Justus-Liebig-Universität Giessen

The Analysis of Unconventional Economic Datasets

Doctoral Thesis submitted to Justus-Liebig-Universität Gießen

Faculty of Economics and Business Studies

October 2016

(Disputation & Revision: February 2017)

by

Jochen Lüdering

Center for international Development and Environmental Research (ZEU)

Senckenbergstraße 3, 35390 Gießen, Germany

jochen.luedering@zeu.uni-giessen.de

Supervisors:

Prof. Dr. Peter Winker
Department of Statistics and Econometrics

Prof. Dr. Matthias Göcke
Department of Economics of Transformation
and Integration

Dedicated to my wife Silke, who made me take weekends off and thus, preserved my sanity.

Acknowledgments

In the process of writing this thesis I have benefited from advice offered by a large number of people, as a consequence any list would likely be incomplete. However, I want to single out a few of the people without whom I would have had a much more difficult time writing this thesis.

I like to express my deepest gratitude to my supervisor Prof. Dr. Peter Winker. He granted me all the necessary freedoms, and always found time for appointments with me in his tight working schedule. I also want to thank him for the fruitful cooperation in our joint publication and project work. I am also grateful that Prof. Dr. Matthias Göcke was willing to serve as my second advisor. His advice in particular for the first paper has been very helpful. I appreciate that Prof. Dr. Peter Tillmann and Prof. Dr. Christina Bannier agreed to also serve as committee members.

Further, I was lucky for my supportive colleagues at the Department of Statistics and Econometrics, the Center for international Development and Environmental Research (ZEU) and the rest of the university. At the department, discussions with Dr. Björn Fastrich, Dr. Alexandru Mandes, Daniel Grabowski and Johannes Lips have been a great asset. At ZEU I benefited from the interdisciplinary environment and the fruitful interaction with Dr. Matthias Höher, Dr. Iris Gönsch, Dr. Matthias Staudigel, Dr. Svetlana Feedoseva, Jennifer Heiny, Yvonne Dernedde, Laura Werner, Martin Wiesmair, Sarah Hüller, Björn Weeser and Jan Welsch. Apart from my immediate colleagues I also want to thank Prof. Georg Götz, Prof.

Jürgen Meckl (all Gießen University) and Prof. Guido Bünstorf (Kassel University) for their comments on my research.

I also thank my parents Änne and Heiner Lüdering for their support through-out my educational and academic career. Last but not least I want to mention my loving wife Silke who has always been there there for moral support. I am curious what the future will bring for us, now that we are both concluding our education.

Even though I was a public servant during my doctoral studies, I received financial support from the German Academic Exchange Service (DAAD) to attend the TPRC conference in Arlington, USA. The German Statistical Society (DStatG) supported my visit to CMStatistics 2016 in Sevilla, Spain and the publisher Lucius & Lucius financed student assistants for data cleaning. Of equal importance to me were the data supplied by third parties. Thus, I am grateful to Carna Botnet (`internetcensus2012.bitbucket.org`), Les Cottrell at SLAC (Stanford), `digizeitschreiften.de` and `IAFD.com`.

Contents

List of Abbreviations

2SLS	Two Stage Least Squares
3SLS	Three Stage Least Squares
DSL	Digital Subscriber Line
ICMP	Internet Control Message Protocol
ICT	Information and Communications Technology
ITU	International Telecommunication Union
IP	Internet Protocol (Version or Version 6)
FOMC	Federal Open Market Committee
FTTH/B	Fiber To The Home/Building
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
MNC	Maximum Neighbour Centrality
OLS	Ordinary Least Squares
OCR	Optical Character Recognition
PWT	Penn World Tables
QE	Quantitative Easing
SEM	Simultaneous Equation Model
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
VAR	Vector Autoregressive Model

1 Preface

The doctoral thesis at hand encompasses five research papers on three subject areas. Two manuscripts discuss the suitability of latency as measure for Internet quality across countries. The following two are concerned with the application of topic models and automatic classification of texts in an economic context, while the last paper suggests to combine social network analysis with survival analysis in order to estimate the impact of centrality on professional success.

At first glance the three research areas are very different and have little in common in terms of content. While the study of Internet quality fits into the macro-development and growth literature, the two papers on topic models are only similar in terms of applied method but address questions on monetary policy and economic history respectively. Finally, the study on social networks and success belongs in the field of labor economics, sociology or business economics.

At second sight one may realize that there are nonetheless some issues, which are common to the individual manuscripts: the datasets used in this analysis all consist of secondary data. This means that they were not originally intended to be used in economic analysis. Consequently, a lot of data preparation and cleaning was necessary before any econometric methods could be applied. As the data had not been used for this kind of research before, their economic analysis provides interesting new insights, which may not have been possible with conventional data.

Further, the datasets stand out for their complexity and size. Which, along with fast-speed of change, are the classical criteria for *Big Data*. This made it necessary to carefully select methods and tools to work around the associated difficulties.

The choice of the title "The Analysis of Unconventional Economic Datasets" shall emphasize the complexity and secondary nature of the data, the latter being a feature rather than a criterion of *Big Data*. Laney (2001) came up with three dimensions along which data might be big, which could also serve as criteria for a definition of *Big Data*. According to Laney (2001) the data can be changing fast, be large in size and/or of high complexity due to which the use of conventional tools and methods will be challenging. Based on the aforementioned criteria the manuscripts in this thesis deal with *Big Data* problems. However, I refrain from including *Big Data* in the dissertations title, as it has become a widely used buzzword, whose meaning has been diluted in the public perception. In addition, a comprehensive overview of Big Data applications in economics would be beyond the scope of this thesis.

1.1 Research Area I: The use of Internet latency data in Economic Analysis

The first area of research deals with a novel indicator for Internet quality, based on latency times of Internet communication. I argue that the widespread use of the penetration rate in economics might bias the results and lead to the impression that the *Digital Divide* was narrowing, due to the neglect of the quality dimension of Internet service.

In a first paper (Chapter 2) I start out by comparing existing measures of Internet quality and usage with the proposed indicator based on latency and analyze the demand and supply dynamics at work. Therefore, I employ data from Carna Botnet (2013), who collected latency data for every single host on the Internet and calculated per country median values. In a second step these values serve as inputs in a simultaneous equation model, where a certain Internet quality, respectively quantity, is explained by demand and supply. It turns out that there are some similarities but also notable differences between the introduced latency measure and the conventional measures found in other datasets.

A second paper (Chapter 3) builds upon these findings and introduces the latency measure in a growth model. As the original dataset did not have a time dimension I resorted to latency data collected by the PingER facility at Stanford University, which originally started to provide precise latency measures between research institutes from the field of particle physics. In terms of methodology it builds upon Röller and Waverman (2001) and Koutroumpis (2009), who use Telecommunication penetration and Broadband penetration rates respectively to determine the effect of these technologies on economic growth. I found huge effects of latency improvements on economic growth. These findings are in line with the existing literature on broadband penetration and economic growth.

1.2 Research Area II: Quantitative Text-Analysis

The second pillar of the dissertation project is concerned with the application of topic models for quantitative text-analysis. It consists of two separate works, both relying on topic models (more precisely *Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003)*) in terms of methodology. In topic models it is assumed that documents are generated from a predetermined set of topics. Hence, an algorithm is used to reverse the creation process and determine the underlying topics. These topics are generated endogenously, where the researcher only has to supply the number of topics the algorithm is generating. As each document is composed of words from several topics, the topics weights can be obtained and compared across time.

The first paper (Chapter 4), a joint research project with Peter Winker, analyses 51 years of volumes of the *Journal for Economics and Statistics (Jahrbücher für Nationalökonomie und Statistik)* where we examine the interplay between the economic discourse and observed economic data. Thereby, we try to get an impression whether economists can anticipate changes in key economic indicators or primarily discuss these changes after they have occurred. One of the economic measures we examine is the inflation rate. If economists were able - by the use of some models - to predict the future inflation rate, there should be an increase of publications concerned with inflation prior to an increase of the inflation rate. The German statistical office provides the long series on prices (from 1881 onward) on request, which serves as our benchmark. Unfortunately, in the case of inflation the scientific discourse appears to be following the actual economic developments. For other time series (e.g. unemployment and debt) we find that the debate precedes changes in the observed data.

Chapter 5 is a joint project with Peter Tillmann in which we use LDA to dissect the discussion regarding the Taper Tantrum period on Twitter. We analyze how the discourse on Twitter influences a set of US asset prices, by introducing the topic weights from the LDA estimation into a VAR model. We show that shocks to single topics have significant effects on U.S. bond yields, exchange rates and stock prices. Hence, we can conclude that the debate on social media matters for U.S. asset prices.

1.3 Research Area III: Network Centrality and Survival

The third area of research (Chapter 6) is concerned with the combination of social network analysis and survival analysis. To the best of my knowledge this combination of methods is a novel approach in econometrics.

In particular I want to use network centrality in a survival model in order to answer the question whether being well-connected is beneficial for a career. The “connectedness” is operationalized as a) a person’s betweenness centrality and b) the maximum centrality score of a person’s immediate neighborhood. Thus, even if an individual is not on any shortest path between any two nodes in a network, it counts as *well-connected* if one of its direct neighbors has a high value.

As the title “Standing and Survival in the Adult Film Industry” suggests, the analysis utilizes data on the pornography industry obtained from the *Internet Adult Film Database* (`IAFD.com`). The industry has the appealing features that inter-personal connections are very visible there, while hard to observe in other industries. Moreover, collaborations tend to be of short durations, as a consequence also changes in the career path become apparent.

The results indicate that being central in the collaborative network dramatically reduces the risk to leave the industry. While one may question the external validity of these findings due to "pornography" not being "just like any other job", it confirms the findings in the existing literature using a novel methodology and dataset.

Bibliography

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). "Latent dirichlet allocation". In: *Journal of machine Learning Research* 3, pp. 993–1022.

Carna Botnet (2013). "Internet Census 2012: Port scanning /0 using insecure embedded devices". Available via `http://internetcensus2012.bitbucket.org/`.

Koutroumpis, P. (2009). "The economic impact of broadband on growth: A simultaneous approach". In: *Telecommunications Policy* 33.9, pp. 471–485.

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Tech. rep. META Group.

Röller, L.-H. and L. Waverman (2001). "Telecommunications Infrastructure and Economic Development: A Simultaneous Approach". In: *American Economic Review* 91.4, pp. 909–923.

2 The Measurement of Internet Availability and Quality in the Context of the Discussion on Digital Divide

Jochen Lüdering

This article is available as J. Lüdering (2015). *The measurement of internet availability and quality in the context of the discussion on digital divide*. Discussion Papers 65 [rev.] Justus Liebig University Giessen, Center for international Development and Environmental Research (ZEU).

It has been presented at the annual congress of the German Statistical Association (Deutsche Statistische Gesellschaft) "Statistische Woche 2015", September 17th 2015 at HSU Hamburg. Early versions were presented at the MAGKS Colloquium August 23th 2014 in Rauischholzhausen and in the framework of the class on "Advanced Topics in Industrial Organization" by Prof. Dr. Goerg Götz, February 6th at Gießen University.

Acknowledgments As this has been my first paper, I was particularly thankful for comments and guidance from my supervisors Prof. Peter Winker and Prof. Matthias Göcke. Moreover, I want to thank my former colleagues at ZEU Dr. Svetlana Fedoseeva and Dr. Matthias Staudigel and the helpful comments offered by the participants in the MAGKS Class on Industrial Organization offered by Prof. Dr. Georg Götz.

Keywords Digital Divide, Internet, Telecommunication Infrastructure, Measurement, Bandwidth, Latency

JEL L96, O33, O57

Abstract

The operationalization of Internet quality and availability is of great importance when discussing the *digital divide*. The usage of the penetration rate (the share of Internet users in the population) – widely used in economic analysis – can easily be misleading in this debate, suggesting that the digital divide is narrowing. This appears to be an artifact of the data, as some industrialized countries are already close to a penetration rate of 100%, while it is still growing for developing countries.

I argue that one should focus on the study of Internet quality in a country rather than the number of users. To this end, I introduce a new latency-based measure to judge the quality of Internet, based on a novel dataset, and compare it to related measures. The results indicate that it may indeed be useful to measure Internet quality across countries. In particular the availability of the indicator for 247 countries and semi-autonomous regions makes it an interesting tool for policy analysis.

The possibility to examine the effects of different determinants on individual quantiles is particularly interesting. ICT investment appears to be strongly correlated with lower latency (better Internet quality) in the lower part of the distribution, while there appears to be little explained variation in the top of the latency distribution. In line with the theoretical discussion the results indicate that population density is an important determinant of latency.

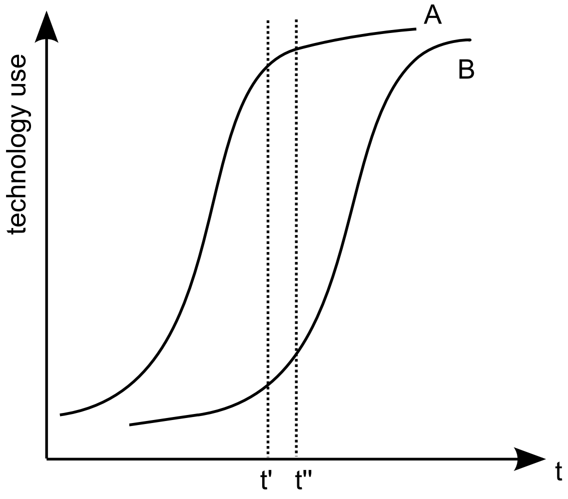
2.1 Introduction

The Internet is an infinite source of knowledge and an important tool of communication. It constitutes a potential input for economic development, as ideas spread easily and transaction costs in many fields are drastically reduced. Therefore, one could suspect that differences in availability and usage of Internet lead to differences in economic outcomes. This phenomenon, dubbed the "digital divide", provoked a fair amount of research in economics and related sciences. The studies of the consequences are related to the relationship between information and communications technology (ICT) use and growth (Dasgupta, Lall, and Wheeler 2001; Koutroumpis 2009; Czernich et al. 2011), inequality (DiMaggio and Hargittai 2001), and political participation (Sylvester and McGlynn 2010) on the one hand. On the other hand, there are important firm specific questions about the impact of ICT usage on productivity and innovation (Bertschek, Hogrefe, and Rasel 2015).

To mitigate the potential adverse effects of the digital divide the study of its determinants is important. One widely used approach is the study of Internet diffusion, which is based on the share of population in a country that uses the Internet *at all* (penetration rate). This measure has some weaknesses, as it disregards any information on connection quality, mean of access and utilization of Internet.

Based on this measure one could get the idea that developing countries are somehow catching up as suggested by Cuberes et al. (2010). The Internet penetration rate is approaching the upper bound of 100% in industrialized countries and Internet usage in developing countries is still increasing (indicated by the shift from t' to t'' in Figure 2.1). While the interpretation that the digital divide is narrowing might be a measurement artifact due to the ratio of Internet users in the population approaching the upper bound, it neglects important aspects of Internet access quality in terms of speed (latency and bandwidth), as well as reliability and availability. This is particularly troublesome as connections in developing countries tend to be unstable and the availability of access is often limited to a few international hotels and universities. When measuring the share of Internet users, the indicator does not reflect whether the users have occasional or regular Internet access.

Figure 2.1: The process of Internet diffusion in an industrialized country (A) and a developing country (B) following an S-shape



The aim of this paper is twofold: In a first step I address the question how the digital divide should be measured. For that purpose I discuss the suitability of latency as a measure of Internet quality and how it compares to the penetration rate and international bandwidth. For that purpose I introduce a novel dataset constructed from Carna Botnet (2013). In the second step I analyze the determinants of Internet provision and point out how these determinants differ across different indicators.

The remainder of this paper is structured as follows: Section 2.2 summarizes the debate on Internet diffusion and discusses different measures of Internet usage and quality in their applicability in this context. Section 2.2.2 in particular explains the technological and conceptual background of *latency* in relationship to (computer) networks. Moreover, I try to disentangle the two related terms of latency and bandwidth when it comes to Internet speed. Section 2.2.3 describes the nature of the data and the process of aggregation and closes with some descriptive results on the distribution of latency times across countries. Finally, in Section 2.3 I set up a simple model of demand and supply and estimate it for different measures of Internet quality and usage using three stage least squares (3SLS).

2.2 The Study of Internet diffusion

Research in the field of digital divide is strongly connected with the theory of technological diffusion. The epidemic models around which the theory of technology diffusion is based are dating back to Griliches (1957). The basic idea of these models is that exposure to a new invention in a neighboring region will lead to the adoption of the technology in the home region. The usage of the new technology grows exponentially at first and is later only slowly adopted by the more conservative producers, which leads to the famous S-shape depicted in Figure 2.1.

Examples include Chinn and Fairlie (2010), who apply Blinder-Oaxaca decomposition on data of Internet adoption and computer ownership, finding that income differences are the main source of the digital divide. Unfortunately, many of their explanatory variables are correlated with GDP and there are potential issues of endogeneity (e.g. the inclusion of electric power consumption). Other authors try to explain Internet penetration by introducing different socio-economic explanatory variables. Cuberes et al. (2010) test for network effects through the inclusion of lagged values of Internet usage. They try to address the resulting endogeneity concern by using an Arellano-Bond estimator. They claim to have found evidence of network effects, through the significant predictive power of the lagged number of Internet users. Wunnava and Leiter (2008) try to explain Internet penetration through income inequality (measured by a Gini-coefficient) in addition to the standard explanatory variables like telecommunication infrastructure, constructed from telephone and computer penetration.

However, in the context of epidemic models technologies are related to narrow applications (see Griliches (1957) original application to a new kind of hybrid corn). In contrast the Internet is perceived as a *General Purpose Technology* (Harris 1996) as it is very universal in its scope and just sets out a foundation for other technologies to be used on top. For this to function it requires substantial investment in ICT infrastructure to yield any returns. The applications build on-top of the infrastructure including simple technologies like Internet-based time synchronization (via ntp - network time protocol) as well as more contemporary inventions like Voice over IP telecommunication, BitCoin transactions and video conference systems.

The availability and the limits to the utilization of Internet, depend to a large extent on governments and telecommunication providers. The situation is in many cases similar to road infrastructure: I can connect my front door to the road, which is in most cases financed by the government. Nonetheless, whether my shoes get dirty on the way to work depends more heavily on whether the municipal road is paved, rather than on my own investment in the three meters between pavement and doorstep.

Therefore, epidemic models do not very well reflect the provision of Internet infrastructure in countries. A better foundation is the model underlying Röller and Waverman (2001) and Koutroumpis (2009), where Internet provision is determined by demand and supply for telecommunication and, respectively, broadband infrastructure to estimate the impact of ICT on economic growth.

2.2.1 The Different Facets of the "Digital Divide"

The choice of *measure* of the "Digital Divide" is of great importance. Using the number of users as a proxy for Internet infrastructure, is problematic. It omits any measure of quality but includes users regardless of their mean of access. The latter could be important as countries are very heterogeneous in terms of the composition of technologies used to access the Internet. Dial-up connections are used in areas where fixed-line phones are common. Wireless technology is - at least for telecommunication - very common in developing countries. Each technology has its own advantages in terms of availability and reliability on the one hand, and bandwidth and latency on the other hand. Moreover, the focus on users rather than hardware is likely to result in an underestimation of the digital divide, as private possession of computers is more pronounced in industrialized countries. The mean of access differs as well across countries. In industrialized countries, every user tends to have his or her own computer or Internet capable device, as well as their own broadband connection. In developing countries most users can only gain Internet access from libraries, universities, Internet cafes or at the workplace rather than at home.

Having or not-having an uplink does not fully reflect the access to information either. In the absence of net neutrality the flow of informationn may even be artificially constraint. There is anecdotal evidence (Mirani 2015) that there are more Facebook users than Internet users in developing countries, because Facebook is offering subsidized data plans which only allow the Facebook-App to access the Internet while its data plans prohibit the use of the free Internet.

Measuring the IT dispersion in terms of hosts or servers would result in even larger gaps - as the majority share of infrastructure is hosted in the United States and Western Europe, while its users, administrators and owners might be spread all over the world. Despite these potential limitations the measure of the number of hosts is used in the literature. The number of hosts (Kiiski and Pohjola 2002; Hargittai 1999) and the number of IPs (IP addresses) (Miner 2015) are, in this discussion, two sides of the same coin. IPs have the additional drawback that address space was allocated freely in the early days of Internet development and is scarce today. As a consequence of the previous generous allocation and acquisitions, the US company Hewlett-Packard currently holds two blocks of 16 million IP addresses compared to 28 million IPs allocated to all of Spain.[1] Depending on the actual measurement technique this might also bias the number of hosts. In some environments every printer might have a public IP and show up as a host, reachable from the outside. While in cases where IPs are scarce people increasingly use *network address translation*, where several computers or even households and institutions only receive one single public IPv4 address.[2] This critique

[1] According to `http://www.nirsoft.net/countryip/es.html`, accessed January 2014.

[2] For one current example from Germany see the recent policy of the cable provider Unitymedia who do

might be more relevant in a comparison across countries, than within a single country, as in the case of Miner (2015). But even there, it is likely that some institutions and firms receive IPs more generously than normal users. Additionally, servers tend to have several IP addresses, while workstations usually only have a single (often non-public) IP address.

The last dimension of interest in this discussion is the extend and way of Internet usage. While the discussion before was centered around capabilities, at the end of the day the actual application is what matters. On the micro-level there is one strain of literature (Pantea and Martens 2014; Goolsbee and Klenow 2006) concerned with the time spent online as measure of Internet availability. In these papers utility is derived from the product of time and capital investment in IT. However, today the marginal costs of Internet usage is approaching zero in industrialized countries and is equal for all users, due to common flat-rate tariffs. Consequently, the variability results only from differences in time constraints. In addition, there are countless application specific studies on the micro level measuring adoption of a specific technology. One of these is Hitt and Tambe (2007), who study the access to different categories of websites.

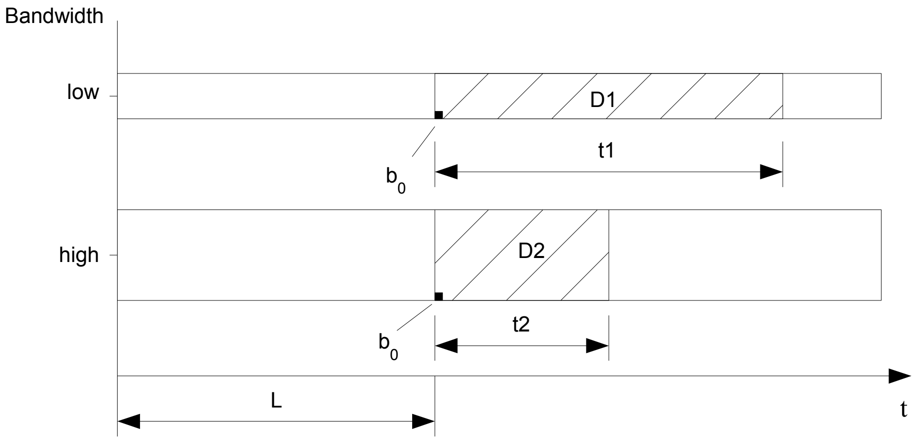
2.2.2 The Latency and Bandwidth Relationship

If one wants to measure the quality of Internet infrastructure, rather than its application, the usage of bandwidth and latency are plausible alternatives. These two values add up to the *experienced* Internet speed and are closely related. Figure 2.2 shows the relationship between the two measures, for a download of files of the same size (D1=D2) using a hypothetical low and a high bandwidth connection. Both horizontal bars in the diagram are hypothetical cables with high and respectively low bandwidth. The horizontal axis shows the flow of data through the cable over a time period t. After a user has requested data (D1/D2) time L (the latency) passes before the data starts to arrive. The actual transmission takes transmission

no longer provide a IPv4 Address per connection for consumers http://www.onlinekosten.de/news/artikel/51398/0/Unitymedia-Neukunden-erhalten-nur-noch-IPv6-Adressen.

times t1 and t2, which depend on the bandwidth. All other data transfer before and after the data blocks (D1/D2) are neglected in this example.

Figure 2.2: Transmission of equally sized data via low and high bandwidth and constant lantency (L)



Source: Own work loosely based on
`http://zoompf.com/blog/2011/12/i-dont-care-how-big-yours-is`

Latency (L) is the time for the first bit (e.g. b0) to reach its recipient, it is independent of the bandwidth. Its determinants are the technology used for transmission, distance and number of routers on the way and their respective load. The lower bound is given by the speed of light in a fiber optic cable. Consequently, if one wants improvement on that end, the only possibility are shorter, more direct cables. On the other end there are improvements to be made by increasing router capacity, which would potentially hold packages longer if the throughput is insufficient.

Bandwidth is the throughput of data usually measured in (mega)bits per second and is commonly the measure associated with the term "Internet connection speed". It has greatly expanded in recent years. Latency on the other hand has only gradually improved. For most ordinary applications the user will receive his disutility from the sum of latency and transmission time. Hence, improvements in latency and bandwidth are to some degree substitutes. Improving latency is rather costly, while increasing the bandwidth is comparatively cheap. As a consequence, latency increasingly matters for the transfer of small amounts of data in high bandwidth networks, as the actual transmission time tends towards zero as bandwidth increases. and only latency remains as "waiting time", that the user experiences when surfing the web.[3]

The ability to substitute these two inputs depends on the application. In particular synchronous communication relies on (reasonable) short latency. In particular voice communication relies on instant feedback for the speaker. Also for financial transactions (in particular high-frequency trading) low latency is of uttermost importance.[4] When watching a TV stream online, it may not be important for the viewing experience itself, whether one receives the data a few seconds later. However, hearing the neighbors (who might use a classic terrestrial antenna for TV reception) cheer before one does even sense an attempt by a striker in live football broadcast, might diminish one's own enjoyment from watching the world cup final.[5]

When measuring bandwidth, the method of aggregation is crucial. The international bandwidth per country as it is used in this paper is available from the ITU (International Telecommunication Union). It is a good measure to reflect potential technological bottlenecks, by comparing the bandwidth between countries. The international bandwidth is important as the majority of content providers reside in single countries (e.g. the United States or Ireland). On the other hand, Halavais (2000) finds that a lot of connections (in his hyperlinks on the web) are links to content in geographical proximity. Hence, for a lot of applications (e.g. surfing the web) the rest of the world does not matter very much, while for centralized services

[3] The share of latency in total transmission time is $\frac{L}{L+t}$. With technological advances the transmission time of small amounts of data tends towards zero. Hence, the relative importance of latency approaches unity.

[4] Some background information are available at http://www.informationweek.com/wall-streets-quest-to-process-data-at-the-speed-of-light/d/d-id/1054287? accessed 11.02.2015

[5] Zota (2014) showed the latency differences for Internet-based broadcasts in the wake of the of the 2014 Fifa World Cup.

(i.e. YouTube) it might be of great importance. However, the relative importance of the local hosting industry might differ between developing and developed countries. In developing countries, where domestic hosting services are unreliable, international bandwidth is likely of greater relative importance. This is due to the fact that users tend to use foreign provided ICT services if the local options are limited or unreliable. One example is the popularity of French E-Mail providers in Africa.

2.2.3 Description of the Latency Data

In the first parts of this paper the use of latency data as a proxy for Internet quality was proposed. In order to analyze the suitability, the empirical part of this paper mainly employs data from Carna Botnet (2013). The authors used a program to gain access to thousands of embedded computers with trivial default passwords settings, which were used to scan the whole Internet. The usage of compromised devices gave them access to a huge bandwidth, which allowed to perform bandwidth intensive tests and contact every host multiple times from different places around the earth throughout the last quarter of 2012.

This analysis focuses on the measure of ICMP[6] echo-requests, which yields the latency for a transmission between two clients. The requesting host sends out an echo-request (Ping) and the recipient answers with an echo-reply (Pong). The measured round-trip-delay is the latency between request and reply. It depends on the electrical signal transmit time, hence on distance, and on queues and processing in routers on route to the destination. The target hosts were assigned randomly and contacted multiple times from different sources. This means that the latency between one host and one random host on the Internet, should guarantee representative measures for the Internet as a whole. A small limitations stems from the fact that it can not be guaranteed that ICMP, which is used for control messages is

[6] The Internet Control Message Protocol, is used to transmit error and control messages in an IP based network.

treated exactly like its TCP and UDP equivalents used for data transmission. However, it should provide a good approximation of quality of TCP and UDP transmission.

Data Preparation and Aggregation In a first step the ICMP data has been purged of records indicating no response from the host. This could be for two reasons, either the IP Address is not assigned or the host was off-line at the time of the connection attempt. As there were several attempts to connect a certain hosts, chances are that it has been reached at least once. Nonetheless, it is likely that machines which are always on, are over-represented in the sample. Moreover, these machines are likely to have a faster connection (e.g. at government offices, telecommunication companies or universities) than those connected via dial-up. As this pattern would be the case for most countries, it should not influence the results on a cross-country basis.

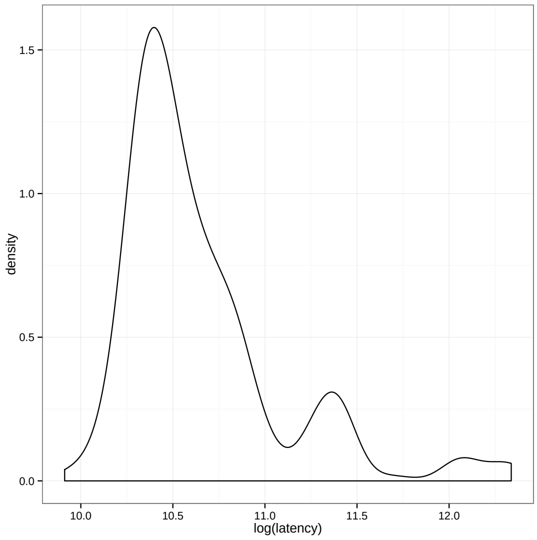
I aggregated the data on a per-IP basis and using Maxmind's GeoLite database[7] linked it to the country of origin. Out of the 594,050,059 hosts it was impossible to determine the location of 194,415 hosts in addition to 63,000 hosts associated with *Anonymous Proxy* service and no clear location. In order to reduce the number of observations to a manageable dataset I sampled the data on a per country basis and drew a random sample of 100,000 hosts per country. For countries with fewer observations, all observations are included. The distribution of latency in the sample is positively skewed. Hence, I used the median in the process of aggregation to mitigate the influence of outliers.

Visualization and Descriptive Statistics The skewed distribution found within countries (see Figure 2.5 on page 29) prevails for the country medians on inter-country level (see Figure 2.3).

The map in Figure 2.4 shows the geographical distribution of latency. As one would expect latency is high in Africa, South America, and parts of Asia, reflecting the general level of

[7] `http://dev.maxmind.com/geoip/legacy/geolite/` accessed June 10th 2013

Figure 2.3: Distribution of median-Latency across countries

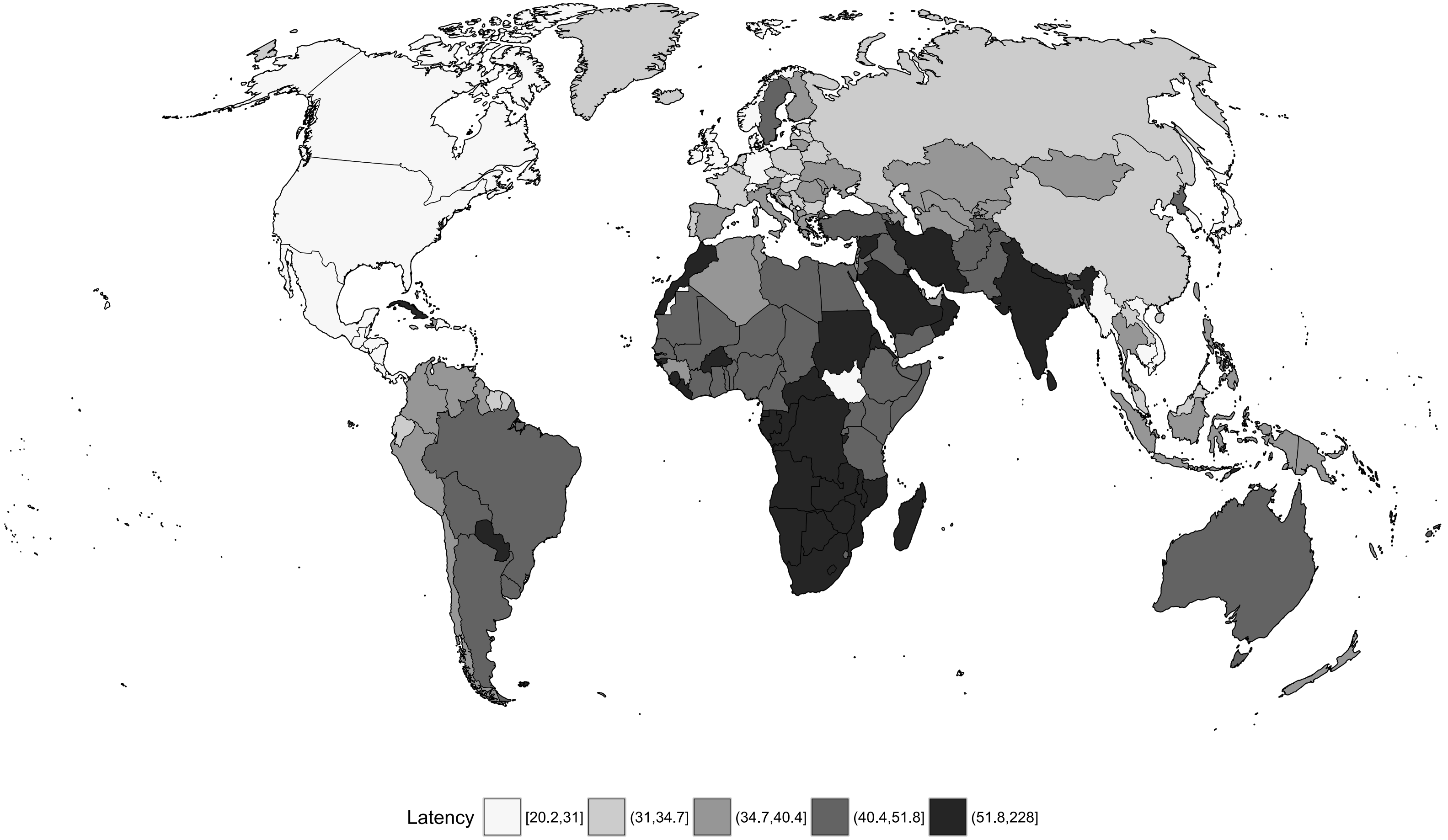


economic development in these regions. More surprising is the low latency found in Western Sahara and South Sudan. These findings coincide with a very low number of observations for these countries. As a consequence measurement errors are likely i.e. the computers in question might not even be located inside the border of the territories in question. Internet quality within these countries are likely comparable to (or slightly worse than) Morocco and respectively Sudan, who are or were controlling the territories.

Data Quality and Ethics Krenc, Hohlfeld, and Feldmann (2014) discuss data quality of the dataset. They point out that the methodology used to collect the data, was the novel part of the project and the reason for the media buzz, as complete scans of the Internet had been conducted earlier. I argue that only the used random assignment and various computers used as probes for the scan, allow to reflect the connectivity of a country to the whole Internet rather than to single reference points.

The data quality issue which they find, might be a concern, depending on the use of the data. In particular, they explained that it is difficult to disentangle the different waves of scans. I believe that on an aggregate level, e.g. the comparison of country medians in this

Figure 2.4: Geographic distribution of Latency



Note: For the sake of readability, latency is reported in milliseconds here, instead of representing the raw data in microseconds.

paper, the uneven number of scans (resulting from incomplete waves) and unequal intervals are of minor importance. However, it remains a concern if one would pursue an analysis using the number of ip-adresses, hosts or computers.

A last important question Krenc, Hohlfeld, and Feldmann (2014) touch upon is whether the usage of such "illegally obtained" data might be used from an ethical point of view, and what ethical codes of conduct apply when dealing with the data. They find that there is no consensus on this topic yet. I feel confident that this data may be used as no harm was done and the data is available publicly. Moreover, in the domain of econometrics of crime data gathered by criminals appears to be publishable. A famous example is Levitt and Venkatesh (2000), who discuss the finances of a drug selling gang. In this publication I deal with highly aggregated per country medians, implying that the publication of the results and data used in the analysis does not interfere with the privacy of individuals.

Comparison An interesting comparison can be made between median latency, bandwidth per user and share of Internet users. In Table 2.1 the countries are ranked according to each indicator and the top and bottom fifteen are shown. There are notable differences between the three rankings. It is striking that countries in Middle America are doing well in terms of latency, while the top twenty list for bandwidth per user is dominated by European countries. A peculiarity is the case of Cambodia, which ranks 10th in terms of latency but has the fifth lowest share of Internet Users. Unfortunately, there is no easy explanation for the different performance. But there is anecdotal evidence for a governmental investment which does not reach the majority of people yet.[8] In later parts of this paper I will examine the relationship between infrastructure and users. Of the countries with bad Internet, most are located in Sub Saharan Africa, South Asia and the Middle East.

[8] The Phom Penh Post reported on the 16. July 2009 that 2/3 of the country are now covered with fiber optical cable (See: `http://www.phnompenhpost.com/business/fibre-optic-cable-links-regions-data-networks`, accessed 18.06.2014)

Table 2.1: Country ranking according to three indicators

		“Best” Internet	
	Bandwidth per User	Share of Users	Latency
1	LUXEMBOURG	ICELAND	MACAO
2	HONG KONG	NORWAY	HONG KONG
3	MALTA	SWEDEN	JAPAN
4	SINGAPORE	DENMARK	KOREA, REPUBLIC OF
5	ICELAND	NETHERLANDS	MEXICO
6	SWITZERLAND	LUXEMBOURG	CANADA
7	SWEDEN	FINLAND	UNITED STATES
8	PORTUGAL	NEW ZEALAND	BELIZE
9	NORWAY	QATAR	BAHAMAS
10	UNITED KINGDOM	BAHRAIN	CAMBODIA
11	BELGIUM	UNITED KINGDOM	DOMINICA
12	DENMARK	CANADA	GUATEMALA
13	NETHERLANDS	ANDORRA	CURACAO
14	FINLAND	SWITZERLAND	DENMARK
15	ROMANIA	UNITED ARAB EMIRATES	SWITZERLAND

		“Worst” Internet	
	Bandwidth per User	Share of Users	Latency
1	IRAQ	MADAGASCAR	OMAN
2	GHANA	COTE D’IVOIRE	LESOTHO
3	CAMEROON	LESOTHO	SOUTH AFRICA
4	NIGERIA	MOZAMBIQUE	SUDAN
5	MADAGASCAR	CAMBODIA	PARAGUAY
6	ANGOLA	AFGHANISTAN	SYRIAN ARAB REPUBLIC
7	UZBEKISTAN	CAMEROON	KUWAIT
8	AFGHANISTAN	BANGLADESH	SAUDI ARABIA
9	TANZANIA	IRAQ	INDIA
10	SUDAN	RWANDA	NEPAL
11	NEPAL	PAKISTAN	MOROCCO
12	MOZAMBIQUE	LAO	ANGOLA
13	LAO	NEPAL	SRI LANKA
14	YEMEN	INDIA	ZAMBIA
15	ZAMBIA	NAMIBIA	IRAN

2.3 Determinants of Internet Adoption

After the discussion on the suitability of different indicators in order to measure Internet usage and quality, as well as the introduction of the data, in the previous section, the current section is dedicated to the determinants explaining Internet use and provision, as measured by the different indicators. The focus of the analysis will be on the novel latency measure, which I proposed in earlier parts of this paper.

2.3.1 A Simple Model of Demand and Supply

In order to identify the determinants of Internet infrastructure, I formulate a simple model illustrating the effects of demand and supply factors. The two forces jointly determine the equilibrium level of infrastructure provisioned. The scope of the model lies on providing a framework, which can be utilized the discussed indicators and compare their determining factors.

Demand is determined by income and access costs, which is the relative price of broadband service with respect to income. In the literature there is a debate on other potential other factors influencing demand. Wunnava and Leiter (2008) make a point that language barriers and education influence the utility gained from using the Internet, as these factors influence the understanding of online material and the amount of information available to the individual user. Nonetheless, in simple economic models it is usually assumed that demand for a good is independent from the utility gained from its consumption.

Assuming the simple case the demand equation only depends on prices and income and takes the form of:

$$y_{Di} = f(\text{Income}_i, \text{Price}_i) \tag{2.1}$$

Supply is determined by the amount of investment and associated costs of construction, the revenue from selling Internet services (Price) and market structure determining the pricing strategy of the firm. Leading to following supply equation.

$$\begin{aligned} y_{Si} = g(&\text{Investment}_i, \text{Cost of Construction}_i, \\ &\text{Price}_i, \text{MarketStructure}_i) \end{aligned} \tag{2.2}$$

In equilibrium demand and supply will be equal leading to: $Y^* = Y_{Si} = Y_{Di}$, which is the value we are likely to observe in the data.

Market Structure The market structure and the role of governments varies greatly across countries making it difficult to reflect it accordingly in the supply equation (Röller and Waverman 2001, p. 917). On the basis of oligopoly theory one would expect market structure to have an impact on prices and quantities. As an example assuming Cournot competition, oligopolists would reduce supply in order to charge a mark-up over marginal costs. With increasing competition one would expect increased supply and lower prices. On the other hand, a smaller number of firms in the market, might also increase the potential for governments influence.

The government objectives of involvement might also differ across countries and across time. In the past telecommunication had been regarded as a natural monopoly in the past, due to its high fixed costs. Hence, only governmental investments made it possible to supply Internet services. Only during the course of the 1990s governments began to liberalize the telecommunication market (for an overview see DICE Database 2009), after its operation became economically viable. Alternatively, governments may also artificially prohibit private operators to enter the market. In particular countries with democratic deficits may want to "control" the flow of information.

Consequently, it is convincing that government controlled monopolies differ from monopolized markets with private enterprises as governments often follow policy objectives, rather than operate profit maximizing enterprises. Due to anti-trust regulations, common in market based economies, and the objective of liberalization true profit maximizing monopolies are unlikely to exist.

Röller and Waverman (2001) dummy-out the US and Canada in the supply equation "Given the private market driven telecommunications suppliers" (Röller and Waverman 2001, p. 917). Following their approach, I also include a dummy variable to treat liberalized and non liberalized countries differently. The monopoly-dummy marks countries with just one single provider in order to capture the effect of government intervention in the market.

2.3.2 Empirical Analysis

The model described above can be estimated by simultaneous equation modeling, where the demand and supply equation are estimated jointly. In the empirical specification, demand is determined by income (as GDP per capita) and prices, measured by the *monthly charge for broadband connectivity*. Prices are considered to be endogenously determined by demand and supply, while income is considered to be exogenous.

Turning to the supply equation, the specification is the following: Cost is reflected by a countries population density (people per km^2). The idea is that a lower population density would lead to longer cables and, depending on the mean of access, more antennas and other equipment to serve the same number of people. In a recent paper, Götz (2013) showed that it may indeed be an important determinant of ICT infrastructure provision. As in the demand equation the monthly charge for broadband connectivity is included as a measure of Internet prices in the supply equation.

The investments in ICT infrastructure are aggregated over time, using a perpetual inventory method. The original data from the ITU database only includes investments flows rather than capital stocks. Certain ICT equipment deteriorates fast while some hardware remains in use for a long time. This is reflected by the discount function e^{-ax}. The calculation is described in more detail in Appendix 2.A.

The data on the monthly charge for broadband connectivity, the share of Internet users in the population and investment in ICT infrastructure originate from the World Telecommunication/ICT Indicators database 2013 (16th edition), while GDP per capita and population density were taken from the World Development Indicators online in February 2014.

Information on the market structure is scarce and the specificities differ between countries with respect to the number of competitors and their respective strategies. The level of competition is approximated by the number of providers active in the data. However, the data quality on the number of providers might not be very high.[9]

The following system of equations is jointly estimated by 3SLS treating the price as endogenous, instrumenting it by the exogenous variables not present in the respective equation. Thereby, one accounts for the fact that prices are jointly determined by demand and supply.

$$\begin{aligned} Y_{Di} &= \alpha_0 + \alpha_1 \text{MonthlyCharge}_i \\ &\quad + \alpha_2 \text{log}(\text{GDPCap}_i) + \varepsilon_{Di} \\ Y_{Si} &= \beta_0 + \beta_1 \text{MonthlyCharge} + \beta_2 \text{log}(\text{PopDensity}) \\ &\quad + \beta_3 \text{log}(\text{Stock}) + \beta_4 \text{log}(\text{providers}_i) \\ &\quad + \beta_5 \text{monopoly}_i + \varepsilon_{Si} \end{aligned}$$

Results

Table 2.2 shows the regression results of the SEM model across the different indicators, with OLS results provided for comparison. When comparing the regressions for the different indicators one should keep in mind that latency is a “negative‘” measure, with 0 ms representing instantaneous transmission. If the covariates had the same effect on the indicator, one would

[9] Taken from the 2008 issue of the CIA World Factbook (Central Intelligence Agency 2008), as current issues do not include information on providers

Table 2.2: Regression Results (SEM)

	log(Latency)				Log(Bandwidth/User)				log(Users/Pop)			
	OLS		3SLS		OLS		3SLS		OLS		3SLS	
	D	S	D	S	D	S	D	S	D	S	D	S
(Intercept)	11.030^{***}	10.817^{***}	10.936^{***}	10.862^{***}	6.013^{***}	7.815^{***}	3.866^{***}	1.147	-4.729^{***}	-3.172^{***}	-5.158^{***}	-6.699
	(0.161)	(0.165)	(0.503)	(0.483)	(0.490)	(0.695)	(0.818)	(7.770)	(0.290)	(0.454)	(0.579)	(4.362)
monthlycharge	0.001^{**}	0.000^{**}	0.003^{**}	0.001	-0.003^{***}	-0.003^{***}	-0.005^{**}	0.030	-0.001^{**}	-0.000	-0.004^{**}	0.021
	(0.000)	(0.000)	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)	(0.027)	(0.000)	(0.000)	(0.002)	(0.016)
log(gdpcap)	-0.049^{***}		-0.061		0.482^{***}		0.747^{***}		0.431^{***}		0.504^{***}	
	(0.018)		(0.052)		(0.055)		(0.084)		(0.033)		(0.059)	
log(PopDensity)		-0.017		-0.051^{**}		0.046		0.301		0.047		0.237
		(0.012)		(0.022)		(0.052)		(0.399)		(0.035)		(0.190)
log(Stock/Pop)		-0.061^{***}		-0.133^{**}		0.438^{***}		1.558^{*}		0.415^{***}		1.143^{**}
		(0.019)		(0.059)		(0.078)		(0.919)		(0.049)		(0.540)
log(providers)		-0.009		-0.012		0.107^{*}		0.435		0.090^{**}		0.134
		(0.012)		(0.028)		(0.054)		(0.494)		(0.037)		(0.254)
monopoly		0.053		0.244^{**}		0.103		-2.172		0.007		-2.259^{**}
		(0.060)		(0.111)		(0.265)		(2.068)		(0.180)		(0.973)
Adj. R^2	0.130	0.185	-1.249	0.300	0.565	0.436	0.583	-12.127	0.672	0.547	0.399	-14.426
Num. obs.	105	105	105	105	105	105	105	105	105	105	105	105

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

expect the signs of the coefficients to be reversed compared to the bandwidth and Internet penetration rate.

The price (MonthlyCharge) appears to influence Internet use and quality in a similar manner. The effects appear to influence the results primarily from the demand side, where the coefficients are very similar. A change of one dollar in subscription prices results in a change of 0.3% in latency, 0.5% in bandwidth per capita and 0.4% in the penetration rate (Users/Pop). Income has *ceteris paribus* no effect on effect on latency, while a 1% increase in income leads to 0.7% higher bandwidth per capita and 0.5% increase in the penetration rate.

A significant effect of population density can only be observed on the latency, where the coefficient is even significant at the 95% confidence level. A one percent higher population density leads to a reduction of latency times of 0.5%.

As expected, there is a significant influence of the accumulated stock of ICT capital on User/Pop as well as Bandwidth/User (only on the 10% level) – and the coefficient has, as expected, the opposite sign for the model explaining latency. The magnitude of the effect differs across the three indicators while change in 1.5% Bandwidth per User for a 1% change in ICT capital, it is only -0,13% for latency. The difference in magnitude of two coefficients gives some support to the fact that latency improvements are more difficult to achieve than improvements of bandwidth.

When interpreting the penetration rate - an increase of 1.15% for a 1% increase in ICT capital, the question of reverse causality arises. However, I am convinced that the number of users today does have little influence on the *accumulated* ICT capital of the past years.

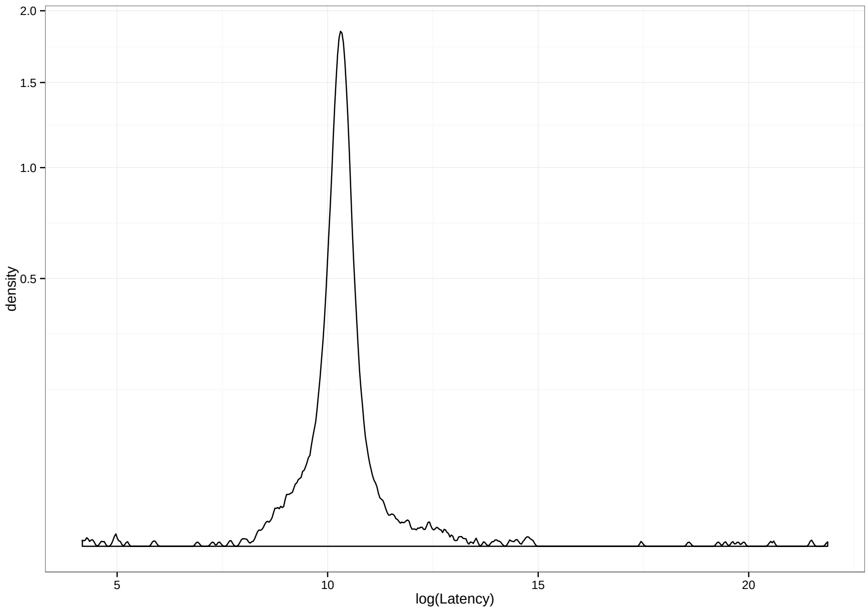
The log number of providers in a country appears to be significant in the OLS specifications. While the monopoly dummy for a non competitive market structure is significant for two specification. Having a monopoly or non-liberalized market leads to a 27.6% higher latency and a hypothetical reduction of the share of Internet users in the population by -89.9%.

As the R^2 is not very useful to interpret in the 3SLS estimation. One can only argue that the R^2 from the OLS estimation hints at the fact the model explains more of the variation of Bandwidth per User and the User share of the population then of latency. This results likely from the strong correlation of income with the first two measures and no significant correlation between income and the latter measure.

Different Quantiles

It is obvious that one possibility for differences in the results between different indicators might be a consequence of the choice of aggregation methods. While penetration rate (User/population) and bandwidth per user are the means, I used the median to analyze the effects on latency. This choice was necessary due to the skewed distribution of latency times (See Figure 2.5 as an example). Hence, the question arises if the effect of the covariates on latency differs if one regards quantiles other than the median.

Figure 2.5: Distribution of log(Latency) in the Germany



Note: In order to improve readability, the square root is used to scale density.

The results for the 10th and 90th percentile, as well as the 1st and 3rd quartile in comparison to the median is shown in Table 2.3. For the 3rd quartile and the 90th percentile only price

Table 2.3: Regression Results (Quantiles)

	10th		25th		50th		75th		90th	
	D	S	D	S	D	S	D	S	D	S
(Intercept)	10.691^{***}	10.116^{***}	10.856^{***}	10.488^{***}	10.936^{***}	10.862^{***}	11.046^{***}	10.995^{***}	11.208^{***}	11.192^{***}
	(0.245)	(0.546)	(0.309)	(0.489)	(0.503)	(0.483)	(0.596)	(0.754)	(0.706)	(1.117)
monthlycharge	0.002^{**}	−0.001	0.002^{**}	−0.001	0.003^{**}	0.001	0.003^{*}	0.002	0.004^{*}	0.004
	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)	(0.003)	(0.002)	(0.004)
log(gdpcap)	-0.072^{***}		-0.070^{**}		−0.061		−0.033		0.004	
	(0.025)		(0.032)		(0.052)		(0.061)		(0.072)	
log(PopDensity)		-0.105^{***}		-0.087^{***}		-0.051^{**}		−0.028		0.002
		(0.028)		(0.025)		(0.022)		(0.018)		(0.018)
log(Stock/Pop)		-0.141^{**}		-0.142^{**}		-0.133^{**}		−0.072		0.009
		(0.065)		(0.058)		(0.059)		(0.100)		(0.151)
log(providers)		−0.026		−0.026		−0.012		−0.004		0.002
		(0.035)		(0.031)		(0.028)		(0.035)		(0.049)
monopoly		0.203		0.305^{**}		0.244^{**}		0.086		−0.005
		(0.144)		(0.129)		(0.111)		(0.081)		(0.064)
R^2	−0.137	−0.713	−0.614	−0.266	−1.206	0.333	−0.724	−0.122	−0.573	−0.645
Adj. R^2	−0.159	−0.799	−0.646	−0.330	−1.249	0.300	−0.758	−0.179	−0.604	−0.728
Num. obs.	210	210	210	210	210	210	210	210	210	210

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

remains significant. In the lower quantiles the effect of population density and the stock of ICT capital becomes stronger compared to the median case. Moreover, the coefficient for income becomes highly significant.

Again, it is important to bear in mind that higher percentiles, imply longer latency times and, hence, worse Internet quality. The empirical analysis shows that the model explains supply and demand for "'high quality" Internet, while only prices and the intercept remain significant for 75th and 90th percentile.

2.4 Conclusion

After a brief survey of the existing literature on investment in Internet infrastructure, I introduced a novel measure of Internet quality based on latency. This measure has advantages over existing ones, in particular the widespread use of the Internet penetration rate. As pointed out, latency is closely related to infrastructure quality. Moreover, its relative importance with respect to bandwidth increases when bandwidth becomes large, even for day-to-day activities like surfing the web. Additional advantages include the possibility to measure it directly over the Internet, compared to the survey-based collection of bandwidth and user data. Having data for 247 countries and territories yielded little additional benefit as the analysis is constraint by a large share of missing observations in the explanatory variables, in particular ITU database. Nonetheless for descriptive purposes it is nice to have data for as many regions as possible.

The model developed in this paper explains more of the variation of bandwidth per user and the user share as compared to latency. There are notable differences in the correlation between the measures of Internet availability and quality and the explanatory variables, which supports the idea that each measure is related to a distinct aspect of Internet quality. Consequently, the measures also differ in terms of policy implications. Latency can only be improved by shorter fiber optic cables, which require a certain population density to be cost effective. The fact that bandwidth and user share are strongly correlated with income appears reasonable, as both measures can easily be improved by additional connections to neighboring countries, which might be the result of a higher level of competition in the market. The share of users could be increased by supporting Internet Cafes, supporting Internet access in public institutions or subsidizing private Internet connections. Both goals might be easy to achieve in a country's capital, whereas improving median latency is costly. The empirical analysis has shown, how one could improve Internet quality for lower quantiles (e.g. where latency is relatively low), we learn rather little about the long tail of high latency. Either these cases are rather heterogenous or other determinants are at work which determine high latency.

It appears that low population density, which increases the infrastructure costs per household is hindering improvements in Internet quality and as a consequence may lead to a growing digital divide. Considering the evidence from the literature subsidizing backbone infrastructure in less densely populated areas likely yields significant economic return. It is left for future research to validate these findings with the newly introduced latency indicator.

Unfortunately, it is impossible to make any inference about causality in the cross section. Nonetheless, I hope that my contribution provokes additional research in the field of measures of Internet quality, in order to put the discussion on digital divide on a more solid footing.

Bibliography

Bertschek, I., J. Hogrefe, and F. Rasel (2015). “Trade and technology: new evidence on the productivity sorting of firms”. In: *Review of World Economics (Weltwirtschaftliches Archiv)* 151.1, pp. 53–72.

Carna Botnet (2013). “Internet Census 2012: Port scanning /0 using insecure embedded devices”. Available via `http://internetcensus2012.bitbucket.org/`.

Central Intelligence Agency (2008). *The CIA World Factbook*. old edition retrieved from Project Gutenberg (http://www.gutenberg.org/).

Chinn, M. D. and R. W. Fairlie (2010). “ICT Use in the Developing World: An Analysis of Differences in Computer and Internet Penetration”. In: *Review of International Economics* 18.1, pp. 153–167.

Cuberes, D., L. Andres, T. Serebrisky, and M. A. Diouf (2010). *The diffusion of Internet: a cross-country analysis*. Working Papers. Serie AD 2010-07. Instituto Valenciano de Investigaciones Economicas, S.A. (Ivie).

Czernich, N., O. Falck, T. Kretschmer, and L. Woessmann (2011). “Broadband Infrastructure and Economic Growth”. In: *Economic Journal* 121.552, pp. 505–532.

Dasgupta, S., S. Lall, and D. Wheeler (2001). *Policy reform, economic growth, and the digital divide - an econometric analysis*. Policy Research Working Paper Series 2567. The World Bank.

DICE Database (2009). *History of Telecommunication Liberalization*.

DiMaggio, P. and E. Hargittai (2001). *From the 'Digital Divide' to 'Digital Inequality': Studying Internet Use as Penetration Increases*. Working Papers 47. Princeton University, Woodrow Wilson School of Public, International Affairs, Center for Arts, and Cultural Policy Studies.

Goolsbee, A. and P. J. Klenow (2006). “Valuing Consumer Products by the Time Spent Using Them: An Application to the Internet”. In: *American Economic Review* 96.2, pp. 108–113.

Götz, G. (2013). "Competition, regulation, and broadband access to the internet". In: *Telecommunications Policy* 37.11, pp. 1095–1109.

Griliches, Z. (1957). "Hybrid Corn: An Exploration in the Economics of Technological Change". In: *Econometrica* 25.4, pp. 501–522.

Halavais, A. (2000). "National borders on the world wide web". In: *New Media and Society* 2.1, pp. 7–28.

Hargittai, E. (1999). "Weaving the Western Web: explaining differences in Internet connectivity among OECD countries". In: *Telecommunications Policy* 23.10-11, pp. 701–718.

Harris, R.-G. (1996). *The Internet as a GPT : Factor Market Implications*. Discussion Papers dp97-01. Department of Economics, Simon Fraser University.

Hitt, L. and P. Tambe (2007). "Broadband adoption and content consumption". In: *Information Economics and Policy* 19.3-4, pp. 362–378.

Kiiski, S. and M. Pohjola (2002). "Cross-country diffusion of the Internet". In: *Information Economics and Policy* 14.2, pp. 297–310.

Koutroumpis, P. (2009). "The economic impact of broadband on growth: A simultaneous approach". In: *Telecommunications Policy* 33.9, pp. 471–485.

Krenc, T., O. Hohlfeld, and A. Feldmann (2014). "An Internet Census Taken by an Illegal Botnet: A Qualitative Assessment of Published Measurements". In: *SIGCOMM Comput. Commun. Rev.* 44.3, pp. 103–111.

Levitt, S. D. and S. A. Venkatesh (2000). "An Economic Analysis Of A Drug-Selling Gang's Finances". In: *The Quarterly Journal of Economics* 115.3, pp. 755–789.

Miner, L. (2015). "The unintended consequences of internet diffusion: Evidence from Malaysia". In: *Journal of Public Economics* 132.C, pp. 66–78.

Mirani, L. (2015). *Different World: Millions of Facebook users have no idea they're using the internet*. Quarz, published 10. February 2015. Blog post. accessed 11.02.2015.

Pantea, S. and B. Martens (2014). "Has the digital divide been reversed? – Evidence from five EU countries". In: *electronic International Journal of Time Use Research* 11.1, pp. 13–42.

Röller, L.-H. and L. Waverman (2001). "Telecommunications Infrastructure and Economic Development: A Simultaneous Approach". In: *American Economic Review* 91.4, pp. 909–923.

Sylvester, D. E. and A. J. McGlynn (2010). "The Digital Divide, Political Participation, and Place". English. In: *Social Science Computer Review* 28.1, pp. 64–74.

Wunnava, P. V. and D. B. Leiter (2008). *Determinants of Inter-Country Internet Diffusion Rates*. IZA Discussion Papers 3666. Institute for the Study of Labor (IZA).

Zota, V. (2014). "Anpfiff - Technik für eine ungetrübte Fußball-WM". In: *c't - Magazin für Computertechnik* 13, p. 70.

2.A Estimation of accumulated capital stock

Regarding ICT investments, the ITU database only includes the investments flows in a given year. However, ICT capital can be used for a number of years until it is depreciated. ICT hardware is not homogeneous, as certain equipment last only a short period of time while others (e.g. cables) are used over several years or even decades. To reflect these features the stock of ICT capital is estimated using the following exponential function:

$$\text{Stock}_0 = \sum_{-\bar{T}}^{t=0} e^{0.1t} \times \text{Investment}_t \tag{2.3}$$

Ideally, one would aggregate the data from the beginning of ICT investment in order to estimate the current capital stock. Ideally, for equation 2.3 one would set $\bar{T}$, the point in the past from where one would calculate the capital stock, to ∞. Due to shortcomings of the data, one has to weigh the number of included periods against the loss of observations, as in particular in early periods the data are very scarce. For my estimation I included investment over 10 years ($\bar{T} = 10$). If one would set no cut-off and include all countries regardless of missing observations one would bias the results in favor of countries who have good statistical data.

In order to mitigate the issue of missing data one has to impute the missing values or suffer from bias or loss of observations. The following steps were undertaken to fill in the missing data:

1. For missing observations on the current edge (as well as for the beginning of the time series), annual investments are assumed constant since the last observation.
2. "Holes", missing values inside a time series, were imputed linearly. In a panel analysis one should use a multiple imputation method, as standard errors will be to small otherwise. Nonetheless, in this static setting were only the cumulative values are used, this issue can be neglected.

2.B Supplementary Tables

Table 2.4: List of variables

Variable Name	Description
Bandwidth/User	International bandwidth per user, as available from the ITU
gdpcap	GDP per Capita
Investment	Investment in ICT capital per capita
latency	Latency (mean and quantile as indicated) from the "Internet Census 2012"
monthlycharge	monthly charge for broadband Internet
monopoly	Dummy for providers=1
providers	The number of providers per country
Stock/Pop	The calculated accumulated stock of ICT capital
UserShare	The Share of Internet users in the population (also penetration rate)

Table 2.5: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
log(Bandwidth/User)	105	9.908	1.815	5.068	15.224
log(gdpcap)	105	8.564	1.582	5.536	11.290
log(Latency), median	105	10.650	0.481	9.912	12.338
monthlycharge	105	69.564	198.633	6.137	1,760.449
monopoly	105	0.143	0.352	0	1
log(providers)	105	2.542	1.859	0.000	8.854
PopDensity	105	−9.543	1.508	−13.243	−3.937
Stock/Pop	105	5.602	1.308	0.156	7.624
User/Pop	105	−1.086	1.069	−4.263	−0.041

Table 2.6: Correlation between Variables

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
log(Bandwidth/User) (1)								
log(gdpcap) (2)	0.73^{***}							
log(Latency), median (3)	-0.48^{***}	-0.40^{***}						
monthlycharge (4)	-0.43^{***}	-0.14	0.24^{*}					
monopoly (5)	-0.31^{**}	-0.18	0.33^{***}	0.12				
log(providers) (6)	0.43^{***}	0.46^{***}	-0.30^{**}	-0.19	-0.5^{***}			
log(PopDensity) (7)	0.16	0.15	-0.26^{**}	-0.04	0.03	-0.03		
log(Stock/Pop) (8)	0.62^{***}	0.79^{***}	-0.48^{***}	-0.24^{*}	-0.17	0.37^{***}	0.11	
log(User/Pop) (9)	0.66^{***}	0.84^{***}	-0.40^{***}	-0.21^{*}	-0.32^{***}	0.46^{***}	0.18	0.74^{***}

3 Low Latency Internet and Economic Growth: A Simultaneous Approach

Jochen Lüdering

This article is available as J. Lüdering (2016a). *Low Latency Internet and Economic Growth: A Simultaneous Approach*. MAGKS - Joint Discussion Paper Series in Economics 34/2016. Philipps-Universität Marburg.

The article was presented at the Research Conference on Communications, Information and Internet Policy (TPRC44), Arlington, VA, USA, September 30 – October 1, 2016 and the 27th European Regional Conference of the International Telecommunications Society (ITS Europe 2016), Cambridge, UK, September 7th – 9th 2016.

Acknowledgments I am grateful for valueable comments by Steffen Viete (ZEW) and the participants of ITSEurope, as well as the 44th TPRC Conference. Also, I want to thank Les Cottrell at Stanford for providing the latency data. In addition, a DAAD travel grant allowed me to attend the TPRC conference in Arlington, VA, USA, which has been very helpful.

JEL: O47, O57, L96

Keywords: Economic Growth, Simultaneous Equations, Internet, Latency

Abstract

Given the quality of the available data on Internet access across several countries, it is necessary to evaluate alternative measures to assess the effect of Internet access on economic outcomes.

The research at hand builds up on an earlier paper (Chapter 2), which introduced a novel measure of Internet quality. A logical consequence has been to introduce the new indicator (average latency for a country) into established models of economic growth. The data used in this analysis spans the period from 2008 to 2014 and covers 155 countries. The findings largely confirm previous results, that Internet access is beneficial to economic growth and emphasize the appropriateness of technical measures of Internet quality for economic analysis. Apart from providing insight into the quality dimension these measures do not rely on survey data, but can be obtained directly requiring only a low level of investment, making the data collection process viable even for smaller institutions.

3.1 Introduction

The reliance on Internet connectivity has become common place in most industrialized economies, and there is empirical evidence that broadband access contributes to economic growth. Unfortunately, the empirical studies in this area suffer from limited data availability to assess long-run growth effects and usually cover only small sets of countries. The aim of this study is to introduce a novel measure for Internet quality and show that it can be used in a conventional growth model. The latency of an Internet connection can be measured directly and one does not rely on intergovernmental agencies for collecting the data.

Apart from the lack of availability some existing measures also have conceptual problems. Hence, it becomes crucial to discuss the operationalization of Internet usage, availability or quality when interpreting results in a context of economic growth or the digital divide. For example, the penetration rate, i.e. the share of population using the technology, is easily available from the World Bank and therefore widely used (Czernich et al. 2011; Koutroumpis 2009). Comparing the penetration rate across countries is dangerous as the survey methodology differs across countries and it neglects any information about the quality and frequency of Internet access. Moreover, the fact that the ratio is bound by 0 and 1, one may falsely "discover" that the digital divide is closing, as some (in particular Scandinavian) countries have reached a penetration rate of close to the hundred percent bound. Consequently, any increase in the penetration rate in a developing country will result in a smaller "divide" between the industrialized and the developing country, despite an increasing divide in qualitative terms.

Thus, new measures are needed to quantify the phenomenon and provide a sound foundation for the discussion of Internet and growth, as well as the extend of the digital divide. One possibility is the use Internet bandwidth per user (Rohman and Bohlin 2012). This provides information to one aspect of quality, but it remains difficult to estimate. Hence, I propose the use of latency data, which can be directly measured from every computer on the Internet. Therefore, it is not subject to the same difficulties when aggregating data from country specific survey sources. Bandwidth (usually referred to when discussing Internet speed) and latency are two concepts that are best explained jointly. Bandwidth is the amount of information which can be transported at a single point in time between two points. In contrast latency is the actual time of transportation between two points. A shipping container full of hard disks is an illustrative example of a very high bandwidth connection between two ports. However, transporting the data in a truck from Melbourne to Atlanta may well take days on a cargo vessel, before the first bit of information arrives at the destination. In contrast, transmitting Morse codes via ham radio covers the distance within milliseconds but only providing enough bandwidth for a few characters per minute. Modern broadband communication combines a low latency with a high bandwidth, with slight differences depending on the technology used.

Even though, bandwidth is often perceived as the most important measure when it comes to Internet speed. Nonetheless, latency has real implications on economic outcomes. Most prominent in economic research is the discussion of high-frequency trading and low latency strategies (Hasbrouck and Saar 2013). Outside the financial industry the scientific evidence of the effect of latency becomes scarce. However, there is anecdotal evidence from statements by prominent figures in the IT industry, that latency has a profound effect on their business. For example Greg Linden (back then at Amazon) reports[1] about Marissa Meyer's (back then a Google Engineer) talk at the *Web 2.0* conference stated that an increase in latency of 0.5s lead to a reduction in traffic and revenue at Google. Linden reports similar findings from A/B testing at Amazon: an increase in latency of as little as 100ms would result in "substantial and costly drop in revenue". Similar evidence has been presented by other companies at *Velocity 2009*.[2]

After Lüdering (2015) discussed the suitability of latency as a proxy for Internet quality, this paper tries to assess the suitability of this indicator by introducing it into a growth model. Consequently, it contributes additional evidence on the causal relationship between information and communications technology (ICT) infrastructure and economic growth by combining a dataset (Zennaro et al. 2006) which is novel to economics with the established methodology from Röller and Waverman (2001) and Koutroumpis (2009). The dataset used here does not cover quite as many countries as Lüdering (2015) but also contains a time dimension. The dataset used in the analysis spans a period of 6 years from 2008 to 2014.

The remainder is organized as follows. Section 3.2 provides a brief overview of the literature on the relationship between economic growth and ICT infrastructure. Subsequently, the data used in this paper is described in Section 3.3 while Section 3.4 elaborates on the empirical approach. The results are discussed in Section 3.5 and Section 3.6 closes with a conclusion.

[1] See Linden's blog article `http://glinden.blogspot.de/2006/11/marissa-mayer-at-web-20.html` accessed 2016-09-27.

[2] See Souders's summary `http://radar.oreilly.com/2009/07/velocity-making-your-site-fast.html`, accessed 2016-07-27.

3.2 Economic Growth and Telecommunication Technology

In neoclassical growth models (Solow 1956; Swan 1956) technological progress is the sole driver of economic growth in the steady state. Endogenous growth models (Romer 1990) endogenized the creation of new, "non-rival" technology from human capital. The utilization of these new technologies eventually leads to economic growth. Hence, any mean, such as Internet access and economic integration that facilitates access to new technologies is growth enhancing.

The nature of ICT as a General-Purpose Technology implies that an investment in ICT capital leads to improvements in productivity across many fields of the economy leading to growth in total factor productivity. However, in order to realize the associated productivity gains complementary investements in other capital (e.g. knowledge) are required. These additional investments contribute to economic growth as Basu and Fernald (2007) show.

The empirical work on the relatonship between communication means and economic development dates back to Hardy (1980), who finds evidence that landline telephones are an important contributor to economic development. However, his results proved not to be robust to alterations in the sample of countries. More recent work by Röller and Waverman (2001) used an updated methodology that endogenized demand and supply for telecommunication. The authors find strong evidence for a link between telecommunication and economic growth. By differentiating between three levels of telecommunication infrastructure Röller and Waverman find evidence of positive network externalities. The necessary critical mass for increasing returns appears to be close to universal service. Sridhar and Sridhar (2007) refine the approach and specifically address the case of developing countries. Along these lines Lee, Levendis, and Gutierrez (2012) also conduct an empirical analysis of the impact of telecommunication infrastructure on economic growth in developing countries. They find that there is a particular large effect for mobile telecommunication on economic growth.

This hints at the possibility of leap-frogging and skipping the costly investment into landline infrastructure.

Subsequent analysis have turned towards the relationship between economic growth and Internet. Using a simple linear estimation approach Qiang and Rossotto (2009) find a large effect of broadband Internet connectivity on economic development. Applying the more sophisticated simultaneous estimation approach by Röller and Waverman (2001) Koutroumpis (2009) manages to establishes a causal link between Internet usage and economic development using a panel of the EU-15 countries over the duration of three years. He finds significant returns to ICT investments in particular for countries with a high initial penetration rate (e.g. the Scandinavian countries). Czernich et al. (2011) use an instrumental variable approach to estimate the impact of broadband adoption on economic growth. In order to solve issues of endogeneity the authors construct a theoretical broadband penetration rate, which they employ in the estimation. A recent contribution by Clarke, Qiang, and Xu (2015) took a closer look on the effects of Internet usage on the level of the individual firm. They find a robust link between Internet use and labor productivity for firms of different sizes but in particular for small and medium enterprises.

3.3 Data

The data used in this analysis is compiled from a variety of sources. Details of the origin of specific variables is provided in Table 3.1. While most of the variables are standard and the data are taken from the specified sources, there are a few specificities which are illustrated in this section.

It is a common issue in the empirical growth literature that one needs to calculate the stock of capital. In many cases, it is sufficient to revert to databases such as the Penn World Tables (PWT). When conducting this analysis version 9 of Penn World Tables was not yet available with no date fixed for publication. The PWT 8.1 of the database does not cover the sample

Table 3.1: Variables used in the analysis

Name	Description	Source
GDP	GDP	World Bank
GDPC	GDP per capita	World Bank
K	stock of capital	Penn World Tables and World Bank
L	size of labor force	World Bank
ICT	average round trip time	PingER
P	Broadband Price	ITU
EDU	spending on education (% of GDP)	World Bank
RD	R&D investment (% of GDP)	World Bank
Urban	Urban Population (%)	World Bank
ICTI	ICT Investment	ITU
InterPlatform	Herfindahl index for inter-platform competition	own calculation based on ITU data

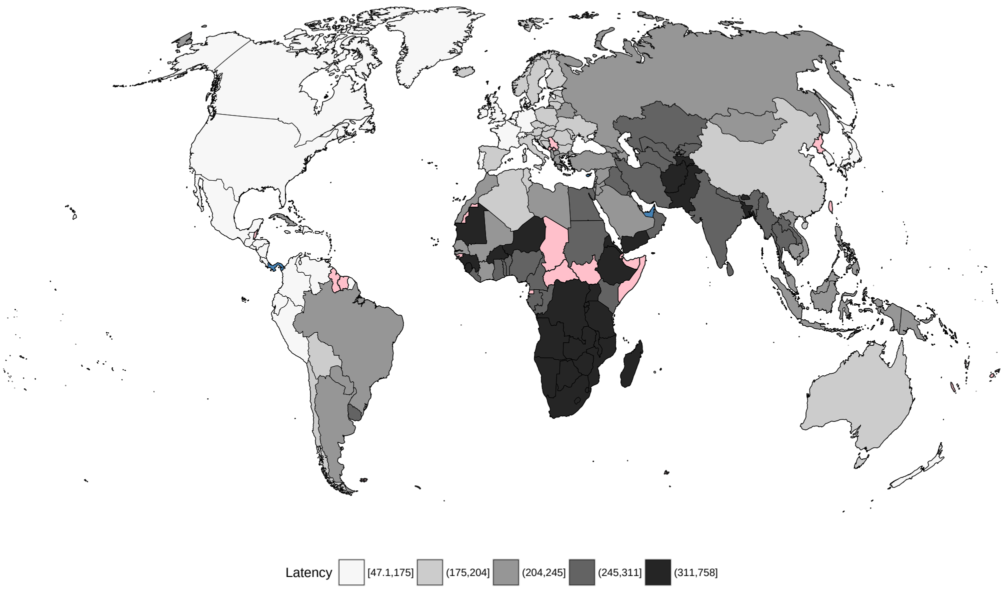
period in this article. As a workaround I calculate the capital stock for the analysis period using the initial capital stock from the Penn World Tables and add the investments available from the Worldbank, assuming a constant deprecation rate of 4.5% (based on Berlemann and Wesselhöft 2012) for all countries.[3] Ideally, one would deduct ICT capital from the general capital stock. However, the limited availability of data on ICT investments and the stock of ICT capital does not permit this. As the stock of ICT capital is small in contrast to general capital, the effects on the analysis should be negligible. If there was any effect, it would lead to an underestimation of the effect of ICT.

As a measure for Internet quality this paper uses latency data (specifically average round-trip-time). The data is provided by the PingER (Ping End-to-end Reporting) project (`http://www-iepm.slac.stanford.edu/pinger/`) run by SLAC National Accelerator Laboratory at Stanford University. An introduction to the PingER facility and dataset can be found in Zennaro et al. (2006). The *average round trip time* is the time that has passed between sending a request to a remote server (usually at other renowned research institutes) and receiving the answer at the monitoring server at Stanford University. This implies that lower

[3] The results appear to be not sensitive to the assumed value of the depreciation rate.

values indicate a better Internet quality. In order to aid interpretation as Internet quality in the regression analysis the values are multiplied by -1 after taking the logarithms, consequently positive coefficients will imply positive influence of Internet quality.

Figure 3.1: Spatial Distribution of Average Round Trip Time in 2014

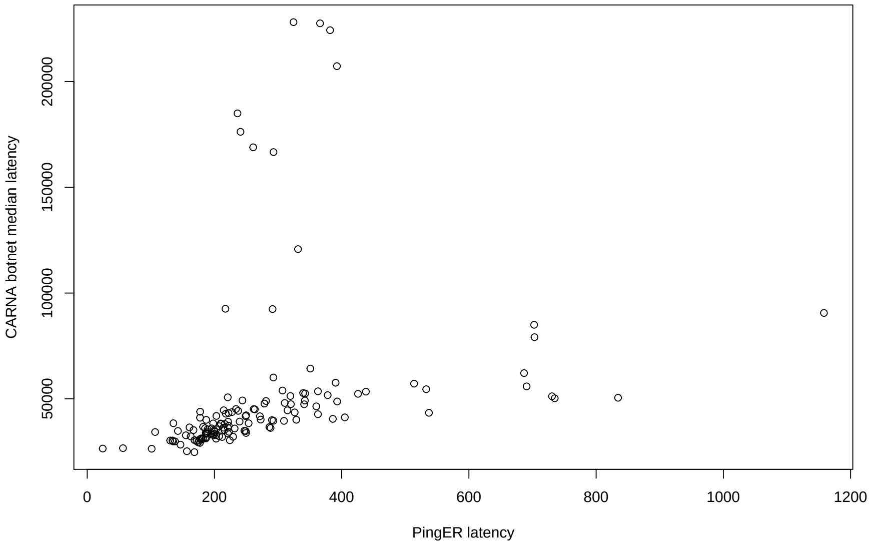


There is no data (in 2014 or not at all) available for countries colored pink

The measurements are aggregated per country and year, as the data is collected for several nodes in a single country and at an hourly frequency. The resulting dataset is available for a large set of 165 countries (see Figure 3.1). Unfortunately these sites have been added to Pinger consecutively. For example, there is no data on Panama and the United Arab Emirates before 2015. Other shortcomings include that the split of "Serbia and Montenegro" has not been incorporated into the data to date. The countries completely missing from the dataset due to the lack of remote-sides include small countries like Suriname and Guyana, and the closed off or crises stricken countries Somalia, Chad, Central African Republic, South Sudan and North Korea.

As PingER measures the latency between research institutes and universities it appears questionable whether it really reflects the circumstances across a country. Therefore, Figure 3.2 shows the latency measure used here plotted against the data from Lüdering (2015). The latter are per country median latency measurements of all Internet hosts in a country. Based on the Figure one may get the idea that there is indeed a positive correlation between the two latency datasets. Nonetheless, some countries have a rather high latency for their general population given the low latency enjoyed at their research institutions.

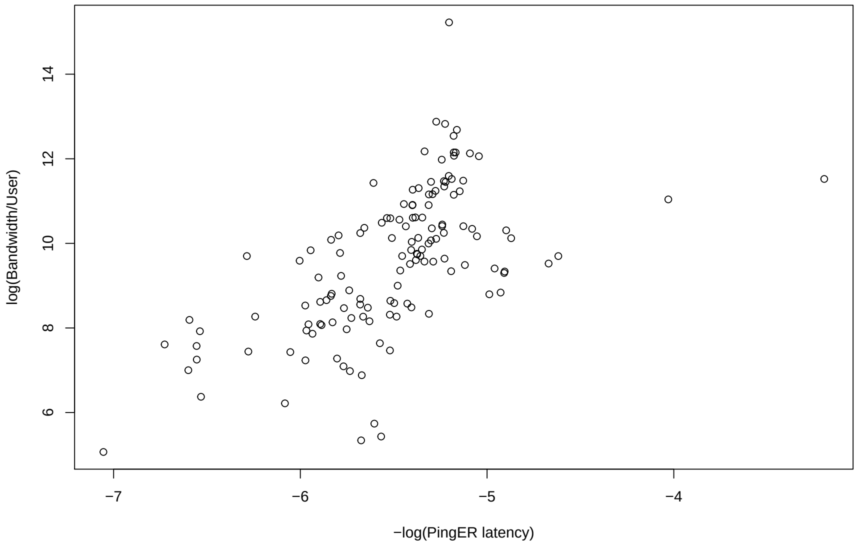
Figure 3.2: Two latency measures



Note: The CARNA Botnet latency measure (see Lüdering 2015) is the measured median latency across a country, the PingER latency is measured between well-connected universities and research institutions only.

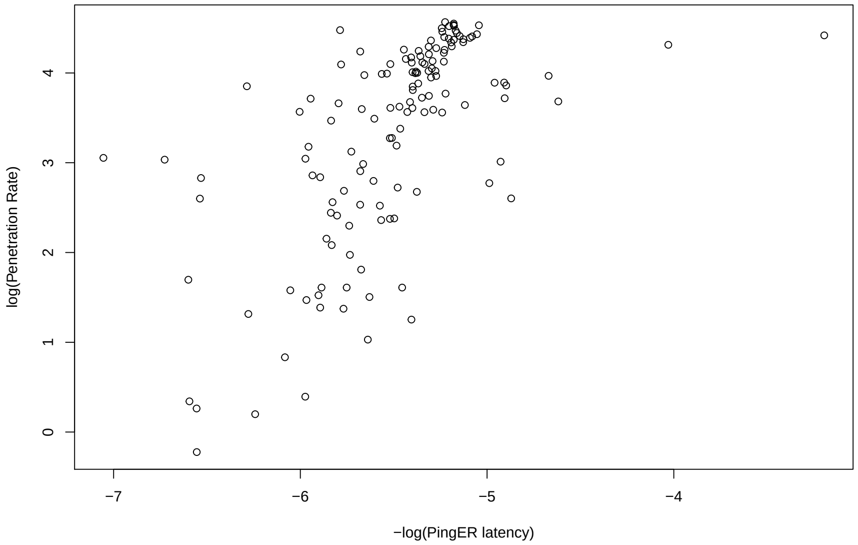
In order to support the argument that latency also provides some good overall summary of Internet quality in a country Figure 3.3 and 3.4 plot latency against bandwidth and the penetration rate respectively. From these plots one may take away that there is a positive relationship between (negative) latency and the other measures of Internet quality, while this may be a little vague in the case of the penetration rate.

Figure 3.3: Relationship between latency and bandwidth per user



Note: The negative log of latency is used in order for quality to increase along both axis.

Figure 3.4: Relationship between latency and penetration rate



Note: The negative log of latency is used in order for quality to increase along both axis.

Due to the absence of direct measures of competition in the market for telecommunication (e.g. number of companies providing Internet access), the level of competition between different technologies to access the Internet is used as a proxy for competition. The Herfindahl index for the inter-platform competition in the telecommunication market measures the concentration of the industry by summing up the market shares of each considered platform. Analogous to Koutroumpis (2009) for country i in year t it is given by:

$$\text{InterPlatform}_{it} = \sum_{1}^{n} \left(\frac{\text{Platform}_m}{\text{Total connections}} \right)^2 \tag{3.1}$$

The approach in the paper at hand, differs slightly in the scope. It includes data on all types of Internet connections where Koutroumpis (2009) is limited to broadband connections. While the primary focus lies on FTTH (Fiber optic cable to the home / to the building), DSL (Digital Subscriber Line) and (TV-) Cable in both cases, the "other" platform in this paper sums up all means of Internet connections in the ITU database which includes wireless (i.e. via Wi-Fi), satellite, dail-up and "other". Considering all means of Internet access should also be reflected by the used prices. Ideally one would use an aggregate price index of all access technologies. Due to reasons of data availability prices for broadband connectivity are used as proxy for general Internet pricing.

3.4 Empirical Analysis

Missing values in the dataset, in particular information on prices and investment into ICT infrastructure makes it necessary to apply imputation methods. For this exercise the *Amelia*[4] package for R is employed. After generating five imputed datasets the coefficient values are aggregated using Rubin's Rules (Rubin 1987). In addition one may also aggregate the R^2 values for regressions on multiple imputed datasets (Harel 2009). However, since the aim of instrumental regression lies in precise estimation of the effect of covariate x on dependent

[4] `https://cran.r-project.org/web/packages/Amelia/`

variable y when x is correlated with the error term, goodness of fit is in general not of interest. Moreover, in this case R^2 also lacks a natural interpretation and may also become negative. Nonetheless, R^2 is reported here by convention. While it is straightforward to perform multiple imputation on the dataset and estimate the system of equations the application of the aggregation rules remains largely untested for simultaneous estimation procedures.

A lot of missing data is due to a small set of countries: Angola, Eritrea, Faroe Islands, Greenland, Liechtenstein, Myanmar, French Polynesia, Puerto Rico, San Marino, Syria contain a large amount of NA values. However, removing these countries before imputation does not have a large effect on the results.

The empirical approach of this analysis builds up on the simultaneous estimation approach originally developed by Röller and Waverman (2001) and later refined Koutroumpis (2009). By following the model by Koutroumpis (2009) as close as possible I attempt to make the results comparable and infer whether round trip time can be used as a proxy for Internet quality. The differences that exist between the analysis at hand and Koutroumpis (2009) are due to issues of data availability. Working with a larger sample of countries some information was not available in a consistent manner. Consequently, the model was slightly altered. For example Koutroumpis (2009) included a measure of regulation of the telecommunication industry. This measure has been omitted from the model in this paper as the data is not available for the countries and timespan used in the analysis.

The model relies on a classical aggregate production function where a country's economic output is determined by capital and labor. The model is augmented with a measure of Internet quality, as an additional factor:

$$\text{GDP}_{it} = f(K_{it}, L_{it}, ICT_{it})$$

K is the level of capital (not including ICT capital) L is the size of the labor force and ICT is ICT capital.

Again following Koutroumpis (2009) a micro model is used to endogenize the creation of ICT capital by modeling demand, supply and production equations of ICT capital.
The demand for Internet quality is given by

$$\mathsf{ICT}_{it} = h(\mathsf{GDPC}_{it}, \mathsf{P}_{it}, \mathsf{URB}_{it}, \mathsf{EDU}_{it}, \mathsf{RD}_{it})$$

stating that the Internet quality demanded depends on income per head (GDPC), prices (P), the share of the population living in urban agglomerations (URB) and the expenditure on education (EDU) and R&D (RD).
The supply of Internet quality consists of the investments in ICT capital which is solely determined by prices and market structure

$$\mathsf{ICTI}_{it} = g(P_{it}, \mathsf{InterPlatform}_{it})$$

and the resulting improvement in quality (production function)

$$\Delta\mathsf{ICT}_{it} = k(\mathsf{ICTI}_{it})$$

. Using a log-linear approximation this model gives rise to the following system of equation:

$$\log(\mathsf{GDP}_{it}) = a_0 + a_1 \log \mathsf{K}_{it} + a_2 \log \mathsf{L}_{it} - a_3 \log \mathsf{ICT}_{it} + \varepsilon^1_{it} \tag{3.2}$$

$$\log(\mathsf{ICT}_{it}) = b_0 + b_1 \log \mathsf{GDPC}_{it} + b_2 \log \mathsf{P}_{it} + b_3\mathsf{EDU}_{it} + b_4\mathsf{URB}_it + b_5\mathsf{RD}_{it} + \varepsilon^2_{it} \tag{3.3}$$

$$\log(\mathsf{ICTI}_{it}) = c_0 + c_1 \log \mathsf{P}_{it} + c_2\mathsf{InterPlatform}_{it} + \varepsilon^3_{it} \tag{3.4}$$

$$-\log\left(\frac{\mathsf{ICT}_{it}}{\mathsf{ICT}_{i,t-1}}\right) = d_0 + d_1 \log \mathsf{ICTI}_{it} + \varepsilon^4_{it} \tag{3.5}$$

3.5 Results

Following Koutroumpis (2009) the empirical model is estimated applying a three-stage-least-squares GMM approach. In addition two-stage estimates are provided. Both methods are used to jointly estimate a system of equations, the third step takes the interdependence of the error terms into account and provides more accurate coefficient estimates. The estimation results are in detail provided in Table 3.2. There is a strikingly huge positive significant effect of ICT quality on GDP. The size of the effect is huge over all specifications ranging from an increase of 0.45% to 2.5% in GDP for a 1% increase in Internet quality. The other coefficients are not central to this paper. The estimation of the equation system aims at providing a consistent estimate of the effect of ICT quality on GDP growth, rather than examining the specificities of demand and supply effects. Nonetheless, the coefficients largely have the expected signs. Demand is reduced by prices and increased by urbanization, education and R&D investments. Supply is reduced by a higher Herfindahl index (*i.e.* less inter-platform competition). The negative sign of price in the supply equation is a little bit surprising. Similarly, in the ICT production equation one would expect the investments in ICT effect to have a positive effect on the change in Internet quality.

Including dummy variables for each country and each year reduces the size of the coefficient of interest and leads to a loss of significance for many variables in the system. In particular urbanization education spending and research and development investments are likely time-invariant country specificities, which are completely captured by the included country specific effects. In addition, the depressed growth rates due to the global financial crises are to a large extent captured by the time fixed effects.

After having obtained an estimate for the coefficient one can address the question what the economic effect of improvements in ICT quality has been over the period of six years.

Table 3.2: Statistical models

	(1)	(2)	(3)	(4)
	2SLS	2SLS-fixed	3SLS	3SLS-fixed
Aggregate Production				
(Intercept)	19.903^{***}		15.349^{***}	
	(3.780)		(4.294)	
K	0.515^{***}	0.030	0.655^{***}	0.034
	(0.094)	(0.031)	(0.111)	(0.030)
L	0.326^{***}	0.332^{**}	0.169^{*}	0.324^{**}
	(0.074)	(0.172)	(0.106)	(0.169)
ICT	2.502^{***}	0.477^{**}	1.902^{***}	0.449^{**}
	(0.440)	(0.230)	(0.534)	(0.221)
Demand				
(Intercept)	-5.154^{***}		-5.488^{***}	
	(0.285)		(0.379)	
GDPC	0.008	-1.231	0.036	-4.810
	(0.028)	(22.069)	(0.034)	(23.580)
P	-0.307^{***}	-1.118	-0.258^{***}	-1.098
	(0.050)	(3.247)	(0.058)	(3.023)
URBAN	0.007^{***}	0.013	0.006^{***}	0.069
	(0.001)	(0.288)	(0.001)	(0.293)
EDU	0.017^{**}	0.040	0.019^{***}	0.044
	(0.010)	(0.174)	(0.006)	(0.160)
RD	0.109^{***}	0.107	0.105^{***}	0.054
	(0.021)	(0.379)	(0.034)	(0.358)
Supply				
(Intercept)	32.666^{***}		32.815^{***}	
	(2.376)		(2.330)	
P	-1.691^{**}	-0.521	-1.809^{***}	-0.475
	(0.755)	(0.491)	(0.720)	(0.430)
InterPlatform	-11.896^{**}	0.627	-11.478^{**}	1.133
	(6.699)	(1.793)	(6.444)	(1.473)
Infrastructure Production				
(Intercept)	0.184^{**}		0.202^{**}	
	(0.095)		(0.091)	
ICTI	-0.007^{*}	0.447	-0.008^{**}	0.657
	(0.005)	(0.547)	(0.005)	(0.675)
Adj. R^2				
Aggregate Production	0.750	0.996	0.838	0.997
Demand	0.397		0.435	
Supply		0.829		0.835
InfraProd				

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

K, L, ICT, GDPC, P, ICTI are log transformed

Model 2 and 4 include country and year dummies to control for specificities

The growth in ICT quality is approximated by the log differences, which is multiplied by the estimated coefficient.

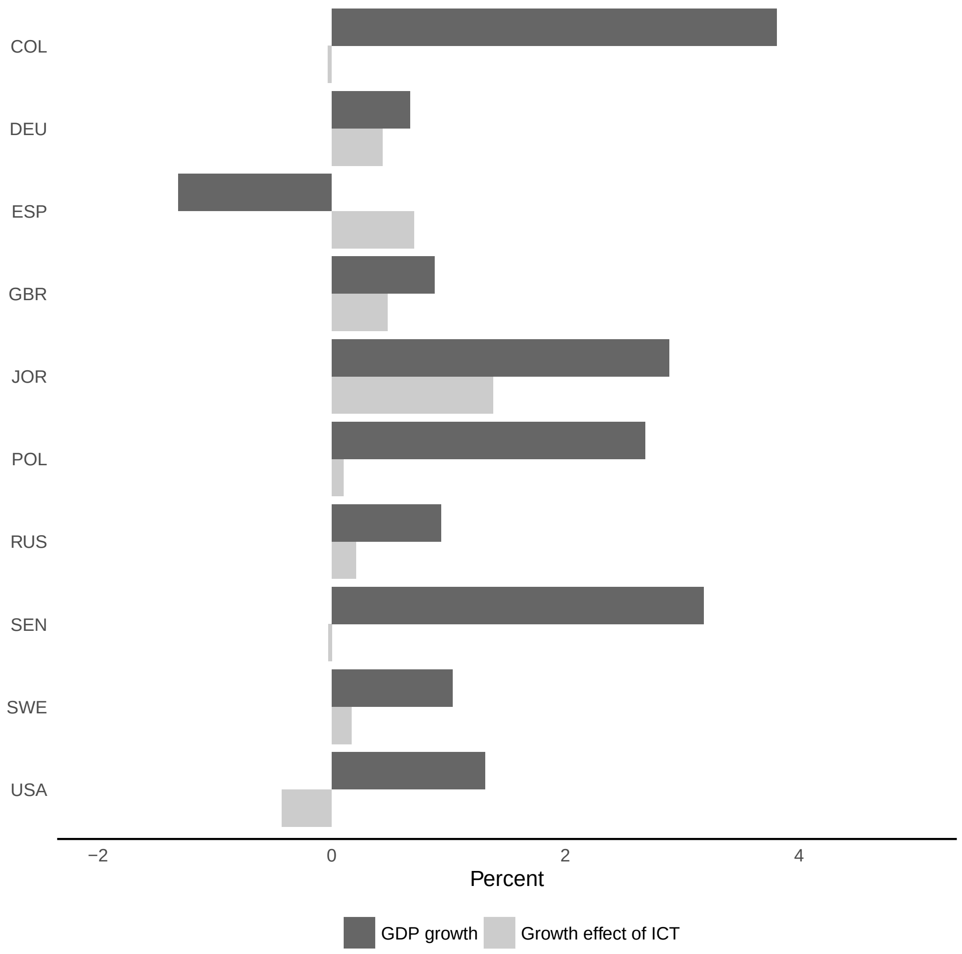
$$\mathsf{GE} = [(\log \mathsf{ICT}_{2014} - \log \mathsf{ICT}_{2008}) \times (-\hat{a}_3) + 1]^{\frac{1}{6}} \tag{3.6}$$

Figure 3.5 shows the average growth effects (GE) and countries GDP growth rate over the whole period for a subset of the included countries (The results for all countries can be found in Figure 3.7 in the appendix). Columbia had a growth rate of GDP of 3.8% annually, while the effect of improvements in ICT had a small negative effect. For the US the lack of improvements in ICT quality had an even stronger negative effect. This illustrates, that economic growth is a net-effect resulting from a variety of influences. For most countries growth contribution of ICT improvement and economic growth have the same sign: Germany, growing 0.67% annually, had a contribution of 0.43 percentage points from ICT improvements. However, in Spain there was a positive growth effect of 0.7% from ICT in spite of suffering from negative economic growth during the period, due to the European sovereign debt crises. The reported effects are very large, in terms of elasticity of GDP with respect to Internet quality, but also in terms of the growth contribution of ICT. The effects reported here are of the same magnitude as reported by Koutroumpis (2009). For example, the research at hand finds a growth contribution of 0.7% of ICT for Spain, while Koutroumpis reports 0.39% for Spain.[5] Due to differences in methodology comparing the results of Czernich et al. (2011) is not as straightforward, as the authors examine the changes in GDP *per capita*. They find that a change of 1 percentage point in the broadband penetration rate leads to an increase of annual GDP per capita growth of 0.9 to 1.5 percentage points. Which is of the same magnitude as the median growth effect of 0.71% (mean: 1.57%) of Internet quality.

As outlined earlier the combination of simultaneous equation modeling with multiple imputation appears to be untested. In order to check the robustness of the results the approach is

[5] Unfortunatelly, Koutroumpis (2009) does not provide compound annual growth effects for more countries

Figure 3.5: Growth contribution of ICT



twofold. On the one hand, regression results for complete cases and, respectively a dataset generated by single imputation are provided in Table 3.5. In particular the single imputation case yields coefficients of very similar size as in the multiple imputation case. As expected, the standard errors are smaller when relying on single imputation for missing values. If one only considers the complete cases the number of observations is substantially reduced, leading to several differences. It is notable that the coefficient of interest to this analysis on ICT remains similar in size and significant at the 10% level. On the other hand, the stability of the estimation results is confined using a simulation method. The applied simulation procedure is outlined in greater detail in Appendix 3.B. The results indicate that the application of multiple imputation reduces the variance of the estimated coefficient at the cost of a small bias.

3.6 Conclusion

The preceding analysis shows that the direct measure of latency can be used as a proxy of Internet quality. The estimation technique builds up on the methodology by Röller and Waverman (2001) and Koutroumpis (2009) to mitigate the potential for simultaneity.

Using latency as a proxy for ICT quality makes it possible to obtain information on the infrastructure quality on a variety of countries without relying on data collected by a country's authorities. This increases the number of countries on which consistent information on Internet quality is available. Thus, the sample covers 155 countries compared to the subset of OECD countries usually used in previous papers.

The evidence from the analysis confirms the strong effect of ICT infrastructure on economic development established in Koutroumpis (2009) and Czernich et al. (2011). While these previous studies have established that Internet *usage* is an important factor for growth, my contributions finds that it is also the quality of the infrastructure that matters. This implies that despite a narrowing digital divide in terms of users, the qualitative dimension is also important.

Bibliography

Basu, S. and J. Fernald (2007). "Information and communications technology as a general-purpose technology: evidence from US industry data". In: *German Economic Review* 8.2, pp. 146–173.

Berlemann, M. and J.-E. Wesselhöft (2012). *Estimating Aggregate Capital Stocks Using the Perpetual Inventory Method: New Empirical Evidence for 103 Countries*. Working Paper 125/2012. Helmut Schmidt University, Hamburg.

Clarke, G. R., C. Z.-W. Qiang, and L. C. Xu (2015). "The Internet as a general-purpose technology: Firm-level evidence from around the world". In: *Economics Letters* 135, pp. 24–27.

Czernich, N., O. Falck, T. Kretschmer, and L. Woessmann (2011). "Broadband Infrastructure and Economic Growth". In: *Economic Journal* 121.552, pp. 505–532.

Hardy, A. P. (1980). "The role of the telephone in economic development". In: *Telecommunications Policy* 4.4, pp. 278–286.

Harel, O. (2009). "The estimation of R 2 and adjusted R 2 in incomplete data sets using multiple imputation". In: *Journal of Applied Statistics* 36.10, pp. 1109–1118.

Hasbrouck, J. and G. Saar (2013). "Low-latency trading". In: *Journal of Financial Markets* 16.4, pp. 646–679.

Koutroumpis, P. (2009). "The economic impact of broadband on growth: A simultaneous approach". In: *Telecommunications Policy* 33.9, pp. 471–485.

Lee, S. H., J. Levendis, and L. Gutierrez (2012). "Telecommunications and economic growth: an empirical analysis of sub-Saharan Africa". In: *Applied Economics* 44.4, pp. 461–469.

Lüdering, J. (2015). *The measurement of internet availability and quality in the context of the discussion on digital divide*. Discussion Papers 65 [rev.] Justus Liebig University Giessen, Center for international Development and Environmental Research (ZEU).

Qiang, C. Z.-W. and C. M. Rossotto (2009). “Information and Communications for Development 2009: Extending Reach and Increasing Imapct”. In: Washington DC.: World Bank. Chap. Economic impacts of broadband, pp. 35–50.

Rohman, I. and E. Bohlin (2012). “Does broadband speed really matter as a driver of economic growth? Investigating OECD countries”. In: *International Journal of Management and Network Economics* 2.4, pp. 336–356.

Röller, L.-H. and L. Waverman (2001). “Telecommunications Infrastructure and Economic Development: A Simultaneous Approach”. In: *American Economic Review* 91.4, pp. 909–923.

Romer, P. (1990). “Endogenous Technological Change”. In: *Journal of Political Economy* 98.5, S71–102.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Solow, R. M. (1956). “A Contribution to the Theory of Economic Growth”. In: *The Quarterly Journal of Economics* 70.1, pp. 65–94.

Sridhar, K. S. and V. Sridhar (2007). “Telecommunications Infrastructure And Economic Growth: Evidence From Developing Countries”. In: *Applied Econometrics and International Development* 7.2, pp. 37–56.

Swan, T. W. (1956). “Economic Growth and Capital Accumulation”. In: *The Economic Record* 32.2, pp. 334–361.

Zennaro, M., E. Canessa, K. R. Sreenivasan, A. A. Rehmatullah, and R. L. Cottrell (2006). “Scientific Measure of Africa’s Connectivity”. In: *Information Technologies and International Development* 3.1, pp. 55–64.

3.A Additional Tables

Table 3.3: Summary Statistics

Variables	N	Mean	St. Dev.	Min	Max
GDP	1,082	24.396	2.069	19.999	30.325
K	936	26.230	1.932	22.080	31.441
L	1,071	15.409	1.563	12.007	20.508
ICT	933	−5.529	0.499	−7.582	−3.194
GDPC	1,082	8.239	1.611	4.989	11.356
P	986	3.330	0.932	−0.056	7.473
EDU	559	4.809	2.085	1.100	19.258
RD	473	1.057	1.025	0.013	4.387
URBAN	1,085	58.459	23.022	8.550	100.000
InterPlatform	465	0.557	0.161	0.226	1.000
ICTI	668	20.167	1.839	12.062	25.117

Note: Summary statistics calculated before imputation.
K, L, ICT, GDPC, P, ICTI are log transformed

Table 3.4: Countries in PingER dataset

Afghanistan
Albania
Algeria
Andorra
Angola
Argentina
Armenia
Australia
Austria
Azerbaijan
Bahamas
Bahrain
Bangladesh
Belarus
Belgium
Benin
Bhutan
Bolivia
Bosnia Herzegovina
Botswana
Brazil
Brunei
Bulgaria
Burkina Faso
Burundi
Cambodia
Cameroon
Canada
Cape Verde
Chile
China
Colombia
Costa Rica
Croatia
Cuba
Cyprus
Czech Republic
DR Congo
Denmark
Dominican Republic
Ecuador
Egypt
El Salvador
Eritrea
Estonia
Ethiopia
Faroe Islands
Finland
France
French Polynesia
Gabon
Gambia
Georgia
Germany
Ghana
Greece
Greenland
Guatemala
Guinea
Haiti
Honduras
Hong Kong
Hungary
Iceland
India
Indonesia
Iran
Iraq
Ireland
Israel
Italy
Ivory Coast
Jamaica
Japan
Jordan
Kazakhstan
Kenya
Korea Rep
Kuwait
Kyrgyzstan
Laos
Latvia
Lebanon
Lesotho
Liberia
Libya Arab Jamahiriya
Liechtenstein
Lithuania
Luxembourg
Macedonia
Madagascar
Malawi
Malaysia
Maldives
Mali
Mauritania
Mauritius
Mexico
Moldova
Mongolia
Morocco
Mozambique
Myanmar
Namibia
Nepal
Netherlands
New Zealand
Nicaragua
Niger
Nigeria
Norway
Oman
Pakistan
Palestine
Panama
Papua New Guinea
Paraguay
Peru
Philippines
Poland
Portugal
Puerto Rico
Qatar
Republic of the Congo
Romania
Russia
Rwanda
San Marino
Saudi Arabia
Senegal
Seychelles
Sierra Leone
Singapore
Slovak Republic
Slovenia
Solomon Islands
South Africa
Spain
Sri Lanka
Sudan
Swaziland
Sweden
Switzerland
Syria
Tajikistan
Tanzania
Thailand
Timor-Leste
Togo
Trinidad and Tobago
Tunisia
Turkey
Turkmenistan
Uganda
Ukraine
United Arab Emirates
United Kingdom
United States
Uruguay
Uzbekistan
Venezuela
Vietnam
Yemen
Zambia
Zimbabwe

Table 3.5: Complete cases and single imputation regression

	complete cases (3SLS)	single Imputation (3SLS)
Aggregate Production		
(Intercept)	16.46	18.08^{***}
	(11.68)	(3.33)
K	0.83^{***}	0.60^{***}
	(0.30)	(0.08)
L	0.06	0.20^{***}
	(0.25)	(0.06)
ICT	2.69^{*}	2.25^{***}
	(1.39)	(0.38)
Demand		
(Intercept)	-4.37^{***}	-5.32^{***}
	(0.85)	(0.24)
GDPC	0.14^{***}	0.02
	(0.05)	(0.02)
P	-0.72^{***}	-0.27^{***}
	(0.27)	(0.04)
URBAN	0.00	0.01^{***}
	(0.00)	(0.00)
EDU	-0.03^{*}	0.01^{***}
	(0.02)	(0.00)
R&D	0.18^{***}	0.13^{***}
	(0.07)	(0.02)
Supply		
(Intercept)	21.61^{***}	32.56^{***}
	(2.32)	(0.88)
P	1.08^{*}	-1.01^{**}
	(0.63)	(0.41)
InterPlatform	-8.22^{***}	-15.27^{***}
	(2.10)	(2.16)
Infrastructure Production		
(Intercept)	0.13	0.26^{***}
	(0.14)	(0.07)
ICTI	-0.00	-0.01^{***}
	(0.01)	(0.00)
Adj. R^2		
Aggregate Production	0.82	0.79
Demand	-0.16	0.41
Supply	-0.26	-1.61
Infrastructure Production	-0.02	-0.01
Num. obs.	167	1085

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

K, L, ICT, GDPC, P, ICTI are log transformed

3.B Simulation

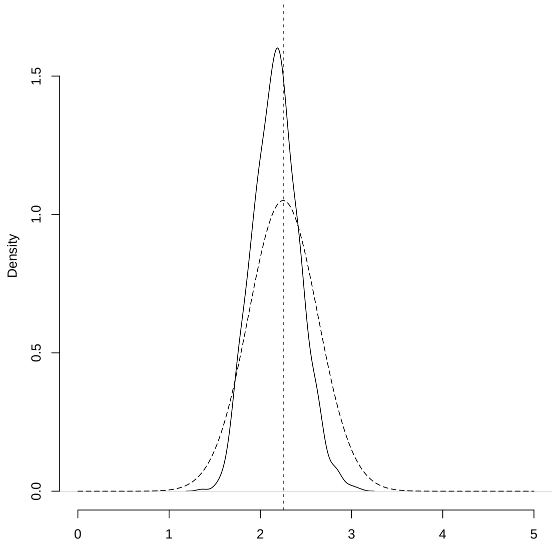
The robustness of the estimation results is assessed by simulating the imputation process and the subsequent estimation for 1000 times.

- Taking an imputed dataset as given and estimating coefficients which are considered to be the “real” coefficients, a new set of dependent variables are generated. Random errors are drawn from a normal distribution (with the covariances being estimated from the original residuals) and added to the newly generated variables.

- In a loop a number of “holes”, equal the missing share in the original dataset, is added to the dataset and the imputation algorithm is run to impute five datasets. The estimation is run for each of the datasets and the coefficients are combined using Rubin’s rules.

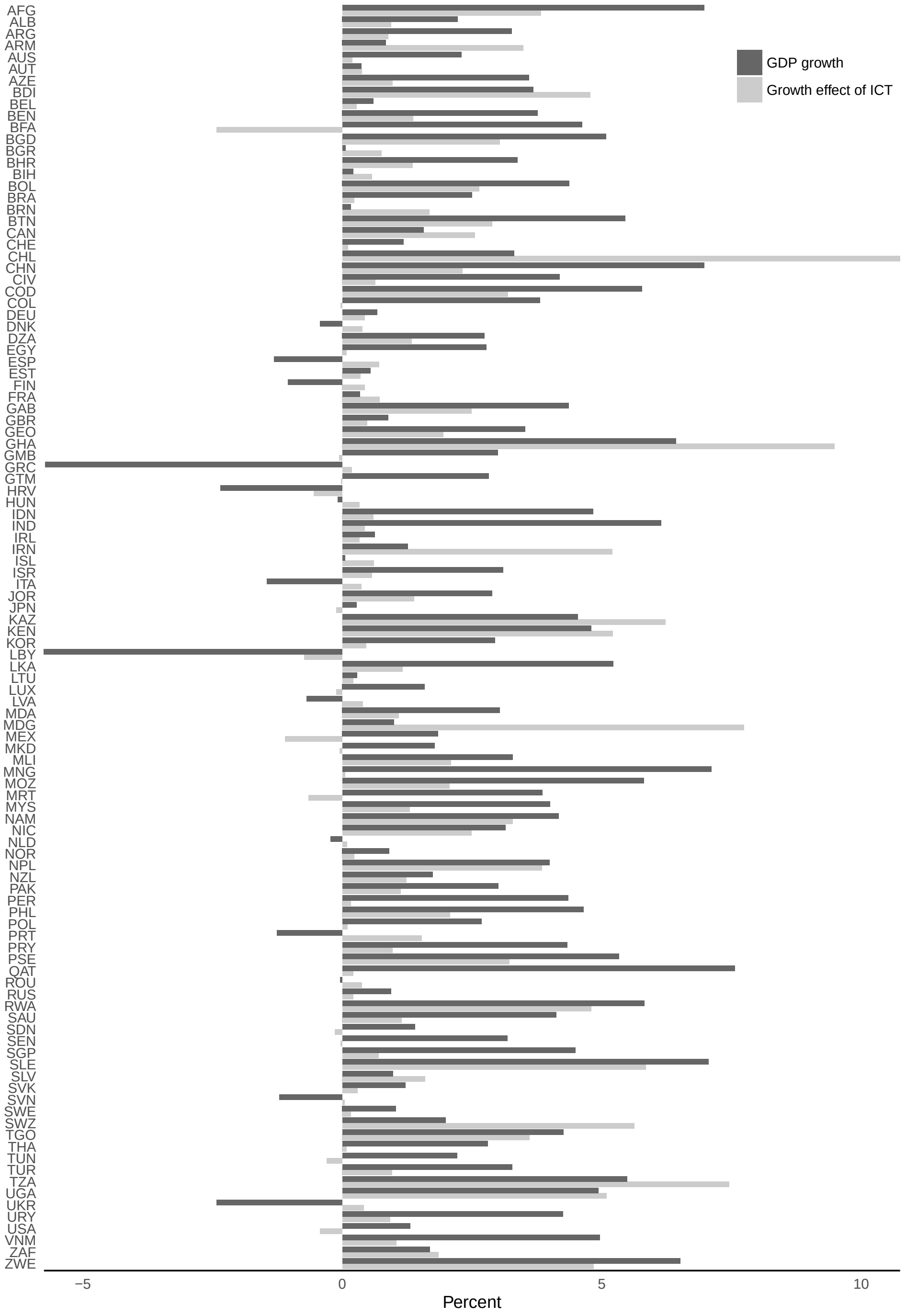
- The distribution of the coefficients over several simulations can be analyzed. As an example the coefficient for ICT is shown in Figure 3.6. Note: Variance for the “real” coefficient is obtained from original standard errors

Figure 3.6: Simulated coefficient (solid) and “real” coefficient (dashed)



3.C Growth effects

Figure 3.7: ICT growth effects



Note: Bars are truncated at -5% and +10%

4 Forward or Backward Looking? The Economic Discourse and the Observed Reality

Jochen Lüdering[a] & Peter Winker[b]

This article has been published as J. Lüdering and P. Winker (2016). "Forward or backward looking? The economic discourse and the observed reality". In: *Jahrbücher für Nationalökonomie und Statistik* 236.4, pp. 483–516. An early version has been presented at the MAGKS Colloquium March 20th 2015 in Rauischholzhausen. The current version has been presented at CMStatistic 9-11 December 2016 in Sevilla, Spain.

Acknowledgments The authors want to thank Andreas Schiermeier, Peter Reifschneider, and Christoph Funk for assistance with obtaining and preprocessing the data. We are grateful for the funding provided by Lucius & Lucius and to Jochen Kothe at Niedersächsische Staats- und Universitätsbibliothek Göttingen, Georg-August-Universität Göttingen for providing access to `digizeitschriften.de`. Last but not least, we want to thank three anonymous referees for their comments and the editor Joachim Wagner for facilitating the refereeing process.

JEL classification: B23, C22, C49

Keywords: Automatic text analysis, economic history, Latent Dirichlet Allocation, time series analysis

[a] Own contribution 50%

[b] Department of Statistics and Econometrics and Center for International Developmental and Environmental Research (ZEU), Justus-Liebig-Universität Gießen, e-Mail: peter.winker@wirtschaft.uni-giessen.de

Abstract

Is academic research anticipating economic shake-ups or merely reflecting the past? Exploiting the corpus of articles published in the Journal of Economics and Statistics (Jahrbücher für Nationalökonomie und Statistik) for the years 1949 to 2010, this pilot study proposes a quantitative framework for addressing these questions. The framework comprises two steps. First, methods from computational linguistics are used to identify relevant topics and their relative importance over time. In particular, Latent Dirichlet Allocation is applied to the corpus after some preparatory work. Second, for some of the topics which are closely related to specific economic indicators, the developments of topic weights and indicator values are confronted in dynamic regression and VAR models. The results indicate that for some topics of interest, the discourse in the journal leads developments in the real economy, while for other topics it is the other way round.

4.1 Introduction

What drives the selection of topics in economic research? Given freedom of research in the public higher education system, the research agenda is not a result of political decision making. Thus, it might be driven by personal interest, perspectives of gains in reputation, traditions handed-down from the doctoral supervisor, networks, job perspectives, tasks in economic policy advice etc. While all listed arguments might be relevant for almost all fields of science, empirical social sciences such as economics might be subject to a further driver – reality. Following the financial crisis, we have seen a regained interest in financial market stability and credit rationing (e.g., Turner et al. 2010), in the sequel, public debt came back on the agenda (e.g., Burret, Feld, and Köhler 2013). Given the current influx of refugees in Germany, it does not appear too far-fetched to predict that the economic analysis of causes and effects of migration will see a renaissance in the near future.

The link between reality and economic research might be triggered by different mechanisms including some of the aforementioned ones. We do not strive to identify these drivers and their relative importance, but address a much more modest intellectual goal, namely the identification of the evolution of research interests over time and its interaction with developments in the real economy. Possibly, economists have rational expectations about future economic developments and, consequently, focus their research on topics which are to become relevant. Alternatively, they just observe the economic situation and try to explain it ex post. Besides identifying relevant research topics, our aim is to find out which direction of the links between economic science and economic reality is prevailing.

Given that the evolution and deployment of research fields takes time, such an analysis requires a sufficiently long observation period. For the quantitative research approach taken in this contribution, it implies that a long sample of data is required. Long time series on key economic indicators become increasingly available (see e.g., Rahlf 2016), but have to be treated carefully given a substantial number of structural breaks over the last 150 years.

The task of finding quantitative information about what topics economists focused over time or, at least, what their expectations might have been about key indicators, is even more challenging. Besides some business cycles indicators including qualitative information about business expectations at the individual level, such time series are not available neither at individual nor at aggregate level.

To close this gap we employ a quantitative analysis on the discourse in scientific journals in economics. We start with a pilot study on the debate in the Journal of Economics and Statistics for the period between 1949 and 2010. Thereby, we assume that a foreseeable development, in particular one which is considered as a problem, results in an increase of scientific publications with a focus on the particular problem prior to the actual development. This holds true if economists became aware of the problem early enough, given that there might be a noticeable publication lag. If this is not the case, publications in economic journals will only be published after the problem has occurred. Obviously, the latter also applies for shocks which are hardly predictable.

Our empirical approach comprises two stages: First, we have to identify topics discussed in the Journal of Economics and Statistics and their relative importance over time. Second, we have to establish a link between the importance attached to certain topics and the actual development of the economic reality which they might reflect. While the second step makes use of standard approaches from econometric time series analysis, the first step relies on tools from computational linguistics, which more recently, also made their way into economic analysis. However, establishing a link between topic weights and real data appears to be a novel contribution. To the best of our knowledge, only the recent contributions by Hansen, McMahon, and Prat (2014) and Larsen and Thorsrud (2015) follow a similar approach. Hansen, McMahon, and Prat analyze the impact of increased transparency on the functioning of central banks and ultimately on monetary policy using the minutes and transcripts of the Federal Open Market Committee. Larsen and Thorsrud examine the impact of "news" on the business cycle, based on a Norwegian business newspaper followed over a period of 9000 days.

In contrast to our analysis, both articles consider a rather short time span and, in the case of Hansen Hansen, McMahon, and Prat (2014) are based on a narrowly defined text corpus.

Topic models are a mean to classify the content in a large text corpus. The algorithms endogenously identify so called *topics*, which are not necessarily topics in the semantic sense, but rather clusters of words which often appear jointly in a text. Our approach to topic modeling closely follows Griffiths and Steyvers (2004), who also work on a corpus of scientific literature. In their contribution, the authors introduce Gibbs sampling as an algorithm to Latent Dirichlet Allocation (LDA). They classify articles in the *Proceedings of the National Academy of Science of the United States* (PNAS) using LDA. Similarly, Hall, Jurafsky, and Manning (2008) analyze the history of ideas in the field of computational linguistics. In their paper, they provide a convincing visualization of the rise of probabilistic topic models in computational linguistics. Grün and Hornik (2011) conduct an analysis for the Journal of Statistical Software and provide the R Package *topicmodels*[1] along with programming examples. Our own implementation builds partly on their code.

In the second step of the analysis, we use the probabilities assigned to each volume of the Journal for specific topics resulting from the LDA as input for dynamic regression models. In a univariate model for explaining an economic time series related to a topic, both leads and lags of this input variable are used as potentially explanatory variables. This allows us to assess the dynamic dependencies between the relevance of a topic in papers of the Journal with the development of real economic data related to the topic. As a robustness check, we also specify and estimate bivariate VAR-models for the two variables of interest and conduct Granger causality tests on these models.

The remainder of the paper is organized as follows. First, Section 4.2 introduces the text corpus obtained from the Journal of Economics and Statistics and the economic time series used for the further analysis. The following Section 4.3 provides a short explanation of topic modeling. The application of this method and the results obtained for the Journal of

[1] See `https://cran.r-project.org/web/packages/topicmodels/index.html`

Economics and Statistics are subject of Section 4.4. In Section 4.5, the dynamic interrelation between the importance of topics and the respective economic time series is analyzed. Section 4.6 provides concluding remarks and an outlook for further research.

4.2 Text Corpus and Economic Data

4.2.1 Text Corpus

For the purpose of our pilot study, we concentrate on a single economic journal with a close link to economic science in Germany. Furthermore, in order to identify developments over sensible time spans, a journal existing already for a long time period was required. For these reasons, we selected the Jahrbücher für Nationalökonomie and Statistik (also Journal of Economics and Statistics). The journal has been appearing regularly since 1863, with a few exceptions (e.g. during the second World War). There have been 235 volumes to date (2015), with currently one volume being published per year. Due to the effort required for the preparation of text data used in the quantitative analysis, the analysis is restricted to the period from 1949 to 2010. It will be left for future research to include more volumes.

An analysis of a subsample of 250 articles (5%) supports the assumption that scientists publishing in this journal focus mainly on the German economy. Out of the 250 Articles examined only eleven focus on other countries, while fourteen are multi-country studies including Germany, 60 discuss only Germany. The remaining articles have a methodological or theoretical focus and therefore are not linked to empirical findings for a particular country. The focus of the journal on a single country simplifies the second step of the analysis.

Over the years the use of the English language in the economic discourse has been increasing. The language use in the Journal of Economics and Statistics reflects this development, as it accepted articles in English and German. Starting in the 1990s English language articles became more common, but it took until the late 2000s before whole issues appeared in English. In total 80% of the 2500+ documents are in German. For the analysis we do not differentiate between German and English articles, but leave it to the algorithm to differentiate or join topics in both languages.

We had access to the scanned images of all volumes except the most recent ones via `digizeitschriften.de`. The meta data regarding the volumes considered for the present analysis are provided in Table 4.3 in Appendix 4.B. For obvious reasons, the index volumes are excluded from the analysis. No volumes of the journal appeared in 1957 and 1974. Each of the two years was succeeded by a year with two volumes (1958 and 1975). For the periods 1967 to 1970 and 1971 to 1973, there were volumes covering two years. To disentangle the volumes covering multiple years, we made use of the dates on the covers of the single issues to assign the included articles to a calender year. As a consequence, volumes 181 to 183, as well as 186, 187, 191 and 192 were allotted on two years. It turned out that all issues of volume 190 appeared in 1976 and volume 188 covered a period of three years. Allocating the individual issues to calender years solved the case of missing data for 1974. Unfortunately, this did not provide a solution for the case of missing data in 1957. The observations for this year were later imputed by calculating the mean of the values for the preceding and succeeding year.

The source format differed among the volumes. From the year 2000 onward, we had access to digital publications. Older volumes were obtained only as scanned PDF files from `digizeitschriften.de`. We used *Abby Finereader 12 Corporate* to perform Optical Character Recognition (OCR) and turn the documents into text files as the quality of the already existing text files was not sufficient for the purpose of our analysis. The OCR Software retained the formatting of the headlines which we used to break the journal up into single articles which are the units of our analysis. Finally, we used manual labor to clean the texts, e.g. by removing tables, footnotes and equations.

A few further preparatory steps were necessary to come up with the final text corpus for the topic modeling algorithm. For the implementation of these steps, we closely followed Grün and Hornik (2011) by employing the text mining infrastructure supplied in the *R* package *tm*.[2] In particular, the following steps were performed:

- The German language features a large number of grammatical forms of words. Considering every single case would greatly inflate the vocabulary (the set of words which forms the basis for the application of LDA). By stemming of words the different grammatical forms of the same word are reduced to an identical stem. *"The stem is the part of the word that is common to all inflected variants"*, as wikipedia[3] puts it. We apply the stemming algorithms *SnowballC* (setting "german") to produce final word stems. The main feature is the removal of suffixes of words. Consequently, *kaufen*, *kaufe*, *käufer* are all reduced to the stem *kauf*. The algorithm removes the umlaut *ä* and replaces it by the vowel *a* as umlauts are often used in forming the plural form (Example: Ball → Bälle). Unfortunately, these transformations come with a certain loss of information, which to some degree is intended (e.g. removal of plural forms) but also has unintended consequences (*fordern* (request) and *fördern* (support) become indistinguishable). The German stemming also produces useful results for English words, as both languages share a Germanic root.
- All superfluous blanks, newline and tabulator codes, numbers and punctuation marks were removed.
- We removed all German stopwords, i.e. the most common words that would otherwise dominate most of the topics without being linked to specific content. We use the list of stopwords shown in Appendix 4.A as supplied by the R package *tm*.
- Finally, we only considered terms which, after stemming, consisted of five to twenty characters to further reduce the size of the vocabulary. This measure also helped to remove foreign language stopwords, in particular English words (e.g. he, she, it,

[2] See `https://cran.r-project.org/web/packages/tm/index.html` for details about the package.

[3] `https://en.wikipedia.org/wiki/Word_stem`, retrieved December 8th, 2015.

you) and long compounded words (e.g. *Nasenspitzenwurzelentzündung*[4]), which could potentially bias the result.

As output of these preparatory steps we obtain a document term matrix, providing the number of occurrences f_{ij} of each selected term i (column) in each article j (row). In order to only consider the most important terms in the analysis, terms are selected by their relative importance for explaining specific articles (Blei and Lafferty 2009). The relative importance is measured by the *term frequency–inverse document frequency (tf-idf)*. Even though a variety of weighting options exist we adopt the definition used in Grün and Hornik (2011):

$$\begin{aligned} (\text{tf-idf})_{ij} &= \text{tf}_{ij} \cdot \text{idf}_{ij} \text{ where} && (4.1)\\ \text{tf}_{ij} &= \frac{f_{ij}}{\sum\limits_i f_{ij}} && (4.2)\\ \text{idf}_{ij} &= \log_2 \left(\frac{D}{\sum\limits_j \mathsf{I}(f_{ji} > 0)} \right). && (4.3) \end{aligned}$$

$\sum_i$ and $\sum_j$ indicate summation over all terms i and documents j respectively. tf$_{ij}$ is the number of occurrences of term i in relation to the total number of occurrences of all terms in document j. The idf$_{ij}$ value is the logarithm to base 2 of the ratio of the total number of documents in the corpus (D) to the number of documents containing term i, with I being an indicator function for $f_{ij} > 0$. For the analysis we select all terms which are prominent in individual documents, rather than exhibiting a high overall frequency as measured by the idf-value. Hence, for every term i the mean of its *tf-idf*$_{ij}$ values across all documents is calculated. Following Grün and Hornik (2011) we use the median (in our case 0.004) over all *tf-idf*$_{ij}$ values ($\forall i, j$) as a cut-off and only include terms with mean values across documents larger than this median.

For a total of $D = 2\,675$ articles, the number of terms (word stems) considered for the

[4] Example for the "intractable problems of compound words" in German from the original description of the Snowball algorithm `http://snowball.tartarus.org/texts/germanic.html`.

further analysis is reduced by this selection to $|V| = 22\,171$. The resulting matrix $\mathbf{F} = (f_{ij})$ serves as input for the topic modeling algorithm.

4.2.2 Economic Data

As will be described in more detail in Section 4.4, topic modeling results in a substantial number of topics. While many of them can be given an intuitive interpretation as they refer to specific economic theories or institutions, only a small number corresponds closely to some key economic indicators which are also available as long time series, and hence qualify for the second step of our quantitative approach. For this second step, we select data for the five economic issues inflation, trade (net-exports), public debt, unemployment rate and interest rates. These real-world economic time series were chosen due to their availability over a long time period, as well as our ability to identify corresponding topics and their prominent role in the economic debate and theory.

Inflation

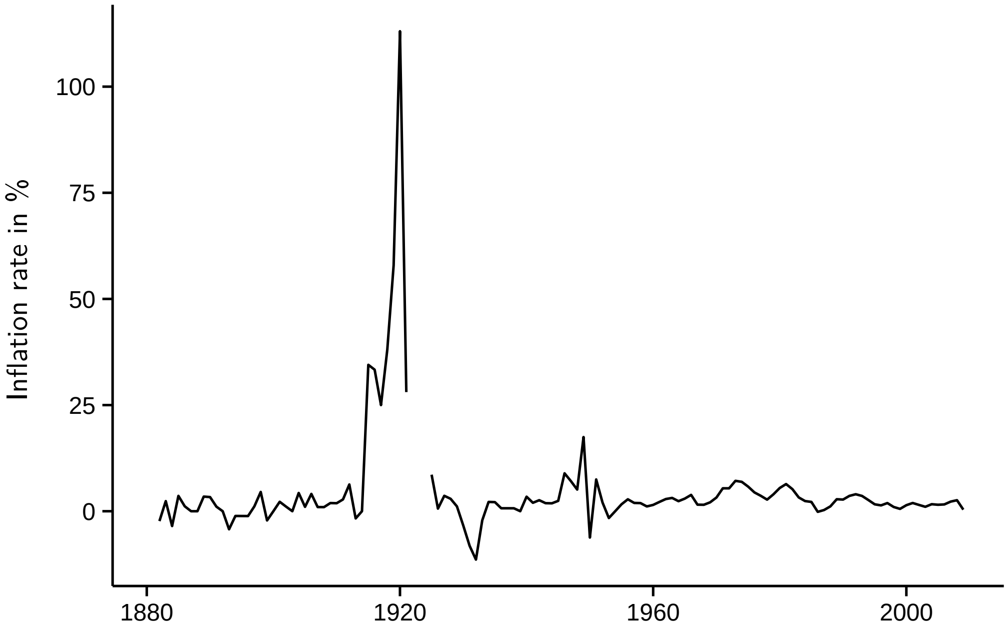
The longest time series we could obtain out of these five domains is the German inflation rate. The German statistical office compiled a very long time series of the consumer price index (CPI), which, due to limited data quality for the early years including the run-up to the hyperinflation after 1918, is only available on request.[5] The price index data covers the period from 1881 to 2009. Due to the period of hyperinflation, there is no price data for 1922 and 1923. In 1948, there are separate values for the first and second half of the year, following the introduction of the D-Mark in June 1948.

Note that we make use of a traditional growth rate given that the approximation by log-differences might deviate substantially for the periods of high inflation or hyperinflation. Figure 4.1 provides a plot of the inflation rate. Obviously, the plot is dominated by the period of German hyperinflation during the 1920s, which dwarfs all other spikes in inflation rates.

[5] Statistisches Bundesamt (2013). Preise – Verbraucherpreisindex Lange Reihe von 1881.

However, this does by no means imply that inflation was not an issue at any other period in time. A second period of monetary instability followed after the second world war. Afterwards, the high inflation period in the 1970s stands out. In the early years, there have been frequent periods of deflation, which were most pronounced in the interwar period. Since the 1940s deflation has become very rare.

Figure 4.1: The German inflation rate 1881 – 2009

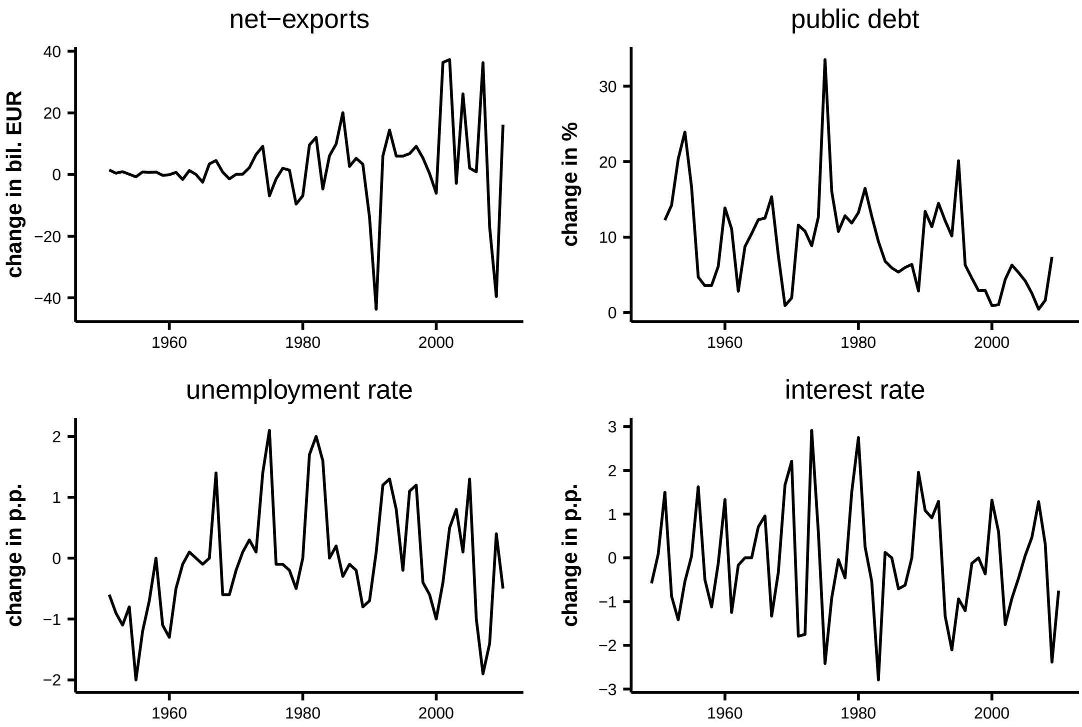


Trade

Trade is operationalized as the German net exports, for which data are available from the German statistical office at a yearly frequency since 1950.[6] We rescaled the original data to billion Euro in order to operate with a similar scale as for the other economic indicators. As the data series is non-stationary, the econometric analysis will be based on the differentiated

[6] *Außenhandel: Zusammenfassende Übersichten für den Außenhandel (Endgültige Ergebnisse)* Fachserie 7 Reihe 1, Issue 2013 from December 2014.

Figure 4.2: Real-world time series



series. Along with debt, the unemployment rate and the interest rates, a time series plot of net-exports is shown in Figure 4.2.

Debt

Data for the German public debt is available since 1950.[7] To mitigate structural breaks in the time series due to changes in the data collection methodology and the presence of non-stationarity, the relative changes in total debt are used in this analysis.

Unemployment Rate

The German unemployment rates (in percent) were retrieved from the GENESIS Database of the German Statistical Office. From 1950 to 1990, the data coverage is limited to West

[7] Statistisches Bundesamt Fachserie 14 Reihe 5: Finanzen und Steuern - Schulden des Öffentlichen Gesamthaushalts Table 1.1.1.

Germany. From 1991 onward, data for the whole of Germany are used. Since the series is non-stationary, the first differences are used for the econometric analysis.

Interest Rates

The time series of interest rates is available from *Deutsche Bundesbank*.[8] From July 1st, 1948 until 1998, the “Diskontzinssatz”, which widely served as a base for financial contracts, was used. It is available at monthly frequency. With the transfer of the authority of the monetary policy to the ECB, this rate was replaced by the so called *Basiszins*. It is still available at a monthly frequency, which is only adjusted every six months. Therefore, we use the yearly average of the interest rate for the econometric analysis.

4.3 Topic Modeling the Economic Discourse

4.3.1 Methods for Quantitative Text Classification

Methods for quantitative text classification originate from the field of information retrieval, where these methods were developed in order to render text electronically searchable. Early applications relied on the tf-idf (term frequency - inverse document frequency) classification, which is still used as a preprocessing step in modern approaches. The classification is based on a simple counting procedure to represent the importance of a term in a document (the term frequency) in relation to the importance of the term in the entire corpus. This method represents a text corpus comprising an arbitrary number of documents as a matrix (term-by-document matrix), with values for any given term in a vocabulary. A standard reference for these early methods is Salton and McGill (1986).

Modern methods also include spherical k-means (Dhillon and Modha 2001; Hornik, Feinerer, et al. 2012) and the related mixtures of Mises-Fisher distributions (Banerjee et al. 2005;

[8] `http://www.bundesbank.de/Navigation/DE/Statistiken/Zeitreihen_Datenbanken/Makrooekonomische_Zeitreihen/makrooekonomische_zeitreihen_node.html` retrieved 01.11.2015.

Hornik and Grün 2014), which might be considered as problem specific clustering algorithms. Our work is based on the widely used class of probabilistic topic models, leaving a comparison with some of the alternatives for future research. While both approaches are unsupervised learning algorithms clustering algorithms assign a single topic to a document instead of determining the mixtures of topics of a single document as in the case of topic models.

In the following we will introduce the basic idea of probabilistic topic modeling in its application to economic discourse. The method exhibits two major aspects. First, it analyzes text without imposing a priori keywords or categories. Instead, clusters of terms appearing together frequently (topics) emerge endogenously. Second, the method has a sound statistical background, allowing the application of standard estimation and inference procedures. We will briefly sketch the historical background of the method, the theoretical model, estimation procedures and the evaluation of modeling outcomes.

Modern topic models can be traced back to Deerwester et al. (1990), who developed *Latent Semantic Analysis (LSA)*[9] as a sophisticated method to overcome shortcomings of the simple tf-idf classification. Subsequently, *Probabilistic Latent Semantic Analysis (pLSA)* introduced a sound statistical foundation to topic modeling (Hofmann 1999). Being based on the likelihood principle, it also defines a generative model of the data (on term level). Building up on pLSA, Blei, Ng, and Jordan (2003) extended the method to *Latent Dirichlet Allocation* (LDA), by adding a probabilistic generative model at the level of the documents. In spite of several extensions made in recent years, e.g. the *correlated topic model (CTM)* (Blei and Lafferty 2007) allowing for correlation of topics across documents and the recent introduction of *TopicMapping* by Lancichinetti et al. (2015), which adds the idea that documents can be described as networks of terms, LDA remains the state of the art in topic modeling. The original estimation method by Blei, Ng, and Jordan (*Variational Expectation Maximization*) was rather slow, which proved to be problematic in some areas (e.g. commercial applications

[9] In the context information retrieval, the method is often called Latent Semantic Indexing (LSI).

in information technology). As a consequence, Bayesian methods as suggested by Griffiths and Steyvers (2004) became widely used.

The recent introduction of structural topic models (stm) by Roberts, Stewart, and Airoldi (2016) represents an interesting extension to topic modeling allowing for the inclusion of document-level meta information at the level of the generative model. This approach makes it possible to make inference about the underlying social processes influencing the creation of documents. While it would provide additional insight into the development of the economic discourse, our dataset contains little in terms of consistent meta information which would be usable for such an analysis. In future research considering a more heterogeneous set of journals, *stm* might be an approach to incorporate the editor's preferences in selecting articles for the journal.

4.3.2 The Theoretical Model of LDA

At the core of LDA lies an abstract theoretical model which describes how documents are created. Thereby, it is assumed that a document is "a mixed bag of words", which implies that the order of words within a document is ignored and just the frequencies of words are considered. In practice, the documents are obtained from machine readable sources (either short as a tweet or long like a journal article) and are usually altered by a preparatory data cleaning step.[10] Each document is assumed to be made up from several topics which determine the probability of each term from the vocabulary to be included in the document.

To be more precise about this data generating mechanism, we introduce some notation. First, we have a *vocabulary* V comprising all terms considered in the analysis. The size of this vocabulary is denoted by $|V|$. Each document is given by a vector of terms $\boldsymbol{w} = (w_1, \dots, w_N)$, where N denotes the length of the document. We do without a document specific index, as the following discussion will focus on a single document. Only in the final step, the results for single documents will be aggregated for the corpus comprising all documents.

[10] See Subsection 4.2.1 for a description of this preprocessing step for the current application.

It is assumed that each term in the text has its origin in some topic. Thereby, topic k, $k = 1, \ldots, K$ is represented by a vector of probabilities $\boldsymbol{\beta_k} = (\beta_{k,1}, \ldots, \beta_{k,|V|})$ assigned to each term in the corpus, i.e. it is characterized by some terms being more frequent than others. For example, if the corpus includes the terms "inflation", "debt" and "growth", a topic with probabilities $(0.9, 0.05, 0.05)$ might be associated with inflation, while a topic with probabilities $(0.05, 0.6, 0.345)$ might be focused on the nexus between public debt and growth. The $K \times |V|$ matrix resulting from stacking all $\boldsymbol{\beta_k}$ vectors is denoted as $\boldsymbol{\beta}$. The vector $\boldsymbol{z} = (z_1, \ldots, z_N)$ denotes the vector of topics giving rise to the terms in $\boldsymbol{w}$. All z_n, $n = 1, \ldots, N$ come from the set of all topics of size K, which is the same for all documents in the corpus. Typically, the z_n are not all different, but rather concentrate on a few topics for a specific document. Furthermore, also the assignment of a term to a topic is not unique as the same term might belong to several topics.

Now, the generative process of a document in the LDA model can be described as follows (see also Grün and Hornik 2011): First, a categorial probability distribution $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ is randomly chosen which describes the relevance of topics within the document.[11] In particular, for all $k = 1, \ldots, K$, $\theta_k \in [0, 1]$ and $\sum_{k=1}^{K} \theta_k = 1$. Next, for each term w_n, $n = 1, \ldots, N$ in the document, a single topic z_n is randomly selected according to the probability distribution $\boldsymbol{\theta}$. Then, according to the probability distribution on the terms of z_n, i.e. $\boldsymbol{\beta_{z_n}}$, the term w_n is drawn.

In conclusion, given all topics and the probabilities of these topics for a document, the random process of generating the document can be described. However, the aim of the analysis is rather the reverse: Only the documents forming the corpus and, consequently, the vocabulary are available, while we are interested in identifying topics and, for each document, the relevance (probability) of each topic. How this can be achieved by adding some assumptions on the generative process is described in the following subsection.

[11] In the literature on LDA modeling, this distribution is often labeled as a multinomial distribution which is adequate assuming just the outcome of one draw.

4.3.3 Estimation of LDA models

The theoretical model presented in the previous subsection requires a substantial number of parameters, namely $K \cdot |V|$ probabilities $\beta_{k,i}$ and the number of documents times K probabilities θ_k. Given these parameters, the probability of the observed documents can be calculated. However, as in a standard maximum likelihood setting, we are interested in "reversing" the argument and obtaining estimates of the parameters given the observed documents, i.e., we are searching those parameter settings making it most likely to observe our documents and, consequently, determine topics and their relevance for individual documents endogenously. Given the number of parameters and the functional interdependencies between the $\beta_{k,i}$ and the θ_k, it turns out that a straightforward maximum likelihood approach is not feasible without imposing additional constraints and, possibly, using alternative optimization/estimation procedures (Griffiths and Steyvers 2004, p. 5229).

A first simplifying assumption consists in considering the categorial distribution $\boldsymbol{\theta}$ as a random draw from a uniform Dirichlet distribution with scaling parameter α, i.e., $\boldsymbol{\theta} \sim \text{Dir}(\alpha)$ (Blei, Ng, and Jordan 2003). Then, for given parameters α and $\boldsymbol{\beta}$, the probability of observing a specific $\boldsymbol{\theta}$, a set of N terms $\boldsymbol{w}$ and corresponding topics $\boldsymbol{z}$ is given by (Blei, Ng, and Jordan 2003, p. 996):

$$p(\boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{z}) = p(\boldsymbol{\theta}|\alpha) \prod_{n=1}^{N} p(w_n|z_n, \boldsymbol{\beta}) p(z_n|\boldsymbol{\theta}) \, . \tag{4.4}$$

Integrating over the random vector $\boldsymbol{\theta}$ and summing over the components of $\boldsymbol{z}$ results in the marginal distribution for a single document. Finally, by calculating the product of the marginal properties of all documents of the corpus, the probability of the corpus is obtained. Despite the simplification by considering $\boldsymbol{\theta}$ as a random draw from $\text{Dir}(\alpha)$, maximum likelihood estimation still does not appear to be feasible (Griffiths and Steyvers 2004, p 5229).

To overcome this problem, a variety of (approximate) estimation procedures have been suggested. The original procedure is a variant of the *expectation maximization* (EM) algorithm, the so called *variational expectation maximization* (VEM), which is the method suggested by Blei, Ng, and Jordan (2003) when introducing LDA. Shortly afterwards, Griffiths and Steyvers (2004) proposed the use of Gibbs sampling, which – according to the authors – exhibits faster convergence. This method has become widely applied and is used for our empirical application. For a comparison of the two methods see Welling, Teh, and Kappen (2008).

The method requires using a further assumption regarding the data generating process. It is assumed that the term distribution is a random draw from a Dirichlet distribution with parameter δ. Using this assumption, the probability for a single document according to Equation (4.4) could be obtained by integrating out $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, which might be done separately as $\boldsymbol{\beta}$ appears only in the first and $\boldsymbol{\theta}$ in the second term. As is shown by Griffiths and Steyvers (2004, p 5229), the resulting expressions still do not allow for a direct calculation. Therefore, they propose to apply a Markov Chain Monte Carlo approach. They provide details on how the probability for each topic is updated based on the distribution for all other topics. The Markov chain is constructed to converge to the target distribution by repeated sampling from the target distribution.

While the sampling algorithm provides direct estimates of the association of a term to a topic, we are interested in the predictive topic distribution across documents. After the Markov Chain has converged, using a set of samples, as an estimate of the predictive distribution ($\boldsymbol{\theta}$) can be obtained. The approximation relies on the number of times ($\eta_k^{\boldsymbol{w}}$) the algorithm associated document $\boldsymbol{w}$ with topic k:

$$\hat{\theta}_k^{\boldsymbol{w}} = \frac{\eta_k^{\boldsymbol{w}} + \alpha}{\sum\limits_k \eta_k^{\boldsymbol{w}} + K\alpha} \tag{4.5}$$

4.3.4 Model Validation

There are two approaches to compare topic models with respect to the choice of the number of topics. One approach focuses on fitting the model on a subset (e.g. 90%) of data and evaluating the fit for the remaining data. However, Chang et al. (2009) argue that this approach does not result in models which are appealing to human judgment.

A more convenient method is introduced by Griffiths and Steyvers (2004, p. 5231). Ideally one would compare the different models based on the likelihood as a function of the number of topics K, which involves summing over all possible assignments of words to topics. Griffiths and Steyvers (2004, p. 5231) circumvent the resulting computational issue by approximating the likelihood by the harmonic mean of a set of values which are calculated from samples provided by the Gibbs sampling algorithm. For reasons of completeness it shall be mentioned that there is also some criticism regarding this method of approximation. See Buntine (2009), as well as Wallach et al. (2009) for an overview on the critique and alternative approximation methods.

While yielding good results in our analysis, selecting a model by maximization of the likelihood occasionally leads to an unreasonable large number of topics, which become hard to interpret. Consequently, authors sometimes deviate from the estimated number of topics according to the likelihood approach to allow for a more straightforward interpretation of the topics. Nonetheless, in our case estimating the number of topics according to this method results in topics allowing for a meaningful interpretation and stable results.

4.4 Taking LDA to the Data

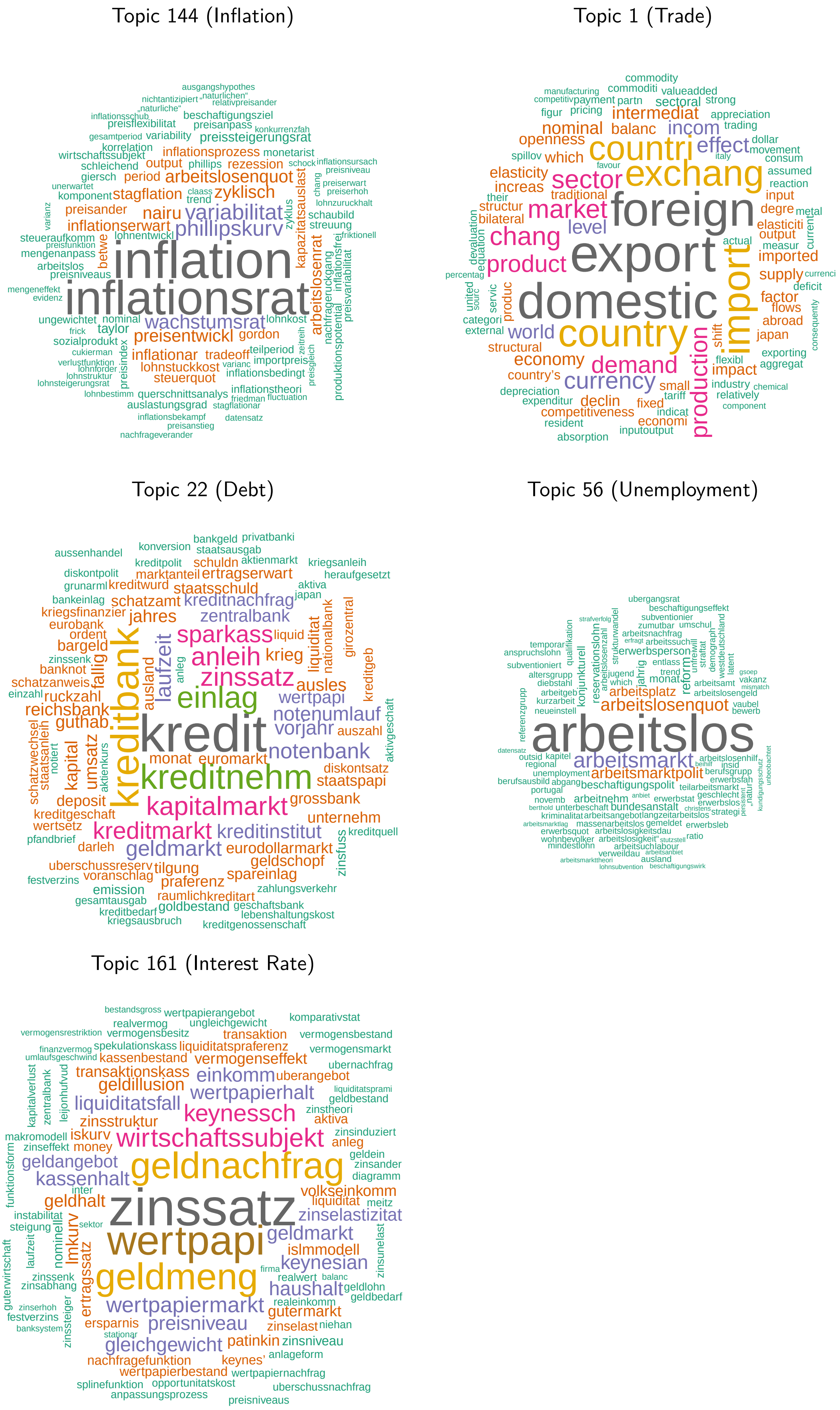
Due to the size of the dataset we estimated the topics using Gibbs sampling. Repeating the estimation for a number of topics (K) between 2 and 1000, we found K=165 to be the optimal choice based on the harmonic mean method. Finding an optimal number of topics through maximization is subject to some difficulties as the function is not smooth. It shall be noted that the original application by Griffiths and Steyvers (2004) was aimed at providing a rough estimate of the magnitude of the value for K.

Apart from K, some parameters had to be chosen a priori. We stick to $\alpha = 1/K$ and

$\delta = 0.1$, chosen according to the literature (Griffiths and Steyvers 2004). Afterwards the Markov chain is run for 2000 iterations, which we, following Grün and Hornik (2011, p 10), assume sufficient for it to converge. From the resulting 165 topics we select five topics that, to our understanding, are the ones most closely related to the economic indicators introduced in Section 4.2.2. These topics are presented in Figure 4.3, where the font sizes of the terms indicate their relative importance within a topic. The discussion of trade (topic 1) is the only topic that is primarily discussed using the English language, which might be explained to some degree by the international interest in trade itself. The German language equivalent of the topic can be found in Appendix 4.D. The topic concerned with debt (topic 22) appears to be centered on loans given to companies and individuals. The terms describing sovereign debt (e.g. *Staatsschulden*) are part of the topic but do not show up with the same frequency. The stem *arbeitslos* is the single most significant term in topic 56 (unemployment), all other somewhat significant words are compound words closely related to employment. Topic 144 is concerned with the discussion of inflation and the inflation rate, also the term *Phillipskurve* shows up prominently describing a theoretical framework in which inflation is often discussed. The fifth and last topic considered is based around the term *zinssatz* [en: interest rate]. The discussion also encompasses finance (*geldmarkt* [en: money market]) and macroeconomics (Preisniveau [en: price level], Liquiditätsfalle [en: liquidity trap]).

Even though Figure 4.3 shows only a small subset of the 165 topics, each of the topics can be attributed to a particular idea or debate in economics. Appendix 4.D shows additional topics which appear to be closely related to the ones used here. Any further attempt to derive the stories behind these topics in greater detail should also involve qualitative analysis, i.e. a careful reading of those documents exhibiting high probabilities for the topic of interest.

Figure 4.3: Identified key topics



Topic 144 (Inflation)

Topic 1 (Trade)

Topic 22 (Debt)

Topic 56 (Unemployment)

Topic 161 (Interest Rate)

4.5 The Relationship Between Discourse and Economic Data

4.5.1 Univariate Dynamic Model

The relationship between a real economic indicator and the logarithm of the sum over probabilities for the corresponding topic of all documents in a given year is estimated by linear regression models. The results are shown in Table 4.1. The augmented distributed lag model includes both leads and lags of the topic indicator as explanatory variables for the current value of the economic indicator. Statistically significant parameters for lagged values indicate that the scientific discussion on the topic precedes changes in the economic variable, while statistically significant leads point at a scientific discussion following the developments of economic indicators. The model selection procedure considers all models with lagged values up to three years and leading values for up to three years. From all possible 1024 subsets of these potential explanatory variables, the selected model is the one minimizing the Akaike Information Criteria (AIC).

In case of the relationship between inflation and topic 144 the regression relationship takes the form of:

$$\mathrm{infl}_t = \beta_0 + \beta_1 \times \mathrm{t144}_t + \beta_2 \times \mathrm{t144}_{t+3} + \beta_3 \times \mathrm{t144}_{t-3} + \beta_4 \times \mathrm{infl}_{-1} + u_t \tag{4.6}$$

The inflation rate $infl_t$ is explained by the weight of topic 144 (t144) at time t, $t+3$ and $t-3$. The other regressions differ with respect to the economic time-series and explanatory topics used and the lags and leads included.

Table 1 summarizes the estimation results. Each column provides the estimated coefficients for a particular economic variable exhibited in the row labeled "dependent variable", while the number of the corresponding topic is shown in the next row. The row labeled topic without a number corresponds to the parameter for the instantaneous effect. For the following rows, positive numbers in parentheses indicate leads, while negative numbers indicate lags. We also allow for lagged values of the endogenous variables and a deterministic linear trend.

For all models, we find a link between the importance of the topic in the scientific discussion and the observed economic indicator with the adjusted R^2 ranging from 0.23 to 0.61. However, it turns out that there are differences across models with respect to the role of lagged, leading and contemporaneous effects. The use of logarithms for the topic probabilities allows us to interpret the coefficients[12] β as short-run semi-elasticities: A 1% percent increase in the discussion on the topic is associated in a change by $0.01 \times \beta$ units of the indicator.

One needs to be careful when interpreting the time lag as indicated by the regression analysis. While the analysis treats the time of publication and the time of writing as identical, in reality there is, in most cases, a notable publication lag between those two dates. A long publication lag would make it less likely that one finds evidence of the hypothesis that the discourse leads the real-world economic time-series. Hence, our analysis may be described as conservative as it under- rather than overestimates the support for the hypothesis that economic discourse leads real developments.

Without taking into account a potential publication lag, the results of the dynamic regression models in Table 4.1 allow for the following conclusions. More discussion of the topic related to inflation in the past has no significant effect, and contemporaneous discussion appears even negatively related to the actual inflation rate. In contrast, the link between an increase in inflation and future discussion of the topic is statistically significant, but the absolute size of the effect is moderate: An increase in current inflation by 0.058 percentage points would correspond to an increase of the topic weight by 10% three years later. For net-exports, a statistically significant positive effect of past discussion of the topic on current values is found, while an increase in net exports today rather seems to reduce future discussion of the topic as indicated by the statistically significant negative coefficient for Topic (1). For debt, only a

[12] Note: We use β by convention, it is not to be confused with the parameter of the same name from the LDA model.

Table 4.1: Regression results

Number	(1)	(2)	(3)	(4)	(5)
Dependent Var.	Inflation	d(NetExp)	$\frac{\text{debt}_t - \text{debt}_{t-1}}{\text{debt}_{t-1}}$	d(unemp)	d(interest)
Topic	Topic 144	Topic 1	Topic 22	Topic 56	Topic 161
Constant	3.4544^{***}	-86.9928^{**}	4.8936	0.8935	-0.1028
	(1.2546)	(40.197)	(0.0478)	(5.3723)	(0.8609)
Topic	-0.3077^{*}	-5.7066^{**}		0.2593^{**}	
	(0.1778)	(2.4449)		(0.1218)	
Topic (1)		-5.6708^{**}			0.2587^{*}
		(2.1887)			(0.1455)
Topic (2)				0.3052^{**}	0.1977
				(0.1172)	(0.1452)
Topic (3)	0.5810^{***}	-0.7422			
	(0.1807)	(2.1619)			
Topic (-1)				-0.3956^{**}	-0.2402
				(0.1247)	(0.1598)
Topic (-2)		5.2182^{**}			
		(2.5121)			
Topic (-3)	0.1385		-2.3133^{***}		-0.2309
	(0.1675)		(0.8411)		(0.1514)
Endogenous (-1)	0.5653^{***}		0.5606^{***}	0.7148^{***}	0.1874
	(0.0971)		(0.1140)	(0.1140)	(0.1164)
Endogenous (-2)				-0.3032^{**}	-0.4741^{***}
				(0.1297)	(0.1250)
lin. Trend		0.5482^{***}	-0.1449^{***}		
		(0.2024)	(0.0492)		
adj R^2	0.606	0.2275	0.4024	0.4505	0.2635
F	22.174	4.299	13.7981	10.0214	4.3407
p(F)	0.0	0.002	0.0	0.0	0.0013
N	56	57	56	56	57

Note: standard errors in parenthesis

*, ** and *** indicate significance at 10%, 5% and 1% levels respectively

All topic probabilities in logarithms

d() indicates first differences

statistically significant positive effect of past discussion of the topic (3 years ago) with its actual change is obtained, while for unemployment again links in both directions are found. While past discussion appears to be negatively correlated with the unemployment rate, the nexus becomes positive for the discussion in the future. There is only a weak link between an increase in the interest rate and an increase of the relevance of the corresponding topic in the future.

4.5.2 Robustness Checks

In the review process two issues regarding our estimation procedure have been pointed out. Hence, we provide additional robustness checks on the stability of our results. In the baseline model we used natural logarithms to transform the approximately log-normal distributed topic probabilities. This scales the values between $-\infty$ and 0. Using instead the inverse cumulative standard normal distribution function, providing a mapping from $[0, 1]$ to $[-\infty, +\infty]$, to transform the data leads to results which are only marginally different in qualitative terms.

The linear relationship chosen for our univariate regression model might also be challenged. In general, it is assumed that it represents a fairly good first order approximation for a monotonic relationship. However, alternative functional forms might fit better if scientific interest is raised to the same extent by negative and positive (expected) changes of an economic indicator or negative and positive (expected) deviations from some benchmark, respectively. As a further robustness check, we consider the squared deviation from the sample mean for the inflation rate, and the squared (relative) differences for the other economic indicators as dependent variables. The qualitative findings appear robust for the inflation rate, the relative change of debt and the change of interest rates. In contrast, for the regressions for net-exports and the unemployment rate, the significance of the coefficients of interest disappears providing support for the linear form rather than for a symmetric reaction of scientific interest to positive and negative (expected) changes.

4.5.3 VAR-Model

As a robustness check for our empirical results, we also estimate VAR-models. A major advantage of this model class is that both variables under consideration are treated as jointly endogenous, while all explanatory variables are lagged values of these endogenous variables. Thus, one does not need to make arbitrary choices on which of the time-series are endogenous and exogenous. On the downside, the model does not allow for an explicit modeling of contemporaneous dependencies, which only show up through a correlation of error terms. Moreover, VAR models allow conducting tests for Granger causality (Lütkepohl 2007, pp 102f), which assess the usefulness of each time-series to predict the other.

For the VAR model, we do not consider holes in the lag structure (Winker 2000) to check whether the subset selection procedure applied for the dynamic model in Subsection 4.5.1 has a qualitative impact on the findings. The lag length for the VAR model is selected based on the AIC with maximum lag length of six years. The lag lengths used for the VAR models for the different variables are reported in the last row of Table 2.

The VAR model and the corresponding Granger causality tests are illustrated using the first pair of variables from our application, i.e. inflation rate (infl_t) and the weight of topic 144 ($\mathsf{t144}_t$) over time. The current values of both, the inflation rate and the topic weight are modeled as depending on their own past values and the past values of the other variable. Given that the optimum lag length according to AIC for this pair is three years, the VAR model is given by:

$$\mathsf{infl}_t = \alpha_{1,1}\mathsf{t144}_{t-1} + \ldots + \alpha_{1,3}\mathsf{t144}_{t-3} + \alpha_{1,4}\mathsf{infl}_{t-1} + \ldots + \alpha_{1,6}\mathsf{infl}_{t-3} + \varepsilon_{1,t} \tag{4.7}$$

$$\mathsf{t144}_t = \alpha_{2,1}\mathsf{t144}_{t-1} + \ldots + \alpha_{2,3}\mathsf{t144}_{t-3} + \alpha_{2,4}\mathsf{infl}_{t-1} + \ldots + \alpha_{2,6}\mathsf{infl}_{t-3} + \varepsilon_{2,t} \tag{4.8}$$

As the explanatory variables are the same for both equations, the system of equations can be estimated by conducting a simple OLS regression for each equation separately (Lütkepohl 2007, p 72). The estimated parameters jointly reflect the intertemporal dependencies between the two variables. Therefore, considering individual parameters and testing their statistical significance is not very informative. Instead, tests for Granger causality are performed. A variable A Granger-causes variable B if its lagged values have a statistical significant influence on B beyond the dynamics of B already reflected by the lagged values of the endogenous variable B itself. For example, in equation (4.7), testing the null hypothesis that $\alpha_{1,1}$, $\alpha_{1,2}$ and $\alpha_{1,3}$ are all zero corresponds to the statement that the past development of the topic weights has no additional explanatory power for the current inflation rate going beyond the information already contained in the past development of the inflation rate itself (i.e., in the parameters $\alpha_{1,4}$, $\alpha_{1,5}$ and $\alpha_{1,6}$). This null hypothesis is labeled as "topic 144 is not Granger causal for inflation". It is tested by means of a Wald test. The test statistic asymptotically follows a χ^2-distribution with the number of degrees of freedom corresponding to the lag length of the selected model. Table 2 provides the test statistics and the marginal p-values for all variable pairs and both directions of potential Granger causality.

Table 4.2: Test for Granger Causality

	(1)	(2)	(3)	(4)	(5)
Econ. Var.	inflation	d(netexp)	$\frac{\text{debt}_t - \text{debt}_{t-1}}{\text{debt}_{t-1}}$	d(unemp)	d(interest)
Topic	144	1	22	56	161
			topic → reality		
χ^2	3.0901	2.4045	11.416	6.2377	0.3179
p-value	0.3779	0.3005	0.0096	0.0442	0.8530
			topic ← reality		
χ^2	9.4342	1.4324	0.882	2.5116	4.9948
p-value	0.0240	0.4885	0.8297	0.2848	0.0822
Lag length (max. 6)	3	2	3	2	2

Overall, the results of Granger causality testing are consistent with those found for the single equation models with the exception of net exports. For the link between the inflation rate and topic 144, we find a lag length of three which is identical to the one in the single equation setting. We cannot reject the null hypothesis that scientific discussion (i.e. topic 144) does not Granger cause the inflation rate, while the reverse hypothesis has to be rejected at the 5% level. This implies that scientific discussion is affected by developments in the inflation rate, but not vice versa. The model for the change of net exports and topic 1 differs from the single equation model by a lag length of two instead of three. Furthermore, for the single equation model, a deterministic time trend and the current value of the topic weight were found to be influential. This contemporaneous effect is reflected in a high correlation of error terms between the two equations of the VAR model, but does not affect the Granger causality test which might explain the missing evidence for a link in both directions. The VAR model for the change of debt includes three lags, which is in line with the single equation model exhibiting no leads and three lags. Nevertheless, again a significant impact of past discussion in economic science of topic 22 on current changes in debt level in the sense of Granger causality is found, while the actual development of debt does not exhibit Granger causality on the discourse. For unemployment, both models suggest a lag length of two. While the single equation model suggests dependencies in both directions, the null hypothesis of no Granger causality could be rejected at the 5% level only for the influence of past discussion regarding the topic on current changes of unemployment. Finally, for the interest rate, the VAR model comprises only two lags, while the maximum lag and lead length found in the single equation model was three. Nevertheless, the qualitative findings are again similar. While the past discourse on the interest topic is not Granger causal for the current changes in interest rates, the null hypothesis of no Granger causality running from interest rate changes to the extent of scientific discussion about interest rates is rejected at the 10% level though not at the 5% level. Two out of five times we find evidence in favor of our hypothesis that the discussion is leading the economic development and two times the evidence is supporting the hypothesis that the discussion is reacting to the economic development. In the case of net-exports we find evidence for neither hypothesis.

4.6 Conclusion

It was demonstrated, how probabilistic topic modeling of scientific publications in economics and actual economic developments can be put in perspective. To this end, first the corpus of articles published in the Journal of Economics and Statistics between 1949 and 2010 was analyzed by means of LDA resulting in an endogenously created list of relevant topics. Most of the topics found make sense from a semantic point of view and a substantial part of them can be given an immediate interpretation related to economic theory, economic institutions or developments of economic variables.

The second step of the analysis concentrated on those topics which are found to be closely linked to economic indicators available for a long enough time period to allow for a dynamic econometric modeling at annual frequency. This econometric analysis was conducted both with single equation models allowing for lags and leads of the topic weights to have an impact on the current value of the economic variable of interest as well as with VAR models. For all five variables under consideration (inflation rate, net-exports, debt, unemployment rate and interest rate), a relevant – and mostly also statistically significant – link between scientific discussion in the journal and real developments could be found. However, the direction of the influence along the time dimension is not uniform across the models. While a lead of economic discussion with respect to the realization of variables is found for debt and unemployment, the temporal dependency is the other way round for inflation and interest rates, while the dependency appears to be most pronounced contemporaneously for trade. Given a potential publication lag, one may interpret this as additional evidence in favor of the discussion leading the real world developments.

While the proposed two-stage quantitative approach appears promising as an additional tool for analyzing the development of economic thought over time, it will have to be extended in future work in various directions. First, the constraint on a single journal caused by available resources for digitalization, text recognition and data preparation has to be overcome by extending the analysis to other journals being published over a long period. Second, although the application of a specific implementation of LDA using Gibbs sampling for the estimation works well for the present corpus, it has been reported in the literature that the robustness of these methods is limited. Therefore, further research should be devoted on improving the modeling and estimation procedure, and testing alternative approaches. Third, our econometric analysis at the second step of the analysis does not take into account the fact that the topic weights are generated data, which might have an impact on the inference in the second step. It does not appear obvious to us, however, how the uncertainty from the first step might be modeled statistically without using a bootstrap approach which does not appear to be feasible with available computational resources given the high computational complexity of the first step. Hence, any extension in this direction poses an important challenge for future research. Fourth, one should consider a comprehensive multivariate analysis and consider topic correlation and cointegration at the stage of the topic model and the subsequent regression analysis. In addition one may also introduce document level meta data through an STM approach (Roberts, Stewart, and Airoldi 2016) in order to shed light on the economic debate and identify potential effects of individual editors. Finally, and most importantly, the purely quantitative approach used here does not represent a substitute to classical hermeneutic analysis, it rather provides a complementary method to detect relevant fields of research (topics) and how they developed over time putting them in perspective to real economic developments.

Bibliography

Banerjee, A., I. S. Dhillon, J. Ghosh, and S. Sra (2005). “Clustering on the unit hypersphere using von Mises-Fisher distributions”. In: *Journal of Machine Learning Research* 6, pp. 1345–1382.

Blei, D. M. and J. D. Lafferty (2007). “A correlated topic model of Science”. In: *Annals of Applied Statistics* 1.1, pp. 17–35.

Blei, D. M. and J. D. Lafferty (2009). “Topic models”. In: *Text mining: classification, clustering, and applications*. Ed. by A. N. Srivastava and M. Sahami. Data Mining and Knowledge Discovery Series. Boca Raton, Florida, USA: CRC Press. Chap. 4, pp. 71–94.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). “Latent dirichlet allocation”. In: *Journal of machine Learning Research* 3, pp. 993–1022.

Buntine, W. (2009). “Estimating Likelihoods for Topic Models”. In: *Advances in Machine Learning*. Vol. 5828. Lecture Notes in Computer Science, pp. 51–64.

Burret, H. T., L. P. Feld, and E. A. Köhler (2013). “Sustainability of Public Debt in Germany – Historical Considerations and Time Series Evidence”. In: *Journal of Economics and Statistics (Jahrbücher fuer Nationalökonomie und Statistik)* 233.3, pp. 291–335.

Chang, J., J. Boyd-Graber, C. Wang, S. Gerrish, and D. M. Blei (2009). “Reading Tea Leaves: How Humans Interpret Topic Models”. In: *Neural Information Processing Systems*.

Deerwester, S. C., S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman (1990). “Indexing by latent semantic analysis”. In: *Journal of the American Society for Information Science* 41.6, pp. 391–407.

Dhillon, I. S. and D. S. Modha (2001). “Concept Decompositions for Large Sparse Text Data Using Clustering”. In: *Machine Learning* 42.1, pp. 143–175.

Griffiths, T. L. and M. Steyvers (2004). “Finding scientific topics”. In: *Proceedings of the National Academy of Sciences* 101.suppl 1, pp. 5228–5235.

Grün, B. and K. Hornik (2011). "topicmodels: An R Package for Fitting Topic Models". In: *Journal of Statistical Software* 40.13, pp. 1–30.

Hall, D., D. Jurafsky, and C. D. Manning (2008). "Studying the History of Ideas Using Topic Models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '08, pp. 363–371.

Hansen, S., M. McMahon, and A. Prat (2014). *Transparency and Deliberation within the FOMC: A Computational Linguistics Approach*. CEP Discussion Papers dp1276. Centre for Economic Performance.

Hofmann, T. (1999). "Probabilistic latent semantic indexing". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 50–57.

Hornik, K., I. Feinerer, M. Kober, and C. Buchta (2012). "Spherical k-means clustering". In: *Journal of Statistical Software* 50.10, pp. 1–22.

Hornik, K. and B. Grün (2014). "movMF: An R Package for Fitting Mixtures of von Mises-Fisher Distributions". In: *Journal of Statistical Software* 58.10, pp. 1–31.

Lancichinetti, A., M. I. Sirer, J. X. Wang, D. Acuna, K. Körding, and L. A. N. Amaral (2015). "High-Reproducibility and High-Accuracy Method for Automated Topic Classification". In: *Physical Review X* 5.011007.

Larsen, V. H. and L. A. Thorsrud (2015). *The Value of News*. Working Papers 0034. Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School.

Lütkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*. Berlin: Springer.

Rahlf, T. (2016). "The German Time Series Dataset 1834–2012". In: *Journal of Economics and Statistics (Jahrbücher fuer Nationalökonomie und Statistik)* 236.1, pp. 129–143.

Roberts, M. E., B. M. Stewart, and E. M. Airoldi (2016). "A model of text for experimentation in the social sciences". In: *Journal of the American Statistical Association. forthcoming.*

Salton, G. and M. J. McGill (1986). *Introduction to Modern Information Retrieval*. New York, USA: McGraw-Hill.

Turner, A., A. Haldane, P. Woolley, S. Wadhwani, C. Goodhart, A. Smithers, A. Large, J. Kay, M. Wolf, P. Boone, S. Johnson, and R. Layard (2010). *The Future of Finance: The LSE Report*. London School of Economics & Political Science.

Wallach, H. M., I. Murray, R. Salakhutdinov, and D. Mimno (2009). "Evaluation Methods for Topic Models". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09, pp. 1105–1112.

Welling, M., Y. W. Teh, and B. Kappen (2008). "Hybrid variational/MCMC inference in Bayesian networks". In: *Proceedings on the 24th Conference on Uncertainty in Artificial Intelligence*.

Winker, P. (2000). "Optimized Multivariate Lag Structure Selection". In: *Computational Economics* 16.1-2, pp. 87–103.

4.A German stopwords

The following *stopwords* are removed from the vocabulary. The list is supplied by the r package *tm*.

aber alle allem allen aller alles als also am an ander andere anderem anderen anderer anderes anderm andern anderr anders auch auf aus bei bin bis bist da damit dann der den des dem die das daß derselbe derselben denselben desselben demselben dieselbe dieselben dasselbe dazu dein deine deinem deinen deiner deines denn derer dessen dich dir du dies diese diesem diesen dieser dieses doch dort durch ein eine einem einen einer eines einig einige einigem einigen einiger einiges einmal er ihn ihm es etwas euer eure eurem euren eurer eures für gegen gewesen hab habe haben hat hatte hatten hier hin hinter ich mich mir ihr ihre ihrem ihren ihrer ihres euch im in indem ins ist jede jedem jeden jeder jedes jene jenem jenen jener jenes jetzt kann kein keine keinem keinen keiner keines können könnte machen man manche manchem manchen mancher manches mein meine meinem meinen meiner meines mit muss musste nach nicht nichts noch nun nur ob oder ohne sehr sein seine seinem seinen seiner seines selbst sich sie ihnen sind so solche solchem solchen solcher solches soll sollte sondern sonst über um und uns unse unsem unsen unser unses unter viel vom von vor während war waren warst was weg weil weiter welche welchem welchen welcher welches wenn werde werden wie wieder will wir wird wirst wo wollen wollte würde würden zu zum zur zwar zwischen

4.B Tables

Table 4.3: List of volumes

Vol	Year	Vol	Year	Vol	Year	Vol	Year	Vol	Year	Vol	Year	Vol	Year
1	1863	38	1882	74	1900	111	1918	147	1938	184	1970	221	2001
2	1864	39	1882	75	1900	112	1919	148	1938	185	1971	222	2002
3	1864	40	1883	76	1901	113	1919	149	1939	186	71/72	223	2003
4	1865	41	1883	77	1901	114	1920	150	1939	187	72/73	224	2004
5	1865	42	1884	78	1902	115	1920	151	1940	188	1975	225	2005
6	1866	43	1884	79	1902	116	1921	152	1940	189	1975	226	2006
7	1866	44	1885	80	1903	117	1921	153	1941	190	~~75/~~76	227	2007
8	1867	45	1885	81	1903	118	1922	154	1941	191	76/77	228	2008
9	1867	46	1886	82	1904	119	1922	155	1942	192	77/78	229	2009
10	1868	47	1886	83	1904	120	1923	156	1942	193	1978	230	2010
11	1868	48	1887	84	1905	121	1923	157	1943	194	1979		
12	1869	49	1887	85	1905	122	1924	158	1943	195	1980		
13	1869	50	1888	86	1906	123	1925	159	1944	196	1981		
14	1870	51	1888	87	1906	124	1926	160	1944	197	1982		
15	1870	*r*	*1888*	88	1907	125	1926	161	1949	198	1983		
16	1871	52	1889	89	1907	126	1927	162	1950	199	1984		
17	1871	53	1889	90	1908	127	1927	163	1951	200	1985		
18	1872	54	1890	91	1908	128	1928	164	1952	201	1986		
19	1872	55	1890	92	1909	129	1928	165	1953	*202*^{r}	*1986*		
20	1873	56	1891	93	1909	130	1929	166	1954	203	1987		
21	1873	57	1891	94	1910	131	1929	167	1955	204	1988		
22	1874	58	1892	95	1910	132	1930	168	1956	205	1988		
23	1874	59	1892	96	1911	133	1930	169	1958	206	1989		
24	1875	60	1893	97	1911	134	1931	170	1958	207	1990		
25	1875	61	1893	98	1912	135	1931	171	1959	208	1991		
26	1876	62	1894	99	1912	*r*	*1931*	172	1960	209	1992		
27	1876	63	1894	100	1913	136	1932	173	1961	210	1992		
28	1877	64	1895	101	1913	137	1932	174	1962	211	1993		
29	1877	65	1895	102	1914	138	1933	175	1963	212	1993		
30	1878	66	1896	103	1914	139	1933	176	1964	213	1994		
31	1878	67	1896	104	1915	140	1934	177	1965	214	1995		
32	1879	68	1897	105	1915	141	1935	178	1965	215	1996		
33	1879	69	1897	106	1916	142	1935	179	1966	216	1997		
34	1879	70	1898	107	1916	143	1936	180	1967	217	1998		
35	1880	71	1898	108	1917	144	1936	181	67/68	218	1999		
36	1881	72	1899	109	1917	145	1937	182	68/69	219	1999		
37	1881	73	1899	110	1918	146	1937	183	69/70	220	2000		

Note: The volumes marked *r* are index volumes, only 202 carries a volume number.

Notes on the list of volumes

181 Issue 4 was the first to appear in 1968 (March)

182 Issue 4–5 was the first to appear in 1969 (March)

183 Issue 5 was the first to appear in 1970 (February)

184 completely appeared in 1970

185 completely appeared in 1971

186 Issue 3 was the first to appear in 1972 (February)

187 Issue 2 was the first to appear in 1973 (January)

188 Issue 1 Appeared in 1973 (December), Issue 2–5 appeared in 1974 (January to November), Issue 6 appeared in 1975 (February)

190 All issues appeared in 1976 (contrary to the available meta data)

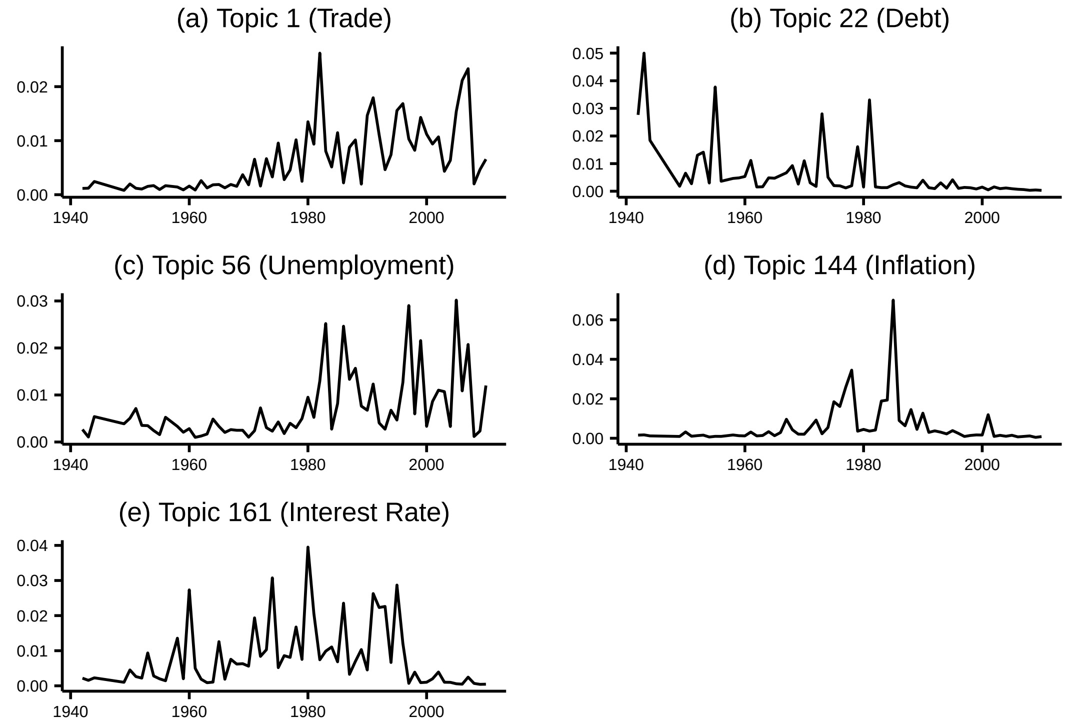
191 Issue 4 was the first to appear in 1977 (February)

192 Issue 5 was the first to appear in 1978

4.C Topic Probabilities

Figure 4.4 shows the development of probabilities for the key topics between 1948 and 2010.

Figure 4.4: Topic probabilities



(a) Topic 1 (Trade)

(b) Topic 22 (Debt)

(c) Topic 56 (Unemployment)

(d) Topic 144 (Inflation)

(e) Topic 161 (Interest Rate)

4.D Further Topics

The following pages show additional topics identified by the LDA algorithm. In addition to the key topics used in the analysis, there are further topics in the field of inflation (Figure 4.5), trade (Figure 4.6), debt (Figure 4.7) unemployment (Figure 4.8) and interest rates (Figure 4.9). This list of of fields is far from being exhaustive. There are a variety of other topics discussed in the journal (see examples in Figure 4.10), which are not easily operationalized as the discussion of capitalism and Marxism (Topic 100) or may not very interesting from an economic point of view (e.g. "terms describing a table" in Topic 165).

While Topic 144, which we used in the analysis, is narrowly focused on inflation and the inflation rate, there are further topics related to inflation (Figure 4.5), Topic 119 is concerned with *geldpoliti* [en: monetary policy], as well as money supply and expansionary policy. Topic 134 is concerned with shocks, with inflation being a prominent term. Topic 142 is the English language equivalent to Topic 119 (monetary policy). Figure 4.6 shows further topics associated with international trade. The German equivalent (topic 36) to the topic we selected (Topic 1) is centered around "ausland" and "inland" [en: foreign and domestic] and not as narrow as the english original. Topic 44 is loosely concerned with trade, with terms "handelspoliti" [en: trade policy] and "aussenhandelstheori" [en: theory of international trade] popping into the eye. Price differentiation [ger: preisdifferenzier], product [ger: erzeugnis] as well as terms relating to foreign and domestic are at the center of topic 86. Figure 4.7 and Figure 4.8 show additional topics related to debt and unemployment respectively. Apart from topic 191, which is concerned with interest rates in the narrow sense and consequently used in our analysis, only Topic 120 (Figure 4.9) appears to be somewhat related but talks more about central banking.

In the regression analysis it would be possible to combine two or more topics, which makes the analysis broader. Prior research has shown that this does not improve our results. It can be assumed that narrow topics are best at reflecting narrow economic ideas.

Figure 4.5: Estimated topics related to inflation

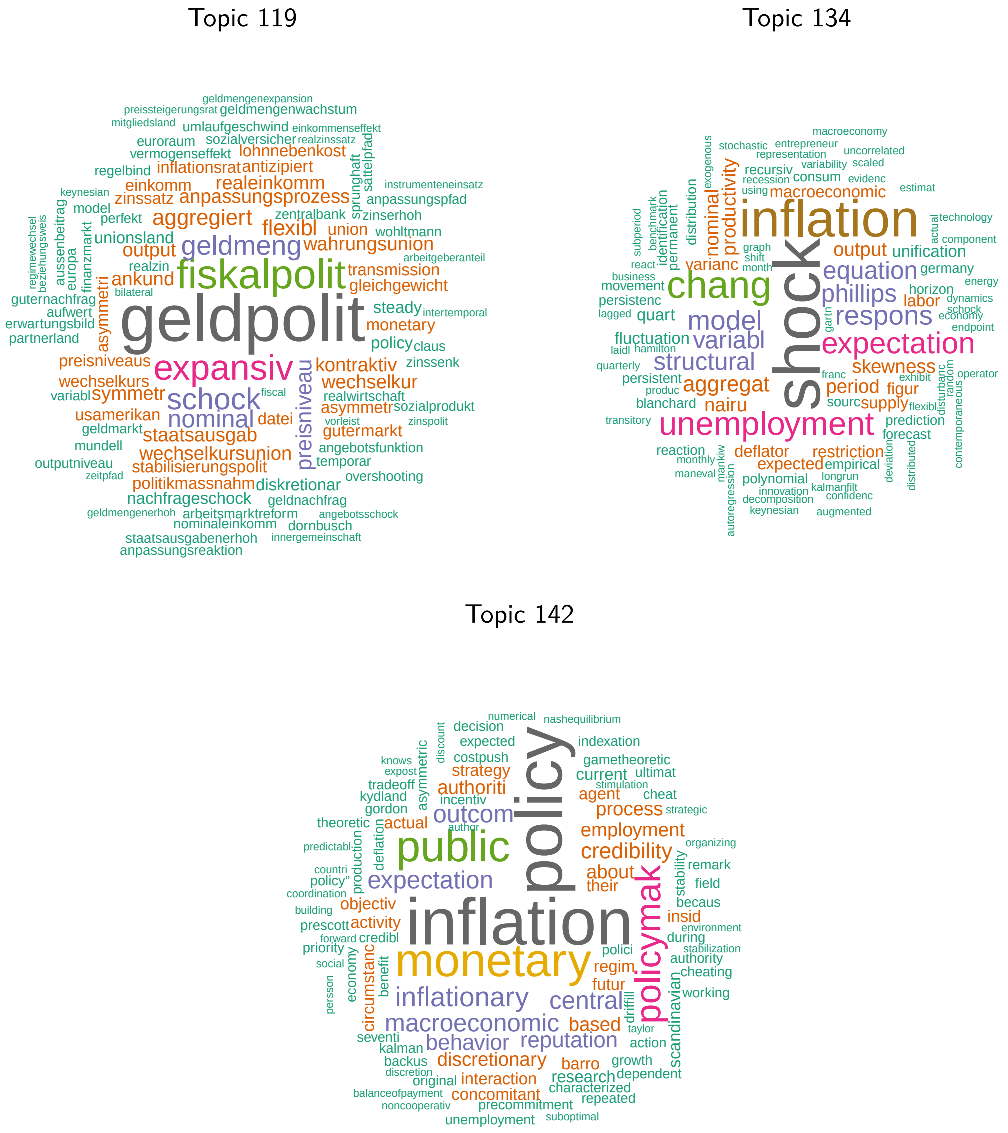
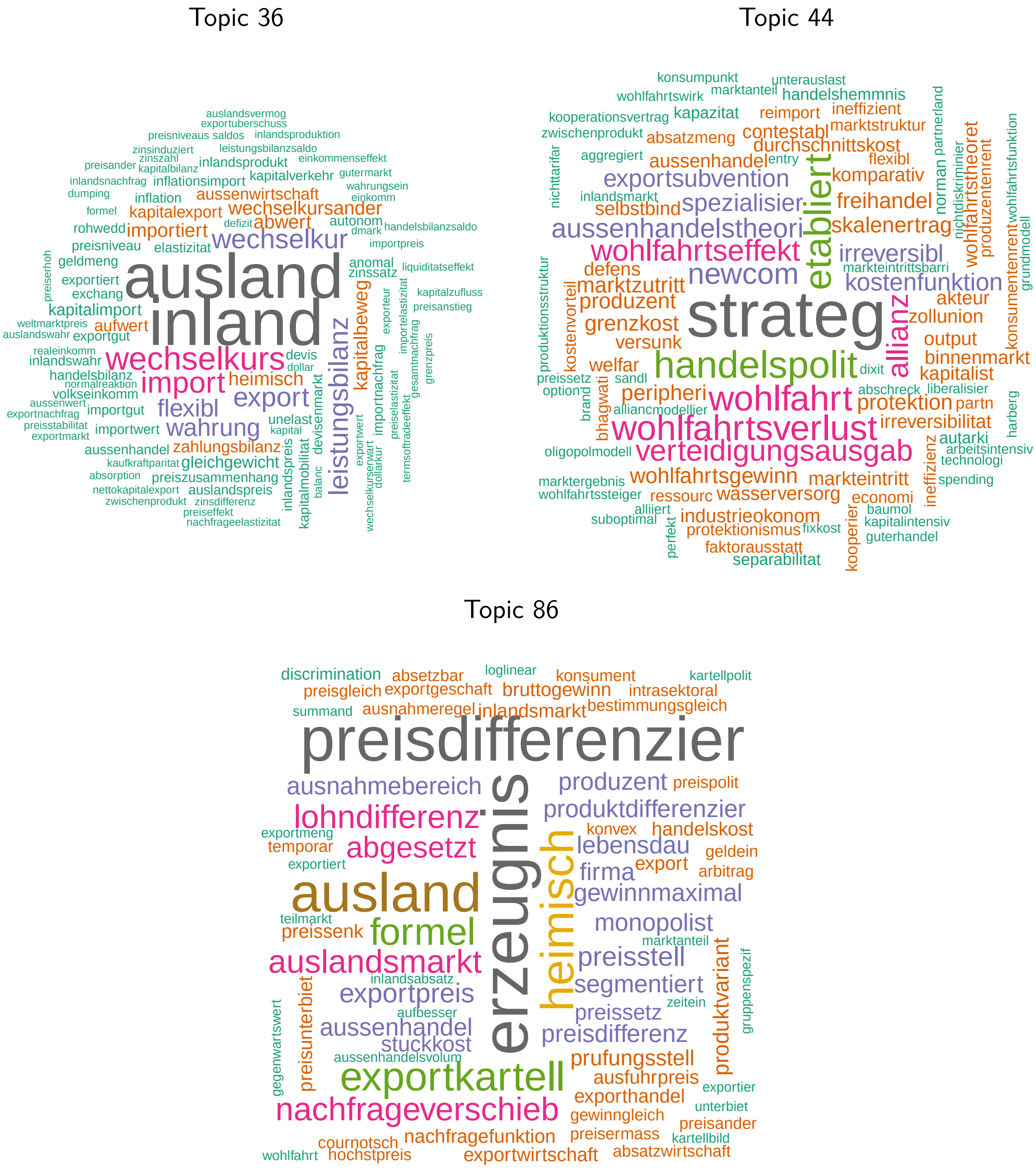


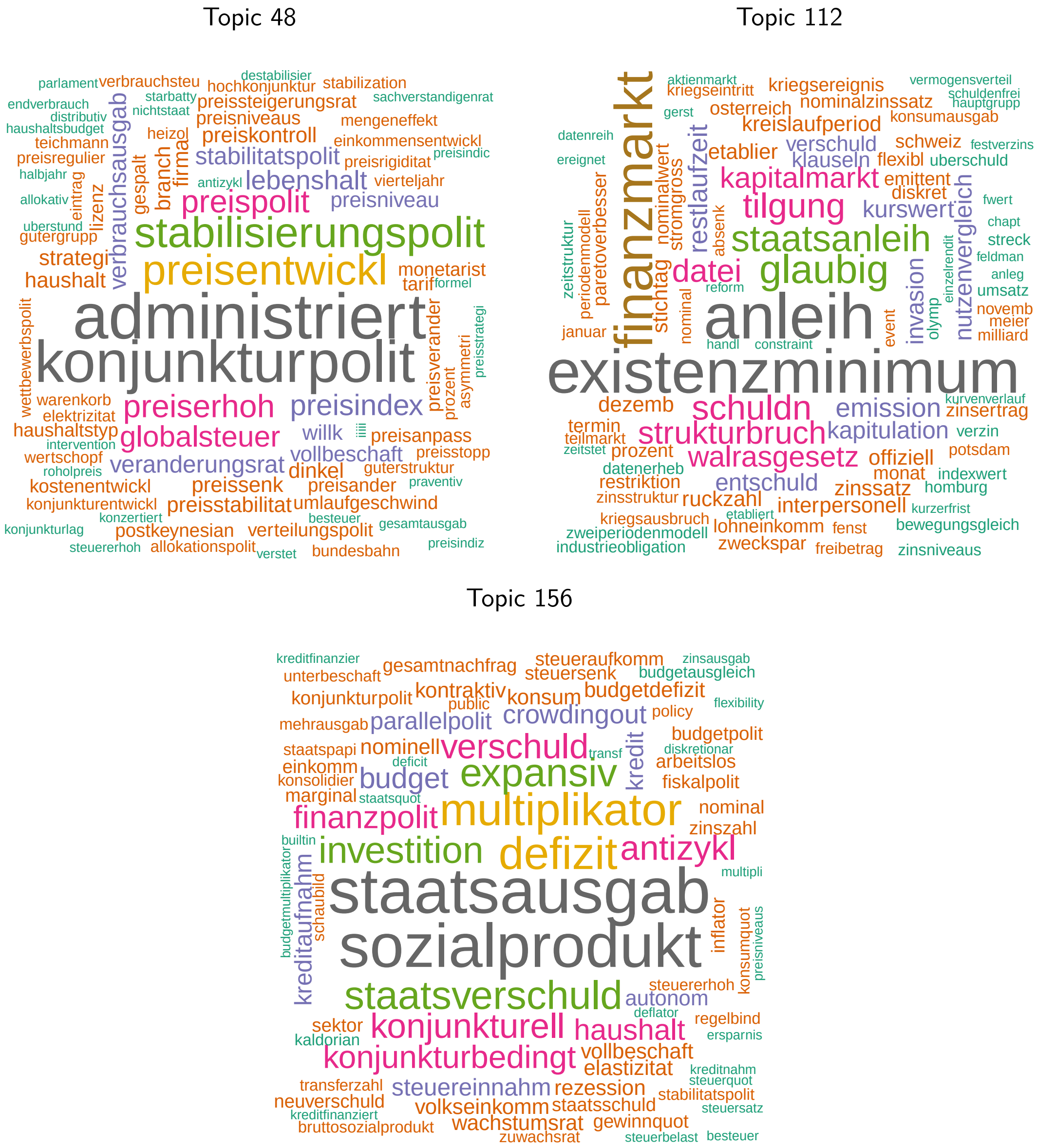
Figure 4.6: Estimated topics related to trade


Topic 36

Topic 44

Topic 86

Figure 4.7: Estimated topics related to debt


Topic 48

Topic 112

Topic 156

Figure 4.8: Estimated topics related to unemployment

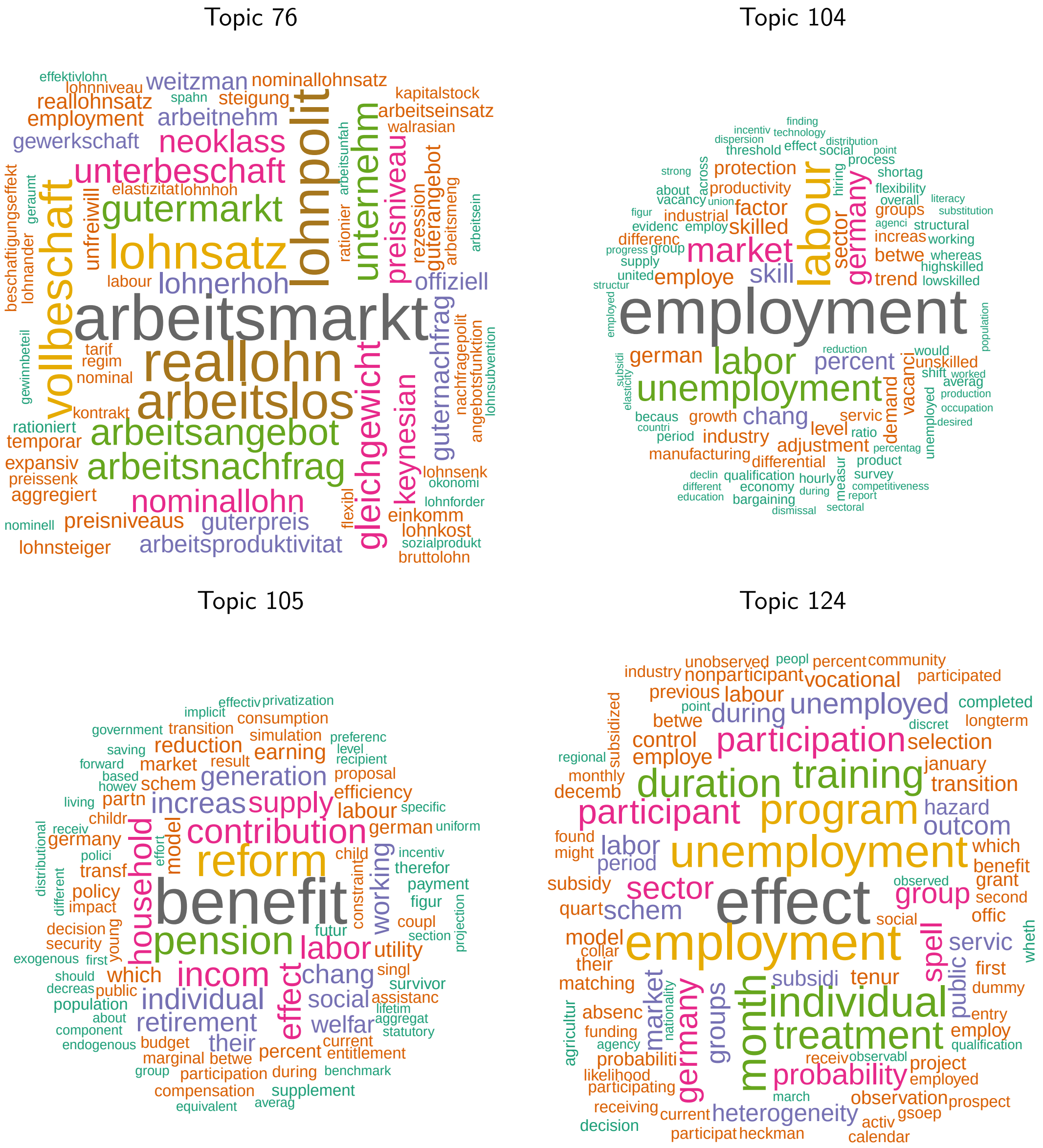


Figure 4.9: Addtional estimated topic related to interest rates

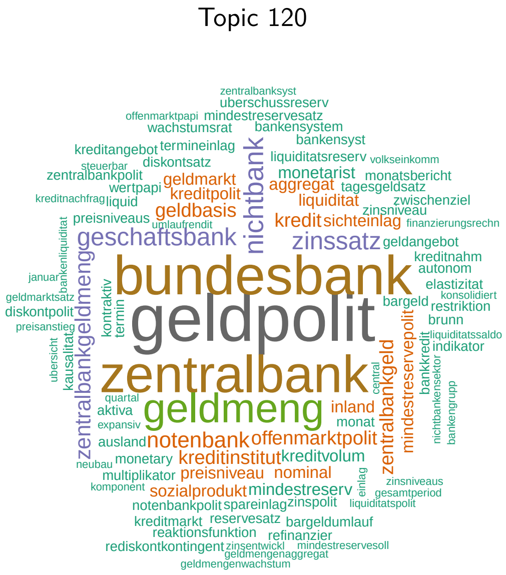
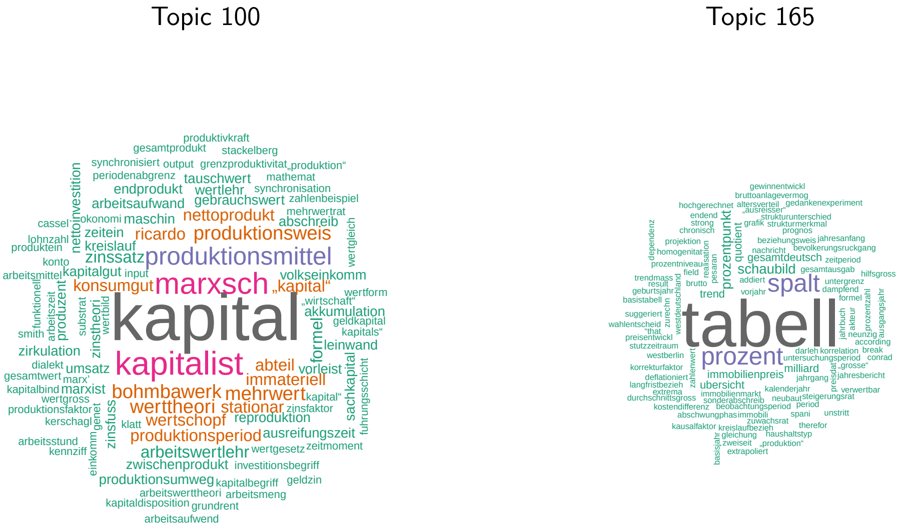

Topic 120

Figure 4.10: Example for “unrelated topics” estimated by the algorithm


Topic 100

Topic 165

5 Monetary Policy on Twitter and its Effect on Asset Prices: Evidence from Computational Text Analysis

JOCHEN LÜDERING[c] & PETER TILLMANN[d]

This article is available as J. Lüdering and P. Tillmann (2016). *Monetary Policy on Twitter and its effect on asset prices: Evidence from computational text analysis*. MAGKS - Joint Discussion Paper Series in Economics 12/2016. Philipps-Universität Marburg.

The paper has been presented by Peter Tillmann at COMPSTAT (Oviedo, 2016), the 48th Money, Macro and Finance Research Group Annual Conference (Bath, 2016) and CMStatistic 9-11 December 2016 in Sevilla.

Acknowledgments We thank Matthias Neuenkirch, Peter Winker and seminar participants at the Bundesbank Big Data group and COMPSTAT (Oviedo, 2016) for very helpful comments. We are grateful to Annette Meinusch, Lucas Hafemann and Florian Viereck for preparing the dataset.

Keywords: Monetary Policy, Fed, Latent Dirichlet Allocation, Text Analysis, VAR

JEL classification: E32, E44, E52

[c] Own contribution 50%

[d] Department of Monetary Economics, Justus-Liebig-Universität Gießen, e-Mail: peter.tillmann@wirtschaft.uni-giessen.de

Abstract

In this paper we dissect the public debate about the future course of monetary policy and trace the effects of selected topics of this discourse on U.S. asset prices. We focus on the "taper tantrum" episode in 2013, a period with large revisions in expectations about future Fed policy. Based on a novel dataset of 90,000 Twitter messages ("tweets") covering the entire debate of Fed tapering on Twitter we use Latent Dirichlet Allocation, a computational text analysis tool to quantify the content of the discussion. Several estimated topic frequencies are then included in a VAR model to estimate the effects of topic shocks on asset prices. We find that the discussion about Fed policy on social media contains price-relevant information. Shocks to the frequencies of "tantrum"-, "QE"- and "evidence"-related topics are shown to lead to significant asset price changes. We also show that the effects are mostly due to changes in the term premium of yields consistent with the portfolio balance channel of unconventional monetary policy.

5.1 Introduction

The formation of monetary policy expectations by market participants is at the core of the monetary policy transmission process. While there is a large body of research on how central banks communicate with financial markets (Blinder et al. 2008), there is little evidence about how, given the communication of the central bank, the discourse about monetary policy by market participants shapes market expectations. This lack of research is most likely due to the lack of data about individual views on future policy.

In this paper we study the changing policy expectations of market participants and the resulting change in forward looking financial variables such as asset prices. We focus on an episode in recent U.S. monetary policy which has been characterized by a major shift in market expectations: the "taper tantrum" period in 2013. After Fed chairman Ben Bernanke mentioned an eventually exit from the Fed's asset purchase programs in May 2013, markets changed their assessment of the future course of policy resulting in a phase of unusual volatility, an increase in long-term U.S. interest rates and an appreciation of the U.S. dollar. Markets quickly coined the term "tapering" to describe the Fed's exit from QE3 (3rd round of quantitative easing). Given the exaggerated market reaction, this period is referred to as the "taper tantrum".

We analyze this episode based on a dataset that contains the entire traffic on Twitter.com, the social media network, on Fed tapering. The dataset consists of 90,000 text messages ("tweets") between April and October 2013 and reflects the debate among market professionals during the tantrum period. Twitter data offers several advantages over alternative datasets: First, in contrast to news articles or analyst reports, which are written and read be relatively few people, Twitter allows us to exploit the views of the crowd of financial professionals. Second, while it is unclear whether news reports are actually read, Twitter messages appear as push messages on mobile phones and are actively shared, discussed, endorsed or refuted. Hence, the tweets give more reliable evidence about individuals' views than the consumption of news reports. Third, the high frequency of observation allows us to trace the public debate in real time.

Since our aim is to model the changing beliefs about future monetary policy, we need to quantify the information content of the Twitter data.[1] In this paper we employ a tool set taken from computational linguistics. Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) is used to extract latent topics out of the Twitter conversation on tapering. LDA dissects a text document into different topics based on Markov Chain Monte Carlo estimation. It estimates a certain number of different topics which, based on each topic's most frequent words, will be labeled manually.

Examples of topics are "tight financial conditions", "data", "fear", "stimulus" and many more. The resulting topic frequencies, which express the likelihood that a given tweet contains this specific topic, are then included in a vector autoregression (VAR) together with a daily series of macroeconomic fundamentals and asset prices. We show that a shock to selected topics frequencies leads to a significant change in asset prices. This finding supports the notion that the public debate about future monetary policy, which is reflected by the discourse on Twitter.com, contains information that is relevant for pricing financial assets. A shock to the likelihood of the topic describing "premature tightening", for example, raises bond yields and leads to an appreciation of the dollar.

This paper is closely related to the recent literature on text mining and computational text analysis, respectively, for financial and monetary policy applications. This rapidly growing literature is summarized by Loughran and McDonald (2016).[2] For the purpose of this paper, the existing research applying models of latent information in textual data is particularly interesting.

[1] In a companion paper Meinusch and Tillmann (2016), we construct a dictionary with keywords describing a certain policy path. The drawback of this approach is that we have to specify a list of keywords in advance, which is likely to disregard important information from expressions not on our list of keywords. Furthermore, we can focus on two alternative policy paths only, an early or a later tapering decision, and have to leave out other dimensions of the discussion.

[2] Bholat et al. (2015) provide a survey of text mining applications relevant for central banks.

Hansen and McMahon (2016) apply LDA modeling to the entire history of policy statements issued by the Federal Open Market Committee (FOMC) of the Federal Reserve. Thus, they are able to identify at which point in time the FOMC spent time discussing a specific topic. Selected topic frequencies are included in a Factor-Augmented VAR (FAVAR) model. The authors find that forward guidance related topics and topics reflecting the current economic situation affect real and financial variables.[3] While our modeling framework is similar, we do not study central bank communication but the discourse of the market about what the Fed is likely to do.[4] Hendry and Madeley (2010) and Hendry (2012) use latent semantic analysis for the communication of the Bank of Canada. They identify "themes" of communication, which are used to explain interest rate changes. Although they do not address a financial application but instead focus on fluctuations in the business cycle, the paper by Larsen and Thorsrud (2015) is also relevant for our work. They use a dataset with Norwegian newspaper articles to construct an aggregate news index by employing topic modeling. The news indices are related to economic activity within a Bayesian regression framework. Using a similar methodology but covering a time horizon of 61 years Lüdering and Winker (2016) examine whether economist anticipate changes in the state of the economy or merely discuss these events *ex-post* by looking at the relationship between between economic publications and real-world time-series. The LDA model identifies key topics for "debt", "trade", "Inflation", "Unemployment" and "interest rate". A subsequent regression analysis finds a significant link between the economic discussion and the time-series, only in two out of five cases (debt and unemployment) the economic discussion precedes the economic developments, while for the other time-series the discussion is following the economic developments.

We show that topics related to the "tantrum" notion of the tapering discussion, the implications of "QE" and the debate about incoming "evidence" affect bond yields and

[3] Using the same dataset, Hansen, McMahon, and Prat (2014) model the effect of increased transparency on the policy debate in FOMC meetings.

[4] Lucca and Trebbi (2009) and Schonhardt-Bailey (2013) are other recent papers on textual analysis of FOMC communication.

exchange rates. While the effect of "tantrum"-related topics is long lasting, the impact of the other topics is short-lived. This suggests "tantrum"-related topics indeed contain information about the policy *path*, while the other topics contain mostly noisy information in the sense of Tetlock (2007).

We also decompose yields into the expectations component and the term premium. Based on this decomposition we show that the response of bond yields is mostly due to responses of term premia. This findings lends support to the balance sheet channel as the transmission channel of shifting tapering expectations.

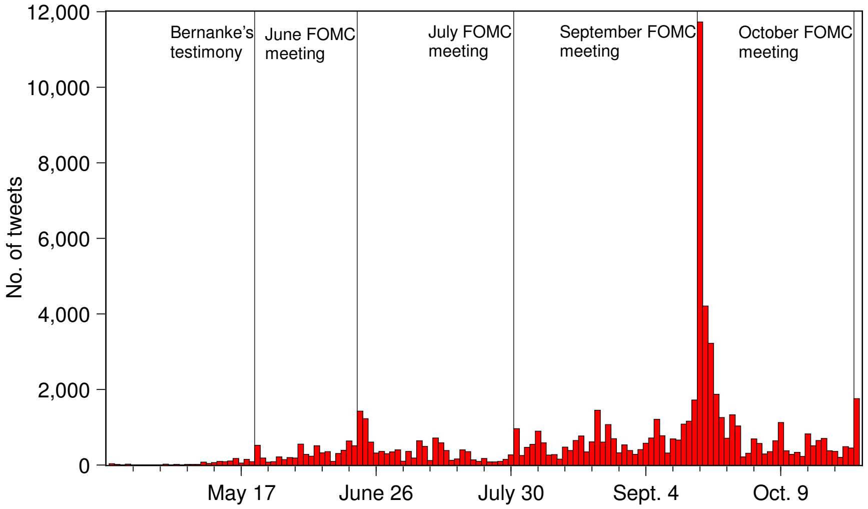
This paper is organized as follows: In Section 2 we introduce the Twitter dataset. Section 3 gives some background on computational text analysis and presents the LDA approach used in this paper. The estimation of a VAR model that includes asset prices and selected topics frequencies is described in Section 4. The results and some robustness checks are discussed in Section 5 and Section 6 draws conclusions.

5.2 The Dataset

The dataset used in this study consists of all Twitter messages containing the words "Fed" and "taper" sent between April 15 and October 30, 2013. The data has been purchased from `Gnip.com`. Because it is highly likely that any tweet on the Fed's exit from QE contains both filter words, we are certain to have a comprehensive dataset that reflects the entire tapering debate on Twitter. After deleting a few tweets in languages other than English, we are left with 87024 tweets. Re-tweets, i.e. Twitter messages forwarded by users, are left in the dataset because a forwarded tweet is likely to be a relevant tweet and, as a result, the forwarding of a tweet also contains information.

For each tweet we know the content, the sender and the time the tweet was sent. We normalize the timing of each tweet to New York time. While trading hours end at 4 pm Eastern Time, twittering continues even after markets have closed. To account for tweets sent after markets closed, tweets sent after 4 pm are attributed to the next trading day. Likewise, weekends and holidays have been excluded due to the lack of asset price data. Finally, tweets are aggregated to daily frequency.

Figure 5.1: Number of Tweets



Notes: The vertical lines indicate the testimony of chairman Bernanke on May 22 and the subsequent FOMC meetings.

Figure (5.1) plots the daily number of tweets and, as vertical lines, the most important monetary policy events. Tweeting on the Fed's tapering decision gradually picks up before the testimony of chairman Bernanke on May 22. During this testimony, Bernanke mentioned the possibility of exiting from QE3, a statement that triggered the markets' subsequent tantrum reaction. A first peak is reached prior to the June meeting of the Federal Open Market Committee (FOMC), for which some market participants expected more detailed information about the pace of the tapering. After the FOMC postponed the tapering decision, the discourse among market participants intensified before each subsequent FOMC meeting. The peak was reached before the September FOMC meeting, for which the vast majority of Twitter users expected the decision about a reduction of monthly purchases of securities. However, market again misjudged the Fed as the FOMC again postponed the decision. The Fed eventually announced a reduction in its monthly asset purchases in the January 2014 FOMC meeting.

From Figure (5.1) we see that the number of tweets is systematically higher on FOMC meeting days.[5] This is not surprising given the market's interest in monetary policy decisions. For our empirical analysis below this implies that we should control for FOMC meeting days and, in addition, for days on which FOMC minutes are published.

Since the monetary policy debate on Twitter captures the overall discussion among market participants well, we will now use topic models to dissect the discussion into policy-relevant topics. We want to see which topic was most relevant on selected days and, in particular, how the information contained in these topics is reflected by asset prices.

The "taper tantrum" it subject to a relatively small empirical literature. Most studies, such as Eichengreen and Gupta (2015), Aizenman, Binici, and Hutchison (2016) and Mishra et al. (2014) ask whether the macroeconomic vulnerability of emerging market countries determines how strongly these countries were hit in 2013. However, these authors typically do not quantify market expectations and their revision directly. Rather, they argue that the changes in U.S. asset prices in 2013 are appropriate indicators of shifts in expectations. Meinusch and Tillmann (2016) use the same Twitter dataset that is used in this paper. Based on a dictionary of words they built proxy variables for the beliefs of an early and a late tapering, respectively. Here we extend and broaden this line of research: we use several dimensions of the debate on Twitter, not just early or late tapering, and relate them to asset prices.

[5] Due to the convention of the software used, the vertical lines in Figure (5.1) are drawn at the beginning of each day, while the bars are plotted from the beginning to the end of a given day.

5.3 Applying Topics Models

In this section we apply topics models to dissect the discussion on Twitter regarding the unwinding of QE in its most important parts, which we then relate to asset pricing. The recent introduction of topic modeling into the field of economics enables researchers to automatically classify texts and obtain underlying topics which constitute the document, given the assumed generative process and several predetermined parameters. It should be noted that these topics are not necessarily coherent topics in the semantic sense, but rather clusters of *terms* which repeatedly appear together over several documents. Similar to clustering methods, it is up to the researcher to make sense of the topics based on the words of which they are comprised.

For the application, we follow Grün and Hornik (2011) in using the *R* package *tm*[6] for pre-processing the data and subsequently *topicmodels*[7] for the fitting of the topic models.

5.3.1 Preliminary Steps

Starting out with the corpus of Twitter messages described in the previous section, the first step consists of cleaning the data to obtain the vocabulary V. The vocabulary is the set of different terms selected from the corpus, our entire set of tweets, because the terms are well suited to explain the content of individual documents. It is the goal to omit all terms that are not helpful in differentiating between topics.

As topic models are only concerned with the joint appearance of words in individual documents and are not influenced by grammar, we can remove all punctuation and redundant space characters from the tweets. A specificity of twitter messages is the appearance of hyperlinks and twitter usernames (i.e. *@username*), which we remove in their entirety. Further, all words are decapitalized and the stemming algorithm *SnowballC* is applied to create word stems. A stem is the part of a word which is common to a variety of grammatical forms. These two measures, stemming and decapitalization, lump together different grammatical

[6] See `https://cran.r-project.org/web/packages/tm/`.

[7] See `https://cran.r-project.org/web/packages/topicmodels/`.

forms and remove the differentiation due to a word being capitalized at the beginning of a sentence.

Words which are frequent but add little meaning to a document are called *stopwords* and are removed from the corpus and thus excluded from the vocabulary. We use the list of English language stopwords provided by the R package *tm*. In addition, the terms "Fed" and "taper" are removed from the corpus as these words have been used to select the dataset in the first place. Hence, they should be included in each tweet. Finally, all terms with a length between 4 and 20 characters appearing in at least five tweets are used to create the document-term matrix $\boldsymbol{F}$, which holds the frequencies $f_{i,j}$ of $|V| = 5082$ different terms in $D = 87024$ tweets and is the basis for the subsequent LDA estimation.

In applications of topic models, it is common practice to further reduce the vocabulary by selecting only terms which are important to describe individual documents. This is usually done based on the *tf-idf* (term frequency–inverse document frequency) value (Blei and Lafferty, 2009), which implies weighting the frequency in a single document against the overall frequency. The particularities of Twitter messages, e.g. the choice of words and the length of tweets limited to 140 characters, results in particularly high tf-idf values (median > 1).[8] As the importance of the individual words for explaining the different documents appears to be particularly high and the vocabulary is rather short (5082), in particular with respect to the number of documents in the corpus (87024), we do not remove any further terms.

5.3.2 Latent Dirichlet Allocation (LDA)

This section provides a brief overview of LDA models and the estimation method behind our analysis. Before their recent arrival in economics, topic models have been used since the 1990s (Deerwester et al. 1990) to address issues in the area of information retrieval. Hofmann (1999) introduced probabilistic theory to topic models, providing a sound statistical background. His

[8] Other datasets, using scientific articles, result in median *tf-idf* values of 0.004 (Lüdering and Winker 2016) and 0.1 (Grün and Hornik 2011).

approach (probabilistic Latent Semantic Analysis) has later been extended to Latent Dirichlet Allocation (LDA) by Blei, Ng, and Jordan (2003). Although LDA has subsequently been refined, e.g. time varying topics have been suggested by Wang and McCallum (2006) and the model has been extended to allow for topic correlation by Blei and Lafferty (2007), their underlying theoretical model remains the state of art in topic modeling up to today.

In LDA the creation of documents is described by an abstract generative process. It is assumed that all documents in a corpus are generated from a fixed set of K different topics. The topics consist of a set of terms w from a vocabulary V. Each term w in a document $\boldsymbol{w}$ is generated by first drawing a topic given a vector of topic probabilities $\theta_{\boldsymbol{w}}$ and afterwards drawing a term, given its probability β_k in a topic. Hence, the probability of word w_i is given by

$$P(w_i) = \sum_{j=1}^{K} P(w_i|k_j)P(k_j). \tag{5.1}$$

In order to estimate the matrices of predicted probabilities, $\hat{\theta} = K \times D$ and $\hat{\beta} = K \times |V|$, our variables of interest, an algorithm is used to reverse the generative process. Due to the complexity of the model, standard maximum likelihood procedures do not proof suitable. Hence, a number of sophisticated methods have been developed to estimate the model nonetheless. In modern applications of LDA, the original estimation algorithm (variational expectation maximization, VEM) has largely been replaced by Gibbs sampling, a Markov Chain Monte Carlo approach suggested by Griffiths and Steyvers (2004). Instead of estimating the topic distribution θ and the term distribution β directly, the distribution of $\boldsymbol{z}_i$, the assignments of words to topics is estimated. Based on these assignments of words to topics, approximations of θ and β can be computed. In order to perform the computation, it is necessary to make the simplifying assumptions that θ and β are random draws from the Dirichlet distributions Dir(α) and, respectively, Dir(δ). The parameters on the Dirichlet distributions are chosen according to the literature (Griffiths and Steyvers 2004) and set to $\alpha = 50/K$ and $\delta = 0.1$.

In applied work, the choice of the optimal number of topics K remains an important issue. In order to obtain an estimate for K, we apply the harmonic mean method as suggested by Griffiths and Steyvers (2004).[9] It consists of taking a number of samples as estimates for $P(\boldsymbol{w}|K)$ from the Markov Chain, and the subsequent computation of the harmonic mean across the values. The resulting function of the relationship between K and $P(\boldsymbol{w}|K)$ is not smooth. Thus, simple maximization does not necessarily lead to useful results.[10] However, we end up with an unreasonably large number of topics given our data. As a consequence the obtained topics are very narrow and difficult to interpret. Hence, we follow the pragmatic approach by Hansen, McMahon, and Prat (2014) of choosing a lower value in order to produce topics which are more appealing to human judgment. By setting $K = 30$, we take into account that Twitter messages contain 140 characters as a maximum and the tweets in our sample have already been pre-selected to cover a specific area, the tapering decision of the Fed.

Following Griffiths and Steyvers (2004) the Markov Chain is constructed to converge to the "true" distribution of $\boldsymbol{z}_i$, which is the vector of assignments of words to topics. After 2000 iterations of sequential updating, from which the first 100 are discarded, the Markov Chain is assumed to have converged. The approximation for document probabilities θ is calculated based on the number of times document $\boldsymbol{w}$ has been associated with topic k, measured by the count variable $\eta_k^{\boldsymbol{w}}$, relative to the sum of all associations of document $\boldsymbol{w}$ to any topic:

$$\hat{\theta}_k^{\boldsymbol{w}} = \frac{\eta_k^{\boldsymbol{w}} + \alpha}{\sum\limits_k \eta_k^{\boldsymbol{w}} + K\alpha} \tag{5.2}$$

Analogous the vector of topic probabilities β is approximated from the number of times

[9] For an alternative procedure that determines the number of topics as a result of a trade-off between the salience of topics and the load on an given topic see Goldsmith-Pinkham, Hirtle, and Lucca (2016).

[10] Griffiths and Steyvers (2004) circumvent this issue evaluating K at large steps.

that term w_i has been associated with topic k as indicated by count variable $\eta_k^{w_i}$, relative to the sum of all associations of any word to topic k.

$$\hat{\beta}_k^{w_i} = \frac{\eta_k^{w_i} + \delta}{\sum\limits_{w_i} \eta_k^{w_i} + |V|\delta} \tag{5.3}$$

The resulting matrix of term probabilities $\hat{\beta}$ reveals the contents of the 30 different topics, which reflect the discussion on Twitter. Table (5.1) lists the five most frequent words of each topic.

Table 5.1: Content of topics

Topic	Most frequent words	Topic	Most frequent words
1	will, next, announc, yellen, increas	16	meet, FOMC, minut, call, close
2	gold, price, drop, concern, future	17	high, sampp, wont, record, asia
3	market, worri, debt, want, syria	18	dollar, forex, specul, trade, boost
4	market, good, emerg, forecast, talk	19	bernank, bank, might, chairman, back
5	market, week, ahead, tantrum, caus	20	data, soon, time, economi, depend
6	economi, reason, weak, dudley, strong	21	bullard, need, octob, inflat, possibl
7	septemb, lockhart, begin, expect, goldman	22	decis, surpris, market, post, notap
8	economist, street, expect, wall, time	23	year, decemb, evan, still, like
9	like, delay, make, policy, move	24	start, today, sept, just, june
10	risk, look, take, paush, statement	25	feder, reuter, reserv, office, gain
11	bond, keep, point, refrain, plan	26	think, wont, will, back, cant
12	view, decis, federalreser, william, georg	27	month, treasuri, purchas, yield, asset
13	talk, fear, share, fall, data	28	stock, talk, investor, market, amid
14	rate, rise, mortgag,time, interest,long	29	new, delay, busi, washington, money
15	stimulus, readi, world, shock, program	30	will, report, gross, break, bubbl

5.4 The Empirical Model

In this section we relate the discussion on Twitter as reflected in a selected number of identified topics on U.S. asset prices. For that purpose we use a VAR model in which we include not only information from the topics but also asset prices and macroeconomic conditions. Because we want to model a parsimonious VAR system, we cannot include all 30 topics in the VAR model jointly, which would leave too few degrees of freedom for estimation. Instead, we focus on selected topics only.

5.4.1 Selected Topics

In particular, we identify three sets of topics, which are directed towards particularly important aspects of the assessment of future monetary policy and the tapering decision, receptively. Each set includes two topics, $Topic_t^i$ and $Topic_t^j$. The following table summarizes the topics we use in the estimation:

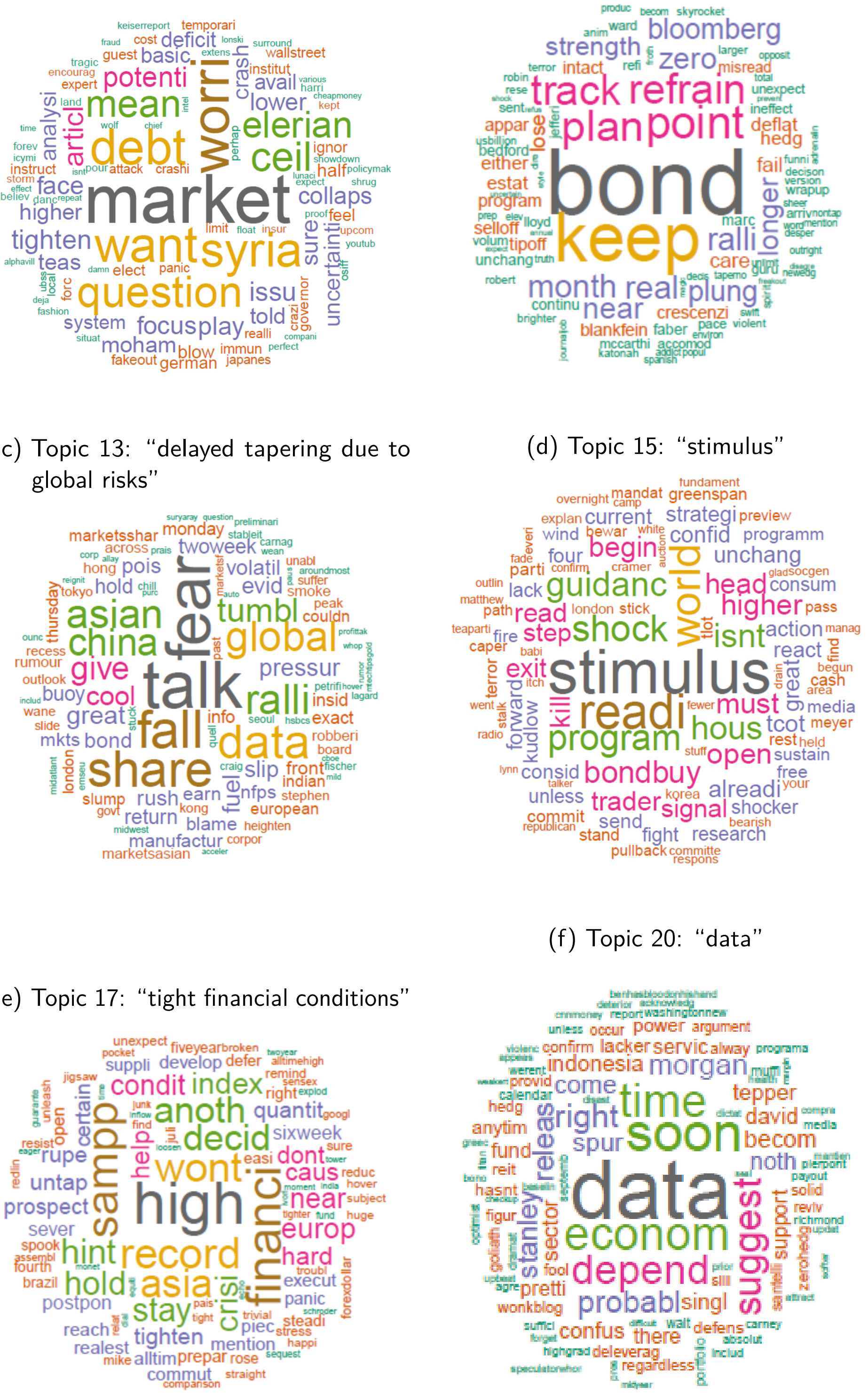
Table 5.2: Selected sets of topics

	set	$Topic_t^i$	$Topic_t^j$
1	"tantrum"	T3 "premature tightening"	T13 "delayed tapering due to global risks"
2	"QE"	T11 "bond market"	T15 "stimulus"
3	"evidence"	T17 "tight financial conditions"	T20 "data"

Figure (5.2) visualizes each of the selected topics by a word cloud, where the size of each word in the cloud reflects the significance (probability) of the word for a given topic. While the word list in Table (5.1) gives a broad assessment of each topic, the word clouds help interpreting the content of each topic in light of the debate about monetary policy.

The first topic set captures the notion of the market "tantrum" occurring in 2013. Topic 3 contains the discussion of a possible tightening that is considered premature in light of the ongoing U.S. debt crisis and the conflict in Syria, see panel (a) of Figure (5.2). Therefore, we label topic 3 "premature tightening". Topic 13, in contrast, see panel (c) of the figure, reflects the discussion of a tapering decision that might be delayed due to the risks in Asian

Figure 5.2: Selected topics



(a) Topic 3: "premature tightening'

(b) Topic 11: "bond market"

(c) Topic 13: "delayed tapering due to global risks"

(d) Topic 15: "stimulus"

(e) Topic 17: "tight financial conditions"

(f) Topic 20: "data"

financial markets and other global considerations. We refer to this topic as "delayed tapering due to global risks".

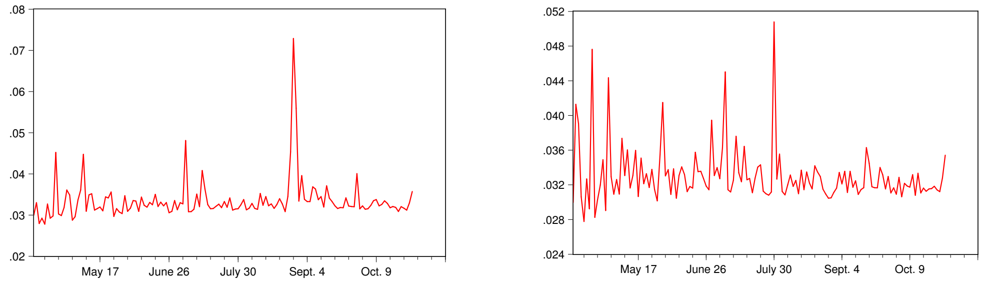
The second set reflects the debate about the U.S. bond market. Topic 11, see panel (b) in Figure (5.2) lends itself to an interpretation in terms of the state of the "bond market". Likewise, topic (15) is about a general monetary "stimulus", see panel (d).

The third set includes the debate among Twitter users about evidence-based monetary policy decisions. Topic (17), see panel (e), highlights "tight financial conditions". Finally, topic (20) contains the discussion of Twitter users about incoming economic data as input for monetary policy decisions. We label this topic "data".

We are agnostic with regard to the signs of the effects shocks to these topics have on U.S. asset prices and rely on the VAR model introduced below to highlight the market impact of each topic.

The importance of the topic over time can be assessed by the matrix of topic probabilities across tweets. In order to make meaningful interpretation possible we need to account for the fact that the number of tweets per day is not constant across the observation period. Thus, for each topic we calculate the mean topic probability over all tweets in a single day. These aggregate values can easily be compared over the course of the whole sample period. Figure (5.3) shows the frequency for each topics.

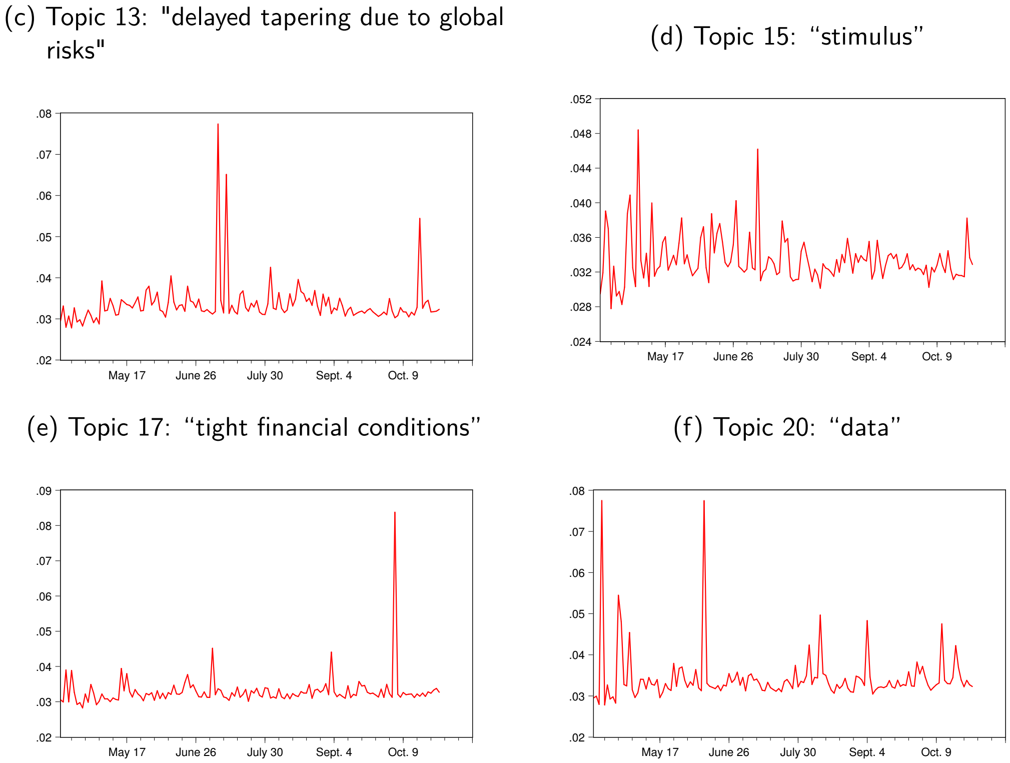
Figure 5.3: Selected topic frequencies



(a) Topic 3: "premature tightening'

(b) Topic 11: "bond market"

Figure 5.3: Selected topic frequencies (cont.)


(c) Topic 13: "delayed tapering due to global risks"

(d) Topic 15: “stimulus”

(e) Topic 17: “tight financial conditions”

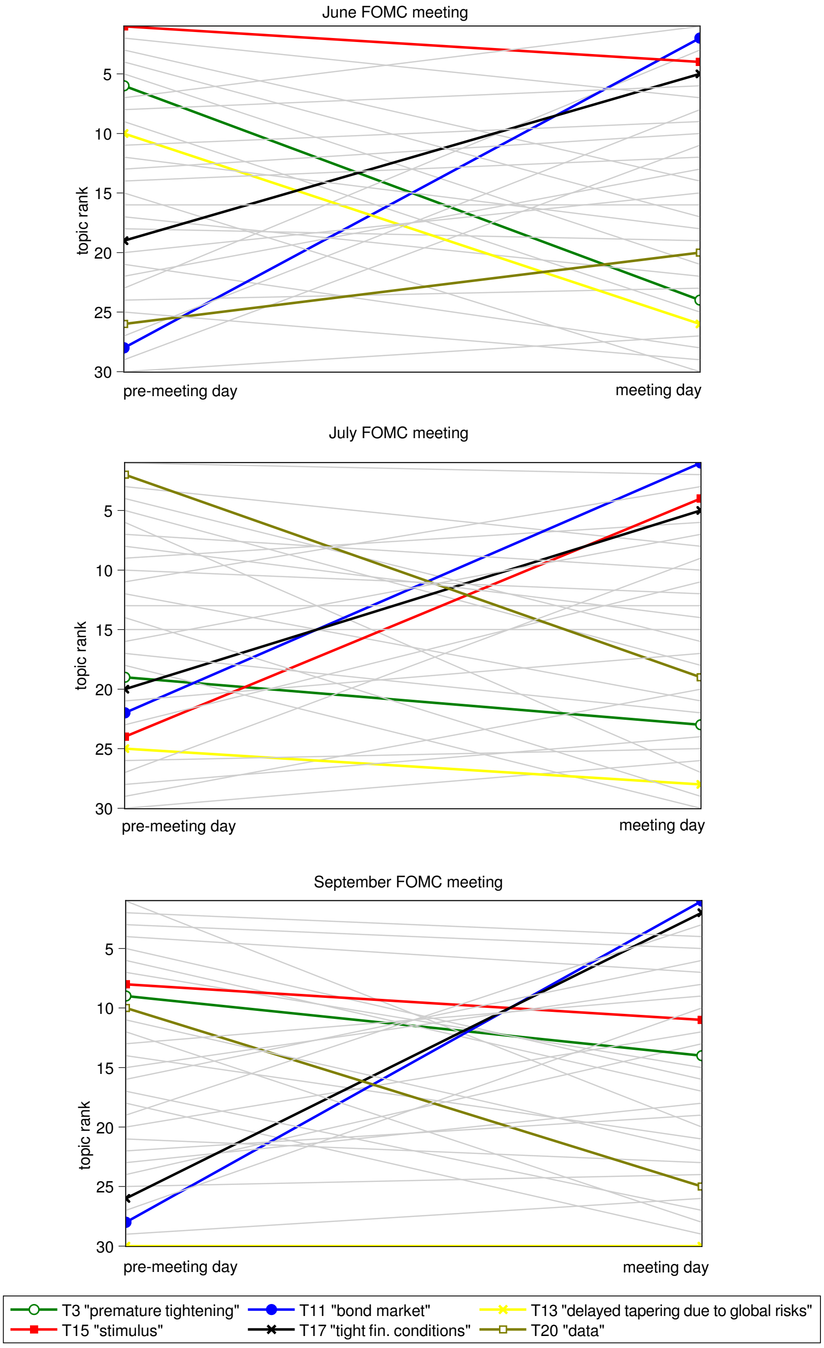
(f) Topic 20: “data”

5.4.2 The Evolution of Topics around FOMC Meetings

To shed light on the behavior of topic frequencies, we focus on the three most important FOMC meeting days in our sample and order each topic according to its frequency on pre-meeting and meeting days, respectively. Figure (5.4) shows the rank of each selected topic on both days. Thus, we can see how the relevance of each topic changes when new information from the meeting outcome becomes available.

It can bee seen that on all three meeting days the discussion on tight financial conditions and the bond market gains importance. This is reflecting the fact that immediately after the press release market commentators discuss the meaning of the meeting outcome for financial markets. Worries on a premature tightening, in contrast, fall on meeting days compared to pre-meeting days. This is intuitive as on each meeting in our sample the eventual tapering decision has been delayed.

Figure 5.4: Evolution of topics on FOMC meeting days



Notes: We order each topic according to its frequency on the pre-meeting day and the meeting day, respectively. The graph presents the rank of each topic on both days. The selected topics used for the empirical analysis are highlighted by colors, all other topics are shown in gray.

5.4.3 The VAR Model

In the empirical model each selected topic is scaled by the sum of the remaining topics. The reason for this is that all topics must add to one. An increase in one topic frequency must be associated with a decrease in the sum of all other topic frequencies. To gauge the relative importance of a topic, we have to normalize it by the importance of the remaining topics. Thus, each topic $Topic_t^n$ with $n \in (i, j)$ is included as the odds ratio, i.e.[11]

$$\overline{Topic}_t^n = \frac{Topic_t^n}{1 - Topic_t^n} \text{ for } n = i, j \text{ and } t = 1, ..., T. \tag{5.4}$$

We estimate a VAR model in order to derive the dynamic responses of asset price changes to the two summary indicators. The reduced-form representation of the VAR is

$$Y_t = A_0 + A(L)Y_t + D_t^{FOMC} + u_t \text{ for } E[u_t u_t'] = \Sigma_u, \tag{5.5}$$

where $A(L)$ reflects the matrix polynomial in the lag operator of order p, Y_t is a vector of endogenous variables and u_t constitutes a white noise process with variance-covariance matrix Σ_u. We also add a constant, A_0, to the model. To control for monetary policy events, we include separate dummies contained in D_t^{FOMC} for Bernanke's testimony on May 22, each FOMC meeting and each release date of FOMC minutes in our sample period. Most likely many tweets comment on FOMC meeting outcomes or releases of FOMC minutes. We control for these days in order to see whether tweets contain information even in the inter-meeting or inter-release day period, respectively. It is important to allow for each FOMC event to be reflected in separate dummies in order to prevent that FOMC meetings with, say, tightening and easing policy steps cancel out. In our discussion of the results below, we show impulse responses derived from VAR models with and without these FOMC-related dummies.

[11] Using the log of the odds ratio results is almost identical results.

The vector Y_t contains the following variables

$$Y_t = \left(Y_t^{CESI} \quad \overline{Topic}_t^i \quad \overline{Topic}_t^j \quad A_t^{US} \right)'. \tag{5.6}$$

In this model, Y_t^{CESI} is the daily Citigroup Economic Surprise Index (CESI) for the U.S. economy. The CESI is defined as the weighted historical standard deviations of data surprises, that is, actual data releases minus the median survey expectations from the Bloomberg survey. Thus, the index captures positive and negative surprise realizations of macroeconomic data releases. It is important to control for data releases in our empirical model as market expectations reflected in Twitter messages also reflect macroeconomic news. The weights of economic indicators are derived from relative high-frequency impact on spot exchange rates.

Asset prices in the U.S. are reflected by A_t^{US}. We use two alternative asset prices: the 10-year yield and the nominal USD-EUR exchange rate. The latter two asset prices are taken from the St. Louis Fed's FRED database. The yield series are fitted yields from the Adrian et al. (2013) term structure model. Note that an increase in the USD-EUR exchange rate implies a depreciation of the dollar. The estimated VAR model includes four lags of the endogenous variables and is estimated on daily data for $T = 139$.

Estimating a VAR model in order to derive impulse response functions necessitates the identification of structural shocks from the estimated reduced form residuals. Here we impose a Cholesky identification on our VAR system which implies that on a given day each variable affects only those variables ordered behind it in the VAR model. The ordering of the variables corresponds to the order given in the description of the model, i.e. $Y_t^{CESI}, \overline{Topic}_t^i, \overline{Topic}_t^j, A_t^{US}$. This implies that a change in $\overline{Topic}_t^i$, for example, has an effect on asset prices on the same day, while a change in asset prices can affect the Twitter topics only on the following day. We believe it is important in our context to allow Twitter messages to have a contemporaneous effect on asset prices. The attractiveness from Twitter as a medium of exchange stems from the speed by which users can respond to news and interact with others. Any information contained in the Twitter exchange should be allowed to move markets instantaneously.

At the same time, we accept the restriction that users cannot contemporaneously respond to asset price developments on a given day. In fact, this restrictions helps us controlling our results for those tweets that simply comments on ups and downs in asset markets and focusing on tweets that contain views about future policy.[12] In the robustness section presented below, we will also present results based on alternative identification schemes.

5.5 Results

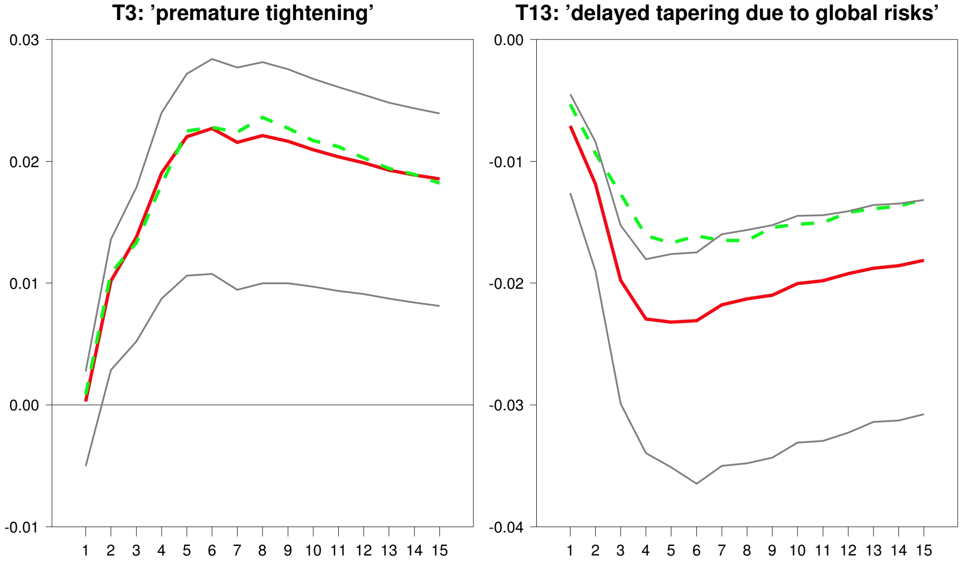
In the following subsections, we present the resulting impulse response functions describing the responses to topic shocks. In each impulse response graph, we show bootstrapped confidence bands reflecting the 16th and the 84th percentiles of the draws. We also show the responses from a model in which we include dummies for major FOMC events (red line) and a model without FOMC dummies (green, dotted line). We focus on the responses of asset prices, which are the focus of this paper.

5.5.1 Baseline Results

Figures (5.5) and (5.6) report the responses of the two selected U.S. asset prices to an unexpected increase in the frequencies of the two “tantrum” topics. An increase in the importance of topic 3 (“premature tightening”) leads to a strong increase in bond yields, see Figure (5.5). A one standard deviation increase raises yields by two basis points. An increase in the frequency of topic 13 (“delayed tapering due to global risks”), however, lowers Treasury yields by two basis points. For the shock to topic 3 both impulse responses, the one in red based on the baseline model and the one in green coming from a model without FOMC dummies, exhibit virtually identical dynamics. For shocks to topic 13, however, controlling

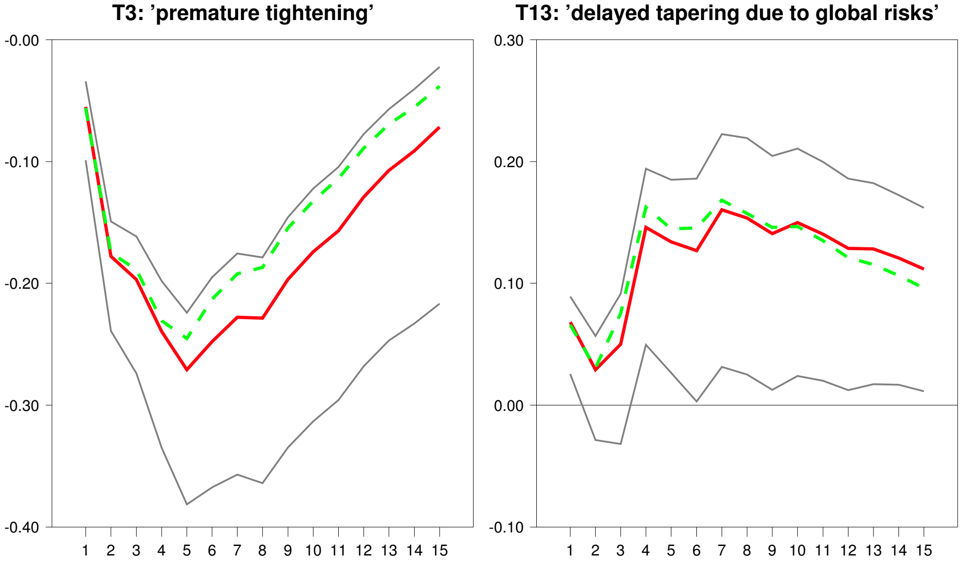
[12] The problem with alternative identification schemes, e.g. sign restrictions or heteroscedasticity-based approaches, is that we lack the theory-based restrictions needed to generate reliable impulse responses.

Figure 5.5: Response of 10-year yields



Notes: Yields are ordered last. The red (solid) line is the baseline result from a VAR with dummies for FOMC meetings and minutes releases and the the green (dotted) line is derived from a VAR without dummies. The confidence band indicates the 16th and 84th percentiles.

Figure 5.6: Response of USD-EUR exchange rate



Notes: The exchange rate is ordered last. The red (solid) line is the baseline result from a VAR with dummies for FOMC meetings and minutes releases and the the green (dotted) line is derived from a VAR without dummies. The confidence band indicates the 16th and 84th percentiles.

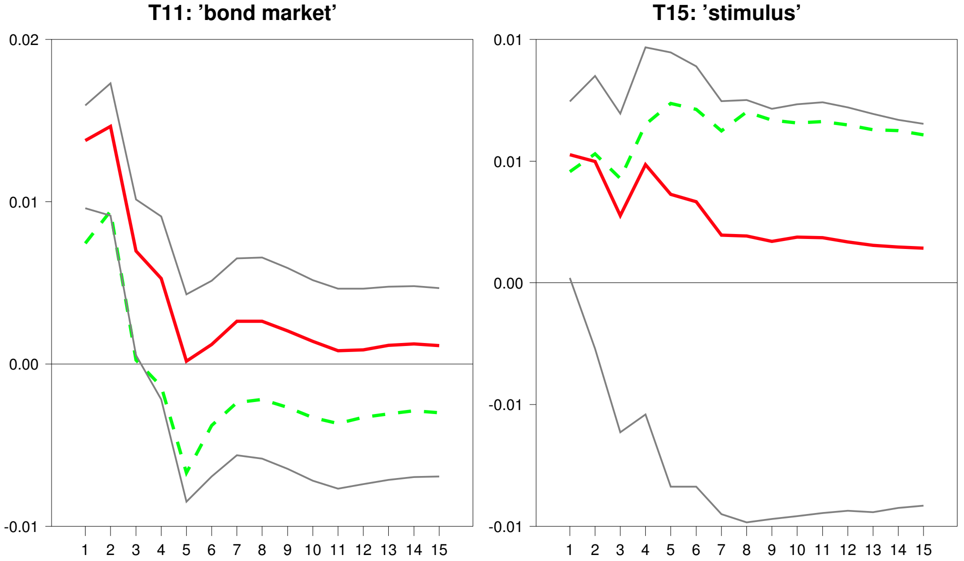
for FOMC events is particularly important as the green line indicated a weaker response to the shock. The US dollar appreciates following an increase in the frequency of topic 3 (“premature tightening”) and depreciates after a shock to topic 13 (“delayed tapering due to global risks”) as shown in Figure (5.6).

These results show that a heightened discourse of the Fed’s tapering decision, which leads to larger shares on topics 3 and 13 in the discussion, leads to high interest rate volatility. It is important to note that these estimated effects of the discussion on Twitter on asset prices do not stem from changes in the macroeconomic environment that could make a policy tightening or easing more likely. Since we control for the business cycle by including the CESI measure, the effects are driven by views of Twitter users alone. Likewise, we dummy out the days with FOMC meetings and releases of FOMC minutes. Thus, the results are not driven by tweets that simply reflect the information contained in official Fed communication on these policy days.

The asset prices responses to “QE”-related topic frequencies are shown in Figures (5.7) to (5.8). An increase in the share of topic 11 (“bond market”) or topic 15 (“stimulus”) raises bond yields. Hence, the discussion about both topics reflects the market assessment that the Fed is concerned about the overheated bond market and will likely reduce its monetary stimulus. The exchange rate response is consistent with this interpretation: the dollar appreciates following a surprise increase in topic 11 and topic 15, respectively.

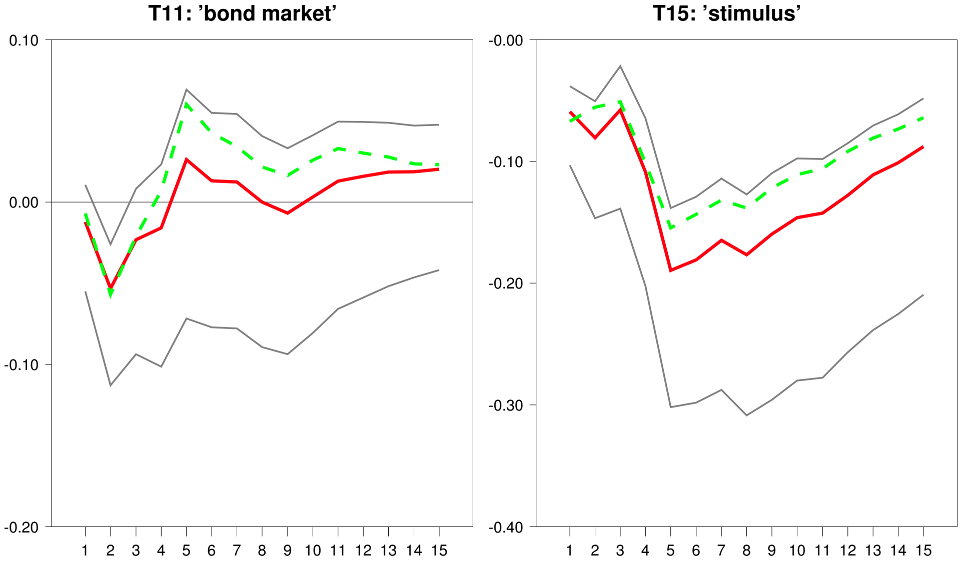
The responses to “evidence”-related topic shocks, see Figures (5.9) to (5.10), give rise to a consistent interpretation: as the Twitter discussion of topic 17 (“tight financial conditions”) intensifies, Treasury yields increase and the dollar appreciates. Thus, this topic reflects the public’s understanding of the Fed’s concern about overheating financial conditions. An increase in the likelihood of topic 20 (“data”) depresses yields and leads to a depreciation of the dollar. The more the Twitter discussion centers around the macroeconomic environment, the less likely an early tapering decision seems to be.

Figure 5.7: Response of 10-year yields



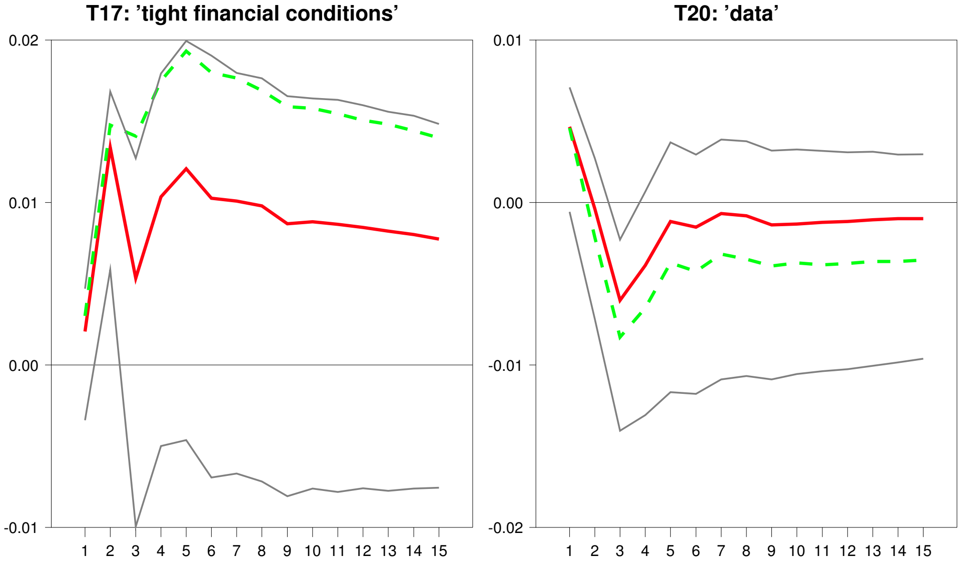
Notes: Yields are ordered last. The red (solid) line is the baseline result from a VAR with dummies for FOMC meetings and minutes releases and the the green (dotted) line is derived from a VAR without dummies. The confidence band indicates the 16th and 84th percentiles.

Figure 5.8: Response of USD-EUR exchange rate



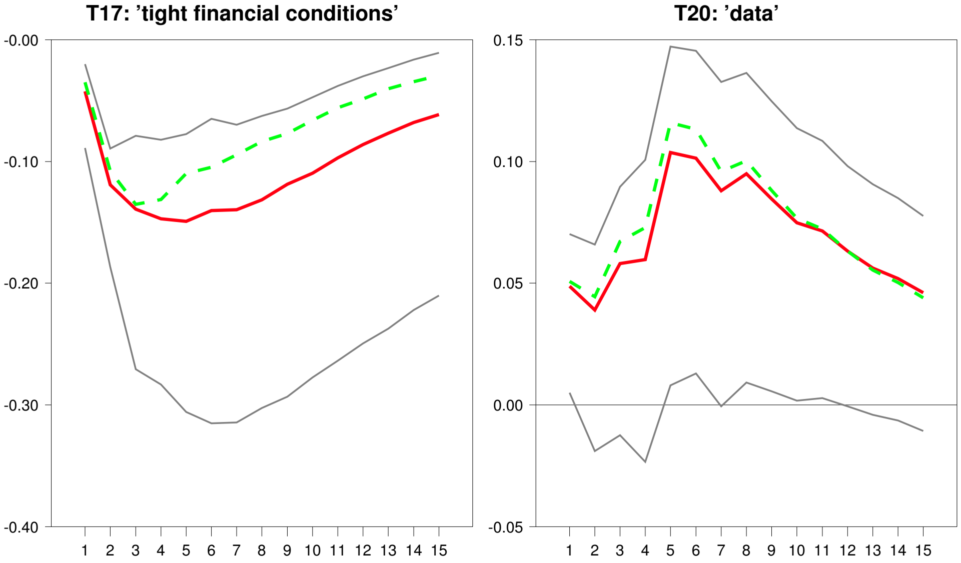
Notes: The exchange rate is ordered last. The red (solid) line is the baseline result from a VAR with dummies for FOMC meetings and minutes releases and the the green (dotted) line is derived from a VAR without dummies. The confidence band indicates the 16th and 84th percentiles.

Figure 5.9: Response of 10-year yields



Notes: Yields are ordered last. The red (solid) line is the baseline result from a VAR with dummies for FOMC meetings and minutes releases and the the green (dotted) line is derived from a VAR without dummies. The confidence band indicates the 16th and 84th percentiles.

Figure 5.10: Response of USD-EUR exchange rate

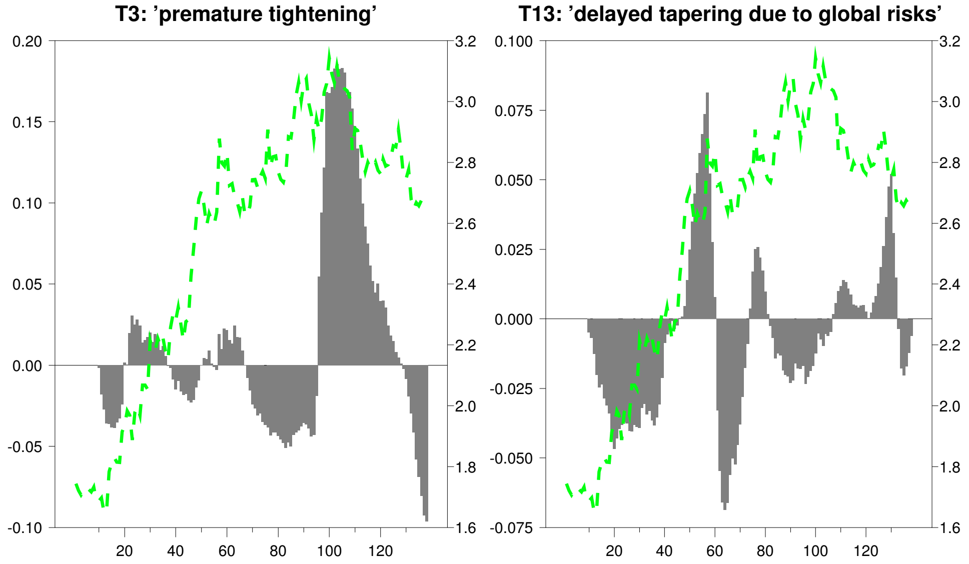


Notes: The exchange rate is ordered last. The red (solid) line is the baseline result from a VAR with dummies for FOMC meetings and minutes releases and the the green (dotted) line is derived from a VAR without dummies. The confidence band indicates the 16th and 84th percentiles.

The results show a long-lasting response of yields and the exchange rate to the topics that are immediately reflecting the tapering decision, that is, topic 3 and 13. This suggests that tweets do indeed contain information about the policy *path*. The other topics studied here, most notably topic 17 and topic 20, have, at best, a short-lived impact. This is in line with Tetlock's (2007) finding that most stock price responses to media news is driven by noise rather than fundamental information.

5.5.2 The Contribution of Shocks

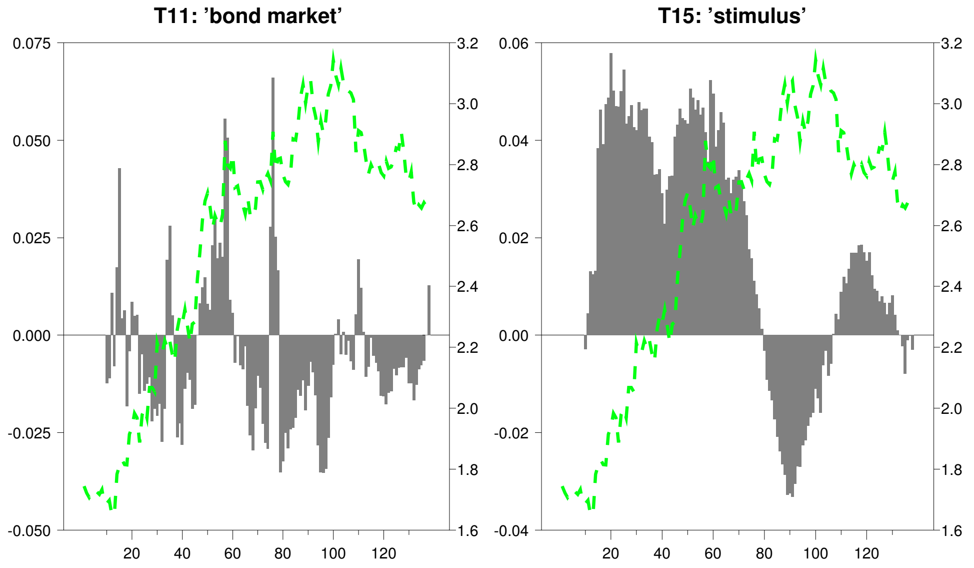
Figure 5.11: Contribution to 10-year yields



Notes: The gray bars indicate the contribution of the topic to the evolution of bond yields. The green (dotted) line (right scale) gives the 10-year bond yield. The horizontal axis shows the numbered days in the sample.

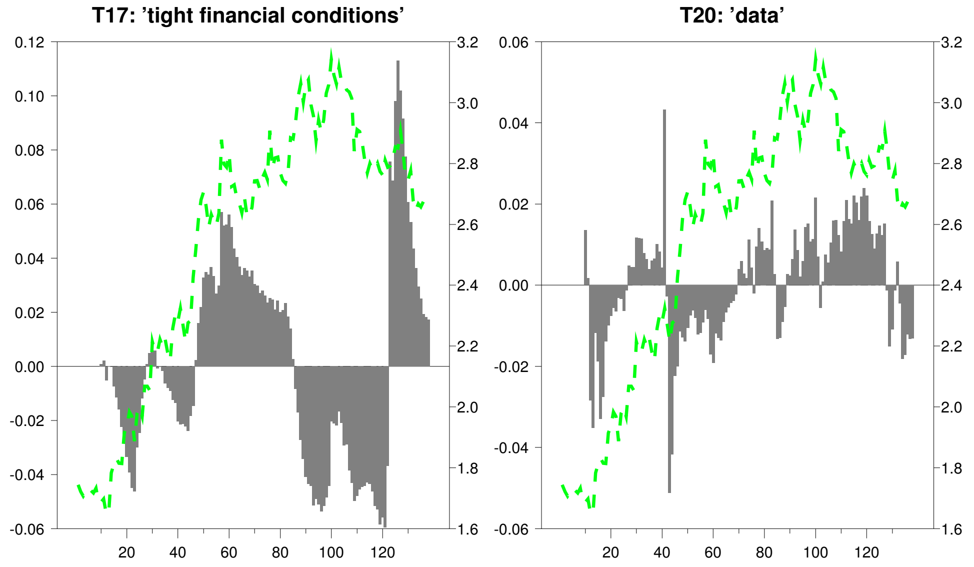
A historical decomposition of the baseline model for 10-year bond yields shows the contribution of each topic shock to the evolution of bond yields over time. Figures (5.11) to (5.13) present the historical decomposition for each set of topics based on the long-term interest rate. The magnitude of the contribution is relatively small such that we have to plot the shock contribution on a separate axis in order to visualize the bars more clearly. An increase

Figure 5.12: Contribution to 10-year yields



Notes: The gray bars indicate the contribution of the topic to the evolution of bond yields. The green (dotted) line (right scale) gives the 10-year bond yield. The horizontal axis shows the numbered days in the sample.

Figure 5.13: Contribution to 10-year yields



Notes: The gray bars indicate the contribution of the topic to the evolution of bond yields. The green (dotted) line (right scale) gives the 10-year bond yield. The horizontal axis shows the numbered days in the sample.

in topic 3 (“premature tightening”) explains a large fraction of the yield increase before the September 2013 FOMC meeting. Likewise, the steep increase in bond yields prior to the June FOMC meeting is driven by topic 13 (“delayed tapering due to global risks”). Another interesting finding is that topic 17 (“tight financial conditions”) is responsible for the drop in bond yields before and after the September 2013 FOMC meeting. A negative contribution means that the increase in topic 17 is smaller than expected, leading to a negative surprise component. Hence, compared to the tightening of market conditions in the run-up to the September FOMC meeting, the debate about financial conditions remains subdued.

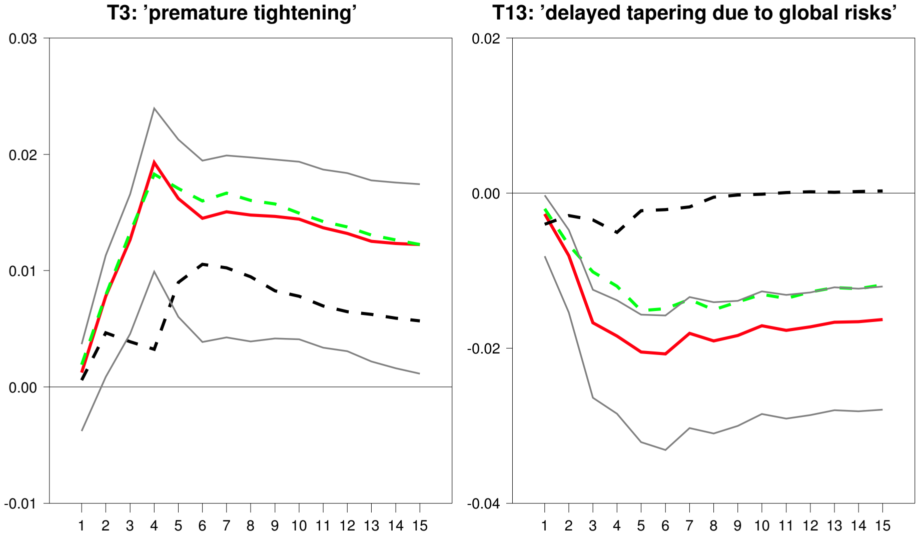
5.5.3 Decomposing Transmission Channels

The previous result shed light on the overall effects of topic shocks. The model is not able to disentangle different transmission channels. At the same time, unconventional monetary policy such as asset purchases is often believed to work through two main channels: first, to the extent different asset classes are imperfect substitutes, asset purchases by the central bank raise bond prices and, through portfolio readjustments of investors, also other asset prices. This channel is referred to as the *portfolio balance channel* of asset purchases. Second, by purchasing assets the central bank conveys information about persistently low policy rates in the future. This affects market expectations and, as a result, asset prices. The latter effect is known as the *signaling channel* of unconventional monetary policy.

Based on an estimated term structure model, any change in Treasury yields can be decomposed into changes in expected short rates and changes in the term premium. In the context of quantitative easing this is particularly important as changes in the term premium are often associated with the *portfolio balance channel* and changes in the expectation component are reflecting the *signaling channel* (see, among others, Thornton (2012), Bauer and Rudebusch (2014), and Wu (2014)).

To shed light on the two main transmission channels of asset purchases and, as a consequence, tapering, we substitute the Treasury yield used before by the estimated expectation component and, as a a separate variable, the estimated term premium. All three variables are taken from the model of Adrian, Crump, and Moench (2013).[13] The results for decomposed 10-year yields are shown in Figures (5.14) to (5.16). In each figure the red and green lines are depicting the impulse responses of the term premium. The black dotted line is the response of the expectations component of 10-year yields to a topics shock.

Figure 5.14: Response of 10-year yields with different transmission channels

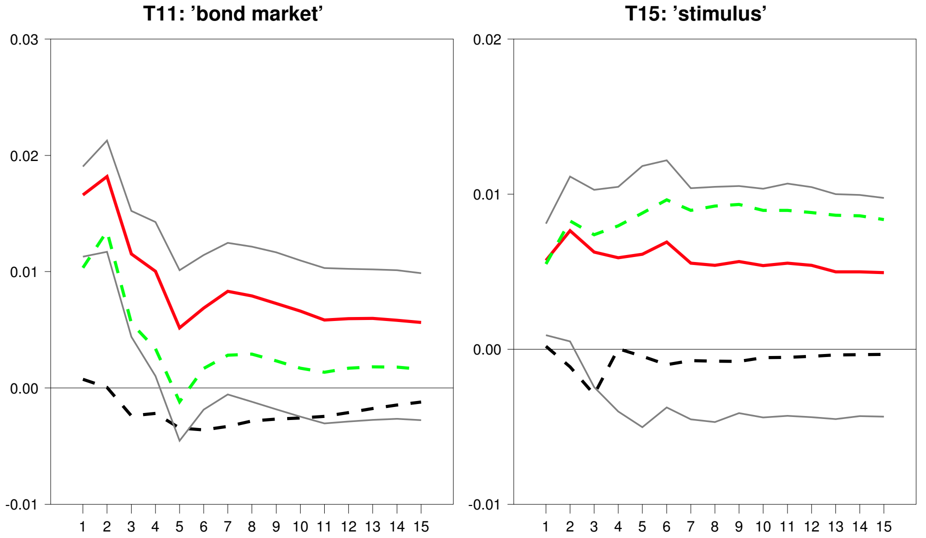


Notes: The expectations component and the term premium are ordered second last and last, respectively. The red (solid) line plots the response of the term premium for a VAR with dummies for FOMC meetings and minutes releases, the green (light dotted) line depicts the response of the term premium for a model without dummies and the black (dotted) line is the response of the expectations component of yields.

We find that the effects shown before were driven by the response of the term premium, which exhibits a significant response to the topics shocks. The expectations component shows a very weak response only. This difference between the responses of the two components of bond yields is most clearly visible for the "tantrum" and the "QE" topics. This finding is intuitive as the tapering decision of the Fed, which is discussed in our Twitter data, pertains to the exit from asset purchases under its QE3 program and not the eventual "lift-off" of

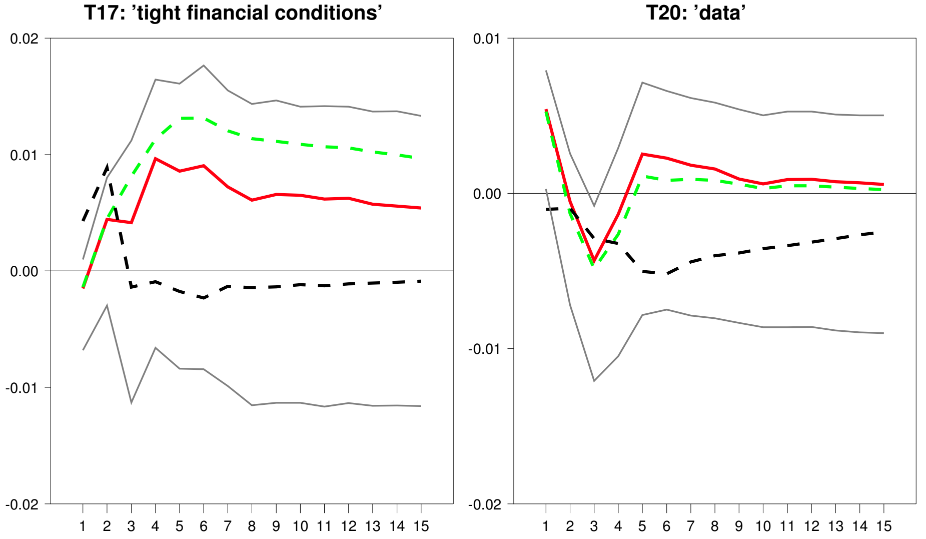
[13] The fitted yields, the estimated expectation component the term premium are available at `https://www.newyorkfed.org/research/data_indicators/term_premia.html`.

Figure 5.15: Response of 10-year yields with different transmission channels



Notes: The expectations component and the term premium are ordered second last and last, respectively. The red (solid) line plots the response of the term premium for a VAR with dummies for FOMC meetings and minutes releases, the green (light dotted) line depicts the response of the term premium for a model without dummies and the black (dotted) line is the response of the expectations component of yields.

Figure 5.16: Response of 10-year yields with different transmission channels

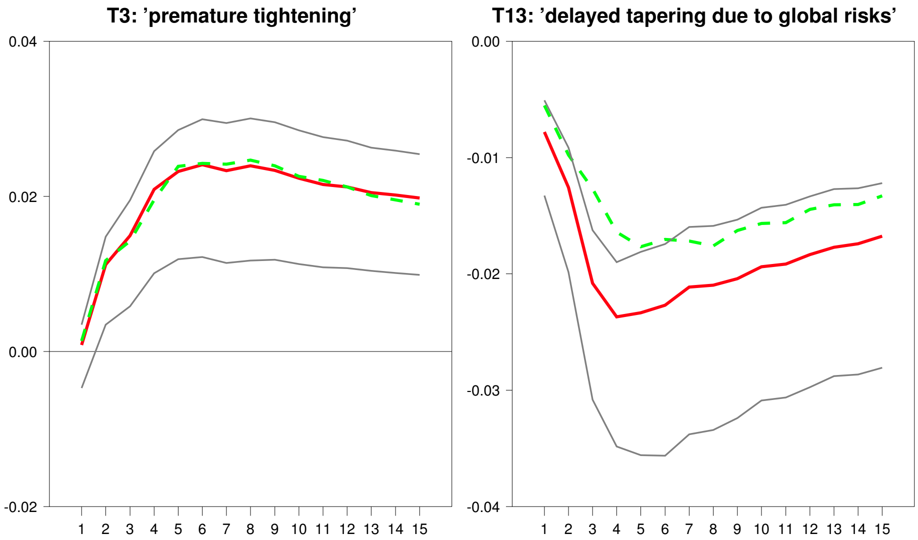


Notes: The expectations component and the term premium are ordered second last and last, respectively. The red (solid) line plots the response of the term premium for a VAR with dummies for FOMC meetings and minutes releases, the green (light dotted) line depicts the response of the term premium for a model without dummies and the black (dotted) line is the response of the expectations component of yields.

short term interest rates. As a matter of fact, unwinding QE is a tightening policy action that makes an eventual "lift-off" more likely, but the question of ending asset purchases clearly dominates the tapering discussion. Thus, our results point to a reversed *portfolio balance channel* during the taper tantrum episode.

5.5.4 Including the Number of Tweets

Figure 5.17: Response of 10-year yields with number of tweets



Notes: Yields are ordered last. The red (solid) line is the baseline result from a VAR with dummies for FOMC meetings and minutes releases and the the green (dotted) line is derived from a VAR without dummies. The confidence band indicates the 16th and 84th percentiles. The model includes the log of the daily number of tweets which is ordered fourth.

In the VAR models used so far, we derived impulse responses following a shock to each topic frequency. This shock reflects an increase in the share of a given topic on a particular day. The model does not include, however, a measure of how much Twitter activity there is on this particular day. It should matter whether we see a shift in topic probabilities in a day with only 100 tweets or on a day with 10.000 tweets. The number of tweets contains information about the attention the debate receives on Twitter. To control for the number of tweets, we include the log-level of tweets in the vector Y_t. This variable is ordered fourth, i.e. after the topic frequencies and before the asset price. The results for 10-year yields are shown in Figures (5.17). We find results which are broadly similar to our baseline findings. Thus, our findings are independent from the overall number of tweets sent on a given day. Here and in the robustness exercises below we do not show the responses to the other shocks to save space. All other results remain also unchanged.

5.5.5 Changing the Ordering in the VAR Model

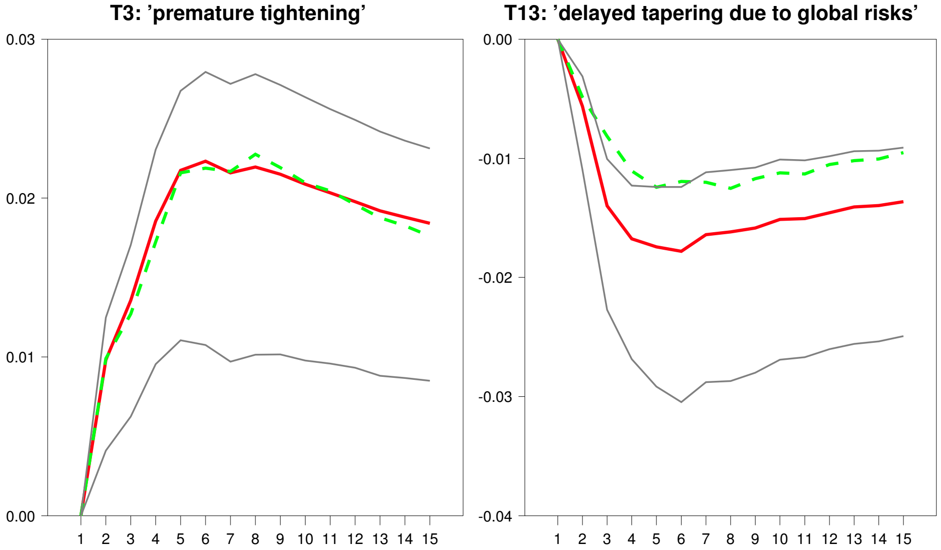
Our identification of shocks to topic frequencies is based on the assumption that on a given day the topic frequencies affect asset prices but not vice versa. This identifying assumption is reflected in the ordering of the endogenous variables in Equation (5.6). To corroborate the robustness of our results, we reverse this ordering, that is, we order asset prices second, i.e. after the macroeconomic conditions, and the topics third and fourth. If the impulse responses do not change, our baseline results are robust with regard to the identification assumption.

Figure (5.18) reports the impulse response of 10-year yields to shocks in “tantrum”-related topics. As in the baseline model, bond yields rise after a shock to topic 3 (“premature tightening”) and fall after a shock to topic 13 (“delayed tapering due to global risks”). We can conclude that using a specific ordering of the variables in our Choleski identification scheme does not drive our findings.

5.5.6 Exogenous Fundamentals

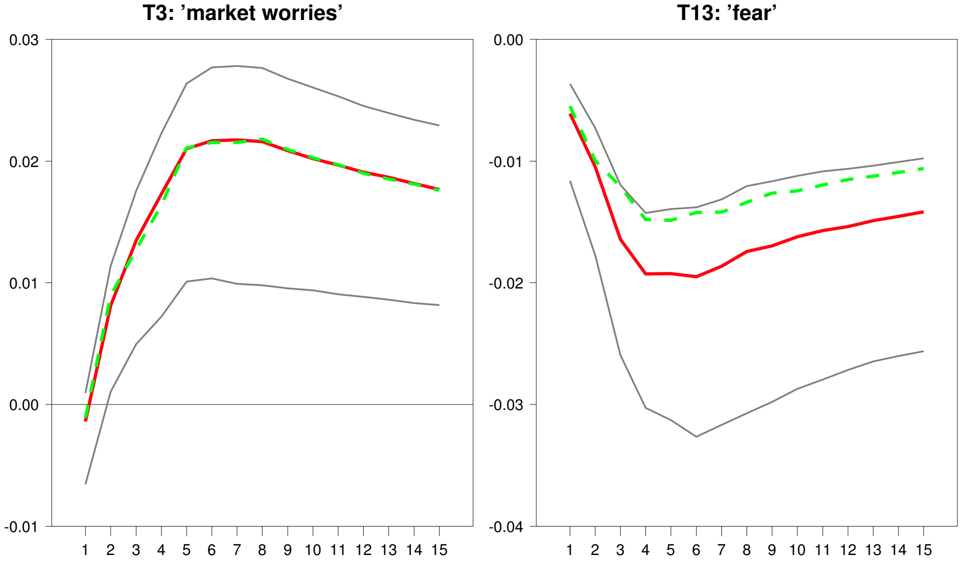
In our baseline model the daily proxy for macroeconomic fundamentals is ordered first, that is, changes in fundamentals have a contemporaneous effect on all other variables. However, with a delay of one day, changes in topics or in financial conditions are also allowed to affect fundamentals. To prevent this latter feedback effect, we estimate a model, in which fundamentals enter as a purely exogenous variable. They still affect all other variables

Figure 5.18: Response of 10-year yields for alternative VAR ordering



Notes: Yields are ordered second, i.e. before topics. The red (solid) line is the baseline result from a VAR with dummies for FOMC meetings and minutes releases and the the green (dotted) line is derived from a VAR without dummies. The confidence band indicates the 16th and 84th percentiles.

Figure 5.19: Response of 10-year yields for exogenous fundamentals



Notes: Yields are ordered first, i.e. before topics. The red (solid) line is the baseline result from a VARX, i.e. a VAR with exogenous fundamentals, with dummies for FOMC meetings and minutes releases and the the green (dotted) line is derived from a VAR without dummies. The confidence band indicates the 16th and 84th percentiles.

contemporaneously, but the feedback from financial conditions on fundamentals is absent. Thus, the VAR model becomes a VARX model

$$Y_t = A_0 + A(L)Y_t + D_t^{\mathsf{FOMC}} + A_Y Y_t^{\mathsf{CESI}} + u_t \text{ for } E\left[u_t u_t'\right] = \Sigma_u \tag{5.7}$$

where Y_t^{CESI} enters with a coefficient vector A_Y. We restrict fundamentals to affect the endogenous variables contemporaneously, that is we do not include lags of the exogenous regressor. The results, see Figure (5.19), suggest that our results are also robust with respect to the treatment of the macroeconomic fundamentals.

5.6 Conclusions

The expectations about future monetary policy matter for asset prices. The process in which expectations are formed, however, is opaque. A key contribution to expectations formation is the public debate about future monetary policy among households and investors. This paper dissects the debate about monetary policy for a period with large swings in policy expectations - the "taper tantrum" episode in 2013. Based on a large dataset containing all Twitter messages on the Fed's unwinding of asset purchases ("tapering") we use computational linguistic methods (LDA) to slice the debate into different topics. The frequencies of selected topics are then modeled in a VAR framework. We show that shocks to selected topic frequencies have significant effects on U.S. bond yields, exchange rates and stock prices.

The results are robust to the specification of the VAR model and suggest that the discourse about policy in social media matters for asset prices. With the help of social media we can shed light on the black box of expectations formation, that is, how people share and comment on information and how an aggregate market view evolves. For applications for which expectations play an important role, such as monetary policy, asset pricing and central bank communication, social media offers interesting opportunities. In particular, questions related to central bank communication and expectations management, respectively, could be addressed by using high-frequency social media data.

In future research the cross-section or network dimension of the data can be used. The present paper employs daily aggregates of the Twitter exchange. It might also be fruitful to exploit the high-frequency flow of information in the network of Twitter users and the resulting formation of expectations.

Bibliography

Adrian, T., R. K. Crump, and E. Moench (2013). “Pricing the term structure with linear regressions”. In: *Journal of Financial Economics* 110.1, pp. 110–138.

Aizenman, J., M. Binici, and M. M. Hutchison (2016). “The Transmission of Federal Reserve Tapering News to Emerging Financial Markets”. In: *International Journal of Central Banking* 12.2, pp. 317–356.

Bauer, M. D. and G. D. Rudebusch (2014). “The Signaling Channel for Federal Reserve Bond Purchases”. In: *International Journal of Central Banking* 10.3, pp. 233–289.

Bholat, D., S. Hans, P. Santos, and C. Schonhardt-Bailey (2015). *Text mining for central banks*. Handbooks 33. Centre for Central Banking Studies, Bank of England.

Blei, D. M. and J. D. Lafferty (2007). “A correlated topic model of Science”. In: *Annals of Applied Statistics* 1.1, pp. 17–35.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). “Latent dirichlet allocation”. In: *Journal of machine Learning Research* 3, pp. 993–1022.

Blinder, A. S., M. Ehrmann, M. Fratzscher, J. D. Haan, and D.-J. Jansen (2008). “Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence”. In: *Journal of Economic Literature* 46.4, pp. 910–45.

Deerwester, S. C., S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman (1990). “Indexing by latent semantic analysis”. In: *Journal of the American Society for Information Science* 41.6, pp. 391–407.

Eichengreen, B. and P. Gupta (2015). “Tapering talk: The impact of expectations of reduced Federal Reserve security purchases on emerging markets”. In: *Emerging Markets Review* 25.C, pp. 1–15.

Goldsmith-Pinkham, P., B. Hirtle, and D. O. Lucca (2016). *Parsing the content of bank supervision*. Staff Reports 770. Federal Reserve Bank of New York.

Griffiths, T. L. and M. Steyvers (2004). "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101.suppl 1, pp. 5228–5235.

Grün, B. and K. Hornik (2011). "topicmodels: An R Package for Fitting Topic Models". In: *Journal of Statistical Software* 40.13, pp. 1–30.

Hansen, S. and M. McMahon (2016). "Shocking language: Understanding the macroeconomic effects of central bank communication". In: *Journal of International Economics* 99, Supplement 1. 38th Annual NBER International Seminar on Macroeconomics, S114–S133.

Hansen, S., M. McMahon, and A. Prat (2014). *Transparency and Deliberation within the FOMC: A Computational Linguistics Approach*. CEP Discussion Papers dp1276. Centre for Economic Performance.

Hendry, S. (2012). *Central Bank Communication or the Media's Interpretation: What Moves Markets?* Staff Working Papers 12-9. Bank of Canada.

Hendry, S. and A. Madeley (2010). *Text Mining and the Information Content of Bank of Canada Communications*. Staff Working Papers 10-31. Bank of Canada.

Hofmann, T. (1999). "Probabilistic latent semantic indexing". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 50–57.

Larsen, V. H. and L. A. Thorsrud (2015). *The Value of News*. Working Papers 0034. Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School.

Loughran, T. and B. McDonald (2016). "Textual Analysis in Accounting and Finance: A Survey". In: *Journal of Accounting Research* 54.4, pp. 1187–1230.

Lucca, D. O. and F. Trebbi (2009). *Measuring Central Bank Communication: An Automated Approach with Application to FOMC Statements*. NBER Working Papers 15367. National Bureau of Economic Research.

Lüdering, J. and P. Winker (2016). "Forward or backward looking? The economic discourse and the observed reality". In: *Jahrbücher für Nationalökonomie und Statistik* 236.4, pp. 483–516.

Meinusch, A. and P. Tillmann (2016). *Quantitative Easing and tapering uncertainty: evidence from Twitter*. unpublished. Giessen University.

Mishra, P., K. Moriyama, P. M. P. N'Diaye, and L. Nguyen (2014). *Impact of Fed Tapering Announcements on Emerging Markets*. IMF Working Papers 14/109. International Monetary Fund.

Schonhardt-Bailey, C. (2013). *Deliberating Monetary Policy*. MIT Press, Cambridge.

Tetlock, P. C. (2007). "Giving Content to Investor Sentiment: The Role of Media in the Stock Market". In: *Journal of Finance* 62.3, pp. 1139–1168.

Thornton, D. L. (2012). *Evidence on the portfolio balance channel of quantitative easing*. Working Papers 2012-015A. Federal Reserve Bank of St. Louis.

Wang, X. and A. McCallum (2006). "Topics over Time: A non-Markov Continuous-time Model of Topical Trends". In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '06, pp. 424–433.

Wu, T. (2014). *Unconventional Monetary Policy and Long-Term Interest Rates*. IMF Working Papers 14/189. International Monetary Fund.

6 Standing and "Survival" in the Adult Film Industry

Jochen Lüdering

This article is available as J. Lüdering (2016b). *Standing and Survival in the Adult Film Industry*. MAGKS - Joint Discussion Paper Series in Economics 26/2016. Philipps-Universität Marburg.

It has been presented at the annual congress of the European Economic Association August 24th, 2016 in Geneva. Early versions were presented at the MAGKS Colloquium March 17th, 2016 in Rauischholzhausen and in the framework of the class on "Survival Analysis and Industrial Dynamics" taught by Prof. Dr. Guido Buenstorf, January 7th 2014 at Kassel University.

Acknowledgments I am indebted to Pascal Hofmann and Julian Kessel for their assistance in obtaining and pre-processing the data. Further, I want to thank the participants of the class on Survival Analysis taught by Guido Buenstorf (Kassel University) and participants in the EEA session on Labor Economics for their helpful comments. Last but not least I am grateful for the data collection by IAFD.com and their contributors, which has enabled this work.

JEL classification: C41, D80, C55

Keywords: Duration Analysis, Professional Success, Network Analysis, Centrality, Social Capital

Abstract

It is conventional wisdom that knowing the right people is essential for one's career, which is supported in the literature on *social capital*. However, the empirical evidence in this field remains ambiguous. While the literature recognizes that "connections" certainly help finding any job at all, it remains unclear if there are long-term career benefits.

While it is difficult to record a network structure in most industries, in the adult film industry collaborations between performers are easily observed. Consequently, a collaborative network can be constructed for which centrality measures can be calculated in order to estimate the effect of a person's centrality on individual success. Unfortunately, success is not easily observed either. Hence, in this manuscript, the survival in the industry is used as a proxy for professional success. This assumption is justified by the economic argument that, in the absence of lock-in effects, performers will remain in the industry as long as it remains profitable. The profitability will not only depend on monetary aspects, but, in addition to economic costs, it takes potential costs from loss of reputation and personal well-being into account.

The research at hand stands out by pioneering the use of centrality measures in duration models which, to the best of my knowledge, is a novel approach in the field. The results indicate that there is a strong correlation between network centrality and industry survival for the adult film industry.

6.1 Introduction

It is widely accepted that networking is an important aspect to get ahead within ones career. There is a whole industry existing around the idea of facilitating networking between individuals. Online Social Networks, in particular career oriented ones like Xing and Linkedin, are the most visible intermediaries in this industry, which promise to formalize the social interaction and allow people to exploit this important resource more efficiently.

The importance of a social network (in the interpersonal sense) for professional success is an integral part of social capital theory (Lin 1999a). It has provoked a substantial amount of research in economics and sociology. The relevant sociological literature is summarized by Gorman and Marsden (2001), who document the empirical evidence to be inconclusive whether networking leads to any long lasting advantages. However, in the short run networking seems to be beneficial as it helps to find any job at all (See Granovetter (1995) for a discussion and recent evidence by Brady (2015)). Moreover, it appears to be problematic that a majority of studies work on cross-sectional data and make inference about dynamic developments. Wolff and Moser (2009) who study the benefit of networking from a sociological point of view based on longitudinal data are a noteworthy exception to this.

The early economic literature on the effects of social networks is surveyed by Ioannides and Loury (2004). Similar to the sociological literature, the authors find ambiguous evidence regarding the effect of using personal contacts to find a job. Their findings suggest that job specificities may be determining the sign of the effect. More recent findings by Brown, Setren, and Topa (2016) and Burks et al. (2015) find small positive effects on wages. Fernandez, Castilla, and Moore (2000) find it to be particularly beneficial for the employer rather than for the employee to utilize referrals in their hiring process.

In empirical work it appears to be a central issue that personal interactions are hard to track in most industries and researchers therefore often rely on surveys. An exemption to this is the film industry, where joint movie appearances are easily observed. Early examples of using screen credit data includes Faulkner and Anderson (1987), these information on collaborations have subsequently been used in descriptive studies of the film industry (Ahmed et al. 2007; Herr et al. 2007). Further, the collaborative network of the film industry has been used to illustrate the *small-world experiment*, which became part of popular culture with the advent of "Six Degrees of Kevin Bacon" [1] in 1994. An actor's Bacon number is the collaborative distance from Kevin Bacon. Older and more relevant for economists is the Erdös Number, a similar measure for scientific publications. The first record of the Erdös number dates back to Goffman (1969). A person's Erdös number measures the collaborative distance between himself and the Hungarian mathematician Paul Erdös.[2] Despite the popularity of these measures, the selection of Erdös and Bacon as reference points in the collaborative networks is rather *ad hoc*. At least in the case of Bacon we know that he is far from being the most central person in the actor network.[3] More "sensible", classical centrality measures of prominence in the collaborative network in economics are provided by RePEc (`http://collec.repec.org`). To my knowledge these data are still waiting to be used in an economic publication. Research regarding the relationship between networking and professional success in academia has to a large extent been based on the impact of co-authorship networks on "citation performance" (Abbasi, Altmann, and Hossain 2011; Bordons et al. 2015; Cimenler, Reeves, and Skvoretz 2014). The results indicate that centrality is correlated with the author's publication performance measured by their g-index (Egghe 2006). However, in the light of the existence of so called "citation clubs" (Kostoff 1998), where the same authors regularly cite members of their clubs due to sympathy, friendship or strategic career considerations, problems of endogeneity become likely in these type of settings.

My contribution is primarily a methodological one, but it also uses an interesting set of

[1] `https://en.wikipedia.org/wiki/Six_Degrees_of_Kevin_Bacon` accessed 2015-08-03.

[2] My own Erdös number is four, which I obtained through my recent collaboration with Peter Winker (3). Winker had previously collaborated with Dennis K. J. Lin (2), who in turn has a joint publication with Erdös coauthor Gutti Jogesh Babu (1). (Determined using `http://www.ams.org/mathscinet/collaborationDistance.html`)

[3] `http://www.randalolson.com/2015/03/04/revisiting-the-six-degrees-of-kevin-bacon/` accessed 2015-08-03.

novel data. In order to address the question whether collaborating with the "right" people will lead to "professional success", I combine centrality measures derived from social network analysis with methods of survival analysis. To the best of my knowledge, this combination of methods is a novel approach and has not been used in economics so far.

The adult film industry, which is the subject of this analysis constitutes a classical freelancer economy. Performers are usually employed on a project basis. In that respect my research is similar to Nadler (2016), who analyze the hiring patterns of lightning technicians and grips (rigging technicians on a film set) on film sets. As a first step I construct a network of the adult film industry, where personal interactions are easily observed. From this rather descriptive exercise, the position of each performer in the industry is obtained. In a second step these centrality measures are used as a proxy for connectedness in an empirical analysis to address the central question whether being connected is beneficial for professional success. Contrary to the movie industry, where statistics about box office revenues are published, there are no information on the revenues from adult films or other direct measures of success. Hence, this paper relies on *survival* in the industry as a measure of economic *success* and argues that one person stays in the industry as long as (economic) benefits exceed costs. The latter are not limited to financial burdens but also include social stigmatization and marginalization resulting from involvement in the industry. Hence, a person will only remain in the industry if the condition holds that expected utility is larger than expected costs:

$$E_t(U_{t+1}) > E_t(C_{t+1}). \tag{6.1}$$

While there is ample evidence that individuals remain in a position longer if they obtained the position through networking, it becomes a much better proxy for success if one considers the whole industry and pattern of freelance employment.

The particularities of duration data require the use of specialized econometric methods: survival analysis. The application of this method has some tradition in economics (Kiefer

1988) and is widely used in the field of industrial dynamics (Mata and Portugal 1994; Disney, Haskel, and Heden 2003; Buenstorf 2007), employment (Hunt 1995; Card, Chetty, and Weber 2007; Kuhlenkasper and Kauermann 2010) and education (Edwards and Ureta 2003; Gury 2011).

The remainder of the paper is organized as follows: Section 6.2 covers the theoretical background. The adult film industry and the dataset used in the analysis are portrayed in more detail in Section 6.3. Section 6.4.1 and 6.4.2 are concerned with the definition of the two concepts *success* and *being connected*. In Section 6.5 the survival model is employed to analyze the effect of centrality for success. The results are discussed and put into perspective in Section 6.6 before Section 6.7 closes with a conclusion.

6.2 Social Networks in the Theory of Social Capital

In an early definition Bourdieu (1983, own translation) described social capital as encompassing "the entirety of present and potential resources, which are connected to the possession of a durable network of more or less institutionalized connections of mutual acquaintance and recognition. In other words, these resources are derived from membership of social group." Subsequently, the concept of social capital has been extended by various authors to include several means or services which a society provides for its members, including institutions, a feeling of "belonging" as well as social interaction. Lin (1999a) is one attempt to structure these different concepts along several dimensions.

One strain of the sociological literature, dating back to Granovetter (1973) and Granovetter (1983), analyses the structure of a social network arguing that it is important to have a large social network to derive the largest benefit. This kind of thinking gives rise to the "Strength of weak ties" argument (Granovetter 1973). It says that it is the weak ties between people that are of importance for individuals. Weak ties exist between colleagues, teachers and students, and acquaintances. In contrast, close friends and family members are connected through strong ties. The argument is that individuals connected through a strong link are to a large extent connected to the same people. However, an individual i to whom a person j is connected through weak links only, likely has strong links to persons outside of the immediate neighborhood of j. Hence, a single weak link between i and j increases the size of the social network of i, and hence access to information, by a larger extent, than a single strong link to a close friend f whose network largely overlaps.

A competing approach, the social resource theory, developed by Lin (e.g. described in Lin 1999b), argues that resources are contained inside a social network, which enables an individual to access the resources through the network. Consequently, it is not only important how the network is structured but also with whom a person is connected.

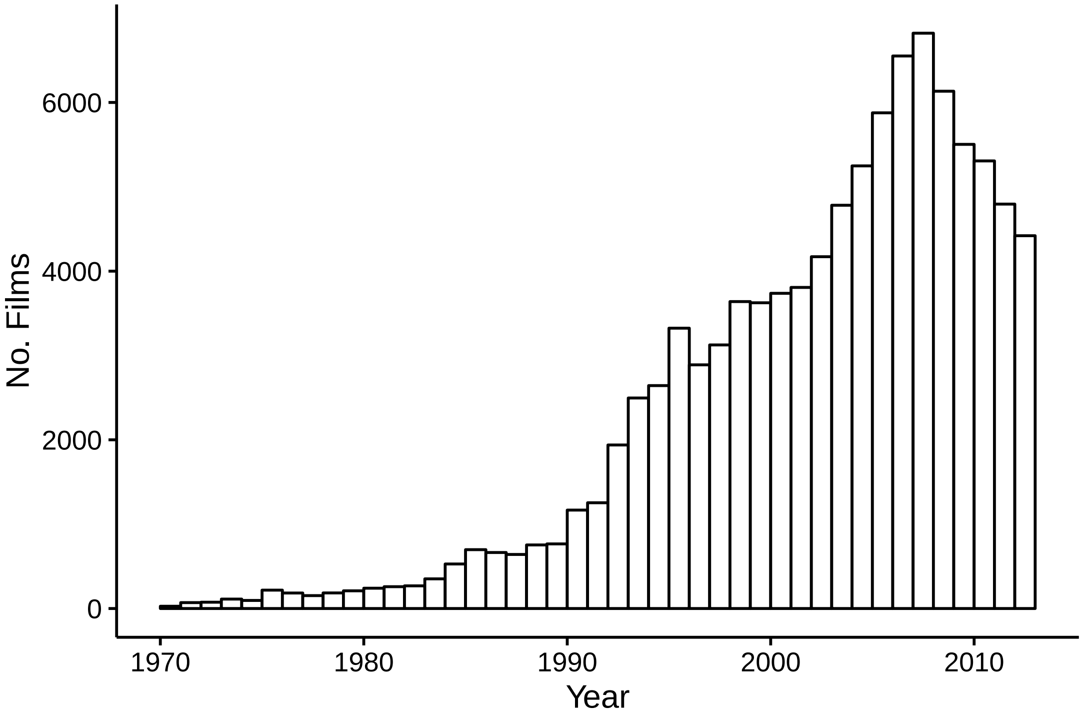
Whether it is link strength or its destination, there is an agreement that social capital is a resource on which an individual can draw upon through the usage of a social network. This research takes up both theoretical concepts in the empirical estimation, considering not only own centrality (Granovetter) but also the centrality of the persons in the neighborhood (Lin), as sources for potential success.

The economic debate is concerned with the value of a social network in the narrow sense as a mean to facilitate the exchange of information which is the basis for the research at hand. An early application in labor economics is Montgomery (1991), who shows that in a labor market with adverse selection, well-connected workers are better off than those without large social networks, because firms rely on referrals for hiring to overcome adverse selection. Hence, an individual without contacts to a company would need to exercise substantial effort for signaling and self-marketing. More contemporaneous theoretical works include Calvó-Armengol and Jackson (2004) and Ioannides and Soetevent (2006), who explicitly model a network of social contacts and analyze the effect of networking on employment, wages and resulting inequality.

6.3 The Adult Film Industry

The adult film industry is geographically centered in San Fernando Valley, located 30km from Los Angeles. Danta (2009) estimates that in 2009 about 71% of the adult film industry was located in “the valley”, with an additional 12% in other parts of the US and only 2% of the industry located abroad. The location provided excellent conditions for the industry, due to the proximity to Hollywood (and thus access to fortune seeking young actors) and the 1988 ruling of the California Supreme Court[4] that the production of adult films is protected as *free speech* under the First Amendment.

Figure 6.1: Number of Films per Year



Statistics on adult entertainment are scarce and accuracy of the available data is highly debated. The industry’s trade journal, Adult Video News (AVN), estimated 12.8 billion US$

[4] The People v. Harald Freeman: see `https://en.wikipedia.org/wiki/People_v._Freeman` accessed February 1st 2016.

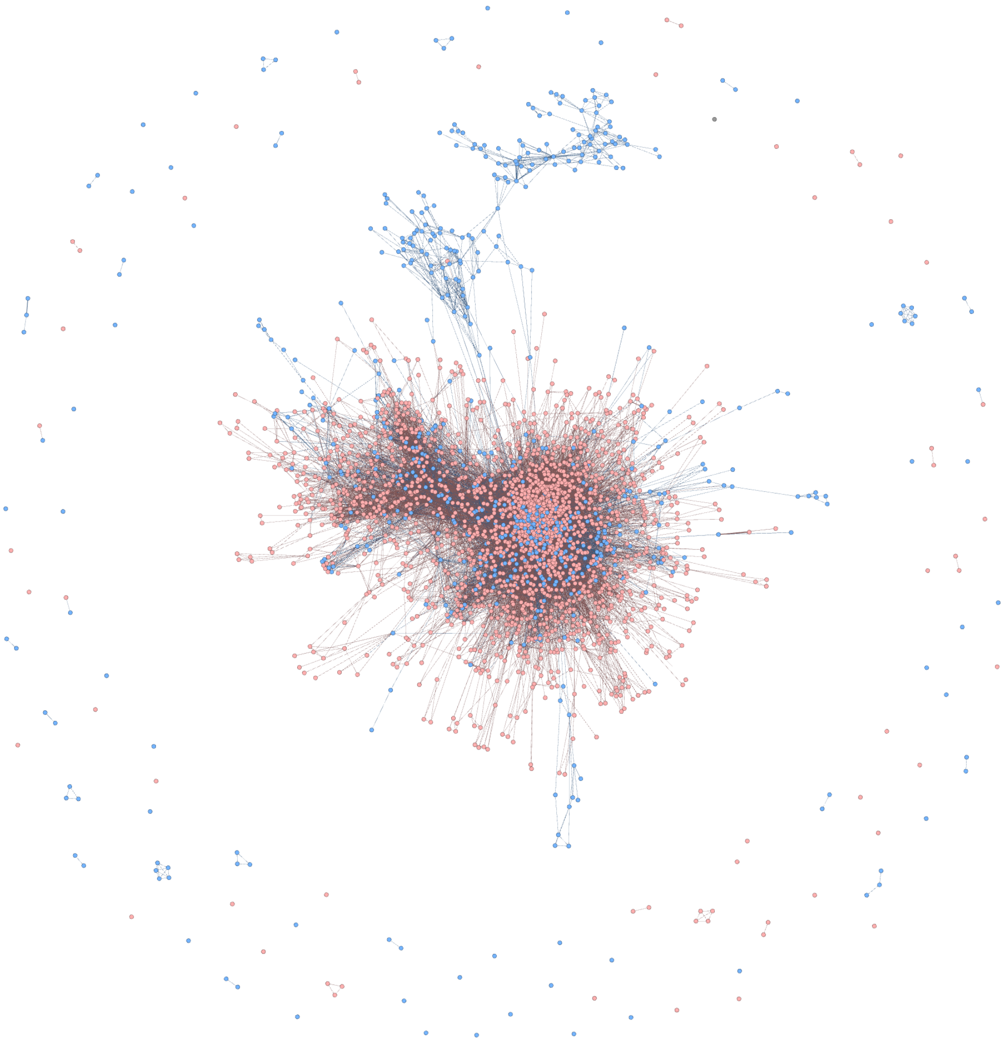
of retail sale in the *Adult Entertainment Subsector* (as cited in Edelman 2009) for the year 2006.[5] The largest part of the proceeds is accounted for by "Video sales and rentals". Still, the revenue from video sales were 15% less than in the previous year, which is in line with the data used in this paper, where the number of films released per year have been declining since 2008 (see Figure 6.1).

The analysis at hand uses a dataset obtained from the Internet Adult Film Database (`iafd.com`). It consists of 7001 female and 2886 male performers who appeared in 102,871 films (which excludes compilations and web scenes). A comprehensive overview on the wealth of the data on `IAFD.com` can be found in Millward (2013), who goes in great detail about the distribution of places of birth, measurements and hair color. The data used here was downloaded in the last quarter of 2013. Despite the extraction date in 2013 there were already three films included which were to be released in 2014. Although the first film in the database was published in 1951, the analysis focuses on the period between 1970 and 2013, when a sufficient number of films appeared every year. The number of films per year are shown in Figure 6.1. One observes that the number of films released was growing from the beginning up until 2007, when 6821 films were published. The successive demise can likely be explained by the substitution of DVD-style releases with Internet pornography (for the history of the interplay between pornography and technology see Coopersmith 1998), which is not (to the same extent) captured in the database. To avoid issues of selection bias all web scenes are excluded from the analysis.

Figure 6.2 shows the network structure in 2013. In the graph, nodes represent individual performers which were active in that year. Two performers are connected by an undirected link if they appeared together in at least one film in 2013. While the largest share of female (red) performers are centered in the middle, there are a few communities of performers completely disconnected from the main network. Moreover, it is interesting to see that there are also

[5] For the debate on the validity of this number see Rich (2001), Ackman (2001) and Silverstein (2006).

Figure 6.2: Visualization of the performer network in 2013



Note: Red indicates a female performer, blue indicates a male performer. A link between two dots indicates a joint film appearance.

groups consisting largely of men (blue) constituting the gay pornography industry sector. There are only very few individuals who connect the homosexual and heterosexual markets.

The question of data quality is a delicate one. In particular, information provided by the actors and studios on performer characteristics are potentially flawed. It is not unlikely that the year of birth is misreported for marketing purposes, as it remains questionable if performer *Rose Agree* really joint the porn industry at the age of 87. In total there are 11 performers who joined the industry above the age of 65 and 50 who were below the age of 18. The latter cases can roughly be split in three groups. There are at least two cases were performers were added to a film later when it was re-released on DVD, e.g. resulting in the (obviously wrong) calculated age of 8, even though the scenes were filmed when the actress was in her twenties. There are also the cases of Tracy Lords (born 1967) and Alexandra Quinn (born 1973) who were later confirmed to be underage at the time of making their first films, which prompted legal action. The recent cases mostly are young Brazilian men, who state to be younger than 18. Compared to the age, the information on which actor appeared in which film is likely much more reliable.

Figure 6.3: Kaplan-Meier estimate for survival of male and female performers

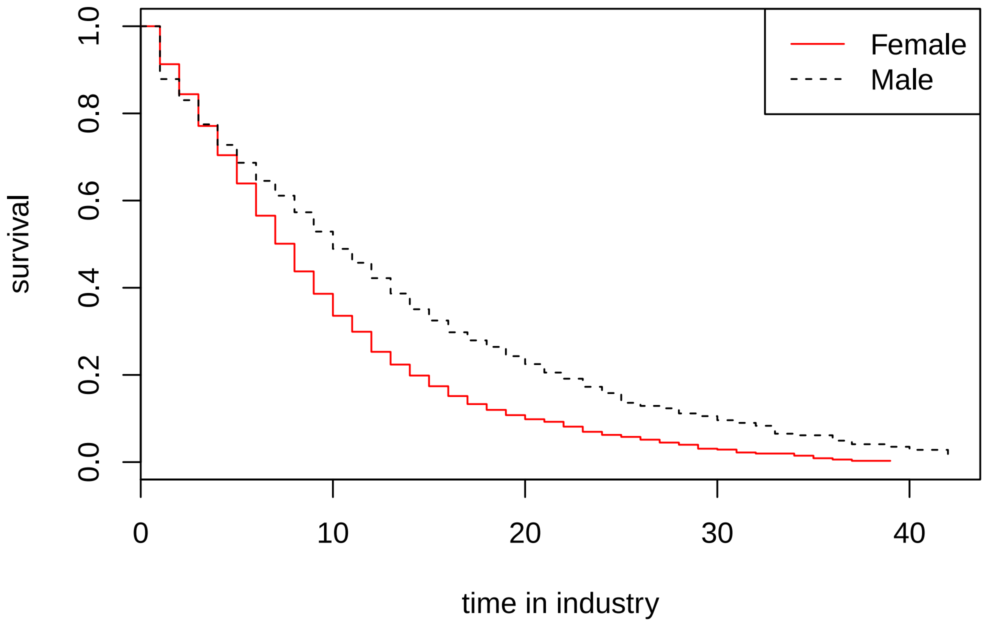


Figure 6.3 shows the Kaplan-Meier (Kaplan and Meier 1958) estimates of the survival function for male and female performers, the lines indicate the share of performers remaining in the industry after a given number of years. The graph shows that more male than female performers drop out within the first three years. Afterwards, the “survival” of men is higher than that of their female counterparts. A large gap arises after ten years, when only about 30% of women and 50% of men remain in the industry. There are no observations of women who have been in the industry for more than 40 years.

6.4 Concepts of Success and Connectedness

This section will further elaborate on the two central concepts and their operationalization in this paper. It starts out with the discussion of *success* and afterwards elaborates on the metric of *network centrality* used in this analysis.

6.4.1 Survival as Success

Due to absence of direct measures of economic success, like sales, box-office revenues and actors salaries, I adopt the notion that remaining in the business is a success by itself. Based on economic intuition one would argue that a performer leaves the industry, if an activity is no longer worthwhile economically. These changes could result from increases in costs, reduced revenues or arising outside options. Hence, survival in the industry may serve as a proxy for economic success.

For this argument to hold, an actor needs to be, at least to a certain extent, rational and has to have a good understanding of the costs associated with his involvement in the adult film industry. These are not necessarily limited to financial costs, but also include societal costs resulting from social stigma or adverse effects on personal health.

Further, the absence of lock-in effects must be assumed. These exists in industries with high fixed costs, which — at least financially — is not the case in the adult film industry.

Given an imperfect labor market, lock-in effects could also stem from the lack of outside options due to high search costs.

6.4.2 Network metrics

As outlined in the discussion on the theoretical nature of social networks, the career effects of centrality is likely twofold. On the one hand a person can be central in the network herself and thus be well connected. On the other hand it may also be beneficial to be linked to a central person, without necessarily having a very prominent role on the network oneself. Both effects are included in the analysis and measured by betweenness centrality of the performer herself and the centrality in her neighborhood.

Betweenness centrality

The position of a performer within the network is measured by *betweenness* centrality. It has a meaningful finite value for vertices being outside the network (0), as it is a measure of the ratio of "shortest pathes" leading through a given vertex. The alternative "closeness", measuring the number of steps to every other vertex, would imply a value of infinity for vertices outside the network, which would pose difficulties, when using the data in a subsequent regression analysis.

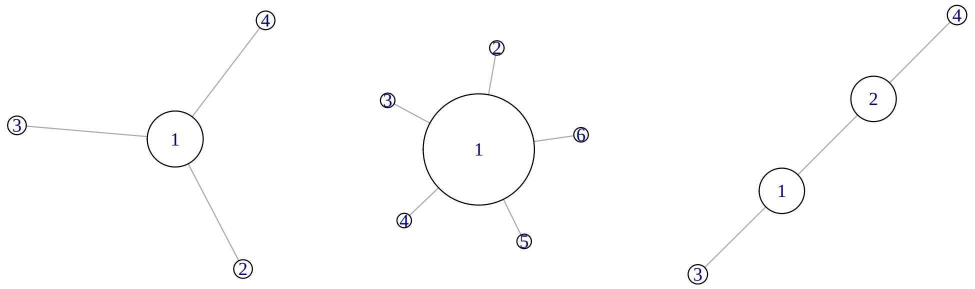
According to Freeman (1978) the betweenness centrality (C_B) of node i is given by

$$C_B(i) = \sum_{k \neq j; i \notin \{k,j\}} \frac{P_i(k,j)}{P(k,j)} \tag{6.2}$$

$P(k,j)$ is the total number of shortest paths between the nodes k and j. $P_i(k,j)$ indicates the number of shortest paths between k and j leading through node i. Hence the sum over P_i/P for all k and j gives the share of shortest paths leading through node i.

Figure 6.4 (a and b) shows that betweenness centrality (C_B), indicated by the size of the

Figure 6.4: Visualization of Betweenness Centrality



(a) Betweenness of (1): 3 (b) Betweenness of (1): 10 (c) Betweenness of (1): 2

vertices, for identical shaped networks is influenced by network size. Hence, betweenness (C_B) must be normalized according to

$$C_B^{\text{norm}}(i) = \frac{C_B(i)}{(n^2 - 3 \cdot n + 2)/2}. \tag{6.3}$$

The denominator is the maximum possible value for C_b, which is the betweenness value in a star-shaped network (Freeman 1977) as a function of the number of nodes (n) in the graph. As shown in Figure (6.4) the vertex at the center of a star-shaped network (a and d) has the maximum betweenness score, given a constant number of nodes, compared to other shapes. This is rather obvious if one compares (a) and (c) in Figure 6.4, which consists of four nodes each. In the first case all shortest paths between vertex 2, 3 and 4 lead through the central vertex 1. In (c) only two shortest paths lead through vertex 1 (3–2, 4–3).

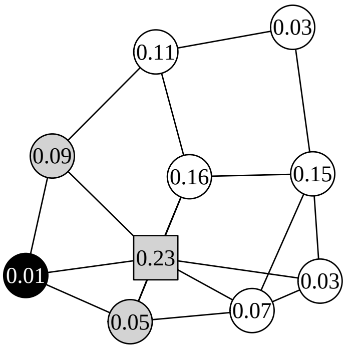
Neighborhood

The neighborhood of a vertex i consists of all the other vertices h it has a direct connection with. Henceforth, the *maximum neighbor centrality* of a vertex i (MNC$_i$) is given by

$$\text{MNC}_i = \max_{h \in N(i)} C_B^{\text{norm}}(h) \tag{6.4}$$

Figure 6.5 visualizes an example of this measure in a graph. The vertices are labeled with their normalized betweenness scores. For example, the *MaxNeighborCentrality* is to be determined for the black vertex. Among the direct neighbors (grey) the square shaped vertex has the highest value for betweenness ($0.23 > 0.09 > 0.05$). Hence, the MNC for the black node is 0.23, the betweenness value of the square neighbor.

Figure 6.5: Network showing MaxNeighborCentrality



6.5 The Relationship between Networking and Success

The empirical analysis in this section examines the relationship between the probability to leave the industry in the next period (hazard rate), a number of variables indicating the position of an individual in the social network and control variables for gender and age. Table 6.1 provides the descriptive statistics for the dataset. “Event” is a dummy variable indicating if the person has left the industry yet – the mean of 0.571 tells us that about 57% of the actors in the sample had left the industry in 2013. Year of birth indicates the year of birth, StartAge the age at joining the industry. StartYear and EndYear show the period of activity in the adult film industry. If the person is male, the dummy variable male is equal to 1. Centrality, MNC and the “disconnected” dummy are measures of the performers position in the network.

Table 6.1: Summary Statistics

	mean	sd	median	min	max
Event	0.571	0.495	1.000	0	1
Year of Birth	1977.180	10.978	1980	1900	1995
StartAge	24.488	6.671	23.000	*	87
StartYear	2001.668	7.983	2004	1951	2013
EndYear	2007.405	6.736	2010	1969	2013
male	0.265	0.441	0	0	1
Centrality	0.001	0.005	0.000	0.000	0.187
MNC	0.018	0.020	0.012	0.000	0.187
disconnected	0.033	0.179	0	0	1

Note: Variables in the top panel are constant across time, whereas the variables in the bottom panel are time varying. For the latter ones descriptive statistics are calculated over all observations in individual and time dimension.

*: Calculated minimal starting age is 8 (Person was added to the film years later). Confirmed minimum age when joining the porn industry is 16 (Traci Lords, Alexandra Quinn).

In order to analyze the duration data, it is necessary to resort to methods of survival analysis for two reasons: Duration data has a very characteristic distribution a) it is never negative and b) it is usually not normally distributed. Additionally, the methods can be used to deal with censored and truncated data. The most widely used model is the Cox regression models, as it appeals to many applied researchers, because the semi parametric nature does not require any assumptions about the functional form of the (unknown) baseline hazard.

The hazard function in a Cox model is given by:

$$h(t,x) = h_0(t)e^{\beta \cdot x} \tag{6.5}$$

The unobservable baseline hazard $h_0(t)$ is the risk to experience an event in the absence of any effects of the covariates. As usual x is a vector of explanatory variables, and β is a vector of the associated coefficients. A common approach is to interpret $e^{\hat{\beta}_x}$ as the ratio of hazards of two individuals differing by one unit in x.

The model makes a proportionality assumption for the hazards, which means that the *effect* of a covariate can not change over time. One should note that time-varying covariates and non-proportional hazards are two different extensions of the Cox model, even though both are deemed "extended" Cox models. However, both extensions, allowing for time varying *values* and/or time varying *effects* of covariates, can be incorporated in the Cox model. While it is straightforward to introduce a duration dependence for the Xs through episode splitting, it may sometimes be difficult to come up with the "correct" functional form for the duration dependence of the βs if one cannot come up with a theoretical explanation for the duration dependence. However, if the functional form is found the duration dependence is simply introduced as an interaction effect with the function of time.

However, according to Allison (2010, p. 422) the effects of violating the proportional hazard assumption are, depending on the research question, often not as grave. The violation of the assumption for a specific covariate would mean that the coefficient represents "some sort of average effect" over the period of observation.

An alternative approach, popular in applied work, to circumvent the problem of non-proportional hazards is the use of parametric survival models. On the downside the fully parametric survival models come with their own assumptions which are at least as restrictive as the proportional hazard assumption of the Cox regression model. They assume that the distribution of survival times is known, e.g. from previous research. The simplest form of hazard function is obtained for the so called exponential model, which only depends on the coefficients and covariate values. The more general Weibull model, allows for a parameter σ different from 1 on the log-exponential distribution of the error terms. Hence, the hazard function becomes:

$$h(t,x,\boldsymbol{\beta},\lambda) = \lambda\gamma t^{\lambda-1} e^{e-\lambda\beta_1 x} \tag{6.6}$$

$$\text{with} \qquad \gamma = e^{-\beta_0/\sigma} \tag{6.7}$$

$$\text{and} \qquad \lambda = 1/\sigma \tag{6.8}$$

The shape of the hazard function is usually reported by providing the *shape* (λ) and *scale* (γ) parameters, which are functions of σ, "the variance-like parameter on the log-time scale" (Hosmer, Lemeshow, and May 2008, p 261). The fully parametric nature has the advantage that the model can be estimated by full maximum likelihood, in contrast to the partial likelihood estimation of Cox models. Additionally, one can obtain fitted values to predict survival times.

In Table 6.2 coefficients for five differently specified survival models are presented. The risk to leave the industry in the following period is explained by being *male*, *startage* (age at the time of joining the industry), *MNC* (the maximum centrality value among the neighbors), a dummy to account for *disconnected*ness from the network, and own *centrality*. Column (1) to (3) are semi-parametric Cox regression models. Model (1) is the base specification, which is altered in Model (2) by stratification based on gender (i.e. allowing for men and women to differ in their baseline hazard), justified by the fact that the Kaplan-Meier graph (Figure 6.3) showed that the two groups (male and female) were affected differently. In (3) interaction terms with time (indicated by ·t) are introduced in order to reflect the issue of non-proportionality. The multiplication of the MNC value with the time (observation time lies in the interval [1,42]) causes the smaller size of the coefficients. Column (4) and (5) are fully parametric hazard models, where the functional form of survival times is assumed to follow a Weibull and exponential distribution respectively. In order to preserve comparability with the Cox models the parametric models are also presented in their proportional hazard notation, thus the signs of the coefficients retain their interpretation. Using a likelihood ratio

Table 6.2: Results (Survival Models)

	Cox			*parametric*	
	prop. hazards			*prop. hazards*	
		strat.	strat. & TT	Weibull	exponential
	(1)	(2)	(3)	(4)	(5)
male	−0.307***			−0.337***	−0.229***
	(0.034)			(0.034)	(0.034)
StartAge	0.014***	0.014***	0.025***	0.014***	0.013***
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)
MNC	−20.906***	−20.834***	−15.238***	−18.751***	−20.194***
	(1.458)	(1.459)	(2.153)	(1.427)	(1.442)
centrality	−1,730.909***	−1,728.950***	−2,117.128***	−1,646.382***	−1,608.111***
	(78.680)	(78.723)	(114.598)	(76.170)	(76.170)
disconnected	0.174***	0.162***	0.167***	0.203***	0.123**
	(0.063)	(0.063)	(0.063)	(0.062)	(0.062)
StartAge·t			−0.002***		
			(0.0005)		
MNC·t			−0.896***		
			(0.285)		
centrality·t			55.260***		
			(10.062)		
log(scale)				2.253***	
				(0.040)	
log(shape)				0.306***	
				(0.010)	
Observations	51,686	51,686	51,686	51,686	51,686
R^2	0.058	0.054	0.055		
Max. Possible R^2	0.789	0.764	0.764		
Log Likelihood	−38,610.690	−35,852.540	−35,829.740	−14,829.660	−15,227.250
LR Test	3,085.291***	2,857.632***	2,903.237***	2,965.14***	2,767.36***
	(df = 5)	(df = 4)	(df = 7)	(df = 5)	(df = 5)

Note: *p<0.1; **p<0.05; ***p<0.01

Standard errors in parenthesis

strat: model uses straticiation, allowing for the baseline hazard to differ for male female performers.

TT: interaction with time is included

test it can be shown that the Weibull model provides the better fit to the data than the exponential model.

The coefficients in a survival model represent the conditional change in the risk to experience an *event* (e.g. leaving the industry) in the following year. Both variables representing the network (centrality and MNC) are significant and lower the probability to experience an event (i.e. drop out of the industry) in the following period. While the centrality of the neighbor has a rather small effect, there is a large effect of own centrality on the survival. Being disconnected from the network increases the risk to experience an event. These results are consistent across the different specifications of the survival model, which gives strong support to the hypothesis that having worked with well-connected people and being central in the network helps ones own success. As expected from the descriptive work the risk is reduced for male performers and is increasing in *startage*. While the results appear robust across all specifications, the model explains only a fraction of the variation in the hazard rate. Potential further factors influencing entry an exit in the adult film industry include the social background including heritage, upbringing and outside options, for which information is scarce.

As the size of the effects appear to be similar across model specifications, the effect sizes are discussed exemplary for model 1. Figure 6.6 shows the relative hazards for the variables of interest. The top panel shows the relative hazard in comparison to a person with median centrality. To aid readability in the light of the skewed distribution only the interval between minimum and mean is shown. For a person being twice as central as the median performer the relative hazard is reduced by about 10%. This increase in centrality is rather small. Instead of 4×10^{-5} of all connections in the network running through the node, 8×10^{-5} of all connections would run through the node. Due to the skewed distribution the median and the mean lie rather far apart in distribution. Hence, the relative hazard at mean (0.0022) centrality is reduced by over 90%, compared to the median.

Considering the effect of MNC in the middle panel (showing the whole interval minimum to maximum), the effects are strong. Increasing ones own MNC by about 3 percentage points will reduce the hazard relative to the median to half. This is a substantial change considering that the maximum lies at 18,7%.

The bottom panel shows the effect of the age at which the performer joins the industry. The risk to leave the industry in the following year increases by 25% when the performer starts at 33 instead of 18.

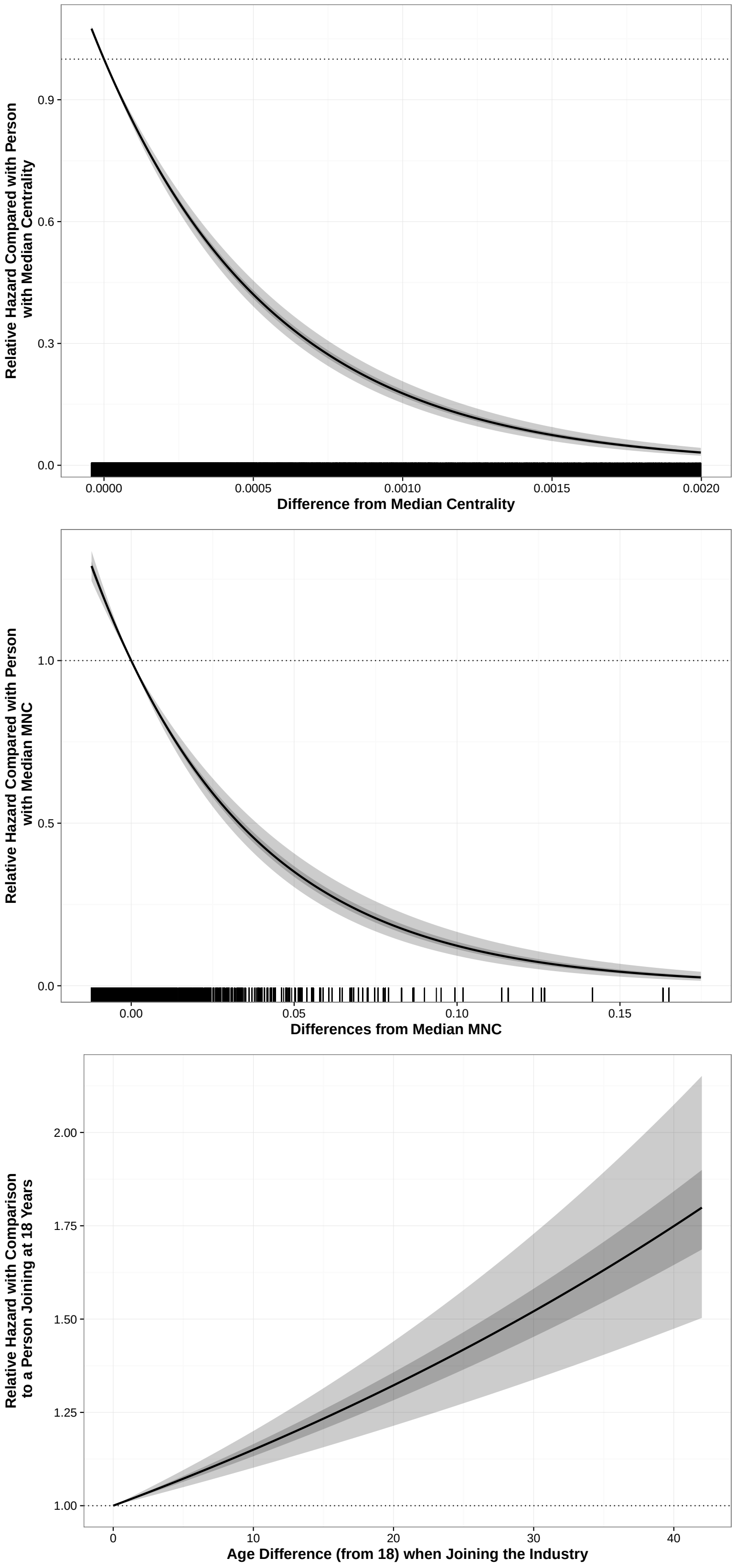
A potential critique concerns the question of endogeneity in the survival models employed. It is no guaranteed that the causality runs from centrality to the likelihood to leave the industry. Unfortunately, there is little one can do in terms of refined estimation techniques as there is a lack of established methods for using instrumental variables in duration models. Otherwise the position in the network could be instrumented by the film studios. However, the problem might not be as severe as it appears on first sight. Taking a closer look at the construction of the network one realizes that the network is not accumulated over time. The network at any given time consists of the collaboration in that year only. Additionally, there is a small time lag of one year between the possible event (in t+1) and the value of the covariate value (in t).

6.6 Discussion

The results found in the empirical analysis are rather strong compared to the empirical literature (Brown, Setren, and Topa 2016; Burks et al. 2015; Gorman and Marsden 2001). However, several studies show that, while networking does not necessarily lead to good jobs, it generally helps to find any job at all (Gorman and Marsden 2001, p. 485). Whether the eventual returns are positive or negative may depend on the specific characteristics and the "quality" of the position in question.

The evidence at hand supports the argument that social capital is indeed important for one's professional success. Both measures, own centrality and the centrality of the collaborators,

Figure 6.6: Effects of Coefficients



have a significant negative impact on the probability to leave the industry. The effect of own centrality is larger by several magnitudes than the effect of the position of the collaborator in the network. These findings hint at that it is easier to utilize resources from ones own social network rather than relying on a person in ones neighborhood, with whom one may only be loosely connected to, for the social network.

Obviously, the measure of success used in the paper at hand differs from those used in the literature, where the measures are usually very direct, e.g. pay rises and promotions. An argument brought up in defense of the empirical literature in Granovetter (1995, pp. 153–154) is that a good initial social network yields an even larger number of ties in the long-term. Thus, the use of social ties today yields even more long-term benefits rather than the immediate outcome. This argument is a strong point in favor of survival analysis, which looks at the lifespan of activity rather than immediate pay-offs.

The adult film industry might be different from other industries. Long contracts are rather unlikely and the data shows that performers appear in films produced by different studios over their career. Hence, one should ask the question whether the results hold any external validity. I argue that similar conditions also exists in other industries which are characterized by a freelance structure like the "ordinary" film industry, scientific collaborations and (to some degree) the music industry.

From a purely econometric perspective, there is no strong argument in favor of a causal relationship. It may well be that one's own centrality, is influenced by survival in the industry. The argument becomes less plausible when considering the centrality of the most prominent "partner" a performer appeared with in a movie: Does someone appear in films with more prominent people the longer he is in the industry? I would argue that in pornography this is not necessarily the case, as there is always a demand for new "faces" (Slattery 2001, p. 243).

6.7 Conclusion

In this paper I have presented empirical evidence on the relationship between networking and professional success. The study is based on a novel dataset on the adult film industry. To the best of my knowledge, it is also the first work that combines survival analysis with methods from social network analysis. It is a drawback of survival analysis that it is difficult to deal with endogeneity, as no established IV-method exists for survival models. Nonetheless, there is consistent evidence, robust over all specifications, that adult performers remain longer in the industry, thus are more successful if they are well connected by collaboration with central actors in the industry and/or are very central themselves. As expected the age at joining the industry has a negative effect on the time active in the industry.

While there is strong support in favor of a positive relationship of networking and success in the analysis, one might expect the adult film industry to be substantially different from other industries. Consequently, one should exercise caution when discussing the external validity of the results. While there is a large body of literature with ambiguous results on various industries and obtained with different methods. It might be wise to consider applying the same method to other freelance industries (e.g. academia), to confirm the findings or determine how the adult film industry differs from other fields.

Bibliography

Abbasi, A., J. Altmann, and L. Hossain (2011). "Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures". In: *Journal of Informetrics* 5.4, pp. 594–607.

Ackman, D. (2001). "How Big Is Porn?" In: *Forbes*. `http://www.forbes.com/2001/05/25/0524porn.html` accessed 22.09.2015.

Ahmed, A., V. Batagelj, X. Fu, S.-H. Hong, D. Merrick, and A. Mrvar (2007). "Visualisation and analysis of the Internet movie database". In: *Visualization, 2007. APVIS'07. 2007 6th International Asia-Pacific Symposium on*. IEEE, pp. 17–24.

Allison, P. D. (2010). "The Reviewer's Guide to Quantitative Methods in the Social Sciences". In: ed. by G. R. Hancock and R. O. Mueller. Routledge. Chap. Survival Analysis, pp. 413–424.

Bordons, M., J. Aparicio, B. González-Albo, and A. A. Díaz-Faes (2015). "The relationship between the research performance of scientists and their position in co-authorship networks in three fields". In: *Journal of Informetrics* 9.1, pp. 135–144.

Bourdieu, P. (1983). "Ökonomisches Kapital, kulturelles Kapital, soziales Kapital". In: Kreckel, R. *Soziale Ungleichheiten*. Vol. 2. Soziale Welt : Sonderband. Göttingen: Schwartz, pp. 183–198.

Brady, G. (2015). "Network Social Capital and Labour Market Outcomes: Evidence For Ireland". In: *The Economic and Social Review* 46.2, pp. 163–195.

Brown, M., E. Setren, and G. Topa (2016). "Do Informal Referrals Lead to Better Matches? Evidence from a Firm's Employee Referral System". In: *Journal of Labor Economics* 34.1, pp. 161–209.

Buenstorf, G. (2007). “Evolution on the Shoulders of Giants: Entrepreneurship and Firm Survival in the German Laser Industry”. In: *Review of Industrial Organization* 30.3, pp. 179–202.

Burks, S. V., B. Cowgill, M. Hoffman, and M. Housman (2015). “The Value of Hiring through Employee Referrals”. In: *The Quarterly Journal of Economics* 130.2, pp. 805–839.

Calvó-Armengol, A. and M. O. Jackson (2004). “The Effects of Social Networks on Employment and Inequality”. In: *The American Economic Review* 94.3, pp. 426–454.

Card, D., R. Chetty, and A. Weber (2007). “The Spike at Benefit Exhaustion: Leaving the Unemployment System or Starting a New Job?” In: *American Economic Review* 97.2, pp. 113–118.

Cimenler, O., K. A. Reeves, and J. Skvoretz (2014). “A regression analysis of researchers’ social network metrics on their citation performance in a college of engineering”. In: *Journal of Informetrics* 8.3, pp. 667–682.

Coopersmith, J. (1998). “Pornography, Technology and Progress”. In: *Icon* 4, pp. 94–125.

Danta, D. (2009). “Ambiguous Landscapes of the San Pornando Valley”. In: *Yearbook of the Association of Pacific Coast Geographers* 71, pp. 15–30.

Disney, R., J. Haskel, and Y. Heden (2003). “Entry, Exit and Establishment Survival in UK Manufacturing”. In: *Journal of Industrial Economics* 51.1, pp. 91–112.

Edelman, B. (2009). “Markets: Red Light States: Who Buys Online Adult Entertainment?” In: *Journal of Economic Perspectives* 23.1, pp. 209–20.

Edwards, A. C. and M. Ureta (2003). “International migration, remittances, and schooling: evidence from El Salvador”. In: *Journal of Development Economics* 72.2, pp. 429–461.

Egghe, L. (2006). “Theory and practise of the g-index”. In: *Scientometrics* 69.1, pp. 131–152.

Faulkner, R. R. and A. B. Anderson (1987). “Short-Term Projects and Emergent Careers: Evidence from Hollywood”. In: *American Journal of Sociology* 92.4, pp. 879–909.

Fernandez, R. M., E. J. Castilla, and P. Moore (2000). “Social capital at work: Networks and employment at a phone center”. In: *American Journal of Sociology* 105.5, pp. 1288–1356.

Freeman, L. C. (1977). “A Set of Measures of Centrality Based on Betweenness”. In: *Sociometry* 40.1, pp. 35–41.

Freeman, L. C. (1978). “Centrality in social networks conceptual clarification”. In: *Social Networks* 1.3, pp. 215–239.

Goffman, C. (1969). “And What Is Your Erdös Number?” In: *The American Mathematical Monthly* 76.7, p. 791.

Gorman, E. H. and P. V. Marsden (2001). “Social Networks, Job Changes, and Recruitment.” In: *Sourcebook on Labor Markets: Evolving Structures and Processes*. Ed. by I. Berg and A. L. Kalleberg. New York: Kluwer Academic/Plenum, pp. 467–502.

Granovetter, M. (1973). “The Strength of Weak Ties”. In: *American Journal of Sociology* 78 (6).

Granovetter, M. (1983). “The strength of weak ties: A network theory revisited”. In: *Sociological theory* 1.1, pp. 201–233.

Granovetter, M. (1995). “Getting a job: A study of contacts and careers”. In: University of Chicago Press. Chap. Afterword 1994: Reconsiderations and a New Agenda, pp. 139–182.

Gury, N. (2011). “Dropping out of higher education in France: a micro-economic approach using survival analysis”. In: *Education Economics* 19.1, pp. 51–64.

Herr, B. W., W. Ke, E. Hardy, and K. Borner (2007). “Movies and actors: Mapping the internet movie database”. In: IEEE, pp. 465–469.

Hosmer, D. W., S. Lemeshow, and S. May (2008). “Parametric Regression Models”. In: *Applied Survival Analysis*. John Wiley & Sons, Inc. Chap. 8, pp. 244–285.

Hunt, J. (1995). “The Effect of Unemployment Compensation on Unemployment Duration in Germany”. English. In: *Journal of Labor Economics* 13.1, pp. 88–120.

Ioannides, Y. M. and L. D. Loury (2004). “Job Information Networks, Neighborhood Effects, and Inequality”. In: *Journal of Economic Literature* 42.4, pp. 1056–1093.

Ioannides, Y. M. and A. R. Soetevent (2006). “Wages and Employment in a Random Social Network with Arbitrary Degree Distribution”. In: *American Economic Review* 96.2, pp. 270–274.

Kaplan, E. L. and P. Meier (1958). “Nonparametric Estimation from Incomplete Observations”. In: *Journal of the American Statistical Association* 53.282, pp. 457–481.

Kiefer, N. M. (1988). “Economic Duration Data and Hazard Functions”. In: *Journal of Economic Literature* 26.2, pp. 646–79.

Kostoff, R. N. (1998). “The use and misuse of citation analysis in research evaluation”. In: *Scientometrics* 43.1, pp. 27–43.

Kuhlenkasper, T. and G. Kauermann (2010). “Duration of maternity leave in Germany: A case study of nonparametric hazard models and penalized splines”. In: *Labour Economics* 17.3, pp. 466–473.

Lin, N. (1999a). “Building a network theory of social capital”. In: *Connections* 22.1, pp. 28–51.

Lin, N. (1999b). “Social Networks and Status Attainment”. In: *Annual Review of Sociology* 25, pp. 467–487.

Mata, J. and P. Portugal (1994). “Life Duration of New Firms”. In: *Journal of Industrial Economics* 42.3, pp. 227–45.

Millward, J. (2013). *Deep Inside - A Study of 10,000 Porn Stars and Their Careers*. English. URL: http://jonmillward.com/blog/studies/deep-inside-a-study-of-10000-porn-stars/.

Montgomery, J. D. (1991). “Social Networks and Labor-Market Outcomes: Toward an Economic Analysis”. In: *American Economic Review* 81.5, pp. 1407–18.

Nadler, C. (2016). “Networked Inequality: Evidence from Freelancers”. Job Market Paper, UC Berkeley.

Rich, F. (2001). “Naked Capitalists”. In: *The New York Times*. http://www.nytimes.com/2001/05/20/magazine/naked-capitalists.html retrieved 22. September 2015.

Silverstein, J. (2006). "Is Porn a Growing or Shrinking Business?" In: *ABC News*. http://abcnews.go.com/Technology/story?id=1522119 accessed 23.09.2015.

Slattery, T. (2001). *Immodest Proposals: Through the Pornographic Looking Glass*. iUniverse.

Wolff, H.-G. and K. Moser (2009). "Effects of Networking on Career Success: A Longitudinal Study." In: *Journal of Applied Psychology* 94.1, pp. 196–206.

6.A Technical Note

The computations in the document where conducted using R. The social network analysis was conducted using the package *igraph*[6]. The survival analysis relies on the packages *survival*[7] and *eha*[8]. The pretty tables were created using *stargazer*[9] and *xtable*. The relative hazards are illustrated using simPH[10]

[6] igraph Team (2015), Network Analysis and Visualization. R package version 1.0.1 http://irgraph.org.

[7] Therneau,Terry M (2015), survival: Survival Analysis. R package Version 2.38-1.

[8] Broström,Göran (2015). eha: Event History Analysis. R package version 2.4-3.

[9] Hlavac, Marek (2014). stargazer: LaTeX code and ASCII text for well-formatted regression and summary statistics tables. R package version 5.1. http://CRAN.R-project.org/package=stargazer.

[10] Gandrud, Christopher. 2015. simPH: An R Package for Illustrating Estimates from Cox Proportional Hazard Models Including for Interactive and Nonlinear Effects. Journal of Statistical Software. 65(3)1-20.

7 Concluding Remarks

The previous five chapters constitute my research on the usage of unconventional datasets in order to address questions arising in economic research, which are difficult or expensive to answer when relying on official statistics or the collection of own survey data.

One cross-cutting issue of the individual papers is the complexity of the data used. In Chapter 2 the raw data was immensely large in size, as a consequence I had to resort to sampling, aggregation and outside-sorting. After this preparatory step the aggregated data could be analyzed using established methods in order to replicate results from the literature. In Chapter 4 the scanning and OCR of 60 years of a scientific journal led to long processing times and manual cleaning. Afterwards, Latent Dirichlet Allocation was used to quantify the content of the whole corpus of articles. The application of LDA yields topic weights which can subsequently be employed as a variable in a conventional regression analysis. Analyzing the network data in Chapter 6 the data is not very large in terms of size, but it is inherently complex due to the high number of linkages between the different nodes in the network.

In terms of methodology my research pioneered the combination of methods which – to the best of my knowledge – have not been used jointly so far. This includes the combination of social network analysis as well as the application of simultaneous equations models to multiple imputed datasets. Moreover, topic models have seldom been used in conjunction with short tweets. Thus, a lot of fine tuning of parameters and procedures was necessary.

It turns out that the analysis of unconventional data offers new insights in old debates and therefore holds a large potential for the future, requiring creativity and a good understanding of empirical methods.

Affidavit

Erklärung gemäß §10 Absatz 7 der Promotionsordnung

Ich erkläre hiermit, dass ich die vorgelegten und nachfolgend aufgelisteten Aufsätze selbstständig und nur mit den Hilfen angefertigt habe, die im jeweiligen Aufsatz angegeben oder zusätzlich in der nachfolgenden Liste aufgeführt sind. In der Zusammenarbeit mit den angeführten Koautoren war ich mindestens anteilig beteiligt. Bei den von mir durchgeführten und in den Aufsätzen erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis niedergelegt sind, eingehalten.

Gießen den 7. Februar 2017

Jochen Lüdering

Liste der Aufsätze

Lüdering, J. (2015). *The measurement of internet availability and quality in the context of the discussion on digital divide*. Discussion Papers 65 [rev.] Justus-Liebig-University Gießen, Center for international Development und Environmental Research (ZEU).

Lüdering, J. (2016a). *Low Latency Internet and Economic Growth: A Simultaneous Approach*. MAGKS - Joint Discussion Paper Series in Economics 34/2016. Philipps-Universität Marburg.

Lüdering, J. und P. Winker (2016). "Forward or backward looking? The economic discourse and the observed reality". In: *Jahrbücher für Nationalökonomie und Statistik* 236.4, S. 483–516.

Lüdering, J. und P. Tillmann (2016). *Monetary Policy on Twitter and its effect on asset prices: Evidence from computational text analysis*. MAGKS - Joint Discussion Paper Series in Economics 12/2016. Philipps-Universität Marburg.

Lüdering, J. (2016b). *Standing and Survival in the Adult Film Industry*. MAGKS - Joint Discussion Paper Series in Economics 26/2016. Philipps-Universität Marburg.