

ICT, Search Behavior and Market Outcomes

Inauguraldissertation zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften
der Universität Mannheim

Michael Kummer

vorgelegt im Frühjahrssemester 2014

Abteilungssprecher: Prof. Dr. Eckhard Janeba

Referent: Prof. Dr. Martin Peitz

Korreferent: Prof. Philipp Schmidt-Dengler, Ph.D.

Verteidigung: 20.5.2014

Acknowledgements

First I would like to express my gratitude to my advisors Martin Peitz and Philipp Schmidt-Dengler for their guidance and advice. The motivating curiosity of Martin Peitz constantly forced me to improve my work further. His merciless, but constructive criticism taught me how to scrutinize ideas at an early stage and his ability to make me see the big picture greatly benefitted my research. I thank Philipp Schmidt-Dengler for his continued advice. Discussing my work with him on a regular basis constantly pushed me forward, whenever direction was needed. It taught me how economic research works, how to dig deeper, but how to still keep the focus on the central issues.

Special thanks also go to Irene Bertschek who, together with Martin Peitz, gave me the opportunity to pursue my studies at the Center for Doctoral Studies in Economics at the University of Mannheim. Her continuing encouragement and the great research environment she has created at ZEW's department for the economics of ICT have been truly inspirational. Her continued advice, her patience and her faith in analyzing contributions to Wikipedia even in the dark hours of database problems were pivotal.

Marianne Saam was both a great coauthor and project leader. Her valued feedback helped me to further improve my work. I learned a lot from her foresight in managing the project, which kept things together and made further collaborations possible. I also thank George Giorgidze and Iassen Halatchliyski, co-authors of chapter two. Our interdisciplinary discussions challenged me to broaden my perspective and leverage the insights from other disciplines.

I thank Franz Hackl, Christine Zulehner and Rudi Winter-Ebmer, who coauthored the chapters on electronic commerce. I am grateful for their continued support and encouragement, which, for several years already, were and continue to be a great source of motivation. The joint research was a great pleasure and I learned from the collaboration how to rigorously scrutinize a question and analyze the data from several angles, until all doubts are cleared.

Particularly, I thank them (and some of their colleagues) for making me feel welcome over uncountable coffee capsules and inspiring discussions throughout the repeated research stays at University of Linz. I also thank them, together with Jesus Crespo-Cuaresma, Georg Dürnecker, Avi Goldfarb, Stephen Kastoryano, Jan-Peter

Siedlarek, Yuya Takahashi and Mike Ward, for several bits of help and extra advice that provided valuable orientation. Moreover, different chapters of this thesis have further profited from comments and suggestions from many more people. Their feedback is acknowledged at the beginning of each chapter separately.

I am also thankful for administrative and technical help without which my dissertation would hardly have been possible. In particular, I would like to thank Ulrike Merkel for her endless efforts and infinite patience in organizing research stays, conference visits, and reminding me of deadlines. Every day, she takes great burden off our shoulders. I also thank the ZEW's administrative staff in general for making the administrative side of things so easy for the researchers. I thank Robert Bräutigam for helping to keep the computer "Oskar" happy and helping us to teach that machine new tricks. Also my gratitude for Oskar's peaceful computing efforts should be mentioned. I also want to thank Marion Lehnert and Sandro Holzheimer for their support. Financial support from SEEK and from the Wissenschaftscampus Tübingen is gratefully acknowledged.

I want to thank participants of the first MaCCi - IO day for valuable feedback and my colleagues both in ZEW's department for the Economics of ICT and in the PhD Seminar in IO at the University of Mannheim for their patience, when repeatedly exposed to progressing versions of the same ideas. Their feedback has helped me to continuously improve my work.

Apart from those already mentioned, I want to thank my friends and colleagues at the University of Mannheim and at the ZEW for their help and/or many inspiring discussions about side issues in "the real world" or simply company over coffee. I thank Thomas for the help when sharing the joys of programming and Olga for joining forces to fight with the Wikipedia database and the great cooperation. Specifically, but not exhaustively, I want to mention Malin, Christoph, Daniél, Dirk, Ben, Jana, Hannes, Max, Gordon, Christian, Christian, Fabienne, Alex, Johannes, Patrick, André, Steffen, Stephan, Michelle, Philipp and Aiyong.

I am indebted to my family and friends who supported me continuously throughout my this thesis and my life, even in recent times when I basically stopped answering to e-mails and calls. Above all, I would like to express my gratitude to two people: I thank Zhenia Bogdanova-Kummer for being my source of energy and inspiration - the person with whom I feel at home, no matter where. And I thank my father for everything he has taken upon himself to make my childhood work - against all odds - and for helping this naggingly curious boy grow up.

Contents

Acknowledgements	iii
List of figures	ix
List of tables	xii
1 Introduction	1
1.1 Links, Search and Content Production	2
1.2 Firm Behavior on a Price Comparison Site	4
2 Spillovers in Content Networks	7
2.1 Introduction	7
2.2 Literature	10
2.3 The Empirical Model	12
2.3.1 Basic Intuition - Throwing Stones into a Pond	13
2.3.2 Identifying Assumptions for the Treatment Effects	15
2.3.3 Reduced Form Analysis	18
2.3.4 Structural Form Analysis and Bounds	19
2.4 Data	25
2.4.1 Data Preparation - Treated and Control Group	26
2.4.2 A Closer Look at the Datasets	28
2.5 Estimation Results	30
2.5.1 Large Events	31
2.5.2 Neighbors of Featured Articles	33
2.5.3 Bounds for the Structural Estimator	36
2.5.4 Aggregate Effects and Heterogeneities in the Spillover.	37
2.6 Conclusions, Limitations and Further Research	40
2.7 Descriptive Statistics	43
2.8 Additional Regression and Figures	46
3 Network Centrality and Content Creation	49

3.1	Introduction	49
3.2	Related Research	52
3.3	Data	54
3.3.1	The Anatomy of the Data Set	54
3.4	Relationships of Interest and Methodology	57
3.4.1	Network Position and User-Generated Content	57
3.4.2	Getting Connected to the Category of Economics	60
3.5	Results	61
3.6	Conclusion	63
3.7	Tables	65
3.7.1	Summary Statistics	65
3.7.2	Regression Results	68
3.7.3	Figures	72
4	Market Structure in E-Commerce	73
4.1	Introduction	73
4.2	Theoretical Predictions	76
4.3	Data and Empirical Strategy	79
4.4	Results	86
4.4.1	Market Structure and Market Performance	86
4.4.2	Life Cycle Effects	89
4.4.3	Substitutes over the Product Life Cycle	90
4.4.4	Robustness	92
4.5	Conclusions	93
5	99 Cent: Price Points in E-Commerce	107
5.1	Introduction	107
5.2	Literature	109
5.2.1	Theories for price points	109
5.2.2	Empirical evidence	111
5.3	Data and prevalence of price points	113
5.4	Price rigidity and firm behavior	116
5.4.1	Pooled analysis	118
5.4.2	Stratified analysis	118
5.4.3	Price rigidity for price leaders	119
5.4.4	Price jumps after focal prices	120
5.5	Price points and consumer behavior	122
5.6	Interpretation and conclusions	125

5.7 Tables and Graphs	127
A Additional Materials - Spillovers	137
A.1 Details: Dataset	138
A.1.1 Preparation and Extraction	138
A.1.2 Choice of Treated Articles and Neighborhood	138
A.1.3 Choice of Control Group Articles and Neighborhood	139
A.2 Details: Estimation of Spillover Parameter	140
A.2.1 Introductory remarks	140
A.2.2 Setup and Basic Intuition	140
A.2.3 Proof of Result 1	141
A.2.4 Estimating Bounds for the Parameters of Interest	145
A.3 Aside: Reaction to treatment of the neighbor	150
B Additional Materials - Network Position	153
B.1 Preparation of the Data and Definition of the Category Economics . . .	154
B.2 Robustness	155
C Additional Materials - Market Structure	159
C.1 Screenshot of the Price Comparison Site	160
C.2 Construction of Substitutes	161
Bibliography	163
Erklärung der Urheberschaft	172
Short CV of Candidate	174

List of Figures

2.1	Schematic representation of a local treatment, which affects only one of the two subnetworks and there only a single node directly.	14
2.2	Schematic representation of the two extreme networks, used to compute the upper and lower bound estimates of the parameters of interest. . .	24
2.3	Catastrophes: Comparing average clicks (new edits) of treated pages (and neighbors) to three comparison groups.	27
2.4	“Today’s featured articles”: Comparing average clicks (new edits) of treated pages (and neighbors) to three comparison groups.	29
2.5	Figure contrasting the mean of clicks on featured articles, with the aggregated clicks on all neighboring pages.	46
2.6	Figure showing the aggregated new revisions on all neighboring pages.	46
3.1	Development of the median of the outcomes and the indegree over time.	72
4.1	Median markup plotted against the number of firms and age of product	103
4.2	Instrument using firm’s listing behavior in earlier lifecycles	104
4.3	Minimum markup in different phases of the product life cycle	105
4.4	Median markup in different phases of the product life cycle	105
5.1	Distribution of the cent digits	127
5.2	Distribution of the last euro digits of 00c-ending offers	128
A.1	Schematic representation of the two extreme networks, used to compute the upper and lower bound estimates of the parameters of interest. . .	145
C.1	Typical screenshot of the price comparison that is shown at geizhals.at	160

List of Tables

2.1	Large events: clicks/added revisions over time for indirect neighbors. . .	31
2.2	'Featured articles': clicks/added revisions over time for direct neighbors.	34
2.3	Spillover of views and translation to edits.	35
2.4	Clicks/added revisions over time, when including article heterogeneity.	38
2.5	Conversion of attention to action: Content vs. money contributions. . .	39
2.6	Summary statistics: direct neighbors of shocked 'featured articles'. . .	43
2.7	Indirect neighbors of shocked 'large events articles' (2 clicks away) . .	43
2.8	Included "featured articles" and associated articles (1 clicks away). . .	44
2.9	Included disasters and associated articles (2 clicks away).	45
2.10	Robustness check: Reduced number of events.	47
3.1	Summary statistics of main variables. Connected articles.	65
3.2	Summary statistics of main variables. Articles that get connected to category.	66
3.3	Summary statistics of the frequency of changes of main variables. . . .	66
3.4	Weekly changes of main variables.	67
3.5	Relationship of page length and centrality.	69
3.6	Relationship of the growth of page length and the page becoming connected.	70
3.7	Growth of page length when page gets connected: Excluding 2 periods before and after.	71
4.1	Summary statistics of the collapsed two-dimensional panel-data with the info on the level of goods and time	96
4.2	First stage regressions for instrumenting the number of firms	97
4.3	Minimum Markup	98
4.4	Median Markup	98
4.5	Coefficient of Variation	99
4.6	Shipping Cost	99
4.7	The Importance of Substitutes over the Life Cycle	100

4.8	Markup and price dispersion weighted by clicks.	101
4.9	Median markup and the composition of shops	102
5.1	Descriptive statistics	129
5.2	Frequency of focal price endings for (i) low-price and high-price offers and (ii) low-price vs. high-price products.	130
5.3	Focal prices and price stickiness: all price offers	131
5.4	Focal prices and price stickiness: price-leading offers.	132
5.5	Focal prices and price stickiness: size of price jumps	133
5.6	Demand and focal prices	134
5.7	Demand and price points for different types.	135
B.1	Robustness checks for the relationship of page length and centrality. . .	156
B.2	Relationship of number of authors and centrality.	158

Chapter 1

Introduction

Information and Communication Technologies (ICT) help to reduce transaction costs in several ways. The study of this phenomenon and the resulting economic consequences are the underlying common theme in the four subsequent chapters of my dissertation. Each of them is a self-contained paper that contributes to the field of Industrial Organization. In particular, I contribute to the study of two important phenomena that arose as a consequence of the ICT-enabled reduction in transaction costs, which modern societies have seen over the last two decades: electronic commerce (E-Commerce, in what follows) and (commons based) “peer production” (Benkler and Nissenbaum (2006)). Both of these have gained importance, because ICT reduced the cost of storing information and retrieving it at a later point in time. Thus it became easy to search and find information once it was stored digitally. Moreover, ICT reduced the cost of communication, which allows remotely located partners to coordinate and jointly resolve complex tasks, be it a commercial transaction or the collaboration on a complex joint project, such as writing a scientific paper or an encyclopedia.

Chapters 2 and 3 focus on a cost reduction that might have very drastic long-term consequences: New digital media enable new communication platforms, where *multiple* agents can *interactively* coordinate or contribute thoughts and ideas. They can split up work that is necessary for achieving a large goal and perform tasks in an independent and modular fashion Lerner and Tirole (2002). Under this new production regime, called (commons based) “peer production” (Benkler and Nissenbaum (2006)), highly valuable and complex services can be produced: Successful examples are Open Source Software or Wikipedia. Chapters 2 and 3 of my thesis study the contribution flows to Wikipedia, which is the most shining example of successful peer production.¹ It has been created by thousands of volunteers and is now world’s most important platform for documenting and storing encyclopedic knowledge. Both chapters focus on the role

¹Chapter 3 is coauthored with Marianne Saam, Iassen Halatchliyski and George Giorgidze.

of the hyperlinked citation network between Wikipedia’s articles and shed light on how it influences users’ search and contribution behavior.

The second part of my thesis (chapters 4 and 5) is related to electronic commerce. This new form of retailing is a second phenomenon that was enabled by ICT-induced reductions in search and coordination costs. In E-Commerce, several important transaction costs that used to be borne by one of the two sides, are no longer important. Shops do not have to maintain costly “brick and mortar” facilities any longer and they can ship on demand, rather than anticipating where the customers will desire which product. Moreover, they can create user profiles and tailor which offers specific clients see to their previously revealed preferences. Customers, in return, can economize on costly expeditions to the stores, by ordering from remote locations. Most importantly, consumers can compare many prices much faster and more easily. This fact is the main theme of the two papers focusing on E-Commerce (chapters 4 and 5), where my coauthors and I study retailer behavior on Austria’s largest price comparison site, *www.geizhals.at*. Both chapters are coauthored with Franz Hackl and Rudolf Winter-Ebmer. Christine Zulehner coauthored chapter 4.

Additional materials for the chapters 2, 3 and 4, which were not included in the main text (e.g. additional results, data descriptions, proofs or robustness checks) are provided in separate appendices after the four main chapters. The bibliography with the references of all four chapters can be found at the end.

1.1 Links, Search and Production in Networks of User Generated Content.

This part of my dissertation (chapters 2 and 3) studies the production patterns on the German Wikipedia, which is a production setting of (commons based) “peer production” (Benkler and Nissenbaum (2006)). This term refers to multiple agents coordinating via an online platform to organize and distribute work, because they wish to achieve a large goal. Under this new production regime highly valuable and complex services, such as Open Source Software or Wikipedia have been produced.

Chapter 3 (coauthored with Marianne Saam, Iassen Halatchliyski and George Giorgidze) focuses on the role of an article’s network position on contribution patterns. This is motivated by the observation that producers of user-generated content have to decide where to contribute, before they can contribute content to large and highly structured online platforms like Wikipedia. This decision is expected to depend on the way the content is organized. We analyse whether the hyperlinks on Wikipedia channel the attention of producers towards more central articles. We observe a sample

of 7,635 articles belonging to the category economics on the German Wikipedia over a period of 153 weeks and we measure their centrality both within this category and in the network of over one million German Wikipedia articles. Our analysis reveals that an additional link from the observed category is associated with around 140 bytes of additional content and with an increase in the number of authors by 0.5. The relation of links from outside the category to content creation is much weaker.

In chapter 2 I exploit exogenous shocks to the attention on individual articles to precisely disentangle the effect of a link from other background factors that might simultaneously influence both the network structure and the contributions to an article. I ask how networks generate externalities, such as spillovers or peer effects and focus on the challenge of quantifying these externalities in the face of endogenous network formation. I tackle the problem by exploiting local exogenous shocks on a small number of nodes in the network and investigate spillovers of attention on the German Wikipedia. I show how the link network between articles influences the attention that articles receive and how the additional attention is converted into content.

Exogenous variation is generated by natural and technical disasters or by articles being advertised on the German Wikipedia's start page. The effects on neighboring pages are substantial: They generate an increase in views of almost 100 percent and content generation is affected similarly. Aggregated over all neighbors, a view on a treated article converts one for one into a view on a neighboring article. This approach applies even if, absent network data, identification through partial overlaps in the network structure fails and thus helps to bridge the gap between the experimental and social network literatures on peer effects.

I conclude this section by pointing out that the importance of understanding peer production in the context of knowledge production cannot be overstated, because the process of building, documenting and transferring knowledge is characterized by enormous frictions: Transmitting ideas has always been extremely costly, involving face to face interaction or a media for storing knowledge and a lot of time. Moreover, the successful transmission of knowledge is not guaranteed: New and even existing knowledge can be inaccurately disfigured (and even lost), if it is not sufficiently well documented or transmitted to following generations. However, despite all these frictions a civilization's knowledge and its innovations have always been a fundamental input to its productivity. How well an economy can build new knowledge or transmit the existing one to the public or to future generations is a central determinant for the success of a culture. Thus, many resources have typically been devoted to the generation and transfer of knowledge, often involving an entire sector of a civilization's economy.

Peer produced Wikipedia was a first showcase, where thousands of volunteers took only a decade to collect, digitize and document virtually all encyclopedic knowledge

in unprecedented detail. This highlights how ICT-enabled reductions of transaction costs could transform the societal process of knowledge generation, documentation and transmission, and such developments are likely to have deep and long lasting effects on education and innovation. However, given the importance of the knowledge generation and knowledge conservation process for society, it is crucial to understand these long run effects as early as possible. Each chapter in this part uncovers a relevant aspect of peer produced content networks: Chapter 3 highlights the importance of local semantic structures, such as links from articles that belong to the same category. In chapter 2 I precisely quantify how much attention spills across links and point out the crucial role of a high visiting frequency for the success of this type of public good production.

1.2 Firm Behavior on a Price Comparison Site

Price comparison sites offer a technology that allows to compare prices and immediately find the cheapest offer with only a few clicks. This technology has the potential of revolutionizing the market for retail and it seriously challenges the shops that advertise online. The question arises, how shops can continue to make a profit on such a site, and the shops' owners might have to devise new strategies to do so. In two separate chapters (4 and 5) we study firm behavior on Austria's market dominating site for price comparisons: *www.geizhals.at*.

In chapter 4 (coauthored with Franz Hackl, Rudolf Winter-Ebmer and Christine Zulehner) we analyze the interaction between market structure and market performance and how it varies over the product cycle. To account for the potential endogeneity in this relation, we use a novel instrumental variable approach. We furthermore investigate the relationship of market structure and price dispersion. We combine data from *geizhals.at* with retail data on wholesale prices provided by a major hardware producer for consumer electronics. We observe firms' retail and input prices, and all their moves in the entry and the pricing game over the whole product life cycle. Our results show that instrumenting is important for estimating the empirical effect of competition on the markup of the price leader. One more firm in the market is associated with a reduction of the price leader's markup which is equivalent to competition between existing firms for an additional three weeks in the product life cycle. Our results support search theoretic models and contradict models of monopolistic competition. Moreover our results support the existence of price dynamics over the product cycle. They also highlight the substitutability between newly innovated and old expiring technologies and how it varies with respect to competitors' and own brand innovations.

In chapter 5 (coauthored with Franz Hackl and Rudolf Winter-Ebmer) we analyze

the phenomenon of quoting prices that end in a special ending, such as “.99c.” Setting prices ending in nines is a common feature of many markets for consumer products. Their prevalence has been explained either by a specific image of such price points or by the exploitation of rational inattention on the part of the consumers who want to economize on the cost of information processing. We use data from the largest Austrian online market for price comparisons (www.geizhals.at), where any disproportionately high frequency of such endings would be expected to disappear if marginal cost pricing prevailed. We analyze how frequently special pricing patterns emerge and we explore the impact of these price points on the consumers’ demand. We find a remarkable prevalence of such prices. Moreover, prices ending with nine are also sticky: price-setters change them with a significantly lower probability, rivals underbid these prices more seldom if they represent the cheapest price on the market, and we observe higher price jumps by price leaders for these price points.

Both chapters together show that certain phenomena, which would be expected to disappear in online markets continue to exist. They have not disappeared even several years since the site has been introduced. This suggests that shops might be able to continue making small profits as a consequence of specific behavioral patterns, such as rounding down a price that ends in “.99c”.

Chapter 2

Spillovers in Networks of User Generated Content* -

Evidence from 23 Natural and 34 Pseudo-Experiments

2.1 Introduction

Over the last decade, it was surprising to witness how large numbers of volunteers coordinated to produce Wikipedia. It is now the world's most consulted reference for encyclopedic information, highlighting the potential of collaborative production.² Consequently, the amount and quality of voluntary contributions to online public goods ("peer production"), such as Wikipedia or open source software, is of great economic interest: Successfully leveraging this potential in other settings could be very beneficial to society, but requires understanding how exactly peer production works. In this paper I analyze spillovers of attention, transmitted through links in the German Wikipedia and how attention affects contribution effort.

*I thank the selection committee of the IIIrd ICT Conference in Munich for awarding a very early version of this paper the 2nd Prize in the "Best PhD Paper Award." The author is grateful to the Wikimedia Foundation and Frédéric Schütz for access to the Wikipedia data. Thanks to Thorsten Doherr and Manfred Knobloch for support with data processing and to George Giorgidze for help with DSH. I benefitted from discussions with Irene Bertschek, Christoph Breunig, Luis Cabral, Jörg Claussen, Ulrike Cress, Habiba Djebbari, Neil Gandal, Avi Goldfarb, Sanjeev Goyal, Shane Greenstein, Thorsten Grohsjean, Maximilian Kasy, Francois Laisney, Jose Luis Moraga-Gonzalez, Kathy Nosal, Gal Oestreicher-Singer, Martin Peitz, Marianne Saam, Philipp Schmidt-Dengler, Olga Slivko, Yossi Spiegel, Joel Waldfogel, Michael Ward and Andrea Weber. The participants in several seminars at ZEW and U Mannheim, U Linz, U Kaiserslautern and WU Wien and of the 11th ICT-Conference (ZEW, Mannheim), the EARIE 2013 (Évora), the 11th conference of Media Economics (Tel Aviv) and the IIIrd ICTCM (Munich) provided valuable input. James Binfield, Timo Schäffer and Daniel Bergman provided outstanding research assistance. Financial support from the WissenschaftsCampus Tübingen is gratefully acknowledged.

²Its quality is sufficient to almost completely drive previous incumbents out of existence: Encyclopedia Britannica was the most prominent English encyclopedia and the "Brockhaus" dominated in Germany. Both have suffered considerable losses in sales and market share.

How much attention can be channeled by links and how much of this attention is converted to action? These questions are key to understanding peer production and matter in many other contexts, such as public decision making or advertisement, where it is essential to channel attention. However, they cannot easily be analyzed empirically, because most outcome variables of interest might themselves drive the network structure.³ I circumvent this endogeneity problem by exploiting local exogenous shocks in the attention to single nodes in Wikipedia’s article network. The resulting attention spillovers to neighboring nodes generate exogenous variation to attention that is independent of the production process. As initial shocks I use natural and technical disasters or when a neighbor is advertised on Wikipedia’s start page for 24 hours. Both are shown to affect traffic and allow applying difference-in-differences to measure direct and indirect treatment effects. Moreover, I formally show how to relate these treatment effects to the structural parameter that measures attention spillovers.

Considering the network formed by Wikipedia’s articles (as nodes) and the hyperlinks between articles (as directed links), I obtain a dataset on 57 primary news and attention shocks, which contains daily information on views and content generation of almost 13,000 articles and more than 700,000 observations.⁴ I document a large initial attention spillover, independent of whether the initial shock is generated by a disaster or by advertisement on the start page. I find that the initial increase of attention to neighboring pages of featured articles translates to substantial content generation (= editing activity). Views of neighbors doubled on average, and editing activity almost doubled. Furthermore, the number of authors increased, indicating that new authors contributed.

Distinguishing articles by their length I find that the spillovers of attention do not depend on the length of the link’s target, whereas content generation *does*. Like the average article, short articles were visited 35 times more, on average, but these additional visits resulted in substantially fewer new edits and a smaller increase in length than in the full sample. In short, citation links matter for the attention that nodes receive, but much less for the content that is generated on such nodes. This may be justified given the maturity the German Wikipedia had reached by 2007.

To relate the measured effects to the structural spillover parameter, I extend a

³In my case the outcomes of interest are attention (= clicks) and new edits, but the problem applies in general. Both processes, like supply and demand, may be driven by unobserved dynamics. The resulting methodological issues are a constant obstacle in a wide range of applications that try to measure peer effects or the role of social networks in generating externalities. Examples are interpersonal connections and the take up of micro-finance, or peer effects in schooling, aid programs or health interventions.

⁴23 large-scale media events such as natural disasters and 34 articles that were advertised on Wikipedia’s main page for 24 hours, and all their respective network neighbors. The information was obtained 14 days before and after the events. Details are provided in Section 2.4 and Appendix 2.4.1.

standard model of peer effects by Bramoullé et al. (2009) to allow for the incorporation of local exogenous shocks. I show how the spillover parameter can be uncovered in two steps: (i) by applying a difference-in-differences strategy to obtain estimates for the indirect treatment effect (as in Kuhn et al. (2011)) and (ii) by discounting for higher order spillovers in the network. I also show that bounds can be derived if the network information cannot be used to account for higher order spillovers. This illustrates why the estimation strategy is robust to both endogenous network formation and Manski's (1993) reflection problem. The resulting model provides a notation to nest approaches for identifying social effects that are based on exploiting exogenous (pseudo-) experimental variation⁵ into a framework which considers network structure. I apply these techniques to my data and obtain an interval estimator for the structural spillover parameter of interest. I find that an average increase of ten views on the neighboring pages results in an increase of 2.22 to 2.92 views on the page in the center. These bounds are computed using extreme (benchmark) assumptions on the network, and can be computed even when no information on the link structure is available. My method for deriving these bounds is an additional contribution to the literature.

My findings allow for a more abstract reading. The hyperlink network between articles can be interpreted as a citation network and Wikipedia as a peer production tool for the documentation of human knowledge. Consequently the relevance of my findings extends to other settings of peer production including open source software or scientific research. While it is true that my strategy requires a lot from the data⁶, recent advances in data handling techniques and the increasing availability of data on social interactions will provide further applications for this strategy.

In the next section (2.2) I discuss the relevant literature and this paper's contributions. The methodological approach is discussed in Section 2.3, which extends the linear peer effects model and describes identification through local treatments in networks. Detailed derivations of the estimator and the bounds are in Appendix A.2. Section 2.4 discusses the data collection and the relevant variables. The empirical results and how to relate my reduced form estimates to the structural model are described in Section 2.5. Concluding remarks, limitations and avenues for further research are offered in Section 2.6. The appendices contains summary statistics, robustness checks, additional figures and a discussion of why network neighbors should not react to their neighbor's treatment.

⁵Partial population treatments (Moffitt (2001)) or impact evaluation studies based on a two-stage randomization over sub-populations (villages) and then individuals inside sub-populations.

⁶Exogenous treatments of individuals in networks (or groups) could rarely be observed in previous studies. Researchers often have the network structure and no exogenous source of identification, or exogenous variation yet no information on the network structure. However, such data are increasingly available from field experiments or online sources.

2.2 Literature

Coase's (1937) insight that production should either be organized in a free market if market frictions are low, or in a firm if they are high, was fundamentally challenged by the success of Open Source Software production and Wikipedia. The new coordinating principle, by which large numbers of people distribute small modules of the total workload via the web is referred to as commons-based peer production (Benkler 2002 and 2006). The extraordinary past achievements of this production mode illustrate the deep impact its emergence might have on the economic process and even society as a whole. This paper contributes to the literature in several ways. First, I document the role of the network for spillovers of attention and for content production in a relevant setting of peer production - the German Wikipedia. Second, I measure attention spillovers and quantify how attention is converted into action (contributing content). Finally, I analyze the heterogeneities in the spillovers in the network. In what follows, I discuss the streams of the literature that each of these contributions add to.

By analyzing the role of the network for content production in Wikipedia, I add to previous research, which has analyzed the correlation between a node's position in a network and the outcomes of interest (Fershtman and Gandal (2011), Claussen et al. (2012) or Kummer et al. (2012)). Economists have asked how social networks influence economic real world outcomes for (at least) two reasons: First, it is important to understand how a network's structure affects individuals' outcomes and to quantify the resulting overall value of a network and its links. Second, it matters whether the outcome of our connections or peers influences us, be it positively or negatively. My paper quantifies the causal effect of the average attention of a focal article's neighbors on the attention of the focal article. Previous research has struggled with the following empirical problems: The outcome variable might itself drive network position, thus giving rise to the classic endogeneity problem. Moreover, the reflection problem laid out by Manski (1993) applies, since nodes influence each other like peers (Bramoullé et al. (2009)). This paper circumvents both problems by exploiting local exogenous treatments of single nodes in Wikipedia's article network.

A second contribution to the literature is the econometric approach to quantifying attention spillovers between Wikipedia articles. My formal framework combines existing approaches and extends it in a novel way. I analyze treatments of neighbors *in a network*, but instead of focusing on the effect of treatment I focus on the *spillovers* of these treatments and use them as sources of exogenous variation in the *attention* to such articles. Moreover, I use the fact that exogenous treatment sometimes affects only a single node. Such local treatments are analogous to the Partial Population Treatment that Moffitt (2001) suggested for the analysis of peer effects - not in the context of

network analysis - to solve the reflection problem Manski (1993).⁷ There is also a close relationship to studies that added a higher layer of randomization, which allows the computation of *indirect* treatment effects.⁸ An example is Crépon et al. (2013), who randomize over cities and vary the treatment intensity to study whether labor market programs have a negative impact on the non-eligible. Studies that use exogenous local shocks to single individuals could be called “Mini Population Treatments” and this idea is used increasingly often in recent studies that use network information (Aral and Walker (2011), Banerjee et al. (2012), Carmi et al. (2012)).

Following the analysis of attention spillovers, I analyze how attention translates into action. I find a conversion rate of 1000 clicks for 1 edit. These findings highlight the need of adding an important extra ingredient to modularity and strong leadership (Benkler and Nissenbaum (2006), Lerner and Tirole (2002)), to guarantee the success of peer production: If the individuals contribute infrequently, a high overall frequency of visits is necessary. This reaffirms the potential of ICTs to enable peer production through their ability to drastically reduce coordination costs. To guarantee that the content production is exclusively due to attention, I exploit sudden exogenous spikes in the attention to a neighbor, which affect not only the shocked nodes in a network, but are also transmitted across links (Carmi et al. (2012)). Such spikes are generated by large-scale events like natural disasters and accidents or the advertisement of featured articles on Wikipedia’s start page. Little is known about how attention influences the decision to contribute to a public good. Several papers show that attention through blogs or reviews, even negative, can be positively related to purchase and investment decisions (Barber and Odean (2008), Berger et al. (2010), Hu et al. (2013)). However, it is typically impossible to measure the amount of attention generated by the publicity and how it is converted to action.⁹

⁷Dahl et al. (2012) provide an example of such an experiment. An alternative approach is to exogenously vary the composition of peer groups: Zhang and Zhu (2011) uses the fact almost all Chinese Wikipedia users in mainland China were blocked by the government’s “Chinese fire wall”, to measure the effect on the incentives to contribute. Also disasters or fatal accidents are frequently used in similar settings. (Sacerdote (2001), Imberman et al. (2009)), Ashenfelter and Greenstone (2004)). Keegan et al. (2013), who analyze the structure and dynamics of Wikipedia’s coverage of breaking news events.

⁸When social effects or spillovers are present, a violation of Stable Unit Treatment Value Assumption (SUTVA) compromises the validity of the control group (Ferracci et al. (2012)). Depending on the application, a second layer (classrooms, villages, districts etc.) can remedy the issue. (Miguel and Kremer (2003), Angelucci and De Giorgi (2009), Kuhn et al. (2011) and many more).

⁹Altruism and social image concerns are important drivers of voluntary provision (non-monetary) of a public good in offline contexts (Carpenter and Myers (2010)). Social effects and attention to the individual contribution also matter in peer productivity (Shang and Croson (2009), Huberman et al. (2009)). Yet, studies that precisely quantify how attention converts to contributions and that disentangle this effect from the other relevant drivers of contributions are rare.

The last contribution of this paper consists of analyzing whether attention spill uniformly or whether there are large heterogeneities. I analyze the drivers of the attention spillovers and subsequent content generation. Only a few papers study what determines whether spillovers take place or not. Carmi et al. (2012) pioneering analysis finds that the network structure does well in predicting spillovers on Amazon's recommendation network, but their findings are challenged by the fact that spillovers are also an important driver of Amazon's algorithm that places and sorts the links.¹⁰ I distinguish articles by their length, by how they are generally linked, and by how closely they are linked to the shocked articles. Concerning drivers of individual attention to an item, only few studies have analyzed which items receive collective attention. (Hoffman and Ocasio (2001), Wu and Huberman (2007)).¹¹ I contribute by analyzing what users choose, when presented with several options for a click and the subsequent conversion of awareness to making a voluntary (non-monetary) contribution to a public good.

In conclusion, this paper provides new insights into the dynamics of user activity in the world's largest knowledge repository, measures how users allocate their attention, and how these effects are mediated by node characteristics. It shows how treatments diffuse across networks if the content items are linked and how attention is converted into contributions of effort. To the best of my knowledge my results are the first to show how a citation network influences users' contributions through channeling attention.

2.3 The Empirical Model

In this section I discuss the empirical model. I first give a basic and informal intuition of my estimation approach (Subsection 2.3.1). Next I discuss the assumptions made to identify the effect of the exogenous treatments I use (Subsection 2.3.2) and the reduced form estimation of the regressions (Subsection 2.3.3). The last and most extensive subsection (2.3.4) describes the extended linear peer effects model: I discuss how and under which assumptions the researcher can identify the structural parameter that measures spillovers from the reduced form estimates if she observes the network information. In the same section I also show how to compute an upper and a lower bound for the coefficient when the network information is not available. An important case where my arguments do not apply are situations where the neighbors of the treated

¹⁰Carmi et al. (2012) analyze the effect of the external shocks of recommendations by Oprah Winfrey on the product network of books on Amazon. They find that a recommendation not only triggers a spike in sales of the recommended book but also of books adjacent in Amazon's recommendation network.

¹¹Viral Marketing studies are concerned with the diffusion of information in a social network, i.e. mediated by social propagation, rather than repeated individual decisions.(e.g. Aral and Walker (2011), Ho and Dempsey (2010), Hinz et al. (2011))

nodes/individuals observe the treatment and adjust their outcome as a reaction.¹² Appendix A.3 shows, how the model would have to be extended to include such a possibility and which challenges to identification of the spillover parameter would emerge as a result.

2.3.1 Basic Intuition - Throwing Stones into a Pond

This subsection provides an intuitive explanation of the data structure and the estimation approach. The basic idea of the research approach can be imagined as “throwing stones into a pond and tracing out the ripples”. The design of this paper uses the fact that certain nodes were affected by a large increase of attention, that this was exogenous, and that ex-post it is known to the researcher when exactly the pseudo-experiment occurred. Moreover, since the link structure is also known, it is possible to observe what happens to the directly or indirectly neighboring nodes. As in a pond, we would expect the largest effect on the directly hit node and a decreasing amount of additional attention the further away an article is from the epicenter.

The schematic representation in Figure 2.1 shows how the data is structured. Wikipedia articles are the nodes of the network. They are represented by a circle with a letter inside. Each circle represents a different article in the German Wikipedia. Articles are connected to each other via links, which are visible on Wikipedia as highlighted blue text. Clicking on such text forwards the reader to the next article and these links form the edges of my directed network. Such links are represented by a line between two nodes. An important aspect of my identification strategy requires the observation of two disconnected subnetworks at the same time. This is represented by network L and network C shown facing each other. I maintain this notation in all derivations that follow.

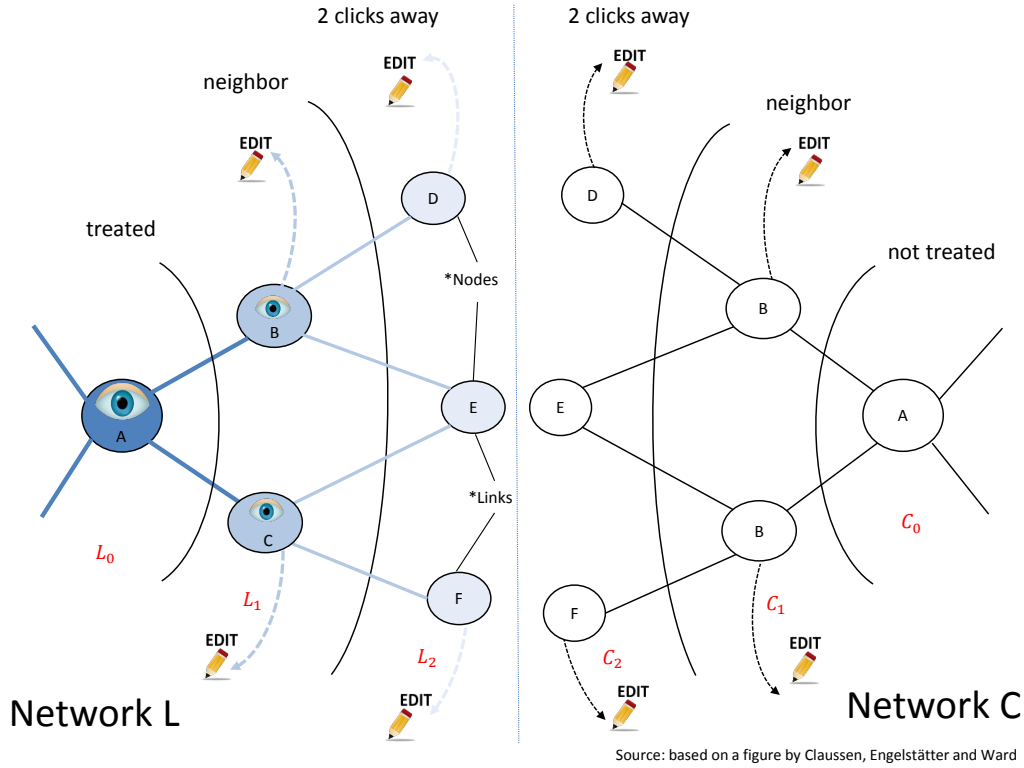
I focus on subnetworks around a start node. These start nodes are denoted by subscript 0. Hence, the start node of the two networks are denoted by ℓ_0 and c_0 . Consider “today’s featured articles” for a moment: start nodes ℓ_0 and c_0 could both be featured articles, so both are eligible for treatment. However, only one of them can be selected to become “today’s featured article” on any given day. The nodes that receive a direct link from a start node (direct neighbors) in network L form the set of direct neighbors L_1 and a focal node from that set is sometimes denoted as ℓ_1 .¹³ The set of indirect neighbors¹⁴ in the network L forms L_2 and so on. Analogously the set C_1 is

¹²E.g. classmates, that react with protest to an unfair punishment of their peer.

¹³While the set L_0 consists only of one node ($L_0 = \{\ell_0\}$), set L_1 consists of multiple nodes.

¹⁴Indirect neighbors are defined as receiving at least one link from a node in set L_1 without themselves being in L_1 . Hence the shortest path from the start node to an indirect neighbor is via two clicks.

Figure 2.1: Schematic representation of a local treatment, which affects only one of the two subnetworks and there only a single node directly.



NOTES: The figure illustrates the structure of the data. Wikipedia articles are the nodes of the network. Each circle with a letter inside represents a different article in the German Wikipedia. The eye icons represent attention, while the pencil illustrate a decision to contribute an edit. Articles are connected to each other via links, which are represented as lines. The design of this paper uses the fact that certain nodes were affected by a large and exogenous increase of attention, and that it is known to the researcher when the pseudo-experiment occurred. In this setting this is represented by the two subnetworks L and C. Both, nodes in L_0 and C_0 , could be hit by a disaster (or are featured articles). Hence they are eligible for treatment. Yet, only one is actually hit (or becomes “today’s featured article”) at any given day. The coloring illustrates the effect of such a large local shock on Wikipedia, which affects only subnetwork L. The shocked node is colored in dark blue, the direct neighbors are colored in light blue and so on. If we observe a valid second network from which it is possible to infer what the outcomes would have been if no treatment had taken place, we can use these outcomes for comparing the size of the outcomes layer by layer. In general the network may be directed or undirected (Wikipedia articles are directed). The figure draws on a representation in a working paper on network formation by Claussen, Engelstaetter and Ward.

the direct neighbors of the start node in network C, and C_2 is the indirect neighbors of node c_0 .

In a typical network in which the outcome of the individual nodes depends on the outcome of their neighbors we would observe many correlations and cross influences. However, it would be difficult to discern where they originate from or whether they are due to underlying and unobserved background factors which merely affect the nodes in similar ways. The coloring in Figure 2.1 illustrates the mechanism of local exogenous shocks (“the stone in a pond”). The shocked node is colored in dark blue, the direct neighbors are colored in light blue and so on. As I will show formally in the next sections, identification of the spillover hinges on the ability to observe a valid counterfactual from which to infer what the outcomes would have been if no treatment had taken place. If this is possible, we can compare the outcomes layer by layer. More

information about how the layers are identified and obtained is provided in Section 2.4.

2.3.2 Identifying Assumptions for the Treatment Effects

The spillover parameter will be evaluated by estimating difference-in-differences for each layer separately. To clarify the assumptions in the reduced form estimation by layer, I use the control-treatment notation from the impact evaluation literature (cf. Angrist and Pischke (2008)). First, this highlights similarities to the Partial Population Treatment (cf. Moffitt (2001)). Second, it aids the understanding of the assumptions made here. Terminology and notation are inspired by Kuhn et al. (2011).¹⁵

Direct Effect of Treatment: Consider a node in network $i \in \{\ell, c\}$ in period t . The direct treatment effect measures the effect of treatment on the treated. We would like to compare the observed outcome after treatment to the unobservable outcome of the same individual if we had not treated them.

$$\mathbf{E}[y_{\ell 0,t}^1 | d_{\ell 0,t} = 1] - \mathbf{E}[y_{\ell 0,t}^0 | d_{\ell 0,t} = 1] \quad (2.1)$$

ℓ denotes the subnetwork which is treated in period t and c the subnetwork that is not. $d_{i,t}$ indicates if node i itself was directly treated. Superscript 1 denotes the outcome of a treated observation and 0 the outcome of the untreated counterpart. $\mathbf{E}[y_{\ell 0,t}^0 | d_{\ell 0,t} = 1]$ is the unobservable counterfactual.

We estimate this counterfactual term using a comparable node¹⁶ in a period where it is not treated. I take two approaches to obtain such an estimate: (i) a simple approach compares the observation “before and after” the treatment. It attributes all observed changes in outcomes to the treatment.¹⁷

Assumption Direct Treatment Effect before-after:

$$\mathbf{E}[y_{\ell 0,t}^0 | d_{\ell 0,t} = 1] = \mathbf{E}[y_{\ell 0,t-1}^0 | d_{\ell 0,t-1} = 0] \quad (2.2)$$

The “before and after” counterfactual estimate maximizes the similarity between the treated and untreated observations. However, it will fail to capture any period-specific effects that would have affected all nodes even absent treatment. Any weekday

¹⁵Readers who know the estimation of direct and indirect treatment effects might wish to merely browse the formulas or skip this subsection. The identifying assumption will be: Absent treatment, the control observations have a similar rate of change across time to the treated subnetworks. They must grow at similar rates and be affected similarly by any Wikipedia wide dynamics such as weekdays etc.

¹⁶A node, which is believed to be affected by treatment in similar ways.

¹⁷If the object/individual was observed more than once before treatment it might be possible to further improve this approach by accounting for trends in the outcomes etc.

fluctuations, shocks etc. will be attributed to the treatment. (ii) Alternatively, “difference-in-differences” uses individuals in the same populations, which were not eligible for treatment, or here, eligible individuals in untreated subpopulations. The unobservable counterfactual outcomes of the treated nodes are assumed to be the treated nodes’ pre-treatment outcome *plus* the *change* of the non-treated control observation.

Assumption Direct Treatment Effect-DiD:

$$\begin{aligned} \mathbf{E}[y_{\ell 0,t}^0 | d_{\ell 0,t} = 1] &= \mathbf{E}[y_{\ell 0,t-1}^0 | d_{\ell 0,t-1} = 0] + \\ &+ \{ \mathbf{E}[y_{c0,t}^0 | d_{c0,t} = 0] - \mathbf{E}[y_{c0,t-1}^0 | d_{c0,t-1} = 0] \} \end{aligned} \quad (2.3)$$

Note that in the context of Wikipedia articles, the crucial assumption is not that articles *are* very similar but that they *evolve* in a similar way. On average they have similar growth in readership and edits and are subject to similar intertemporal fluctuations.

Indirect Treatment Effects: The *ITE* measures the spillover or externality effect of treatment of eligibles on the outcomes of non-eligibles. As for the direct treatment effect, we cannot observe the outcome of the non-eligibles had the eligibles not been treated. Knowing the distance to the treatment’s epicenter, I can compare the nodes of the subnetworks by layer. ITE_1 refers to the effect on direct neighbors, ITE_2 for indirect neighbors and so on.¹⁸ My dataset requires even more involved notation because I differentiate nodes along four dimensions (treatment, time, distance and subnetwork). I use $D_{xr,t}$ as shorthand that takes the value 1 if both of the following conditions are simultaneously satisfied: (i) subnetwork x was treated and (ii) there exists a treated node exactly r steps away by the shortest route. For direct neighbors we have:

$$ITE_1 = \mathbf{E}[y_{\ell 1,t}^1 | D_{\ell 1,t}^1, d_{i,t}^0] - \mathbf{E}[y_{\ell 1,t}^0 | D_{\ell 1,t}^1, d_{i,t}^0] \quad (2.4)$$

As before, $d_{i,t}^1$ indicates if node i was directly treated in period t .¹⁹ $y_{\ell r,t}^1$ is now the outcome if *some neighbor* in D_{xr} was treated in t , and $y_{\ell r,t}^0$ denotes the outcome if nobody in that set was treated.

¹⁸Well known papers that estimate ITE_1 s are Angelucci and De Giorgi (2009), Kuhn et al. (2011) or Crépon et al. (2013), to name a few. Miguel and Kremer (2003) include distance layers in the estimation to incorporate a similar notion of distance to treatment in a real world setup.

¹⁹To save space treatment status is indicated by superscripts, $d_{i,t}^0$ otherwise. Notation has to be more involved here, because it is no longer possible to talk of a single node, as the treated nodes can have many different neighbors.

The object of interest is the ITE_r , Generally, for any range r :

$$ITE_r = \mathbf{E}[y_{\ell r,t}^1 | D_{\ell r,t}^1, d_{i,t}^0] - \mathbf{E}[y_{\ell r,t}^0 | D_{\ell r,t}^1, d_{i,t}^0] \quad (2.5)$$

As in the direct treatment effect, I have to estimate the counterfactual outcome using two methods: (i) a “before and after” comparison (ii) a difference-in-differences between neighbors in the comparison subnetwork.

Assumption ITE_r -before-after:

$$\mathbf{E}[y_{\ell r,t}^0 | D_{\ell r,t}^1, d_{i,t}^0] = \mathbf{E}[y_{\ell r,t-1}^0 | D_{\ell r,t-1}^0, d_{i,t-1}^0] \quad (2.6)$$

Estimating an ITE_r from a “before and after” estimation has the same advantages and drawbacks as the direct treatment effect. Analogously, the drawbacks can be accounted for by computing a difference-in-differences estimator. In the context of an ITE , we need to observe comparable, but untreated, subpopulations.²⁰

Assumption ITE_r -DiD:

$$\begin{aligned} \mathbf{E}[y_{\ell r,t}^0 | D_{\ell r,t}^1, d_{i,t}^0] &= \mathbf{E}[y_{\ell r,t-1}^0 | D_{\ell r,t-1}^0, d_{i,t-1}^0] + \\ &+ \{ \mathbf{E}[y_{cr,t}^0 | D_{cr,t}^0, d_{i,t}^0] - \mathbf{E}[y_{cr,t-1}^0 | D_{cr,t-1}^0, d_{i,t-1}^0] \} \end{aligned} \quad (2.7)$$

The counterfactual is estimated by last period’s value *plus* the comparison group’s rate of change. The same assumptions apply as for the direct treatment effect difference-in-differences. Before moving on to the econometric specification, I conclude this section by summarizing the identification result in terms of the difference-in-differences estimator:

Conclusion ITE_r DiD: If Assumption ITE_r -DiD holds, the difference below identifies the ITE_r .

$$\begin{aligned} ITE_r &= \mathbf{E}[y_{\ell r,t}^1 | D_{\ell r,t}^1, d_{i,t}^0] - \{ \mathbf{E}[y_{\ell r,t-1}^0 | D_{\ell r,t-1}^0, d_{i,t-1}^0] + \\ &+ (\mathbf{E}[y_{cr,t}^0 | D_{cr,t}^0, d_{i,t}^0] - \mathbf{E}[y_{cr,t-1}^0 | D_{cr,t-1}^0, d_{i,t-1}^0]) \} \end{aligned} \quad (2.8)$$

Hence, our estimator of the ITE_1 is based on the pre-treatment outcomes and comparing the *change* in the outcomes of direct neighbors of the eligible nodes in a *treated* subnetwork to the direct neighbors of the eligible nodes in the *non-treated* subnetwork. Thus, (indirectly) treated and control observations must grow at similar rates and be affected similarly by any dynamics that affect the entire Wikipedia

²⁰ Ideally we would like to observe a random selection of the subpopulations in which any treatments are to be administered, and in the second step we administer treatment to the eligible nodes. Moreover, we observe both subpopulations before the treatment of one takes place.

(weekday dynamics etc.). Note that this conclusion also applies to the direct treatment effect, when setting r to 0.

2.3.3 Reduced Form Analysis

To obtain the *ITEs* for each layer, I apply reduced form regressions which allow the understanding of the impact of the local treatment on both the treated pages and their neighbors. These are very similar in spirit to the analysis in Carmi et al. (2012). The idea is to compare pages grouped by their distance to the page which experiences treatment to their analogue in the control group (L_0 to C_0 , L_1 to C_1, \dots). I denote all reduced form coefficients by ϕ . Furthermore, I define “treatment” for each set of pages along the lines of the indirect treatment effects (ITE_r) in the previous section.²¹ I let s indicate the day relative to day 0, the day when the treatment is administered. Hence s runs from -14 to 14. λ_s is an indicator, which takes the value 1 if $t = s$ and 0 otherwise. Each set of pages that corresponds to one layer in the network is regressed separately. So if I focus on the treated nodes, the neighbors and the indirect neighbors, it results in the following system of fixed effect regression equations, which all are based only on dummy variables:

L_0 .) Difference in Differences specification at level L_0 ²²:

$$y_{it} = \phi_i^{L_0} + \sum_{s \in S} \phi_{1,s}^{L_0} \lambda_s + \sum_{s \in S} \phi_{2,s}^{L_0} (\lambda_s * treat_{L_0,i}) + \xi_{it} \quad (2.9)$$

... $treat_{L_0}$: treatment on the very page; $S = \{-14, \dots, 14\}$

L_1 .) At level L_1 ($treat_{L_1}$ means the shock is 1 click away):

$$y_{it} = \phi_i^{L_1} + \sum_{s \in S} \phi_{1,s}^{L_1} \lambda_s + \sum_{s \in S} \phi_{2,s}^{L_1} (\lambda_s * treat_{L_1,i}) + \xi_{it} \quad (2.10)$$

L_2 .) At level L_2 ($treat_{L_2}$ means the shock is 2 clicks away):

$$y_{it} = \phi_i^{L_2} + \sum_{s \in S} \phi_{1,s}^{L_2} \lambda_s + \sum_{s \in S} \phi_{2,s}^{L_2} (\lambda_s * treat_{L_2,i}) + \xi_{it} \quad (2.11)$$

²¹The dummy in the regression for the neighbors (sets L_1 and C_1) takes the value 1, not if the node was itself treated, but if the corresponding start node (ℓ_0) was treated in t (and 0 otherwise).

²²The specifications I use are fairly standard “regression difference in differences” similar to Jacobson et al. (1993) or as described in Angrist and Pischke (2008).

In words, I run the same difference-in-differences on three levels (on L_0 , L_1 and L_2 (shown only for large events)). $treat_{L0,i}$ is an indicator variable for a page that is (going to be) featured on Wikipedia’s main page, $treat_{L2,i}$ takes the value of 1 for pages that are two clicks away from pages that are (going to be) affected by such a shock. The cross terms correspond to this indicator variable multiplied with the time dummies. Thus, a cross term captures whether treatment has occurred at a given point in time or not. For an observation in the control-group this variable will always take a value of 0, while for an observation in the treated group this variable will take a value of 1 if it corresponds to the event time the time-dummy aims to capture. Hence, if the treatment is effective, the coefficients of the cross terms are expected to be 0 before treatment occurs and positive for the periods after the treatment. The *ITEs* from the previous subsection are captured by the ϕ_2 coefficient that corresponds to day 0 in the regressions above. I look at $\phi_{2,0}^{L_1}$ for the *ITE*₁, which corresponds to L_1 and analogously at $\phi_{2,0}^{L_0}$ for L_0 and $\phi_{2,0}^{L_2}$ for L_2 .

Other than the cross terms I also include page fixed effects and another full set of time dummies (event time) to control for general (e.g. weekday-specific) activity patterns in Wikipedia. This procedure is crude because it does not consider several important factors such as how well neighbors are linked among each other or how large the peak of interest is on the originally shocked page. Yet, it is useful, since the results from the reduced form analysis are based on minimal assumptions and provide guidance as to whether attention spillovers exist at all. They also allow us to see how far they carry over, and whether they result in increased production. Finally, they allow me to provide a lower bound and an upper bound estimate of the aggregate spillover effects to be expected.

2.3.4 Structural Form Analysis and Bounds

Beyond measuring the size of the *ITE*, I am interested in quantifying the size of the spillovers of attention that exist between Wikipedia articles on normal days. In this section, I augment the well known linear-in-means model for peer effects, as formulated in Manski (1993), with exogenous shocks. Departing from the version that was used by Bramoullé et al. (2009),²³ I show how exogenous shocks can be exploited to identify spillovers (or peer effects). This is possible in my modification of the model, even if the nodes characteristics or the network structure are endogenous. In other words, exogenous shocks are used as a focal lens to identify the spillovers, which is usually very challenging. In this section I provide only the point of departure and the main

²³They show how identification of peer effects can be achieved in social networks, using an IV-strategy.

results. The details and derivations can be found in Appendix A.2.²⁴

Recall that the underlying relationship of interest is the role of links.²⁵ How much attention spills via links can be modeled using the well known linear-in-means model of the type discussed in Manski (1993), who shows that the coefficient of interest is generally very hard to identify. I start from the same form of model.²⁶

$$y_{it} = \alpha \frac{\sum_{j \in P_{it}} y_{jt}}{N_{P_{it}}} + X_{it-1}\beta + \gamma \frac{\sum_{j \in P_{it}} X_{jt-1}}{N_{P_{it}}} + \epsilon_{it}$$

y_{it} denotes the outcome of interest in period t and X_{it-1} are i 's observed characteristics at the end of period $t - 1$.²⁷ P_{it} is the set of i 's peers and $N_{P_{it}}$ the number of i 's peers. The coefficient of interest is α : It captures the effect of the performance of i 's peers and in the present context it measures how the views of an article are influenced by the views of the adjacent articles. The coefficient vector β accounts for the impact of i 's own characteristics and γ measures the effect of the peers' average characteristics on i 's performance. In the setting of this paper β accounts for how the page's own length or quality might affect how often it is viewed and γ captures how length and quality of neighboring pages affect views of page i . Bramoullé et al. (2009) suggest a more succinct representation based on vector and matrix notation:

$$\mathbf{y}_t = \alpha \mathbf{G} \mathbf{y}_t + \beta \mathbf{X}_{t-1} + \gamma \mathbf{G} \mathbf{X}_{t-1} + \epsilon_t \quad \mathbf{E}[\epsilon_t | \mathbf{X}_{t-1}] = 0$$

A few remarks: \mathbf{G} is a $N \times N$ matrix. $G_{ij} = \frac{1}{N_{P_i}-1}$ if i receives a link from j and $G_{ij} = 0$ otherwise. Clearly this model and, specifically, measuring the social parameter α is of interest to a very large literature. To incorporate exogeneous variation, I augment this model by including a vector of treatments, which for simplicity, is assumed to take the value of 1 for treated nodes and the value of 0 otherwise.

$$\mathbf{y}_t = \alpha \mathbf{G} \mathbf{y}_t + \mathbf{X}_{t-1}\beta + \gamma \mathbf{G} \mathbf{X}_{t-1} + \delta_1 \mathbf{D}_t + \epsilon_t \quad \mathbf{E}[\epsilon_t | \mathbf{D}_t] = 0 \quad (2.12)$$

For the treated side \mathbf{D}_t is a vector consisting of zeros and ones that indicates which nodes are treated. On the untreated subnetwork we have $\mathbf{D}_t = \mathbf{0}$, a vector of zeros. In some of the proofs and in my application I will assume a local treatment that affects only a single node. This captures the notion of a local treatment condition,

²⁴The derivations involve quite heavy notation, but are otherwise relatively straightforward.

²⁵The mechanism we have in mind, is that attention from article A can be diverted to article B if a link exists. This is interesting, since some of the users who get to see B might later start to edit it.

²⁶Note that it is easy to add a fixed effect to the model, but that it will be eliminated when taking differences. Consequently, I omit it for ease of notation.

²⁷Note, that I can observe the current state of a Wikipedia article once a day at a fixed time.

under which only one node is exposed to treatment (a “mini population treatment”). Formally this is written as $\mathbf{D}_t = \mathbf{e}_{\ell 0}$; that is, a vector of zeros and a unique one in the coordinate that corresponds to the treated node.

Note that I do not require that the structure of the network (\mathbf{G}) to be the result of an exogenous network formation process. Rather only the selection which of the eligible node that gets treated must be exogenous.²⁸ It is worth stressing that my setup is fundamentally different from Bramoullé et al. (2009) because it will use an entirely different source of identification. Moreover, there will be no requirements needed concerning the linear independence of \mathbf{G} and \mathbf{G}^2 .

In this model, the reduced form expectation conditional on “treatment” is given by:

$$\mathbf{E}[\mathbf{y}_t | \mathbf{D}_t] = (\mathbf{I} - \alpha \mathbf{G})^{-1}[(\beta + \gamma \mathbf{G})\mathbf{E}[\mathbf{X}_{t-1} | \mathbf{D}_t] + \delta_1 \mathbf{D}_t] \quad (2.13)$$

Define the set of observations in the subnetwork where treatment occurs in t by the subscript ℓ and a comparison group in which no node is treated by the subscript c . If these sets of nodes can also be observed one period earlier, a difference-in-differences estimator can be computed.

Result 1: Denote the difference in differences estimator as

$$\text{DiD} := \{\mathbf{E}[\mathbf{y}_{\ell,t} | \mathbf{D}_{\ell,t}] - \mathbf{E}[\mathbf{y}_{\ell,t-1} | \mathbf{D}_{\ell,t-1}]\} - \{\mathbf{E}[\mathbf{y}_{c,t} | \mathbf{D}_{c,t}] - \mathbf{E}[\mathbf{y}_{c,t-1} | \mathbf{D}_{c,t-1}]\}$$

and assume that the treatment affects only the contemporary outcome of the treated node and not its exogenous characteristics.²⁹ Then the DiD contains the following quantity:

$$\text{DiD} = \delta_1 \mathbf{D}_t (\mathbf{I} + \alpha \mathbf{G} + \alpha^2 \mathbf{G}^2 + \alpha^3 \mathbf{G}^3 + \dots)$$

Proof: For a proof please refer to Appendix A.2.3.

In words, this result means that the node is affected by both treatment and second and higher order spillovers, the positive feedback loop that ensues as the neighbors increase their performance in sync with their peers. Instances of higher order effects³⁰

²⁸In the present application, all “eligible” nodes (featured articles) are equally likely to be treated. They are the nodes in the group L_0 . Neighbors (in L_1) are typically not featured. Hence they are not eligible and naturally less likely to be themselves treated.

²⁹The independent characteristics X should not be immediately affected by treatment because this would threaten the identification of the spillover. However, they may adjust over time. As long as we can observe one period where only the outcome is affected, but not the characteristics, the result holds.

³⁰Note that I am considering the homogeneous network, so all spillovers have the same magnitude.

are $\alpha^2\delta_1$ in the second round or $\alpha^3\delta_1$ in the third round and so on. The other important factor is whether and how often spillovers of a given order q arrive. This depends on the number of indirect paths of length q that go from the shocked node $\ell 0$ to any focal node j .³¹

Note the close relationship to the Bonacich centrality in the paper by Ballester et al. (2006), who aim at identifying the “key player” of a network. Like in their framework, the *number of channels* for indirect spillovers matters. Yet, for measuring spillovers in a “mini population treatment” we care about the reverse direction, the quantity that spills from the shocked node to any other node.

My result shows that the difference-in-differences approach alone will not directly reveal α , the social parameter of interest, because nodes might have a feedback effect on each other. The neighbor’s change in performance (due to the original impulse) will affect the neighbors’ neighbors, but also feed back to the originally treated $\ell 0$ -node. The estimator will observe all the changes in outcome at the end of this process, when all higher order spills have taken place.

In some applications this will be the object of interest to the researcher. However, in the present context, the research is motivated by the desire to know the effect of the link structure and not of the treatment per se. Consequently it is warranted to dig deeper in order to understand the structural parameters.

Computing the parameters is not necessarily feasible, because it requires knowledge of the complete link structure. However, a closer look at the nodes independently reveals that limited information about the link structure suffices to acquire additional information about the parameters. In the following two subsections I show how to get the point estimate for the peer effects coefficient if the network is known and I show how to derive upper and lower bound estimates for the parameter if no information about the network is available.

Estimator of the Peer Effects Parameter if the Network Structure can be Observed

If the network structure can be observed, the peer effect parameter α can be backed out by computing the higher orders of the network graph (\mathbf{G} -matrix). To know how many spillovers arrive in each round, it suffices to focus on the entries $\mathbf{G}_{ij}, \mathbf{G}^2_{ij}, \mathbf{G}^3_{ij}, \text{etc.}$ ($i = \ell 0$) that document the number of paths via 1, 2, 3 and more links from the treated node to the neighboring node in question. With this information it is straightforward to

³¹In the proof I need to assume that the network formation *process* is not affected by the treatment. I checked this assumption in my “today’s featured article application” and verified, that link formation remains on low levels. If anything, there is an increase by 0.2 in-links per article in sync with the peak in edits, but not with clicks. I conclude that this is an acceptably small source of potential bias.

compute by how much the observed effect at the node in question has to be discounted and to use this information to compute the true average effect.

Upper and Lower Bound Estimates of the Peer Effects Parameter if the Network Structure is Unobserved

If the network structure cannot be observed, it is still possible to obtain boundary estimates for the peer effects based merely on two separate comparisons of (i) the directly treated nodes and their counterparts (L_0 vs. C_0) in the control group and (ii) their neighbors (L_1 vs. C_1). This is relevant in many empirical settings, because randomization and information on the network *together* are rarely available. In contrast, a separate comparison of eligible and non-eligible nodes in randomly treated communities or networks (without network information) can frequently be observed. Also with this restricted knowledge it is possible to obtain a lower bound estimate for the coefficient α , if the researcher is willing to make more rigorous, but sensible, assumptions. In what follows I briefly show how to obtain the bounds. The idea behind this derivation is to select two specific “extreme” types of network which either minimize or maximize second and higher order spillovers. These benchmark networks are schematically represented in Figure 2.2. I use a directed network with only “outward bound” links emanating from ℓ_0 to $\ell_1 \in L_1$ to obtain the upper bound estimate of the social/spillover parameter α .³² The opposed benchmark is a fully connected network, where every node is the direct neighbor of every one of its peers. From there I obtain the lower bound estimate of the social parameter. A more detailed account is provided in Appendix A.2.4.

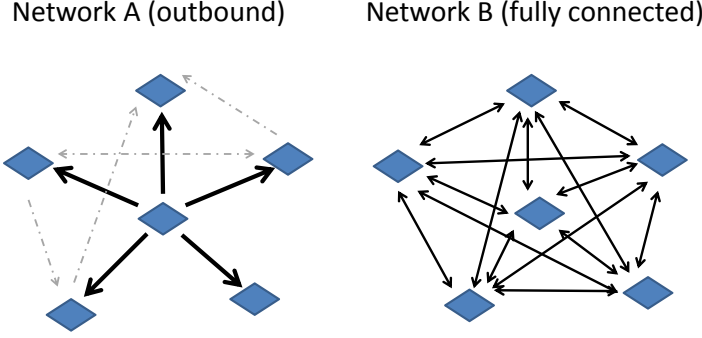
Upper Bound: If we ignore higher order spillovers,³³ we can obtain an upper bound estimate for the direct treatment effect ($\bar{\delta}_1$) by applying the difference-in-differences estimator on the level of directly treated nodes (L_0) and a suitable comparison group (C_0). After that I can move on to estimate the upper bound for the parameters for spillovers ($\bar{\alpha}$) by combining it with a second difference-in-differences estimator at the neighbor level. Let $DiD_{(\ell a - ca)}$ denote such a difference-in-differences, ($a \in \{0, 1\}$), where the nodes are either in the center of the network (L_0 or C_0) or are the neighbors of the start nodes (L_1 vs. C_1):

$$\begin{aligned}\hat{\bar{\delta}}_1 &= \widehat{DiD_0} = \hat{\Delta\ell_0} - \hat{\Delta c_0} \\ \hat{\bar{\alpha}} &= \frac{\widehat{DiD_1}}{\widehat{DiD_0}} NP_{\ell_1}\end{aligned}\tag{2.14}$$

³²For this benchmark we ignore any existing links among L_1 nodes.

³³Or maintain the assumption that we can observe the nodes’ performance before any higher order spillovers arrive at the treated node

Figure 2.2: Schematic representation of the two extreme networks, used to compute the upper and lower bound estimates of the parameters of interest.



NOTES: The “outbound network” (left) is used to obtain the upper bound estimate. It is a directed network with only “outward bound” links. This implies ignoring any existing links among L_1 nodes. Holding the number of nodes and the observed ITEs fixed, the social parameter will be estimated to be largest in this type of network. The fully connected network (right), is the benchmark case from which the lower bound of the social parameter can be estimated.

- $\Delta\hat{\ell}0 := \frac{1}{NP_{\ell 0}} * \sum_i (y_{i,\ell 0,t=1} - y_{i,\ell 0,t=0})$
- $\Delta\hat{c}0 := \frac{1}{NP_{c0}} * \sum_i (y_{i,c0,t=1} - y_{i,c0,t=0})$

with $\widehat{DiD_1} = \Delta\hat{\ell}1 - \Delta\hat{c}1$ and the definitions of $\Delta\hat{\ell}1$ and $\Delta\hat{c}1$ paralleling those of $\Delta\hat{\ell}0$ and $\Delta\hat{c}0$. In my application’s reduced form estimations of the previous section DiD_1 corresponds to $\phi_{2,0}^{L_1}$ and DiD_0 is estimated by $\phi_{2,0}^{L_0}$. This upper bound estimator would be suitable under the potentially quite strong assumption that higher order spillovers are negligible. I proceed to show how to compute the lower bound estimates under the assumption of *maximal* second order spillovers. The lower bound gives an idea of the maximal size of the problem that might result from trusting the easily computed upper bound estimates.

Lower Bound: It is also possible to compute a lower bound estimate for α and δ_1 . This bound can be obtained by imagining that the network is fully connected, i.e. every node links to every other node, assuming that all effects are of the same sign, strictly ordered and (w.l.o.g) positive.³⁴ Further computations in Appendix A.2 show that in a network with N nodes, the lower bound of the estimator for α is characterized

³⁴The precise assumption is $DiD_0 > DiD_1 > HO^B > 0$, as stated and explained in Lemma A.1

by the solution to the following quadratic equation:

$$\underline{\alpha}^2 - \left[\frac{DiD_0}{DiD_1} + (N - 1) \right] \underline{\alpha} + (N - 1) = 0 \quad (2.15)$$

This equation has two solutions, one of which lies between 0 and 1. The closed form solution for $\underline{\alpha}$ is hence given by:

$$\underline{\alpha} = \frac{1}{2} \left[\frac{DiD_0}{DiD_1} + (N - 1) \right] - \sqrt{\frac{1}{4} \left[\frac{DiD_0}{DiD_1} + (N - 1) \right]^2 - (N - 1)} \quad (2.16)$$

Recall that all the quantities required are readily available from the reduced form estimations. DiD_1 corresponds to $\phi_{2,0}^{L_1}$ and DiD_0 is estimated by $\phi_{2,0}^{L_0}$. In Appendix A.2.4 I provide a proof for my claims and explain how this bound is derived. Which of the estimates is more accurate will depend on the size of the spillover effect, but to a very large extent also on the real network structure and the number of nodes.

A closer examination of Result 1 reveals that the upper bound estimator would be quite suitable if the researcher has reasons to make the (potentially quite strong) assumption that higher order spillovers are negligible. It would also be appropriate in networks with very sparse connections among its members. The lower bound estimator might be more suitable if the researcher believes the network to be highly connected and expects the spillover coefficient to be relatively large.³⁵ The bounds have several limitations (cf. Appendix A.2.4) and for some applications the bounds might turn out to be too wide to be actually informative. Still, taken together, the bounds can provide a useful first characterization of the spillover parameters in question.

2.4 Data

This section briefly surveys the data collection procedure in Subsection 2.4.1, and describes the datasets used both for disasters (large events) and “today’s featured article” (Subsection 2.4.2). A more detailed description of how the underlying database was put together and the procedure I used to extract the dataset is provided in Appendix A.1.

³⁵So large that α^2 and α^3 are still sizeable.

2.4.1 Data Preparation - Treated and Control Group

To obtain my dataset I augment the publicly available data dumps provided by the Wikimedia Foundation³⁶ with data on the link structure between articles, data on the download frequency of pages and information on major media events which occurred during our period of observation. The data I use are based on 153 weeks of the entire German Wikipedia’s revision history between December 2007 and December 2010.³⁷

I use 23 large-scale events, 34 articles that were featured on Wikipedia’s main page, and all their respective network neighbors. “Featured articles” were found by consulting the German Wikipedia’s archive of pages that were selected to be advertised on Wikipedia’s main page (“Seite des Tages”). To identify major events, I consulted the corresponding page on Wikipedia. I focus on the content provision that results from attention spillovers and which is a consequence of the spike in interest and the resulting improvements to the linked pages. Hence, I will not focus on the treated pages where content generation might be related to the events directly. Instead I obtained data on the direct and indirect network neighbors.

For each primary shock I obtain two sets of control observations. The first set is based on pages which are similar but unlikely to be affected by the treatment. I selected other featured articles and neighbors thereof that were advertised either later or earlier in time. This gives me a set $C1_{control}$ which is similar in both size and characteristics to the sampled pages (before the shock). The second set is obtained by extracting the data based on treated pages a second time, but 42 days before the actual shock occurred. I refer to the articles in this “placebo-treatment” as $C1_{placebo}$.

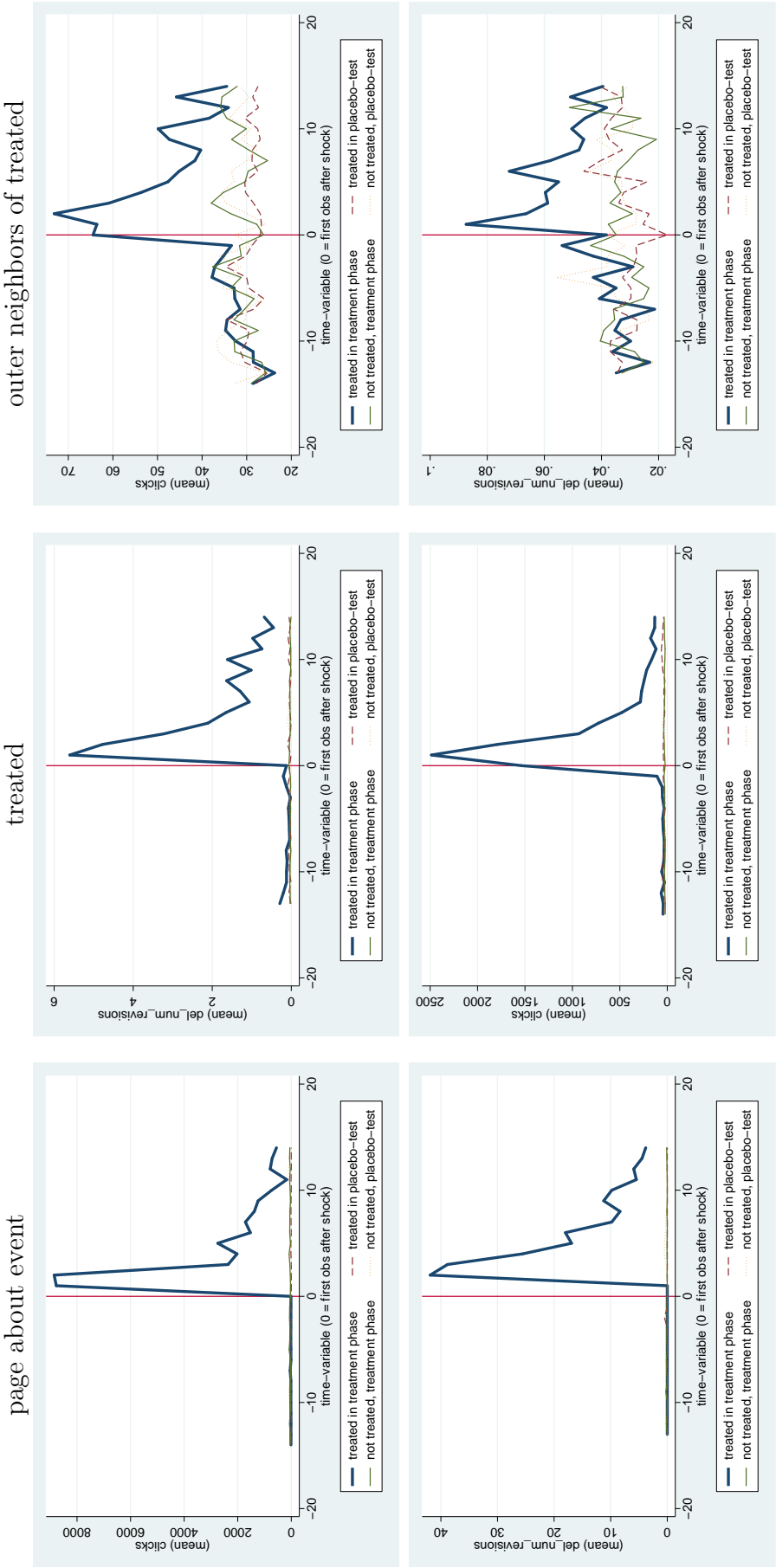
The resulting dataset contains information on views and content generation of almost 13,000 articles, 14 days before and after the events (more than 750,000 observations). Table 2.8 shows which events were included in the “today’s featured article” dataset and the associated number of observations for each of the conditions. Column 1 shows the number of articles that belong to each featured article. Columns 2-4 show the corresponding number of observations, separated by treatment status.³⁸ Table 2.9 shows the information for the data on large events which includes both natural disasters as well as technical or economic catastrophes. More details about the data preparation and selection of the events are provided in Appendix A.1.

³⁶I have access to a database that was put together in a joint effort of the University of Tübingen, the IWM Tuebingen and the ZEW Mannheim. It is based on data from the German project, which currently has roughly 1.4M articles and thus provides us with a very large number of articles to observe.

³⁷The data were stored in a relational database (disk-based) and queried using Database Supported Haskell (DSH) (Giorgidze et al. (2010)).

³⁸Observations range from 2,088 to 33,872, covering various topics such as innovations (CCD-sensor), art history (Carolingian book illustrations) or soccer clubs (Werder Bremen).

Figure 2.3: Catastrophes: Comparing average clicks (new edits) of treated pages (and neighbors) to three comparison groups.



NOTES: The figure shows the results for natural disasters and large accidents. The left column shows the average effect on the pages about the disaster (“event pages” - by definition, they were created after the event), the middle column the directly treated pages, that users turn to, until the event gets a page of its own (“L1”, with reciprocal link to the future event page), and the right column for the pages that are one click away from L_1 . The upper row shows the average number of clicks the lower row shows the average number of edits. The outcomes are shown for the treated articles and the control groups separately. Directly hit pages received up to 8,500 additional clicks and up to 40 new revisions on average. Pages that will have a reciprocal link received up to approx. 2,500 clicks and up to 5 additional revisions. However, not only the treated pages, but also their neighbors received 35 additional clicks and up to 0.04 additional revisions on average.

2.4.2 A Closer Look at the Datasets

Summary statistics for the data on large events are shown in Table 2.7. The data contains 425,981 observations from 7,379. From the table it can be seen that the average page contains 5658 bytes of content and has undergone 84 revisions. However, the median is substantially lower at 3885 bytes and only 40 revisions. Also, the summary statistics of the first differences (variables starting with “Delta:”) reveal that on a typical day nothing happens on a given page on Wikipedia. This highlights the necessity of using major events as a focal lense for analyzing activity on Wikipedia,³⁹ which is confirmed by the visual inspection of the direct and indirect effect of treatments.

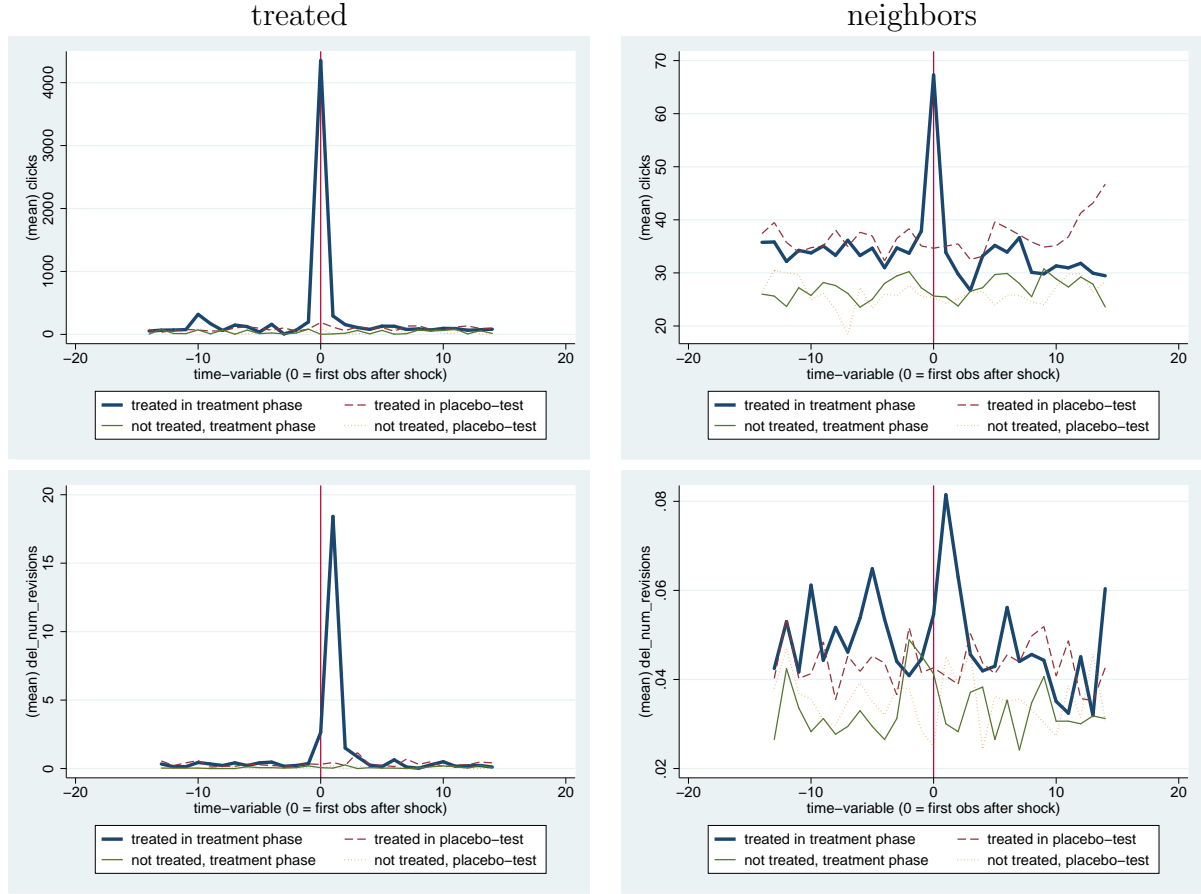
In Figure 2.3 I plot the average clicks (top row) and the average number of added revisions (bottom) for the three groups of pages zero clicks away (left column), one click away (middle column) and two clicks away (right column). Each of the plots features four lines. The bold blue line represents the treated group or its neighbors when they were actually treated, hence “treated in treatment phase”. The dashed red line represents the same group but during the placebo treatment at an earlier point in time. The thin green line (“not treated, treatment phase”) shows the control group at the time when the real shock occurred and the thin dotted yellow line represents the “unrelated” observations, which are never treated and observed in the placebo period.⁴⁰ The left column shows the control group and the article about the incident (“event page”; L_0), which are created only after the onset of the event. Most of these 23 pages did not exist at all before the onset of the event and therefore only a few have a placebo condition available. The upper row shows that the directly affected pages experience a very large spike of 8,500 clicks per day on average. The number of additional revisions peaks on the first days of treatment, when the pages are created: an average of almost forty revisions are added to a page on the first day. On the pages that are to share a reciprocal a link from the treated page the effect is quite large: while the clicks on the average L_1 page increase by 2,500, the absolute value of the average increase in revision activity is only five. When I look at pages that are two clicks away, the effects are much smaller, especially for revisions, but quite pronounced. The clicks on the average adjacent page go up by 35 and the absolute value of the average increase in revision activity is already no more than 0.04.

A summary of the data from “featured articles” are shown in Table 2.6. The data

³⁹Further descriptive analyses that compare treated and control groups before and during treatment show that the groups are very similar in their activity levels before the shocks occurred and that the control group did not change its behavior during treatment. These tables and their description were omitted for reasons of brevity. They are available from the author upon request.

⁴⁰For greater ease of representation I included a graphical representation of only two variables. The summary statistics for these groups before and after treatment are also available as tables upon request.

Figure 2.4: “Today’s featured articles”: Comparing average clicks (new edits) of treated pages (and neighbors) to three comparison groups.



NOTES: The figure shows the results for featured articles that were advertised for a full day on Wikipedia’s main page. The left column shows the average outcome on the directly treated pages (set “ L_0 ” containing 63 pages total), the lower row for the pages one click away (set “ L_1 ”, which contains 5,489 pages). The upper row shows the average number of clicks the lower row shows the average number of edits.

contains 317,550 observations from 5,489 pages⁴¹ on the main variables. Note that this corresponds to a much smaller number of pages per treatment, which is due to the fact that I focus on the directly linked pages in this condition. The table shows that the median page contains 4833 bytes of content and has undergone 48 revisions. In this sample, the mean is substantially higher at 6794 bytes and 95 revisions. As before, the summary statistics of the first differences show how little activity occurs on a normal day on any given page on Wikipedia.

Figure 2.4 plots the aggregate dynamics around the day when the start node was shown on Wikipedia’s main page and corresponds to Figure 2.3 for the large event

⁴¹Since pages were observed also in the placebo condition, each page is sampled twice, and hence I observe 10,950 distinct time series.

condition. I plot the average clicks (left column) and the average number of added revisions (right columns), but now only for the treated pages and direct neighbors. As before, each of the four figures contains four lines, one for each condition that can be obtained by combining treatment (yes/no) and placebo (yes/no). The major difference to the large events condition is the brevity of the treatment. Attention rises from typical levels, below 50 views, to more than 4200 views on average, but immediately returns to the old levels the day after treatment is administered. A very similar pattern can be observed for the neighbors where attention is almost twice as high as on a usual day and then falls back to the old levels. A similar pattern can be observed for the number of revisions. Excepting large events, activity rises already before $t = 0$. Nevertheless, on the day of treatment the spike of activity is also pronounced for the neighbors.⁴²

2.5 Estimation Results

In what follows I present my estimation results and discuss their interpretation. Before I proceed with the details of my estimations, it is worth recalling a few important facts. The point of departure of the estimations in this paper is estimating Equation 2.11 (Section 2.3.3) for large events and Equation 2.10 for “featured articles”. This is due to two reasons: first, the two conditions differ in how local the treatments are. Second, only the “featured articles” exist at the time of treatment, while the page at the center of a large event treatment typically does not exist and will instead be created in the following days.

Moreover, I avoid potentially endogenous link formation during treatment by considering only links that had been in place *a week before the treatment*. When a page is found to lie in both the treatment and control groups it is excluded from the estimation, because including such pages will bias the estimated coefficients towards zero. Extremely broad pages with a very large number of links (e.g. pages that correspond to years) were excluded from estimation to avoid biases from oversampling. Finally, I use the seven observations from two weeks before treatment (days -14 through to -8) as the reference group in the estimations and I include only flow variables such as views, new revisions, new authors etc. to guarantee that my results are not driven

⁴²Note that I cannot cleanly estimate the direct treatment effect if the treatment drastically increased the number of links. (cf. Comola and Prina (2013)). This may be a minor issue for disasters and, if important, introduces noise in the quantification of the conversion rates. I checked this for “today’s featured articles” and found that it is a minor issue. Link formation increases by 0.2 new in-links over 120 in-links per article on average on the day after treatment. It moves in sync with the peak in edits, but not with clicks. This is a small source of potential bias resulting from the edit activity, but is unlikely to affect viewership. Hence, the result from “today’s featured articles” are my preferred estimates.

by any anticipation effects.⁴³ The following two subsections report the results for both conditions.

2.5.1 Large Events

Table 2.1: Large events: clicks/added revisions over time for indirect neighbors.

	clicks			new edits		
	(1) before after	(2) compare control	(3) compare placebo	(4) before after	(5) compare control	(6) compare placebo
t = -2	4.442 (4.372)	3.172 (4.709)	3.487 (4.545)	0.011 (0.008)	0.010 (0.009)	0.015 (0.010)
t = -1	2.639 (3.040)	0.978 (3.993)	3.144 (3.742)	0.022*** (0.007)	0.010 (0.010)	0.026*** (0.008)
t = 0	33.661** (14.471)	37.391** (14.421)	36.047** (14.386)	0.006 (0.010)	0.003 (0.011)	0.021* (0.011)
t = 1	32.794*** (11.075)	35.020*** (11.098)	35.397*** (11.113)	0.055** (0.023)	0.049** (0.024)	0.062** (0.023)
t = 2	42.375*** (13.671)	38.767*** (13.650)	44.730*** (13.589)	0.034*** (0.012)	0.037** (0.014)	0.043*** (0.012)
t = 3	30.066*** (8.283)	22.069** (9.168)	30.895*** (8.730)	0.027*** (0.009)	0.021* (0.011)	0.025** (0.011)
t = 4	22.871*** (6.850)	17.601** (7.065)	21.918*** (6.917)	0.027** (0.012)	0.026** (0.012)	0.028** (0.013)
Constant	30.795*** (2.296)	29.900*** (1.001)	29.994*** (1.289)	0.032*** (0.004)	0.033*** (0.002)	0.032*** (0.002)
All cross	Yes	Yes	Yes	Yes	Yes	Yes
Time Dummies	No	Yes	Yes	No	Yes	Yes
Observations	52360	162338	104214	49980	154959	99477
Number of Pages	2380	7379	4737	2380	7379	4737
Adj. R ²	0.003	0.003	0.003	0.002	0.001	0.001

NOTES: The table shows the relationship of clicks/added revisions and time dummies for indirect neighbors of shocked articles (2 clicks away from epicenter) in the large events condition. The reduced form regressions estimate the ITE. Columns (1)-(3) show the results for clicks and Columns (4-6) for new edits to the articles. Specification (1) and (4) show a simple 'before and after'; (2) and (5) contrast treated and comparison group; Columns (3) and (6) show the comparison of treated articles with themselves but seven weeks earlier (placebo treatment). Fixed Effects Panel-Regressions with heteroscedasticity robust standard errors. The unit of observations is the outcome of a page i on day t . The time variable is normalized and runs from -14 to 14.; Only crossterms closer to treatment are shown, but all were included. Reference group t-14 to t-5; standard errors in parentheses: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$; no. of obs. = 323158; no. of clusters = 44; no. of articles = 7379.

For this group the estimation concerns the set of L_2 pages that are two clicks away from the epicenter: the future page about the disaster. This is not because closer pages

⁴³Anticipation effects are impossible for disasters but cannot be entirely ruled out in the “featured articles” condition, where sophisticated users, who can obtain the information about pages that are going to be presented soon. In fact the editors of the daily featured article, have to edit the article in the week before it is advertised, to make sure it fits into the corresponding box on Wikipedia’s main page. This alone results in increased activity during the week before treatment. After carefully studying this process, I am not very concerned about this feature of the data, because the magnitude of the day-0 effect suggests that the vast majority of attention influx is due to readers who do not anticipate which page is to be advertised.

are uninteresting, but because the shock of the analyzed events is very big and likely to directly affect pages that will eventually be directly and bidirectionally linked. If, for example, a city in the province under consideration was hit by the earthquake, the added content on that page might simply cover this very fact. In such a case, this is not an improvement that arose from the increased attention that results from the adjacent event, but a change that is directly caused by the treatment. As explained above, this is not the effect I am primarily interested in. Consequently I focus on pages that were indirectly linked at the time of the shock and that never became directly linked. These articles are no longer likely to be directly affected by the treatment on the page two clicks away.⁴⁴ Moreover, to ensure that my L_2 pages are not directly related to the event I checked whether a page that was in L_2 when I evaluated the network a week before the shock was going to be linked to the page of the disaster later. Since this indicates a potential direct relationship, I eliminated such pages from the sample.

The results of the estimation of the model for L_2 nodes are shown in Table 2.1. The table shows the results for clicks in the first three columns and the results for the number of added revisions in Columns 4, 5 and 6. All the specifications are OLS panel regressions which include a fixed effect for the page and standard errors are clustered on the event level (23 clusters). Column (1) and (4) shows the results of a simple before and after. Columns 2, 3, 5 and 6 show the contrast in the difference-in-differences. Note that I run each regression twice to take advantage of my two comparison groups: first I contrast the treated pages against the control group and then I contrast it with the placebo treatment, i.e. with the treated articles themselves, but simulating a (placebo) treatment 42 days (i.e. 7 weeks) before the real shock.

For ease of representation the table only shows the coefficients for the cross terms from two periods before the shock until four periods after the shock. Until the onset of the event (periods -2 to -1), we would expect insignificant effects for the cross terms and after the event has occurred a positive effect would imply that some form of spillover is present. As in the visual evidence, the average increase in clicks relative to the control group (Column 1) amounts to 35-38.7 more clicks on average. For the placebo treatment (Column 2) this effect is almost equal, but a bit larger from the second day onwards.

Does the spillover in attention also translate to additional content generation? Obviously, this question matters for the relevance of the spillovers I find in this paper. If it does, spillovers of attention have far-reaching implications for other peer production

⁴⁴The results for the L_1 group are included in the appendix. The effects are very large and statistically significant. The estimated coefficients for the L_0 group (not reported) are close to 4,500 for clicks and between 20 and 25 for revisions. However, due to the lack of sufficient observations, even these very large coefficient estimates are not statistically different from zero.

settings.⁴⁵ Generally, the effects are somewhat different for the number of revisions (in line with the graphical analysis), since the effects are much smaller. An effect is consistently revealed from the first day after the treatment. It is small in absolute terms, since roughly one in twenty-five pages gets an additional revision. Yet, given the low levels in average activity on a given page and day, this is still a noteworthy effect. Comparing the pages with the placebo treatment I observe a small increase in editing activity before the onset of the event, which is however not confirmed by comparison with the control group. The size of the effect still more than doubles after day 1, at which point the comparison with the control group suggests a drastic increase in editing activity.⁴⁶

2.5.2 Neighbors of Featured Articles

Table 2.2 shows the results for the “featured articles”. For this reduced form estimation I consider the model for L_1 nodes (Equation 2.10) in Section 2.3.3. This is the relevant group here because the treatment takes place entirely inside Wikipedia⁴⁷ and it is “completely local” since no two articles can be featured simultaneously. Hence, only the treated page is directly affected and any variation in the neighbors is almost certainly a result of the processes that take place inside Wikipedia.

The first three columns of the table show the results with clicks as the dependent variable. The estimation is the same as in Table 2.1 and the clustering is implemented on the level of events as before. The main insight of this table is that it confirms the statistical significance of the effect in the graphical analysis and provides a quantification of its size. The size of the effect is estimated to be 33.1 to 34.6 additional clicks on the average neighbor page on the day of treatment. In Columns 4-6, I observe an important effect of about 0.032 additional revisions one day after the treatment of the neighbor page. Note two things here: First, the effect is very small in absolute terms and corresponds to one additional edit per thirty pages. Second however, this is

⁴⁵If more attention leads to better or more contributions, the importance of link networks for channeling attention would have important implications for open source software, research activities and innovation.

⁴⁶I verified that the result is not driven by running a robustness check, where I exclude four events: the event which was associated to most pages in my dataset, Tunisia and those where the starting date or the most important page of the event was most difficult to identify: the bankruptcy of Lehman, the eruption of Eyjafjallajökull and the plane crash in Smolensk. In this specification, the results are confirmed. The most notable difference is the increased magnitude of the effect in the clicks, as for the remaining events, the average increase is close to 15 additional clicks. Despite the fact, that there are still more than 6,000 pages included in both comparisons, the effects for the number of revisions are no longer significantly different from zero, except in the fourth period of one specification.

⁴⁷Unlike in the disaster case, when an article is advertised on German Wikipedia’s start page this is usually not covered by media or anything of the like.

Table 2.2: 'Featured articles': clicks/added revisions over time for direct neighbors.

	clicks			new edits		
	(1) before after	(2) compare control	(3) compare placebo	(4) before after	(5) compare control	(6) compare placebo
t = -2	-0.709 (2.644)	-5.064 (4.051)	-2.629 (3.477)	-0.010 (0.007)	-0.028** (0.011)	-0.018* (0.011)
t = -1	3.454 (2.668)	2.149 (3.082)	4.792 (4.187)	-0.006 (0.006)	-0.021* (0.012)	-0.004 (0.008)
t = 0	32.888*** (9.073)	33.128*** (9.162)	34.638*** (9.294)	0.004 (0.006)	-0.006 (0.009)	0.004 (0.008)
t = 1	-0.572 (1.799)	-0.158 (2.266)	0.773 (3.214)	0.030** (0.011)	0.032** (0.012)	0.033** (0.014)
t = 2	-4.639* (2.511)	-2.523 (2.965)	-3.700 (3.144)	0.012 (0.008)	0.015* (0.008)	0.017 (0.011)
t = 3	-7.705** (3.114)	-8.373** (3.371)	-3.807 (5.435)	-0.005 (0.008)	-0.011 (0.012)	-0.012 (0.013)
t = 4	-1.225 (2.178)	-2.557 (2.766)	2.038 (5.615)	-0.009 (0.007)	-0.016 (0.014)	-0.009 (0.009)
Constant	34.421*** (1.060)	31.982*** (0.816)	35.354*** (0.768)	0.051*** (0.003)	0.043*** (0.002)	0.046*** (0.002)
All cross	Yes	Yes	Yes	Yes	Yes	Yes
Time Dummies	No	Yes	Yes	No	Yes	Yes
Observations	83424	120758	166518	79632	115269	158949
Number of Pages	3792	5489	7569	3792	5489	7569
Adj. R ²	0.005	0.004	0.003	0.000	0.000	0.000

NOTES: The table shows the relationship of clicks/added revisions and time dummies for direct neighbors of shocked articles in the 'featured articles' condition. The reduced form regressions estimate the ITE. Columns (1)-(3) show the results for clicks and Columns (4-6) for new edits to the articles. Specification (1) and (4) show a simple 'before and after'; (2) and (5) contrast treated and comparison group; Columns (3) and (6) show the comparison of treated articles with themselves but seven weeks earlier (placebo treatment). Fixed Effects Panel-Regressions with heteroscedasticity robust standard errors. The unit of observations is the outcome of a page i on day t . The time variable is normalized and runs from -14 to 14.; Only crossterms closer to treatment are shown, but all were included. Reference group t-14 to t-5; standard errors in parentheses: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$; no. of obs. = 240900; no. of clusters = 63; no. of articles = 5489.

an increase in contribution activity of eighty to one hundred per cent.

Table 2.3 summarizes the results in view of the spillover. The first two columns show the results of the difference-in-differences in views for the treated article in the epicenter. Columns 3 and 4 repeat the before-after estimates from Columns 1 and 4 in Table 2.2 for the change in views and edits of the neighbors. In Column 5 I add a new dependent variable, the change in the number of editors. This serves as a robustness check of whether the shock's effects on edits are some sort of artifact or whether this actually brings in new knowledge. As can be seen in the first two columns, the estimated direct effect of treatment is approximately 4200 views depending on the comparison group. The number of authors, shown in Column 5, experiences a spike paralleling the one for edits (Column 4). Usually less than 1 in 50 articles (on average) is edited by an author, who never edited the article before. During treatment 1 in 30 of the neighbors are edited by a new author (a 72% increase).

Table 2.3: Spillover of views and translation to edits.

	epicenter (L0)		direct neighbors (L1)		
	(1) clicks	(2) clicks	(3) clicks	(4) new edits	(5) new authors
t = -2	-32.686 (59.224)	-25.611 (58.298)	-0.709 (2.644)	-0.010 (0.007)	-0.005** (0.002)
t = -1	36.002 (85.090)	76.750 (78.919)	3.454 (2.668)	-0.006 (0.006)	-0.003 (0.002)
t = 0	4269.573*** (1421.399)	4121.472*** (1422.301)	32.888*** (9.073)	0.004 (0.006)	-0.000 (0.003)
t = 1	206.540* (118.136)	138.306 (132.641)	-0.572 (1.799)	0.030** (0.011)	0.013*** (0.005)
t = 2	58.453 (68.899)	56.972 (68.261)	-4.639* (2.511)	0.012 (0.008)	0.002 (0.003)
t = 3	-31.991 (58.213)	-50.944 (54.872)	-7.705** (3.114)	-0.005 (0.008)	-0.005* (0.002)
t = 4	-7.338 (59.781)	-49.194 (63.384)	-1.225 (2.178)	-0.009 (0.007)	-0.003 (0.003)
Constant	80.235* (41.869)	93.085** (39.295)	34.421*** (1.060)	0.051*** (0.003)	0.018*** (0.001)
All cross	Yes	Yes	Yes	Yes	Yes
Time Dummies	Yes	Yes	No	No	No
Observations	1474	1584	83424	79632	79632
Number of Pages	67	72	3792	3792	3792
Adj. R ²	0.182	0.182	0.005	0.000	0.000

NOTES: The table summarizes the results of the reduced form regressions to estimate the spillovers in clicks and edits. Columns (1)-(2) show the results for the direct effect of treatment on clicks (ATE) and Columns (3) the spillover to direct neighbors of the articles (ITE). and Columns (4) show the conversion of the spillover in clicks to new edits at the direct neighbors and Column (5) shows the number of new author's that contributed to the articles (at the neighbors). All Specifications are a simple before-after comparison. Fixed Effects Panel-Regressions with heteroscedasticity robust standard errors. The unit of observations is the outcome of a page i on day t . The time variable is normalized and runs from -14 to 14.; Only crossterms closer to treatment are shown, but all were included. Reference group t-14 to t-5; standard errors in parentheses: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$;

I further test the robustness of my difference-in-differences results by excluding the first third of the “featured articles”. Results are shown in Table 2.10 and reveal the same patterns as Table 2.1, but at lower significance levels.⁴⁸ The number of authors moves largely in parallel with the number of revisions, indicating that twice as many new authors as usual edit the article due to the treatment of its neighbor. On the one hand this is a large effect in relative terms, on the other hand it means that only one in seventy articles is edited by a new author. More robustness checks included regressing against all comparison groups simultaneously and using different samples

⁴⁸The coefficients are not significant for the placebo-condition, but note that I apply extremely rigorous clustering and further note that all surrounding point estimates are negative, while the one at $t = 1$ is positive.

or resolutions. They do not convey additional insight and are available in the online appendix.⁴⁹

2.5.3 Bounds for the Structural Estimator

Unfortunately I cannot compute the precise structural estimator because the full matrix \mathbf{G} formed by the German Wikipedia is too large to be computed in memory. Hence I cannot solve for \mathbf{G}^2 and higher orders of the link matrix.⁵⁰ However, it is possible to present upper and lower bound estimates of the structural parameters that are discussed in Subsection 2.3.4 and derived formally in Appendix A.2.

To compute these values the researcher has to decide where to evaluate the number of peers. I choose to evaluate the coefficients at the median which is 31 for indirect neighbors of disaster pages and 36 for neighbors of “featured articles”. This is a crude first evaluation which primarily serves to highlight how easy it is to retrieve the structural parameters once this decision is made. The rest reduces to a back of the envelope calculation for the upper bound of the social/spillover parameter α and the shock δ_1 . I use Equation A.22: $\overline{\delta_1}$ is directly estimated to be 4,190 (2,440 in the disaster condition). The estimate of $\overline{\alpha}$ is 0.292 based on “featured articles” (and based on disasters, 0.483).⁵¹

Computing the lower bound estimates is not much more involved: It suffices to plug the estimates and the number of nodes into the closed form solution given in Equation A.28. This gives the point estimator for the lower bound of α , which is estimated to be $\hat{\underline{\alpha}} = 0.222$ for “featured articles” and $\hat{\underline{\alpha}} = 0.320$ for disasters.

To conclude this section I attempt to quantify the meaning of these results: literally they mean that if the average clicks on the neighboring pages are increased by ten, this alone would result in an increase of 2.215 to 2.92 clicks on the page, which all come from the neighbors. Even though caution is needed to make the following claim, the results suggest that placing links has an effect, but that it is small. Provided this out of equilibrium thought experiment is warranted, creating additional links from neighbors that increase aggregate viewership of the neighbors by 200 is predicted to result in 1.61 additional views on the target page.⁵² While this absolute effect in clicks

⁴⁹<https://sites.google.com/site/kummersworkingpapers/spilloversonlineappendix>

⁵⁰Ongoing work is attempting to solve this issue. If these efforts are fruitful, the results might be included in a revised version of this paper.

⁵¹I briefly illustrate how simple this computation is: merely divide the estimated effect on the neighbors (34) by the estimated effect on the treated (4,190) and multiply by the median in-degrees (36). For disasters the analogous computation is $(38/2,440 \cdot N=31)$.

⁵²As before I use the median number of neighbors for these thought experiments. Consequently 200 aggregate view correspond to five more views on average. The quantification is based on the upper bound estimates of α in the “featured article” condition (and would be 3.31 for disasters).

is very small, the conversion to content is even smaller than that since even huge shocks did not generate many revisions on neighboring articles. This suggests that placing links strategically will only generate large effects if the pages that link out are very frequented. However, for the normal traffic on a typical Wikipedia page we would expect very small effects.

2.5.4 Aggregate Effects and Heterogeneities in the Spillover.

In this subsection I first discuss the aggregate spillover effects. Second, I offer the results of a first analysis of how article and network characteristics mediate the spillovers of attention and the associated conversion into content generation. This serves as a test of the assumption that attention spillovers are homogeneous, as is assumed in my model. Moreover, it is interesting for understanding the factors that mediate how attention spills across links and also how attention is converted into content.

First I aggregate the changes in clicks and revisions over all neighboring articles and then average over the 34 different “featured article” clusters. This is done in Figures 2.5 and 2.6 in order to summarize and illustrate the insights from the “featured articles” condition. I find that on average there are 4000 clicks on all neighbors taken together (Figure 2.5). Given that the average treated articles received an additional 4000 clicks this corresponds to a one to one conversion of clicks on the treated page to clicks on one of the neighbors. In other words, the average visitor clicks on exactly one of the links. The total number of revisions on the neighboring pages (Figure 2.6) increases approximately from 4.5 to 8.5. This means that the 4000 initial additional clicks are converted into 4000 additional clicks on neighbors and four new revisions or a ratio of 1000:1000:1. On the level of the individual article, where usually one in 30 gets an edit, it is still only one in 15 on the day of treatment. Note that these findings are in line with the average “facebook-engagement rate”, which is typically just below 0.01.⁵³

Second, I analyze how article characteristics influence the spillovers. I add additional control variables that account for differences in the articles’ characteristics. The results of this analysis are shown in the first three columns of Table 2.4, where I added variables that account for an article’s length, how well it is generally linked and how closely they are linked to the shocked articles (by counting closed triads with other neighbors). Columns 4-6 show the results of an analysis that considers short articles (“stubs”) separately. A word of caution is in place here: The explanatory variables at hand are subject to many unobserved influences, such as relevance, challenging topics, etc. Hence, these regressions might introduce endogeneity problems. Nevertheless these

⁵³<http://www.michaelleander.me/blog/facebook-engagement-rate-benchmark/>. The benchmark measures, how many of a user’s friends and followers react to their posts.

Table 2.4: Clicks/added revisions over time, when including article heterogeneity.

	joint estimation			short articles only		
	(1) clicks	(2) revisions	(3) length	(4) clicks	(5) revisions	(6) length
t = -2	-5.064 (4.051)	-0.028** (0.011)	-2.484 (3.012)	-2.241 (1.564)	0.004 (0.006)	-0.105 (0.202)
t = -1	2.177 (3.278)	-0.022* (0.012)	7.659 (6.390)	5.178* (3.014)	0.006 (0.005)	0.204 (0.283)
t = 0	38.770*** (11.776)	-0.017* (0.009)	0.419 (2.756)	35.748*** (12.929)	0.002 (0.007)	0.009 (0.197)
t = 1	-0.675 (2.781)	0.041** (0.018)	1.972 (1.681)	1.559 (2.571)	0.017* (0.010)	1.151 (1.210)
t = 2	-2.523 (2.965)	0.015* (0.008)	6.698* (3.957)	-0.134 (2.136)	0.003 (0.002)	-0.089 (0.198)
t = 3	-8.373** (3.371)	-0.011 (0.012)	-2.057 (1.370)	-0.813 (2.258)	0.005 (0.009)	-0.744 (0.541)
t = 4	-2.557 (2.766)	-0.016 (0.014)	-2.308 (1.722)	7.067 (4.435)	-0.005 (0.007)	0.927 (1.027)
short article on t = 0	-2.631 (5.745)	0.003 (0.008)	-1.386 (2.343)			
short article on t = 1	-0.962 (2.726)	-0.037** (0.017)	-0.159 (1.029)			
many L1 links article on t = 0	24.027 (19.759)	0.009 (0.025)	19.834 (18.849)			
many L1 links article on t = 1	-6.963 (7.463)	-0.020 (0.024)	-3.385 (2.825)			
long article on t = 0	-8.770 (10.807)	-0.001 (0.016)	-4.732 (3.983)			
long article on t = 1	3.164 (4.333)	-0.000 (0.029)	10.428 (6.505)			
well linked article on t = 0	-15.565 (14.763)	0.030** (0.014)	2.261 (5.132)			
well linked article on t = 1	1.484 (4.648)	-0.003 (0.023)	-3.044 (4.871)			
Constant	31.982*** (0.823)	0.043*** (0.002)	2.959*** (0.524)	10.297*** (0.994)	0.003*** (0.001)	0.137* (0.071)
All cross	Yes	Yes	Yes	Yes	Yes	Yes
Time Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	120758	115269	115269	19932	19026	19026
Number of Pages	5489	5489	5489	906	906	906

NOTES: The table shows the results of the reduced form regressions estimating the ITE for neighbors. Columns (1)-(3) show additional control variables that provide a deeper characterization of the articles to analyze if spillovers vary across these groups. Columns (4-6) analyze the reduced sample that only contains articles that were shorter than 1500 bytes ("stubs"). Specification (1) and (4) show the results for clicks, (2) and (5) for new edits, and Columns (3) and (6) show added content. All estimations contrast the treated with both comparison groups (based on other featured articles and the same articles but seven weeks earlier (placebo treatment)). Fixed Effects Panel-Regressions with heteroscedasticity robust standard errors. The unit of observations is the outcome of a page i on day t . The time variable is normalized and runs from -14 to 14.; Only crossterms closer to treatment are shown, but all were included. Reference group t-14 to t-5; standard errors in parentheses: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$; no. of obs. = 120758; no. of clusters = 63; no. of articles = 5489.

Table 2.5: Conversion of attention to action: Content vs. money contributions.

	donations		contributions	
	donations high	donations low	all articles	short articles
click-through	0.024	0.002	0.008	0.008
conversion to action	0.124	0.004	0.001	0.0005

NOTES: As the click-through rate for donations I used the “Banner-Klick pro Impressions”-statistic, which measures the ratio of how often the banner was shown, vs. how often it was clicked on. As analogue to the conversion to an edit (action) I used “Anzahl pro Banner Klick”, which measures the ratio of actually completed donations to banner-clicks. I report the minimum (25% quartile attention: 0.004, action: 0.01) and the maximum (75% quartile attention: 0.009, action: 0.019) of the 2011 campaign. Source: 2011 Wikipedia campaign tests and own calculations.

questions deserve to be studied further, so I report the results with the caveat that they are correlations that they cannot necessarily afford a causal interpretation.

Splitting the sample into well-connected articles with many links and poorly connected ones with few, I do not find a significant difference between the two groups. The same is true for a variable that captures whether a page is very long or not. I get a positive but insignificant point estimate for page views. By the same token, there is no statistically significant effect for nodes that show high clustering relative to the shocked node (neighbors, that get many links also from other neighbors). The point estimate seems to indicate a higher click-through rate onto those articles, but these clicks might be indirect click-throughs (forwarded from neighbors of the shocked node). Generally, no effect on clicks in Column 1 is statistically significant which indicates that the spillovers do not vary systematically by the four mentioned properties.

An interesting pattern emerges, when I consider only “stubs”, i.e. pages that do not exceed a length of 1500 bytes. I find that the increase is the same as on average, but that attention is less likely to convert into edits (Columns 4-6 of Table 2.4). Short articles are viewed 32 times more often than on an average day, but in absolute terms the increase in edits is smaller (.017 vs. .034, in separate estimation). A remark is in order: Since stubs generally have a much lower probability to be viewed or edited, these increases are *larger in relative terms* than for normal articles.⁵⁴ Finally I report results of an extended analysis which is omitted here, for reasons of brevity.⁵⁵ I include the number of clicks on the treated page in the regression and, as expected, the number of links on the neighboring pages is positively related to that value.

It is exciting to contrast my findings for the contribution of content with the

⁵⁴The probability of being viewed increases by 300% (vis a vis 100%) for the average neighbor. The chances for an edit increases by 500% vis a vis an 80% increase over baseline levels. This much stronger relative effect in the number of edits indicates that contributors do make a greater effort (in comparative terms) to contribute to pages where the existing content is limited.

⁵⁵Available online at <https://sites.google.com/site/kummersworkingpapers/spilloversonlineappendix>.

donations of money during Wikipedia’s annual campaign for donations.⁵⁶ I do so in Table 2.5, which shows my main results for content contributions and contrasts them with the click-through rates during the campaign. When doing so I find a striking similarity in the size of the click-through rates between content items and the click-through to the donation banners during the campaign. In other words, it turns out that the spillover of attention is almost the same between articles and from articles to the donations-banner. However, the conversion into action, once on the donating page, is even higher for monetary donations than for making an edit on the new page. Even though the intention behind clicking on a banner is certainly different from the one behind clicking on any link, this fact gives an idea of just how costly it is for most visitors to make an edit.

2.6 Conclusions, Limitations and Further Research

This paper investigates how the network of links between articles on the German Wikipedia influences the attention and subsequent content generation individual articles receive. To the best of my knowledge my results are the first to provide a causal quantification of how the link network influences users’ contributions in an important online network of content pages. By studying exogenous short term shocks to attention I can measure attention spillovers. I find substantial spillovers in terms of both views and editing activity. Articles in the neighborhood of shocked articles received 35 more visits on average - an increase of almost 100 percent.

Moreover, I am able to isolate the effect of attention from other determinants of public good contribution such as reputation, social image and altruism. I find that on average 1000 views are needed before a Wikipedia edit occurs, before additional content is generated. I also find that this rate is lower than the conversion rate from clicks on funding campaign banners to donations, indicating that it is less costly to donate money than to contribute content. Taken together, my findings suggest that (i) making an edit is very costly and (ii) that contributions due to attention alone will ensure sufficient provision of public goods only, whenever large amounts of traffic are available, which can be the case online. Offline or with platforms that lack many visitors, other important drivers of public goods contributions, such as social image and altruism may be needed. (cf. Carpenter and Myers (2010))

Even though my design allows a causal interpretation of the reduced form estimates of the spillover, I incorporate exogenous treatments of individual nodes in networks into a workhorse model to formalize peer effects (or spillovers) in networks

⁵⁶ Available at http://de.wikipedia.org/wiki/Wikipedia:Fundraiser_2011/Tests (last: Feb. 27, 2014)

(Bramoullé et al. (2009), De Giorgi et al. (2010)). Exogenous treatments can serve as a new and complementary source of identification of the structural spillover effect, which does not depend on assuming an exogenous network structure. The model I suggest is quite general and nests also two-layered randomized control settings that rely on exogenous variation over subpopulations.⁵⁷

My structural estimates suggest that an article will receive 30 percent of the number of average views on neighboring articles. Hence, by placing links to oft frequented nodes and thus increasing the average daily views on their neighbors by ten, one could obtain three additional daily visits to an article. I show that upper and lower bounds for the structural parameters can be computed even if the underlying network structure is unknown. The bounds are easily computed for settings where only one node is treated in each subpopulation. I conjecture that they can easily be generalized to treatments that affect more than one node. It is thus no longer necessary to neglect the network structure in an experiment that aims at identifying social effects, merely because the information on links is not available. Finally, the basic modeling approach of exploiting open triads in the network structure is formally similar to spatial Durbin models in spatial econometrics (cf. LeSage (2008)).

So what do we learn for advertising on the web, setting up a firm wiki or for realizing the Wikimedia Foundation’s vision?⁵⁸ Additional views translate one for one into additional views on a neighbor. The significance of this result deserves emphasis: On average, *every* visitor of “today’s featured article” clicks on one of the links to acquire further information. The click-through rates I find are very similar to the clicks-throughs to the donation banner. Moreover, the spillover does not depend on the targets’ characteristics - all contents have a fair chance of getting some attention. However, what happens once the attention is there *does* depend on the items’ characteristics - much less content is generated *shorter* pages.

My findings highlight the importance of citation networks for channeling human attention, but suggest that using the link network is an expensive and inefficient strategy for channeling contribution flows. My results also indicate that many users only look up information. I cannot say whether the low levels of conversion from attention to an action are the same for Wikipedia (editing) and adverts (purchase), but the similarities of the click-through rates to donations are striking. Further research could study if similar results apply to and to young wikis with less content.

There are some limitations to the presented approach. Most importantly, the strategy of exploiting local exogenous treatments will not allow the identification of

⁵⁷Moffitt (2001), Angelucci and De Giorgi (2009), Kuhn et al. (2011), Crépon et al. (2013) etc.

⁵⁸A world where all “can freely share in the sum of all knowledge” (Wikimedia-Foundation (2013))

the social spillover parameter if neighbors of the treated nodes observe the treatment and adjust their outcome as a reaction to the mere fact that their neighbor was treated. An example would be a teacher who selectively punishes or favors a single student: if other pupils react to the special treatment, e.g. by changing their motivation to study for the subject, then their performance change reflects the sum of the spillover and their behavioral adjustment. In Appendix A.3, I outline such a case and illustrate formally why the spillover parameter can no longer be identified. Another limitation is the assumption that the network formation *process* is not affected by the treatment. This assumption is warranted for Wikipedia’s “today’s featured article” but less so for disasters. Generally, if the process is affected by treatment all estimates of indirect treatment effects will reflect a sum of the treatment on the existing network and new spillovers due to the changes in the link network which might lead to upward biases (cf. Comola and Prina (2013)).

A promising area for further analysis would investigate whether the new contributions, especially the ones by new authors, add substantive knowledge or rather focus on improving small details. Future research should also exploit the heterogeneity in intensity of direct treatment effects more thoroughly. In particular, it would be interesting to analyze how attention, here measured as average effect, is distributed across neighbors. Is it evenly distributed or do users herd to only a few of the linked pages? Another promising area would use the methodology based on exogenous local treatments alongside that based on the network structure and the exploitation of open triads (Bramoullé et al. (2009), De Giorgi et al. (2010)). The approaches are complementary; research along these lines will result in valuable insights. Finally, it was not yet possible to surmount the computational hurdle of exploiting the detailed network information when obtaining the structural estimates. Future research should include this information and investigate which population parameter should be optimally included for relating reduced form and structural parameters.

2.7 Descriptive Statistics

Table 2.6: Summary statistics: direct neighbors of shocked 'featured articles'.

	mean	sd	min	p10	p50	p90	max
Length of page (in bytes)	6794	6784	17	51	4833	15262	81585
Number of authors	33	35	1	2	21	77	324
Clicks	33	131	0	0	0	77	20384
Number of Revisions	95	130	1	3	48	237	1382
Links from Wikipedia	118	301	0	6	36	286	9484
Dummy: literature section	.3	.46	0	0	0	1	1
References (footnotes)	1.3	4.5	0	0	0	4	182
Links to further info	2.3	4.2	0	0	1	6	155
time variable (normalized)	0	8.4	-14	-12	0	12	14
Delta: Number of Revisions	.042	.39	0	0	0	0	42
Delta: Length of page	2.1	159	-31473	0	0	0	31462
Delta: Number of authors	.015	.13	0	0	0	0	9
Delta: Links from Wikipedia	.054	1.1	-90	0	0	0	438
Delta: References	.0014	.097	-18	0	0	0	18
Delta: Links further info	.00078	.1	-19	0	0	0	16

NOTES: The table shows the distribution of the main variables. The unit of observations is the outcome of a page i on day t . The time variable is normalized and runs from -14 to 14.; no. of obs. = 317550; no. of start pages = 63; no. of articles = 5489.

Table 2.7: Indirect neighbors of shocked 'large events articles' (2 clicks away)

	mean	sd	min	p10	p50	p90	max
Length of page (in bytes)	5658	6287	16	33	3885	13210	76176
Number of authors	29	34	1	1	18	71	435
Clicks	33	174	0	0	0	70	29865
Number of Revisions	84	133	1	2	40	211	2083
Links from Wikipedia	123	447	0	5	31	269	27611
Dummy: literature section	.2	.4	0	0	0	1	1
References (footnotes)	1.3	4.2	0	0	0	4	150
Links to further info	2.7	5.1	0	0	1	7	130
time variable (normalized)	0	8.4	-14	-12	0	12	14
Delta: Number of Revisions	.035	.35	0	0	0	0	44
Delta: Length of page	1.8	106	-22416	0	0	0	27500
Delta: Number of authors	.013	.12	0	0	0	0	11
Delta: Links from Wikipedia	.049	2.5	-1148	0	0	0	216
Delta: References	.0014	.13	-32	0	0	0	29
Delta: Links further info	.0011	.12	-15	0	0	0	31

NOTES: The table shows the distribution of the main variables. The unit of observations is the outcome of a page i on day t . The time variable is normalized and runs from -14 to 14.; no. of obs. = 425981; no. of start pages = 44; no. of articles = 7379.

Table 2.8: Included “featured articles” and associated articles (1 clicks away).

name of event	pages	observations		
	no.	control	treated	Total
Afrikaans	128.0	5,481.0	1,943.0	7,424.0
Alte_Synagoge_(Heilbronn)	52.0	1,885.0	1,131.0	3,016.0
Banjo-Kazooie	125.0	5,191.0	2,030.0	7,221.0
Benno_Elkan	139.0	5,133.0	2,900.0	8,033.0
Bombardier_Canadair_Regional_Jet	92.0	4,205.0	1,073.0	5,278.0
CCD-Sensor	586.0	31,001.0	2,871.0	33,872.0
Charles_Sanders_Peirce	258.0	11,716.0	3,219.0	14,935.0
Das_Kloster_der_Minne	51.0	1,827.0	1,102.0	2,929.0
Deutsche_Bank	343.0	10,005.0	9,860.0	19,865.0
Eishockey	162.0	4,698.0	4,698.0	9,396.0
Ekel	270.0	10,295.0	5,336.0	15,631.0
Fahrbahnmarkierung	44.0	1,276.0	1,276.0	2,552.0
Geschichte_Ostfrieslands	235.0	7,453.0	6,177.0	13,630.0
Geschichte_der_deutschen_Sozialdemokratie	306.0	9,599.0	8,033.0	17,632.0
Glanzstoff_Austria	270.0	14,094.0	1,537.0	15,631.0
Glorious_Revolution	153.0	6,206.0	2,668.0	8,874.0
Granitschale_im_Lustgarten	83.0	3,857.0	928.0	4,785.0
Gustav_Hirschfeld	142.0	6,438.0	1,740.0	8,178.0
Hallenhaus	71.0	2,117.0	2,001.0	4,118.0
Helgoland	228.0	8,120.0	5,104.0	13,224.0
Jaroslavl	321.0	12,789.0	5,829.0	18,618.0
Jupiter_und_Antiope_(Watteau)	36.0	1,160.0	928.0	2,088.0
Karolingische_Buchmalerei	162.0	4,843.0	4,553.0	9,396.0
Katholische_Liga_(1538)	37.0	1,682.0	464.0	2,146.0
Martha_Goldberg	55.0	1,595.0	1,595.0	3,190.0
Naturstoffe	320.0	9,338.0	9,222.0	18,560.0
Paul_Moder	61.0	1,798.0	1,682.0	3,480.0
St._Martin_(Memmingen)	59.0	1,653.0	1,711.0	3,364.0
Stabkirche_Borgund	40.0	1,421.0	899.0	2,320.0
Taiwan	167.0	5,017.0	4,669.0	9,686.0
USS_Thresher_(SSN-593)	90.0	3,712.0	1,479.0	5,191.0
Visum	56.0	1,624.0	1,624.0	3,248.0
Wenegnebt	55.0	1,798.0	1,363.0	3,161.0
Werder_Bremen	292.0	8,555.0	8,323.0	16,878.0
Total	5,489.0	207,582.0	109,968.0	317,550.0

NOTES: The table shows the “featured articles” in the dataset that were advertised on German Wikipedia’s start page. Column 1 shows the number of associated articles that are one click away from one of the corresponding start pages (be it treated or control). Columns 2-4 show the number of observations. Observations associated with actually treated articles are shown separately from control observations. Pages can be accessed by pasting the title behind the last slash in: <http://de.wikipedia.org/wiki/>

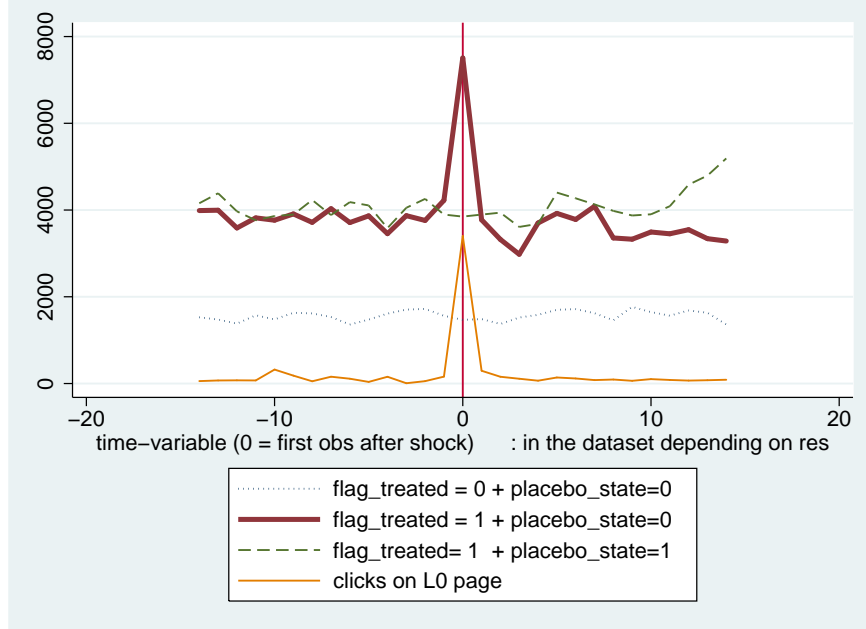
Table 2.9: Included disasters and associated articles (2 clicks away).

name of event	pages	observations		
	No.	control	treated	Total
Air-France-Flug_447	102.0	4,495.0	1,392.0	5,887.0
Air-India-Express-Flug_812	369.0	19,662.0	1,711.0	21,373.0
Amoklauf_von_Winnenden	74.0	2,088.0	2,146.0	4,234.0
Bahnunfall_von_Halle_(Belgien)	52.0	2,436.0	580.0	3,016.0
British-Airways-Flug_38	144.0	6,699.0	1,624.0	8,323.0
Buschfeuer_in_Victoria_2009	33.0	928.0	957.0	1,885.0
Deepwater_Horizon	203.0	8,178.0	3,596.0	11,774.0
Erdbeben_in_Haiti_2010	379.0	15,602.0	6,322.0	21,924.0
Erdbeben_in_Sichuan_2008	227.0	11,571.0	1,508.0	13,079.0
Erdbeben_von_L'Aquila_2009	96.0	3,654.0	1,885.0	5,539.0
Flugzeugabsturz_bei_Smolensk	368.0	12,412.0	8,758.0	21,170.0
Grubenunglück_von_San_Jose	149.0	8,033.0	551.0	8,584.0
Josef_Fritzl	129.0	6,264.0	1,044.0	7,308.0
Kaukasuskrieg_2008	346.0	18,705.0	1,276.0	19,981.0
Kolontár-Dammbruch	99.0	4,669.0	1,073.0	5,742.0
Luftangriff_bei_Kunduz	2,107.0	113,767.0	7,772.0	121,539.0
Northwest-Airlines-Flug_253	1,151.0	65,279.0	1,276.0	66,555.0
Sumatra-Erdbeben_vom_September_2009	116.0	4,002.0	2,726.0	6,728.0
US-Airways-Flug_1549	226.0	7,888.0	5,220.0	13,108.0
Unglück_bei_der_Loveparade_2010	499.0	15,283.0	13,572.0	28,855.0
Versuchter-Anschlag_am_Times_Square	202.0	10,353.0	1,334.0	11,687.0
Wald-_und_Torfbrände_in_Russland_2010	273.0	13,485.0	2,204.0	15,689.0
Zugunglück_von_Castelldefels	35.0	1,508.0	493.0	2,001.0
Total	7,379.0	356,961.0	69,020.0	425,981.0

NOTES: The table shows the events in the dataset. Column 1 shows the number of pages that are two clicks away from one of the two associated start pages (be it treated or control). Columns 2-4 show the number observations associated with the articles. Observations associated with actually treated articles are shown separately from control observations. Pages can be accessed by pasting the title behind the last slash in: <http://de.wikipedia.org/wiki/>

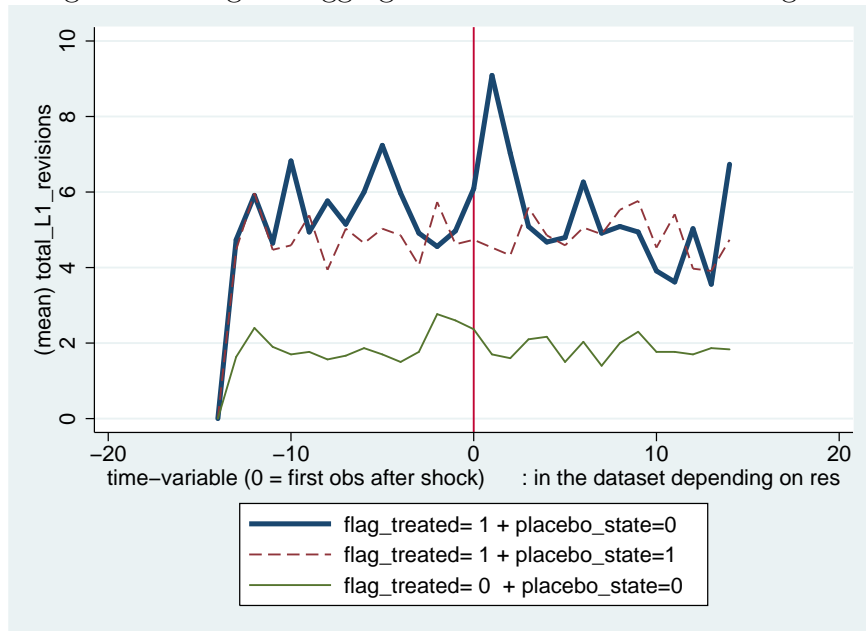
2.8 Additional Regression and Figures

Figure 2.5: Figure contrasting the mean of clicks on featured articles, with the aggregated clicks on all neighboring pages.



NOTES: The figure shows the aggregated effect on the pages that are one click away. The average treated page received up to 4000 additional clicks, all neighbors together received approx. the same number of additional clicks

Figure 2.6: Figure showing the aggregated new revisions on all neighboring pages.



NOTES: The figure shows the aggregated effect on the pages that are one click away. All neighbors of treated articles together received approx. four additional revisions.

Table 2.10: Robustness check: Reduced number of events.

	clicks		del revisions		del authors	
	(1)	(2)	(3)	(4)	(5)	(6)
	compare control	compare placebo	compare control	compare placebo	compare control	compare placebo
t = -2	-2.117 (5.554)	4.304 (4.147)	-0.026** (0.012)	-0.020 (0.015)	-0.014** (0.006)	-0.009* (0.005)
t = -1	2.953 (4.448)	11.074* (5.974)	-0.012 (0.013)	-0.004 (0.011)	-0.002 (0.005)	-0.005 (0.004)
t = 0	34.625** (13.296)	40.149*** (13.572)	-0.013 (0.011)	0.004 (0.011)	-0.006 (0.005)	-0.008 (0.005)
t = 1	-1.463 (2.685)	2.145 (4.649)	0.037* (0.018)	0.033 (0.021)	0.015** (0.007)	0.012 (0.008)
t = 2	-3.262 (4.308)	-0.427 (4.076)	0.012 (0.010)	0.005 (0.015)	0.002 (0.004)	-0.006 (0.006)
t = 3	-10.195** (4.874)	-3.046 (4.977)	-0.009 (0.012)	-0.033* (0.019)	-0.005 (0.006)	-0.010* (0.005)
t = 4	-3.023 (4.074)	5.034 (3.968)	-0.031* (0.016)	-0.019 (0.012)	-0.009 (0.005)	-0.003 (0.005)
Constant	30.930*** (1.291)	34.526*** (1.135)	0.047*** (0.003)	0.050*** (0.003)	0.016*** (0.001)	0.016*** (0.001)
All cross	Yes	Yes	Yes	Yes	Yes	Yes
Time Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	73084	98252	69762	93786	69762	93786
Number of Pages	3322	4466	3322	4466	3322	4466
Adj. R ²	0.004	0.003	0.000	0.000	0.001	0.001

NOTES: The table presents a robustness check. It shows the relationship of clicks/added revisions and time dummies for direct neighbors of shocked articles in the 'featured articles' condition for only a reduced number of events. Standard errors in parentheses. Fixed Effects Panel-Regressions with heteroscedasticity robust standard errors. Only crossterms closer to treatment are shown, but all were included. Reference group t-14 to t-5. * p<0.10, ** p<0.05, *** p<0.01

Chapter 3

Centrality and Content Creation in Networks^{*}

- The Case of German Wikipedia

3.1 Introduction

User-generated content has proven to be a cheap and surprisingly accurate source of information. Still, little is known about how its producers select the content to which they contribute and how platform administrators may influence this choice. While Wikipedia has been the most successful prototype of a wiki, wikis in other contexts, e.g. private businesses, often struggle to encourage and manage activity. Administrators of platforms face three challenges: motivating potential first-time users, making them connect to the platform and encouraging the contribution of content that is useful to others (Lerner and Tirole (2002), Jian and MacKie-Mason (2012)).

In order to encourage contributions, it is important to understand how authors select articles. In this paper, we study one mechanism that possibly channels their activity. We start from the hypothesis that the hyperlink network between Wikipedia articles attracts the attention of authors towards more central articles. In particular, we analyze how the position of an article in the network is related to the amount of content contributed and to the number of new authors joining the article. This question is situated in the more general context of understanding how producers in peer production of information goods select their tasks.

^{*}This Paper is coauthored with Marianne Saam (Centre for European Economic Research (ZEW)), Iassen Halatchliyski (Knowledge Media Research Center (IWM-KMRC)) and George Giorgidze (University of Tübingen). We thank the Wikimedia Foundation for granting access to the Wikipedia data, Thorsten Doherr and Manfred Knobloch for support with the data processing, and Frédéric Schütz for providing us with the data on page views. We benefitted from discussions with Irene Bertschek, Ulrike Cress, Benjamin Engelstätter, Avi Goldfarb, Francois Laisney, Jose Luis Moraga-Gonzalez, Martin Peitz, Philipp Schmidt-Dengler, Michael Ward, the participants of the ICT Conference 2012 at ZEW in Mannheim and of the annual conference of the EARIE 2012 in Rome. Benedikt Achatz, Sergiy Golovin, Burak Tuerkoglu and Fabian Trottner provided helpful research assistance. We acknowledge financial support from the WissenschaftsCampus Tübingen.

On Wikipedia, there are three main possibilities for finding articles of interest: categories, text search and hyperlinks. Frequent authors use additional devices such as lists of new articles, watchlists or lists of articles classified as needing improvement. Hyperlinks constitute an organizing principle that is indispensable to online peer production of a vast amount of information. They enable a non-hierarchical access and a nonlinear reading experience that are characteristic for wikis (Greenstein and Devereux (2009)).

Meanwhile little research has been undertaken on the question how hyperlinks influence contributions in wikis. Wikipedia's rules determine hyperlinks between articles to be semantic links, that means links that are set according to important connections in the attributes of the two subjects. The links need not be reciprocal and the guidelines on the German Wikipedia stipulate that an article must be readable without the information from the linked pages. It is not compatible with Wikipedia's rules to set links just to attract attention to an article or without embedding its subject into the text pointing to it. Finally, within Wikipedia, links should point only to pages about technical terms or to pages that contain further information on topics that might be of particular interest to readers of the originating article.² Hyperlinks on Wikipedia are generally regarded as a reliable source of information on semantic relations between words. They have been used extensively in linguistic research (see for example Medelyan et al. (2009)). Adafre and de Rijkje (2005) propose a procedure that automatically detects missing links between pages that should be linked given their relevance to each other. Taken together, this research suggests that hyperlinks on Wikipedia are generally set in accordance with the guidelines (see also Priedhorsky et al. (2007) on rapid detection of vandalism), but that the topics of articles on Wikipedia do not completely predetermine their link structure. The actual links depend on the dynamic content of an article and on the accuracy of linking. This implies that variations in centrality occur regularly and affect the navigation of readers and potential authors on a given set of articles. Our main hypotheses are that higher centrality is positively related to (i) the length of an article's content and (ii) the number of new authors joining the article.

Economic research considers spillovers to be a central feature of knowledge production. They arise when the production of new knowledge relies on existing knowledge, which can be used without paying for it and without diminishing anyone else's use of it (see for example Romer (1990) in the context of growth theory). Studies on R&D have highlighted that the strength of spillovers depends on the distance between the knowledge that is available and the knowledge that is being produced.

²<http://de.wikipedia.org/wiki/Wikipedia:Verlinken>, accessed on July 23, 2012.

This distance may be defined in various ways, for example geographically or according to sectors of economic activity (Griliches (1992), Audretsch and Feldman (1996)).

In the context of Wikipedia, we also consider spillovers occurring in the production of knowledge. The channel of spillovers that we are analyzing consists in the hyperlinks pointing from one article to another. However, we are not looking for knowledge spillovers in the classical sense, but for spillovers in the level of production activity. On Wikipedia, this approach is based on the hypothesis that links placed on page A pointing to page B may attract the attention to page B. Consequently, the existence of an additional link may trigger the contribution of authors who might not have contributed in its absence. These spillovers affect the level of content provision on a page and also increase the knowledge contributed to the page. Note however, that the dimension to which the notion of spillover applies in our context is not the knowledge itself but the attention and effort that authors direct to a particular page after they read another one pointing to it.

When analyzing the relation between centrality and content provision, we exploit different dimensions of proximity that exist between articles. In particular, we compare the links from articles that are semantically close to links which are on average less close. We also compare direct links to an article, measured by the number of incoming links (the indegree), to indirect links, measured by the closeness centrality. We chose a sample of more than 7,000 articles belonging to a particular category (economics; German: "Wirtschaft"). For this sample, we compute centrality measures both within the category and on the entire German Wikipedia.

We find that an increase in the number of links from within the category is strongly associated with an increase in page length. It is also associated with new authors contributing to the article. The strongest relation between centrality and content generation is found for direct links from the category network. The relation to links from other pages of German Wikipedia is weaker and insignificant in our main specification. The additional influence of indirect links appears negligible. Social network analysis reveals that the category economics is, like many networks, constituted by one large cluster and single articles or small network components that are disconnected from it. We find that getting connected to the large component raises the page length and its rate of change sizeably in the following weeks.

Our research is inspired by two strands of work on user-generated content: First, we are interested in knowing whether evidence generated inside a limited category of a network (e.g., Kittur and Kraut (2008), Ransbotham et al. (2012)) holds when taking into account links outside the category. Second, we follow work by Fershtman and Gandal (2011) and Claussen et al. (2012) on direct and indirect spillovers in networks of software producers. Contrary to these strands of work, we do not consider the

network of authors but the hyperlink network of user-generated content.

3.2 Related Research

The empirical analysis of (social) networks has been of interest to scientists of different disciplines for several decades, resulting in a vast literature and in an established methodology based on the analysis of graphs. This tool has been widely used in empirical applications that are relevant to economics, so that we are forced to restrict ourselves to discussing only large overarching themes.³ Some studies center around the existence and the structure of social networks, applying a variety of formally defined network measures. Other applications have analyzed the prevalence of homophily in networks, the importance of weak ties and social capital (for example in job-market outcomes), or the benefits associated with filling structural holes in networks.

Social networks have since then been at the heart of a variety of theoretical and empirical studies in economics. Diffusion in networks was originally studied in medicine and biology, but the methods can also be used in economics to study technology adoption or viral marketing. Moreover, economists became interested in citation networks. Goyal et al. (2006) analyze the evolution of the collaboration network of economists from the 1970s until the 1990s. They find that a structure of separated 'small islands' of researchers is increasingly replaced by a 'small world' network where every pair of nodes (authors) is connected by a short path. In fact, citation networks of scientific papers had been analyzed as early as the 1960s.⁴ More recently, Albert et al. (1999) have undertaken a similar endeavor for web pages.

Particularly relevant to this paper are studies focussing on spillovers in production through social networks. Fershtman and Gandal (2011) analyze direct and indirect knowledge spillovers in the production of open source software and Claussen et al. (2012) in the electronic gaming industry. Both papers consider the relationship between developers' network positions and the success of the projects they are working on. Our research considers a different network in a similar context, namely the hyperlink network of articles. Thus, we borrow from the approach used by Halatchliyski et al. (2010) who analyze authors' contributions in two related knowledge domains considering the article network.

Several previous papers have studied collaboration between authors on Wikipedia. Denning et al. (2005) discuss problems associated with the collaboration of volunteers in

³For a more detailed summary of the literature (until 2008), cf. Jackson (2008).

⁴Without using the more recently developed measures of network position, de Solla Price (1965) evaluates citation data and provides several interesting statistics on average references and citations in the network.

Wikipedia, such as the unknown quality of articles or accidental inaccuracies. Focusing on a non-monetary reward tool at Wikipedia, “Barnstars”, which can be awarded to hard working authors, and its contribution to content creation, Kriplean et al. (2008) offer a theoretical lens for understanding how wiki software can be designed to support the contribution of good work. In his dissertation, Soto (2009) reviews further existing research based on Wikipedia data and (among other things) quantitatively analyzes the ten largest Wikipedias, finding that the patterns concerning the composition of authors on the platform, as well as production patterns, are highly similar.

Other empirical analyses focus on the determinants of the quality of articles. Kittur and Kraut (2008) examine how the number of collaborating editors in Wikipedia and the coordination methods they use affect article quality measured by peer evaluations in Wikipedia’s quality assessment project. Their empirical results show that adding more editors to an article improves article quality only when the editors use appropriate coordination techniques. Zhang and Zhu (2011) empirically examine the potentially inverse relationship between the incentives to contribute and the size of the group of contributors. Based on exogenous variation in group size at the Chinese Wikipedia due to access blocks issued by the government, their analysis shows that contributors receive social benefits increasing with both the amount of contribution and group size. Accordingly, the result confirms that the more contributors value these social benefits, the more they tend to reduce their contributions after the block.

Ransbotham et al. (2012) analyze the relation between the network of authors associated with the collaborative writing of articles and the content value measured as article views. Their results based on social network analysis reveal a curvilinear relationship between the number of distinct contributors to user-generated content and viewership. They conclude that network effects are stronger for newer articles. Gorbatai and Piskorski (2012) and Piskorski and Gorbatai (2010) also test hypotheses related to the author network underlying Wikipedia. They ask whether the density of their individual social networks is related to both norm violations of authors and the likelihood of their easy discouragement after deletions and reverts of their work.

Ransbotham and Kane (2011) analyze the duration until an article on Wikipedia is promoted to a featured article or demoted. They find that an article is most likely to be promoted if the average experience of authors is close to the mean. Articles written by relatively “young” and relatively “old” teams face a longer time span until they are promoted. Halatchliyski et al. (2010) analyze contributions of authors that contributed to articles in two related but different domains of knowledge. They find that the authors that are most central in the author network also contribute to integrating the two fields. Greenstein and Zhu (2012*a* and 2012*b*) investigate the language bias of articles and its evolution over time. Comparing articles in the English Wikipedia to two reference

corpora taken from publications of the Democrat and the Republican party in the U.S. Congress, they find that an early bias of Wikipedia towards Democrat language has gradually disappeared over time. Yet, this erosion of the overall bias comes from new articles, which use Republican vocabulary, while articles which used to be biased appear to stay biased. Gorbatai (2011) employs data from Wikipedia to highlight how demand and supply can be aligned in the absence of market prices. She shows that “professional” editors of Wikipedia strongly react to (attempted) contributions of “unexperienced” users, as they are a sign of increased demand.

3.3 Data

We downloaded a full-text dump from the Wikipedia toolserver and constructed the time-varying graph of the article network on a weekly basis. In addition to the measures of an article’s network position, which lie at the heart of our analysis, we extracted data on additional variables like the length of the page, the number of authors or the categories to which the article belongs. These variables are described in detail in the next section. Details about the data extraction and the descriptive statistics can be found in Appendix B.

3.3.1 The Anatomy of the Data Set

In the data set we find approximately 7,000 articles that were inexistent at the beginning of our period of observation or ceased to exist before the end and are, hence, excluded from the analysis. Using network analysis we identify one large cluster within the category that can be reached via the directed network of incoming links. Following a typical classification that Capocci et al. (2006) apply to Wikipedia, we observe that these pages are either part of the one strongly connected component (set of pages mutually reachable via hyperlinks) or of the out-component (pages reachable from the strongly connected component) of the subnetwork formed by pages associated to the category of economics. We observe 7,635 pages that are always part of this cluster, which we refer to as the “connected component in the category economics” (or just “connected” or “reachable” articles). The other pages could not always be reached via the category network. During the period of observation, 1,237 of these pages received an incoming link from the connected component in the category economics, and thus became part of that component.

Consequently, we use two data sets for our analysis. The first data set is a balanced panel observing the 7,635 articles that remain in the connected component during 153

weeks. It contains in total 1,168,155 observations.⁵ In the second data set we use only those articles that get connected to the economics category during the period of observation. In total we count 1,237 such pages and observing them weekly results in 203,031 observations of this group. In this sample we discarded a small portion of articles that are not only disconnected (in the sense of not linked to the major cluster in the network) from the economics category but also from the entire German Wikipedia at some point in time.

Table 3.1 provides summary statistics of our variables for the balanced panel of articles that are always reachable from the category.⁶ The unit of observation is an article in a given week and we observe the network position of each article in terms of incoming hyperlinks. We observe the length of a page in bytes, how many authors it has and when it was created. One byte corresponds roughly to one letter. The median length is 3630 bytes and the median article was written by 16 authors. Our main centrality measures are indegree and closeness centrality. Both are calculated for the entire Wikipedia and for the articles belonging to the category economics. The indegree is calculated as the number of direct links pointing to a page from the entire German Wikipedia and from the category the article was drawn from. Since articles from the category are also contained in the entire Wikipedia, we report the difference of the two indegrees. By sample construction, every page is connected to the category and hence receives at least one link from it. The median page has eleven links from Wikipedia, four of which are from within the category. Articles usually belong to more than one category, but we do not observe these additional categories.⁷ The distributions of the centrality variables show that for many articles half or more of the links come from economics. Consequently we consider that this category is central to the majority of the articles we observe. Maximal values of page length, the number of authors and indegree lie far above the 90th percentile.

The closeness centrality measures are based on the inverse average distance of one article to all other articles in the relevant network.⁸ Again, the directed centrality measure is computed on both the network made up by pages in the category and the

⁵In ongoing research we analyze articles that come to existence during the period of observation.

⁶Since many distributions are strongly left-shaped while having a long right tail, we prefer tables with percentiles to a graphical illustration

⁷Except for the category sociology that we use for sensitivity analysis.

⁸Closeness centrality in terms of incoming links for an article i on a network containing N articles is defined as the inverse of the sum of shortest paths (geodesic distances) D_{ij} to that article multiplied by the maximal path length $N - 1$. Articles j from which no path leads to i ($j \notin M$) are assigned the distance N , which exceeds the longest possible distance by one:

$$C_i = \frac{N - 1}{\sum_{j \in M} D_{ij} + \sum_{j \notin M} N}.$$

entire German Wikipedia. We observe in our data that the original closeness measures are mainly driven by the variations in the share of disconnected articles and in the network size over time (not reported). In order to abstract from these effects, we compute the relative closeness ranks for our balanced panel. This procedure may be useful in work on dynamic networks in general. In the econometric estimation we use age and dummies for redirect pages and pages containing a literature section as control variables. Several other variables come to mind, that could be included as controls. Some of them are the number of references, the number of distinct authors that contributed in the past, etc. However, on the page level, fundamental differences in the averages of the levels or the growth rates in these variables are all captured by the page fixed effect we include in every regression. We therefore opted for a succinct specification with only the three controls we already mentioned. Age captures whether the article has been on the wiki for a long time or whether it is still “under construction”. The indicator variable for redirect pages flags pages that were converted to a linkage, which merely redirects the reader to the page of a synonym. The presence of a literature section, finally, points to an article that draws extensively on scientific, literary or journalistic sources outside Wikipedia and therefore tends to be longer. The median age of articles is 217 weeks, that is roughly four years. Only around ten percent of the articles are less than two years old, so the majority of articles in our sample are mature articles. Table 3.2 shows the same summary statistics as Table 3.1, but for the sample of articles that get connected to the category of economics during the period of observation. We consider the sample over the entire 153 weeks. In the beginning, none of the articles can be reached from the main component but all become connected later on. The page length and the number of authors are generally a bit smaller, but otherwise show a rather similar distribution, except for the 90th percentile and the maximum. The median page length of 3,044 bytes is about 600 bytes shorter than the median page length of articles that are always part of the connected component. The number of links within the category is smaller by sample design, since most of the articles are disconnected from the main component of the category for many weeks. The number of links from outside the category is similar in median in both samples but considerably smaller in the upper percentiles of the sample of articles that are initially disconnected. We do not report the closeness in this sample because it is mainly driven by the fact of being connected or disconnected. The articles are slightly younger than those in the main sample, but the median age still lies far above three years.

Figure 3.1 shows the development of median values of page length, the number of authors and indegree over the 153 weeks observed. The figure documents the growth that articles experience over time and hence the need to control for time effects in our estimation.

To see how often the variables typically change for individual pages, we aggregate the frequency of changes in the network and content variables over time. This is shown in Table 3.3, where the unit of observation is a page observed throughout the 153 weeks and the table displays the number of changes in variables. The changes are reported for our main sample of articles that are always reachable from the large component of the category. Less than 25 percent of the pages never experience any change in their number of incoming links and less than ten percent are never edited nor receive any additional author. At the same time we see that most articles do not change in any given period, since the frequency of changes of 90 percent of the articles lies at or below 15 to 36 out of 153. An exception are the closeness measures, which change nearly every week for every page. They depend on the structure of the entire network, which is subject to almost permanent change, especially when the entire German Wikipedia is being considered.

Finally, Table 3.4 displays the magnitude of changes for all observations with non-zero change. The reason not to keep the balanced panel here is to make the distribution of changes more visible, which is otherwise dominated by zeros. The median change in page length is 18 bytes in a week, which corresponds to about two words. This highlights that small changes are frequent in the work that many authors contribute to Wikipedia in order to improve the quality of articles. The 75th and the 90th percentile lie at 70 and 309 bytes, which corresponds to a short sentence and a very short paragraph. The median and also most frequent change in incoming links per week is equal to one. The maximal values of changes in page length and links seem to correspond to reverts of entire articles and lie far above the 99th percentile. Changes in closeness are quite symmetrically distributed around zero, which is not surprising, since we use a relative closeness measure. As much as 80 percent of the changes amount to an increase of far less than one point (of 100) in the relative closeness position per week. The distribution of changes is important for interpreting the strength of the effects obtained in our regressions.

3.4 Relationships of Interest and Methodology

3.4.1 Network Position and User-Generated Content

We are interested in analyzing whether a higher centrality in the article network is associated with (i) more content being generated and (ii) contributions by new rather than by previous authors of a page. Our main explanatory variables are measures of centrality in the network of incoming hyperlinks. As described in the previous section, we have four centrality measures: the number of incoming links within the category

economics (indegree within category) and from the entire German Wikipedia (global indegree) as well as the closeness rank in the network of the category and in the global network. As further control variables we add dummies for an article being a redirect, for the presence of a literature section and for article age. We assume that the relation between outcomes and indegrees may be linear or quadratic while the other variable enter our estimation only in a linear way. Data from Wikipedia pages are generated inside two network contexts, the authors network, analyzed in several previous studies, and the hyperlink network formed by the pages, which we are investigating. The skewness and the long tails in the distributions of the number of incoming links, the page length and the number of authors underline that the data show similar properties as other network data. Like with almost all dynamic network data, at least three sources of endogeneity play a role in potentially affecting our estimates.

Firstly, articles differ substantially in their relevance to the wider audience and in other unobserved dimensions. Particularly the difference in their relevance is likely to affect both the network position and the content generation in the same direction, thus generating correlation between these two variables. Secondly, Wikipedia is a collaborative site where the content matter of certain pages is subject to unobserved exogenous shocks and seasonalities. Sudden spikes of interest in certain issues might lead to more authors contributing to single pages or to the entire platform. Moreover, since contributions to Wikipedia continuously grow and inevitably generate some hyperlinks, page length and hyperlinks may both have a time trend. The third source of endogeneity stems from editors who simultaneously edit page B and set a link from page A to page B. Such activity will also lead to a correlation between the network position of a page and its content, but the author's attention will not have been attracted to editing page B via the link from page A. Other and observationally equivalent problems are caused by temporal variations in other unobserved factors such as authors' idiosyncratic preferences, or article popularity in general, which influence both content creation and links. Note that measuring the position of articles based on a two-mode author-article network suffers from similar problems.

Like Kittur and Kraut (2008) and Ransbotham et al. (2012) we use the temporal structure of the data to track the variation within one and the same article by using article fixed effects. Moreover the data are rich enough to allow controlling for systematic temporal variation or particularities of singular weeks by employing time fixed effects. We estimate two-way fixed effects panel regressions based on the following equations:

$$(page\ length)_{it} = \alpha_i + \alpha_t + \beta * (centrality_{it}) + \gamma * X_{it} + \epsilon_{it} \quad (3.1)$$

$$(num.\ authors)_{it} = \alpha_i + \alpha_t + \beta * (centrality_{it}) + \gamma * X_{it} + \epsilon_{it} \quad (3.2)$$

where $centrality_{it}$ is a vector of the four centrality measures mentioned above. X_{it} includes the three control variables indicating redirects, literature sections and age (weeks since the first edit), i designates the article and t the week. Fundamental differences between pages in the averages (of levels and growth rates) of other relevant variables, such as the number of references, the number of distinct authors that contributed in the past, will all be captured by the page fixed effects, which we include in every regression. Since the data allow observing an article's network position in a panel design, we can effectively tackle the first two sources of endogeneity, which are constant heterogeneity specific to articles and time trends or time-dependent shocks that affect the entire network.

Tackling the third source of endogeneity, reverse causality from content to links, is more difficult in our data of connected articles as it cannot be dealt with by fixed effects alone. Unfortunately, we are also not aware of any completely exogenous and quasi-experimental source of variation of articles' network position that allows us to set up an empirical analysis based on the period before and after of such a variation.

However, we can implement a research design, where we look at a large variation in network position and analyze the growth of that article before and after this event. To do so, we can make use of a special type of pages in order to shed more light on the effect of network position on content provision. These are the articles that are initially disconnected from the large economics cluster and that got connected in our period of observation. In order to understand why looking at these articles may be useful, note that authors in general do not observe whether an article is connected to a large component or not. Experienced users may look at the option that allows to display the direct links pointing to a page. Yet, users will not necessarily employ it when linking from another page and, more importantly, they will not see how the linking articles themselves are connected. Most authors will thus not consciously decide to link an article from a large cluster of several thousand articles from which it was previously not accessible. The length of the page may influence the creation of links towards this page. But we expect that there is no systematic relation between page length and whether new links come from outside the category, from isolated pages within the category (which leaves the article disconnected from the cluster economics) or from the main cluster of the category. If we find an effect of getting connected to the large cluster of the category economics that is strong and lasting compared to the coefficients of the indegrees found in the sample of always connected articles, we consider that it plausibly results from the sudden sharp increase in connectedness. This sharp increase is reflected in a discontinuity in the closeness centrality.

When looking at the articles that get connected to the category, we examine both the effects on the page length (level) and on the growth in page length. If we find a

significant effect of getting connected on page growth, we consider it to be unlikely to rely on a correlation between connectedness and the error term, since this unobserved effect on the error term would have to coincide with connectedness not only in the period of getting connected but also in future periods.

3.4.2 Getting Connected to the Category of Economics

In order to analyze the effect of becoming part of the connected component in the category of economics, we put together a sample that includes articles that are at first not connected, but become connected to the category at some point during our period of observation. There are in total 1,237 of these articles. Since the change in closeness centrality is very similar for all of them, we just consider a dummy for becoming connected. We do not consider additional changes in indegree, since we know that most articles change by one link at maximum in a given week and do not change in most weeks. Therefore accounting for getting connected and indegree simultaneously may result in overcontrolling. We analyze both the length and the rate of change of a page from five weeks before the page becomes connected until five weeks after. In a few cases we observe that a page was connected more than once. In those cases we consider only the last time when the page gets connected in our sample.

For the eleven weeks in the sample, we regress page length on an indicator variable that takes the value of one if the page can be reached via the links from the main component of economics and zero otherwise. This means it takes the value zero in the five weeks before connection and the value one in the week when connection occurs as well as in the five weeks after. Furthermore we regress the first difference of page length over time on the same indicator variable. The two-way fixed effects regressions thus take the form:

$$(page\ length)_{it} = \alpha_i + \alpha_t + \beta * \iota(page\ connected)_{it} + \epsilon_{it} \quad (3.3)$$

$$\Delta(page\ length)_{it} = \alpha_i + \alpha_t + \beta * \iota(page\ connected)_{it} + \epsilon_{it} \quad (3.4)$$

with $t = 0$ at the period of the jump into the category and $t \in \{-5, \dots, 5\}$.

In order to alleviate the concern that becoming connected is rather the effect than the cause of simultaneous editing of the target page and the pages pointing to it around week 0, we compare weeks -7 to -3 with weeks 3 to 7 in a further specification (reported in Table 3.7). While our approach reduces the vulnerability to simultaneity issues in important aspects, fully disentangling the factors that might drive simultaneity would require exogenous instruments or the ability to explicitly account for the identity of the linking articles and their properties, which we believe to be a fruitful avenue for further research.

3.5 Results

Table 3.5 shows the two-way fixed effects regressions corresponding to equation 3.1, where page length is regressed on several sets of network variables, article fixed effects and time fixed effects.⁹ The table shows the result for 7,635 articles from the category economics that belong to the large cluster in that category throughout the entire 153 weeks. The first column shows the coefficients for the number of links that the page receives from the entire Wikipedia and a squared term. Our estimates indicate that an additional link pointing to a page is associated with 13 more bytes of text. This corresponds to one or two words. The insignificant coefficient on the quadratic term indicates no curvature. A main question of our investigation is whether the effect of links from the category is different from the mean effect of all links. In the second column we add the number of links that the page receives from other pages of the category economics. These links represent a subset of the global links. The effect can be interpreted as the additional effect from a link being a category link. The coefficient for a category link is more than ten times higher than the coefficient obtained when not differentiating between the two groups of links. Moreover, the new variables render the coefficient for a link that comes from outside the category small and insignificant, suggesting that the explanatory power mostly stems from the category network. Since we run regressions with article fixed effects, the coefficients apply to deviations from the averages that are specific to the article. If the number of incoming links from the category exceeds this average by one, the target page is by 141 bytes longer (considering the sum of the two linear coefficients). For links from the category we estimate significant declining effects, with the coefficient for the quadratic term taking, however, a rather low value of -0.13 .

Column 3 and 4 add the relative closeness rank, which measures whether a page is located rather in the center of the network or rather in its periphery. Column 3 shows the specification of column 1 augmented with the relative rank in closeness on the entire Wikipedia. Given that we scaled the rank variable such that it ranges from 0 to 100, the coefficient indicates that a ten points improvement in the relative closeness position is associated with 150 additional bytes of content. In the descriptive statistics we saw that the closeness of most articles changes by less than one point in any given week. From this point of view the effect looks small. Moreover, the size of the coefficient for indegree is barely affected and the added explanatory power of the new variable is rather low. Finally, Column 4 brings together all available network variables, including the measure of the closeness rank both on Wikipedia and inside the category. The

⁹Time fixed effects were implemented manually by adding a dummy for each point in time in the regression.

coefficient of the closeness rank inside the category is insignificant and the coefficient of the closeness rank on the entire German Wikipedia is even smaller than in column 3. The coefficient of the number of links from the category remains very close to its value in column 2. The control dummies for redirects and a literature section have the expected signs. Older articles tend to be longer.

To analyze the robustness of the last result we replace the contemporaneous measures of centrality by the ones from the week before and we ran the analysis after eliminating more outliers from the sample. We also perform the estimation for a different category, sociology, and we repeated the estimation when including a proxy for how often a page was clicked in the last week. Finally we checked, whether the higher centrality is not only associated with more content but also with more authors. All of these checks confirm the main results from above. All robustness checks are presented in Appendix B. We report the four robustness checks on content generation in Table B.1, and the results on new authors in Table B.2

Summing up our results for the connected component in the category of economics, we find that a higher number of links from articles in the same category is associated with more content generation and additional authors. The increase in page length related to an additional link from the category may look small since it corresponds to a short sentence. From the descriptive statistics we saw, however, that small changes are an essential ingredient of the development of Wikipedia. Consequently, we consider the effect as non-negligible. The effect of links from outside the category is insignificant in our main specification and significant but about three times smaller in some robustness checks. The effect of closeness centrality is negligible.

The regressions in Tables 3.6 and 3.7 use the information on the 1,237 pages that get connected to the main cluster of economics during the period of observation. This is associated with a discontinuous jump in closeness centrality at the time of connection, which can be identified and used to contrast the level (and the growth) of the content before and after this event. Table 3.6 shows the results when we consider 5 periods before and after the jump, also including the period of the jump itself. The first two columns show the results from a simple pooled OLS regression, whereas columns 3 and 4 show the two-way fixed effects results when including both time and article dummies. The coefficients affecting the level of the page length (column 1 and 3) indicate that getting connected is associated with an increase in page length by approximately 400 bytes. This effect is both significant and sizeable compared to the effect of one additional link in the previous sample. The explanatory power of the regression is, however, very low. The cumulative effect over five weeks is even stronger for the first differences of page length (columns 2 and 4), ranging from 66 bytes per week in the pooled regression to 195 bytes per week when including time and article

fixed effects. These are sizeable effects which cannot be expected to last forever. It might be that a share of the additional content is provided in the same week as the article gets connected.

In Table 3.7 we account for that possibility, by excluding the week of the “jump” into the connected component and the two weeks before and after. Instead we consider two five-week intervals that are separated by the interval two weeks before and after the jump (i.e., week -7 to -3 vs. week 3 to 7). As expected, the coefficients get smaller, which indicates that a substantial fraction of the newly generated content is provided within the weeks after the new connection was established. However, the effects remain by and large positive and indicate that an article grows by 9 (pooled) to 21 bytes per week (fixed effects) faster during weeks 3 to 7 after being connected. We still observe not only a level but also a growth effect.

3.6 Conclusion

The creation of user-generated content in a peer production setting requires mechanisms that help producers to identify content they want to contribute to. We consider the network of hyperlinks between Wikipedia articles as a possible channel of spillovers in production activity that attracts more producer effort to more central articles. We find that the page length of an article is positively associated with the number of links pointing to it after controlling for time-invariant unobserved heterogeneity, time effects and several other variables.

On average, one more link is associated with a page length that is 13 bytes higher, which corresponds roughly to one or two words. When differentiating between links within the category economics, which we selected as sample, and links from other Wikipedia pages, we find a large discrepancy in effects. One more link from an article from inside the category is related to an increase in page length of around 140 bytes. This is a sizeable effect given that the median weekly change in page length, excluding observations without any change is only 18 bytes. At the same time, the coefficients for links from outside the category becomes insignificant. The importance of links from the same category is corroborated in several robustness checks which persistently confirm that the effect of links from outside the category is much smaller. Moreover, links from the category are strongly related to new authors’ contributions. On average every second additional link from the category is associated with a new author contributing to the page. These results are all obtained in a balanced sample of articles that are always connected to the large cluster of the category. Articles that are initially not connected increase by more than 300 bytes in length during the five weeks after connection.

Taken together the results suggest that adding missing hyperlinks to Wikipedia or

extending the content of articles in a way that it connects better to other articles may not only improve the quality of the information but also foster further contribution by authors that have not yet contributed to the newly linked articles. While the size of the additional contributions that may be expected is not very high, these changes of a few words or one sentence constitute a large part of contributions to Wikipedia. This strategy is expected to work best within a cluster of thematically related articles. Links from articles that do not share a central category with the target article seem to enhance content generation much less. Thus we find new evidence that semantical relatedness may matter more than the mere presence of direct links between pages in generating spillovers in content provision.

From a researcher's perspective, our results suggest that it may be an acceptable strategy in the context of content networks to use only a smaller group of articles (nodes) for network computations, which share a common category, as long as one does not extrapolate the result to the unobserved nodes. This should not be said without adding a word of caution: First, our results are not based on a two-mode author-article network considered in several other studies but on the link network of Wikipedia articles. Whether they extend to two-mode contexts remains to be tested. Second, our conclusions are obtained based on data from relatively mature articles and should be reexamined for newly created articles.

3.7 Tables

3.7.1 Summary Statistics

Table 3.1: Summary statistics of main variables. Connected articles.

	min	p10	p25	p50	p75	p90	max
Length of page (in bytes)	20	1049	1872	3630	7470	14089	229379
Number of authors	1	6	9	16	30	56	821
Links from Wikipedia	1	2	5	11	28	76	7981
Links from Wikipedia excl. categ.	0	0	2	6	17	53	7750
Links from category	1	1	2	4	10	23	667
Rel. closeness rank (Wikipedia)	.013	10	25	50	75	90	100
Rel. closeness rank (category)	.013	10	25	50	75	90	100
Dummy: literature section	0	0	0	0	0	1	1
Dummy: page is redirect	0	0	0	0	0	0	1
Age (in months)	1	113	162	217	271	316	492

Articles that were always connected to econ. main component. Number of observations: 1168155

Table 3.2: Summary statistics of main variables. Articles that get connected to category.

	min	p10	p25	p50	p75	p90	max
Length of page (in bytes)	19	915	1653	3044	5207	9231	67988
Number of authors	1	5	8	12	20	33	267
Links from Wikipedia	1	2	4	7	13	24	3914
Links from Wikipedia excl. categ.	0	1	2	5	10	21	3910
Links from category	0	0	1	1	2	4	122
Dummy: literature section	0	0	0	0	0	1	1
Dummy: page is redirect	0	0	0	0	0	0	1
Age (in months)	1	84	129	181	236	283	451

Number of observations included: 203031.

Table 3.3: Summary statistics of the frequency of changes of main variables.

	min	p10	p25	p50	p75	p90	max
Length of page (in bytes)	0	3	5	11	22	36	136
Number of authors	0	2	4	7	14	24	123
Links from Wikiped (excl. categ.)	0	0	1	4	12	34	152
Links from categ.	0	0	1	3	7	15	121
Rel. closeness rank (Wikipedia)	152	152	152	152	152	152	152
Rel. closeness rank (categ.)	149	151	152	152	152	152	152

The unit of observation is a page over entire period. Number of pages included: 7635

Table 3.4: Weekly changes of main variables.

	Min	p1	p10	p25	p50	p75	p90	p99	Max	Obs.
Length of page (in bytes)	-95,222	-868	-42	-1	18	70	309	2,739	83,235	124,771
Number of authors	1	1	1	1	1	1	2	3	76	82,260
Links from Wikipedia	-439	-2	-1	1	1	1	2	8	1,455	121,589
Links from Wikiped (excl. categ.)	-439	-2	-1	1	1	1	2	9	1,455	90,214
Links from category	-130	-2	-1	1	1	1	1	4	80	46,304
Rel. closeness rank (Wikipedia)	-92	-1.6	-0.58	-0.23	-0.02	0.18	0.51	1.8	91	1,137,528
Rel. closeness rank (category)	-99	-0.98	-0.20	-0.11	-0.04	0.03	0.13	2	85	1,090,973

Articles that were always connected to econ. main component. Only observations with non-zero changes.

3.7.2 Regression Results

Table 3.5: Relationship of page length and centrality.

	(1) Wiki degree	(2) Wiki & cat.	(3) add closeness	(4) all vars
Links from Wikipedia	13.333*** (3.18)	2.958 (1.22)	12.934*** (3.14)	2.931 (1.22)
(Links from Wikipedia) ²	-0.000 (-0.54)	0.001** (2.04)	-0.000 (-0.47)	0.001** (2.07)
Links from category		138.129*** (8.80)		135.871*** (8.47)
(Links from category) ²		-0.130*** (-5.24)		-0.127*** (-5.02)
Rel. closeness rank (Wikipedia)			15.216*** (6.17)	7.505*** (3.08)
Rel. closeness rank (category)				-1.230 (-0.67)
Dummy: literature section	1295.963*** (6.11)	1249.985*** (5.95)	1287.521*** (6.07)	1248.055*** (5.94)
Age (in months)	10.648*** (21.55)	8.361*** (22.76)	10.692*** (21.85)	8.416*** (22.46)
Dummy: page is redirect	-546.408 (-0.57)	-742.157 (-0.77)	-590.851 (-0.59)	-767.075 (-0.77)
Constant	3336.571*** (30.10)	2803.789*** (22.18)	2582.005*** (16.28)	2501.686*** (15.73)
Time dummies	Yes	Yes	Yes	Yes
Observations	1168155	1168155	1168155	1168155
Groups	7635	7635	7635	7635
Adj. R ²	0.107	0.130	0.109	0.131

t statistics in parentheses

2-way fixed effects OLS regressions with both time and article dummies (robust stand. errors)

Only articles connected over entire period were included. Dependent variable: page length.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.6: Relationship of the growth of page length and the page becoming connected.

	(1)	(2)	(3)	(4)
	OLS Levels	OLS Differences	2-Way FE Levels	2-Way FE Differences
Dummy: page is connected to cat.	439.133*** (5.49)	66.343*** (6.06)	317.699*** (5.72)	194.809*** (5.48)
Constant	4059.235*** (70.60)	10.458*** (4.10)	2584.101*** (6.22)	-2056.589*** (-4.76)
Time dummies	No	No	Yes	Yes
Observations	14376	14324	14376	14324
Groups			1327	1327
Adj. R ²	0.002	0.002	0.037	0.007

t statistics in parentheses

Columns 1 and 2 show pooled OLS-Regressions, Columns 3 and 4 include article and time fixed effects.

All Regressions use heteroscedasticity-robust standard errors. Dependent Variable: page length.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.7: Growth of page length when page gets connected: Excluding 2 periods before and after.

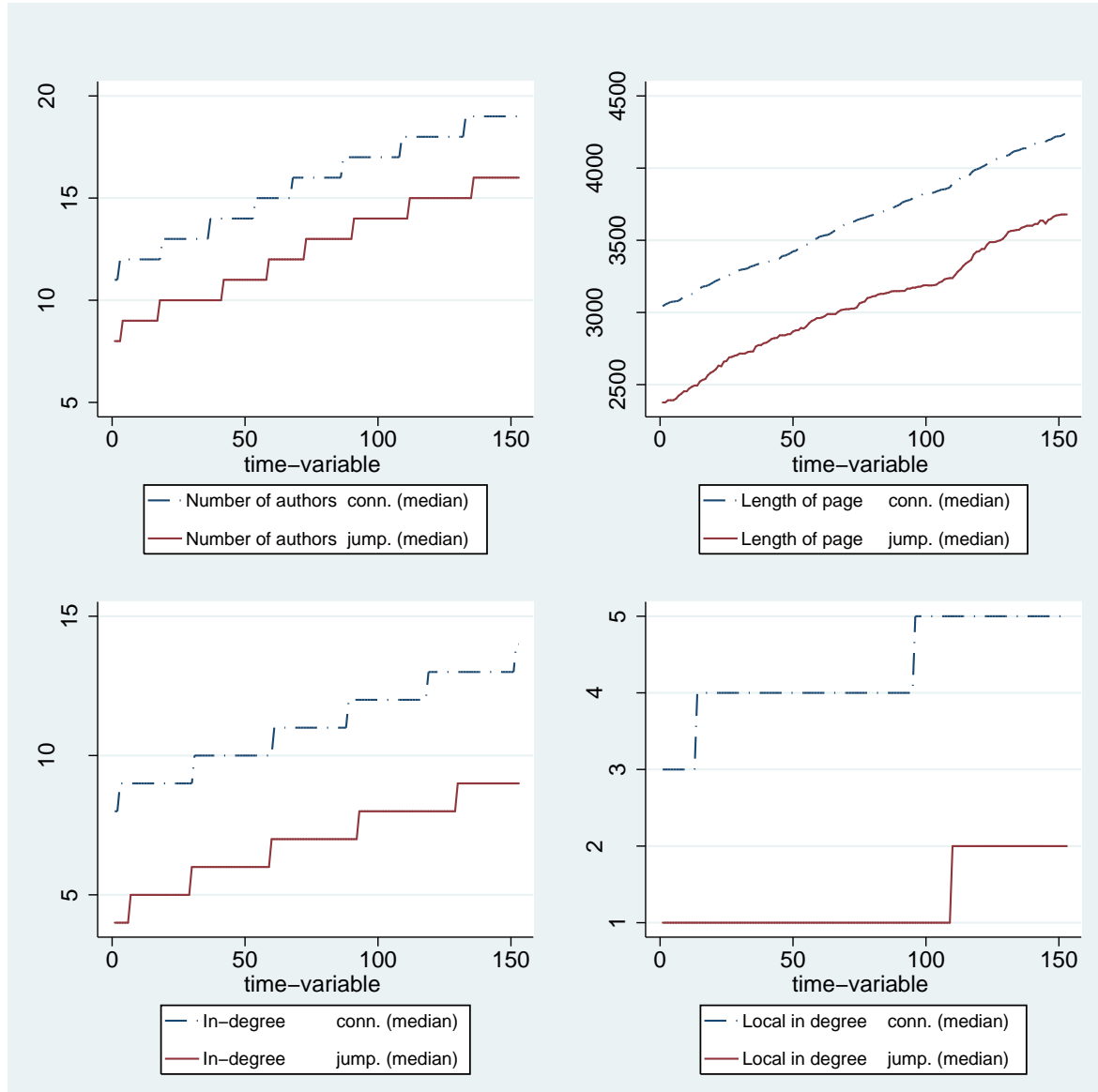
	(1) OLS Levels	(2) OLS Differences	(3) 2-Way FE Levels	(4) 2-Way FE Differences
Dummy: page is connected to cat.	369.197*** (4.38)	8.650** (2.17)	255.683*** (3.61)	21.334** (2.02)
Constant	4049.740*** (69.20)	7.293*** (4.50)	3654.610*** (12.47)	-116.975 (-1.30)
Time dummies	No	No	Yes	Yes
Observations	12283	12237	12283	12237
Groups			1268	1268
Adj. R ²	0.001	0.000	0.042	0.002

NOTES: The table shows the relationship of the growth of page length and the page becoming connected, excluding the period of the jump itself and the 2 periods before and after. *t*-statistics in parentheses. Columns 1 and 2 show pooled OLS-Regressions, Columns 3 and 4 include article and time fixed effects. All Regressions use heteroscedasticity-robust standard errors. Dependent Variable: page length. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.7.3 Figures

Development of variables over time

Figure 3.1: Development of the median of the outcomes and the indegree over time.



NOTES: The figure shows the development of the median of the outcomes and the indegree in both groups, the always connected sample and the smaller jumping sample.

Chapter 4

Market Structure and Market Performance in E-Commerce*

4.1 Introduction

Under reasonably general conditions, the consequences of an increase in the number of market participants are lower prices and lower markups. The empirical assessment of this relation is however not an easy task. Markups are not readily available, and prices and market structure are endogenous: firms may enter in response to perceived profit opportunities or may exit in response to realized losses.^{2,3}

In this paper, we use a novel instrumental variables strategy to investigate the interaction between market structure and market performance in e-commerce. We use data for digital cameras from an Austrian online price-comparison site (price search engine). We observe the firms' retail and input prices as well as all their moves in the entry and the pricing game. When we measure the rate at which markups decline towards zero, we account for the endogenous timing decision to list a specific product by using previous listing decisions as instruments. In addition, we include product fixed effects to capture unobserved quality and design features of the specific cameras

*This chapter is coauthored with Franz Hackl (University of Linz), Rudolf Winter-Ebmer (University of Linz & IHS, Vienna) and Christine Zulehner (Goethe University Frankfurt & WIFO, Vienna). A paper with the same title is forthcoming in the *European Economic Review*. We would like to thank the editor and two anonymous referees for very helpful comments. Furthermore, we thank Irene Bertschek, Sara Fisher-Ellison, Avi Goldfarb, Gautam Gowrisankaran, Michelle Haynes, Jose Luis Moraga-Gonzalez, Ariel Pakes, Martin Peitz, Philipp Schmidt-Dengler and the participants of the SEEK 2010 Workshop on online markets for valuable comments. We are indebted to geizhals.at and a producer of consumer electronics for providing the data. Andreas Lumesberger and Sergiy Golovin provided invaluable research assistance. This project was supported by ZEW's SEEK research program.

²One way to account for endogeneity is developing a structural model of market structure, entry, and exit. The pioneering study on entry into local markets by Bresnahan and Reiss (1991) shows that the first two or three entrants have the largest impact on market price, and that later entrants do not significantly reduce market price any further.

³Experimental evidence of this relation goes back to Selten (1973), who coined the statement "four are few and six are many."

as these might be correlated with both markups and firms' entry. To obtain a full picture of the underlying model of competition, we then follow Baye et al. (2004) and Haynes and Thompson (2008a) and analyze measures of price dispersion as well.⁴

We further analyze the relation of the number of firms and markups across the product life cycle. Products in e-commerce are very often only listed for a short time, which allows us to observe products from birth to death.⁵ This is important for three reasons: i) Entry in such a market is particularly easy because an existing firm only has to decide whether to list a new camera or not. This low entry cost makes the number of firms volatile and provides an optimal testing field.⁶ ii) Several researchers have claimed that competition or the absence thereof is particularly important at the beginning of a product life cycle, while later on, competition may matter less.⁷ In particular, when a new product emerges on the market and consumers are uncertain about their tastes, they may postpone their purchasing decision. Firms react to this uncertainty of demand and various price dynamics might be the consequence.⁸ iii) Finally, we investigate the effect of substitutes on the markup over the product life cycle and are interested in differences between newly innovated and old expiring technologies as well as between own brand and rivals' brand products.⁹

We are not the first to investigate the relation of market structure and market performance in e-commerce. Previous studies such as Brynjolfsson and Smith (2000), Baye et al. (2009), Baye et al. (2003) and Haynes and Thompson (2008a), however, do not take the endogeneity of the number of sellers and product life cycle effects into account.¹⁰ Baye et al. (2003) and Baye et al. (2004) look at price dispersion using

⁴For example, monopolistic competition predicts markups and price dispersion to go down when the number of firms increases (Perloff and Salop, 1985); while in a model with heterogeneity in consumers' search cost and producers' marginal cost the latter would go up (Carlson and McAfee, 1983).

⁵The average span of the product life cycle of digital cameras amounts to 167 days in our data.

⁶In a recent survey, Martin (2012) argues that market structure may adapt only slowly to long-run equilibrium levels and many entering firms may be atypical fringe firms unable to influence market structure at the core. While this describes well-established markets with market leaders and high advertising requirements, market structure in e-commerce is different: due to the cheap and easy establishment of online shops, many such shops operate only online.

⁷Examples include Berry (1992), Campbell and Hopenhayn (2005), Carlton (1983), Davis (2006), Dunne, Roberts and Samuelson (1988), Geroski (1989), Mazzeo (2002), Seim (2006), and Toivanen and Waterson (2000, 2005). For a survey see Berry and Reiss (2007).

⁸See, for example, Bergemann and Välimäki (2006a,b), who analyze dynamic price paths in monopolistic settings and find that in mass markets prices should decrease over the product cycle.

⁹Klepper (1996, 2002) describes the evolutionary pattern of birth and maturity of technologically progressive industries and we apply and extend the predictions of his model to the market of consumer electronics.

¹⁰Barron et al. (2004) analyze the relationship of markups and price dispersion and the number of firms using data from gasoline retail markets. They find that both markups and price dispersion decrease as the number of firms increases and interpret this as evidence in favor of models of monopolistic competition.

various metrics. Baye et al. (2004), for example, analyze price dispersion measured by the relative price gap (the difference of the first and second price) and show that it decreases as a function of the number of firms, but not over time. Haynes and Thompson (2008a) use data on 400 digital cameras in the US and show that with more firms in the market prices go down and dispersion increases. Ellison and Ellison (2005, 2009) examine the competition of internet retailers and identify different strategies that are applied in online markets to cope with the increased price sensitivity.

The empirical literature investigating the market structure along the life cycle of a consumer product is rather small. Haynes and Thompson (2008b) take a first step to explain entry and exit behavior in a shopbot. To do so, they estimate an error-correction model and show that entry into and exit from a market are correlated with a measure of lagged price-cost margins and the number of competitors. Barron et al. (2004) mention the life cycle, but use it only as a control variable. In the marketing literature, Moe and Yang (2009) analyze the product life cycle in e-tailing. However, their data did not allow them to consider the endogeneity of entry and exit. Hitsch (2006) considers the dynamic decision problem of a single firm that is uncertain about the demand for a new product and shows that in the ready-to-eat cereal industry the value of reducing uncertainty is large. This indicates that there are product cycle effects which should be accounted for.

For e-commerce in Austria, we find a highly significant results of the number of firms on markups. *Ten* additional competitors in the market are associated with a reduction of the median markups by 0.23 percentage points and the minimum markup by 0.55 percentage points. However, accounting for the potential endogeneity of markups and the number of firms in the market, we see a substantially higher negative outcome: *ten* additional retailers tend to reduce the markup of the median firm by 0.95 percentage points and the markup of the cheapest firm by 1.24 percentage points. We also find that having one more firm in the market apparently reduces the markup of the price leader by the same amount as the competition between existing firms in a period of three additional weeks in the product life cycle. If we abstract from any dynamic or product life cycle effects, our results support the validity of search theoretic models such as Carlson and McAfee (1983) or Baye and Morgan (2001) and contradict models of monopolistic competition.

We use firms' past listings decisions as an instrument. We argue that this is a valid instrument as products offered in different markets some time ago should have no direct influence on prices and sales of current products. Potential threats to this identification strategy are the timing of past listing decisions and the similarity of products. We thus run robustness checks on the instrument by varying the timing of firms' past behavior and using markets farther away in terms of time or model

specification from our chosen product market. The results of these robustness checks show qualitatively and quantitatively very similar results to our preferred specification.

Furthermore, we find that there is a highly significant age effect, i.e., the longer a product is on the market the lower are markups. Although Bergemann and Välimäki (2006a) only consider monopolistic markets, this result could be interpreted in support of their price dynamics when consumers are uncertain about their tastes of a product which newly emerged on the market and when there is social learning.

Our main results on the effects of competition are robust to the inclusion of varying numbers of substitutes over the life cycle of the product. For products with a higher number of substitutes we measure lower markups, in particular if the substitutes are newer products and come from rivals rather than from the same brand.

The remainder of the paper is organized as follows. We present the theoretical predictions derived in the literature in Section 4.2 and describe the data as well as the empirical strategy in Section 4.3. We discuss our estimation results in Section 4.4 and conclude in Section 4.5.

4.2 Theoretical Predictions

Our paper aims at explaining the effect of market structure on markups and on price dispersion. While under reasonably general conditions the direction of the effect is clear, search cost may complicate things. Therefore, we now discuss the various models and potential hypotheses. As our empirical analysis also includes the effects over the product life cycle, we discuss the literature on price dynamics and industry dynamics with overlapping technologies as well.

As earlier studies have argued, models allowing for price dispersion in a homogenous market have been classified into three groups:¹¹ i) First, search-theoretic models (Varian (1980), Rosenthal (1980)), which evoke price dispersion by introducing heterogeneity in the search costs of consumers. These models predict that an increased number of sellers results in a larger price dispersion and – somewhat counter-intuitively – a higher average price. Baye and Morgan (2001) is an example that directly considers e-commerce. They get rid of the counterintuitive result on market structure by allowing for different groups of consumers (a price-insensitive group and price-sensitive consumers who take advantage of zero online search costs) which makes the retailers' randomization over prices rational. ii) Models of monopolistic competition (Perloff and Salop, 1985) can account for price dispersion when extended by asymmetries across firms such as heterogeneous producer cost or heterogeneous producer demand (Barron

¹¹See, for example, Barron et al. (2004) or Haynes and Thompson (2008a).

et al., 2004). These models would predict that a larger number of sellers is associated with a lower average price and smaller price dispersion. iii) Carlson and McAfee (1983) present a search-theoretic model that accommodates two sources of heterogeneity by assuming a non-degenerate distribution of producers' marginal cost and heterogeneous visiting cost of consumers. Here, the prediction is that average prices would go down while price dispersion would rise.

In the purely search theoretic approach as it was coined by Varian (1980) two types of consumers are present in the market. One type has very low search cost and will hence always buy at the cheapest shop, while the other one with high search cost will buy from a random shop. As a consequence shops have to strike a balance between aiming to be the cheapest shop and selling to all the price-sensitive consumers or confining themselves to their share of price-insensitive consumers but selling to them with a higher markup. In such a setup, everything else being equal an increased number of sellers results in a larger price dispersion and – somewhat counter-intuitively – a higher average price.

Baye and Morgan (2001) theoretically investigate the market for information gatekeepers. They analyze the behavior of firms listed on a price-comparison site as well as the behavior of the monopolistic shopbot. Shops, which have a local monopoly in their town, have to choose between serving only the uninformed population of their own town or advertising on the price-comparison site to potentially serve informed customers in all other towns as well. Consumers, on the other hand, have the option to subscribe to the price-information site or to remain uninformed. In the first case, they can choose from among all shops, but in the latter case they can only buy locally. Given the site's behavior and the share of consumers using the site, the model predicts that the shops will randomize over prices in the price setting equilibrium. They do so in order to maintain positive markups without being undercut by their opponents *with certainty*. Thus, they generate price dispersion in the market for this homogenous product. The impact of more competition on the platform is not explicitly analyzed in the model. Yet it is relatively easy to see that the minimum price (the lower bound of the support of the price distribution in their model) is decreasing in the number of firms, whereas the range of the distribution (price dispersion) increases with an increasing number of firms.

In models of monopolistic competition it is assumed that consumers perceive products to be different across sellers. If all sellers have the same marginal cost, each consumer draws her valuation for the good offered by each seller from a common distribution, and demand is symmetric, then Perloff and Salop (1985) show that an increase in the number of sellers yields an increase in the price elasticity of individual firm demand, lowers the markup and the equilibrium price. If demand is asymmetric

and the number of different seller types is constant, an extension of the analysis (Barron et al., 2004) indicates that the increase in the number of sellers of each type will yield a reduction of markups and prices through an increased price elasticity across sellers. Because of the reduction in markups for sellers and common marginal cost, the variance in markups decreases with an increase in competition. As a consequence the price dispersion diminishes.¹²

Carlson and McAfee (1983) assume monopolistic competition with heterogeneous firms in market where heterogeneous buyers search. Buyers look for the best price until their expected return from visiting one more shop is smaller than their search cost. Shops, which have heterogeneous marginal cost, use pricing rules which depend on the average price in the market. While price dispersion is due to the heterogeneous marginal cost, they explicitly conjecture, that reputation or heterogeneous visibility in advertisement would also generate equilibrium price dispersion. In equilibrium all shops in the market make positive profits (those who do not are predicted to leave the market), yet the most efficient ones make larger profits than the others. An increased number of shops, all else equal, leads to lower prices and a modest increase in price dispersion, which is bounded from above by the heterogeneity of the shops. As far as the informational requirements of the model are concerned the model's assumption that every change of a shop's pricing policy is perceived fits well with the setting on a price comparison site. However, the reasoning of sequential search is somewhat at odds with such a market setting.

The models discussed so far are static models. In a dynamic context, Bergemann and Välimäki (2006a,b) investigate the intertemporal incentives of a new buyer who is uncertain about her tastes for an experience good. Their model of optimal pricing assumes a monopolist that sells a new experience good over time to a population of heterogeneous buyers. While oligopolistic competition is not analyzed, these models provide insights into the intertemporal pricing effects per se. For example, Bergemann and Välimäki (2006a)'s results show that markets can be classified into mass and niche markets. The dynamic equilibrium prices of mass markets decrease over time and buyers purchase in all periods. In a niche market, however, not all consumers buy at the static monopoly price. Therefore, the monopolist initially offers low prices to capture a larger share of consumers. This is at the expense of targeting the more solvent consumer group of the market. The results also hold in the context of social learning which fits our market of digital cameras of one brand producer best. From their analysis we expect price paths that could be either decreasing or increasing, at

¹²See Barron et al. (2004) for an extensive discussion on the models of monopolistic competition and their predictions.

least at the beginning of the product cycle. As we investigate the effect of the number of firms over the product life cycle, our empirical results may also provide stylized facts theoretical models could incorporate.

In his seminal work Klepper (1996) describes the evolutionary pattern of birth and maturity of technologically progressive industries in form of overlapping product life cycle: Innovation, entry, growth, decline and exit are driven by the way new technologies evolve over time.¹³ Consumers are assumed as the driving force behind this evolution: as they experiment with alternative technological variants they form a view on their preferred variation and decide on the success and failure of different offered technologies. As a consequence overlapping product life cycles emerge in which existing technology rivals with newly innovated and old expiring technologies. Although Klepper has its focus on major technological innovations his model can also be applied to our product markets on consumer electronics in which the product innovations manifest as additional or increasingly powerful camera features. Our empirical model differs from the work of (Klepper, 1996) and (Klepper, 2002) in various ways. Whereas Klepper focuses on the innovational process of manufacturers our focus lies on the retailers' markup. Moreover, Klepper does not explicitly distinguish whether the manufacturers' innovation (new product) competes with the competitors' or the own products' life cycle. While it typically always makes sense to skim off the rivals' rents by introducing new and better products, an early launch of new product generations might cannibalize the sales of the the firms' own and earlier introduced products.

4.3 Data and Empirical Strategy

Price search engine: In our analysis we use data from the only Austrian price comparison site, www.geizhals.at. At the time of our analysis Geizhals.at listed on average price offers from 1,200 firms for 200,000 products.¹⁴ The business model of Geizhals is as follows: the retailers have to pay a fixed fee for each referral request of a customer to the respective e-shop.¹⁶ If the retailer agrees to embed the Geizhals logo

¹³Further literature concentrating on market growth as the major driving force that explains market entry and exit are Schmalensee (1989), Scherer and Ross (1990), Shaked and Sutton (1987), Sutton (1991), Sutton (1998), Dasgupta and Stiglitz (1980), Bresnahan and Reiss (1991) and Asplund and Sandin (1999).

¹⁴Recently, Geizhals has expanded to other European countries, including Germany, Poland, and the United Kingdom.¹⁵ This recent internationalization led to a substantial expansion of products and supplying retailers. At present, the website offers more than 723,000 products with 64 million price quotes that can be actualized by the retailers several times per day. In January 2012, Geizhals registered 3.1 million "unique clients." The number of unique clients is calculated from the number of different terminal devices (PCs, PDAs, etc.) used to access a website.

¹⁶A referral request is a click by the customer on the link of an online shop at Geizhals.at. After the click, the online shop of the retailer opens in a new browser window.

and link on its website, a reduced fee is paid. If the total of these click-dependent fees does not exceed a certain limit, the online store has to pay a flat fee. The electronic retailers can list as many products they want and can change the prices as often they want, free of charge. There is also no cost if retailers decide to suspend a certain price quote temporarily. Hence, apart from a relatively small flat rate and the click-dependent fee retailers are not confronted with entry or exit costs in the different product markets.¹⁷ By this construction, Geizhals has the incentive to increase its profits by permanently extending the number of retailers. However, at least in Austria, Geizhals has already acquired such a strong reputation among customers that online stores cannot afford not to be listed at geizhals.at as the market is dominated by this price search engine. Hence, our data cover the whole electronic retailing market in Austria.

Available data: For the study in this paper, we use *daily* data on 70 items (mainly digital cameras) from a major hardware manufacturer,¹⁸ which were listed during the period from January 2007 through December 2008.¹⁹ We define a camera's birth by its appearance on Geizhals.at. The cameras were offered by up to 203 sellers from Austria and Germany. Fixed effects for the different products will control for the unobserved heterogeneity of goods and traders on the varying product markets.

For time t (measured in days), we observe for each product i and retailer j the $price_{ijt}$,²⁰ the $shipping\ cost_{ijt}$ posted at the website,²¹ and the $availability_{ijt}$ of the product.²² Additionally, we observe the customers' referral requests ($clicks_{ijt}$) from the Geizhals.at website to the retailers' e-commerce website as a proxy for consumer demand.²³ Customers have the option to evaluate the (*service*) *quality* of the firms on a five-point scale, the average of which is listed together with the price information on Geizhals.at. *Wholesale prices* for each product i at time t were obtained from the Austrian representative of the international manufacturer. We do not claim that these wholesale prices correspond perfectly to the retailers' marginal cost. Even though the

¹⁷Of the 1,200 retailers at the time of our analysis, only a very small number of retailers have other contracts with Geizhals.at, e.g., they pay only for products actually sold.

¹⁸The hardware manufacturer is a multinational corporation specializing in manufacturing electronic equipment in several areas. The manufacturer asked to keep its name anonymously.

¹⁹For our instrumentation strategy, we use also the product life cycle of cameras entering the market starting from May 2006.

²⁰We would like to stress that in our market transaction prices are equal to posted prices. Consumers of digital cameras are most often no firms and have no bargaining power.

²¹Shipping cost is the only variable that has to be parsed from a text field. We use the information on "cash in advance for shipping to Germany," which is the type of shipping cost most widely quoted by the shops. Missing shipping costs are imputed with the mean shipping cost by the other retailers and controlled for with a dummy for imputed values.

²²We coded two dummies whether the product was available immediately (or at short notice) or within 2-4 days.

²³See Dulleck et al. (2011b) for a description of the data.

manufacturer's distribution policy indicates that the retailers should be served by the local representative, it may happen that single retailers procure commodities from, for instance, the Asian market. Moreover, the local representative might offer special promotions including lower wholesale prices in exceptional cases, e.g. if a retailer commits to promoting the manufacturer's good in a special way. Finally, it must be mentioned that the retailers in e-commerce might have additional costs each time they order in addition to the wholesale price. We assume this additional cost to be constant for all stores. Despite these qualifications, our measures are a very good proxy for the actual marginal cost of the retailers.²⁴ $Price_{ijt}$ and $wholesale\ price_{it}$ were used to calculate the firms' $markup_{ijt}$ ²⁵ according to the Lerner index and the markets' $price\ dispersion_{it}$.

Organization of data: We reorganized the data in such a way that the product life cycles of all digicams start at the same day 1. Hence, we have shifted the product life cycles of the digicams so that we can analyze the impact of market structure on markup and price dispersion in each of the different stages of the product life cycle²⁶. Our panel consists of 70 product life cycles. We define the end of a product life cycle as the point when the number of referral requests diminishes to less than 500 *clicks*. Thus, we use a daily unbalanced panel with information on the *products' age*, the *number of firms*, *average markups*, the *markup of the price leader*, different measures for price-dispersion, and the *number of clicks* for product market i over time t .

Descriptives: Table 1 contains summary statistics of the data. Our dataset includes 70 complete product life cycles. The average length of a model's life cycle is 240 days with a standard deviation 133 days. Each observation in the descriptives refers to a single product i at a given day t in the product life cycle. We will use the markup (Lerner index) and the price dispersion as endogenous variables. Whereas the median markup amounts to 18% on average, the mean markup for price leaders is only 4.8%. These numbers are of comparable size to those in Ellison and Snyder (2011), who report an average markup of 4% for memory modules on Pricewatch.com. We use different measures for price dispersion: the *coefficient of variation* and the *standard deviation of the distribution of prices*, as well as the *absolute price gap between the price*

²⁴According to the Austrian distributor the Austrian and German lists of wholesale prices are almost identical. Note the manufacturer's incentive to keep cross-border sales between distributors and retailers as low as possible - an argument which supports the reliability of our wholesale prices as indicator for marginal cost.

²⁵To account for the problem that high markups might be economically irrelevant, we run all our regressions also weighted by $clicks_{ijt}$. We do not observe quantities sold, only the clicks when a consumer goes from the Geizhals.at website to the retailers' e-commerce website. The clicks do, of course, not reflect the actual demand, but we think is a valid indicator.

²⁶We control for the contemporaneous structure of calendar time with dummies for each different month in the dataset and with the number of substitutes at the respective calendar times.

leader and the second cheapest price. On average, a product life cycle amounts to 166 days with a mean of 104 firms offering the digicams.

Figure 4.1 shows that the estimated markup declines with age, and, more importantly, as the number of firms increases. The decline in markups with the number of firms is a rather smooth phenomenon, and not as quick as one might expect in perfectly transparent e-commerce markets. Even with 70 and more firms in the market, there is a positive markup. In the top left panel, the *median markup_{it}* is scattered against the *number of firms_{it}* in the corresponding market; the top right panel shows the average pattern. The number of firms ranges from 0 to more than 200 and the median markup ranges from 0 to 35%.

There are also some negative markups, especially for the minimum price firms, where the average markup of the price leader is only 4.8%. In our dataset, we observe negative markups for 26.9% of all best-price offers. This is in line with Ellison and Snyder (2011), who also report a substantial number of price offers with negative markups in the case of Pricewatch.com. Negative markups might have several possible causes: they might simply point to sellouts after overstocking, they might hint to cases where retailers are not procuring products via official retail channels, or they might indicate the use of loss leader strategies where a digicam is offered at a price below marginal cost to attract new customers or to make profits with complementary goods.

In the middle row, the median markup is plotted against the age of the product. We typically observe a camera between seven and 15 months. Again, the markets' median markups fall with the duration of the product life cycle. Our assumption of constant variable costs for e-tailers squares well with the flattening of the markup after 3-4 months. In the lower row, the number of retailers is plotted against the age of the camera: there is a steep increase in the number of listing firms at the very beginning of the life cycle, whereas after 12 months the number of firms is declining again. This average pattern hides some heterogeneity, which can be seen on the left-hand side. Some cameras are listed only by a small number of firms (between 20 and 60). Then there is a group of products which is offered by about 60 shops, and finally, the third and largest group of cameras is listed by roughly 150 shops and more. This segmentation can be explained by a specialization of shops on certain product categories. Whereas some shops are focused in their assortment on mass products (simple digital cameras) others restrict their range of products on highly specialized digital SLR cameras for professionals. This heterogeneity over the products provides a first indication on the importance of product-fixed effects in our estimations.

Empirical strategy: To estimate the impact of market structure on markups and price dispersion, we estimate the following fixed-effects regression as our baseline model:

$$markup_{it} = \gamma + \alpha_1 age_{it} + \alpha_2 age_{it}^2 + \beta_1 numfirms_{it} + \beta_2 numfirms_{it}^2 + \omega_i + \tau_t + \epsilon_{it}$$

This model will be estimated for the market's *i median markup* and the *minimum markup*; a similar strategy is used for price dispersion which is measured as the *coefficient of variation* and other measures. Life cycle effects are captured by a quadratic age trend. In a later specification, we compute separate splines for each phase of the life cycle to capture varying competition effects over the life cycle of the product. We included month-specific dummies to account for calendar-time fixed effects (τ_t) and product fixed effects (ω_i). Note that the product fixed effects control for observable and unobservable product characteristics on the basis of which the e-tailers build their profit expectations and thus make their decisions on the offered product portfolio. These product fixed effects filter out product specific characteristics which remain constant over the lifespan of the camera ²⁷. Hence, for our analysis we exploit only the time varying information within the life cycle of each product. However, we have to control for potential endogeneity issues which are based on the time dimension of firms' listing decisions over the product life cycle.

Sources of endogeneity: In all markets, particularly in an e-tailing shopbot market, it is important to treat market structure as endogenous; due of simple and low-cost market entry and exit, e-tailers can easily adapt to changing circumstances by listing a particular product or not. If, for example, unobserved factors temporarily drive up markups for some item, shops that did not sell the item before might move into this market. Thus, we would expect to observe more shops in markets where higher markups can be reaped and vice versa: reverse causation. This, in turn, will result in an upward bias: an estimated OLS coefficient showing the correlation of the number of firms with markups will be less negative than the true causal parameter.

On the other hand, an OLS estimate might suffer from omitted variables bias: variables related to demand, like consumer preferences or actual sales are unobserved, but might be correlated with both prices and market structure. Again, a positive correlation between demand and market structure and at the same time a positive correlation between sales and prices will lead to an upward bias of an OLS estimate.

In order to overcome both problems, we suggest an instrumentation strategy that can explain market structure but which is both unrelated with demand and has no direct influence on prices.

Instrumentation strategy: In the Geizhals.at data, we observe the complete life

²⁷For example, they control for the shops' expectations in best-seller products with a large number of sellers versus non-selling products which will be listed by only few shops

cycle of many products together with the firms' decisions to carry the products in the shop. Therefore, we use previous listing decisions as instrument to cope with the endogeneity of the number of firms offering a specific product at time t during the life cycle. There are two dimensions for the endogeneity problem: One dimension refers to the timing decision over the product life cycle. Does a shop list these products from the beginning or at a later point in time? For markets with brand names, part of the listing decisions can be explained by common patterns, such as an established supply relationship, a shop might have with a producer or a wholesale importer, or variations in the availability of the product. These patterns will be independent from the type of camera the shop may supply. Thus we use the timing of previous listing decisions of e-tailers for other brand products of our manufacturer as an instrument for current listing decisions. This is a strong instrument: statistically it does influence current listing decisions strongly. For our instrument to be valid, an exclusion restriction must hold: the listing decision of a series of different products in the past will have no direct impact on markups for another product in the future. This is a plausible assumption because we are using very different markets. We will discuss some threats to this identification strategy below.

Another dimension refers to the assortment the e-tailer offers. There might be shops, which specialize only on products that promise ex ante large markups (or variables correlated with the markup like sales). There might be some heterogeneity of the cameras in the market in terms of aspects as quality and design features that might be correlated with both markups and entry of firms. In our estimation, we will use product fixed effects to capture these unobserved features of the specific cameras. To cope with this dimension of endogeneity we control throughout the paper with product fixed effects for observable and non-observable product features which might have influence on the assortment decision. Given these product fixed effects on the first and second stages of the IV regressions potential selection effects by varying product assortments are controlled for. Therefore, in our analysis we are only exploiting markup variations over the individual product life cycle.

The instrumentation strategy is illustrated with an example in Figure 4.2. The figure shows the product life cycles for cameras A to H introduced at different points in time. In a first step we are only interested in the listing decision of a single e-tailer, which we shall call an E-shop for the sake of illustration. The vertical dashes indicate the listing decisions (either zero or one) on the third (tenth) day of product life cycle of camera D (camera F). Let us consider whether our E-shop will list a product D on the third day after introduction. This decision is represented by the encircled line on item D . We predict the probability of this event by the E-shop's general probability of listing a similar item that has been on the market for three days. We consider only the

last three items that have been introduced *before* product D has entered the market. We then calculate how many of those items were listed by the E-shop on the third day after they appeared. Taking the share gives us an estimate of E-shop's probability of listing product D on its third day of existence. In a second step, we aggregate these probabilities across shops to obtain the predictor of the number of shops that will offer item D on a given day.

This strategy can easily be extended for each day in the product life cycle, giving us a predicted market structure for each day of the product cycle. For the E-shop's listing decision of product F on the tenth day, for example, we use the respective decisions on the tenth day for products E , G , and H . To guarantee the validity of the instrument, we use only products that were introduced before the introduction of the camera in question. Note that for instrumenting F , we ignored products A through D because those cameras lay too far in the past. When calculating the instrument, we fixed the number of earlier introduced cameras to a constant number of three products.²⁸ In contrast to a constant time interval our approach of fixing a constant number of products guarantees valid standard errors that can be calculated without bootstrapping methods.

First-stage regressions: As we use the time patterns of previous listing decisions in completely different markets, our instrument should not have a direct causal implication for today's markups and price dispersion. Moreover, the listing of a particular type of camera on a particular day in the past should have no influence on sales today. Therefore, the instrument will comply with the necessary exclusion restrictions. Table 4.2 presents the first stage regressions and shows that the instrument is strong enough to explain the market's actual entry decisions, which are depicted by the number of firms at each point in the product life cycle. Columns (1) and (2) compare the contribution of the instrument to explaining the number of firms, and columns (3) and (4) show the contribution to its quadratic term. The instruments are strongly significant and have the expected sign. The marginal R^2 – due to the large number of fixed effects – amounts from 0.0016 up to 0.028 with F-values close to or above 400.

Could past listing decisions have a direct impact on markups of current products and thereby threaten our identification mechanism? Potential threats concern the fact that markets for past products might be close – either in calendar time or in the type of product – to the current product. Therefore, we tested variations of the instrument to see whether our findings are robust. We applied three systematic variations: i) We vary the number of previous products forming the baseline of previous listing decisions from 3 to 5 or 8. This changes also the length of time used for previous listings. ii) We

²⁸In the next subsection robustness checks on the first-stage regressions are discussed.

use the listing decisions of different brand names – instead of our brand name. iii) Our cameras can be roughly divided into several subsubcategories²⁹: simple digicams and SLR cameras. We introduce instruments which are only from the same subsubcategory as our product or only from the other subsubcategory. Both previous listings from other brand names as well as those from other subcategories should be seen as exerting even less influence on demand for the current product and, therefore, the exclusion restriction should be easier fulfilled. All these variations do not change our results much.³⁰

Caveats with respect to the instrumentations strategy are dynamic aspects of pricing decisions of firms: if there is a remaining correlation between the market structure of past products at a particular day in their life cycles and current pricing decisions of the firm on the same days of the life cycle, then our identification would fail and we would get biased IV coefficients.

4.4 Results

4.4.1 Market Structure and Market Performance

Tables 4.3 and 4.4 show our basic results for the interrelation of market structure and markups. These baseline specifications are parsimonious, as they consider only the number of firms on the market – either linearly or in quadratic terms – and the product life cycle. We also account for calendar time and product fixed effects. Columns 1 and 3 show OLS estimations, whereas in Columns 2 and 4, our instrumental variables approach is used.

Our results indicate a highly significant and relatively strong correlation of the number of firms with markups. Not accounting for the endogeneity of the number of firms and using OLS, we would estimate the effect of ten additional competitors in the market to reduce minimum markup by 0.55 percentage points and median markups by 0.23. The cheapest firm would react significantly more strongly than the median firm, which might be explained by the high frequency with which prices are changed in online markets, where the cheapest price is a focus of considerable attention from both consumers and firms.

If we instrument for the number of firms, we see a substantially larger negative effect: 10 additional retailers tend to reduce the markup of the cheapest firm by 1.24 percentage points and the markup of the median firm by 0.95 percentage points. These figures are large in economic terms considering the standard deviation of 57 firms in

²⁹The categorization has been done by geizhals.at.

³⁰Results are available from the authors.

our sample. As discussed above, OLS is likely to underestimate the true absolute effect of an additional firm on the markup, as it does not account for the fact that attractive items also attract more firms. Again, the cheapest firm reacts considerably stronger than the median firm.

In Columns 3 and 4 we use a quadratic specification of the number of firms: it turns out that there is a solid negative but decreasing influence of the number of retailers on markup, both for the cheapest and the median firm. There is virtually no turnaround given the maximum number of firms in our sample of 203. For the minimum markup, the negative influence of the number of firms ceases at 375, and that for the median markup with 195 firms.

Looking at the impact of the product cycle on markups, we observe that markups decrease with time in all regressions. In the 2SLS regressions markups decrease more slowly at product introduction, then more steeply as the product life cycle advances. This decline is stronger for minimum markups – relative to median markups. One reason for this phenomenon could be market saturation.

To investigate the impact of the number of sellers on price dispersion we concentrate on the coefficient of variation (Table 4.5). While the OLS regressions show a small negative relation between the number of firms and price dispersion, in the 2SLS results (Columns 2 and 4), we see a positive relationship. In the linear case, increasing the number of firms by 10 increases the coefficient of variation by 0.011. For the quadratic case (Column 4) we observe an even stronger increase in price dispersion.

Ignoring dynamic effects, the combined results on markups and price dispersion are compatible with the search theoretic model of Carlson and McAfee (1983) that accommodates two sources of heterogeneities by assuming a non-degenerate distribution of producers' marginal cost and heterogeneous visiting costs of consumers. In addition, the augmented search theoretic model by Baye and Morgan (2001), which features the firms' randomization over prices as a consequence of different user groups, are well in line with our findings. The other search theoretic models are, however, not in line with our findings of a decreasing median markup. Moreover, models of monopolistic competition predict a decreasing price dispersion, a hypothesis which is not supported by our data.

Baye et al. (2004) analyze price dispersion, using a very similar dataset to the one we use here. They focus on the relative price gap (the difference of the first and second price) and show that it decreases as a function of the number of firms, but not over time. Given this finding and a brief analysis of the average price they use a calibration to discriminate the predictions of several clearing house models. We build on their findings, but estimate both markups and price dispersion. Moreover, we analyze the lifecycle dynamics and we can instrument the number of firms based on the lifecycle of

previous models. Our OLS estimates corroborate their finding of a decreasing average price, since even the median markup is decreasing in the number of firms. The negative relationship is even stronger, when accounting for the endogeneity in the number of firms. When instrumenting for the number of firms, we do not find a statistically significant decreasing relationship between the number of firms and the coefficient of variation (price dispersion).³¹ Note however, that their baseline measure of dispersion is the price difference between the first two offers, whereas we focus on the coefficient of variation (a measure they used in robustness checks).

Shipping cost: While sellers are ranked at Geizhals.at according to prices *net* of shipping costs by default, the price ranking including shipping cost is only accessible via a detour. Figure C.1 in the appendix shows a screenshot of the price comparison site. As can be seen there, a quick and easy comparison of shipping costs alone is not possible, because shipping costs can be reported in different ways and there is no automatic ranking possible.³² As shipping costs are often used as part of an obfuscation strategy (Ellison and Ellison, 2009) it is interesting to see whether shipping costs react to market pressure as well. In Table 4.6 we report the effects of the market structure on shipping costs divided by median price.³³ As there are different shipping costs available, we concentrate on those mostly observed in the data: shipping costs to Germany when paying cash in advance. Interestingly, the IV patterns are largely the same as in Tables 4.3 and 4.4. While OLS predicts a positive relationship between shipping costs and the number of firms, Columns 2 and 4 reveal a robust negative relationship with a small and positive quadratic term.

It is remarkable that more competition seems to decrease also shipping costs. Our estimation shows that ten more firms actually decrease average shipping cost in that market by 62 cents (the mean of the shipping cost is 7.7 Euro). This market structure effect of shipping costs is economically significant, but smaller than the effects on markups: around a quarter of the effect on the median and one-eighth of the effect on minimum markups. These results confirm the visibility argument, that consumers have a much harder time comparing shipping costs than actual prices.³⁴ This raises interesting questions about whether and how different market structures may result in a different role for price transparency. As this is beyond the scope of the paper we left

³¹This finding persists even after introducing a click-weighted measure of price dispersion.

³²Different possibilities of shipping costs are e.g. standard shipping, shipping to Germany or Austria, and different shipping costs depending on the payment options.

³³In order to make the results comparable with the percentage values of the markup we are also using a percentage value for the shipping cost.

³⁴In a robustness check we computed the total markup including shipping costs to Germany when paying cash in advance. In line with the separate results for markups and shipping cost, we find an even higher effect of market structure on “gross markups”: both the minimum and median markups tends to decline significantly with the number of firms. The dispersion of “gross prices” increases.

it for further research.

4.4.2 Life Cycle Effects

In this section we investigate whether the profit-squeezing effect of a higher number of firms is the same in different phases of the product life cycle. Several authors claim that competition might be particularly important at the beginning of the life cycle of a product (e.g. Toivanen and Waterson (2005) and Berry (1992)). On the other hand, at the beginning of the product life cycle pioneer consumers might react less to prices and therefore a higher markup can be achieved. If they are uncertain about their preferences, the opposite may also hold (see Bergemann and Välimäki (2006a)).

To check for different effects of market structure on markups over the product cycle we extend our linear baseline model with crossterms. These crossterms interact the number of firms with four dummy variables for the life cycle of the product (Phase1: days 0-45, phase 2: days 46-105, phase 3: 106-225, phase 4: days 226-800). For ease of interpretation of the coefficients, in Figures 4.3 and 4.4 we plot these results for the minimum and median markup, respectively.

In these plots, each line represents a product of a certain age; we plot the curve for products right after their introduction and after 1, 2, 5, and 9 months on the market³⁵. Our plots show a consistent pattern. In Figure 4.3, we observe the pattern for minimum markups. Throughout the life cycle of the product, an increasing number of firms is associated with a fairly similar reduction in markups. The picture is similar for median markups in Figure 4.4, but here the reaction to an increasing number of firms tends to fall over the life cycle - the splines become flatter.

A simple Cournot model would predict that the markup is inversely related to the number of firms, which would lead to a flattening out reaction to increased competition. Our detailed analysis of competition effects over the life cycle shows a more nuanced pattern. For median markups, we do see some flattening out: after the first months of product introduction, median markup reacts still negative, but somewhat less to an increased number of firms. For minimum markups, this is not the case: regardless of the phase of the life cycle of the product, the reaction to more intense competition is the same. This may be due to the higher importance of minimum mark-ups (prices) for consumer demand in online price-comparison sites. Actual transactions are much more concentrated towards the lowest prices. An intuitive argumentation would be that in online price-comparison sites, where prices are very transparent, it does not make sense for newcomers to start with median prices. Only very low prices will catch the

³⁵The dots represent the median of the empirically observed distribution of the number of firms within each phase.

attention of customers. In addition online stores on Geizhals.at can follow the prices of their competitor over the platform. Although we do not know whether the retailers actually follow all their competitors, we are convinced that they know when they lost their leading position and to whom. An increased competition effect might be the results. In contrast to existing literature stressing that the first two or three entrants have the largest impact on markets prices (e. g. Bresnahan and Reiss (1991)) our empirical results show a different picture: enforced by the transparency of e-commerce markets additional firm entries in all phases of the product life cycle have the same effect on markups of the price-leader.

4.4.3 Substitutes over the Product Life Cycle

So far, we considered all markets for cameras independently, implicitly assuming no interrelations between cameras of different type which are offered at the same time and on the same platform. This simplification allowed us to describe the market in general terms. In this section, we enlarge our empirical model by allowing for substitute products. The availability of substitutes may offer an additional channel for competition in such a market; not considering it may seriously bias measured market structure or competition effects.

On the one hand, the availability of substitutes may drive down profits and markups as such; on the other hand, our measure of competition, the number of firms offering the *same* camera, may be misleading: the number of competitors offering a *similar* camera may be important as well. As the technology of these cameras is quickly improving over time, it is important to distinguish between substitutes with an older technology – which we define as products brought earlier on the market – and a newer technology, i.e. cameras which are introduced later. Moreover, the brand of the camera may be decisive: cameras from a rival producer may be stronger substitutes as compared to new cameras from the same producer. The former may target a new camera towards successful rival products, whereas in the latter case firms may fear cannibalization effects in the introduction of successor products and may, therefore, be more careful in the choice of design or timing of a new product introduction.

As there is no natural definition of substitutes, we identify substitutes by a conclusive behavior of searchers on the price comparison site. We will follow the general idea to identify and analyze different search spells (search cluster) for each user of the website geizhals.at. Each search spell should represent the customers' search and information process during the purchase of a specific product. We assume customers to consider all clicked products during the search spell as potential substitutes. The analysis of frequencies and the identification of most frequently clicked pairs of products

over all customers give us a statistical foundation for the identification of substitutes. See Appendix A for a description how we have calculated substitutes in our data.

Table 4.7 extends our 2SLS estimations for minimum markups³⁶ (compare the benchmark case from Table 4.4, Column 4) by controlling for available substitutes in its different forms. Note, that the number of substitutes is time varying as it counts the number of available substitutes during each day of the life cycle separately. As expected, the coefficient for the number of firms is still significantly negative, but a little bit smaller than in the simple specification. Moreover, the quadratic terms disappears. This may be due to the additional competition effect coming from the substitutes. As compared to the simple specification, the life cycle effect is completely unchanged.

In general, an additional substitute product reduces the markup of the cheapest firm by 0.77 percentage points. Note, that we have defined substitutes in a very narrow sense, with a mean of only 0.63. This markup-reducing effect of additional substitutes is substantial, its economic importance is difficult to judge, though. Bringing a new substitute to the market may open up a new field of competition, which is not easily compared with an additional number of firms: Many firms may offer this new substitute, but the new product does not operate at exactly the same market. To make the effect of substitutes comparable with the direct competition effect we can increase substitutes and number of firms each by 10%: increasing the number of firms by 10% will reduce minimum markup by 1.19 percentage points, whereas 10% more substitutes will reduce markup only by 0.04 percentage points.

Splitting up substitutes into older and younger ones we see our presumption confirmed that technologically more advanced (newer) products represent a larger threat to the minimum markup compared to older substitutes. Furthermore, substitutes of rival brands have a substantially larger negative outcome – minus 2.3 percentage points of the markup. On the contrary, same brand substitutes even have a positive association with the minimum markup. One explanation for this result might be found in the fact that we observe the retailers' and not the manufacturer's markup (although we would assume a high correlation between both). When introducing new products manufacturers apparently leave retailers larger margins if potential and good running older products from the same brand are still on the market. In that way manufacturers might convince retailers to better promote the new brand product with new technology or features.

A look at the most detailed level in Column (4) confirms this presumption. Newer substitutes from competitors have a larger negative effect on the markup than older

³⁶Similar results can be obtained for the median markup.

and more outdated products of rivals. Although we do not measure any effect of newer substitutes of the same brand – as manufacturers may understand their business not to cannibalize the old products’ rents – we measure a significant positive association with the markup if older same brand substitutes are still seen as potential substitutes by the customers.

4.4.4 Robustness

We perform several robustness checks. First, we test the robustness of the basic results by using varying definitions of price dispersion and by using other definitions of markups. Second, we account by weighting for the fact that some of the price offers may attract less attention from potential buyers. Finally, we consider characteristics of shop specialisation and quality.

We experiment with different definitions of price dispersion: apart from the coefficient of variation (the benchmark case from Table 4.5, Column 4) we use the standard deviation of prices and a coefficient of variation calculated in such a way that the prices are weighted with the number of clicks received. All these variations show a similar pattern: price dispersion increases with the number of firms, in some cases with a decreasing rate.³⁷

We investigate further whether our results are influenced by the fact that we treat all product offers symmetrically in our regressions. In particular, in questions of price dispersion researchers typically mistrust the validity of price offers that are much too high (cf. Baye et al. (2004)). This suggests weighing price offers with the number of clicks they receive to give low-ranked and perhaps less reliable price offers less weight. We do this in Table 4.8, which weighs each offer by how often it was clicked. This also implies that any offers that did not attract clicks by consumers do not enter this specification at all. Our main results are confirmed in this specification. In the case of the minimum markup weighting reduces significance of the linear term to some extent while the squared term gains importance.

Finally, we want to see whether our results are due to changes in the composition of the shops offering an item over the life cycle. In particular, the presence of larger shops, cheaper, or more reliable shops or a higher presence of shops that sell not only online but also have a brick and mortar outlet might affect the outcomes. Therefore, in Table 4.9 we include the composition of shops in the regression. In this table the first column shows the benchmark estimation (compare Table 4.4, Column 4). We then add the share of firms that have the item stocked (i.e., immediately available), the share of firms with low reputation (measured by customer feedback), the share of low-price

³⁷Results are available on request.

firms (firms offering generally lower prices in other markets), the share of large firms, and the share of shops with a brick and mortar facility. All of these shares are scaled on a range from 0 to 100: for example, if, in Table 4.9, the share of large firms increases by 10%, this is associated with a drop in median markups by 0.55 percentage points.

When we introduce these measures of market heterogeneity one by one, because they are to a large extent multicollinear, we find our general results completely unaffected. Both, minimum and median markups fall as the share of larger firms (Column 5) and the share of firms with low reputation (Column 3) increase. The sign of the statistically significant variables are reasonable: we would expect lower markups if there are more firms with low reputation and stronger competition in case of an increasing share of large firms in which undercutting of prices might have a larger impact. The decrease of markups is more pronounced for the size of the firms. The share of low-price firms is not correlated with the median markup (Column 4). Finally, the share of firms that have the item in stock (Column 2) and the share of firms also having a brick-and-mortar facility (Column 6) are related to an increase in markups; again, the positive signs confirm the expected price-setting behavior.³⁸

4.5 Conclusions

In this paper, we investigate the interaction between market structure and market performance in e-commerce using detailed data for digital cameras from an Austrian online price-comparison site. We analyze the empirical association of competition with the markup of the price leader and of the median firm. We account for potential endogenous timing decisions to list a specific product by using previous listing decisions as instruments and include product fixed effects to capture the products' unobserved quality and design features. We further investigate the relation of market structure and measures of price dispersion as well as the development of markups over the product cycle.

Our estimation results show a significant empirical association of markups and the number of retailers in the market. Median markups are lower by 0.23 percentage points and the minimum markup by 0.55 percentage points once ten additional competitors have entered the market. We also find that instrumenting is important for estimating the relation between competition and the markup and we see a substantially higher negative effect. With ten additional retailers, the markup of the median firm is reduced by 0.95 percentage points and the markup of the cheapest firm by 1.24 percentage

³⁸Analogous regressions for minimum markups show the same signs, except for the presence of low-price shops, having a positive effect.

points. Ignoring dynamic pricing effects, we may interpret our results as support for search theoretic models (Carlson and McAfee, 1983). They contradict models of monopolistic competition (Perloff and Salop, 1985).

Our results are also in line with the theoretical predictions of Baye and Morgan (2001) as well as the results of recent empirical papers by Haynes and Thompson (2008a) for the US online market for cameras and by Campbell and Hopenhayn (2005) for the US brick-and-mortar retail industry. In both cases, the competitive effects of an increasing number of firms persist in a homogenous goods market. Even with more than one hundred retailers we find markups still decreasing.

The analysis of markups over the product cycle further shows significantly lower markups the longer a product is on the market. Our results refer to e-tailing in the presence of a price-search engine with very narrowly defined products. In such a situation, consumers can very easily collect information about prices and seller reliability. Still, it takes a large number of sellers and a relatively long time for firm markups to dissipate. We may interpret this result in support of price dynamic models when consumers are uncertain about their tastes of a product newly emerged on the market and there is social learning (Bergemann and Välimäki, 2006a).

The markup of the price leader diminishes over the life cycle of the product. This allows us to compare the competitive effect of the number of firms to the effect of time: having one more firm in the market reduces the markup of the price leader by the same amount as three additional weeks in the product life cycle. In other words, by waiting three more weeks a consumer will get the same price reduction she would get if she went to a market with one additional firm, *ceteris paribus*. In reality, waiting longer will typically also increase the number of firms, thus increasing the advantage of waiting.

Finally, our results also indicate support for substitutability between newly innovated and old expiring technologies. The inclusion of potential substitutes in our estimations reveals interesting stylized facts. The amount of substitutes tends to reduce the firms' mark ups. We distinguish between older and younger substitutes as well as own brand or competitors' brand products. Newer substitutes by competitors are associated with larger reductions in markups compared to older substitutes by competitors. Whereas an increasing amount of older substitutes of the same brand leads to higher markups for the younger products,³⁹ we do not observe changing markups for the older substitutes if more new own brand products are introduced. Lacking testable theoretical hypothesis we do not account for the price setting game which

³⁹Manufacturers might, for example, incite retailers with higher markups on the new product if a high number of older own brand substitutes are on the market

might be involved in the listing behavior of online shops but just focus on the quantity of competing products.

Our results highlight the usefulness of this very specific market for consumer electronics, where product life cycles are particularly short and thus can be fully observed. Thus, analyses of such environments have great potential to shed light on phenomena of markups over the product life cycle, early adopters, and inter-temporal price discrimination.

Table 4.1: Summary statistics of the collapsed two-dimensional panel-data with the info on the level of goods and time

	count	mean	sd	min	p10	p90	max
Average price in EUR	15827	949	1399	99	153	2101	7864
Median price in EUR	15827	938	1387	98	151	2085	7990
Minimum price in EUR	15827	853	1294	78	127	1928	7085
Number of sellers	15827	104	58	1	11	168	203
Age in days	15827	166	110	1	35	328	444
Age at death of model in days	70	240	133	35	89	439	444
Wholesale price in EUR	15827	764	1114	79	123	1715	5801
1=item was clicked at least once	15827	.91	.29	0	1	1	1
Aggregate clicks at product i	15827	27	40	0	1	65	646
Average clicks per shop offering product i	15827	.33	.66	0	.0079	.76	21
Markup of price leader (percent)	15827	4.8	7.9	-26	-5	16	35
Median markup for product i in (percent)	15827	18	3	0	15	21	35
Markup of price leader incl. shipping cost in EUR	15445	7.8	6.8	-21	-16	16	36
Median markup incl. shipping cost* in EUR	15445	19	3.7	-3.8	15	23	36
Coefficient of variation of prices	15735	.087	.16	0	.047	.11	5
Standard deviation of prices in EUR	15735	67	170	0	14	130	6051
Coefficient of variation of prices incl. shipping cost*	15120	.09	.16	0	.048	.12	4.9
Absolute price gap between best and second price in EUR	15735	11	27	-0.000098	0	29	516
Average shipping cost in EUR	15445	7.7	1.4	0	6.2	9.2	24
Average reputation on a scale from 1.0 (best) to 5.0 (worst)	15744	1.7	.16	1.1	1.6	1.8	3.5
Average availability (1: in stock, 2: within 2-4 days)	15142	1.5	.15	1	1.4	1.7	2
Share of German shops	15827	.65	.11	0	.57	.74	1
Share of shops with brick and mortar facility	15806	.48	.11	0	.41	.57	1
Click-weighted markup of price leader (percent)	14339	4.9	8	-26	-4.8	15	44
Click-weighted median markup for product i (percent)	14339	9.1	7.3	-26	0	17	96
Click-weighted coefficient of variation of the prices	13577	.054	.091	0	.014	.087	4.1
Substitutes younger same	15827	.011	.1	0	0	0	1
Substitutes older same	15827	.44	1.4	0	0	1	8
Substitutes younger competitors	15827	.014	.12	0	0	0	1
Substitutes older competitors	15827	.087	.41	0	0	0	4
All substitutes	15827	.63	2.1	0	0	1	14

NOTES: The unit of observation is product i at time t (product-time panel). The time variable is days since market introduction.

Table 4.2: First stage regressions for instrumenting the number of firms

VARIABLES	(1) (# of firms/10)	(2) (# of firms/10)	(3) (# of firms/10) ²	(4) (# of firms/10) ²
Instrument for number of firms/10	0.19*** (0.012)	0.85*** (0.028)	1.31*** (0.241)	15.27*** (0.542)
Instrument for number of firms/10 squared		-0.05*** (0.002)		-0.99*** (0.035)
Age in months	1.96*** (0.072)	1.75*** (0.071)	34.59*** (1.404)	30.13*** (1.378)
Age in months squared	-0.12*** (0.002)	-0.11*** (0.002)	-2.36*** (0.037)	-2.17*** (0.037)
Constant	0.59 (0.451)	-0.70 (0.445)	-52.64*** (8.768)	-79.82*** (8.602)
Monthly Dummies	Yes	Yes	Yes	Yes
Product Dummies	Yes	Yes	Yes	Yes
Observations	15,893	15,893	15,893	15,893
<i>Marginal</i> R ²	0.0073	0.0283	0.0016	0.0276
Number of goods	70	70	70	70
F-test	420.7	417.6	388.1	392.5

NOTES: Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1; The table shows the first stage regressions for instrumenting the number of firms by the instrumental variable that is based on listing behavior of shops over the life-cycles in earlier product markets. Columns (1) and (2) show the first stage regressions for instrumenting the number of firms. Columns (3) and (4) show the corresponding estimation for instrumenting the *square* of the number of firms. Note that the coefficients are much larger in these columns, since the variance of the predicted variable is much larger after taking the square. The R² of the baseline regression without the instrument (not included in the table) amounts to 0.4867. The F-Statistics amount to 231.1 for testing Column (1) against the baseline model and to 400.7 for testing Column (4) against Column (3).

Table 4.3: Minimum Markup

	(1) OLS	(2) 2SLS	(3) OLS	(4) 2SLS
Number of firms/10	-0.548*** (0.012)	-1.243*** (0.110)	-0.768*** (0.030)	-1.366*** (0.159)
Number of firms/10 squared			0.012*** (0.002)	0.018** (0.009)
Age in months	-3.702*** (0.109)	-2.140*** (0.274)	-3.661*** (0.109)	-2.517*** (0.166)
Age in months squared	0.027*** (0.003)	-0.068*** (0.015)	0.028*** (0.003)	-0.040*** (0.008)
Constant	-0.017 (0.672)	1.416* (0.773)	0.963 (0.682)	2.430** (0.981)
Monthly Dummies	Yes	Yes	Yes	Yes
Product Dummies	Yes	Yes	Yes	Yes
Observations	15893	15893	15893	15893
Number of products	70	70	70	70
Adj. R ²	0.559		0.561	

NOTES: The unit of observations is the outcome of product i on day t . The first two columns show the results without a squared term, Columns C and D include the squared number of firms. Columns A and C show OLS panel regressions with product fixed effects. In columns B and D the number of firms has been instrumented. The dependent variable is shown above the columns. Standard errors in parentheses: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 4.4: Median Markup

	(1) OLS	(2) 2SLS	(3) OLS	(4) 2SLS
Number of firms/10	-0.232*** (0.007)	-0.951*** (0.076)	-0.596*** (0.018)	-1.159*** (0.097)
Number of firms/10 squared			0.021*** (0.001)	0.030*** (0.005)
Age in months	-2.008*** (0.065)	-0.392** (0.189)	-1.941*** (0.064)	-1.028*** (0.101)
Age in months squared	-0.005*** (0.002)	-0.103*** (0.011)	-0.004** (0.002)	-0.057*** (0.005)
Constant	8.544*** (0.398)	10.026*** (0.533)	10.164*** (0.398)	11.734*** (0.597)
Monthly Dummies	Yes	Yes	Yes	Yes
Product Dummies	Yes	Yes	Yes	Yes
Observations	15893	15893	15893	15893
Number of products	70	70	70	70
Adj. R ²	0.274		0.296	

NOTES: The unit of observations is the outcome of product i on day t . The first two columns show the results without a squared term, Columns C and D include the squared number of firms. Columns A and C show OLS panel regressions with product fixed effects. In columns B and D the number of firms has been instrumented. The dependent variable is shown above the columns. Standard errors in parentheses: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 4.5: Coefficient of Variation

	(1) OLS	(2) 2SLS	(3) OLS	(4) 2SLS
Number of firms/10	-0.003*** (0.000)	0.011*** (0.004)	-0.006*** (0.001)	0.016** (0.007)
Number of firms/10 squared			0.000** (0.000)	-0.001 (0.000)
Age in months	0.024*** (0.005)	-0.008 (0.010)	0.024*** (0.005)	0.002 (0.007)
Age in months squared	-0.000 (0.000)	0.002*** (0.001)	-0.000 (0.000)	0.001*** (0.000)
Constant	0.404*** (0.028)	0.372*** (0.030)	0.416*** (0.028)	0.338*** (0.041)
Monthly Dummies	Yes	Yes	Yes	Yes
Product Dummies	Yes	Yes	Yes	Yes
Observations	15801	15801	15801	15801
Number of products	70	70	70	70
Adj. R ²	0.042		0.043	

NOTES: The unit of observations is the outcome of product i on day t . The first two columns show the results without a squared term, Columns C and D include the squared number of firms. Columns A and C show OLS panel regressions with product fixed effects. In columns B and D the number of firms has been instrumented. The dependent variable is shown above the columns. Standard errors in parentheses: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 4.6: Shipping Cost

	(1) OLS	(2) 2SLS	(3) OLS	(4) 2SLS
Number of firms/10	0.009*** (0.002)	-0.104*** (0.015)	-0.065*** (0.004)	-0.240*** (0.023)
Number of firms/10 squared			0.004*** (0.000)	0.014*** (0.001)
Age in months	0.018 (0.014)	0.250*** (0.035)	0.025* (0.014)	0.019 (0.023)
Age in months squared	-0.001* (0.000)	-0.014*** (0.002)	-0.000 (0.000)	0.002** (0.001)
Constant	1.303*** (0.084)	1.583*** (0.104)	1.618*** (0.084)	2.383*** (0.129)
Monthly Dummies	Yes	Yes	Yes	Yes
Product Dummies	Yes	Yes	Yes	Yes
Observations	15441	15441	15441	15441
Number of products	70	70	70	70
Adj. R ²	0.207		0.227	

NOTES: The unit of observations is the outcome of product i on day t . The first two columns show the results without a squared term, Columns C and D include the squared number of firms. Columns A and C show OLS panel regressions with product fixed effects. In columns B and D the number of firms has been instrumented. The dependent variable is shown above the columns. Standard errors in parentheses: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 4.7: The Importance of Substitutes over the Life Cycle

	Minimum Markup			
	(1)	(2)	(3)	(4)
	all	older vs. newer	own brand vs. other brands	all interactions
Number of firms/10	-1.143*** (0.147)	-1.102*** (0.151)	-0.972*** (0.149)	-0.922*** (0.154)
Number of firms/10 squared	0.005 (0.008)	0.003 (0.008)	-0.005 (0.008)	-0.009 (0.009)
Age in months	-2.562*** (0.163)	-2.532*** (0.165)	-2.566*** (0.164)	-2.523*** (0.165)
Age in months squared	-0.042*** (0.008)	-0.043*** (0.008)	-0.042*** (0.008)	-0.044*** (0.008)
All substitutes	-0.765*** (0.112)			
Substitutes newer		-1.496*** (0.279)		
Substitutes older		-0.618*** (0.121)		
Substitutes same brand			0.383** (0.173)	
Substitutes other brands			-2.276*** (0.236)	
Substitutes newer same brand				0.077 (0.502)
Substitutes newer other brands				-3.912*** (0.578)
Substitutes older other brands				-2.038*** (0.253)
Substitutes older same brand				0.512*** (0.196)
Constant	1.783* (0.951)	1.436 (0.971)	0.722 (0.965)	0.438 (0.998)
Monthly Dummies	Yes	Yes	Yes	Yes
Product Dummies	Yes	Yes	Yes	Yes
Observations	15827	15827	15827	15827
Number of products	70	70	70	70

NOTES: The table is based on the main results in the paper, but includes the number of substitutes for the product. The unit of observations is the outcome of product i on day t . The first columns shows the results when including all substitutes. Columns B differentiates between newer and older substitutes. Column C shows the results for distinguishing same brand substitutes from competitors' substitutes. Column D distinguishes the substitutes along both dimensions. In all columns the number of firms has been instrumented. The dependent variable is the minimum markup. Standard errors in parentheses: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 4.8: Markup and price dispersion weighted by clicks.

	(1) clw. min. markup	(2) clw. med. markup	(3) clw. coeff. var.
Number of firms/10	-0.228 (0.241)	0.758*** (0.284)	0.019*** (0.006)
Number of firms/10 squared	-0.030** (0.012)	-0.052*** (0.014)	-0.000* (0.000)
Age in months	-3.104*** (0.192)	-3.545*** (0.227)	-0.016*** (0.004)
Age in months squared	-0.015* (0.009)	0.014 (0.010)	0.001*** (0.000)
Constant	-0.639 (1.276)	-0.848 (1.508)	0.035 (0.029)
Monthly Dummies	Yes	Yes	Yes
Product Dummies	Yes	Yes	Yes
Observations	14401	14401	13639
Number of products	70	70	70

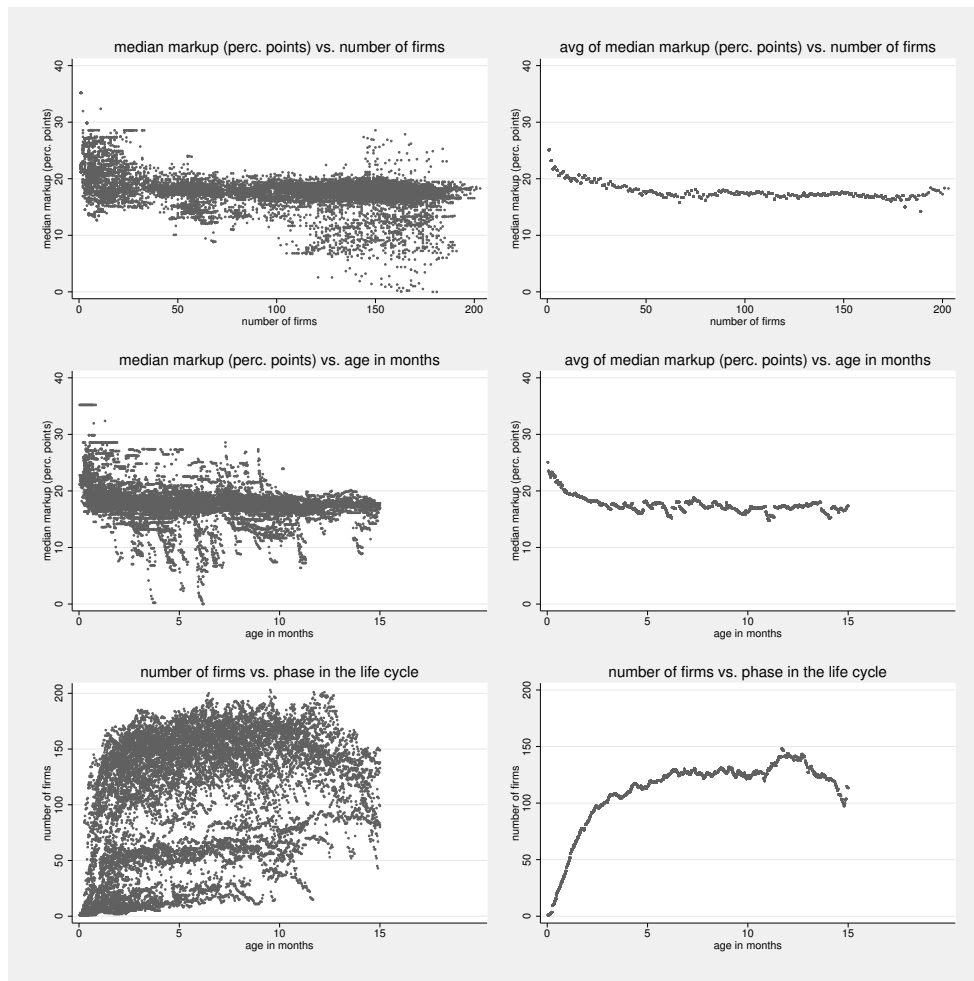
NOTES: The table weighs prices by clicks on the respective product before computing the moments of the price distribution to reproduce the main results in the paper. The unit of observations is the outcome of product i on day t . Each column shows the same estimation with a different dependent Variable. The dependent variable is shown above the columns. The number of firms has been instrumented. Standard errors in parentheses: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 4.9: Median markup and the composition of shops

	(1) benchmark	(2) availability	(3) reputation	(4) price level	(5) size	(6) brick/mortar
Number of firms/10	-1.159*** (0.097)	-1.323*** (0.136)	-1.128*** (0.097)	-1.146*** (0.119)	-1.197*** (0.092)	-1.128*** (0.105)
Number of firms/10 squared	0.030*** (0.005)	0.032*** (0.006)	0.028*** (0.005)	0.030*** (0.005)	0.038*** (0.005)	0.028*** (0.006)
Age in months	-1.028*** (0.101)	-0.977*** (0.112)	-1.022*** (0.101)	-1.043*** (0.115)	-1.197*** (0.095)	-1.007*** (0.099)
Age in months squared	-0.057*** (0.005)	-0.063*** (0.006)	-0.058*** (0.005)	-0.056*** (0.005)	-0.044*** (0.004)	-0.059*** (0.005)
Share on stock		0.034*** (0.008)				
Share low rep			-0.010*** (0.002)			
Share low price				-0.004 (0.008)		
Share larger shops on					-0.055*** (0.003)	
Share brick/mortar						0.011*** (0.003)
Constant	11.734*** (0.597)	10.883*** (0.539)	12.090*** (0.599)	11.909*** (0.468)	14.190*** (0.547)	11.096*** (0.729)
Monthly Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Product Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	15893	15893	15893	15893	15893	15872
Number of products	70	70	70	70	70	70

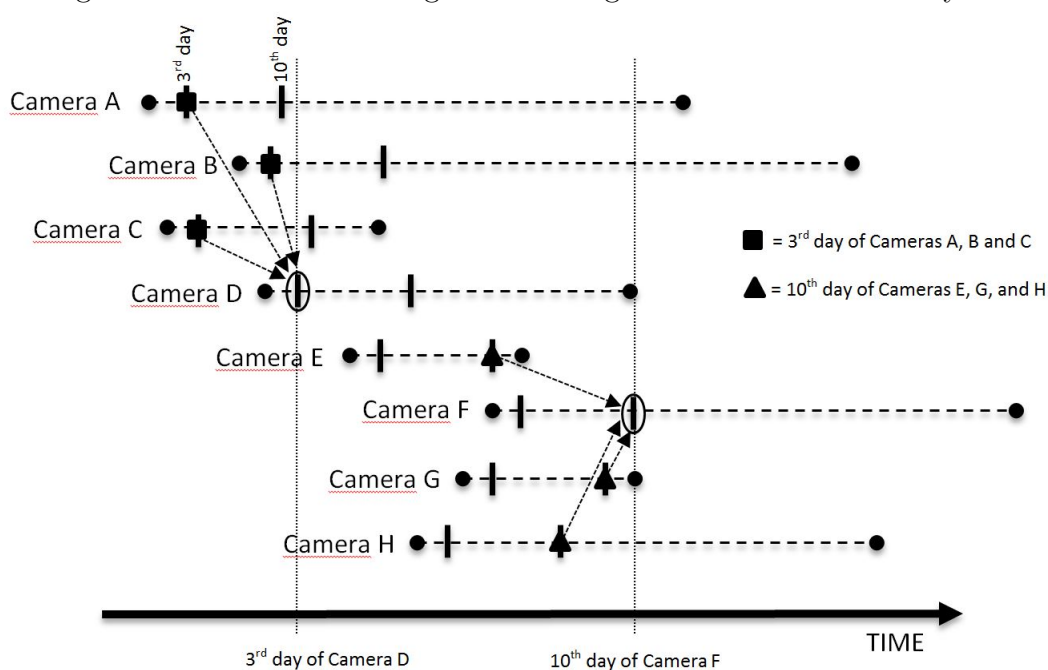
NOTES: The unit of observations is product i on day t . Each column shows the same estimation with a different variable capturing shop composition. Other than in the previous tables, the dependent variable in all columns is the median markup, as it is more relevant in the context of shop composition. The number of firms has been instrumented. Standard errors in parentheses: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure 4.1: Median markup plotted against the number of firms and age of product



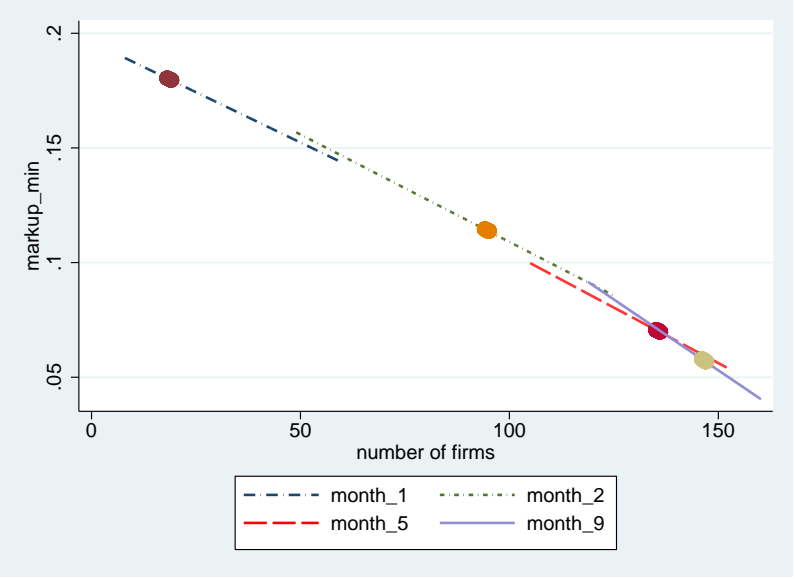
NOTES: The plot shows the empirically observed distributions of the number of firms, age and median markup plotted against each other. In the top left panel, the $median\ markup_{it}$ is scattered against the $number\ of\ firms_{it}$ in the corresponding market and the right column shows the corresponding averages. In the middle row the $median\ markup_{it}$ is plotted against the *age of the product*. In the lower row, the *number of firms* is plotted against the *age of the product* (in months).

Figure 4.2: Instrument using firm's listing behavior in earlier lifecycles



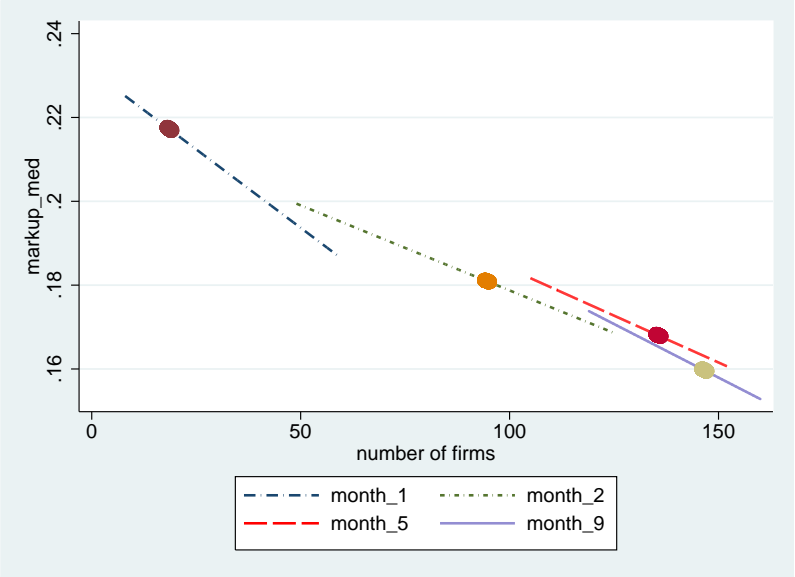
NOTES: If we want to predict how many shops will list a product on any given day q after introduction we use a shop's general probability of listing one of the three items that entered the market before product j , q days after they were introduced. Examples: to predict listing behavior for camera D on day 3 (encircled dash), we would use information on the three cameras A , B , and C on their respective third days of existence (black squares). However, we would not use the information from the cameras that saw light after D was introduced. To predict how many shops listed camera F on day 10 (encircled dash), we would use the information from the three cameras E , G , H which saw light before F was introduced. Cameras A , B , C (on day 10) are ignored for the computations of the instrument for camera F on day 10. These considerations can be applied to each day in the product lifecycle of a camera.

Figure 4.3: Minimum markup in different phases of the product life cycle



NOTES: Each spline shows the estimated relationship of number of firms and markup at a different point in time (after 1,2,5 and 9 months). The curves are plotted on the range from the 33rd to 67th percentile and the dots represent the median of the empirically observed distribution of the number of firms at the point in time it corresponds to.

Figure 4.4: Median markup in different phases of the product life cycle



NOTES: Each spline shows the estimated relationship of number of firms and markup at a different point in time (after 1,2,5 and 9 months). The curves are plotted on the range from the 33rd to 67th percentile and the dots represent the median of the empirically observed distribution of the number of firms at the point in time it corresponds to.

Chapter 5

99 Cent: Price Points in E-Commerce*

5.1 Introduction

This paper analyzes the role of price points or focal prices (used synonymously) in the sellers' price setting behavior of e-commerce markets and the consumers' purchasing decisions. Price points are defined as values with special price endings that are frequently used, i.e., ending in zero (also referred to as “even prices”) and 9-ending prices (“just-below prices” or “odd prices”). Early explanations for this phenomenon rely on the customers' perceptions of these price points, i.e., particular price endings may convey a particular image of a product (image effect). Examples for these prices are 1 €100.00 or 6 €.00. With these even prices firms might signal that the high quality of their products does not make it necessary for the selling firm to engage in a Bertrand competition with declining prices. In a more recent study, Basu (2006) explains the price-setting of odd prices as a rational strategy of oligopolists in a retail market. In his model, consumers disregard the right-most digits of the price, due to the cost of processing very detailed information. Hence, in equilibrium, it might be rational for firms to set 9-ending prices. By doing so, the firms can increase their prices somewhat and escape the zero-profit forecast of the Bertrand equilibrium in a retail market (level effect). Examples for these prices are 9 €.99 or also 1 €.999,00.

*This chapter was coauthored with Franz Hackl (University of Linz) and Rudolf Winter-Ebmer (University of Linz and IHS, Vienna). A paper with the same title was published in *Information Economics and Policy* (2014, Vol. 26, pp. 12 - 27.) We thank the editor and two anonymous referees for valuable suggestions. We are grateful to Martin Peitz, Christine Zulehner, two referees and seminar participants at Istanbul (EARIE), Milano (Bocconi), Madrid (CEMFI), Essen (RWI), Hannover, Heidelberg, Linz, and Vienna for helpful comments. We thank the Austrian National Bank's Jubilee Fund for supporting the research (grant 12444) and geizhals.at for giving us access to the data and providing useful advice.

From a managerial perspective price setting is an important entrepreneurial task. This is especially true for the area of e-commerce in which the success of retailing depends on the firms' behavior in Bertrand competition. Taking the right decisions when it comes to setting prices is a crucial factor for a retailer's economic success. But what are the right prices for maximizing a firm's profits? Do customers react fully rational on the lowest price or do we observe some kind of bounded rationality (such as left-to-right price-processing in which the cents get irrelevant). Do some price endings (price points) have more impact in the consumers' demand than others? In a setting of Bertrand competition it is equally important to know how other firms (re)act and how other firms set their prices: can we observe that certain price points are undercut more often than other ones? This paper sheds light on the role that price points play for the behavior of firms and customers and thereby provides important managerial implications.

Whereas previous research analyzes these price points in offline markets (scanner data from supermarkets and real estate markets), we focus on online markets. The lack of studies in emerging e-commerce markets is surprising given that the digital revolution might change our understanding of the pricing process of firms. One might argue that the lack of comprehensive studies in online markets can be explained by the less pronounced problems of price comparison: less cognitive difficulties to memorize and compare products of different retailers. In particular, at a price comparison site, all price offers for a specific product can be seen by one mouse-click. On the other hand, price dispersion in such markets is still high Baye et al. (2006). We also observe service differentiation between e-tailers so that the idea of strictly ascending price listings loses importance. We will show that, although less than in brick-and-mortar stores, we do see a considerable proportion of odd as well as even prices in online shopping as well.

Most of the studies in offline markets analyze price points either on the demand side in the form of field or laboratory experiments or on the supply side in the form of price rigidity analysis. The focus and innovation of our paper lies in the consistent and comprehensive analysis of both sides of an online market. A careful analysis of the pricing strategies with price points has, on the one hand, to consider the price-setting decisions of firms, and, on the other hand, regularities in consumer demand. In order to draw correct conclusions, both market sides should be analyzed within a single set of data and it should be checked whether the reactions of both sides of the market are consistent. If the price points turn out to be equilibrium outcomes, they should prevail longer as comparable non-focal prices. The fact that we are the first to observe an entire market place enables us not only to look at the price stickiness for prices set by random firms but also to see the prices set by price leaders (i.e., the retailer that offers the good at the lowest price).

We use comprehensive data from Austria's largest price comparison site to explore theories on price points. While most previous studies consider only buying intentions, small samples, and a limited number of products, our data has the advantage in that we can observe the entire online market place with all competing shops. We use the price information on 23,317 products posted by 698 sellers together with the information about referral clicks and last-click-throughs, which are typically used as demand indicators in e-tailing. As many products are more expensive, we not only concentrate on 99 cent endings but also on prices ending in 9 euros, which will carry the same flavor of the argument.

In the beginning, the Internet was seen as the embodiment of perfect competition with instantaneous and comprehensive information of customers leading to fierce price competition, dwindling product differentiation, and vanishing brand loyalty Brynjolfsson and Smith (2000). In such an environment of perfect information, special price patterns like 99-cent endings are not expected to be important for the consumers' decisions. In contrast to this presumption, we show that price points are also prevalent in e-commerce markets. Our empirical results are generally supportive of Basu's theory: consumers are prone to disregard ordinary cent endings in their shopping behavior; price points are, therefore, more stable if they end with 9; and the best-prices ending in 9 or 99 are not changed by the price-setting firms and are less likely to be underbid by the rivals of the shop. Moreover, we observe higher price jumps for prices ending in 9. On the other hand, there is some role for image effects in the perception of consumers as well.

5.2 Literature

5.2.1 Theories for price points

Researchers have focused mainly on two consumer-oriented explanations for the phenomenon of price points² Stiving and Winer (1997):

(i) The first approach has been called the level effect, left digit effect, left to right processing (Thomas and Morwitz (2009), Basu (2006), Thomas and Morwitz (2005)). The basic idea is the assumption that the consumers use a heuristic to calculate,

²There is a wealth of operational or ad-hoc explanations. The most famous example is the anecdotal account Stiving and Winer (1997) of shop owners, who posted prices that would force the clerks to give back some small change in order to force their staff to register the transactions rather than pocket the money. Among other ad-hoc theories, Monroe (1973) mentions (and refutes) views that the number 9 might be considered a magical number with special properties. Clustering has also been considered as a tool to maintain tacit collusion as has been shown in Christie and Schultz (1994) and Christie et al. (1994).

compare, and memorize prizes due to their limited brain-capacity to process prices exactly: they read prices from the left, and in particular, they disregard cent prices. As a result of this boundedly rational behavior, the consumers tend to overestimate the gap between prices differing only by a small amount, if the lower price has a smaller left digit (e.g., €3.00 vs. €2.99).

This theory has only recently been formalized in a Bertrand-Equilibrium model Basu (2006). The paper assumes boundedly rational consumers, who do not bother to take into account what's after the comma. Instead, they "guess" that it is the average of all the last digits of the products in the market. This setting is used to analyze the market equilibria in a Bertrand setting where the firms can post prices but cannot affect quantities.

The model is relevant for two reasons. First, the mechanism attacks the Bertrand paradox. Since the consumers ignore the cent ending, undercutting a 9-ending price by less than a full euro does not generate additional turnover; on the other hand, it reduces the profit margin. As a consequence, the firms who want to undercut will undercut by a full euro; undercutting with small amounts does not make sense. Second, Basu's model generates an equilibrium in 99-prices, which is perfectly rational on the part of the firms and results in positive profits for the competing firms.

Basu's (2006) model predicts that if a market equilibrium results in focal prices ending in 99 cents, other firms might be less willing to undercut this price point. Let's assume marginal costs of €10.00. Is a price of, say, €11.47 equally stable than a price of €10.99? In the former case, Bertrand behavior of firms will lead to some downward pressure below €11.00; not so in the latter case due to left-to-right price-processing. Furthermore, even though the theory would not predict price-endings other than 99 in equilibrium,³ it predicts similar demand for such items as long as they are priced with the same euro digit; in other words: cent prices should not have an effect in an estimation of a demand function. Given this theoretical equilibrium, we predict the prices that end in 9 to be maintained longer than any other price-endings.

A different explanation for the level effect is given by Monroe (1973). The hypothesis postulates that consumers, when they plan to purchase an item, have already formed an expectation of the price they are going to pay, a so-called target price.⁴ If such a target price is memorized with an even number, it would typically be beneficial to set prices below these even-numbered thresholds. Gedenk and Sattler

³Basu (2006) not only establishes an equilibrium in 99 cent prices but also a further result with two equilibria, one with 99 cent, and the other at marginal cost.

⁴In a similar version of the story, consumers might have a binding budget set, because they have only one or two cash bills with them and while a price just below this threshold lies within their budget set, a price just above does not.

(1999) argue that it might be rational for profit-maximizing firms to set 9 ending prices when facing some uncertainty about the type of customers: there might be some consumers who show a discontinuous reduction in demand if a price passes a certain (even) threshold (type 1) and others who do not (type 2). In case the firm raises its price above this threshold, the firm might gain only little by charging a higher price from type 2 types, but lose a lot of revenue from type 1 consumers. As these explanations are usually seen observationally equivalent to the boundedly rational behavior from above, we do not discriminate between the two explanations any longer.

(ii) The other important consumer-based approach is called the image effect (Stiving and Winer (1997), Thomas and Morwitz (2009), Schindler (1991)). Price setters may use the pattern of a price to transmit specific information about the item that is being offered. An example would be a clothing company that uses 00c endings in regular times and 99 endings for items on sale. If this was true and a 9-ending communicated a good bargain, we would also expect price clustering, higher demand, and maybe some rigidity on 9-ending prices. Yet, not all the prices are predicted to end in 9 and a similar pattern might emerge for prices ending in 00 (or any other special number), if zeros were to transmit positive properties such as product quality.⁵

It should be noted that the “image effect” predicts higher demand at focal prices, whereas the “level effect” predicts that consumers do not react to the differences in the cent digits with their demand behavior. Moreover, the “image effect” might be prevalent not only in 99-ending prices, but also in even prices. In terms of price stability, Basu (2006) predicts that 99-ending prices are more stable; in particular, they are not underbid by the rivals. As for the “image effect” theory the model has no clear prediction for the equilibrium outcome: image prices could be less stable, because rivals might, for instance, want to destroy the image of a cheap price by simply underbidding it; on the other hand, they might be more stable if the additional demand at the image price is so high that any changes are not worthwhile.

5.2.2 Empirical evidence

As for the demand effects of focal prices, there exists a broad experimental literature, going back to Ginzberg (1936), but only a few studies look at real markets and actual demand on a larger scale. In one larger study on single products, Stiving and Winer (1997) use scanner data on yoghurt and tuna and find a large and generally positive relationship between a 9-ending price and sales. Moreover, they find that consumers do not process prices holistically. Aalto-Setälä (2005) analyzes the utilization of special

⁵See also Palmon et al. (2004) for a survey.

price points before and after the currency change to the euro in Finland and finds a slow adjustment process of price endings over two years. Sehity et al. (2005) were using the introduction of the euro to show that the retailers all across Europe quickly converged to focal prices again, which had been overturned on January 1, 2002. The field experiments by Schindler and Kibarian (1996) and Anderson and Simester (2003) observe the demand for items in a mail-order catalogue, where the prices were manipulated to show 9-endings. Generally, these experiments find an overall positive effect on demand. Yet, the significance of the effect depends on a variety of other factors, such as how much information was available on a product and whether the item has been introduced only recently. Laboratory experiments, in general, have to rely on purchase intentions instead of actual demand.⁶ An example of such an experiment is Gendall et al. (1997), suggesting that 9-ending prices affect purchase intentions positively.⁷ Lacetera et al. (2011) find substantial evidence for the tendency to focus on the left-most digit of odometer values in the purchasing decisions of car buyers.

Studies looking at price setting and rigidity in an online environment are highly infrequent.⁸ Levy et al. (2011) find evidence that the prices ending in 9 are sticky in the sense that the price setters are more reluctant to change them. Once these prices are changed, though, the ultimate price jump is larger what might be due to a larger adjustment need.

Only a few studies differentiate explicitly between the level (left-digit effect) and image effects. The aforementioned study by Stiving and Winer (1997) finds that the consumers treat the pre- and post-comma digits in a different manner. However, apart from this study, the evidence on the image effect stems from lab experiments and surveys. In an intention-to-purchase experiment, Liang and Kanetkar (2006) obtain similar results as Stiving and Winer (1997). However, both studies report that even pricing plays a role, which they tend to take as evidence for the image effect. Using the experimental evidence from Italian consumers, Guido and Peluso (2004) find evidence for even target prices. Moreover, when analyzing how prices are memorized, they find that consumers recall odd prices smaller than they really are, especially when the left-

⁶There are exceptions to this rule, however, such as the lab study by Thomas and Morwitz (2005), who measure dependent variables beyond purchase intentions.

⁷Liang and Kanetkar (2006) provide an extensive review of the existing literature on price endings and discuss the literature on numerical processing and memory-effects of odd prices; Kauffman and Lee (2005) review the issues of price rigidity in e-commerce in general.

⁸Macroeconomists look at the price points or reference prices when they inquire the existence or prevalence of nominal price rigidities; e.g., Eichenbaum et al. (2011), Levy et al. (2011), and Konieczny and Rumler (2006). The 9-ending prices can act as price points and might contribute to the price stickiness in general.

most digit is manipulated in the experiment.⁹ Thomas and Morwitz (2005) finally conclude that the scale of the level effect depends on which digit to the left is affected, which is an indication that the left-digit-effect plays a role.

Specific evidence for the image effect is provided by e.g. Naipaul and Parsa (2001), who compare restaurant menus and find that menu makers use even prices to suggest higher quality and just-below prices to indicate bargains. Lynn et al. (2013) use data from prices which are set by consumers themselves – pay-what-you-want prices, tips in restaurants and amounts in self-pumped gasoline purchases – to argue that consumers do have a preference for even numbers.

5.3 Data and prevalence of price points

For our empirical analysis, we use the database of <http://www.geizhals.at>. This website is a price search engine collecting the price offers via standardized protocols from retailers and presents them electronically via its website. Due to the broad Austrian market penetration of Geizhals.at, this price search engine practically covers the whole Austrian online market including suppliers from other countries (especially Germany) that are interested in the Austrian e-commerce business.

We use an inflow sample of *all* price spells starting in an arbitrary week in 2007. A price spell is defined as the offer from a specific firm j for a specific good i at a specific price p_{ij} . Each spell has a starting and ending time so that we can exactly measure the spell durations in seconds. Price spells end because firms change their price or stop offering the product. We can observe all price spells between Monday, June 4, 2007, 03:00:00 to Monday, June 11, 2007, 02:59:59. We used these data to compute aggregate product specific or shop specific variables such as the relative price of a firm's price spell.¹⁰ Considering all prices in this step is necessary to avoid a sampling bias in favor of volatile prices. Next, we moved on to consider only offers that were newly quoted in the given week for the analysis, that is we focus on the inflow sample of prices. This is particularly important for the duration analysis, since looking at the stock sample would introduce a sampling bias in favor of long lived prices.¹¹ Additionally, we know for each and every price spell, the respective referral requests

⁹See also Guéguen and Legohérel (2004).

¹⁰The relative price is calculated as the offers' price relative to the week's mean price of all price offers on geizhals.at. Hence, for these aggregate data not only the spells of the inflow dataset are used.

¹¹Note, that by construction a specific offer of a firm with highly volatile prices might be represented several times in the data set. A reduction of the data set in a way that only one price spell per firm and product remains in the sample does, however, not change our empirical results concerning the duration analysis and the estimated effects of price points on consumers' demand.

from the customers. The referral requests are customer clicks on the firm's product offer at the Geizhals.at webpage resulting in a forward from the website of the price search engine to the online shop of an e-tailer. We normalize the referral requests to clicks per week to cope with the different offer durations. As very cheap products are not expected to be bought online due to relatively high shipping cost, we consider only those product offers whose average price is larger than €25. In 2007, purchases below that threshold were still rather atypical in Austrian e-commerce and thus very likely to occur as a part of a bundle. In total, we have 805,949 price spells for 698 e-tailers and 23,317 products. On average, we measure 24.8 referral requests per day and product. Due to the multi-faceted structure of the data, we have control variables at three different levels. More detailed descriptives on the variables of interest can be found in Table 5.1.

(i) The e-tailer specific variables that are constant for the product offers of a specific web-shop j are the (a) country of origin, (b) a dummy variable, indicating whether the firm provides a pick up facility for customers who want to avoid the cost and time it takes to ship an item, and (c) all quality reviews of previous customers. The reviews consist of a free text section and a standardized grading scheme (five-grade scale: 1 denotes best, 5 denotes worst). We use this information to construct the average grade as a variable that captures the perceived service quality of shops. (d) Moreover, we can control for the number of customer evaluations.

(ii) Furthermore, we have information on three product-specific controls. (a) The number of competing shops measures how many shops were active in market i , and is thus a measure of competition. (b) Another control is the activity in the market that is generated by counting the number of price changes that were observed during the week of observation to control for the general turbulence in market i . (c) The third control is the quality indicator for the product: customers can recommend the product on the Geizhals.at website for purchase or avoidance. From these valuations, we know the percentage of positive recommendations for the product.

(iii) Apart from price, referral clicks and the time stamps (the start and end times) of a price spell, we observe the following offer-specific (ij) information: (a) shipping cost and (b) whether the product can be shipped immediately. If some control variables (mostly availability, shipping-cost, pick-up facility, and product reviews) were not available, we imputed those variables at the mean and used the imputed variables together with a missing flag for imputations.

Even though the referral clicks are available in the data, the actual act of purchasing a product is unknown, because actual purchases happen at the e-tailer's own web site. This is unobservable for Geizhals.at, and thus, for us. Therefore, following Smith and Brynjolfsson (2001), we use the concept of last-click-through (LCT) as a

proxy for the purchasing decision. If a customer is searching for a product, he/she might meander around different websites, comparing the characteristics of the shops, but will finally settle for the preferred shop and buy from there. The last click to a shop selling the product is usually identified as the click with the highest purchase probability. We construct the LCT from the referral clicks using a procedure based on hierarchical clustering and Grubbs' test for outlier detection.¹² As LCTs have, by definition, a higher purchase probability, the comparison of normal clicks with LCTs enables us to make predictions of the impact of price points on actual demand.

Figure 5.1 shows the frequency distribution of price-endings to check for the prevalence of price points.¹³ The four graphs depict the distribution of price endings in different samples of interest: (a) all prices, (b) offers that were clicked at least once, (c) price-leading offers, and (d) best-selling offers (defined as most frequently clicked offers with a market share greater than 10%).¹⁴

Note that in any of the four samples, special price endings are relatively more frequent. Given that we observe 100 different cent-endings, a uniform distribution would imply a frequency of 1 percent of the observations for each digit. Clearly the 00c-, 99c-, 90c-, and 50c-endings are more frequent than the other price endings. This pattern is even more pronounced, when we restrict our attention to price-leading or best-selling offers. More than 15% of the online best-price quotes end with 00c. Almost 4% of the price-leading offers (fourfold the expected frequency) were quoted in 99c and another 5% were quoted in 90c. Taken together, the four focal price endings make up more than 35% of the bestselling price-endings.

This is not surprising given the fact that about 30% of all price offers end with a 9 at the unit position or at the ten's place (e.g., 9€990,00).

It might be surprising that the percentage for price quotes ending with 00c is relatively high. As the average price in our sample is more than €400, many prices can be expected to end by even numbers. In this case, a typical focal price will be a

¹²In practice, the construction of the LCT is a lengthy procedure because a time span for observation as well a product span has to be chosen. See Dulleck et al. (2011a) for details. Here, we define LCTs on the product level where a time interval of one week initiates a new search period. Additional results for the other definitions of LCTs are available upon request.

¹³We focus on the endings of prices as they are quoted (without shipping cost), since this is the main view for consumers and also the sole criterion for sorting offers on geizhals.at. A series of tests including gross prices (sum of price and shipping cost) did not change the estimates for the coefficients of specific endings of the net prices. Looking at the endings of the gross prices we found no effect at all. We therefore conclude, that consumers do not perceive specific endings that result from the sum of price and shipping cost.

¹⁴Market share was defined as the number of referral requests for offer ij relative to the total number of clicks on product j .

price like €599.00 or €739.00.¹⁵ Hence, even if the prices end with 00c, the left-digit effect might apply. Figure 5.2 shows the distribution of the unit position for prices ending with 00c. The four graphs show the same four groups of interest we used above. It seems that in all groups, a cumulation of prices ending in 9 can be observed, and as before, the pattern is even more pronounced for the clicked and best-selling offers. E.g., in the sample of best-price offers, almost a third of the offers end with nine euro, zero cent.

Table 5.2 shows how focal prices vary across different types of goods. In the first two columns the offers for the given products are separated according to the relative product price. The lowest price offers for the given products are listed in the lowest tercile, the distribution of focal prices for the highest price offers can be seen in the highest tercile. In column four and five products are distinguished according to the mean price of all offers per product in different terciles. The columns list the shares of the focal prices for the cheap (lowest tercile) and the expensive (highest tercile) products. In the last column the distribution of focal prices for the total sample is shown. The comparison of the different terciles reveals that, on average, price points can be observed more often for higher-priced offers as well as more expensive products. Moreover, endings in zero cent (including €9.00) are relatively more common for high-price offers and expensive products whereas the relative importance of price endings in nine increases for low-price offers and cheap products.

5.4 Price rigidity and firm behavior

In this section, we analyze the relationship between price endings and the duration of price-quotes.¹⁶ If the equilibrium is on 99-ending prices, then the maximizing firms have a low inducement to change these price points. On the other hand, if a price ending is non-focal, e-tailers should have an incentive to switch to a nearby price point to increase demand. Hence, the price point theory predicts that shops maintain offers at price points significantly longer than non-focal prices.

We use a duration analysis to compare the survival time of offers with price points to the other offers' survival time. First, we look at *individual* firm behavior as it is measured by the time span that elapses until a firm changes its *own* price. We model the hazard of ending an individual spell at duration t using a semi-parametric Cox

¹⁵Given such high prices, it would stretch the model a bit too far to expect the firms to bid prices up to €599.99.

¹⁶Kauffman and Lee (2005) discuss the hypotheses about price rigidity in e-tailing relating to market concentration, product quality, and size of the market. Kashyap (1995) gives a comprehensive overview of the price stickiness in retail (catalogs).

model that allows for a fully flexible baseline hazard $h_{j0}(t)$ and adds other variables proportionally (j is the firm index and i is the product index):

$$h(t|j, D_{ij}, \theta_{ij}, x_i) = h_{j0}(t) \exp(\delta D_{ij} + \beta x_i + \gamma \theta_{ij}).$$

We use two different definitions for price points. In the first definition, we use the most frequent price points as dummy variables: cent prices ending in 99c (1=ends in 99c), 90c (1=ends in 90c), and 50c (1=ends in 50c). For prices with 00c we distinguish those with a nine before the comma (1=ends in 9.00) from those with other digits (1=ends in \neg 9.00) because the former should be considered as 9-ending prices. Note, that the marketing literature defines odd prices as the practice of pricing just below the nearest round number, where “nearest” is left somewhat ambiguous and context or level specific Holdershaw et al. (1997).

To convey the flavor of odd and even prices, we use a second, more encompassing, definition. We set the dummy *9-clustered* to 1 if the price is either ending with 9 neglecting the right-most zeros (e.g., €576.90, €39.00, €590.00) or has at least one nine in the last four digits including the cent (e.g., €91.81, €899.11, €59.92). This definition for *9-clustered* prices combines differing variants of 9-ending as an additional check. It is broader and covers also prices that at first sight might not be considered as focal prices: obviously a price of 1€9.98 sends out a certain signal — and most observers would consider it as a kind of “odd” price. While we present this combination to catch also such special prices, we are aware, that the concept is not so clear-cut as the version with the individual focal price endings. The advantage of the definition is that it is quite generally capturing the notion of just below pricing. Moreover, this encompassing concept of *9-clustered* prices may lead to a potential downward bias because the concept may be too broad.

The dummy *even price (without 9)* takes the value 1 whenever a price’s ending is 00c without a nine being present in the four right-most digits (e.g., €345.00, €100.00). Additionally, we add controls at the offer level and the product level:¹⁷ θ_{ij} captures four offer-specific controls, which include relative shipping cost, immediate availability (1 if the product is immediately available), clicks, and price. x_i captures three product-invariant characteristics such as the number of competing shops, activity in the market in form of the number of price changes, and an indicator for recommendations of the product by consumers.

¹⁷Since we later stratify with respect to shops, there is no need to add shop-specific controls.

5.4.1 Pooled analysis

Table 5.3 (column 1-3) presents the results from a pooled analysis where the baseline hazard $h_{j0}(t)$ is assumed to be equal for all firms (i.e., $h_{j0}(t) = h_0(t) \forall j$). Standard errors are clustered at the product level. The base regression in column (1) uses no additional control variables, and thus, corresponds to estimating simple Kaplan-Meier survivor functions for the five different price points: €0.99, €0.90, €0.50, €9.00, and €-9.00. All offers with price points have a lower hazard: they last longer. The effect is strongest for the odd price ending with €9.00 followed by that ending with €-9.00 and €0.99. We see the lowest impact for €0.50. A coefficient of -0.764 for €9.00 corresponds to a reduction of 53.33% in the hazard rate. Likewise, a coefficient of -0.239 for the 50c dummy translates to a reduction of 21.18%. When a non-focal price ending has a 50% chance of surviving 54 hours, an offer that is priced in 50c has a 58% chance to make it to the same point in time. A 90c offer's odds to survive 54 hours or more are 72%.

When adding control variables in column (2), we see that our results basically remain unchanged: both the numerical values as well as the ranking of coefficients for our price points are fairly stable. Although there are no strict theoretical predictions for our control variables, the coefficients fit into a reasonable picture. More expensive articles and offers with higher shipping cost tend to have longer offers. The fact that offers for products that are immediately available hold longer is consistent with obfuscation strategy: e-tailers that do not have the product in stock might use short-run bargain offers in order to attract the customers' attention. At a later point in time, when the product can be delivered immediately, they switch to a higher price. The more clicks an offer generates, the lower is the incentive to change the price. The same applies for a low number of average price changes from all other firms on the market, which is an indicator of the market's intensity of competition.

Column (3) uses the alternative specification of price points. Again, the price points have significant and substantial longer durations than other prices: here, the highest effect is for *even price (without 9)*, with a somewhat smaller effect for the *9-clustered* dummy.

5.4.2 Stratified analysis

It may be that particular shops have specific pricing strategies: one shop is changing prices routinely each week, while another might change them every day. In order to deal with these differences, we use a stratified analysis, allowing for firm-specific baseline hazard rates. Our identification of the effect of price points stems only from the within-firm variation in the duration of the price offers. As in any fixed-effects

model, we expect lower coefficients in this stratified analysis.

In columns (4-6) of Table 5.3, we report the results from the stratified estimation, again including product-specific clustered error terms. While, as expected, our numerical coefficients are smaller as compared to the pooled analysis, all the previous patterns are reinforced. The focal prices have a longer average duration and the effect is the strongest for the prices ending in €9.00 in column (5) and *even price (without 9)* prices in column (6). Again, the weakest effect is found for the price quotes ending in 50c. This stratified analysis shows that our price point effects cannot be explained by firm-specific pricing policies: even within the same shop, offers with price points are kept longer than other comparable offers.

5.4.3 Price rigidity for price leaders

Up until now, we looked at the temporal stability of *any* price offers. To investigate the market outcomes or equilibrium outcomes, the equilibrium price should be studied. As there is always a significant amount of price dispersion in online markets, we take the offer with the lowest price, the “price-leading” offer, as the equilibrium price. We thus draw a sub-sample of offers, which held the lowest price at one point in time. The time for which a price-leading offer is valid can be reduced in different ways: either another firm actively undercuts this price or the price-leading firm itself discontinues the offer by going out of the market or charging a higher price. For an analysis of the competitive actions in such a market, undercutting by rivals is the decisive feature. Basu (2006) would argue that an equilibrium at 99-cent prices would not be undercut by a rival because it is not profitable to do so.

The reduced sample consists of 41,764 offers and 14,933 incidents, where a price-leading offer was undercut by an opponent. Our analysis uses a competing risk Cox model stratified by firm in order to allow for two outcomes: underbidding of opponents and own price changes. Table 5.4 shows the results. Columns (1-3) show the top offers that ended by undercutting while columns (4-6) show the top offers that were withdrawn or changed by the price-leader itself. We use the same specifications as in Table 3. Our main interest is in the undercutting part. The focal prices are less likely to be undercut, the effects for the prices ending in €9.00 and €0.99 (columns 1 and 2) are the strongest, and the effects for even prices are smaller (€-9.00) or insignificant (€0.50). When we use our alternative specification (column 3), we see a similar pattern: a strong effect for odd prices and a weaker effect for even prices (*even price (without 9)*). The results for own changes in columns (4-6) are less clear: only prices ending in 90c live longer before they are withdrawn by the firms.

The controls in Table 5.4 basically show a consistent picture. The higher the

number of clicks while in lead, the higher is the probability that the offer is undercut by a competitor. The more active a market is (measured with the number of price changes of all other firms), the shorter are the best-price spells. In markets with a higher number of competing shops, we observe that the shops change their prices more frequently.

5.4.4 Price jumps after focal prices

If the price points, like the 99c prices, are a possible equilibrium outcome and are more sticky, price changes, once they occur, should be larger Kashyap (1995). Levy et al. (2011) report higher price jumps for the price points in both brick and mortar stores as well as for the Internet prices. As we can observe the complete market we can improve the strategy for online markets by observing only the price jumps of the price leaders. This is of special importance in the online business in which the price search engines provide higher levels of transparency, and therefore, price leading offers describe the equilibrium phenomena better by simply considering any price quotes in the market. We test the hypothesis of higher price jumps for focal prices using all downward price changes, where we can also distinguish between undercutting by other firms and own price reductions. If the best price is undercut by another firm, the average price jump is minus 27c, and if the own firm is reducing the price, the reduction is smaller, amounting only to 7c on average.

Table 5.5 shows the results using a log change of the best price as the dependent variable in a product-fixed effects panel regression. In the first two columns for undercutting by rivals, we find very strong price-point effects, in particular, for price endings €0.99 and €9.00, but also somewhat smaller effects for price endings €0.90, €0.50, and €-9.00. If we, in turn, use our shortcut for suspicious 9 prices, *9-clustered* (column 2), we see significantly larger price jumps, whereas for even prices (*even price (without 9)*), there is no effect.

For own price reductions in columns (3-4) the effects are generally smaller, but they do confirm the theory that 99c prices are an equilibrium. We do find significant positive effects only for price endings €0.99 and €9.00; all other price points are insignificant. For *9-clustered* price endings (column 4), contrary to undercutting, we do not find significant effects, which might be due to the broader and more hazy concept of focal prices that is used here.

A similar picture can be shown for upward changes in the bestprice caused by the price leader raising its price or withdrawing the offer (in this case either the same or another firm will have the new lowest price for the product). Considering these upward jumps we again observe that focal prices are generally more stable. Price jumps (in

the upward direction) are larger if the previous best price was either ending in 99, 9.00 or in zero Cents.

5.5 Price points and consumer behavior

Finally, we look at the impact of price points on consumer behavior. Our demand estimation in Table 5.6 looks at the consumers' clicks (q_{ij}) on the website www.geizhals.at. A consumer click is a referral request to retailer j for product i .

$$q_{ij} = a_0 + bD_{ij} + a_1 \text{rel. price}_{ij} + cX_{ij} + a_2 \text{product}_i + \epsilon_{ij}.$$

As the duration of the price offers varies, we standardize the consumer clicks as the number of clicks per week. The vector D_{ij} includes various dummy variables for the price points, like odd or even prices. The variables of interest are the indicators for the respective price endings €0.99, €0.90, €0.50, €9.00, €−9.00, *9-clustered*, and *even price (without 9)*.

Additionally, to directly test Basu (2006), we artificially split up the price into its euro digits (multiplied by 100) and the remaining cent digits. Fully rational consumers would pay the same attention to the changes in both the euro and the cent digits, provided the change is the same; the coefficients for the two variables should be equal and negative. Basu (2006), on the other hand, argues that the consumers do not consider the cent digits in their demand, and hence, its coefficient should be zero.

The variable *rel. price* measures the price of product i of retailer j relative to the average price of product i over all retailers ($\text{rel. price}_{ij} = \frac{p_{ij}}{\sum_{j=1}^N p_{ij}/N}$). The additional control variables in vector X include *rel. shipping cost*, *German shop* (equal to 1 if the online shop is located in Germany, and 0 otherwise), *immed. availability* (equal to 1 if the product is deliverable at a short notice), *reviews (grade)* that measures the service and reliability evaluation of a shop j by the costumers on a scale from 1 to 5 (very good to very bad), and *reviews (num.)* that counts the number of customers who have given an evaluation of the retailers' service characteristics. As we are merging the markets for different products, we include the product-fixed effects in the estimation. We are using the price offers for a week in June 2007. As different price offers are valid for varying time periods, we calculate a standardized number of clicks per week as our dependent variable q_{ij} .

The price setting may be endogenous to demand. Therefore, we instrument the relative price of the product by the mean relative price a firm has in all markets except in the markets where the products belong to the same subsubcategory as the product in question.¹⁸ The identifying assumption is that the price setting is not-related

¹⁸Geizhals.at maps products hierarchically into subsubcategories, subcategories, and categories describing the substitutional relationship between the products. As an example, the category "Video/Photo/TV" contains the subcategory "TV sets" and the subsubcategory "30-39 inch LCD TV sets". In total, 358 subsubcategories and 40 subcategories are given.

markets should not influence demand in our model.¹⁹ In the first stage regression, the instrument has a strong and significant impact on the prices resulting in a marginal R^2 of 0.093. The exclusion restriction is particularly convincing because the customers in a price-comparison site cannot look at the complete price list for all products in the other subsubcategories of one particular firm. As the organization of the website allows only price comparisons for one particular product across firms, the consumers have a very hard time to get an image of the overall pricing behavior of one particular firm.

Table 5.6 looks at the influence of the price points on demand using customer clicks. Columns (1-3) of Table 5.6 use all referral requests as demand indicators, whereas columns (4-6) employ only LCTs that are considered as better indications for real purchasing decisions. Both indicators for demand give fairly similar results. As expected, those on LCTs are numerically considerably smaller.

There is no clear indication that the usage of price points (column 1) boosts demand. The prices ending with 99c are unremarkable. Whereas the prices ending with 90c or 50c attract somewhat less clicks, those ending with 00c, and in particular, those ending with 9 euro, attract more clicks. These results may be due to our sample that primarily consists of higher-priced items: setting prices ending with 9 euros seems to boost demand. The regression results for the LCTs in column (4) confirm this picture, albeit on a lower scale.²⁰ The pattern is generally confirmed when we use our definitions for the odd and even prices in column (2): we have somewhat higher demand for *9-clustered* price endings, but no effects for even prices that do not contain any nines. The relatively small effect of *9-clustered* prices vanishes once we introduce LCTs in column (5).

Our test for left-to-right processing also brings interesting results: contrary to a fully rational consumer, our shoppers behave apparently only boundedly rational. The euro part of the price has a strong negative effect on demand, whereas the effect of the cent is either positive for the number of clicks or zero for LCTs.²¹ This is a direct affirmation of Basu's (2006) hypothesis of consumers disregarding the cent, which can

¹⁹Kaiser and Minjae (2009) use the same instrumentation strategy for the demand for consumer magazines.

²⁰Note that the LCTs indicate, by definition, a higher purchase probability. Therefore, the comparison of normal clicks with LCTs allows us a prediction of the impact of price points on actual demand.

²¹We do not use the instrumented relative price in columns (3) and (6) as we did in the other regressions. Note, that we cannot instrument the absolute euro value, as we want to compare the coefficients for the euro value and the cent value directly: perfectly rational consumers should indicate a significant hundredfold coefficient for the euro value as compared to the cent value. We cannot use our instrument for the absolute euro value since it is not possible to meaningfully instrument the absolute price across the huge variety of completely different products. However, the comparison of IV-regression with simple OLS regressions does not show significant differences that might question our results in columns (3) and (6) of Table 5.6.

also explain the heterogenous results concerning the impact of cent endings on demand. We do not find convincing arguments supporting the theories on the image effect of price points (note, for instance, the negative impact of 50c on the demand).

For more expensive products, not only the cent but also the unit digits should be irrelevant for the consumer's decision. Therefore, we repeat this exercise for the products with prices above €100 for which we shifted our test procedure one digit to the left: a price of €234.67 is thus split up into a "relevant variable" with a value of 23.000 and an "irrelevant variable" with value 0.467. Again, the rational consumers would show the same coefficients for both variables. For all clicks, we find coefficients of -0.028^{***} (0.002) for the first and $+0.376^{***}$ (0.081) for the second variable; for the LCTs, we find -0.002^{***} (0.0002) and $+0.034^{***}$ (0.010) (there are no other substantial changes for the other variables). Although the consumers are price sensitive at the relevant price digits (the relevant variable), the positive significant coefficient for the "irrelevant variable" shows that the shoppers tend to choose price endings with high digits (9s). These patterns with the *non-negative* signs for the "irrelevant variable" again strongly confirm Basu's bounded rationality hypothesis.

The additional control variables are in line with what is to be expected in online markets. The relative price of the shop decreases demand.²² The positive influence of the higher shipping cost on market demand can be seen as evidence for obfuscation strategies, where the online shops compensate their lower prices with higher extra charges.²³ E-tailers with immediately available offers and better and more customer evaluations attract more customers. The positive sign of the Germany dummy points out that this highly successful Austrian price search engine is increasingly used by German users.²⁴

To test the robustness of our results and also to better understand which forces drive them, we replicated the specification in our estimation of demand (Table 5.6) for different subgroups of consumers, different types of products or different types of shops.²⁵ The results of these experiments are shown in Table 5.7. In the first two columns of panel A we distinguished extensive searchers, who spend a lot of time on the site and look at many offers and brief searchers, who look at only a few offers, before either leaving the site or making their purchase. In columns (3) and (4) of the

²²This effect is not statistically significant in the LCT model, which might be due to the instrumentation strategy. We also estimated the model with OLS, and there, the coefficients for the relative price are both significant and somewhat higher than the ones presented in Table 5.6.

²³Hossain and Morgan (2006) show that the buyers are inattentive to the shipping costs in eBay auctions as well.

²⁴In November 2008, 72 percent of the online shops were located in Germany and 62 percent of the clicks were from German IP addresses.

²⁵We thank the anonymous referees for valuable suggestions for these sample splits.

Table we distinguished between expensive and cheap products. Columns (5) and (6) juxtapose frequent users of geizhals.at and users who request the price comparison site only sporadically. In the last two columns, finally, we compare the results when we use only “pure play” webshops with those from a sample where we use shops having both a web-store and a brick and mortar shop.²⁶

For brevity, panel A shows only the coefficients for the price points. Considering the interpretation of the coefficients, one has to take into account that between some subgroups the mean of the dependent variable is different in size. Between intense and brief searchers – as well as between frequent and infrequent searchers – there are not many differences, except that the infrequent and the brief searchers buy more at 99-ending prices. In the case of cheaper products prices ending in 50 or 90 cents attract less demand, whereas in the case of more expensive products, even prices attract more demand. Apparently, the importance of images varies. Finally, price points are perceived fairly different in “pure play” webshops as compared to combined web and brick and mortar stores: When a brick and mortar store exists as well, consumers reward typical odd prices – those ending in 99 cents or 9 euros; whereas consumers shopping in a web-only store reward even euro endings and dislike cent endings, like 90 or 50 cents.

In panel B we test whether consumers attach different weight to euro vs. cent digits – and whether these weights differ across subgroups. The results overwhelmingly confirm the hypothesis that consumers read prices from the left to the right and attach less weight to cent digits. While the euro digits are consistently negative, the cent digits are mostly positive and in many cases significantly so. While this significantly positive coefficient may be due to some image effects, it indicates, indeed, that consumers have a peculiar awareness of prices: if the difference of a higher price lies only in the cent endings, it does not reduce demand.

5.6 Interpretation and conclusions

In this paper, we present the first comprehensive and consistent analysis of price points on the supply and demand side of a market. We analyze the issue in the context of e-commerce, an environment that is, a priori, not very favorable to price points. We can show that price points (special prices ending in 9 or 0) are non-the-less prevalent in e-commerce. Our results are consistent with Basu (2006) who assumes boundedly rational shoppers ignoring the rightmost digits due to limited processing capacity.

²⁶We define “pure play” webshops as shops that do not have a real world shop and click and mortar shops as stores which also have offline facilities, where consumers can inform themselves and make their purchases.

Profit-maximizing firms will adapt to this bounded rational behavior of consumers by setting prices ending in 9.

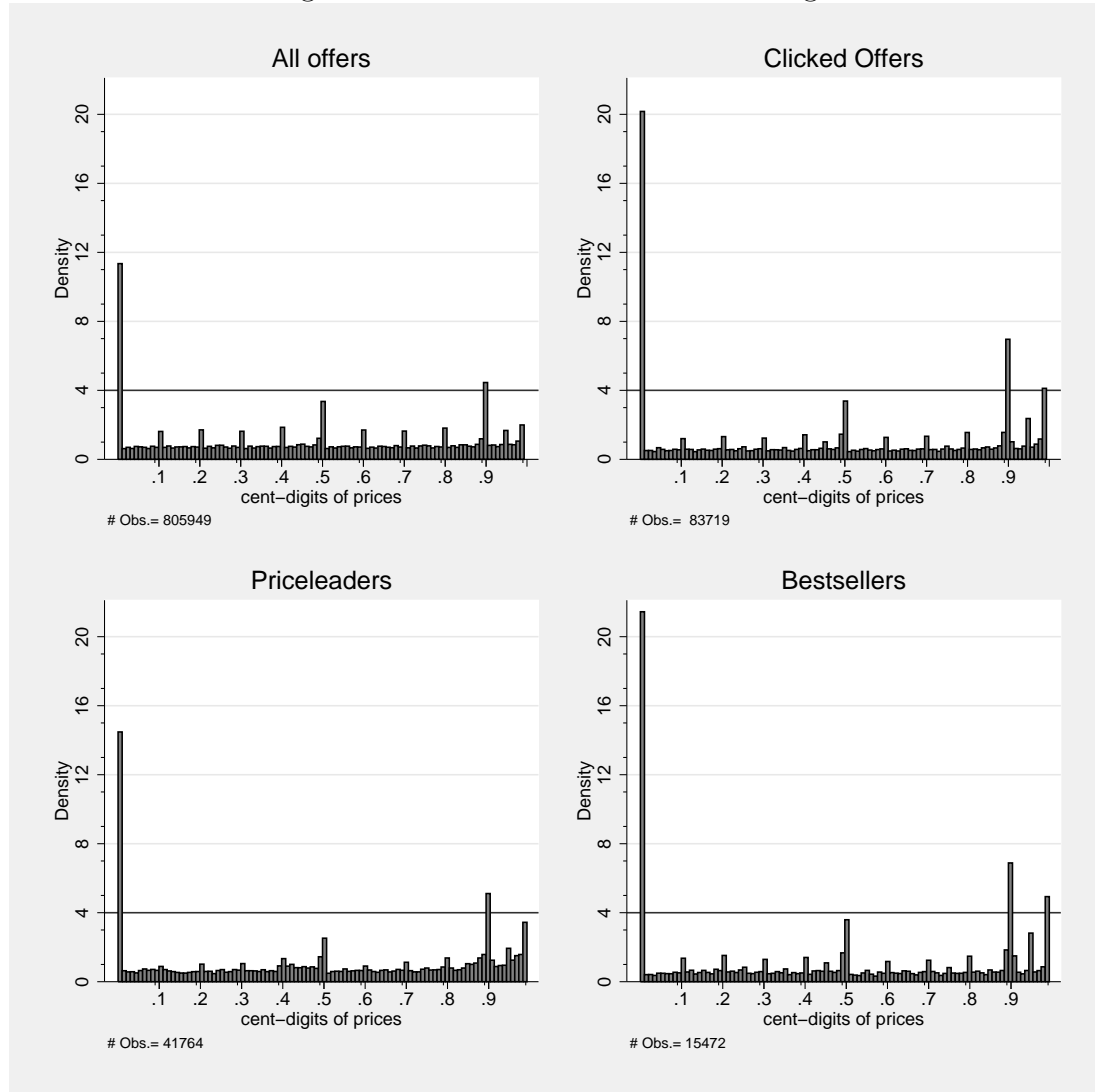
We found in our data covering the Austrian e-commerce market that the prices ending in 9 show typical equilibrium characteristics: they are more sticky than regular or even prices. This is particularly so for best-price offers: when these offers end with 9, they are less likely to be undercut by the rivals, and are also less often changed by the firm itself as compared to non-focal prices.

Although there is no clear-cut consistent impact of price points on the consumers' demand we find that the market does not react to the differences in the cent digits, as Basu (2006) is assuming. From a more general perspective, our results are more in line with Basu's level effect explanation for price points rather than the image effect theories even though we cannot rule out the existence of image effects for certain price patterns.

From a managerial perspective, price setting is one of the most important tasks. It turns out that even in highly competitive and transparent settings, like e-commerce with price-comparison sites, price points seem to matter. Which prices should firms set to maximize profits? We found clear-cut empirical evidence that consumers do display some form of left-to-right price-processing. Setting 9-ending prices might therefore be profitable: the price might lie below consumers' even-numbered thresholds for the maximal willingness to pay and competitors might leave such a best prices unchallenged a little longer. Managerial implications are less clear when it comes to other focal prices for which theoretical predictions and empirical results reflect ambiguous evidence: While it seems that consumers reward firms which set prices at zero cents, they punish those with prices ending in 90 or 50 cents. Image theory, however, is still too unspecific to give a clear reason for this.

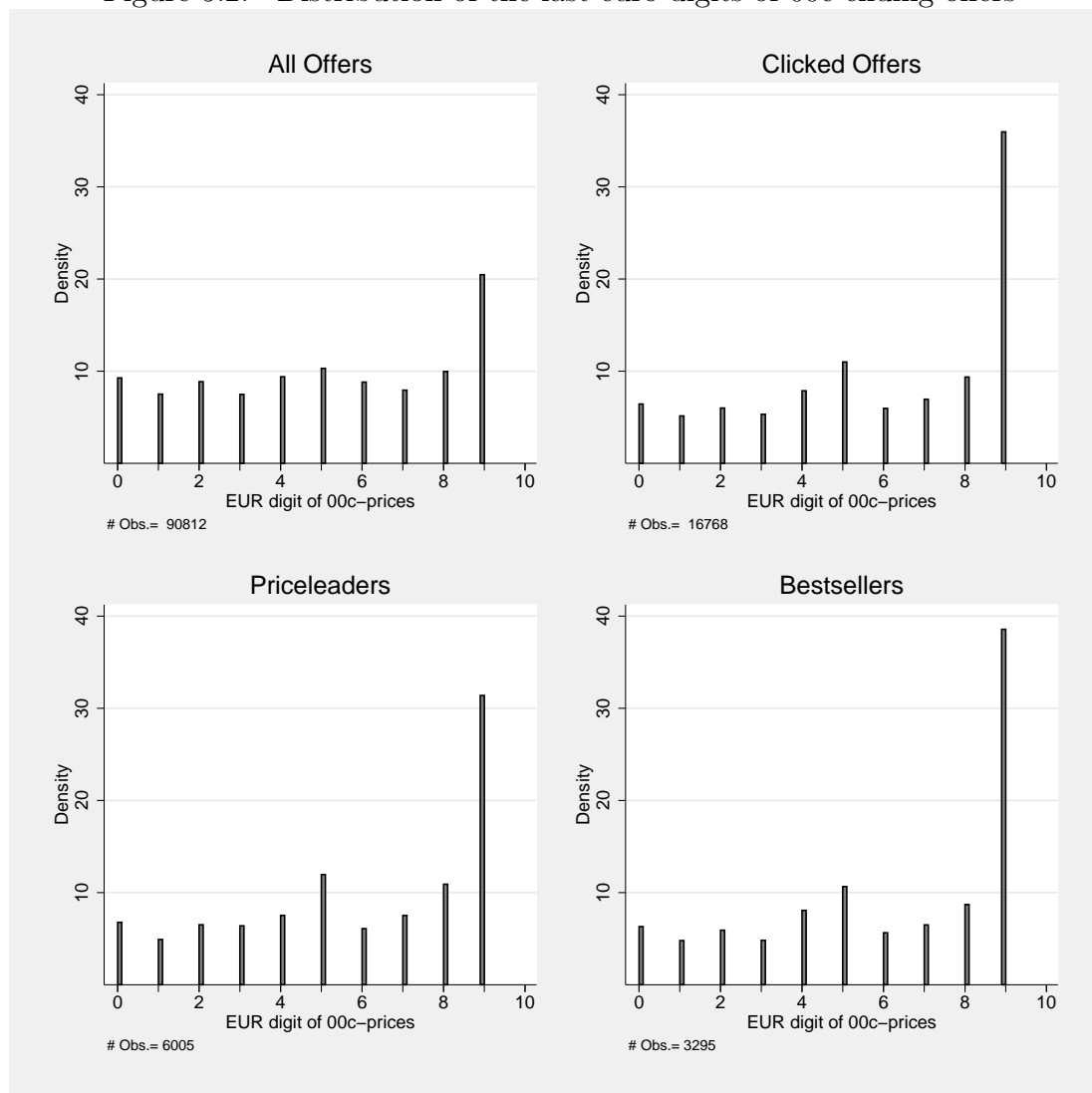
5.7 Tables and Graphs

Figure 5.1: Distribution of the cent digits



NOTES: No. of observations: 805,949 (83,719 clicked offers; 41,764 price-leaders; 15,472 best-sellers)

Figure 5.2: Distribution of the last euro digits of 00c-ending offers



NOTES: No. of observations: 90,812 (16,768 clicked offers; 6,005 price-leaders; 3,295 best-sellers)

Table 5.1: Descriptive statistics

Variable	count	mean	sd	min	max	Description
CENT ENDINGS (DEFINITION 1)						
1=ends in 99c	805949	0.02		0.00	1.00	dummy: price ending in 99c
1=ends in 90c	805949	0.04		0.00	1.00	dummy: price ending in 90c
1=ends in 50c	805949	0.03		0.00	1.00	dummy: price ending in 50c
1=ends in 9.00	805949	0.02		0.00	1.00	dummy: price ending in €9.00
1=ends in -9.00	805949	0.09		0.00	1.00	dummy: price ending in €-9.00
CENT ENDINGS (DEFINITION 2)						
9-clustered	805949	0.36		0.00	1.00	dummy: at least one of the last four digits on 9
even price (no 9)	805949	0.11		0.00	1.00	dummy: for even ending and no 9 in the last four digits
VARS FOR E-TAILER J						
reviews (grade)	805949	1.70	0.63	1.00	5.00	avg. customer evaluation of shop (1 best, 5 worst)
reviews (num.)	805949	187.03	227.79	0.00	1850.00	number of evaluations for shop i
German shop	805949	0.61		0.00	1.00	dummy: shop located in Germany (not Austria)
VARS FOR PRODUCT I						
number of shops	805949	90.29	44.03	1.00	257.00	number of shops listing product i
price changes (excl. i)	805949	2.00	0.76	0.00	10.00	number of price changes on item i (in 100s excl. j's own)
recommendations	805949	0.68	0.21	0.00	1.00	share of reviewers who recommended the item
LCTs on product	805949	10.61	31.94	0.00	857.00	total LCTs on product i (unweighted)
clicks on product	805949	173.40	881.08	0.16	25684.10	clicks per day and product on offer i
unweighted clicks on product	805949	118.85	323.74	1.00	8312.00	total clicks on product i (unweighted)
VARS FOR OFFER IJ ^{a)}						
euro digits	805949	404.86	899.08	0.00	98369.00	euro digit of the price ^{c)}
cent digits	805949	0.47	0.32	0.00	0.99	cent digits of the price
relative price	805949	0.98	0.13	0.00	4.99	price relative to the avg. price on prod. i' market ^{b)}
rel. shipping cost	805949	1.01	0.47	0.00	185.36	ship. cost rel. to the avg. ship. cost on prod. i's market
immed. availability	805949	0.17		0.00	1.00	dummy: item can be shipped within a day
duration	805949	6.07	9.21	0.06	43.07	duration of the offer (in days)
duration top (undercut)	14933	2.52	5.06	0.06	42.17	duration of best-price offers, undercut by competitors
duration top (own change)	26817	4.28	8.05	0.06	43.07	durat. of best-price offers, the shop changed itself
OTHERS						
clicks per week	805949	1.04	15.04	0.00	4673.58	clicks per week (clicks/duration)
LCTs per week	805949	0.09	1.83	0.00	644.81	last clicks per week (LCTs/duration)
clicks	805949	0.47	4.58	0.00	651.00	total clicks on ij (not weighted)
LCTs	805949	0.04	0.41	0.00	74.00	total LCTs on ij (unweighted)

NOTES: ^{a)} Multiple offers for product i of e-tailer j are possible within our observation period. ^{b)} Note that we calculated this price based on all observations (the stock sample), and hence, the mean may differ from 1. ^{c)} This surprisingly high maximum price stems from a 2.5m wide TFT-screen with cinema-sound equipment.

Table 5.2: Frequency of focal price endings for (i) low-price and high-price offers and (ii) low-price vs. high-price products.

	low-price vs. high-price offer		low-price vs. high-price product		total sample
	lowest tercile	highest tercile	lowest tercile	highest tercile	
1=ends in 00c	0.0720	0.147	0.0743	0.144	0.109
1=ends in -9.00	0.0543	0.126	0.0659	0.107	0.0863
1=ends in 9.00	0.0177	0.0209	0.00843	0.0369	0.0227
1=ends in 50c	0.0285	0.0342	0.0317	0.0369	0.0343
1=ends in 90c	0.0361	0.0524	0.0418	0.0427	0.0423
1=ends in 99c	0.0215	0.0173	0.0183	0.0216	0.0200

NOTES: Number of observations: 523,748 offers; The figures represent shares of the focal price in the respective terciles.

Table 5.3: Focal prices and price stickiness: all price offers

	pooled			stratified		
	(1)	(2)	(3)	(4)	(5)	(6)
1=ends in 99c	-0.428*** (0.009)	-0.297*** (0.009)		-0.098*** (0.010)	-0.053*** (0.010)	
1=ends in 90c	-0.408*** (0.007)	-0.370*** (0.007)		-0.098*** (0.008)	-0.087*** (0.008)	
1=ends in 50c	-0.239*** (0.007)	-0.172*** (0.007)		-0.055*** (0.008)	-0.037*** (0.008)	
1=ends in 9.00	-0.764*** (0.009)	-0.533*** (0.010)		-0.195*** (0.011)	-0.131*** (0.011)	
1=ends in -9.00	-0.445*** (0.005)	-0.375*** (0.005)		-0.132*** (0.007)	-0.103*** (0.007)	
9-clustered			-0.138*** (0.004)			-0.012*** (0.003)
even price (no 9)			-0.334*** (0.005)			-0.055*** (0.006)
euro digits in 100s		-0.003*** (0.000)	-0.003*** (0.000)		-0.002*** (0.000)	-0.002*** (0.000)
rel. shipping cost		-0.065*** (0.005)	-0.047*** (0.005)		0.005** (0.002)	0.005** (0.002)
immed. availability		-0.166*** (0.005)	-0.173*** (0.005)		0.042*** (0.008)	0.042*** (0.008)
clicks		-0.030*** (0.002)	-0.034*** (0.002)		-0.021*** (0.001)	-0.021*** (0.001)
number of shops		0.001*** (0.000)	0.001*** (0.000)		0.001*** (0.000)	0.001*** (0.000)
price changes (excl. i)		0.450*** (0.004)	0.456*** (0.005)		0.424*** (0.005)	0.424*** (0.005)
recommendations		0.017 (0.014)	0.013 (0.014)		0.027** (0.012)	0.026** (0.012)
Observations	805949	805949	805949	805949	805949	805949
Number of Failures	778521	778521	778521	778521	778521	778521
Log L	-9720432	-9720432	-9720432	-5942823	-5942823	-5942823
χ^2	13812	30442	22454	626	9883	9241
DF	5	14	11	5	14	11

NOTES: Cox hazard model with cluster-robust standard errors, columns 4-6 include a stratification with respect to shops. Dependent variable: duration of price spell; no. of distinct offers (j offering i) = 410245; no. of e-tailers j = 698; no. of products i = 23317; standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1.

Table 5.4: Focal prices and price stickiness: price-leading offers.

	undercut			own change		
	(1)	(2)	(3)	(4)	(5)	(6)
1=ends in 99c	-0.227*** (0.050)	-0.242*** (0.051)		0.067 (0.056)	0.066 (0.057)	
1=ends in 90c	-0.119** (0.049)	-0.103** (0.049)		-0.128*** (0.040)	-0.118*** (0.041)	
1=ends in 50c	-0.086 (0.063)	-0.090 (0.064)		0.009 (0.047)	0.008 (0.047)	
1=ends in 9.00	-0.336*** (0.055)	-0.312*** (0.056)		-0.047 (0.044)	-0.048 (0.044)	
1=ends in -9.00	-0.139*** (0.044)	-0.144*** (0.045)		0.026 (0.035)	0.020 (0.035)	
9-clustered			-0.093*** (0.025)			-0.018 (0.014)
even price (no 9)			-0.080** (0.037)			0.018 (0.027)
euro digits in 100s		-0.002 (0.001)	-0.002 (0.001)		0.001 (0.000)	0.001 (0.000)
rel. shipping cost		-0.007 (0.032)	-0.008 (0.032)		-0.040** (0.016)	-0.041** (0.016)
immed. availability		0.028 (0.041)	0.022 (0.041)		0.006 (0.023)	0.005 (0.022)
clicks while in lead		0.020*** (0.002)	0.020*** (0.002)		0.014*** (0.002)	0.014*** (0.002)
number of shops		0.006*** (0.000)	0.006*** (0.000)		0.001*** (0.000)	0.001*** (0.000)
price changes (excl. i)		2.463*** (0.375)	2.473*** (0.376)		3.322*** (0.444)	3.324*** (0.444)
recommendations		0.146** (0.058)	0.141** (0.058)		-0.011 (0.030)	-0.011 (0.030)
Observations	41764	41764	41764	41764	41764	41764
Number of Failures	14933	14933	14933	26817	26817	26817
Log L	-80060	-80060	-80060	-152829	-152829	-152829
χ^2	51	609	566	16	208	185
DF	5	13	10	5	13	10

NOTES: Competing risk Cox hazard model with cluster-robust standard errors, stratified at the firm level; dependent variable: duration the offer marked the lowest price; failures in the group of undercut (actively changed) offers: 14933 (26817; and 14 offers that were valid for more than 43 days); no. of firms j in estimation = 519 ; no. of products i =13341; no. of distinct firm-product combinations ij = 21629; standard errors in parentheses: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 5.5: Focal prices and price stickiness: size of price jumps

	undercut		own change down		upward change	
	(1)	(2)	(3)	(4)	(5)	(6)
1=ends in 99c	0.731** (0.289)		0.296*** (0.084)		0.564** (0.232)	
1=ends in 90c	0.387*** (0.134)		0.099 (0.131)		-0.165 (0.106)	
1=ends in 50c	0.489*** (0.170)		-0.008 (0.118)		0.114 (0.144)	
1=ends in 9.00	1.254*** (0.177)		0.866*** (0.158)		1.461*** (0.165)	
1=ends in -9.00	0.503*** (0.147)		0.192 (0.160)		0.658*** (0.158)	
9-clustered		0.261*** (0.074)		0.053 (0.037)		-0.047 (0.084)
even price (no 9)		0.119 (0.106)		-0.066 (0.118)		0.060 (0.101)
price in 100s	0.055*** (0.010)	0.057*** (0.010)	0.022*** (0.002)	0.022*** (0.002)	0.023*** (0.008)	0.025*** (0.008)
number of shops	-0.005*** (0.001)	-0.005*** (0.002)	-0.001 (0.001)	-0.001 (0.001)	-0.004*** (0.001)	-0.004*** (0.001)
clicks while in lead	-0.006* (0.003)	-0.006* (0.003)	0.005 (0.005)	0.005 (0.005)	0.013*** (0.004)	0.015*** (0.004)
immed. availability	-0.245* (0.138)	-0.225 (0.138)	-0.009 (0.094)	-0.001 (0.099)	-0.063 (0.087)	-0.039 (0.087)
price changes (excl. i)	1.406 (0.940)	1.299 (0.929)	-0.026 (2.689)	-0.179 (2.820)	4.799 (5.155)	4.516 (5.180)
recommendations	0.004 (0.090)	0.012 (0.090)	0.009 (0.071)	0.016 (0.071)	-0.307** (0.123)	-0.304** (0.122)
1=no recommendation	0.114* (0.067)	0.107 (0.067)	0.011 (0.041)	0.010 (0.043)	-0.016 (0.065)	-0.025 (0.067)
rel. shipping cost	-0.101 (0.111)	-0.100 (0.115)	0.030 (0.058)	0.032 (0.060)	-0.022 (0.089)	-0.000 (0.096)
Constant	-1.377*** (0.207)	-1.317*** (0.220)	-2.842*** (0.086)	-2.818*** (0.093)	-0.985*** (0.145)	-0.875*** (0.153)
Observations	14933	14933	11399	11399	12809	12809
Number of Firms	444	444	246	246	340	340
Mean of Dep. Var.	-1.32	-1.32	-2.70	-2.70	-1.13	-1.13
Log L	-32400	-32400	-17239	-17239	-24357	-24357
DF	13	10	13	10	13	10

NOTES: Fixed effects panel regression with heteroscedasticity robust standard errors; dependent variable in columns (1) to (4): $\log(\text{bestprice}_t - \text{bestprice}_{t+1})$; due to the logarithmic transformation the dependent variable in columns (5) and (6) is: $\log(\text{bestprice}_{t+1} - \text{bestprice}_t)$; no. of distinct firms j in estimation = 481; no. of products i =9804; no. of individual offers ($i \times j$) = 15581; standard errors in parentheses: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 5.6: Demand and focal prices

	clicks per week			LCTs per week		
	(1)	(2)	(3)	(4)	(5)	(6)
1=ends in 99c	0.082 (0.124)			-0.020 (0.014)		
1=ends in 90c	-0.288*** (0.084)			-0.019** (0.010)		
1=ends in 50c	-0.582*** (0.095)			-0.039*** (0.011)		
1=ends in 9.00	0.580*** (0.127)			0.042*** (0.014)		
1=ends in -9.00	0.270*** (0.063)			0.016** (0.007)		
9-clustered		0.088** (0.038)			0.007 (0.004)	
even price (no 9)		0.064 (0.059)			0.003 (0.007)	
euro digits in 100s			-0.301*** (0.019)			-0.017*** (0.002)
cent digits			0.139*** (0.053)			0.005 (0.006)
relative price	-3.908*** (0.470)	-3.992*** (0.468)		-0.265*** (0.053)	-0.273*** (0.053)	
rel. shipping cost	0.092** (0.038)	0.103*** (0.037)	0.096** (0.037)	0.008* (0.004)	0.009** (0.004)	0.008** (0.004)
immed. availability	1.602*** (0.050)	1.583*** (0.049)	1.664*** (0.048)	0.116*** (0.006)	0.114*** (0.006)	0.120*** (0.005)
reviews (grade)	-0.120*** (0.029)	-0.127*** (0.029)	-0.092*** (0.028)	-0.009*** (0.003)	-0.010*** (0.003)	-0.007** (0.003)
reviews (num.)	0.002*** (0.000)	0.002*** (0.000)	0.003*** (0.000)	0.000*** (0.000)	0.000*** (0.000)	0.000*** (0.000)
German shop	0.213*** (0.049)	0.218*** (0.049)	0.236*** (0.048)	0.008 (0.006)	0.008 (0.006)	0.010* (0.005)
Constant	4.161*** (0.482)	4.198*** (0.480)	1.382*** (0.114)	0.287*** (0.054)	0.290*** (0.054)	0.085*** (0.013)
Observations	793507	793507	793507	793507	793507	793507
Number of Products	22722	22722	22722	22722	22722	22722
Log L			-3244456			-1513480
χ^2	7152	7064		2490	2458	
DF	13	10	9	13	10	9

NOTES: Columns (1), (2), (4), and (5) show IV-Panel regressions, columns (3) and (6) show simple panel regressions; dependent variable in columns (1-3): clicks per week, dependent variable in columns (4-6): LCTs per week; no. of distinct offers (j offering i) = 388655; no. of firms = 569; no. of products = 22722; standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1.

Table 5.7: Demand and price points for different types.

Panel A: Price Points								
	search type		product category		use of page		on- vs. offline	
	(1) intense	(2) brief	(3) expensive	(4) inexp.	(5) freq.	(6) infreq.	(7) pure play	(8) click+mortar
1=ends in 99c	-0.028 (0.079)	0.036*** (0.012)	0.269 (0.186)	0.074 (0.188)	-0.018 (0.090)	0.023** (0.011)	0.012 (0.141)	0.678*** (0.190)
1=ends in 90c	-0.200*** (0.054)	-0.017** (0.008)	-0.036 (0.135)	-0.583*** (0.125)	-0.236*** (0.061)	-0.016** (0.007)	-0.288*** (0.107)	-0.226*** (0.075)
1=ends in 50c	-0.337*** (0.061)	-0.026*** (0.010)	-0.538*** (0.145)	-0.387*** (0.141)	-0.401*** (0.069)	-0.019** (0.008)	-0.710*** (0.113)	0.100 (0.102)
1=ends in 9.00	0.264*** (0.081)	0.025** (0.013)	0.841*** (0.156)	-0.248 (0.270)	0.267*** (0.092)	0.022** (0.011)	0.406** (0.163)	0.609*** (0.109)
1=ends in -9.00	0.122*** (0.040)	0.014** (0.006)	0.839*** (0.091)	-0.084 (0.102)	0.137*** (0.045)	0.012** (0.005)	0.248*** (0.084)	-0.149*** (0.054)
relative price	-2.183*** (0.299)	-0.335*** (0.047)	-11.261*** (1.836)	-2.105*** (0.410)	-2.511*** (0.341)	-0.286*** (0.040)	-3.622*** (0.568)	-7.522*** (0.610)
Constant	2.365*** (0.306)	0.360*** (0.048)	11.618*** (1.872)	2.431*** (0.445)	2.713*** (0.349)	0.308*** (0.041)	3.831*** (0.581)	8.697*** (0.614)
further controls	yes	yes	yes	yes	yes	yes	yes	yes
Observations	793507	793507	261796	261952	793507	793507	656026	137481
Number of Products	22722	22722	8184	7707	22722	22722	22177	14177
Mean of Dep. Var.	0.59	0.07	1.32	0.84	0.70	0.05	1.22	0.55
χ^2	4952.29	2872.74	4220.77	1811.45	5389.50	2175.27	5763.76	3319.26
Df	13	13	13	13	13	13	13	13
Panel B: BASU DEMAND								
	search type		product category		use of page		on- vs. offline	
	(1) intense	(2) brief	(3) expensive	(4) inexp.	(5) freq.	(6) infreq.	(7) pure play	(8) click+mortar
euro digits in 100s	-0.161*** (0.012)	-0.018*** (0.002)	-0.236*** (0.017)	-6.548*** (0.345)	-0.184*** (0.014)	-0.014*** (0.002)	-0.382*** (0.024)	-0.137*** (0.017)
cent digits	0.089*** (0.034)	0.012** (0.005)	-0.076 (0.083)	0.070 (0.080)	0.108*** (0.039)	0.008* (0.005)	0.208*** (0.065)	0.133** (0.055)
Constant	0.769*** (0.073)	0.089*** (0.011)	2.806*** (0.216)	3.553*** (0.246)	0.865*** (0.083)	0.072*** (0.010)	1.565*** (0.141)	1.990*** (0.149)
further controls	yes	yes	yes	yes	yes	yes	yes	yes
Observations	793507	793507	261796	261952	793507	793507	656026	137481
Number of Products	22722	22722	8184	7707	22722	22722	22177	14177
Mean of Dep. Var.	0.59	0.07	1.32	0.84	0.70	0.05	1.22	0.55
Df	9	9	9	9	9	9	9	9

NOTES: IV-Panel regressions. Columns (1) and (2): consumers with different search intensity. Columns (3) and (4): high and low-cost items. Columns (5) and (6): frequent and infrequent buyers. Column (7) shows the results for pure play stores and column (8) for firms that also have a shop offline. Dependent Variable: clicks per week; standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1; max. no. of distinct offers (j offering i) = 388655; max. no. of firms = 569; max. no. of products = 22722.

Appendix A

Additional Materials to Chapter 2:

“Spillovers in Networks of User Generated Content”

A.1 Details about Data Preparation, the Treated and the Control Groups

This section gives detailed information about the preparation and storage of the dataset. The subsequent subsections explain how the database was put together and the procedure I used to extract the dataset that I use.

A.1.1 Preparation and Extraction

The dataset is based on a full-text dump of the German Wikipedia from the Wikimedia toolserver. To construct the history of the articles' hyperlink network for the entire encyclopedia, it was necessary to parse the data and identify the links. From the resulting tables, I constructed a time-varying graph of the article network, which provided the foundation for how I sample articles in my analysis. Furthermore, information about the articles, such as the number of authors who contributed up to a particular point in time or the existence of a section with literature references was added. Hence, the data I use are based on 153 weeks of the the entire German Wikipedia's revision history between December 2007 and December 2010. Since the data are in the order of magnitude of terabytes, it was not be possible to conduct the data analysis using only in-memory processing. We therefore stored the data in a relational database (disk-based) and queried the data using Database Supported Haskell (DSH) (Giorgidze et al. (2010)).¹

A.1.2 Choice of Treated Articles and Neighborhood

"Featured articles" were found by consulting the German Wikipedia's archive of pages that were selected to be advertised on Wikipedia's main page ("Seite des Tages") between December 2007 and December 2010. To reduce the computational burden and to avoid the risk of temporal overlaps of different treatments, I focus on pages that were selected on the 10th of a month. I identified all the pages that received a direct link (L_1) or an indirect link (L_2) from such a featured article more than a week before treatment. I evaluated links with this time gap before the shock actually occurred to make sure that the results are not driven by endogeneous link formation.² Having fixed the set of pages to observe, I extracted daily information on the contemporary state of the articles (page visits, number of revisions, number of distinct authors that contributed, page length, number of external links etc.). I determine these variables on a daily basis, 14 days before the event occurred (on a neighboring page) and 14 days after the shock (giving a total of 29 observations per page).

To identify major events, I consulted the corresponding page on Wikipedia and selected the 26 largest events with spontaneous onset. For each of these events we identified the page that corresponds to the event, which are considered to be in the set " L_0 " (sometimes also called "start pages"). Note that this page is typically created *after the event* occurred³, which obliges me to identify the pages, that user will most likely turn to until the disaster's page is in place. To achieve this, I used the link data to identify the set of pages that later shared a reciprocal link with the start page. Such

¹This is a novel high-level language allowing the writing and efficient execution of queries on nested and ordered collections of data.

²I thus only include pages that had a link before it was known that the start page will be hit. I furthermore exclude pages that receive their indirect (L_2) link via a page that has more than 100 links, since such pages are very likely either pure "link pages" very general pages (such as pages about a year), that bare only a very weak relationship to the shocked site.

³Usually it takes up to two days until the event receives its own page.

a reciprocal link indicates that they were closely related to the event. After the event page came in to existence they were only one click away (set " L_1 "). Next, we identified those pages that received a link from an L_1 page (unidirectional) (2 clicks away set " L_2 ")

I am most interested in attention spillovers and content provision, which are not directly related to the events but rather a consequence of the spike in interest and the resulting improvements to the linked pages. Hence, I will not focus on the treated pages directly, but on the set L_1 that are "one click away", in my analysis of the "featured articles".⁴ For disasters the shock is very large and the event page usually does not exist at the time of the shock, so the L_1 pages might have been treated themselves.⁵ Hence, I focus on the indirectly linked set of pages (L_2) in the analysis below.

A.1.3 Choice of Control Group Articles and Neighborhood

The approach I take in this paper hinges on the availability of a valid control group. To obtain such observations I pursue two distinct strategies. The first approach uses pages which are similar but unlikely to be affected by the treatment. For a first comparison I selected other featured articles and neighbors thereof that were advertised featured either later or earlier in time. Given such a similar page, I identified their direct and indirect neighbors when the event occurred on the treated page. This gives me a set $C1_{control}$ which is similar in both size and characteristics to the sampled pages (before the shock). Yet, the choice of the start pages in the comparison group is somewhat arbitrary.⁶ I address this issue by simulating a treatment on the treated pages 42 days before the disaster or event occurred. I refer to the articles in this "placebo-treatment" as $C1_{placebo}$, because for them $t = 0$ when no actual treatment occurred. By design, this comparison group consists of the same set of articles (treated and their neighbors). This comes at the cost of observing the articles at a different point in time. A third control group of "unrelated" observations results from applying a placebo to the control group.⁷

Table 2.8 (in the data appendix) shows which featured articles were chosen by my procedure and included in the data. In general, they cover various topics such as innovations (e.g. the CCD-sensor), places (Helgoland), soccer clubs (Werder Bremen) or art history topics (Carolingian book illustrations). The first column of the table shows the number of articles that belong to each featured article. The last three columns show the number of observations that received a link from an article before it was advertised featured, separated by whether or not they belong to a time-series with actually treated observations.⁸ The numbers range from 2,088 to 33,872.

⁴Effects on the pages that are 2 clicks away were too small to be measured.

⁵Some of the consequences of major events, such as earthquakes, might change the state of the world and thus trigger a change in content, which is merely *due to the event* (e.g. destruction of an important monument). Consequently, I do not emphasize the change in activity on the pages that are only one click away for disasters. I also exclude pages if they were later directly linked to the event page.

⁶Ideally the selection of comparison pages should be based on matching procedures, which is unfortunately not possible without computing the characteristics of all the 1,000,000 nodes. My approach is however quite robust independently of how I specify the control group. I also compared to the neighbors around articles of similar size and relative importance, about similar topics, but in a remote geographic space or technical domain. Such a change in the specification of the control group does not affect my results. (available upon request).

⁷This set of observations actually emerged as an artifact from the data extraction. Nevertheless it provides yet another group that can be compared to the treated group.

⁸Note, that each page shows up 29 times in the raw data and was sampled twice (placebo and real treatment), so that the number of corresponding pages (treatment or control) can be inferred by dividing the number of observations by 58.

For disasters I proceeded along similar lines. I focused on the network around older catastrophes that occurred at a different point in time and were not from exactly the same domain, to avoid overlaps in the link network ($C_{control}$). Alternatively, I observe the same set of pages seven weeks before the disaster ($C_{placebo}$). Table 2.9 shows which events were included in the data and shows the associated number of observations for each of them. The dataset includes both natural disasters as well as technical or economic catastrophes.

A.2 The empirical model and structural identification of the parameter of interest.

This section presents the structural model and discusses the parameters of interest, the challenges in identifying them and the approach taken to tackle them.

A.2.1 Introductory remarks

I depart from the well known linear-in-means model as formulated by Manski (1993).⁹

$$y_{it} = \alpha \frac{\sum_{j \in P_{it}} y_{jt}}{N_{P_{it}}} + X_{it-1}\beta + \gamma \frac{\sum_{j \in P_{it}} X_{jt-1}}{N_{P_{it}}} + \epsilon_{it} \quad (\text{A.1})$$

y_{it} denotes the outcome of interest in period t and X_{it-1} are i 's observed characteristics at the end of period $t-1$ (beginning of period t).¹⁰ P_{it} is the set of i 's peers and $N_{P_{it}}$ represents the number of i 's peers. β measures the effect of i 's own characteristics and γ accounts for how i 's performance is affected by the peers' average characteristics. The coefficient of interest is α . In the present context it measures how the clicks on page i are influenced the clicks on the adjacent pages. Bramoullé et al. (2009) suggest a more succinct notation based on vector and matrix notation:

$$\mathbf{y}_t = \alpha \mathbf{G} \mathbf{y}_t + \beta \mathbf{X}_{t-1} + \gamma \mathbf{G} \mathbf{X}_{t-1} + \epsilon_t \quad \mathbf{E}[\epsilon_t | \mathbf{X}_{t-1}] = \mathbf{0}$$

Note that the linear in means model provides the weakest basis for identification. I conjecture that the insights carry over to other linear models and less weakly identified non-linear models.

A.2.2 Setup and Basic Intuition

Augment the model (eq. A.1) by observable and locally applied treatments (shocks):

$$y_{it} = \alpha \frac{\sum_{j \in P_{it}} y_{jt}}{N_{P_{it}}} + X_{i,t-1}\beta + \gamma \frac{\sum_{j \in P_{it}} X_{j,t-1}}{N_{P_{it}}} + \delta_1 D_{it} + \epsilon_{it} \quad (\text{A.2})$$

where the new coefficient δ_1 measures the direct effect if a node(page) is treated.

Note that $X_{it-1}\beta$ may contain an individual fixed effect and an additively separable age-dependent part: $X_{it-1}\beta = \beta_i + \widetilde{X_{i,t-1}}\beta_1 + \beta_2 f(\text{age})$. To see how local treatments

⁹Note that it is easy to add a fixed effect to the model, but that it will be eliminated when taking differences. Consequently, I omit it for ease of notation.

¹⁰The choice of the temporal structure depends on the application that the researcher has in mind. In the present application many independent variables are stock variables (articles' characteristics such as page length), while the dependent variables are typically flows (clicks or new revisions).

can be used as a source of identification, consider two pairs of nodes.

Local application of treatment: First, consider two connected nodes, where one is treated ($\ell 0$) in period t and the neighbors are not treated ($\ell 1 \in L1$). Assume for simplicity that $\ell 0$ is the only treated node in $\ell 1$'s neighborhood.

$$\ell 0 :: \textcolor{blue}{y}_{\ell 0 t} = \alpha \frac{\sum_{j \in P_{\ell 0 t}} y_{j t}}{N_{P_{\ell 0 t}}} + X_{\ell 0 t-1} \beta + \gamma \frac{\sum_{j \in P_{\ell 0 t}} X_{j t-1}}{N_{P_{\ell 0 t}}} + \delta_1 \textcolor{red}{1} + \epsilon_{\ell 0 t} \quad (\text{A.3})$$

$$\ell 1 \in L1 :: y_{\ell 1 t} = \alpha \frac{\textcolor{blue}{y}_{\ell 0 t} + \sum_{j \in P_{\ell 1 t}/\ell 0} y_{j t}}{N_{P_{\ell 1 t}}} + X_{\ell 1 t-1} \beta + \gamma \frac{\sum_{j \in P_{\ell 1 t}} X_{j t-1}}{N_{P_{\ell 1 t}}} + \delta_1 \textcolor{red}{0} + \epsilon_{\ell 1 t} \quad (\text{A.4})$$

If we now consider a comparison group of two connected nodes ($c0$ and $c1$) where nobody gets treated, D_t would take the value 0 for both $c0$ and $c1$. The newly introduced term would simply drop out. It can easily be seen, how the local treatment will allow to measure the spillover or peer effect. This will be possible despite the richness in other sources of variation, provided (i) the shocks are large enough and (ii) the “control network” allows to credibly infer the dynamics in the “treated network”, had no treatment taken place.

Condensed Notation: I use the matrix notation suggested by Bramoullé et al. (2009) and incorporate the newly proposed vector of treatments¹¹:

$$\mathbf{y}_t = \alpha \mathbf{G} \mathbf{y}_t + \mathbf{X}_{t-1} \beta + \gamma \mathbf{G} \mathbf{X}_{t-1} + \delta_1 \textcolor{blue}{D}_t + \epsilon_t \quad \mathbf{E}[\epsilon_t | \mathbf{D}_t] = \mathbf{0} \quad (\text{A.5})$$

\mathbf{G} is a $N \times N$ matrix, which captures the link structure in the network. $G_{ij} = \frac{1}{N_{P_i}-1}$ if i receives a link from j and $G_{ij} = 0$ otherwise. Note that I do not require \mathbf{G} to be exogenously given, but only \mathbf{D}_t , a vector which is 1 at the treated nodes (if they are *currently* treated) and 0 otherwise. In some of the proofs and in my application I will assume a local treatment that affects only a single node. Formally this is written as an elementary vector $\mathbf{D}_t = \mathbf{e}_{\ell 0}$ with the 1 in the coordinate that corresponds to the treated node. On the untreated subnetwork we have $\mathbf{D}_t = \mathbf{0}$, a vector of zeros.

Unlike Bramoullé et al. (2009), I do not look for an instrument for $\mathbf{G} \mathbf{y}$. Since I rather use exogenous shocks that affect only one part of the network, there will be no requirements on the linear independence of \mathbf{G} and \mathbf{G}^2 .

A.2.3 Proof of Result 1

I shall now proceed to provide the formal argument for Result 1. To increase the readability I will make a few assumptions to keep things simple. Most importantly I assume the network \mathbf{G} to be stable over time but I allow \mathbf{X}_t to change dynamically. I set the comparison group (which was indexed by c) to be the group itself S periods earlier, which results in an S -period difference-in-differences.¹²

¹¹ \mathbf{X} might include a time-dependent component (e.g. a linear function of age) as well.

¹²Importantly the nodes in the network have to be observed over time and have to evolve in a stable fashion, to ensure that the first differences are the same at t and $t - S$. This setting corresponds to comparing the evolution of nodes in a very stable network during a post and a pre-treatment stage. It is also reasonably close to the “placebo condition” of my application below. At the end of the formal derivations I will discuss the consequences of relaxing the requirement of a stable network or the consequences of adding the assumption that \mathbf{X}_t does not change between the periods of observation.

Result 1: A difference-in-differences estimator contains the following quantity:

$$\text{DiD} = \delta_1 \mathbf{D}_t (\mathbf{I} + \alpha \mathbf{G} + \alpha^2 \mathbf{G}^2 + \alpha^3 \mathbf{G}^3 + \dots)$$

Proof. The reduced form corresponding to equation A.5 is given by:

$$\mathbf{y}_t = (\mathbf{I} - \alpha \mathbf{G})^{-1} [\mathbf{X}_{t-1} \beta + \gamma \mathbf{G} \mathbf{X}_{t-1} + \delta_1 \mathbf{D}_t + \epsilon_t] \quad (\text{A.6})$$

and the expectation conditional on the “treatment” is:

$$\begin{aligned} \mathbf{E}[\mathbf{y}_t | \mathbf{D}_t] &= (\mathbf{I} - \alpha \mathbf{G})^{-1} [(\beta + \gamma \mathbf{G}) \mathbf{E}[\mathbf{X}_{t-1} | \mathbf{D}_t] + \delta_1 \mathbf{D}_t + \mathbf{E}[\epsilon_t | \mathbf{D}_t]] =^{b.A.} (\text{A.7}) \\ &=^{b.A.} (\mathbf{I} - \alpha \mathbf{G})^{-1} [(\beta + \gamma \mathbf{G}) \mathbf{E}[\mathbf{X}_{t-1} | \mathbf{D}_t] + \delta_1 \mathbf{D}_t] \end{aligned}$$

Taking the first difference, we obtain:

$$\begin{aligned} \Delta_t \mathbf{E}[\mathbf{y} | \mathbf{D}] &= \mathbf{E}[\mathbf{y}_t | \mathbf{D}_t] - \mathbf{E}[\mathbf{y}_{t-1} | \mathbf{D}_{t-1}] = \quad (\text{A.8}) \\ &= (\mathbf{I} - \alpha \mathbf{G})^{-1} [(\beta + \gamma \mathbf{G}) \{ \mathbf{E}[\mathbf{X}_{t-1} | \mathbf{D}_t] - \mathbf{E}[\mathbf{X}_{t-2} | \mathbf{D}_{t-1}] \} + \delta_1 \Delta \mathbf{D}_t] = \\ &= (\mathbf{I} - \alpha \mathbf{G})^{-1} [(\beta + \gamma \mathbf{G}) \{ \mathbf{E}[\mathbf{X}_{t-1} | \mathbf{D}_t] - \mathbf{E}[\mathbf{X}_{t-2} | \mathbf{D}_{t-1}] \} + \delta_1 \mathbf{D}_t] \end{aligned}$$

...where $\Delta \mathbf{D}_t = \mathbf{D}_t - \mathbf{D}_{t-1}$ and the second equality holds, because treatments are assumed to start in period t , but not before.¹³

Now consider the control group formed by the same network, but S periods earlier:

$$\mathbf{y}_{t-S} = \alpha \mathbf{G} \mathbf{y}_{t-S} + \mathbf{X}_{t-S-1} \beta + \gamma \mathbf{G} \mathbf{X}_{t-S-1} + \delta_1 \mathbf{D}_{t-S} + \epsilon_{t-S}$$

The first difference of the reduced form’s conditional expectations are:

$$\begin{aligned} \Delta_{t-S} \mathbf{E}[\mathbf{y} | \mathbf{D}] &= \mathbf{E}[\mathbf{y}_{t-S} | \mathbf{D}_{t-S}] - \mathbf{E}[\mathbf{y}_{t-S-1} | \mathbf{D}_{t-S-1}] = \\ &= (\mathbf{I} - \alpha \mathbf{G})^{-1} [(\beta + \gamma \mathbf{G}) \{ \mathbf{E}[\mathbf{X}_{t-S-1} | \mathbf{D}_{t-S}] - \mathbf{E}[\mathbf{X}_{t-S-2} | \mathbf{D}_{t-S-1}] \} + \delta_1 \Delta \mathbf{D}_{t-S}] = \\ &= (\mathbf{I} - \alpha \mathbf{G})^{-1} [(\beta + \gamma \mathbf{G}) \{ \mathbf{E}[\mathbf{X}_{t-S-1} | \mathbf{D}_{t-S}] - \mathbf{E}[\mathbf{X}_{t-S-2} | \mathbf{D}_{t-S-1}] \} + 0] \end{aligned}$$

with $\Delta \mathbf{D}_{t-S} = 0$, since treatments are assumed to start in period t , but not earlier. Proceeding to take the Difference in Differences, we obtain:

$$\begin{aligned} \text{DiD} &:= \Delta \mathbf{y}_t \mathbf{E}[\mathbf{y} | \mathbf{D}] - \Delta \mathbf{y}_{t-S} \mathbf{E}[\mathbf{y} | \mathbf{D}] = \\ &= (\mathbf{I} - \alpha \mathbf{G})^{-1} [(\beta + \gamma \mathbf{G}) \{ \mathbf{E}[\mathbf{X}_{t-1} | \mathbf{D}_t] - \mathbf{E}[\mathbf{X}_{t-2} | \mathbf{D}_{t-1}] \} + \delta_1 \mathbf{D}_t] - \\ &\quad - (\beta + \gamma \mathbf{G}) \{ \mathbf{E}[\mathbf{X}_{t-S-1} | \mathbf{D}_{t-S}] - \mathbf{E}[\mathbf{X}_{t-S-2} | \mathbf{D}_{t-S-1}] \} \end{aligned}$$

Denoting the change in the expectation of \mathbf{X}_{t-1} conditional on \mathbf{D}_t more concisely by $\{ \mathbf{E}[\mathbf{X}_{t-1} | \mathbf{D}_t] - \mathbf{E}[\mathbf{X}_{t-2} | \mathbf{D}_{t-1}] \} = \Delta_t (\mathbf{E}[\mathbf{X} | \mathbf{D}])$ and rearranging gives:

$$\text{DiD} = (\mathbf{I} - \alpha \mathbf{G})^{-1} [(\beta + \gamma \mathbf{G}) \{ \Delta_t (\mathbf{E}[\mathbf{X} | \mathbf{D}]) - \Delta_{t-S} (\mathbf{E}[\mathbf{X} | \mathbf{D}]) \} + \delta_1 \mathbf{D}_t] \quad (\text{A.9})$$

which reduces to:

$$\text{DiD} = (\mathbf{I} - \alpha \mathbf{G})^{-1} \{ \delta_1 \mathbf{D}_t \} \quad (\text{A.10})$$

¹³That difference contains the time-dependent component and the effect of any changes in the independent variables. If βX_{it} is modeled to contain an additively separable age-dependent part as in our example above, $\Delta X_{it-S} \beta$ would contain $\frac{df(\text{age})}{dt}$ (to be eliminated by taking the Difference in Differences).

if $\Delta_t(\mathbf{E}[\mathbf{X}|\mathbf{D}]) = \Delta_{t-S}(\mathbf{E}[\mathbf{X}|\mathbf{D}])$. Thus, the identifying assumption is that the expected changes of the pages between $t-1$ and t are the same as from $t-S-1$ and $t-S$. This is satisfied if $\Delta X_t|D_t$ is stationary of order one.

Provided $(\mathbf{I} - \alpha\mathbf{G})^{-1}$ is invertible we can use the property that $(\mathbf{I} - \alpha\mathbf{G})^{-1} = \sum_{s=0}^{\infty} \alpha^s \mathbf{G}^s$ ¹⁴, the general impact of a local treatment is:

$$\text{DiD} = \delta_1 \mathbf{D}_t (\mathbf{I} + \alpha\mathbf{G} + \alpha^2\mathbf{G}^2 + \alpha^3\mathbf{G}^3 + \dots) \quad (\text{A.11})$$

which completes the proof. \square

Discussion of the assumptions used:

- (i) $\mathbf{E}[\epsilon_t|\mathbf{D}_t] = \mathbf{0}$
- (ii) α is smaller than the norm of the inverse of the largest eigenvalue of \mathbf{G} . A regularity condition to ensures that the expression $(\mathbf{I} - \alpha\mathbf{G})^{-1} = \sum_{s=0}^{\infty} \alpha^s \mathbf{G}^s$ is well defined.
- (iii) I assumed the network to be stable over time and used it's earlier state as control observation. Formally this is written as $\mathbf{G}_{\ell,t} = \mathbf{G}_{\ell,t-1} = \mathbf{G}$ and $\mathbf{G}_{c,t} = \mathbf{G}_{\ell,t-S} = \mathbf{G}$. This assumption could be relaxed, but only at the expense of strengthening the following assumption.
- (iv) $\Delta_t(\mathbf{E}[\mathbf{X}|\mathbf{D}]) - \Delta_{t-S}(\mathbf{E}[\mathbf{X}|\mathbf{D}])$, which means that the expected changes of the pages between $t-1$ and t are the same as from $t-S-1$ and $t-S$ ¹⁵. This is the analogue of the well known common trends assumption.
- (v) SUTVA on the level of subnetworks: the non-treated subnetwork is not affected by treatment of the treated subnetwork. In the present context SUTVA holds for my placebo condition and, given the size of the Wikipedia network, it is also plausibly satisfied for the control group formed by a remote part of the network.

The proof for the control group consisting of remote nodes is analogous. It relaxes the third assumption and requires a more general formulation of the fourth. The qualitative meaning of the generalized assumption will be the same: Absent treatment the treated *network* and the control *network* must “*evolve* in the same way.”¹⁶ However, I have to maintain the assumption that the network formation *process* is not affected by the treatment.¹⁷ I do not consider this assumption warranted for disasters and I checked this assumption in my “today’s featured article application”: Link formation remains on low levels. On normal days, articles’ degree grows steadily by about 0.1 links per day, with total in-links averaging at 120. There is a short increase by 0.2 in-links per article (or 0.2% of the link stock). Yet, first this is in sync with the peak in edits, but

¹⁴ \mathbf{G} is invertible if $\alpha < 1$ (Bramoullé et al. (2009)) and the infinite sum is well defined if α is smaller than the norm of the inverse of the largest eigenvalue of \mathbf{G} (Ballester et al. (2006)).

¹⁵Particularly, any time trends or other dynamics, is to be eliminated by the Differences in Differences, if $\frac{df(\text{age})}{dt}$ is the same evaluated at $t-S$ and at t .

¹⁶To be more precise, the link formation and the way in which the characteristics of the nodes change over time have to be the same (common trends) in both networks. This guarantees that the counterfactual outcome of the treated network can be inferred from its own past and the evolution in the control network. The derivations require a lot of notational overhead and the resulting conditions are quite unwieldy. Assumption 4 would refer not only to $\Delta\mathbf{X}$, but to $\Delta\mathbf{G}\mathbf{X}$ to allow for relaxing Assumption 3.

¹⁷If this is the case, all estimates of indirect treatment effects, will reflect a sum of the treatment on the existing network and new spillovers due to the changes in the link network (cf. Comola and Prina (2013)), which will lead to upward biases if not accounted for.

not with the peak in clicks, and second, like for edits, the peak is large in relative, but small in absolute terms. I conclude that this is an acceptably small source of potential bias.

Estimating α : Analysis on the Node Level

Above we have shown what is measured by the difference-in-differences. From now on I shall refer to a node in the control condition by c and to a node in the treated condition by ℓ . If \mathbf{D}_t denotes the vector of treatments which is 1 at the treated nodes and 0 otherwise, estimation of the difference-in-differences identifies:

$$\mathbf{DiD} = \delta_1 \mathbf{D}_t (\mathbf{I} + \alpha \mathbf{G} + \alpha^2 \mathbf{G}^2 + \alpha^3 \mathbf{G}^3 + \dots) \quad (\text{A.12})$$

When we take the analysis back from the level of treated networks and look at the nodes individually, all that matters for each focal node j is its own row in this set of equations. To simplify this analysis I now introduce the local treatment assumption, exploiting the fact that only a single node in my network is treated each day. This is like a partial population treatment Moffitt (2001) with only one single node (a mini population) being treated.

Local Treatment Assumption: *Under the local treatment assumption $\mathbf{D}_t = \mathbf{e}_i$, where \mathbf{e}_i is an elementary vector with node i being the only treated node.*

If only one node is treated, the spillover dynamic is greatly simplified. With $\mathbf{D} = \mathbf{e}_i$, the only factor to be evaluated for each node is its corresponding ji element in the matrix \mathbf{G} , \mathbf{G}^2 and its higher orders.¹⁸ We distinguish a shocked node $\ell 0 \in L0$, a neighbor $\ell 1 \in L1$ and the indirect neighbors (2 clicks away, 3 clicks away etc.) as follows:

$$\begin{aligned} \ell 0 : \mathbf{DiD}_0 &= \delta_1 (1 + \mathbf{0} + \alpha^2 G_{ii}^2 + \alpha^3 G_{ii}^3 + \dots) \\ \ell 1 : \mathbf{DiD}_1 &= \delta_1 (\mathbf{0} + \alpha G_{ij} + \alpha^2 G_{ij}^2 + \alpha^3 G_{ij}^3 + \dots) \\ \ell 2 : \mathbf{DiD}_2 &= \delta_1 (\mathbf{0} + \mathbf{0} + \alpha^2 G_{ik}^2 + \alpha^3 G_{ik}^3 + \dots) \\ &\quad \text{etc.} \end{aligned} \quad (\text{A.13})$$

Sorting the nodes with respect to their distance from $\ell 0$ and estimating these strata separately results in as many estimation equations as can reasonably be traced and two parameters to be estimated. This fact is the basic idea of this paper, because it enables the researcher to back out the estimates for the structural parameters α and δ_1 . All that is needed is a sequence of reduced form difference-in-differences estimates for increasingly large link distances. If the precise information on \mathbf{G} and its higher orders is available the parameters can be directly estimated.¹⁹ If not, it is possible to compute an upper and a lower bound for the parameters α and δ_1 . In the next subsection I proceed to show how the boundary estimates can be computed.

¹⁸The information in the higher orders of the adjacency matrix \mathbf{G} is the same as the information from the sampling strategy in combination with knowing who was affected by the local treatment. Some nodes ($L0$) are known to be directly treated. Neighbors ($L1$) have a direct link so that the entry in \mathbf{G} that links them to the treated node is positive. However, for those who only have an indirect link, the corresponding entry in \mathbf{G} takes the value 0 and only the relevant element of \mathbf{G}^2 will be greater than 0.

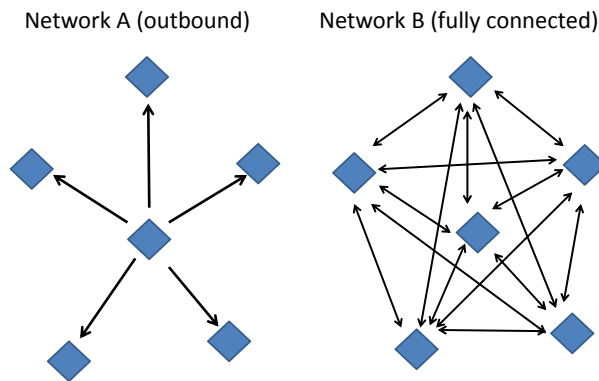
¹⁹To do this use all the ij values that correspond to each individual focal node j as weights for α , α^2 , α^3 , etc. and minimize a quadratic loss function. Unfortunately I cannot show this here, because the full matrix \mathbf{G} formed by the German Wikipedia is too large to be computed in memory.

A.2.4 Estimating Bounds for the Parameters of Interest

If the researcher lacks information on \mathbf{G} it is possible to compute an upper and a lower bound for the social parameter α and the treatment effect δ_1 . The goal in this section is to back out a lower and an upper bound estimate for α and δ_1 , that is based only on the estimated DiD 's and the number of nodes. This is useful, since the precise information on \mathbf{G} is often not easy to obtain.²⁰ In my proofs I use the local treatment assumption (only one individual in the network is treated), for both ease of notation and understanding. It applies to “today’s featured articles”.²¹

In what follows I will show how to obtain these bounds. In Subsection A.2.4, I will give an intuitive account of the underlying ideas. In Subsection A.2.4, I will set up the preliminaries, including a Lemma that will be used. Subsection A.2.4 obtains the upper bound and Subsection A.2.4, finally, provides the proof for the lower bound.

Figure A.1: Schematic representation of the two extreme networks, used to compute the upper and lower bound estimates of the parameters of interest.



NOTES: The “outbound network” (left) is used to obtain the upper bound estimate. It is a directed network with only “outward bound” links. Holding the number of nodes and the observed ITEs fixed, the social parameter will be estimated to be largest in this type of network. The fully connected network (right), is the benchmark case from which the lower bound of the social parameter can be estimated.

Intuition for obtaining Bounds

To see why we can bound the parameter, even without knowing the details of the network structure, we can select two ‘specific ‘extreme’ types of networks which either minimize or maximize the higher order effects. For greater convenience, I repeat the illustration of such networks in Figure A.1.

The network that minimizes higher order spillovers is a directed network with only “outward bound” links from ℓ_0 to $\ell_1 \in L_1$ ²². This implies no links between the

²⁰The information might either not be available, or so big that computing its higher orders might confront the researcher with substantial computational challenges.

²¹I conjecture that extending the proof to partial population or randomized treatments will be straight forward. It merely means taking into account that more than one node gets treated and that the effects from the treated can also spill to the other treated, which will render the formulas quite unwieldy.

²²and possibly further on to $\ell_2 \in L_2$, $\ell_3 \in L_3$ and so on.

nodes in $L1$ and will serve as upper bound. The opposite type of network is a network, where every node is the direct neighbor of every one of its peers.²³ The fully connected network simplifies the analysis, because it has only two types of nodes (treated or not). Higher order spillovers are the same for every node of the same type. Moreover, given α and N , the fully connected network has the greatest second and higher order spillovers.²⁴ This allows to derive a closed form solution for the lower bounds of the relevant parameters.

Preliminaries

Before I proceed to characterize the bounds of the coefficient, it is useful to point out a fact that will be important in the argument that follows. Start by rewriting the formulas in equation A.13 without explicit characterization of the higher order spills:

$$DiD_0 = \delta_1 + HO_{\ell_0} \quad (\text{A.14})$$

$$DiD_1 = \frac{\alpha}{NP_{\ell_1}}\delta_1 + HO_{\ell_1} \quad (\text{A.15})$$

where $HO_{\ell_0} = \delta_1(\alpha^2 G_{ii}^2 + \alpha^3 G_{ij}^3 + \dots)$ and $HO_{\ell_1} = \delta_1(\alpha^2 G_{ij}^2 + \alpha^3 G_{ij}^3 + \dots)$. These effects are typically not trivial. They depend on the underlying network of peers and need to take into account the network structure. However, I can use a simple insight concerning the size of the higher order effects.

Lemma A.1. *Given the total effect, larger higher order effects, imply smaller coefficients, i.e. for $DiD_0 > DiD_1 > HO^B > HO^A \geq 0$: for any $HO^A < HO^B$, $\alpha^A > \alpha^B$ and $\delta_1^A > \delta_1^B$.²⁵*

Proof. We have to make the following two comparisons:

$$\begin{aligned} DiD_0 &= \delta_1^A + HO^A & vs. & & DiD_0 &= \delta_1^B + HO^B \\ DiD_1 &= \frac{\alpha^A}{NP_{\ell_1}}\delta_1^A + HO^A & vs. & & DiD_1 &= \frac{\alpha^B}{NP_{\ell_1}}\delta_1^B + HO^B \end{aligned}$$

This can be transformed as follows:

$$\delta_1^A = DiD_0 - HO^A \quad vs. \quad \delta_1^B = DiD_0 - HO^B \quad (\text{A.16})$$

$$\alpha^A = \frac{(DiD_1 - HO^A)}{\delta_1^A} NP_{\ell_1} \quad vs. \quad \alpha^B = \frac{(DiD_1 - HO^B)}{\delta_1^B} NP_{\ell_1} \quad (\text{A.17})$$

From equation A.16 it is immediately obvious that $HO^A < HO^B$ implies $\delta_1^A > \delta_1^B$. For comparing α substitute the corresponding δ_1 from A.16 into A.17, define $HO^A := HO^B - \varepsilon$ (for $\varepsilon > 0$) and rewrite equation A.17 as

$$\alpha^A = \frac{a}{b} NP_{\ell_1} \quad vs. \quad \alpha^B = \frac{a - \varepsilon}{b - \varepsilon} NP_{\ell_1} \quad (\text{A.18})$$

²³I will sometimes refer to this network as “classroom” network.

²⁴Every node affects every other node via a direct link and everybody will get second and higher round spillovers from *every* other node.

²⁵Note that the requirement $DiD_1 > HO^B$ has bite, since it implies $\alpha < 0.5$. This assumption need not be satisfied in all applications, but it applies well to settings where the spills dissipate quickly and to settings where the direct effect on the treated is much larger than on the neighbors ($DiD_0 \gg DiD_1$). This is the case in most applications and certainly so in the present one.

where $a = (DiD_1 - HO^A)$ and $b = DiD_0 - HO^A$. Comparing α^A vs. α^B is equivalent to comparing $\frac{a}{b}$ vs. $\frac{a-\varepsilon}{b-\varepsilon}$. Since we have $a, b, \varepsilon > 0$, $\varepsilon < b$ and $\varepsilon < a$:

$$\begin{aligned} \frac{a}{b} - \frac{a-\varepsilon}{b-\varepsilon} > 0 &\Leftrightarrow a(b-\varepsilon) - b(a-\varepsilon) > 0 \\ &\Leftrightarrow a\varepsilon < b\varepsilon \\ &\Leftrightarrow^{b.A.} a < b \end{aligned}$$

The last inequality holds by the initial assumptions, which completes the proof. \square

With this lemma in hand we can now proceed to derive benchmarks (upper and lower bound estimates) for the parameters of interest.

Upper Bound: Network without higher order spillovers.

In the “outbound” network higher order spills back to the originating nodes do not exist²⁶: $HO_{\ell 0}$ and $HO_{\ell 1}$ would be 0. This is equivalent to assuming:

$$DiD =^{b.A.} \bar{\delta}_1 D_t (I + \bar{\alpha} G + \mathbf{0} + \mathbf{0} + \dots) \quad (A.19)$$

which is equivalent to having²⁷:

$$\begin{aligned} DiD_0 &= \bar{\delta}_1 && \text{for treated } L0 - \text{nodes} \\ DiD_2 &= 0 && \text{for } L2 \\ &&& \dots \text{analogously for } L3 \text{ and higher} \end{aligned} \quad (A.20)$$

By Lemma A.1 this assumption leads to an upper bound of both coefficients. If all effects are of the same sign and $DiD_0 > DiD_1 > HO > 0$ ²⁸, the difference-in-differences for a node $\ell 1 \in L1$ ²⁹ would simply reduce to:

$$DiD_1 = \frac{\bar{\alpha}}{NP_{\ell 1}} \bar{\delta}_1 \quad (A.21)$$

A consistent estimator of $\bar{\delta}_1$ and the observed difference-in-differences will be enough to estimate $\bar{\alpha}$. In the “outbound network”, I apply difference-in-differences on the level of directly treated nodes to obtain such an estimate. Then I move on to estimate $\bar{\alpha}$:

$$\begin{aligned} \hat{\delta}_1 &= \widehat{DiD_0} = \Delta \hat{\ell} 0 - \Delta \hat{c} 0 \\ \hat{\alpha} &= \frac{\widehat{DiD_1}}{\widehat{DiD_0}} NP_{\ell 1} \end{aligned} \quad (A.22)$$

with $\Delta \hat{\ell} 0 := \frac{1}{NP_{\ell 0}} * \sum_i (y_{i,\ell 0,t=0} - y_{i,\ell 0,t=1})$, $\Delta \hat{c} 0 := \frac{1}{NP_{c0}} * \sum_i (y_{i,c0,t=0} - y_{i,c0,t=1})$. The definition of $\Delta \hat{\ell} 1$ and $\Delta \hat{c} 1$ for the $\widehat{DiD_1}$ parallels the definition of $\Delta \hat{\ell} 0$ and $\Delta \hat{c} 0$.

²⁶Admittedly, in such a network, endogeneity would not be a problem in the first place.

²⁷ $D_{\ell 0}$ denotes the value of D at the central node, that is related to the focal node.

²⁸ DiD_0 (DiD_1) denotes the difference-in-differences for treated nodes (neighbors). For the reverse relationships ($DiD_0 < DiD_1 < HO < 0$) the estimate based on assuming an “outward bound” network gives a lower bound, if the effects go in opposite directions, my claims do not necessarily hold and will have to be verified by the researcher. Slightly more involved assumptions will be needed.

²⁹Which corresponds to an Indirect Treatment Effect or an “Externality”

Discussion: The assumption in equation A.19 implies no “multiplication-effects” or “feedback-loops” between the nodes.³⁰ In the light of the formalization presented here, this is a strong assumption. However, in the impact evaluation literature with fixed and stable classroom sizes or villages, this assumption is almost taken implicitly, whenever the researchers report merely the ATE and ITEs. (cf. Angelucci and De Giorgi (2009), Carmi et al. (2012), Dahl et al. (2012), etc. etc.).

Having said that, the upper bound estimator is quite suitable if higher order spillovers are negligible. In what follows I compute the lower bound estimates under the assumption of maximal higher order spillovers. This will give a sense of the maximal size of the bias that might result from assuming away the higher order complexities of a network.

Lower Bound: Network with maximum higher order spillovers.

In this subsection I derive the lower bound estimates under the assumption of a fully connected network. Formally, consider the matrix $\underline{\mathbf{G}}$, that corresponds to a fully connected network:

$$\underline{\mathbf{G}} = \begin{pmatrix} 0 & \frac{1}{N-1} & \frac{1}{N-1} & \cdots & \frac{1}{N-1} \\ \frac{1}{N-1} & 0 & \frac{1}{N-1} & \cdots & \frac{1}{N-1} \\ \frac{1}{N-1} & \frac{1}{N-1} & 0 & \cdots & \frac{1}{N-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{N-1} & \frac{1}{N-1} & \frac{1}{N-1} & \cdots & 0 \end{pmatrix}$$

First, observe that all nodes are direct neighbors, i.e. $NP_{\ell 0} = NP_{\ell 1} = NP_{\ell} = N - 1$. Next, note that there are only two types of nodes: Directly treated nodes and neighbors. Let us now characterize the higher order spillovers that arrive at the treated node. From equation A.13 we know that the spillovers that arrive at a node in L0 are given by:

$$\ell 0 : DiD_0 = \delta_1(1 + \mathbf{0} + \alpha^2 G_{ii}^2 + \alpha^3 G_{ii}^3 + \dots)$$

The formula above points out that no spillovers of order 1 arrive at the treated node, since i does not link on to himself.³¹ But in a network characterized by $\underline{\mathbf{G}}$, (and maintaining local treatment) the second order spillovers arrive from *every* neighbor, i.e. NP_{ℓ} times, third order spillovers arrive $(N - 1)^2 - (N - 1)$ times etc.³² The number of channels for spillovers of order S is given by:

$$\begin{aligned} \#channels_{ii,S} &= (N - 1)^{S-1} - (N - 1)^{S-2} + (N - 1)^{S-3} + \dots \\ &= \sum_{s=1}^{S-1} (N - 1)^s (-1)^{(S-1)-s} \quad S \geq 2 \end{aligned}$$

³⁰Neglecting higher-order spillovers is like implicitly introducing a temporal structure where a spillover takes time to occur and taking a snapshot after the first order effect. This is possible if, for example, spillovers are slow or if the temporal structure of the available data is fine grained enough.

³¹Note that this is precisely the point where the local treatment assumption is most useful, because had we treated $T > 1$ nodes, then we would have to count $T-1$ direct spillovers that arrive at i , which obviously would render the following considerations less tractable.

³²Counting the number of channels for third and higher order spillovers is a matter of combinatorics: The number of channels for higher order increases at an almost exponential rate, leading to potentially very large effects, that are moderated only by the decrease of the primary effects during transmission.

The sum of second and higher order spillovers arriving at the treated node is:

$$\begin{aligned} HO_{ii} &= \sum_{S=2}^{\inf} \delta_1 \frac{\alpha^S}{(N-1)^S} \#channels_{ii,S} \\ &= \sum_{S=2}^{\inf} \delta_1 \frac{\alpha^S}{(N-1)^S} \sum_{s=1}^{S-1} (N-1)^s (-1)^{(S-1)-s} \end{aligned}$$

All non-treated neighbors are the same and the number of channels for spillovers of order S from node i to node j is computed almost³³ in the same way:

$$\begin{aligned} \#channels_{ij,S} &= (N-1)^{S-1} - (N-1)^{S-2} + (N-1)^{S-3} + \dots \\ &= \sum_{s=0}^{S-1} (N-1)^s (-1)^{(S-1)-s} \quad S \geq 2 \end{aligned}$$

Again the sum of second and higher order spillovers at the neighboring nodes is:

$$\begin{aligned} HO_{ij} &= \sum_{S=2}^{\inf} \delta_1 \frac{\alpha^S}{(N-1)^S} \#channels_{ij,S} \\ &= \sum_{S=2}^{\inf} \delta_1 \frac{\alpha^S}{(N-1)^S} \sum_{s=0}^{S-1} (N-1)^s (-1)^{(S-1)-s} \end{aligned} \quad (\text{A.23})$$

Before we can move on to derive the lower bound estimates, note that we have $\sum_{s=1}^{S-1} (N-1)^s (-1)^{(S-1)-s} < (N-1)^{S-1}$ which will be a convenient fact for simplifying the estimation of the lower bound.

$$\begin{aligned} HO_{ii} &= \sum_{S=2}^{\inf} \delta_1 \frac{\alpha^S}{(N-1)^S} \sum_{s=1}^{S-1} (N-1)^s (-1)^{(S-1)-s} < \\ &< \sum_{S=2}^{\inf} \frac{\alpha^S}{(N-1)^S} (N-1)^{S-1} = \\ &= \frac{1}{(N-1)} \sum_{S=2}^{\inf} \alpha^S = \frac{\alpha^2}{(N-1)} \frac{1}{1-\alpha} \end{aligned} \quad (\text{A.24})$$

Let us call this expression $\overline{HO_{ii}}$. Analogously we obtain $\overline{HO_{ij}} = \frac{\alpha^2}{(N-1)} \frac{1}{1-\alpha}$. Plug these values into the equations A.14 and A.15 from above. With Lemma A.1 at our disposal, we can use $\overline{HO_{ii}}$ and $\overline{HO_{ij}}$ to back out the lower bounds of the coefficients α and δ_1 :

$$DiD_0 = \hat{\delta}_1 + \overline{HO_{\ell 0}} \quad (\text{A.25})$$

$$DiD_1 = \frac{\hat{\alpha}}{NP_{\ell 1}} \hat{\delta}_1 + \overline{HO_{\ell 1}} \quad (\text{A.26})$$

It is somewhat tedious, but straight forward to show, that solving this system of

³³s now starts at 0.

equations results in a quadratic equation for $\hat{\alpha}$:

$$\hat{\alpha}^2 - \left[\frac{DiD_0}{DiD_1} + (N-1) \right] \hat{\alpha} + (N-1) = 0 \quad (\text{A.27})$$

The closed form solution for $\hat{\alpha}$ is hence given by:

$$\widehat{\alpha_{1/2}} = \frac{1}{2} \left[\frac{DiD_0}{DiD_1} + (N-1) \right] + / - \sqrt{\frac{1}{4} \left[\frac{DiD_0}{DiD_1} + (N-1) \right]^2 - (N-1)} \quad (\text{A.28})$$

Under weak regularity conditions³⁴ one solution is above 1 and another one between 0 and 1. The latter one is the solution for $\hat{\alpha}$ and it can easily be used to retrieve $\hat{\delta}_1$ from equation A.14

Discussion: Note that this closed form solution requires only the number of nodes, and the two estimates from the difference-in-differences (for treated nodes and neighbors). It can be computed when nothing is known about the network, except how many agents and who was treated. It is thus as readily available as the upper bound estimators.

Clearly, one would immediately wish for more.³⁵ Having more information about the network structure or even the link strength between nodes is certainly desirable and, generally, will allow for more interesting additional results. Finally, while the proof here advantageously uses the local treatment assumption, I conjecture, that it is straightforward to extend it to treatments of more than one node.

A.3 Aside: Reaction to treatment of the neighbor

Everything above was derived under the assumption that nodes do not observe or at least do not react to the local treatment of their neighbors. This is appropriate for neighbors of Wikipedia articles that get advertised on the start page.³⁶ In general however, subjects might observe treatment of their neighbors and react to the fact.

An example are children at school, who get annoyed or jealous when their peer was treated in a nice way and they were not.³⁷ In such situations the students/villagers might react to *merely observing* the treatment of their neighbors by selecting a different value for the outcome variable. To model such a situation we need to further augment the model in equation A.2 by both the observable treatments (shocks) that are locally applied, and a term that captures the possible reaction to the treatment of the neighbor.

$$y_{it} = \alpha \frac{\sum_{j \in P_{it}} y_{jt}}{N_{P_{it}}} + X_{it}\beta + \gamma \frac{\sum_{j \in P_{it}} X_{jt}}{N_{P_{it}}} + \delta_1 D_{it} + \delta_2 \frac{\sum_{j \in P_{it}} D_{jt}}{N_{P_{it}}} + \epsilon_{it} \quad (\text{A.29})$$

Where δ_1 measures the direct treatment effect and the new coefficient δ_2 measures reactions of the node, when it “observes” treatment of one (or several) of its peers. Consider again two connected nodes, where one is treated ($\ell 0$) in period t and the

³⁴ $DiD_0 > DiD_1$, which is to be expected for most treatments and follows from $\alpha < 0.5$ and $N > 1$

³⁵ Note that if there is reason to believe that α is greater than 0.5 an analogue of Lemma 1 that relaxes my assumption of $\alpha < 0.5$ is required.

³⁶ For two reasons: (i) Wikipedia articles cannot react and (ii) the advertisement is not associated with any changes in the real world, so there is no reason for any updates.

³⁷ Other examples entail economic agents in a village, who observe that their neighbor was refused a social service for failure to comply with a requirement (e.g. sending their kids to school) or commuters in a city, who observe when their friends got caught (after the local transport authority increased the frequency of controls and the punishment for failure to present a valid ticket).

neighbors are not treated ($\ell 1 \in L1$). Assume for simplicity that $\ell 0$ is the only treated node in $\ell 1$'s neighborhood. Similarly, but different, we have:

$$\ell 0 :: y_{\ell 0 t} = \alpha \frac{\sum_{j \in P_{\ell 0 t}} y_{j t}}{N_{P_{\ell 0 t}}} + X_{\ell 0 t} \beta + \gamma \frac{\sum_{j \in P_{\ell 0 t}} X_{j t}}{N_{P_{\ell 0 t}}} + \delta_1 \mathbf{1} + \delta_2 \frac{\sum_{j \in P_{\ell 0 t}} \mathbf{0}}{N_{P_{\ell 0 t}}} + \epsilon_{\ell 0 t} \quad (\text{A.30})$$

$$\ell 1 \in L1 :: y_{\ell 1 t} = \alpha \frac{\mathbf{y}_{\ell 0 t} + \sum_{j \in P_{\ell 1 t} / \ell 0} y_{j t}}{N_{P_{\ell 1 t}}} + X_{\ell 1 t} \beta + \gamma \frac{\sum_{j \in P_{\ell 1 t}} X_{j t}}{N_{P_{\ell 1 t}}} + \delta_1 \mathbf{0} + \delta_2 \frac{\mathbf{1} + \sum_{j \in P_{\ell 1 t} / \ell 0} D_{j t}}{N_{P_{\ell 1 t}}} + \epsilon_{\ell 1 t} \quad (\text{A.31})$$

Now we get two types of spillover effects in this model: First the “pure spillover” α , due to the effect of treatment on the outcome of $\ell 0$. But second, also the “behavior change” of the node, δ_2 , when it “observes” treatment of its peer kicks in.

Applying a Difference in Differences strategy alone will measure the joint effect of these two “spillovers”. It will not identify α separately, unless δ_2 is believed to be 0. If this assumption is not warranted only the total “treatment-of-peer”-effect can be measured. Depending on the application we might care about the effect of treatments, in which case this aggregate effect will be interesting. It is simply important to be aware that it is not possible to identify the pure spillover effect in such a setting.

Appendix B

Appendix to Chapter 3:

*“Centrality and Content Creation
in Networks”*

B.1 Preparation of the Data and Definition of the Category Economics

We downloaded a full-text dump from the Wikipedia toolserver and constructed the time-varying graph of the article network on a weekly basis. In addition to the measures of an article’s network position, which lie at the heart of our analysis, we extracted data on additional variables like the length of the page, the number of authors or the categories to which the article belongs. Before computing the values of the variables, we accounted for the revisions that were made by small programs, so-called “bots”, which automatically make small formal changes to ensure that a consistent style is maintained throughout Wikipedia. We did not consider the revisions that were carried out by bots and we also excluded bots from the author count. In our analysis we use data on 153 weeks between December 2007 and December 2010. While articles have been selected from one category, network measures account for links between these articles and the entire German Wikipedia.

Because of the scale of the data in the order of magnitude of terabytes, it would be unthinkable to conduct the data analysis using only in-memory processing. We stored the data in a disk-based, relational database and queried the data using Database Supported Haskell (DSH) (Giorgidze et al. (2010) and (2011)), a novel high-level language allowing for formulation and efficient execution of queries on nested and ordered collections of data.¹

With this tool we sampled all the articles belonging to the categories and subcategories of economics (“Wirtschaft” - which may mean both “economy” and the discipline of economics in German) from this relational database. The choice of articles sampled was based on Wikipedia’s category tree. Even though the ordering is not purely hierarchical, articles that belong to a category are usually allocated among specific subcategories. The more general category is often not reported on the article page. Therefore we had to account also for subcategories if we wanted to ensure that our definition of a category is not too narrow. Consequently, to sample the pages belonging to economics, we extracted a list of the subcategories of that category and eliminated those which were too remotely related to economics. This procedure left us with a list of 380 subcategories. We then proceeded to identify all pages that were linked to one of the categories on the list during at least one week that lies within our period of observation, which resulted in a sample of roughly 19,000 articles. Sampling articles based on categories of content is an approach that has been used in previous papers dealing with large content networks like Wikipedia (cf. Halatchliyski et al. (2010)). However, when evaluating the information from the network formed by directed hyperlinks between articles, we do not rely exclusively on the subset of articles that we sampled. While we compute the social network measures only for the

¹DSH queries are automatically translated into efficient lower-level query languages that the underlying database system understands. For this study, we utilised DSH’s capability of translating high-level queries on nested and ordered collections of data to efficient bundles of SQL queries. For comparison, we have formulated several DSH queries used for the Wikipedia data analysis directly in SQL and found that the equivalent DSH queries were more concise, easier to write and easier to maintain. This was mostly due to DSH’s support for order, nesting, abstractions for query reuse and concise comprehension notation.

articles inside the category (i.e., for roughly 10,000 nodes), we use the links from all pages in the entire network (i.e., more than one million nodes) to compute them. This is different from previous work, where network measures are often computed only on subnetworks, that means abstracting from the existence of all the other articles. We therefore consider it to be of methodological interest to see whether estimating the effect of the network position on such a reduced network leads to a big or a small error. Hence, we define the category network as the set of nodes that remain within the category economics and the global network as the one that is set up by the entire German Wikipedia.

As we are interested in the network position within the entire Wikipedia, we have to handle the large mass of more than a million articles. We compute the number of incoming links and, using the *igraph* library by Csardi and Nepusz (2006), the closeness centrality for each article at every week. We do this for both the network corresponding only to the pages in the category and for the global network of all articles in the German Wikipedia. It is important to note that we carry out the entire analysis using the *directed* network formed via incoming hyperlinks. These links are observed and edited on those pages from which they direct away, but considered in our analysis as features of the pages which they are pointing to. On the latter pages they are only visible when using a tool provided on the side bar. Finally, Wikipedia collects all content about a topic on one single article and creates “redirect pages” for widely used synonyms that users might be looking for. These pages redirect users, who search for synonyms of the Wikipedia entry, almost silently to the main page. Before computing the network measures, we accounted for the existence of redirect pages, by counting a link to a redirect page also as a link to its target page.

B.2 Additional Descriptive Statistics and Robustness Checks

In Table B.1, we report four robustness checks of the last result: In the first column, we replace the contemporaneous measures of centrality by the ones from the week before, which cannot be influenced by current editing behavior. This tests whether our result is mainly driven by a reverse effect of content generation on incoming links created in the same week. We find virtually no change in the results and thus consider this effect not to be important. In the second column, we eliminate outliers from the sample. We observe two kinds of outliers visible in Tables 3.1 and 3.4: articles that gain a lot of attention in the form of long contributions, many authors and many links (both from the entire Wikipedia and within the category), and articles that experience very high changes in these variables in at least some periods. We compute maxima of levels and changes per article. We eliminate articles that lie in the extreme two percent for any maximal change. Of the remaining articles, we eliminate those lying above the 95th percentile of the maximal levels of any variable. In total this eliminates 15 percent of the articles. The results show that both indegrees are estimated to have even larger coefficients, which sum to 222 bytes for a link from within the category. The quadratic specification now better captures a positive but declining influence of links from outside the category.

Table B.1: Robustness checks for the relationship of page length and centrality.

	(1) Lagged cent.	(2) Excl. outliers	(3) Sociology	(4) Add clicks
Links from Wikipedia	2.946 (1.22)	62.605*** (8.01)	31.015*** (4.19)	2.940 (1.22)
(Links from Wikipedia) ²	0.001** (2.04)	-0.264*** (-5.07)	-0.011*** (-7.50)	0.001** (2.07)
Links from category	134.937*** (8.42)	159.688*** (7.07)	59.778* (1.71)	135.463*** (8.46)
(Links from category) ²	-0.125*** (-4.95)	-1.230** (-2.14)	-0.104*** (-2.74)	-0.126*** (-5.03)
Rel. closeness rank (Wikipedia)	7.505*** (3.10)	1.818 (1.23)	10.421 (1.28)	7.473*** (3.07)
Rel. closeness rank (category)	-1.182 (-0.65)	-4.608*** (-3.86)	-8.406* (-1.96)	-1.205 (-0.66)
Dummy: literature section	1248.652*** (5.92)	1002.577*** (9.60)	338.438 (0.77)	1247.455*** (5.93)
Age (in months)	8.396*** (22.30)	3.576*** (17.61)	11.154*** (9.62)	8.502*** (22.56)
Dummy: page is redirect	-718.429 (-0.74)	110.313 (0.39)	0.853 (0.00)	-771.635 (-0.77)
Clicks				0.233** (2.38)
Constant	2516.272*** (15.89)	1933.826*** (21.14)	3633.877*** (6.95)	2481.048*** (15.47)
Observations	1160520	994041	195381	1168155
Groups	7635	6497	1277	7635
Adj. R ²	0.130	0.227	0.095	0.131

t statistics in parentheses

2-way fixed effects OLS regressions with both time and article dummies (robust stand. errors)

Only articles connected over entire period were included.

* p<0.10, ** p<0.05, *** p<0.01

In the third column, we perform the estimation for a different category, sociology, excluding those articles that overlap with economics. As in our main sample, links from within the category have a much stronger effect on page length (roughly three times as large as a link from outside the category). However, this coefficient is significant only at the ten percent level, which may be a consequence of the smaller sample size. The coefficient for links from the entire Wikipedia is now significant, which was not the case for economics. Column 4 finally reports how the results change when including a proxy for how often a page was clicked in the last week.² Not surprisingly, as articles are clicked more often they get longer. However, the relationship of an article's network position and its length remains unaffected by the inclusion of this variable.³

Next we turn to the question whether the higher centrality is not only associated with more content but also with more authors. Table B.2 shows the two-way fixed effects regressions corresponding to equation 3.2. It mirrors the specifications from Table 3.5, but the regressions have now the number of authors as the dependent variable. Columns 1 and 3 show the results when using the centrality measures from the entire Wikipedia. The results indicate that an additional link is associated with roughly 0.11 more authors, with a very weak curvature of the slope. Similarly to our results for page length, the effect is much stronger for links from the category: an additional link from the category corresponds to approximately 0.54 more authors (considering the sum of Wikipedia and category coefficients). The coefficient for links from outside the category is much smaller but remains significant in all specifications. The closeness rank has a negligible effect in column 3, which turns insignificant in column 4.

²We measured the clicks in the 24 hours before the next due date in our weekly panel.

³We performed further robustness checks that did not affect the main conclusions. We excluded pages that merely redirected the reader to a different page and explanation pages. We also included a measure of how often pages linking to the page under consideration were viewed. Next we included several other (potentially endogenous) measures that better describe the pages (number of revisions, number of references). We repeated everything for authors, where some effects are somewhat reduced, but they are always in the same direction and continue to matter. The results are available from the authors upon request.

Table B.2: Relationship of number of authors and centrality.

	(1) Wiki degree	(2) Wiki & cat.	(3) add closeness	(4) all vars
Links from Wikipedia	0.112*** (4.25)	0.073*** (3.24)	0.111*** (4.23)	0.072*** (3.23)
(Links from Wikipedia) ²	-0.000** (-2.51)	-0.000** (-2.05)	-0.000** (-2.50)	-0.000** (-2.04)
Links from category		0.468*** (6.39)		0.476*** (6.38)
(Links from category) ²		-0.000*** (-3.06)		-0.000*** (-3.18)
Rel. closeness rank (Wikipedia)			0.017** (2.29)	-0.007 (-1.22)
Rel. closeness rank (category)				-0.009* (-1.65)
Dummy: literature section	1.552*** (4.78)	1.393*** (4.53)	1.543*** (4.76)	1.406*** (4.57)
Age (in months)	0.072*** (26.10)	0.064*** (44.22)	0.072*** (26.07)	0.064*** (43.66)
Dummy: page is redirect	0.269 (0.13)	-0.399 (-0.19)	0.220 (0.10)	-0.434 (-0.20)
Constant	6.127*** (13.05)	4.376*** (11.30)	5.291*** (13.73)	5.140*** (13.00)
Time dummies	Yes	Yes	Yes	Yes
Observations	1168155	1168155	1168155	1168155
Groups	7635	7635	7635	7635
Adj. R ²	0.463	0.495	0.463	0.495

t statistics in parentheses

2-way fixed effects OLS regressions with both time and article dummies (robust stand. errors)

Only articles connected over entire period were included. Dependent variable: number of authors.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Appendix C








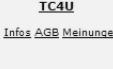



Additional Materials to Chapter 4:

*“Market Structure and Market
Performance in E-Commerce”*

C.1 Screenshot of the Price Comparison Site

Figure C.1 displays a screenshot of the price-comparison site www.geizhals.at. By default the shops are sorted according to their price (the lowest price listed on the first place). The first column (“Preis in EURO”) shows the price, followed by the shop’s name (“Anbieter”) the average consumer ratings for the shop (“Händlerbewertung”). Shipping cost and how readily the item is available (“Verfügbarkeit Versand”) are indicated in the fourth column. Several prices are usually quoted, depending on destinations and method of payment. Consumers can easily see how much they will be charged for shipping (shipping costs are quoted on the same line as prices). The screenshot also shows that consumers themselves have to build the sum of price and the relevant shipping cost, if they use the standard settings. In the rightmost column, most shops provide additional information on the product (“Artikelbezeichnung des Händlers”).

Figure C.1: Typical screenshot of the price comparison that is shown at geizhals.at

Preis in €*	Anbieter	Händler-Bewertung	Verfügbarkeit Versand*	Artikelbezeichnung des Händlers
311,-- 	 Hinweis: Firmensitz in Deutschland Infos AGB Meinungen	 Note: 2.02 214 Bewertungen	versandfertig in 3 Tagen Vorkasse € 9,90 Nachnahme € 15,90 Kreditkarte € 9,90	Nokia E 61 Blackberry Silber - Mobiltelefon ohne Vertrag ! (23.11.2007, 10:11)
336,--	 Infos AGB Meinungen	 Note: 1.20 850 Bewertungen	versandbereit in 1-2 Werktagen Vorkasse € 5,88 Nachnahme € 9,38 Express ab € 13,80 Abholung möglich (A-5550 Radstadt)	Nokia 0040446 Nokia - E61 UMTS-Mobiltelefon grau (E61) (0040446) (Art# 7A31909) (23.11.2007, 10:13)
348,24  Infos AGB Meinungen		 Note: 1.38 200 Bewertungen	Lagernd im Außenlager, 1-2 Werktage Abhol/Versandbereit (Ab Bestellbestätigung) Osterreich: Vorkasse € 5,- Nachnahme € 11,- Kreditkarte € 18,93 Express € 25,50 Deutschland: Vorkasse € 13,40 Nachnahme € 26,40 Kreditkarte € 27,33 Abholung nach Vereinbarung möglich (A-1170 Wien)	Nokia 05no0061 Nokia E61 grey/silver (E-)GPRS, HSCSD • WAP/MMS • Push-to-talk • OS: Symbian Series 60 (3rd Edition) • Quadband • Farbdisplay (16.7 Mio. Farben, 320x240 Pixel) • Vibracall • polyphone Klingeltöne • Video-/Mp3-Player • J.. (23.11.2007, 10:15)
363,--	 Infos AGB Meinungen	 Note: 1.70 22 Bewertungen	2-4Werktage Osterreich: Vorkasse € 5,- Nachnahme € 9,- Deutschland: Vorkasse € 10,-	Nokia E61 (23.11.2007, 10:11)
364,95	 Infos AGB Meinungen	 Note: 1.70 22 Bewertungen	Versandbereit in 2 - 4 Werktagen Stand: 23.11.2007, 14:56 Uhr Osterreich: Vorkasse € 4,80 Nachnahme € 9,90 Deutschland:	Nokia Nokia E61 ohne Bindung Nokia E61 grey/silver (E)GPRS/HSCSD (A) WAP/MMS Quadband RS-MMC-Slot Video/Mp3 144g (23.11.2007, 10:11)

NOTES: The Figure shows which information is shown to consumers on the price comparison web site. The shops' price quotes for a specific item are ordered by price net of shipping cost. The shops' consumer rating, the shipping cost, availability and the shop's own information about the item are provided with the price.

C.2 Construction of Substitutes

Based on the idea to identify the most frequently clicked pairs of products during the customers' search processes our calculation of substitutes is operationalized with the following steps:

(1) First, we identify the different consumers according to their individual `geizhals.at`-cookie which was downloaded to the users' computer by the Website and which uniquely identifies the user of the price search engine.

(2) Some users might search for very different products at the same time (e. g. a vacuum cleaner and a digicam), others are interested in similar products at different points in time (e. g. a consumer buys another camera one year later). We therefore have to sort the customers' click sequence (=referral requests) on `geizhals.at` into different search clusters.

(a) A search cluster includes only clicks in the same subsubcategory¹ and have to contain at least three clicks. To detect whether consumers search identical products at different points in time we apply Grubbs' outlier detection test with a significance level of 95% to identify separate time-separated search cluster. Moreover, we define a period of at least one week as a minimum time span between two clicks to separate a sequence of clicks into two different search clusters.

(3) All clicked products within such a search cluster are considered by the customer as potential substitutes. We exploit this information by measuring the incidence how often certain product pairs are clicked together. The resulting frequency tables for each subsubcategory depict that some product pairs are clicked very frequently together and other pair combinations can be observed only very rarely. We define the list of potential substitutes as the top two percent percentile of these frequency tables. This is a rather conservative measure of potential substitutes, which gives us a relatively low number of substitutes. Repeating the exercise with a lower cutoff point resulted in very similar results.

¹Geizhals maps its products hierarchically into categories, subcategories and subsubcategories to describe the similarity between the goods. Using the lowest hierarchical level in our analysis guarantees that only very similar products get into the customers' search spells.

Bibliography

- Aalto-Setälä, V. (2005), 'How do markets behave? The adjustment of price endings', *Journal of Product & Brand Management* **14**(7), 455–459.
- Adafre, S. F and M. de Rijke (2005), Discovering missing links in wikipedia, in 'Proceedings of the 3rd International Workshop on Link Discovery', pp. 90–97.
- Albert, R., H. Jeong and A.L. Barabási (1999), 'Internet: Diameter of the world-wide web', *Nature* **401**(6749), 130–131.
- Anderson, E. and D. Simester (2003), 'Effects of \$9 price endings on retail sales: Evidence from field experiments', *Quantitative Marketing and Economics* **1**(1), 93–110.
- Angelucci, Manuela and Giacomo De Giorgi (2009), 'Indirect effects of an aid program: How do cash transfers affect ineligibles' consumption?', *The American Economic Review* **99**(1), 486–508.
- Angrist, Joshua D and Jörn-Steffen Pischke (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.
- Aral, Sinan and Dylan Walker (2011), 'Creating social contagion through viral product design: A randomized trial of peer influence in networks', *Management Science* **57**(9), 1623–39.
- Ashenfelter, Orley and Michael Greenstone (2004), 'Using mandated speed limits to measure the value of a statistical life', *Journal of Political Economy* **112**(S1), S226–S267.
- Asplund, M. and R. Sandin (1999), 'Competition in interrelated markets: An empirical study', *International Journal of Industrial Organization* **17**(3), 353–369.
- Audretsch, D.B. and M.P. Feldman (1996), 'R&D spillovers and the geography of innovation and production', *The American Economic Review* **86**(3), 630–640.
- Ballester, Coralio, Antoni Calvo-Armengol and Yves Zenou (2006), 'Who's who in networks. wanted: The key player', *Econometrica* **74**(5), 1403–17.
URL: <http://ideas.repec.org/a/ecm/emetrp/v74y2006i5p1403-1417.html>
- Banerjee, Abhijit, Arun G Chandrasekhar, Esther Duflo and Matthew O Jackson (2012), The diffusion of microfinance, Technical report, National Bureau of Economic Research.
- Barber, Brad M and Terrance Odean (2008), 'All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors', *Review of Financial Studies* **21**(2), 785–818.
- Barron, J.M., B.A. Taylor and J.R. Umbeck (2004), 'Number of sellers, average prices, and price dispersion', *International Journal of Industrial Organization* **22**(8-9), 1041–1066.
- Basu, K. (2006), 'Consumer Cognition and Pricing in the Nines in Oligopolistic Markets', *Journal of Economics & Management Strategy* **15**(1), 125–141.
- Baye, M., J. Morgan and P. Scholten (2006), Information, search and price dispersion, in T.Hendershott, ed., 'Handbook Economics and Information Systems', Elsevier Press: Amsterdam.
- Baye, M.R. and J. Morgan (2001), 'Information gatekeepers on the internet and the competitiveness of homogeneous product markets', *American Economic Review* **Vol. 91, No. 3**, 454–474.
- Baye, M.R., J. Morgan and P. Scholten (2003), 'The value of information in an online consumer electronics market', *Journal of Public Policy & Marketing* **22** (1), 17–25.

- Baye, M.R., J. Morgan and P. Scholten (2004), 'Price dispersion in the small and in the large: Evidence from an internet price comparison site', *The Journal of Industrial Economics* **52**(4), 463–496.
- Baye, M.R., J.R.J. Gatti, P. Kattuman and J. Morgan (2009), 'Clicks, discontinuities, and firm demand online', *Journal of Economics & Management Strategy* **18**(4), 935–975.
- Benkler, Y. (2002), 'Coase's penguin, or, linux and" the nature of the firm"', *Yale Law Journal* pp. 369–446.
- Benkler, Y. (2006), *The wealth of networks: How social production transforms markets and freedom*, Yale Univ Pr.
- Benkler, Y. and H. Nissenbaum (2006), 'Commons-based peer production and virtue*', *Journal of Political Philosophy* **14**(4), 394–419.
- Bergemann, D. and J. Välimäki (2006a), 'Dynamic price competition', *Journal of Economic Theory* **127**(1), 232–263.
- Bergemann, D. and J. Välimäki (2006b), 'Dynamic pricing of new experience goods', *Journal of Political Economy* **114**(4), 713–743.
- Berger, Jonah, Alan T Sorensen and Scott J Rasmussen (2010), 'Positive effects of negative publicity: When negative reviews increase sales', *Marketing Science* **29**(5), 815–827.
- Berry, S. and P. Reiss (2007), 'Empirical models of entry and market structure', in: *M. Armstrong & R. Porter (ed.), Handbook of Industrial Organization, Volume 3, Chapter 29, Elsevier* pp. 1845–1886.
- Berry, Steven T. (1992), 'Estimation of a model of entry in the airline industry', *Econometrica* **60**(4), 889–905.
- Bramoullé, Yann, Habiba Djebbari and Bernard Fortin (2009), 'Identification of peer effects through social networks', *Journal of Econometrics* **150**(1), 41–55.
- Bresnahan, T.F. and P.C. Reiss (1991), 'Entry and competition in concentrated markets', *Journal of Political Economy* **99**(5), 977–1009.
- Brynjolfsson, E. and M.D. Smith (2000), 'Frictionless commerce? a comparison of internet and conventional retailers', *Management Science* **46**(4), 563–585.
- Campbell, J. R. and H. A. Hopenhayn (2005), 'Market size matters', *Journal of Industrial Economics* **53**, 1–25.
- Capocci, A., V.D.P. Servedio, F. Colaiori, L.S. Buriol, D. Donato, S. Leonardi and G. Caldarelli (2006), 'Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia', *Physical Review E* **74**(3), 036116.
- Carlson, J.A. and R.P. McAfee (1983), 'Discrete equilibrium price dispersion', *The Journal of Political Economy* **91**(3), 480–493.
- Carlton, D.W. (1983), 'The location and employment choices of new firms: An econometric model with discrete and continuous endogenous variables', *Review of Economics and Statistics* **65**(3), 440–449.
- Carmi, Eyal, Gal Oestreicher-Singer and Arun Sundararajan (2012), 'Is oprah contagious? identifying demand spillovers in online networks', *NET Institute Working Paper. No. 10-18.*, Available at SSRN: <http://ssrn.com/abstract=1694308> or <http://dx.doi.org/10.2139/ssrn.1694308> (August 3, 2012).
- Carpenter, Jeffrey and Caitlin Knowles Myers (2010), 'Why volunteer? evidence on the role of altruism, image, and incentives', *Journal of Public Economics* **94**(11), 911–920.
- Christie, W., J. Harris and P. Schultz (1994), 'Why did NASDAQ market makers stop avoiding odd-eighth quotes?', *Journal of Finance* **49**(5), 1841–1860.
- Christie, W. and P. Schultz (1994), 'Why do NASDAQ market makers avoid odd-eighth quotes?', *Journal of Finance* **49**(5), 1813–1840.
- Claussen, J., O. Falck and T. Grohsjean (2012), 'The Strength of Direct Ties: Evidence from the Electronic Game Industry', *International Journal of Industrial Organization* **30**(2), 223–30.
- Coase, Ronald H (1937), 'The nature of the firm', *economica* **4**(16), 386–405.
- Comola, Margherita and Silvia Prina (2013), 'The impact of exogenous interventions on networks: A dynamic peer effect model'.

- Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot and Philippe Zamora (2013), ‘Do labor market policies have displacement effects? evidence from a clustered randomized experiment’, *The Quarterly Journal of Economics* **128**(2), 531–80.
- Csardi, G. and T. Nepusz (2006), ‘The igraph software package for complex network research’, *InterJournal Complex Systems* **1695**.
- Dahl, Gordon B, Katrine V Løken and Magne Mogstad (2012), Peer effects in program participation, Technical report, National Bureau of Economic Research.
- Dasgupta, P. and J. Stiglitz (1980), ‘Industrial structure and the nature of innovative activity’, *Economic Journal* **90**(358), 266–293.
- Davis, Peter (2006), ‘Spatial competition in retail markets: Movie theaters’, *Rand Journal of Economics* **forthcoming**.
- De Giorgi, Giacomo, Michele Pellizzari and Silvia Redaelli (2010), ‘Identification of social interactions through partially overlapping peer groups’, *American Economic Journal: Applied Economics* **2**(2), 241–75.
- de Solla Price, D.J. (1965), ‘Networks of scientific papers’, *Science* **149**(3683), 510.
- Denning, P., J. Horning, D. Parnas and L. Weinstein (2005), ‘Wikipedia risks’, *Communications of the ACM* **48**(12).
- Dulleck, U., F. Hackl, B. Weiss and R. Winter-Ebmer (2011a), ‘Buying online: An analysis of shopbot visitors in Austria’, *German Economic Review* **12**(4), 395–408.
- Dulleck, Uwe, Franz Hackl, Bernhard Weiss and Rudolf Winter-Ebmer (2011b), ‘Buying online: An analysis of shopbot visitors’, *German Economic Review* **12**(4), 395–408.
URL: <http://dx.doi.org/10.1111/j.1468-0475.2011.00535.x>
- Dunne, T., M.J. Roberts and L. Samuelson (1988), ‘Patterns of firm entry and exit in U.S. manufacturing industries’, *Rand Journal of Economics* **19**(4), 495–515.
- Eichenbaum, M., N. Jaimovich and S. Rebelo (2011), ‘Reference Prices, Costs, and Nominal Rigidities’, *American Economic Review* **101**(1), 234–262.
- Ellison, G. and S.F. Ellison (2005), ‘Lessons about markets from the internet’, *Journal of Economic Perspectives* **19**(2), 139–158.
- Ellison, G. and S.F. Ellison (2009), ‘Search, obfuscation, and price elasticities on the internet’, *Econometrica* **77**(2), 427–452.
- Ellison, S.F. and C.M. Snyder (2011), ‘An empirical study of pricing strategies in an online market with high frequency price information’, MIT, *Department of Economics, Working Paper 11-13*.
- Ferracci, Marc, Grégory Jolivet and G. J. van den Berg (2012), ‘Evidence of treatment spillovers within markets’, online. last retrieved on Nov. 11, 2013 at http://www.efm.bris.ac.uk/ecgrjeval_resubmit.pdf.
- Fershtman, C. and N. Gandal (2011), ‘Direct and indirect knowledge spillovers: The “social network” of open-source projects’, *RAND Journal of Economics* **42**(1), 70–91.
- Gedenk, K. and H. Sattler (1999), ‘The impact of price thresholds on profit contribution - should retailers set 9-ending prices?’, *Journal of Retailing* **75**(1), 33–57.
- Gendall, P., J. Holdershaw and R. Garland (1997), ‘The effect of odd pricing on demand’, *European Journal of Marketing* **31**(11/12), 799–813.
- Geroski, P.A. (1989), ‘The effect of entry on profit margins in the short and long run’, *Annales d’Economie et de Statistique* **15-16**, 333–353.
- Ginzberg, E. (1936), ‘Customary prices’, *The American Economic Review* **26**, 296.
- Giorgidze, G., T. Grust, N. Schweinsberg and J. Weijers (2011), Bringing back monad comprehensions, in ‘Proceedings of the 4th ACM SIGPLAN Haskell Symposium, Tokyo, Japan’, ACM, ACM, pp. 13–22.
- Giorgidze, G., T. Grust, T. Schreiber and J. Weijers (2010), Haskell boards the Ferry: Database-supported program execution for Haskell, in ‘Revised selected papers of the 22nd international symposium on Implementation and Application of Functional Languages, Alphen aan den Rijn, Netherlands’, Vol. 6647 of *Lecture Notes in Computer Science*, Springer. Peter Landin Prize for the best paper at IFL 2010.

- Gorbatai, A. (2011), 'Aligning collective production with demand: Evidence from wikipedia', *Working Paper*.
- Gorbatai, A.D. and M. Piskorski (2012), 'Social structure of contributions to wikipedia', *Working Paper* downloaded from <http://www.wjh.harvard.edu/hos/papers/AndreeaGorbatai/AndreeaGorbatai.pdf>.
- Goyal, S., M.J. Van Der Leij and J.L. Moraga-González (2006), 'Economics: An emerging small world', *Journal of Political Economy* **114**(2), 403–412.
- Greenstein, S. and F. Zhu (2012a), 'Collective intelligence and neutral point of view: The case of wikipedia', *Working Paper*.
- Greenstein, S. and F. Zhu (2012b), Is wikipedia biased, in 'American Economic Review, Papers and Proceedings'.
- Greenstein, S. and M. Devereux (2009), Wikipedia in the spotlight, Technical Report 5-306-507, Kellogg School of Management.
- Griliches, Z. (1992), 'The search for r&d spillovers', *Scand. J. of Economics* **94**, 29–47.
- Guéguen, N. and P. Legohérel (2004), 'Numerical encoding and odd-ending prices', *European Journal of Marketing* **38**(1/2), 194–208.
- Guido, G. and A. Peluso (2004), 'Consumers' perception of odd-ending prices with the introduction of the Euro', *Journal of Product and Brand Management* **13**, 200–210.
- Halatchliyski, Iassen, Johannes Moskaliuk, Joachim Kimmerle and Ulrike Cress (2010), 'Who integrates the networks of knowledge in wikipedia?', *Proceedings of the 6th International Symposium on Wikis and Open Collaboration* pp. 1–10.
- Haynes, M. and S. Thompson (2008a), 'Price, price dispersion and number of sellers at a low entry cost shopbot', *International Journal of Industrial Organization* **26**(2), 459–472.
- Haynes, M. and S. Thompson (2008b), 'Entry and exit behavior at a shopbot: E-sellers as kirznerian entrepreneurs', *Unpublished Working Paper*.
- Hinz, Oliver, Bernd Skiera, Christian Barrot and Jan U Becker (2011), 'Seeding strategies for viral marketing: an empirical comparison', *Journal of Marketing* **75**(6), 55–71.
- Hitsch, G. (2006), 'An empirical model of optimal dynamic product launch and exit under demand uncertainty', *Management Science* **25**(1), 25–50.
- Ho, Jason YC and Melanie Dempsey (2010), 'Viral marketing: Motivations to forward online content', *Journal of Business Research* **63**(9), 1000–1006.
- Hoffman, Andrew J and William Ocasio (2001), 'Not all events are attended equally: Toward a middle-range theory of industry attention to external events', *Organization Science* **12**(4), 414–434.
- Holdershaw, Judith, Philip Gendall and Robert Garland (1997), 'The widespread use of odd pricing in the retail sector', *Marketing Bulletin* **8**, 53–58.
- Hossain, T. and J. Morgan (2006), '... Plus shipping and handling: Revenue (non) equivalence in field experiments on ebay', *Advances in Economic Analysis and Policy* **6**(2).
- Hu, Nan, Yi Dong, Ling Liu and Lee J Yao (2013), 'Not all that glitters is gold: the effect of attention and blogs on investors' investing behaviors', *Journal of Accounting, Auditing & Finance* **28**(1), 4–19.
- Huberman, Bernardo A, Daniel M Romero and Fang Wu (2009), 'Crowdsourcing, attention and productivity', *Journal of Information Science* **35**(6), 758–765.
- Imberman, Scott, Adriana D Kugler and Bruce Sacerdote (2009), Katrina's children: Evidence on the structure of peer effects from hurricane evacuees, Technical report, National Bureau of Economic Research.
- Jackson, Matthew O (2008), *Social and economic networks*, Princeton Univ Pr.
- Jacobson, Louis S, Robert John LaLonde and Daniel G Sullivan (1993), 'Earnings losses of displaced workers', *American Economic Review* **83**(4), 685–709.

- Jian, L. and J. MacKie-Mason (2012), Incentive-centered design for user-contributed content, *in* M. Peitz and J. Waldfogel, eds, 'The Oxford Handbook of the Digital Economy', Oxford University Press, Oxford, pp. 399–433.
- Kaiser, U. and S. Minjae (2009), 'Do media consumers really dislike advertising? An empirical assessment of the role of advertising in print media markets', *International Journal of Industrial Organization* **27**(2), 292–301.
- Kashyap, A. (1995), 'Sticky prices: New evidence from retail catalogs', *Quarterly Journal of Economics* **110**(1), 245–274.
- Kauffman, R. and D. Lee (2005), Should we expect less price rigidity in the digital economy? MIS, Carlson School of Management, University of Minnesota.
- Keegan, Brian, Darren Gergle and Noshir Contractor (2013), 'Hot off the wiki structures and dynamics of wikipedia's coverage of breaking news events', *American Behavioral Scientist* **57**(5), 595–622.
- Kittur, Aniket and Robert E. Kraut (2008), Harnessing the wisdom of crowds in wikipedia: quality through coordination, *in* 'Proceedings of the 2008 ACM conference on Computer supported cooperative work', CSCW '08, ACM, New York, NY, USA, pp. 37–46.
URL: <http://doi.acm.org/10.1145/1460563.1460572>
- Klepper, S. (1996), 'Entry, exit, growth, and innovation over the product life cycle', *The American Economic Review* pp. 562–583.
- Klepper, S. (2002), 'Firm survival and the evolution of oligopoly', *RAND Journal of Economics* **33**(1), 37–61.
- Konieczny, J. and F. Rumler (2006), Regular adjustment - theory and practice, Working Paper Series 669, European Central Bank.
URL: <http://ideas.repec.org/p/ecb/ecbwps/20060669.html>.
- Kriplean, T., I. Beschastnikh and D.W. McDonald (2008), Articulations of wikiwork: uncovering valued work in wikipedia through barnstars, *in* 'Proceedings of the ACM 2008 conference on Computer supported cooperative work'.
- Kuhn, Peter, Peter Kooreman, Adriaan Soetevent and Arie Kapteyn (2011), 'The effects of lottery prizes on winners and their neighbors: Evidence from the dutch postcode lottery', *The American Economic Review* **101**(5), 2226–47.
- Kummer, Michael E, Marianne Saam, Iassen Halatchliyski and George Giorgidze (2012), 'Centrality and content creation in networks – the case of german wikipedia', *ZEW-Centre for European Economic Research Discussion Paper* (12-053).
- Lacetera, N., D. Pope and J. Sydnor (2011), Heuristic thinking and limited attention in the car market, NBER Working Papers 17030, National Bureau of Economic Research, Inc.
URL: <http://econpapers.repec.org/RePEc:nbr:nberwo:17030>.
- Lerner, J. and J. Tirole (2002), 'Some Simple Economics of Open Source', *Journal of Industrial Economics* pp. 197–234.
- LeSage, James P (2008), 'An introduction to spatial econometrics', *Revue d'économie industrielle* (3), 19–44.
- Levy, D., D. Lee, H. Chen, R. Kauffman and M. Bergen (2011), 'Price points and price rigidity', *Review of Economics and Statistics* **93**(4), 1417 – 1431.
- Liang, J. and V. Kanetkar (2006), 'Price endings: magic and math', *Journal of Product & Brand Management* **15**(6), 377–385.
- Lynn, Michael, Sean Flynn and Chelsea Helion (2013), 'Do consumers prefer round prices? evidence from pay-what-you-want decisions and self-pumped gasoline purchases', *Journal of Economics Psychology*.
- Manski, Charles F (1993), 'Identification of endogenous social effects: The reflection problem', *The Review of Economic Studies* **60**(3), 531–42.
- Martin, S. (2012), 'Market structure and market performance', *Review of Industrial Organization* **40**(2), 87–108.
- Mazzeo, M.J. (2002), 'Product choice and oligopoly market structure', *RAND Journal of Economics* **33**(2), 221–242.

- Medelyan, O., D. Milne, C. Legg and I. H. Witten (2009), 'Mining meaning from Wikipedia', *International Journal of Human-Computer Studies* **67**(9), 716–754.
- Miguel, Edward and Michael Kremer (2003), 'Worms: Identifying impacts on education and health in the presence of treatment externalities', *Econometrica* **72**(1), 159–217.
- Moe, W.W. and S. Yang (2009), 'Inertial disruption: The impact of a new competitive entrant on online consumer search', *Journal of Marketing* **73**(1), 109–121.
- Moffitt, Robert A (2001), 'Policy interventions, low-level equilibria, and social interactions', *Social Dynamics* pp. 45–82.
- Monroe, K. (1973), 'Buyers'subjective perceptions of price', *Journal of Marketing Research* **10**(1), 70–80.
- Naipaul, S. and H. Parsa (2001), 'Menu price endings that communicate value and quality', *Cornell Hotel and Restaurant Administration Quarterly* **42**(1), 26.
- Palmon, O., B. Smith and B. Sopranzetti (2004), 'Clustering in real estate prices: determinants and consequences', *Journal of Real Estate Research* **26**(2), 115–136.
- Perloff, J.M. and S.C. Salop (1985), 'Equilibrium with product differentiation', *The Review of Economic Studies* **52**(1), 107.
- Piskorski, M.J. and A. Gorbatai (2010), 'Testing coleman's social-norm enforcement mechanism: Evidence from wikipedia', *Working Paper*.
- Priedhorsky, R., J. Chen, S. K. Lam, K. Panciera, L. Terveen and J. Riedl (2007), 'Creating, destroying and restoring value in wikipedia', *Proceedings of the 2007 International ACM Conference on Supporting Group Work* pp. 259–268.
- Ransbotham, S. and G. Kane (2011), 'Membership turnover and collaboration success in online communities: Explaining rises and falls from grace in wikipedia', *MIS Quarterly* **35**(3), 613–627.
- Ransbotham, S., G.C. Kane and N. Lurie (2012), 'Network characteristics and the value of collaborative user-generated content', *Marketing Science* **31**, 387–405.
- Romer, P.M. (1990), 'Endogenous technological change', *Journal of Political Economy* **98**, Number 5 (2)(5), 71–102.
- Rosenthal, R.W. (1980), 'A model in which an increase in the number of sellers leads to a higher price', *Econometrica: Journal of the Econometric Society* pp. 1575–1579.
- Sacerdote, Bruce (2001), 'Peer effects with random assignment: Results for dartmouth roommates', *The Quarterly Journal of Economics* **116**(2), 681–704.
- Scherer, F.M. and D. Ross (1990), *Industrial Market Structure and Economic Performance*, Third Edition, Boston: Houghton Mifflin Company.
- Schindler, R. (1991), 'Symbolic meanings of a price ending', *Advances in Consumer Research* **18**(1), 794–801.
- Schindler, R. and T. Kibarian (1996), 'Increased consumer sales response though use of 99-ending prices', *Journal of Retailing* **72**(2), 187–199.
- Schmalensee, R. (1989), 'Inter-industry studies of structure and performance', in: *R. Schmalensee & R. Willig (ed.), Handbook of Industrial Organization, Edition 1, Volume 2, Chapter 16, Elsevier* pp. 951–1009.
- Sehity, T., E. Hoelzl and E. Kirchler (2005), 'Price developments after a nominal shock: Benford's Law and psychological pricing after the euro introduction', *International Journal of Research in Marketing* **22**(4), 471–480.
- Seim, K. (2006), 'An empirical model of firm entry with endogenous product-type choices', *The RAND Journal of Economics* **37**(3), 619–640.
- Selten, R. (1973), 'A simple model of imperfect competition, where 4 are few and 6 are many', *International Journal of Game Theory* **2**(1), 141–201.
- Shaked, A. and J. Sutton (1987), 'Product differentiation and industrial structure', *Journal of Industrial Economics* **36**(2), 131–146.

- Shang, Jen and Rachel Croson (2009), 'A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods', *The Economic Journal* **119**(540), 1422–1439.
- Smith, M. and E. Brynjolfsson (2001), 'Consumer decision-making at an internet shopbot: brand still matters', *Journal of Industrial Economics* **49**(4), 541–558.
- Soto, J. (2009), Wikipedia: A quantitative analysis, PhD thesis.
- Stiving, M. and R. Winer (1997), 'An empirical analysis of price endings with scanner data', *Journal of Consumer Research* **24**(1), 57–67.
- Sutton, J. (1991), *Sunk Costs and Market Structure: Price Competition, Advertising and the Evolution of Concentration*, Cambridge, MA: MIT Press.
- Sutton, J. (1998), *Technology and Market Structure*, Cambridge, MA: MIT Press.
- Thomas, M. and V. Morwitz (2005), 'Penny wise and pound foolish: the left-digit effect in price cognition', *Journal of Consumer Research* **32**(1), 54–64.
- Thomas, M. and V. Morwitz (2009), 'Heuristics in numerical cognition: implications for pricing', *Handbook of pricing research in marketing* pp. 132–149.
- Toivanen, Otto and Michael Waterson (2000), 'Empirical research on discrete choice game theory models of entry: An illustration', *European Economic Review* **44**, 985–992.
- Toivanen, Otto and Michael Waterson (2005), 'Market structure and entry: Where's the beef?', *Rand Journal of Economics* **36**, 680–699.
- Varian, H.R. (1980), 'A model of sales', *The American Economic Review* pp. 651–659.
- Wikimedia-Foundation (2013), 'Frequently asked questions', online. last retrieved on Sept. 9, 2013 at <http://wikimediafoundation.org/wiki/FAQ/en>.
URL: http://wikimediafoundation.org/wiki/FAQ/en#What_are_your_plans.3F_Where_is_this_going.3F
- Wu, Fang and Bernardo A Huberman (2007), 'Novelty and collective attention', *Proceedings of the National Academy of Sciences* **104**(45), 17599–17601.
- Zhang, X. and F. Zhu (2011), 'Group size and incentives to contribute: A natural experiment at chinese wikipedia', *The American Economic Review* **101**, 1601–15.

Erklärung der Urheberschaft

Hiermit erkläre ich, die vorliegende Dissertation selbständig angefertigt und mich keiner anderen als der in ihr angegebenen Hilfsmittel bedient zu haben. Insbesondere sind sämtliche Zitate aus anderen Quellen als solche gekennzeichnet und mit Quellenangaben versehen.

Mannheim, August 22, 2014

Michael KUMMER

CURRICULUM VITAE

MICHAEL KUMMER

EDUCATION

09/2009–06/2014(*exp.*) **Economics:** PhD - Candidate at U. Mannheim (CDSE) and ZEW (Dept. for Economics of ICT)

10/2006–07/2007 **Economics:** Master 2 Recherche: Économie mathématique et économétrie at Toulouse School of Economics

10/2000–06/2005 **Economics:** Magister-degree at University of Vienna

PROFESSIONAL EXPERIENCE

Since 09/2009 **Researcher:** at ZEW in Mannheim, Germany (Department for Economics of Information and Communication Technologies); Field of Investigation: Empirical Analysis of Price Comparison Sites, Content Networks and Social Media.

10/2007–08/2009 **Researcher:** at Johannes Kepler University in Linz, Austria: Field of Investigation: Empirical Analysis of Market Structure, Entry and Competitive Behavior in Online Markets

TEACHING EXPERIENCE

Spring 2011–2014 Teaching: Assistant for the course "Markets and Strategies", Master Course taught by Prof. Martin Peitz

10/2004–06/2005 Tutorial for a course in applied econometrics at Vienna University. (three hours of teaching per week besides course work)

RESEARCH INTERESTS

Industrial Economics: structure of online markets, firm and consumer behavior online.
Web 2.0: Network effects and knowledge transfer in user generated content.
Social and Citation Networks, Causal Effects

PUBLICATIONS IN REFEREED JOURNALS

forthcoming Hackl, Franz, Michael Kummer, Rudolf Winter-Ebmer and Christine Zulehner (forthcoming), Market Structure and Market Performance in E-Commerce, *European Economic Review*, DOI: 10.1016/j.eurocorev.2014.03.007

2014 Hackl, Franz, Michael Kummer and Rudolf Winter-Ebmer (2014), 99 Cent: Price Points in E-Commerce, *Information Economics and Policy* 26,, 12-27.

REFEREED CONFERENCE PUBLICATION

2013 Giorgidze, George , Torsten Grust, Iassen Halatchliyski and Michael Kummer (2013), Analysing the entire Wikipedia history with Database Supported Haskell, in: Kostis Sagonas, Practical Aspects of Declarative Languages, 15th International Symposium, PADL 2013, Rome, Italy, January 21-22, 2013, Proceedings Series: Lecture Notes in Computer Science Bd. Vol. 7752, Rome, Italy

DISCUSSION PAPERS

11/2013 "Spillovers in Networks of User Generated Content - Evidence from 23 Natural Experiments on Wikipedia", 2013

12/2012 "Centrality and Content Creation in Networks - The Case of German Wikipedia" (R&R at IEP), Discussion Paper No. 12-05, jointly with Marianne Saam; Iassen Halatchliyski; George Giorgidze, Iassen Halatchliyski, 2012

WORK IN PROGRESS

"Customer Reviews as Public Information Good in Online Markets", jointly with Franz Hackl (U. Linz), 2010

"Market Entry in E-Commerce" NET Institute Working Paper No. 08-23, jointly with Maximilian Kasy (U. Harvard), 2008

PRESENTATIONS AT CONFERENCES AND WORKSHOPS

2014	5th Annual Conference on Internet Search and Innovation, SEARLE Center, Chicago 12th Annual International Industrial Organization Conference (IIOC), Chicago 7th ICT Conference at ParisTech, Paris
2013	3rd Conference on ICT, Munich 11th Workshop on Media Economics, Tel Aviv 11th Annual International Industrial Organization Conference (IIOC) - Rising Star Sessions, Boston 11th ZEW Conference on the Economics of ICT, Mannheim; EARIE Annual Conference, Évora; Sunbelt, Hamburg; 3rd SEEK - Conference, Mannheim; SEEK Workshop on Social Network Formation and Peer Effects (Mannheim);
2012	EARIE, Rome, Centre for Competition and Regulatory Workshop, Vienna, GOR, Mannheim
2011	Annual Conference: EARIE
2010	Annual Conferences: EEA, EARIE, German Economic Association
2009	ICT Workshop at ParisTech, 2009

SEMINAR PRESENTATIONS

MIT, Media Center (Boston), Northeastern U (Boston), WU Wien, U Kaiserslautern, U Linz, U Mannheim, ZEW Mannheim

RESEARCH GRANTS

2014	SEEK-Project: Side Effects of Economic Crises in Europe and the Provision of Online Public Goods. (Partners: M. Hinnosaar (Col. Carlo Alberto), T. Hinnosaar (Col. Carlo Alberto), Olga Slivko (ZEW), M. X. Zhang (HKUST), Peter Gloor (MIT))
2010	SEEK-Project: Firm Behavior over the Product Life Cycle. Entry, Exit and Pricing Strategies on Online Price-Comparison Sites (Partners: F. Hackl (U. Linz), R. Winter-Ebmer (U. Linz & IHS, Vienna) and C. Zulehner (U. Vienna))
2008	NET Institute grant for a paper on Firm Entry in E-Commerce Industries (Partners: M. Kasy (U. Berkeley))

GRANTS, PRICES AND AWARDS

"Best PhD Paper Award" at the IIIrd ICT Conference in Munich - 2nd Prize (2013)

Postgraduate Grant for non-German-speaking countries of the Federal Ministry of Science and Research. (finance for the Master at TSE; 2006-2007)

Scholarship for excellent course work in Economics ("Leistungsstipendium"), granted by the Austrian Government. (2003)

LANGUAGE AND SOFTWARE SKILLS

German (native), English, Spanish, French, STATA, Python, SQL, GAUSS, Matlab, E-Views, L^AT_EX, Microsoft-Office