



## What topic modeling could reveal about the evolution of economics

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

To cite this article: Angela Ambrosino, Mario Cedrini, John B. Davis, Stefano Fiori, Marco Guerzoni & Massimiliano Nuccio (2018) What topic modeling could reveal about the evolution of economics, Journal of Economic Methodology, 25:4, 329-348, DOI: [10.1080/1350178X.2018.1529215](https://doi.org/10.1080/1350178X.2018.1529215)

To link to this article: <https://doi.org/10.1080/1350178X.2018.1529215>



Published online: 18 Oct 2018.



Submit your article to this journal [↗](#)



Article views: 613



View related articles [↗](#)




View Crossmark data [↗](#)



Citing articles: 3 View citing articles [↗](#)



# What topic modeling could reveal about the evolution of economics\*

Angela Ambrosino<sup>a</sup>, Mario Cedrini <sup>a</sup>, John B. Davis<sup>b,c</sup>, Stefano Fiori<sup>a</sup>, Marco Guerzoni<sup>a,d</sup> and Massimiliano Nuccio<sup>a</sup>

<sup>a</sup>Dipartimento di Economia e Statistica “Cognetti de Martiis”, Università di Torino, Turin, Italy; <sup>b</sup>Department of Economics, Marquette University, Milwaukee, WI, USA; <sup>c</sup>Department of Economics, University of Amsterdam, Milwaukee, WI, USA; <sup>d</sup>ICRIOS, Bocconi University, Milan, Italy

## 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 a map of the discipline over time, and illustrate the potentialities of the technique for the study of the shifting structure of economics in a time of (possible) fragmentation.

## ARTICLE HISTORY

Received 1 October 2017  
Accepted 13 July 2018

## KEYWORDS

Topic modeling; economics as science; economics literature; text analysis

## JEL CODES

B4; B1; B2; A1

## 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 & 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, p. 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)

**CONTACT** Mario Cedrini  [mario.cedrini@unito.it](mailto:mario.cedrini@unito.it)

\*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](http://www.despina.unito.it)).

© 2018 Informa UK Limited, trading as Taylor & Francis Group

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, behavioural 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 & Fontana, 2018) 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, & 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, p. 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 & Toivanen, 2016, p. 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 & 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, Azoulay, Ellison, Hill, and Lu (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, and in view of the general limitations of existing analytical methods, this paper proposes a radically new quantitative approach to the history of economics, which – by use of a topic modeling technique called Latent Dirichlet Allocation (LDA) – aims to detect the hidden, or 'latent' structure of the discipline. Section 2 first exposes the general philosophy of topic modeling and the concrete working of the now widely diffused LDA technique, and then the assumptions and qualifications upon which we have applied LDA to the corpus of economics articles published in the academic journals of the JSTOR database (whose characteristics are explored in the second part of the section). The aim, which is illustrated in Section 3, is to clarify how to construct the abovementioned map and investigate the historical evolution of economics and its changing structure by looking at the most relevant topics dealt with by economists in each decade of the twentieth century. The section introduces readers to the complexity of interpreting topics, on one side, and to the problems that derive from their peculiar (heterogeneous, or even multidimensional) nature, on the other, and illustrates how such problem can be addressed by means of *ad hoc* instruments. Section 4 presents some concluding reflections drawn from the study on topic modeling as possible complementary approach to a purely qualitative analysis of the discipline in historical perspective.

## 2. A topic-modeling analysis of economics

### 2.1. 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<sup>1</sup>), 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. Applied to the full text of archive documents (in the form of a list of all words appearing in each of them, with related frequency), LDA calculates probabilistic regularities, or trends in language texts, recurring themes in the form of co-occurring words. It is based on a 'bag-of-words' assumption (see Blei, Ng, & Jordan, 2003): texts are represented as the bag of their words – that are the primary entities

considered by the algorithm – with no consideration for grammar or word order (words, like documents, are therefore technically interchangeable), only for multiplicity. LDA considers exclusively frequency and co-occurrence of single words (monograms). Presupposing that words referring to similar subjects appear in similar contexts, LDA groups them 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.<sup>2</sup> 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 & Bogdanov, 2013).<sup>3</sup> 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<sup>4</sup>:

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

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, 2012; Rhody, 2012) for labelling and hermeneutically interpreting topics, *ex post*, and developing new possible theories based on the latent structure identified by the algorithm<sup>6</sup>. 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 & Bogdanov, 2013, p. 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 1000 times bigger’, since ‘the new scale changes our relationship to our object, and in fact *it changes the object itself*’ (Moretti, 2017, p. 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 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, & Tribe, 1997). And what economists do, in primis, is writing texts, and 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 nineteenth and early twentieth 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.

## 2.2. Dataset and methodology

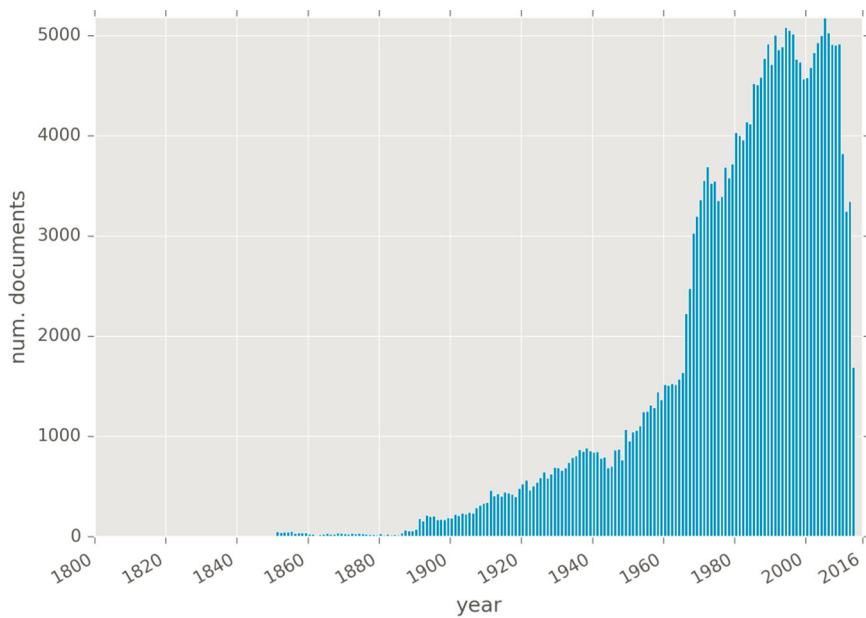
The dataset explored in this research programme, then, are 250,846 articles published from 1845 to 2013 in 188 journals stored in the digital library JSTOR. JSTOR Data for Research (DfR) provides datasets of content on JSTOR for use in research that can be automatically processed. Data available upon agreement on the use of the data itself<sup>7</sup> 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'<sup>8</sup> 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).

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).

Knowledge development is a complex process, based on the continuous emergence of new ideas and research programmes 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



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

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, Guerzoni, Nuccio, and Siragusa (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 programme.

**Table 1.** The JSTOR dataset.

Collection JSTOR database		Number of articles 250,846	Frequency distribution 100%	Cumulative frequency 100%
1	Economic and Political Weekly	45,118	18,0%	18,0%
2	The American Economic Review	15,408	6,1%	24,1%
3	Annals of the American Academy of Political and Social Science	13,380	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%



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 & 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. The hidden structure of economics

#### 3.1. Topic interpretation

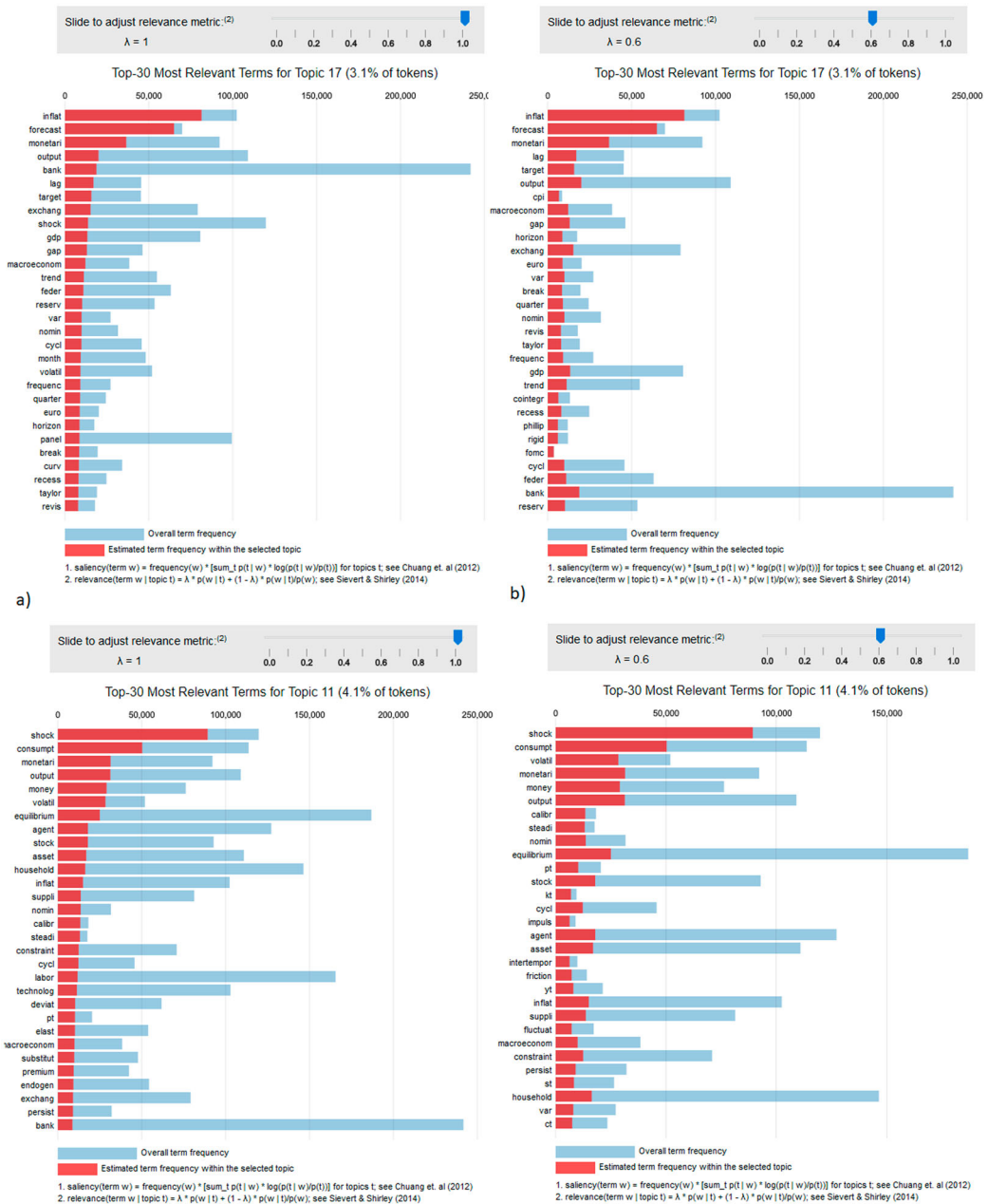
By identifying the most relevant topics in the various decades considered, LDA can serve to detect shifts in the changing structure of the economics discipline over time. However, it does not assist in labelling topics. Consider, for instance, the two following topics selected among the 27 of the time window '2010–2014'<sup>9</sup>:

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
inflat	macroeconomi
suppli	trend
nomin	feder
calibr	reserv
stead	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 Figure 2(a,b), 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 2(a). The absolute width of the red bar makes 'bank' appear as one of the most important words in defining Topic 17. Yet, it is quite a common term (as its blue bar shows) which





**Figure 2.** (a) Topic interpretation (topic 17): « Macroeconomic Theory », 2010–2014. (b) Topic interpretation (topic 11): « Macroeconomic Models », 2010–2014.

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  $\phi$  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  $\lambda$ , 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  $\lambda$  and takes the following form:

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

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

With  $\lambda = 0.6$  (right side of Figure 2(b)), the word list changes substantially, to include words that, although they do not figure among the commonest in the topic, are highly 'relevant' to it, that is, they most contribute to define it. 'Topic 11' can thus appear as one of 'Macroeconomic models', shaped by the lexicon of DSGE in particular (as is evident from the appearance of terms like 'calibr', 'stead', 'inter-tempor', '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.

### 3.2. A map of economics over time

Remarkably, LDAvis provides important details about the specific topics detected.

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).

Second, by selecting a term – for example, 'calibr' in Figure 3 – 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 4).

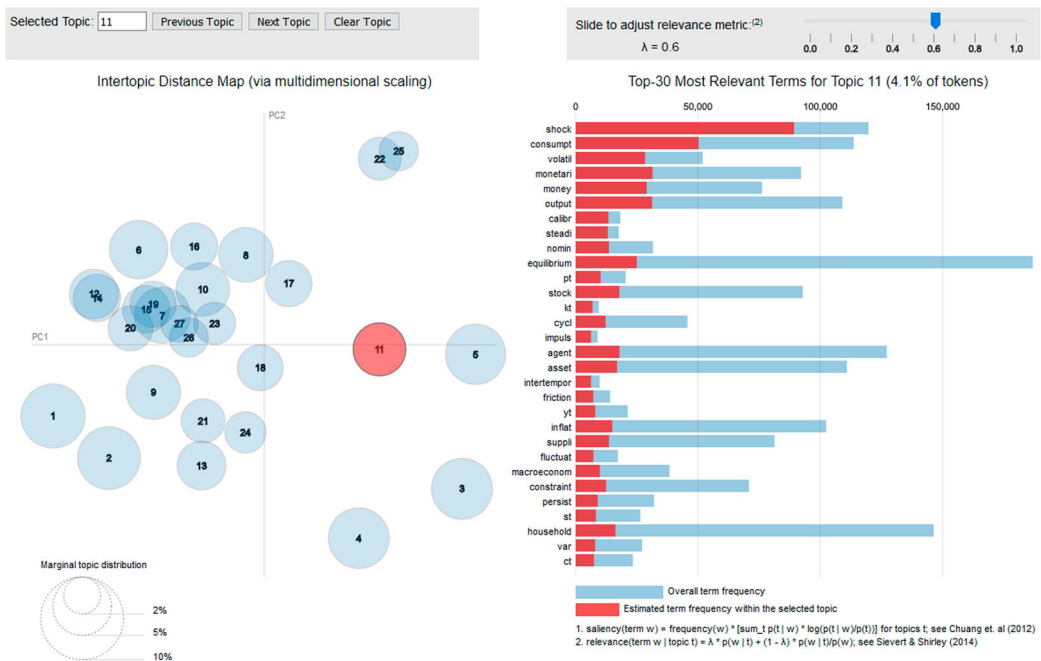


Figure 3. Degree of competition between topics within documents.

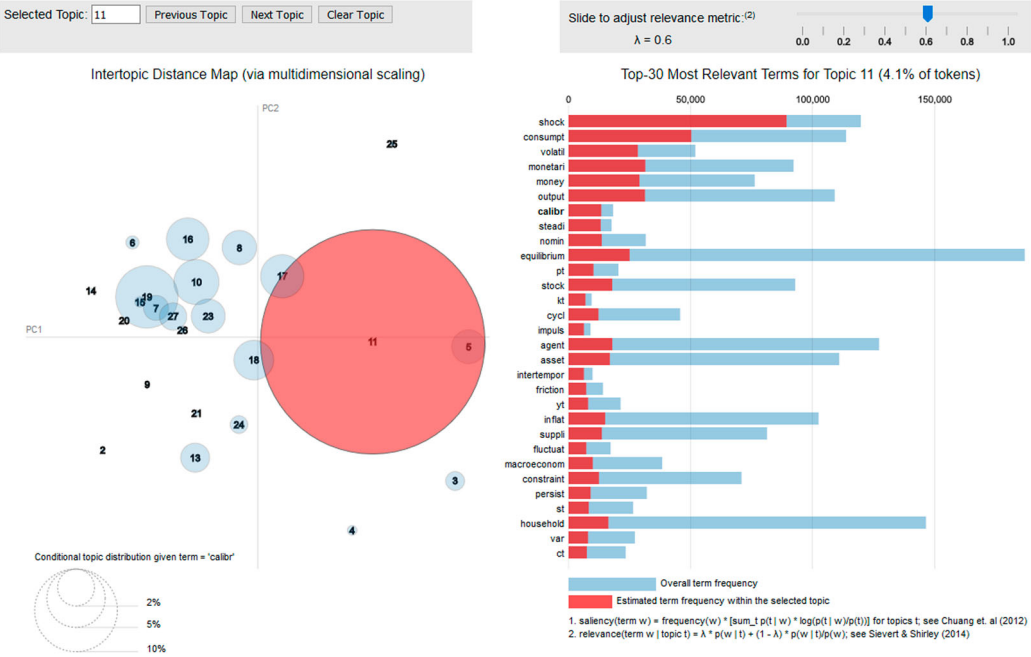


Figure 4. Terms' 'conditional distribution' over topics.

By labelling all topics in all decades, supported by LDAvis, it then becomes possible to visualize a 'map' of economics over time. Tables 2 and 3 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'.

Below follows the 'map' of economics as discipline since the Sixties (Tables 2 and 3).

As with any map, the one proposed here is in some sense arbitrary. This arbitrariness is due to the above mentioned 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, p. 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)?

3.3. The nature of topics (the case of law and economics)

To further analyse the nature of topics, one has to zoom in closer. Consider the decade 1940–1950 (Table 4): two different topics are detected showing terms (with  $\lambda = 1$ ; yellow column in the table) in the first ten positions that are broadly connected to regulation and law. Remarkably, both

**Table 2.** 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. <sup>1</sup>	4.5	Industrial organization: market structures, etc. <sup>1</sup>	4.9
Industrial organization: market structures, etc. <sup>1</sup>	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. <sup>2</sup>	4.3
Britain’s economic history	3.8	Insurance	3.6	Taxation, fiscal policy, fiscal behaviour	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	<i>French terms</i>	1.3	Africa	1.8
<i>Non Anglo-Saxon words</i>	1.5	<i>Non Anglo-Saxon words</i>	1.3	<i>Non Anglo-Saxon words</i>	1.0

<sup>1</sup>Industrial organization: market structures, firm strategies, and market performance; <sup>2</sup>Industrial organization: firm objectives, organization, and behaviour.

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 4 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  $\lambda$  chosen. As observed,  $\lambda$  can assume values between 0 and 1. While assuming  $\lambda = 1$  amounts to considering the most probable terms under each topic as generated by LDA, shifting the value of  $\lambda$  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 (we term it ‘Law and Economics’), whereas Topic 10 appears more regulation-oriented (‘Regulation’). As a matter of fact, with  $\lambda = 0.6$ , 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  $\lambda = 1$  is possibly due to the fact that they are both evidently involved in discussions concerning various issues generically linked to law, or to the fact that the approach which will become known as ‘Law and Economics’ has not reached the maturity required to stand alone.

After repeating the exercise for all decades and all possible sources of confusion, we find that the hidden structure of economics, so to speak, presents a topic whose words are quite evidently related

**Table 3.** 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	<i>Econometrics parameters</i>	5.3
Prediction/cycles (econometrics, applied)	4.7	Industrial organization: firm objectives, etc. <sup>2</sup>	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. <sup>2</sup>	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. <sup>1</sup>	4.2	Theoretical economics	3.7	Macroeconomic models (Dsge)	4.1
Demography	3.8	<i>Unusual terms</i>	3.7	Globalization	3.6
Monetary policy (open ec. macroeconomics)	3.7	Labour	3.6	Behavioural 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	<i>Unusual terms</i>	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
<i>Non Anglo-Saxon words</i>	1.8	Energy and environment	2.1	<i>Cyrillic letters</i>	2.2
East Asia (politics, military)	1.3	Public choice	1.9	Automotive industry, China	2.2
<i>Non Anglo-Saxon letters</i>	0.9	<i>Unusual terms</i>	1.2	Energy and environment	2.0

<sup>1</sup>Industrial organization: market structures, firm strategies, and market performance; <sup>2</sup>Industrial organization: firm objectives, organization, and behaviour.

to the typical lexicon used when applying an economic approach to legal issues, and share many of the words appearing in the lists that define each of them when visualized through LDAvis (and  $\lambda = 0.6$ ). We thus consider such topics for the four decades 1910–20, 1960–70, 1970–80, and 1980–90 (see Table 5), and the sequences of the first 20 terms that define them with  $\lambda = 1$  first, and  $\lambda = 0.6$  then. The focus is therefore on an early decade (1910–1920) when, as the JSTOR database shows, the field was quite important and highly represented in economics journals; and on the three decades (1960–70, 1970–80, and 1980–90) which roughly correspond to the institutionalization of the economic approach to law as a proper subdiscipline of economics.

Table 5 reports two columns for each decade, according to the value set for  $\lambda$ : terms in colored boxes are common to the two lists of terms generated by LDA with, respectively,  $\lambda = 1$  and  $\lambda = 0.6$ , while terms in white boxes appear in just one of the two lists. In other words, terms in yellow (with  $\lambda = 1$ ) become blue when  $\lambda = 0.6$ , whereas terms in white boxes 'define' the topic only when  $\lambda$  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  $\lambda = 1$  are also present in the  $\lambda = 0.6$  list, despite changes in their relative position. Looking at the lower part of the ranking, however, might help us see things differently. With  $\lambda = 0.6$ , the lists of terms in each decade signal a specific focus on 'law and trial' issues.

**Table 4.** Topic interpretation. ‘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	
	$\lambda = 1$	$\lambda = 0,6$	$\lambda = 1$	$\lambda = 0,6$
Terms	union board wage worker bargain law employe court member collect legisl strike local committe part right feder administr manag agreement constitut elect disput offic decis presid contract execut congress vote	<b>union</b> <b>board</b> <b>bargain</b> <b>employe</b> <b>wage</b> <b>court</b> <b>worker</b> <b>strike</b> <b>collect</b> <b>legisl</b> <b>law</b> <b>member</b> <b>disput</b> <b>elect</b> <b>parti</b> <b>arbitr</b> <b>membership</b> <b>local</b> <b>jurusdict</b> <b>execut</b> <b>vote</b> <b>committe</b> <b>presid</b> <b>constitut</b> <b>right</b> <b>agreement</b> <b>manag</b> <b>negro</b> <b>congress</b> <b>administr</b>	commiss competit regul transport railroad carrier retail law court sale util compani administr commerc charg traffic manufactur decis freight agenc feder interest commod ship air depart class sell consum store	<b>commiss</b> <b>carrier</b> <b>regul</b> <b>transport</b> <b>competit</b> <b>retail</b> <b>railroad</b> <b>traffic</b> <b>interest</b> <b>freight</b> <b>antitrust</b> <b>commerc</b> <b>court</b> <b>air</b> <b>regulatori</b> <b>charg</b> <b>sale</b> <b>store</b> <b>ship</b> <b>util</b> <b>vessel</b> <b>passeng</b> <b>manufactur</b> <b>law</b> <b>decis</b> <b>administr</b> <b>territori</b> <b>wholesal</b> <b>rail</b> <b>pilot</b>

Let us now glance at the ‘prevalence’ of these topics. 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. The topic of the decade 1910–1920 is ranked 6th, covering 4.9% of the articles in the corpus. The topic of the decade 1960–1970 is ranked 18th (2.9%). In the Seventies, the topic made up of words related to the economic approach to law is ranked 24th (2.2%): while the topic of the decade 1980–1990 is ranked 12th (3.5%). In itself, ‘prevalence’ is of little help in further interpreting topics. 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 the prevalence, in the corpus, of three of the four topics here focused on (the exception is the topic of the decade 1910–1920), is quite high (6.1%), while the weight of topics identified by a list of words related to economic analysis of law appears to be significantly lower in other decades (see Table 6).

Although an ‘economic analysis of law’ topic is always present with a non-negligible weight, it is never among the most representative in term of prevalence. As to the concentration index (see Figure 5(a,b)), such topics exhibit 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. In other decades, the topic is more concentrated (its HHI value is high in relative terms with respect to

**Table 5.** Topics related to the economic approach to legal issues, 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	
Terms	$\lambda = 1$	$\lambda = 0.6$	$\lambda = 1$	$\lambda = 0.6$	$\lambda = 1$	$\lambda = 0.6$	$\lambda = 1$	$\lambda = 0.6$
court		<b>court</b>	law	<b>law</b>	law	<b>law</b>	law	<b>law</b>
hous		<b>commiss</b>	act	<b>court</b>	court	<b>court</b>	right	<b>court</b>
commiss		<b>hous</b>	court	<b>act</b>	right	<b>legal</b>	court	<b>legal</b>
regul		<b>regul</b>	commiss	<b>commiss</b>	legal	<b>right</b>	rule	<b>right</b>
legisl		<b>legisl</b>	right	<b>legisl</b>	act	<b>crime</b>	legal	<b>act</b>
build		<b>tenement</b>	legisl	<b>legal</b>	rule	<b>enforc</b>	act	<b>regul</b>
provis		<b>judici</b>	legal	<b>crime</b>	properti	<b>act</b>	regul	<b>legisl</b>
corpor		<b>licens</b>	regul	<b>right</b>	crime	<b>defend</b>	legisl	<b>rule</b>
tenement		<b>build</b>	agreement	<b>regul</b>	contract	<b>supra</b>	properti	<b>enforc</b>
enforc		<b>statut</b>	rule	<b>enforc</b>	enforc	<b>crimin</b>	feder	<b>supra</b>
licens		<b>enforc</b>	administr	<b>amend</b>	defend	<b>accid</b>	parti	<b>damag</b>
constitut		<b>room</b>	committe	<b>justic</b>	parti	<b>rule</b>	action	<b>plaintiff</b>
district		<b>evil</b>	member	<b>agreement</b>	judg	<b>damag</b>	enforc	<b>litig</b>
statut		<b>justic</b>	decis	<b>crimin</b>	protect	<b>judg</b>	protect	<b>defend</b>
privat		<b>provis</b>	feder	<b>polic</b>	regul	<b>judici</b>	contract	<b>crime</b>
railroad		<b>suprem</b>	section	<b>suprem</b>	damag	<b>injuri</b>	damag	<b>protect</b>
judici		<b>polic</b>	action	<b>statut</b>	compens	<b>suprem</b>	liabil	<b>liabil</b>
york		<b>decis</b>	crime	<b>judici</b>	accid	<b>punish</b>	constitut	<b>amend</b>
investing		<b>liquor</b>	board	<b>jurisdict</b>	legisl	<b>amend</b>	claim	<b>victim</b>
decis		<b>sanitari</b>	power	<b>judg</b>	polic	<b>litig</b>	commiss	<b>justic</b>
room		<b>corpor</b>	offic	<b>rule</b>	person	<b>polic</b>	defend	<b>constitut</b>
evil		<b>privat</b>	control	<b>committe</b>	liabil	<b>contract</b>	privat	<b>action</b>
legislatur		<b>legislatur</b>	enforc	<b>administr</b>	constitut	<b>justic</b>	supra	<b>crimin</b>
justic		<b>prevent</b>	constitut	<b>violat</b>	crimin	<b>properti</b>	provis	<b>statut</b>
prevent		<b>district</b>	issu	<b>board</b>	supra	<b>lawyer</b>	plaintiff	<b>suprem</b>
board		<b>sallon</b>	procedur	<b>bill</b>	action	<b>plaintiff</b>	litig	<b>feder</b>
bill		<b>constitut</b>	restrict	<b>provis</b>	justic	<b>trial</b>	common	<b>properti</b>
administr		<b>health</b>	provis	<b>hear</b>	amend	<b>statut</b>	effici	<b>neglig</b>
commette		<b>judg</b>	judg	<b>constitut</b>	claim	<b>victim</b>	person	<b>commiss</b>
protect		<b>amend</b>	bill	<b>action</b>	priva	<b>compens</b>	practi	<b>parti</b>

other topics): it covers a relatively smaller number of papers, in which however it appears as the most prevalent one, or one of the few most prevalent. In such decades, the topic is a distinctive one in the discipline.

Remarkably, while such topics exhibit low HHI values in the early decades (1890–1900, 1900–1910, 1910–1920), the situation is reversed in the next ones, and their 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 topics concerning the economic approach to law attained high levels of concentration at the time when ‘Law and Economics’ reaches its maturity stage as specific subfield of the discipline.

Looking at the corpus of articles, it becomes evident that in the last three decades (excluding therefore the five years between 2010 and 2014), the topic defined by words related to ‘Law and Economics’ appears as the ‘first’ topic in a relatively high number of articles: the HHI value increases exactly in the ‘boom’ phase of the research programme. 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

**Table 6.** Prevalence of topics on the economic analysis of law, 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



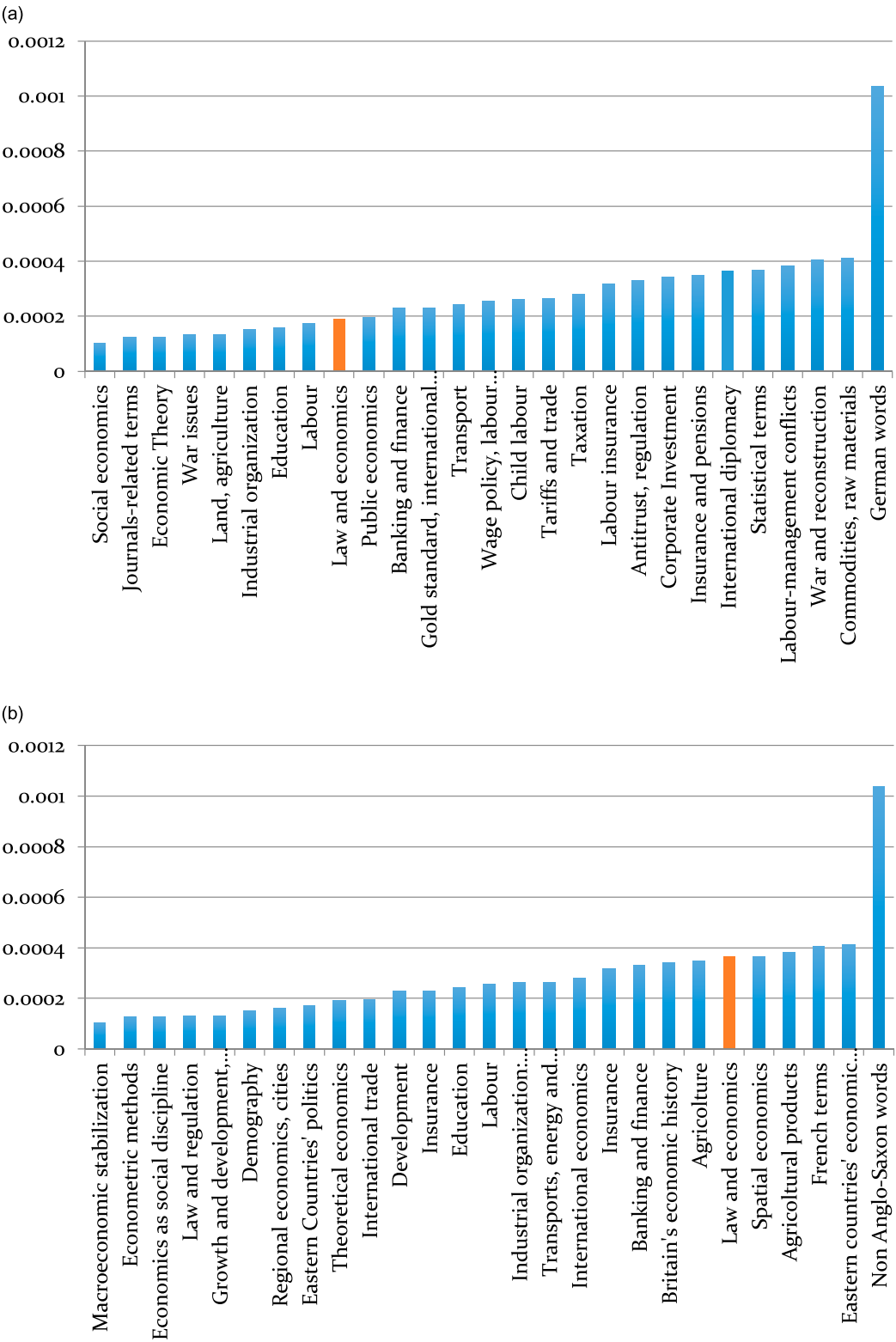


Figure 5. (a) Concentration indexes (HHI) for each topic, 1910–1920. (b) Concentration indexes (HHI) for each topic, 1970–1980.

values, one can infer that the topics related to 'Law and Economics' can be characterized as quite 'general' for much of the considered time period. Good performances in terms of prevalence are compensated by relatively low levels of concentration: in any time window, the topic labeled as '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: in this decade, the topic defined by words related to economics and law shows lower levels of prevalence accompanied by high concentration, concomitant with the development of 'Law and Economics' as specific, specialized and compact subdisciplinary approach in economics.

#### 4. Concluding remarks on topic modeling as analytical tool to investigate 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 and its changing structure. We have however shown that LDA can help to identify the salient topics dealt with by economists in a specific time window, and that specific instruments have been developed for zooming in on topics themselves, which allow further investigation of the 'hidden' structure of economics and, consequently, the possibility to use this latter to complement history-of-economics analyses.

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<sup>10</sup> clearly makes a difference, but the fundamental novelty of the approach lies, first, in considering the *full text* of economic articles (under the 'bag of words' assumption, as seen), rather than citations, bibliographical references, metadata, or any other specific feature<sup>11</sup>. 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 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,<sup>12</sup> 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, 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.

It seems reasonable to assume that 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 & 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 & Pahre, 1989, p. 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, 2018; see also Suominen & Toivanen, 2016) 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, p. 28).

The assumption made here that changes in the semantic content of topics follow the evolution of knowledge in the field is 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 & Toivanen, 2016). 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 & 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 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. 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 programmes of today’s mainstream pluralism.

## Notes

1. 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 (2012), Blei et al. (2003).
2. LDA is the state of the art for probabilistic topic modeling (Blei et al., 2003). It can be implemented in any programming language. This study uses the toolkit ‘Gensim’, implemented in Python: the toolkit has been specifically developed and successfully employed for large-scale datasets including huge collections of electronic documents. LDA provides, as output, estimations of: a distribution of topics which describes the corpus of documents; a distribution of words defining each topic; a distribution of topics for each article. The technical assumptions behind the LDA generative probabilistic model are explained at length in Blei (2012), Blei et al. (2003), to which we remind interested readers. LDavis (see below in the text, Section 3) is a visualization tool for topic models intended to simplify exploration of the output produced by LDA; it is based on the programming language R (library: LDavis).
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, p. 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.
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 & Bogdanov, 2013, p. 560)
7. Data was obtained upon agreement from JSTOR DfR (<https://www.jstor.org/df/r/>) 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.
9. The last time window considered is not a decade: the JSTOR database does not include articles published after 2014.
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).
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.

## Acknowledgement

We gratefully acknowledge financial support from the European Society for the History of Economic Thought (ESHET grant 2015).

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

We gratefully acknowledge financial support from the European Society for the History of Economic Thought (ESHET grant 2015).

## Notes on contributors

**Angela Ambrosino** is adjunct Professor of Microeconomics and Macroeconomics at the University of Turin. Her areas of specializations are institutional economics, history of economic thought and economic methodology. She is a Member of the Executive Committee of the Italian Association for the History of Political Economy (STOREP) since 2015.

**Mario Cedrini** is Assistant Professor of Economics at the University of Turin, Italy, where he currently teaches macroeconomics, international economics and economic methodology. His research has explored John Maynard Keynes's thought and 'method', and has focused in particular on the contemporary relevance of Keynes's international economics. His research interests also include economics as science, and the relationships between economics and other disciplines. He is Secretary-elect (since 2015) of the Italian Association for the History of Political Economy (STOREP), and editor (together with F. Cassata and R. Marchionatti) of the journal *Annals of the Fondazione Luigi Einaudi*. He is co-author with Roberto Marchionatti of *Economics as social science. Economics imperialism and the challenge of interdisciplinarity* (Routledge, 2017).

**John B. Davis** is Professor Emeritus of Economics at Marquette University and Professor Emeritus of Economics at the University of Amsterdam, and Fellow of the Tinbergen Institute. He is author of *Keynes's philosophical development* (Cambridge, 1994), *The theory of the individual in economics* (Routledge, 2003), *Individuals and identity in economics* (Cambridge, 2011), and co-author with Marcel Boumans of *Economic methodology: Understanding economics as a science* (Palgrave, 2010). He has been a visiting professor at the Sorbonne, Cambridge University, Erasmus University, and Duke University. He is a former editor of the *Review of Social Economy* (1987–2005), is a co-editor of the *Journal of Economic Methodology*, and an editor of the 'Routledge Advances in Social Economics' book series. He is a past president or chair of the History of Economics Society, the International Network for Economic Method, and the Association for Social Economics, and is a past vice-president of the European Society for the History of Economic Thought.

**Stefano Fiori** is Associate Professor of Economics at the Department of Economics and Statistics 'Cognetti de Martiis' (University of Turin). His research fields are focused on the history of economic thought and on the connection among philosophy, economics, and other social sciences, viewed in historical perspective. His scientific interests include pre-classical and classical economics, Austrian economics, institutional and new institutional economics, economic methodology, and theories of bounded rationality. He is the author of *Ordine, mano invisibile, mercato. Una rilettura di Adam Smith* (Utet, Torino 2001).

**Marco Guerzoni** is Associate Professor of Applied Economics at the University of Turin (Department of Economics and Statistics 'Cognetti de Martiis'), where he teaches principle of economics, data journalism, economics of culture, economics of innovation and technology policy. His research area covers management and economics of innovation, technology policy, and economics of culture. He has been recently working on the methodological implications of big-data and machine learning for business and in social science with a focus on the issues of model selection, inference, and hypotheses mining.

**Massimiliano Nuccio** (PhD in Information Economics) is a research coordinator at Despina Big Data Lab at the Department of Economics and Statistics 'Cognetti de Martiis', University of Turin, where he previously held a Marie Curie Fellowship. He is also deputy director of the Master in Data Science for Complex Economic Systems (MADAS) at Collegio Carlo Alberto, Torino. He has lectured in various universities in Italy and abroad, including Bocconi University, Milan, University of Bologna, IMT Lucca, American University of Dubai and Leuphana Universität Lüneburg. His research interests focus on consumption practices from the perspective of social sciences, combining approaches and theories from economics, sociology and geography. Recently, he has conducted research on the impact of digital transformation and data analytics on consumer behaviour, cultural industries and urban and regional development.

## ORCID

Mario Cedrini  <http://orcid.org/0000-0003-1059-4458>

## References

- Angrist, J., Azoulay, P., Ellison, G., Hill, R., & Lu, S. F. (2017). Economic research evolves: Fields and styles. *American Economic Review*, 107(5), 293–297.
- Backhouse, R. E., & Cherrier, B. (2014). *Becoming applied: The transformation of economics after 1970* (Working Paper 2014–15). Center for the History of Political Economy.
- Backhouse, R. E., Middleton, R., & Tribe, K. (1997, September 3–5). 'Economics is what economists do', but what do the numbers tell us? Paper for the Annual History of Economic Thought Conference, University of Bristol. Retrieved from [https://seis.bristol.ac.uk/~hirm/Downloadpapers/Backhouse,%20Middleton%20and%20Tribe%20\(1997\)%20Economics%20is%20what%20economists%20do%20con%20ver.pdf](https://seis.bristol.ac.uk/~hirm/Downloadpapers/Backhouse,%20Middleton%20and%20Tribe%20(1997)%20Economics%20is%20what%20economists%20do%20con%20ver.pdf)
- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84.
- Blei, D. M., & Lafferty, J. D. (2006). *Dynamic topic models*. Proceedings of the 23rd international Conference on Machine Learning. ICML '06, ACM, pp. 113–120.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., ... Boyack, K. W. (2012). Design and update of a classification system: The UCSD Map of science. *PLoS ONE*, 7(7), e39464.
- Card, D., & DellaVigna, S. (2013). Nine facts about top journals in economics. *Journal of Economic Literature*, 51(1), 144–161.
- Cedrini, M., & Fontana, M. (2018). Just another niche in the wall? How specialization is changing the face of mainstream economics. *Cambridge Journal of Economics*, 42(2), 427–451.
- 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. Retrieved from <https://beatricecherrier.wordpress.com/2015/06/23/is-there-a-quantitative-turn-in-the-history-of-economics-and-how-not-to-screw-it-up/>
- Cherrier, B. (2017). Classifying economics: A history of the JEL codes. *Journal of Economic Literature*, 55(2), 545–579.

- Claveau, F., & 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 & B. Caldwell). Abingdon: Routledge.
- Colander, D., Holt, R., & 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., & Siragusa, G. (2017). A bimodal network approach to model topic dynamics. mimeo. Retrieved from <https://arxiv.org/abs/1709.09373>
- Dogan, M., & 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. (2018). Economics: The view from below. *Swiss Journal of Economics and Statistics*, 154(5). doi:10.1186/s41937-017-0019-2
- Fourcade, M., Ollion, E., & Algan, Y. (2015). The superiority of economists. *Journal of Economic Perspectives*, 29(1), 89–114.
- Heap, H. S. P., & Parikh, A. (2005). The diffusion of ideas in the academy: A quantitative illustration from economics. *Research Policy*, 34(10), 1619–1632.
- Klaes, M. (2017, May 18–20). *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.
- 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* (pp. 13–35). Cheltenham: Edward Elgar.
- 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.
- Kosnik, L.-R. (2018). A survey of JEL codes: What do they mean and are they used consistently? *Journal of Economic Surveys*, 32(1), 249–272.
- Kuhn, T. S. (2000). *The road since structure: Philosophical essays, 1970–1993, with an autobiographical interview* (J. Conant & J. Haugeland). Chicago, IL: University of Chicago Press.
- Marchionatti, R., & Cedrini, M. (2017). *Economics as social science*. London: Routledge.
- 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.
- Mimno, D., & Blei, D. (2011). Bayesian checking for topic models. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 227–237). Stroudsburg, PA: Association for Computational Linguistics.
- Mohr, J. W., & Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. *Poetics*, 41(6), 545–569.
- Moretti, F. (2005). *Graphs, maps, trees: Abstract models for a literary history*. London: Verso.
- Moretti, F. (2013). *Distant reading*. London: Verso.
- Moretti, F. (2017, January 16–18). Patterns and interpretation. Speech given at the "Distant Reading and Data-Driven Research in the History of Philosophy" Conference, Università di Torino.
- Moretti, F., & Pestre, D. (2015). Bankspeak: The language of World Bank reports, 1946–2012. *New Left Review*, 92, 75–99.
- Morris, S. A., & Van der Veer Martens, B. (2008). Mapping research specialties. *Annual Review of Information Science and Technology*, 42(1), 213–295.
- Panhans, M. T., & Singleton, J. D. (2017). The empirical economist's toolkit: From models to methods. *History of Political Economy*, 49(5), 127–157.
- Pencavel, J. (1991). Prospects for economics. *The Economic Journal*, 101(404), 81–87.
- Reay, M. J. (2012). The flexible unity of economics. *American Journal of Sociology*, 118(1), 45–87.
- Rhody, L. M. (2012). Topic modeling and figurative language. *Journal of Digital Humanities*, 2(1). Retrieved from <http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/>
- Rodrik, D. (2015). *Economics rules. Why economics works, when it fails, and how to tell the difference*. New York: W.W. Norton.
- Sievert, C., & Shirley, K. E. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70). Baltimore, MD: Association for Computational Linguistic.
- Suominen, A., & 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), 2464–2476.
- Trautwein, H.-M. (2017). The last generalists. *The European Journal of the History of Economic Thought*, 24(6), 1134–1166.