BGPE Discussion Paper

No. 177

Economic History Goes Digital: Topic Modeling the Journal of Economic History

Lino Wehrheim

November 2017

ISSN 1863-5733

Editor: Prof. Regina T. Riphahn, Ph.D.
Friedrich-Alexander-University Erlangen-Nuremberg

B G P E Discussion Paper Series

Economic History Goes Digital: *Topic Modeling the* Journal of Economic History

Lino Wehrheim[*]

November 21, 2017

Abstract

Digitization and computer science have established a whole new set of methods to analyze large collections of texts. One of these methods is particularly promising for economic historians: topic models, statistical algorithms that automatically infer themes from large collections of texts. In this article, I present an introduction to topic modeling and give a very first review on the research using topic models. I illustrate their capacity by applying them on 2.675 articles published in the *Journal of Economic History* between 1941 and 2016. This contributes to traditional research on the JEH and to current research on the cliometric revolution.

JEL-Classification: A12, C18, N01

Keywords: Economic History, Topic Models, Latent Dirichlet Allocation, Cliometrics, Digitization, Methodology

Lino Wehrheim
University of Regensburg
Department of Economics, Department of History
Universitätsstraße 31, 93053 Regensburg, Germany
lino.wehrheim@ur.de, +49 941 943 2715

[*] I thank Manuel Burghardt, Mark Spoerer, Tobias Jopp, and Katrin Kandlbinder as well as the participants in the research seminar in economic and social history at University of Regensburg for invaluable advice. This paper was presented in the lecture series on digital humanities at University of Regensburg.

Introduction

The turn of economic history towards economics and quantitative methods during the 1960s can be at least partially explained by the technological changes which facilitated the dissemination of computers (Haupert 2016). With digitization, economic history (as every other field in science) is again confronted with far-reaching technological changes. Despite the many uncertainties concerning the effects of digitization on economic history, one thing seems to be indisputable:[1] falling costs of digitization leading to large collections of digitized records (like *Chronicling America*) and advanced methods for analyzing them will change the way economic historians carry out their research in the future (Abramitzky 2015; Collins 2015; Mitchener 2015).

Looking at a different discipline, the narrative changes from future into present tense. In digital history, and digital humanities in general, scholars have already adapted to the growing mass of digital resources by incorporating methods from computer science.[2] Standing at the forefront of these methods, so-called topic models enjoy rapidly increasing popularity (Meeks and Weingart 2012, p. 2). The term "topic model" refers to statistical algorithms that automatically infer themes, categories, or topics from texts and which are the state-of-the-art in automated text analysis (Matthew Jocker (2013, p. 123) calls them the "mother of all collocation tools").

The idea behind topic modeling is quite simple: instead of reading texts and manually categorizing their topics (which for some collections of texts can require a great amount of resources or even be impossible), the distributions of words across documents are used to infer

[1] Questions on the future of economic history were discussed on a special panel at the 75th anniversary of the Economic History Association, see Journal of Economic History, Volume 75 Issue 4.

[2] For an assessment of the status quo in digital history, see the white paper "Digital History and Argument", the Arguing with Digital History working group, Roy Rosenzweig Center for History and New Media.

the inherent categories. This way, text can be quantified, a process that allows integrating qualitative sources into quantitative research.[3]

Although Ran Abramitzky (2015) and Kris Mitchener (2015) already mention them, to the author's best knowledge there is no paper published in an economic history journal explicitly covering or using topic models. Therefore, this paper intends to shed some light on a "exciting new trend" (Abramitzky 2015, p. 1248) and illustrate that topic models are a tool that promises to be of great utility especially for economic historians with their affinity with quantitative analysis. One reason topic models are rather unknown outside the community of digital humanities may be that this is a rather young discipline and much or even most of its research is not published in traditional print journals. Rather, research is communicated on blogs and websites, especially when it comes to tutorials, what may function as a "barrier of entry" to scholars from other disciplines (Meeks and Weingart 2012, p. 3).[4] In this paper, I will provide a mainly non-technical description of topic modeling as its (Bayesian) statistics are explained in detail by others.[5] Rather, the aim is to give insights into the general principles of topic modeling from a user's perspective and to address the questions to be considered before starting a topic model project, for instance: How do the texts have to be processed in order to be analyzed by a topic model? Which parameters have to be specified? Which potential problems have to be addressed? I will provide an overview of the literature using topic models which illustrates their disciplinary versatility, followed by a practical application: I use the most prominent topic model – *Latent Dirichlet Allocation* – to extract topics from all articles

[3] Abramitzky (2015, p. 1248) calls this "turning books into data".

[4] Scott Weingart's blog gives a helpful overview of blogs writing about topic modeling, see http://www.scottbot.net/HIAL/index.html@p=19113.html.

[5] See Blei et al. (2003), Blei and Lafferty (2009), Griffiths and Steyvers (2004), and Steyvers and Griffiths (2007) for formal descriptions.

published in the *Journal of Economic History* (JEH) between 1941 and 2016. The results will demonstrate that topic models are just the right tool for research like the work by Robert Whaples (1991, 2002), who has done a topic analysis of the JEH in a more traditional fashion. Furthermore, it will be shown that topic models can contribute to current research by Claude Diebolt and Michael Haupert (2017) and Robert Margo (2017) on the disciplinary shift in economic history known as the cliometric revolution.

The Principles of Topic Modeling

Topic models are one part in the field of text mining, which again is a melting pot of different disciplines like data mining, computational linguistics, and machine learning (Grimmer and Stewart 2013, p. 268; Miner 2012, pp. 31–34).[6] They are algorithms that analyze word occurrences to discover inherent categories and were developed in the field of computer science, machine learning, and information retrieval (Meeks and Weingart 2012, p. 2). Strictly speaking, they should be called probabilistic topic models, as they build on the assumption that a document can exhibit different topics and therefore work with probability distributions of words and topics (Steyvers and Griffiths 2007, pp. 430–32). There are different kinds of topic models depending on the statistical assumptions of the algorithms (Steyvers and Griffiths 2007). The one most commonly used and "state of the art in topic modeling" (Lüdering and

[6] To describe the origins of topic modeling in the context of digital humanities is quite challenging as this touches several disciplines that all have different histories. For example, the 'history of humanities computing' can be traced back to Father Roberto Busa, who indexed the work of Thomas Aquinas in the late 1940s, see Hockey (2004) and Jockers (2013). For a brief description of the recent development of topic models, see Lüdering and Winker (2016).

Winker 2016, p. 493) is *Latent Dirichlet Allocation* (LDA), which was introduced by David Blei et al. (2003).[7]

What do we expect from topic models? Basically, we want to know what our documents are about without reading all of them. Topic models provide an answer to this question by giving the topic composition of every document in our corpus. But, what is a topic? Topic models treat topics as distribution over words, so the second kind of output are lists of words that the model identified as having a high probability of occurring together.

Topic models build on two basic assumptions: Firstly, they assume that the semantic meaning of a text is created by the joint occurrence of words, although not all word clusters produce what we would call meaning (for example prepositions and articles). The idea of topic modeling is to use statistical methods to identify the relevant word clusters. They can be interpreted as topics "because terms that frequently occur together tend to be about the same subject" (Blei 2012b, p. 9). In other words, this assumption implies that meaning is relational, so the meaning of one single word depends on its co-occurrence with other words (Mohr and Bogdanov 2013, pp. 546–47). Topic models account for this polysemy by allowing a word to belong to different topics (Steyvers and Griffiths 2007, p. 429).

Secondly, topic models assume that a document is generated in a process which can be described by the following model (Blei 2012a, p. 80).[8] The corpus consists of D documents, each of which consisting of N words $w_{d,n}$, where $w_{d,n}$ is the nth word in document d. The overall vocabulary V is fixed. Documents exhibit a share of every topic k (although some might be infinitesimally small) with θ_d describing document d's distribution across topics. The overall number of topics K is assumed to be fixed. As stated above, topics are treated as

[7] There have been several extensions of the original model covering different assumption of LDA (Blei 2012a, pp. 82–84). In the following, the terms LDA and topic model will be used synonymously.

[8] That is why topic models are also called generative models (Steyvers and Griffiths 2007, p. 427).

distributions over words with β_k representing the distribution of topic *k*. Every word in a document is assigned to one or multiple topics, which is represented by topic assignment $z_{d,n}$ for word *n* in document *d*. A graphical representation of this model is provided in Figure 1. The only observed variable is words, which is represented by a shaded node. All other variables are hidden.

[Figure 1 about here]

The generative process of a document itself is assumed to be as follows (Blei 2012a, pp. 78–82; Blei and Lafferty 2009, pp. 73–75). First, choose a distribution over topics θ_d. From this, draw a topic *k*. Finally, choose a word $w_{d,n}$ from this topic. This is repeated for every word in every document. In other words, it is assumed that first, the author decides what topic the text should be about by determining the topic shares (step one). The actual writing is interpreted as choosing words from a topic-specific vocabulary according to the topic-shares (steps two and three). The reader cannot observe the generative process but only the output (the words).

The basic idea behind LDA is that the generative process corresponds to a joint probability distribution of the hidden variables (topic vocabulary and topic shares) and the observed variables (words). This distribution is used to answer the question: "What is the likely hidden topical structure that generated my observed documents?" (Blei 2012b, p. 9). The conditional distribution of the hidden variables given the observed variables called *posterior distribution* is given by (Blei 2012a, p. 80):

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \quad (1)$$

This posterior is what we are looking for because it tells us the probability of topics and topic assignments of words given the words that form our corpus. Unfortunately, the conditional distribution cannot be computed directly (Griffiths and Steyvers 2004, p. 5229). There are several techniques for estimating the posterior (Blei and Lafferty 2009, pp. 76–78) and

explaining all of them would go beyond the scope of this paper. The most common one, Gibbs sampling, can be outlined as follows.[9] Technically, LDA assumes the two steps of generating the documents to happen randomly (Blei 2012a, p. 78). Starting from a random topics assignment, Gibbs sampling resamples the topic assignment for every word in every document by asking two questions: Which topics can be found in the document and which topics is this word assigned to in other documents? It calculates the topic assignment with the highest probability given the assignments of the other words in the document and given the topic assignment of the word under consideration in other documents and updates the word's topic assignment accordingly. How many times this is done is determined by the researcher with more iterations leading to more coherent topics, although this effect will level off at some point (Jockers 2014, p. 147).[10]

So far, the name LDA has not been explained. The distribution over topics in the first step θ_d is assumed to follow a *Dirichlet* distribution (Blei 2012a): a distribution over another distribution that is specified by the Dirichlet parameter α, which is a vector over $(\alpha_1; \alpha_2; \ldots \alpha_K)$ (Steyvers and Griffiths 2007, pp. 430–32; Wallach et al. 2009). The topics β_k are assumed to follow a Dirichlet distribution over words with the parameter η. Both α and η may be interpreted as concentration parameters that can be modelled symmetrically, implying that the topics are distributed equally over the corpus and words contribute equally to the topics. Alternatively, they can be modelled as asymmetric and be estimated, implying that some topics are more

[9] The following description is inspired by a lecture given by David Mimno at the Maryland Institute for Technology in the Humanities in 2012 (available at https://vimeo.com/53080123) and Ted Underwood's description of topic modeling on his Blog (available at https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/). For a technical description, see Griffiths and Steyvers (2004) and Steyvers and Griffiths (2007).

[10] There is a tradeoff between topic coherence and the time it takes to train the model. Finding many topics in large corpora can keep the computer busy for hours.

prevalent (asymmetric α) and some words are more "important" for a topic (asymmetric η) than others. Finally, the model *allocates* words to different *latent* (i.e. not observable) topics (Blei 2012a).

Topic Models in Practice

As stated above, topic models treat topics as distributions over words. Accordingly, the results are groups of words that have a high probability of occurring together. However, these groups lack any kind of label (Blei 2012a, p. 79). They might or might not be recognizable as a theme at first glance. Anticipating the results from the topic models on the JEH, two examples shall be given. The ten most probable words for topic 1 are *japanese, japan, china, chinese, rice, land, period, government, meiji,* and *tokugawa* (words are ranked in a decreasing order). For topic 7, it is *bank, banks, banking, deposits, reserve, national, notes, system, state,* and *credit.* Both topics are quite coherent, and reasonable labels could be *Japan and China* and *Banking*. It is important to note that the topic model found these topics without any prior information on entities like countries or financial institutions.

Topics do not necessarily have to describe what the documents "are about". They can also be clusters of methodological words or days of weeks (Boyd-Graber et al. 2015, p. 240). In general, interpretability (or coherence) can be regarded as the linchpin in topic modeling: only if we can identify its meaning we can process the topic further. The degree of coherence depends on model specification (especially the number of topics), the characteristics of the corpus, and the level of granularity one is interested in (Jockers 2013, pp. 127–28). In general, decreasing the number of topics results in more coherent but also less specific topics. Running several topic models with different numbers of topics will be the most practical solution for identifying the right number given the research question.

Interpreting and labelling the topics is of course quite subjective as the interpretation of words can differ from one reader to another (Jockers 2013, p. 130), which makes transparency an

important condition. Nevertheless, subjectivity is a familiar problem: individual judgments also must be made when coding a text manually. In other words, by using topic models, we can postpone the moment when subjective assessments become necessary: from the ex-ante subjectivity of specifying categories to the ex-post subjectivity of interpreting them. There are some metrics to diagnose the "quality" of the topics (Boyd-Graber et al. 2015), but in the end, it depends on human interpretation to identify their common denominator. The aspect of human interpretation is discussed by Jonathan Chang et al. (2009) who pursue an experimental approach to investigate how humans interpret topics.

There is one technical remark especially important for historical research: the standard LDA topic model does not capture changes in the use of language. For example, sources from the early 18th and the late 19th century might describe the same subject with different vocabularies, which would probably lead to two different topics. There are extensions of LDA accounting for this (Blei and Lafferty 2006), but this problem theoretically could be solved by combining those topics covering the same subject in different "languages" or by creating sub-corpora. Changes in terminology will be a potential problem in some cases, in others they might be what we are looking for,[11] so controlling for them depends on the corpus as much as on the research question.

Some remarks have to be made on the data: topic modeling works with so called "bag of words" representations of texts, which means that the word order does not matter (Blei 2012a, p. 82). The sentences "France was industrialized after Great Britain" and "Great Britain was industrialized after France" are treated as identical. This may look somewhat unrealistic, but still both sentences suggest that their content is about industrialization, France, and Great Britain.[12] Furthermore, the text is converted into so-called tokens which means separating a

[11] See for example McFarland et al. (2013).

[12] For an extension relaxing the bag of words assumption, see Wallach (2006).

string of text into pieces (Boyd-Graber et al. 2015, p. 230). This is done most simply by using whitespace as a mark. Depending on the tokenizer, this can also imply cutting at punctuation, converting to lower case, and removing numbers.

There are several further steps that can be applied to preprocess the corpus (Boyd-Graber et al. 2015, pp. 227–31). It is common to remove words that occur frequently and have no semantic meaning (like *the, and, a, or*). These so-called *stopwords* are removed based on a fixed list, but sometimes it might be necessary to further remove corpus-specific words if they occur too often and therefore only produce noise (Jockers 2013, p. 131). A common measure of relative importance is the *term frequency – inverse document frequency (tf-idf)* (Blei and Lafferty 2009). Depending on where the documents come from, they may contain words that do not belong to the text itself, so-called *boilerplate* (Boyd-Graber et al. 2015, p. 228). This could be HTML-tags when the text has been directly received from a website, download signatures or text fragments from other texts caused by missing page breaks.

Another step is the normalization of the text itself: removing capitalization, reducing the words to their stem, or lemmatizing (reducing words to their basic forms) can help to remove noise from the data (Boyd-Graber et al. 2015).[13] What kind of preprocessing steps should be taken depends on the corpus and the research question. For example, it can be helpful to concentrate only on nouns, which can be achieved by using a part-of-speech-tagger (Jockers 2013, p. 131).[14]

Topic models come with a caveat especially important for historians: their results crucially depend on the quality of the documents. Most texts used by historians will be either transcriptions or optical character recognition (OCR) treated scans. As both can be prone to misspellings, one has to check the documents carefully before applying a topic model

[13] For tools to carry out these steps, see Graham et al. (2016). For a discussion of the effect of stopword-removal, see Schofield et al. (2017).

[14] For a general discussion of texts as data for economic research see Gentzkow et al. (2017).

(otherwise it could be a typical case of garbage in, garbage out). In some cases, the road of digital scholarship can already end here as the text quality might be too poor and correcting the texts would be too time consuming or costly. In others, as Daniel Walker and William Lund (2010) show, systematic errors like repeated OCR-mistakes can be treated, and some amount of random errors might be tolerated.[15]

There are several applications for topic modeling, inter alia an extension for *R* (Graham et al. 2016). In this paper, I used MALLET, a user-friendly tool developed by Andrew McCallum in 2002 (McCallum 2002) which implements LDA and Gibbs sampling.

Concluding this chapter, the main strengths of topic modeling shall be emphasized. Topic modeling is primarily about reducing complexity by finding and applying categories. The crucial point of topic modeling is that no classification scheme has to be specified in advance. Rather, the documents speak for themselves and define their own categories, which is a major advantage compared to using classification schemes like dictionaries or JEL-codes as used in Abramitzky (2015), McCloskey (1976), and Whaples (1991; 2002), which only in the rarest case perfectly fit the data. This especially holds true for historical research where the usage of contemporary categories may miss the point as they might not fit historical sources.

This touches some fundamental epistemological considerations: scientists approach their sources with some a priori framework in mind, which results from their prior knowledge, their personal interest, their socialization, and so forth. Accordingly, they examine their sources according to their individual concepts of relevance, which in itself is neither good nor bad as long as their reasoning is comprehensible. Still, this contains the risk of biasedness, as for instance, people tend to select information that confirms their beliefs. In contrast, the topic model is agnostic, that is, it works without any a priori understanding of the sources. Rather, it

[15] OCR-mistakes can build their own topic, see Jockers (2013).

identifies their inherent structure according to the statistical algorithm. In this respect, topic models are also different from so-called supervised learning approaches where algorithms are iteratively trained by human intervention. Furthermore, human coding is prone to imprecision and mistakes to which the computer is immune. A practical benefit of their unbiasedness is that topic models can help to identify relevant sources, including also those that might be overlooked by using search terms. This way, topic models can also be used for browsing data bases.

The second strength of topic models is their ability to automatically create quantitative representations of texts including their semantic meaning which goes far beyond traditional methods of quantification. The automated nature of the process allows to enlarge the database as far as computing power allows. The fact that words and documents can be assigned to multiple categories enables a degree of granularity that would be infeasible in manual coding. Besides, topic models can be applied to other input than texts, like images (Blei 2012a, S. 83), which opens new possibilities for quantitative research.

Using topic models, we can integrate textual sources into a quantitative framework and this way combine texts with traditional data. There is a myriad of conceivable applications for economic historians. For instance, topic models allow us to get insights into the reasoning of economic agents as we now can use textual resources on a completely different scale. Especially, combining topic models with other text mining approaches like measuring sentiment seems very promising. Minutes of central banks, ministries, cabinets, or executive boards seem to be ideal candidates for a topic modeling application. Furthermore, the ambiguous notion of impact can be investigated much more tangibly. Giving just one example from some current research, it can be studied how decision-makers are influenced by economic policy advice. The following section will give an overview of the literature that uses topic models.

Literature Review

By now, there is a considerable amount of research using topic models. In Table 1, the literature of potential interest for economic historians is presented using the old-fashioned way of categorizing texts. In the following, some of them will receive special notice.

David Newman and Sharon Block (2006) have been among the firsts to apply topic models on historical sources. They use several kinds of topic models for finding themes in a colonial US newspaper, the *Pennsylvania Gazette*. Their analysis of 80,000 documents published between 1728 and 1800 impressively shows the potential of topic modeling for large scale research.

Newspapers are also the database for Tabitha Bonilla and Justin Grimmer (2013) who investigate the influence of several raises of the terror alert level under the Bush administration between 2002 and 2005 on public debate in the media. Paul DiMaggio et al. (2013) apply topic models on newspapers in order to find how the shrinking public support of the arts in the US between 1986 and 1997 was framed by the media coverage. Carina Jacobi et al. (2015) examine the coverage of nuclear technology in the *New York Times* between 1945 and 2013. The project "Mining the Dispatch" by Robert Nelson applies topic models on the *Richmond Daily Dispatch*, a Confederate daily newspaper, between 1860 and 1865 to investigate social and political life in Civil War Richmond.[16]

Jockers (2013) uses topic models for a corpus that at first glance does not seem to be especially relevant for economic historians (19th century novels from Great Britain and the US). Nevertheless, his work illustrates how topic models can be combined with metadata of the documents and this way be further refined. Particularly, he records the authors' sex and nationality, which allows him to show that, for example, female authors write more about "Affection and Happiness" than their male counterparts.

[16] Availabe at http://dsl.richmond.edu/dispatch/pages/home.

Neil Fligstein et al. (2014) use topic models to answer the question why the Federal Reserve (Fed) failed to see the financial crisis in 2008. Particularly, they use the topics in the Federal Open Market Committee (FOMC) minutes to measure how the Fed perceived the US economy between 2000 and 2008. They can show that the Fed was neither aware of a housing bubble nor of the entanglement of the housing and the financial markets. Stephen Hansen et al. (2014) use the same database to investigate how transparency affects the deliberation of monetary policymakers.

That topic models can be combined with econometrics and economic data is shown by Hansen and McMahon (2016). Studying also the FOMC, they investigate the effects of central bank communication on macroeconomic and financial variables. Another example of the integration of topic models into economic analysis is given by Jochen Lüdering and Peter Winker (2016). They study the question as to whether economic research anticipates changes in the economy or merely looks at the economy from an ex post viewpoint. They apply a topic model on the *Journal of Economics and Statistics* and compare the temporal occurrence of topics connected to the inflation rate, net-exports, debt, unemployment and the interest rate to their corresponding economic indicators. Scientific journals are also the sources of David Hall et al. (2008), David Mimno (2012), and Allen Riddell (2014).

How topic models can be used for research in finance is shown by Vegard Larsen and Leif Thorsrud (2015), Larsen and Thorsrud (2017), Thorsrud (2016a), and Thorsrud (2016b), which all build on the same corpus (articles published in a Norwegian business newspaper between 1988 and 2014). Here, the topics of the newspaper are used to predict asset prices (Larsen and Thorsrud 2017) and economic variables (Larsen and Thorsrud 2015). Furthermore, they are used to construct a real-time business cycle index for so-called nowcasting (Thorsrud 2016a).[17]

[17] In finance, there seems to be an affinity towards text as data, which can be traced back to Tetlock (2007) being the first to use text analysis in order to measure market sentiment.

[Table 1 about here]

Topic Modeling the JEH*: Whaples Reloaded*

When testing something new, it can be helpful to know what the results ideally should look like. That is why in the following, topic models will be used to identify themes in the *Journal of Economic History*. The JEH is chosen as a case study because with Whaples (1991, 2002) there are two works that deliver an invaluable benchmark: Whaples classified the content of the JEH according to a modified version of *Journal of Economic Literature's* (JEL) codes, counting the percentage of pages published in a given category. In contrast, the topic model works without an ex ante classification scheme, and it works automatically.

A topic model is applied on two samples of texts: The first one includes all articles from Volume 1 to Volume 50 Number 2 using 41 topics just as in Whaples (1991).[18] The second sample extends the analysis into the present, consisting of all articles published between 1941 and 2016. Here, a topic model with 25 topics is used, which corresponds to the number of subjects in Whaples (2002).

In both samples, the topics were generated with MALLET, using 2,000 iterations and allowing for hyperparameter optimization (that is, modelling the Dirichlet parameters asymmetrically). The corpus was preprocessed in the following manner. Regular expressions from the header on the first page and the copyright section of every paper were deleted. The documents consist of bibliographical text to a large degree, which distorted the topics in the first trials. Therefore, the most frequent expressions related to bibliographical references were removed. This mainly concerns places of publications. For instance, every variation of "university press" was removed, as was every occurrence of New York, Cambridge, London and Oxford in a

[18] That is: all articles published in regular and Task issues except regular book reviews and dissertation summaries, see Whaples (1991).

bibliographical reference.[19] Names of universities were not removed as they might be part of a subject like disciplinary history. Furthermore, the expressions “per cent” and “New York” (if not in a reference) were collapsed into “percent” and “newyork” as “per” and “new” are part of the stoplist.[20] Furthermore, download signatures had to be removed.

For the stoplist, the MALLET built-in list was used, as was the built-in tokenizer that removed capitalization and numbers. Further preprocessing steps like stemming (reducing words to a common stem) were not applied to keep the process as transparent as possible. Eventually, the overall database consists of 2,675 articles or 19.8 million tokens, which approximately equals 35 times of “War and Peace”.[21]

MALLET basically provides two kinds of output: First, it produces the topic keys, which show the most probable words for every topic (their number can be varied). Second, it generates a file containing the topic shares (or distributions) for every file that add up to one. This makes it possible to identify the most prominent topics for every article, to calculate average topic shares for every topic and to compute time series of topic prevalence.

[Table 2 about here]

The topics of the first sample are shown in Table 2. The first column states the topic number randomly given by MALLET. In the second column, the 30 most probable words for every topic are given in descending order. For example, in topic 1 *japanese* is the most probable word,

[19] In the first trials, almost all topics contained the word *Cambridge*. Other cities that occur in the final topics could not be found to turn up regularly in bibliographical references except in combination with “university press”.

[20] As ‘york’ and ‘cent’ occurred in several early topics, it became clear that actually New York and per cent was meant, so this step was done for reasons of clarity and esthetics.

[21] The stopwords can be received upon request. For sample one, the database consists of 1,728 documents.

followed by *japan*.[22] The relative importance of words for a topic might be better visualized using word clouds (as in Figure 2).[23]

[Figure 2 about here]

In most cases, the topics seem to clearly exhibit what we would expect when thinking of topics and show a great degree of coherence: for example, in topic 11 the words *agricultural, agriculture, wheat, grain, farmers, crops* suggest that this topic is most likely about agriculture. That topic 18 can be labelled *Slavery and Servitude* is not only justified by words like *slaves, slave,* and *slavery* but also by the reference to Robert Fogel and Stanley Engerman.

A topic that stands out is topic 36, which (at least for the author) does not have an intuitive interpretation. Looking at the articles that show the highest share of topic 36, it becomes clear this topic covers research concerning people: either they cover individuals, like the article by Walters and Walters (1944) on David Parish (48%), or groups of people like the article by Freeman Smith (1963) on the international bankers committee on Mexico (48%). The numerous occurrences of months seem to derive from the fact that those articles are largely based on the interpretation of letters.

Topics 6 and 31 show that topics can also represent a different kind of theme, in this case the use of technical expressions typical for quantitative methods: topic 6 contains words that can be attributed to basic descriptive statistics. Particularly, words like *period, year(s), series, annual, time, index* are terms connected to time series. Topic 31 contains words that could be found in the glossary of a textbook on econometrics. Obviously, the topic model differentiates between rather descriptive and econometric methods. Topic 23, having the highest average

[22] If a stemmer had been used, these words would have been collapsed into *japan*.

[23] Depicting every topic as a word cloud would exceed the available space of this article.

share of all topics, seems to contain general expressions which might be typical for an economic historian's jargon.

Wherever they seem appropriate, subjects from Whaples (1991) were added as labels.[24] If this was not the case, a new label was given.[25] The overall impression is that the topics seem to match the subjects used in Whaples (1991) quite well. From the 41 subjects, 26 can be identified including nearly all the major ones.

In some cases, the topics seem to be more differentiated than the subjects: topics like *Japan and China* (1), *Germany* (26) and *France* (35) could of course be assigned to "Country Studies", but they are identified as independent subjects by the topic model.[26] The same holds true for "Trade": the topic model finds different subcategories like *Slave Trade* (15) or topic *North Atlantic* (24). The subject "Economic Growth" seems to be split into two topics, one describing growth (topic 5) and one explaining it (topic 34). Topic 38 could be attributed to "Imperialism/Colonialism", but a label like *Westward Movement* seems to be more appropriate.

The topic model also makes a difference between geographical and sectoral aspects of industrialization: topic 33 contains words relating to Great Britain as the first country to industrialize, whereas topic 39 shows words referring to the textile industry as a central sector concerning industrialization. Furthermore, some topics are connected to different subjects: for example, topic 30 contains words that could belong to both "Public Finance" and to "War", which is not surprising as one major part of public spending is on military purposes.

[24] These subjects are based on JEL codes, see Whaples (1991, pp. 289–90).

[25] Of course, this assignment is somewhat subjective, but it is not more subjective than assigning pages to subjects by hand.

[26] Except for Canada, which shares a topic with other countries (Topic 20), every country analyzed in Whaples (1991) constitutes a special topic. These countries are Britain (33), France (35), Italy (16), Germany (26), Japan (1), Russia/Soviet Union (29), and the United States (9).

The topics about individual countries draw attention to the question of different languages: words like *der, die, das,* or *des, les, sur* would be regarded as stopwords in a German or French corpus.[27] Of course, these words could be removed by expanding the stoplist. Anyways, they facilitate the identification of documents that build on sources in languages other than English, which for example might support research concerning geographical coverage. In the topic on France, the words *annales* and *histoire* may be regarded as a reference to the *Annales School* and its major journal *Annales d'histoire économique et sociale* (Burguière 2009).

The topic model did not identify some subjects from Whaples (1991), which could have several reasons. The subject might just be too small compared to the corpus (like in the case of "Minorities/Discrimination"), what could be solved by increasing the number of topics or by reducing the corpus into a subsample. Here, the agnostic nature of the model again comes into play. Searching for a subject on minorities might be completely reasonable in a certain framework. Still, the model did not identify this topic as substantial at the given level of granularity. That is, given the number of topics, the model assesses "Minorities" as irrelevant.

Another reason could be that different subjects share a similar type of vocabulary (or meaning) and therefore cannot be separated by the topic model (like "Business Cycles" and "Recessions/Depressions"), what again might imply that they are not clearly specified. In Whaples (2002), several subjects are combined what also hints at that direction. Another theoretical, although not very likely reason might be that a subject does not have any specific vocabulary and therefore is untraceable for a topic model.

[27] These words most probably stem from the bibliographical references which were not translated. The problem of corpora that comprise of different languages is discussed in Mimno et al.(2009).

Expanding the analysis into the present, another topic model is run on all articles between 1941 and 2016, this time with 25 topics as in Whaples (2002).[28] The results are shown in Table 3 (the development of all topics can be found in Appendix 1). Again, the labels where chosen as in Whaples (2002) wherever they fit the topics. From the 24 subjects used in Whaples (2002), 17 can be attributed to topics.[29] In principle, reducing the number of topics should lead to more coherent but also more general topics. Of course, both tables cannot be compared directly as the reduction does not happen ceteris paribus. Nevertheless, some observations can be made: the topic *Industrialization* in the second sample is again spread over two topics. One topic comprises references to Great Britain as the place of the first industrialization (topic 14). This time, the second one is much broader: topic 3 contains references to the textile as well as other early industries. Looking at the documents with the highest share of topic 3, it turns out that their common theme is technology.

[Table 3 about here]

Reducing the numbers of topics creates a subject that is left out in Whaples (2002) but was used in its predecessor: topic 16 seems to cover several countries, bringing back the subject "Country Studies". Again, there is one topic containing econometric vocabulary (topic 6), although the words are slightly different. In the case of the *Descriptive Language* (8), there now seems to be a stain of words connected to economic growth.

Topic models can be used to describe some general trends in the JEH: judging by the topic shares of sample two across time (see Appendix 1), *Methodology and Disciplinary History* (20) has experienced a major decline right from the start (with the exception of 1959/60), a finding

[28] A direct comparison of the results in Whaples (2002) like in sample one could have been carried out as well, but was relinquished due to space restrictions.

[29] The 25th subject in Whaples (2002) is the residual "Other".

consistent with Whaples (1991, 2002).[30] The same holds true for the topic *People* (11), and *Economic Growth* (21). The most prominent topic is *Economic Growth* (21) with an average topic share of 16.6%. During its heyday in the 1960s, *Economic Growth* reached nearly 27% (1967). In other words, every article in 1967 consisted on average to more than one quarter of words connected to this topic.[31] Since then, the academic interest has constantly declined (see Appendix Figure 3), a finding that is in line with Whaples (1991, 2002).

Technically, every document comprises a share of every topic, even though it might be vanishingly small. Defining a topic as "substantial" if it has a share of 10% or more, articles in the JEH contain on average 3.2 topics, which since 1941 has changed only marginally. The development over time and the results for a 5% and 20% threshold can be found in Appendix 2. The same continuity can be stated for topic concentration with an average Herfindahl index of 0.24 per article. These findings are most likely due to the nature of the JEH as a specialists' journal. In general, looking at the topic distribution of a document can give an idea of what it is about. To give a prominent example, the topic distribution of Robert Fogel's 1962 railroad paper is given in Figure 3. Not very surprisingly, the most prominent topic is *Transportation* with a topic share of 41%.

[Figure 3 about here]

The topic shares can be used to investigate topic correlation. Calculating the topic correlation based on individual documents yields mostly uncorrelated topics, except of *Econometric Language* which shows some negative correlation with *People*, *Methodology and Disciplinary*

[30] The peak of 1960 can be attributed mostly to the Task issue; a nice punchline: the article with the highest share of topic 20 is Goodrich (1960) discussing how the use of quantitative methods affects economic history.

[31] The use of annual averages is of course prone to outliers. If one is interested in the long term development, a moving average seems to be more appropriate. On the other hand, the outliers might be what we are looking for if we are interested in identifying special events.

History and *Economic Growth*.[32] The low correlation probably results from to the low number of topics within the documents. Nevertheless, topics might correlate across time, in terms of some topics occurring together resulting from larger topical trends. Computing the correlation coefficients based on annual topic shares increases the number of correlated topics and confirms what reasonably can be expected: for example, there is a high correlation between *Standard of Living and Health* (2) and *Econometric Language* (6). The topics *People* and *Methodology & Disciplinary History* stand out as they are correlated negatively to almost every other topic. On the positive side, *Econometric Language* is the topic most correlated to others, which confirms its quality as a meta topic. A network representation of topic correlation can be found in Figure 4, which illustrates the connection between topics based on their correlation. The question of correlation should at the heart of further research, for example by using a type of topic model that explicitly accounts for correlation (Blei and Lafferty 2007). Another option for future research might be the inclusion of article metadata like author information (as in Whaples 1991, 2002) and the comparison to other economic history journals.

[Figure 4 about here]

The Cliometric Revolution in Topics

The methodological topics lead us to a subject recently addressed by Diebolt and Haupert (2017) and Robert Margo (2017) which is also covered by Whaples (1991, 2002): the turn of economic history towards economic theory and quantitative/econometric methods during the 1960s, known as the cliometric revolution.[33] Can the spread of economic methods be observed in the topics? Looking at the distribution of the methodological topics over time, the answer is clearly yes.

[32] The correlation matrix is available upon request.

[33] For a comprehensive history of cliometrics see Haupert (2016) and the cited papers.

[Figure 5 about here]

Figure 5 shows the annual average topic shares of the methodological topics.[34] There is a continuous rise of the econometric topics starting in the 1960s, a finding completely in line with Diebolt and Haupert (2017), although the rise in the econometric topics is not as steep as in their measure and their first peak in the early 1970s cannot be found.[35] This might be due to the fact that Diebolt and Haupert (2017) do not include the Task issues: until the late 1960s, the Task issues contain more disciplinary reflections and less cliometrics than regular issues (Whaples 1991, p. 293).[36]

The finding of a more or less continuous rise in the econometric topics to date is in line with Margo (2017) who uses a dictionary approach. His search terms can be regarded as corresponding to the econometric language topic.[37] It is important to note that by using a topic model, we can come to the same conclusion, but without tying down the search terms *ex ante*. Rather, the topic model identified this theme without any prior knowledge, which again stresses the agnostic nature of the model. This also highlights one advantage of topic models over dictionary approaches: with a dictionary, it is necessary to use unambiguous terms which limits the search list. The topic model also includes words that also have a non-econometric meaning (like *test* or *significant*).

[34] The econometric language topics exhibit almost identical shares in both samples indicating that they are quite congruent. The descriptive topics show some difference because in sample 2 this topic seems to be less coherent than in sample 1.

[35] Diebolt and Haupert (2017) count equations, tables and graphs per page.

[36] Until 1996, papers presented at the annual meetings of the *Economic History Association* were published in a fourth issue, which was devoted to the "Tasks of Economic History" (Diebolt and Haupert 2017, p. 22; Margo 2017, p. 12).

[37] Margo (2017) uses an index based on the terms *regression, logit, probit, maximum likelihood, coefficient, standard,* and *error*.

Furthermore, the topic model shows that looking only at econometrics in the narrow sense does not capture the whole extent of the cliometric revolution. As Margo (2017, pp. 27–28) points out, although early cliometric work did apply quantitative methods, it discussed the results only briefly. Accordingly, only a low degree of econometrical terminology can be expected. At this point, the topic model again shows its convincing strength: identifying a second, more descriptive topic, it can be shown that the JEH already was quantitative when econometrics just began to expand. The development depicted in Figure 5 can be interpreted as the gradual integration of ever more advanced quantitative methods over time which is mirrored in a linguistic shift. The descriptive topic is also present well before 1960 indicating that at a low level, quantitative methodology was used before the arrival of econometrics, which is in line with Diebolt and Haupert (2017).

Figure 5 can be interpreted as describing the intensity of the use of quantitative methods. Papers on average became more cliometric during the 1960s. But was this development accompanied by an increase in the number of cliometric articles? To measure the extent of the cliometric revolution, another feature of topic models is applied. Topic models can be used to classify articles according to their content. An example is given in Figure 6: articles were classified as "quantitative" if their share of either topic 6 or 8 (the two topics related to economic methods) amounts to at least 5%. Compared to an average topic share of 4% in the overall corpus (median 0.04%), the 5% threshold seems appropriate (additionally, a narrower definition of "quantitativeness" was used by increasing the threshold up to 10%). In both cases, the development of the 1960s is now much more similar to the one described by Diebolt and Haupert (2017). Still, there is a continuous rise after the 1970s, a difference which might be again due to the different database. To account for the Task issues effect, another topic model is run on all articles excluding the Task issues. The result is that now the share of quantitative articles reaches a peak of 100% at the 5%-level in 1971 and then stays above 80% most of the time (see Appendix 3).

[Figure 6 about here]

Conclusion: Writing Digital Economic History

This article presents a state-of-the-art method from digital humanities: topic models, which are statistical algorithms that extract themes (or, more generally, categories) from large collections of texts. I introduce the basic principles of topic modeling, give a very first review of the existing literature, and illustrate the capability of topic models by decomposing 2,675 papers published in the *Journal of Economic History* between 1941 and 2016. Comparing my results to traditional scholarship on the JEH and to current research on the cliometric revolution, I can show that topic models are a sophisticated alternative to established classification approaches. Without any prior specification, the topic model identifies two topics containing terms connected to quantitative research. Using the temporal distribution of these topics, the model can retrace economic history's turn towards economics during the 1960s. Further research could include a topic model analysis of purely economic and historical journals in order to infer topical reference points, and of course of other journals from economic history to gain a more comprehensive perspective on the discipline.

For economic history, the three main strengths of topic models are efficiency, unbiasedness and quantification: they provide the means for analyzing a myriad of documents in a short amount of time avoiding the risk of human negligence; they are agnostic in terms of waiving *ex ante* classification schemes like JEL codes; and they deliver quantitative representations of texts which can be integrated into existing econometric frameworks.

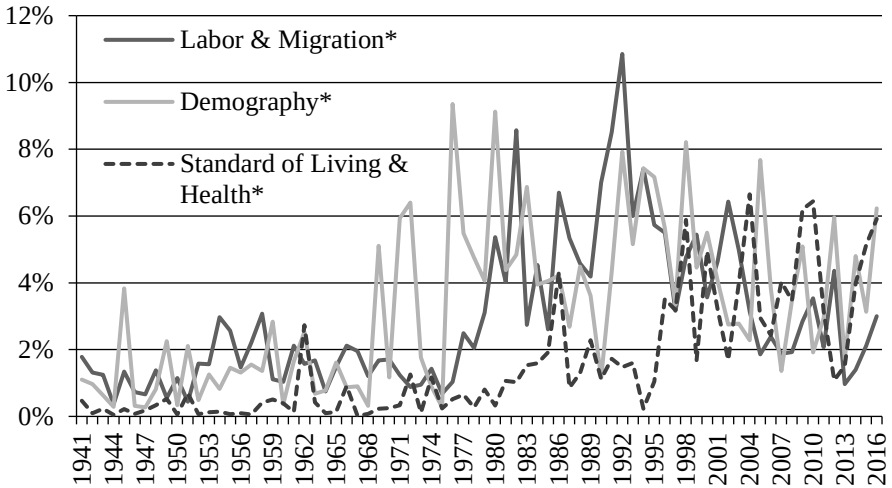
Especially the latter point makes topic models a worthwhile approach for economic historians. As part of the wider approach of distant reading (Moretti 2013), they provide the opportunity of re-integrating textual sources into economic historians' research. One conceivable application could be the generation of historical data. As the research in finance described in the literature review has shown, topic models can be used to predict developments on financial

markets and short-term economic development. Instead of predicting the future, this approach could be transferred to settings with a lack of historical data. For instance, applying topic models on historical newspaper could yield surrogates for financial and macroeconomic data.

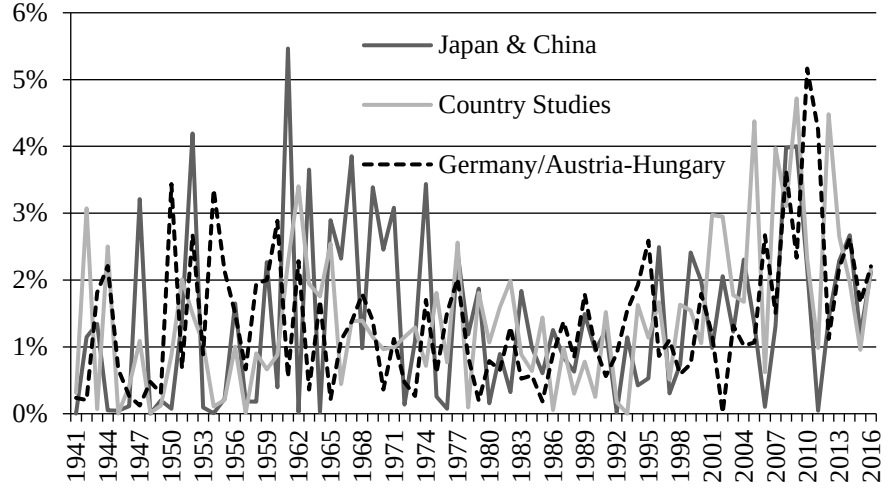
Topic models provide a useful tool for reducing complexity, identifying relevant sources, and generating new research questions. By their very nature, they possess the unifying potential of interdisciplinary scholarship. As "the future of economic history must be interdisciplinary" (Lamoreaux 2015, p. 1251), topic models are one step in securing the significance of economic history. If it is true that "our tribe has been particularly adept at drawing on metaphors, tools, and theory from a variety of disciplines" (Mitchener 2015, p. 1238), economic history should use this ability and integrate digital tools like topic models in its toolkit. Building on the distinct propensity to empirical work, digitization will not be a threat but rather a chance for economic history to become a role model for uniting traditional quantitative analysis, digital methods, and, by a return to some "old" economic historians' virtues, thorough text analysis. As Collins (2015, p. 1232) puts it: "It may […] improve the economic history that we write by ensuring our exposure to state-of-the-art methods and theory." This article hopes to contribute some of this exposure.

Appendix 1: Annual Average Topic Shares Sample Two

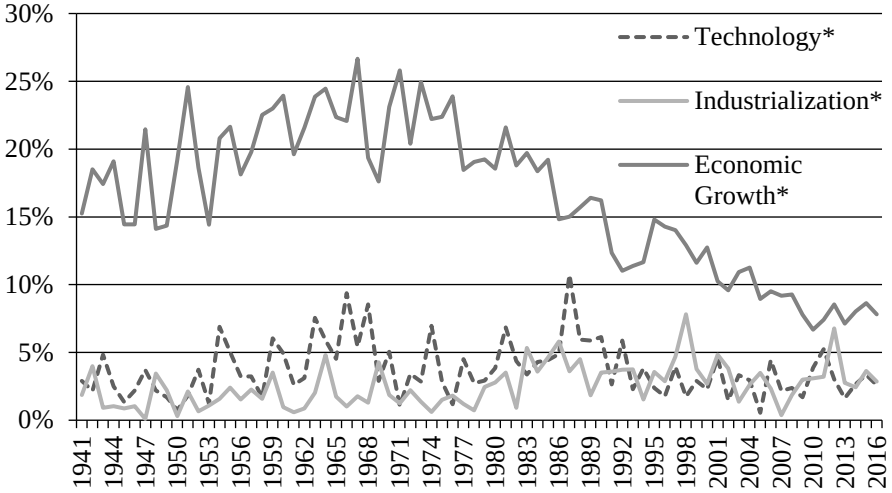
APPENDIX FIGURE 1



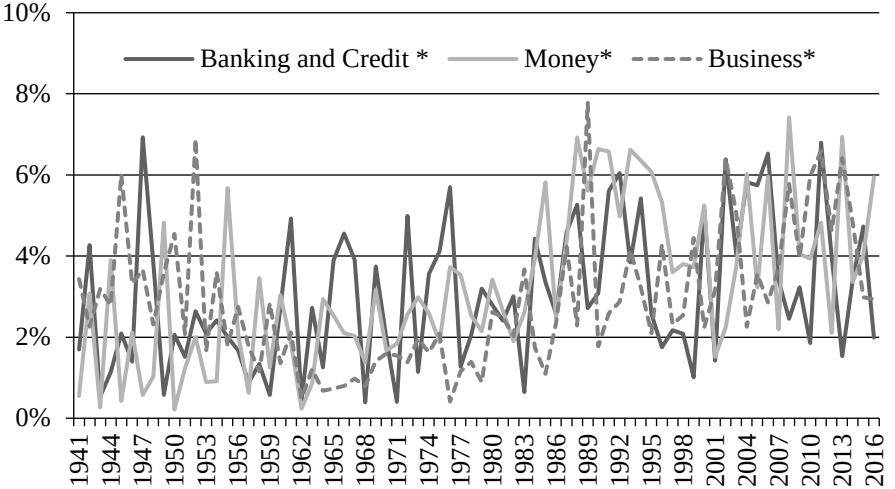
APPENDIX FIGURE 2



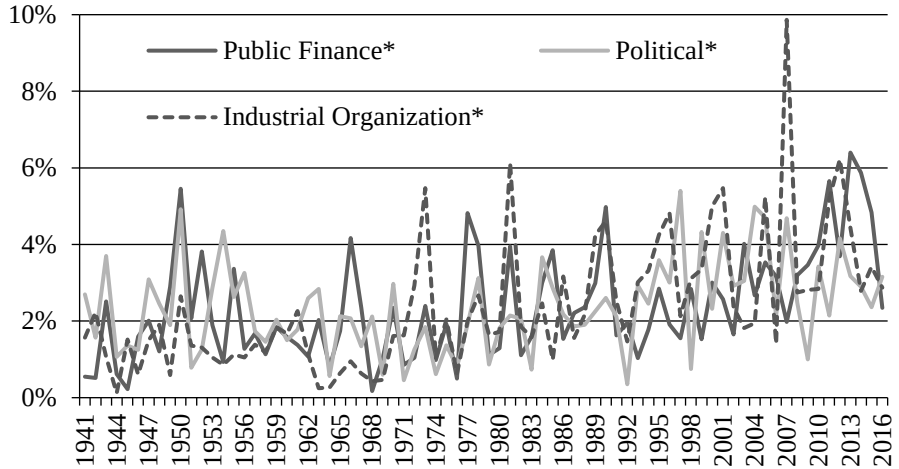
APPENDIX FIGURE 3



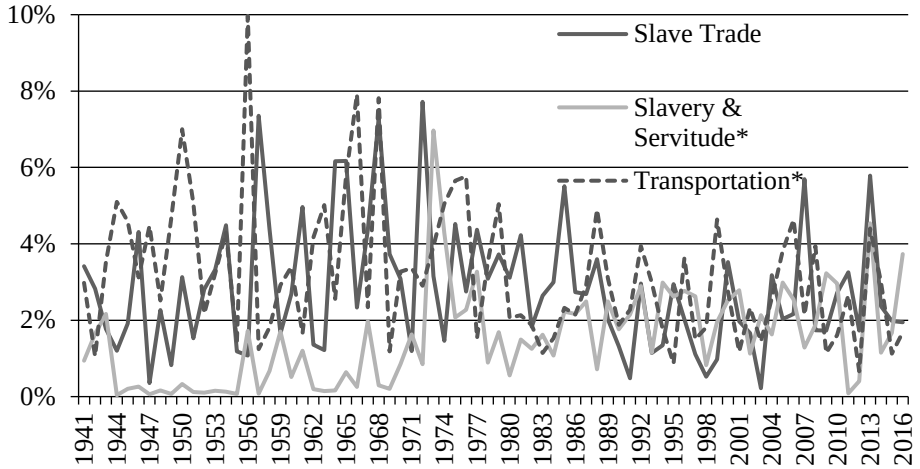
APPENDIX FIGURE 4



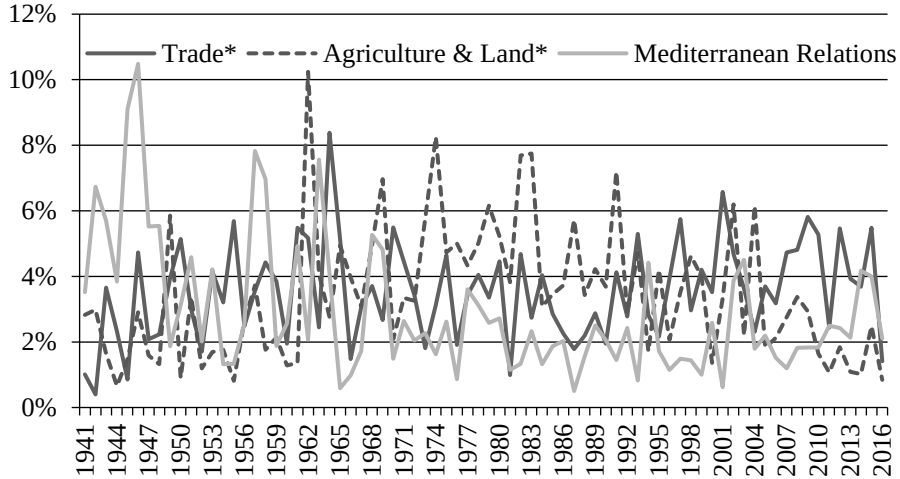
APPENDIX FIGURE 5



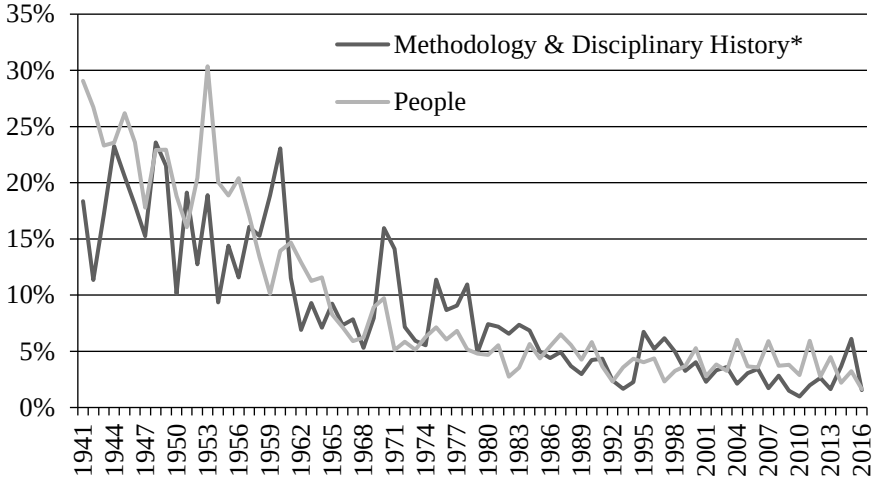
APPENDIX FIGURE 6



APPENDIX FIGURE 7



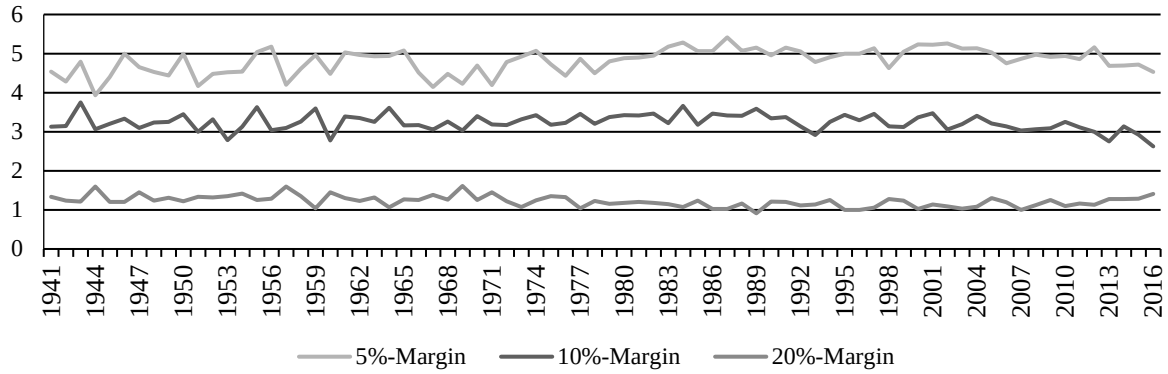
APPENDIX FIGURE 8



Notes: Asterisks mark labels used in Whaples (2002)
Sources: Author's own computations.

Appendix 2: Substantial Topics

APPENDIX FIGURE 9
ANNUAL AVERAGE NUMBER OF SUBSTANTIAL TOPICS PER DOCUMENT



Notes: A topic was defined as substantial if the corresponding per document share reaches 5% or more (10%, 20%).

Sources: Author's own computations.

Appendix 3: Excluding Task Issues

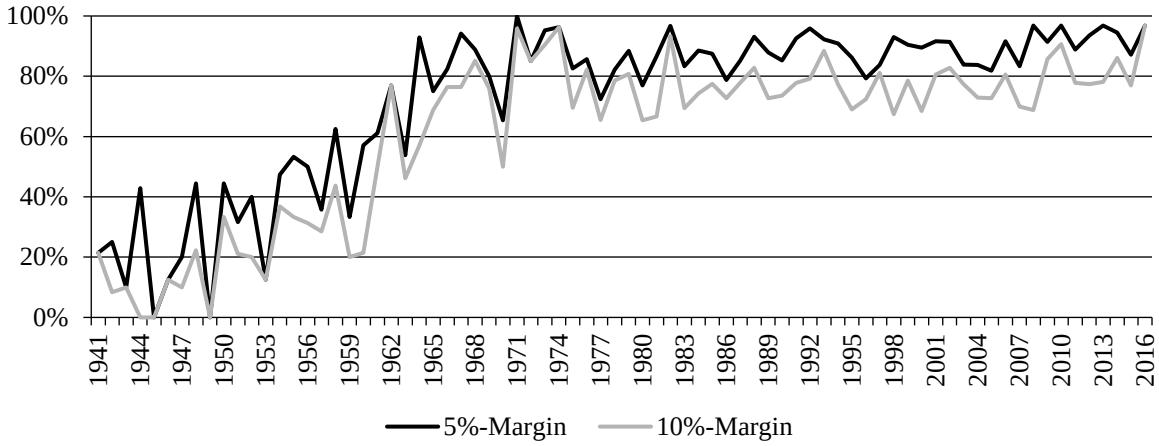
A topic model with 25 topics is applied on all articles published between 1941 and 2016 excluding Task issues as identified by Diebolt and Haupert (2017) which reduces the corpus from 2,675 to 1,885 documents. Again, the topic model identifies two topics that can be interpreted as representing quantitative methods. The 30 most probable words of the quantitative topics are shown in Appendix Table 1. Articles are defined as quantitative if they exhibit a share of one of the two quantitative topics of 5% (10%) or more (Appendix Figure 13).

APPENDIX TABLE 1
QUANTITATIVE TOPICS

percent price prices table period rate data average total years rates year increase series estimates time demand Figure index Figures income decline cost annual relative estimate change evidence supply ratio
data results variables variable table significant effects effect model economic sample level time percent coefficient regression analysis equation coefficients average number test standard change year market positive regressions estimated dummy

Notes: 30 most probable words in descending order.
Sources: Author's own computations.

APPENDIX FIGURE 10
SHARE OF QUANTITATIVE ARTICLES



Notes: Number of quantitative documents per year divided by the overall number of documents per year.
Sources: Author's own computations.

REFERENCES

Abramitzky, Ran. "Economics and the Modern Economic Historian." *Journal of Economic History* 75, no. 4 (2015): 1240–51.

Arguing with Digital History working group. *Digital History and Argument* white paper, Roy Rosenzweig Center for History and New Media (November 13, 2017). https://rrchnm.org/argument-white-paper/.

Bellstam, Gustaf, Sanjai, Bhagat, and Cookson, J. A. "A Text-Based Analysis of Corporate Innovation." SSRN Working Paper No. 2803232, May 2017.

Blei, David, Ng, Andrew Y., and Jordan, Michael I. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (2003): 993–1022.

Blei, David M. "Probabilistic Topic Models." *Communications of the ACM* 55, no. 4 (2012a): 77–84.

Blei, David M. "Topic Modeling and Digital Humanities." *Journal of Digital Humanities* 2, no. 1 (2012b): 8–11.

Blei, David M., and Lafferty, John D. "Dynamic Topic Models." *Proceedings of the 23rd international Conference on Machine Learning* (2006): 113–20.

Blei, David M., and Lafferty, John D. "A Correlated Topic Model of Science." *The Annals of Applied Statistics* 1, no. 1 (2007): 17–35.

Blei, David M., and Lafferty, John D. "Topic Models." In *Text mining. Classification, Clustering, and Applications*, edited by Ashok N. Srivastava, and Mehran Sahami, 71–93. Boca Raton, FL: CRC Press, 2009.

Bonilla, Tabitha, and Grimmer, Justin. "Elevated Threat Levels and Decreased Expectations. How Democracy Handles Terrorist Threats." *Poetics* 41, no. 6 (2013): 650–69.

Boyd-Graber, Jordan, Mimno, David, and Newman, David J. "Care and Feeding of Topic Models." In *Handbook of Mixed Membership Models and Their Applications*, edited by Edoardo M. Airoldi, David M. Blei, Elena A. Erosheva, and Stephen E. Fienberg, 225–74. Boca Raton: Taylor & Francis, 2015.

Burguière, André. *The Annales school: An Intellectual History.* Ithaca: Cornell University Press 2009.

Chang, Jonathan, Boyd-Graber, Jordan, Wang, Chong, Gerrish, Sean, and Blei, David M. "Reading Tea Leaves: How Humans Interpret Topic Models." *Advances in Neural Information Processing Systems (2009)* (2009): 288–96.

Collins, William J. "Looking Forward: Positive and Normative Views of Economic History's Future." *Journal of Economic History* 75, no. 4 (2015): 1228–33.

Diebolt, Claude, and Haupert, Michael. "A Cliometric Counterfactual: What if There Had Been Neither Fogel nor North?" *Cliometrica (forthcoming)* (2017).

DiMaggio, Paul, Nag, Manish, and Blei, David. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture. Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41, no. 6 (2013): 570–606.

Fligstein, Neil, Brundage, Jonah S., and Schultz, Michael. "Why the Federal Reserve Failed to See the Financial Crisis of 2008: The Role of "Macroeconomics" as Sense-Making and Cultural Frame." IRLE Working Paper No. 111-14, Berkeley, September 2014.

Freeman Smith, Robert. "The Formation and Development of the International Bankers Committee on Mexico." *Journal of Economic History* 23, no. 4 (1963): 574–86.

Gentzkow, Matthew, Kelly, Bryan T., and Taddy, Matt. "Text as Data." NBER Working Paper No. 23276, Cambridge, MA, March 2017.

Goodrich, Carter. "Economic History: One Field or Two?" *Journal of Economic History* 20, no. 4 (1960): 531–38.

Graham, Shawn, Milligan, Ian, Weingart, Scott B. *Exploring Big Historical Data: The Historian's Macroscope.* London: Imperial College Press 2016.

Grajzl, Peter, and Murrell, Peter. "A Structural Topic Model of the Features and the Cultural Origins of Bacon's Ideas." CESifo Working Paper No. 6643, October 2017.

Griffiths, Thomas L., and Steyvers, Mark. "Finding scientific topics." *PNAS* 101, no. 1 (2004): 5228–35.

Grimmer, J., and Stewart, B. M. "Text as Data. The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21, no. 3 (2013): 267–97.

Grimmer, Justin. "A Bayesian Hierarchical Topic Model for Political Texts. Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18, no. 1 (2010): 1–35.

Hall, David, Jurafsky, Daniel, and Manning, Christopher D. "Studying the history of ideas using topic models." *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2008): 363–71.

Hansen, Stephen, and McMahon, Michael. "Shocking Language: Understanding the Macroeconomic Effects of Central Bank Communication." CAMA Working Paper No. 4-2016, Canberra, January 2016.

Hansen, Stephen, McMahon, Michael, and Prat, Andrea. "Transparency and Deliberation within the FOMC: a Computational Linguistics Approach." CEPR Discussion Paper No. 9994, London, June 2014.

Haupert, Michael. "History of Cliometrics." In *Handbook of Cliometrics*, edited by Claude Diebolt, and Michael Haupert, 3–21. Berlin, Heidelberg: Springer Reference, 2016.

Hockey, Susan. "The History of Humanities Computing." In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 3–19. Malden, Mass.: Blackwell Publ., 2004.

Jacobi, Carina, van Atteveldt, Wouter, and Welbers, Kasper. “Quantitative analysis of large amounts of journalistic texts using topic modelling.” *Digital Journalism* 4, no. 1 (2015): 89–106.
Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History.* Urbana Ill.: Univ. of Illinois Press 2013.
Jockers, Matthew L. *Text Analysis with R for Students of Literature.* Cham: Springer 2014.
Lamoreaux, Naomi. “The Future of Economic History Must Be Interdisciplinary.” *Journal of Economic History* 75, no. 4 (2015): 1251–57.
Larsen, Vegard H., and Thorsrud, Leif A. “The Value of News.” CAMP Working Paper No 6/2015, Oslo, October 2015.
Larsen, Vegard H., and Thorsrud, Leif A. “Asset Returns, News Topics, and Media Effects.” CAMP Working Paper No 5/2017, Oslo, September 2017.
Lüdering, Jochen, and Tillmann, Peter. “Monetary policy on Twitter and its effect on asset prices: Evidence from computational text analysis.” Joint Discussion Paper Series in Economics No. 12-2016, Marburg, March 2016.
Lüdering, Jochen, and Winker, Peter. “Forward or Backward Looking? The Economic Discourse and the Observed Reality.” *Journal of Economics and Statistics* 236, no. 4 (2016): 483–515.
Margo, Robert A. “The Integration of Economic History Into Economics.” NBER Working Paper No. 23538, Cambridge, MA, June 2017.
McCallum, Andrew 2002. *"MALLET: A Machine Learning for Language Toolkit*.
McCloskey, Donald. “Does the Past Have Useful Economics.” *The Journal of Economic Literature* 14, no. 2 (1976): 434–61.
McFarland, Daniel A., Ramage, Daniel, Chuang, Jason, Heer, Jeffrey, Manning, Christopher D., and Jurafsky, Daniel. “Differentiating language usage through topic models.” *Poetics* 41, no. 6 (2013): 607–25.
Meeks, Elijah, and Weingart, Scott B. “The Digital Humanities Contribution to Topic Modeling.” *Journal of Digital Humanities* 2, no. 1 (2012): 2–6.
Miller, Ian M. “Rebellion, Crime and Violence in Qing China, 1722–1911. A topic Modeling Approach.” *Poetics* 41, no. 6 (2013): 626–49.
Mimno, David. “Computational Historiography: Data Mining in a Century of Classics Journals.” *ACM Journal on Computing and Cultural Heritage* 5, no. 1 (2012): 1–19.
Mimno, David, Wallach, Hanna M., Naradowsky, Jason, Smith, David A., and McCallum, Andrew. “Polylingual Topic Models.” *EMNLP 2009* (2009): 880–89.
Miner, Gary. *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications.* Amsterdam: Elsevier/Academic Press 2012.
Mitchener, Kris J. “The 4D Future of Economic History: Digitally-Driven Data Design.” *Journal of Economic History* 75, no. 4 (2015): 1234–39.
Mohr, John W., and Bogdanov, Petko. “Introduction - Topic models. What They Are and Why They Matter.” *Poetics* 41, no. 6 (2013): 545–69.
Moretti, Franco. *Distant Reading.* London, New York: Verso 2013.
Newman, David J., and Block, Sharon. “Probabilistic Topic Decomposition of an Eighteen-Century American Newspaper.” *Journal of the American Society for Information Science and Technology* 57, no. 6 (2006): 753–67.
Quinn, Kevin M., Monroe, Burt L., Colaresi, Michael, Crespin, Michael H., and Radev, Dragomir R. “How to Analyze Political Attention with Minimal Assumptions and Costs.” *American Journal of Political Science* 54, no. 1 (2010): 209–28.

Riddell, Allen B. “How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models.” In *Distant readings*. *Topologies of German Culture in the Long Nineteenth Century*, edited by Matt Erlin, and Lynne Tatlock, 91–113. Suffolk: Boydell & Brewer, 2014.

Schofield, Alexandra, Magnusson, Mans, and Mimno, David. “Pulling Out the Stops: Rethinking Stopword Removal for Topic Models.” *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, no.* 2 (2017): 432–36.

Shirota, Yukari, Hashimoto, Takako, and Sakura, Tamaki. “Topic Extraction Analysis for Monetary Policy Minutes of Japan in 2014. Effects of the Consumption Tax Hike in April.” In *Advances in Data Mining: Applications and Theoretical Aspects*, edited by Petra Perner, 141–52. Cham: Springer, 2015.

Steyvers, Mark, and Griffiths, Tom. “Probabilistic Topic Models.” In *Handbook of Latent Semantic Analysis*, edited by Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch, 427–48. Hoboken: Taylor and Francis, 2007.

Tetlock, Paul C. “Giving Content to Investor Sentiment: The Role of Media in the Stock Market.” *The Journal of Finance* 62, no. 3 (2007): 1139–68.

Thorsrud, Leif A. “Nowcasting Using News Topics. Big Data versus Big Bank.” Norges Bank Working Paper 20/2016, Oslo, December 2016a.

Thorsrud, Leif A. “Words are the New Numbers: A Newsy Coincident Index of Business Cycles.” Norges Bank Working Paper 21/2016, Oslo, December 2016b.

Wallach, Hanna M. “Topic Modeling: Beyond Bag of Words.” *Proceedings of the 23rd international Conference on Machine Learning* (2006): 977–87.

Wallach, Hanna M., Mimno, David, and McCallum, Andrew. “Rethinking LDA: Why Priors Matter.” *Advances in Neural Information Processing Systems 22* (2009): 1973–81.

Walters, Philip G., and Walters, Raymond. “The American Career of David Parish.” *Journal of Economic History* 2, no. 2 (1944): 149–66.

Whaples, Robert. “A Quantitative History of the Journal of Economic History and the Cliometric Revolution.” *Journal of Economic History* 51, no. 2 (1991): 289–301.

Yang, Tze-I, Torget, Andrew J., and Mihalcea, Rada. “Topic Modeling on Historical Newspapers.” *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (2011): 96–104.

TABLE 1

LITERATURE REVIEW SUMMARY

PAPER	DATABASE	TIME	TOPIC	DEPARTMENT
Bellstam et al. (2017)	Analyst reports on S&P 500 firms	1990-2012	Measuring firms' inventive activities	Finance
Blei and Lafferty (2007)*	*Science*	1990-1999	Topic scouting/TM methodology	Computer Science
Bonilla and Grimmer (2013)[a]	Multiple U.S. newspapers, transcripts of newscasts	2002-2005	Influence of terror alerts on public opinion	Political Science
DiMaggio et al. (2013)	Multiple U.S. newspapers	1986-1997	Coverage of U.S. public financial assistance to arts in newspapers	Sociology, Computer Science
Fligstein et al. (2014)	Federal Open Market Committee minutes	2000-2008	FOMCs perception of the financial crisis in 2008	Sociology
Grajzl and Murrell (2017)*	Multiple writings by Francis Bacon	ns	Identifying features and origins of Francis Bacon's ideas	Economics
Grimmer (2010)*	US Senate press releases	2007	Identifying politicians' agendas	Government
Hall et al. (2008)	Association for Computational Linguistics (ACL) Anthology	1978-2006	Disciplinary History	Symbolic Systems, Linguistics, Computer Science
Hansen et al. (2014)	Federal Open Market Committee minutes	1987-2006	Effects of transparency on monetary policy	Economics
Hansen and McMahon (2016)	Federal Open Market Committee statements	1998-2014	Effects of central bank communication on macroeconomic and financial variables	Economics
Jacobi et al. (2015)	New York Times	1945-2013	Press coverage of nuclear technology	Communication Science
Jockers (2013)	Fiction from the U.S. and Great Britain	1750-1899	Topic scouting/topic analysis	English
Larsen and Thorsrud (2015)	Business Newspaper (*Dagens Næringsliv*)	1988-2014	Forecasting macroeconomic data	Economics

Larsen and Thorsrud (2017)	Business Newspaper (*Dagens Næringsliv*)	1988-2014	Effects of news on asset prices	Economics
Lüdering and Tillmann (2016)	Twitter messages referring to the Fed	2013	Measuring expectations of monetary policy and its effects on asset prices	Economics
Lüdering and Winker (2016)	Journal of Economics and Statistics	1949-2010	Time perspective of economic research/Disciplinary History	Economics
Miller (2013)	Crime Reports from Chinese Administration	1722-1911	Analysis of the nature of unrest and violence in Qing China	East Asian Languages and Civilizations
Mimno (2012)*	Multiple Classics Journals	1850-2006	Topic scouting/TM methodology	Computer Science
Newman and Block (2006)*	Colonial newspaper (*Pennsylvania Gazette*)	1728-1800	Topic scouting/TM methodology	Computer Science, History
Quinn et al. (2010)*	Speeches in the U.S. Senate	1995-2004	Measuring political attention	Law, Political Science, Computer Science
Riddell (2014)	Multiple US-based German Studies Journals	1928-2006	Topic scouting/Disciplinary History	Computational Science
Shirota et al. (2015)	Minutes of Meetings of the Bank of Japan	2014	Effects of consumption tax increase on monetary policy	Economics
Thorsrud (2016a, 2016b)	Business Newspaper (*Dagens Næringsliv*)	1988-2014	Estimating business cycles based on news	Economics
Yang et al. (2011)	Multiple Texan Newspapers	1989-2008	Topic scouting/TM methodology	Computer Science/Engineering, History

Notes: "Topic scouting" refers to papers which apply topic models with the primary goal if identifying topics. "TM methodology" refers to papers that apply topic models to discuss methodological issues. Asterisks mark papers using a different topic model than LDA. Department is recorded according to authors' affiliations.

TABLE 2

TOPICS IN THE JOURNAL OF ECONOMIC HISTORY 1941-1990

#	Most Probable Words	Label	ATS
0	company firms industry oil companies production firm industries coal industrial steel market american research competition standard u.s electric manufacturing large sales small business corporation plant size plants largest petroleum gas	Other Industry Studies*	1.75 %
1	japanese japan china chinese rice land period government meiji tokugawa tokyo development agricultural tax modern economic taiwan osaka irrigation samurai merchants rural century village modem population history han shanghai traditional	Country Studies, Japan and China	0.9
2	prices price trade demand goods exports market supply imports production export index period products terms years commodities consumption decline increase rise year real percent rose increased markets century domestic commodity	Prices*	2.7
3	economic history historical work historians theory study analysis studies point question discussion problem research view questions problems book review fact historian data professor past approach time evidence general economics recent	History of Economic History*	8.7
4	railroad railroads canal transportation construction cost railway costs canals social miles western freight river railways pacific water road ohio rail roads lines erie traffic transport improvements line system central rates	Transportation*	1.3
5	growth income capita economic real rate output population agricultural estimates national product labor percent sector agriculture economy consumption increase change share century productivity rates force farm gross distribution index relative	Economic Growth*	2.6
6	percent table data total average period year years estimates number rate annual rates source series Figures Figure sources time estimate large index based estimated statistics appendix percentage made increase ratio	Descriptive Language/Time Series	7.9
7	bank banks banking deposits reserve national notes system state credit financial money loans deposit federal commercial capital assets states reserves bankers private savings newyork specie country monetary bank's currency free	Banking*	1.6
8	capital investment long series united growth depression british cycle fluctuations cycles migration movements states period economic american population swings business construction building emigration economy demand australia expansion unemployment net great	Business Cycles*	1.5
9	states american united newyork state massachusetts boston america u.s philadelphia washington pennsylvania england war north journal early john national historical history dollars report americans james connecticut robert d.c william thomas	Country Studies*, U.S.	2.1
10	labor workers union unions strike national industrial employers strikes trade welfare insurance industry members work hours social wages association benefits organization committee management collective bargaining worker unemployment employer a.f local	Labor*, Labor Relations	0.9
11	agricultural agriculture wheat grain farmers crops yields crop farm land production farming food output yield productivity bushels corn dairy harvest animals cattle livestock acre enclosure milk animal meat labor grains	Agriculture*	1.6
12	social political society class theory capitalism wealth life men man revolution thought classes power human state marx labor economy keynes economics great capital economists ideas capitalist principles religious natural free	History of Economic Thought*	2.7
13	law state public government rights laws property political private act interests legal court legislation policy power protection constitution regulation general economic acts courts support vote crown corporations interest cases issue	Law*	1.9
14	capital interest market rates investment rate financial stock percent loans debt loan credit funds bonds securities return company companies million assets mortgage investments finance shares london markets money exchange investors	Finance	2.2
15	trade british ships african slave africa ship coast vessels traders slaves voyage shipping century european liverpool west sailing profits freight cargo port voyages goods ports shipbuilding dutch gold herring sea	Trade*, Slave Trade	1.2
16	italian italy genoese century medieval venice medici florence merchants del venetian bruges genoa merchant fourteenth della storia business florentine rome fifteenth milan ages roover commercial thirteenth wool insurance cloth branch	Country Studies*, Italy	1.0

17	economic countries foreign industrial development industry british world growth capital trade international europe country united european domestic investment britain production industries policy national states industrialization war great period france germany	International Investment*	3.3
18	slaves slave slavery labor free south contract southern servants north engerman sugar white fogel emancipation cost plantation indentured war civil negro work black costs servant planters history market freedom plantations	Slavery and Servitude*	1.0
19	iron steam steel power engine machine industry engines production coal patent patents invention water machines tons pig machinery fuel technology technological products diffusion inventions cost furnaces furnace early process technical	Power/Energy Industries*	1.5
20	canadian canada mexico spain spanish latin america brazil madrid mexican brazilian ontario toronto quebec american century royal chile wool rio government mining castile reciprocity indigo percent seville del crown toledo	Country Studies*, Colonies	0.8
21	labor capital productivity output costs production cost factor change american rate technical relative manufacturing scale industry input efficiency inputs technology wage united technological prices british price higher states function economies	Manufacturing*	2.3
22	gold money exchange monetary currency silver specie standard rate foreign price paper treasury coins inflation coin notes mint circulation dollar real supply bills international market interest prices period series rates	Money*	1.6
23	time made part years large great fact found important small general make long system number end place good times people high early brought order period case means hand country set	not specified	12.3
24	colonial colonies trade tobacco british merchants england sugar american west london english indies shipping britain tonnage planters pounds maryland virginia vessels merchant chesapeake america exports great revolution north middle south	Trade*, Imperialism/Colonialism*, North Atlantic	1.4
25	labor workers wage wages women force work earnings employment percent census men children occupations skilled age school female immigrants male job black unskilled immigration occupational occupation education jobs participation schooling	Labor*	2.2
26	german der germany und die des industry von berlin industrial austria hungary austrian hungarian prussian das zur growth deutschen monarchy tariffs geschichte habsburg protection steel marks iron customs development prussia	Country Studies*, Germany/Austria-Hungary	0.9
27	cotton farm farms farmers south agricultural agriculture labor land southern crop tenants production acreage farmer census acres plantation california tenant tenancy crops georgia counties size states contracts share acre county	Agriculture*, Cotton	1.3
28	south regional regions region cities urban population city north west southern areas states growth central development state local eastern differences market western national east census differentials antebellum northeast interregional atlantic	Regional Studies, Geographic Descriptions	1.9
29	russian land russia peasant peasants labor century serfs serfdom serf europe population village moscow lord medieval estates rubles agricultural system rural estate agrarian feudal peasantry services demesne rent petersburg manorial	Country Studies*, Russia	1.2
30	war government tax expenditures public state policy federal taxes military private income fiscal percent national revenue revenues million budget debt political finance administration controls inflation policies local civil program army	Public Finance*, War*	1.9
31	variables variable model results level equation data significant hypothesis coefficient regression effect coefficients test demand time income positive rate equations expected sample values analysis effects economic u.s evidence estimated function	Econometric Language	3.8
32	population age wealth mortality family fertility children life birth rates families death marriage demographic century number women sample rural living england health household deaths county social rate growth decline households	Demography*	2.0
33	england english century british poor london britain revolution relief wages eighteenth industrial irish history wage evidence counties law wales ireland population early parliamentary laborers nineteenth oxford scotland great parish parishes	Industrialization*, Great Britain	1.4
34	economic development economy growth system change process social market political structure role institutions century institutional major organization systems production markets resources traditional conditions control analysis society problems important individual early	Economic Growth*	7.1
35	french france paris century des les dutch revolution europe eighteenth amsterdam english seventeenth van francs annales histoire sur history archives england livres european economique crisis louis holland vols revue netherlands	Country Studies*, France	1.3

36	company committee march january papers june report december april office letter february october september august city november president john house congress business secretary board letters treasury year received directors plan	People	2.5
37	business history research economic study enterprise american university men years field development entrepreneurs company committee enterprises entrepreneurial businessmen management group public general entrepreneur records individual published entrepreneurship corporation social institutions	Business*	2.5
38	land lands india indian acres settlement iowa county illinois frontier acre western cattle price prairie federal settlers area speculators farm grant grants counties kansas large property state sales history west	Imperialism/Colonialism*,Westward Movement	1.3
39	cotton industry textile mills cloth factory spinning production mill firms workers england textiles looms quality factories silk machinery manufacturing manufacturers labor manufacture industrial weaving industries work woolen weavers learning yarn	Industrialization*, Textile Industry	1.3
40	empire trade merchants greek ancient jewish roman ottoman century egypt merchant economic world balkan jews east greece byzantine commerce palestine traders state evidence silver arab b.c greeks goods mediterranean middle	Trade*, Ancient Trade	0.6

Notes: The table shows the 30 most probable words for every topic in descending order. ATS stands for average topic share over the corpus in percent. Asterisks mark labels used by Whaples (1991). # Marks the topic number randomly given by MALLET.

Sources: Author's own computations.

TABLE 3

TOPICS IN THE JOURNAL OF ECONOMIC HISTORY 1941-2016

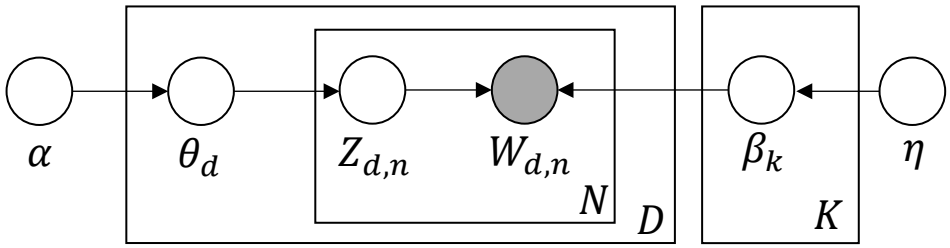
#	Most Probable Words	Label	ATS
0	age wealth population percent family children migration women table household fertility income census families marriage immigrants number households sample rates men years states rural migrants inequality economic data total mortality	Demography*	3.7%
1	japanese japan india china chinese indian rice government period development land asia tokyo century economic population modern history meiji economy asian price early tokugawa prices taiwan osaka agricultural cotton system	Asia	1.4
2	education health mortality school percent height schools disease rates population schooling birth public age states human years high educational children data rate water united income malaria heights diseases life death	Standard of Living and Health*	1.7
3	industry production cotton iron industries technology firms manufacturing power industrial textile mills costs steam machinery american output technological steel technical productivity cost machine british coal machines percent spinning capital cloth	Industrialization*, Technology*	3.8
4	cotton south slaves slave black southern slavery white labor blacks carolina north states free war plantation american whites georgia racial civil race antebellum history negro fogel engerman northern state percent	Slavery and Servitude*	1.7
5	land agricultural farm agriculture farmers wheat production farms labor percent grain crop acres prices crops yields farming acre productivity tenants output cattle harvest average acreage price corn yield number year	Agriculture and Land*	3.7
6	data table results variables variable significant sample effect model percent effects level economic time average coefficient number regression analysis equation coefficients change year estimates test estimated information evidence journal period	Econometric Language	7.0
7	bank banks banking loans credit financial state national reserve deposits capital newyork states interest percent rates loan market system federal notes assets funds deposit commercial money insurance rate total bankers	Banking and Credit*	3.1
8	growth percent income prices output table price estimates data rate series period real index productivity economic total capital labor average capita production rates relative national year industrial united consumption years	Descriptive Language	7.8
9	trade british colonial colonies ships slave percent west african slaves shipping tobacco dutch ship africa vessels merchants century prices price american sugar coast servants london eighteenth america english history north	Slave Trade	2.9

10	railroad states railroads regional west transportation cities south city newyork american region regions united construction costs canal cost state western north central railway ohio urban rates percent east railways miles	Transportation*	3.1
11	made business years time american company general committee great government men part john papers found year public make war office march order fact trade william records good january june april	People	7.7
12	war tax government public state taxes percent military expenditures private fiscal revenues income soviet revenue spending total federal national policy control million local housing political years administration economy taxation budget	Public Finance*	2.4
13	state states political law american canadian united federal congress government laws u.s act vote canada public legislation policy support voting interests party national regulation report reform politics economic bill power	Political*	2.3
14	england revolution century english history eighteenth poor london british wages population early britain industrial living economic evidence irish common enclosure europe ireland medieval modern review society parish wales towns relief	Industrialization*, Great Britain	2.7
15	french france paris century des italy empire les ottoman italian merchants roman trade europe medieval middle early eighteenth egypt merchant centuries venice commercial european history rome genoese livres greek histoire	Mediterranean Relations	2.6
16	mexico russian latin russia spanish spain mexican brazil america government percent century economic sugar colonial madrid brazilian opium cuba development del foreign rio moscow political land argentina serf peasant peru	Country Studies	1.4
17	gold money exchange rate monetary market interest price rates debt percent currency prices silver standard financial stock government foreign period policy real inflation bonds coins specie london paper war crisis	Money*	3.4
18	german germany patents der patent und die des berlin industrial von invention inventors economic austria patenting inventions industry prussia coal inventive hungary percent prussian das zur deutschen deutsche habsburg market	Germany/Austria-Hungary	1.3
19	property rights law land legal contracts costs contract court institutions private cases trade institutional system crown common courts case enforcement rules company apprenticeship apprentices pay claims bay cost masters political	Industrial Organization*	2.5
20	economic history social work historical political business theory research historians society development world industrial study economics economy american studies class economists life great science institutions revolution production capitalism knowledge growth	Methodology and Disciplinary History*	7.5
21	economic growth period system development capital change time important case general fact made economy century part point conditions demand problem paper large analysis question long major evidence process discussion market	Economic Growth*	16.6
22	trade british countries united world foreign states exports britain international domestic economic imports tariff european prices price europe country export american goods percent france war kingdom markets market germany import	Trade*	3.5
23	firms company companies market business firm stock capital investment industry percent oil shares corporate price corporations large financial sales investors profits competition corporation share ownership number information limited insurance private	Business*	2.8
24	labor workers wage wages work employment earnings unemployment force women percent hours worker census working rates employers industrial men industry skilled jobs job unions average number employed time report manufacturing	Labor and Migration*	3.2

Notes: The table shows the 30 most probable words for every topic in descending order. ATS stands for average topic share over the corpus in percent. Asterisks mark labels used by Whaples (2002). # Marks the topic number randomly given by MALLET.

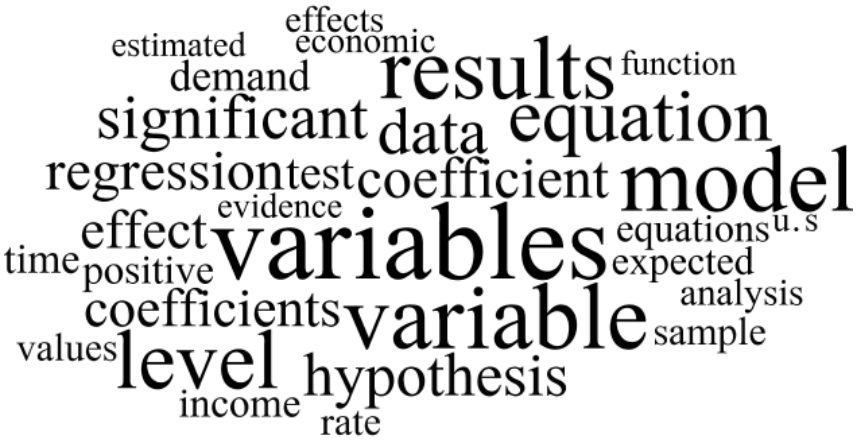
Sources: Author's own computations.

FIGURE 1
GRAPHICAL REPRESENTATION OF THE LATENT DIRICHLET ALLOCATION



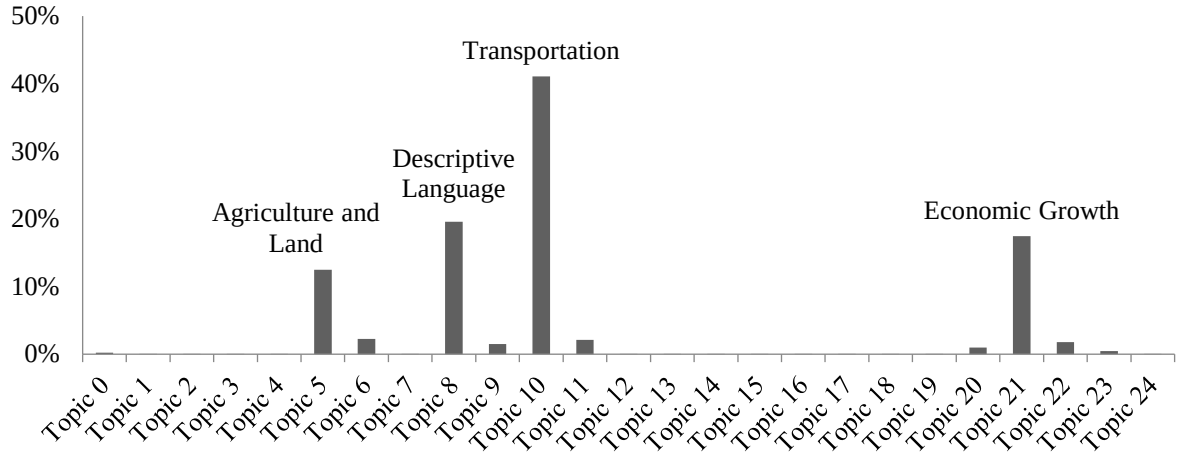
Sources: Blei (2012a)

FIGURE 2
TOPIC 31 ECONOMETRIC LANGUAGE



Sources: Author's own computations.

FIGURE 3
TOPIC DISTRIBUTION OF FOGEL (1962)



Sources: Author's own computations.

FIGURE 4.A
POSITIVE TOPIC CORRELATION

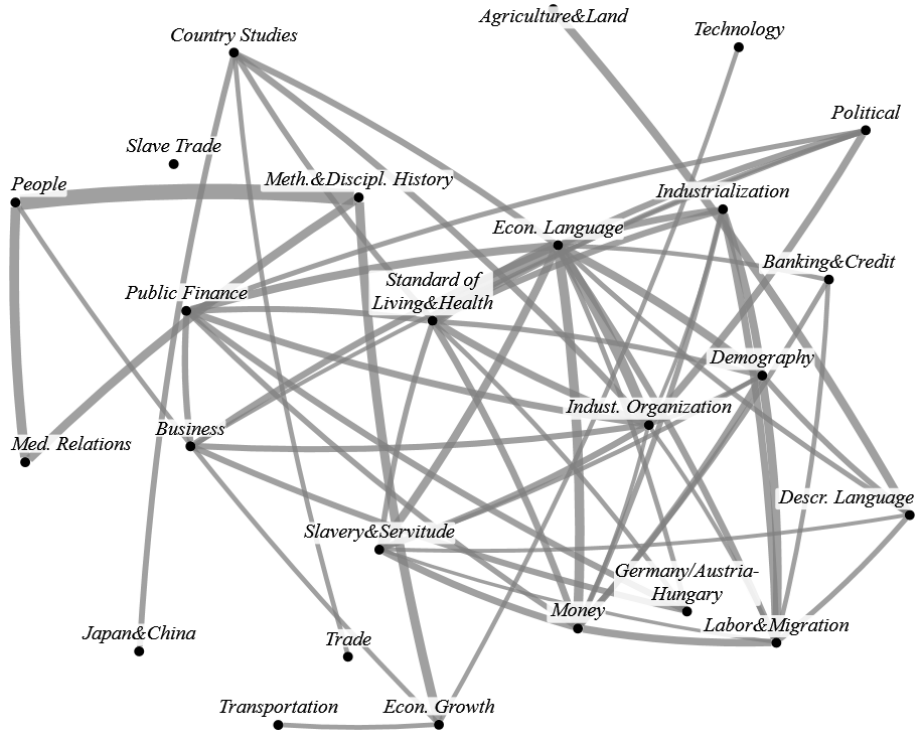
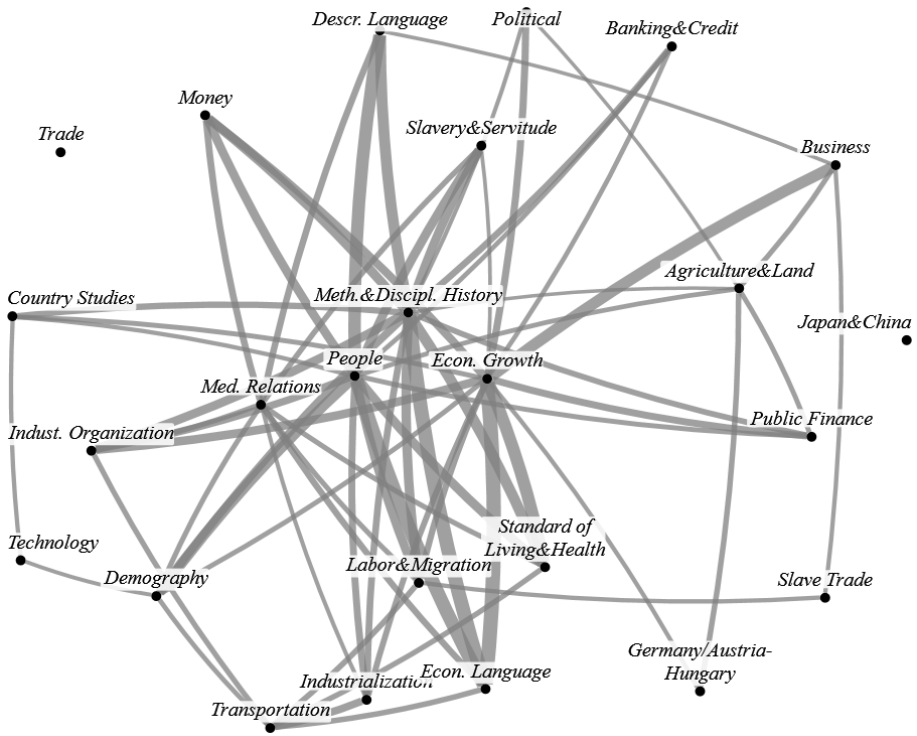
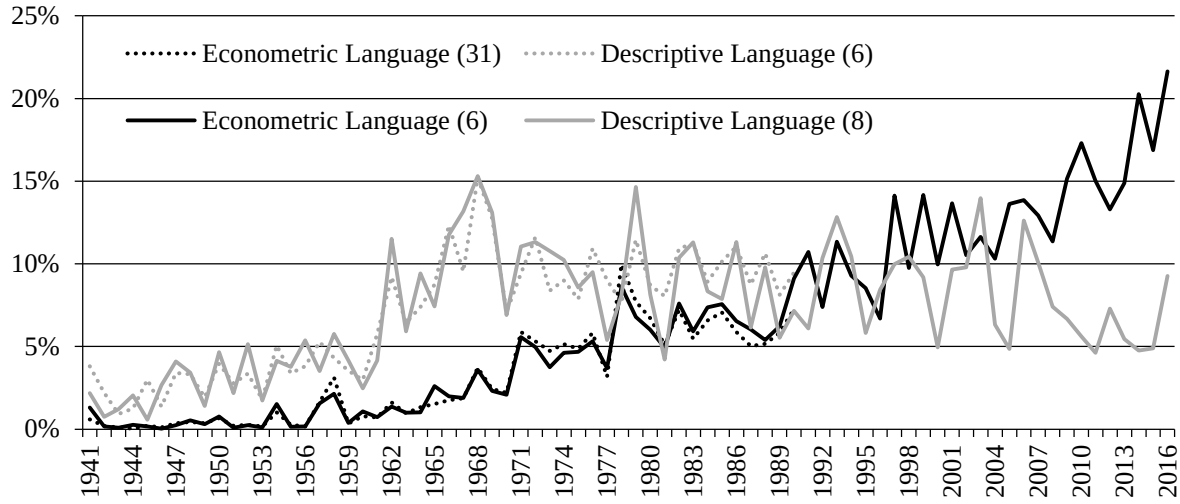


FIGURE 4.B
NEGATIVE TOPIC CORRELATION



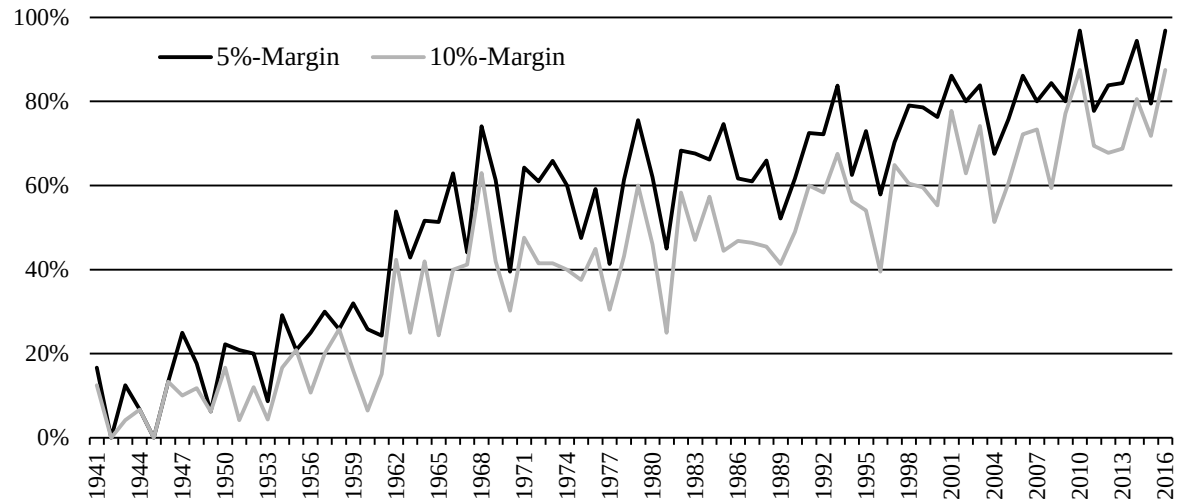
Notes: Width of connecting lines is proportionate to the value of the corresponding correlation coefficient including only coefficients wich are significant at the 5%-level. Correlation coefficients are computed based on annual average topic shares of sample 2.
Sources: Author's own computations.

Figure 5
ANNUAL AVERAGE TOPIC SHARES OF QUANTITATIVE TOPICS



Notes: Doted lines mark topics from sample 1, solid lines mark topics from sample 2.
Sources: Author's own computations.

Figure 6
SHARE OF QUANTITATIVE ARTICLES



Notes: Number of quantitative documents per year divided by all overall number of documents. Documents are classified as quantitative if their share of topic 6 or 8 is amount to 5% (10%) or more.
Sources: Author's own computations.