Scientometrics (2021) 126:5881–5910
https://doi.org/10.1007/s11192-021-03984-1

Semantic and relational spaces in science of science: deep learning models for article vectorisation

Diego Kozlowski[1] · **Jennifer Dusdal**[2] · **Jun Pang**[1] · **Andreas Zilian**[1]

Received: 15 November 2020 / Accepted: 1 April 2021 / Published online: 15 May 2021

Abstract
Over the last century, we observe a steady and exponential growth of scientific publications globally. The overwhelming amount of available literature makes a holistic analysis of the research within a field and between fields based on manual inspection impossible. Automatic techniques to support the process of literature review are required to find the epistemic and social patterns that are embedded in scientific publications. In computer sciences, new tools have been developed to deal with large volumes of data. In particular, deep learning techniques open the possibility of automated end-to-end models to project observations to a new, low-dimensional space where the most relevant information of each observation is highlighted. Using deep learning to build new representations of scientific publications is a growing but still emerging field of research. The aim of this paper is to discuss the potential and limits of deep learning for gathering insights about scientific research articles. We focus on document-level embeddings based on the semantic and relational aspects of articles, using Natural Language Processing (NLP) and Graph Neural Networks (GNNs). We explore the different outcomes generated by those techniques. Our results show that using NLP we can encode a semantic space of articles, while GNN we enable us to build a relational space where the social practices of a research community are also encoded.

Keywords Embeddings · Science of science · Deep learning · Graph neural networks · Semantic space · Relational space

✉ Diego Kozlowski
diego.kozlowski@uni.lu

[1] Faculty of Science, Technology and Medicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

[2] Faculty of Humanities, Education and Social Sciences, University of Luxembourg, Esch-sur-Alzette, Luxembourg

Introduction

The relation between two research articles is a multidimensional phenomenon. Articles can be related because of their topics, authors, or the organisational affiliation of the authors. This implies that there is no unique measurement to compare the relatedness of scientific publications.

When a human compares a pair of articles, he or she can recognise the relatedness in its complexity, given his or her own biases and expertise in the field. The relatedness of a pair of articles is not only interesting for a pairwise comparison, but also to build a holistic representation of a field of research or a discipline. To handle the massively increasing volume and production rate of scientific research articles, we will discuss automatised ways to relate research articles to each other.

One of the most important dimensions of relatedness is the semantic content of articles. Using text as data, Natural Language Processing (hereafter NLP) studies the ways in which textual meaning can be extracted from the documents within a text corpus, in a summarised way (Jurafsky and Martin 2008). This operationalises the concept of semantic relatedness (Mikolov et al. 2013c). One of those techniques is topic modelling, which uses the co-occurrences of words in documents to detect the distribution of topics in a given corpus (Blei et al. 2003). Daenekindt and Huisman (2020) analysed the field of higher education using correlated topic modelling (Blei and Lafferty 2007), which allowed the authors to compare 17,000 abstracts published in journals in the field of higher education research. Schwemmer and Wieczorek (2020) studied the methodological divergence in sociology between quantitative and qualitative analysis in more than 8700 research articles in top journals of the field, using the *wordfish* model (Slapin and Proksch 2008). Another study analysed more than 20 million article titles to investigate the expansion of cognitive boundaries in physics, astronomy, and biomedicine. Findings include that the number of publications grows exponentially, but that the space of ideas expands linearly (Milojević 2015).

Another important aspect in the relatedness of articles is the overarching network structure of science (Boyack and Klavans 2010). A network of articles can be made explicit by their references to previous work. Considering direct citations, two articles are linked if one of them contains a reference to the other. This network has a strong temporal dependency, since as nodes can only have outgoing links to older articles. The links can also be based on co-citations (Kessler 1963; Small 1973). The network structure has the property of defining the distance between two articles by the length of the path between them, i.e., the number of articles need to get from one to another following the reference lists of the corresponding article at each step. Intertwined with the articles' network, a collaboration network of authors can be created (Moody 2004).

New deep learning models promise to revolutionise the way we approach both, text and network data. Nevertheless, there is an ongoing debate in the Artificial Intelligence community on the bias introduced by algorithms (see Bolukbasi et al. 2016; Buolamwini and Gebru 2018; Caliskan et al. 2017; Whittaker et al. 2018). This implies that, when we introduce new methodologies, and specially black-box models like in deep learning, it is essential to study the implicit biases they carry.

Together with other deep learning techniques, embeddings are an important breakthrough in the field of machine learning. An embedding is a low dimensional dense vector, i.e., of real-valued representations that encodes relevant information from the original, high dimensional data (Mikolov et al. 2013a). Nevertheless, not much research has been carried out to bring these techniques closer to the field of Science of Science, except of the

following examples: A recent article used Word2Vec (Mikolov et al. 2013a) to distinguish the most relevant terms in quantitative and qualitative research in the field of Science of Science (Kang and Evans 2020). Paper2Vector (Zhang et al. 2019) is a model that trains word-word, document-word, and document-document relations based on the Skip-Gram model (Mikolov et al. 2013a). This model, although it uses textual content as well as the citation network, cannot use both at the same time, and cannot use other metadata features. Another approach used Graph Neural Networks (hereafter GNN) models, a deep learning model that leverage on the network structure of the data as well as the BERT pre-trained model for the textual embedding, which constitutes the state of the art in NLP. This model does not consider the possibility of using the textual embedding as an input for the GNN (Jeong et al. 2020).

We have selected Science of Science as a case study to explore these new methodological approaches because it is a highly complex and multidisciplinary field of research (Fortunato et al. 2018) that aims to explore the driving forces of science and to develop new methods to better understand the evolution of science over time. The emergence of the field has speed up with the availability of large-scale data sets on science production and the disruption of disciplinary boundaries that encouraged scientists from different disciplines to closely collaborate. An ideal playground to study different types of embeddings as an innovative tool for the methodological development in Science of Science, for an in-depth analysis scientific research articles. We identified this as an important research gap in this developing field of research, and will discuss potential uses and limitations of these new methods. The article focuses on document-level embeddings, based on the semantic and relational aspects of articles using NLP and GNN's to explore semantic and relational spaces in Science of Science. Four different aspects will be analysed: (a) collaboration-patterns (Sooryamoorthy 2009; b) the cumulative effect on citations, i.e., the Matthew effect in science (de Solla Price 1963; Garfield 1972; Garfield and Merton 1979); (c) the position of countries in science global production (King 2004); and (d) the quantitative-qualitative divide (Kang and Evans 2020).

Our main hypothesis is that, while textual embeddings build a representation of the semantic space, GNN embeddings focus on the relational and structural space of a network of research articles. Therefore, textual embeddings help to identify similar content, whereas GNN embeddings are useful to study the embedded social relations in the production of scientific knowledge. An in-depth investigation of this hypothesis is an important step for Science of Science as a field of research, allowing us to pair powerful analysis techniques from computer science with a thematic disciplinary foundation.

The main objective of this methodological contribution is to present an approximation to the use of embeddings in the field of Science of Science. We propose two families of models that use different types of inputs and generate different types of insights: First, we build articles' embeddings based on their textual characteristics, including titles, keywords, and abstracts. For this family of models, we use three different techniques: topic modelling (Blei et al. 2003), Doc2Vec (Mikolov et al. 2013b), and BERT (Devlin et al. 2019). Doc2Vec was selected because it is specifically designed for document-level embeddings. BERT achieves the state of the art for various NLP tasks (Tenney et al. 2019). For a non-deep learning benchmark, we selected the topic modelling approach, which is a widely used framework. Second, we build GNN models that include text, metadata, and the citation network of selected articles. The GNN models are trained on the link prediction task, i.e., predicting whether two articles are linked by a citation, and therefore focus on the network properties. This study does not develop new methodologies from text or network embeddings, but intends to act as a bridge between the new developments made in the field

of Deep Learning, and studies in Science of Science. In our case study, we try the different models on a data set of 22,151 articles from Science of Science, involving different fields, ranging from history and philosophy of science to library and information sciences.

The following two research questions are guiding our analysis: How can we encode the *relational* dimension of articles, and which are its properties? How can we encode the *semantic* dimension of articles, and which are its properties?

Structure of the paper First, we will present the data set characteristics, while in Section Methods we will provide an overview of embedding techniques in both text and networks, the experimental setup and the performance metrics that are used to evaluate the models. In Section Results we will present our results. We will finish the paper with a conclusion and final remarks for future research in Section Conclusion.

Data set and network statistics

The data set was built on a 'journal-based' approach. First, we defined a set of core journals in the field of Science of Science. This selection is based on a recommendation of journals of the 'International Society for Scientometrics and Informetrics' (ISSI)[1]. and has been expanded to include related journals from social sciences and history and philosophy of science to show the wide variability and multidisciplinarity of this field of research. Our main goal was to include all journals that focus exclusively on Science of Science, independently of their disciplinary approach. Second, from this selected set of journals, we included all articles that are available in Elseviers' Scopus journal database. Methodologically, we have used Scopus API[2] to extract the data. By including all articles from these journals, we avoided a potential selection bias towards keywords, and ensure a comprehensive investigation of the field of Science of Science. Given that this study focuses on the methodological aspects of the use of embeddings, we decided to work with this non-exhaustive corpus of research articles, which gives us the advantage to investigate the distribution of embeddings at the journal level. As we carry out a country level analysis in Section Comparing the differences between the relational and semantic spaces, it is important to mention the limitations of the Scopus database. Limitations of the data set include a bias towards English-speaking and Western-oriented journals, and a lack of coverage of social sciences. We are only including peer-reviewed articles and do not investigate other publication formats (e.g., monographs, contributions to edited volumes, conference proceedings, etc.).

Table 1 displays information on the journals in our sample, including the number of articles retrieved, the mean and maximum number of citations by journal, and the year of the first and last publication. We have used 'Scopus Subject Areas and All Science Journal Classification Codes' (ASJC) for a first differentiation of the journals by discipline[3]. Having investigated the repetition of these areas in the different journals, and additionally based on the results of topic modelling (see Section Topic modelling), we assigned each of them to four fields of study: *Management*, *Library and Information Sciences*, *History and Philosophy of Science*, and *Other Social Sciences*, composed by *Education, Communication*

[1] http://www.issi-society.org/links/, last accessed October 6th, 2020

[2] Data collection between April 8 and July 30, 2020, via http://api.elsevier.com

[3] https://service.elsevier.com/app/answers/detail/a_id/15181/

Table 1 Statistics of the data set

Field	Journal	Articles Retrieve	Mean Citations	Max Citations	First Year	Last Year
Management	Research policy	3,221	83.75	4,820	1971	2020
	Science and public policy	1,707	13.27	462	1976	2019
Library and information sciences	Scientometrics	5,136	20.04	1,334	1978	2020
	Journal of informetrics	876	22.63	352	2007	2020
History and philosophy	Synthese	4,151	8.53	910	1946	2020
	Social studies of science	1,069	40.95	4,709	1971	2020
	Science and education	1,034	11.60	298	1992	2020
	Studies in history and philosophy of science	911	8.76	145	1974	2020
	Isis	523	12.47	415	1977	2020
	Science, technology and society	345	6.07	122	1996	2020
	British journal for the history of science	276	9.57	88	1962	2020
	Science and technology studies	111	5.29	39	2012	2019
Other social sciences: education, communication and Anthropology	Public Understanding of Science	977	25.91	518	1996	2020
	Science, Technology and Human Values	757	32.87	828	1982	2020
	Research Evaluation	666	13.15	223	1991	2019
	Minerva	391	16.51	624	1965	2020
	Total	22,151	20.71	4,820	1946	2020

Metric	Value
Number of nodes	16,578
Number of links	68,797
Number of nodes in the giant component	15,615
Number of links in the giant component	68,168
Diameter	37
Average degree	8.3
Max degree	282
Cluster Coefficient (C)	0.081
Mean path length (L)	6.14
Erdös-Renyi average cluster coefficient (C_r)	0.0005
Erdös-Renyi average mean path length (L_r)	4.83
C/C_r	162.45
L/L_r	1.27

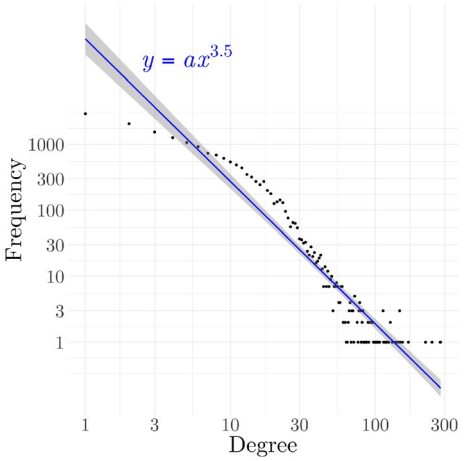

Fig. 1 Network statistics, log-log degree distribution and power law fit

and Anthropology. We consider the repetition of areas and acknowledge that some journals are more multidisciplinary in nature than others, which could lead to an assignment to another field. For example, *Social Studies of Science* incorporates *History*, *Social Sciences (all)* and *History and Philosophy of Science* as subject areas, so it could be assigned to both, *Other Social Sciences* and *History and Philosophy of Science*. This implies that the defined fields cannot be perfectly matched and characterise each journal. In addition, fields are not equidistant from each other. For example, the relation between *Management* and *Library and Information Sciences* is closer than the relationship between those fields and *Philosophy*. Methodologically, these fields are not used as features in the subsequent models, so they do not introduce biases to the models, but are a helpful tool used for the analysis of results, as they allow to visually study the projection of the embeddings in Section Evaluation of embeddings, and for studying the journals epistemic practice division in Section Comparing the differences between the relational and semantic spaces. In Section Topic modelling, we show that we can partially infer these fields from the topic modelling results.

The distribution of the number of research articles and citations per journal is skewed (see also Bornmann et al. 2008), with a tendency towards *History and Philosophy of Science* having less citations per article than the rest of the fields, which corresponds to different citation and publication behaviours across disciplines (Lillquist and Green 2010). Overall, the data set includes 22,151 articles, with an average of 20.7 citations per articles.

From the collection of 22,151 articles, we retrieved the references to build the citation network. 75% of them either cited or were cited by another article in our sample. This subset builds the basis for the citation network[4]. Figure 1 presents a summary of the characteristics of the resulting network the log-log degree distribution of nodes and the fitted power law distribution. The network has 16,578 nodes (articles) and 68,797 links (citations). The

[4] In the Online Resource B we show the summary statistics of the articles inside and outside of the network.

average degree of the network is 8.3 compared to 20.7 connections when using the entire Scopus database. While the network is not connected, the giant component includes most of the edges and vertices. We built 100 replications of a random Erdös-Renyi network (Erdös and Rényi 1960) with the same number of vertices and edges, and computed the average cluster coefficient as well as the average mean path length for comparison. We show that the ratio of the network cluster coefficient with respect to the random Erdös-Renyi network is 162, while the ratio between our network mean path length and that of the random Erdös-Renyi network is 1.27. This means that our network has the properties of *small word* networks (Davis et al. 2003; Iyer et al. 2006).

Methods

This section introduces different approaches for dense vector representations of documents, called embeddings, which will be used for analyses of scientific research articles in Science of Science. We begin with a brief recapitulation of classic machine learning approaches for encoding research articles based on feature engineering, to highlight its differences between those and deep learning based approaches. In Section Methods.1 we present the textual-based embeddings, while in Section The relational space of a research article we present the network-based embeddings. Then, the implementation is detailed in Section Implementation steps. Finally, Section Performance metrics explains the performance metrics used for evaluation.

Classic Machine Learning Approaches

Feature engineering refers to the way in which measurable attributes of an observation can be encoded. In a traditional data set, each observation is defined as a vector, and the full data set is therefore a composition of these vectors, building a matrix of observations as rows and columns as features (Broman and Woo 2018). If the data consists of a network besides the features-matrix, another matrix to describe the network structure has to be implemented, the so-called *adjacency matrix A* (Barabàsi 2016). An article's vector representation can be any metric way to summarise its measurable characteristics. Such a vector describes its metadata features, for example, the year of publication, the number of citations at a moment in time, the number of authors, or a label assigned to the organisation or journal. It can also include descriptors of the network in which the article is embedded: degree, betweenness, or other centrality measures.

The classic treatment of the textual content of an article, for example, the title, abstract, keywords, or full text, is based on the *bag of words*: A Document-Term Matrix (DTM) where each document is represented as a vector, indicating which of the words in the vocabulary is present in a given document (Jurafsky and Martin 2008). In the DTM matrix, the i-th row represents the i-th document and the j-th column represents the j-th word in the vocabulary of the corpus. The value $x_{i,j}$ can either indicate if the j-th word is present in the i-th document as a binary value, the number of times it appears, or a normalised counting, like Term Frequency-Inverse Document Frequency (TF-IDF) (Jurafsky and Martin 2008).

As any n dimensional vector, an article constitutes a point in an n dimensional space. Each of these n dimensions keeps its independent meaning, which is useful for making interpretations over the position of different documents in the space. The representation might be affected by problems of high dimensionality and sparsity. This is

specially true for encoding text, as the vocabulary size can rise to tens of thousands of words with a high probability that most words will not appear in most documents. For research articles this might be true for features like authors, organisational affiliations, or journals, if each value is encoded as a dummy variable, i.e., many variables take only the value 0 or 1 to indicate the absence or presence of each category. In highly dimensional spaces, the notions of distance become blurred and it is hard to generate new insights from the data (Bellman 1966). Therefore, when studying the relations between articles, the analysis tends to focus on a restricted subset of the multiplicity of dimensions that exist. Compared to other methodological approaches, deep learning models can take a multiplicity of dimensions of analysis into account. Applying deep learning methods to Science of Sciences contributes to development on this research gap, as they are able to use multiple features and select those that are more relevant based on the optimisation problem.

Once encoded, descriptive statistics can be used to investigate the relations between observations and features. When the number of dimensions to be considered increases, descriptive statistics can not deal with the studied phenomenon, and some types of models might be necessary to assess the importance of different features. Modelling relations is an entrenched way to reduce the complexity of the problem by selecting the dimensions in focus. One possibility is to use models that measure the relation between two aspects of the phenomena, for example linear models. These types of models demand a lot of time and expert knowledge because they require a careful design and hand-engineered features for a better performance. New developments in machine learning, like deep learning models, can be trained from the raw data to encode the most relevant information, allowing researchers to focus on the analysis of the results.

Deep learning models

are designed in an end-to-end way. Their goal is to minimise the feature engineering steps, and to let the model itself define which the most important latent features are. These models have greater flexibility in terms of the inputs they can receive and the outputs they generate (Goodfellow et al. 2016). In this article, we focus on those models where the output is a subspace projection, where observations are represented. This encoding is defined by the models to maximise an associated task that changes between models. If the model works properly, it will encode the most relevant aspects from the original data to solve the associated task. Using this new, data-driven way of encoding information, it is possible to generate new insights comparing aggregated levels of representation like countries or journals. For this type of modelling, we consider two subtypes: First, the textual embeddings that use textual features as their input. From this, we will infer the *semantic space* of articles. We aim to encode the conceptual sense of each article as a low dimensional vector. Second, we will train models that use both text and the citation networks as input, from which we will infer the *relational space* of research articles. In the relational space, we aim to encode the citation practices as a social phenomenon into a low dimensional vector. Here we consider the semantic and relational spaces as the mathematical objects that represent those properties of articles. The embeddings are the deep learning implementations that we train over our data set to approximate those mathematical objects.

The semantic space of a research article

In this section, we explain the three models that we have used to build the semantic space of research articles. Doc2Vec (Mikolov et al. 2013b) and BERT (Devlin et al. 2019) are deep learning models based on the Word2Vec model (Mikolov et al. 2013a). The Latent Dirichlet Allocation model (hereafter LDA), which is a non deep learning approach, is extensively used for topic modelling, but can also be considered as an embedding.

Word embedding

The representation of documents for training deep learning models is commonly based on word embeddings (Bojanowski et al. 2007; Mikolov et al. 2013a; Pennington et al. 2014). In word embeddings, each word is represented as a dense vector, and documents are a concatenation of those vectors. To build this representation, Mikolov proposes Word2Vec, (Mikolov et al. 2013a), and the *Skip-Gram* implementation. Given a corpus of text and a window size, the model defines the context of a word as the surrounding words within the window size. Then, for each word, it tries to predict its context, internally building a vector for each word. When trained, it learns to project closer words with similar meaning. This means that when we use this model on our Science of Science data set, words like *technology* and *innovation* have a closer representation between each other than with the word *student*. When using word embeddings, the document is represented as a matrix of word vectors. Our goal is to make an embedding of articles, instead of words. Mikolov et al. (2013b) include the identifier of the corresponding document as an additional token to the context window. In this way, it creates both an embedding for words and for documents.

BERT embedding

BERT (Devlin et al. 2019) improves the Word2Vec model using *attention* mechanisms (Vaswani et al. 2017). This means that not every word in the context is equally considered while trying to predict the next word in the context, which implies that the word embedding of a word is also determined by the context. BERT is also based on the principle of transfer learning. This means that the word vectors can be learned on a big general domain corpus, instead of the specific corpus for each task. This is useful because to build a robust representation of concepts, the word embeddings need billions of observations (Mikolov et al. 2013a).

LDA embedding

For comparison, we also use the LDA model proposed by Blei et al. (2003). This model is not based on a deep learning architecture. Instead, it is a generative Bayesian model. It starts from the premise that a corpus is a collection of topics, and that each document is a mixture of those topics. The model takes the distribution of words within documents as input and generates the topics as a probability distribution of words. For example, if *science policy* is a recurrent discussion in our corpus, LDA would define a topic were words like 'technology', 'innovation' or 'policy' are the most relevant. Given these topics as distributions over words, the model will also define each document as a distribution over those topics. LDA is not usually seen as generating an embedding, as this terminology is

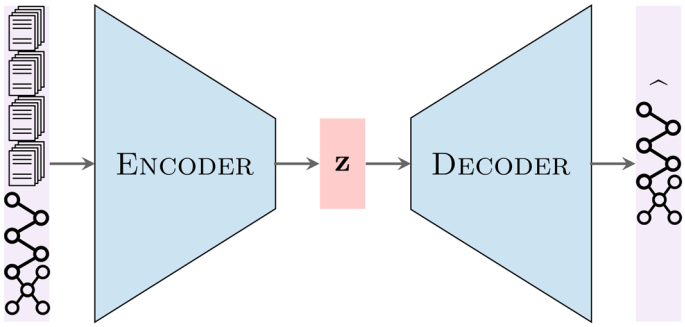

Fig. 2 Graph Encoder-Decoder architecture. Author's representation, based on Allingham (2020)

usually found within the deep learning community. Nonetheless, its output can interpreted of as an embedding for articles, as we can use the articles' distribution over topics as their low-dimensional representation[5].

These three models are compared based on their T-SNE projection (van der Maaten and Hinton 2008), and on how much the GNN model improves when using each of these as input.

The relational space of a research article

The semantic embeddings are built to study exclusively the textual content of documents. An analysis of research articles can consider more than its textual content. We have the opportunity to assess other meta-data features and the citation network to make relationships between articles explicit. GNN's have the potential to assess a more holistic representation of research articles.

GNN is a developing field in the deep learning community that tries to apply techniques that have been proven as useful in computer vision and NLP to solve problems where the data has a network structure. Multi-Layer perceptrons work well with flat inputs, where it learns the compositionality of features, but it does not consider explicitly their specific dependencies (Goodfellow et al. 2016). Recurrent Neural Networks (RNN) are designed for sequential inputs, like text, where the order of the input features is explicitly fed to the network (Sutskever et al. 2011). Convolutional Neural Networks (CNN) are useful when dealing with images, where spatial relations within the image are searched, independently of the specific position in the grid (LeCun et al. 1989). The problem with graphs is that they have explicit relations among nodes which we would like to incorporate to our models, but they are not as regular as relations in images or text. Therefore, traditional RNN and CNN cannot deal with this type of data structure. The two main differences are:

1. Nodes have a variable number of neighbours; and
2. neighbours do not have a predefined order.

[5] In the Online Resource C.1 we give further details of the LDA model.

Deep learning on graphs deals with these issues, trying to generalise RNN and CNN to the complex relations in networks. Although GNN can be used for a number of tasks, like node and graph classification, we approach the embedding generation as an unsupervised problem, aiming to reconstruct the node's neighbourhood. We are not trying to predict some node's features while we train our deep learning models, but we try to rebuild the network structure. Following Hamilton et al. (2017b), we approach the problem with an Encoder-Decoder framework. Figure 2 shows the outline of this architecture.

From the Encoder-Decoder perspective, the model follows two main procedures: The encoder, *ENC*, takes the nodes features and the network structure, and generates a low-level representation of nodes. The decoder, *DEC*, takes the low-level representation as an input and tries to rebuild the network structure. Following Kipf and Welling (2016) proposal of the Graph Autoencoder (GAE), the main element of the decoder function is the *inner product* between nodes' low-level representations:

$$\mathrm{DEC} = ZZ^T,$$

where Z is the low-level matrix representation of nodes with dimension $n \times d$, and with n the number of articles and d the dimension of the article embedding, generated by the encoder. This generates a pairwise decoder. For each node, we reconstruct its relation with all other nodes, generating an $n \times n$ matrix. If nodes share a similar low-level representation, the inner product will give a higher value for their pairwise relation.

If we apply on top of the inner product function we apply a sigmoid activation layer, the decoder will produce the pairwise relation of nodes, expressed as a probability, i.e., the probability of the two nodes being linked in the network:

$$\hat{A} = \sigma(ZZ^T),$$

where $\hat{A}$ is the reconstructed adjacency matrix. The goal is to optimise the encoder in order to minimise the reconstruction loss:

$$\mathcal{L} = \sum_{(u,v)\in\mathcal{V}} \ell(\hat{A}[u,v], A[u,v]),$$

where ℓ is the loss function, in our case the binary cross entropy between the reconstructed link of nodes u and v, and the true value in the adjacency matrix. Optimising this loss function implies training an encoder that will generate an embedding representation of nodes that preserve their similarities in terms of the network structure.

The main difference between the used models is how they define the encoder step. We use as encoders the Graph Convolutional Network (GCN) (Kipf and Welling 2017), GraphSAGE (Hamilton et al. 2017a), Graph Isomorphic Network (GIN) (Xu et al. 2019), Graph Attention Network (GAT) (Veličković et al. 2018), Attention-based Graph Neural Networks (AGNN) (Thekumparampil et al. 2018), and GraphUNet (Gao and Ji 2019) layers, which constitute the current state of the art in the field. In the Online Resource C.2, we present an overview of these models.

Implementation steps

In this section we present the pre-processing, data cleaning, features, hyperparameters, and network architectures that we have used to build and evaluate the models.

All textual models were built using a combination of the title, abstract, and keywords suggested by the authors of each article. Given that the words in the title and keywords are good representations of the content of an article, we use them in triplicate: we build a text that contains three times the title, three times each keyword, and once the abstract. To clean the data, we remove stopwords and trademarks of the journals. After this, we replace the numbers with the special token 'num'. We also did stemming, and later replaced the stem with the most frequent word with that stem.

The Doc2Vec model was implemented using Gensim (Rehurek and Sojka 2010) with the 'distributed memory' learning algorithm, a vector size of twenty and a window size of ten, and using the concatenation of context vectors. To train the BERT embedding, we used the HuggingFace implementation (Wolf et al. 2019) with the 'bert-base-uncased' pre-trained model. Given that BERT generates word vectors, the sentence embedding was built as the mean of the token embeddings in each sentence. The LDA model was implemented using Scikit-learn (Pedregosa et al. 2011) with twenty components, removing tokens that appear in less than five documents or more than 65% of the documents.

The GNN models were trained using the following features:

- First author affiliation ID
- First author ID
- Year of publication
- Journal subject area (three per journal, as dummy variables)
- Topic distribution (from LDA)
- Keywords, title, and abstract information, either as TF-IDF, or using the sentence embedding from Word2Vec or BERT
- Cumulative citations after t years from its publication, for $t \in 1, ..., 10$, and total number of citations[6].

The GNN architectures were implemented using the Pytorch Geometric implementation (Fey and Lenssen 2019). In all cases, we use the Graph Autoencoder (Kipf and Welling 2016) with the inner-product decoder. The main difference between the GNN models (GCN, GraphSage, GIN, GAT, AGNN, GraphUNet) is on the encoder side. To find the best set of hyper-parameters, we replicated the specifications from the original papers. The following hyper-parameters have been used in each model: The GCN has an output dimension of 32 and has been built with two GCN convolutions with a Relu activation layer (Agarap 2018). GraphSage was trained with a similar architecture, using the Sage convolutional layer. The GIN model was trained with five GIN convolutional layers, using ELU activation layer (Clevert et al. 2016), batch normalisation and a normalisation layer in the end. The GAT model was trained with two GAT convolutional layers, with a dropout of 0.6 and a normalisation layer in the end and an ELU activation layer after the first convolution, the embedding dimension was 16. The AGNN model also has two convolutions, starting with a linear projection followed by a Relu activation layer. The GraphUNet has the same dropout as the one mentioned in the original paper (Gao and Ji 2019) with a depth of 4 and an embedding dimension of 16.

[6] When the article has been published less than t years ago, the number of citations has been imputed, based on the previous year number of citations, and the mean variation from $t-1$ to t.

Performance metrics

We choose two traditionally used metrics in GNN for the evaluation of the link prediction task: the Average Precision (AP) and the Area Under the Receiver Operating Characteristic Curve (AUC). To explain AP, we have to introduce some preliminary notions: for a binary classification model an observation can be either positive or negative. A True Positive (*tp*) is an observation predicted as positive that is a real positive case. In the same way we can define true negative (*tn*), false positive (*fp*), and false negative (*fn*) predictions. Furthermore, we define

$$\begin{aligned} \text{Precision} &= \frac{tp}{tp+fp}, \\ \text{Recall} &= \text{TPR} = \frac{tp}{tp+fn}, \\ \text{FPR} &= \frac{tp}{fp+tn}. \end{aligned}$$

Precision measures describe how many of the predicted links are actual links. Recall, or True Positive Rate (TPR) measure, how many of the actual links are predicted by the model as positive cases. The False Positive Rate (FPR) measures the 'false alarm', i.e., the ratio between true positive and true negative cases. Given that the models predict a ranked sequence of link candidates, each potential link is associated with a probability. There is an implicit trade-off between Precision and Recall, or between TPR and FPR. For each mode, using the ranked predictions, we can build a curve of Precision against Recall. The AP is the area under that curve, and can be computed as

$$\text{AP} = \sum_n (R_n - R_{n-1}) P_n .$$

This means, for every candidate it computes the precision at that point, weighted by the change in the Recall. In a similar way, the AUC is the area under the TPR against FPR curve.

After having presented our data and methods as well as the experimental setup of our study, we now present our results.

Results

Topic modelling

For a first characterisation of the data set, we use topic modelling to find the latent space of sub-fields within the corpus, using the LDA model (Blei et al. 2003). LDA provides two different outputs: First, the distribution of word over topics, which is useful for defining the meaning of each topic. Second, the distribution of topics over documents[7]. Using LDAvis, we built an interactive visualisation of the distribution of words over topics[8]. (Sieverta and

[7] With the distribution of topics over documents, we will later take LDA as a way of embedding over the original space.

[8] Available at https://diegokoz.github.io/scisci/

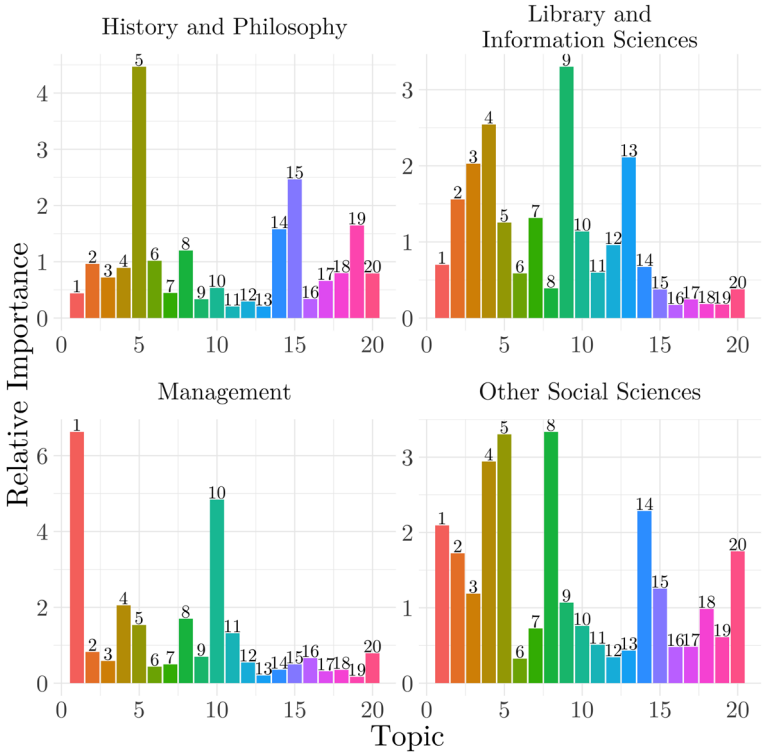

Fig. 3 Relative importance of topics per field. For a list of the most relevant words per topic, see Table D.1 in the Online Resource. An interactive version is available at https://diegokoz.github.io/scisci/

Shirley 2014). Figure 3 shows the relative importance of each topic per field, calculated as the proportion of the topic in the field over the proportion of the topic in the entire data set.

If we compare the the most relevant words per topic[9] with the distribution per field, Figure 3 shows that in *History and Philosophy of Science* the topics discussed are *education* (topic 5), shared also with the field of *Other Social Sciences*, *history* (topic 15) and *logic* (topic 19). In the field of *Library and Information Sciences* many different topics are discussed, with especial emphasis on *bibliometrics* (topic 9), and *universities and scientists* (topic 4), they are also shared with *Other Social Sciences*. *Management* focuses on *technology policy* (topic 1) and *patents and firms* (topic 10). In *Other Social Sciences* we find a variety of topics, and besides the shared topics 4 and 5, there is also attention on *public decision making* (topic 8).

The LDA results show the wide variety of topics covered in the field of Science of Science and also within the different disciplines involved in it. Some topics are shared across fields, and some journals, like *Scientometrics*[10], cover a wide array of thematic studies. This analysis confirms that there is no unique unequivocal way of gathering journals by field.

[9] See Table D.1 in the Online Resource

[10] See Figure D.1 in the Online Resource

Table 2 Link prediction results. Area Under the Curve and Average Precision. Mean result of 10 runs and standard deviation in parenthesis

Text Encoding	Model	AUC	AP
TF-IDF	GAT	0.79 (0.0)	0.78 (0.01)
	GraphUNet	0.79 (0.03)	0.77 (0.04)
	AGNN	0.83 (0.01)	0.78 (0.02)
	SAGE	0.85 (0.0)	0.87 (0.01)
	GIN	0.87 (0.01)	0.88 (0.01)
	GCN	0.87 (0.0)	0.89 (0.0)
D2V	GAT	0.79 (0.0)	0.77 (0.01)
	GraphUNet	0.79 (0.03)	0.76 (0.03)
	AGNN	0.83 (0.02)	0.8 (0.02)
	SAGE	0.85 (0.0)	0.87 (0.01)
	GIN	0.87 (0.01)	0.89 (0.02)
	GCN	0.86 (0.0)	0.88 (0.0)
BERT	GAT	0.78 (0.01)	0.76 (0.01)
	GraphUNet	0.79 (0.03)	0.76 (0.06)
	AGNN	0.84 (0.02)	0.79 (0.02)
	SAGE	0.87 (0.0)	0.89 (0.0)
	GIN	0.87 (0.02)	0.88 (0.02)
	GCN	**0.91** (0.01)	**0.91** (0.0)

Table 3 Link prediction results when removing a feature. Area Under the Curve and Average Precision. Average result of 10 runs and standard deviation in parenthesis

Removed	AUC	AP
First Author	0.91 (0.0)	0.91 (0.0)
Affiliation	0.9 (0.01)	0.9 (0.0)
Subject Area	0.9 (0.01)	0.9 (0.01)
Topic Distribution	0.9 (0.01)	0.9 (0.01)
Year	0.9 (0.0)	0.9 (0.0)
citations at 1:10	0.89 (0.0)	0.89 (0.0)
BERT embedding	0.86 (0.0)	0.87 (0.0)

Evaluation of embeddings

In this section, we perform an evaluation of the multiple models proposed. For this, we quantitatively compare the performance of the models on the link prediction task, including how the different textual embeddings as features improve the results. We also visualise our results based on the T-SNE projection of the embeddings.

Relational space

As the GNN models are trained on a measurable task, we first compare the models results, to analyse the resulting embedding of the one with better performance. In Table 2 we show the Area Under the ROC Curve and the Average Precision, two metrics widely used to evaluate GNNs for the link prediction task (Kipf and Welling 2016; Zhang and Chen 2018). For the six different architectures, we are using three different ways of encoding

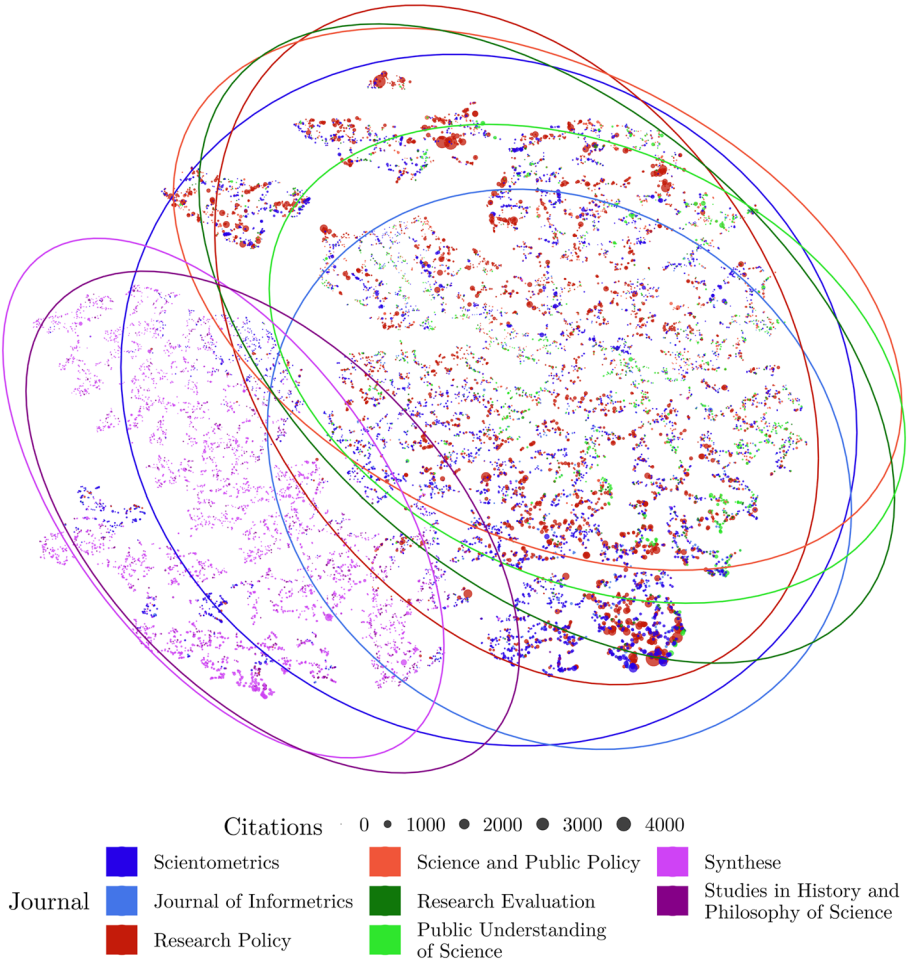

Fig. 4 GNN Embedding, with GCN and BERT encoded text. T-SNE projection. Journals from the same field are presented as variations in the luminosity of a similar chromaticity. The size corresponds to the number of citations. 95% ellipses by journal

text: The traditional TF-IDF features (Kipf and Welling 2017), the sentences embedding using Doc2Vec (Hamilton et al. 2017a), and BERT. As shown above, in every case the best architecture is the GCN (Kipf and Welling 2017), closely followed by the GIN (Xu et al. 2019), and GraphSage (Hamilton et al. 2017a). GCN is also the one that improves the most when using BERT and achives the best performance of 0.91 in both AUC and Average Precision. Using either TF-IDF or D2V achieves approximately equal results.

We performed ablation studies to understand the features relevance. In Table 3 we present the impact of removing one of the features from the GCN model with BERT embeddings. Results show that adding or removing the label of the first author ID does not change the performance, which corresponds with the high cardinality of this feature. Similarly, removing the organisational affiliation, subject area, topic distribution, and year slightly decreases the performance. When removing the cumulative distribution of citations, the model worsens in its predictive power by 2%, and by 5% or 6% if we remove the

Fig. 5 Semantic embeddings. T-SNE projection. Journals from the same field are presented as variations in the luminosity of a similar chromaticity. The size corresponds to the number of citations. 95% ellipses by journal

BERT embedding, which means that these two features are the most relevant. Not having the BERT embedding gives a closer result to those obtained using Doc2Vec and TF-IDF, which also shows the small impact of those ways of encoding text over the GNN. To conclude, the GCN is robust to the use of different features, and it is mostly focused on the network structure. We expect the embedding representation to be highly related with the network properties of the citation patterns and only mildly related with the semantic patterns through BERT.

In Figures 4 and 5, we display the T-SNE projection (van der Maaten and Hinton 2008) for a sample of two journals per field, coloured by field and journal, and sized by number of citations. We can also see the 95% ellipses of each journal, i.e., containing 95% of the articles of the corresponding journal (Fox and Weisberg 2018). Figure 4 is based on the GNN embedding using GCN and BERT. Results show a correlation between the number of citations and the position in space. For example, those articles from the journals *Research Policy* and *Scientometrics* with the highest number of citations are located in the bottom-right and top of the T-SNE projection. Journals from the field of *History and Philosophy of Science*, like *Synthese* and *Studies in History and Philosophy of Science*, where the citing culture and the selection bias of the data set imply lower citations (see chapter 2), are located on the left of the plot. Within this field, articles with a higher number of citations cluster together in the top-left of the T-SNE representation. All fields but *History and Philosophy of Science* form a uniform point cloud that correlates more with the number of citations than the corresponding journal. This can also be observed in the overlapping ellipses of most journals, except for those from *History and Philosophy of Science*. The organisation of the plot by citation patterns rather than thematically highlights when compared with the semantic embeddings in Figure 5. This result is in line with the expectation of the GNN, paying more attention to the relational patterns rather than the semantic content of research articles. Citation rates differ among disciplines. Especially research articles in the social sciences and humanities have a lower citation rate than articles in other disciplines. Further, the selection bias of the data set in favour of specific journals might cause an underestimation of citation rates for those journals which might have a higher number of cross-references to journals that are not included in our sample as well as to other publication formats (monographs, contributions to edited volumes, etc.) that have not been analysed in this study.

Semantic space

Figure 5 shows results for the three textual embeddings. The Doc2Vec model is the only one of the three proposed models that was designed for document-level embeddings. Nevertheless, it does not show any correlation among journals or number of citations; the shape of the point cloud is spherical. This might be an indication for Doc2Vec not being able to learn a good representation of the documents of our data set. This is probably due to the fact that the Skip-Gram model is usually trained on millions of data points. As outlined above, pre-trained embeddings built on a general purpose corpus can improve the words representation, but this is not possible on the document level in Doc2Vec.

The LDA embedding correctly delimitates the four sub-fields of the data set. Articles from the field of *History and Philosophy of Science* are located on the left side of the projection, with articles from *History and Philosophy of Science* located closer to the border with the field of *Other Social Sciences*. This latter field is in between *History and Philosophy* on its left and the field of *Management* to its right, where articles, especially from

Research Evaluation, tend to merge. We can also see a small number of articles from this field in the middle of the cluster of *Library and Information Sciences*. Finally, the fields of *Management* and *Library and Information Sciences*, share a point of contact in the centre of the plot, and some articles from each of these fields can be located in the cluster of the other. Compared to Figure 4, the correlation of highly cited articles is driven by the journals, as within each field we cannot see any clustering of highly cited articles. The BERT embedding, where we were able to use pre-trained word vectors, shows a better performance than Doc2Vec, even when the model is not originally designed for document-level representation and we are averaging the word embeddings in each document. The T-SNE representation shows a stronger delimitation between fields, and a delimitation between journals within each field. Articles from *History and Philosophy of Science* are located on the left side of the figure, but with a stronger delimitation. In the top of the figure we can see the field of *Management*, where articles from *Research Policy* are placed towards the centre of the figure, and those from *Science and Public Policy* are shifted to the top of the figure. The field of *Library and Information Sciences* is split in three groups: Most of the articles from *Scientometrics* lay on the centre of the figure, mixed with those from *Research Evaluation*. On the left, a small proportion of the articles from this journal stays closer to the field of *History and Philosophy of Science*. On the right, another group of articles from this journal, and most the articles from the *Journal of Informetrics*, are located closer to those from *Science and Public Policy*. This might be a reflection of a methodological field that develops technical aspects, but also applies the developed methodologies to thematically specific research questions. Finally, the field of *Other Social Sciences* is found in the bottom of the plot, with articles from *Public Understanding of Science* more delimited, and articles from *Research Evaluation* closer to the field of *Library and Information Sciences*.

Comparing the differences between the relational and semantic spaces

After studying the overall quality of the different models, we now focus on those that have the best performance in both the semantic and the relational space. For the latter, we select the GCN using BERT embeddings as features. For the semantic space, we mainly focus on the BERT model, but also show some results for the LDA model for comparison. In this section, we present how the embedding representation of articles change in the semantic and relational space. We compare the results on four different topics largely studied in the field of Science of Science: First, the representation of collaboration patterns, second, the Matthew effect in science, third, we perform a country level analysis, and fourth, the epistemic practice division in the field. Our goal is to compare how these different phenomenon are encoded in the resulting embedding, and how their representations differ within the proposed models.

Collaboration patterns

The pairwise similarity between articles can be observed through the cosine similarity between their vector. To compare the higher level groups, we can calculate the average cosine similarity between groups.

One possible dimension of analysis are the collaboration patterns in terms of co-authored papers, and whether they are encoded in the embeddings. For this, we the divide articles between four groups by their forms of collaboration: (A) Single author, (B) internal

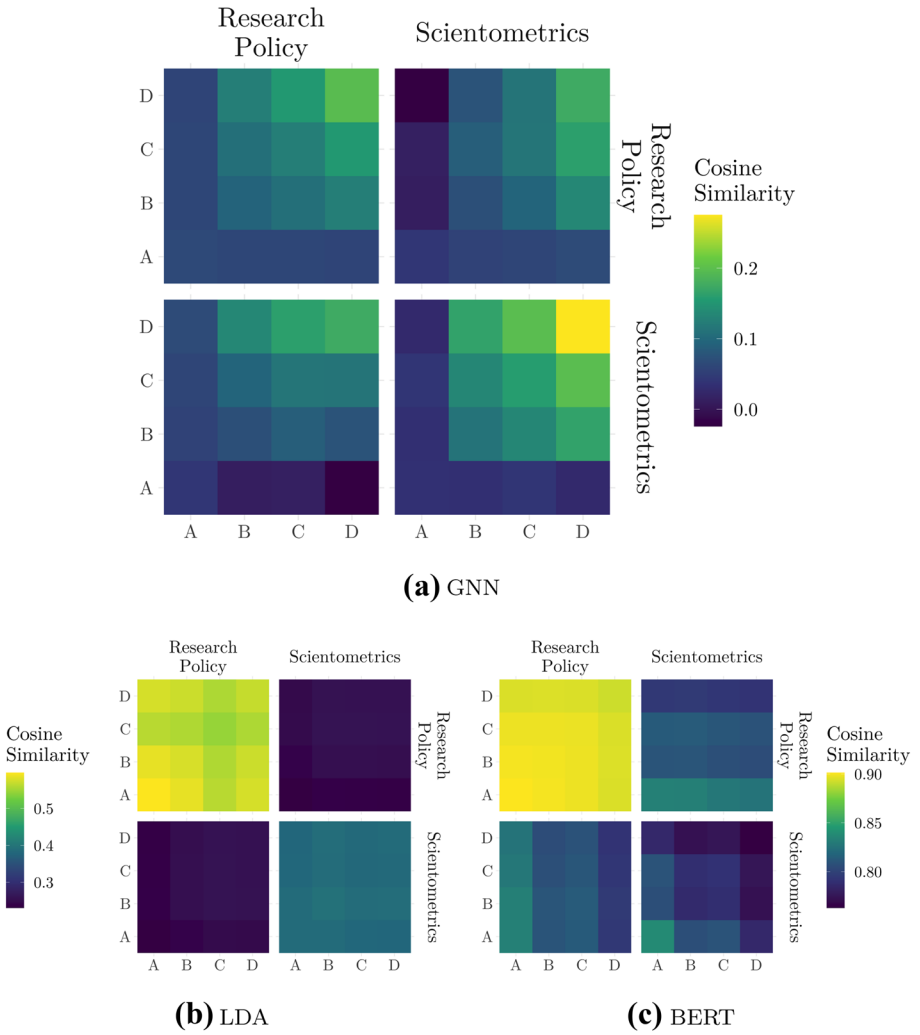

Fig. 6 Cosine similarity by collaboration status, controlled by journal. **A**: Single authorship; **B**: Collaborations between authors from the same organisation; **C**: Collaboration between authors from different organisations from the same country; **D**: Collaborations between authors from organisations in different countries

collaborations within a single organisation, (C) collaborations between authors with different organisational affiliations from the same country, and (D) international collaborations, including authors from different countries and organisational affiliations. To avoid biases due to different collaboration patterns by journal, the results will be illustrated by reference to a comparison of the journals *Research Policy* and *Scientometrics*. For example, we compare the average cosine similarity of articles from one journal to the other, and within each journal. Each of the three embeddings shows different uses of space, and hence for each we define a specific colour scale.

Figure 6a shows the results for GNN. Results shows a higher cosine similarity for international collaboration. the average similarity reduces from this type of collaboration towards single authorship, and this is a pattern that remains independently of the journal.

Articles from *Research Policy* are not closer to other articles in *Research Policy* than to articles in *Scientometrics*, except single authored papers, where we find a small difference. This means that international collaborations are closer to all other articles. If we consider the relational space as a hyper-sphere of articles, i.e., an sphere that lays in more than three dimensions, international collaborations are in the centre of this hyper-sphere, and single authored publications tend to be on the periphery. This is inline with the bibliometric analysis on the higher impact of international collaborations measured through higher citation rates (Adams 2013; Persson et al. 2004; Van Raan 1998). Figure 6b shows the results for the LDA model, while Figure 6c shows the results for the BERT model. In both cases, these semantic embeddings do not show strong correlation between collaboration patterns and article similarity. Articles from *Research Policy* are closer to each other, as well as *Scientometrics* articles between them. In both cases, articles from *Research Policy* tend to be closer together, and in the case of BERT, single authored articles from *Scientometrics* tend to be closer to *Research Policy* publications. This means that while the semantic embeddings build similar representations for articles in the same journal, the GNN builds a representation of articles where the international collaborations are similarly encoded and in a central position.

The Matthew Effect in Science

The Matthew effect in science, introduced by Robert K. Merton Merton (1974), states that articles that are already highly cited have a higher chance of being cited again. Figure 6a reflects on this via the collaboration patterns, while Figure 4 also shows a correlation of articles by the number of citations. To test if the embeddings are able to capture the Matthew effect, we divided the articles by their number of total citations in the Scopus data set, by the quartiles, and a separate group for those with zero citations (five groups in total). Then, we calculated the average Frobenius norm of each articles, and aggregated the results for these five groups. When studying the distribution of the Frobenius norm by citation level on the different models[11], we found that while the GNN generates a higher value for the highly cited articles, the BERT, Doc2Vec, and LDA models do not follow this pattern. This results mean that the GNN is systematically representing highly cited articles differently. This is expected, given that the GNN is trained on a link prediction task, and a higher Frobenius norm is associated with a higher probability of link, i.e., citation link, via the inner product decoder (see Section Methods). Nevertheless, these results show that when designing the GNN embeddings for the link prediction task, instead of trying to predict subject areas like Kipf and Welling (2017), we are able to capture the Matthew effect in our embedding. This is an important conclusion for future research, because it shows the way in which we can use the embeddings framework for studying the Matthew effect in science. In addition, we found that the semantic embeddings do not encode this phenomena. This is an important finding for predictive modelling in which, if the Matthew effect is encoded in the embedding, it would imply reinforcing inequalities.

[11] See Figure E.1 in the Online Resource

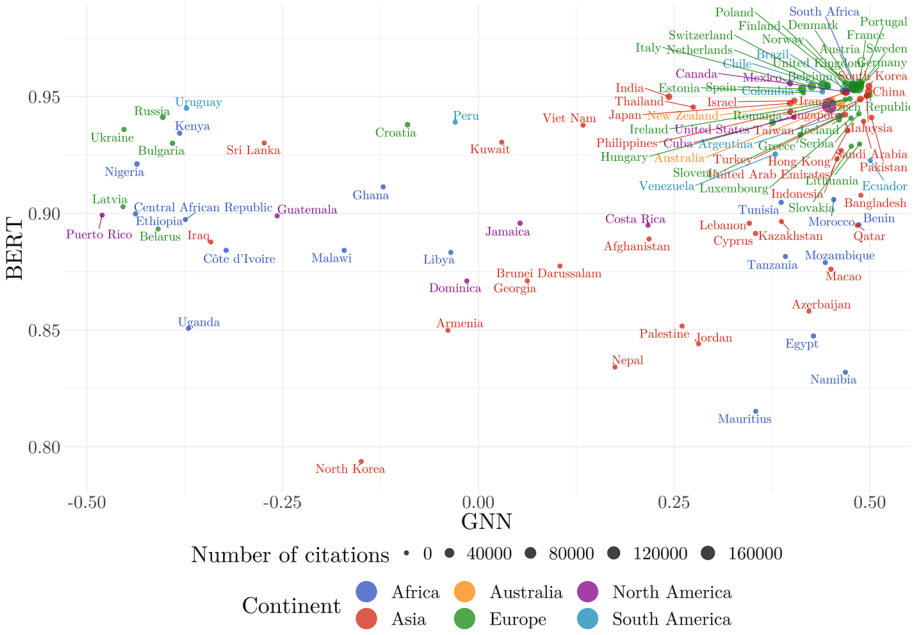

Fig. 7 Average cosine similarity by countries. BERT and GNN. Size by number of citations

Country-level analysis

In the same way we built a BERT embedding by averaging the word embeddings of each document, we can build a hierarchical representation of entities by averaging its components. One of the dimensions of analysis is the role of countries in the field of Science of Science. For this, we took the first author's organisational affiliation to ascribe a geographical location to an article. This does not necessarily mean that an article has been written in that country, but it gives us a proxy for the geographical distribution of scientific work and allows us to reconstruct the average position of countries in the embedding. Using the cosine similarity between countries, Figure 7 shows the average similarity of a country with respect to all others in the GNN (horizontal axis) and BERT (vertical axis) embeddings. This means that we are comparing the semantic proximity on the vertical axis against the structural network-based proximity on the horizontal axis. Results show that there is a centre of gravity of science production (Zhang et al. 2015) that includes most of the English-speaking countries, (Western) Europe and (East) Asia. Close to the core, we can also find some countries from (South) America and (East) Europe. South Africa is the only country from its continent close to the centre, which might be an indication for research activities of the Centre for Research on Evaluation Science and Technology (CREST) at Stellenbosch University[12]. Results show that the BERT cosine similarity is almost always higher than 0.8, while the GNN ranges between −0.5 and 0.5. This means that the semantic representation is in general very similar between all countries, while in

[12] See http://www0.sun.ac.za/crest/

the structural representation countries are never too close, and many of them are even in the opposite direction of most of the other countries. The presented results can be interpreted as follows: While researchers in Science of Science from all countries, within these journals, work on more or less similar content, the relevance that the academic community gives to their work is highly skewed. For example, in the case of Uruguay, the average BERT cosine similarity, i.e., based on the textual content of the articles, from this country and all others, is almost 0.95, a really high value considering cosine similarity moves between −1 and 1. On the other hand, its citation-based cosine similarity is less than −0.35, which means that it is in an opposite direction with respect to most of the countries. As we mentioned in Section Data set and network statistics, this analysis is limited due to the limits of the data set. We cannot fully account for scientific production outside the countries that appear here as peripheral. Including journals from other regions and languages would most probably change the layout of the results, specially for the semantic embeddings (Beigel 2014). In that sense, we have to limit the scope of interpretation to the fact that, *within these journals*, the topics discussed do not vary much. Nevertheless, this result is inline with many other studies in the field on the unequal distribution of citations, at least in international journals (Bonitz et al. 1997; Demeter and Toth 2020; King 2011; Merton 1974).

This analysis answers our first research question. If we use GAE with the GCN, we can encode the *relational* dimension of articles. With the analysis of collaboration patterns, the Frobenius norm, and the country-level analysis, we can confirm that the idea of *prestige* is captured by the GNN embeddings. This concept unfolds into different expressions, such as the different position articles have in the embedding according to their collaboration patterns and citation levels, and also on hierarchical levels of analysis, like the distribution of countries.

Projection of journals on the epistemic spectrum

Word embeddings have shown impressive results on analogy tasks. Mikolov et al. (2013c) show that word embeddings can solve the task 'man is to woman as king is to _____' by doing $\overrightarrow{\text{king}} + \overrightarrow{\text{woman}} - \overrightarrow{\text{man}}$, which returns a vector close to $\overrightarrow{\text{queen}}$. This implies that there is a latent dimension of *gender* in the word embedding, that can be reconstructed by the subtraction $\overrightarrow{\text{woman}} - \overrightarrow{\text{man}}$. Kozlowski et al. (2019) suggest to use analogies for the description of social dimensions. Kang and Evans (2020) applied this to the field of Science of Science. The authors defined two word embeddings and compare how different concepts, like 'theory' and 'measure' were projected into the 'good-bad' dimension, based on a selection of journals assigned to the quantitative and qualitative research communities.

For our analysis, we use the analogies approach proposed by Mikolov et al. (2013c) to study the differences between journals. But instead of building different embeddings based on a pre-clasification of journals like Kang and Evans (2020) did, we defined a single data set and built the analogy on two pivotal journals. The 'quantitative-qualitative' division of science is an open discussion (Leydesdorff et al. 2020; Weber 2004) and a purely quantitative approach does not seem to be able to close it. Given this, our analysis does not intend to prove that there is such a division. A more accurate interpretation is to simply think the analogy as the *latent dimension that separates epistemic practices from two journals*, being

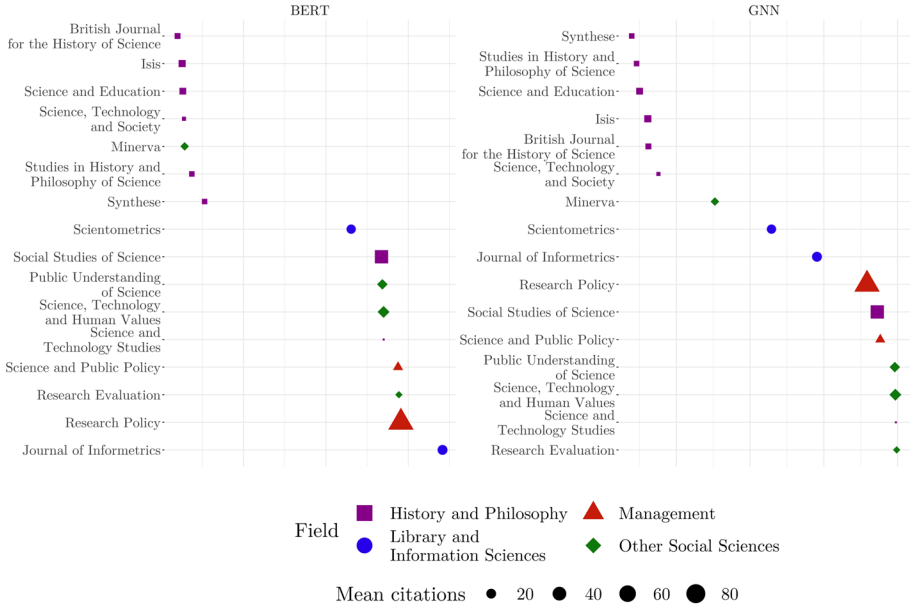

Fig. 8 Cosine similarity with the *Journal of informetrics-ISIS* dimension

this either methodological, epistemic, ontological, or simply related to a different vocabulary typically used in different research fields.

For this, we first generated a vector representation of each journal in our data set, in the same way we built the representation of countries. Then, we selected two journals as pivots for building the latent dimension. The selection of the pivotal journals is necessarily an arbitrary one, but compared with the approach proposed by Kang and Evans (2020), we do not need to previously assign each journal to one of the two poles. After this, we projected the other journals to this latent dimension using cosine similarity, which brings the journals projection closer to one of the two poles. If the way in which journals order themselves in this dimension seems to be random, then there is no latent dimension along the two pivotal journals. In Figure 8, we consider this exercise using the journals *ISIS* and *Journal of Informetrics* as pivots. This latter defines its scope on 'research on quantitative aspects of information science'[13]. *ISIS* is a long standing journal on 'history of science, medicine, and technology and their cultural influences'[14]. In the BERT embedding we can see that the *Journal of Informetrics* is on one side of the extremes, and on the other side we have *British Journal for the History of Science*. We can also see that there is a strong division of journals along the axis, with seven journals really close to the *ISIS* pole and nine close to the *Journal of Informetrics* pole. Almost all journals at the *ISIS* pole are from *History and Philosophy of Science*, except *Minerva*, a multidisciplinary journal. All journals from *Management* and *Library and Information Sciences* are on the *Journal of Informetrics* pole, although the journal *Scientometrics* is not as close to the extreme as we would have expected. Except for *Minerva*, all other journals from *Other Social Sciences* are also

[13] See https://www.journals.elsevier.com/journal-of-informetrics.

[14] See https://www.journals.uchicago.edu/journals/isis/about.

at the *Journal of Informetrics* pole. On the other hand, the same exercise using the GNN embedding, shows a different result. Both journals from the field of *Library and Information Sciences* are together in the middle range, indicating that the *Journal of Informetrics-ISIS* dimension is poorly defined on this embedding, showing that the GNN embedding is not capturing the semantic information as well as the BERT embedding. If we consider the mean citations by journal in the GNN embedding, there seem to be more correlation with this, than that with the epistemic differences by journal. Given this, we conclude that while the BERT embedding correctly capture this phenomenon, the GNN embedding is driven by the relational structure of the citation network, rather than the epistemic content of articles, and therefore is not a proper tool for this type of analysis.

The outlined results answer research question two: Using text-based embeddings, we can encode a *semantic* representation at the article level. Using analogies between journals, we can reflect the epistemic difference between them and project other journals, or articles onto that spectrum.

Conclusion

In this paper, we explored the use of *embeddings* as representations of research articles. We presented an overview of techniques designed on the different elements that compose an article: its text, metadata, and citations. The objective of this article was to study the use of new methodologies that are currently being developed in the field of computer science to analyse journals and research articles in the field of Science of Science.

In Section Methods, we presented two approaches for building an embedding space of articles: the NLP techniques, and the GNN techniques. In Section Evaluation of embeddings, we found that, using textual content in document embeddings enables us to build a semantic space. In Section Evaluation of embeddings we also evaluated the performance of the different models, and concluded that, we can use the network structure in a GNN to define a relational space that embed the social relations underneath. We also presented an extensive comparative analysis within each group of models to find the best performing architecture and the set of hyper-parameters. Our results show that for semantic embeddings, the BERT model gives the most clear results, while for the link prediction task on the direct citation network, the GAE with GCN, using the BERT embedding along with the metadata features, gives the best performance. In Section Comparing the differences between the relational and semantic spaces, we compared the semantic and relational spaces along four different dimensions: collaboration patterns, the country-level analysis, the Mathew effect, and the journals epistemic practice division. We found that the relational space captures the difference on collaboration status, citation levels, and aggregated levels of analysis like differences between countries. The semantic space captures phenomena related with articles topics, their corresponding journal, and epistemological aspects, like the journals epistemic practice division.

These are promising results, because they open the possibility for many different in-depth analysis using this different techniques, according to the respective research question. A responsible use of embeddings as tools for analysis does not imply that we should not use those embeddings that reproduce the existing bias, but that we should acknowledge it. On the one hand, a recommendation system based on GNN would reproduce the biased attention researchers give to different articles. On the other hand, using GNN can be of great help to better understand the biases themselves. These biases, as shown in the

country-level analysis, might not only be a reflection of the scientific community, but are likely related to the chosen Scopus database and the sampling of journals and articles itself. This paper is based on a small data set in the field of Science of Science, and therefore represents a case study. More research on other fields and cross-disciplinary data sets have to be made to explore if the present results are also valid in other fields of research as. The main contributions of this paper are fourfold:

1. **Data**: We have built a data set with 16 core journals from the field of Science of Science, from a range of disciplines. The data set consists of 22,151 research articles, including their metadata (abstract, title, keywords, authors, organisational affiliations of the authors, year of publication, and journals subject areas), the results of the LDA model (the topics distribution of each article), and the cumulative citations from the network. From this data set, we built a network of direct citations with 16,578 nodes and 68,797 links. This data set is suitable for models that work with metadata features, NLP, and networks, allowing us to compare the performance of different approaches[15].
2. **Semantic Modelling**: We performed three different approaches for sentences embedding of titles, abstracts, and keywords: Doc2Vec, LDA and BERT, and compared their performance as features for a link prediction model. We presented the Topic Model results as an interactive plot which give an overview of the topics discussed within the data set. We came to the conclusion that for a data set of these characteristics it is more powerful to use a BERT model, because it can benefit from a pre-training on a large corpus.
3. **Relational Modelling**: We trained the GNN on the link prediction task and made an extensive comparison of multiple possible encoders, using GNN layers that are currently the state of the art in computer sciences. We show that the GCN is the best performing architecture for this task in the present context. We also performed ablation studies to find that the BERT encoding of text and the cumulative citations are the most relevant features for the model.
4. **Model Comparison**: We compared the latent information in the different types of embeddings and arrived to the conclusion that the textual embedding generates a semantic space, while the GNN generates a relational space. This is an important methodological conclusion, as it is a distinction of which type of modelling should be used given the research questions.

Multiple recommendations for future research arise from our study: First, methodological research needs to be done over other GNN architectures. In this article, we presented GNN models for the link prediction task, but node prediction tasks are also suitable for this network. Predicting the article's journal, would generate an embedding representation much closer to the semantic space. Predicting the article's author with GNN can be an important improvement for the author-name disambiguation problem (Schulz et al. 2014). Direct citation networks are not the only network structure to be considered. Second, the network emerging from bibliographic coupling and co-citation can be compared with the results from direct citations. Third, the study of co-authorships and mobility networks are

[15] The DOI of the articles that compose the data set will be available, together with the code for implementing the models in this article, at https://github.com/DiegoKoz/scisci upon publication, and the full data set is available upon request and approval of Scopus. Sharing the code and data aims to improve the reproducibility of this work.

promising lines of research, to explore how communities of co-authors and institutions are organised in the embeddings. Fourth, the use of this methodology can generate important new insights for the field of Science of Science: The flexibility of the low-level representations allows a quantitative approach to many research questions. Problems like the Matthew or the Mathilda effects (Rossiter 1993) can be approached using the cosine similarity analysis presented in this article. The time dynamics in this phenomena could be studied by simply splitting the data set in decades and building multiple embeddings, as in Garg et al. (2018). The field classification on the journal level can be a problematic task. Embeddings methodologies could be used for field classifications on article level. Finally, the development of methodologies based on relational embeddings enable researchers to detect gender and ethnic biases, and have potential leading to policy recommendations.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11192-021-03984-1.

Acknowledgements The authors thank Cassidy R. Sugimoto and Vincent Larivière for their helpful input on the selection of journals for the data set. We also thank Justin J.W. Powell, Marcelo Marques, and Mike Zapp for their valuable comments on previous versions of the manuscript.

Funding This work has been supported by the Doctoral Training Unit *Data-driven computational modelling and applications* (DRIVEN, https://driven.uni.lu), which is funded by the Luxembourg National Research Fund under the PRIDE programme (PRIDE17/12252781).

Data availability The DOI of the articles and their embeddings is available on the website: https://github.com/DiegoKoz/scisci.

Code availability The code used for this project is available at https://github.com/DiegoKoz/scisci.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

Adams, J. (2013). The fourth age of research. *Nature, 497*(7451), 557–60.

Agarap, A. F. (2018). Deep learning using rectified linear units (relu). arXiv: 1803.08375.

Allingham, J. (2020). Latex-tikz-diagrams.github.com/JamesAllingham/LaTeXTikZ-Diagrams.github.com/JamesAllingham/LaTeX-TikZ-Diagrams.

Barabàsi, A.-L. (2016). *Network science*. New York: Cambridge Uiversity Pess.

Beigel, F. (2014). Introduction: Current tensions and trends in the world scientific system. *Current Sociology, 62*(5), 617–625.

Bellman, R. (1966). Dynamic programming. *Science, 153*(3731), 34–37.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics, 1*(1), 17–35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3,* 993–1022.
Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2007). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5,* 135–146.
Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Proceedings of the 30th conference on neural information processing systems, 30,* 4349–4357.
Bonitz, M., Bruckner, E., & Scharnhorst, A. (1997). Characteristics and impact of the matthew effect for countries. *Scientometrics, 40*(3), 407–422.
Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H.-D. (2008). Citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics, 8*(1), 93–102.
Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology, 61*(12), 2389–2404.
Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. *The American Statistician, 72*(1), 2–10.
Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of machine learning research, 81,* 77–91.
Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186.
Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2016). Fast and accurate deep network learning by exponential linear units (ELUs). In Proceedings of the 4th international conference on learning representations, ICLR.
Daenekindt, S., & Huisman, J. (2020). Mapping the scattered field of research on higher education. A correlated topic model of 17,000 articles 1991–2018. *Higher Education, 80*(3), 571–587.
Davis, G. F., Yoo, M., & Baker, W. E. (2003). The small world of the american corporate elite, 1982–2001. *Strategic Organization, 1*(3), 301–326.
de Solla Price, D. J. (1963). *Little science, big science*. New York: Columbia University Press.
Demeter, M., & Toth, T. (2020). The world-systemic network of global elite sociology: The western male monoculture at faculties of the top one-hundred sociology departments of the world. *Scientometrics, 124*(3), 2469–2495.
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pretraining of deep bidirectional transformers for language understanding. In Proceedings of the conference of the north american chapter of the association of computational linguistics (pp. 4171-4186). Minneapolis, Minnesota.
Erdös, P., & Rényi, A. (1960). On the evolution of random graphs. Institute of Mathematics. *Hungarian Academy of Sciences, 5*(1), 17–60.
Fey, M., & Lenssen, J. E. (2019). Fast graph representation learning with PyTorch geometric. In Proceedings of the 7th international conference on learning representations, ICLR.
Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., & Barabási, A.-L. (2018). Science of science. *Science, 359*(6379).
Fox, J., & Weisberg, S. (2018). *An r companion to applied regression*. United States: Sage publications.
Gao, H., & Ji, S. (2019). Graph u-nets. In Proceedings of machine learning research (Vol. 97, pp. 2083-2092). Long Beach, California, USA: PMLR.
Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science, 178*(4060), 471–479.
Garfield, E., & Merton, R. K. (1979). *Citation indexing: Its theory and application in science, technology, and humanities*. New York: Wiley.
Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences, 115*(16), E3635–E3644.
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge: MIT Press.
Hamilton, W. L., Ying, R., & Leskovec, J. (2017a). Inductive representation learning on large graphs. In Proceedings of the 30th neural information processing systems conference (pp. 1024-1034).
Hamilton, W. L., Ying, R., & Leskovec, J. (2017b). Representation learning on graphs: Methods and applications. *IEEE Data Engineering Bulletin, 40*(3), 52–74.
Iyer, B., Lee, C.-H., & Venkatraman, N. (2006). Managing in a "small world ecosystem": Lessons from the software sector. *California Management Review, 48*(3), 28–47.
Jeong, C., Jang, S., Park, E., & Choi, S. (2020). A context-aware citation recommendation model with BERT and graph convolutional networks. *Scientometrics, 124*(3), 1907–1922.

Jurafsky, D., & Martin, J. H. (2008). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (2. ed). Prentice Hall series in artificial intelligence. Upper Saddle River, NJ: Prentice Hall.
Kang, D., & Evans, J. (2020). Against method: Exploding the boundary between qualitative and quantitative studies of science. *Quantitative Science Studies, 1*(3), 930–944.
Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*(1), 10–25.
King, D. A. (2004). The scientific impact of nations. *Nature, 430,* 311–316.
King, R. (2011). Power and networks in worldwide knowledge coordination: The case of global science. *Higher Education Policy, 24*(3), 359–376.
Kipf, T. N., & Welling, M. (2016). Variational graph auto-encoders. In Proceedings of the nips workshop on bayesian deep learning.
Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In Proceedings of the 5th international conference on learning representations (ICLR).
Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review, 84*(5), 905–949.
LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation, 1,* 541–551.
Leydesdorff, L., Ràfols, I., & Milojević, S. (2020). Bridging the divide between qualitative and quantitative science studies. *Quantitative Science Studies, 1*(3), 918–926.
Lillquist, E., & Green, S. (2010). The discipline dependence of citation statistics. *Scientometrics, 84*(3), 749–762.
Merton, R. K. (1974). The sociology of science: Theoretical and empirical investigations (4. Dr.). Chicago: University of Chicago Press.
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). *Efficient estimation of word representations in vector space.*arXiv:1301.3781.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th neural information processing systems conference (pp. 3111-3119).
Mikolov, T., Yih, W.-t., & Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In Proceedings of the 2013 conference of the north American chapter of the association for computational linguistics: Human language technologies (pp. 746-751).
Milojević, S. (2015). Quantifying the cognitive extent of science. *Journal of Informetrics, 9*(4), 962–973.
Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review, 69*(2), 213–238.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12,* 2825–2830.
Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543). Doha, Qatar: Association for Computational Linguistics.
Persson, O., Glänzel, W., & Danell, R. (2004). In ationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics, 60*(3), 4210–432.
Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In Proceedings of the LREC. (2010). *workshop on new challenges for NLP frameworks* (pp. 45–50). Valletta: Malta.
Rossiter, M. W. (1993). The matthew matilda effect in science. *Social Studies of Science, 23*(2), 325–341.
Schulz, C., Mazloumian, A., Petersen, A. M., Penner, O., & Helbing, D. (2014). Exploiting citation networks for large-scale author name disambiguation. *EPJ Data Science, 3*(1), 11.
Schwemmer, C., & Wieczorek, O. (2020). The methodological divide of sociology: Evidence from two decades of journal publications. *Sociology, 54*(1), 3–21.
Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70). Baltimore, Maryland, USA.
Slapin, J. B., & Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science, 52*(3), 705–722.
Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*(4), 265–269.
Sooryamoorthy, R. (2009). Do types of collaboration change citation? collaboration and citation patterns of south african science publications. *Scientometrics, 81,* 177–193.
Sutskever, I., Martens, J., & Hinton, G. E. (2011). Generating text with recurrent neural networks. In Proceedings of the 2011 international conference on machine learning (Vol. 28).

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 4593-4601). Florence, Italy: Association for Computational Linguistics.

Thekumparampil, K. K., Wang, C., Oh, S., & Li, L.-J. (2018). Attention-based graph neural network for semi-supervised learning. arXiv:1803.03735.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9,* 2579–2605.

Van Raan, A. F. J. (1998). The influence of international collaboration on the impact of research results: Some simple mathematical considerations concerning the role of self-citations. *Scientometrics, 42*(3), 423–428.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & . . . Polosukhin, I. (2017). Attention is all you need. In Proceedings of neural information processing systems conference, 30 (pp. 5998-6008).

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. *Proceedings of the International Conference on Learning*. Representations.

Weber. . (2004). Editor's comments: The rhetoric of positivism versus interpretivism: A personal view. *MIS Quarterly, 28*(1), iii.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., & ...Schwartz, O. (2018). Ai now report. AI Now Institute at New York University New York.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., & ...Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. arXiv:1910.03771.

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How powerful are graph neural networks? *Proceedings of the international conference on learn- ing*. representations.

Zhang, L., Powell, J. J., & Baker, D. P. (2015). Exponential growth and the shifting global center of gravity of science production 1900-2011. *Change: The Magazine of Higher Learning, 47*(4), 46–49.

Zhang, M., & Chen, Y. (2018). Link prediction based on graph neural networks. In Proceedings of the 32nd conference on neural information processing systems (Vol. 31, pp. 5171-5181).

Zhang, Y., Zhao, F., & Lu, J. (2019). P2v: Large-scale academic paper embedding. *Scientometrics, 121*(1), 399–432.