# Graph Embedding for Mapping Interdisciplinary Research Networks

Eoghan Cunningham
School of Computer Science, University College Dublin
Dublin, Ireland
eoghan.cunningham@ucdconnect.ie

Derek Greene
School of Computer Science, University College Dublin
Dublin, Ireland
derek.greene@ucd.ie

## ABSTRACT

Representation learning is the first step in automating tasks such as research paper recommendation, classification, and retrieval. Due to the accelerating rate of research publication, together with the recognised benefits of interdisciplinary research, systems that facilitate researchers in discovering and understanding relevant works from beyond their immediate school of knowledge are vital. This work explores different methods of research paper representation (or document embedding), to identify those methods that are capable of preserving the interdisciplinary implications of research papers in their embeddings. In addition to evaluating state of the art methods of document embedding in a interdisciplinary citation prediction task, we propose a novel Graph Neural Network architecture designed to preserve the key interdisciplinary implications of research articles in citation network node embeddings. Our proposed method outperforms other GNN-based methods in interdisciplinary citation prediction, without compromising overall citation prediction performance.

## CCS CONCEPTS

• **Computing methodologies** → Learning latent representations.

## KEYWORDS

graph neural networks, citation graphs, interdisciplinarity

## 1 INTRODUCTION

Globally, the rate of scientific publication is accelerating rapidly. Such unprecedented growth necessitates the development of tools to support the search and recommendation of scientific research papers. Many of these methods hinge on research paper representation learning (or 'document embedding'). This refers to the task of discovering useful representations of scientific articles, typically for use in downstream applications like classification, prediction, and retrieval [4]. These methods rely primarily on Natural Language Processing (NLP) and Network Analysis techniques to learn document embeddings according to a paper's content, a paper's citation relations, or some combination of these sources [12].

In this work, we explore various approaches to scientific document representation from the perspective of mapping *interdisciplinary* research. We posit that document embeddings which preserve a work's interdisciplinary implications (or applications) are more desirable for downstream tasks, such as article recommendation or retrieval. The benefits of interdisciplinary research are well established [13]. However, due to the sheer volume of research published each day, combined with the recognised cognitive strain associated with reading outside of ones chosen and practiced discipline [17], engaging broadly with research from multiple disciplines is challenging. Accordingly, systems for research paper recommendation or retrieval that allow researchers to discover relevant works from beyond their own disciplines or communities must be developed. To this end, we assess a number of document embedding methods with respect to their ability to recognise and preserve the interdisciplinary implications of a research paper in their learned representations. Further, we propose a novel Message-Passing Graph Neural Network architecture, designed specifically for embedding interdisciplinary research. Motivated by work which highlights a relationship between certain citation graph substructures and interdisciplinary research [5, 6], our method permits the aggregation of information from neighbouring nodes in a manner that is sensitive to the local graph structure and additional relations between the neighbours. We evaluate this new method, referred to as Community-Based Sample and Aggregation method (*ComBSAGE*), via an interdisciplinary citation prediction task. The underlying hypothesis here is that those embeddings that are capable of recognising and preserving the *interdisciplinarity* of research articles will better predict citations between two papers that are distant in either the semantic or relational space of research papers.

## 2 BACKGROUND

Representation learning for scientific documents is the task of representing research papers in some dense vector space, such that the important similarities and relations between the papers are preserved i.e., semantically related papers should have similar representations (or *embeddings*) [12]. Scientific articles may be related if they pertain to the same topic (i.e., have similar content), or if there is an application or transfer of knowledge from one to the other (i.e., a citation). Accordingly, article content and citation relations represent the primary sources of information for scientific document representation. We limit our analysis to methods that rely on these sources, and propose that all of the approaches we include can be considered on the spectrum illustrated in Figure 1.
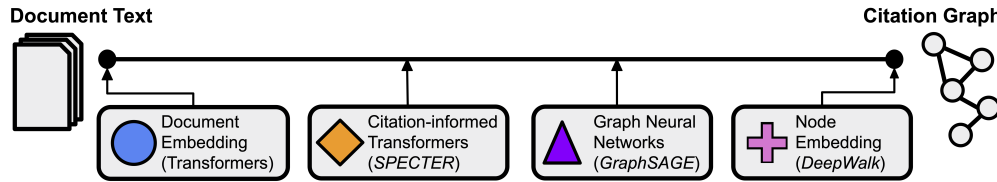
**Document Text**

**Citation Graph**

| Document Embedding (Transformers) | Citation-informed Transformers (*SPECTER*) | Graph Neural Networks (*GraphSAGE*) | Node Embedding (*DeepWalk*) |

**Figure 1: Spectrum of methods for research paper representation learning, with example methods listed.**

We place those methods that rely solely on article text on the left, those methods that rely only on the citation network on the right, and any combined or hybrid methods lie on the continuum in between. In the rest of this section, we discuss some of the important approaches to scientific document representation according to their position on this spectrum.

Large language models like *BERT* [7] were designed to be pre-trained on massive corpora of text, but fine-tuned for some specific task. For example, *SciBERT* [1] is a BERT-based model that is fine-tuned on the full text of a large sample of research papers from Semantic Scholar.

In the field of Network Analysis, *node embedding* refers to the task of learning low-dimensional vector representations for the set of nodes on a graph [2, 8]. These methods are primarily designed to preserve *proximity*, such that similar embeddings are learned for nodes which might belong to the same region or community in the graph. The *DeepWalk* [15] method, which extends the NLP Skip-Gram model [14], learns to predict those nodes that will co-occur on random walks on the graph, so that the resulting representations preserve proximities between nodes.

*SPECTER* [4] uses citation relations between papers as an external *inter-document signal* to fine tune text-based representations. Initialised with SciBERT and using title and abstract text from articles, SPECTER is trained to minimise the distance between the representations of a query paper and an article that it cites, while maximising the distance between the query paper and an article that it does not cite.

Message Passing Graph Neural Networks (MP-GNNs) (e.g. *GraphSAGE* [9], *GCN* [11], *VGAE* [10]) can be applied to graph structured data, where some additional information is available for the nodes. These methods can then learn node embeddings which depend not only on a node's descriptive features, but also the features of its neighbours. MP-GNNs iteratively update the representation of each node by receiving information from neighbouring nodes and learning aggregation functions to combine those messages with a node's current representation. In the context of a citation network, where a node's features are typically derived from the text of the corresponding paper, the representation of some focal paper is then informed not only by its own content, but that of the papers that it cites (the ideas upon which it develops), and even the articles that cite the focal paper (its applications or the work that builds on it).

In the context of interdisciplinary research papers, these may present in citation graphs as *bridges* or *bottlenecks* with both homophilic and heterophilic connections. This makes the task of representing interdisciplinary research more difficult from a Graph Learning perspective. Modelling bottleneck structures and heterophilic interactions remain open challenges in the field, hampered by representation learning phenomena like *over-squashing* [16] and

*over-smoothing* [19]. Therefore, in this paper we further develop the concept of MP-GNNs and propose a novel method of message aggregation, such that important local graph structures (e.g. co-citations) can be considered in an effort to learn article embeddings that will provide better representations for interdisciplinary works.

## 3 METHODS

In this section we present our novel Community-Based Sample and Aggregation (ComBSAGE) method for Message Passing Graph Neural Networks. The goal of this architecture is to incorporate local structural information in the message aggregation stage of MP-GNNs, in order to provide scientific document embeddings that are more appropriate when representing interdisciplinary research, without compromising the quality of paper representations overall.

The motivation behind this approach is as follows. Suppose we wish to aggregate messages from three neighbouring nodes, two of which are connected. We propose that these messages should be combined in a manner that accounts for this connection. Consider, for example, the task of representing a research paper which draws heavily on methods from the fields of mathematics and computer science, and that has a few recent applications in a field like political science. Such a paper presents in a citation network as a node with many connections to a large community of mathematics and computer science-related publications, and a smaller number of connections to works in political science. A traditional MP-GNN may lose much of the signal from the political science papers when aggregating messages. This relates to the *over-smoothing* phenomenon that has been highlighted as a weakness with many existing GNN architectures; where node representations converge and become indistinguishable from their neighbours [3], and equally to the *over-squashing* problem; where information is lost in nodes that represent bottlenecks in the network [16]. Our proposed method of message aggregation splits messages from distinct communities, such that they can be aggregated separately before being combined. We suggest that this approach allows the important interdisciplinary implications of a research paper (e.g. its applications in political science) to be preserved in its representation.

One layer of an MP-GNN is typically described in two functions: *aggregate* and *update*. For a node $v_i$, with features described by a vector $x_i$, we denote the neighbours of $v_i$ as $\{v_j, v_k, ..., v_p\}$ with features $\{x_j, x_k, ..., x_p\}$. An embedding for $v_i$ (denoted by $h_i$) is then generated via the update $h_i \leftarrow \phi(x_i, u_i)$, where $u_i \leftarrow \psi(\{x_j, x_k, ..., x_p\}, x_i)$. Thus, $\psi$ is some (typically permutation invariant) function for *aggregating* information from neighbouring nodes (potentially with consideration of $x_i$), and $\phi$ is some parametric function for *updating* the representation of a node using the aggregated message. MP-GNNs can be trained in a supervised fashion (e.g. in pursuit of some task such as node classification),
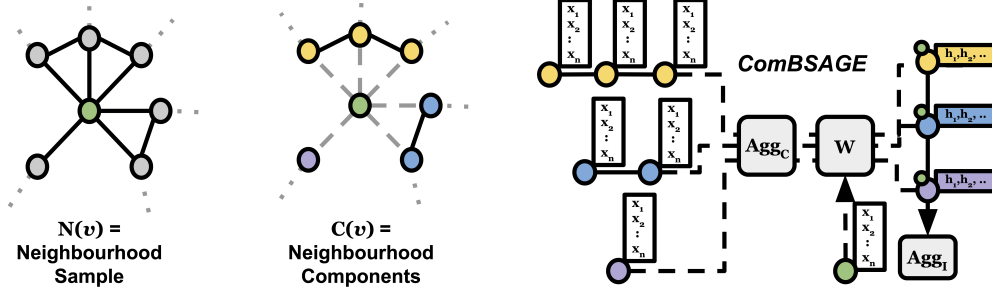
**Figure 2: ComBSAGE. Messages are grouped accord to $C$, and aggregated separately by $AGG_C$, before being combined by $AGG_I$.**

---

**Algorithm 1:** Community-based Aggregation

---

**Input** : Graph $G(V, E)$; input features $\{x_v, \forall v \in V\}$;
weight matrices $W_I$, $W_C$; non-linearity $\sigma$;
differentiable aggregator functions $AGG_C^k$ and
$AGG_I^k, \forall k \in 1...K$; neighbourhood function
$N : v \rightarrow 2^v$; neighbourhood component function
$C : v \rightarrow \{c, c \subseteq N(v)\}$

**Output**: Vector representations $z_v$ for all $v \in V$

1   **for** $k = 1...K$ **do**
2     **for** $v \in V$ **do**
3       **for** $c \in C(v)$ **do**
4         $h_c^k \leftarrow AGG_C^k(\{h_u^{k-1}, \forall u \in C\})$;
5         $h_c^k \leftarrow \sigma(W_I \cdot CONCAT(h_v^{k-1}, h_c^k))$;
6       **end**
7       $h_v^k \leftarrow AGG_I^k(\{h_c^k, \forall c \in C\})$;
8       $h_v^k \leftarrow \sigma(W_C \cdot CONCAT(h_v^{k-1}, h_v^k))$;
9     **end**
10   **end**
11   $z_v \leftarrow h_v^K, \forall v \in V$

---

or in an unsupervised manner where the objective task used for training is graph reconstruction or link prediction.

The proposed ComBSAGE method is outlined in Algorithm 1. The key contribution of our approach is the aggregation function ($\psi$), in lines 3 – 7 of Algorithm 1. The neighbourhood component function $C$ partitions the set of neighbours of some node $v$ into sets of connected components, i.e. $C(v)$ represents the connected components of the subgraph of $G$ induced by the set $N(v)$, (the neighbours of $v$). As such, if some node $v$ is part of a fully connected clique, then $C(v)$ contains only one set – $N(v)$. Conversely, if $v$ represents a hub node connecting $n$ otherwise isolated neighbours, then $C(v)$ contains $n$ sets, each containing a single neighbour. The ComBSAGE method then applies message aggregations first within each of the components of $C(v)$, and then between them. We illustrate the operation of the neighbourhood component function and a single layer of the ComBSAGE aggregation process in Figure 2.

## 4 DATA

To evaluate our proposed method, we require a dense, connected citation network that contains research from a diverse set of scientific disciplines. We construct such a graph using Semantic Scholar
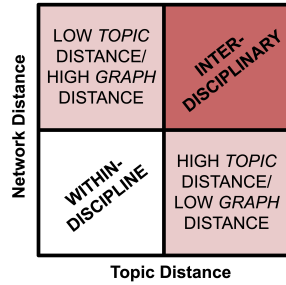
citation information for 58,513 research papers. In order to ensure that the graph contains regions of interdisciplinary research, we collect the research articles according to the following process: (i) Select 8 topics from the All Science Journal Categorisation (ASJC): 'Computer Science', 'Mathematics', 'Chemistry', 'Medicine', 'Social Sciences', 'Neuroscience', 'Engineering', and 'Biochemistry, Genetics and Molecular Biology'. (ii) From each ASJC, select a random sample of up to 1,500 articles published in journals assigned to that topic. (iii) Collect any additional articles that have a 'cited by' or 'citing' relationship with at least 10 articles in the seed set. (iv) Filter the resulting network: remove any articles for which abstract text is not available in English, and retain only the largest connected component in the network. In this manner, we collect a dense, multidisciplinary citation graph with title and abstract text for each article. In the case of each paper, we encode the concatenated title and abstract using the pre-trained SciBERT model and use the final representation of the [CLS] token as an embedding to represent each document (similar to the approach in [4]). The resulting embedding is stored as the only metadata considered on each node.

## 5 EVALUATION

We implement six methods of paper representation for evaluation: (i) *DeepWalk* which does not consider paper content; (ii) *SciBERT* where no citations relations are known between papers; (iii) *SPECTER*; (iv) *GraphSAGE* (MEAN) with a non-parametric mean aggregator; (v) *GraphSAGE* (LSTM) with a learned LSTM aggregator; (vi) our proposed method *ComBSAGE*. Methods (i) and (iii) are trained on our data using the default hyperparameters. (ii) is a pre-trained model from AllenAI. For the GNN methods (iv – vi), each model is implemented with 2 layers, and the batch size, node sampling, and learning rate hyperparameters are all fixed using the best performing set identified in an exhaustive grid-search with the GraphSAGE (MEAN) model. For the ComBSAGE model, we choose a max-pooling aggregator for $AGG_C$ and an LSTM aggregator for $AGG_I$. We also include a *jumping knowledge* [18] connection between the hidden representations in the first and second layer, as this has been shown to be effective in low-homophily networks.

We evaluate each representation method (or paper embedding) based on a citation prediction task. From the data described in Section 4 above, we randomly sample 20% and 5% of the citations to serve as test and validation sets respectively. We learn embeddings from the remaining citation graph according to each of the chosen methods. To treat citation prediction as a binary classification task, we supplement each set of citations (train, validation, and test) with

Figure 3: Quadrant of citation distance. We consider a citation to be *interdisciplinary* if it is between two papers that are distant according to their topic distance (relating to the dissimilarity of their content), and their network distance (relating to their positions in the citation network).



| | Method | Overall AUC |
|---|---|---|
| (i) | *DeepWalk* [15] | 0.95+0.0004 |
| (ii) | *SciBERT* [1] | 0.93+0.0002 |
| (iii) | *SPECTER* [4] | 0.95+0.0048 |
| (iv) | *GraphSAGE (LSTM)* [9] | 0.97+0.0037 |
| (iv) | *GraphSAGE (MEAN)* [9] | **0.98+0.0003** |
| (vi) | *ComBSAGE* (proposed) | **0.98+0.0016** |

**Table 1: Citation prediction performance in terms of Area Under the ROC Curve (AUC). The best performing methods are marked in bold.**

negatively sampled edges/citations that did not occur in the graph. Each edge is then represented as the Hadamard or element-wise product between the embeddings of its endpoints, and classified using a Multi-Layer Perceptron (MLP). We repeat this process for 5 random train:validation:test splits. The overall performance of each embedding approach is reported in Table 1 in Section 6.

In addition to reporting the performance across the entire citation graph, we also wish to understand how each method performs in the interdisciplinary regions of the citation network. Accordingly, we adopt two definitions of *distance* for citations: (i) the *network* distance of a citation, which is the distance between two papers in the citation graph prior to the citation; (ii) the *topic* distance of a citation, which is the dissimilarity between the content of the papers. Using these definitions of distance, every citation can be considered on the quadrant described in Figure 3, with the most interdisciplinary citations positioned in the upper right corner, corresponding to citations between papers positioned in distant regions (i.e., with minimal overlap between their references and citations **and** with minimal similarity between the content of the research papers). In practice, we measure network and topic distance of citations as the cosine distance between the DeepWalk and SciBERT embeddings of the endpoints respectively. We divide all citations in the test set into the four regions of this quadrant and report how paper embedding methods perform at predicting citations in each region. These results are presented in Figure 4.

## 6 RESULTS

Table 1 reports the overall citation prediction performance of each method as measured using the Area Under the ROC Curve. We find that methods which leverage both document text and citation graph information outperform methods which rely on only one modality. Further, we note that our proposed method (*ComBSAGE*) is competitive with the state of the art, with GNN-based methods outperforming more NLP-focused approaches (e.g. *SPECTER*).

Figure 4 shows the balanced accuracy score for citation prediction for each embedding method, in each region of the citation distance quadrant (as per Figure 3). The $y$ position of a marker within each subsection shows the average balanced accuracy score over 5 splits of the data. The markers have been jittered along the $x$-axis for legibility. While the order of the markers along this axis could be considered arbitrary, we have attempted to position methods according to their position on the spectrum in Figure 1. It is apparent from the bottom half of this diagram that citation information is highly important in understanding relations between papers
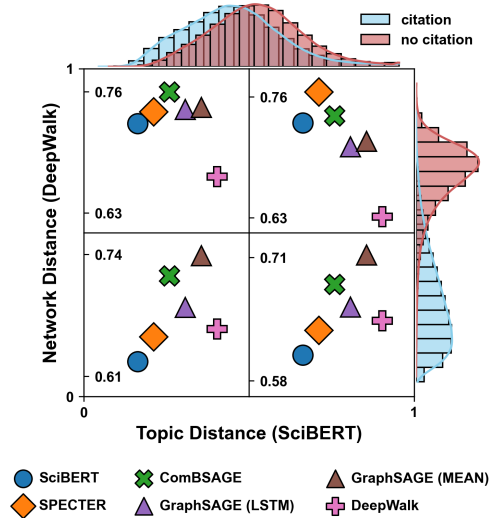


**Figure 4: Citation prediction performance (mean class accuracy), with citations grouped according to the distance between the papers.**

with overlapping citation neighbourhoods. Conversely, when identifying relations between papers that are distant in the citation network, the utility of textual features increases. Our proposed ComBSAGE method outperforms other GNN-based methods when predicting interdisciplinary citations (see also Appendix).

## 7 DISCUSSION & CONCLUSIONS

In this work, we evaluate a variety of document embedding methods on their ability to identify interdisciplinary relations between research papers. We find that those methods that rely more on NLP and the textual content of research papers (*SciBERT*, *SPECTER*) are better able to predict interdisciplinary citations than methods which rely more on network analysis of citation information (*GraphSAGE*, *DeepWalk*). However, these NLP-based methods are not sufficiently capable of leveraging citation information to compete with GNN-based methods in predicting interactions between papers with overlapping citation neighbourhoods or in the task of citation prediction *overall*. Motivated by recent work in citation analysis of interdisciplinary research, we propose a novel method of GNN message aggregation, *ComBSAGE*, that rectifies some of the limitations of MP-GNNs when predicting *interdisciplinary* citations, without compromising *overall* citation prediction performance.

# REFERENCES

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3615–3620. https://doi.org/10.18653/v1/D19-1371

[2] H. Cai, V. W. Zheng, and K. Chang. 2018. A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications. *IEEE Transactions on Knowledge & Data Engineering* 30, 09 (2018), 1616–1637. https://doi.org/10.1109/TKDE.2018.2807452

[3] Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. 2020. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 3438–3445.

[4] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2270–2282. https://doi.org/10.18653/v1/2020.acl-main.207

[5] Eoghan Cunningham and Derek Greene. 2022. Assessing Network Representations for Identifying Interdisciplinarity. In *Companion Proceedings of the Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 758–762. https://doi.org/10.1145/3487553.3524653

[6] Eoghan Cunningham and Derek Greene. 2023. The Structure of Interdisciplinary Science: Uncovering and Explaining Roles in Citation Graphs. In *Complex Networks and Their Applications XI: Proceedings of the 11th International Conference on Complex Networks and Their Applications*, Vol. 1. Springer, 364–376. https://doi.org/10.1007/978-3-031-21127-0_30

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[8] Palash Goyal and Emilio Ferrara. 2018. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems* 151 (2018), 78–94. https://doi.org/10.1016/j.knosys.2018.03.022

[9] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 1025–1035.

[10] Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* (2016).

[11] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=SJU4ayYgl

[12] Diego Kozlowski, Jennifer Dusdal, Jun Pang, and Andreas Zilian. 2021. Semantic and relational spaces in science of science: deep learning models for article vectorisation. *Scientometrics* (2021), 1–30. https://doi.org/10.1007/s11192-021-03984-1

[13] Heidi Ledford. 2015. Team science. *Nature* 525, 7569 (2015), 308.

[14] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1301.3781

[15] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: Online Learning of Social Representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, New York, USA) *(KDD '14)*. Association for Computing Machinery, New York, NY, USA, 701–710. https://doi.org/10.1145/2623330.2623732

[16] Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M. Bronstein. 2022. Understanding over-squashing and bottlenecks on graphs via curvature. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. https://openreview.net/forum?id=7UmjRGzp-A

[17] David N. Wear. 1999. Challenges to Interdisciplinary Discourse. *Ecosystems* 2, 4 (1999), 299–301. http://www.jstor.org/stable/3659022

[18] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. 2018. Representation learning on graphs with jumping knowledge networks. In *International Conference on Machine Learning*. PMLR, 5453–5462.

[19] Y. Yan, M. Hashemi, K. Swersky, Y. Yang, and D. Koutra. 2022. Two Sides of the Same Coin: Heterophily and Oversmoothing in Graph Convolutional Neural Networks. In *2022 IEEE International Conference on Data Mining (ICDM)*. IEEE Computer Society, Los Alamitos, CA, USA, 1287–1292. https://doi.org/10.1109/ICDM54844.2022.00169

# APPENDIX

| Method | AUPRC | F1 (macro average) | Average Precision |
|---|---|---|---|
| (i) *DeepWalk* | 0.345±0.003 | 0.640±0.001 | 0.198±0.002 |
| (ii) *SciBERT* | 0.379±0.004 | 0.654±0.001 | 0.239±0.002 |
| (iii) *SPECTER* | 0.515±0.058 | 0.710±0.020 | 0.303±0.027 |
| (iv) *GraphSAGE (LSTM)* | 0.441±0.026 | 0.698±0.010 | 0.273±0.013 |
| (v) *GraphSAGE (MEAN)* | 0.531±0.003 | 0.730±0.002 | 0.324±0.003 |
| (vi) *ComBSAGE* (proposed) | **0.543±0.008** | **0.738±0.003** | **0.335±0.006** |

**Table 2: Interdisciplinary citation prediction performance according to different metrics. We report Area under the Precision Recall Curve (AUPRC), macro averaged F1 score and Average Precision (AP). In particular, we choose AUPRC and AP as metrics which are especially informative/reliable when the positive class represents a small minority in the dataset; as is the case in the IDR region of our test set.**