

Distinguishing Discourses: a Data-Driven Analysis of Works and Publishing Networks of the Scottish Enlightenment

Iiro Tiihonen¹, Yann Ryan¹, Lidia Pivovarova¹, Aatu Liimatta¹, Tanja Säily¹ and Mikko Tolonen¹

¹University of Helsinki, Finland. Faculty of Arts

Abstract

A key feature of the Enlightenment is the development of a discourse on commerce and economy entangled in larger discussions around politics, morality, and progress. Importantly, this debate was not formalised in the way it may be seen today. Instead, it was an emerging subject incorporating diverse and contested ideas. The objective of this case study, then, is to use various methods to identify the boundaries of these emerging economic, political, and moral disquisitions in a data-driven way, using a unified version of the metadata (e.g. publisher, publication year, format of the print product) of the English Short Title Catalogue (ESTC) and full texts of the Eighteenth Century Collections Online (ECCO). We approach the task iteratively, first making a separation between broadly defined economic documents and other eighteenth century documents by modeling the features which separate samples from two collections of historical economic texts from the wider ECCO data. Then, based on the previous step, we distinguish works similar to Hume's *Political Discourses* (a text at the heart of the Scottish Enlightenment) from other branches of commercial and economic discourse, and analyse this set of works in more detail. We also experiment on how a purely unsupervised approach – a contextualized topic model using BERT encodings – groups our set of economic texts.

Previous historical scholarship has taken the perspective that we ought to identify language use from large corpora of text. The aim has been to contextually understand language from the perspective of a particular group of historical actors or, as is the case with conceptual historians, detect contested and changing concepts. Our approach is different in that we closely link language use to the material and historical circumstances of the individual texts within which these uses can be found. Essentially, we combine computation and social network analysis with the study of changing concepts and word uses over time, taking individual editions rather than language abstracted from them as the basic object of study. This approach allows us to identify additional contextually relevant works by both their linguistic features, and the material and network history of their production.

Jointly, the combination of iterative data-driven discourse detection and the focus on manifested editions allows us not only to extract a significant proportion of the debates forming the Scottish Enlightenment in a data-driven manner, but to link them to the social networks and commercial context in which they were produced and in which they evolved. Thus, this approach allows us to evaluate the existence, scope, and contexts of historical discourses (in this case, economic discourse) in the eighteenth century in a way which is both computationally state-of-the-art and relevant to historical practices and interests. It also demonstrates how data-driven analysis and the traditional hermeneutic approach can be combined to study meanings and their changes over time.

Keywords

computational history, eighteenth-century studies, economic discourse, social network analysis

1. Introduction

The emergence and development of commercial and economic discourse in the eighteenth century is one of the major strands of study for the intellectual history of the Enlightenment era [1, 2, 3]. This article links to that tradition, aiming to detect and analyse the debates of the Scottish Enlightenment that lie at the intersection of entangled political, economic and moral considerations. Compared to most of the existing scholarship, our approach differs in two crucial and interlinked ways. It is data driven, utilising the most comprehensive data sets of full texts and metadata of early modern British print products in a quantitative manner. Previous historical scholarship has focused either on language and its historical context [4] or the contest over and change of concepts. Here we take a different approach and connect the analysis of language to the way it is manifested, namely the production process of the physical editions from which the text comes. Thus, the novelty of our contribution is in the contextually sensitive approach we take to computational intellectual history rather than the application of the computational methods. We also aim to use methods that are interpretable at the level of editions, for example by analysing their term counts and publishers. We especially focus on the participants of the publishing process of different editions and their collaboration networks – the context in which the Enlightenment discourse was physically produced.


In this article we describe the first test case for this approach. First, we produce a computationally derived set of editions similar to Hume’s *Political Discourses*. In a second step, this set is linked to its producers, allowing us to analyse both the discourse itself and the co-operation networks of publishers in which it was produced. In a third step, we find meaningful groups of economic texts with a state-of-the-art topic model, and these subsets also show interesting variation in their presence in different communities of the book trade. The qualitative interpretation of this set of works and the related publisher networks confirms that – despite need for further improvement – computational detection of Enlightenment discourses and the publishing networks that physically produced them is a realistic goal, something that can be achieved with the methods at hand. Our computationally derived data agrees with the expert evaluation, in that the works most similar to *Political Discourses* represent a coherent set of Scottish Enlightenment texts. Furthermore, the clusters of publisher networks which produced them match the communities and co-operation patterns found with traditional approaches.


The term discourse is in itself ambiguous,¹ and instead of fixing its meaning or resolution for the entire article, we start from a very broad category of economic texts and iterate onward to see at which resolution and level of nuance we are currently able to detect and cluster texts that are similar, and whether this similarity is meaningful from the point of view of historians.

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022.

✉ iiro.tiihonen@helsinki.fi (I. Tiihonen); yann.ryan@helsinki.fi (Y. Ryan); lidia.pivovarova@helsinki.fi (L. Pivovarova); aatu.liimatta@helsinki.fi (A. Liimatta); tanja.saily@helsinki.fi (T. Säily); mikko.tolonen@helsinki.fi (M. Tolonen)

🆔 0000-0003-0703-4556 (I. Tiihonen); 0000-0003-1878-4838 (Y. Ryan); 0000-0002-0026-9902 (L. Pivovarova); 0000-0001-9056-1087 (A. Liimatta); 0000-0003-4407-8929 (T. Säily); 0000-0003-2892-8911 (M. Tolonen)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹ Our use of it should not be confused with e.g. historical discourse analysis as it is defined in [5].

The article considers the extent to which definitions of discourse are currently functional for computational intellectual history, and what the level of nuance could be in the future. The aim is that this practical work will also in time make it possible for us to develop a more coherent theoretical position that can be directly compared e.g. to Quentin Skinner’s contextualism [6] or other relevant frameworks of traditional intellectual history. The step of model fitting and evaluation at the beginning is to see to which extent commercial and economic texts can be differentiated from all other types of textual content as a discrete category in a data-driven way, based on existing collections of economic texts partially linked to the ESTC. The motivation for this is twofold: Computational differentiation of an economic corpus from the rest allows us to expand the set of economic texts. It also produces a more interpretable set of terms related to economic matters, and by exploring the variation of this terminology within our set of economic texts, we can try to cluster this data further, which would allow us to speak of discourses at a more precise level. In this article, we attempt to identify works similar to Hume’s *Political Discourses* from the wider set of economic texts based on similarity of term frequencies, and reflect on how this set of works relates to the traditional understanding of which works form part of the Scottish Enlightenment. The reason for picking Hume’s *Political Discourses* as our test case was that it is a work that is specifically about commerce and economy in its widest sense instead of being a technical work focused on trade only.

2. Data

Our data set is derived from the English Short Title Catalogue (ESTC) and the Eighteenth Century Collections Online (ECCO) [7]. The ESTC is the largest catalogue of early modern print products of the Anglosphere, and its machine readable form – enriched, harmonised and refined [8] by the Helsinki Computational History Group – is our main source of information for metadata. This metadata includes information about the place and time of publication, actors related to the print products, physical characteristics and – crucially – references to other collections in which the ESTC record or its copy is part of or referenced. ECCO contains OCR full texts for roughly 200,000 documents, and all of these can be linked to corresponding ESTC records.² Our group has developed an API called Octavo for more convenient information retrieval from ECCO, allowing e.g. the retrieval of a filtered set of terms for each document instead of the full texts that are often heavily affected by OCR noise.

By combining the text data from ECCO with metadata from the ESTC, we created a starting point data set for the analysis of economic discourse. We can form a subset – based on information in the ESTC fields 510a and 533f – of about 11,000 ESTC records belonging to the Goldsmiths’-Kress Library of Economic Literature (GKL), a microfilm collection that covers documents related to economic history in the widest sense, from the primary works of classical economic theory to political pamphleteering and agriculture [9]. Additionally, we know of 4,000 ESTC records that are cited in the Contemporary Printed Sources for British and Irish Economic History 1701–1750 (CSEH) catalogue, which also covers a broad range of topics related to early modern economy. The two collections cover the same eight topics in their classification

² However, it is important to note that this mapping is not one to one, as an ESTC record can e.g. cover multiple volumes that are separate documents in ECCO.

schemes, although GKL adds five additional categories. We then linked these ESTC records to ECCO, which resulted in roughly 8,000 documents from the collections with full text available (this sample is henceforth referred as GKL-CSEH). Having obtained our economic texts, we took a sample of 16,000 ECCO documents that did not link to the GKL-CSEH. In the following steps, our default assumption for these documents was that they were not about economic discourse. The motivation for the random sample was that it could help us detect those linguistic features that differentiate economic discourses from other eighteenth century texts.

Instead of using ECCO texts in their entirety, we aimed for information that would summarize the general trends of term use in the data while not being inflated by OCR noise and the most common words. To balance representativeness and the need to filter, we queried the Octavo API for a vector (e.g. list) of the 5,000 most frequent tokens plus their term frequencies (the number of times they were used in a given document) that were globally (as in the entire ECCO data) found in at least 100 and at most 50,000 documents. This data set of 58 million rows was further filtered with a number of steps to make it easier (both in terms of computation and interpretation) to analyse: tokens with non-standard characters were removed, as were tokens that occurred in less than 200 of the GKL-CSEH documents. Sample documents with less than 250 terms were also removed. The relative presence of a token in a document was normalised by the sum of tokens in the top 5,000 vector of the document (in most cases this is the same as normalising with filtered token count) and this relative frequency of each token per document was scaled with its own standard deviation approximated from the training and test data. The end result of this process was a matrix of roughly 22,000 observations (documents) and 13,000 features (scaled relative frequencies of tokens).

3. Derivation of Economic Vocabulary

Our aim was to use this matrix to detect terms relevant for discourse on commerce, which we did by using it as data for a classifier. We fitted a L_1 penalised logistic regression model with cross validation 20 times, using the features of the matrix as predictors for the class (whether it was about an economy-related topic or not) of the document. The random sample was treated as if it consisted entirely of non-economic texts. Penalised regression is a relatively standard and well-performing approach to text classification with large numbers of features [10]. The specific choices were motivated by the fact that a L_1 penalised model tends to drop features as predictors entirely and with cross validation the parametrisation of the model would be more robust. However, single runs of the cross validated models seemed to have enough variance that further robustness was needed, and this was obtained by running the model 20 times and only keeping those tokens that were non-zero predictors with each run. For these always non-zero parameters, the final parametrisation was the average over these 20 runs. Qualitative evaluation of the 1,060 terms with a non-zero effect size served as a sanity check for our approach, as most of the very impactful words with a positive sign were clearly related to economic discourse. The very top of the predictors (in terms of effect size) is presented in Table 1.

The model was validated with test data of 3,692 observations that were left out of the model fitting steps. In evaluating the model's success in finding "new" instances of economic writings, we did not accept documents that could have been only accepted to the GKL-CSEH categories

Table 1

Top 20 predictors of the model by effect size. The word goldsmiths is sometimes stamped to the beginning of the ECCO copy of a GKL document, which most likely explains the significant effect size of the term. The table also highlights the high amount OCR errors and noise in the data, which motivated the use of the L_1 penalty.

term	effect size	term	effect size
goldsmiths	0.25	interef	0.081
ooo	0.14	ships	0.081
prices	0.14	wrhich	0.077
merchants	0.14	ico	0.071
trading	0.13	traders	0.07
stock	0.12	draining	0.067
woollen	0.087	enrich	0.067
molk	0.087	pound	0.066
pays	0.085	contributed	0.063
individual	0.083	necefary	0.062

Table 2

Confusion matrix of the test data. Columns indicate the real and rows the predicted category.

	Positive	Negative
Positive	850	175
Negative	354	2313

of politics or miscellaneous (e.g. theological works were by default not accepted even though miscellaneous includes some theology). The model performed quite well on this test data, as Table 2 demonstrates. Prior to hand checking of the data, it had a 71 percent true positive and 93 percent true negative rate (Table 2). As a further step of model validation, we looked at the documents labeled as false positives or true negatives, as it was relevant for further steps to understand whether the model had misclassified documents or detected new instances of economic discourse, and to which extent it detects (and GKL-CSEH covers) ECCO's economic documents. All of the false positives were checked, and of these 167 (95 percent) contained at least some content that could justify them belonging to a collections such as GKL-CSEH. A random sample of 100 of the true negatives was checked, and out of these 29 contained at least some content that could justify them belonging to a collection such as GKL-CSEH. The model does not detect all economic documents, but it makes very few false positive mistakes. We concluded that it works well enough to provide material and structure for the historical and linguistic research it serves, but as it fails to detect much of what could be put under its topics, the suitability of its elements (predictions and the obtained list of term features related to economic topics) for analysing any specific discourse must be evaluated separately.

Next we applied the model for two purposes: the extrapolation of our data set of discourse on commerce was the first step, and it resulted in 15,488 new documents classified as being about economic matters. We further extended the data set by including other editions of extrapolated and original GKL-CSEH documents in our data set, which resulted in 33,153 documents, 32,895

having corresponding term frequencies to be used in subsequent analyses.³ The second step was to see whether we could cluster this extended set of economic documents further in a way that allowed us to track specific discourses, which we will discuss in the following section.

4. The Discourse of *Political Discourses*

Hume's *Political Discourses* is a work that sits at the centre of the Scottish Enlightenment and its debates on commerce [11]. It hits the nerve on all the main topics of civil society, greatly paving the way for Adam Smith and Adam Ferguson. Hume's *Political Discourses* is a death punch to mercantilism. It changes the nature of the luxury debate. It penetrates deep into continental discourses with its treatment of rich country, poor country (it was also translated twice into French within a short period of time after its initial publication). And, perhaps most importantly, it engages in a manner that could not be avoided by any eighteenth-century political thinker with the ancient and modern debate and the discussion of population growth in tandem with Robert Wallace. Without exaggeration we may note that the relevance of *Political Discourses* as an individual work for the rise of commercial society is not surpassed even by Montesquieu in the eighteenth century. We may also note that Hume's contemporary reputation and canonization in the eighteenth century relies on *Political Discourses* and not his philosophical works or other essays. These qualities make it an ideal starting point for clustering works of commercial Enlightenment discourse, the next step in our process.

The comparison of the economic documents to *Political Discourses* was done by comparing the distribution of the economic terms (the selected features from the model-fitting step with a positive effect) in it to the corresponding distributions of all the other economic documents, measured using the Jensen-Shannon Divergence.⁴ The motivation of the idea shares some resemblance with the theoretical assumption in topic modeling that texts belonging to the same topic come from the same distribution of term frequencies, but instead of an unsupervised approach, we cluster based on a pre-selected work and consider only the frequencies of a heavily filtered set of terms. The approach was evaluated qualitatively by looking at how the similarity to Hume's *Political Discourses* from a historian's point of view varied as the function of the similarity score. This process was also used to determine the threshold scores for Hume-likeness used to form a subset of the data for further linguistic and network analysis of works that were deemed similar.

Our method for identifying textual similarity is based on word frequencies, hence we also tested the reuses of Hume's *Political Discourses* (e.g. textual overlaps between all the works measured against the work). Text reuses can be full or partial reprints of Hume's essays or direct quotes. The assumption was that if there are a lot of direct quotes or reprints of individual essays by Hume included in the works examined, then obviously this would be at least a partial explanation for a high similarity score between them and Hume's and hence an excellent sanity check for our approach. This indeed turned out to be the case, but importantly it is not the full explanation for similarity. Hume's essay collections that include *Political Discourses* are

³ 42 document term vectors were lost to problems in querying Octavo, the remaining losses did not have full text in ECCO.

⁴ Small smoothing was applied to distributions to be able to compute the metric.

Table 3

Works (with known authors) with the least Jensen-Shannon divergence from Hume's *Political Discourses*. The year given is the year of publication of the first edition, not necessarily the edition with the lowest divergence.

Author	Short Title	First published
Priestley, Joseph	Lectures on history, and general policy	1788
Kames, Henry Home, Lord	Sketches of the history of Man	1774
Steuart, James, Sir	An inquiry into the principles of political oeconomy	1767
Raynal, Abbé	A philosophical and political history of the settlements and trade of the Europeans in the East and West Indies	1775
Mortimer, Thomas	The elements of commerce	1772
Wallace, Robert	Characteristics of the present political state of Great Britain	1758
Michell, Charles	Principles of legislation	1796
Tucker, Josiah	A brief essay on the advantages & disadvantages which respectively attend France and Great-Britain	1749
Williams, John	The rise, progress, and present state of the northern governments	1777
Adams, John	Curious thoughts on the history of man	1789

obviously the ones that score highest evaluated by both of our methods (also left out of Table 3 as trivial). Other large segments of text reuse compared to works included in Table 3 include Priestley's *Lectures* and Thomas Mortimer's *Elements of Commerce*. It should be also noted that Robert Wallace's *Characteristics* is written with Hume as its interlocutor and it includes plenty of quotes from *Political Discourses*. At the same time, the list of works closest to *Political Discourses* based on our method contains several titles that include only few (or no) direct quotes from Hume. Overall, if we study the top-list of highest similarity score to Hume with a qualitative eye (Table 3), we may note that it is a very interesting list of works that in different ways can be said to be similar to *Political Discourses*.

5. Publishing Networks of the Scottish Enlightenment

Another way to study these texts is to look at the networks of actors – printers, booksellers, publishers and so forth – which produced them. Early modern texts were complex commodities and their networks should be taken into account alongside their content. Previous work by the project has shown the validity and potential of historical network analysis using ESTC data [12]. To understand more about the networks which produced the works we deemed relevant using the above methods, we constructed a bipartite graph using the co-occurrences of book trade actors in imprints of all works which were deemed sufficiently close to Hume's *Political Discourses* (a Jensen-Shannon divergence of below .42). This section defines and

describes this graph and its subsequent analysis.

The term network is often used informally by book historians – books in the seventeenth and eighteenth century were ‘networked technologies’ [13], in that they were produced collaboratively and those involved naturally formed networks, had overlapping alliances and fostered both local and international links. Studying these networks – at least qualitatively – has long been a key mode of study of the history of the book. But the term ‘network’ can also be formally defined using network science. In this field, a network is defined as a mathematical graph of entities, known as nodes, and their connections, known as edges. To create a network of book trade actors, we first drew a link between each work and the actors listed in their imprints. Next, a *co-occurrence network* was created by directly connecting actors based on co-occurrences on these imprints.

This approach has been utilised in computational book history elsewhere, for example to analyse the works of John Milton from a collaborative, materialist perspective, to situate authors’ works within a larger marketplace of print, and to look for shifts in the practice of book dedications [14, 15, 16]. This approach typically does not attempt to re-create the complete social network of publishers and their relationships: publishers are sometimes listed on the same imprint without having collaborated in any meaningful way, and many publishers will have known each other but not collaborated, meaning their connections will not show up. Rather, it attempts to represent a specific type of network, one based solely on co-production within texts. By combining a similar approach to constructing networks along with a threshold based on the similarity scores described above, we have been able to take into account both the books’ content and the material circumstances surrounding their production, and understand the data from the dual perspectives of both metadata and full text.

Once the connections between the actors have been extracted in this way, the resulting mathematical graph can be analysed using a range of standard metrics, such as counting a node’s connections or the paths in the network. Another common form of analysis is to detect clusters – or communities – within the network. In network science, a community is usually considered to be a group of nodes which are more densely connected to each other than they are to nodes outside that community [17]. Detecting communities in an actor network of this kind might help to understand the various and overlapping sub-groups involved in the production of a group of closely-related texts. To form these communities we used one of the most widely-used algorithms, known as Louvain [18]. This uses an iterative process to maximise a global network metric – *modularity* – which measures the number of links between nodes within a given set of community labels, and compares it to the same measurement in a random graph of the same size and distribution [19].

To find particularly relevant communities, we retained all pairs of actors who occurred together on at least two separate imprints. The resulting network of 510 nodes and 2,740 edges is divided into 35 separate communities by the algorithm. Not all are significant: only seven of these contain more than ten nodes – many smaller communities are formed by groups of actors who might work on only one text and are therefore disconnected from the wider network. Looking at the works worked on by actors in each of the ten largest communities, we see that they are often geographically or temporally distinct, and sometimes both. Charting the editions in each by year (Figure 1) shows that communities 1 and 2, for example, are large, interconnected, mostly consisting of actors based in London, but clearly temporal. Works

produced by community 2 are gradually superseded by works produced by communities 3 and then 1. Community 4 is almost exclusively Dublin publishers. Works from Scotland are mostly found either in community 5 or community 8 – the former also has significant numbers of works published in London.

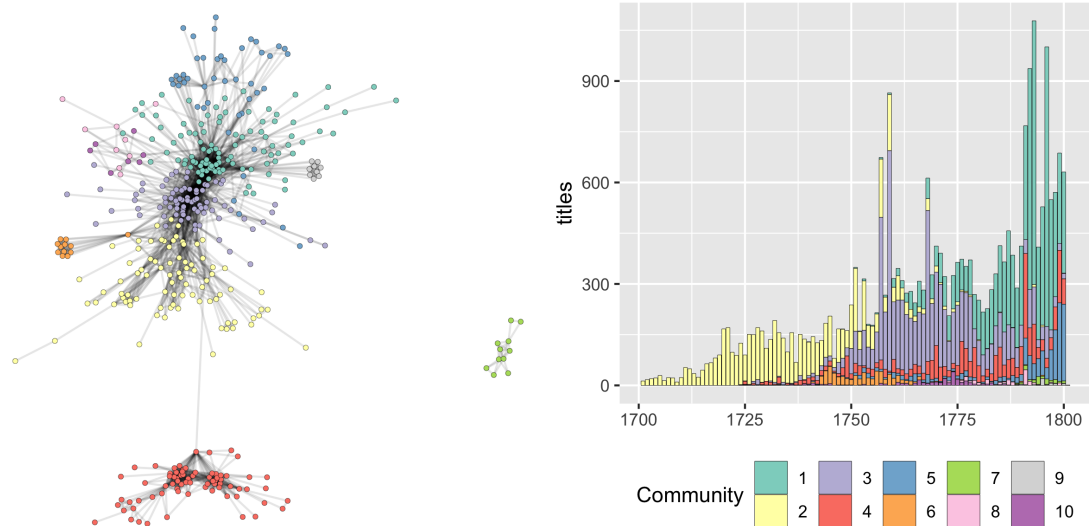


Figure 1: Force-directed diagram showing Louvain community groupings, as well as a plot of works by authors in each community by publication year. For the full network analysis, edges (links) are drawn between pairs of publishers, booksellers, and printers who co-occur on at least two imprints in the ESTC data. In this network diagram, only the strongest links are visualised, for legibility.

These communities therefore allow us to understand more about the distinct but overlapping series of networks responsible for the production and eventual diffusion of works similar to the Hume edition. Drawing the network as a force-directed network graph also helps to grasp its topographical structures (Figure 1). This shows, for instance, a group of American publishers and booksellers (community 7) entirely disconnected from the wider network: unsurprisingly, those separated by the Atlantic were unlikely to directly collaborate with each other. The network also shows the impact and patterns of piracy: there is a mostly disconnected community of Dublin actors, who, in producing pirated editions, naturally did not tend to work with ‘legitimate’ London publishers. However, this network does have some links, and looking at these can point to crucial figures who straddled both established and illegitimate print enterprises – or perhaps moved from one to another. Three figures in particular stand out: Luke White, an ‘obscure bookseller’ based in Dublin who became a wealthy Member of Parliament, winning Government contracts [20], John Archer, often listed as the Dublin agent for London-produced books [21], and the bookseller Hulton Bradley. In the network diagram these can be seen as some of the few figures with connections both to this ‘Dublin’ community as well as to the communities consisting primarily of large London publishers.

Network analysis is also useful in seeing the changing patterns of production of various

editions of particular works, allowing us to trace their evolution from one set of actors to another, and even highlight some of the key figures who facilitated this change. To demonstrate this, we outline the evolution and diffusion of Hume’s *Essays and Treatises on Several Subjects*, a collected volume of various shorter works by Hume (that includes also his *Political Discourses*), most of which had been published initially by the Edinburgh bookseller Alexander Kincaid and his London partner, Andrew Millar. The network and circumstances surrounding the publication of this work forms the basis of a chapter on the Scottish Enlightenment by Richard Sher, and serves as a useful test case in comparing the data-driven results with what is known from the historical record [22]. Tracing the actors involved in the various editions of *Essays and Treatises* illustrates how the text moved through various and overlapping networks of publishers. Imprints of the first editions – in 1753 – show the involvement of Kincaid and his partner Alexander Donaldson, both found in community 5, which consists primarily of actors whose works are printed in Edinburgh, whereas the London edition of the same year lists Kincaid, Donaldson, as well as Millar. The same three actors are attached to most subsequent editions until 1768 where Millar is replaced by Thomas Cadell, to whom he had resigned his business (Millar died in June 1768) [23]. The edition of 1777 lists William Creech, found in community 5, a community notable for containing works almost evenly spread across Edinburgh and London. Subsequent editions also have actors from a London-Edinburgh community: the bookseller Thomas Kay, based in London, and the Edinburgh bookseller Charles Elliot, who opened a shop in London in 1786, and the final editions of the century were worked on by actors found in a mostly London community, community 1. Cases like this have been found by looking for patterns in works deemed interesting from domain expertise, but a future approach will be to look for individuals and works which move from one community to another in a more systematic manner.

Community detection, then, is a useful tool to further our understanding of the various overlapping groups of actors who worked in the book trade in London and Edinburgh. In this case it has helped to see how a work such as *Essays and Treatises* diffused or circulated from one set of closely-connected actors to another, and the intermediate steps it took along the way. Alongside this should be noted the drawbacks of this approach: the sizes of the communities are somewhat arbitrary, and the results give no indication as to whether a given node is well embedded in a particular community, or if it is an edge case sitting between multiple cases. Also worth noting is that the network connections do not capture the extent or quality of any given relationship – just that each pair co-occurred on an imprint together. Despite this, it is clearly a useful way to understand more about the various networks involved in the production of texts within a particular discourse.

6. Categories of Commerce

Following previous work that uses topic modelling to track discourses in historical collections [24, 25, 26], we apply topic modelling to further investigate communities and try to find out whether they are different in terms of the content they are publishing. We trained a Contextualized Topic Model (CTM [27]), which is a neural network that takes as input a contextualized sentence representation (embedding) and predicts a topic distribution. The model is trained using text reconstruction: from the topic distribution it should predict a bag-of-words text

Table 4

Example topics detected by the CTM model trained on the collection of commercial documents.

#	top 10 words in the topic	interpretation
0	men people man army officers persons body troops enemy poor	WAR
1	geo fol car pa andl tha thi ta mi die	-
2	man thing men things people manner number nature place persons	PEOPLE
3	war army french general troops france peace enemy men treaty	HISTORIES
4	miles called river north town south sea fide place built	GEOGRAPHY
5	great britain country large good greater parts england trade land	POLITICAL COMMERCE
6	sea north water miles feet fide river south half called	GEOGRAPHY
7	goods pay paid persons person duties duty hall money ship	SHIPPING
8	england king kingdom france majesty crown parliament queen prince britain	CROWN
9	king year reign time years fame parliament majesty late henry	HISTORY
10	city town county place london country miles places towns church	LONDON / TRAVEL
15	money pounds paid price pay pence gold cent sum pound	MONEY
25	cafe lands law estate plaintiff goods defendant party debt cafes	LAW
33	plants leaves flowers plant flower trees ground fruit grow roots	NATURAL HISTORY / AGRICULTURE
47	people nation war peace government church country power public state	NATION

representation. As a result, each topic is associated with word probabilities, thus resulting in a final topic model similar to standard LDA output, even though it is trained using different principles. A bag-of-words representation is needed only during training, inference is done using a contextualized representation. Thus, the model is able to assign topic probabilities even to sentences that consist of words not seen during training.

To make contextualized text representations we used the ECCO BERT model, which has been made publicly available and is described in a separate publication [28]. Since BERT takes as an input 512 input tokens at most, we use paragraphs as a unit of analysis. Paragraphs are run through the BERT model and the resulting token embeddings are averaged to obtain paragraph embeddings. We use 500,000 random paragraphs from the extrapolated data set of economic documents presented in Section 3 and train a model with 50 topics.

Several topics are exemplified in Table 4. Since all documents in the collection are already relevant to economic discourse (in a broad sense), the resulting topics describe different themes related to this major discourse, such as war (topic 0), shipping (topic 7) or law (topic 25). As is usually the case with topic modelling, a few topics are meaningless, which is exacerbated by OCR errors in historical documents. We try to minimize the problem by introducing a collection-specific list of stopwords, which include the most common non-words produced by OCR – e.g. ‘tbe’, ‘thc’, ‘thle’, which are all misspelled variations of a standard stopword ‘the’. Still, two of the fifty topics consist of meaningless words, e.g. topic 1 in the table. In addition, topics are overlapping, as 0 and 3 shown in the table. In training 50 topics we follow the same logic as [26]: it is always an option to merge topics in further analysis while splitting them is impossible.

After inferring a topic distribution for each paragraph we average them across documents to get a document-topic distribution. Two meaningless topics – 1 and 24 – are the most prominent for approximately 37% of documents. Thus, we use the top 3 topics for each document for further analysis. Then we manually check how documents group together according to their topics. Note that the most probable words, as presented in Table 4, may be not the best representation for human topic interpretation [29]. For our study of changing discourses the most crucial

question is whether topics group together documents in a meaningful way. Manual inspection shows that this is the case with most of the topics. For example, topic 25 is the most prominent in legal reports and groups together law documents, which is not immediately clear from the word list, especially due to the fact that the most salient word in this topic – ‘case’ – is misspelled as ‘cafe’ (mixing long *s* with *f* is one of the most frequent OCR bugs). In Table 4 we add a human interpretation assigned after inspection of documents where the topic was among three most prominent. After the manual inspection we conclude that topics group documents in a meaningful manner and could be used to further explore network communities as detected in the previous step (Section 5).

Looking at the most prominent topics in each work, we can see that they successfully differentiate between different groups of texts. Topic 47, for example, is most prominent for Hume’s *Political Discourses*, as well as *Essays and Treatises*, which reprints it. Other works with this as their most prominent topic have in common the study of history, morality, politics, and the economy in a broad sense, including Lord Henry Home Kames’s *Sketches of the History of Man*, and Joseph Priestly’s *Lectures on history, and general policy*, as well as large numbers of works by Edmund Burke and many of Daniel Defoe’s political essays and pamphlets. Topic 15, on the other hand, can be linked to technical works or those with a narrower focus on trade and the economy. Exemplary texts include Adam Smith’s *An inquiry into the nature and causes of the wealth of nations*, James Steuart’s *An inquiry into the principles of political oeconomy*, and various tracts and texts written on specific economic theory or principles of trade and money. In this way, the topics can be used to get a more fine-grained division between works already identified as part of a more general corpus of economic texts.

The topics also make sense with respect to the way they are distributed across the communities of publisher networks extracted using network analysis techniques. There are significant differences between communities in terms of the distribution of topics within them (Figure 2). In communities 1, 4, 5, and 10, topic 47 is the most prominent: as suggested above, those where this topic is most prominent are texts which deal with the economy in a broad sense. Communities 2 and 3 have a very different profile: in both, topic 15 stands out, one which has a narrower and more technical focus. This suggests (though at this early stage we cannot fully assert this claim) that within those working on economic texts and discourse, there are different – though not distinct – publishing communities, and that these can be found with this approach.

7. Conclusions

The aim we set for ourselves for this article was to develop approaches for analysing early modern discourses on commerce in a data-driven manner, and to link these discourses to the wider context of the physically manifested texts through the publisher networks in which they were produced. We achieved this aim, as we were able to cluster and analyse the relevant context of Hume’s *Political Discourses* – both in terms of works and publisher communities – in a data-driven manner, and the results align with traditional humanities expertise both at a macro and micro level. Similarly, many of the topics produced by the CTM model were meaningful from a historian’s point of view, and were not equally distributed across publishing communities, but varied in their prominence. The main implication of this success is that the

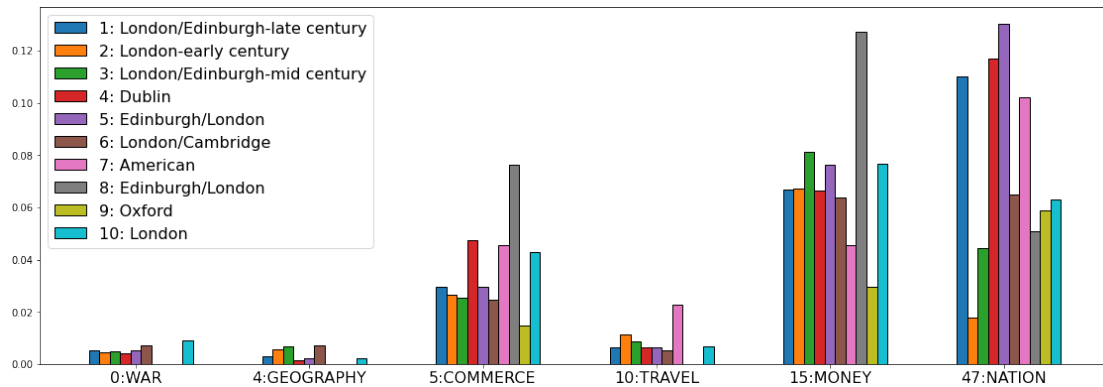


Figure 2: Distribution of selected topics across the publisher communities. We show a relative number of documents within a community where a given topic was among three most prominent.

development of the process and tools applied in this test case and their application for detecting and analysing other historical discourses have the potential to contribute significantly to the historical understanding of the Enlightenment era. It can also be further expanded, as we utilised only some of the metadata related to the materiality aspect of the texts. In the near future we will deepen the analysis of publisher communities and text similarities by considering the variation in the format (closely connected to price and status of the print product) of different editions of the same work, which allows an analysis of the intended audience of a work during its history or the strategies of different clusters of the publishers. The same iterative approach that was applied within this article also applies to the larger aim of developing state-of-the-art ways to approach complex questions such as the detection of soft-edged discourses, and we deem the work presented in this article as a successful first iteration.

Acknowledgments

We thank the Academy of Finland for funding the work on this article (Rise of commercial society and eighteenth-century publishing, grant numbers 333716 and 333717) and Helsinki Computational History Group members for discussions and previous work that made this article possible. We especially thank Eetu Mäkelä for the Octavo API and Ville Vaara for the harmonisation and enrichment of the publisher data. We also thank Richard Sher and Mark Hill for discussions that facilitated the work.

References

- [1] J. Robertson, *The Case for The Enlightenment: Scotland and Naples 1680–1760*, Cambridge University Press, 2005. doi:10.1017/CBO9780511490705.
- [2] I. Hont, *Jealousy of Trade. International Competition and the Nation-state in Historical Perspective*, Harvard University Press, 2005.

- [3] M. Tolonen, *Mandeville and Hume: Anatomists of civil society*, SVEC, University of Oxford, Voltaire Foundation, United Kingdom, 2013.
- [4] P. de Bolla, E. Jones, P. Nulty, G. Recchia, J. Regan, The idea of liberty, 1600–1800: A distributional concept analysis, *Journal of the History of Ideas* 81 (2020), 381–406. doi:10.1353/jhi.2020.0023.
- [5] L. Given, *The Sage encyclopedia of qualitative research methods*, 2008. URL: <https://methods.sagepub.com/reference/sage-encyc-qualitative-research-methods>. doi:10.4135/9781412963909.
- [6] Q. Skinner, Meaning and understanding in the history of ideas, *History and Theory* 8 (1969), 3–53. URL: <http://www.jstor.org/stable/2504188>.
- [7] M. Tolonen, E. Mäkelä, A. Ijaz, L. Lahti, Corpus linguistics and Eighteenth Century Collections Online (ECCO), *Research in Corpus Linguistics* 9 (2021), 19–34. URL: <https://ricl.aelinco.es/index.php/ricl/article/view/161>. doi:10.32714/ricl.09.01.03.
- [8] L. Lahti, J. Marjanen, H. Roivainen, M. Tolonen, Bibliographic data science and the history of the book (c. 1500–1800), *Cataloging & Classification Quarterly* 57 (2019), 5–23. URL: <https://doi.org/10.1080/01639374.2018.1543747>. doi:10.1080/01639374.2018.1543747. arXiv:<https://doi.org/10.1080/01639374.2018.1543747>.
- [9] D. Whitten, Democracy returns to the library: The Goldsmiths’-Kress library of economic literature, *Journal of Economic Literature* 16 (1978), 1004–1006. URL: <http://www.jstor.org/stable/2723473>.
- [10] M. Gentzkow, B. Kelly, M. Taddy, Text as data, *Journal of Economic Literature* 57 (2019), 535–74. URL: <https://www.aeaweb.org/articles?id=10.1257/jel.20181020>. doi:10.1257/jel.20181020.
- [11] M. Tolonen, The Scottish Enlightenment, in: G. Clayes, M. S. Cummings, L. T. Sargent (Eds.), *The Encyclopedia of Modern Political Thought (Volume 2)*, Sage, 2013, 740–745.
- [12] M. J. Hill, V. Vaara, T. Säily, L. Lahti, M. Tolonen, Reconstructing intellectual networks: From the ESTC’s bibliographic metadata to historical material, in C. Navarretta, M. Agirrezabal, B. Maegaard (Eds.), *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, volume 2364 of *CEUR Workshop Proceedings*, CEUR, Copenhagen, Denmark, 2019, 201–219. URL: http://ceur-ws.org/Vol-2364/#19_paper.
- [13] B. Greteman, Making connections with Milton’s *Epitaphium Damonis*, in *Making Milton*, Oxford University Press, 2021, 31–41. URL: <https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198821892.001.0001/oso-9780198821892-chapter-3>. doi:10.1093/oso/9780198821892.003.0003.
- [14] M. Gavin, Historical text networks: The sociology of early english criticism, *Eighteenth-Century Studies* 50 (2016), 53–80. URL: <https://muse.jhu.edu/article/634558>. doi:10.1353/ecs.2016.0041.
- [15] B. Greteman, Milton and the early modern social network: The case of the *Epitaphium Damonis*, *Milton Quarterly* 49 (2015), 79–95. URL: <https://www.jstor.org/stable/26603192>.
- [16] J. R. Ladd, Imaginative networks: Tracing connections among early modern book dedications, *Journal of Cultural Analytics* (2021). URL: <https://culturalanalytics.org/article/21993-imaginative-networks-tracing-connections-among-early-modern-book-dedications>. doi:10.22148/001c.21993.

- [17] M. Girvan, M. E. J. Newman, Community structure in social and biological networks, *Proceedings of the National Academy of Sciences* 99 (2002), 7821–7826. URL: <http://www.pnas.org/cgi/doi/10.1073/pnas.122653799>. doi:10.1073/pnas.122653799.
- [18] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment* 2008 (2008), P10008. URL: <https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008>. doi:10.1088/1742-5468/2008/10/P10008.
- [19] M. E. J. Newman, Modularity and community structure in networks, *Proceedings of the National Academy of Sciences* 103 (2006), 8577–8582. URL: <https://www.pnas.org/content/103/23/8577>. doi:10.1073/pnas.0601602103.
- [20] WHITE, Luke (c.1750-1824), of Woodlands, (formerly Luttrellstown), co. Dublin and Porters, Shenley, Herts. | *History of Parliament Online*, 2022. URL: <http://www.historyofparliamentonline.org/volume/1820-1832/member/white-luke-1750-1824>.
- [21] M. Kennedy, The domestic and international trade of an eighteenth-century Dublin bookseller: John Archer (1782-1810), *Dublin Historical Record* 49 (1996), 94–105. URL: <https://www.jstor.org/stable/30101144>.
- [22] R. B. Sher, *The Enlightenment & the book: Scottish authors & their publishers in eighteenth-century Britain, Ireland, & America*, University of Chicago Press, 2006.
- [23] H. Amory, Millar, Andrew (1705–1768), bookseller, in: *Oxford Dictionary of National Biography*, volume 1, Oxford University Press, 2004. URL: <http://www.oxforddnb.com/view/10.1093/ref:odnb/9780198614128.001.0001/odnb-9780198614128-e-18714>. doi:10.1093/ref:odnb/18714.
- [24] L. Viola, J. Verheul, Mining ethnicity: Discourse-driven topic modelling of immigrant discourses in the USA, 1898–1920, *Digital Scholarship in the Humanities* (2019).
- [25] E. Bunout, Grasping the anti-modern discourse on Europe in the digitised press or can text mining help identify an ambiguous discourse? (2020).
- [26] J. Marjanen, E. Zosa, S. Hengchen, L. Pivovarova, M. Tolonen, Topic modelling discourse dynamics in historical newspapers, in: *Digital Humanities in the Nordic Countries Conference*, Schloss Dagstuhl Leibniz Center for Informatics, 2021, 63–77.
- [27] F. Bianchi, S. Terragni, D. Hovy, Pre-training is a hot topic: Contextualized document embeddings improve topic coherence, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021, 759–766.
- [28] I. Rastas, Y. Ciarán Ryan, I. Tiihonen, M. Qaraei, L. Repo, R. Babbar, E. Mäkelä, M. Tolonen, F. Ginter, Explainable publication year prediction of eighteenth century texts with the BERT model, in: *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, Association for Computational Linguistics, 2022, 68–77. URL: <https://aclanthology.org/2022.lchange-1.7>.
- [29] J. H. Lau, D. Newman, T. Baldwin, Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality, in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, 530–539.