Online Appendix to "The History of Economics Society Bulletin and Journal of the History of Economic Thought (1979-2023)," by José Edwards, Yann Giraud, and Ivan Ledezma

December 2024 issue (DOI: https://doi.org/10.1017/S1053837224000208)

This appendix provides further information of the data collected and methodological details of the quantitative analyses used to explore the content of the *History of Economics Society Bulletin* and *Journal of the History of Economic Thought* (*HESB*/*JHET*).

I. GENERAL DESCRIPTION OF THE CONTENT OF THE *HESB*/*JHET*

As pointed out in the main text, the *HESB*/*JHET* contains 1,825 documents, which we classified into 763 research articles, 660 book reviews, and 402 "other documents" (i.e., editorial notes, guest lectures, presidential addresses, symposia, interviews, etc.). Given the different nature of these documents, and the available metadata, we studied them separately.[1] For the "other documents" we undertook a qualitative historiographic approach. These documents are heterogeneous, connected to the workings of the History of Economics Society (HES) for the most part, and so their analysis went hand in hand with that of the history of the HES. Quantitative analyses were more appropriate to explore the standardized document-samples of book reviews and research articles.

Book Reviews

Unlike research articles, book reviews are not submitted to the *JHET*, but solicited by book review editors. Table A1 lists the most frequent book title-words and reviewers for three editorial periods: 1989-98, 1999-2008, and 2009-23 (more in the main text). This Table will also serve as an introduction to the insights we draw from the research articles.

---

[1] In some cases, the distinction between "research articles" and "other documents" is not clear-cut from either Edwards and Martin (2019) or the Web of Science. Such is the case of long comments or replies on a research article, or of presidential addresses, appearing in the articles section of the journal. In general, we defined research articles as all published documents in the articles section of the journal (1988-2023) containing both an abstract and reference-list. In a few conspicuous cases we completed the indexing of the documents by using the first paragraph of the articles as abstracts.

TABLE A1: Book title-words and reviewers last-names (counts in brackets)

Period	Title-words	Reviewers last-names
1989-1998 (99)	Keynes/Keynesian (12), Monetary/Money (10), Life (7), Modern (7), Classical (5), Development (5), Revolution (5), Political (4), Schumpeter (4), Smith (4)	Heyne (3), Perlman (3), Boland (2), Brandis (2), Egger (2), Grabowski (2), Gram (2), Hollander (2), Klein (2), Landreth (2), Moss (2), Ramstad (2), Rashid (2), Skaggs (2), Steindl (2), Trescott (2), Vaughn (2)
1999-2008 (183)	Keynes/Keynesian (12), Modern (12), Science (12), Smith (11), Legacy (9), Classical (7), Ideas (7), International (7), Marshall (7), Money/Monetary (7), Social (7), Critical (6), Intellectual (6), Market (6), Political (6), Trade (6), British (5), Hayek (5), Models (5), Nature (5), American (4), Biography (4), Companion (4), Development (4), Dictionary (4), Equilibrium (4), Evolution (4), Macreoeconomics (4), Microeconomics (4), Revolution (4), Schumpeter (4), Veblen (4)	Backhouse (5), Middleton (4), Sent (4), Hammond (3), Hands (3), Hoover (3), Peart (3), Young (3), Barber (2), Bateman (2), Boumans (2), Charles (2), Coats (2), Egger (2), Emmett (2), Finkelstein (2), Groenewegen (2), Langholm (2), Lipkes (2), Lowry (2), Madden (2), Maes (2), Marciano (2), McCann (2), Meardon (2), Medema (2), Mehrling (2), Moggridge (2), Pearson (2), Porter (2), Rutherford (2), Tribe (2), Waterman (2), Weintraub (2), Whitaker (2), Worland (2)
2009-2023 (378)	Keynes/Keynesian (26), Smith (25), Money/Monetary (23), Science (20), Markets (17), Finance/Financial (16), Modern (16), American (14), Politics (14), Revolution (13), Intellectual (12), Nature/Natural (12), Marx/Marxian (11), Hayek (10), Biography (9), Development (8), State (8), War (8), World (8), Evolution (7), Knowledge (7), Law (7), Mill (7), Classical (6), Companion (6), Credit (6), Democracy (6), Enlightenment (6), Growth (6), Philosophy (6), Wealth (6)	Bachhouse (9), Paganelli (6), Baruchello (5), Young (5), Boumans (4), Caldwell (4), Couyoumdjian (4), Khan (4), Levy (4), Marciano (4), Mata (4), Medema (4), Bilo (3), Dimand (3), Duarte (3), Emmett (3), Giocoli (3), Hirai (3), Hoover (3), Innocenti (3), King (3), Mehrling (3), Pack (3), Serra (3), Signorino (3), Waterman (3)
All (660)	Keynes/Keynesian (50), Money/Monetary (40), Smith (40), Science (34), Market (27), Social (27), America (24), Revolution (24), Nature (21), Financial (19), Intellectual (19), Life (19), Classical (18), Hayek (18), Revolution (18), Development (17), Biography (15), American (13), Marshall (13), Growth (12), Legacy (12), Policy (12), Evolution (11), Macroeconomics (11), State (11), Trade (11), War (11), Marx (11), Companion (10), International (10), Law (10), Moral (10), Schumpeter (10), Value (10), Austrian (9), British (9), Capitalism (9), Hume (9), Philosophy (9), Wealth (9), World (9) Democracy (8), Equilibrium (8), Europe (8), Public (8)	Backhouse (14), Young (9), Hoover (7), Boumans (6), Marciano (6), Medema (6), Paganelli (6), Smith (6), Waterman (6), Baruchello (5), Caldwell (5), Dimand (5), Emmet (5), Mehrling (5), Middleton (5), Charles (4), Couyoumdjian (4), Egger (4), Fontaine (4), Forget (4), Hammond (4), Hands (4), Heyne (4), Khan (4), Levy (4), Mata (4), Pack (4), Peart (4), Sent (4), Signorino (4), Vaughn (4), Weintraub (4)

The center column in Table A1 lists the most frequent title-words for reviewed books. As expected, J. M. Keynes and Adam Smith dominate among the Great Economists given book-length treatment and reviewed in the *HESB/JHET*. F. Hayek, A. Marshall, K. Marx. J. Schumpeter, and D. Hume are also among the most frequently treated authors.

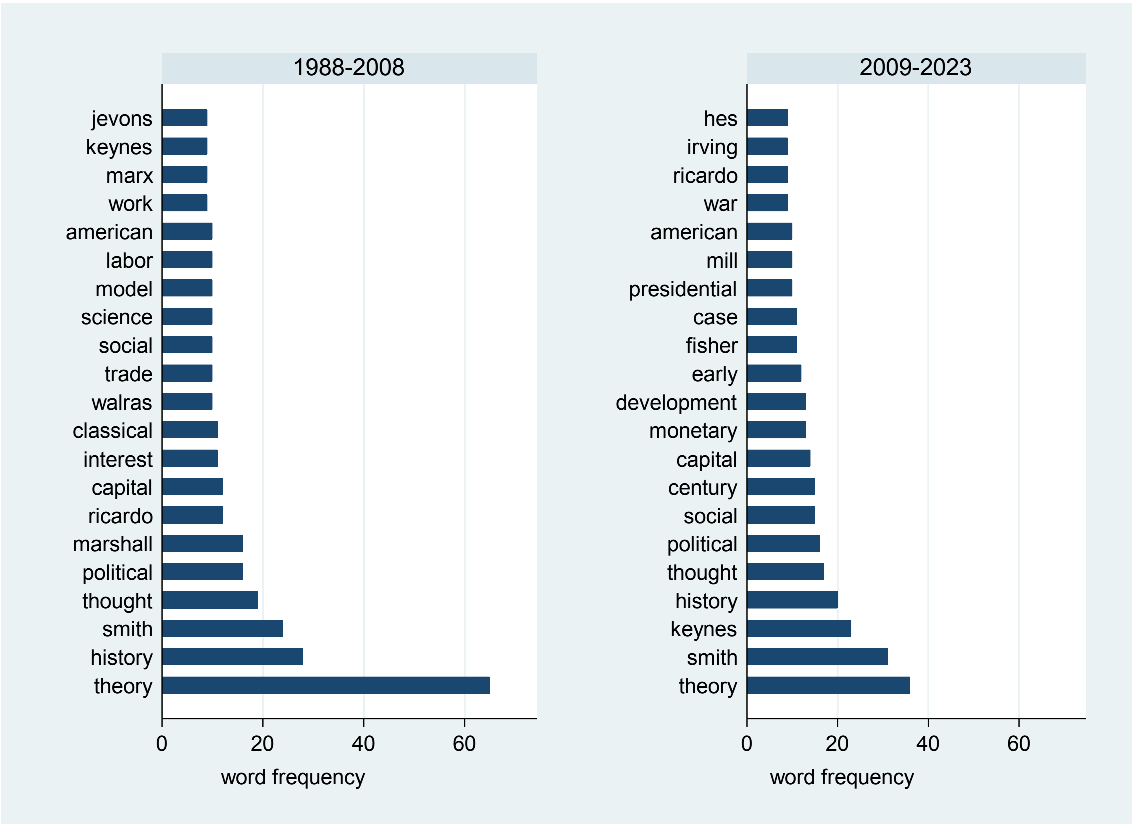
The right column in Table A1 lists the most frequent reviewers for the three periods. Asides from the information of the book review authors, it must be noted that the relative share of book reviews among the different document-types in the *HESB*/*JHET* has increased in time, a trend also visible in Figure 1 (main text). However, at the same time the average number of reviewers has decreased from 0.81 during the first period (i.e., 80 different reviewers for 99 book reviews), to 0.74, and then 0.64. The book review authorship in this journal has thus concentrated in time, as opposed to the authorship of research articles.

Research Articles

Whereas the indexation of book reviews includes mainly just information for “book titles” and “reviewers,” that of the 763 research articles in the *HESB/JHET* also includes standardized metadata on the different authors, titles, abstracts, and reference-lists.[2] However, to exploit these data one must take into account the different indexations of the dataset: by Edwards and Martin (2019) from 1988 to 2008, and by Clarivate’s Web of Science since 2009. This feature is especially critical for the reference-lists (e.g., citations appeared as footnotes in early issues of the journal). As we shall see later, for our network analyses we considered these two indexing periods separately.

Figure A1 lists the most frequent title-words for the two indexing periods, in order to grasp a first few insights about the content of the research articles in the journal.

FIGURE A1. Most frequent title-words in research articles

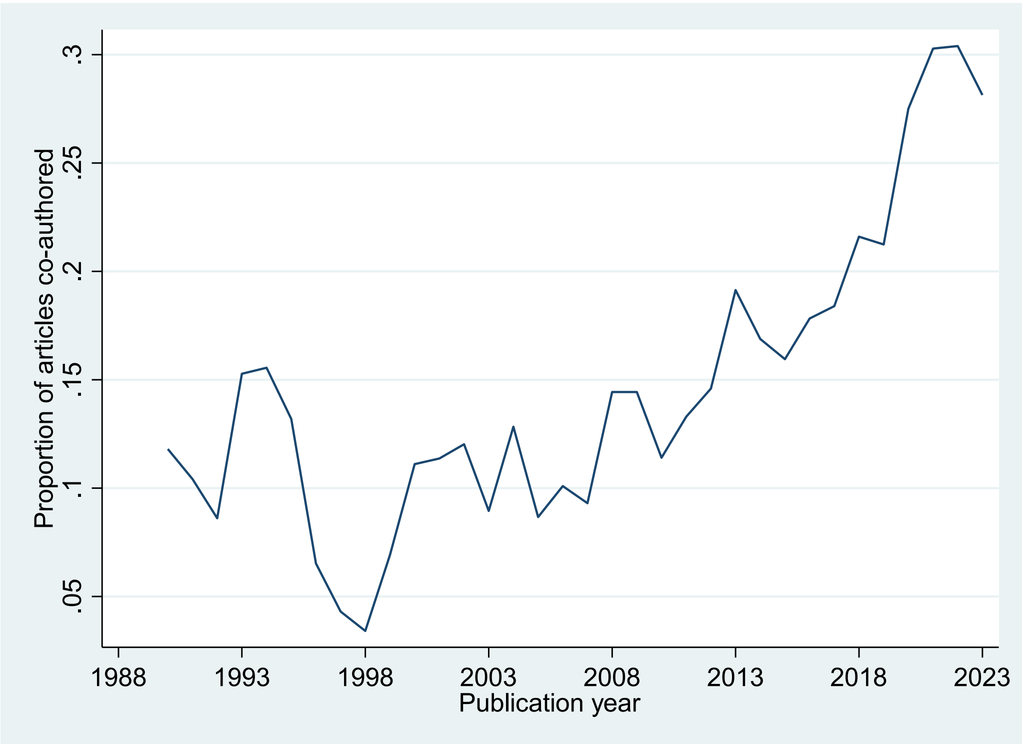


Note: this figure presents the list of words with the 10% highest frequency for each period. Single words and bigrams having at least 3 occurrences were extracted, and frequent words/bigrams adding no additional meaningful information (e.g., first names appearing along with a given last name) were dropped. We also removed stopwords (i.e., prepositions, adverbs, particles, interjections, etc.) and less informative words such as “analysis,” “approach,” “economic,” “economics,” “economist,” “economy.”

[2] For a description of the full metadata, see Edwards and Martin (2019).

The analysis of most frequent words (10% most frequent words for each period) reveals a few comparable patterns relative to book reviews and to our analysis in the main text. We note, for instance, high frequencies for A. Smith and other Great Economists during the first period, and the rise of prevalence of J. M. Keynes or I. Fisher the second period, suggesting a shift in time from published research on classical political economy to monetary/financial themes. Words capturing the methodological orientation of the journal—e.g. "history," "theory," "thought"—are also strongly represented, especially the first period. Our topic model (below) captures these terms together with more detail and context for them.

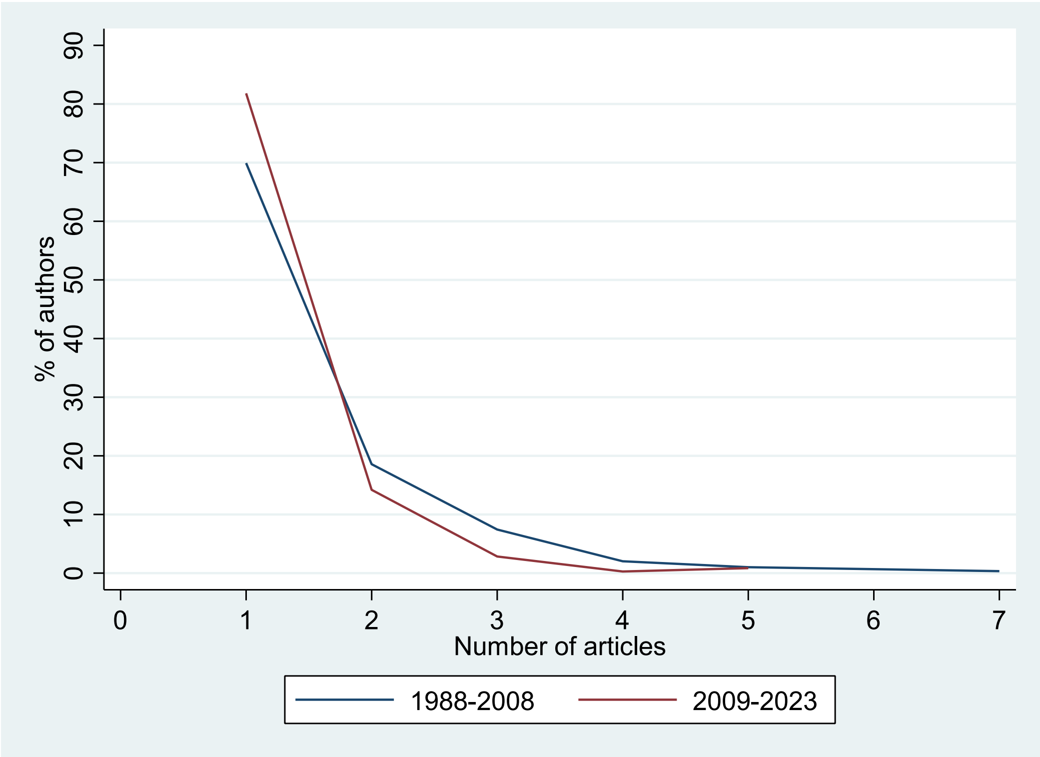
FIGURE A2. Co-authorship over time (1988-2023), 3-year moving averages



Note: this figure shows the evolution of the share of articles written by more than one author. In order to smooth the series, a 3-year moving average is plotted using the full sample period.

Figure A2 represents the tendency of increasing co-authorship over time in the journal, especially since 2008. Figure A3 shows the flattening of the distribution of research articles by author for the second period. Unlike book reviews, the concentration of article-authorship in the journal has been decreasing in time. As pointed out in the main text, we detected 0.73 different authors by article for 1988-2008 vs. 0.99 for 2009-23.

FIGURE A3. Frequency of publication by authors



Note: this figure depicts the frequency of publications by authors for each of our two indexing periods. These distributions are usually referred to as Lotka's law.

II. NETWORK ANALYSES

Beyond word-counts and the analysis of authorship in the journal, one may proceed through network analyses by pairing the different research articles using the similarity of their reference-lists. In line with the methodology of bibliographic couplings, our aim here was to exploit the structure of the resulting networks and to explore them by detecting communities of closely related documents.

Coupling articles using Latent Semantic Analysis

Latent Semantic Analysis (LSA) seeks to reveal the underlying semantic connection between words and documents appearing in large text corpora. Starting from the frequency in which terms are used in different documents, LSA infers the semantic closeness between words so that we can also infer the proximity between documents depending on their "meaning." Going beyond word or n-gram counts is important since we need to take into account that a same word may convey different notions—i.e., polysemy—and that alternative words may be used to express the same underlying concept—i.e., synonymy. The idea is thus to reduce the "semantic space" stemming from the association of words and documents by focusing on common "meanings" rather than on the terms used. To this end, LSA relies on linear algebra operations comparable to those used in principal component analysis, namely truncated singular value decomposition, which extracts "component vectors" capturing the essence of the relationships between documents.[3]

Based on the reduction of a semantic space into its essential components, it is possible to measure the similarity of two documents (i.e., two reference-lists in our case) by computing the "cosine similarity" between the component vectors of them. This amounts to computing the cosine of the angle between the two vectors (their scalar product divided by the product of their length), which goes from -1 to 1.[4] With this information in hand we constructed two weighted undirected networks of research articles, using the similarity of their reference-lists for each of the two periods (i.e., 1988-2008 and 2009-23). In these two networks the nodes are the research articles and their full linkage is characterized by the matrix of cosine similarity between their reference-lists, which we used as weights. We reduced the resulting networks by dropping links indicating dissimilar documents—i.e., links with negative weights—which represented less than 2% in our sample. Unlike standard bibliographic couplings, our method relies on the semantic proximity of the different reference-lists instead of counting exact matches. We expected this method to be robust to differences in the indexation of the reference-lists.

Application

After creating the networks, our next step was to explore them using Blondel et al.'s (2008) community detection algorithm. That algorithm is a heuristic method of community detection relying on "modularity" optimization. It seeks to detect communities (or modules) in the networks featuring high link-densities within communities relative to lower densities between communities. In running the algorithm, we were able to repeatedly detect subsets of research articles having high internal connections—i.e., having similar reference-lists—relative to the rest. It is worth stressing here that we used community detection as a methodology to repeatedly examine the two networks by using several different parameters for the algorithm (see also Edwards 2020).

---

[3] For descriptions of LSA, see Deerwester et al. (1990) and Martin & Berry (2007). For our analyses, we used the STATA implementation of LSA by Schwarz (2019).

[4] A similar mapping between components and words allows for the analysis of relationships between words. For instance, in our analysis of reference-lists, the closest words to "Schumpeter" we found were "Joseph," "innovation," "analysis," and "1954," the year of publication of his History of Economic Analysis.

Table A2 reproduces Table 1 (main text), but now including all 20 communities detected (10 each period), together with a bottom-line of the most frequent authors for each network. These frequencies (i.e., authors with three or more research articles in each network) are consistent with the previously discussed concentration of authorship in the journal.

TABLE A2: 20 communities in two article networks (counts in brackets)[5]

1988-2008		2009-2023	
A (71) 17.5%	Walras (9), Quesnay (6), French (4), Pareto (4) // Barnett (5), Eltis (3), Walker (3), Creedy (2), Daal (2), Frobert (2), Guidi (2), Steiner (2), Tarascio (2), Weber (2)	K (56) 17.1%	Curve (5), Law (5), Friedman (4), Phillips (4), War (4) // Forder (3), Marciano (3), Berta (2), Cherrier (2), Edwards (2), Giraud (2), Medema (2), Rancan (2)
B (61) 15%	American (6), Clark (6), Fisher (6), Institutions (6) // Dimand (5), Rutherford (5), Fiorito (3), Prasch (3), Barber (2), Gunning (2), Henry (2), Leathers (2), Raines (2), Steindl (2), Tilman (2)	L (48) 14.6%	Smith (23), Money/Monetary (4), Banking (3) // Paganelli (3), Evensky (2), Waterman (2)
C (55) 13.6%	Heterodox (5), Knowledge (5), Golinski (4), Science (4) // Backhouse (7), Weintraub (4), Bateman (2), Colander (2), Dow (2), Hands (2), Hoover (2), Hynes (2), Rutherford (2)	M (44) 13.4%	Fisher (6), Financial (5), Model (4), Statistical (3) // Backhouse (2), Biddle (2), Duppe (2), Hoover (2), Schinckus (2)
D (48) 11.8%	Smith (19), Classical (5), Trade (5), Labor (4) // Noell (3), Rashid (3), Young (3), Hueckel (2), Khalil (2), Levy (2), Maneschi (2)	N (44) 13.4%	Mill (9), Ricardo (8), American (4), Trade (4) // Depoortere (3), Baronian, (2), Besomi (2), Bianchini (2), Meardon (2)
E (44) 10.8%	Keynes (10), Interest (5), Wicksell (4), Money (4) // Aslanbeigui (3), Oakes (3), Boianovsky (2), DeVroey (2), Moggridge (2), O'Donnell (2), Tilman (2)	O (41) 12.5%	Keynes (18), Money/Monetary (7), Fisher (5) // Barnett (2), Boiakovsky (3), Kent (2), Rivot (2)
F (42) 10.3%	Jevons (7), Mill (6), Classical (4) // Hollander (4), Moore (4), Peart (3), White (3), Aldrich (2), Churchman (2), Hirsch (2), Kern (2)	P (29) 8.9%	Progressive-ism (5), American (4), Commons (3) // Fiorito (5), Betancourt (2), Chasse (2), Medema (2), Vallois (2)
G (25) 6.2%	Marx (9), Ricardo (9), Value (5), Profit (4) // Caravale (2), Howard (2), Keen (2), King (2)	Q (25) 7.6%	Quesnay (4), Capital (3), Dupuit (2), French (2) // Charles & There (2), Fossati (2), Numa (2)
H (22) 5.4%	Hayek (7), Austrian (2), Evolution (2), Mises (2) // Boettke (2), Leathers (2)	R (24) 7.3%	Hayek (6), Robbins (6), Capital (3), Menger (2) // Cachanosky (2)
I (22) 5.4%	Marshall (12), A. Young (3), Development (3) // Bowman (2), Buchanan (2), Niman (2)	S (12) 3.7%	Development (6), African (2), Case (2), Cycles (2) // Alacevich (2), Boianovsky (2)
J (16) 4%	Capital (7), Bawerk (5), Interest (4) // Ahmad (2), Dorfman (2), Murphy (2), Samuelson (2)	T (5) 1.5%	Coats' legacy (2009) by Augello & Guidi, Cardoso, Dudenhefer, Faucci, Medema
All (406 of 410)	Backhouse (7), Rutherford (7), Barnett (6), Dimand (6), Hollander (6), Leathers (5), Aslanbrigui (4), Khalil (4), Moore (4), Samuelson (4), Tilman (4), Walker (4), Weintraub (4), Bateman (3), Blaug (3), Boianovsky (3), Bowman (3), Coats (3), Creedy (3), Davis (3), DeVroey (3), Eltis (3), Fiorito (3), Goodwin (3), Henderson (3), Howard (3), Kern (3), King (3), Levy (3), Maneschi (3), Mongiovi (3), Noell (3), Oakes (3), Peart (3), Prasch (3), Raines (3), Rashid (3), Samuels (3), White (3), Young (3)	All (328 of 353)	Medema (6), Boianovsky (5), Fiorito (5), Dimand (4), Ahiakpor (3), Backhouse (3), Cherrier (3), Depoortere (3), Duppe (3), Forder (3), Giraud (3), Marciano (3), Numa (3), Paganelli (3), Rivot (3)

[5] Unlike our topic model (below), we were unable to include publication year 2023 in our network analyses, which is the reason why the second network includes just 328 of the 353 research articles. For the first period, our analysis excluded four research articles due to their short reference-lists. For descriptions of this kind of network analysis, see Edwards (2020).

In order to further illustrate our methodology let us consider community "C." A central article in this community is "How should we approach the history of economic thought: fact, fiction or moral tale?" (Backhouse 1992). The five most closely related articles to it are: "Why do evaluative histories matter after all?" (Zouboulakis 2001); "Making economic knowledge: review of Jan Golinski's Making Natural Knowledge" (Sent 2001); "Economic science wars" (Weintraub 2007); "Why teach the history of economics?" (Vaughn 1993); "More economics, please: we're historians of economics" (Moscati 2008). This association found *via* LSA illustrates the orientation of this community towards historiographic discussion, which is also what emerges from reading the six documents. In connection with our argument in the main text, "profession" is the closest word to "historiography" found through our LSA model.

III. TOPIC MODELING

Alternatively, we quantitatively examined the research articles in the *HESB*/*JHET* by associating these documents (i.e., their titles, abstracts, and reference-lists) to "latent" topics, which intricately relate words to texts in a probabilistic model describing the generative process of the text corpus.

Revealing latent topics using Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is an algorithm used for topic modeling, which allows for the automatic clustering of texts using machine-learning techniques. For an exogenously given number of topics, LDA models texts as a likelihood function. Basically, the model assumes that topics are a mixture of words, and texts a mixture of topics.[6] LDA then seeks the optimal topic assignment for each word in each text so as to maximize the likelihood (implied by the model) of observing words within the different texts. Finding this assignment also requires the distribution of word probabilities for topics and the topic probabilities for the texts. LDA assumes a generative process where these probabilities follow Dirichlet distributions.

For feasibility reasons, approximative inference algorithms are used to identify the optimal topic assignment and the parameters of the Dirichlet probability distributions. We rely here on the standard Gibbs sampling algorithm, which iteratively updates the topic assignment of words con-

[6] More precisely, LDA considers the probability of observing a word conditionally on dealing with a specific topic, and the probability of dealing with a topic conditionally on considering a given text. See Blei et al. (2003) for a presentation. Our implementation is based on the STATA routine proposed by Schwarz (2018) completed by R packages for topic modeling.

sidering the rest of assignments as given. To this end the algorithm needs priors about the parameters of the latent Dirichlet allocation of the data generating process, which are the inputs of its iterative Bayesian inference. After convergence of the (approximated) likelihood, each topic is described as a probability distribution over words and each text as a probability distribution over topics.

Given this output, for each topic one may summarize their content using the words associated to them with higher probabilities, and also check the texts for which the topics are highly prevalent—i.e., the texts for which a topic showed high probability. Although we applied the algorithm to the full sample period, we were able to identify the prevalence of the different topics in time by using the metadata of publication year for each text.

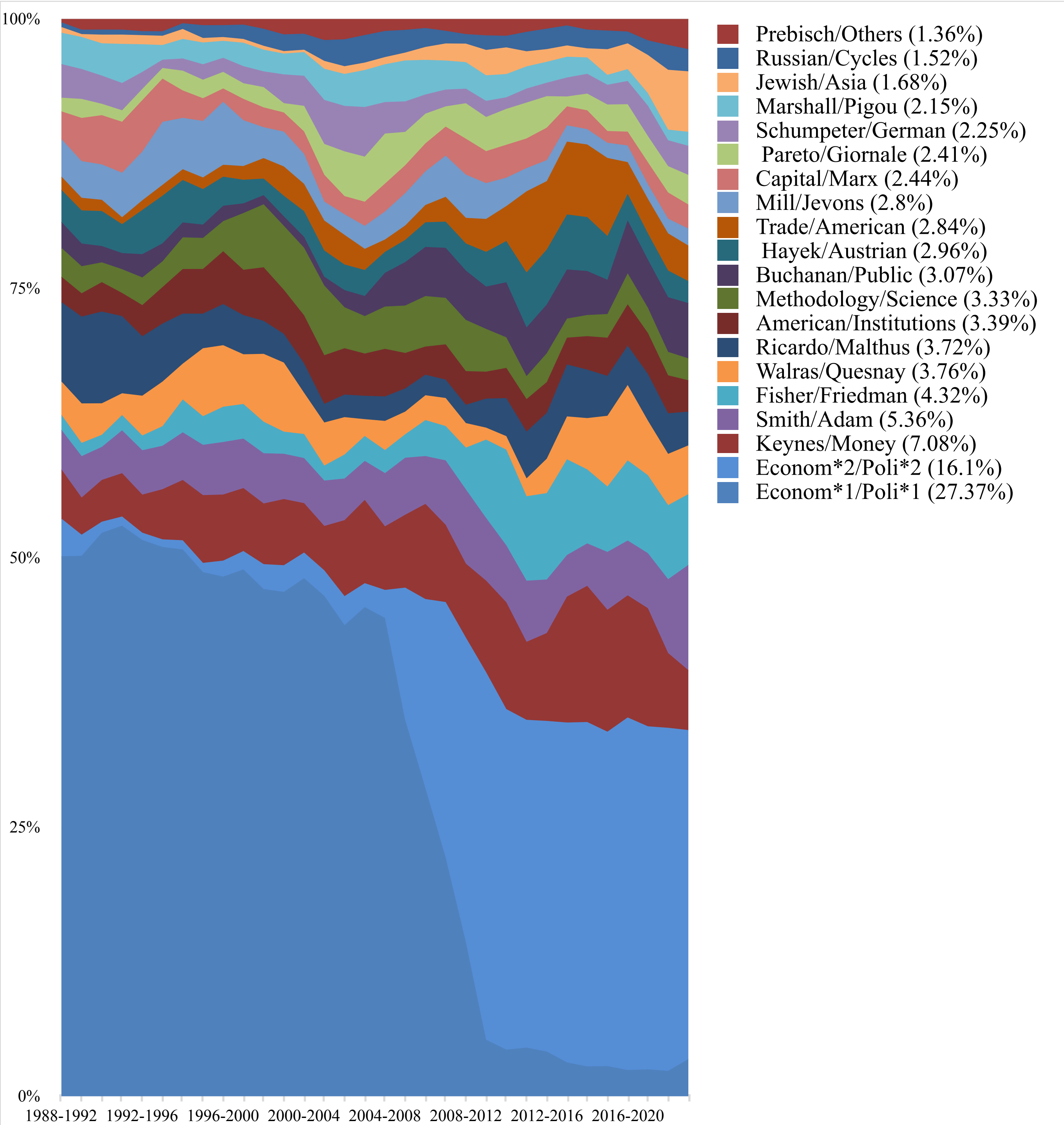
Application

In running our LDA, we used concatenations of the titles, abstracts and reference-lists of the different research articles to create texts. These texts were "cleaned" to remove stopwords, short words, and uninformative terms such as "university," "press," or "doi" from the reference-lists, or "approach," "analysis," or "study" from the abstracts.

An important choice in implementing LDA is the definition of the exogenous number of topics. Our choice of 20 topics was based on standard statistically based metrics, in particular those by Griffiths and Steyvers (2004), Arun et al. (2010), and Deveaud et al. (2014), together with explorative iterations using different topic numbers.[7] As for the iterations, we systematically inspected classifications using 10 to 70 topics, using intervals of five topics each time. At 20 topics, the slope of the likelihood of the corpus proposed by Griffiths and Steyvers (2004) started considerably flattening—i.e., revealing a maximum. At the same time, we observed that numbers greater that 20 produced topics excessively narrow to be useful in this short study. It must be stressed, however, that this iterating process is what allowed us to effectively explore the dataset. In a sense, this process constitutes the essence of our quantitative examination of the *HESB*/*JHET*, together with the previous process of partitioning of the dataset into algorithmically detected communities.

Using Bayesian techniques on posterior sample distribution, Griffiths and Steyvers (2004)'s metric provides an evaluation of the likelihood of the corpus given different numbers of periods. The idea of Arun et al. (2010) is to see the LDA output as splitting the corpus—i.e. the document-word frequency matrix—into a topic-word matrix and a document-topic matrix, which further decomposed should share some observable statistical properties, so that they can be analyzed conditional on the number of topics. Deveaud et al. (2014) proposes a metric seeking to obtain distinct topics.[7] We used the R package LDA tuning for implementation.

FIGURE A4. Evolution of the 20 Topics (5-year moving averages, 1988-2023)



Note: this figure presents the average prevalence of topics within documents by publication year within 5-year moving averages. The average prevalence of each topic appears in brackets.

Figure A4 presents the main outcome of running the LDA, just like Figure 2 in the main text, but this time with the full dataset (i.e., including 2023).[8] It shows the evolving prevalence of the different topics in time, which was obtained by combining the information of topic prevalence in the different texts and computing their average proportions over time, using the information of publication years.

Asides from the first four topics, which remain unchanged, the configuration and ordering of the smaller topics is slightly different here than in our previous model (i.e., Figure 2). Useful to illustrate the workings of our iterating process in adjusting the model, variations in the smaller topics help us further investigating the corpus of research articles in the journal. F. Quesnay and L. Walras, for instance, which appear separate in our first model (main text), are now mixed together into a Walras/Quesnay topic. Similarly, research on finance, which was modeled together with cycles in the first model, is now present within the Fisher/Friedman topic that also includes research on econometrics and recent macroeconomics. All in all, besides the sharp rise of Smith/Adam the last period (i.e., due to the 2023 *JHET* issue including the "Smith at 300" symposium), Figure A4 reproduces the overall picture and argument in our main text: topics on recent economics such as Fisher/Friedman, Buchanan/Public, or Jewish/Asia tend to rise in the last few years, whereas others like Mill/Jevons, Capital/Marx, or Marshall/Pigou (i.e., research on Great Economists other that Smith and Keynes) tend to be prevalent in earlier days of the *HESB*/*JHET*.

IV. REFERENCES

Arun, R., V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy. 2010. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Advances in knowledge discovery and data mining*, Mohammed J. Zaki, Jeffrey Xu Yu, Balaraman Ravindran and Vikram Pudi (eds.). Springer Berlin Heidelberg, 391–402.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. "Latent Dirichlet allocation." *Journal of Machine Learning Research* 3: 993–1022.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10): P10008.

[8] In this case we also display the first two general topics separately, to show their dependence on the indexation the database (i.e., before 2009 the first topic, and from 2009 onwards the second). These two topics contain both words in the reference-lists of the different research articles, and general words, as described in our main text.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. “Indexing by Latent Semantic Analysis.” *Journal of the American Society for Information Science* 41 (6): 391–407.

Deveaud, Romain, Éric SanJuan, and Patrice Bellot. 2014. Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique* 17(1): 61–84.

Edwards, José. 2020. “Fifty Years of HOPE: Changing Priorities in the Historiography of Economics.” *History of Political Economy* 52 (1): 1–46.

Edwards, José, and Felipe Martin. 2019. “The History of Economics Journal Database Complement 1969-2009 User Guide.” *Journal of the History of Economic Thought* 41 (4): 623–29.

Griffiths, Thomas L. and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101, suppl 1: 5228–5235.

Martin, Dian I., and Michael W. Berry. 2007. “Mathematical Foundations behind Latent Semantic Analysis.” In *Handbook of Latent Semantic Analysis*, 35–55. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Schwarz, Carlo. 2018. “ldagibbs: A command for topic modeling in Stata using latent Dirichlet allocation.” *The Stata Journal* 18 (1): 101–117.

Schwarz, Carlo. 2019. “Lsemantica: A Command for Text Similarity Based on Latent Semantic Analysis.” *The Stata Journal* 19 (1): 129–42.