic characteristics of users.

Variable	Observations	Mean	Standard deviation	Min	Max
Age	2653	38.8221	8.4749	13	71.3
Education	2651	13.9481	1.4395	8.9	16.1
Income	2648	47891.6	22718.06	4999	190132
Household size	2648	2.5262	0.3806	1	4
Married	2653	0.4976	0.1133	0	1
Renting	2653	0.1061	0.1262	0	1
Total number of visits	2654	1216.49	2045.201	1	56098
Time spent online (in seconds)	2654	165497.3	238986.3	1	4571000
Average time spent online (in seconds)	2654	147.8715	79.31824	1	2566.202
Total portal visits	1987	411.2654	482.16	1	5002.87
Time spent on portals (in seconds)	1987	47824.82	64005.214	2	194852
Average time spent on portals (in seconds)	1987	116.287	98.0045	2	877

Table 5.3. Summary of online visits per category for December 27, 1999 – March 31, 2000.

site_type	total visits	visits, %	total time	time, %
Adult Products	74	0.027	3552	0.013
Adult Services	3188	1.164	183276	0.672
Airlines	574	0.210	66645	0.244
Arts	36	0.013	4296	0.016
Astrology	82	0.030	10123	0.037
Auction	10954	4.000	1436223	5.265
Banking	542	0.198	87256	0.320
Books	80	0.029	8376	0.031
Business & Companies	1417	0.517	142272	0.522
Business Products	190	0.069	21437	0.079
Business Products & Services	520	0.190	43195	0.158
Chat (general)	14619	5.339	1195700	4.383
Classifieds	316	0.115	23540	0.086
Clothing	404	0.148	85616	0.314
Community	6151	2.246	847759	3.108
Comparison Shopping	742	0.271	53109	0.195
Computers	3461	1.264	303318	1.112
Consulting	13	0.005	296	0.001
Credit	529	0.193	70482	0.258
E-cards	1395	0.509	168041	0.616
E-mail	21057	7.690	1639038	6.008
Education	800	0.292	104169	0.382
Electronics	269	0.098	61183	0.224
Email Subscription/Reminder Services	987	0.360	73698	0.270
Entertainment Services	6363	2.324	839848	3.079
Events	28	0.010	2471	0.009
Finance	3561	1.300	329991	1.210
Financial/Insurance Services	2590	0.946	259504	0.951
Flowers	32	0.012	6391	0.023
Food & Drink	177	0.065	28601	0.105
Forced Content	184	0.067	10481	0.038
Gambling	25	0.009	1735	0.006
Games	10150	3.707	1816850	6.660
Genealogy	1166	0.426	121276	0.445
General Merchandise	549	0.200	118705	0.435
Government	1026	0.375	198415	0.727
Health	671	0.245	87460	0.321
Hosting	3397	1.241	268617	0.985
Hotels	154	0.056	19039	0.070
ISPs	10332	3.773	1021323	3.744
Incentive Site	2872	1.049	121472	0.445
Information Services	2869	1.048	396124	1.452
Insurance	99	0.036	19065	0.070
International	234	0.085	13094	0.048

Table 5.3. (continued)

site_type	total visits	visits, %	total time	time, %
Internet	5918	2.161	383255	1.405
Internet Telephone	733	0.268	82842	0.304
Jobs	674	0.246	107090	0.393
Legal	50	0.018	8596	0.032
Local Portal	524	0.191	54304	0.199
Maps	455	0.166	44401	0.163
Marketing Companies	1554	0.567	72010	0.264
Medical Services	6	0.002	497	0.002
Medicines, Health and Beauty	134	0.049	22233	0.082
Military	105	0.038	22906	0.084
Movies	140	0.051	22808	0.084
Music	2011	0.734	278396	1.021
News	8140	2.973	1132522	4.152
Online Shopping	2987	1.091	508146	1.863
Online Trading	267	0.098	39454	0.145
Organization	128	0.047	17364	0.064
Personal Pages	18	0.007	684	0.003
Places	30	0.011	4050	0.015
Politics	19	0.007	2246	0.008
Portal	85989	31.401	7035045	25.789
Radio	147	0.054	11689	0.043
Real Estate	824	0.301	96578	0.354
Rental Cars	92	0.034	8920	0.033
Science	71	0.026	13391	0.049
Search	8381	3.061	592754	2.173
Security	86	0.031	5284	0.019
Software	19060	6.960	1941364	7.117
Special Interest	1707	0.623	142934	0.524
Sporting Goods	38	0.014	7128	0.026
Sports	4559	1.665	842557	3.089
Streaming Media	3581	1.308	166888	0.612
Sweepstakes	2833	1.035	271513	0.995
Technology	2484	0.907	223933	0.821
Telecommunications	1284	0.469	108340	0.397
Television	791	0.289	182605	0.669
Toys	118	0.043	23074	0.085
Travel/Places	812	0.297	131920	0.484
Vehicle Information	136	0.050	23705	0.087
Vehicles	720	0.263	137533	0.504
Video	105	0.038	16147	0.059
Weather	971	0.355	149041	0.546
Web Design	184	0.067	17520	0.064
Web-based Applications	94	0.034	13001	0.048
Total visits	273839	100	27279730	100

Table 5.4. Summary of visits to portals for December 27, 1999 – March 31, 2000.

site_id	site_type	total visits	visits, %	total time	time, %
com.about	Portal	2636	0.2401	332647	0.2625
com.netaddress	Portal	3832	0.3490	440726	0.3478
com.altavista	Portal	36247	3.3010	3950609	3.1177
com.aol	Portal	77650	7.0715	7444956	5.8752
at.stop	Portal	80	0.0073	3263	0.0026
at.vol.members	Portal	1	0.0001	11	0.0000
au.com.alphalink	Portal	24	0.0022	4594	0.0036
au.com.nettrek	Portal	1	0.0001	217	0.0002
au.com.powerup	Portal	44	0.0040	5167	0.0041
au.com.tig.homepages	Portal	25	0.0023	7119	0.0056
com.bomis	Portal	2866	0.2610	185725	0.1466
com.ceoexpress	Portal	88	0.0080	6726	0.0053
com.clickheretofind	Portal	564	0.0514	38967	0.0308
com.crosswalk	Portal	148	0.0135	23517	0.0186
com.daily1	Portal	95	0.0087	7288	0.0058
com.directhit	Portal	3782	0.3444	315180	0.2487
org.dmoz	Portal	360	0.0328	29776	0.0235
nl.euronet	Portal	117	0.0107	10837	0.0086
com.excite	Portal	71216	6.4855	7594896	5.9936
com.funcoland	Portal	150	0.0137	42362	0.0334
com.galaxy	Portal	149	0.0136	13030	0.0103
com.go	Portal	75183	6.8468	10808148	8.5293
com.handilinks	Portal	18	0.0016	1855	0.0015
com.hotbot	Portal	16084	1.4647	1452725	1.1464
com.hotsheet	Portal	482	0.0439	38348	0.0303
com.infospace	Portal	5354	0.4876	863811	0.6817
com.iwon	Portal	22646	2.0623	3442158	2.7164
com.kanoodle	Portal	66	0.0060	3542	0.0028
com.looksmart	Portal	5438	0.4952	391354	0.3088
com.lycos	Portal	45091	4.1064	5426206	4.2821
ch.lyrics	Portal	632	0.0576	90130	0.0711
com.megaspider	Portal	422	0.0384	18770	0.0148
com.msn	Portal	207923	18.9353	26014508	20.5295
com.myway	Portal	18381	1.6739	2813050	2.2199
com.com.nerdworld	Portal	59	0.0054	6141	0.0048
com.netscape	Portal	96310	8.7708	10943132	8.6359
com.nettaxi	Portal	1399	0.1274	286757	0.2263
com.planetout	Portal	581	0.0529	136870	0.1080
com.snap	Portal	15210	1.3852	1392590	1.0990
com.starmedia	Portal	645	0.0587	122314	0.0965
com.webcrawler	Portal	4912	0.4473	532076	0.4199
org.webzone	Portal	86	0.0078	15859	0.0125
com.www	Portal	2143	0.1952	204857	0.1617
com.yahoo	Portal	378781	34.4951	41237760	32.5431
com.yep	Portal	152	0.0138	16873	0.0133
Potal visits		1098073	100	126717447	100

Chapter 6. Portal competition and the determinants of market shares.

6.1. Introduction.

The purpose of my work is to determine the most important factors that affect consumer preferences towards portals and define portal market shares.

Market shares occupy a prominent role in industrial and marketing research. Contributions to the theory of market shares are as early as Slater (1961) and Fogg (1974). These authors discuss the problem of gaining market share under various competitive conditions. Fogg (1974) discusses the following key means of increasing market shares:

- price, if firm sets it below average to take customers away from competitors;
- new products, introducing innovations or significant modifications of product;
- service, improve the quality of services and support;
- strength and quality of marketing;
- advertising and sales promotion.

For online market with low or no monetary transaction costs involved, only four last factors can be considered to be important parts of market share gain strategy.

For all other markets, where a tradeoff between profit margin and market share exists, several competing hypotheses of the role of market share were developed. There is a significant literature on the implications of market shares on the firms' profit and welfare. Szymanski, Bharadwaj and Varadarajan (1993) and Cook (1985) found the

market share and profitability are positively related and examined the factors that moderate the magnitude of this relationship. They argue that most managers must focus on building market share as the mean of increasing profits.

In contrast, Boulding and Staelin (1990) showed that very high market share derive no additional profit, and Schwalbach (1991) found optimal market share as between 65 and 70 percent on service and retail markets.

Still, strategic importance of a strong market position – in the form of market share – as a key performance factor is undoubted.

6.2. Modeling the portal market share.

6.2.1. Defining the market share for online portal.

There are two traditional approaches in defining a market size and, therefore, market shares (Fogg (1974), Kotler (1999)):

- total amount of sales or revenue on the market;
- number of consumers.

When making a decision about how to define the online market share, it is

imperative to keep in mind the specifics of revenue process for online firms. Online portal does not receive most of its revenue from the user of its services; instead, the size of customer base and number of views is used to obtain money from advertizing. In these circumstances market share, defined in terms of sales or revenue is not directly related to customers' choice and preferences. As a result, it is preferable to define market and market shares in terms of customer number.

In turn, customer side for the online market can be calculated in three different ways:

- total number of customers registered with portals;
- number of visits made by users;
- total time spent online (Jesdanun (2007)).

The first of the above measures provides good measure for total awareness about the portal, but is of a little help in terms of determining the present popularity of online firm. Since portals do not offer an option to "unsubscribe" from portal usage, customer base contains the information of all users ever subscribed for portal services. In case where a researcher is interested in analyzing the present activity on the portal, it may provide misleading information.

Number of visits and total time spent online gives better understanding of current trends on portal market. However, total time spent online may be affected by the personal characteristics of user (slow learning ability and, as a result, slow browsing behavior; high activity in electronic communication leading to long time spent at e-mail, chats or forums; and others), and provide inconsistent metric. Therefore, the number of visits makes the best metric for portal market share, and I will use it as such in my analysis.

The market shares of portals are constructed in the following way: for each week of observations total number of visits is considered to be 100% of shares or the entire market. The ratio of visits to a particular portal to all portal visits each week will constitute its market share. This way all market shares satisfy the basic property of market shares: they lie between 0 and 1, and sum to 1 for all competitors.

6.2.2. Data.

In order to analyze a change in market shares as a function of changes in different portal characteristics, I need to rearrange my original data to create a panel dataset.

We keep track of market shares defined in terms of total number of visits to all 45 portals in the dataset during 14 weeks of observations, for a total of 630 observations. Figure 6.1 demonstrates the trends of market share changes over the observed period of time and Table 6.1 presents the correlation coefficients between portal attributes.

6.2.3. Conceptual and econometric model.

The literature presents a number of approaches concerning modelling of market shares, and predicting effects of different variables on changes in market shares. Additive market shares and market share attraction models are the most common (Buzzell and Wiersema (1981), Cook (1985))[4]. Owing to the nature of panel estimation, additive model looks more attractive and will be used in this section.

[4] Basic linear additive model is expresses as $MS_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + ... + \beta_n Y_{t-1} + \beta_{n+1} Y_t + ...$, where *MS* stands for market share in period *t*, and *X*, *Y*, ... are decision variables. Another way of stating this model is $MS_t = \beta_0 MS_{t-1} + \beta_1 X_t + \beta_2 Y_t + ...$, where $MS_{t-1}$is included to capture the effects of lagged variables and concentrate directly on the difference in market shares. These basic formulations were applied in several

In order to implement the theoretical hypotheses and verify the prior empirical findings I included different groups of variables into my estimation:

- variables related to the portal brand recognition and customer awareness;
- portal attributes' variables;
- variables reflecting demographic characteristics of portal service users.

Definitions of variables introduced into the model are presented in Table 6.2. Following Goldfarb (2002b) I also include the effect of distributed Denial of Service attack[5] on Yahoo! on February 7, 2000, the 7th week in our panel.

To begin, I consider the classical linear regression model, in which we partition the conditional expectation into time-variant and time-invariant components, $x_{nt}^{'}\beta_0$ and $z_n^{*'}\eta_0$, respectively:

$$E[y_{nt} \mid X, Z^*] = x_{nt}^{'}\beta_0 + z_n^{*'}\eta_0, \qquad \begin{matrix} n = 1,...,N \\ t = 1,...,T \end{matrix} \tag{6.1}$$

where n indexes the individual portals of the panel and t indexes the period of observation. The matrix X contains observable portal attributes and the matrix Z^* contains the unobserved characteristics of the individual portals that are constant from a time period to a time period; vector z_n^* represents unobserved characteristics of the portals that are constant from time period to time period. Since z_n^* are unobserved, we will treat the $\alpha_n = z_n^{*'}\eta_0$ as additional unknown parameters. As such, the α_n are

---

ways. Variables, whose effect on market shares were believed to be nonlinear have been included and linearized; multiplicative models that allow for joint effect of variable were used after the appropriate transformations; simultaneous equation models were applied..
Attraction models that are based on the theorem of market share determination, stating that market shares of competitors will be proportional to their shares of total market effort. $MS_{ij} = \frac{Effort_{ij}}{\sum Effort_{ij}}$, where brand j is on the market with n competitors. Attraction models always satisfy the basic property of market shares: they lie between 0 and 1, and sum to 1 for all competitors. Their practical application, however, has a disadvantage of a greater complexity (when linearization is not possible) and likely presence of multicollinearity. Despite the theoretical advantage attraction models are less used due to computational complexity and limited data availability.

[5] Denial of Service (DoS) attack is an attack on a computer system or network that causes a loss of service to users, typically the loss of network connectivity and services by consuming the bandwidth of the victim network or overloading the computational resources of the victim system.

usually called fixed effects. Each is a distinct intercept for the regression function of an individual portal in the panel. Fixed effects regression is the model to use when we want to control for omitted variables that differ between cases but are constant over time. It lets use the changes in the variables over time to estimate the effects of the independent variables on our dependent variable.

When there is a reason to believe that some omitted variables may be constant over time but vary between cases, and others may be fixed between cases but vary over time, then both types can be included by using random effects. The latent variables model is extended to treat the α_n as random variables, or random effects. In addition to

$$E[y_{nt} \mid X, \alpha] = x_{nt}^{'}\beta_0 + \alpha_n, \qquad \begin{matrix} n = 1,...,N \\ t = 1,...,T \end{matrix} \tag{6.2}$$

$E[y_{nt} \mid X] = x_{nt}^{'}\beta_0 + \alpha_0$ is specified, assuming that the conditional mean of every α_n given $X \equiv [X_1^{'},...,X_N^{'}]'$ is equal to the same constant α_0.

Fixed effects model always give consistent results but it may not be the most efficient model to run. The specification test for choosing between fixed- and random-effects models was devised by Hausman. Test checks a more efficient model against a less efficient but consistent model to make sure that the more efficient model also gives consistent results. The chi-square test is based on Wald criterion:

$$W = \chi^2[K] = [b - \hat{\beta}]'\Sigma^{-1}[b - \hat{\beta}] \tag{6.3}$$

6.3. Empirical results.

6.3.1. General panel estimation.

I start with the estimation of the naïve regression of market share function of the number of features presented by the portal. The results are summarized in Table 6.3, and suggest that number of portal features strongly affect the size of market share.

General specification for the market share estimation includes all the available variables.

The results of this estimation are summarized in Table 6.4. I test the null hypothesis that the differences in coefficients estimated by the efficient random effects estimator and the ones estimated by the consistent fixed effects estimator are not systematic. For this estimation null hypothesis is rejected, so fixed effect estimation is the preferred specification.

The results suggest only quality of mail and search services have major effect on the market shares. The adopted log-linear specification allows us to make direct inferences: increase in Mail and Search quality can lead to an increase of market share, for 5.6% and 4.4% respectively. Due to the little variability in some of the portal attributes over the observed period, some of them were dropped from the analysis.

In this situation, I try to get some helpful insights from random effects specification. It also suggests that Mail and Search can be treated as major determinants of market shares, but also point that Greetings, News service, Messenger and Weather service may play some important role in forming the consumer preferences towards the portal.

The adoption of fixed effects model also tells that the heterogeneity between portals is constant over time.

6.3.2. Aggregated model estimation.

In order to overcome the data problems – limited variability of certain characteristics, I run the estimation of the model with reduced number of variables, some of the new variables will be combinations of old ones.

Here I divide all portal characteristics into several groups that have the most effect on customer decisions. In this estimation I distinguish the following groups of factors: Age, assuming that if portal exist for longer time, the larger customer base can be accumulated, this variable can also serve as a proxy for brand recognition; Mail should be included as individual variable since we have reasons to assume that users separate this service from others (Nie and Erbring (2000)), the availability and quality of mail influence customer retention rate; Virtual community factor (Gallaugher and Downing (2000)); Personalization as a possibility to combine mainly used services on the customized portal page; and Search quality (Goldfarb(2000)).

For this reduced model, I also run fixed and random effect regression (results are summarized in Table 6.5) and perform the specification test. Based on the specification test again we should adopt the fixed effect regression.

Results again suggest that the quality of mail service is most significant in user decisions. This can be explained by the fact that primary goal of portal users is e-mail service and the quality of mail at this stage is the major determinant of their choice; Nie

and Erbring (2000) questionnaire informally supports this idea. Along with the mail service, every user obtains a unique name that can be used in other portal services, such as games, chat, forums, and gets involved into virtual community. Virtual community factor does not appear to be significant in determining general market share, owing this to the limited variability in these services. Still, as we discussed above, the virtual community factor can become the substantial part of the switching costs, and some users rank the volume and quality of virtual community high in their preferences towards certain portal.

6.3.3. Estimation with lagged dependent variable.

The lack of the variability in certain portal features results in lower significance of the coefficients; portal characteristics do not vary much in a short period of time, while the number of unique visits to portal varies from day to day. However, in case of web portal I can suggest one more characteristic, which affects the market share at any given period of time.

There are number of reasons to assume that current market share depends on market share at previous time period; we have two main reasons to suggest this relationship:

- as a result of the force of habit (inertia) people do not change their consumption preferences immediately, following the price increase or income decrease. When there is no monetary cost involved, users have no immediate disutility from using the same product, but switching involves some learning cost and

transitional cost, so substantial fraction of customers would use the same portal as on previous day;

- in addition, higher customer base can generate higher rate of attraction since current users share their knowledge with potential ones; the higher customer base is, the more information spillover can occur.

Taking all this into consideration, I create a model with lagged dependent variable. For this purpose I assume that only previous period market share affects the current one, which suggests the lag of one period only. The simple dynamic specification is formulated as following:

$$E[y_{nt} \mid X, \alpha, y_{n,0}, ..., y_{n,t-1}] = \phi_0 y_{n,t-1} + x_{nt}' \beta_0 + \alpha_n, \qquad \begin{matrix} n = 1, ..., N \\ t = 1, ..., T \end{matrix} \qquad (6.4)$$

Substantial complications arise in estimation of such a model, however. In both the fixed and random effect settings, the difficulty is that lagged dependent variable is correlated with the disturbance, and coefficients are generally inconsistent. The general solution approach, developed in the literature, relies on instrumental variables estimators. For this, I start with simple fixed effect estimator and then consider two instrumental variables. View time of the portal and amount of information (measured by number of bytes exchanged with the portal) are used as instruments. Both of them should be highly correlated with market share, but I assume them to be independent from the disturbances. The results for these estimations are presented in Table 6.6. However, this estimation does not produce any significant coefficients. This can be partially explained by the fact that Age of portal partially controls for the above effects; indeed, if portal exists longer time the customer base accumulated and the level of knowledge about this particular portal will increase.

6.3.4. Seemingly unrelated regression estimation.

The above analysis gives the basic understanding of overall portal popularity. However, in order to understand the key factors affecting market share I would like to consider the shares of separate services. For certain portals, that specialize on offering a particular service (e.g. games or chat), only the market share of this service matters. Estimation of the overall market shares cannot capture this effect.

I continue the analysis applying seemingly unrelated regression (SUR) with primitive coefficients to three different markets:

$$E[y_{tj} \mid X] = x_t^{'}\beta_{0j}, \qquad \begin{matrix} t = 1,...,T \\ j = 1,...,J \end{matrix} \tag{6.5}$$

where the additional subscript j denotes the regression equation for jth dependent variable.

Here, I assume that the market shares of mail service, virtual communities and search are not defined independently. Indeed, when user starts using one service of the portal he can easier start using others as well. Again, I assume that share of certain service is determined by service quality, the availability of other services on the particular portal and demographic characteristic of service users (I assume that demographic characteristics of users of different services can vary).

64.4% of portals participate in the market for e-mail, 73.3% present at least one service, associated with virtual community, search market receives the attention of 86.6% of all portals; and 44.4% of portals participate in all three markets. Figures 6.2 – 6.4 demonstrate the trends of market share changes for three markets.

Since not all portals provide the same set of services, I have to control for it when markets are separated. For the purpose of this estimation, in order to distinguish zero market share of existing service from nonexistent services, I treat the latter ones as missing variables.

The results are summarized in Table 6.7 and matrix of residuals is presented in Table 6.8.

Most signs of our estimated coefficients for the separate markets correspond to intuitive expectations. My prior assumption is that large number of services attracts users and increase market shares. Positive coefficients of the Mail quality in mail equation, Search quality in search equation, and the size of Virtual community in virtual community equation are expected and show clear correlation between quality service and the market share increase. As before, Mail quality and Search show strong as the determinants of market shares (for all three markets).

There are some notable differences in the ways, in which shares for the three markets are determined.

Virtual community share is positively correlated with Age of the portal (0.002%), Mail (0.02%) and Search quality (0.006%), Shopping (0.03%), Weather (0.05%) and Personalization (0.04%); and negatively correlated with Auction (-0.08%), Finance (-0.06%) and existence of personal Page (-0.04%). For portals whose main specialization is games, forums or chat, mail is the additional convenient feature for the members of virtual community that allows to contact each other easily. Visiting the Auction and Finance services are time-consuming and reduce the possibility of future engaging into virtual communication, justifying the negative relationship.

Demographic variables appear significant for the virtual community market share estimation. Positive coefficient of User age and negative coefficient of User age^2 could be expected, and are related to the fact that younger people tend to engage more and more to the virtual communication, but at certain point of time (starting the first job, for example) reduce time spent there. More educated people spent less time in virtual communities (-0.02%) as well as married individuals (-0.18%).

From the estimation of search share, we observe positive relationship with Auction (0.06%), Finance (0.07%), Messenger (0.12%). The explanation is that these services are analogous to the search process in the sense that they are used with a specific goal or target in mind. Again, I found several demographic variables significant for the search market shares. Positive coefficients of User education (0.01%) and negative coefficient of User age (-0.02%) could be expected and can be easily explained. More educated

people engage in larger amount of various searches, but older individuals spend less time doing so. Household size and Income variables appear to be significant as well.

Demographic variables do not play an important role in defining market share for mail service. Mail share depends mostly on own Mail quality (9.2%). Age of the portal, Weather and Personalization services positively correlate with mail market shares. Surprisingly, Auction and Sport services demonstrate strong negative correlation with mail market share (-3.33% and -3.99% respectively).This may be possible due to the fact that some of the major players on the mail market do not offer these features, which creates the major distortion in the structure of the estimated coefficients. Indeed, hotbot, iwon, aol and altavista (ranked #4, #5, #7 and #8 by the size of mail market share) did not offer Auction feature; also, hotbot did not feature Sport on the portal page.

In order to eliminate the effects of such singularities, I run the additional SUR estimation without Auction and Sport attributes. The results of this refined estimation are presented in Table 6.9 (matrix of residuals appears in Table 6.10).

Again, the most important factor determining Mail share is Mail quality, 8.83% increase in share can be achieved by improving this feature. The existence of News, Weather and Messenger features may improve share by 3.82, 4.85 and 4.53% respectively. Possibility of Personalization (1.22%) and Search quality (0.19%) also contribute to gaining additional mail market share.

Virtual community share in this estimation found to be positively correlated with Mail (0.015%), Virtual community size (0.014%), Search quality (0.005%) and Weather (0.04%); and negatively correlated with Shopping (-0.027%) and News (-0.04%). Existence of personal Page (0.048%) and Messenger (0.15%), that allows member of virtual community to share more information improves virtual community market share. As before, User age (0.007%) and User education (-0.017%) influence is significant.

Positive relationship with Mail (0.022%) and Finance (0.026%), and negative relationship with Virtual community (-0.01%) holds for Search market share. Again, this estimation suggests that older and less educated people has lower tendency to engage into online search process.

Demographic variables play an important role in determining the shares for search

and virtual communities, but not for the mail service. I argue that the important difference between mail and other services exist; e-mail service is used by all types of customers whereas only certain demographic groups are interested in using other services such as search, chats, forums and online games. Coefficients suggest that user's age and education play important role in engaging into online search or online virtual communities.

The hypothesis of independence of the three market share equations is rejected in Breusch-Pagan test, which give me the reason to suggest more complicated relationships between different portal attributes, and their role in determining market shares.

6.3.5. Heckman selection estimation.

From the above analysis, I have reasons to suggest that portal attributes are chosen non-randomly; instead, portal characteristics are grouped in a certain manner to attract more customers to the services provided. In order to check the data for sample selectivity, I utilize the Heckman two-stage procedure, which is based on the following:

$$y_t = X_t\beta + \rho\sigma\upsilon_t + e_t \qquad (6.6)$$

The unobserved error term υ_t is replaced by its mean conditional on $z_t = 1$ and explanatory variables W_t:

$$E[\upsilon_t \mid z_t = 1, W_t] = E[\upsilon_t \mid \upsilon_t > W_t\gamma, W_t] = \frac{\phi(W_t\gamma)}{\Phi(W_t\gamma)} \qquad (6.7)$$

Ordinary probit is used to obtain a consistent estimates $\hat{\gamma}$, and on the second step the

unobserved υ_t is replaced with selectivity regressor $\frac{\phi(W_t\hat{\gamma})}{\Phi(W_t\hat{\gamma})}$ and equation (6.6) becomes:

$$y_t = X_t\beta + \rho\sigma\frac{\phi(W_t\hat{\gamma})}{\Phi(W_t\hat{\gamma})} + residual \qquad (6.8)$$

Tables 6.11 – 6.13 present the results of the Heckman selection estimation. Since coefficient ρ is significantly different from zero, I can reasonably infer that I have selectivity in my dataset. The regression results for coefficients are inconclusive, maybe due to a small sample.

Selectivity of portal characteristics is closer related to the decisions of portal management than to consumer choice, making its way outside of the area of research of this dissertation. However, it points to an important factor of portal profitability and success, and makes an excellent topic for future research. Data needed for this research must be more detailed in terms of portal characteristics. Not only the number and composition of portal attributes, but also the order of their appearance on different portals are necessary. Such study will lead to important implications for the online firms' managers, guiding their decisions about web page attributes.

6.4. Conclusions.

After including all of the variables of objective portal attributes together with several demographic characteristics, I found no significant evidence of brand loyalty influence on market shares, in contrast to Goldfarb (2000) and Bucklin and Sismeiro (2001). Together with findings of Ray (2007) it means that most of the loyalty online may be explained by unobservable heterogeneity. Unobservables may include wide range of portal characteristics from color palette and ease-of-use to number of clicks required for various tasks as well as be connected to the emotional attachment of users to certain portal brands or services.

I also found that possibility of personalization does not affect market shares significantly, but the existence of messenger has positive impact on customers' retention. For the early stages of portal market development, mostly the mail and community services were in demand and thus quality of these exact services determined the customer retention. I did not found any significant evidence of DoS attack on the market share; in contrast, the number of visits increased after the attack, mainly due to mail usage increase.

My estimations have demonstrated that individual portal features such as portal age, mail and search quality, are very important in explaining the overall market share, but less powerful in the explaining the market shares of separated services. In contrast, demographic characteristics of users did not have significant influence on overall market shares, but could affect the market share of virtual community and search.

This data also suggests that there exists a clear separation between the market for the mail service and other services produced by portals. Markets for search and virtual communities demonstrate high volatility in market shares, and the peaks of visits appear at different times for different portals. This indicates that customers of different portals

engage into dissimilar types of searches and online discussions and probably represent different demographic groups. These different demographic groups are tied to different services, portal characteristics, and, as a result, to different portals. Taking this into account, I can find some common features that help to attract and retain diverse customers. Targeting and maintaining the good relationship with current users also should become the important part of portal market policy. This provides additional motivation to study and model the consumer preferences toward different portal attributes, as well as studying the effects on advertising and media mentions on change in market shares, which will be beneficial for the full understanding the phenomenon of portal popularity.

Table 6.1. Correlation coefficients for the portal attributes.

	share	age	mail	auction	shopping	sport	chat	greetings	games
share	1								
age	0.4638	1							
mail	0.6988	0.2835	1						
auction	0.3106	0.188	0.3653	1					
shopping	0.1879	0.0377	0.202	0.3115	1				
sport	0.1688	-0.0255	0.1589	0.3287	0.5343	1			
chat	0.2151	0.2657	0.3401	0.075	0.0478	0.0278	1		
greetings	0.3446	0.0801	0.4301	0.191	0.4911	0.3923	0.4536	1	
games	0.3075	0.1164	0.3887	0.4177	0.4939	0.405	0.3617	0.5664	1
finance	0.3465	0.181	0.2948	0.1817	0.4648	0.4684	0.4204	0.6003	0.5974
news	0.2147	0.0508	0.3117	0.3851	0.6166	0.4899	0.4225	0.5	0.4729
search_quality	0.5731	0.3447	0.4267	0.4391	0.2739	0.3774	0.1725	0.3809	0.3943
messenger	0.7596	0.3681	0.5088	0.2129	0.2058	0.1963	0.2327	0.2656	0.3238
personal	0.2835	0.1579	0.1915	0.3267	0.019	0.2294	0.2699	0.1879	0.2854
weather	0.2754	0.0341	0.3344	0.2983	0.3806	0.362	0.229	0.4202	0.465
page	0.3814	0.3019	0.365	0.2874	-0.0119	0.1029	0.3563	0.3986	0.3771

	finance	news	search_quality	messenger	personal	weather	page
finance	1						
news	0.6327	1					
search_quality	0.3523	0.3523	1				
messenger	0.3391	0.2098	0.6147	1			
personal	0.3283	0.1897	0.4041	0.3238	1		
weather	0.594	0.5511	0.1157	0.0683	0.0971	1	
page	0.3152	0.2	0.5475	0.4315	0.4718	0.1279	1

Table 6.2 Definition of variables introduced into the model.

Market share: MS	Market share of individual portal, measured as natural logarithm of the ratio of total visits to this particular portal to total portal visits
Portal attributes	
Portal age	Measured in weeks since portal first appeared online
Mail	Volume of the mailbox provided by the portal
Auction	Dummy variable, equals to 1 if service is provided and zero otherwise
Shopping	Dummy variable, equals to 1 if service is provided and zero otherwise
Sport	Dummy variable, equals to 1 if service is provided and zero otherwise
Chat	Measures the volume of chat community
Greetings	Dummy variable, equals to 1 if service is provided and zero otherwise
Games	Dummy variable, equals to 1 if service is provided and zero otherwise
Finance	Dummy variable, equals to 1 if service is provided and zero otherwise
News	Dummy variable, equals to 1 if service is provided and zero otherwise
Search	Measures quality of search engine on a given portal, based on the outside published rankings.
Messenger	Dummy variable, equals to 1 if service is provided and zero otherwise
Weather	Dummy variable, equals to 1 if service is provided and zero otherwise
Page	Dummy variable, equals to 1 if service is provided and zero otherwise
Personalization	Personalization possibility, measured by the number of services that can be included into personalized portal page
Virtual community	Combined variable, that measures the overall size of all services associated with virtual communities, such as chats, forums, discussion groups, games, greetings.
Demographic characteristics of users[6]	
User age	Age of the user of certain service
User education	The number of years of education of the user
User income	The income of the portal user
Household size	The household size of the portal user
Marital status	User's marital status

[6] Please, note that for market share estimations averages of consumers characteristics are used.

Table 6.3. Naïve estimation of market shares (errors in parenthesis)

Variable	Fixed Effects	Random Effects	Maximum Likelihood Random Effects
Portal age	.114716 **	.1246642***	.1250441***
	(.0562211)	(.0387487)	(.0381184)
Number of features	3.581731**	4.168477***	4.174866***
	(1.75412)	(.7624962)	(.7430848)
Constant	3.692048	-1.572111	-1.661525
	(12.38253)	(6.987677)	(6.82586)

*** significant at a 1% level ** significant at a 5% level * significant at a 10% level

Hausman test: Ho: difference in coefficients not systematic

$\chi^2(2) = 0.27$

Prob> $\chi^2 = 0.8739$

Table 6.4. General estimation coefficients. (errors in parenthesis)

Variable	Fixed Effects	Random Effects	Maximum Likelihood Random Effects
Portal age	-.0000111	.0000152	.000009
	(.0000293)	(.000028)	(.0000281)
Mail	.056468 ***	.046293 ***	.048707 ***
	(.010693)	(.01045)	(.01043)
Auction	dropped	.0180771	.018223
		(.011626)	(.013116)
Shopping	.0003912	-.0000459	.0000573
	(.0020913)	(.0021073)	(.0020637)
Sport	.0016779	.000936	.0011093
	(.001077)	(.001084)	(.0010677)
Chat	.0000574	.0003706	.002773
	(.0028545)	(.0028194)	(.5675034)
Greetings	dropped	.0236696 *	.0237162 *
		(.012983)	(.0145379)
Games	dropped	-.0167847	-.0168863
		(.0125487)	(.0141586)
Finance	dropped	-.0042648	-.0042086
		(.015131)	(.0170706)
News	dropped	-.0237903 *	-.0239668
		(.0134989)	(.0152123)
Search	. 044313 ***	.018471 *	.018925 *
	(.005103)	(.010865)	(.01295)
Messenger	dropped	.1586065***	.1583554 ***
		(.0216061)	(.0244928)
Personal	dropped	.0032888	.0028424
		(.0185019)	(.0210415)
Weather	dropped	.0395516 **	.0397845 ***
		(.0121548)	(.0137019)
Page	dropped	-.0030183	-.03029
		(.0123003)	(.0251094)
Constant	.0071102 *	-.0019639	-.0015191
	(.0038159)	(.0073173)	(.0080424)

*** significant at a 1% level ** significant at a 5% level * significant at a 10% level

Hausman test: Ho: difference in coefficients not systematic

χ^2 (7) = 72.26

Prob> χ^2 = 0.000

Table 6.5. Estimation coefficients for grouped factors (errors in parenthesis).

Variable	Fixed Effects	Random Effects	Maximum Likelihood Random Effects
Portal age	-2.04e-06	.0000324	-.0000951
	(.0000287)	(.0000282)	(.0000521)
Mail	. 050404***	. 048044***	.0333178
	(.009957)	(.009999)	
Virtual community	3.43e-06	-.0003219	-.0264391***
	(.0027482)	(.0024727)	(.0089534)
Personalization	dropped	.0221946***	.0283749 **
		(.0071195)	(.0135234)
Search	. 043217*** .	.0015117	-.0073986 ***
	(.007083)	(.0010819)	(.0021923)
Constant	.0073081*	-.0033845	.0496147 ***
	(.0038907)	(.0067776)	(.0129264)

*** significant at a 1% level ** significant at a 5% level * significant at a 10% level

Hausman test: Ho: difference in coefficients not systematic

χ^2 (4) = 44.37

Prob> χ^2 = 0.0000

Table 6.6. Estimation coefficients for grouped factors with lagged dependent variable (errors in parenthesis).

Variable	Fixed Effects	IV (viewtime)	IV (information)
Previous share	-.0132885	.6157867	1.116115
	(.0370786)	(1.326211)	(1.541322)
Portal age	-.0053252	.013574	.0276071
	(.0095374)	(.0389227)	(.0458518)
Mail	-.0428925	-.0215094	-.0624058
	(.3105964)	(.3750725)	(.495065)
Virtual community	dropped	dropped	dropped
Personalization	dropped	dropped	dropped
Search	dropped	dropped	dropped
Constant	-1.451188	-2.376962	-3.087203
	(1.133514)	(2.320639)	(2.8391)

*** significant at a 1% level ** significant at a 5% level * significant at a 10% level

Table 6.7. Estimation coefficients for seemingly unrelated regression (errors in parenthesis).

Variable	Mail share	Virtual community share	Search share
Portal age	.0138491 **	.0002003 ***	.0001374
	(.0063659)	(.000058)	(.000101)
Mail volume	9.28002 ***	.0240953 ***	.017704 **
	(.4630624)	(.004287)	(.007369)
Virtual community	-.4905094	.0076595 **	-.0111545
volume	(.4530295)	(.004194)	(.0072103)
Search quality	.3045536 ***	.0062138 ***	0.1759 ***
	(.1169211)	(.001082)	(0.0497)
Auction	-3.335893 ***	-.0868088 ***	.0665612 ***
	(.9743802)	(.009022)	(.0155079)
Shopping	2.050117	.0314872 **	-.0223635
	(1.57984)	(.014628)	(.025144)
Sport	-3.997614 ***	-.006046	-.0451486 ***
	(.9239716)	(.008555)	(.0147056)
Finance	-.7191751	-.0613806 ***	.0745455 ***
	(1.144949)	(.010601)	(.0182226)
News	.0618535	.0176259	.0094498
	(1.743341)	(.016142)	(.0277464)
Messenger	1.480178	.0930142 ***	.1202496 ***
	(1.744976)	(.016157)	(.0277725)
Weather	5.100973 ***	.0580092 ***	.0026737
	(.8754027)	(.008105)	(.0139326)
Page	-.5508563	-.0456993 ***	-.0015481
	(1.654924)	(.0153237)	(.0263392)
Personalization	2.366064 **	.0450295 ***	-.005825
	(.9499139)	(.008795)	(.0151185)
User age	-.2076327	.0090494 **	-.0220699 ***
	(.4381178)	(.004056)	(.0069729)
User age ^ 2	.0054169	-.0001047 *	.0003041 ***
	(.0060035)	(.0000556)	(.0000956)
User education	-.4824033	-.0157727 ***	.016724 *
	(.5416363)	(.005015)	(.00862)
User income	.000014	.0000008 ***	-.000001 *
	(.000034)	(.0000003)	(.000005)
Household size	3.745692	-.0357367	.128605 **
	(3.28816)	(.0304466)	(.052333)
Marital status	-8.72016	-.1798767 **	-.2202831
	(13.5373)	(.087625)	(.215455)
Constant	-16.07253 ***	-.0967714 ***	-.0577697 *
	(2.16882)	(.020082)	(.034518)
R^2	0.8247	0.7946	0.5012
χ^2	1510.14	1241.78	322.52

*** significant at a 1% level ** significant at a 5% level * significant at a 10% level

Table 6.8. Correlation matrix of residuals for SUR.

Variable	Mail share	Virtual community share	Search share
Mail share	1		
Virtual community share	0.4329	1	
Search share	0.1646	-0.1398	1

Breusch-Pagan test of independence: χ^2 (3) = 75.139

Table 6.9. Estimation coefficients for seemingly unrelated regression without Auction and Sport features (errors in parenthesis).

Variable	Mail share	Virtual community share	Search share
Portal age	.0088804	.0000368	.0002802 ***
	(.006435)	(.0000646)	(.0000996)
Mail volume	8.835614 ***	.0156564 ***	.0225865 ***
	(.4831619)	(.0048469)	(.007479)
Virtual community	-.1562821	.0140053 ***	-.0148253 **
volume	(.4763895)	(.0047789)	(.0073741)
Search quality	.1974838*	.0059401 ***	.088005 ***
	(.1218081)	(.0012219)	(.0269625)
Shopping	2.085136	-.0276476 *	-.0016521
	(1.522167)	(.0152697)	(.023562)
Finance	.3960118	-.0119248	.0262414 *
	(1.002403)	(.0100557)	(.0155164)
News	3.827316 **	-.0434778 **	.035925
	(1.714807)	(.0172022)	(.0265439)
Messenger	4.536233 ***	.1500027 ***	.0011602
	(1.741851)	(.0174735)	(.0018855)
Weather	4.859173 ***	.0405 ***	.0217983
	(.8904622)	(.0089327)	(.0137837)
Page	-.9195774	.0488004 ***	-.00247
	(1.758398)	(.0176395)	(.0272187)
Personalization	1.227431 **	.0139481	.0187442
	(.9372612)	(.0094022)	(.0145081)
User age	-.2390817	.0077392 *	-.0208154 ***
	(.4654727)	(.0046694)	(.0072052)
User age ^ 2	.0065582	-.0000789	.0002863 ***
	(.0063762)	(.000064)	(.0000987)
User education	-.6372893	-.0171392 ***	.0164184 *
	(.5752709)	(.0057709)	(.0089048)
User income	.0000271	9.36e-07 **	-9.27e-07
	(.000037)	(3.71e-07)	(5.73e-07)
Household size	4.305223	-.0254754	.122922 **
	(3.494185)	(.0350521)	(.0540873)
Marital status	-10.50174	.0924294	-.2340444
	(14.38845)	(.1443386)	(.2227222)
Constant	-11.37215 ***	-.0062658	-.1110019 *
	(2.070288)	(.0207682)	(.0320465)
R^2	0.8018	0.7275	0.4666
χ^2	1298.81	856.82	280.81

*** significant at a 1% level ** significant at a 5% level * significant at a 10% level

Table 6.10. Correlation matrix of residuals for SUR without Auction and Sport features.

Variable	Mail share	Virtual community share	Search share
Mail share	1		
Virtual community share	0.4864	1	
Search share	0.1385	-0.2074	1

Breusch-Pagan test of independence: χ^2 (3) = 95.899

Table 6.11. Heckman selection estimation. Selection into mail service (errors in parenthesis).

Variable	Equation	Selection Equation
Auction	-.1180752	-.3666605
	(.6822809)	(1.198583)
Shopping	.5138239	9433113
	(.6716644)	(.8976336)
Sport	.015732	.0297045
	(.5911634)	(1.006687)
Finance	-.2914747	-.9102253
	(.8539004)	(1.072763)
Messenger	.4664836	6.076128
	(1.418686)	
Weather	.3313947	.9496311
	(.6324662)	(.83813)
Page	.7759082	2.111691
	(1.524858)	(1.435161)
Personalization	-.1845603	-.5562825
	(.8537364)	(.9431831)
News		6.859507
		(.8834671)
Constant		-7.009001

Two-step estimate of ρ = 1.7288011 is being truncated to 1

Wald χ^2 (15) = 12.61

Prob > χ^2 = 0.6324

Table 6.12. Heckman selection estimation. Selection into virtual community (errors in parenthesis).

Variable	Equation	Selection Equation
Auction	.0048705	.0758555
	(6.65e+07)	
Shopping	4299227	-11.51437
	(7.84e+07)	(7052.877)
Sport	-.2598806	-.1746025
	(7.16e+07)	
Finance	.5405348	11.88563
	(1.01e+08)	(4284.413)
Messenger	-.5397044	-.2954422
	(1.02e+08)	
Weather	-.3169115	-11.75569
	(8.38e+07)	(5408.575)
Page	-.1078739	-.0338687
	(1.29e+08)	
Personalization	.2574624	.0738498
	(7.36e+07)	(1883.521)
Constant		-6.108409
		(6384.983)

Two-step estimate of ρ = 6801.1559 is being truncated to 1

Wald χ^2 (12) = 0.00

Prob > χ^2 = 1.0000

Table 6.13. Heckman selection estimation. Selection into search sevice (errors in parenthesis).

Variable	Equation	Selection Equation
Auction	-.028135	.4704107
	(9.47e+07)	
Shopping	.6374233	12.21205
	(1.02e+08)	
Sport	.0669373	-.4535455
	(8.64e+07)	
Finance	.1223237	.8587899
	(1.13e+08)	
Messenger	-.133232	-11.22639
	(1.57e+08)	
Weather	.1451755	.4140289
	(8.66e+07)	
Page	.0111155	11.21783
	(1.85e+08)	
Personalization	.0850294	.0210149
	(1.09e+08)	
Constant		-6.079707

Two-step estimate of ρ = 14116.891 is being truncated to 1

Wald χ^2 (8) = 0.00

Prob > χ^2 = 1.0000

Chapter 7. Analysis of switching behavior.

7.1. Introduction.

There are a number of studies of attrition and retention in marketing research. What drives customers from the well-known goods and services, even if dropping and eventual switching might be costly? How changes in quality of the product can help retain customers and survive on the market? Duration analysis helps to answer these questions.

The analysis of the duration data and hazard rates came fairly recently to the economic and business studies; before, it was a prerogative of the biomedical research. Analysis of the probability of failure is based on calculation of regression equations in which the regressand is dichotomous and regressors are the value and trend of selected variables. There are two major reasons why this research issue cannot be addressed via ordinary least squared analysis (OLS) regression techniques: first, the dependent variable of interest (survival/failure time) is most likely not normally distributed; second, there is a problem of censoring, that is, some observations will be incomplete.

From the earliest publications in 1970s, researchers tried to build models for determination and prediction of timing of an event. A common research question in medical, biological, engineering or economic research is to determine whether or not certain continuous variables are correlated with survival or failure times.

Cox in 1972 presented an approach to hazard model, which became a very popular

method of analyzing the effects of covariates on the hazard rate. The proportional hazard model was developed later and contained no constant term; it became a common choice for modeling survivals, but was applied mostly to medical and biological research (see Farenwell (1978), Sleeper and Harrington (1990) for review). Kiefer (1985) and Lancaster (1990) present useful notes about the application of this approach.

Estimations of hazard or failure rates were used widely for the financial markets. Thies and Gerlovski (1993), Barr, Seiford and Siemens (1994) tested the solvency of national and state banks in their papers. Hwang, Lee and Lian (1997) were calculating the probability and timing of bank failures. These works were using the OLS analysis and linear function to obtain their results. Jacobson and Mode (1985) used the Poisson distribution for human survival simulation.

Later, a regression-like approach was used in the field of Industrial Organization. Cole and Gunther (1995) used a parametric estimation of survival model applied to determinants of bank failures using loglogistic underlying distribution; they extended survival model to separate estimation of hazard rate and survival time. Later, this technique was successfully applied by Gonzales-Hermosillo, Pazarbasioglu and Billings (1996) to the case of Mexican financial crises. Shumway (1999) developed discrete-time hazard model with a logit model estimation program to estimate bankruptcy predictors.

Failure or bankruptcy analysis allows for pure hazard estimation while switching between goods and services is a two-way process: initial drop (that can be characterized by hazard function) and future adoption of different producer of goods and services. Another line of researchers analyzed this switching behavior.

Borenstein (1991) explored switching between retailers of gasoline; Knittel (1997) had a similar investigation on a long distance telephone carriers market. They both concluded that only considerable changes in prices or fall of quality can change the customers' attitude and lead to potential drop. For the online market, when the cost of dropping/switching is relatively low, companies do not have much flexibility. In order to retain their customers, they must provide quality services and make sure that their users are satisfied all the time. In that perspective, study of dropping from Internet portals becomes very interesting. Online markets allow to explore this question with great detail because of the nature of data. With detailed data that keeps track of all movements

online, we can see not only the web page that the surfer ended up at, but also his original path to this page (which is impossible even for very detailed grocery data that was often used in hazard and switching models). Goldfarb (2002b) concentrated on how bad experience with the certain webpage can affect user's decision to leave it, using the data for the DoS attack on Yahoo! and CNN and several Internet shopping websites. The results show strong survival rate at Yahoo, but not the other websites.

In this study I am interested in modeling the hazard function for the online markets in order to understand what factors can lead to potential users' drop off, and will do that in three model specifications. I will use both nonparametric and parametric estimates with different underlying distribution to find which can better fit and explain the situation on online portal market. Later, I allow for user heterogeneity and use Kaplan-Meir model for the analysis.

Sometimes switches may be triggered by the household members, who wish to share their online experience and knowledge about new / better portals. Using my data, I explore the question whether users from multiple member households produce higher switching rates.

7.2. Duration model of switching.

7.2.1. Definition of hazard.

The most important question is how to identify the failure. Different metrics can be developed to approach this problem:

- using the entire portal visits data. If the portal visits are not followed by the visit to the same portal, we may assume that user dropped. Potential problem with this approach is that for certain services (News, Games, Search, etc.) an individual may prefer to use several portals instead of just one. In this case, movement between different web portals is not an indicator of drop;
- reasonable assumption was made that, in contrast to other services, user normally visits only one virtual mailbox; and data supports this hypothesis. Therefore, changing the mailbox on one portal to another is a strong indication of intentional drop from the portal;
- significant changes in time user spent on different portals. Starting the new portal, user needs additional time to learn how to navigate in the new environment and usual tasks take more time than it was on the familiar portal; some learning is required before consumer can switch completely. If individual starts spending more time on the new portal, than on the old one, I positively identify it as an indicator of switching.

Given all the above, I will use two different metrics as the definition of failure for the hazard function: change of the virtual mailbox and change in individual's usage time of portals.

7.2.2. Data.

Based on our metrics, we analyze our data in two ways:

- Model I uses the change of virtual mailbox as the definition of failure;
- Model II uses the change in usage time as the definition of failure.

There is sufficient number of failures for the above definitions. For Model I, I identify 2629 failures, and this number is equal to 55 for Model II.

7.2.3. Conceptual and econometric model.

In the model specification, I concentrate on consumer characteristics and also include portal attributes following the same ideas used for market share estimations. Based on the previous discussion, I will use three different model specifications for the survival model estimations:

- Specification A includes demographic characteristics of users and full information about portal attributes;
- Specification B includes demographic characteristics of users and aggregated information about portal attributes[7];
- Specification C includes demographic characteristics of users and Portal Age to capture the possible lock-in and spillover effects[8].

I start with the simple logit analysis to evaluate the probability of switching and then estimate the duration models of switching on portal market.

[7] Aggregated model is defined the same way as for the market share estimation, subchapter 6.3.2.

[8] Possible spillover effect was previously discussed in subchapter 6.3.3.

7.2.3.1. Logit model of probability of switching.

The behaviors of all dynamic systems are dependent upon their initial conditions. Given dataset does not provide any information about the individuals' experience with Internet portals before the observed period, creating the left-censoring[9] problem. In this situation a model of the portal switching with the discreet dependent variable will be coherent exploratory strategy for the beginning. Combining the results of such estimation with the parametric survival estimates will provide better understanding of users' switching behavior.

To estimate the probability of switching as a function of portal attributes and user characteristics I first code the dependent variable as a dummy variable:

$$Y_n = \begin{cases} 0 \text{ if individual } n \text{ switches from the given portal} \\ 1 \text{ if individual } n \text{ stays with the given portal} \end{cases} \quad (7.1)$$

and then

$$\begin{aligned} prob[Y = 1] &= F(\beta' X) \\ prob[Y = 0] &= 1 - F(\beta' X) \end{aligned} \quad (7.2)$$

The set of parameters β reflects the impact of the individual characteristics and portal attributes on the probability of switching Y. I use the logistic distribution to complete the logit model.

$$prob[Y = 1] = \frac{1}{1 + e^{-\beta x}} \quad (7.3)$$

[9] I use the term censoring to refer to the situation when an individuals spells of interest may end before the observation period and thus is not observed in the data.

7.2.3.2. Cox's proportional hazard model.

The proportional hazard model is the most general of the regression models because it is not based on any assumptions concerning the nature or shape of the underlying survival distribution. The model assumes that the underlying hazard rate (rather than survival time) is a function of the independent variables (covariates); no assumptions are made about the nature or shape of the hazard function. Thus, in a sense, Cox's regression model may be considered to be a semiparametric method. We will use it first before going to the parametric estimation.

The model specifies that:

$$\lambda(t_i) = e^{-\beta X_i}\lambda_0(t_i) \tag{7.4}$$

The function λ_0 is a baseline hazard. Cox's partial likelihood estimation allows to obtain β without requiring estimation of λ_0. For the simplest case, the partial log-likelihood is

$$\ln L = \sum_{i=1}^{K}[\beta' X_i - \sum_{j \in R_i} e^{\beta' X_j}] \tag{7.5}$$

While no assumptions are made about the shape of the underlying hazard function, the model equations shown above do imply the following. First, we specify a multiplicative relationship between the underlying hazard function and the log-linear function of the covariates. Basically, it is assumed that, given two observations with different values for the independent variables, the ratio of the hazard functions for those two observations does not depend on time (proportionality assumption). Second, we impose a log-linear relationship between the independent variables and the underlying hazard function.

7.2.3.3. Parametric estimations of hazard model.

The most popular choices for the parametric estimations of the hazard rate are exponential, Weibull and lognormal distributions. Exponential distribution models the hazard rate that does not change over time. Since portals are experienced goods, we assume that attrition rate or drop rate must vary over time; this implies using the distributions with different behavior. Bennett (1999) presents a good discussion about functional forms for the duration analysis.

There are two opposing effects that may influence the behavior of portal users (Bolton (1998), Bucklin and Sismeiro (2001)):

- lock-in effect: the longer an individual uses the portal the less likely he is to drop because of switching costs - this implies the decreasing hazard function.
- bad experience effects: when user first start using the portal, he may be dissatisfied and this causes higher attrition rates among the new users – this may lead to the increasing hazard for the beginning of the time period.

In the present context, the lognormal and loglogistic distributions are likely candidates, because they can generate a hazard which first rises and then falls (Figure 7.1), as it can be expected with experienced goods. I am also going to use the Weibull estimation since it allows for the monotonically increasing/decreasing hazard function – if lock-in effect is strong it may override the influence of learning and bad experience effects.

7.3. Empirical Results.

7.3.1. Logit estimation of the probability of switching.

The results of logit probability estimates are presented in Tables 7.1-7.4. For model I, main factors contributing to the survival probability are Mail quality, which increases the probability of survival in 1.22 times[10]; existence of such portal features as Shopping, Finance, News, which raise the probability of survival in 1.93, 1.64 and 1.21 times respectively. The existence of personalization possibility on the portal increases its survival probability in 1.48 times, Page and Weather features increase the odds of survival in 1.17 and 1.15 times. Among the demographic characteristics, User age and Household size increase the probability of survival by .7% and 4%, and higher user education reduces the probability of survival by 3.6%. From the aggregated model we can see that virtual communities positively contribute to the probability of portal survival (18.4%).

For model II, Shopping, Finance, Search quality, Weather, personal Page become major positive determinant of survival, contributing up to 71% to the probability. Again, User age, Education and Household size influence the survival probability, all of them have positive impact (0.5, 8 and 10% respectively).

---

[10] If probability of survival P_i is given by $\frac{1}{1+e^{-Z_i}}$, then probability of switching is $1-P_i = \frac{1}{1+e^{Z_i}}$, therefore the odds ratio in favor of survival is $\frac{P_i}{1-P_i} = \frac{1+e^{Z_i}}{1+e^{-Z_i}} = e^{Z_i}$, and can be obtained from the logit estimation by the simple transformation.

7.3.2. Cox's proportional hazard estimation.

The results of proportional hazard model are summarized in the Tables 7.5 and 7.6.

The hazard ratios reported correspond to a one-unit change in the corresponding variable. For Model I, I find that married users have higher hazards and hazard rate may increase together with the size of their household (by 1.28%); that more educated people tend to have lower dropping rate by 1.07%, and that users with lower income stay longer (by 0.01%). Whether a user rents a house does not seem to make much difference as well as user age. The effect of the Portal age on the survival function is positive but small (1.0008%) for Model I as well as the effect of such portal attributes as Shopping (1.73%), News (0.58%) and Personalization (1.37%). The model as a whole is statistically significant.

For Model II, only Household size (7.86%) and Marital status (0.0004%) appear to be significant among the demographic characteristics, and quality of search (1.16%) increases survival rates.

Both models produce the hazard function declining over time (Figure 7.2).

7.3.3. Lognormal regression.

Lognormal survival distribution is represented by the following hazard and survival functions:

$$f(t) = (p/t)\phi[p\ln(\lambda t)]$$
$$S(t) = \Phi[-p\ln(\lambda t)] \qquad (7.6)$$

with $\ln t$ normally distributed with mean and standard deviation $1/p$.

This is a fully parametric model (as opposed to Cox's proportional hazard model) and estimates are presented in Tables 7.7 and 7.8.

Estimated hazard functions depend on t, p and X. Unfortunately, the actual magnitudes of the effects of the covariates are difficult to interpret for the hazard function. The signs of the estimated coefficients suggest a direction of this effect of the variable on the hazard function if the hazard is monotonic such as Weibull. In case of a non-monotonic hazard (lognormal and loglogistic functions), even the direction of influence is ambiguous. Still, we try to give some regression-like interpretation of coefficients.

Education and Household size affect the survival rates in Model I significantly and negatively, Income affects it positively. Here, more portal characteristics appear to be significant determinants of portal survivals, in particular, Shopping, Sport, Games, Finance, News, Search and Weather. The more attributes portal has, the higher the survival of such portal.

Age, Household size and Marital status influence customers' dropping decisions in Model II, none of the portal attributes found to be significant.

Figure 7.3 illustrates the survival curves.

7.3.4. Loglogistic estimation.

For loglogistic regression, the hazard function and the survival function are respectively:

$$\lambda(t) = \lambda p(\lambda t)^{p-1}/[1+(\lambda t)^p]$$
$$S(t) = 1/[1+(\lambda t)^p] \qquad (7.7)$$

where $\ln t$ has a logistic distribution with mean $-\ln\lambda$ and variance $\pi^2/(3p^2)$.

The results of loglogistic regressions are summarized in Tables 7.9 and 7.10. Again, for Model I, Education and Income are found to significantly affect survival rates. More educated people are 0.08 less likely to drop from the portal and users with higher income are 0.0001 times more likely to switch.[11] A positive effect of Shopping (0.78) among portal characteristics contributes to the survival odds. Age, Household size and Marital status have effect on survival odds in Model II, by factors of 0.14, 2.83 and -14.84.

Survival curves are presented on Figure 7.4. For both models, peaks of the switching happen between third and fourth week, and most of the users switch by the eighth week, the middle of the observed timeframe.

[11] The corresponding coefficients for the accelerated-time failure and proportional odds models are related by $\beta_j = -\alpha_j p$ for the *j*th covariate. Stata provides the estimates for α and γ, the reciprocal of *p.* An estimate for the odd ratio is found by $\hat{r} = \exp(\hat{\beta}_j) = \exp(-\hat{\alpha}_j p) = \exp(-\hat{\alpha}_j \frac{1}{\hat{\gamma}})$. For discussion see David G. Kleinbaum and Mitchell Klein, "Survival Analysis: A Self-Learning Text." Springer, 2005.

7.3.5. Weibull hazard rate model.

$$\lambda(t) = \lambda p(\lambda t)^{p-1}$$
$$S(t) = e^{-(\lambda t)^p} \tag{7.8}$$

specify the survival distribution for Weibull hazard rate estimation, the results are presented in Tables 7.11 and 7.12. Same tendency is observed in the coefficients of Model I: Education, Income, Household size and Marital status produce most of the positive effect on portal survival rates, adding 2.8, 2.7, 3.8 and 10% to the survival probability.[12] Portal attributes, such as Mail quality and Search quality may increase the probability of survival by 3.4 and 2.8%; together with the existence of Greetings (1.79%), Games (1.42%), News (1.58%) and Weather (5.67%) features. For Model II, only the variables representing Age, Education, Household size and Marital status appear to be significant. All the above indicate that older people tend to drop from portal less often as well as less educated people. This can be an indication of the existence of learning costs. Survival curves are presented on Figure 7.5, the parameter p>0 suggests the increasing rate of survival.

Generally, in Model I, Education and Income produce the most significant coefficients among demographic characteristics for all model specifications, accompanied by several portal attributes; while Age, Household size and Marital status are the most important in Model II. This result is robust to different model specifications and different underlying distributions. This, in fact, highlights the interesting distinction between estimations of two models. When we identify switch as change of the mail

[12] The Weibull hazard function in the form $h(t) = \lambda p t^{p-1}$, a Weibull proportional hazard model is defined by reparametrizing $\lambda = \exp(\beta_0 + \beta_1 x)$; and hazard ratio is obtained as $HR\ [x=1\ vs\ .x=0] = \frac{\exp(\beta_0 + \beta_1 x) p t^{p-1}}{\exp(\beta_0) p t^{p-1}} = \exp(\beta_1 x)$. For discussion see David G. Kleinbaum and Mitchell Klein, "Survival Analysis: A Self-Learning Text." Springer, 2005.

service, the survival function depends both on user personal characteristics and portal attributes. At the same time, if switches identified through portal usage time, only demographic characteristics of customers produce significant coefficients for the survival function. Also, the number of switches in Model II is notably lower than number of switches in Model I. Overall, Model II that uses change in portal usage time as a definition of failure, tends to produce less significant results than Model I.

One of the easy explanations would be the consumer heterogeneity. Group of people, who use at least two portals at the same time for a certain time period, i.e. demonstrate distinctive behavior, is different from other population in their demographic characteristics. In this situation, our metric for failure will mostly affect individuals with these distinctive characteristics. By choosing the time spent on portal as our metric, I selected a specific group of people, which is not representative for the entire sample.

7.4. Consumer heterogeneity.

The parametric models used above do not allow for any individual heterogeneity, because hazard function does not vary across individuals.

Internet users choose which website to visit as they make all the other choices: given all the information, they choose the best alternative. However, the factors they consider while making this choice can vary from person to person. One of the important determinants of consumer behavior is the level of this consumer's activity.

Kotler (1999) stresses the importance of consumer differentiation on the basis of their activity level; Nie and Erbring (2000) have found demographic and behavioral

differences in all users who have Internet access. Depending on the demographic and socio-economic characteristics, user activity differs from person to person. Internet users have different habits in terms of total Internet usage (total time spent online, total number of web sites visited during one session, number of sessions per day/per week, number of different web sites visited); they also differ in their preferences toward one or another web site and the likelihood to switch or change their preferences.

7.4.1. Identification of different user types.

Based on previous research, we would like to separate at least two groups of Internet users: intensive users and regular users. The definition “intensive user” is applied to experienced users, who use Internet actively and heavily; we will use this definition in this paper. I already discussed that learning costs and virtual communities are the two parts that mostly constitute switching costs between different portals. For active users, who surf Internet intensively, switching constantly from site to site, trying to find best possible place, the learning component of the switching costs can be considerably lower than for non-active users. It may also be the case that some of them get additional utility from searching and/or combining different services. Regular users, for whom learning costs are relatively high due to their insufficient experience in using different systems and interfaces, are less likely to switch and prefer to stay with the same service.

I find this separation necessary, because intensive users and regular users can evaluate all their options differently: some regular users can overestimate the level of their learning costs and be locked-in to some particular portal even though it does not suit their needs in the best possible way. On the other hand, intensive users, while searching for the best possible choice, can use Internet portals in a more efficient way

and even combine services from different portals to get the highest possible utility level.

Several variables are used to describe the level of activity online. Following Nie and Erbring (2000) and Goldfarb (2000a), I constructed and analyzed the following variables: total number of visits, total time spent online, average time of visit (Figure 7.6).

Most investigators use the total time spent online as an indication of heavy / intensive usage. After we expanded the definition of intensive user, this variable is no longer relevant. High values for the total time spent online variable can not only be an indicator of intensive usage of Internet resources, but also a result of inexperienced and uneducated search and usage. A person who is not familiar with particular designs and systems can spent a considerable amount of time online before he achieves his goal; thus, total time spent will not be an indicator of intensive use, in contrast, it is an indicator of inefficiency. The same argument can be applied to average time online (per visit) since this variable is connected with the above. In this situation, we assume that a total number of visits can be the best possible indicator of user's activity. Individual going online often and performing different tasks can be identified as a heavy / intensive user as I defined it earlier.

I use natural cutoffs in the sample in order to determine the number of visits that can be an indicator of intensive usage. This cutoff point is 800 visits.

Using these indicators, I found that proportion of heavy Internet users, contrary to Nie and Erbring (2000), is relatively high and constitutes about 10% of all users in the dataset.

7.4.2. Heterogeneity test.

For the heterogeneity test, I estimate random coefficients and fixed coefficients models (Table 7.13). For the estimation, the following household characteristics were included: User age, User education, User income, Household size, Marital status and Renting. I expected the coefficient of education to be positive since more educated people usually have higher learning ability, therefore learning costs for them tend to be lower and they are more likely to be intensive users. Based on the results of Nie and Erbring (2000), I also expected the coefficient of age to be negative since learning costs usually increase for older people. These hypotheses hold for our sample, however, the coefficients are insignificant except for the User age in the fixed effects estimation.

Using the estimated random effect and fixed effect models, I performed the Hausman test with the null hypothesis H_0: differences in coefficients are not systematic. This hypothesis was rejected, which indicates that there are systematic differences between two groups of users. I use this result to support the assumption that Internet users' population is heterogeneous and the behavior of different groups of users should be modeled separately.

7.4.3. Kaplan-Meier estimation.

This section uses the Kaplan-Meier product limit estimators to further explore the differences between consumer groups. First, graphs of the Kaplan-Meier hazard curves are generated to show how the probability of dropping from the portal changes over time. Then, I use Kaplan-Meier estimators to generate a graph with separated hazard curves for intensive and regular users. Finally, Kaplan-Meier hazard curves are used to examine the effect of multiple spells on the probability of leaving the portal.

The Kaplan-Meier product limit estimator is a non-parametric method for estimation of hazard function. Let n_k denote the number of individuals whose observed duration is at least T_k, and let h_k denote the number of observed spells completed at time T_k. An empirical estimate of the hazard rate will be:

$$\hat{\lambda}(T_k) = \frac{h_k}{n_k} \tag{7.9}$$

And the estimate of the survivor function is:

$$\hat{S}(T_k) = \prod_{i=1}^{k} \frac{n_i - h_i}{n_i} \tag{7.10}$$

Figure 7.7 shows the Kaplan-Meier estimates of the hazard curve for dropping out from the portal. Both curves indicate that probability of dropping decreases over time; for Model II all exits appear before the week 6, which is the middle of our timeframe.

It is reasonable to assume that intensive users should have higher probability of switching and dropping from the portal. In order to verify this hypothesis, I use Kaplan-Meier estimates of hazard rates of intensive versus regular users (Figure 7.8). Model I demonstrates the clear distinction between hazard rates of different user groups. I detect 25% to 55% differences between dropping probabilities during the observed period of time. Log-rank test and Wilcoxon (Breslow) test were used to test the hypothesis of

equality of survivor functions (Tables 7.14 and 7.15); this hypothesis was rejected for Model I. Survival function comparison is presented in Table 7.16. This result, however, does not hold for the Model II, which does not demonstrate significant differences between hazard functions, and the hypothesis of equality of survivor functions cannot be rejected (Tables 7.17-7.19). Table 7.17 demonstrates that number of intensive users among switchers in Model II is about 88%, which is clearly not representative for the studied sample. For comparison, this number in Model I is only 59%.

To further extend this analysis, I also use Kaplan-Meier approach to the effect of multiple spells on the hazard function estimate. For Model I, the hypothesis of equality of survivor functions can be rejected with 2% significance level (Tables 7.20 and 7.21). One of the reasons can be the limited number of observations on multiple spells (1.26%). Table 7.22 and Figure 7.9 display survival functions for single and multiple spells. In Model II, no multiple spells are observed.

7.5. Knowledge spillover within the household.

On the way towards making his portal selection, Internet user is constantly bombarded by information from multiple sources: Internet banners, TV commercial, printed ads, as well as opinions and suggestions shares in online communities. Additional source of information he may find in his own home. Household presents an ideal environment for the information sharing and influencing. Within his own household, individual can get information as detailed as he want, and this information is coming from a trusted source.

Due to the nature of such information spillover, I hypothesize that increased number of switches will result from it. For example, a recommendation or even a strong opinion that certain portal is superior to others may lead to near automatic switching.

Parametric estimations did not suggest high degree of correlation between the number of household members and survival rates. It is important to understand the critical distinction between single- and non-single households, and incorporate this idea into the model. Here, I use Kaplan-Meier estimators to generate a graph with separated survival curves for users from single- and multiple-member households. The results, presented on Figure 7.9, suggest the difference in the switching and survival rates. Survival function comparison demonstrate that survival rates differ from 5 to 41% for the users from single- and multiple-member households (Table 7.28).

However, data does not suggest any patterns in switching for the users from multiple-member households. For example, in household X, user A switched to go.com on the 3rd week of observations and user B switched to the same portal on week 11; in household Y, user C switched to yahoo.com on the 3rd week of observations, user D switched there on the 4th week, and user E switched to yahoo.com on the 6th week of observations. At the same time, in the household Z, user F switched to yahoo.com on the 8th week of observations, and used G switched to msn.com on the 9th week of observations. The nature of the communication between the household members is unknown to us: it may well be that in the first two cases the individual who switched first led others to switching to the same portal; and in the third case, individuals explored different portals and used this shared knowledge to find the portal that fits them best.

The detailed study of the search patterns for different household members may provide interesting insights and contribute to the question of knowledge spillover. This created a beautiful new direction for the research and click-stream data provides the rich source for such study.

7.6. Conclusions.

In this chapter I use the survival analysis to address the question of how portal attributes and individual characteristics drive the droppings of customers from portals. Parametric estimations with different underlying hazard help identify the declining hazard rates for the dropping, and establish the connection between certain demographic characteristics, such as age and education, and probability of staying with the same portal. I also found that larger number of quality portal attributes improves portal's survival rates.

Using nonparametric estimation, I identify significant factors which were previously omitted from the studies of online behavior. Here I broaden the approach usually taken for the analysis of consumers' online behavior by defining two different user groups. Data confirms that regular users are less likely to switch from one portal to another than intensive users. This finding creates an interesting direction for future research of online behavior although it requires more information about individuals. These results also give important information for portal managers since better knowledge of customer base eventually leads to establishing closer connections between portal and customer, and following revenue increase.

Due to limited observations in the available data I was unable to fully explore the effect of multiple spells, but there are indicators suggesting that users under multiple spells are less likely to drop from the portal.

Also, patterns in the data provide the evidence of the knowledge spillover occurring between the household members. Individuals from multiple-member households demonstrate different switching pattern from singles.

Figure 7.1. Generic hazard rate.

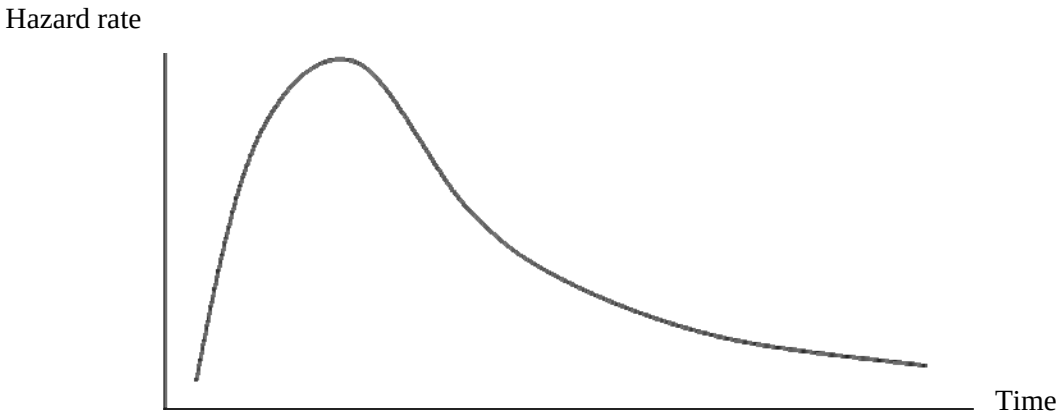
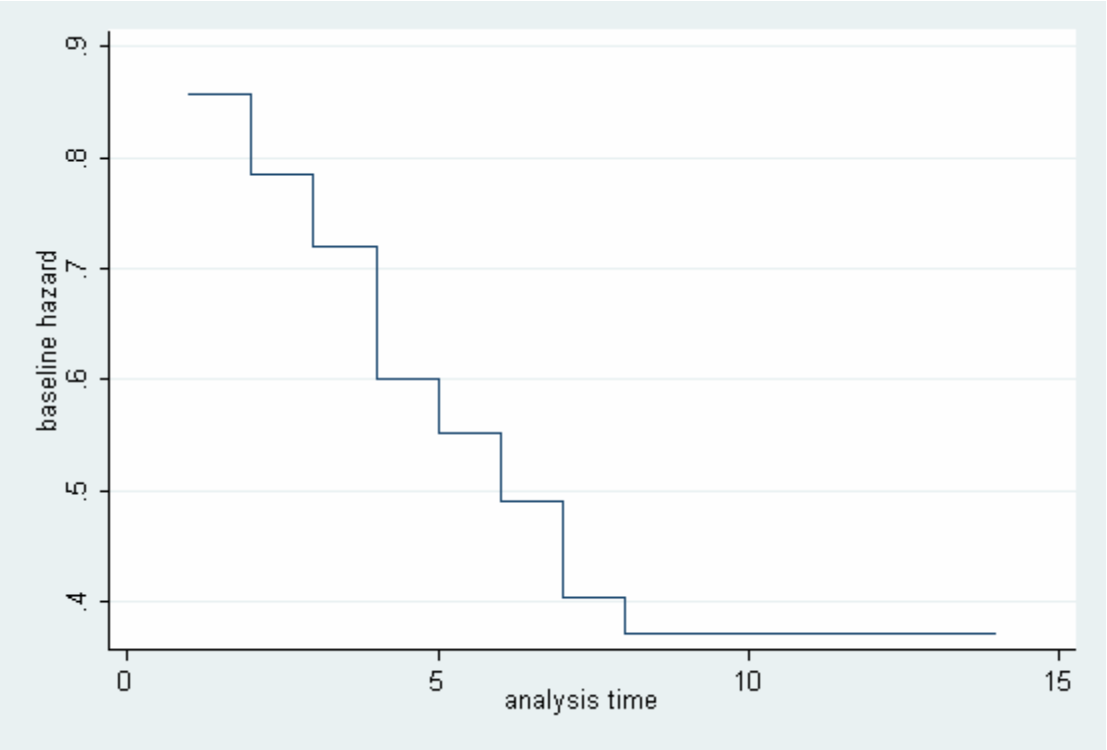


Figure 7.2. Cox proportional hazard.

a) Model I.



b) Model II.

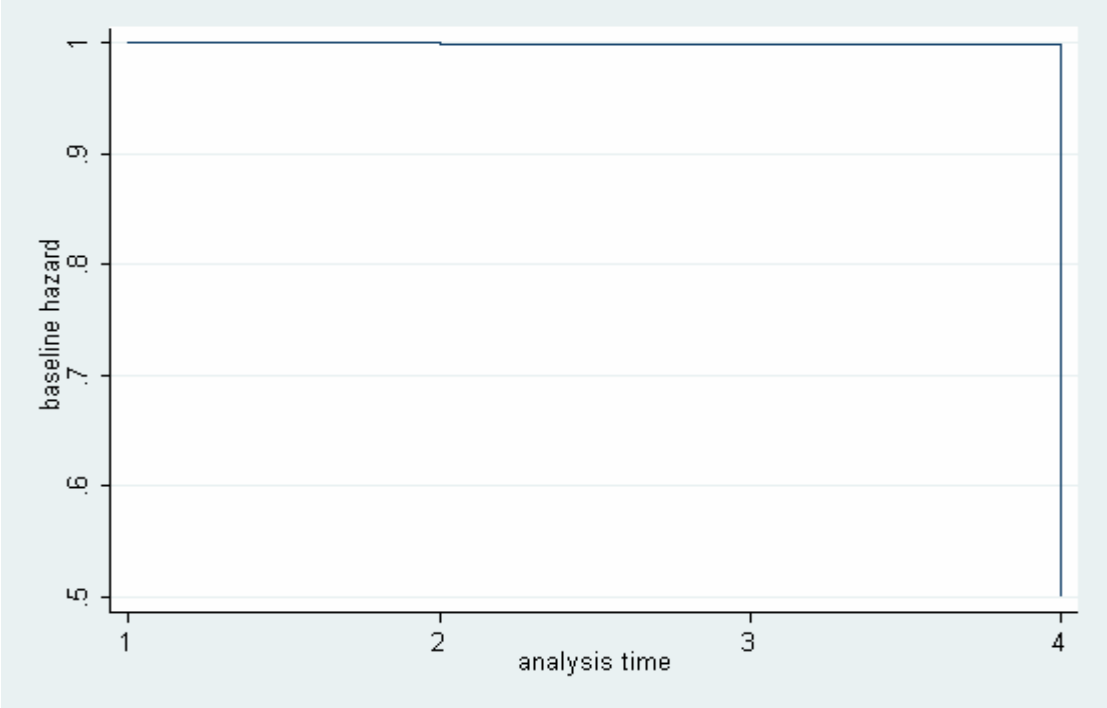
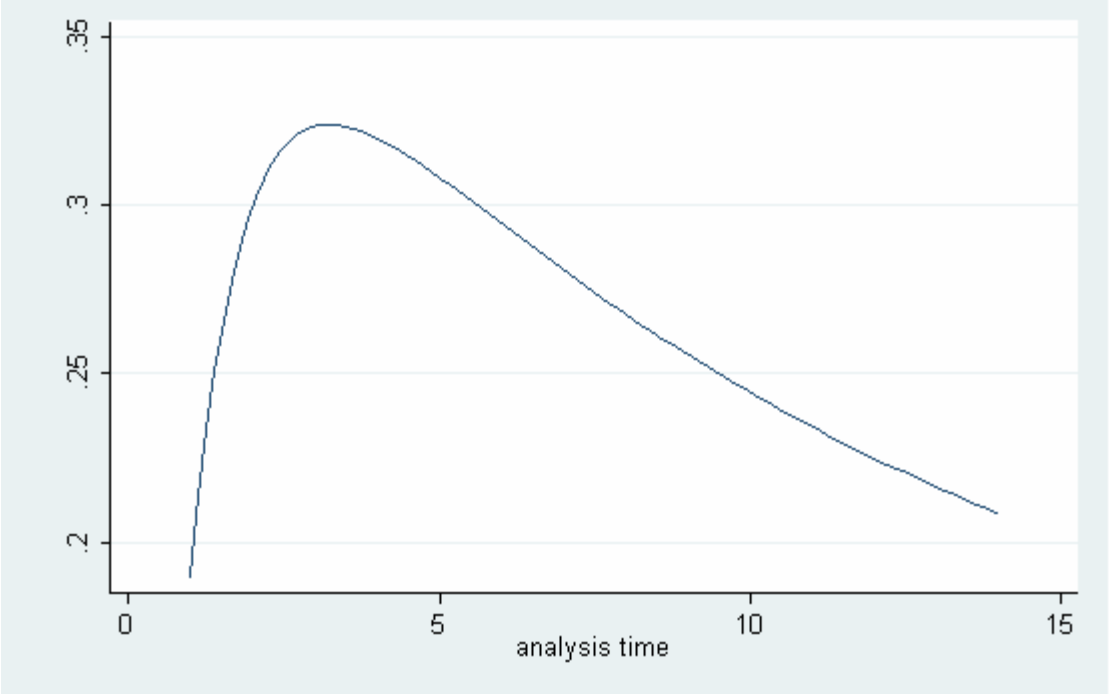


Figure 7.3. Lognormal estimated survival rates.
a) Model I.



b) Model II

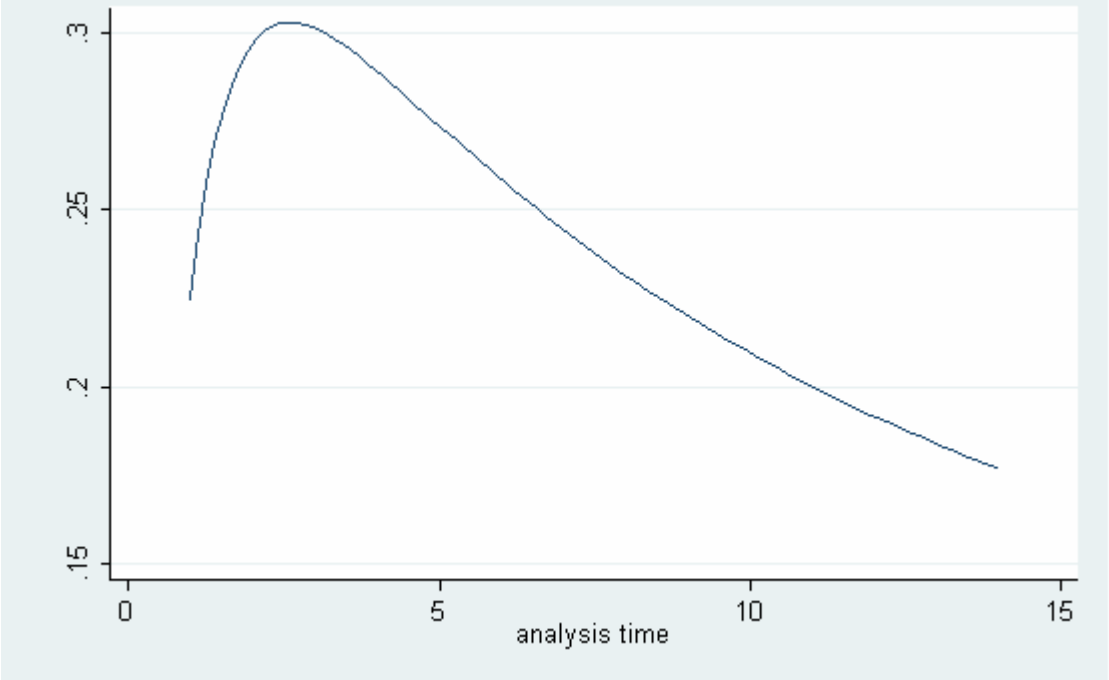
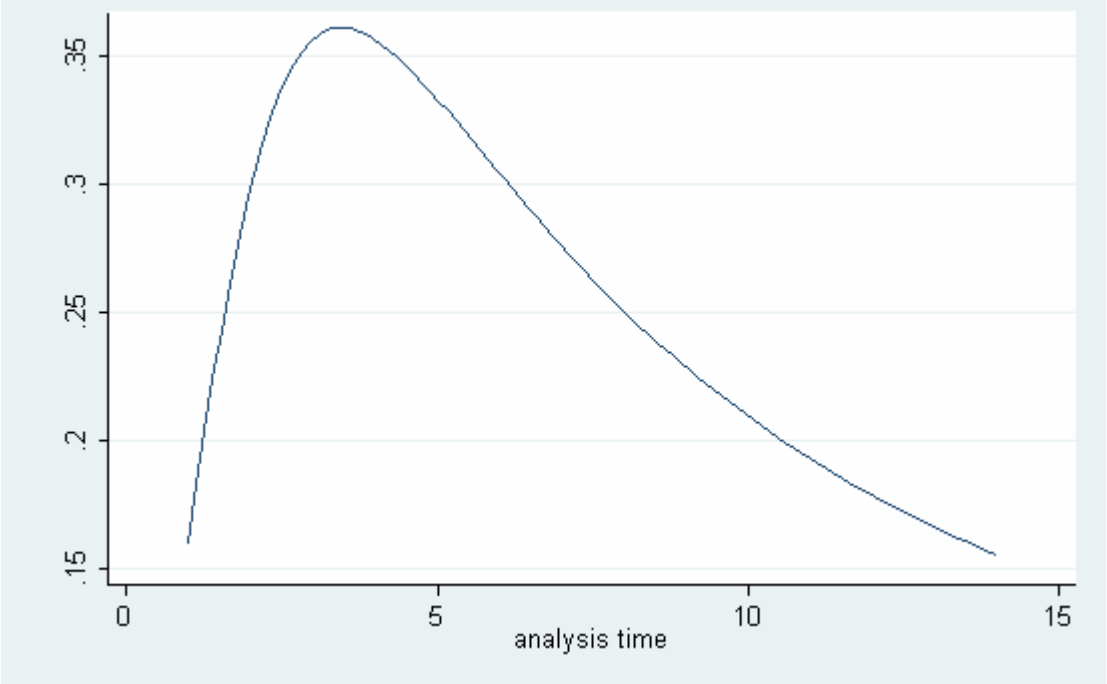


Figure 7.4. Loglogistic estimated survival rates.
a) Model I.



b) Model II

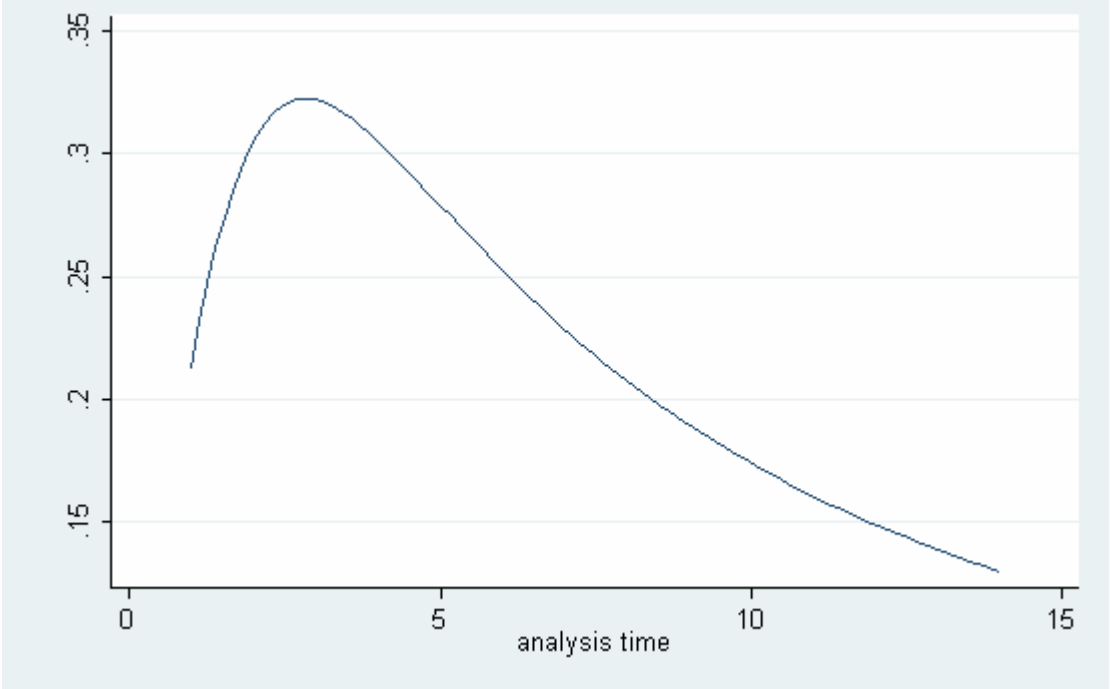
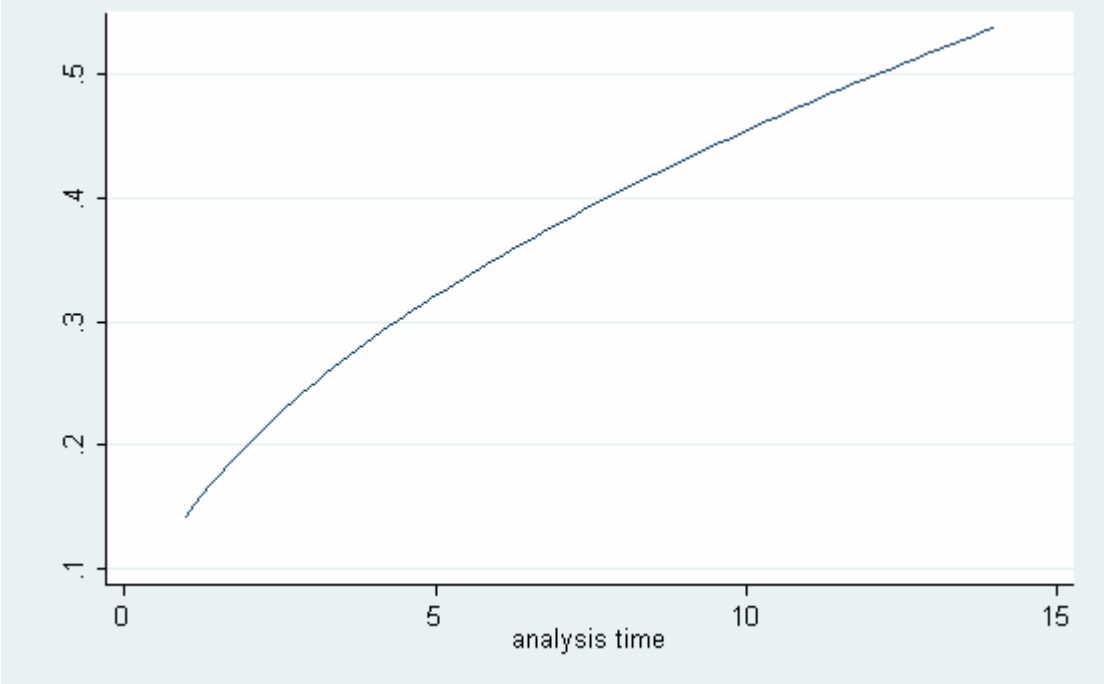


Figure 7.5. Weibull estimated survival rates.
a) Model I.



b) Model II

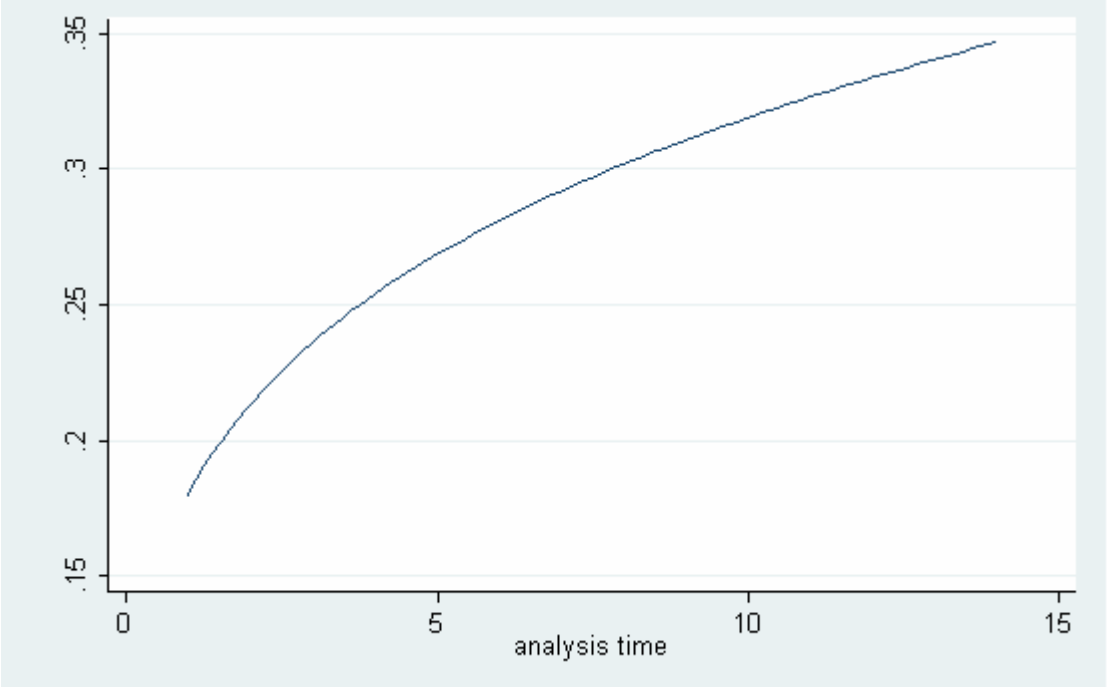
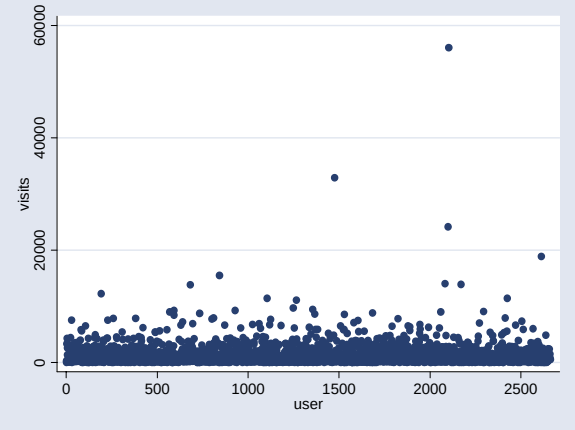
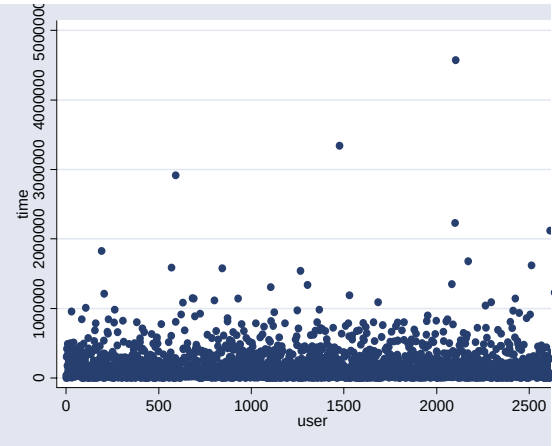


Figure 7.6. Indicators of user activity online.

a) Total number of visits



b) Total time online



c) Average time per session

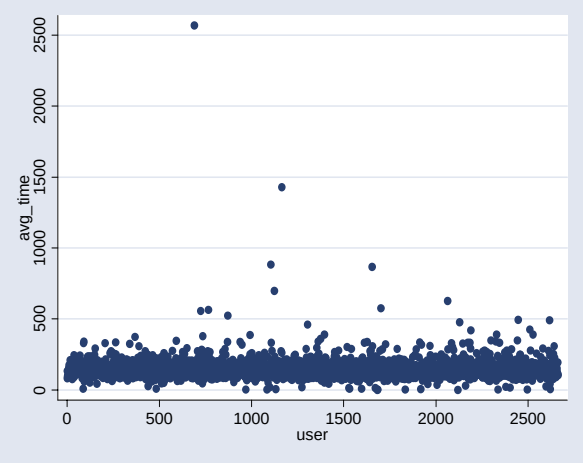
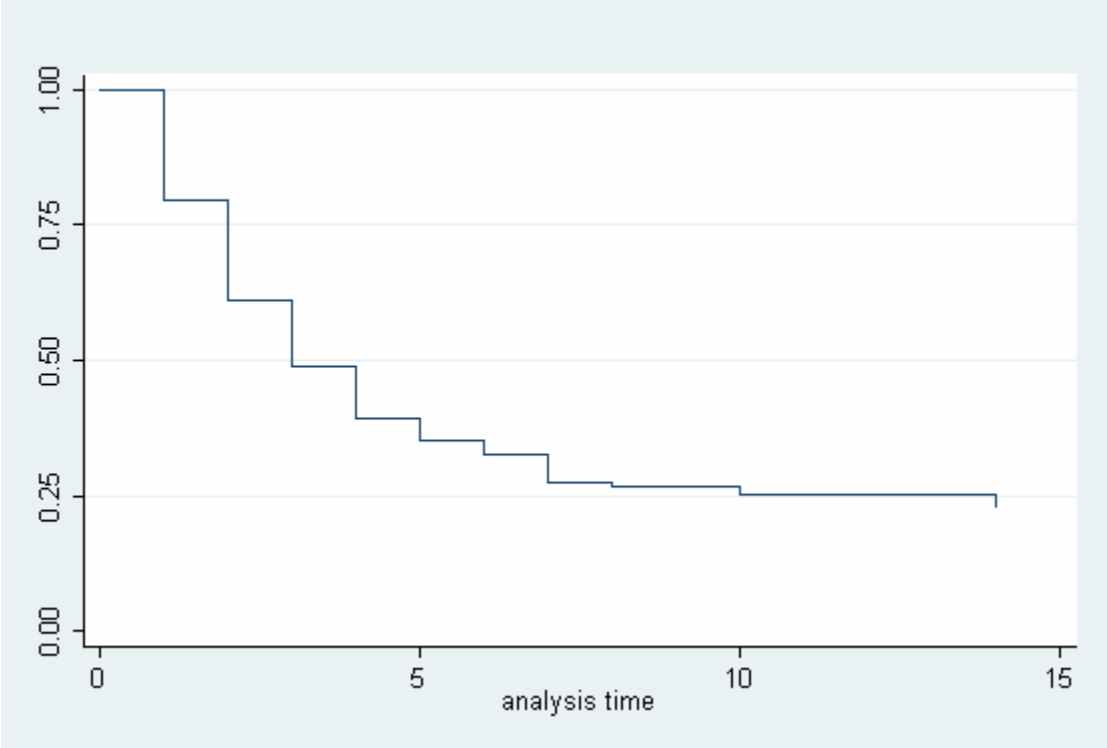


Figure 7.7. Kaplan-Meier survival estimate

a) Model I.



b) Model II

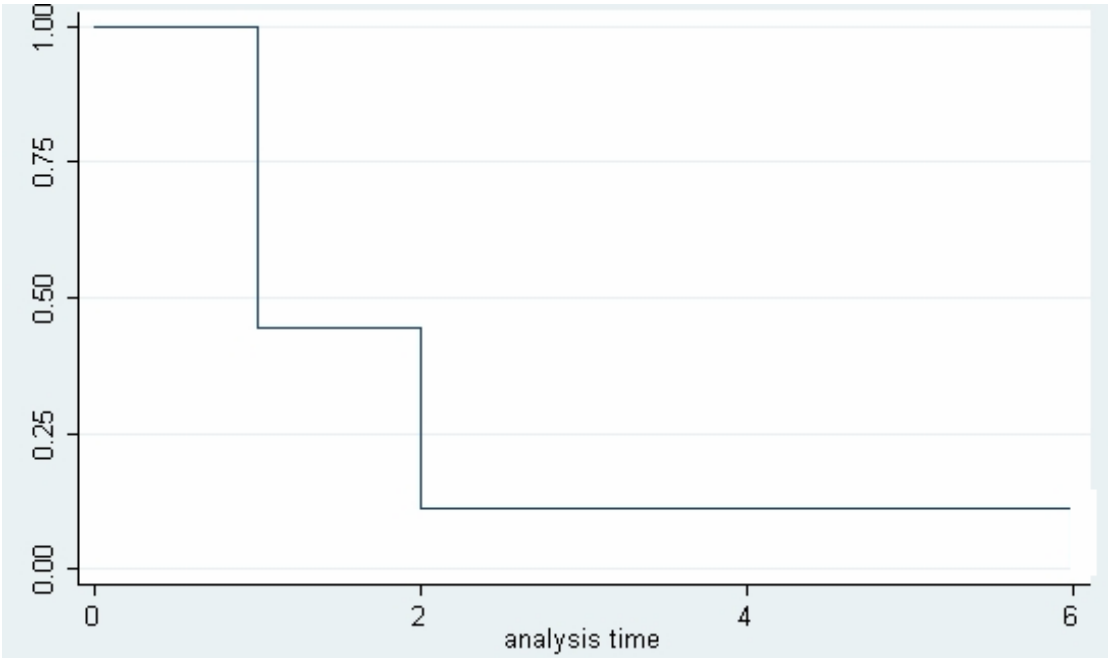


Figure 7.8. Kaplan-Meier survival estimates by user activity, Model I.

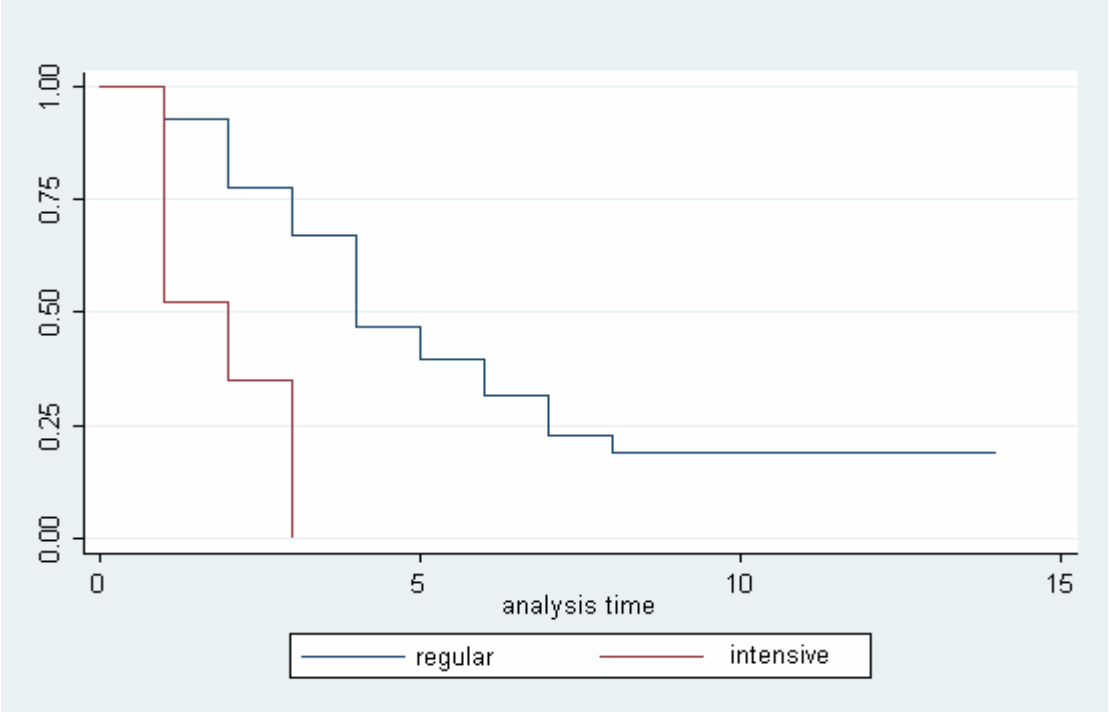


Figure 7.9. Kaplan-Meier survival estimates for multiple spells, Model I.

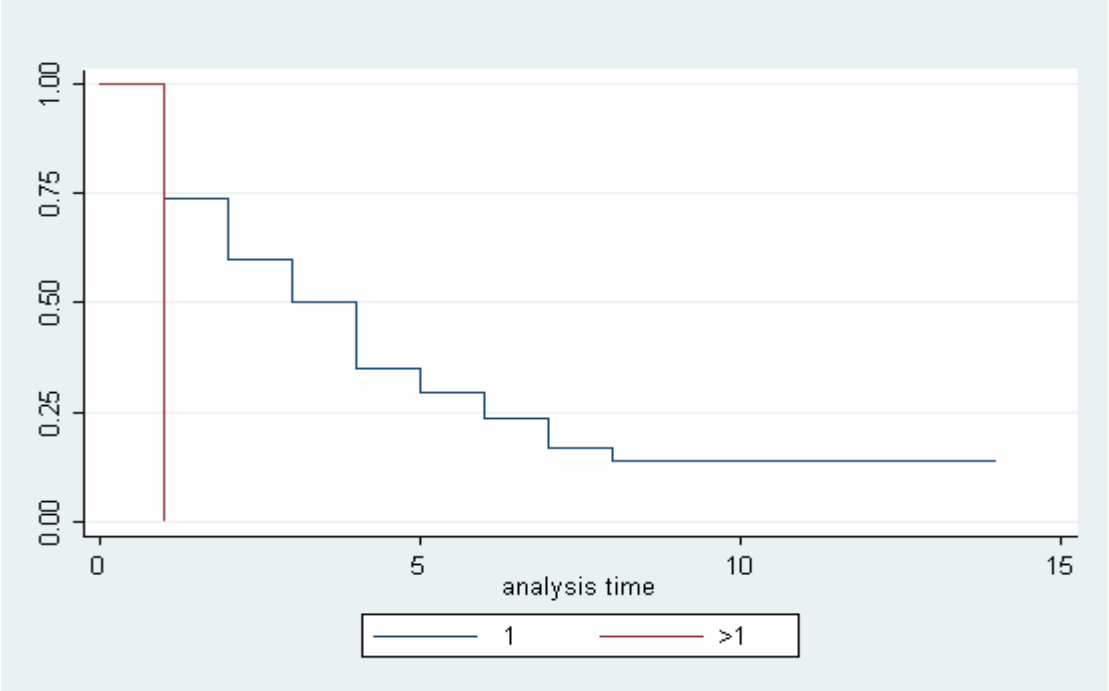


Figure 7.10. Kaplan-Meier survival estimates for single- and multiple-member households, Model I.

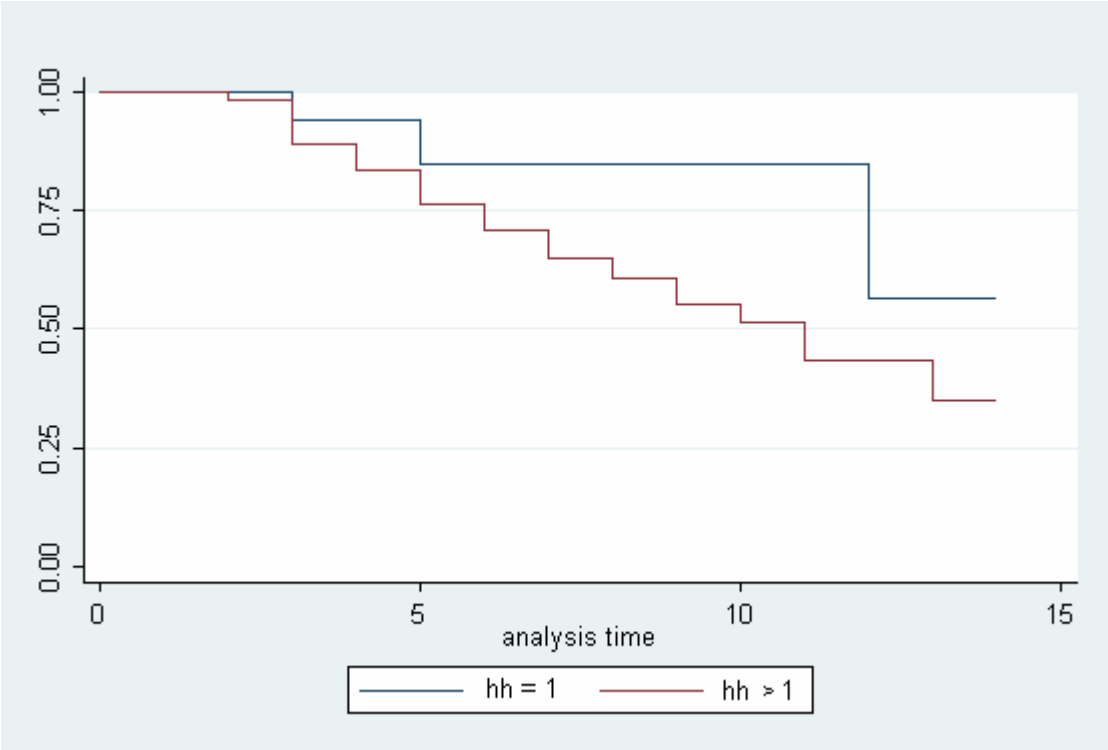


Table 7.1. The results of logit estimation, Model I (errors in parentheses).

Variable	Specification A	Specification B	Specification C
User age	.0075075***	.0053766 ***	.0062462 ***
	(.0005824)	(.0005764)	(.0005669)
User education	-.023665 ***	-.0339738 ***	-.0202157 ***
	(.0026586)	(.0026349)	(.0026012)
User income	8.64e-07 ***	1.37e-06 ***	-2.10e-07
	(1.85e-07)	(1.84e-06)	(1.80e-07)
Household size	.0444623 ***	.0519154 ***	.133543 ***
	(.0137378)	(.0136193)	(.0134394)
Marital status	.0727855	.1244754 ***	.0413253
	(.044968)	(.0445686)	(.0440167)
Renting	.3709179 ***	.3762524 ***	.3685617 ***
	(.0382296)	(.0378637)	(.0373927)
Portal age	.000174	. 001465 ***	. 0079701 ***
	(.0002981)	(.0000781)	(.0000425)
Mail	.2022451 ***	.2523511 ***	
	(.0072661)	(.0033118)	
Auction	.1510316 ***		
	(.0263139)		
Shopping	.6607609 ***		
	(.0254703)		
Sport	-.3056854 ***		
	(.0107633)		
Chat	-.2912507 ***		
	(.0326344)		
Greetings	-.0916447 ***		
	(.0282181)		
Games	-.3575351 ***		
	(.0281882)		
Finance	.4959419 ***		
	(.0365255)		
News	.1963918 ***		
	(.0333078)		
Search	.0215099 ***	. 0230896 ***	
	(.0008247)	(.0006435)	
Messenger	-.018844		
	(.0294434)		
Personal	.3931209 ***		
	(.0181102)		
Weather	.1399696 ***		
	(.0256308)		
Page	.1604911 ***		
	(.0204074)		
Virtual		. 1840279 ***	
		(.0032312)	
Personalization		-.0870895 ***	
		(.0029711)	
Constant	-1.599541 ***	-.9326521 ***	-1.237428 ***
	(.0772376)	(.0645327)	(.0633697)
Log - likelihood	-685163.94	-691928.78	-701208.98
LR χ^2	71181.11	57651.43	39091.03
Prob > χ^2	0.0000	0.0000	0.0000

*** significant at a 1% level ** significant at a 5% level * significant at a 10% level

Table 7.2.Logit odds ratios, Model I.

Variable	Specification A	Specification B	Specification C
User age	1.007536	1.005391	1.006266
User education	.9766128	.9665969	.9799873
User income	1.000001	1.000001	.9999998
Household size	1.045466	1.053287	1.14287
Marital status	1.0755	1.132554	1.042191
Renting	1.449064	1.456815	1.445654
Portal age	1.000174	1.001466	1.008002
Mail	1.224148	1.287048	
Auction	1.163033		
Shopping	1.936265		
Sport	.7366183		
Chat	.7473283		
Greetings	.9124293		
Games	.6993982		
Finance	1.642044		
News	1.217004		
Search	1.021743	1.023358	
Messenger	.9813324		
Personal	1.481598		
Weather	1.150239		
Page	1.174087		
Virtual		1.202049	
Personalization		.9165951	

Table 7.3. The results of logit estimation, Model II (errors in parentheses).

Variable	Specification A	Specification B	Specification C
User age	.0057452 ***	.0071595 ***	.0103928 ***
	(.0005473)	(.0005453)	(.00054)
User education	.0976407 ***	.0968859 ***	.1051905 ***
	(.0024899)	(.002472)	(.0024319)
User income	-4.60e-06 ***	-4.56e-06 ***	-5.47e-06 ***
	(1.78e-07)	(1.77e-07)	(1.73e-07)
Household size	.0811871 ***	.1010962 ***	.1611859 ***
	(.0130902)	(.0130304)	(.012846)
Marital status	.3935501 ***	.3716682 ***	.3073924 ***
	(.042445)	(.0422054)	(.0417746)
Renting	.287136 ***	.3292793 ***	.4275823 ***
	(.035758)	(.0355989)	(.0354147)
Portal age	-.0038255 ***	.0028982 ***	.0094692 ***
	(.0004719)	(.0001132)	(.0000413)
Mail	-.2401215 ***	.0804501 ***	
	(.013315)	(.0047492)	
Auction	-.163161 ***		
	(.0461967)		
Shopping	.9340096 ***		
	(.0481739)		
Sport	-.2280224 ***		
	(.0190127)		
Chat	.9643612 ***		
	(.0599907)		
Greetings	-.7298846 ***		
	(.0389534)		
Games	.5963828 ***		
	(.0531457)		
Finance	.7839588 ***		
	(.0685106)		
News	-.5788971 ***		
	(.0679219)		
Search	.0514873 ***	.0368534 ***	
	(.0010768)	(.000833)	
Messenger	1.330753 ***		
	(.0430122)		
Personal	-1.713185 ***		
	(.0255303)		
Weather	1.444843 ***		
	(..0329434)		
Page	1.71177 ***		
	(.0264966)		
Virtual		.6143679 ***	
		(.0050008)	
Personalization		.0267876 ***	
		(.0036379)	
Constant	-5.382378 ***	-6.264906 ***	-4.847968 ***
	(.0952913)	(.0626059)	(.0596654)
Log - likelihood	-620204.2	-626591.84	-648136.48
LR χ^2	115050.58	102275.31	59186.03
Prob > χ^2	0.0000	0.0000	0.0000

*** significant at a 1% level ** significant at a 5% level * significant at a 10% level

Table 7.4.Logit odds ratios, Model II.

Variable	Specification A	Specification B	Specification C
User age	1.005762	1.007185	1.010447
User education	1.102567	1.101735	1.110922
User income	.9999954	.9999954	.9999945
Household size	1.084574	1.106383	1.174903
Marital status	1.482234	1.450152	1.359874
Renting	1.332605	1.389966	1.533545
Portal age	.9961818	1.002902	1.009514
Mail	.7865323	1.083775	
Auction	.8494544		
Shopping	2.544692		
Sport	.7961064		
Chat	2.623111		
Greetings	.4819646		
Games	1.81554		
Finance	2.190125		
News	.5605162		
Search	1.052836	1.037541	
Messenger	3.783891		
Personal	.1802907		
Weather	4.241188		
Page	5.538755		
Virtual		1.848488	
Personalization		1.02715	

Table 7.5. The results of Cox proportional hazard estimation, Model I (errors in parentheses).

Variable	Specification A	Specification B	Specification C
User age	1.002756	1.002611	1.002227
	(.0046014)	(.0045973)	(.0045646)
User education	1.07955 ***	1.078953 ***	1.080176 ***
	(.0247914)	(.0246474)	(.0246435)
User income	.9999915 ***	.9999916 ***	.9999916 ***
	(1.74e-06)	(1.74e-06)	(1.74e-06)
Household size	1.289642 **	1.286906 **	1.283686 **
	(.149748)	(.1494457)	(.1486328)
Marital status	.882264 *	1.857555 *	1.881118 *
	(.6885853)	(.6761351)	(.6817539)
Renting	1.14878	1.202618	1.197114
	(.35053)	(.3658021)	(.3635192)
Portal age	1.001335	.9999246	1.000815 **
	(.0028522)	(.0009141)	(.0004148)
Mail	1.073152	1.04578	
	(.0818637)	(.0386977)	
Auction	.7922891		
	(.2382252)		
Shopping	1.73801 **		
	(.474012)		
Sport	1.176833		
	(.1624421)		
Chat	1.382391		
	(.4300414)		
Greetings	.6988392		
	(.1679053)		
Games	.7470849		
	(.2027455)		
Finance	1.157613		
	(.4046887)		
News	.5811116 *		
	(.1821)		
Search	1.006448	.9994643	
	(.0103742)	(.0076935)	
Messenger	.8347783		
	(.2269264)		
Personal	1.37102 *		
	(.2324055)		
Weather	1.276301		
	(.3014161)		
Page	.8560421		
	(.169309)		
Virtual		.9619538	
		(.0325587)	
Personalization		1.058883 **	
		(.0295068)	
Log - likelihood	-17996.998	-18011.385	-18013.848
LR χ^2	66.63	37.85	32.93
Prob > χ^2	0.0000	0.0001	0.0000

*** significant at a 1% level ** significant at a 5% level * significant at a 10% level

Table 7.6. The results of Cox proportional hazard estimation, Model II (errors in parentheses)

Variable	Specification A	Specification B	Specification C
User age	1.103491	1.094696	1.08944
	(.0679502)	(.0614019)	(.0590821)
User education	.9821421	.9222166	.8942754
	(.171865)	(.1502483)	(.1408451)
User income	.9999978	1.000001	1.000001
	(.0000103)	(9.48e-06)	(9.28e-06)
Household size	7.868902 **	5.811764**	5.290134 **
	(7.03618)	(4.778346)	(4.179189)
Marital status	.0003973 *	.0007358 *	.0009187 *
	(.001634)	(.0027705)	(.0033078)
Renting	.8838316	1.211202	1.56449
	(2.05386)	(2.553814)	(3.341275)
Portal age	.9946249	.9970977	1.004049
	(.0237637)	(.0103247)	(.0038639)
Mail	1.010771	1.041266	
	(1.531176)	(.3467389)	
Auction	.3903155		
	(1.495133)		
Shopping	4.557499		
	(7.523843)		
Sport	.4335273		
	(.6441456)		
Chat	.8507554		
	(1.063699)		
Greetings	dropped		
Games	dropped		
Finance	dropped		
News	dropped		
Search	1.159176 *	1.171248**	
	(.1009562)	(.0823752)	
Messenger	1.023186		
	(2.960949)		
Personal	3.240249		
	(11.05647)		
Weather	.6025411		
	(.9701235)		
Page	.1517549		
	(.361408)		
Virtual		.7011265	
		(.2543153)	
Personalization		.7988659	
		(.291089)	
Log - likelihood	-157.87369	-160.50345	-163.31088
LR χ^2	20.91	15.65	10.03
Prob > χ^2	0.2305	0.1547	0.1867

*** significant at a 1% level ** significant at a 5% level * significant at a 10% level

Table 7.7. Fit parametric survival models using lognormal underlying distribution, Model I (errors in parentheses)

Variable	Specification A	Specification B	Specification C
User age	-1.21e-06	-1.08e-06	-1.04e-06
	(3.65e-06)	(3.67e-06)	(3.67e-06)
User education	-.0000312 *	-.0000312 *	-.0000331 *
	(.000017)	(.0000171)	(.0000171)
User income	3.82e-09 ***	3.75e-09 ***	3.80e-09 ***
	(1.17e-09)	(1.18e-09)	(1.18e-09)
Household size	-.0001531 **	-.0001563 *	-.0001536 *
	(.0000882)	(.0000885)	(.0000885)
Marital status	-.000388	-.0003597	-.0003599
	(.0002886)	(.0002894)	(.0002898)
Renting	-.0001025	-.0001242	-.0001168
	(.000236)	(.0002367)	s(.000237)
Portal age	9.10e-07	3.48e-07	-5.71e-07 *
	(2.15e-06)	(7.04e-07)	(3.17e-07)
Mail	-.0000813	-.0000386	
	(.0000562)	(.000028)	
Auction	.0000771		
	(.0002159)		
Shopping	-.0004504 **		
	(.0002073)		
Sport	-.000228 **		
	(.0001)		
Chat	-.0000842		
	(.000239)		
Greetings	.0003027		
	(.0001862)		
Games	.0005268 **		
	(.0002133)		
Finance	-.0005052 *		
	(.0002645)		
News	.0004932 **		
	(.000242)		
Search	-.0000156 **	-3.56e-06	
	(7.89e-06)	(5.95e-06)	
Messenger	.000029		
	(.0002017)		
Personal	-.0001486		
	(.0001281)		
Weather	-.000304 *		
	(.0001835)		
Page	4.35e-06		
	(.0001475)		
Virtual		.0000414	
		(.0000259)	
Personalization		-.0000533 **	
		(.0000215)	
Constant	9.591088 ***	9.590847 ***	9.590855 ***
	(.0005188)	(.000416)	(.0004112)
Log - likelihood	15115.478	15101.729	15098.331
LR χ^2	52.04	24.54	17.75
Prob > χ^2	0.0001	0.0063	0.0069

*** significant at a 1% level ** significant at a 5% level * significant at a 10% level

Table 7.8. Fit parametric survival models using lognormal underlying distribution, Model II (errors in parentheses)

Variable	Specification A	Specification B	Specification C
User age	-.0000433 ***	-.0000404***	-.0000415*
	(.000016)	(.0000155)	(.0000154)
User education	.0000252	.0000314	.0000457
	(.0000487)	(.0000485)	(.0000483)
User income	9.75e-10	2.06e-10	. -3.27e-10
	(3.56e-09)	(3.45e-09)	(3.48e-09)
Household size	-.0008149 ***	-.0007328 ***	-.0007231***
	(.0002735)	(.0002654)	(.0002603)
Marital status	.0036561 ***	.0034726 ***	.0036094 ***
	(.0012805)	(.0012369)	(.0012152)
Renting	.0003352	.0003119	.0002169
	(.0006504)	(.0006444)	(.0006515)
Portal age	3.60e-07	-4.18e-07	-5.82e-07
	(7.43e-06)	(2.55e-06)	(1.04e-06)
Mail	.0000916	.0000283	
	(.000451)	(.0000946)	
Auction	-.0001996		
	(.0011339)		
Shopping	-.0003856		
	(.0004677)		
Sport	.0002709		
	(.0004337)		
Chat	.0001218		
	(.0003821)		
Greetings	dropped		
Games	dropped		
Finance	dropped		
News	dropped		
Search	-.0000251	-.0000262	
	(.0000265)	(.0000198)	
Messenger	-.0001004		
	(.000861)		
Personal	-.0001036		
	(.0010402)		
Weather	.0001278		
	(.0004898)		
Page	.0003332		
	(.0007461)		
Virtual		.0000753	
		(.0000982)	
Personalization		.0000363	
		(.0000929)	
Constant	9.591031***	9.590579***	9.590581***
	(.0013636)	(.0010493)	(.0010078)
Log - likelihood	365.24963	364.20733	363.10753
LR χ^2	19.63	17.54	15.34
Prob > χ^2	0.2938	0.0929	0.0319

*** significant at a 1% level ** significant at a 5% level * significant at a 10% level

Table 7.9. Fit parametric survival models using loglogistic underlying distribution, Model I (errors in parentheses)

Variable	Specification A	Specification B	Specification C
User age	1.09e-07	2.31e-07	2.28e-07
	(2.10e-06)	(2.11e-06)	(2.11e-06)
User education	-.000022 **	-.0000218 **	-.0000223 **
	(9.68e-06)	(9.68e-06)	(9.66e-06)
User income	2.49e-09 ***	2.45e-09 ***	2.46e-09 ***
	(6.65e-10)	(6.65e-10)	(6.63e-10)
Household size	-.0000611	-.0000584	-.0000582
	(.0000503)	(.0000503)	(.0000503)
Marital status	-.0002433	-.0002253	-.0002251
	(.0001654)	(.0001656)	(.0001657)
Renting	-.0000328	-.0000258	-.0000246
	(.0001347)	(.0001345)	(.0001345)
Portal age	-8.31e-07	-5.58e-09	-2.19e-07
	(1.17e-06)	(3.96e-07)	(1.76e-07)
Mail	-9.24e-06	-.0000109	
	(.0000326)	(.0000159)	
Auction	.0001129		
	(.000117)		
Shopping	-.0002101*		
	(.0001176)		
Sport	-.0000268		
	(.0000562)		
Chat	-.0001504		
	(.0001368)		
Greetings	.0001269		
	(.0001051)		
Games	.0000639		
	(.0001233)		
Finance	.0000105		
	(.0001523)		
News	.0001712		
	(.0001327)		
Search	-1.12e-06	. 5.58e-08	
	(4.47e-06)	(3.34e-06)	
Messenger	.0000746		
	(.000109)		
Personal	-.0000981		
	(.0000706)		
Weather	-.0000865		
	(.000102)		
Page	.0000531		
	(.0000807)		
Virtual		8.95e-06	
		(.0000145)	
Personalization		-.0000118	
		(.0000121)	
Constant	9.590352 ***	9.590181 ***	9.590191 ***
	(.000292)	(.0002365)	(.00023)
Log - likelihood	15982.169	15974.526	15973.883
LR χ^2	34.99	19.71	18.42
Prob > χ^2	0.0201	0.0321	0.0053

*** significant at a 1% level ** significant at a 5% level * significant at a 10% level

Table 7.10. Fit parametric survival models using loglogistic underlying distribution, Model II (errors in parentheses).

Variable	Specification A	Specification B	Specification C
User age	-.0000172 *	-.0000171 *	-.0000199 **
	(.0000101)	(9.54e-06)	(9.90e-06)
User education	-4.96e-06	-3.70e-06	4.67e-06
	(.0000292)	(.0000287)	(.0000288)
User income	1.86e-09	1.53e-09	1.44e-09
	(2.09e-09)	(2.04e-09)	(2.13e-09)
Household size	-.0003309 *	-.0003043*	-.0003514
	(.0001903)	(.0001803)	(.0001857)
Marital status	.0016622 **	.0017062 **	.0020102**
	(.0008348)	(.0007915)	(.0008064)
Renting	.0004727	.0005217	.0004399
	(.0004112)	(.0004109)	(.0004312)
Portal age	1.81e-06	7.89e-07	
	(4.06e-06)	(1.48e-06)	
Mail	-.0000964	-.0000238	-3.01e-07
	(.0002674)	(.0000578)	(6.27e-07)
Auction	.0001145		
	(.0006485)		
Shopping	-.0000697		
	(.000303)		
Sport	.0000948		
	(.000252)		
Chat	-.0000412		
	(.0002216)		
Greetings	dropped		
Games	dropped		
Finance	dropped		
News	dropped		
Search	-.0000165	-.0000223**	
	(.0000167)	(.0000117)	
Messenger	.0000947		
	(.000492)		
Personal	-.0003031		
	(.0006095)		
Weather	.0000965		
	(.0002667)		
Page	.0003645		
	(.0004132)		
Virtual		.0000553	
		(.0000586)	
Personalization		.0000395	
		(.0000528)	
Constant	9.589937 ***	9.589775***	9.589937 ***
	(.0008513)	(.0006793)	(.0008513)
Log - likelihood	382.78118	381.9988	380.1117
LR χ^2	16.06	14.49	10.72
Prob > χ^2	0.5197	0.2068	0.1513

*** significant at a 1% level ** significant at a 5% level * significant at a 10% level

Table 7.11. Fit parametric survival models using Weibull underlying distribution, Model I (errors in parentheses)

Variable	Specification A	Specification B	Specification C
User age	1.005332	1.00417	1.003218
	(.0045494)	(.0045809)	(.004534)
User education	1.053842 **	1.045901*	1.040609 *
	(.0245462)	(.0241071)	(.0237528)
User income	.9999941 ***	.9999946 ***	.9999944 ***
	(1.76e-06)	(1.75e-06)	(1.76e-06)
Household size	1.340051 **	1.332279 **	1.356389 **
	(.1611962)	(.1613105)	(.1637139)
Marital status	2.304086 **	2.432974 **	2.421583 **
	(.822743)	(.870294)	(.8614341)
Renting	1.372012	1.505305	1.610015
	(.4526559)	(.4946774)	(.5247387)
Portal age	.9960099	.9990327	1.001175 ***
	(.0031741)	(.000859)	(.0004178)
Mail	1.231335 ***	1.098939 ***	
	(.099784)	(.0385333)	
Auction	1.16273		
	(.3968055)		
Shopping	1.543599		
	(.4472222)		
Sport	1.827264 ***		
	(.2814804)		
Chat	.9497247		
	(.3082574)		
Greetings	.5847372 **		
	(.1393341)		
Games	.3509712 ***		
	(.1000242)		
Finance	3.142621 ***		
	(1.20374)		
News	.4580261 **		
	(.1549461)		
Search	1.041824 ***	1.007988	
	(.0107874)	(.0075083)	
Messenger	1.083096		
	(.3217067)		
Personal	1.337564		
	(.2373816)		
Weather	1.736736 **		
	(.3983344)		
Page	.9347668		
	(.1951302)		
Virtual		.8888395 ***	
		(.0291651)	
Personalization		1.160912 ***	
		(.031727)	
Log - likelihood	13717.985	13679.921	13663.524
LR χ^2	142.25	66.12	33.33
Prob > χ^2	0.0000	0.0000	0.0000

*** significant at a 1% level ** significant at a 5% level * significant at a 10% level

Table 7.12. Fit parametric survival models using Weibull underlying distribution, Model II (errors in parentheses)

Variable	Specification A	Specification B	Specification C
User age	1.293793 ***	1.257861***	1.248988***
	(.071901)	(.0663063)	(.0637055)
User education	.8083016	.7537501	.7326656 *
	(.159420)	(.140426)	(.1299292)
User income	1.000002	1.000006	1.000008
	(.0000114)	(.0000106)	(9.95e-06)
Household size	84.94872 ***	50.25047 ***	40.59485***
	(80.57678)	(44.6416)	(34.58548)
Marital status	5.23e-10 ***	4.51e-09 ***	8.01e-09***
	(2.36e-09)	(1.93e-08)	(3.32e-08)
Renting	.0399234	.089708	.1104196
	(.114988)	(.2431945)	(.290322)
Portal age	1.010491	1.00473	
	(.0240727)	(.0091475)	
Mail	.2835797	.8269278	1.002927
	(.4097874)	(.2649703)	(.003566)
Auction	6.309033		
	(23.11333)		
Shopping	17.31607 *		
	(26.20436)		
Sport	.238479		
	(.3360934)		
Chat	.2445067		
	(.3002129)		
Greetings	dropped		
Games	dropped		
Finance	dropped		
News	dropped		
Search	1.106078	1.079699	
	(.0947162)	(.0784136)	
Messenger	3.362926		
	(9.302731)		
Personal	.3924086		
	(1.306822)		
Weather	.6840612		
	(1.09687)		
Page	.593319		
	(1.406131)		
Virtual		.7370979	
		(.2617957)	
Personalization		.9830376	
		(.3024899)	
Log - likelihood	349.53211	346.4464	344.62999
LR χ^2	54.32	48.15	44.51
Prob > χ^2	0.0000	0.0000	0.0000

*** significant at a 1% level ** significant at a 5% level * significant at a 10% level

Table 7.13. Regression of activity indicator on household characteristics (errors in parentheses).

Variable	Fixed effects	Random effects
User age	12.9529**	13.1376
	(6.104168)	(8.527912)
User education	3.039871	10.25005
	(15.8493)	(22.13671)
User income	.0229072	.1821005
	(.187614)	(.26187)
Household size	-3.279436	-8.116515
	(4.790543)	(6.684059)
Marital status	-.004886	-.0409043
	(.0417416)	(.0582608)
Renting	-.0233558	-.1846721
	(.1878096)	(.2621372)
Constant	177.5104	282.5443
	(294.3113)	(411.1058)

*** significant at a 1% level ** significant at a 5% level * significant at a 10% level

Hausman test: Ho: difference in coefficients not systematic

χ^2 (5) = -2.33

$\chi^2 < 0$ => model fitted on these data fails to meet the asymptotic assumptions of the Hausman test

Breusch and Pagan Lagrangian multiplier test for random effects:
visits[t,t] = Xb + u[t] + e[t,t]
Estimated results:

	Variance	Standard deviation
visits	1235498	1111.53
e	636182.3	797.6104
u	0	0

Test: Var (u) = 0

χ^2 (1) = 2917.19

Prob > χ^2 = 0.0000

Table 7.14. Log-rank test for equality of survivor functions by user activity, Model I.

User activity	Events observed	Events expected
regular	186	244.47
intensive	130	71.53
Total	316	136.00

χ^2 (1) = 133.21

Pr > χ^2 = 0.0000

Table 7.15. Wilcoxon (Breslow) test for equality of survivor functions by user activity, Model I.

User activity	Events observed	Events expected	Sum of ranks
regular	186	244.47	-16680
intensive	130	71.53	16680
Total	316	136.00	0

χ^2 (1) = 133.21

Pr > χ^2 = 0.0000

Table 7.16. Kaplan-Meier survival function comparison by user activity, Model I.

		Survival function	
User activity		regular	intensive
time	1	0.9268	0.5250
	2	0.7765	0.3500
	3	0.6694	0.0000
	4	0.4686	.
	5	0.3965	.
	6	0.3172	.
	7	0.2266	.
	8	0.1888	.
	9	0.1888	.
	10	0.1888	.
	11	0.1888	.
	12	0.1888	.
	13	0.1888	.
	14	0.1888	.

Table 7.17. Log-rank test for equality of survivor functions by user activity, Model II.

User activity	Events observed	Events expected
regular	8	8.44
intensive	1	0.56
Total	9	9.00

χ^2 (1) = 0.80

Pr > χ^2 = 0.3711

Table 7.18. Wilcoxon (Breslow) test for equality of survivor functions by user activity, Model II.

User activity	Events observed	Events expected	Sum of ranks
regular	8	8.44	- 4
intensive	1	0.56	4
Total	9	9.00	0

χ^2 (1) = 0.80

Pr > χ^2 = 0.3711

Table 7.19. Kaplan-Meier survival function comparison by user activity, Model II.

		Survival function	
User activity		regular	intensive
time	1	0.5000	0.00
	1.625	0.5000	.
	2.25	0.1250	.
	2.875	0.1250	.
	3.5	0.1250	.
	4.125	0.1250	.
	4.75	0.1250	.
	5.375	0.1250	.
	6	0.0000	.

Table 7.20. Log-rank test for equality of survivor functions of multiple spells estimation, Model I.

Spells	Events observed	Events expected
1	312	314.23
2	4	1.77
Total	316	316.00

χ^2 (1) = 5.08

Pr > χ^2 = 0.0242

Table 7.21. Wilcoxon (Breslow) test for equality of survivor functions of multiple spells estimation, Model I.

Spells	Events observed	Events expected	Sum of ranks
1	312	314.23	-704
2	4	1.77	704
Total	316	316.00	0

χ^2 (1) = 5.08

Pr > χ^2 = 0.0242

Table 7.22. Kaplan-Meier survival function comparison of multiple spells estimation, Model I.

		Survival function	
Spells		1	2
time	1	0.7375	0.0000
	2	0.6003	.
	3	0.5002	.
	4	0.3502	.
	5	0.2963	.
	6	0.2370	.
	7	0.1693	.
	8	0.1411	.
	9	0.1411	.
	10	0.1411	.
	11	0.1411	.
	12	0.1411	.
	13	0.1411	.
	14	0.1411	.

Table 7.23. Log-rank test for equality of survivor functions of multiple spells estimation, Model II.

Spells	Events observed	Events expected
1	9	9.00
Total	9	9.00

χ^2 (0) = 0.00

Pr > χ^2 = .

Table 7.24. Wilcoxon (Breslow) test for equality of survivor functions of multiple spells estimation, Model II.

Spells	Events observed	Events expected	Sum of ranks
1	9	9.00	0
Total	9	9.00	0

χ^2 (0) = 0.00

Pr > χ^2 = .

Table 7.25. Kaplan-Meier survival function comparison of multiple spells estimation, Model II.

		Survival function
Spells		1
time	1	0.4444
	1.625	0.4444
	2.25	0.1111
	2.875	0.1111
	3.5	0.1111
	4.125	0.1111
	4.75	0.1111
	5.375	0.1111
	6	0.0000

Table 7.26. Log-rank test for equality of survivor functions of estimation for single- and multiple-member households, Model I.

Household members	Events observed	Events expected
1	3	5.79
>1	33	30.21
Total	36	36

χ^2 (1) = 1.72

Pr > χ^2 = 0.19

Table 7.27. Wilcoxon (Breslow) test for equality of survivor functions of estimation for single- and multiple-member households, Model I.

Household members	Events observed	Events expected	Sum of ranks
1	3	5.79	-200
>1	33	30.21	200
Total	36	36	0

χ^2 (1) = 1.48

Pr > χ^2 = 0.2240

Table 7.28. Kaplan-Meier survival function comparison of estimation for single- and multiple-member households, Model I.

		Survival function	
Household members		1	>1
time	1	1.0000	1.0000
	2	1.0000	. 0.9835
	3	0.9412	. 0.8883
	4	0.9412	. 0.8336
	5	0.8471	. 0.7642
	6	0.8471	. 0.7096
	7	0.8471	. 0.6504
	8	0.8471	. 0.6056
	9	0.8471	. 0.5529
	10	0.8471	. 0.5161
	11	0.8471	. 0.4367
	12	0.5647	. 0.4367
	13	0.5647	. 0.3493
	14	0.5647	. 0.3493

Chapter 8. Conclusions and policy implications.

Over the last decade Internet size has grown tremendously, as well as its role in everyone's daily life. In this thesis I explored different aspects of customer behavior online and its influence on portal competition.

First, I utilized the panel data analysis to define the determinants of portal market shares. This is essential because market shares are an important factor of online firm profitability. The results suggest that both portal features and individual user characteristics affect the overall market share, but portal attributes exert a higher degree of influence. In general, a larger number of portal features raises its market share; each additional feature may lead to up to 3-4% increase.

By separating the market shares for most popular portal services, I explored the differences between consumers using these services. I found that e-mail service is universal and its market share is not influenced by user characteristics, but is strongly affected by mail quality. This result suggests that wishing to attract more customers to mail service portal must improve its quality, and control the quality of other portal attributes since mail is interconnected with some of them. Appearance of Gmail on the mail market did not change the leadership on it much, although Google presented the largest mailbox volume and the best mail search tools. But the main existing players (Yahoo, AOL and MSN) matched the size of the mailbox and were able to keep their leading positions due to the number of features connected to their mail service and never presented by Google.

The quality of search is the major determinant of search service market share, which is also influenced by several demographic characteristics such as user age, education and income. Improving the ease of search process together with advancing the search quality are crucial for gaining market share; and current success of Google on the search engine

market confirms this finding.

Market shares of virtual communities are driven by the consumer demographic characteristics to a large extent. Moreover, the volatility of these market shares suggests that communities are different from portal to portal, producing an important implication for online firms. By finding more information about members of a particular community, a portal can establish tighter connection with them; also, online advertising can be targeted for the members of this community.

Later I apply the survival analysis approach to study the switching behavior of individual users. This dissertation is among the first to explore online consumer heterogeneity. I divide Internet users into two distinct groups and find that they tend to demonstrate different hazard rates of dropping. Having the higher dropping rates, intensive users generate more revenue for the online firm by being exposed to the larger number of posted banners. One of the goals of portal managers should be attracting such users by increasing the number and quality of portal characteristics.

Older and less educated people have lower tendency to drop from the portal and eventually come to another one. This result indicates learning costs exist on portal market, imposing barriers to switching. Online firm may overcome this with increasing ease-of-use and user friendliness of the portal, thus making itself more attractive for new users.

Hazard rates of dropping from the portal are found decreasing over time, suggesting that with time, users are going to be harder to attract, either because of the switching costs or due to the fact that they settled at their “ideal” portal.

The results of the panel data estimation and the ideas developed in this part of my dissertation can be successfully applied to the growing market of the mobile portals. As mobile connectivity grows at a dramatic pace worldwide, many companies have invested into development of mobile portals, which lead to an emergence of a separate mobile portal marketplace. Such mobile portals are created by cellular phone providers, major e-commerce sites (which are moving into mobile commerce or “M-commerce”), “classic” portal companies as well as new startups with dedicated mobile portal technology. Mobile portals are trying to capitalize on the ongoing trend to perform many tasks not just online, but online via a mobile device, such as a cell phone or a PDA.

Since widespread availability of mobile connectivity came a few years after fast traditional Internet connections became popular, the state of the mobile portal market and technology lags behind that of classic portals, such as Yahoo!, MSN, Google and others by at least five years. As a result, many of the trends that affected the Internet portal market in 2000 are occurring in the mobile portal market today.

Finally, the methods developed in this dissertation can be effectively applied on new, more detailed click stream data from today's portals. Such data contains a specific advertising message clicked by the user, a full URL link accessed, as well as information about the referring site that sent the traffic to the advertiser. For the previous years, when portals generated most of the profit from the banners featured at their page, the number of views was vital source of profit. As Internet advertisement revenue model further shits from paying per-ad display to paying per-click and per-purchase, combining the above detailed data with this approach developed in this work will allow us to study specific revenue sources, track successful ad campaigns and thus make more accurate policy recommendations.

References

Bharat N. Anand and Ron Shachar (2000). “Brands: Information and Loyalty.” Harvard Business School Working Paper 00-096.

J. Yannis Bakos (1997). “Reducing Buyer Search Costs: Implications for Electronic Marketplaces.” Management Science, Vol.43, No.12.

Barr, R., Seiford, L., Siems, T., 1994, “Forecasting Bank Failure: A Non-Parametric Frontier Estimation Approach.” Researches Economiques de Louvain, 60, pp. 417-429.

Alan Beggs (1989). “A Note on Switching Costs and Technology Choice.” The Journal of Industrial Economics, Vol.XXXVII, No.4, pp.437-439.

A. Beggs and Paul Klemperer (1992). "Multiperiod Competition with Switching Costs." Econometrica, 60

D. Scott Bennett, 1999, “Parametric Models, Duration Dependence and Time-Varying Data Revisited.” American Journal of Political Science, Vol. 43, No. 1, 256-270.

Ruth N. Bolton (1998). “A Dynamic Model of the Duration of the Customer’s Relationship with a Continuous Service Provider: The Role of Satisfaction.” Marketing Science, Vol.17, Issue 1, pp.45-65.

Severin Borenstein (1991). “Selling Costs and Switching Costs: Explaining Retail Gasoline Margins”. RAND Journal of Economics, Vol. 22, No.3.

William Boulding and Richard Staelin (1990). “Environment, Market Share, and Market Power”, Management Science, Vol. 36, No. 10, Focussed Issue on the State of the Art in Theory and Method in Strategy Research, pp. 1160-1177

Eileen Bridges, Chi Kin (Bennett) Yim and Richard A Briesch (1995). “A High-Tech Product Market Share Model with Customer Expectations.” Marketing Science, Vol.14, No.1, pp.61-81.

Erik Brynjolfsson and Michael D. Smith (2000). “The Great Equalizer? Consumer Choice Behavior at Internet Shopbots.” Working paper, MIT Sloan School of Management.

Randolph E. Bucklin and Catarina Sismeiro (2001). “A Model of Web Site Browsing Behavior Estimated on Clickstream Data.” Working paper, UCLA, Anderson School.

Enid Burns (2006). “MySpace Rules the Web.” ClickZNetwork, July 11.

Robert D. Buzzell and Frederick D. Wiersema (1981). “Modelling Changes in Market Share: A Cross-Sectional Analysis.” Strategic Management Journal, Vol.2, No.1, pp. 27-42.

Pei-Yu (Charon) Chen and Lorin M. Hitt (2001). “Measuring Switching Costs and Their Determinants in Internet -Enabled Business: A Study of the Online Brokerage Industry”. Working paper, University of Pennsylvania, Wharton School.

Cole, R. and Gunther, J., 1995, “Separating the Likelihood and Timing of Bank Failure.” Journal of Banking and Finance, 19,1073-89.

Victor J. Cook, Jr (1985). “The Net Present Value of Market Share.” Journal of Marketing, Vol. 49, No. 3, pp. 49-63

Cox, D., 1972, “Regression Models and Life Tables.”Journal of the Royal Statistical Society, series B, 34, 187-220.

TüLin Erdem and Michael P. Keane (1996). “Decision-making Under Uncertainty: Capturing Dynamic Brand Choice Processes in Turbulent Consumer Goods Market.” Marketing Science, Vol.15, No.1, pp.1-20.

Farenwell, V., 1978, “An Application of the Cox’s Proportional Hazard Model to Multiple Infection Data.” Applied Statistics, 28, 2, 136-143.

C. Davis Fogg (1974). “Planning Gains in Market Share”. Journal of Marketing, Vol. 38, No. 3, pp. 30-38.

Tommy S. Gabrielsen and Steinar Vagstad (2000). “Consumer Heterogeneity and Pricing in a Duopoly with Switching Costs.” Working paper, University of Bergen.

John M. Gallaugher and Charles E. Downing (2000). “Portal Combat: An Empirical Study of Competition in the Web Portal Industry”. Journal of Information Technology Management, Vol.11, No.1-2.

Avi Goldfarb (2002a). “Using Household-Specific Regression to Estimate True State dependence at Internet Portals.” Mimeo, Northwestern University.

Avi Goldfarb (2002b). “Advertising, Profits, Switching Costs, and the Internet.” Mimeo, Northwestern University.

Avi Goldfarb (2003). “Swithching Costs or Changing Preferences? Understanding the Effects of Denial of Service Attacks.” Mimeo, University of Toronto, Rotman

School of Management.

Gonzalez-Hermosillo, B., Pazarbasioglu, C., Billings, R., 1996, “Banking System Fragility: Likelihood Versus Timing of Failure - An Application to the Mexican Financial Crisis.” International Monetary Fund Working Paper: WP/96/142.

William H. Greene (1997). Econometric Analysis, 3rd ed. Prentice Hall, NJ

Hackers, Hits and Chats: An E-Commerce and Marketing Dictionary of Terms. http://www.udel.edu/alex/dictionary.html#dig

Hwang, D., Lee, C., Liaw, T., 1997, “Forecasting Bank Failures and Deposit Insurance Premium.” International Review of Economics and Finance , 6, 317-334.

Mark E. Jacobson, Charles J. Mode, 1985 “A Computer-Generated Model of Human Survival Functions”, The American Statistician, Vol 39, No. 2, p.145

Robert Jacobson (1988). “Distinguishing among Competing Theories of the Market Share Effect”. Journal of Marketing, Vol. 52, No. 4, pp. 68-80

Robert Jacobson and David A. Aaker (1985). “Is Market Share All That It's Cracked up to Be?” Journal of Marketing, Vol. 49, No. 4, pp. 11-22

A. Jesdanun (2007). “Nielsen scraps Web page view rankings.” Associated Press, July 9.

Eric J. Johnson, Wendy W. Moe, Peter S. Fader, Steven Bellman and Gerald L. Lohse (2001). “On the Depth and Dynamics of Online Search Behavior.” Working paper, University of Pennsylvania, Wharton School.

Robert Jaques (2007). “Microsoft revamps MSN Mobile portal”, iTnews.com.au, June 19.

Ashkan Karbasfrooshan (2007). “Merger of the Titans: Should Google Acquire Yahoo?” SeekingAlpha, July 2.

Peter Kennedy (2001). A Guide to Econometrics, 4th ed. MIT press, MA

Kiefer, N., 1985, ”Econometric Analysis of Duration Data.” Journal of Econometrics , 28, 1, 1- 169.

Moshe Kim, Doron Kliger and Bent Vale (2001). “Estimating switching costs and oligopolistic behavior.” Mimeo, The University of Haifa.

Paul Klemperer (1987). “Markets with Consumer Switching Costs.” The Quarterly Journal of Economics, 102.

Paul Klemperer (1995). “Competition when Consumers have Switching Costs: An

Overview with Applications to Industrial Organization, Macroeconomics, and International Trade." Review of Economic Studies, 62.

C. R. Knittel (1997). "Interstate Long Distance Rate: Search Costs, Switching Costs, and Market Power." Review of Industrial Organization, 12.

Philip Kotler (1999). "Principles of Marketing." Prentice Hall College Div., 8th edition.

Kryzanowski, L. and Roberts, G.S., 1993, "Canadian Banking Solvency, 1922-1940." Journal of Money, Credit and Banking, 25, 3, 1, 361-376.

Lancaster, T., 1990, "The Analysis of Transition Data." New York: Cambridge University Press.

John G. Lynch, Jr. and Dan Ariely (2000). "Wine Online: Search Costs Affect Competition on Price, Quality and Distribution." Marketing Science, Vol.19, No. 1.

Richard B. Maffei (1960). "Brand Preferences and Simple Markov Process." Operational Research, Vol. 8, Issue 2.

Daniel L. McFadden (1974) "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka (ed.), Frontiers in Econometrics, 105-142, Academic Press: New York.

John B. Meisel (1981). "Entry, Mulitiple-Brand Firms and Market Share Instability". The Journal of Industrial Economics, Vol. 29, No. 4, pp. 375-384

Jason Lee Miller (2007). Google's Better, But They Like Yahoo. WebProNews.com, June 28.

Aneel Karnani (1982). "Equilibrium Market Share-A Measure of Competitive Strength". Strategic Management Journal, Vol. 3, No. 1 (Jan., 1982), pp. 43-51

Niall Kennedy (2006). "Yahoo! is top portal", Niall Kennedy's Weblog, May 19

Nitin Mehta, Surendra Rajiv and Kannan Srinivasan (2001). "Active Versus Passive Loyalty: A Structural Model of Consideration Set Formation."

Wendy W. Moe and Peter S. Fader (2002). "Capturing Evolving Visit Behavior in Clickstream Data". Working paper, University of Pennsylvania, Wharton School.

Nickolay Moshkin and Ron Shachar (2000) "Switching Costs or Search Costs?" The Foerder Institute for Economic Research Working Paper No. 3-2000.

Phillip Nelson (1970). "Information and Consumer Behavior." The Journal of Political Economy, Vol. 78, Issue 2.

Norman H. Nie and Lutz Erbring (2000). “Internet and Society: A Preliminary Report.” Mimeo, Stanford Institute for the Quantitative Study of Society.

Yung-Hoon Park and Peter S. Fader (2002). “Modeling Browsing Behavior at Multiple Websites”. Working paper, University of Pennsylvania, Wharton School.

Bill Ray (2007). “Yahoo! and MSN up the mobile browsing stakes.” The Register, June 20.

M. Reardon (2007). “Broadband providers looking for sweeter deals?” ZDNetnews, March 20.

Keith Regan (2006). “Report: Google Leads in Search but Trails Yahoo in Portal Efforts”, E-Commerce Times, May 22

Paul A. Ruud (2000). An Introduction to Classical Econometric Theory. Oxford University Press, NY

Joachim Schwalbach (1991). “Profitability and Market Share: A Reflection on the Functional Relationship”. Strategic Management Journal, Vol. 12, No. 4 (May, 1991), pp. 299-306

Tyler Shumway (1999). “Forecasting Bankruptcy More Accurately: A Simple Hazard Model.” Working Paper, University of Michigan Business School.

Charles C. Slater (1961). “The Most Profitable Market Share Objectives”. Journal of Marketing, Vol. 25, No. 4, pp. 52-57.

Sleeper, L. and Harrington, D., 1990, “Regression Splines in the Cox Model with Application.” Journal of the American Statistical Association, 85, 412, 941-949.

Tom Spring (1999). “Portal Angst: A Good Site Is Hard to Find.” PCWorld, February 23.

D. Sullivan (2003). “Nielsen NetRatings Search Engine Ratings.” Nielsen NetRatings, February 25.

David M. Szymanski, Sundar G. Bharadwaj and P. Rajan Varadarajan (1993). “An Analysis of the Market Share-Profitability Relationship”. Journal of Marketing, Vol. 57, No. 3 (Jul., 1993), pp. 1-18

Bill Tancer (2006a). “Google Finance - If it Looks Like a Portal...” Hitwise US, March 29.

Bill Tancer (2006b). “Google, Yahoo! and MSN: Property Size-up.” Hitwise US, May 19.

Thies, C. and Gerlovski, D., 1993, “Bank Capital and Bank Failure, 1921-1932:

Testing the White Hypothesis." The Journal of Economic History, Vol.53, No.4, pp. 908-914.

Jean Tirole (1988). The Theory of Industrial Organization. MIT Press, MA

C. Christian von Weizsäcker (1984). "The Cost of Substitution." Econometrica, Vol.52, No.5

Joel Waldfogel and Lu Chen (2003). "Does Information Undermine Brand? Information Intermediary Use and Preference for Branded Web Retailers." Working paper.

Birger Wernerfelt (1991), "Brand Loyalty and Market Equilibrium." Marketing Science, Vol.10, No.3.

Jeffrey M. Wooldridge(2002). Econometric Analysis of Cross Section and Panel Data. MIT Press, MA

(2005) "The Battle of the Portals." The Economist, October 22, pp.65-66.

(2006) "Google Tops Halfway Mark for Search-Engine Market Share." DomainMagazine, May 28.

(2007) "Mobile Portal Market Structure", E-Business Strategies

(2007) "Yahoo goes mobile", Australian IT, June 20.

http://www.archive.org

http://compnetworking.about.com/library/weekly/aa011900a.htm

Appendix

A1.Collinearity diagnostics

Table A1. Collinearity diagnostics for portal attributes.

Variable	VIF	Sqrt (VIF)	Tolerance	R^2
Portal age	1.34	1.16	0.7479	0.2521
Mail	2.07	1.44	0.4825	0.5175
Auction	1.90	1.38	0.5276	0.4724
Shopping	2.44	1.56	0.4091	0.5909
Sport	1.80	1.34	0.5552	0.4448
Chat	2.09	1.44	0.4793	0.5207
Greetings	2.36	1.54	0.4238	0.5762
Games	2.25	1.50	0.4439	0.5561
Finance	3.27	1.81	0.3057	0.6943
News	2.59	1.61	0.3860	0.6140
Search quality	2.53	1.59	0.3956	0.6044
Messenger	2.54	1.59	0.3933	0.6067
Personal	1.57	1.25	0.6378	0.3622
Weather	2.10	1.45	0.4772	0.5228
Page	2.10	1.45	0.4753	0.5247
Mean VIF	2.20			

Table A2. Eigenvalues and condition index for portal attributes.

	Eigenvalue	Condition index
1	10.1307	1.0000
2	1.4721	2.6233
3	0.8380	3.4769
4	0.6213	4.0381
5	0.5215	4.4077
6	0.4452	4.7702
7	0.4119	4.9594
8	0.2896	5.9149
9	0.2548	6.3050
10	0.2499	6.3676
11	0.1864	7.3724
12	0.1585	7.9952
13	0.1432	8.4102
14	0.1336	8.7092
15	0.0784	11.3706
16	0.0651	12.4760
Condition number		12.4760
Determinant of correlation matrix	0.0007	

Table A3. Collinearity diagnostics for demographic characteristics of users.

Variable	VIF	Sqrt (VIF)	Tolerance	R^2
User age	4.31	2.08	0.2318	0.7682
User education	2.67	1.63	0.3749	0.6251
User income	3.14	1.77	0.3181	0.6819
Household size	5.09	2.26	0.1964	0.8036
Marital status	4.71	2.17	0.2122	0.7878
Renting	4.01	2.00	0.2491	0.7509
Mean VIF	3.99			

Table A4. Eigenvalues and condition index for demographic characteristics of users.

	Eigenvalue	Condition index
1	6.1555	1.0000
2	0.6458	3.0874
3	0.1452	6.5113
4	0.0393	12.5193
5	0.0074	28.8900
6	0.0060	32.1545
7	0.0009	82.4457
Condition number		82.4457

Determinant of correlation matrix 0.0167

Condition number derived from the eigenvalues for multicollinerity diagnostics. Normally, we suggest that if condition index is between 10 and 30, there is moderate to strong collinearity; and if it exceeds 30, there is severe multicollinearity.[13]

Variance inflation factor (VIF) shows how variance of estimation will be inflated by the presence of multicollinearity. As a rule of thumb, if the variable's VIF exceeds 10 (this will happen if R_j^2 exceeds 0.90), that variable considered to be highly collinear.[14]

Portal attributes in the data do not demonstrate high degree of collinearity, although demographic characteristics are collinear as it could be expected.

[13] See D.A. Belsley, E.Kuhn and R.E.Welsch, "Regression Diagnostics:identifying Influential Data and Sources of Collinearity", John Wiley&Sons, New York, 1980

[14] See David G.Kleinbaum, Lawrence L. Kupper and Keith E. Muller, "Applied Regression Analysis and Other Multivariate Methods", PWS-Kent, Boston, Mass.,1988