The Journal of
HISTORICAL
NETWORK
RESEARCH

7 | 2022

C²DH
LUXEMBOURG CENTRE FOR
CONTEMPORARY AND DIGITAL HISTORY

UNIVERSITY OF
LUXEMBOURG

Imprint

Université du Luxembourg 2022

Luxembourg Centre for Contemporary and Digital History (C²DH)
Université du Luxembourg
Belval Campus
Maison des Sciences Humaines
11, Porte des Sciences
L-4366 Esch-sur-Alzette

Editors
Asst. Prof. Dr. Marten Düring (Luxembourg Centre for Contemporary and Digital History | C²DH)
apl. Prof. Dr. Robert Gramsch-Stehfest (Friedrich-Schiller-Universität Jena)
Dr. Christian Rollinger (Universität Trier)
Dr. Martin Stark (ILS – Institut für Landes- und Stadtentwicklungsforschung, Dortmund)
Clemens Beck, M. A. (Friedrich-Schiller-Universität Jena)

ISSN 2535-8863

Contact
Principal Contact
JHNR-editors@historicalnetworkresearch.org
Support Contact
Dr. Marten Düring (Université du Luxembourg)
JHNR-support@historicalnetworkresearch.org

Cover Design and Typesetting
text plus form, Dresden, Germany
Cover image
Martin Grandjean
Copyediting
Andy Redwood, Barcelona, Spain

Published online at
https://doi.org/10.25517/jhnr.v7i1

The Journal of HISTORICAL NETWORK RESEARCH

ANDREA SANGIACOMO/
RALUCA TANASESCU/
HUGO HOGENBIRK/
SILVIA DONKER

Recreating the Network of Early Modern Natural Philosophy: A Mono- and Multilingual Text Data Vectorization Method

Journal of Historical Network Research 7 (2022) 33–85

Keywords History of philosophy, early modern natural philosophy, network analysis, text data vectorization, semantic features

Abstract How could one create a network representation of a book corpus which spans over two hundred years? In this paper, we present a method based on text data vectorization for a complex and multifaceted network representation of an early modern corpus of 239 natural philosophy textbooks published in Latin, French, and English. We use unsupervised methods (namely, topic modeling, term frequency – inverse document frequency, and multilingual word embeddings) to represent the broader features of this corpus, such as its homogeneity in style and linguistic usages, both among works written in the same language, and across multiple languages. We call this the 'textual dimension.' We also use a collocate analysis of specific keywords to explore how certain concepts were understood, reshaped, and disseminated in the corpus. We call this the 'semantic dimension.' Each of these two dimensions provides a different way of correlating the books via text data vectorization and of representing them as a network. Since these dimensions are complex and multifaceted, the network we construct for

C²DH LUXEMBOURG CENTRE FOR CONTEMPORARY AND DIGITAL HISTORY

UNIVERSITY OF LUXEMBOURG

each of them is a multiplex, made from several layer-graphs. Furthermore, using existing bio-bibliographical information, this research provides the grounds for further expanding the described network representation in such a way as to create a third multiplex, one that explores some of the social features of the authors in question.

eISSN: 2535-8863
DOI: 10.25517/jhnr.7i1.129

1. From Authors and Books to Multiplex Networks*

Current scholarship on the history of philosophy and science (Garber 2016) acknowledges that early narratives about the seventeenth-century Scientific Revolution (Butterfield 1957; Koyré 1957; Hall 1966; Westfall 1992) were overly simplified and limited in terms of the range of materials and authors they considered. Today's scholars working in this field thus struggle to enlarge the corpus of works they study and expand the canon of authors they consider. In previous research (Sangiacomo *et al.* 2021a and 2021b), we managed to compile a corpus of 239 early modern printed books (first editions only), containing approximately twenty million words, written in Latin (54%), French (27%) and English (19%), which are all concerned with providing a systematic and encompassing account of the changing field of natural philosophy between 1587 (Abraham de la Framboisière's (1560–1636) *Methodicae Institutiones*) and 1832 (John Robison's (1739–1805) *A System of Mechanical Philosophy, vol. 4.*).[1] This corpus is available at a sufficiently high OCR quality (with a minimum of 90% word-accuracy per page) as to allow for reliable automated text mining (Sangiacomo *et al.* 2022; Holley 2009).

Let us describe briefly the general information about the authors and works included in this corpus. First, they can be divided roughly into three 'nationalities,' corresponding to the national labels provided by the bio-bibliographical *Dictionaries*[2] used to compile the corpus (Sangiacomo *et al.* 2021 ibid.). This reference to 'nationality' should not be taken here to reflect historical categories; rather, it can be used as a working label to identify the broad political environment in which

* **Acknowledgements:** This article is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 801653).
Corresponding author: Andrea Sangiacomo

1 The corpus in question was compiled by first manually selecting relevant authors and titles from the bio-bibliographical dictionaries of early modern philosophers (cf. note 2), then expanding this list using keywords derived from the selected titles to scrape *WorldCat* for additional relevant works and authors. We extensively described the workflow behind the corpus collection and expansion, as well as the rationale behind exclusively using first editions, in Sangiacomo *et al.* 2021 a and b. The data set is available on Zenodo [Sangiacomo *et al.* 2021c].

2 Wiep van Bunge, Henri Krop, Bart Leeuwenburgh, Paul Schuurman, Han van Ruler and Michiel Wielema, *Dictionary of Seventeenth- and Eighteenth-Century Dutch Philosophers* (London: Bloomsbury, 2003); John Yolton, Valdimir Price and John Stephens. *Dictionary of Eighteenth-Century British Philosophers* (London: Bloomsbury, 1999); Andrew Pyle. *Dictionary of Seventeenth-Century British Philosophers* (London: Bloomsbury, 2000); Luc Foisneau. *Dictionary of Seventeenth-Century French Philosophers* (London: Bloomsbury, 2008); Heiner F. Klemme and Manfred Kuehn. *Dictionary of Eighteenth-Century Philosophers* (London: Bloomsbury, 2011). Whenever possible, we integrated these sources with other sources available, such as the Virtual Internet Authority File (viaf.org), EEBO/ECCO, and BnF Data (data.bnf.fr).

each author was predominantly embedded. Second, 23.5% of the titles included in this corpus contain a direct reference to one or more of three main authorities (scholastic, Cartesian and Newtonian), or are hybrid examples between these three authorities. A few more orientations or authoritative figures are mentioned, but they do not compare, quantitatively speaking, with these three. The above-mentioned percentage rose to 38.7% after further annotation on the grounds of the biographical information provided by the dictionaries. However, for the most part, the titles and authors in our corpus are mostly obscure or entirely forgotten within today's scholarship. We also lack explicit information about how particular authors or works were connected to one another (e.g., personal relationships or correspondences between authors,[3] direct references between works).

From a historical point of view, we might assume that the included authors and works did form some sort of network, in the loose, non-technical sense of the term, since they were all engaged to some degree or other with teaching, discussing, and exploring the same discipline, namely natural philosophy; they were also arguably aware of each other's work to some extent and did share a number of common philosophical and scientific sources, both old and more recent. Furthermore, all the titles in our corpus are what we call 'primary' sources, in the sense that they endeavor "to offer a systematic exposition of natural philosophy that could be used to teach it to new generations," (Sangiacomo *et al.* 2021b, p. 8) with many of them having been actually used in universities as teaching material. However, translating their connectedness into an actual network – in the more technical sense of the term (a representation of a dataset based on graph theory) – poses a number of challenges. Besides the issue of the lack of actual information about any direct links between these authors and works, we also face another serious challenge, related to the fact that these books were written in several languages (Latin, French and English). Dealing with multilingual corpora is a topic that is both relevant and thorny in past and current discussions in Natural Language Processing (hereafter NLP) (Schmidt and Wörner 2012; Lind *et al.* 2021), because it is accompanied by a number of technical difficulties (Kivelä *et al.* 2014).

In this paper, we provide a method based on text data vectorization for a complex and multifaceted network representation of the corpus at hand. By text data (or text document) vectorization, we refer to the idea of representing the distinctive features of text documents (in our case books) through quantitative and qualitative aspects associated with how words are used within them (Singh 2022; Shahmirzadi, Lugowski, and Younge 2019). In particular, the proposed method combines multiple computational tools and approaches to tackle both global

3 We checked the correspondence networks *EMLO* and *ePistolarium,* but the majority of our authors did not appear in these databases. Among those present, we found mostly canonical figures.

and specific features of the texts based on correlation scores between text vectors. First, we employ unsupervised methods (topic modeling, term frequency – inverse document frequency, and multilingual word embeddings) to represent style homogeneity and linguistic similarities in the overall corpus, both among works written in the same language, and across multiple languages. The aspects discussed will thus be more relevant to a 'textual dimension.' Second, we use a collocate analysis of specific keywords to explore how certain concepts were understood, reshaped, and disseminated in the corpus. The aspects discussed will thus be more conceptual in nature and they pertain to what we call 'the semantic dimension.'

Each of these two dimensions provides a different way of correlating the documents, and thus representing them as a network. Since each dimension is in itself complex and multifaceted, the networks we construct will not be single-layered (monoplexes), but rather multiplex networks composed of several layer-graphs. Therefore, this approach allows us to move from a simple initial inventory of books to the construction of two multiplex networks that represent the relationships between these books from the point of view of textual and semantic correlations. In addition, provided that we have enough information about the authors of the works included in our inventory, this paper provides the grounds for further expanding the initial network representation in such a way as to create a third multiplex network that explores some of the social features of the authors of the books we study. Our future work will incorporate a crucial characteristic of historical research, namely diachrony, which is currently restricted within the textual and semantic multiplexes due to the symmetrical relationship between word vectors.

While our main aim here is to present how this data-vectorization method works for one particular corpus of early modern science works, as well as to illustrate some of the technicalities and limitations behind it, we also emphasize a few remarkable observations that readily emerge from its implementation, and which require further investigation. While examining textual similarities, our network analysis allows us to identify a group of works that can be singled out as models for the 'average textbook' in early modern natural philosophy, at least from the point of view of style and use of language. Remarkably, the authors of these paradigmatic textbooks are mostly late scholastics working in the same context – the mid-seventeenth century Dutch Republic – and sometimes verging towards Cartesianism (in the case of authors like Johann Sperlette (1661–1725), Johannes Clauberg (1622–1665), and Antoine LeGrand (1629–1699)). Moreover, we can also note how the importance of later Dutch professors like Willem 's Gravesande (1688–1742) and Pieter van Musschenbroek (1692–1761) in the spreading of Isaac Newton's (1643–727) ideas depends on how they rewrote them in a style that was more akin to the gold standard of mid-seventeenth century scholastic textbooks. These are remarkable observations, which illustrate the heuristic potential of the method we describe.

Our discussion goes as follows. Section two describes the general rationale behind building our multiplex networks. Sections three and four present the two textual and semantic multiplexes. Section five concludes with a few reflections on how this work can be expanded upon by generating a third social multiplex.

2. Designing the Multiplex Networks

Multilayer network analysis has a prominent place in sociology, where it is used to depict and investigate the complexity of human relationships (Dickison et Rossi 2016). It is also widely used in biology and biomedicine to model the dynamics of complex biological systems and, generally, the dynamics of multifaceted real-world systems (Boccaletti *et al.* 2006), in which the uncertainty and complexity of the relationships between elements would be extremely difficult to render solely on the grounds of qualitative analyses (Sayama 2015). As a result, there is a lack of consensus regarding the attendant terminology, which varies widely (Kivelä *et al.* 2014). For the purpose of this paper, we will use the following terminology: 'multiplex network' to refer to a sequence of graphs with the same nodes (or multirelational graphs), in which the nodes belong to different layers and have different types of relationships in each layer (Wasserman and Faust 1994; Cozzo *et al.* 2016); and 'multilayer network' to refer to a graph in which the nodes and the relationships between them are (partially) different from layer to layer.[4] Therefore, the different types of reciprocal relationships rendered via undirected intralayer links between books will result in two multiplex networks (textual and semantic) while, at a later stage (presented in section 5), the intralayer links between authors will result in a social multiplex network.

The use of multiplex networks for representing historical data has recently been gaining attention. It has been applied so far to early modern correspondences in the Republic of Letters (Van den Heuvel 2015; Van Vugt 2017) as a tool to add further complexity and depth to the study of the rich information surrounding epistolary exchanges. This paper builds and expands upon these previous successful attempts by also showing how multiplex networks can be used to create a network representation for a corpus that, by itself, does not already come with clearly discernable edges among nodes (as in the case of correspondence).

In this respect, it should also be noted that our departing corpus surely allows for a diachronic analysis, but this paper *does not* primarily aim to offer such an analysis. Diachronicity is directly relevant in naturally directed corpora like correspondences or controversies, in which the succession of the exchanges be-

4 A complex graph made of different multiplex graphs is also known as a 'network of networks' (Kenett, Perc, and Boccaletti 2015); however, in this paper we will use the general term 'multilayer network'.

tween the actors has a direct impact on how the information flows and must be interpreted. Our corpus is different, however, since it consists of books that are not necessarily written in reply to one another, but are simply published in a certain temporal succession. At this point, the main task of our method is to create a sufficiently robust network representation for this corpus, and this goal is met primarily by using text data vectorization, that is, by building a symmetrical relationship between word and document vectors. While directionality can be further added to this representation as a way of exploring relevant nuance and aspects in the corpus, this is a step that we will take at a later stage in our project.

In the departing corpus, diversity is particularly apparent from two perspectives: in the multiplicity of different languages used (Latin, French, English) and the multiplicity of the 'nationalities' of the authors working in these languages (Dutch authors working mostly in Latin, and French and British authors working both in Latin and French or English). How these works correlate will be studied from a textual and semantic point of view, which will enable us to determine how similar (or different) the works are. Each kind of correlation will thus be the foundation of a different multiplex network, and in each of these multiplex networks we handle the presence of multiple languages and nationalities in different ways, adapted to the kind of correlation we plan to investigate.

The 'textual correlation multiplex network' ('textual multiplex,' for short) studies similarity from the more general perspective of the textual features of the entire books and the entire corpus. In this context, we directly take into account how textual correlations emerge by considering the corpus according to four different scenarios: (i) mono-lingual and mono-national (e.g., English, French, or Dutch books written Latin); (ii) mono-lingual and trans-national (e.g., books written in Latin irrespective of the nationality assigned to the authors); (iii) multilingual and mono-national (e.g., books by English authors written in English and Latin,); and (iv) multi-lingual and multi-national (namely, the corpus analyzed in its entirety). Since there is no computational technique that works or is adequate to cover all these scenarios, we use three main computational approaches: topic modeling (TM) and term frequency – inverse document frequency (hereafter tf-idf) when we deal with monolingual sub-corpora (scenarios i and ii), and multilingual word-embeddings (hereafter WE) when we broach multilingual corpora (scenarios iii and iv). In the textual multiplex, ties among books represent their potential for sharing a linguistic style and the use of language.

The 'semantic similarity multiplex network' ('semantic multiplex,' for short) provides a collocate analysis (hereafter CA) of specific keywords and then compares them in order to detect convergences or divergences in their usages across sub-corpora composed of books written in the same language. In this context, similarity is taken to express the linguistic and conceptual usage of specific key terms, hence different books are associated via similarities in their linguistic or conceptual features. National diversity can be represented here in terms of the

different ways in which the same monolingual corpora can be split based on national differences. In this semantic multiplex, computational methods do not seem to offer reliable tools for multilingual analysis. Thus, we provide a manual, human-based analysis of how results associated with keywords in different languages converge based on our own assessment of how they translate to one another. In the semantic multiplex, ties among books represent their potential for sharing conceptual usages associated with specific keywords.

Table 1 provides an overview of how different sub-corpora are tackled by using different methods in each multiplex described so far; we also include the potential 'social multiplex' that we will mention in the conclusions.

Note that the potential complexity of the corpus is not approached homogeneously across the multiplex networks. The textual multiplex tackled all the ways in which the corpus itself can be articulated, including across linguistic and national divides. It thus creates a more comprehensive background for the other two networks. The semantic multiplex instead represents the corpus only from a monolingual point of view, doing away with national divisions. The social multiplex does not discriminate between authors publishing in different languages, nor does it include sublayers based on 'nationality' (which is an artificial division introduced by how the corpus of works was created), although (as we will explain in the conclusions) national divisions can be derived from the layers that consider the geographical locations of affiliations and publishers.

In formally analyzing these networks, we take into account three centrality measures, interpreted as follows: (a) degree centrality: the number of connections a work established in the network (Newman 2010, 169); (b) eigenvector centrality: a measurement which establishes the most prominent nodes in the network, heavily dependent on the degree of the respective node (work), as well as on the degree of the works connected to that node (ibid.); (c) betweenness centrality: a measure of the centrality of a work based on the number of shortest paths between other works going through the respective work (ibid., 186). Thus, most of the works with the highest number of links will appear to have the highest eigenvector while, typically for the corpus in question, the node with the lowest degree will rank highest in terms of betweenness centrality if it is placed on the path between other nodes. While in social networks centrality measures are often applied to extrapolate a node's power or influence position, here they are better understood as a *potential for centrality,* with their actual fulfillment to be assessed through (qualitative) hermeneutic analysis (Düring 2016).

	LAYERS AND METHODOLOGIES			
Type of corpus/ Scenario	**Textual multiplex**		**Semantic multiplex**	**Social multiplex**
i. *Mono-national Mono-lingual*	NL_Latin (TFIDF)	NL_Latin (TM)	Not Applicable (N/A)	N/A
	FR_Latin (TFIDF)	FR_Latin (TM)	N/A	N/A
	FR_French (TFIDF)	FR_French (TM)	N/A	N/A
	UK_Latin (TFIDF)	UK_Latin (TM)	NA	NA
	UK_English (TFIDF)	UK_English (TM)	N/A	N/A
ii. *Multi-national Mono-lingual*	NL, FR, UK_Latin (TFIDF)	NL, FR, UK_Latin (TM)	NL, FR, UK_Latin (CA)	N/A
	NL, FR, UK_French (TFIDF)	NL, FR, UK_French (TM)	NL, FR, UK_French (CA)	N/A
	NL, FR, UK_English (TFIDF)	NL, FR, UK_English (TM)	NL, FR, UK_English (CA)	N/A
iii. *Mono-national Multi-lingual*	NL_Latin and French (WE)		N/A	N/A
iv. *Multi-national Multi-lingual*	NL, FR, UK_Latin, French, English (WE)		N/A	NL, FR, UK_Latin, French, English (SNA)

Tab. 1 Network configurations and corresponding methodologies.

3. The Textual Multiplex

3.1 Method

Our early natural philosophy corpus required a multifaceted methodological approach that took into consideration mono- and multilingualism. Most of the existing work related to text representation techniques has been carried out in monolingual contexts, and it is only recently that researchers have started to work in multilingual contexts, mostly due to the growing plurilingual databases available online and in historical archives (Zosa and Granroth-Wilding 2019). We used three established text mining methods to explore general linguistic similar-

ities in the corpus at hand. The first is topic modeling, an unsupervised machine learning technique that discovers topic models in large sets of monolingual unstructured texts and clusters the texts accordingly (Blei, Ng, and Jordan 2003; DiMaggio *et al.* 2013; Jockers and Mimno 2013; Suominen and Toivanen 2015; Allen and Murdoch 2020; Blanke and Aradau 2021). The second is tf-idf vectorization, a numerical statistic which clusters monolingual texts on the grounds of their textual similarity by evaluating how important a word is to a document within a collection of documents (Aizawa 2003; Bafna *et al.* 2016; Jabri *et al.* 2018). The third is multilingual word-embeddings, a technique to represent words in a text as vectors based on each word's context and 'embed' these vectors in a vector space (Bjerva and Praet 2015; Joulin *et al.* 2016; Alaux *et al.* 2019). More details on the specifics of each approach are offered in the next corresponding subsections.

In the following, we discuss how we implemented each of these three methods on different sub-sections of the corpus:

1) English, French, Dutch-Latin sub-corpora (mononational, monolingual): topic modeling and tf-idf (cf. § 2, scenario i)
2) English, French, Latin corpora (multinational, monolingual): topic modeling, tf-idf (cf. § 2, scenario ii)
3) English, French (mononational, multilingual): multilingual word-embeddings (cf. § 2, scenario iii)
4) The whole corpus (multinational, multilingual): multilingual word-embeddings (cf. § 2, scenario iv)

In practice, we can compare the results obtained by applying topic modeling and tf-idf from two perspectives: from that of languages and from that of nationalities. We can also compare the results obtained by using word embeddings according to two perspectives: specific nationalities, and the whole corpus together. These types of slicing allow us to investigate this collection of books from various angles which, together, will arguably offer a multifaceted perspective on the landscape of early modern science as a budding discipline.

3.2 Topic modeling

Topic modeling uses a probabilistic model (the Latent Dirichlet Allocation, LDA) that estimates probability distributions for topics in documents and words in topics related to natural philosophy (Blei and Jordan 2002). However, since our corpus consists of natural philosophy works that we handpicked for their systematic nature, understanding the way the algorithm clustered them on the basis of a topic that appears to be very similar to the others was bound to be difficult. An important shortcoming of topic modeling is the fact that researchers need to provide the algorithm with a certain number of topics, without actually knowing what the most relevant number for a certain corpus is. We overcame this pit-

fall by algorithmically calculating coherence values (Röder, Both, and Hinneburg 2013), which establishes the optimal number of topics for a given collection of texts.[5]

Four main open-source Python libraries were used for the building of the pipeline: CLTK for the lemmatization of the Latin corpus; NLTK for the lemmatization of the English and French corpora and for document pre-processing (cleaning and tokenization); Gensim for the calculation of coherence values, for the extraction of the topic models, and for the vectorization of the topics and documents; and finally, NetworkX for representing the corpora as networks based on their Numpy similarity matrices.

In order to more concretely illustrate how this method works, let us consider each monolingual sub-corpus. The topic modeling-based layer connects works on the grounds of their adherence to each of the topics identified (Gretarsson *et al.* 2012). After we computationally calculated the number of most suitable topics, in each case using a coherence value algorithm, we distilled the following topics:

Latin **Topic 1:** *corpus, pars, motus, ratio, moveo, aqua, locus, tempus, ignis, terra, radius*
Topic 2: *ratio, pars, deus, species, forma, corpus, materia, homo, anima, locus, causa, potentia*
Topic 3: *deus, ratio, homo, pars, species, genus, forma, natura, causa materia, corpus*
Topic 4: *senatus, romanus, aequilibrium, corpus, deus, ratio*

French **Topic 1:** *corps. partie, chose, nature, dieu, homme, cause, matière, terre, raison, mouvement*
Topic 2: *corps, partie, terre, air, point, raison, soleil, mouvement, cause*
Topic 3: *corps, partie, point, force, mouvement, air, terre, égal, ligne, eau, rayon, vitesse*

English **Topic 1:** *force, body, motion, time, water, velocity, part, distance, earth*
Topic 2: *thing, part, body, motion, reason, cause, time, nature, place, particles*

5 Being aware of the conceptual limitations of topic modeling and of their demographical (or statistical) relevance, we carried out a thorough cleaning of the corpus by means of a very large list of stopwords which contained conceptually spurious words, such as Lat. *nimirum, propterea, solum* or even verbs like *esse* or *facere.* These words are not typically listed among stopwords, but proved to have a great impact on the relevance of the outputted topics.

Topic 3: *motion, parts, body, particle, atoms, time, nature, reason, part, water, matter, space*
Topic 4: *angle, latitude, part, line, degree, place, circle, time, meridian, star, radius, distance*
Topic 5: *form, part, water, time, animal, body, angle, specie, surface, light*
Topic 6: *body, part, water, nature, motion, light, thing, fire, time, matter, cause, color*

These topics very broadly represent how certain keywords are mostly used in several groups of texts. Each keyword in each topic has a different weight, which also determines its relevant importance within that topic.

In the case of the Latin sub-corpus, we can begin to note a difference between topics 1 and 4 on one side, and topics 2 and 3 on the other. Topics 1 and 4 are predominantly composed of keywords most directly associated with physical reality, the subject matter of natural philosophy. These keywords are also relatively neutral concerning any ideological or philosophical orientation, since all works about natural philosophy will arguably have to discuss bodies (*corpus*), motion (*motus*), proportions (*ratio*), and so on. In topics 2 and 3 we can observe that other keywords gain prominence, associated with a more specific scholastic way of treating natural philosophy. Scholastic natural philosophy is (a) usually presented within a more systematic treatment of other traditional disciplines (including metaphysics, logic and ethics), and (b) heavily relies on the technical jargon of Aristotelian origins. The first point is exemplified by the greater prominence that keywords like God (*deus*), soul (*anima*), man (*homo*) acquire, while the second can be seen from the presence of keywords that can easily be connected with scholastic methods, like matter (*materia*), form (*forma*), cause (*causa*), power (*potentia*).

Note that topics for the vernacular languages are relatively more homogeneous among themselves, and consistent with the Latin topics 1 and 4. In fact, they suggest an even stronger focus on discussions about natural phenomena, and leave little room for terminology that might have a more distinctive metaphysical or even scholastic feel. This suggests that natural philosophy in vernacular languages tends to be more concerned with descriptions of natural phenomena themselves, experimental reports, natural curiosities, or handbooks for relatively younger students, while Latin would remain available for the sort of more encompassing discussion of natural philosophy within a metaphysical and systematic context, as is regularly the case with scholastic approaches to this discipline. This provides a clue to explain the data presented below in § 4.3 about the relative greater homogeneity of French- and English-writing authors compared to Latin-writing authors. The reason for such homogeneity may be due to the nature of the works we are considering, which are more diverse (in terms of the range of topics they cover) in Latin than they are in French or English.

The following step consisted of vectorizing each document on the basis of the topic vectors (that is, converting topic vectors into numerical representations), in order for the machine to process the textual contents of our corpus. Every single document in the corpus features all k topics in various (probabilistic) ratios (or scores, or simply values). That is, a certain document is x% about topic 0, y% about topic 1, and so on and so forth. x, y, z, etc. make up a vector representing the respective document from the point of view of these k topics. Every document in the corpus is thus represented by a different vector (with an identical structure). These vectors are correlated, in the sense that values are computed that quantify the similarity between any and every two of them. The quantifications are read as weights of the edges uniting the vectors (and therefore, the documents) within a network graph. We then formalized the relationship between the vectors into an undirected network, in which the books in their textual data form are the nodes and the links are the correlation scores between the topics shared by most of any two nodes.

Figure 1 exemplifies the network of books written in Latin and connected on the grounds of the four above-mentioned topics. The homogeneity of the network should not be striking, since any two book-nodes are linked by four links (topics) of various weights.

In spite of the visually opaque structure of the network, the various network metrics prove to be revealing. The ten most connected nodes in the topic modeling layer of multi-national Latin works are presented in Table 2.[6] They are each linked to 66 other nodes, all on the grounds of topic 1. These are also the nodes which have the highest eigenvector centrality, and which established the strongest correlations in the entire Latin corpus. Also, although we have listed the ten most connected nodes, there are actually 30 nodes with the same score (66 connections), which attests to the homogeneity of the network. When we create the same Latin network on the grounds of 5 topics, for instance, there will be 26 highly connected nodes, with 53 connections each. The same degree of homogeneity can easily be noticed in the case of the French and English layers, in which the most connected nodes are linked to 45 and 26 other titles, respectively.

Note that Newtonians ('s Gravesande, van Musschenbroek, Robison, Colin MacLaurin (1698–1746), David Gregory (1659–1708)) immediately appear as the most connected nodes, both in Latin and British subcorpora. Of all the nodes,

6 The years indicated in the table do not point to a diachronic analysis of the corpus. Rather, they indicate the linguistic composition of the sample – with Latin and French writings spanning over (almost) the whole period and English gaining prominence starting in mid-seventeenth century; these are presented as a timeframe for the most connected ten nodes in each linguistic sublayer.

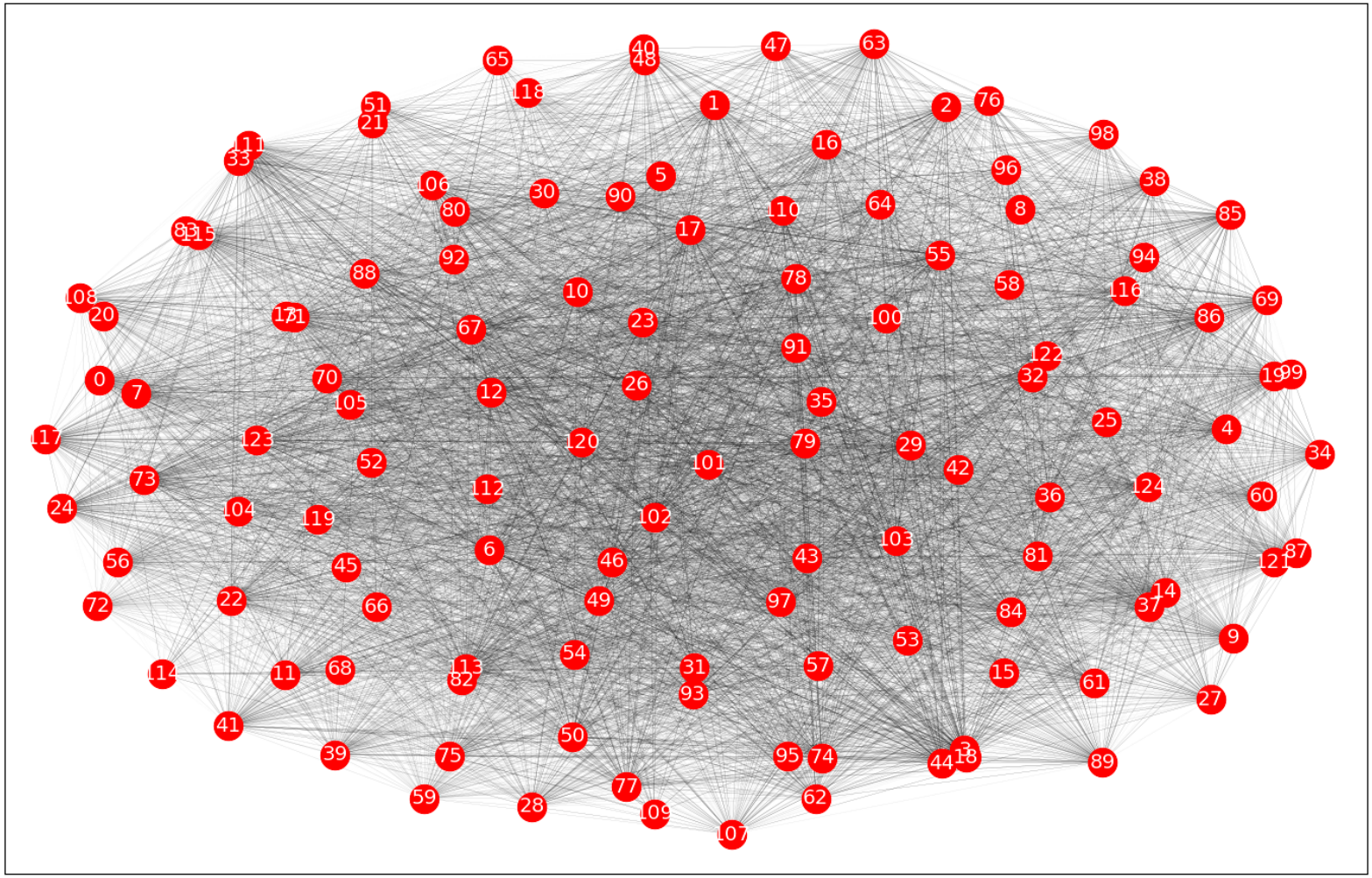

Fig. 1 Network representation of the Latin corpus cf. topic models (spring layout).

eISSN: 2535-8863
DOI: 10.25517/jhnr.v7i1.129

LATIN (123 nodes, 1587–1800)	
1. 1723_GRAVESAND *Philosophiæ Newtonianaes institutiones*	6. 1695_LeCLERC *Physica sive de rebus corporeis*
2. 1702_GREGORY *Astronomiae physicae et geometricae elementa*	7. 1720_GRAVESANDE *Physices elementa mathematica*
3. 1727_ODÉ *Principia philosophiae naturalis*	8. 1757_DeLaCAILLE *Lectiones elementares astronomicae*
4. 1734_MUSSCHENBROEK *Elementa physicae*	9. 1681_DuHAMEL *Philosophia vetus et nova, vol. 5*
5. 1748_MUSSCHENBROEK *Institutiones physicae*	10. 1762_MUSSCHENBROEK *Compendium physicae experimentalis*
FRENCH (61 nodes, 1606–1799)	
1. 1793_MONGE *Encyclopédie*	6. 1760_DUFIEU *Manuel physique*
2. 1781_PARADuPHANJAS *Elémens de physique*	7. 1789_PAULIAN *Dictionnaire, vol. 3*
3. 1762_PAULIAN *Dictionnaire, vol. 2*	8. 1777_TAITBOUT *Abregé élémentaire*
4. 1773_PAULIAN *Dictionnaire, vol.1*	9. 1772_PARADuPHANJAS *Théorie des êtres sensibles*
5. 1769_PAULIAN *Système général de philosophie, vol. 3*	10. 1787_SIGAUD-LAFOND *Eléments de physique*
ENGLISH (55 nodes, 1644–1822)	
1. 1748_RUTHERFORTH *System of Natural Philosophy*	6. 1812_PLAYFAIR *Outlines of Natural Philosophy*
2. 1804_ROBISON *Elements of Mechanical Philosophy*	7. 1705_DITTON *General Laws of Nature and Motion*
3. 1822_ROBISON *A System of Mechanical Philosophy, vol. 1*	8. 1822_ROBISON *A System of Mechanical Philosophy, vol. 2*
4. 1803_WOOD *The Principles of Mechanics*	9. 1775_MACLAURIN *Newton's Philosophical Discoveries*
5. 1822_ROBISON *A System of Mechanical Philosophy, vol. 3*	10. 1822_ROBISON *A System of Mechanical Philosophy, vol. 4*

Tab. 2 Top 10 highest ranking works in terms of degree, eigenvector centrality, and connection strength in the monolingual topic layers.

Node #	Author	Dominant topic(s)	# of shortest paths
114	Columbus	Topic 2 (99%)	13,854
47	's Gravesand	Topic 1 (99%)	1,890
107	Swinden	Topic 1 (67%), Topic 4 (32%)	1,594
8	Boyvin	Topic 3 (99%)	714
118	Burgersdijk	Topic 1 (16%), Topic 3 (62%), Topic 4 (21%)	238
105	Pourchot	Topic 3 (78%), Topic 4 (15%)	232

Tab. 3 Nodes presenting betweenness centrality above 0 in the Latin corpus.

they also establish the strongest connections between themselves. However, if we consider betweenness centrality, we observe that the most strategically positioned works are actually scholastic. Table 3 presents the top six nodes with positive betweenness centrality, in connection with the most significant topic to which they are associated.

Columbus (fl. 1635–1665) is a French late scholastic author, and his work (published in 1669) is a fairly standard scholastic textbook of Scotist orientation. This work has by far the highest number of shortest paths connecting other nodes in the network, although the node itself has the lowest number of links in the network. This means that this work is the nexus between many other authors, although not directly related to them. We interpret this fact as suggesting that our corpus mostly 'talks' in a scholastic-like fashion, although this does not mean that it necessarily, or even predominantly, endorses a scholastic approach in natural philosophy. This point is corroborated by topic distribution. The topic distribution for nodes with significant betweenness centrality indicates that the most influential topic is Topic 2 (the one with greatest scholastic flavor), although the best represented is Topic 1 (the more neutral). The sub-corpus that best represents Topic 1 is a very cohesive and strongly connected group of thirty texts, which are further connected to 66 other works.

However, the works in which Topic 2 is the most dominant are those that are most likely to influence the whole corpus, given their high betweenness. Based on the suggestion above, this means that scholastic ways of phrasing and presenting natural philosophy discussions will have a significant impact on how *all* works in our corpus deal with the field. This explains in what sense Columbus's text, which is placed on the highest number of shortest paths between two other nodes (13,854), is the most representative of the corpus under scrutiny topic-wise, although it is directly connected (that is, it shares a similar topic distribution) to only 6.89 other nodes.

Overall, the topic modeling analysis of monolingual sub-corpora reveals that our corpus is relatively homogeneous from a textual point of view, which is expected given that it includes systematic works that focus on the same subject matter. Nonetheless, we also begin to discern some subtle currents that shape the corpus, like the distinction between a more 'neutral' use of language, or a more 'scholastic' approach. These differences are further explored via the second method, tf-idf vectorization.

3.3 Tf-idf

The second approach we use to link the works relies on tf-idf, the algebraic model that analyses a document at the word-level and determines how representative a word is of a collection of documents (Lavin 2019). This statistical metric multiplies how many times a word appears in a document by the inverse document frequency of the word across a set of documents. Since topic models revealed a very homogenous natural philosophy collection of books, the tf-idf-based layer provided the grounds for further differentiations in clustering, because it not only takes into account the most frequent words, but also considers the least frequent ones and their distribution over the whole corpus. Tf-idf is a good indicator for textual similarity and it does so for each word in isolation (using word-vectors), not contextually. For example, from the point of view of topic modeling a book that mentions *corpus* and another that focuses on *pars* will appear very close together, as keywords of the same (or very similar) topic, while from the point of view of tf-idf scores they may appear further apart, because the two may be statistically different.

In applying tf-idf, we use the same correlation score methodology as above, except that each document is represented by a tf-idf vector. The vectorization in this case was carried out on the basis of the scikit-learn open-source library in Python, and was used to link the nodes (the books) based on a similarity matrix. Again, we use this to underscore elements of continuity among our books, but we also begin to discern differences. Tf-idf identifies related bodies of text across a large data set by taking into account both the most frequent and the least frequent terms. Existing literature shows that this scales better than topic modeling with large corpora (Nguyen *et al.* 2015; Carrera-Trejo *et al.* 2015; Venkatesaramani 2019). However widely used, the method is not without its own shortcomings, of which the most relevant for our purposes are the flattening out of polysemy and the fact that it does not capture co-occurrences across documents. Working with multiple layers thus becomes even more relevant: topic modeling makes up for the sheer statistical nature of tf-idf, as tf-idf statistically nuances the conceptual fogginess of topic models.

The network thus realized (Figure 2) emphasizes once again the visual homogeneity of the corpus, to an even higher degree than in the case of the monolingual topic models. And again, this is not surprising, since the number of word

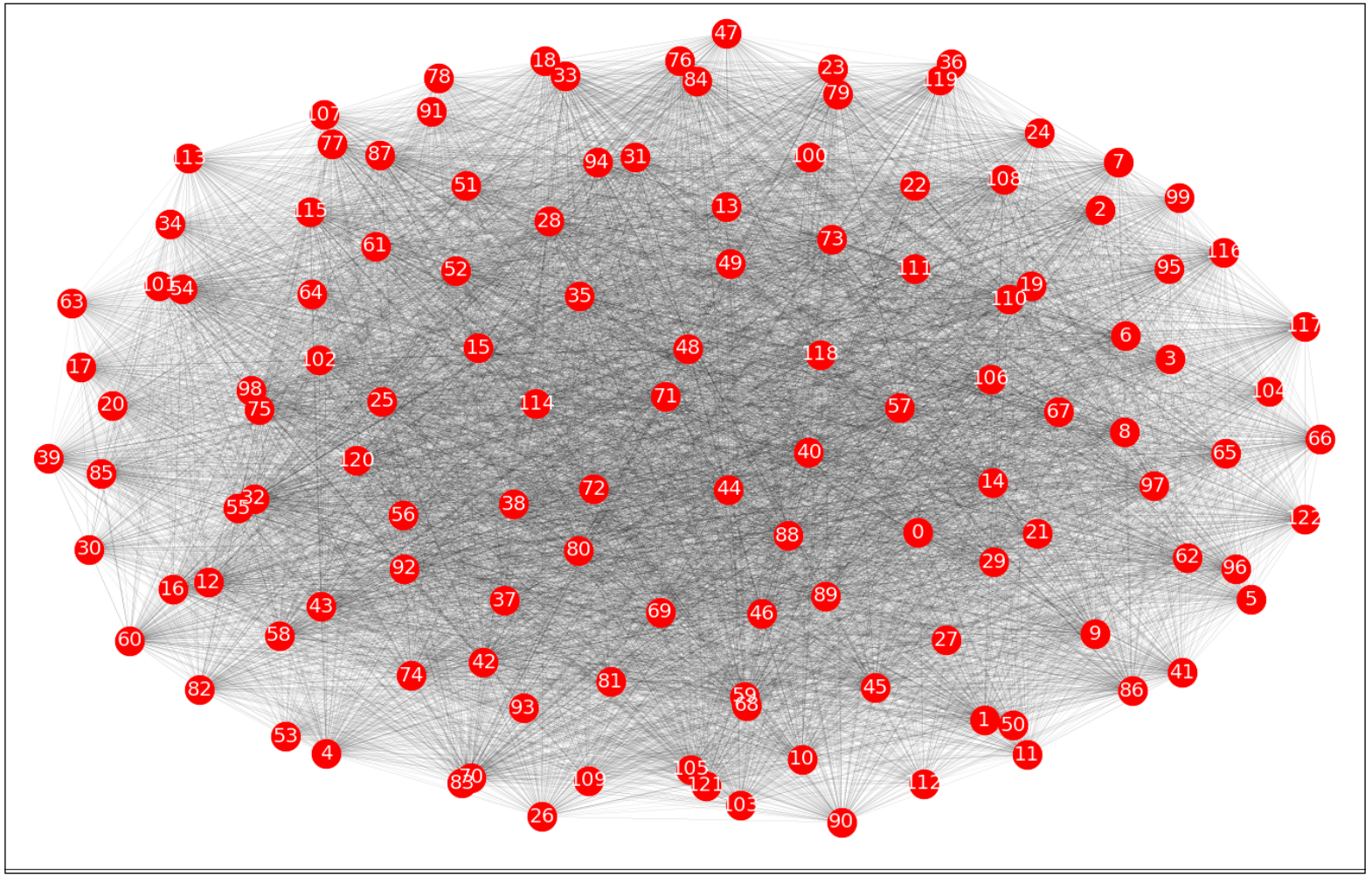

Fig. 2 Network representation of the Latin corpus cf. tf-idf vectors (spring layout).

eISSN: 2535-8863
DOI: 10.25517/jhnr.v7i1.129

Highest Degree	Highest Eigenvector	Highest Betweenness
1. 1739_FRASSEN *Philosophia academica* **76.047 (node 28)**	1. 1739_FRASSEN *Philosophia academica* **0.114 (node 28)**	1. 1762_DeLaCAILLE *Ad lectiones elementares astronomiae* **0.04 (node 91)**
2. 1672_LeGRAND *Institutio philosophiae* **75.152 (node 115)**	2. 1672_LeGRAND *Institutio philosophiae* **0.113 (node 115)**	2. 1783_SEGUY *Philosophia ad usum scholarum accommodata* **0.005 (node 112)**
3. 1664_CLAUBERG *Physica* **74.136 (node 60)**	3. 1664_CLAUBERG *Physica* **0.111 (node 60)**	3. 1660_CHABRON *Philosophia per breviter argumenta explicata* **0.004 (node 50)**
4. 1694_SPERLETTE *Physica nova* **73.173 (node 88)**	4. 1694_SPERLETTE *Physica nova* **0.110 (node 88)**	4. 1800_VanDerEYCK *Institutiones physicae* **0.0001 (node 20)**
5. 1647_STIER *Praecepta doctrinae logicae* **72.215 (node 86)**	5. 1647_STIER *Praecepta doctrinae logicae* **0.108 (node 86)**	**Lowest degree:**
6. 1644_DEUSING *Naturae theatrum universale* **71.106 (node 117)**	6. 1644_DEUSING *Naturae theatrum universale* **0.107 (node 117)**	1832_JACQUIER *Institutionum philosophicarum synopsis* **25.60 (node 93)**
7. 1655_FOURNENC *Universae philosophiae synopsis* **71.084 (node 116)**	7. 1655_FOURNENC *Universae philosophiae synopsis* **0.106 (node 116)**	
8. 1652_SENGUERDArnold *Collegium physicum* **70.560 (node 122)**	8. 1652_SENGUERDArnold *Collegium physicum* **0.1066 (node 122)**	
9. 1645_KYPER *Institutiones physicae* **70.489 (node 40)**	9. 1645_KYPER *Institutiones physicae* **0.1063 (node 40)**	
10. 1649_BASSON *Philosophiae naturalis adversus Aristotelem Libri XIII* **70.336 (node 43)**	10. 1649_BASSON *Philosophiae naturalis adversus Aristotelem Libri XIII* **0.1062 (node 43)**	
Strongest correlations		
1726_MUSSCHENBROEK *Epitome elementorum physico-mathematicorum* & 1722_CLERC *Opera philosophica, vol. 4* **0.999 (nodes 15 & 34)**	1688_LANGENHERT *Compendium physicae* & 1688_GEULINCX *Compendium physicae* **0.988 (node 84 & 85)**	1722_CLERC *Opera philosophica, vol. 4* & 1734_MUSSCHEN-BROEK *Elementa physicae* **0.867 (nodes 34 & 51)**

Tab. 4 Node ranking in the tf-idf layer.

vectors shared by any two document nodes are higher than 4 (the optimized number of topic models).

Nevertheless, various network metrics indicate a different configuration of the most central nodes. Table 4 presents the measurements resulting from the tf-idf analysis of the whole Latin corpus, according to which documents with similar relevant words will group together.

The highest degree indicates the ten most connected works in terms of textual similarity, each textually similar to between 76 and 70 other works, which also makes them the most prominent in the network. However, although they are connected to a large number of other works, the correlations they establish are not among the strongest. As far as the strength of correlation is concerned, works pertaining to the same author tend to naturally rank the highest. Nevertheless, in spite of the many cases of multiple works by the same author in the corpus, only van Musschenbroek and Francis Bacon (1561–1626) rank high in this respect, which means that they are the most consistent in terms of writing. Another general observation is that French authors tend to establish the least connections (e.g., Jacquier scores the lowest degree), indicative of the fact that they have a different writing style from most of the other authors. This observation is also supported by the fact that four of the nodes with the highest betweenness centrality are French. It has been shown that the highest betweenness points towards hybrid or interdisciplinary modes of writing (Evans 2016), which in our case may be an indication of French authors using an eclectic discourse, informed by their scholastic predecessors, as well as references to new trends (e.g. Cartesian). Finally, the Dutch tend to be the most representative of a textual style (the way language is used across the corpus), as we can see from the list of the most strongly correlated nodes.

As mentioned above, tf-idf is a way of representing the similarity among a number of works by taking into account their linguistic homogeneity. In this context, degree centrality is a direct measure of similarity, since works will be more connected if they are more similar among each other. When looking at the list of works that score highest in terms of degree centrality, a very apparent feature is that they are all published in a relatively short range of time. Six works are published in a 10-year window, between 1644 and 1655. Another three works are published in the second half of the seventeenth century (1664, 1672, 1694), and only one work is published in the eighteenth century (1739). In the first 10-year cluster, one work belongs to the British corpus (Stier, 1599–1648), two to the French corpus (Sébastien Basson, cc. 1573–?? and Jacques Fournenc, 1609–1665), and three to the Dutch corpus (Antonius Deusing, 1612–1666; Albertus Kyper, 1614–1655, and Arnold Senguerd, 1610–1667). Outside of this cluster, two belong to the French corpus (Claude Frassen, 1620–1711 and Jean Sperlette, 1661–1740), one to the Dutch corpus (Johannes Clauberg, 1622–1665), and one to the British corpus (LeGrand).

This distribution invites two main considerations. First, some of the most similar works in the corpus are also published across different countries in a rather limited range of time (the 10-year cluster). They also share a common scholastic orientation (even in the case of Basson who, while being critical of Aristotelian philosophy, directly engages with it). Note that the production of scholastic works is not limited to this specific decade. And yet, tf-idf seems to find here the greatest linguistic homogeneity among scholastic works. We can then identify a group of works that can be singled out as models for the 'average textbook' in early modern natural philosophy, at least from the point of view of their style and use of language. Remarkably, the authors of these paradigmatic textbooks are mostly late-scholastics working in the same context, the mid-seventeenth century Dutch Republic, who sometimes incline towards Cartesianism (Sperlette, Clauberg, LeGrand). Second, in terms of national corpora, the French and the Dutch are equally the most represented, leaving the British corpus with only two works (Stier and Frassen), which are both scholastic. This might be correlated to the fact that Latin is a more prominent language in the French and Dutch corpora in comparison to the British (32% compared to, for instance, 94% in the Dutch corpus), and with the fact that scholastic authors tend to write in the most similar and standardized way, but are relatively less present in the British corpus.

What is interesting to note here is that the only Newtonian work that ranks high from the point of view of tf-idf is one of van Musschenbroek's, a Dutch Newtonian writing in Latin. The fact that very few Newtonian works appear to score highly in this layer might be an indicator that, overall, our corpus contains more works akin to scholastic and Cartesian orientations, while Newtonian works are present but relatively isolated. This prompts an intriguing hypothesis. We surmise that one aspect of the importance played by Dutch professors like 's Gravesande and van Musschenbroek in the spreading of Newton's ideas depends on how they rewrote them in a style that was more similar to the gold and accepted standard style of mid-seventeenth century scholastic textbooks. They might thus have contributed to popularizing Newton, especially in Latin, by conveying Newtonian ideas through a way of writing and dealing with the subject that was already well-established and widely accepted. They contributed to the acceptance of Newton by translating his view into a language that looked less original and idiosyncratic with respect to the average norm established in the discipline, and which was mostly represented by scholastic textbooks.

3.4 Multilingual word-embeddings

While the LDA and tf-idf models are suitable for clustering monolingual documents, the documents in the multilingual corpora required a different approach. One of the very few available methods of clustering multilingual documents using word representations is *fastText* (Joulin *et al.* 2016), an open-source Python library that allows users to represent texts based on 157 pre-trained language models. Since work on unsupervised multilingual text representation and

classification without external sources (such as machine translation, parallel corpora, or bilingual dictionaries) is still in progress, and what has been proposed so far uses, for most part, the same technique (multilingual word-embeddings), we limited ourselves to one single starting layer (corresponding to the entire primary corpus in the three languages).

The layer created using multilingual word-embeddings is just as tightly knit as the other layers, although nodes tend to group together more than in the other two (cf. Figure 3), and they do so primarily on the grounds of language: a certain text will establish the strongest connections with another text written either in the same language or between two well represented languages, because performance across languages in *fastText* is unequal (i.e., the alignment of vectors between English and French will be more accurate than that between Latin and French). The highest degree is 12, represented by a group of seven works (six Dutch: Gilbertus Jacchaeus, 1578–1628; David Gorlaeus, 1591–1612; Ruard Andala, 1665–1727; Adriaan Heerebord, 1613–1661; Albertus Kyper; David Gorlaeus, and one British: John Cook, unknown life dates), and the lowest degree is 3 (Margaret Cavendish's *Observations*). Again, Dutch authors tend to be the most representative of the corpus in terms of how language is used, while Cavendish (1623–??) is one of the most original authors and establishes the weakest connection with another British author, Cook (*Clavis Natura*).

In terms of centrality measures, the two authors that rank highest are Jacchaeus (a rather traditional scholastic) and Andala (a Cartesian). This confirms some of the observations we made previously: that the corpus is dominated by works written in a scholastic style, although Cartesians are also successful in widely disseminating their approach. Noticeably, a few Newtonian authors are also relatively high in eigenvector centrality: Cook ranks 13th and Hugh Hamilton (1729–1805) 26th. These Newtonians thus managed to somehow articulate and shape their views using a style and language that remains relatively connected and similar to those of the majority of the other authors. However, many Newtonians (Benjamin Wilson, 1721–1788; MacLaurin; James Wood, 1760–1839; Thomas Rutherford, 1712–1771; Benjamin Worster, fl. c. 1722–1730); Edward Peart, 1756?–1824) rank extremely low in terms of eigenvector centrality, showing that (especially by the second half of the eighteenth century) Newtonians usually did not share the same use of language as the majority of the other authors in the corpus. In other words, a Newtonian way of writing tended to isolate itself and establish a new style in its own right.

This point is confirmed by looking at the most strongly correlated works, namely those that have similar discourses irrespective of their language. The titles presented in Table 5 are the most strongly connected between any two of them.

Except for Walter Charleton (1619–1707), an eclectic author conversant with Cartesianism, and LeGrand (a Cartesian), all the works included here somehow

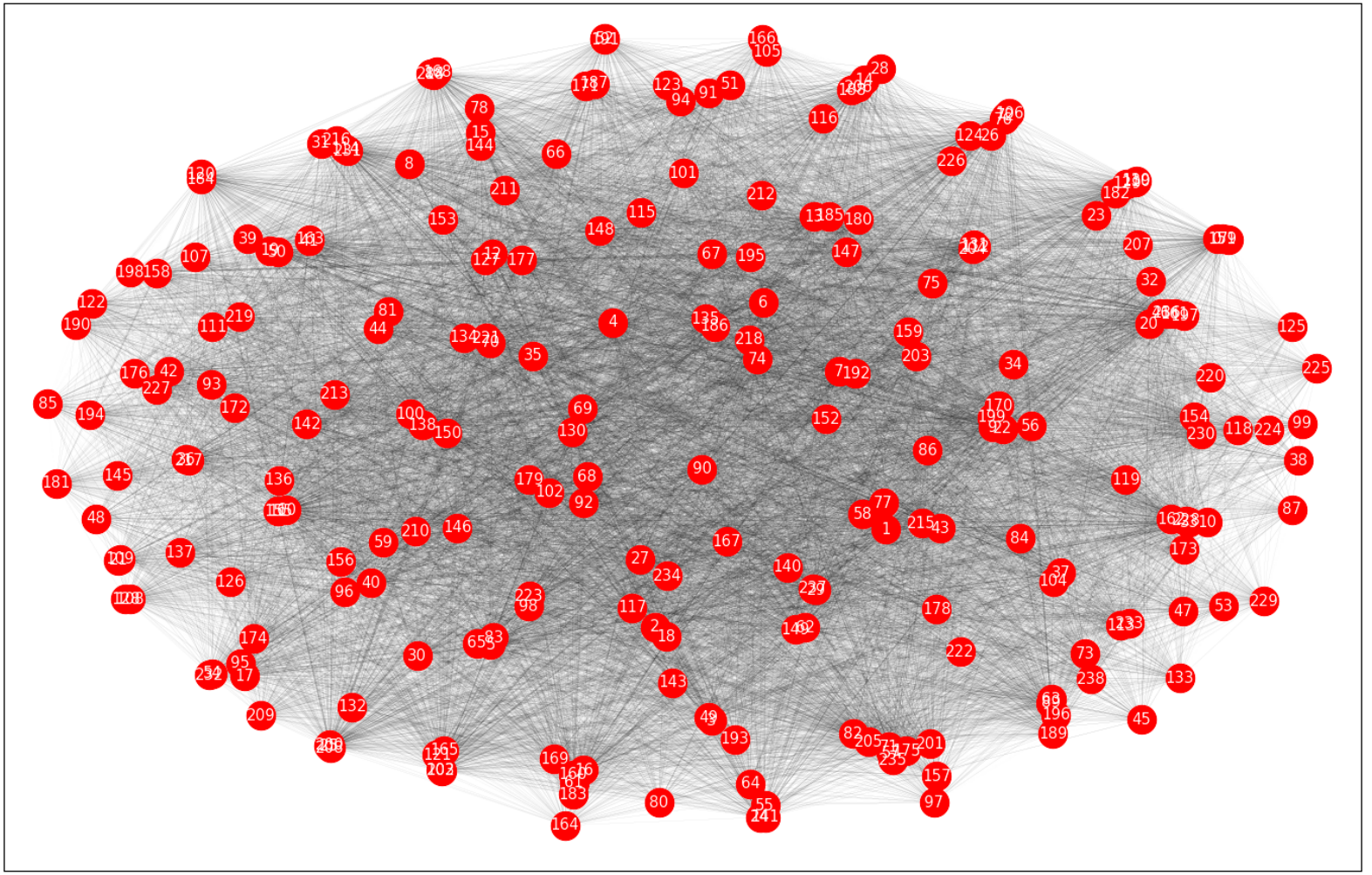

Fig. 3 Network representation of the whole multilingual corpus cf. fastText vectors.

eISSN: 2535-8863
DOI: 10.25517/jhnr.v7i1.129

1654_CHARLETON *Physiologia Epicuro Gassendo Charltoniana*	1730_WORSTER *Principles of Natural Philosophy*
1733_COOK *Clavis natura*	1694_LeGRAND *An Entire Body of Philosophy*
1774_HAMILTON *Four Introductory Lectures*	1803_WOOD *The Principles of Mechanics*
1754_WILSON *The Principles of Philosophy*	1748_RUTHERFORTH *System of Natural Philosophy*
1775_MACLAURIN *Newton's Philosophical Discoveries*	1789_PEART *On the Elements*

Tab. 5 The most strongly correlated works in the whole multilingual corpus.

built upon, expanded, propagated, or developed Newton's approach. This means that, from the point of view of the overall multilingual corpus, English Newtonian works established the most consistent style of writing. While scholastic works tend to share a similar style, we already noted that they diverge quite significantly on various details. Newtonian works in English show a different pattern: they segregate themselves more significantly with respect to all the other works in the corpus and tend to use language in a consistent way, with relatively less divergences among themselves. This provides further support to what we will observe below (§ 4.3) while studying the use of specific keywords among Newtonians.

Existing scholarship knows well that, by the end of the eighteenth-century, Newtonian natural philosophy tends to replace the scholastic approach. Our data suggests that, from the point of view of the use of language and style, this transformation was arguably correlated with two important aspects: (1) the need to accommodate Newtonian ideas within the previously accepted scholastic style of writing natural philosophy; (2) subsequently, the need to establish a new style, distinctly Newtonian, that can become a new norm in the discipline. This twofold process entails that, at their first appearance, Newton's own ideas and way of writing were idiosyncratic, and the success of their dissemination largely depended on the efforts of subsequent generations of natural philosophers in mediating this originality with the received standards of style and writing already established in the discipline.

4. The semantic multiplex

4.1 Method and Tools

Within the semantic multiplex, we investigate the books in our corpus in a fine-grained fashion, making use of only some of the semantic information available. This is done by comparing all pairs of books to each other with respect to one single keyword, which is deemed to be important to the field. Keywords are derived from the computational analysis of topics in the whole corpus, which constituted a key component of the textual multiplex described previously (§ 3). This process is repeated for 20–30 keywords (depending on the number of keywords that were extracted in the topics) and each of these keywords generates a layer within the semantic multiplex. Since the tools available for similarity analysis that allow for multilingual analysis are not robust for application on these smaller areas of investigation, the only corpora investigated in this layer are the monolingual, transnational sub-corpora.

There are two avenues available for generating similarity scores between books indexed on specific keywords. The first and more common method is by generating vector models (or word-embeddings) of the investigated keyword by generating scores between each wordtype (a word forming a distinct item in a vocabulary) in the investigated text and the keyword, based on the relative frequency of their distribution in each other's vicinity. Each word then receives a vector representation in an n-dimensional space, where *n* stands for the total number of wordtypes in all of the texts in the corpus. The vectors can be compared by using a closeness metric. A second method (derived from the former) is called collocate analysis (CA), which extracts the highest scoring wordtypes in the n-long vector to provide a list of collocates; this list summarizes the most salient semantic connections the wordtype makes in the corpus (Brezina *et al.* 2015). It is this second method that will be used to construct the semantic multiplex. The advantages of this are twofold: (i) the method picks up on more salient aspects of a term, in contrast to the more general linguistic background; and (ii) it provides a potential way to bridge the multilingualism gap in this method by hand.

Collocate analysis does not make use of previously learned models. Although it depends on a process of word-embedding, the WEs used for these layers are generated wholly based on the corpus itself. Using a pre-learned model has many advantages. A pre-learned model (such as the one described in the textual similarity word-embedding layer; see above, § 3.2) for example, can capture multilingual data. In addition, it encodes distances between words that are commonly used in similar ways within the training corpus. That is to say, the model takes the already learned synonyms in the corpus that are used for constructing the model into account, in order to construct similarity scores between texts using these synonyms. Depending on the context of usage, this can be an advantage or

disadvantage. The corpus on which the model is trained is much larger than the corpus investigated in this paper, so it is able to generate more accurate predictions.

However, the goal of the semantic analysis of individual concepts-keywords is that of investigating highly specific, technical vocabulary that can be expected to deviate in its usage and contexts from the corpus used for generating the model. In this case, relying on pre-learned models might introduce infelicitous accounts of the semantic distance between words that do not hold for the corpus under investigation here. Since within this semantic multiplex each layer is generated per keyword, where each of these keywords is one of the central topics to natural philosophy, in this layer the models will be generated from scratch based only on our own corpus.

In order to compute similarity scores, we assign to every work in one of the corpora (one for every language) a sparse vector representation of the distributional features of the word we are investigating. Between each pair of words, a value – the 'pointwise mutual information,' or PMI – is calculated, following similar applications in Brezina *et al.* 2015 and de Bolla *et al.* 2019.[7] From this sparse vector representation, we extract a list of collocates based on the PMI threshold value. All word combinations that meet this threshold can be assumed to be saliently connected with one another. We then calculate the similarity score obtained between each pair of works based on the overlap between their collocate lists, and we generate a network using these scores as edge weights and the works as nodes. The scores signify how similar the two works are in regard to their salient usage of the investigated word.

The (monolingual) layers of the semantic multiplex are constructed by defining the edges as the similarity score (normalized between 0 and 1) and generating a similarity score for each pair of works, and for each monolingual corpus. This process is repeated for each of the keywords. Each of these layers will be of interest, as they split up the corpus in different ways for different words. Nonetheless, we also introduce a 'summary layer,' which provides an average of the similarity scores for each of the keywords, in addition to an overview of the amount of divergence in these results.

7 PMI is a normalization procedure which ensures that pairs of words that are highly frequent in a text are not scored disproportionately highly merely due to their frequency. This is done by dividing the actual chance of finding a word x in the windows surrounding word y by the chance of finding word x around word y on the assumption of a random distribution of words. The higher the found chance than the base chance, the higher the PMI score.

This summary layer does not replace the analysis of the individual word-based layers, but rather complements it by offering a more distant perspective. If (and insofar as) the similarity score in the summary layer is similar to the similarity score of various individual sublayers, this points to the fact that the works are consistently similar in how they use multiple keywords. The greater the divergence between the summary layer score and the individual layers, the more idiosyncratic are the connections between the works based on their use of different keywords. We can then use individual layers to provide a sort of 'close observation' of how a single keyword is used within the corpus, and use summary layers to derive a more 'distant observation' about average features of the overall subcorpora written in the same language.

Similarly to the scripts created for topic modelling, tf-idf, and multilingual word-embeddings, the scripts for the semantic multiplex use open-source Python packages: the CLTK and NLTK packages for lemmatization and preprocessing of the texts and NetworkX for the generation of networks. Besides this, the algorithms have been implemented making use of the basic functionalities of the Python programming language.

4.2 From Sparse Vectors to Collocation Analysis

Vectors of *n* length, where *n* is the number of wordtypes in the entire corpus, provide a way to model certain semantic features of a keyword via its contexts. We take into account two key values: (i) the number of times the investigated wordtype has occurred in windows (in our case, we used a windowsize of 12) surrounding all instances of the keyword; and (ii) a normalization of these counts using an appropriate normalization algorithm (PMI). Scores are obtained for every other wordtype for the investigated keyword. We can thus obtain a vector that, in total, defines the semantic features of the keyword and its contexts by taking all of these scores together. Each of the individual values in the vector has a different meaning: when high, it signifies a strong and salient connection between the keyword and the wordtype associated with the value; when low, it signifies no connection, or a very weak one.

From these vectors of PMI scores, one can extract all the scores that exceed a certain given threshold. It is via this threshold that we get at the 'collocates' of the word; a word is thereby characterized by a list of the words that it is most strongly connected to. In this construction we used a threshold of 5; the higher the threshold, the shorter the derived lists of collocates become, and vice versa. The threshold has been set relatively low in comparison to other studies (Brezina *et al.* 2015), since when investigating overlap, the results become more stable when the lists of collocates are on average longer.

These lists of collocates can be compared to one another in order to observe how similar the words are from the point of view of their saliently connected

wordtypes. For this comparison we use the *Jaccard-Index* (Leydesdorff 2008). The Jaccard-Index defines the similarity of two unordered collections by dividing the size of the overlap between the two collections by the size of the union of the two collections.[8]

Figure 4 provides a visualization of the networks produced by this method with respect to the keywords 'corpus' (similar results have been generated for all other keywords):

Every set of works receives a connection, but not every connection is of similar strength. We find, for example, that Musschenbroek's 1726 *Epitome Elementorum Physico Mathematicorum* and Leclerc's[9] 1722 *Opera Philosophica Vol. 4* (on physics) are relatively poorly connected works that are very strongly connected to each other. This signifies that van Musschenbroek and Leclerc's works use *'corpus'* in a way that is non-standard within this corpus of natural philosophy, but this distinct usage is very similar in the two author's works. Similar analyses can be extracted for other groups in the corpus. Figures 5 and 6 below show the results for *'corps'* and 'body' used as keywords in the transnational French and English corpora, respectively.

4.3 Summary layers

Each of the three monolingual layers consists of 20–30 sublayers (one for each keyword, as above), and we can generate summary layers for each of them. A common way of doing this is by creating a multi-relational graph, where multiple edges can exist between two nodes, and each of these edges is labelled with the relation it denotes. The multi-relational graph, however, remains difficult to parse for a human investigator. Therefore, we use a more reductive method for achieving a summary layer in this case. Our summary layers are generated by averaging over each of the edge scores and taking the average score for this edge. These averaged layers are thus representative of the underlying wordtype specific layers. Figures 7, 8, and 9 illustrate the results of the summary layers of the Latin, French and English corpora.

Many of the relations we have seen for the wordtype-specific results resurface in the summarized results. For example, van Musschenbroek and LeClerc were identified in Figure 4 as works which are relative outliers in the network, but are strongly connected with each other, when their relation is indexed on the word

8 The scores are multiplied by ~2500 to generate more easily readable numbers. This aestheticization leaves the ordering and relative distances between score-pairs (the relevant properties of the scores) fully intact, but means that the results should not be read as 'percentage points' where a score of a hundred would indicate perfect similarity.

9 Jean Le Clerc (1657–1736).

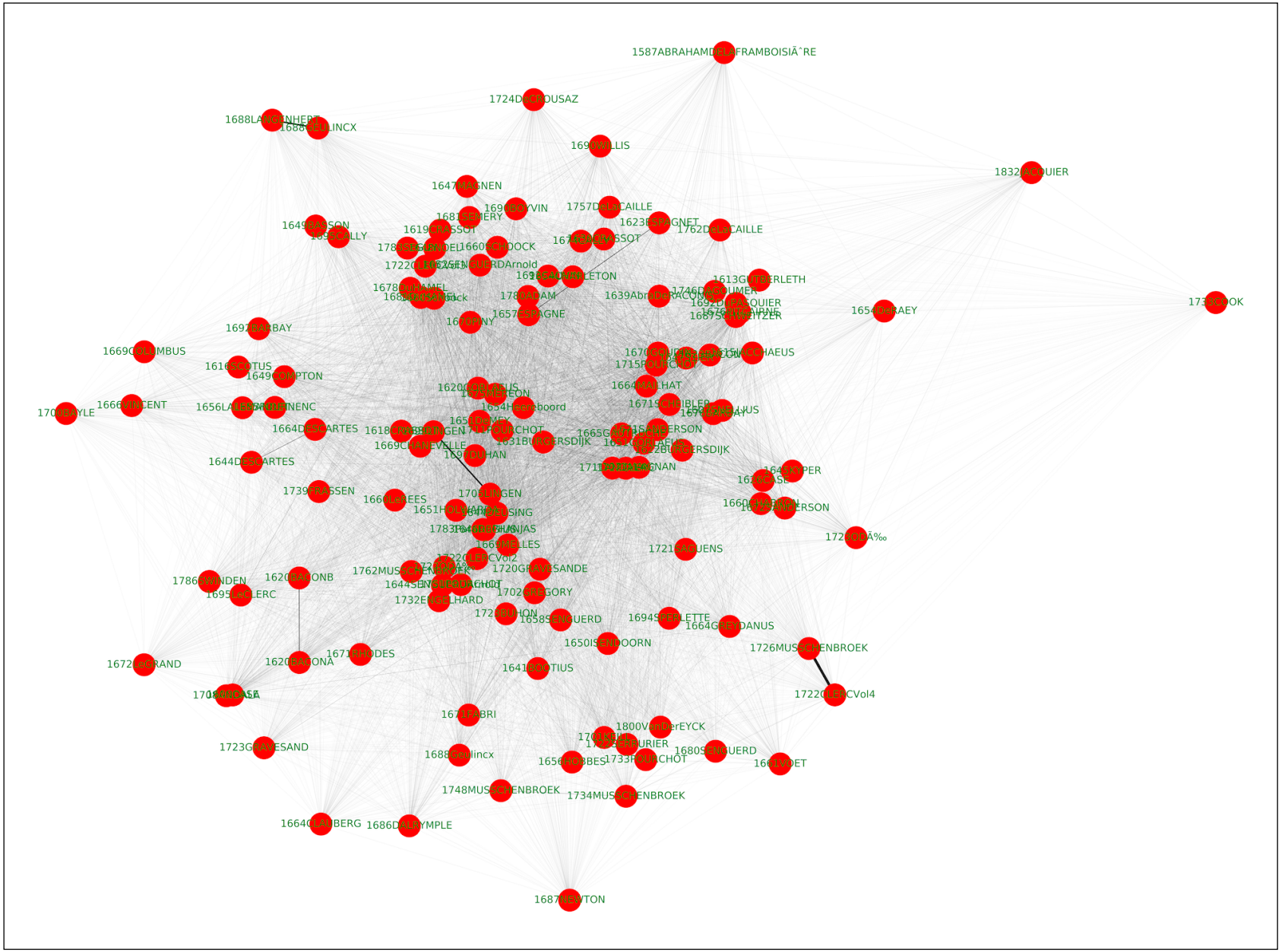

Fig. 4 Network representation of the word 'corpus' in the transnational Latin corpus using a windowsize of 12.

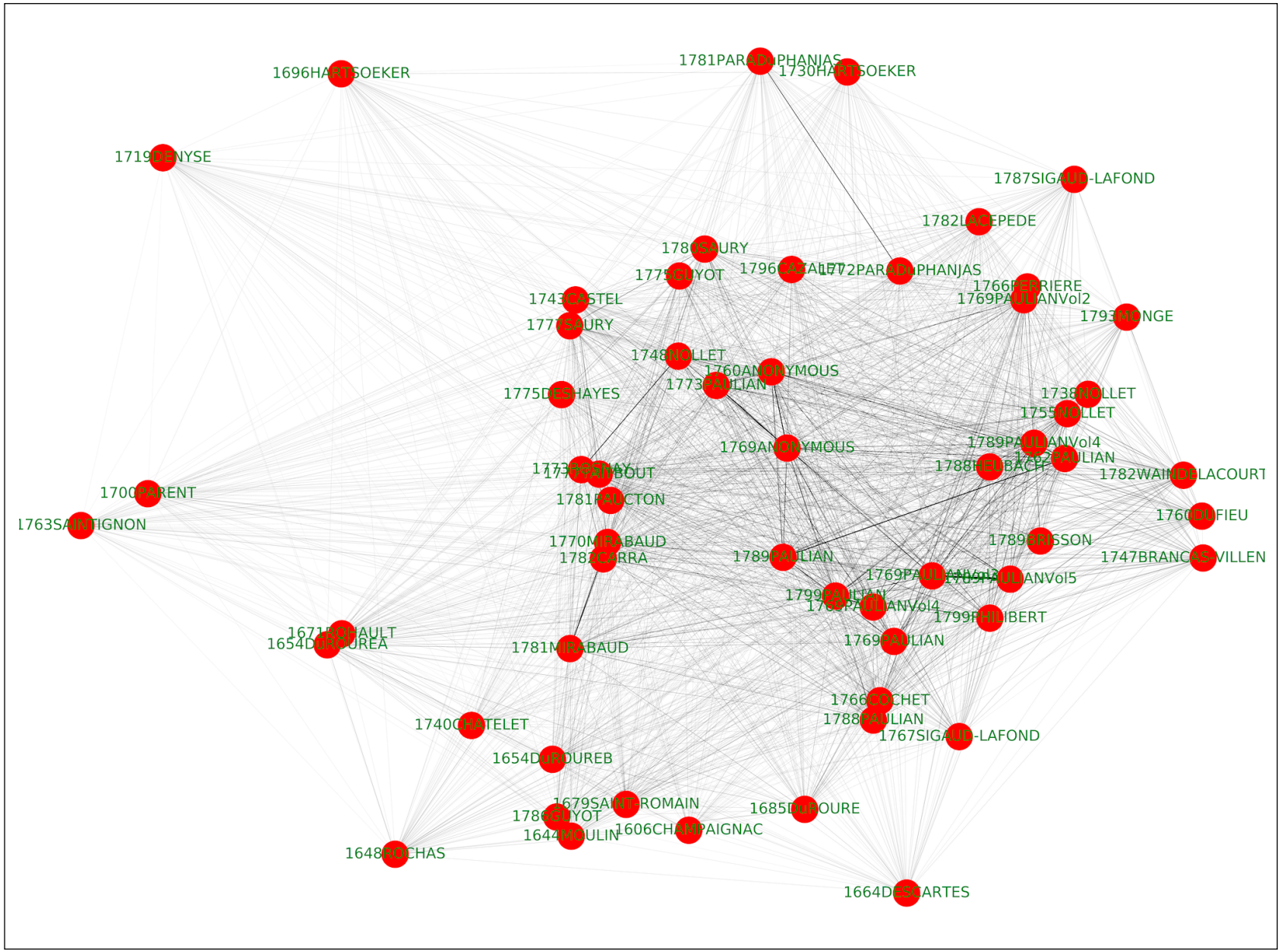

Fig. 5 Network representation of the word 'corps' in the transnational French corpus using a windowsize of 12.

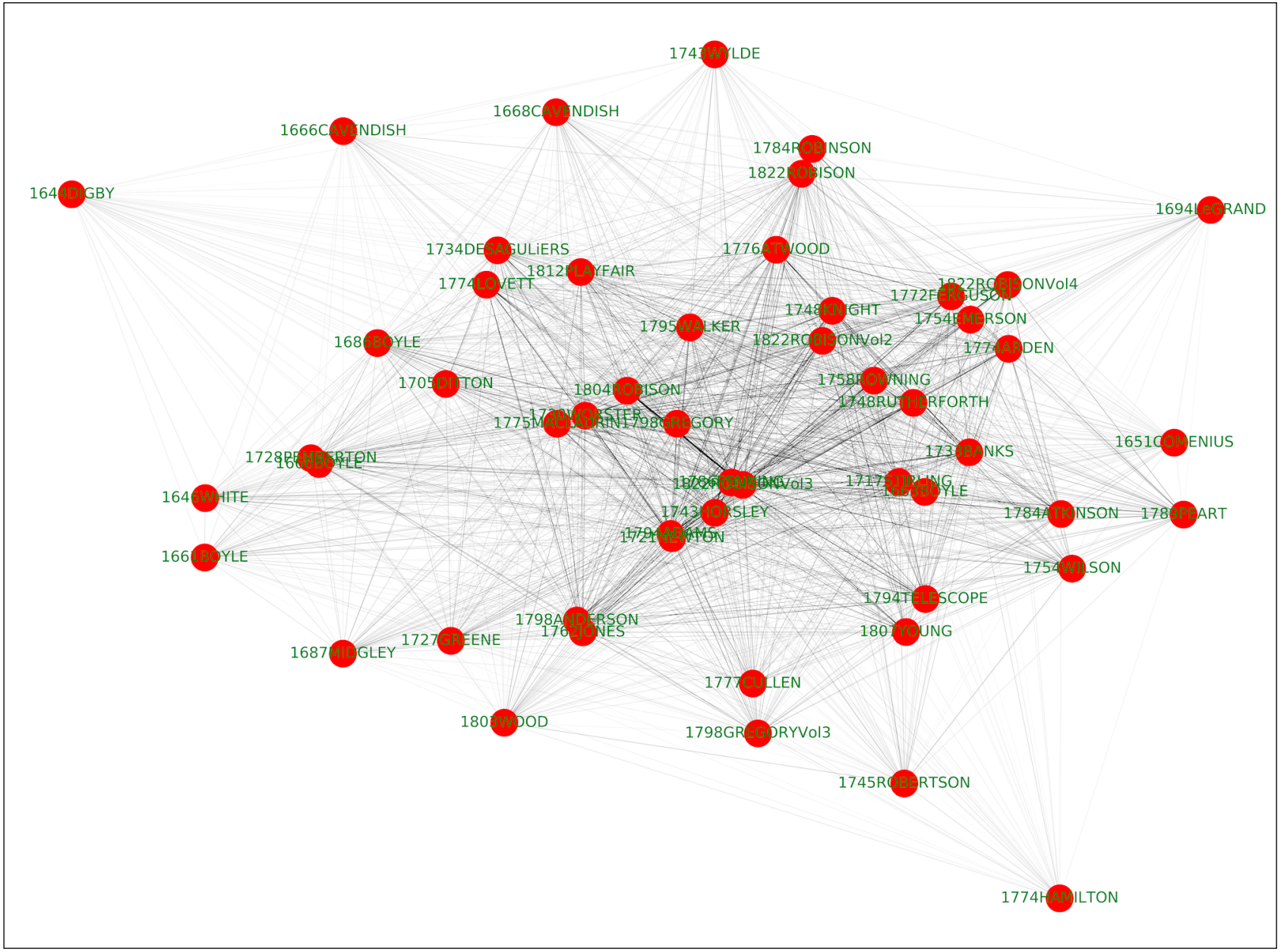

Fig. 6 Network representation of the word 'body' in the transnational English corpus using a windowsize of 12.

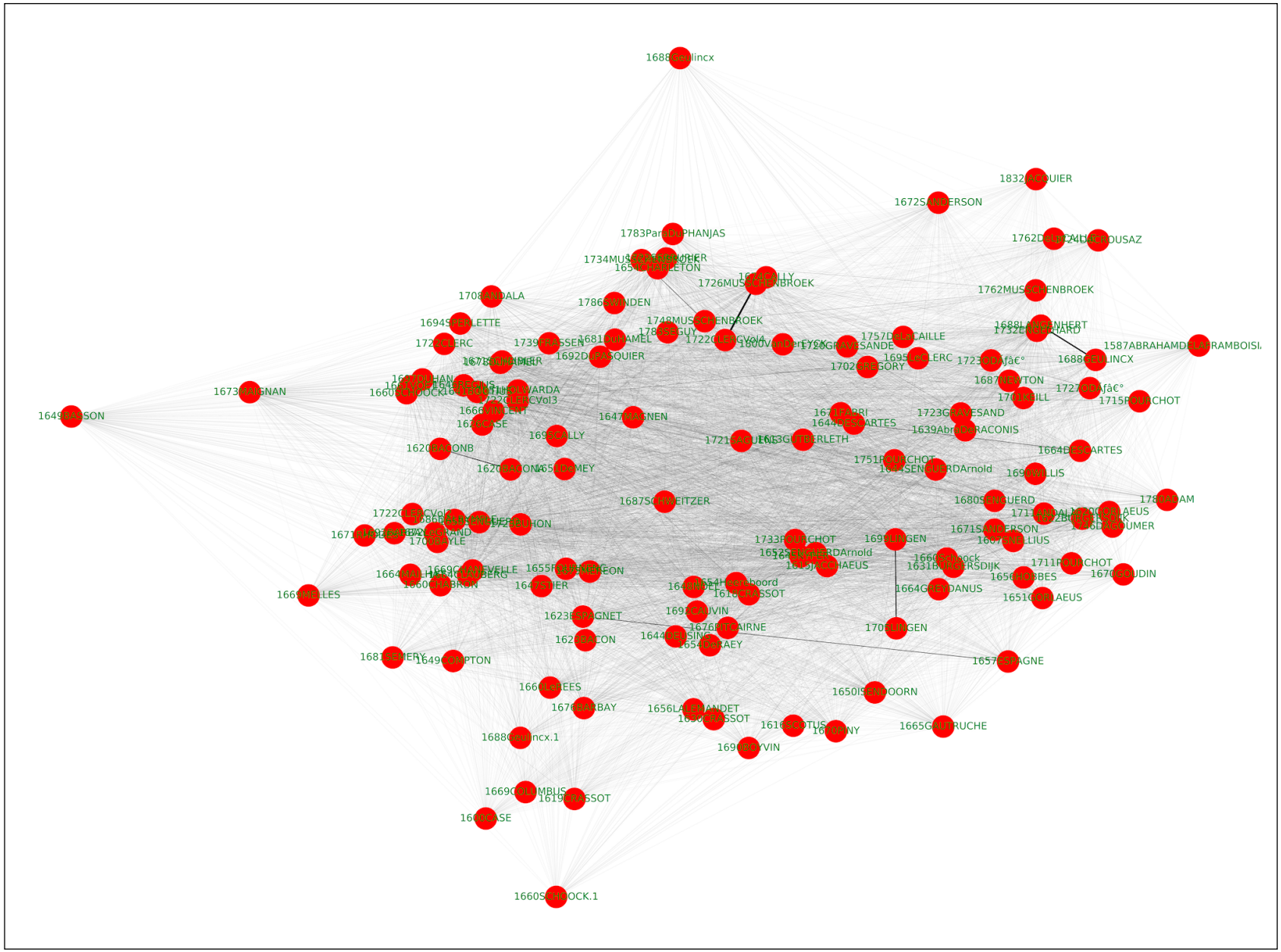

Fig. 7 Network representation of all of the investigated words in the Latin transnational corpus.

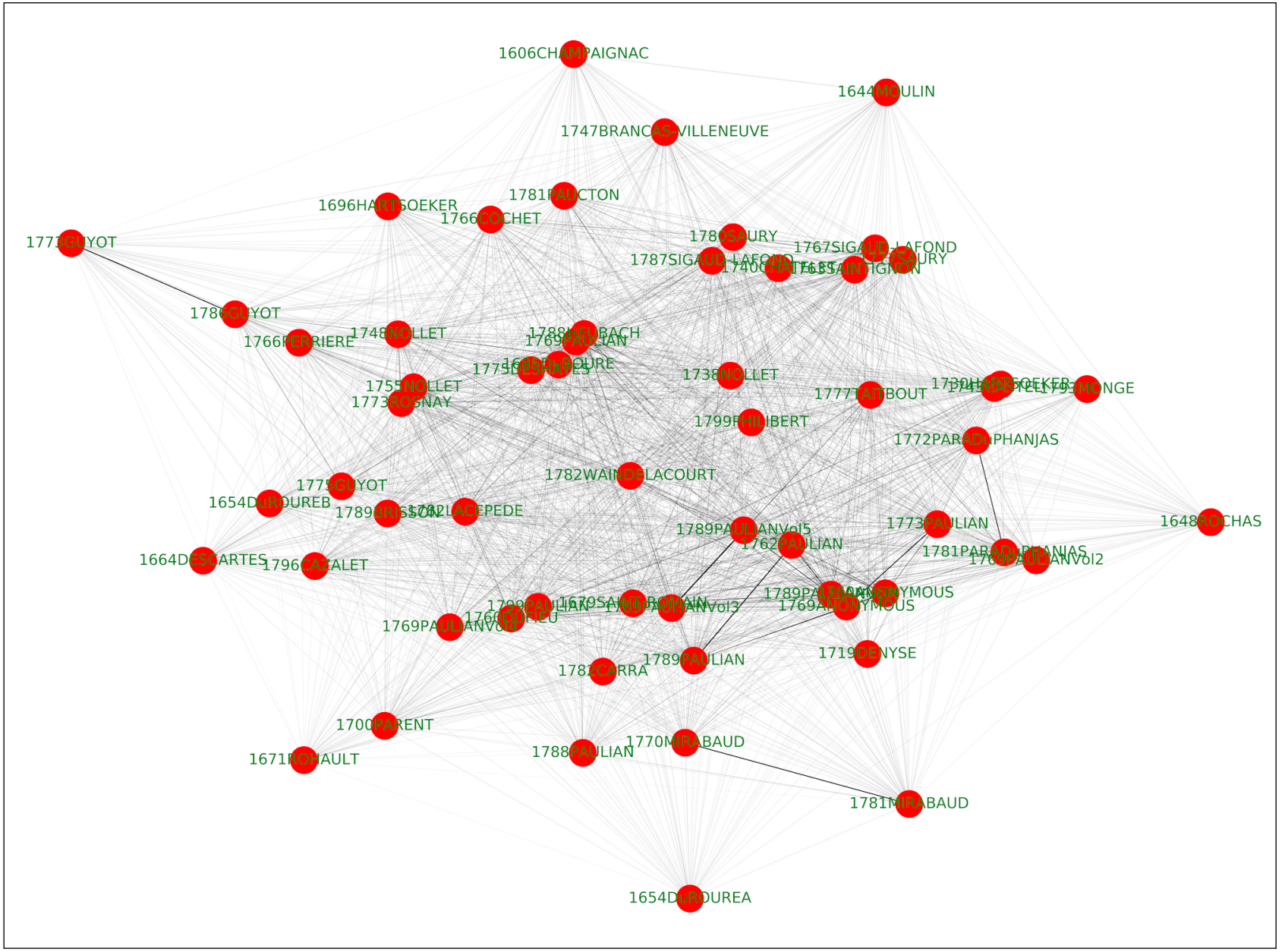

Fig. 8 Network representation of all of the investigated words in the French transnational corpus.

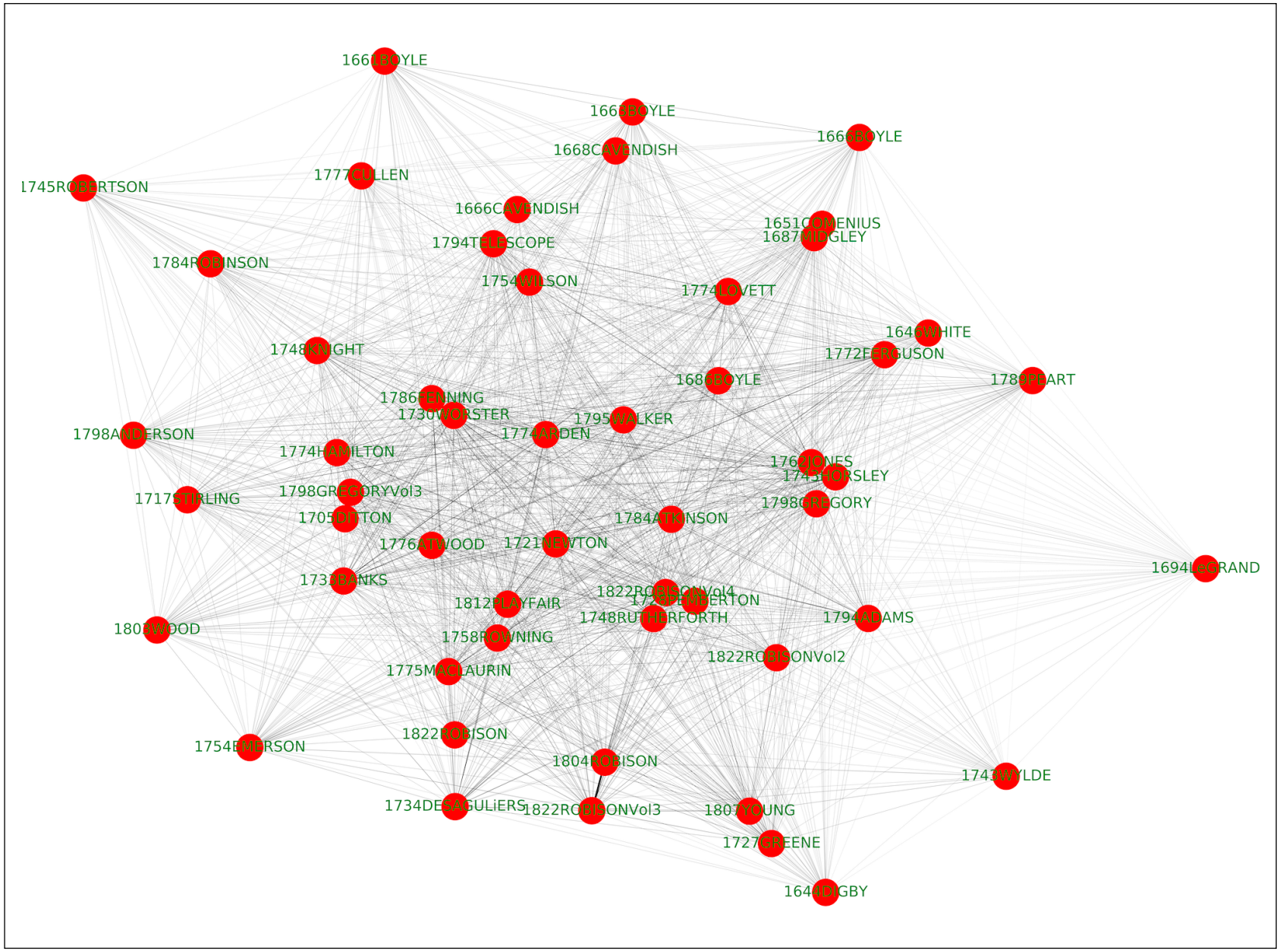

Fig. 9 Network representation of all of the investigated words in the English transnational corpus

corpus. This relation is retained in the summary layer, as we can see at the top of Figure 7, although they are now also placed in a tightly knit group of other works. However, not all of the results previously found are recoverable in the summary layer. In the English results indexed on 'body' (Figure 6), we find that Wood's 1803 *The Principles of Mechanics* and Atwood's 1776 *Description of the experiments* are strongly connected and placed in proximity (the tie strength found is 64, which is significantly higher than the average tie strength for the English corpus of 26.66), whereas a significantly reduced connection and proximity occurs in the summarized English corpus (the tie strength found is 35.09).

The advantage of the summary layer is that analysis is easier, since less data needs to be investigated by the researcher. However, this is based on the assumption that the summarized scores represent their underlying word-indexed scores. The more extreme scores there are in the word-indexed results, the less representative the average score is of the underlying results. This means that the more diverse and heterogeneous the individual layers are, the less representative the summary layer will be. An indicator of how well the summary layer represents the underlying results is the average tie strength and the amount of divergence between this average and the population results. Each of the word-indexed layers will have an average tie strength, as well as the average divergence from the average tie strength by all the connections in the layer. We take the average of these divergence values for the word-indexed layers as providing an average divergence across the corpus. This value is indicative of how well a summary-layer represents the underlying results. By comparing the divergence to the average tie strength, we can see whether the divergence is relatively high. For example, if the average tie strength turns out to be 100, and the divergence 2, then there is relatively little divergence between the different keywords, which means the keywords can be well represented by the summary layer. However, the values we found tell a different story. Consider Table 6.

In the English and French subcorpora, the average divergence is higher than the average tie strength. In the Latin subcorpus, the average divergence is half that of the average tie strength. In both these cases, the average divergence is high. The average tie strength does not allow us to easily extrapolate conclusions regarding the heights of the underlying scores of the word-indexed sub-layers. This is because the scores found here are wildly divergent. However, this tells us that a lot of information is contained in the way the underlying word-indexed

	Latin	**English**	**French**
Average strength	38.37	26.66	19.64
Average divergence	20.89	36.75	34.54

Tab. 6 Tie strength divergence and averages in the three languages

layers differ from one another and, in effect, in the ways the corpus gets split up based on particular keywords. This means that meaningful results can be derived from the analysis of word-based layers and that results based on a summary layer need to be carefully qualified.

To put it in other terms, high similarity scores with low divergence across layers translates more accurately into a single layer net, whereas low similarity scores do not. This means that the greater the similarity in the corpus, the greater the chance of collapsing the multiplex into a single-layer network. While this remains a theoretical option available from the point of view of the method we employed, our particular corpus does not warrant this reduction. In this respect, the use of the summary layer also serves as a further justification for the need to represent the semantic dimension of the corpus as a multiplex, rather than a plain graph.

Besides the generally high divergence, what is also interesting to note is that the Latin corpus scores quite differently from the two other corpora. It scores both lower on the average divergence and significantly higher on the average tie strength. A higher average tie strength signals that within the Latin subcorpus, works are generally better connected. This in turn tells us that the Latin corpus contains works that, generally, are more similar in their use of terminology, leading to higher connection scores between works. The Latin corpus is generally more stable in its use of terminology overall, whereas the other corpora are more turbulent and dissimilar to each other in terms of how their technical terms are used. The lower average divergence of the Latin corpus tells us that the spread of scores is more centered around the average. In addition to the high average telling us that the Latin works are relatively similar, the low divergence also tells us to expect less, and less extreme, outliers. In general, the Latin corpus shows a higher level of homogeneity than the French and English corpora do. However, this observation is only an average observation. A more fine-grained analysis of specific keywords and philosophically unified subsets of the corpora reveals that, in various particular cases, vernacular sub-corpora happen to show greater homogeneity.

To provide an example of this fact, we consider the relation between four works by four British authors: Newton, Hamilton, MacLaurin, and Robison (the last three of whom are known as Newtonians). Tables 7 and 8 show the average strength between these four English Newtonian authors and the same average strength in the use of the same keyword, *body*.

The four English authors are on average connected significantly more strongly than the entire English corpus (78.91 vs 26.65, respectively). The higher average connection than the rest of their linguistic corpora does not translate to a higher connection on their use of the keyword *body*. In this case, their average connection is slightly below the sub corpus' average connection for 'body' (38 instead of

	1721_Newton	1774_Hamilton	1775_MacLaurin	1804_Robison
1721_Newton *Opticks*	X	94.2	86.97	56.33
1774_Hamilton *Four Introductory Lectures*	94.2	X	77	50.33
1775_MacLaurin *Newton's Philosophical Discoveries*	86.97	77	X	111.63
1804_Robison *Elements of Mechanical Philosophy*	56.33	50.33	111.63	X

Tab. 7 Tie strength for the averaged results on the four Newtonian authors to one another.

	1721_Newton	1774_Hamilton	1775_MacLaurin	1804_Robison
1721_Newton *Opticks*	X	15	73	71
1774_Hamilton *Four Introductory Lectures*	15	X	16	4
1775_MacLaurin *Newton's Philosophical Discoveries*	73	16	X	49
1804_Robison *Elements of Mechanical Philosophy*	71	4	49	X

Tab. 8 Tie strength for 'body' on the four Newtonian authors to one another

39.53). This suggests that the English Newtonians are relatively consistent among each other in their use of this particular keyword, and this consistency does not seem to deviate significantly from the average consistency that we can find between their works by considering multiple keywords.

The data concerning the English Newtonians also single out Hamilton as a different case. In the multilingual multiplex, we noted the same eccentricity in relation to Hamilton: out of the 10 most connected nodes, Hamilton established the strongest connections with four Newtonian authors (Wilson, Worster, Wood, and MacLaurin). The scores relative to *body* (Table 8) between Hamilton and the other three works are significantly lower than the connections amongst themselves, which are on average higher than the scores for *body* in the entire English sub-corpus. This suggests that Hamilton's text (although strongly connected to Newtonian texts, as can be seen from Table 5 in section 3.3) might be less distinctively Newtonian in his use of the term *body*. In fact, Hamilton is not usually presented as a leading Newtonian. Our network representation supports this observation and corroborates the findings of the multilingual multiplex, in which Hamilton scores high in betweenness centrality, but extremely low in eigenvector centrality. Hamilton uses the term *body* in a relatively less standardized way when compared to the other well-known Newtonian writers. In the case of English-writing authors, then, our initial hypothesis (at the end of § 3) may be partially confirmed. A greater sharing of Newtonian orientation among English-writing authors might account for the relatively greater homogeneity in the use of the keyword *body* in the English corpus, whereas Hamilton may stand out, via further qualitative research, as a less standard Newtonian, as his relatively high betweenness centrality in the multilingual multiplex suggests.

We can derive three conclusions from this comparison. First, our method corroborates our expectation that works written by authors who we assessed to belong to a similar philosophical orientation usually reveal a stronger average connection among themselves than with others, as the results from Table 7 compared to the average tie strength in the entire corpus show. Second, this stronger connection does not exclude a degree of difference and diversity, which is variable. This variability depends on the specific relation that a philosophical orientation maintains to the specific keyword investigated. In some cases (Hamilton, for instance), we conjecture that there are Newtonians that adhere to a less standard use of language, and thus potentially offer a more independent or original (or perhaps just a more hybrid and less normalized) approach, which would be worth exploring further to form a qualitative point of view. Thirdly, *one* possible reason for the greater homogeneity in the use of particular keywords in vernacular languages can be located in the presence of a more dominant and cohesive philosophical orientation among multiple authors who write in the same vernacular language.

5. Adding Social Depth

In this section we would like to suggest a way to deepen our corpus representation and analysis by integrating it within a further social multiplex. The corpus under scrutiny is mostly composed of books. Each book has a corresponding author, and for some of these authors we can recover biographical information based on existing scholarship (mostly the bio-bibliographical *Dictionaries* of seventeenth- and eighteenth-century philosophers used to compile the corpus itself (Sangiacomo *et al.* ibid.), or other available online resources such as *WorldCat* and *Google Books*. The reason why our approach becomes more tentative here is due to the current difficulty in gathering sufficient social data.

In terms of the social profile of the authors of the books in our corpus, the most easily available information is concerned with their academic affiliations and the scientific institutions with which they were associated. Further socially relevant information can also be derived by considering information included in the books themselves, like their publishers and place of publication, which can be used to reconstruct potential scenarios for interactions among authors. The ensuing social multiplex thus offers a glimpse of what Valleriani *et al.* (2019) call 'epistemic communities' that shape the corpus, and which we use to refer to spaces of production of scientific knowledge.

However, as noted in the introduction of this paper, our current information about the authors is fragmentary. Especially for those names not featured in the existing *Dictionaries,* the author information needs thorough cleaning, checking and normalization before we can start to consider using other biographical sources. Titles that were scraped from *Worldcat* often index a plurality of names as 'author', which can include (but is not limited to) illustrators, editors, and publishers as well as libraries, which are sometimes stated as such, but are most often not[10]. Each entry has to be examined before we can even have an author list from this corpus section to begin with.

The following discussion of the social multiplex will thus be illustrative, and will be based only on a subset of the corpus that is directly covered by the *Dictionaries,* because these currently offer more reliable and consistent bio-bibliographical information available. Given this restriction, the total number of authors considered in the social multiplex is 142. The number of their publications is 196.

---

10 Examples include, for one title, the list of authors: *"Seguy, Antoine; professeur de philosophie); Paul-Denis Brocas; Pierre-Théophile Barrois; Joseph-Gérard Barbou"* (French corpus) and *"Robison, John; John Thompson Exley; University of Bristol. Library. Exley Bequest.; University of Bristol. Library. Exley Collection.; University of Bristol. Library. John Thompson Exley Bequest."* (English corpus).

For each sublayer, ties are formed based on the authors' shared circles, using indirect social cues to determine social similarity. The binding elements throughout the social multiplex are thus an attempt at outlining how authors included in our departing catalogue might have belonged to the same 'knowledge environments.' In this preliminary step, time and duration remains outside the scope of our inquiry, so two authors being tied does *not* mean they were physically connected (although in many cases this is possible), but that they shared the same knowledge environment. Adding diachronic dimensions should help to further ascertain whether the authors might have had actual historical relationships. However, adding this diachronic dimension at this stage (in which our access to social data is still relatively scarce) risks making the networks too small and sparce to be representative, while taking a more distant perspective focused on knowledge environments allows us a glimpse into the sort of territories that most authors gravitated towards throughout the time period covered by our corpus.

The relation between the social multiplex and the other two multiplex networks consists of the fact that nodes in the social multiplex are directly related (via authorship) to the books in the other two multiplexes. As a minimalist proposal, we thus suggest a social multiplex made of four main layers:

1) *Affiliation.* This layer connects authors on their professional working environment. It shows a tie when authors share a working environment that is important for knowledge exchange. We consider several affiliations that can be regarded as meaning they belonged to a knowledge environment. Most obviously and most frequently occurring are universities and schools, but we also include learned societies (such as the Royal Society of London) and other places of knowledge exchange, like the Royal Court of The Hague. We include post- and pre-graduate affiliations, since these are both important in the shaping of ideas and are often overlapping (i.e., an author has a position while also studying).
2) *Place of Affiliation.* This layer reflects co-affiliation based on spatial coordinates; the connectivity among the authors based on the location (city) of the institution they were affiliated with. This is a deepening of the first layer that will disclose more connections, since one city can, and often has, more than one institute. To look at this separately makes sense, considering that geographical proximity allows for easier interaction.
3) *Publisher.* This sub-layer connects authors based on a shared publisher, as another indication of a shared knowledge environment. With their influence on the composition of books, publishers can play an important role in the circulation of knowledge, sometimes more so than the authors themselves (Valleriani *et al.* ibid., 81). Self-publishing authors and authors whose publishers are unknown remain unconnected.
4) *Place of publication.* This layer, like affiliation place, shows geographical proximity, now based on the location where the author had their works published.

Our approach involves two steps. The first step consists of building affiliation networks with two node types. This is a preliminary stage in which the resulting networks consists, for instance, of both authors and institutions, or authors and publishers. While visually revealing certain principal structural properties, the bipartite network is unfit for most analysis tools because of the two different node types. While workarounds are possible (see i.e., Borgatti 2009, Bonacich 1991), our second steps consisted of connecting the authors by deriving the so-called co-affiliation networks, in which case an author is connected to another if they share a common attribute, such as when they taught at the same university, or published with the same publisher. From the original bipartite network (author-to-affiliation) we derive the projected one-mode (actor-to-actor) networks, which are suitable for regular network analysis and help us to understand any existing tie pattern (Borgatti and Halgin 2011, 420). The networks have been built, visualized and analyzed in Python's NetworkX, unless stated otherwise. Figure 10 illustrates the four sublayers in the social multiplex. The authors have ties to 144 affiliations in 96 different places. Using the publication information, we listed 134 publishers located in 34 different locations. Each of the constructed networks contain all authors. However, some authors remain entirely disconnected, either because they had no affiliation, or an unknown affiliation or publisher.

Although illegible on the node level, we can easily see that each layer has its own distinct shape and structure. Layer 1 (affiliation) and 2 (affiliation place) stand out for having one giant, densely connected component. Layer 3 (publisher) shows few connections and can hardly be considered a cohesive structure at all. Layer 4 (publication place) is more spread out, with a few smaller components, but also one large, connected component. Within the biggest connected components of layers 1, 2 and 4, we find a core of high degree nodes and a periphery of lower degree nodes.

Looking at the affiliations layer, centrality measures (degree, betweenness and eigenvector) are relatively homogeneous. Two British authors top every score: Henry Pemberton (1694–1771) and George Gregory (1754–1808). Both authors were strongly involved with Newton's natural philosophy. Looking at the top 5, we see that the next most central authors concerning degree and eigenvector centrality are all Dutch: 's Gravesande, the first prominent Dutch Newtonian; Henricus Regius (1598–1679), an independent scholastic author influenced by Descartes; and Martin Schoock (1614–1669), a more traditionalist scholastic author opposing Descartes. Betweenness centrality instead has two French scholastic authors (Pierre Du Moulin, 1568–1658, and Philippe De La Très Sainte Trinité, 1603–1671) and a Dutch scholastic (Nicolaas Hartsoeker, 1656–1725). This suggests that Newtonian and Cartesian authors tend to be the most strongly and diversely connected, and the most similar in terms of their affiliations.

In general, we observe clustering by nationality, noting that French authors are the least densely connected because they are more spread over affiliations in

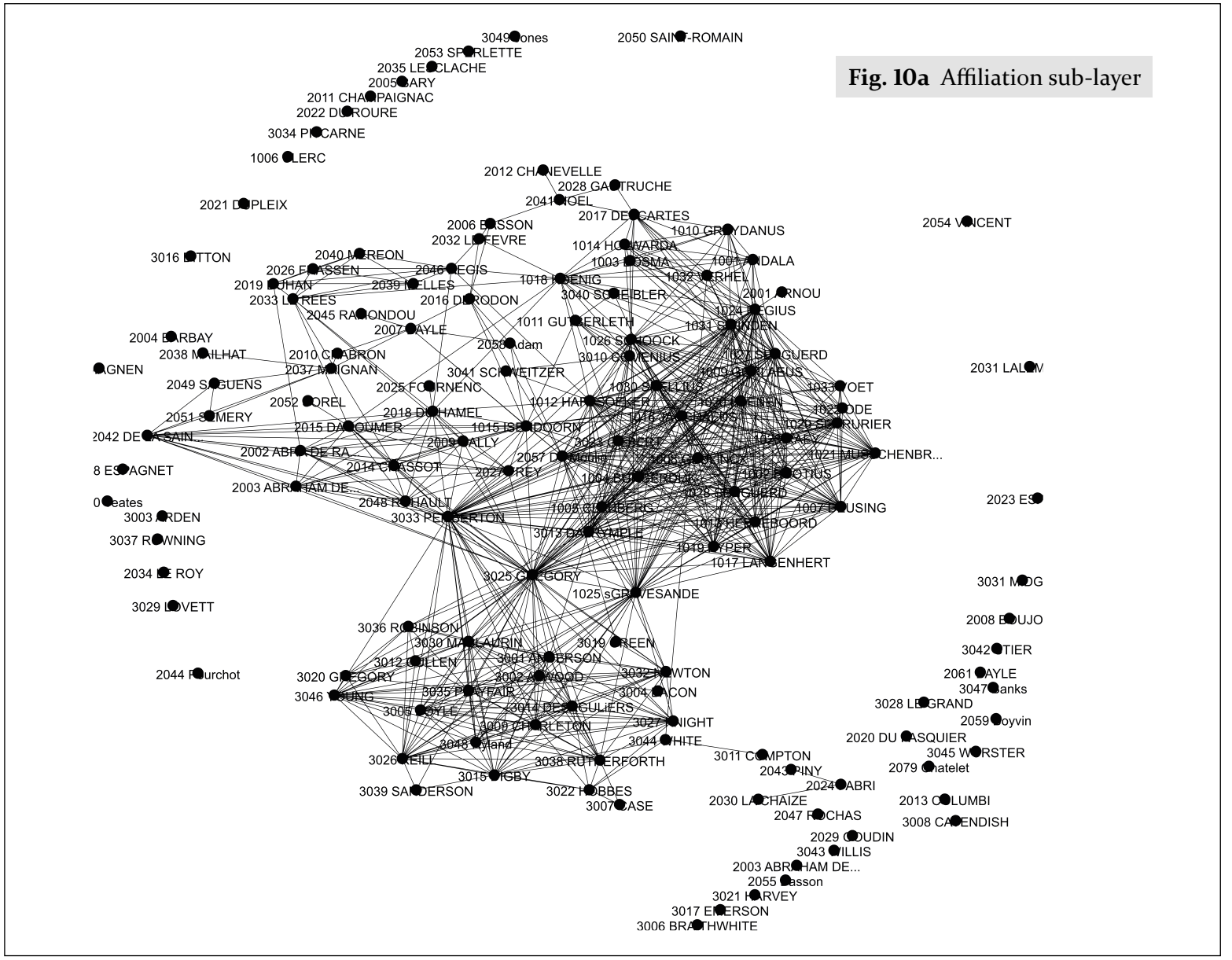

Fig. 10a–d Network representations of the four layers of the social multiplex (spring layout). These are the projected author-to-author graphs. The core of each of the layers shows the connected components, where nodes have ties to others. The unconnected nodes are floating in the periphery around the connected parts.

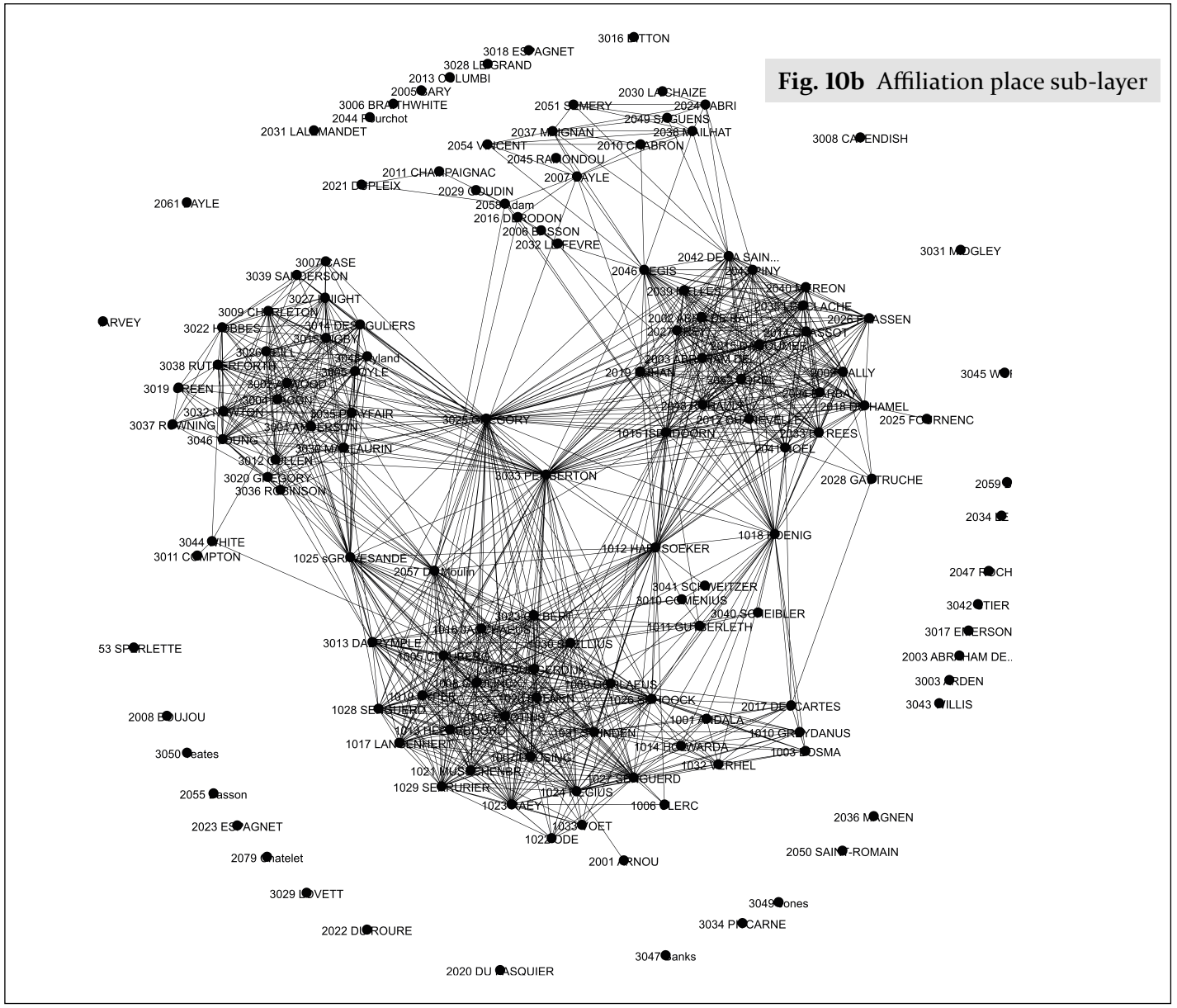

Fig. 10b Affiliation place sub-layer

eISSN: 2535-8863
DOI: 10.25517/jhnr.v7i1.129

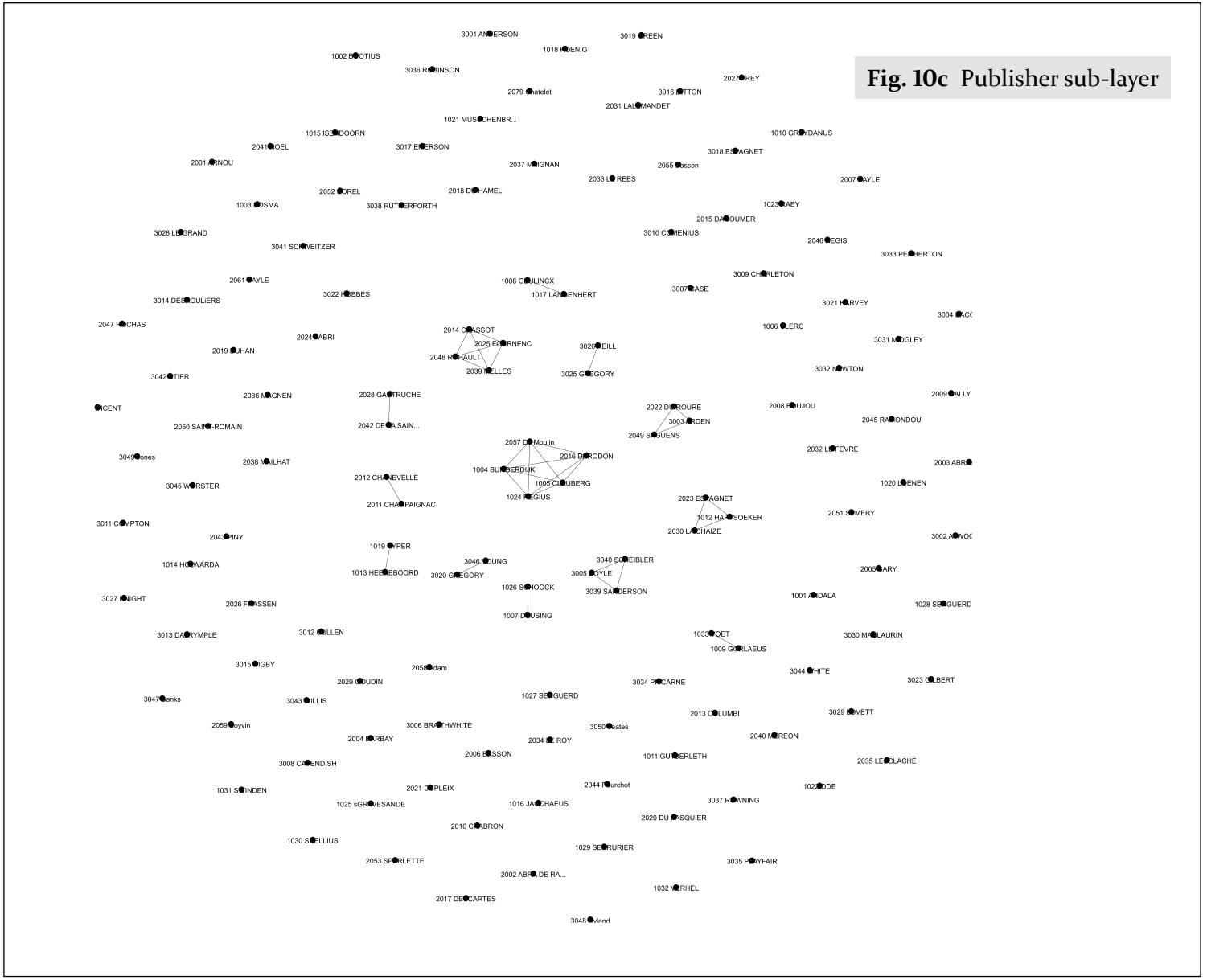

Fig. 10c Publisher sub-layer

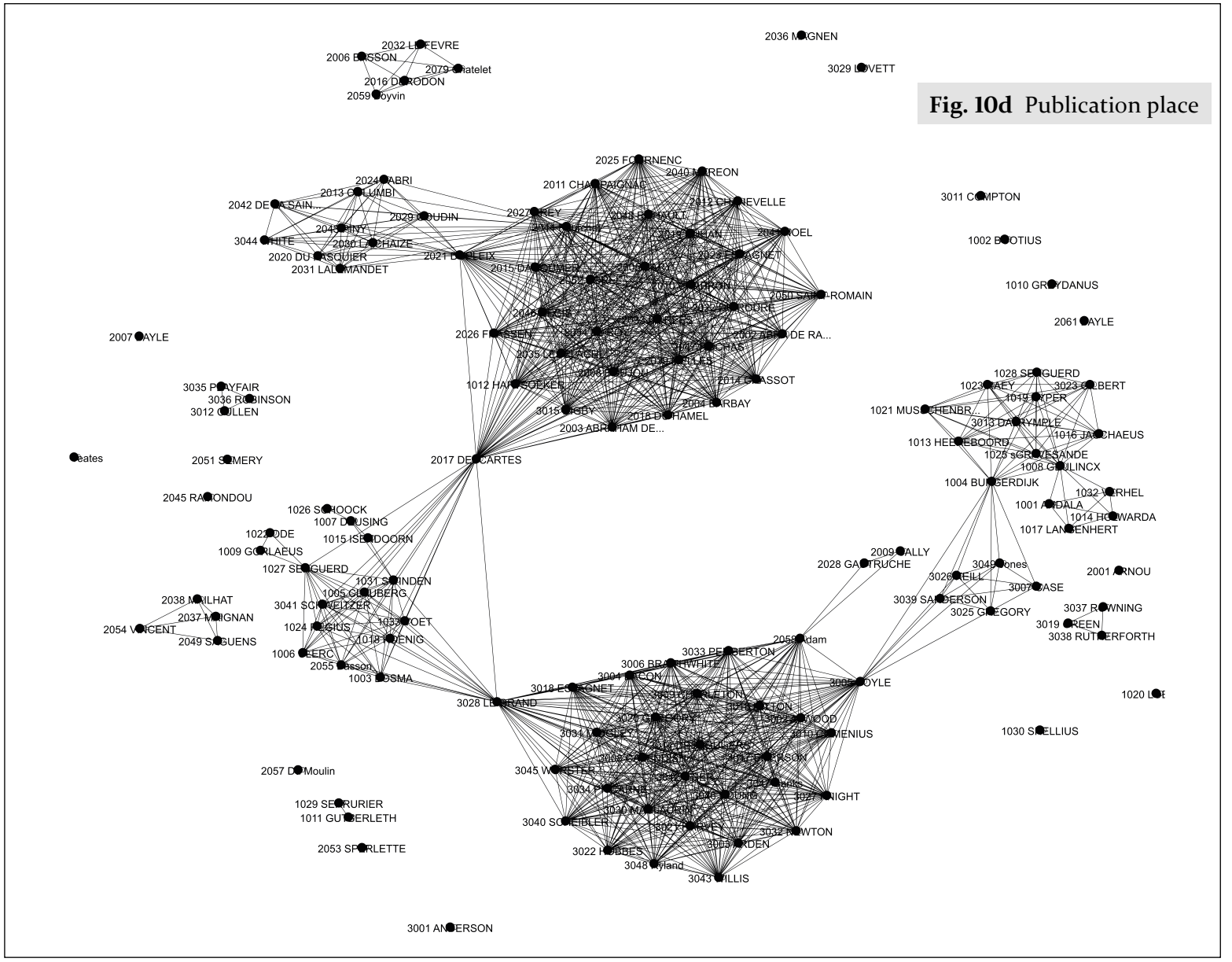

Fig. 10d Publication place

the country than the other two nationalities. Dutch affiliations show a bit more variety, since Dutch institutions host authors from various other backgrounds, including French and British. Authors of high betweenness are those that appear to be the bridges between the nationality clusters, such as Gregory and Pemberton (linking the three national areas together), and Du Moulin and Gisbertus Ab Isendoorn (1601–1657) (linking the French and Dutch areas). Gregory and Pemberton are remarkable since, given their important positions in their academic field, they might have been able to transfer knowledge from one geographical position to another. Moreover, since the same authors are also highly connected degree-wise, their influence could have been significant in the spread of Newtonian thought from the English to the French and Dutch contexts.

Layer 2 represents affiliation locations, and is similar in structure and characteristics to layer 1. However, layer 2 highlights 'propinquity,' or geographical nearness. To be geographically close makes it more likely that the authors would have been in actual contact – either physically in person, or that they were in contact with each other's ideas. While not all authors would have been able to physically interact (because of the large timeframe), their works would have been available to each other over the years. One thing these connections therefore denote is the opportunity to have access to the systematic sources of scientific knowledge that were the basis of learning and teaching natural philosophy at the time. The top 5 authors are almost the same as in layer 1. Still, there are also a few other authors that stand out, like the Dutch Newtonian Samuel Koenig (1712–1757), who is 5th highest in degree and betweenness, and who is the only Newtonian between the French and Dutch contexts without any ties to English authors. In terms of betweenness centrality, Pierre-Sylvain Régis (1632–1707) and François Bayle (1622–1709) are new on the list, and are interesting due to the fact that they are Cartesians in an otherwise mostly scholastic environment. Their high betweenness position, as with Koenig's high degree, could have fostered the dissemination of new thoughts into the more conservative environments. While the top scores for eigenvector are higher than in layer 1, they decrease at a higher rate, which means that having a high score carries a certain significance in this respect, with Gregory and Pemberton having the most important positions again.

Moving to layers 3 and 4, it should again be noted that our temporal perspective encompasses two centuries; the very sparse structure of layer 3 can thus be explained by the fact that, over time, authors and publishers associated differently, and mostly without creating large, enduring hubs. However, we do observe a few interesting exceptions. The two publishing houses that published the most works from our corpus are the Dutch Elzevier (five) and the French Denys Thierry (four). The Elzevier clique includes authors from the Netherlands and France, whose titles acknowledged either Aristotle, such as Frank Pieterzoon Burgerdijk (1590–1635) and Basson, or Descartes, such as Clauberg and Regius, but also Descartes himself, who published with Elzevier. This suggests that Elzevier played an important role in disseminating and to some extent foster-

ing a dialogue between late-scholastic and Cartesian thought. Further exploring the role of individual publishers in the future, including re-editions and reprints, could be illuminating.

Layer 4 offers a highly interesting and complementary perspective, illustrated below in Figure 11.

This is a dense layer, and the upshot that emerges from it is that there is a higher concentration of publishers than institutes in each city. Moreover, a few capital cities hold a monopoly of most of the publications, since they host most of the publishing houses. Layer 4 also emphasizes the bridging role of authors who publish in several of these clusters, and here the most important are Descartes himself, some Cartesians (such as Le Grand, who's active across France and Britain), and some later French scholastics, such as Scipion Dupleix (1569–1661) and Edmond Pourchot (1651–1734). This national aspect is reflected in the fact that the high eigenvector nodes are all authors that published in Paris, this being the cluster that holds the most central actors in all respects. Interestingly, we also observe two separate Dutch clusters, one for publishers in Amsterdam and one for Leiden-Franeker. It is notable that the Leiden cluster has no Cartesian works (these appear mostly in Franeker), suggesting the existence of segregated groups of authors based on their philosophical orientation.

6. Conclusion

The main result of our method is that we can now represent the starting corpus from the point of view of various (at least two, possibly three) multiplex networks, which are connected with one another by virtue of the fact that they are derived from the same entities (ultimately, the 239-book corpus). These multiplex networks witness the basic fact that the corpus is constituted of works dealing with the same subject: natural philosophy. This is expressed by the certain degree of homogeneity that we could observe in all three multiplex networks. Nonetheless, our method is also capable of distinguishing between various shades of difference between the authors and works included in the network. We could differentiate between the overall features of entire layers (e.g., the relatively greater degree of homogeneity in the use of vernacular languages when compared with Latin), the features of certain groups (e.g., the existence of a group of late-scholastic authors whose work exemplifies what an 'average' natural philosophy textbook might look like), and even features of the small group of individuals (e.g., the correlation and relative insulation of Newtonian natural philosophers who carve up their expanding niche), and of course the diachronic dimensions of the corpus (e.g., that they produce 'time-slices' which will offer a more temporally situated image of a subset of texts or authors). Despite the relative homogeneity, our method is thus also suitable to spot discontinuities and differences. These differences do not appear to be random, but usually correlate with known philosoph-

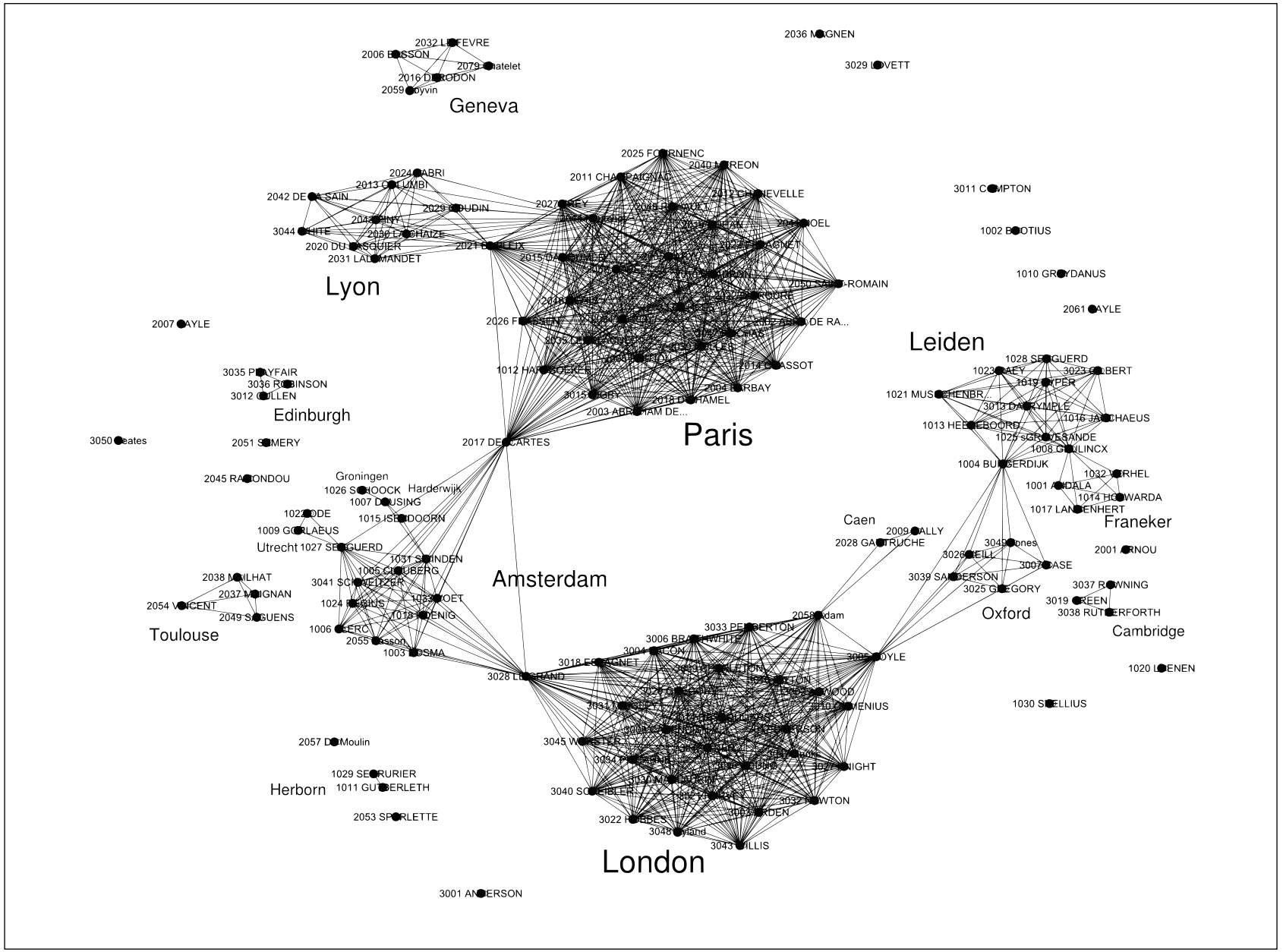

Fig. 11 Network representation of sub-layer 4 (publication place) shows authors clustering by cities (spring layout using visualization software Gephi (We used Gephi for this figure because it offered a better visualization of this specific layer.)

ical divides that we expect to be represented in the corpus, despite having little information about the great majority of the works included. Hence, we are confident that this method has a promising heuristic potential for complementing more traditional research on the evolution of early modern natural philosophy, but also more broadly for the integration of quantitative methods in the study of history of ideas and science.

7. References

Aizawa, Akiko. "An Information-Theoretic Perspective of Tf-Idf Measures." *Information Processing & Management,* vol. 39 (1, 2003): 45–65, https://doi.org/10.1016/S0306-4573(02)00021-3.

Alaux, Jean; Edouard Grave, Marco Cuturi, and Armand Joulin. "Unsupervised Hyperalignment for Multilingual Word Embeddings." 2019, https://arxiv.org/abs/1811.01124.

Allen, Colin and Jaimie Murdock. "LDA Topic Modeling: Contexts for the History & Philosophy of Science." In *The Dynamics of Science: Computational Frontiers in History and Philosophy of Science,* Ramsey, G., De Block, A. (Eds.). Pittsburgh University Press; Pittsburgh, 2021. http://philsci-archive.pitt.edu/17261/.

Bafna, P., D. Pramod and A. Vaidya, "Document Clustering: TF-IDF Approach," 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016, pp. 61–66, DOI: 10.1109/ICEEOT.2016.7754750.

Bjerva, Johannes and Raf Praet. "Word Embeddings Paving the Way for Late Antiquity." *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH).* Association for Computational Linguistics (2015): 53–57.

Blanke, Tobias and Claudia Aradau. "Computational Genealogy: Continuities and Discontinuities in the Political Rhetoric of US Presidents." *Historical Methods. A Journal of Quantitative and Interdisciplinary History* 54 (1, 2021): 29–43. DOI:10.1080/01615440.2019.1684859.

Bonacich, Phillip. "Simultaneous Group and Individual Centralities." *Social Networks* no. 13 (1991): 155–168.

Blei, David M, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet allocation." *Journal of Machine Learning Research* no. 3 (2003): 993–1022.

Blei, David and Michael I. Jordan. "Modeling Annotated Data." *Technical Report UCB//CSD-02-1202,* U.C.Berkeley Computer Science Division, 2002.

Boccaletti, Stefano, Vito Latora, Yamir Moreno, Mario Chavez, and Dong-Uk Hwang. "Complex Networks: Structure and Dynamics," *Phys. Reps.,* no. 424 (2006): 175–308.

Borgatti, Stephen P. 2009. "Social Network Analysis, Two-Mode Concepts in," in *Encyclopedia of Complexity and Systems Science,* ed. Robert A. Meyers (2009), DOI: https://doi.org/10.1007/978-0-387-30440-3_491.

Borgatti, Stephen P. and Daniel S. Halgin. "On Network Theory." *Organization Science* 22 (5, 2011). DOI: https://doi.org/10.1287/orsc.1100.0641.

Brezina, Vaclav, Tony McEnery, and Stephen Wattam. "Collocations in Context: A New Perspective on Collocation Networks." *International Journal of Corpus Linguistics,* 20 (2, 2015): 139–173.

Butterfield, Herbert. *The Origins of Modern Science.* London: Bell, 1957.

Carrera-Trejo, Víctor; Grigori Sidorov, Sabino Miranda-Jiménez, Marco Moreno Ibarra, and Rodrigo Cadena Martínez. "Latent Dirichlet Allocation Complement in the Vector Space Model for Multi-Label Text Classification." *International Journal of Combinatorial Optimization Problems and Informatics,* 6 (1, 2015): 7–19.

Cozzo Emanuele; Guilherme Ferraz de Arruda, Francisco A. Rodrigues, and Yamir Moreno. "Multilayer Networks: Metrics and Spectral Properties." *Interconnected Networks. Understanding Complex Systems,* ed. Antonios Garas (Cham: Springer, 2016): 17–35. DOI: https://doi.org/10.1007/978-3-319-23947-7_2.

De Bolla, Peter; Ewan Jones, Paul Nulty, Gabriel Recchia, and John Regan. "Distributional Concept Analysis: A Computational Model for History of Concepts." *Contributions to the History of Concepts,* no. 14 (1, 2019): 66–92.

Dickison, Mark E.; Matteo Magnani, and Luca Rossi. *Multilayer Social Networks.* New York: Cambridge University Press, 2016.

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

Düring, Marten. "How Reliable Are Centrality Measures for Data Collected from Fragmentary and Heterogeneous Historical Sources? A Case Study." In *The Connected Past. Challenges to Network Studies in Archaeology and History,* ed. Tom Brughmans, Anna Collar, and Fiona Coward, 85–102. Oxford: Oxford Publishing, 2016.

Evans, Eliza D. "Measuring Interdisciplinarity Using Text." Socius Sociological Research for a Dynamic World 2 (7) (2016), DOI:10.1177/2378023116654147.

Garber, Daniel. "Why the Scientific Revolution Wasn't a Scientific Revolution, and Why It Matters," in *Kuhn's Structure of Scientific Revolutions at Fifty: Reflections on a Science Classic,* eds. Robert J. Richards and Lorraine Daston, 133–150. Chicago: University of Chicago Press, 2016. DOI: 10.7208/Chicago/9780226317175.001.0001.

Gretarsson, Brynjar, Jason O'Donovan, Svetlin Bostandjiev, Tobias Hans Höllerer, Arthur Uy Asuncion, David J. Newman, and Padhraic Smyth. "TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling." *ACM Transactions on Intelligent Systems and Technology* no. 3 (February 2012): 1–26. DOI: https://doi.org/10.1145/2089094.2089099.

Hall, Rupert A. *The Scientific Revolution, 1500–1800: The Formation of the Modern Scientific Attitude.* Boston: Beacon Press, 1966.

Holley, Rose. 2009. "How Good Can it Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs." *D-Lib Magazine* no. 15 (3/4): unpaginated.

Jabri, S., A. Dahbi, T. Gadi, and A. Bassir, "Ranking of Text Documents Using TF-IDF Weighting and Association Rules Mining," 2018 4th International Conference on Optimization and Applications (ICOA), 2018, pp. 1–6, DOI: 10.1109/ICOA.2018.8370597.

Jockers, Matthew L. and David Mimno. "Significant Themes in 19th-Century Literature." *Poetics,* vol. 41 (6, 2013): 750–769.

Joulin, Armand; Edouard Grave, Piotr Bojanowski, Matthujs Douze, Hérve Jégou, Tomas Mikolov. "FastText.zip: Compressing Text Classification Models." 2016. arXiv:1612.03651.

Kenett, Dror Y., Matjaz Perc, Stefano Boccaletti. "Networks of Networks – An Introduction." *Chaos, Solitons & Fractals,* no. 80 (November 2015): 1–6.

Kivelä, Mikko, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. "Multilayer Networks." *Journal of Complex Networks,* no. 2 (3, 2014): 203–271, DOI: https://doi.org/10.1093/comnet/cnu016.

Koyré, Alexandre. *From the Closed World to the Infinite Universe.* Baltimore: The Johns Hopkins Press, 1957.

Leydesdorff, Loet. "On the Normalization and Visualization of Author Co-Citation Data: Salton's Cosine versus the Jaccard Index." *Journal of the American Society for Information Science and Technology* no. 59 (January 2008): 77–85.

Lind, Fabienne; Jakob-Moritz Eberl, Olga Eisele, Tobias Heidenreich, Sebastian Galyga, and Hajo G Boomgaarden. "Building the bridge: topic modeling for comparative research." *Communication Methods and Measures,* vol. 6 (2121, 2): 96–114, DOI: 10.1080/19312458.2021.1965973.

Nguyen, Thang; Jordan Boyd-Graber, Jeffrey Lund, Kevin Seppi, and Eric Ringger. "Is Your Anchor Going Up or Down? Fast and Accurate Supervised Topic Models." *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2015): 746–755. DOI: 10.3115/v1/N15-1076.

Röder, Michael, Andreas Both, and Alexander Hinnenburg. "Exploring the Space of Topic Coherence Measures." *WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (February 2015): 399–408, DOI: https://doi.org/10.1145/2684822.2685324.

Sangiacomo, Andrea; Raluca Tanasescu, Silvia Donker, and Hugo Hogenbirk. "Expanding the Corpus of Early Modern Natural Philosophy: Initial Results and A Review of Available Sources." *Journal of Early Modern Studies* vol. 10 (2021a, 1): 107–115.

Sangiacomo, Andrea; Raluca Tanasescu, Silvia Donker, and Hugo Hogenbirk. "Mapping the evolution of early modern natural philosophy: corpus collection and authority acknowledgement." *Annals of Science* 79 (2021b, 1): 1–39, DOI: https://doi.org/10.1080/00033790.2021.1992502.

Sangiacomo, Andrea; Raluca Tanasescu, Silvia Donker, and Hugo Hogenbirk. *Normalisation of Early Modern Science: Inventory of 17th- and 18th-Century Sources (1.0.0)* [Data set]. Zenodo, 2022. DOI: https://doi.org/10.5281/zenodo.5566681.

Sangiacomo, Andrea; Hugo Hogenbirk, Raluca Tanasescu, Antonia Karaisl, and Nick White. "Reading in the Mist: High-Quality Optical Character Recognition Based on Early Modern Digitized Books." *Digital Scholarship in the Humanities,* 2022. DOI: https://doi.org/10.1093/llc/fqac014.

Sayama, Hiroki. *Introduction to the Modeling and Analysis of Complex Systems.* Geneseo, NY: Open SUNY Textbooks, 2015.

Schmidt, Thomas and Kai Wörner. *Multilingual Corpora and Multilingual Corpus Analysis.* Amsterdam, New York: John Benjamins, 2012.

Shahmirzadi, Omar, Adam Lugowski, Adam, and Kenneth Younge. "Text Similarity in Vector Space Models: A Comparative Study," *18th IEEE International Conference on Machine Learning and Applications (ICMLA),* 2019: 659–666, DOI: 10.1109/ICMLA.2019.00120.

Singh, Lovedeep. "Clustering Text: A Comparison between Available Text Vectorization Techniques." In *Soft Computing and Signal Processing. Advances in Intelligent Systems and Computing,* eds. Reddy V.S., Prasad V.K., Wang J., Reddy K.T.V., vol 1340. Springer: Singapore, 2022. https://doi.org/10.1007/978-981-16-1249-7_3.

Suominen, Arho and Hannes Toivanen. "Map of Science with Topic Modeling: Comparison of Unsupervised Learning and Human-Assigned Subject Classification." *The Journal of the Association for Information Science and Technology* vol. 67 (10, 2016): 2464–2476. DOI: https://doi.org/10.1002/asi.23596.

Valleriani, Matteo, Florian Kräutli, Maryam Zamani, Alejandro Tejedor, Christoph Sander, Maite Vogl, Sabine Bertram, Gesa Funke, and Holger Kantz. "The Emergence of Epistemic Communities in the Sphaera Corpus: Mechanisms of Knowledge Evolution." *Journal of Historical Network Research,* no. 3 (2019): 50–91. DOI: https://doi.org/10.25517/jhnr.v3i1.63.

Van Den Heuvel, Charles. "Mapping Knowledge Exchange in Early Modern Europe: Intellectual and Technological Geographies and Network." *International Journal of Humanities and Arts Computing,* vol. 9 (1, 2015): 95–114. https://doi.org/10.3366/ijhac.2015.0140.

Van Vugt, Ingeborg. "Using Multi-Layered Networks to Disclose Books in the Republic of Letters." *Journal of Historical Network Research* vol. 1 (1, 2017): 25–51.

Venkatesaramani, Rajagopal; Doug Downey, Bradley Malin, Yevgeniy Vorobeychik. "A Semantic Cover Approach for Topic Modeling." Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (2019): 92–102. DOI: 10.18653/v1/S19-1011.

Wasserman, Stanley and Katherine Faust. *Social Network Analysis: Methods and Applications.* Cambridge: Cambridge University Press, 1994. DOI: https://doi.org/10.1017/CBO9780511815478.

Westfall, Richard S. "The Scientific Revolution of the Seventeenth Century: The Construction of a New World View," in *The Concept of Nature,* ed. John Torrance, 63–93. New York: Oxford University Press, 1992.

Zosa, Elaine and Mark Granroth-Wilding. "Multilingual Dynamic Topic Model." *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019):* 1388–1396.