Noname manuscript No.
(will be inserted by the editor)

Snippets left out from the article of Science of Science

Diego Kozlowski

1 Results

1.1 Comparing Differences between the Semantic and Relational Space

After studying the overall quality of the different models, we 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 show how the embedding representation of articles change in the semantic and relational space. For this, we compare the results on four different topics largely studied in the science of science field: First, the representation of collaboration patterns; second, the Matthew effect; third, we perform a country level analysis; and finally, the quantitative-qualitative divide in the field. The goal is to check if this different phenomenon are encoded in the resulting embedding, and how their representations differs within the different proposed models.

The Matthew effect The Matthew effect (Merton 1974) states that highly cited articles have a higher chance of being cited again. Figure **??** reflects on this via the collaboration patterns, while Figure **??** also shows a correlation of articles by the number of citations. Given that the decoder function of the GNN is the inner product of the embedding (see Section **??**), a higher norm in the vector representation of an article will correspond with a higher link probability, i.e. citation relation, with all other articles. 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 dataset, by the quartiles of the number of citations distribution, and those with zero citations. Those articles with lower citations than the 25% threshold correspond to the group *lower*, those between 25% and 50% to the *mid-low*, between 50% and 75% to *mid-high* and those articles with more than the 75% threshold belong to the group

high. When calculating the average Frobenius norm of articles' embedding at these citation levels in the different models, we confirm that while the GNN generates a higher value for the highly cited articles, the BERT, Doc2Vec and LDA models don't follow this pattern[1]. This implies that the GNN is able to encode in the embedding the Matthew effect, but this cannot be seen when the embeddings are exclusively based on text.

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 science of science community. 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 1 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 semantical 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 (L. 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. Results also 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 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 this 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 the Section **??**, this analysis is limited due to the limits of the dataset. 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 this 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

[1] In Appendix D we show a complementary figure for this analysis.

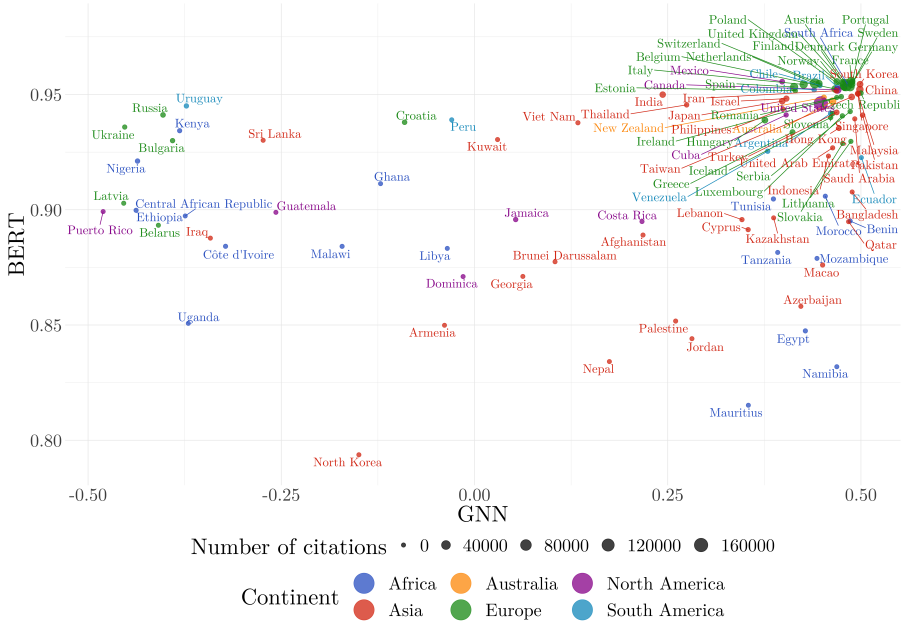

Fig. 1 Average cosine similarity by countries. BERT and GNN. Size by number of citations.

many other studies in the field on the unequal distribution of prestige, at least in the international journals (Bonitz et al. 1997; Demeter et al. 2020; R. King 2011; Merton 1974).

With the analysis of collaboration patterns, the Frobenius norm, and the country-level analysis, we can answer the research question RQ3: How is the concept of *prestige* expressed in the articles' embeddings? The idea of *prestige* is strongly 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.

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*, arXiv:1803.08375.

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

Bacciu, D., Errica, F., Micheli, A., & Podda, M. (2020). A gentle introduction to deep learning for graphs. *Neural Networks*, *129*, 203–221.

Barabási, A.-L. (2016). *Network science*. 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, In *Proceedings of the 30th conference on neural information processing systems*, Barcelona, Spain.

Bonitz, M., Bruckner, E., & Scharnhorst, A. (1997). Characteristics and impact of the matthew effect for countries. *Scientometrics*, *40*(3), 407–422.

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.

Bruna, J., Zaremba, W., Szlam, A., & Lecun, Y. (2014). Spectral networks and locally connected networks on graphs, In *Proceedings of the 2nd international conference on learning representations, ICLR*.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification, In *Proceedings of machine learning research*, New York, NY, USA, PMLR.

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.

Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering, In *Proceedings of the 29th annual conference on neural information processing systems*.

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.

de Solla Price, D. J. (1963). *Little science, big science* (Vol. 5). Columbia University Press New York.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding, In *Proceedings of the conference of the north american chapter of the association of computational linguistics*, Minneapolis, Minnesota.

Erdős, 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*.

Gao, H., & Ji, S. (2019). Graph u-nets, In *Proceedings of machine learning research*, 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* (Vol. 8). Wiley New York.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Grover, A., & Leskovec, J. (2016). Node2vec: Scalable feature learning for networks, In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, San Francisco California USA, ACM.

Hamilton, W. L., Ying, R., & Leskovec, J. (2017a). Inductive representation learning on large graphs, In *Proceedings of the 30th neural information processing systems conference*.

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). 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.

Kelley, H. J. (1960). Gradient theory of optimal flight paths. *Aerospace Research Central*, *30*(10), 947–954.

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. *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.

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature News*, *504*(7479), 211.

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.

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. (2013). Efficient estimation of word representations in vector space. *arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality, In *Proceedings of the 26th neural information processing systems conference*.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). 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*.

Milojević, S. (2015). Quantifying the cognitive extent of science. *Journal of Informetrics*, *9*(4), 962–973.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas,

J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & 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)*, Doha, Qatar, Association for Computational Linguistics.

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: Online learning of social representations, In *Proceedings of the 20th iternational conference on knowledge discovery and data mining - KDD*, New York, New York, USA.

Persson, O., Glänzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, *60*(3), 421–432.

Rehurek, R., & Sojka, P. (2010, May). Software framework for topic modelling with large corpora, In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*, Valletta, Malta.

Rossiter, M. W. (1993). The matthew matilda effect in science. *Social Studies of Science*, *23*(2), 325–341.

Scarselli, F., Gori, M., Ah Chung Tsoi, Hagenbuchner, M., & Monfardini, G. (2009). The graph neural network model. *IEEE Transactions on Neural Networks*, *20*(1), 61–80.

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*, 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*.

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., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need, In *Proceedings of neural information processing systems conference, 30*.

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

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., ... Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv:1910.03771*.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 1–21.

Xie, Y. (2014). Undemocracy: Inequalities in science. *Science*, *344*(6186), 809–810.

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How powerful are graph neural networks?, In *Proceedings of the international conference on learning 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 (S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett, Eds.). In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Proceedings of the 32nd conference on neural information processing systems*.

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

Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2018). Graph neural networks: A review of methods and applications. *arXiv:1812.08434*.

Table A1 Glossary of acronyms

Meaning	Acronym
Natural Language Processing	NLP
Document-Term Matrix	DTM
Graph Neural Networks	GNN
Term Frequency - Inverse Document Frequency	TF-IDF
Latent Dirichlet Allocation	LDA
Recurrent Neural Network	RNN
Convolutional Neural Networks	CNN
Convolutional Graph Networks	CGN
Graph Convolutional Networks	GCN
Graph Isomorphic Network	GIN
Graph Attention Network	GAT
Attention-based Graph Neural Networks	AGNN
Graph Autoencoder	GAE
Area Under the Receiver Operating Characteristic Curve	AUC
Average Precision	AP
True Positive Rate	TPR
False Positive Rate	FPR

Appendix A Glossary

Table B1 Articles out of the network summary statistics.

Field	Journal	Articles Retrived	Mean Citations	Max Citations
Management	Research Policy	643	79.72	3404
	Science And Public Policy	637	10.31	409
Library and Information Sciences	Scientometrics	784	19.73	435
	Journal Of Informetrics	12	13.50	28
History and Philosophy	Synthese	1702	6.70	564
	Studies In History And Philosophy Of Science	372	7.68	63
	Isis	250	9.09	123
	Science And Education	250	10.92	298
	British Journal For The History Of Science	145	9.03	54
	Social Studies Of Science	139	24.71	648
	Science, Technology And Society	134	4.15	47
	Science And Technology Studies	8	2.25	8
Education, Communication and Anthropology	Public Understanding Of Science	170	25.32	416
	Research Evaluation	142	6.15	76
	Science, Technology And Human Values	106	19.51	272
	Minerva	79	7.24	78
	Total	5573	16.00	3404

Appendix B Data statistics. In and Out of Network

In this section, we split the table **??** between those articles that are part of the network in Table B2, and those that have no information about references and are not referenced by any other article, in Table B1. We can see that 75% of articles are part of the network, and compared to those out of the network, they have a higher mean citation. Most of the articles that cannot be included in the network are from 'Synthese' (30%), 'Research Policy'(12%), or 'Science and Public Policy' (11%).

Table B2 Articles in the network summary statistics.

Field	Journal	Articles Retrived	Mean Citations	Max Citations
Management	Research Policy	2578	84.75	4820
	Science And Public Policy	1070	15.03	462
Library and Information Sciences	Scientometrics	4352	20.10	1334
	Journal Of Informetrics	864	22.76	352
History and Philosophy	Synthese	2449	9.80	910
	Social Studies Of Science	930	43.38	4709
	Science And Education	784	11.82	177
	Studies In History And Philosophy Of Science	539	9.50	145
	Isis	273	15.57	415
	Science, Technology And Society	211	7.28	122
	British Journal For The History Of Science	131	10.16	88
	Science And Technology Studies	103	5.52	39
Social Sciences: Education, Communication and Anthropology	Public Understanding Of Science	807	26.04	518
	Science, Technology And Human Values	651	35.04	828
	Research Evaluation	524	15.04	223
	Minerva	312	18.86	624
	Total	16578	21.92	4820

Appendix C Models definitions

C.1 LDA

As a generative Bayesian model, LDA states a generative process in which data is created, and then uses Bayes theorem to fit the parameters. The generative process is as follows:

- Each topic is generated as a multinomial distribution, β_i, over words
- For each document in the corpus, the words are defined in a two-step process:
  1. First, we define the distribution of topics in the document as a multinomial distribution, z,
  2. for each word:
     - randomly select a topic from the given z distribution,
     - given the topic, randomly select the word from the corresponding β distribution.

If we have a dictionary of V possible words, n documents, and k topics. β will be a matrix of kxV, where each row is the realisation of a Dirichlet process, i.e., a stochastic process where its realisations are multinomial distributions. β will indicate the distribution of words over topics. Z will be a matrix of nxk which will indicate the distribution of topics over documents. β and Z are the desired outputs, but the corpus only gives us the actual words of documents. If we consider these documents as a realisation of this chained random processes, we can use the Bayes Theorem to infer the probabilistic distributions:

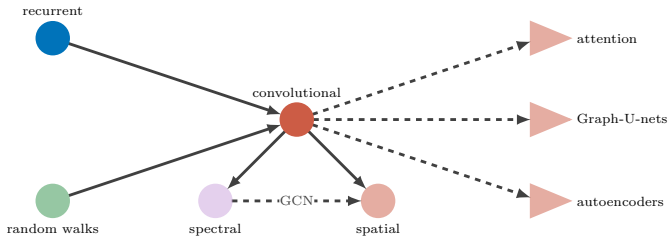

Fig. C.1 Graph Neural Networks framework

$$p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)}$$

where θ is the Dirichlet process that defines the distribution of topics over documents, z, and α is its parameter.

C.2 GNN Models

Figure C.1 shows a synthesis of the different approaches taken to solve this issue. The first approaches were based on building sequences using random walks Perozzi et al. 2014, or recurrent models Scarselli et al. 2009. After this, Graph Convolutions were defined using spectral methods Bruna et al. 2014 and spatial methods Hamilton et al. 2017a. More recently, new architectures were proposed that incorporate attention mechanisms Veličković et al. 2018, U-nets Gao et al. 2019 and autoencoders Kipf et al. 2016. For the remaining of this section, we present these models' intuitions, for an in-depth literature review, we refer the readers to Bacciu et al. 2020; Hamilton et al. 2017b; Wu et al. 2020; Zhou et al. 2018.

C.2.1 First Approaches

In this section, we briefly present the first approaches for GNN. These models are not used in the subsequent experiments, as they are not currently the state of the art. Nevertheless, they are all conceptually important.

For the task of *node embedding*, given the developments on Word Embeddings Mikolov, Yih, et al. 2013, one of the first strategies proposed was to use random walks over nodes to define a sequence that can be later used as the input for a Word2Vec model, as it is normally done with texts on NLP. The first model that proposed this technique was DeepWalk Perozzi et al. 2014. Later, Grover et al. 2016 proposed node2vec, which defines flexible biased random walks, that includes parameters for adjusting the path taken by the random walks to search for structural roles or community structures. The major problem with these approaches is that they do not consider the features of the nodes, so they miss potentially useful information.

Scarselli et al. 2009 proposed the Graph Neural Network model which iteratively updates the nodes' state looking at its neighbours, until it converges. This recurrent model uses a single layer which is iteratively updated.

Convolutional Graph Networks (CGN) models instead use a stack of layers. In this way, the number of updates is fixed, and the parameters of each layer are allowed to change, giving more flexibility to the model. Spectral-based methods were the first type of CGN Bruna et al. 2014. They use the *graph Fourier transform* on the Laplacian matrix (a normalised adjacency matrix), which can be thought of as the effect of a signal over the network. This model, while conceptually important, suffers from many problems. In particular, as it uses the full graph structure, it can only work on transductive settings, i.e., it cannot be used on a different graph on train and test. More importantly, the eigen-decomposition requires $\mathcal{O}(n^3)$ where n is the number of nodes. This is prohibitively expensive when the network has billions of nodes, like social networks.

C.2.2 State Of The Art

In this section, we present the current state of the art in GNN. A combination of these models will be used in the experimental analysis for the task of link prediction.

GCN Many models improve the limitations present in the spectral method proposed by Bruna et al. 2014. Defferrard et al. 2016 build an approximation of the original model with Chebyshev polynomials. Kipf et al. 2017 introduce the *Graph Convolutional Networks* (GCN), which further reduces the model and includes self-connections, this means that in the iterative process of building a representation based on its neighbours, the node will also look at itself, which is a desirable property. The GCN simplifies the model by only looking at the first-order neighbourhood. If the representation of a node is initiated with its feature vector, the GCN update builds an average representation based on itself, due to self-connections, and its neighbours. Instead of iterating the update step until convergence, like recurrent models, in GCN a stack of layers is built. The stacking of GCN layers allows the node representation to be based on more distant nodes. Here, simplicity is the key for building a powerful model.

GraphSage Hamilton et al. 2017a propose several variations over the GCN, in the *GraphSage* model. As the formulation of the problem can be useful for understanding how Graph Convolutional Networks works, we present their algorithm in 1 . The model needs the following inputs:

- The Graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with a list of vertices $\mathcal{V}$ and edges $\mathcal{E}$, and
- the input features x, where x_v is the feature vector of the node v.

We also need to define the number of layers the model will have, K, a set of weight matrices W^k for each layer (that will be later trained with the

Algorithm 1 GraphSAGE Hamilton et al. 2017a

Input: Graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$: input features x ; depth K ; weight matrices $W^k, \forall k \in 1, ..., K$; nonlinearity σ ; differentiable aggregator functions $AGGREGATE_k, \forall k \in 1, ..., K$; neighbourhood function $\mathcal{N} : v \rightarrow 2^{\mathcal{V}}$.

output Vector representations $z_v \forall v \in \mathcal{V}$.

1: $h_v^0 \leftarrow x_v, \forall v \in \mathcal{V}$;
2: **for** $k = 1 \dots K$ **do**
3: **for** $v \in \mathcal{V}$ **do**
4: $h_{\mathcal{N}(v)}^k \leftarrow AGGREGATE_k(\{h_u^{k-1}, \forall u \in \mathcal{N}(v)\})$;
5: $h_v^k \leftarrow \sigma(W^k CONCAT(h_v^{k-1}, h_{\mathcal{N}(v)}^k))$
6: $h_v^k \leftarrow h_v^k / ||h_v^k||_2, \forall v \in \mathcal{V}$
7: $z_v \leftarrow h_v^K, \forall v \in \mathcal{V}$

data), and an activation function σ. We also need a way in which a group node embeddings will be aggregate, and a way of defining the neighbourhood of a node.

The model starts the node embeddings with their feature vectors. After this, in each of the K layers, for each node $v \in \mathcal{V}$, it first defines its neighbours. One of the changes with respect to GCN is that GraphSage samples a fixed amount of neighbours to control the computational footprint. Given the neighbourhood, the $AGGREGATE$ function is used to build a new vectorised representation of those (line 4), and this is later concatenated with the embedding of node v in the previous layer (line 5). A projection is made with the W^k matrix, and also an activation layer is used. This correspond with the typical structure of a deep learning layer. When the model is fitted with back-propagation for a specific task, the W^k are updated in order to optimise the lost function Kelley 1960. Line 6 is simply a regularisation.

Depending on the election of the AGGREGATE and the CONCAT operators, this model is an approximation of GCN from Kipf et al. 2017. If we use the *mean* as an aggregation function, and instead of concatenating h_v^{k-1} and $h_{\mathcal{N}(v)}^k$ we average them, this is exactly the GCN model. The authors also propose two other aggregators: the *LSTM aggregator* which can be more expressive, although the LSTM model (a RNN) requires a sequential order in the inputs, so the authors need to define an ad hoc order for the neighbours. And the *pooling aggregator* in which the neighbours vector representation is fed through a fully connected layer and then the max-pooling is applied.

GIN Recent studies analyse the relation between GNN and the Weisfeiler-Lehman test of isomorphism (Xu et al. 2019). Two graphs are isomorphic if they are topologically identical. Besides the embedding representation of nodes, it is also possible to build embedding representations of the entire graph. In the previous framework, we would need to add a READOUT operation that takes the nodes embedding as input and generates a single embedding representation for the entire network. If for two graphs G_1 and G_2 that are non-isomorphic we can build a different embedding representation, we are able to distinguish between those two. Xu et al. 2019 proves that the GNN

power for discriminating between non-isomorphic networks is at most that of the Weisfeiler-Lehman test. However not every GNN can have this power, the *Graph Isomorphic Network* (GIN) is a proposal that achieves the maximum discriminative capacity by using a specific update function, which replaces the $AGGREGATE$ functions proposed by Hamilton et al. 2017a (*mean*, *LSTM* or *max*) with a *summation* over the neighbourhood. This model has the advantage of a theoretically robust decision on the AGGREGATE function.

GAT In GCN, nodes embeddings are updated using their neighbours embedding. Up to this point, all models consider that every neighbour has the same influence, which might not be true. *Graph Attention Networks* (GAT) Veličković et al. 2018 use attention mechanisms (which are currently state of the art in other problems, like NLP) to assign a different influence to each neighbour. The update of the node representation becomes:

$$h_v = \sigma(\sum_{u \in \mathcal{N}(v)} \alpha_{u,v} W h_u)$$

where W is still a learning weight matrix and $\alpha_{u,v}$ is the normalised attention between nodes u and v.

Following Vaswani et al. 2017, GAT uses *multihead attention*, which implies applying K independent attentions and concatenating their results (except for the final layer, where they are averaged).

This model has the advantage of learning the different importance of the neighbours, given their feature vector. Moreover, the attention mechanisms have the potential of an increasing interpretabilty of the mode. Thekumparampil et al. 2018 propose a variation of this model, *Attention-based Graph Neural Networks* (AGNN), where the *relevance* is defined based on the cosine similarity.

GraphUNet Gao et al. 2019 proposed GraphUNet, based on a new definition of Pooling layers. Convolutional layers in computer vision are normally used with *Pooling layers*. This is because the convolutional filters are trained to detect the presence of a specific feature in a portion of the image, and if the feature is found, the result will have high positive values. The max pooling layer, makes a downsizing of the input and captures the highest values. By doing this, it generates a clear indication whether or not that particular feature was present. However, as it happens with traditional convolutions, the pooling layers are defined based on the regular pattern of the image. Defining a pooling layer on the graph domain could be useful for building better representations of hierarchical patterns. Gao et al. 2019 proposed a new definition of pooling *gPool*. The gPool layer makes a linear projection of the nodes features, and a k-max pooling selection. With the identifier of the selected nodes, it builds the new (reduced) adjacency matrix and feature matrix. Gao et al. 2019 also proposes an unpooling layer, which rebuilds the original network,

and used together they can build an encoder-decoder architecture. The benefit of such an architecture is that it lets the node embedding be built based on the hierarchical properties of their neighbourhood.

Autoencoders The Autoencoder is an architecture in deep learning where the network tries to learn the input, but goes through a compressed state. The network can be divided into two elements:

- the encoder, which can be a regular stack of layers, that ends with a vectorised representation of the input,
- the decoder, where the representation received from the encoder will be sized up to the original form.

The idea is that, if the network is able to reconstruct the original input with a small error margin, then most of the information from the original input is correctly compressed in the low dimension at the end of the encoder.

Kipf et al. 2016 proposed the Graph Autoencoder (GAE) and the Variational GAE (a probabilistic implementation of the GAE), where the encoder can be any way of building a node embedding, like a stack of GCN layers, and the decoder is (for the GAE) the reconstructed adjacency matrix, $\hat{A}$:

$$\hat{A} = \sigma(ZZ^T), \qquad \text{with} \quad Z = GCN(X, A)$$

Where Z represents the embedding matrix built using the GCN and σ is a logistic sigmoid function. The model use the inner product of the node embeddings to reconstruct a probabilistic adjacency matrix. This transforms the embedding into edge probabilities for the given node pairs. This is a particularly useful architecture for the task of *link prediction*.

In this paper, we train our models for this task in a transductive setting. We randomly remove some citation links for test and validation set, and evaluate how well the reconstructed adjacency matrix can predict those removed links. We use as encoders the GCN, GraphSAGE, GIN, GAT, AGNN, and GraphUNet layers defined above, which constitute the current state of the art in the field.

Appendix D Frobenius Norm

In Figure D.1, we can see the Frobenius norm of the article's embedding by their citation level. For the GNN embedding we can see the higher norm of highly cited articles, while in the BERT embedding this does not sustain [2].

[2] Doc2Vec and LDA embeddings do not show significant differences between articles norms.

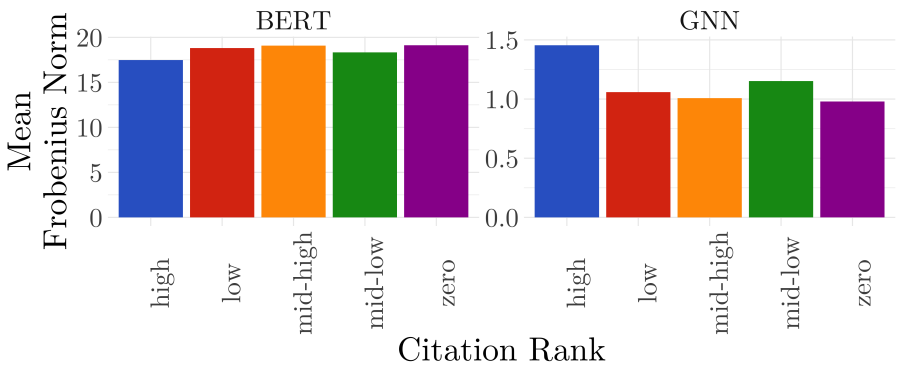

Fig. D.1 Frobenius Norm of articles by citation level.