

1 Introduction

In Germany, the debate about the role of public media and their mission in a rapidly changing media world is a frequently discussed topic. By shifting media content to the Internet, the dual system, which has been shaping the German television and radio landscape since the introduction of private broadcasting in the early 1980s, is facing a radical change. Since 2000, public broadcasting has expanded its range of services, particularly in the digital media sector. There is the concern of commercial providers who are registering an ever-increasing number of digital services financed by fees. They complain about a bias in competition caused by public media offering online text-content because of their fee financing. The advertising-financed business model of the media houses is based on the premise that users visit their websites in order to achieve high advertising revenues. However, if the right to produce digital media content were to be withdrawn from the public media outlets, they would not only be deprived of the possibility of improved information provision. They also threatened to lose their competitiveness with commercial media.

One fundamental question in this debate is whether the offer of public news on the internet is justified. They should only occur where there are clear deficits in private sector supply. A frequently cited argument is that only the public media make it possible to provide information that is free of self-interest. Due to their public mandate and financing, they can afford what private providers cannot or only to a limited extent because of their economic dependency: a journalistic and editorial self-observation of society in the public interest. Due to their constitutional determination, they are obliged to the diversity and representation of the political and social spectrum of opinion in its entire breadth, including minority positions.

On the basis of these justifications, this paper examines whether imperfections in the market of online news exist. In other words, whether socially desired content that meets the (social) needs of society - and is therefore politically desirable - is not, or very poorly provided by the market. In order to examine this, the online news content of public media is compared with the supply of private news providers. More specifically, I use a data set of German online news articles about domestic politics dated from 01.06.2017 to 31.12.2017¹ from six German content providers. In the event that there is no market failure in terms of information provision (the prevalence of topics do not differ), the existence of public media could still be justified by the fact that socially relevant contents of private providers are systematically

¹German federal elections took place on 24th of September 2017.

distorted and the existence of public offerings ensures the diversity of information and opinions transmitted. In order to examine this bias, I analyze how the topics are discussed differently in public and private news.

The research strategy is as follows:

1. **Discovering Topics (Section 4)** To discover the latent topics in the corpus, the structural topic modeling (STM) developed by M. E. Roberts, B. M. Stewart, and E. M. Airoldi (2016) is applied. The STM is an unsupervised machine learning approach that models topics as multinomial distributions of words and documents as multinomial distributions of topics, allowing to incorporate external variables that effect both, topical content and topical prevalence. I will include the business model (public or private) as a control for both the topical content and the topical prevalence. Additional, the month an article was published is included as a control for the topical prevalence. The result of the generative process of STM are the posterior distributions for topic prevalence and topical content, which can then be used to measure the similarity between private and public news providers in the further course of the paper: (1) β is a K -by- V matrix (where K = number of topics and V = vocabulary), where the entry β_{kvc} can be interpreted as the probability of observing the v -th word in topic k for the covariate level c . (2) θ is a D -by- V matrix of the document-topic distributions, where the entry θ_{dk} can be interpreted as the proportion of words in document d which arise from topic k , or rather as the probability that document d deals about topic k .
2. **Measuring Similarity of news content** Different approaches are used to examine similarities and/or differences of content between the news provider:
 - (a) **Differences in topic prevalence:** As I allow prevalence of topics to evolve over time (by month) and vary across business models, the posterior topic-probability θ_{dk} of document can be used to estimate the conditional expectation of topic prevalence for given document characteristics.
 - (b) **Topic correlation** To get an idea of how topics are linked and framed differently across news wires, the correlation of the posterior probability of topics θ_{dk} is calculated.
 - (c) **Differences in word-topic distributions** As each row of β_{kvc} corresponds to the word-distribution of a respective topic for the covariate level c , it can be used to calculate the similarity of the

word-topic distribution between the news provider using established similarity measures.

Approaches 2a and 2b have been used in M. E. Roberts, B. M. Stewart, and E. M. Airoldi (2016), among others, to examine differences in the content of news providers. However, I extend the analysis by calculating the similarity of the word-topic distribution between the news provider, to identify which topics are discussed similar or differently (2c). In addition, this approach is applied to a new dataset covering recent political trends in Germany.

The remaining course of the paper is as follows: In Section 2 I briefly discuss the characteristics of the market for online news in Germany. The data used to conduct the model is described in Section 3. Section 4 explains the generative process of the structural topic model as well as the selected parameters to run the model. The results are evaluated in Section 5.

2 The online news market

The market for media content in Germany is characterized by the coexistence of public and private broadcasters. By shifting media content to the Internet, the dual system, which has been shaping the German television and radio landscape since the introduction of private broadcasting in the early 1980s, is facing a radical change. Since 2000, public broadcasting has expanded its range of services, particularly in the digital media sector. In 2017, 22 own websites and 100 apps were operated by public broadcasters on which they offer their content. As a result, public broadcasting no longer only competes with private television and radio stations, but also enters the market for online news. In the following, the characteristics of the market for online news in Germany are briefly described.

Private media outlets naturally appear as two-sided platforms, that allow interaction between two categories of consumers: audiences and advertisers. As the demand on both consumer-sides are linked via indirect network externalities, the market in which media outlets operate are referred to as two-sided or multi-sided markets. The theoretical literature on two-sided markets originates from the analysis of credit card markets (Rochet and Tirole, 2003) and was later transferred to the concept of other industries, such as dating agencies, real estate agents, and internet “business-to-business” websites (Caillaud and Jullien, 2003). The basic concept of two-sided markets was already discussed decades ago in several economic studies, especially on media markets (Corden, 1952), (Gustafsson, 1978), (Blair and Romano,

1993). However, comprehensive analyses have only been carried out in the last ten years, starting with the works of Rochet and Tirole (2003), Evans (2003) and Armstrong (2006).

Advertising-supported media such as online newspapers are typical examples of two-sided markets where the newspaper can be conceived as platforms that allow interaction between audiences ("eyeballs") and advertisers. The newspaper creates (or buys) content to attract viewers which in turn attract advertisers who pay for readers' attention (Evans and Schmalensee, 2005). The size and characteristics of the audience has a positive effect on the advertisers' willingness to pay, as advertisements are typically sold based on cost per viewer, often expressed in terms of the cost of reaching a thousand viewers (CPM). Advertising can also have an effect on the recipients, which can be either negative or positive, depending on the quality of the advertising. Based on the strength of the indirect network effects, private publishers maximize their revenue by balancing the demand from advertisers and subscribers using different business models (Evans, 2008). Many traditional newspapers follow the subscription/advertising model, where the publisher charges both market sides: The audience pays a fee to obtain access to the content, and advertisers pay to obtain access to the viewers. Many online news agencies provide part of their editorial content for free and hide another, more exclusive part behind a paywall. However, since the Internet has considerably simplified the possibilities for obtaining information and thus reduced the marginal utility of content, such a business model can only be efficient if the content is very exclusive. As a result, many publishers rely on a free-media model, in which the publishers do not charge viewers for access to the media at all, in order to attract as many eyeballs as possible to their platform, and thus, exploit the indirect network effects on the advertising site. In fact, most advertising-financed online magazines earn their gross margin from advertising (Evans and Schmalensee, 2005). In order to maximize their profits, these companies have an interest in attracting as many readers as possible. In addition to the quantity of the audience, the demographic characteristics of recipients also have an influence on the willingness to pay on the advertiser site. Online advertising makes it possible to target ads to particular consumers in real time.

The two-sided market structure of the private news market results in news platforms striving to choose their content in such a way that its reach is as large as possible in order to maximize profits from advertising revenues. Steiner (1952) concluded, that profit-maximizing media owners may choose to offer the same content, i.e. content aligned with the tastes of the majority. (Gabszewicz, Laussel, and Sonnac, 2001) study the problem of diversity of the political content of newspapers. They find that the maximum differenti-

ation only prevails if the readers sufficiently value the political differentiation between the newspapers the advertising market is small enough. On the other hand, advertising may also have a positive impact on the media, as it enables publishers to report independently of political parties. Ellman and Germano (2009) analyze a theoretical framework of a two-sided market for newspaper, where readers value accuracy and advertisers value advert-receptive readers. They found that advertising increases the intensity of competition for readers and therefore raises accuracy of media coverage. However, in the monopolistic case, newspapers under-report news that sufficiently reduces advertiser profits.

Given the crucial role of the media in shaping opinion and promoting democracy, pluralism of opinion and accuracy of information is a major concern of public authorities. Public broadcasting in Germany originated in the post-war period and has always had the task of providing the entire population with independent media. This media offer is intended to guarantee diversity of opinion within the media landscape and to be economically and politically independent. The former is given by the fact that the public media are financed by compulsory fees. To take into account the distinct nature of digital media, the Interstate Broadcasting Agreement (Rundfunkstaatsvertrag) also regulates the scope for action of online services offered by public service broadcasting since 2007. Accordingly, public media are not allowed to distribute purchased content and must - depending on the category of content - set a time limit on its accessibility. In addition, there is a strict advertising ban and prohibition of regional reporting.

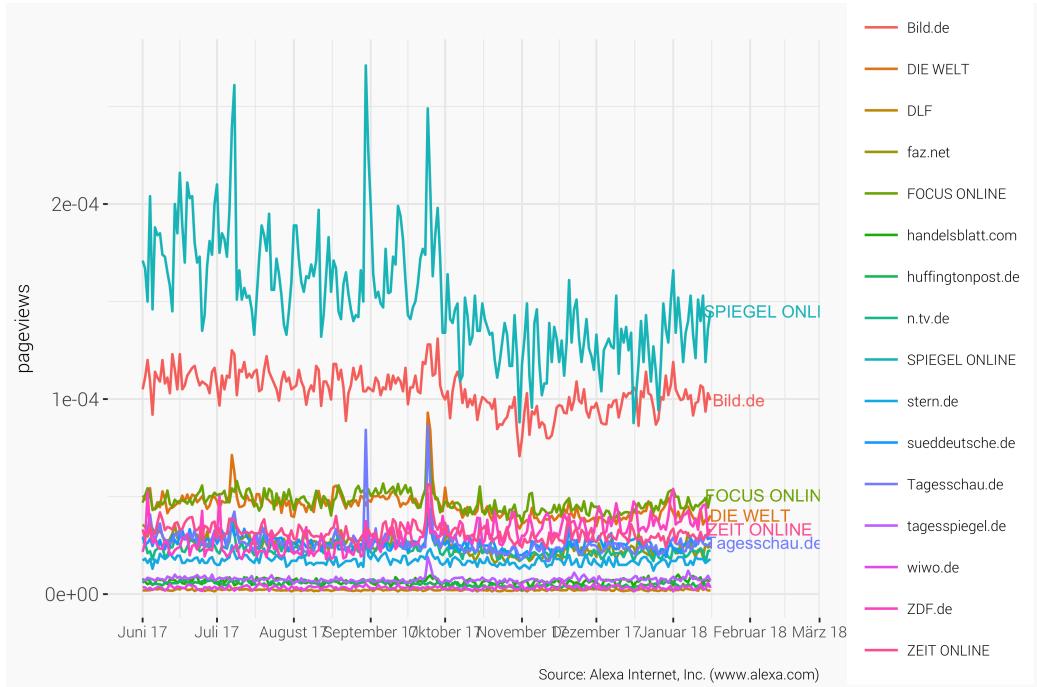
However, given today's media landscape, the role of public media provider and its mandate is questioned. The basic justification for a public media offer can be roughly divided into two categories: On the one hand, possible market failure and on the other hand, ensuring diversity. The former involves the pursuit of non-economic (e. g. democratic, social and cultural) objectives, as well as the provision of information that would not be made available by the market due to a lack of willingness to pay. Ensuring the diversity of media content is intended to counteract possible media bias.

As stated above, this paper examines whether imperfections in the market of online news exist, by comparing the topic prevalence between public and private online news with the supply of private news providers. In the event that there is no market failure in terms of information provision, the existence of public media could still be justified by the fact that socially relevant contents of private providers are systematically distorted and the existence of public offerings ensures the diversity of information and opinions transmitted. In order to examine this bias, the divergence of word-topic probabilities is calculated to analyze how the topics are discussed differently.

3 Dataset and data preparation

Figure 1 shows the largest providers of online news in terms of daily pageviews.² It is striking that SPIEGEL ONLINE has the greatest reach over the entire course of time followed by Bild.de. Public websites like Deutschlandfunk.de or Tagesschau.de have a rather small share in that market.

Figure 1: Percentage of daily pageviews



I conduct the estimation on a sample of 9,393 online news articles from six news provider about domestic politics³, of which only Tagesschau.de belongs to the group of public provider. The reason for this is that the content structure of Tagesschau.de is most similar to that of the private providers. ZDF.de offers predominantly video content and DLF website mainly offers audio content in the form of interviews, which makes it hard to include it in the model. However, as only a part of the private suppliers are included to maintain the proportions between private and public market participants

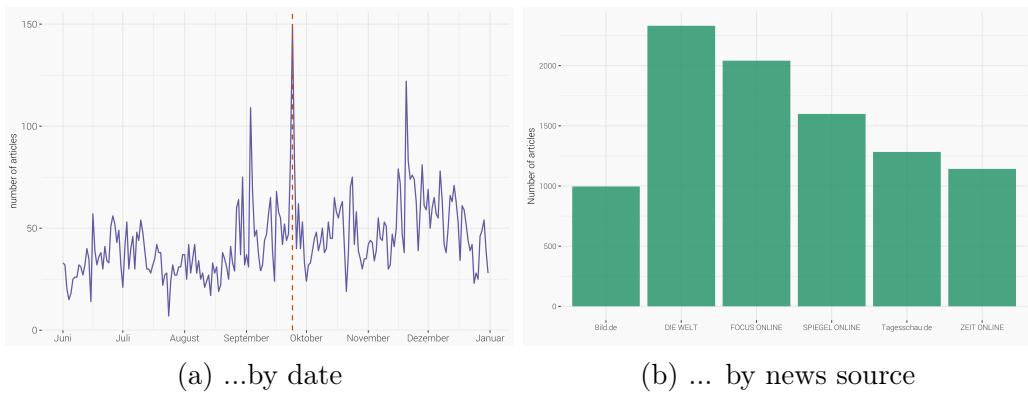
²The figure shows the estimated percentage of daily pageviews on the internet that occurred on a specific website.

³Bild.de, DIE WELT, FOCUS ONLINE, SPIEGEL ONLINE, ZEIT ONLINE, Tagesschau.de

are maintained, thus minimizing possible bias. The articles are dated from 01.06.2017 to 31.12.2017.⁴ I first extract all online articles using the the Webhose.io API.⁵ Then all articles from the section "domestic policy" are filtered by checking the URL structure.

Figure 2a shows the distribution of the number of articles from the respective news sources by date. There is a high peak around the federal elections on September, 24th.

Figure 2: Article distribution...



A central task in text mining is to extract low-dimensional information from documents that are high-dimensional by nature (Bholat et al., 2015). This is related to the task of reducing the number of unique language elements in order to reduce the dimensionality of data (to avoid unnecessary computational complexity and overfitting) while at the same time keeping those words that reflect the content of a document. Any useful representation of text will throw away some information, the trick is to include the relevant information for our needs, and exclude the extraneous information. A common strategy to use text as data and reduce the dimensionality, is to pre-process the text by imposing some preliminary restrictions (stop-word removal, tokenization) based on the nature of the data (twitter text, newspaper articles, speeches, etc.) to reduce the number of language elements (Gentzkow, Kelly, and Taddy, 2017). Intuitively the term frequency (tf) of a word is a measure of how important that word may be. There are words in a document, however, that occur many times but may not be important like

⁴German federal elections took place on 24th of September 2017.

⁵For more information see <https://docs.webhose.io/v1.0/docs/getting-started>. The scraping code was written in Python and can be made available on request.

articles, conjunctions, and so on. These terms, often called "stop words", are important to the grammatical structure of a text, but typically don't add any additional meaning and can therefore be neglected. We use a pre-defined stop word list from the Snowball stemmer project⁶ together with a customized list of stop-words that are redundant superfluous or distorting. We also remove punctuation character (e.g. ., , !, ?, etc.) and all numbers from our corpus. After completing this steps we were left with 55.235 unique terms in our vocabulary.

After pre-processing, each document d is a finite list of terms. Each unique term in the corpus is indexed by some $v \in \{1, \dots, V\}$ where V is the number of unique terms. For each document $d \in \{1, \dots, D\}$ we compute the number of occurrences of term v in document d to obtain the count $x_{d,v}$. The $D \times V$ matrix \mathbf{X} of all such counts is called the document-term matrix. This representation is often referred to as the bag of words model, since the order in which words are used within a document is completely disregarded.

4 The structural topic model

The structural topic model (STM) developed by M. E. Roberts, B. M. Stewart, and E. M. Airoldi (2016) allows to incorporate document specific covariates (e.g. the author or date of a document). STM is a recent extension of the standard topic modeling technique, labeled as "latent Dirichlet allocation" (LDA), which refers to the Bayesian model in Blei, Ng, and Jordan (2003) that treats each word in a topic and each topic in a document as generated from a Dirichlet - distributed prior.⁷ Topic models formalize the idea that documents are formed by hidden variables (topics) that generate correlations among observed terms. Since its introduction into text analysis, LDA has become hugely popular and especially useful in political science.⁸ Wiedmann (2016) uses topic model methods on large amounts of news articles from two german newspapers published between 1959 and 2011, to reveal how democratic demarcation was performed in Germany over the past six decades. **paul_cross-collection_2017** compares editorial differences between media sources, using cross-collection latent Dirichlet allocation (ccLDA), an LDA-based approach that incorporates differences in document metadata.

⁶http://svn.tartarus.org/snowball/trunk/website/algorithms/*/stop.txt

⁷See also Griffiths and Steyvers (2002), Griffiths and Steyvers (2004) and Hofmann (1999). Pritchard, Stephens, and Donnelly (2000) introduced the same model in genetics for factorizing gene expression as a function of latent populations.

⁸see Blei (2012), Grimmer and B. Stewart (2013) and Wiedmann (2016) for an overview in social science and Gentzkow, Kelly, and Taddy (2017) give an overview of text mining applications in economics.

They use a dataset of 623 news articles from August 2008 from two American media outlets - msnbc.com and foxnews.com - to compare how they discuss topics. Reviewing the top words of the word-topic distribution, they find some content differences between the two.

STM has been applied to multiple academic fields: M. E. Roberts, B. M. Stewart, Tingley, et al. (2014) uses STM to analyse open-ended responses from surveys and experiments, Farrell (2016) applies the model to scientific texts on climate change, revealing links between corporate funding and the framing of scientific studies. Mishler et al. (2015) show that "STM can be used to detect significant events such as the downing of Malaysia Air Flight 17" when applied to twitter data. Another study shows how STM can be used to explore the main international development topics of countries' annual statements in the UN General Debate and examine the country-specific drivers of international development rhetoric (Baterno, Dasandi, and Mikhaylov, 2017). Mueller and Rauh (2016) use newspaper text to predict armed conflicts in different regions. They use the estimated topic shares in linear fixed effects regression to forecast conflict out-of-sample. M. Roberts, B. Stewart, and Tingley (2016a) use STM to examine the role of partisanship in topical coverage using a corpus of 13,246 posts that were written for 6 political blogs during the course of the 2008 U.S. presidential election. With the aim of revealing the effect of partisan membership on topic prevalence, each blog is assigned to be either liberal or conservative. To explore the differences between the two, they look at the expected proportion of topic and examine the posts most associated with a respective topic. This approach is similar to M. E. Roberts, B. M. Stewart, and E. M. Airoldi (2016). They also use different measures of distance between the topic-word distributions of the same topic within different models. In section 5.3 a similar approach is applied to measure similarity between the same topic for different covariate levels.

4.1 Generative Process of STM

As mentioned above, the STM allows to incorporate observed document metadata which is able to affect both topical prevalence and topical content. The following description of the generative model - the process of filling a word-position in a document - of the STM is based on M. Roberts, B. Stewart, Tingley, and E. Airoldi (2013) and M. Roberts, B. Stewart, and Tingley (2016b). For each document d and a given number of topics K , a document-specific topic-prevalence vector $d(\boldsymbol{\theta}_d)$ is drawn from a logistic-normal distribution, where the parameters are a function of the covariate

values:

$$\boldsymbol{\theta}_d | \mathbf{x}_{d\gamma}, \boldsymbol{\Sigma} \sim \text{LogisticNormal}(\mu = \mathbf{x}_{d\gamma} \boldsymbol{\Sigma}). \quad (1)$$

$\mathbf{x}_{d\gamma}$ lists the values of all metadata covariates for document d , where γ relates these covariate values to the topic-prevalence. The structure of $\boldsymbol{\Sigma}$ implies the possibility of correlations across documents in the topic-prevalence vector.

According to $\boldsymbol{\theta}$, a specific topic z_{dn} is assigned for the n^{th} word-position in the document through the process:

$$z_{dn} | \boldsymbol{\theta}_d \sim \text{Multinomial}(\boldsymbol{\theta}_d). \quad (2)$$

Conditional in the topic chosen, a specific word, w_{dn} , is chosen from the overall corpus vocabulary V , using the following process:

$$w_{dn} | z_{dn}, \beta_{dkv} \sim \text{Multinomial}(\beta_{dk1}, \dots, \beta_{dkV}), \quad (3)$$

where the word probability β_{dkv} is parameterized in terms of log-transformed rate deviations from the rates of a corpus-wide background distribution m_v . The log-transformed rate deviations can then be specified by a collection of parameters $\{\kappa\}$, where $\kappa^{(t)}$ is a K -by- V matrix containing the log-transformed rate deviations for each topic k and term v , over the baseline log-transformed rate for term v . This matrix is the same for all A levels of covariates. To put it differently, $\kappa^{(t)}$ indicates the importance of the term v given topic k regardless of the covariates. Similarly, $\kappa^{(c)}$ is a A -by- V matrix, indicating the importance of the term v given the covariate level c regardless of the topic. Finally, $\kappa^{(i)}$ is a A -by- K -by- V matrix, collecting the covariate-topic effects:

$$\beta_{dkv} | z_{dn} = \frac{\exp(m_v + \kappa_{kv}^{(t)}, \kappa_{y_d v}^{(c)} + \kappa_{y_d k v}^{(i)})}{\sum_v \exp(m_v + \kappa_{kv}^{(t)}, \kappa_{y_d v}^{(c)} + \kappa_{y_d k v}^{(i)})}. \quad (4)$$

The STM maximizes the posterior likelihood that the observed data were generated by the above data-generating process using an iterative approximation-based variational expectation-maximization algorithm⁹ available in R's `stm` package (M. Roberts, B. Stewart, and Tingley, 2016b). The process gives us two posterior distribution parameter: (1) β is a K -by- V matrix (where K = number of topics and V = vocabulary), where the entry β_{kvc} can be interpreted as the probability of observing the v -th word in topic k for the

⁹A technical description of this maximization process can be found in M. E. Roberts, B. M. Stewart, and E. M. Airoldi (2016)

covariate level c . (2) θ is a D -by- V matrix of the document-topic distributions, where the entry θ_{dk} can be interpreted as the proportion of words in document d which arise from topic k , or rather as the probability that document d deals about topic k . These probability distributions are used to compare the content of public and private news providers in section 5.

4.2 Model and parameter selection

Inference of mixed-membership models, such as the one applied in this paper, has been a thread of research in applied statistics in the past few years (Blei, Ng, and Jordan, 2003) (Erosheva, Fienberg, and Lafferty, 2004) (Braun and McAuliffe, 2010). Topic models are usually imprecise as the function to be optimized has multiple modes, such that the model results can be sensitive to the starting values. Since an ex ante valuation of a model is hardly possible, I compute a variety of different models and compare their posterior probability. This enables me to check how results vary for different model solution (M. Roberts, B. Stewart, and Tingley, 2016a). I then cross-checked some subset of assigned topic distributions to evaluate whether the estimates align with the concept of interest (Gentzkow, Kelly, and Taddy, 2017). These manual audits are applied together with numeric optimization based on the topic coherence measure suggested by Mimno et al. (2011).

This process revealed that a model with 40 topics best reflects the structure in the corpus. Furthermore, the ownership of the news provider of each article (public or private) and the month it was published are used as covariates in the topic prevalence. In other words, the probability distribution of topics depends on the business-model as well as on the month the article was published. The ownership is also included as a covariate affecting topical content, following the assumption that the same topic is discussed in different ways in private and public media respectively. To address problems due to non-convexity, we rely on the spectral initialization approach advocated by M. Roberts, B. Stewart, and Tingley (2016a).

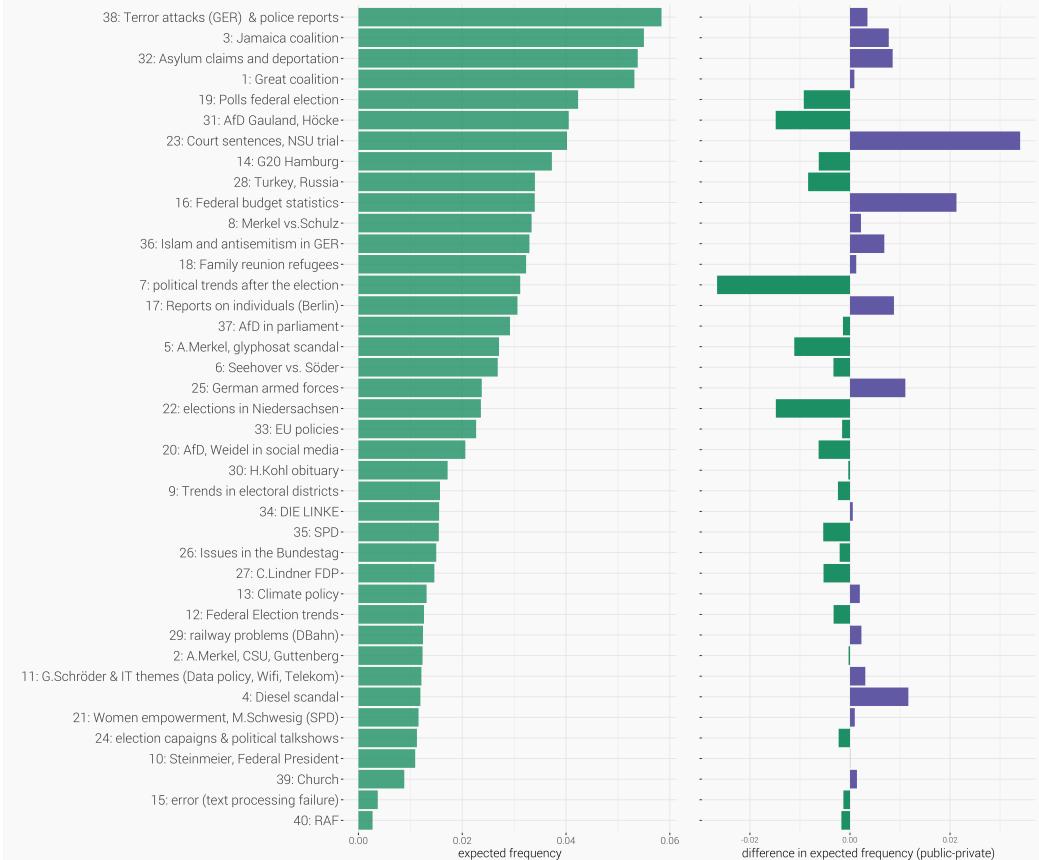
5 Empirical Evaluation

This section summarizes the results of the STM. Subsequently different measures to analyze the content differences between public and private ownership are applied according to the following approaches: (1) To address the question whether market imperfections exist in that sense, that some topics are not covered by private but by public news provider, I use the document-topic probability θ , to estimate the conditional expectation of topic prevalence for

given document characteristics. (2) Next, I find all edges between topics where they exhibit a positive correlation of θ above 0.1 to examine how topics are correlated differently for different covariate levels, indicating how topics are connected and framed differently between private and public media. Approaches (1) and (2) have been used in (M. E. Roberts, B. M. Stewart, and E. M. Airoldi, 2016). However, we extend the analysis by (3) calculating the similarity of the word-topic distribution β between the news provider using various distance measures, to identify which topics are discussed similar or differently.

Figure 3 displays the topics ordered by their expected frequency across the corpus (left side of the Figure) and the expected proportion of a topic in public media minus the expected proportion of topic use in private media (right side of the Figure). Thus topics more associated with public media appear to the right of zero. To assign a label to each topic, we looked at the most frequent words in that topic and the most representative articles (M. E. Roberts, B. M. Stewart, and E. M. Airoldi, 2016).

Figure 3: Expected topic proportion



AUSWERTUNG

5.1 Differences in topic prevalence

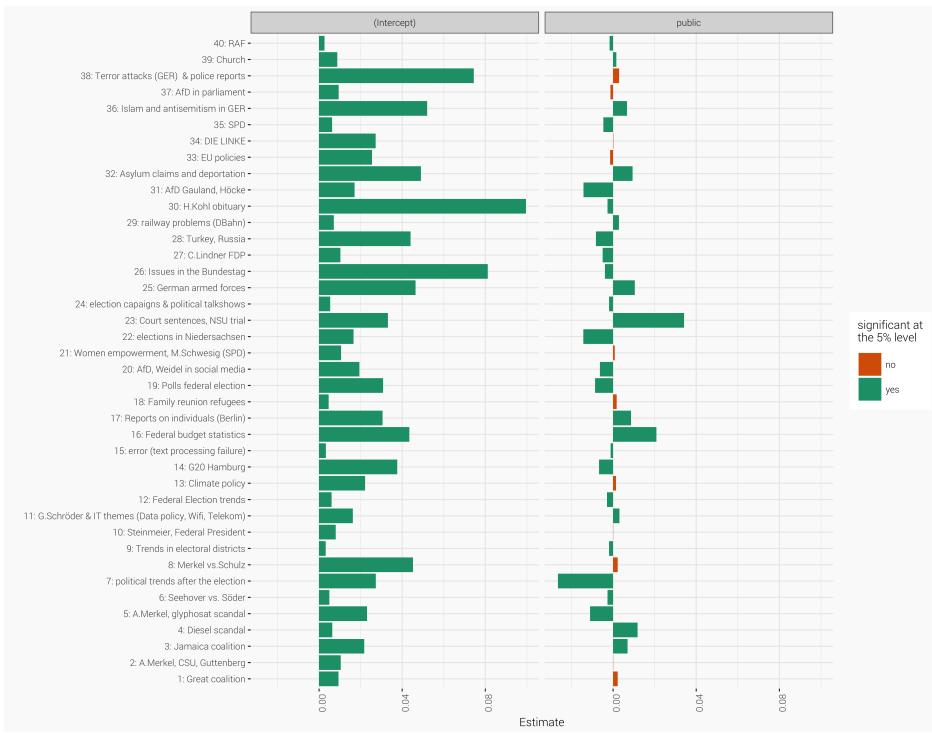
Since news source is included as a covariate in estimating topical prevalence part within the model, the frequency a topic is discussed within a news corpus can be calculated (M. E. Roberts, B. M. Stewart, and E. M. Airoldi, 2016). I estimate the conditional expectation of topic prevalence for given document characteristics. More specifically, I estimate a linear model, where the documents are observations, the dependent variable is the posterior probability of a topic and the covariates are the metadata of documents (see equation 5). The stm-package provides a function that uses the method of composition to incorporate uncertainty in the dependent variable, drawing a set of topic proportions from the variational posterior repeated times and compute

the coefficients as the average over all results (M. Roberts, B. Stewart, and Tingley, 2016b).

$$\theta_d = \alpha + \beta_1 x_{site} + \beta_2 x_{month} + \epsilon \quad (5)$$

Figure 4 displays the regression results for private vs. public. As the month an articles was published is included in the regression, the intercept indicates the mean of private news suppliers in June 2017.¹⁰

Figure 4: Regression results



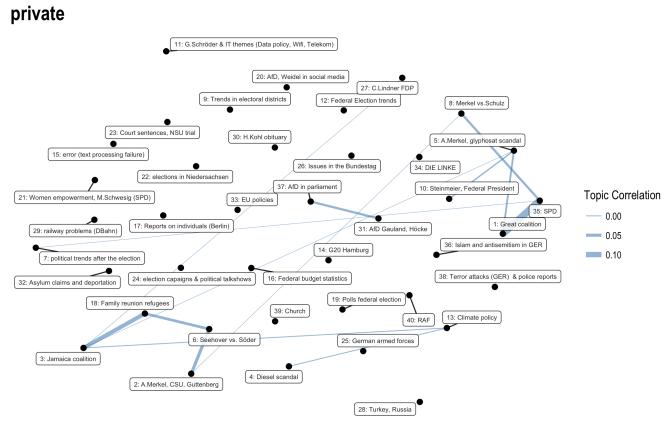
AUSWERTUNG

5.2 Topic correlations

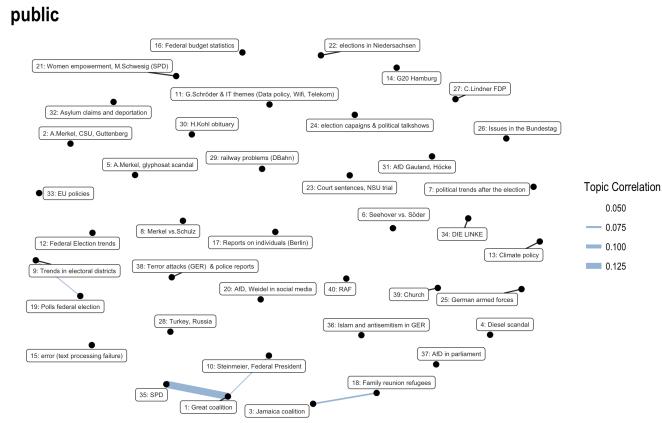
Next, the topic correlation is calculated, indicating how topics are connected and framed in public and private media. In Figure 5, indicates all edges between topics where they exhibit a positive correlation above 0.1 (M. E. Roberts, B. M. Stewart, and E. M. Airoldi, 2016).

¹⁰Table of regression results can be found in appendix.

Figure 5: Topic Correlation



(a) Private news provider



(b) Public news provider

AUSWERTUNG

5.3 Differences in word-topic distributions

Previous topic modeling research try to measure the similarity between topics, comparing the word-topic-distribution between the topics (most of the research uses LDA without metadata, so the word-topic-distribution is the

same within the whole corpus). cosine similarity (He et al., 2009), (Ramage et al., 2009), Kullback-Leibler (KL) (Newman et al., 2009), (Wang et al., 2009) and the average Log Odds Ratio (Chaney and Blei, 2012) are frequently used metrics to compare word probability distributions. Kim and Oh (2011) compare six popular similarity metrics in terms of log likelihood of data, concluding that Jensen-Shannon Divergence (the symmetric variation of KL divergence) is best in terms of performance and generality.

Since a topic is a multinomial distribution over the vocabulary, where β_{kvi} indicates the probability of observing word v in topic k for covariate level i , we can analyze how similar a topic k is for different levels of covariates. More precisely, we can calculate the similarity between β_{kvi} and β_{kvj} . I use each row of that matrix (corresponding to each topic) to calculate the similarity of the word-topic distribution between the public and private media using the square root of the Jensen-Shannon (JS) Divergence. I will compare this metric with a set of other similarity measures recently used in the literature to measure the difference between word-topic distributions: The Kullback-Leibler (KL) divergence (Newman et al. (2009), Wang et al. (2009)), the cosine similarity (He et al. (2009), Ramage et al. (2009)) and the difference of the L_1 norm of the vectors (M. Roberts, B. Stewart, and Tingley, 2016a). Before the results are presented, I will briefly explain these measures before I proceed to the results.

Essentially, the KL divergence is the expectation of the log difference between probabilities a and b . It can be defined as:

$$D_{KL}(a, b) = \sum_{i=1}^T a_i \ln \frac{a_i}{b_i},$$

where $D_{kl} \rightarrow 0$ indicates stronger similarities. However, since the KL divergence is not symmetric, it cannot be used to measure the distance between two distributions, but as a divergence measure.

The Jensen-Shannon Divergence is a positive definite measure, satisfying the following conditions: $D_{js}(a, b) \geq 0$, $D_{js}(a, b) = 0$ iff $(a = b)$. It is also symmetric: $D_{js}(a, b) = D_{js}(b, a)$. The Jensen-Shannon distance $D_{js}(a, b)$ between the two distributions is defined as:

$$D_{js}(a, b) = \frac{1}{2} D_{KL}\left(a, \frac{a+b}{2}\right) + \frac{1}{2} D_{KL}\left(b, \frac{a+b}{2}\right)$$

where D_{KL} is the Kullback-Leibler divergence. Since D_{js} does not satisfy the triangular inequality condition $D_{js}(a, c) \leq D_{js}(a, b) + D_{js}(b, c)$, the JSD

is not considered to be a real distance metric. However, we can use the square root of JSD as a real distance metric (Endres and Schindelin, 2003).

M. Roberts, B. Stewart, and Tingley (2016a) use the L_1 norm of the word-topic distributions to compare topics from different models. The L_1 norm is the sum of the absolute value of the difference between two vectors. Its defined as

$$L_1 = \sum_{i=1} |a_i - b_i|$$

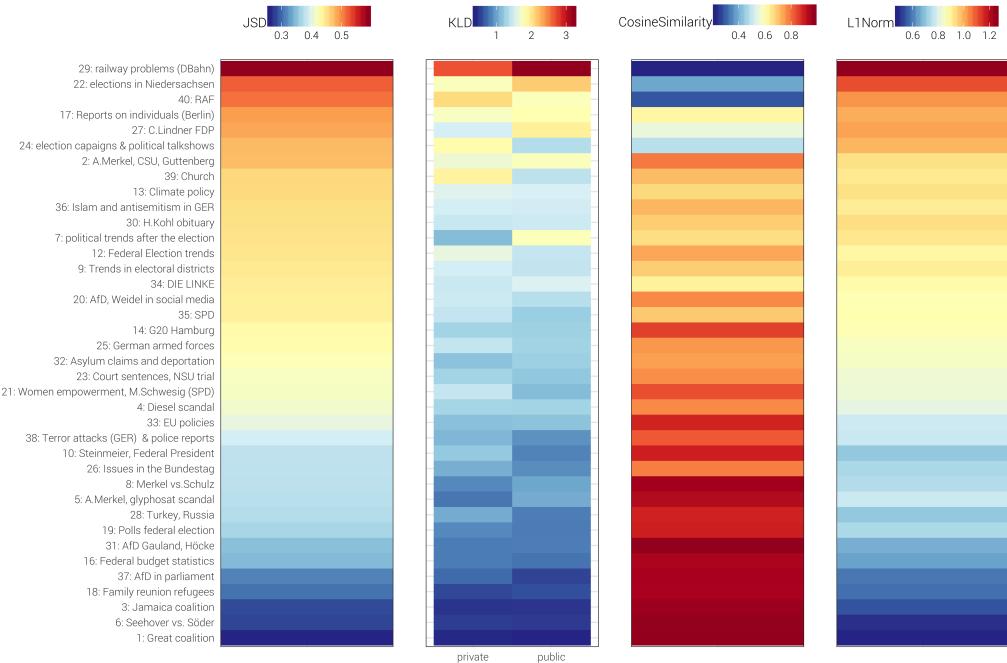
this measure has a range of $[0, 2]$, where $L_1 \rightarrow 0$ indicates strong similarities. M. Roberts, B. Stewart, and Tingley (2016a) compare this measure with a cosine similarity metric, which is essentially the dot product rescaled by the L_2 norm of the vectors and is defined as.

$$\cos(\theta) = \frac{\sum_{i=1} a_i b_i}{\sqrt{\sum_{i=1} a_i^2 \sum_{i=1} b_i^2}}$$

However, different to the above mentioned metrics the cosine similarity has a range of $[0, 1]$ and $\cos(\theta) \rightarrow 0$ indicate less similarity.

Figure 6 displays the different distance measures

Figure 6: Similarity measures of word-topic probabilities



AUSWERTUNG

Appendix