What topic modeling could reveal about the evolution of economics

DOI:10.1080/1350178X.2018.1529215

(Article begins on next page)

What Topic Modeling Could Reveal about the Evolution of Economics[1]

Angela **Ambrosino***, Mario **Cedrini***, John B. **Davis****, Stefano **Fiori***, Marco **Guerzoni***,*** and Massimiliano **Nuccio***

* Dipartimento di Economia e Statistica "Cognetti de Martiis", Università di Torino, Italy

** Marquette University and University of Amsterdam

*** ICRIOS, Bocconi University, Milan

ABSTRACT

The paper presents the topic modeling technique known as Latent Dirichlet Allocation (LDA), a form of text-mining aiming at discovering the hidden (latent) thematic structure in large archives of documents. By applying LDA to the full text of the economics articles stored in the JSTOR database, we show how to construct an intertemporal map of the discipline, and illustrate the potentialities of the technique for the study of the shifting structure of economics in a time of (possible) fragmentation.

KEYWORDS: Topic modeling, economics as science, economics literature, text analysis

JEL codes: B4, B1, B2, A1

[1] This work is part of a more general research project by Despina. Big Data Lab of the University of Turin, Dipartimento di Economia e Statistica "Cognetti de Martiis" (www.despina.unito.it). We gratefully acknowledge financial support from the European Society for the History of Economic Thought (ESHET grant 2015).

1. Introduction

Economics has long been criticized for its “imperialism”, that is for expanding its method and analytical vision into territories traditionally occupied by other disciplines. Starting from the mid-Seventies, the presumed “superiority” of economics – better, of its core, neoclassical approach – imposed itself in a number of fields, discouraging alternative approaches (Marchionatti and Cedrini 2017). An influential article by Fourcade, Ollion and Algan (2015) provides evidence of the persistent “insularity” and dominant position of economics within social sciences. The orthodoxy of the discipline (its dominant school of thought) is united around a recognizable theoretical core (utility maximization, emphasis on equilibrium, neglect of uncertainty) and the common method of mathematical formalism. And the high influence on public policy exerted by economists reflects also (among other factors) the fact that policy-makers tend to perceive the profession as a monolithic whole. Still, Fourcade herself (2018) recognizes that the “unity” of economics, a “truly generalistic form of expertise”, is “flexible” (Reay 2012). Somehow echoing Rodrik’s (2015) argument against the accusation of insularity – resting on the intrinsic variety of economic models, which “admit a wide variety of possibilities”, and on the “diversity” of ideas that exists within the profession – Fourcade argues that mainstream economics can be “malleable enough to incorporate waves of peripheral (and once rejected) ideas and concepts”.

There is ample evidence of the fact that economics is currently unified more by technique and epistemology than by core beliefs. Coats (2014, 383) has recently described economics as a “large and heterogeneous discipline” held together by “formalization and mathematization” but populated by “a number of dissenting or deviant doctrinal schools, rival methodological approaches, and innovative developments designed to remedy its defects and/or overcome its limitations”. Backhouse and Cherrier (2014) document that economics has become more applied since the Seventies, while Panhans and Singleton (2017) argue that economics has moved from a theory-based (key concepts) to a tool-based (admissible empirical practices) discipline. As to the issue of its intrinsic, internal variety of economics, there is a lively debate within the profession on the “pluralism” of today’s mainstream, reflecting increasing perception of the nature of the economic science as fragmented. Economics tends now to appear a more heterogeneous discipline, populated by an unprecedented plurality of research programmes that deviate from the neoclassical core and often originate from other disciplines’ “reverse imperialisms” (evolutionary game theory, behavioral economics, cognitive economics, experimental economics, neuroeconomics, complexity economics, and so on). It has been argued (Davis 2006) that this “pluralistic” state of mainstream economics may be a transitory phase in a Kuhnian cycle of scientific development (shaped by the succession of periods of monism and periods of pluralism) bound to reestablish the dominance of a new, post-neoclassical, mainstream. Others (Cedrini and Fontana 2017) suggest that “mainstream pluralism” is likely to persist over time under the impact of ever-growing specialization, resulting from the necessity of reducing the gap between scholars' competencies and the difficulty of reaching the frontier of economic research.

From this perspective, to use Kuhn's terms, the trend of growth in size and diversity might be transforming economics into an "immature" science. Research would occur within not one, but many local paradigms, which develop their own epochs of "normal science" (that is, of cumulative progress in *local* knowledge) before the discipline might experience (if ever) a revolutionary period. Exactly because it allows for the coexistence between alternative approaches, a Lakatosian framework (see for instance Colander, Holt and Rosser 2004), is now commonly employed to portray the state of economics as science, with the subset of non-"core" research programmes representing not a "protective belt" of applied research but rather the discipline's "periphery". Here, economics encounters other disciplines, shares with them assumptions and theoretical frameworks, and cooperates in the creation of the new research fields in today's "mainstream pluralism" (Davis 2006). As anticipated by the later Kuhn (2000), new knowledge in economics is produced (also, perhaps mainly) at the frontier, in the absence of scientific consensus. Despite the persistence of a "rough pyramidal hierarchy", "minarets … representing local confluences of authority" – borrowing from the "prospects for economics" illustrated by John Pencavel in 1991 (81) – seem to dominate the landscape, while the Kuhnian condition of relative isolation *cum* incommensurability that derives to subfields from specialization likely acts as a main driver of progress in the discipline.

Despite wide agreement (but not consensus: see for instance Dow 2008) that the structure of economics is changing, opinions diverge sharply when it comes to characterizing and explaining this change. There is little doubt that the ambition to investigate the exact nature of the evolution of the discipline is what motivates economists to devote increasing attention to the cartography of economics, starting from the "official" classification system developed by the American Economic Association (AEA) to list economic literature and scholars. The history of JEL codes provides in fact a "relevant proxy to understand the transformation of economics science throughout the twentieth century" (Cherrier 2017, 545) – as confirmed, for instance, by the 1991 revision (under Pencavel's leadership), that aimed to create a virtual map to help economists helping economists "navigate a growing and rapidly changing discipline" (ibid., 577). Still, classification systems of scientific knowledge "are best at monitoring the behavior of known and defined bodies of knowledge, but lend themselves poorly – if at all – to correctly identifying the emergence of truly new epistemic bodies of knowledge" (Suominen and Toivanen 2016, 2464). Whereas science maps – "generated through a scientific analysis of large-scale scholarly datasets in an effort to extract, connect, and make sense of the bits and pieces of knowledge they contain" – should help "identify major research areas, experts, institutions, collections, grants, papers, journals, and ideas in a domain of interest. They can show homogeneity vs. heterogeneity … and relative speed of progress. They allow us to track the emergence, evolution, and disappearance of topics and help to identify the most promising areas of research" (Börner et al. 2012).

Traditionally based on the use of bibliometric techniques such as citation networks, bibliographic coupling, and author co-citation analysis (for surveys, see Morris and Van der Veer Martens 2008 and Börner et al. 2012), the literature about mapping science can now exploit the availability of both databases and new, powerful, quantitative analytical techniques to investigate the changing structure of economics. For instance, Claveau and Gingras' study (2016) combines various algorithmic methods (applied to the Web of Science database) to investigate the shifting boundaries of economic specialties over time. Research fields are identified based on cognitive similarity between articles, which, in turn, derives from bibliographic coupling – similar documents exhibit a high proportion of overlap in their references. A dynamic network analysis then leads to identifying families of specialties and their life cycles – the result being that economics would show, today, fewer divisions than in the past. By using articles' metadata and a machine-learning algorithm trained on the dataset (which rests on Econlit, and Web of Science citation counts), Angrist et al. (2017) come to assign one of 10 fields (preselected on existing JEL codes) to every paper published since 1980 in some 80 journals, and classify articles according to their presumed theoretical, empirical or econometric style. They can thus conclude that microeconomics was and still is the largest field, despite some turbulence within it, and document a turn towards empirical work.

It has been observed that quantitative methods like the ones just mentioned face difficulties of both a methodological and meta-methodological kind – in a nutshell, standards for quantitative history are still to be settled, and a "healthy balance between statistical and qualitative evidence" to be found (see Cherrier 2015). Klaes (2017) notes, for instance, that the intention of letting specialties emerge from the data themselves clashes with the considerable freedom involved in the definition of the areas starting from the clustering technique employed in Claveau and Gingras' study. Likewise, Angrist et al.'s field classification uses as reference a pre-existing one, the JEL codes, and proposes a style classification that is quite arbitrary from any standpoint. Seeking to contribute to the general effort of drawing a multidimensional map of the discipline in historical perspective, this paper proposes a radically new quantitative approach to the history of economics, which – in view of the general limitations of the existing ones – aims to detect the hidden, or "latent" structure of the discipline.

.

2. A topic-modeling analysis of economics

The philosophy of topic modeling

Topic modeling is a form of text-mining aiming at discovering the hidden (latent) thematic structure in large archives of documents. The specific generative statistical model here used, Latent Dirichlet Allocation (LDA[2]), is a scalable basic tool – in machine learning and statistics, it is defined as a dimensionality reduction technique – and a fully probabilistic version of latent semantic analysis. LDA calculates probabilistic regularities, or trends in language texts, recurring themes in the form of co-occurring words. It groups words that compose the archive documents – on the assumption that words referring to similar subjects appear in similar contexts – into different probability distributions over the words of a fixed vocabulary. Being constellations, or sets of groups of words that are associated under one of the themes that run through the articles of the dataset, "topics" constitute the abovementioned latent (meaning inferred from the data; topics do not pre-exist the analysis) structures. The purpose served by LDA is to detect them by "reverse-engineering" the original intentions (that is, to discuss one or more specific themes) of the authors of the documents included in the corpus under examination (Mohr and Bogdanov 2013)[3]. LDA assumes that in the given corpus, all documents share the same set of topics (restricting attention to words with the highest estimated frequency), but that each document exhibits such topics in different proportions depending on words that are present in it (note that LDA generates topics and associates topics with documents at the same time).

For the sake of illustration, McCombie and Pike's (2013) article "No End to the Consensus in Macroeconomic Theory? A Methodological Inquiry", published in the *American Journal of Economics and Sociology*, exhibits five topics defined by the following groups of words[4]:

1. {economist, peopl, societi, challeng, concept}[5],
2. {shock, consumpt, monetari, output, money},
3. {wage, worker, labor, job, unemploy},
4. {inflat, forecast, monetari, output, bank},
5. {debt, fiscal, percent, save, spend}.

---

[2] On topic modeling, see the special issues of *Poetics*, 41(6), 2013, and of the *Journal of Digital Humanities*, 2(1), 2012. On LDA in particular, see Blei 2012a, Blei, Ng and Jordan 2003.

[3] Topic modeling algorithms discover the hidden structure that might be said to have generated the collection of observed documents, the utility of LDA stemming from the fact that "the inferred hidden structure resembles the thematic structure of the collection" (Blei 2012, 79), which thereby becomes manageable.

[4] Stemming and stop word filtering are recommended steps for topic modeling pre-processing. Stop words are some of the most common words, such as "the", "is", "at"; stemming refers to a set of methods used to normalize different tenses and variations of the same word (for example: unemployed and unemployment; inflation, inflates, inflated; etc.).

[5] Each of the five topics is defined by a group of 30 words. The first five words of each topic are those listed in brackets. The 25 words that follow in topic 1, for instance, are: human, crisi, think, labor, action, money, principl, capitalist, great, Marx, say, plan, sociolog, thought, class, profit, regul, common, global, book, object, law, complex, thing, Keyn.

The article discusses the New Neoclassical Synthesis after the global crisis, and explains the exclusion (from the New Consensus itself) of the Keynesian notion of involuntary employment (from deficient demand) on methodological grounds, that is, by throwing light on the "paradigmatic heuristic of the representative agent (497). The topics are defined by the five words that co-occur with high probability – the most probable words from each of the most probable topics. The article is associated with a topic that gathers together terms that vaguely refer to the work of economists (first topic), and four topics that one would associate with macroeconomic theory (second and fourth topic), labour economics (third topic), and debt (fifth topic).

In contrast to other quantitative tools, the LDA process of topic detection is automated, and unsupervised: it minimizes *a priori* intervention – scholars only determine *ex ante* the number of topics (see below). Yet human intervention is fundamental and indispensable (all the more so in the humanities; see Blei 2012b, Rhody 2012) for labelling and hermeneutically interpreting topics, *ex post*, and developing new possible theories based on the latent structure identified by the algorithm[6]. Topic modeling provides "a lens that allows researchers working on a problem to view a relevant textual corpus in a different light and at a different scale" (Mohr and Bogdanov 2013: 560). Unlike search engines and links, topic modeling allows us to 'zoom in' and 'zoom out' to find specific or broader themes; it makes it possible to look at how themes change through time and how they are connected. The idea of zooming can be associated with what Moretti (2005, 2013) calls "distant reading". Digitization, he observes, or being able to "work on 200,000 novels instead of 200", allows us to do "the same thing 1,000 times bigger", since "the new scale changes our relationship to our object, and in fact *it changes the object itself*" (Moretti 2017, 1). Moretti's data-centric approach to novels, plots, and literary genres, based on the use of principal component analysis and clustering techniques used to generate "graphs, maps and trees", redefines a literature in terms of what can be "more easily abstracted, and hence programmed" (ibid.). The aim is to find – to recognize – "patterns", regularities that shape literary fields, otherwise invisible or hidden, and then interpret them. Patterns, Moretti writes, can "bridge the gulf between the empirical and the conceptual; they *make form visible within data*" (ibid.: 7). Even in the written production of the economics discipline, "individual texts in their individuality" (5) are not all that matters. However obscured by the emphasis we usually place on individuality, there is a social element in knowledge production. Moretti reminds us of this when discussing the reasons that motivate hermeneutical work – the idea of uncovering an author's deep but hidden, since unconscious, intentions. This analogy leads to identifying "distant" reading as a

[6] LDA shifts "the locus of subjectivity within the methodological program – interpretation is still required, but from the perspective of the actual modeling of the data, the more subjective moment of the procedure has been shifted over to the post-modeling phase of the analysis" (Mohr and Bogdanov 2013: 560).

possible means of studying the social not within individual works (as is the case of hermeneutics) but as a trait shaping the whole field.

Economics is what economists do, according to a famous dictum attributed to Viner (see Backhouse, Middleton and Tribe 1997). And texts are what economists mainly produce: articles published in the discipline's academic journals - bearing in mind that treatises and books were as important as journal articles as vehicles for the dissemination of economics in the late 19th and early 20th century. A "distant" reading of published articles in economics might therefore help uncover salient traits in the evolution of the discipline over time, and possibly offer insights about the presumed fragmentation of economics.

Dataset and methodology

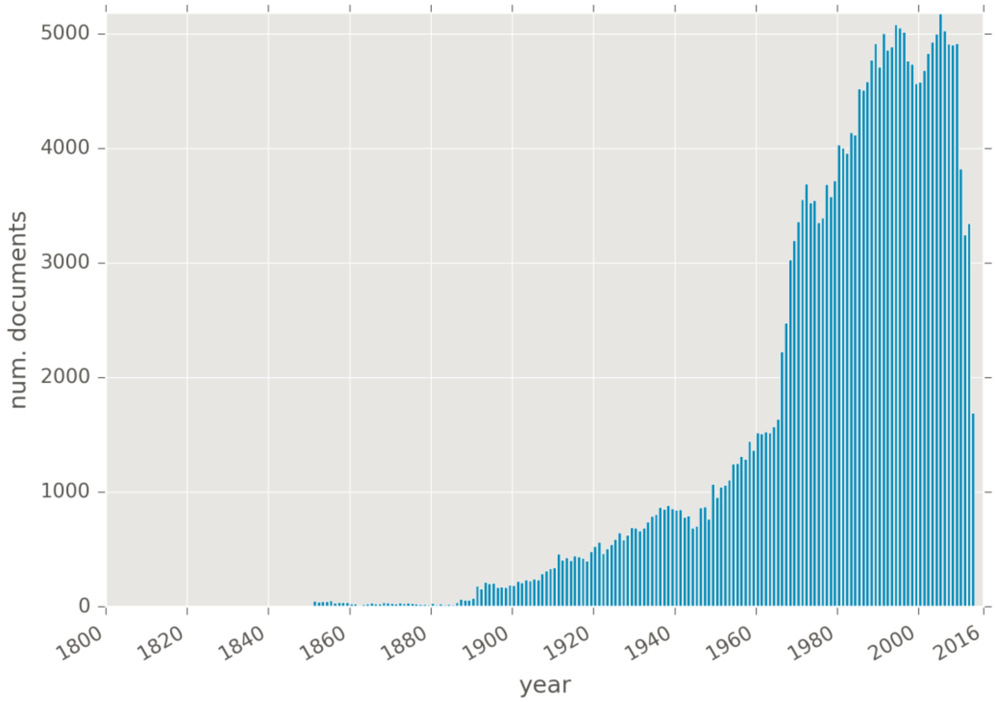
The dataset explored in this research program, then, are 250,846 articles published from 1845 to 2013 in 188 journals stored in the digital library JSTOR. JSTOR Data for Research (DfR) provides datasets of content on JSTOR for use in research that can be automatically processed. Data available upon agreement on the use of the data itself[7] includes metadata, n-grams, and OCR (Optical Character Recognition) full text for most articles, book chapters, research reports and pamphlets on JSTOR. In contrast to other datasets (Scopus, for instance, which is more correctly defined as an abstract and citation database), which are evidently constructed with an emphasis on bibliometrics, DfR allows therefore applying topic modeling techniques to the full content of economics articles, which JSTOR provides in the form of "bags of words" (employed in these documents) with associated frequencies.

To document the evolution of economic research, we limited our analysis to "research articles"[8] published between 1890 and 2013, in view of both the relatively high occurrence of non-English articles included in reviews, news, etc., and the extremely low number (2930 only) of articles published between 1845 and 1890. The number of research articles published per annum becomes significant with the turn of the century (200 in 1900), and linearly increases (800 in the Forties, 1000 ten years later) until the 1960s when it more than doubled in a few years, rising to 5000 items in the last decade of the century. 8220 articles published between 2011 and 2013 appear in the JSTOR database (see figure 1).

Figure 1. Distribution of articles in the JSTOR database by year of publication.

[7] Data was obtained upon agreement from JSTOR DfR (https://www.jstor.org/dfr/) on May 27th, 2015.

[8] The whole dataset includes also "book reviews", "miscellaneous objects", "news", and "editorials", for a total amount of some 460,000 documents.



JSTOR invites selected publications based on historical significance, citation analysis, and relevance to a scholarly audience, while publishers license the contents of their journals to JSTOR to digitize. The dataset is therefore less comprehensive than other databases of academic peer-reviewed journals (see D'Orlando 2013), and is inescapably affected by selection biases (which, however, affect any such database).

The dataset contains all articles published in the list of "elite" journals in economics, the so-called "Blue Ribbon Eight", with the (notable) exception of the *Journal of Economic Theory*. The largest number of entries (over 45,000) belongs to *Economic and Political Weekly*, followed by the *American Economic Review* (with about 15,000). Taken together, the two journals account therefore for about 25% of the articles. The distribution is relatively skewed, since the first 20 journals cover 58% of the sample (see Table 1).

Table 1. The JSTOR dataset

	Collection	**Number of articles**	**Frequency distribution**	**Cumulative frequency**
	JSTOR database	**250846**	**100%**	**100%**
1	Economic and Political Weekly	45118	18,0%	18,0%
2	The American Economic Review	15408	6,1%	24,1%

3	Annals of the American Academy of Political and Social Science	13380	5,3%	29,5%
4	American Journal of Agricultural Economics	6865	2,7%	32,2%
5	The Economic Journal	6666	2,7%	34,9%
6	The Review of Economics and Statistics	5749	2,3%	37,1%
7	Journal of Political Economy	5382	2,1%	39,3%
8	Journal of Farm Economics	5271	2,1%	41,4%
9	The Quarterly Journal of Economics	4994	2,0%	43,4%
10	Econometrica	4542	1,8%	45,2%
11	Southern Economic Journal	4460	1,8%	47,0%
12	Challenge	3954	1,6%	48,6%
13	Public Choice	3480	1,4%	49,9%
14	American Journal of Economics and Sociology	3090	1,2%	51,2%
15	Journal of Economic Issues	2992	1,2%	52,4%
16	The Review of Economic Studies	2936	1,2%	53,5%
17	The Journal of Economic History	2915	1,2%	54,7%
18	Land Economics	2822	1,1%	55,8%
19	Journal of Money, Credit and Banking	2734	1,1%	56,9%
20	Economica	2661	1,1%	58,0%

Knowledge development is a complex process, based on the continuous emergence of new ideas and research programs but also on the disappearance of old ones, while others give birth to processes of knowledge recombination. Under the assumption that changes in the semantic content of topics (in the constellations of words grouped under each topic), follow the evolution of knowledge in the field, the transformation of the topic structure can provide a sound proxy for detecting key macro-developments in the economics discipline. In applying LDA to the JSTOR database of economics articles, although this is a problem of intertemporal topic modeling, we do not employ Blei and Lafferty's (2006) Dynamic Topic Model technique, since their algorithm requires that both per-document topic distribution and per-document per-word topic assignment at time *t* be generated from those very same distributions at time *t-1*. As shown in Di Caro *et al.* (2017), this approach is not able to grasp the birth and death of topics over time and their recombination. We adopt a way of conceptualizing how knowledge evolves between different time-periods by looking at the transformations occurring between the latent topic structures of the time-windows considered, each of them obtained from running a topic modeling program.

As with any unsupervised algorithm, this type of dynamic topic-modeling exercise requires *a priori* definition of both the number of topics and size of time-windows, given that accuracy increases with the number of topics. A more complex and detailed model produces a better fit of the data and reduces biases at work. Still, there is no standardized procedure to derive such a number (see Rhody 2012), or test supporting a precise choice of the parameters, especially when topic modeling is employed to explore the content of a dataset, and not for prediction (Mimno and Blei 2011). We thus follow a research heuristic that combines a sensitivity analysis with the educated opinion of the authors about the meaningfulness of the choices themselves. We began by carefully experimenting with combined estimates of 25, 50, and 100 topics for time spans of 5, 10 and 20 years, for the meaningful, although very large, range of values they cover. As to the size of the time-windows, we opted for 10 years, believing that this represents a reasonable compromise between shorter time-windows, which would have significantly reduced the number of documents, and larger ones, which would have unduly condensed intertemporal variability and the related informational content.

3. An intertemporal map of economics

In contrast to the "zoom out" perspective of the preceding section, which presents the results of applying LDA to the JSTOR database without opening the bags of words that represent topics, we here "zoom in" and look at the changing structure of the discipline by identifying and interpreting topics in the various decades considered. LDA does not assist in labelling topics. Consider, for instance, the two following topics selected among the 27 of the time window "2010-2014"[9]:

Topic 11	**Topic 17**
shock	Inflat
Consumpt	Forecast
monetari	Monetari
output	Output
money	bank
volatil	Lag
equilibrium	target
agent	exchang
stock	shock
asset	Gdp
Household	Gap

[9] The last time window considered is not a decade: the JSTOR database does not include articles published after 2014.

inflat	macroeconomi
suppli	trend
nomin	feder
calibr	reserv
steadi	Var
Constraint	nomin
Cycl	cycle
Labor	month
Technolog	volatil

The two topics include terms from the field of macroeconomics, but it would be quite difficult to distinguish them on the basis of the most probable words of the topic, or their corpus-wide frequency. In general, in fact, topics often tend to display common terms among the first words appearing in the list, words that consequently recur in multiple topics. To bypass this difficulty, we adopt LDAvis, a web-based interactive visualization of topics developed by Sievert and Shirley (2014). On the left side of figures 4a and 4b, words associated with "Topic 17" and "Topic 11" are ranked according to their estimated term frequency within the topic, as shown by the red horizontal barchart, while the blue barchart shows the corpus-wide frequency of the term. Both measures matter: efficiency in differentiating the meanings of topics rises when both the frequency of the terms and their "exclusivity" to the topic, which is a measure of the specificity of the term to the topic, are considered. Consider figure 4a. The absolute width of red bar makes "bank" appear as one of the most important words in defining Topic 17. Yet, it is quite a common term which is generated by the topic in about 10% of its corpus-wide occurrences. Now take "forecast", a much less common word in the corpus: it almost exclusively depends on Topic 17 to generate the term. The measure of "relevance" of a term to a topic, proposed by Sievert and Shirley, rests on the possibility of linearly combining the probability ϕ of a term w to topic k and its exclusivity or "lift", defined as the ratio of a term w's probability p within the topic to its marginal probability across the corpus. Relevance depends on the value to be attributed to a parameter λ, ranging from 0 to 1, that determines the relative weight assigned to the log of the two components, the probability in the corpus and lift (the weight assigned to the probability of the term under the topic relative to its "lift"). Thus, the relevance index r of term w for topic k depends on λ and takes the following form:

$$r(w,k|\lambda) = \lambda \log(\varphi_{wk}) + (1-\lambda)\log(\frac{\varphi_{wk}}{p_w})$$

To capture the relevance of a term for a specific topic, Sievert and Shirley suggest assigning a value of 0.6 to λ, based on a study of the optimal value of the parameter for topic interpretation.

Figure 4a. Topic interpretation (topic 17): « Macroeconomic Theory », 2010-2014.

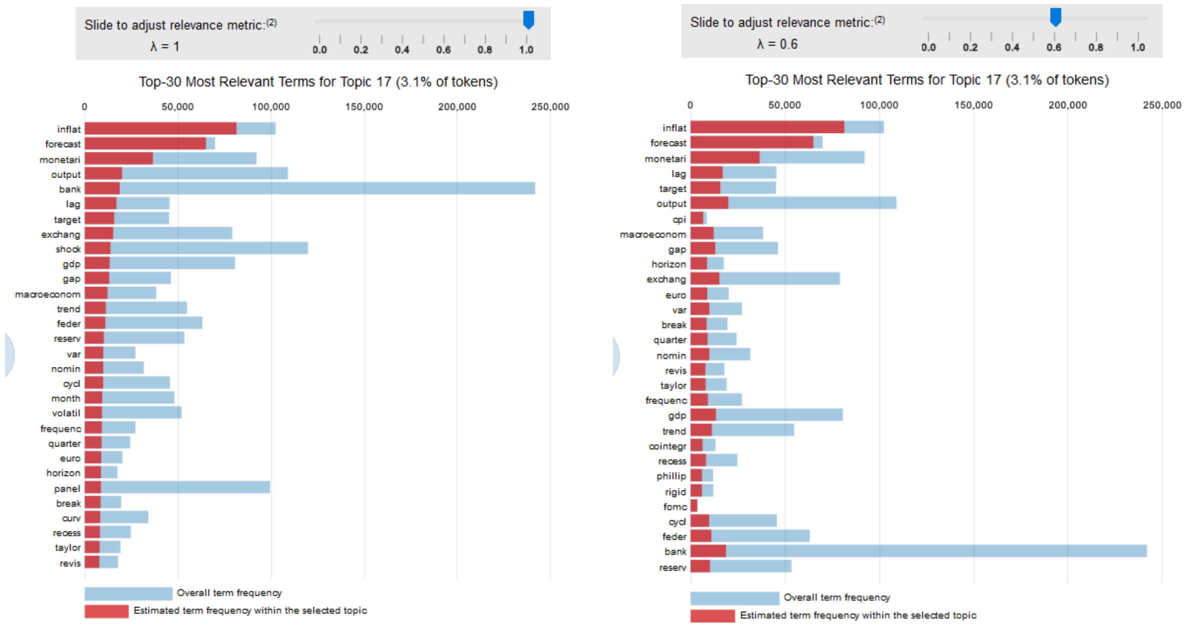
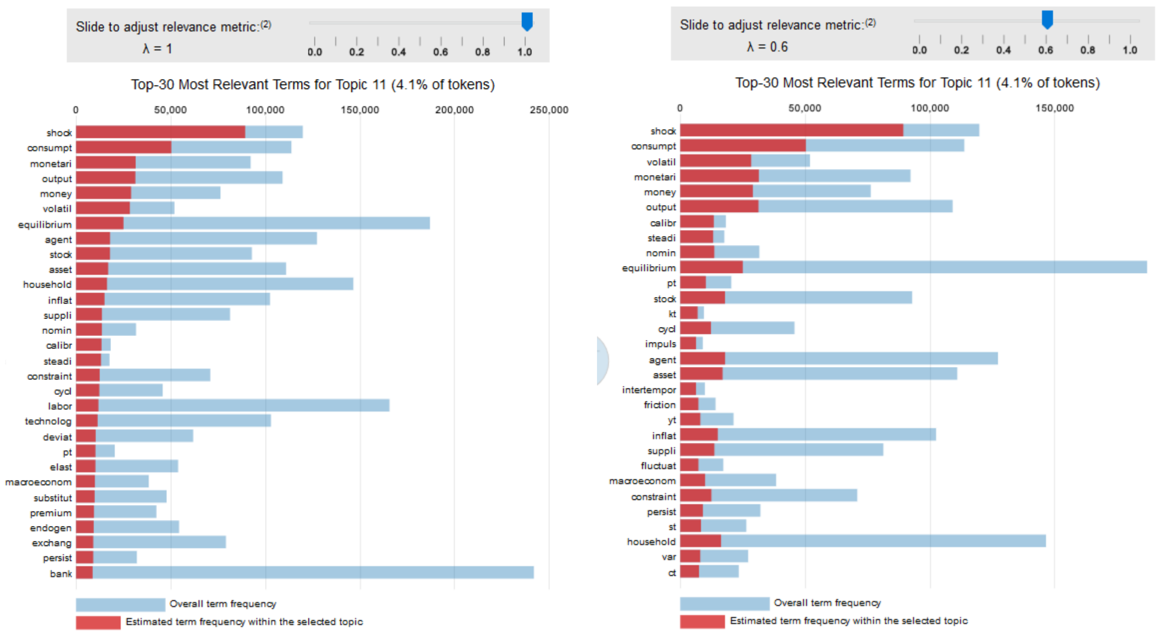


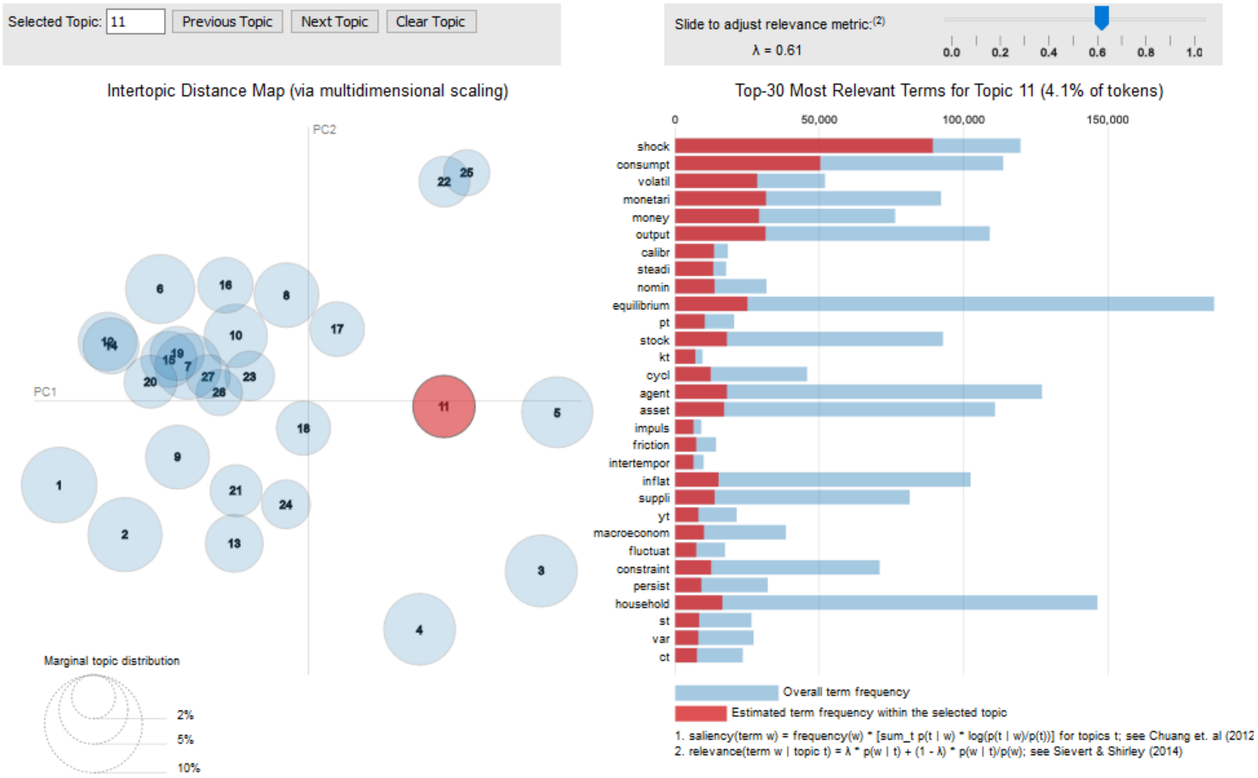
Figure 4b. Topic interpretation (topic 11): « Macroeconomic Models », 2010-2014.



With λ = 0.6, "Topic 11" appears as one of "Macroeconomic models", shaped by the lexicon of DSGE in particular (as is evident from the appearance of terms like "calibr", "steadi", "intertempor", "friction", "persist"). While "Topic 17" can be labeled "Macroeconomic theory", after LDAvis has confirmed that the most "probable" terms of the topic (terms that refer to the common and general lexicon of macroeconomics, as suggested for instance by the appearance of a term connected to the Phillips curve), are also, de facto, the most "relevant" ones.

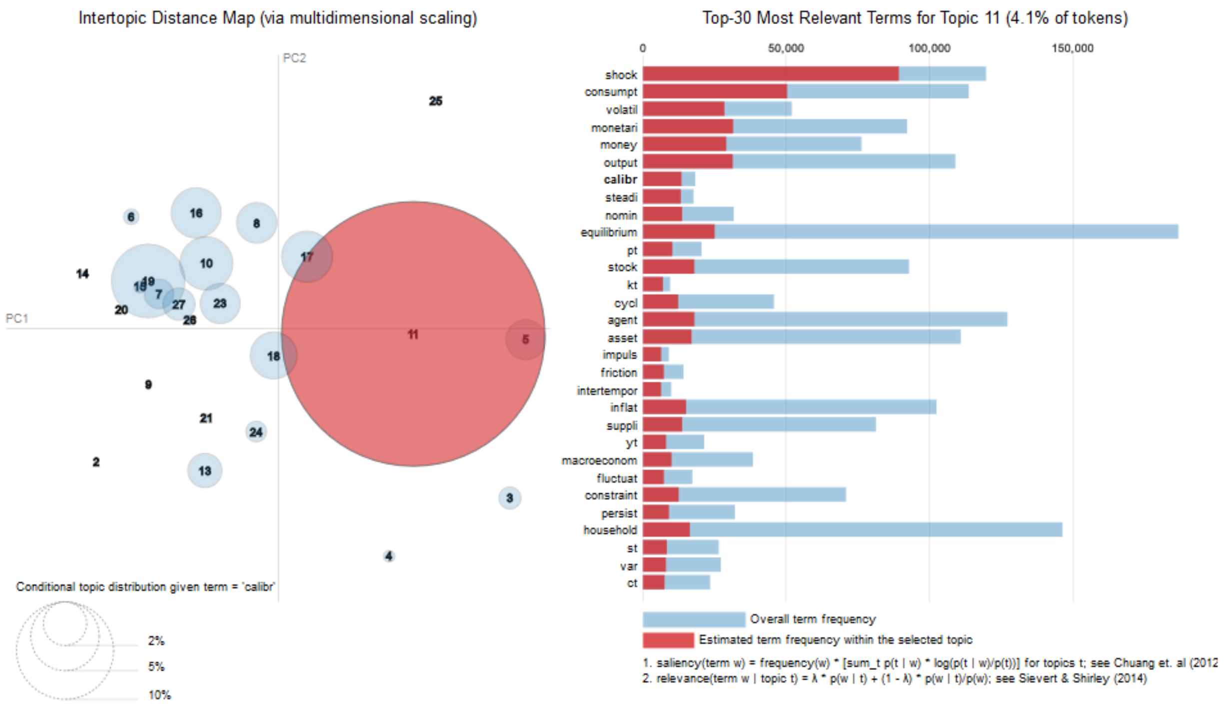
Remarkably, LDAvis provides other important details about topics. First, it measures the relative "prevalence" of the selected topic in the corpus (4.1% for "Topic 11" means that 4.1% of the corpus "comes" from Topic 11).

Figure 5. Topics' "prevalence".



Second, by selecting a term – for example, "calibr" in figure 5 – it becomes possible to visualize its conditional distribution over topics: the areas of the circles becomes proportional to the term-specific frequencies across the database. In the example, the occurrences of the term "calibr" appear as being mainly from "Topic 11", but the figure shows that a significant minority of occurrences come from other topics (like topic 19, 16, 17, 18, and others, see figure 6).

Figure 6. Terms' "conditional distribution" over topics.



By labelling all topics in all decades, supported by LDAvis, it then becomes possible to visualize a "map" of economics over time. Tables 4a and 4b show the 27 topics detected in the LDA model, and an effort in their interpretation – with a view to obtaining more accurate labels than we could otherwise impose by simply looking at frequencies – in the light of topics' "relevance". Remarkably, LDAvis makes it possible to neatly and safely distinguish, in interpreting two topics broadly related to Industrial organization, between two JEL codes, namely "Market structures, firm strategy, and market performance" (L1) and "Firm objectives, organization, and behavior" (L2; see Table 2).

Table 2. Topic interpretation: "Industrial organization" (1980-1990)

Terms (20), LDA	Terms (30), LDAvis (λ=1)	Terms (30), LDAvis (λ=0.6)	Topic interpretation
0.079*firm + 0.027*profit + 0.019*competit + 0.011*contract + 0.01*margin + 0.01*consum + 0.009*equilibrium + 0.009*entri + 0.008*output + 0.007*sale + 0.007*monopoli +	firm profit competit contract margin consum equilibrium entri output sale monopoli buyer qualiti share seller advertis concentr	firm profit competit entri monopoli buyer contract seller advertis monopolist margin bid consum sale qualiti auction equilibrium merger concentr output entrant incent	**INDUSTRIAL ORGANIZATION: MARKET STRUCTURES, FIRM STRATEGY, AND MARKET PERFORMANCE**

0.006*buyer + 0.006*qualiti + 0.006*share + 0.006*seller + 0.005*concentr + 0.005*advertis + 0.005*effici + 0.005*bid + 0.005*regul	effici bid regul incent monopolist offer behavior maxim revenu purchas curv size pp	sell regul brand oligopoli share custom rival	
0.026*manag + 0.012*technolog + 0.011*compani + 0.01*busi + 0.01*organ + 0.009*plan + 0.008*project + 0.008*enterpris + 0.007*perform + 0.007*corpor + 0.004*particip + 0.004*design + 0.004*establish + 0.004*object + 0.004*servic + 0.004*promot + 0.004*strategi + 0.004*integr + 0.004*success + 0.004*manageri	manag technolog compani busi organ plan project enterpris perform corpor particip design establish object servic promot strategi integr success manageri resourc engin firm experi drug technic practic organiz effort improv	manag organ organiz drug manageri enterpris technolog compani project busi ventur electron corpor perform execut engin plan pharmaceut promot personnel machin softwar subcontract design subsidiari consult goal satisfact director task	**INDUSTRIAL ORGANIZATION: FIRM OBJECTIVES, ORGANIZATION, AND BEHAVIOR**

Likewise, it becomes possible to assign the label “Eastern countries’ transition to capitalism” to a topic otherwise not easily interpretable (see Table 3).

Table 3. Topic interpretation: “Eastern countries’ transition to capitalism” (1980-1990)

Terms (20), LDA	Terms (30) LDAvis (λ=1)	Words (30) LDAvis (λ=0.6)	Topic interpretation

0.013*countri + 0.009*soviet + 0.009*growth + 0.009*plan + 0.009*enterpris + 0.007*world + 0.006*reform + 0.006*invest + 0.005*cooper + 0.005*european + 0.005*central + 0.005*socialist + 0.005*foreign + 0.005*percent + 0.004*europ + 0.004*crisi + 0.004*trade + 0.004*sector + 0.004*germani + 0.003*western	countri soviet growth plan enterpris world reform invest cooper european central socialist foreign percent europ crisi trade sector germani western situat materi west german union improv achiev grow export competit	soviet enterpris countri socialist reform plan european crisi europ cooper germani growth hungari hungarian cmea acta world oeconomica ussr german central western poland eastern west gdr billion shortag invest foreign	**EASTERN COUNTRIES' TRANSITION TO CAPITALISM**

Below follows the “map” of economics as discipline since the Sixties (Tables 4a. and 4b.)

Table 4a. A "map" of Economics. Topics, 1960-1970 to 1980-1990

1960-1970		1970-1980		1980-1990	
Topic, Prevalence		TOPIC, Prevalence		TOPIC, Prevalence	
Theoretical economics	8.1	Economics as social discipline	7.9	Mathematical methods	7.9
Economics as social discipline	8.0	Econometric methods	7.0	Econometric methods	7.9
Econometric methods	7.2	Law and regulation	6.9	Economics as social discipline	7.4
Business economics	6.6	Macroeconomic stabilization	5.1	Macroeconomics	5.5
Labour	4.9	Theoretical economics	4.5	Political issues in emerging countries	5.3
Monetary policy	4.5	Industrial organization: market structures, etc.[1]	4.5	Industrial organization: market structures, etc.[1]	4.9
Industrial organization: market structures, etc.[1]	4.2	Eastern countries' politics	4.2	Eastern countries' transition to capitalism	4.3
International politics	4.0	Regional economics, cities	3.8	Industrial organization: firm objectives, etc.[2]	4.3
Britain's economic history	3.8	Insurance	3.6	Taxation, fiscal policy, fiscal behavior	3.9
International trade	3.8	International trade	3.6	Labour	3.6
Taxation, fiscal policy	3.6	Britain's economic history	3.3	International trade	3.6
Agriculture	3.5	Eastern countries' economic organization	3.3	Law and economics	3.5
Regional economic, transports (cities)	3.4	Labour	3.2	Canada: growth and innovation	3.5
Education	3.4	Growth and development, India	3.1	Banking and finance	3.3
Financial markets	3.4	Transports, energy and environment	3.0	Demography	3.3
Welfare, US	3.2	Development	3.0	Growth and agriculture in Partition of India	3.2
Insurance	2.9	International economics	2.8	India	2.9
Law and economics	2.9	Workers' economics	2.8	Agriculture	2.7
Manufactures and raw materials	2.7	Demography	2.8	Education, professions	2.7
Banking and finance	2.5	Banking and finance	2.7	Regional economics (cities)	2.6
Demography	2.4	Education	2.6	Britain's economic history	2.5
Agriculture in Partition of India	2.2	Agriculture	2.4	Insurance	2.5
India's politics	2.2	Law and economics	2.3	Public choice	2.4
Rural economics	1.7	Agricultural products	2.2	Transports	2.0

Oil, energy	1.6	**Spatial economics**	2.2	**Energy and environment**	1.8
Economic history	1.6	*French terms*	*1.3*	**Africa**	1.8
Non Anglo-Saxon words	*1.5*	*Non Anglo-Saxon words*	*1.3*	*Non Anglo-Saxon words*	*1.0*

[1] Industrial organization: market structures, firm strategies, and market performance; [2] Industrial organization: firm objectives, organization, and behavior.

Table 4b. A "map" of Economics. Topics, 1990-2000 to 2010-2014

1990-2000		2000-2010		2010-2014	
TOPIC, Prevalence		TOPIC, Prevalence		TOPIC, Prevalence	
Economics as social discipline	8.3	Econometrics	6.7	Economic history (pre-XX)	6.0
Mathematical methods	6.1	Mathematical methods	6.3	Economics as social discipline (schools of th.)	5.7
India	5.1	Economics as social discipline	6.2	Theoretical economics	5.4
Econometrics	5.1	Education	5.2	Game theory	5.4
Labour	4.9	Econometrics applied to industry	5.0	*Econometrics parameters*	5.3
Prediction/cycles (econometrics, applied)	4.7	Industrial organization: firm objectives, etc. [2]	4.9	Demography	5.0
Game theory	4.6	India	4.8	Banking and finance, debt	4.6
Industrial economics	4.6	Development in Partition of India	4.7	International trade, Fdi	4.4
Industrial organization: firm objectives, etc. [2]	4.5	Game theory	4.4	Managerial economics	4.3
Eastern countries' transition to capitalism	4.4	Macroeconomic stabilization	4.2	Labour	4.2
Industrial organization: market structures, etc.[1]	4.2	Theoretical economics	3.7	Macroeconomic models (Dsge)	4.1
Demography	3.8	*Unusual terms*	3.7	Globalization	3.6
Monetary policy (open ec. macroeconomics)	3.7	Labour	3.6	Behavioral economics	3.5
Agriculture	3.6	Economic history (pre-XX)	3.3	Regional economics (Canada, immigration)	3.3
Taxation, public economics	3.5	Regional economics (cities)	3.3	Health	3.3
Law and economics (law and regulation)	3.4	East Asia, international trade	3.2	Debt	3.2
Banking and finance	3.3	Law and regulation, Canada	3.1	Macroeconomic theory	3.1
Education	3.3	Financial markets	3.0	Firm products, consumption, marketing	3.1

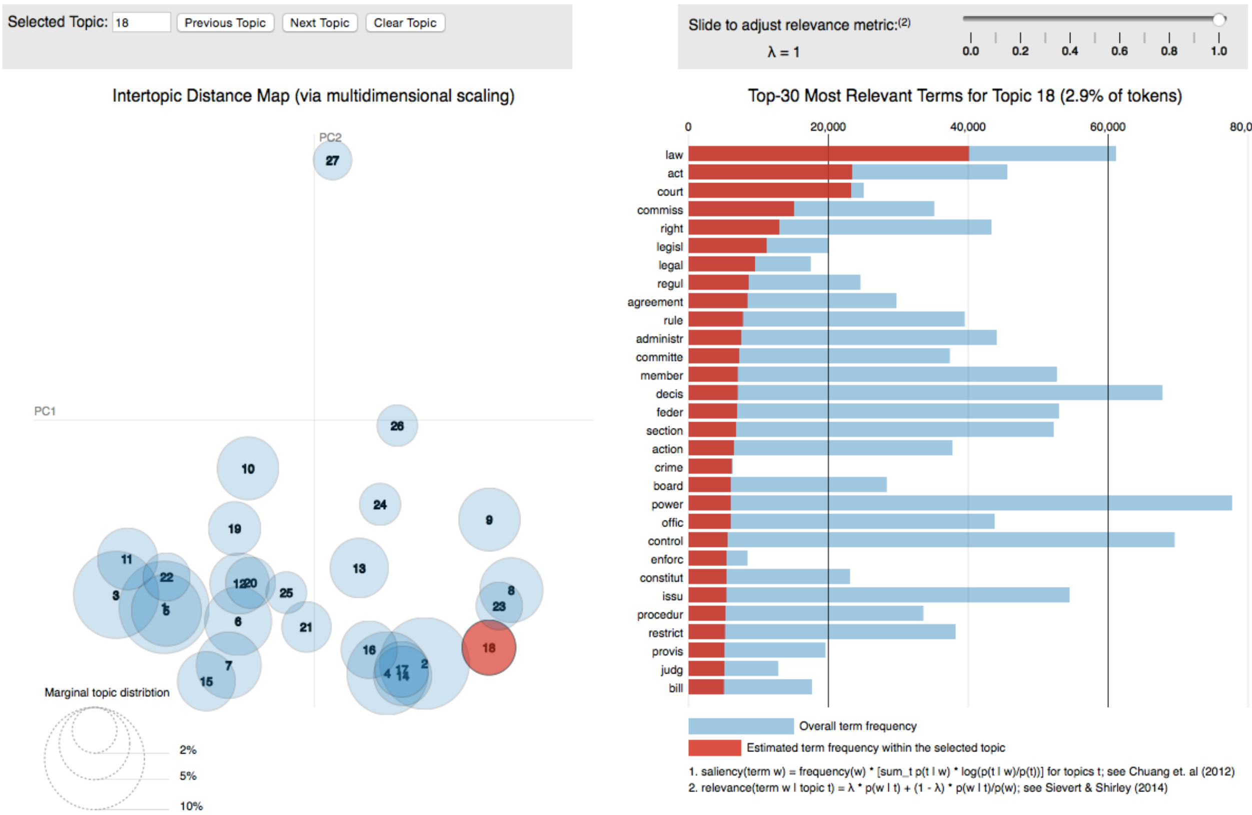
International trade	3.2	**Taxation, public economics**	2.8	**Agriculture, Agricultural insurance**	3.1
Insurance	3.0	**Agriculture**	2.7	**Education**	2.9
Britain's economic history	2.7	**Demography**	2.6	**Public choice**	2.8
Energy and environment	2.4	**Banking and finance, debt**	2.6	*Unusual terms*	2.7
Pharmac. industry, crime, "new" consumption	2.4	**Health**	2.5	**Taxation, public economics**	2.6
Public choice	1.7	**Insurance**	2.3	**Innovation economics**	2.5
Non Anglo-Saxon words	1.8	**Energy and environment**	2.1	*Cyrillic letters*	2.2
East Asia (politics, military)	1.3	**Public choice**	1.9	**Automotive industry, China**	2.2
Non Anglo-Saxon letters	0.9	*Unusual terms*	1.2	Energy and environment	2.0

[1] Industrial organization: market structures, firm strategies, and market performance; [2] Industrial organization: firm objectives, organization, and behavior.

As with any map, the one proposed here is in some sense arbitrary. This arbitrariness is due to the abovementioned selection biases at work in the dataset, but also to the number of topics – which has an important effect on the results of the "zoom in", although this does not create any unsurmountable problem to the analysis – and the nature of topics. Some topics (in *italics* in the tables) are in truth errors which one should attribute to optical character recognition or, more often, to foreign language terms or (mathematical, etc.) symbols. But there is more. To some extent, at least, Rhody's (2012, 15) argument about topic modeling and literary studies – wherein topics can likely be "better understood as a representation of 'discourse' (language as it is used and as it participates in recognized social forms) rather than a thematic string of coherent terms" – is valid for economics as well. Economic papers are, first of all, discourses. While LDAvis can help label "semantically opaque topics", in the terminology used by Rhody, "semantically evident topics" can pose peculiar problems, and induce a careful "back to the papers" approach. Consider the topic "Game theory". How can one ascertain whether the topic denotes a "field" – that is, the topic is *about* game theory – or, conversely, it represents a tendency to use of a "theoretical" style (rather than "empirical" or "econometrical"; to use Angrist et al.'s 2017 categories)?

To further analyse the nature of topics, one has to zoom in closer. Consider a specific topic, one that is quite clearly recognizable in almost all decades (it is missing only in the last time window, 2010-2014). Even a rapid glance at the sequence of ranked terms presented in the left side of Figure 7 allows labelling Topic 18, in the decade 1960-1970, as "Law and Economics". This does not come as a surprise: these were the years when authors like Ronald Coase and Guido Calabresi started applying the approach and toolbox of economics to legal issues.

Figure 7. The topic "Law and Economics" in the decade 1960-1970.



Consider, however, the decade 1940-1950: two different topics are detected showing terms (with λ = 1; yellow column in the table) in the first ten positions that are broadly connected to regulation and law. Remarkably, both topics include terms like *court* and *law*, adding to the difficulty of identifying the topic, between the two, that more clearly expresses the approach known as the economic analysis of law. Table 5 shows, for each topic (called "topic 5" and "topic" 10 respectively, based on their rank), two slightly different lists of terms, according to the value of λ chosen. As observed, λ can assume values between 0 and 1. While assuming λ = 1 amounts to considering the most probable terms under each topic as generated by LDA, shifting the value of λ to 0.6 allows deepening the understanding of the nature of the topic under consideration. Topic 5, in particular, can be interpreted as related to law and trial, whereas Topic 10 appears more regulation-oriented. As a matter of fact, with λ = 0.6, the term "law" even disappears in the ranked terms list of Topic 10, while "court" is now ranked lower. Other terms, like "competit", "traffic", "freight" come into view, suggesting that this peculiar topic has more to do with regulation issues. The overlapping of the ranking of the two topics with λ = 1 is possibly due to the fact that both "Law and Economics" and "Regulation" are evidently involved in discussions concerning various issues generically linked to law, or to the fact that the approach "Law and Economics" has not yet the maturity it will show in later decades.

TABLE 5. Topics "Law and Economics" and "Regulation", 1940-1950.

	1940-1950			
Prevalence	5,1%		3,9	
Rank (and topic)	5		10	
HHI	0,0011		0,00112	
Topic label	Law and Economics		Regulation	
	λ=1	λ=0,6	λ=1	λ=0,6
Terms	union	**union**	Commiss	**commiss**
	board	**board**	Competit	**carrier**
	wage	**bargain**	Regul	**regul**
	worker	**employe**	Transport	**transport**
	bargain	**wage**	Railroad	**competit**
	law	**court**	Carrier	**retail**
	employe	**worker**	Retail	**railroad**
	court	**strike**	Law	**traffic**
	member	**collect**	Court	**interest**
	collect	**legisl**	Sale	**freight**
	legisl	**law**	Util	**antitrust**
	strike	**member**	compani	**commerc**
	local	**disput**	administr	**court**
	committe	**elect**	Commerc	**air**
	part	**parti**	Charg	**regulatori**
	right	**arbitr**	Traffic	**charg**
	feder	**membership**	manufactur	**sale**
	administr	**local**	Decis	**store**
	manag	**jurusdict**	Freight	**ship**
	agreement	**execut**	Agenc	**util**
	constitut	**vote**	Feder	**vessel**
	elect	**committe**	Interest	**passeng**
	disput	**presid**	Commod	**manufactur**
	offic	**constitut**	Ship	**law**
	decis	**right**	Air	**decis**
	presid	**agreement**	Depart	**administr**
	contract	**manag**	Class	**territori**

	execut	negro	Sell	wholesal
	congress	congress	Consum	rail
	vote	administr	Store	pilot

After this exercise has been made for all decades and all possible sources of confusion, is it possible to discuss *the* topic "Law and Economics" as if there existed real continuity over time between bags of words quite evidently related to the typical lexicon used when applying an economic approach to legal issues? Yes, to a certain extent at least. The global view of the topic over the decades, with λ = 0.6, suggests that the topic exhibits a certain degree of stability over time. Consider four different decades (see Table 6), and the related sequences of the first 20 terms that define the topic with λ = 1 first, and λ = 0.6 then. We here focus on the decade 1910-1920, when the approach "Law and Economics" was at its very early stage, and then on three decades roughly corresponding to its becoming a proper subdiscipline of economics.

TABLE 6. "Law and Economics" in four decades (1910-1920, 1960-1970, 1970-1980, 1980-1990)

Decade	1910-20		1960-70		1970-80		1980-1990	
Prevalence	4,90%		2,90%		2,20%		3,50%	
Rank	6		18		24		12	
	λ=1	**λ=0.6**	**λ=1**	**λ=0.6**	**λ=1**	**λ=0.6**	**λ=1**	**λ=0.6**
Terms	court	**court**	law	**Law**	law	**law**	law	**law**
	hous	**commiss**	act	**Court**	court	**court**	right	**court**
	commiss	**hous**	court	**Act**	right	**legal**	court	**legal**
	regul	**regul**	commiss	**Commiss**	legal	**right**	rule	**right**
	legisl	**legisl**	right	**Legisl**	act	**crime**	legal	**act**
	build	**tenement**	legisl	**Legal**	rule	**enforc**	act	**regul**
	provis	**judici**	legal	**Crime**	properti	**act**	regul	**legisl**
	corpor	**licens**	regul	**Right**	crime	**defend**	legisl	**rule**
	tenement	**build**	agreement	**Regul**	contract	**supra**	properti	**enforc**
	enforc	**statut**	rule	**Enforc**	enforc	**crimin**	feder	**supra**
	licens	**enforc**	administr	**Amend**	defend	**accid**	parti	**damag**
	constitut	**room**	committe	**Justic**	parti	**rule**	action	**plaintiff**
	district	**evil**	member	**Agreement**	judg	**damag**	enforc	**litig**
	statut	**justic**	decis	**Crimin**	protect	**judg**	protect	**defend**

	privat	**provis**	feder	**Polic**	regul	**judici**	contract	**crime**
	railroad	**suprem**	section	**Suprem**	damag	**injuri**	damag	**protect**
	judici	**polic**	action	**Statut**	compens	**suprem**	liabil	**liabil**
	york	**decis**	crime	**Judici**	accid	**punish**	constitut	**amend**
	investing	**liquor**	board	**Jurisdict**	legisl	**amend**	claim	**victim**
	decis	**sanitari**	power	**Judg**	polic	**litig**	commiss	**justic**
	room	**corpor**	offic	**Rule**	person	**polic**	defend	**constitut**
	evil	**privat**	control	**Committe**	liabil	**contract**	privat	**action**
	legislatur	**legislatur**	enforc	**Administr**	constitut	**justic**	supra	**crimin**
	justic	**prevent**	constitut	**Violat**	crimin	**properti**	provis	**statut**
	prevent	**district**	issu	**Board**	supra	**lawyer**	plaintiff	**suprem**
	board	**sallon**	procedur	**Bill**	action	**plaintiff**	litig	**feder**
	bill	**constitut**	restrict	**Provis**	justic	**trial**	common	**properti**
	administr	**health**	provis	**Hear**	amend	**statut**	effici	**neglig**
	commette	**judg**	judg	**Constitut**	claim	**victim**	person	**commiss**
	protect	**amend**	bill	**Action**	priva	**compens**	practi	**parti**

Table 6 reports two columns for each decade, according to the value set for λ: terms in colored boxes are common to the two lists of terms generated by LDA with, respectively, λ = 1 and λ = 0.6, while terms in white boxes appear in just one of the two lists. In other words, terms in yellow (with λ = 1) become blue when λ = 0.6, whereas terms in white boxes "define" the topic only when λ assumes the value of the column they belong to. In each decade considered, virtually all words included in the first 10 of the list when λ = 1 are also present in the λ = 0.6 list, despite changes in their relative position. Looking at the lower part of the ranking, however, might help us see things differently. With λ = 0.6, the lists of terms in each decade signal a specific focus on "law and trial" issues.

Let us now glance at the topic's "prevalence". LDAvis plots the topics as circles in the two-dimensional plane, and encodes each topic's overall prevalence using the area of the circles: topics thus appear in decreasing order of prevalence. "Law and Economics" is thus ranked 6th in the decade 1910-1920 (it covers 4.9% of the articles in the corpus), 18th (2.9%) in the decade 1960-1970, 24th (2.2%) in the decade 1970-1980, and 12th (3.5%) in the decade 1980-1990. In itself, "prevalence" is of little help in further interpreting the topic. Things change when this measure of the weight of specific topics in the corpus is considered also in the light of their "concentration index" – the Herfindahl-Hirschman Index, HHI. HHI is a more refined measure of the distribution of the size of individual topics in relation to the corpus, indicating the degree of competition between topics within documents. Now, excepting the first decade, where the prevalence of "Law and Economics" in the corpus is quite high (6.1%), its weight is significantly lower in other decades (see Table 7).

TABLE 7. Prevalence of the topic "Law and Economics", all decades

Decade	**1890-00**	**1900-10**	**1910-20**	**1920-30**	**1930-40**	**1940-50**
Prevalence	6,1	4,9	4,9	2,9	3,9	5,1
Decade	**1950-60**	**1960-70**	**1970-80**	**1980-90**	**1990-00**	**2000-10**
Prevalence	3	2,9	2,2	3,5	3,4	3,1

Although it is always present with a non-negligible weight, in any decade the topic cannot therefore be considered among the most representative in term of prevalence. As to the concentration index (see figures 8a and 8b), "Law and Economics" exhibits heterogeneous HHI values over time. Some decades see the topic diffused in a huge collection of articles in the corpus, many of them including it as one of their non-first topics: it competes with other topics, in other words, in a significant number of articles. "Law and Economics" is more concentrated in other decades (its HHI value is high in relative terms with respect to other topics): the topic covers a relatively smaller number of papers, in which however the topic is the most prevalent one, or one of the few most prevalent. In such decades, the topic becomes a distinctive one in the discipline.

Remarkably, while the topic exhibits low HHI values in the early decades (1890-1900, 1900-1910, 1910-1920), the situation is reversed in the next ones, and its HHI reach high values in the decades 1970-1980, 1980-1990, 1990-2000. This comparison, and the significant differences that emerge between the beginning of the twentieth century and recent decades, may suggest that the topic has attained high levels of concentration when reaching its maturity stage as specific subfield of the discipline.

Figure 8a. Concentration indexes (HHI) for each topic, 1910-1920.

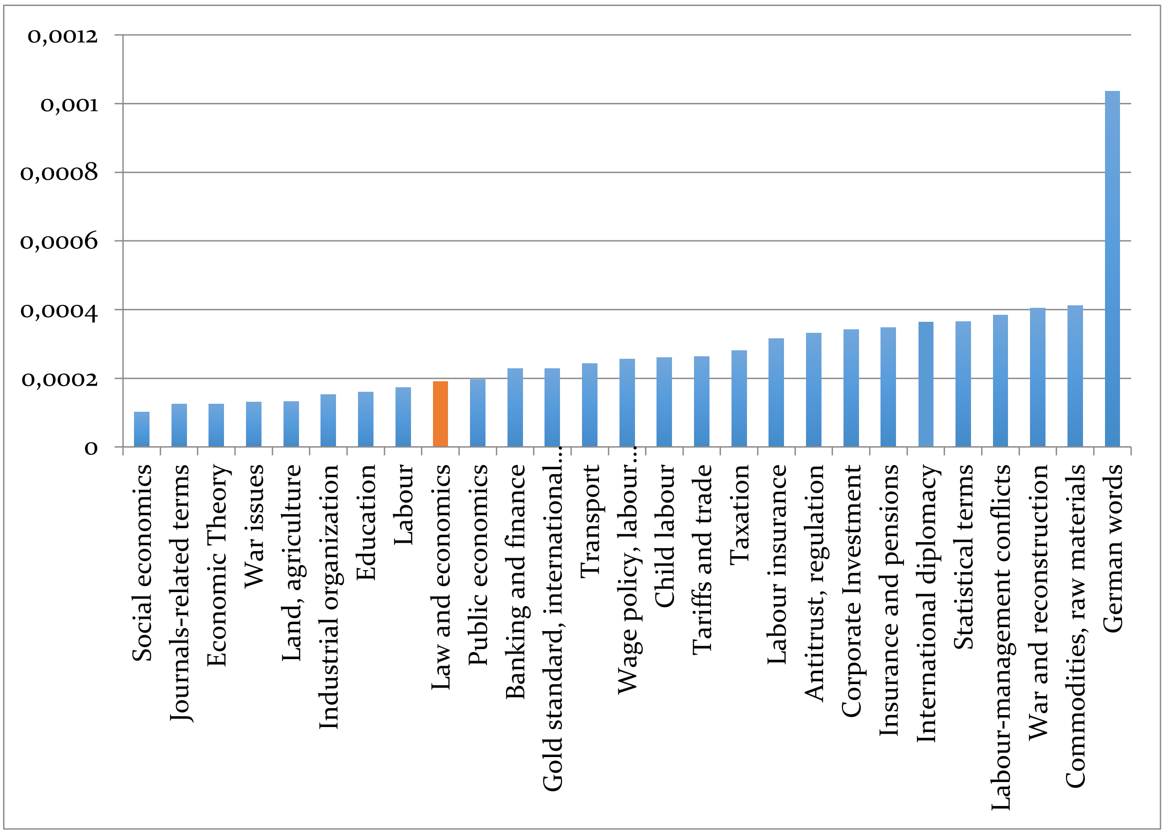
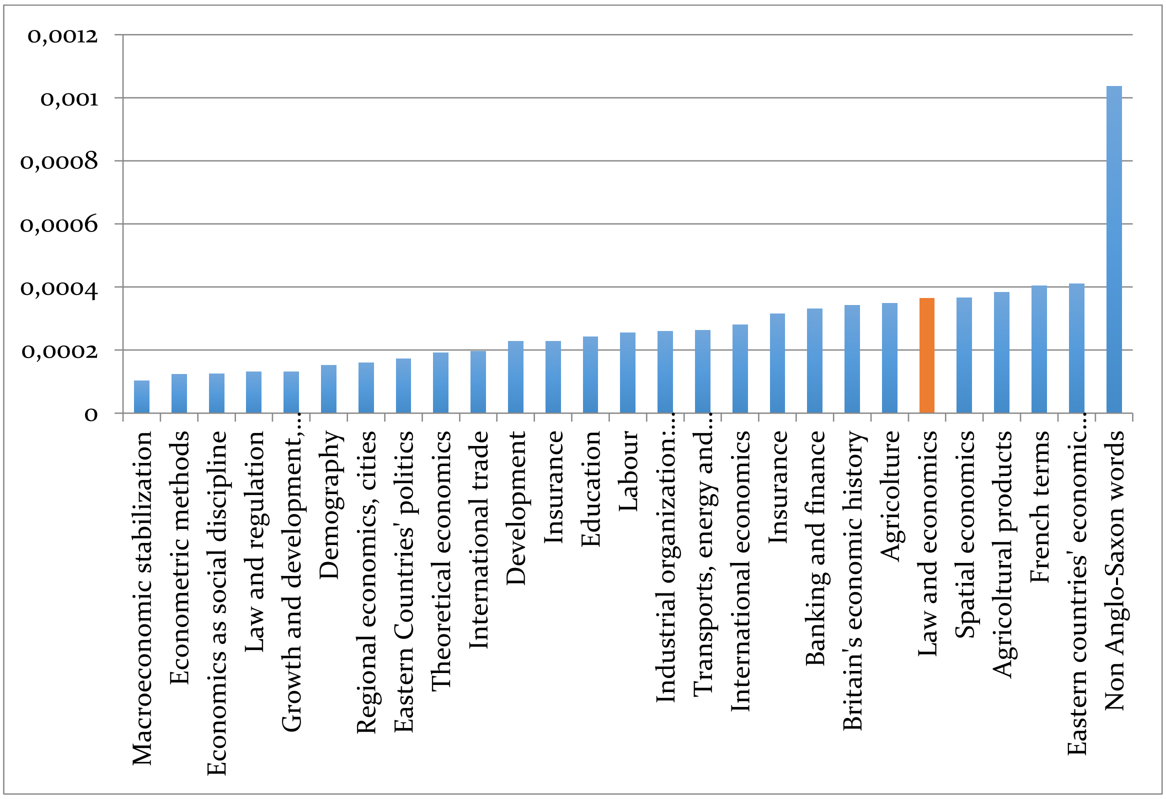


Figure 8b. Concentration indexes (HHI) for each topic, 1970-1980.



Looking at the corpus of articles, it becomes evident that in the last three decades (excluding therefore the five years between 2010 and 2014), a relatively high number of articles show "Law and Economics" as first topic: the HHI value increases exactly in the "boom" phase of the research program. This result acquires greater importance when it is recalled that the JSTOR database does not include the totality of journals strictly associated with this now specialized field and rather only a few (among which the *Journal of Law and Economics* and the *Journal of Law Economics and Organization*). By jointly considering topics' prevalence and HHI values, one can infer that "Law and Economics" can be characterized as a quite "general" topic for much of its history. Good performances in terms of prevalence are compensated by relatively low levels of concentration: "Law and Economics" cannot be portrayed as one of the most representative topics of the collection (and of the discipline), and is rather, often and naturally associated with topics that deal with regulation, taxation, and public choice issues. Yet things seem to change in the Seventies, concomitant with the development of "Law and Economics" as specific subdisciplinary approach: lower levels of prevalence accompanied by high concentration signal the transformation of the field into a highly specialized and more compact one.

4. Topic modeling and the evolution of economics

Space constraints evidently prevent us from offering other than some concrete illustrations of how topic modeling can be put to work and serve as new analytical tool for scrutinizing the evolution of economics. LDA rests on an original methodological approach, applied to a very large corpus of documents. With respect to previous quantitative investigations of the economic literature, the size of the sample[10] clearly makes a difference, but the fundamental novelty of the approach lies, first, in considering the *full text* of economic articles, rather than citations, bibliographical references, metadata, or any other specific feature[11]. Second, in the change of perspective offered by the automated, unsupervised nature of topic detection, creating the possibility of revealing the latent, hidden – and otherwise invisible – structures of economists' works. The purpose topic modeling serves is neither to classify articles (since it presupposes that each article is a collection of topics), nor to embed them strictly into clusters or specialties. Deconstructing articles and "distantly" reading them

[10] For instance, Heap and Parikh's (2005) important study of the diffusion of ideas in academia considers ten journals in economics from 1950 to 1990 (six top journal and four middle ranking ones), and select only articles using specific econometric techniques. The sample investigated by Card and DellaVigna (2013) in their famous articles about top journals in economics includes 13,069 articles (published between 1970 and 2012). Kosnik's (2015) recent exploration reviews about 20,000 academic articles published in seven top research journals from 1960 to 2010.

[11] McCain's 2014 work focuses on the concept of bounded rationality to explore the potentialities for text-mining research of the full-text JSTOR database (3,707 articles are considered).

as bags of words, LDA ultimately generates maps of economic knowledge that do not have the ambition of replicating the territory. It rather leave researchers the task and freedom to focus on alternative points of interest[12], as well as to further analyse the maps themselves (the thematic hidden structure of the corpus) by returning to documents – the elements of the "territory", by means of close inspection of individual texts.

Philosophically speaking, there is a difference, which entails a complementarity, between the "hidden structure" generated by topic modeling and the lines of reasoning that emerge from the analysis of texts. Words reveal a hidden structure of discourses, including when their use does not reflect a theoretical or analytical development of discourse. Simply, they frame discourses. In short, discourses tacitly adopt a certain grid of words, which constitutes the frame of the discourses themselves. In this sense, LDA makes it possible to compare "economics as discourse" (that is, economics as it emerges from the "hidden thematic structure") and economics as set of theories and perspectives (of which HET is part), oriented to solving specific puzzles. In so doing, it allows us also to consider the social dimension tacitly shaping economists' discourse – based on a somewhat more radical concept of "conversation" than the one embedded, for instance, in bibliographical references.

Topic modeling can also assist in analysing the shifts that have recently occurred in the structure of economics. In many ways, the history of social sciences is a complex story of fragmentation and recombination: specialization produces continuous creation of hybrid specialties (Dogan and Pahre 1989), while cross-disciplinary ventures, traditionally considered as attempts to revise disciplinary boundaries, can in truth play a complementary role to such divisions (Fontaine 2015). The pluralistic mainstream landscape created by once "insufficiently hybrid" (Dogan and Pahre 1989, 68) economists may reflect the advent of a new balance, in Knudsen's (2002) terminology, between normal and revolutionary science, between unification and fragmentation. Mainstream "pluralism" is in truth, more correctly, a plurality (Dow 2008), given mainstream economics' attitude to truly alternative methodologies; but the unity of economics is flexible, as said, and allows for the coexistence of incompatible theoretical contributions. The attention economists are currently devoting to the JEL codes classification system (Kosnik 2017; see also Suominen and Toivanen 2015) is also an indirect means of averting a possible "complexity crisis" triggered by fragmentation: in condensing the knowledge structure of economics, maps like the one proposed in this paper can help "increase the absorptive capacity" of the field (Knudsen 2002, 28).

The assumption made here (when proposing some preliminary indexes to investigate topics' life cycles) that changes in the semantic content of topics follow the evolution of knowledge in the field is

12 In this peculiar sense, a qualified comparison between topics generated by LDA and JEL codes (in light also of the historical developments of these latter, see Cherrier 2017) almost imposes itself as a possible future outcome of the research.

evidently a strong one, and needs further investigation (also outside economics, that is, in other social sciences, hard sciences and humanities). Still, topic modeling as (an unsupervised) technique was developed with the aim of facilitating searching, browsing and summarizing large archives, and can be used to challenge, so to speak, human-assigned metadata or subject classification (like JEL codes): comparisons show that automated classification systems are better at identifying novel bodies of knowledge (Suominen and Toivanen 2015). Moreover, topic modeling encourages us to reason about the various, heterogeneous theoretical dimensions of specialization – as well as to identify changing patterns in specialization itself over time. Lastly, and above all, there is evidence that the changing language of economics documents (semantic transformations *in primis*) tend to reflect shifts in approaches and attitudes, and that studies of this kind, if supported by careful research in the history of economics and economic thought, can have a significant impact on our understanding of the evolution of economic knowledge (see, for instance, Moretti and Pestre 2015).

This requirement – that quantitative techniques like topic modeling be employed as a complement, rather than a substitute, for a history-of-economics study of the changing structure of the discipline – might constitute a valuable opportunity for the history of economic thought. Evidently marginalized, in times of ubiquitous specialization, the history of economic thought (HET) can profit from the diffusion of these quantitative analytical tools, and particularly of topic modeling, in view of both the advantages it offers to scholars engaged in the attempt to apply quantitative historical semantics to economics (see Klaes 2017), and of the relative importance that topic modeling induces us to assign to the field. In the general map presented here, the topic "Economics as social discipline" includes both terms related to economics as social science as well as words clearly pertaining to the history of economic thought[13]. As Table 8 shows, its "prevalence" is very high in each decade, but its HHI value is low: the topic is largely diffused but scarcely concentrated, since it generally competes (or is compatible) with others in a significant number of articles.

TABLE 8. "Economics as Social Discipline", 1980-1990 and 2010-2014

	1980-1990	2010-2014
No. articles	43669	15049
EaSD as first topic	3269	1000
Prevalence	7.4	5.7
Rank	3	2
HHI	0.0001076	0.0004012

[13] In part, this owes to the disproportionate weight in the dataset of *Economic and Political Weekly*, a left-leaning forum for exchange of ideas across social sciences.

	λ=1	λ=0.6	λ=1	λ=0.6
Terms	societi	**economist**	economist	**economist**
	economist	**society**	peopl	**marx**
	pp	**scienc**	societi	**capitalist**
	concept	**concept**	challeng	**concept**
	scienc	**keyn**	concept	**challeng**
	world	**marx**	human	**think**
	human	**book**	crisi	**keyn**
	book	**idea**	think	**hayek**
	idea	**human**	labor	**sociolog**
	principl	**thought**	action	**peopl**
	say	**principl**	money	**thought**
	argument	**think**	principl	**societi**
	keyn	**pp**	capitalist	**principl**
	peopl	**logic**	great	**mainstream**
	critic	**say**	marx	**neoliber**
	york	**knowledg**	say	**human**
	argu	**sens**	plan	**keynesian**
	sens	**understand**	sociolog	**great**
	think	**critic**	thought	**thing**
	thought	**argument**	class	**crisi**
	knowledg	**veblen**	profit	**veblen**
	understand	**scientif**	regul	**book**
	marx	**man**	commun	**logic**
	object	**classic**	global	**say**
	behavior	**argu**	book	**heterodox**
	life	**essay**	object	**neoclass**
	theoret	**modern**	law	**action**
	modern	**chapter**	complex	**scientif**
	histor	**world**	thing	**evolutionary**
	ration	**thing**	keyn	**socialist**

The topic embodies many debates on the multiple – social, political, and theoretical – dimensions, which intersect the historical analysis of economics as social discipline. HET, *stricto sensu*, appears as main approach or framework only for those articles that show "Economics as social discipline" as their most prevalent topic. Remarkably, however, in all decades, HET emerges as part of a general topic – "Economics as social discipline" – concerning various kinds of analyses. LDA makes evident that HET is not only the professional field of study exploring theories and thinkers from the past, almost independently from considerations nurtured by current events and disciplinary practices. Over the course of time, HET has demonstrated itself to be an indispensable framework to investigate the foundations of economic theory, and one by means of which different approaches in theoretical and methodological terms can be compared. The topic in the time window 2010-2014 (Table 8, with λ = 0.6) clearly shows[14] that HET has especially served as a home or reference for heterodox and critical approaches, and has been used to legitimize perspectives which support a conception of economic theory as a social and institutional science against decontextualized formalism.

An accurate historical analysis of the complexity and variety of alternative research paths shaping today's fragmentation can provide the theoretical glue (or the big generalist picture lost in the fragmented world of specialization, see Trautwein 2017) needed for the analysis of economics as discipline. This requires historians of economic thought to engage in a close and permanent alliance with economic methodologists and shift their focus from how different the foundations of economics could have been to the different local foundations of the research programs of today's mainstream pluralism.

References

Angrist J., Azoulay P., Ellison G., Hill R., and Lu S.F. (2017), "Economic Research Evolves: Fields and Styles", *American Economic Review*, 107(5): 293-297.

Backhouse R.E. and Cherrier B. (2014) "Becoming Applied: The Transformation of Economics after 1970", Center for the History of Political Economy Working Paper 2014–15.

Blei D. M. (2012), "Probabilistic Topic Modeling", *Communications of the ACM*, 55(4): 77-84.

Blei D. M. and Lafferty J. D. (2006), "Dynamic Topic Models", Proceedings of the 23rd international Conference on Machine Learning. ICML '06, ACM, 113-120.

[14] The bag of words includes thinkers like Marx, Hayek, Veblen; schools of thought like the Keynesian and the neoclassical ones, but also the more general categories of "mainstream" and "heterodox" economics; terms clearly related to the research programs of "mainstream pluralism", like "complex", "evolutionary"; as well as words like "challenge" and "crisi". By looking specifically at the topic's most "relevant" terms, one encounters terms clearly related to the methodology of economics as sub-discipline, like "philosophi", "ontolog", "paradigm".

Blei D.M., Ng A.Y and Jordan M.I. (2003), “Latent Dirichlet Allocation”, *Journal of Machine Learning Research*, 3 (Jan): 993-1022.

Börner K., Klavans R., Patek M., Zoss A.M., Biberstine J.R., Light R.P., et al. (2012) “Design and Update of a Classification System: The UCSD Map of Science”, *PLoS ONE*, 7(7): e39464.

Card D. and DellaVigna S. (2013), “Nine Facts about Top Journals in Economics”, *Journal of Economic Literature*, 51(1): 144-161.

Cedrini M. and Fontana M. (2017), “Just Another Niche in the Wall? How Specialization Is Changing the Face of Mainstream Economics”, *Cambridge Journal of Economics*, 42(2): 427-51.

Cherrier B. (2017). “Classifying Economics: A History of the JEL Codes”, *Journal of Economic Literature*, 55(2): 545-79.

Cherrier B. (2015) “Is There a Quantitative Turn in the History of Economics (and How not to Screw It Up)”, The Undercover Historian. Beatrice Cherrier’s Blog, available at: https://beatricecherrier.wordpress.com/2015/06/23/is-there-a-quantitative-turn-in-the-history-of-economics-and-how-not-to-screw-it-up/.

Claveau F. and Gingras Y. (2016), “Macrodynamics of Economics: A Bibliometric History”, *History of Political Economy*, 48(4): 551-592.

Coats A.W. (2014), *The Historiography of Economics. The Collected Papers of A.W. Coats*, Vol. III, R. E. Backhouse and B. Caldwell (eds.), Abingdon: Routledge.

Colander D., Holt R., and Rosser Jr. J.B. (2004), “The Changing Face of Mainstream Economics”, *Review of Political Economy*, 16(4): 485–499.

Davis J.B. (2006), “The Turn in Economics: Neoclassical Dominance to Mainstream Pluralism?”, *Journal of Institutional Economics*, 2(1): 1-20.

Di Caro L., Guerzoni M., Nuccio M., and Siragusa G. (2017), “A Bimodal Network Approach to Model Topic Dynamics”, mimeo (available at: https://arxiv.org/abs/1709.09373).

Dogan, M. and Pahre R. (1989), “Fragmentation and Recombination of the Social Sciences”, *Studies in Comparative International Development*, 24(2): 56-72.

D'Orlando F. (2013), “Electronic Resources and Heterodox Economics”, *Review of Political Economy*, 25(3): 399-425.

Dow S.C. (2008), “Plurality in Orthodox and Heterodox Economics”, *The Journal of Philosophical Economics*, 1(2): 73-96.

Fontaine P. (2015), “Introduction: The Social Sciences in a Cross-Disciplinary Age”, *Journal of Theoretical Social Psychology*, 51(1): 1-9.

Fourcade M., Ollion E., and Algan Y. (2015), “The Superiority of Economists”, *Journal of Economic Perspectives*, 29(1): 89–114.

Heap S.P.H. and A. Parikh (2005), “The Diffusion of Ideas in the Academy: A Quantitative Illustration from Economics”, *Research Policy*, 34(10): 1619-1632.

Klaes M. (2017), “Quantitative Approaches to Historical Semantics in Economics”, paper presented at the 22nd Annual Conference of the European Society for the History of Economic Thought (ESHET), University of Antwerp, 18-20 May.

Kosnik L.-R. (2018), “A Survey of JEL Codes: What Do They Mean and are They Used Consistently?”, *Journal of Economic Surveys*, 32(1): 249-72.

Kosnik L.-R. (2015), “What Have Economists Been Doing for the Last 50 Years? A Text Analysis of Published Academic Research from 1960–2010”, *Economics: The Open-Access, Open-Assessment E-Journal*, 9 (2015-13): 1–38.

Knudsen C. (2002), “The Essential Tension in the Social Sciences: Between the ‘Unification’ and ‘Fragmentation’ Trap”, in H. S. Jensen, L. M. Richter, M. T. Vendelø (eds), *The Evolution of Scientific Knowledge*, Cheltenham: Edward Elgar: 13-35.

Kuhn T.S. (2000), *The Road since Structure: Philosophical Essays, 1970–1993, with an Autobiographical Interview*, ed. by J. Conant and J. Haugeland, Chicago: University of Chicago Press.

McCain, K.W. (2014), “Assessing Obliteration by Incorporation in a Full-text Database: JSTOR, Economics, and the Concept of “Bounded Rationality”, *Scientometrics*, 101(2): 1445-1459.

Marchionatti R. and Cedrini M. (2017), *Economics as Social Science*, London and New York: Routledge.

Mimno D. and Blei D. (2011). “Bayesian Checking for Topic Models”. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA: Association for Computational Linguistics, 227-237.

Mohr J.W. and Bogdanov P. (2013), “Introduction—Topic Models: What They Are and Why They Matter”, *Poetics*, 41(6): 545-569.

Moretti F. (2017), “Patterns and Interpretation”. Speech given at the “Distant Reading and Data-Driven Research in the History of Philosophy” Conference, Università di Torino, 16-18 January.

Moretti F. (2013), *Distant Reading*, London and New York: Verso.

Moretti F. (2005), *Graphs, Maps, Trees: Abstract Models for a Literary History*, London and New York: Verso.

Moretti F. and Pestre D. (2015), “Bankspeak: The Language of World Bank Reports, 1946-2012”, *New Left Review*, 92: 75-99.

Morris S.A. and Van der Veer Martens B. (2008), “Mapping Research Specialties”, *Annual Review of Information Science and Technology*, 42(1): 213-95.

Panhans M.T. and Singleton J.D. (2017), “The Empirical Economist’s Toolkit: From Models to Methods”, *History of Political Economy*, 49(S): 127-57.

Pencavel J. (1991), “Prospects for Economics”, *Economic Journal*, 101(404): 81-87.

Reay M.J. (2012), “The Flexible Unity of Economics”, *American Journal of Sociology*, 118(1): 45-87.

Rodrik D. (2015), *Economics Rules. Why Economics Works, When It Fails, and How to Tell the Difference*, New York: W.W. Norton, 2015.

Rhody L.M. (2012), “Topic Modeling and Figurative Language”, *Journal of Digital Humanities*, 2(1).

Sievert C. and Shirley K. E. (2014), “LDAvis: A Method for Visualizing and Interpreting Topics”, *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, Baltimore, Maryland: Association for Computational Linguistic, 63-70.

Suominen A. and Toivanen H. (2016), “Map of Science with Topic Modeling: Comparison of Unsupervised Learning and Human-assigned Subject Classification”, *Journal of the Association for Information Science and Technology*, 67(10): 2464-76.

Trautwein H.-M. (2017), “The Last Generalists”, *European Journal of the History of Economic Thought*, 24(6): 1134-66.