

Facilitating Interdisciplinary Knowledge Transfer with Research Paper Recommender Systems

Eoghan Cunningham^{1,2*}, Barry Smyth^{1,2} and Derek Greene^{1,2}

¹School of Computer Science, University College Dublin, Dublin, Ireland.

²Insight Centre for Data Analytics, Dublin, Ireland.

 <https://orcid.org/0000-0002-0435-1962>,

 <https://orcid.org/0000-0003-0962-3362>,

 <https://orcid.org/0000-0001-8065-5418>.

*Corresponding author. E-mail: eoghan.cunningham@ucd.ie;

Abstract

Interdisciplinary knowledge transfer is hindered by information overload and siloed reading and citation practices. Research paper recommender systems, which tend to overemphasise similarity and relevance, can perpetuate information silos due to the so-called ‘filter bubble’ effect. In this work, we argue for the importance of offering novel and diverse research paper recommendations to scientists in order to reduce siloed reading, and facilitate interdisciplinary knowledge transfer. We also identify important RP-Rec-Sys methodologies which serve this purpose. Specifically, we propose a novel framework for evaluating the novelty and diversity of research paper recommendations, drawing on methods from network analysis and natural language processing. Using this framework, we show that the choice of representational method within a larger research paper recommendation system can have a measurable impact on the nature of downstream recommendations, specifically on their novelty and diversity. We describe a paper embedding method that provides more distant and diverse research paper recommendations without sacrificing the relevance of those recommendations compared to other state-of-the-art baselines. By recommending relevant research to users that is distant and dissimilar from their own work, we present a viable method to facilitate interdisciplinary knowledge transfer using research paper recommender systems.

Keywords: Recommender Systems, Diversity, Novelty, Knowledge Transfer

1 Introduction

As the rate of scientific publication continues to accelerate, it is increasingly difficult for researchers to keep abreast of developments in their own fields of study, and even more difficult to follow advancements in areas beyond their primary discipline. Many now rely on research paper recommender systems (RP-Rec-Sys), which combine ideas from recommender systems (Ricci, Rokach, & Shapira, 2015), user modeling personalisation (Ghorab, Zhou, O’connor, & Wade, 2013), information retrieval and discovery (Liu et al., 2019), to support their reading. Various approaches have been adopted for different discovery use-cases (Bai et al., 2019), with each use-case advocating a unique approach to recommendation. For example, Hua, Chen, Li, Zhao, and Zhao (2020) and Hui et al. (2020) consider how researchers engage with a new topic for the first time, whereas Sharma et al. (2020) focus on subject matter experts conducting literature reviews.

A common challenge across all applications of RP-Rec-Sys is how to prevent the so-called *filter-bubble* effect of conventional recommender systems (Portenoy et al., 2022). This can reinforce siloed reading and citation tendencies by prioritising items that are similar to past recommendations. Amid efforts to understand and encourage boundary-crossing research (Qi, Zhou, Sun, Huang, & Zhang, 2024), recommendations that are over-reliant on similarity can be antithetical to the needs of modern interdisciplinary science, where many valuable breakthroughs often come from surfacing the latent or unexpected connections between disciplines (Bornmann, Tekles, Zhang, & Fred, 2019; F. Shi & Evans, 2023; Uzzi, Mukherjee, Stringer, & Jones, 2013). Over the last 20 years, Rec-Sys evaluations have evolved to combat the filter bubble effect with the development of metrics like ‘recommendation novelty’ and ‘recommendation diversity’ (Castells, Hurley, & Vargas, 2021; Smyth & McClave, 2001). However, to date, these evaluation perspectives have not been considered in the context of RP-Rec-Sys (Ali, Ullah, Khan, Ullah Jan, & Muhammad, 2021). In particular, three recent surveys of the RP-Rec-Sys literature highlight recommendation ‘diversity’, ‘novelty’, and/or ‘serendipity’ as open challenges that have not been addressed sufficiently (Ali et al., 2021; Bai et al., 2019; Kreutz & Schenkel, 2022). Addressing this gap in the literature is important as *relevant*, *diverse* and *novel* recommendations (as they are defined by the Rec-Sys community), would serve to expose researchers to relevant works from outside of their established field. Such trans-disciplinary recommendations would be of great value given the importance and impact of interdisciplinary research (S. Chen, Song, Shu, & Larivière, 2022; Larivière, Haustein, & Börner, 2015; F. Shi & Evans, 2023).

Motivated by the above, this work makes two core contributions: we propose a novel framework for evaluating research paper recommendations, and hence highlight a RP-Rec-Sys method capable of making *relevant*, and *interdisciplinary* recommendations. The framework is designed to bridge the divide between RP-Rec-Sys evaluations and the broader Rec-Sys literature, and to facilitate and promote far-reaching, cross-discipline research recommendations and interdisciplinary knowledge transfer. Specifically – to supplement established measures of recommendation relevance – we propose novel metrics for research paper recommendation *diversity* and *novelty* (or *unexpectedness*) according to citation network distances and research topic dissimilarities. As the motivation of this work is to facilitate interdisciplinary knowledge transfer, we demonstrate our proposed framework with an experiment in which we evaluate four recommenders employing different representation (or *paper embedding*) methods, in order to identify a method that can provide more distant yet still relevant recommendations for the user. We show that the choice of paper embedding method upstream of an RP-Rec-Sys has a measurable impact on both the quality (as measured by relevance) and interdisciplinarity (as measured by diversity and novelty) of downstream recommendations. In particular we highlight a recent embedding method – ComBSAGE (Cunningham & Greene, 2023) – that provides more far-reaching, interdisciplinary recommendations without compromising relevance when compared to other state-of-the-art approaches.

The remainder of this work is structured as follows. Section 2 formalises the tasks of research paper recommendation and embedding, and outlines important existing works in the area. Section 3 explains the framework we propose for evaluating research paper recommender systems. In particular, we emphasise our approach for measuring diversity and novelty in the context of RP-Rec-Sys. In Section 4, we describe an RP-Rec-Sys evaluation designed to demonstrate our proposed framework, and thus to highlight promising RP-Rec-Sys methodologies that are capable of generating more novel and diverse recommendations. First, we summarise the dataset compiled for our analysis in Section 4.1. Second, we detail the implementation of the paper embedding methods that we evaluate. Third, we describe our approach for generating recommendations given these article representations, and finally, the evaluation metrics used. The results of our evaluations are presented in Section 5 and we offer some final conclusions in Section 6.

2 Background

We begin by formalising the task of research paper recommendation in Section 2.1 where we outline several important aspects of RP-Rec-Sys evaluations, before discussing the concepts of *diversity* and *novelty* as they relate to both science-of-science and recommender systems in Section 2.2. Finally, we discuss research paper representation learning (or embedding) in Section 2.3, as it is a crucial aspect of RP-Rec-Sys and the subject of our experiments.

2.1 Research Paper Recommendation

Research paper recommender systems (RP-Rec-Sys) are designed to suggest research papers that align with the interests or the needs of an end-user. Consistent with the field of Rec-Sys and its broader applications, approaches to research paper recommendation can be categorised into two groups: Content-Based Filtering (CB) and Collaborative Filtering (CF). The former methods involve generating recommendations for users based on descriptions or representations of items (Lops, De Gemmis, & Semeraro, 2011; Smyth, 2007), whereas CF methods primarily depend on interactions between users and items to provide recommendations (Koren, Rendle, & Bell, 2021; Y. Shi, Larson, & Hanjalic, 2014). Many systems have been implemented from a variety of task-oriented perspectives – recommending papers to authors (Jiang et al., 2023), recommending authors to authors (Portenoy et al., 2022), and even recommending citations for papers (Ali et al., 2021). Within each of these perspectives, many possible use-cases of RP-Rec-Sys exist. For example, Hua et al. (2020) and Hui et al. (2020) focus specifically on researchers engaging with a new subject for the first time, while Sharma et al. (2020) develop their system to support researchers conducting in-depth literature review. All CB methods of RP-Rec-Sys (and many CF/hybrid methods), rely on some fundamental underlying form of paper representation or embedding.

The diverse approaches to RP-Rec-Sys are reflected in the broad range of evaluation methods proposed in the literature. One of the most reliable means of evaluating RP-Rec-Sys is through user studies: users interact directly with the system, and their satisfaction is assessed either through explicit surveys or deduced from factors like session duration (Portenoy et al., 2022). Although these so-called ‘online’ evaluation methods can be highly informative, they come with considerable costs, particularly in terms of recruiting and surveying users. Therefore, ‘offline’ evaluations are often preferred as a more feasible option or employed as a preliminary step before proceeding with online evaluations. Offline evaluations typically require some (usually external) ground truth used to define which papers are relevant to a given user or query paper. In the context of scientometrics, existing relationships which have been used as ground truths include ‘paper-cites-paper’ (Ali et al., 2022), ‘author-cites-paper’ (Jiang et al., 2023), and ‘paper-co-read-with-paper’ (Singh, D’Arcy, Cohan, Downey, & Feldman, 2023) information. Of course, these choices will depend on the specific use-case and availability of appropriate data.

The choice of ground truth in RP-Rec-Sys evaluations has been identified as a core challenge in the domain (Ali et al., 2021). For example, in those cases of RP-Rec-Sys where the user/author view is not considered, direct citation information (‘paper-cites-paper’) is commonly used as a ground truth for relevance. However, implementing such an evaluation while respecting the inherent temporal organisation of a citation network can be challenging. Most methods involve excluding certain references from the bibliography of some query or test papers, and require the system to predict these citations. This process resembles the standard link prediction problem frequently posed in the field

of network analysis. Consequently, only the most recent publications can be considered as queries. Otherwise, a user is tasking the model with predicting early citations given the knowledge of later citations, thus disregarding the real temporal ordering of the data. Avoiding this temporal data leakage limits evaluations to testing a single use-case of RP-Rec-Sys: ‘complete the bibliography of this unpublished research paper’.

Of course, there are several other use-cases of RP-Rec-Sys that should be also considered, and it is important to be able to evaluate systems in such scenarios. Co-share and co-view relationships between papers have been employed as alternatives to direct citation (Singh et al., 2023). Such information is gathered by services like Semantic Scholar and can offer a highly valuable indication of relatedness or similarity between papers. However, the scope of any evaluations that rely on these user-based relations will be restricted to the small sample of papers for which they are available. Moreover, we have concerns that encouraging models to fit to this signal may only serve to reinforce established reading and citation behaviours. Given the above limitations of these popular methods, we rely on *co-citation* relations in our evaluations as a ground truth for establishing relevance between two papers. Further details about our evaluation framework and its advantages can be found in Section 3.

Regardless of the choice of ground truth, (e.g. ‘paper-cites-paper’, ‘paper-co-saved-with-paper’), most evaluations employ metrics designed to measure the proportions of recommendations that are relevant to the user or to their query (e.g. precision), the quality of the ordering of the recommendations (e.g. DCG, MRR), the proportion of possible relevant papers that were recommended (e.g. recall), or some combinations of these metrics (e.g. MAP, F1). Exploring these different metrics may permit the system to be evaluated with respect to a small number of different use-cases. For example, a researcher who makes regular use of an RP-Rec-Sys may value high precision for small recommendation sets (i.e., the top 5–10 papers), while an author undertaking some large survey or literature review may prefer high recall on a larger recommendation set (30+ papers). However, limiting the scope of evaluations to these metrics can exclude other valid RP-Rec-Sys use-cases from consideration. To date there has been limited attention directed towards recommendation *diversity* and *novelty* (Ali et al., 2021; Bai et al., 2019), despite the emphasis that has been placed on these objectives in the broader recommender systems literature over many years (Castells et al., 2021; Smyth & McClave, 2001). Systems that perform well with respect to these metrics may facilitate another use-case – authors seeking exposure to novel ideas, challenges, and methodologies from other research domains.

2.2 Novelty and Diversity

Novelty and diversity are important concepts in both recommender systems research and scientometric analysis. As Castells et al. (2021) note, “*novelty can be generally understood as the difference between present and past experience whereas diversity relates to the internal difference within parts of an*

experience". Each concept has been formalised and understood independently in these two fields, and while usage of these concepts differs across the two domains, both perspectives are closely related and remain relevant to this work. In the following section, we will outline the concepts of novelty and diversity as they relate separately to scientometric analysis and to recommender systems, and crucially, how *research* novelty and diversity may be supported and encouraged by emphasising and promoting *recommendation* novelty and diversity.

In the field of scientometrics, a research output may be deemed *novel* if it involves a 'surprising' or 'unlikely' integration of ideas, methods, or models (F. Shi & Evans, 2023; Shin, Kim, & Kogler, 2022). It has been shown that the most impactful and ground-breaking advancements are often led by works which maximise this definition of novelty, and thus research novelty is highly valued. For example, the work of Altschul et al. (1997) presented a highly-novel combination of search methods from computer science, with data and challenges from biomedical sciences, to build a tool for protein and DNA search. It is now among the most cited research outputs of all time. Research diversity has been considered and evaluated from a number of similar perspectives. For example, many studies measure the diversity of topics, disciplines, or subject categories referenced within an article, using this as a metric to gauge the intellectual breadth of a single work. (Larivière et al., 2015; Okamura, 2019; Porter, Cohen, David Roessner, & Perreault, 2007). Others evaluate this across the entire body of work of a researcher to determine the breadth of their research diversity (Porter et al., 2007; Yu, Wang, Zhang, Bu, & Xu, 2023). In this manner, research diversity serves as an indicator for multi- and interdisciplinary work, which has been shown to correlate with research impact (S. Chen et al., 2022; Larivière et al., 2015). Consider Marine Science as a pertinent example of the importance of disciplinary diversity in scientific research. As environmental and geostrategic conflicts over ocean resources accelerate, there is a growing need for Marine Science research that integrates environmental, legal, social, economic, and political perspectives (Markus, Hillebrand, Hornidge, Krause, & Schlüter, 2018).

In the area of recommender systems, diversity and novelty take on different interpretations. For the purposes of this work, we will consider the following formulations of the *diversity* and *novelty* of a set of recommendations (Castells et al., 2021). For a set of k recommendations given for some query, we denote the novelty of the set of recommendations as the mean cosine distance between the representation of a recommended item and that of the query. Similarly, the diversity of the set of k recommendations can be defined as the mean pairwise cosine distance between the representations of recommended items. This is sometimes referred to as the Average Intra-List Distance (Castells et al., 2021; Smyth & McClave, 2001). Both concepts are considered desirable across many recommender systems applications (Kaminskas & Bridge, 2016; Nakatsuji et al., 2010; Vargas & Castells, 2011; L. Zhang, 2013). For example, enhancing recommendation diversity can be a means of reducing the redundancy of a set

of recommendations, as is the case with travel recommender systems, which usually seek to avoid suggesting similar properties in the same destination resort (Smyth & McClave, 2001); in this case, recommending a broader set of options is an important way for a recommender system to ‘hedge its bets’ especially when the precise needs and preferences of the user are uncertain. Likewise, when recommending movies or music to consumers, novelty is often desirable according to the inherent satisfaction of unexpectedness, change, and complexity, and also as a foil to satiation and the decreased satisfaction associated with repeated consumption (Castells et al., 2021; Nakatsuji et al., 2010).

When recommending research papers, it is crucial to recognise the difference between the respective formulations of *research* novelty and diversity, and *recommendation* novelty and diversity. We consider a recommended research paper to be a *novel recommendation*, if it is unexpected with respect to the user or query (as per Equation 1), and not simply if the recommended paper represents novel work according to definitions of (F. Shi & Evans, 2023) or (Shin et al., 2022). Of course, the former does not exclude the latter – a recommendation may be novel according to both definitions. In fact, it follows that suggesting recommendations of high *recommendation* novelty could increase the *research* novelty of a researcher’s future outputs, as they go on to integrate the ideas from surprising or otherwise distant sources. We argue that the same is true for research diversity. A set of recommended papers with high *recommendation diversity* may not contain any works which are themselves diverse or interdisciplinary according to existing scientometric definitions (Okamura, 2019; Porter et al., 2007). However, a crucial motivation for increasing *recommendation* diversity in RP-Rec-Sys is to promote *research* diversity in future works.

To date, while recommendation diversity, novelty, and the related concepts of ‘serendipity’ and ‘surprise’ have been explored by the wider recommender systems community (Castells et al., 2021), they are less well understood in the context of research paper recommender systems. For instance, in a recent survey by Ali et al. (2021) of 67 models for RP-Rec-Sys (published between 2013–2021), only 3 papers were found to evaluate recommendation ‘novelty’. Other surveys have highlighted diversity, novelty, and serendipity as key challenges (Bai et al., 2019; Kreutz & Schenkel, 2022) in the domain and high-priority future work.

Diverse and novel recommendations are particularly important in the field of research paper recommendation, to broaden the horizons of researchers, by exposing them to novel ideas, methodologies and challenges, usually from other fields, and by ultimately removing barriers to interdisciplinary research. This is especially valuable given that interdisciplinary research outputs, which integrate ideas from several distinct disciplines have been consistently found to be more impactful than the more conventional, incremental, within-discipline outputs (Larivière et al., 2015; Okamura, 2019; F. Shi & Evans, 2023).

Existing works that evaluate research paper recommendation diversity and novelty, are limited and varied. Nishioka, Hauke, and Scherp (2019) attempted to use additional information from authors’ *twitter* profiles to provide novel or ‘serendipitous’ recommendations, and relied on a small user-study as an ‘online’ evaluation. Other studies have developed offline evaluations, for example: Rodriguez-Prieto, Araujo, and Martinez-Romo (2019) measured recommendation serendipity, according to how a set of recommendations differ from a set of recommendations provided by some ‘primitive’ model, while Chaudhuri, Sinhababu, Sarma, and Samanta (2021) assessed the *research* or *disciplinary* diversity of individual papers recommended, rather than the *recommendation* diversity of the set of recommended articles.

2.3 Scientific Document Representation

Representation learning for scientific documents refers to the task of representing research papers in a vector space so that the important relationships between the papers are preserved – i.e., semantically related papers should have similar representations (or *embeddings*) (Kozłowski, Dusdal, Pang, & Zilian, 2021). These research paper embeddings can then be used to support downstream tasks such as literature review, classification, or (as is the focus of this work) recommendation. Scientific articles can be related in different ways. For example, two papers may be related if they have similar content, or two papers might be related if there is a transfer of knowledge or ideas between them; that is, there is some citation or chain of citations from one to the other (Jo, Liu, & Wang, 2022). Accordingly, article text and citation relationships are the main sources of information for scientific document representation (Kozłowski et al., 2021). Some approaches to document embedding rely on additional article metadata, such as information about the authors or the publication venue. We exclude this information from our discussion of paper embedding as we believe such metadata can surface relations between papers that are only superficial in the context of some downstream use-cases. For example, in the RP-Rec-Sys use-case of exposing researchers to relevant works from outside the scope of their current reading patterns, author or venue relations may only serve to reinforce their existing behaviours. Of course, many use-cases exist where authorship information is relevant (e.g. seeking works written by some author) or even vital (e.g. personalised recommendations based on an author’s bibliography). In such cases, methods have been developed for including these relations downstream in the recommender system and not in the article embedding stage (Ali et al., 2022; Jiang et al., 2023; Mei et al., 2022). Thus, while additional relations based on metadata can be employed downstream according to the use-case or application, we prefer to rely only on article content and citation relations during the article embedding stage. We focus our analysis on methods that make use of these two modalities, and propose that all of the approaches we include can be considered on the spectrum of techniques illustrated in Figure 1. We position methods that rely only on the text of an article

on the left-hand side, those that rely only on citations on the right, while hybrid methods (combining text and citations) lie in between these extremes.

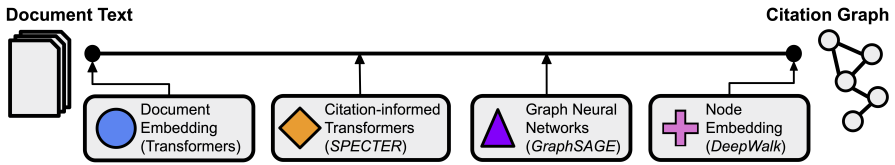


Fig. 1 The spectrum of scientific document representation methods. Methods for paper embedding can be positioned along this spectrum, with those methods exclusively reliant on article text placed on the left, and those methods exclusively reliant on the citation graph placed on the right. Architectures that make use of both sources of information are situated between these two extremes.

Simple text-based approaches to research paper representation learning (including those applied downstream to recommendation tasks) use ‘bag-of-words’ vectors, which count the occurrence of each word in each document such that documents which contain similar vocabularies will have similar representations. These approaches can be extended to re-weight or transform the features according to a features utility in discriminating between documents relating to different topics (e.g. TF-IDF, LDA, LSA, etc.) (Achakulvisut, Acuna, Ruangrong, & Kording, 2016; Amami, Pasi, Stella, & Faiz, 2016; Dai et al., 2018). Transformer-based language models, such as BERT (Devlin, Chang, Lee, & Toutanova, 2018) (Bidirectional Encoder Representations from Transformers), leverage attention mechanisms and self-attention layers to transform the textual content of research papers into document embeddings (Nogueira, Jiang, Cho, & Lin, 2020). Crucially, models like *SciBERT* (Beltagy, Cohan, & Lo, 2019) have been pre-trained on large corpora of research papers in order to capture the specific vocabulary and semantics of academic writing, and are commonly applied to an article’s title and abstract to produce paper embeddings.

At the other end of the spectrum, some methods exclusively depend on information from the citation network. Given a graph $G = (V, E)$, where V is the set of nodes or vertices on the graph, and E is the set of edges, or interactions between those nodes, *Graph Representation Learning* refers to the task of learning latent features/low-dimensional vectors to describe nodes on the graph G . These *node embeddings* offer a compact and informative representation of the nodes which preserve important properties, such as proximities and community structure in the graph. In the case of a network of research papers, where nodes represent research articles and the interactions between the nodes represent citation relations for example, node embeddings can be used as the source of paper embedding. Thus, papers which are nearby to each other in the citation network, (e.g. they share many citations or references), will have similar representations. In addition to the purely network-based methods for article embedding (e.g. *DeepWalk*) and the purely NLP-based approaches

(e.g. *Sci-BERT*), the complex nature of research papers and their relations necessitates approaches that are capable of leveraging (and combining) both modalities, to provide a single article representation.

Naïve approaches have been developed which simply combine or concatenate text and network embeddings. However, more sophisticated techniques have been suggested to effectively merge the semantic and relational dimensions of research articles. For example, *SPECTER* (Cohan, Feldman, Beltagy, Downey, & Weld, 2020) 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. Thus, in addition to the vocabulary and semantics of a research article’s content, *SPECTER* uses the knowledge transfer described by citation relations to inform similarities in the embedding space, making it a good choice for downstream RP-Rec-Sys applications (Ali et al., 2022).

Message Passing Graph Neural Networks (MP-GNNs) (e.g. GraphSAGE, GCN, VGAE) can be applied to graph structured data, where some additional information is available for the nodes. Given a graph, $G = (V, E)$ with node features X , one layer of an MP-GNN is typically described with 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 $\{v_j, v_k, \dots, v_p\}$ with features $\{x_j, x_k, \dots, x_p\}$. An embedding for v_i (denoted by h_i) is then generated via the update $h_i \leftarrow \phi(x_i, u_i)$, where $u_i \leftarrow \psi(\{x_j, x_k, \dots, x_p\}, x_i)$. Thus, ψ is some (typically permutation invariant) function for *aggregating* information from neighbouring nodes (potentially with consideration of x_i), and ϕ 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), or in an unsupervised manner where the objective task used for training is graph reconstruction (or link prediction). MP-GNNs are designed to 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 informed not only by its own content, but also by the content of the papers that it cites (the ideas upon which it develops), and even by the content of the articles that cite the focal paper (its applications or the work that builds on it). For these reasons MP-GNNs are commonly applied to document representation and RP-Rec-Sys (Ali et al., 2021; Jiang et al., 2023; Kozłowski et al., 2021; Kreutz & Schenkel, 2022).

3 Method: Evaluating Research Paper Recommendations

In this section, we present the primary technical contribution of our work: a novel approach to RP-Rec-Sys evaluations. Our framework draws on contemporary practices in the wider Rec-Sys literature. It is consistent and compatible with existing metrics employed in RP-Rec-Sys evaluations (i.e., measures of recommendation relevance), while adding new metrics for recommendation novelty and recommendation diversity calculated from multiple perspectives. Additionally, our method is intended to avoid a number of common pitfalls associated with RP-Rec-Sys evaluations. Specifically, it is designed to be robust and consistent with the many diverse use-cases of RP-Rec-Sys and crucially, to respect the inherent temporal ordering of a citation network. Our proposed method can be considered in two stages. First, we describe our approach for offline evaluation of RP-Rec-Sys *relevance*. Secondly, we present four metrics for evaluating RP-Rec-Sys *novelty* and *diversity*.

3.1 Evaluating Recommendation Relevance

Given a query paper q , we first evaluate a set of recommendations r , according to their relevance. We adopt co-citation relations as our ground-truth for relevant papers. Formally, a recommended paper is relevant to the query paper if and only if both papers are co-cited in some third publication. Co-citation relations have been widely used as a measure of relevance in scientometric analyses (C. Chen, McCain, White, & Lin, 2002; Gmür, 2003; Small, 1973; J. Zhang & Zhu, 2022). A co-citation may occur between two papers for a number of reasons; (i) an author lists/compares/evaluates two papers that are methodologically/conceptually similar; (ii) an author outlines an important transfer of knowledge from one paper to the other; (iii) an author highlights an interesting problem and a potential solution.

Crucially, co-citation between papers is a strong and flexible signal for relation that is established and supported by the authors active in the area. Thus, co-citation analysis has been used to identify connections between research across separate disciplines (Trujillo & Long, 2018), and map the disciplinary structure of large research corpora (Boyack & Klavans, 2010). In many offline evaluations, co-citation may prove a more appropriate ground truth compared to other paper-paper relations (such as those outlined in Section 2.1). For example, direct citation between papers is often implemented as the ground truth for relevant recommendations (Ali et al., 2022; Mei et al., 2022; Nogueira et al., 2020). However, it is challenging to evaluate such a system while still adhering to a strict temporal split and avoiding any potential data leakage. As direct citations can only occur at the time of publication, the set of query papers in evaluations that rely on direct citations must be restricted to contain only unpublished articles. This specific case of RP-Rec-Sys is commonly referred to as ‘citation recommendation’ (Ali et al., 2021). Direct citation ground truths should not be applied outside of this use-case, as they disobey

the inherent temporal structure of a citation network. Unlike direct citations, a co-citation relationship between two papers can appear long after they were initially published. As such, we propose the use of co-citation relations to allow users to query systems with published research, such as examples of their own works, (i.e., for personalised recommendations), or queries relevant to their ongoing works (i.e., for information retrieval or supporting literature reviews).

Accordingly, the relevance of a set of recommendations can be measured using simple metrics such as *precision*, the proportion the articles recommended in r that are relevant to q , and *recall*, the proportion of articles relevant to q that are recommended in r . Alternatively, relevance could be measured using more complex metrics which consider the ordering of recommended articles within the set r , such as Area Under the ROC Curve (AUC) or Discounted Cumulative Gain (DCG).

3.2 Evaluating Recommendation Diversity and Novelty

In addition to understanding the relevance of recommended articles, it is pertinent in many use-cases to also evaluate the diversity and novelty of a set of recommendations. Diverse and novel recommendations have many benefits across different applications (see Section 2.1). In the specific domain of research paper recommender systems, diverse and novel recommendations (that remain relevant to the user) can expose researchers to otherwise unknown methods/-solutions/challenges from distant fields of study. In this way, RP-Rec-Sys can promote the exchange of ideas across disparate disciplines and burst scientific filter bubbles (Portenoy et al., 2022).

To assess the diversity and novelty of the recommendations generated, we present four metrics: (i) citation network recommendation novelty, (ii) citation network recommendation diversity, (iii) article content recommendation novelty, and (iv) article content recommendation diversity. In our calculations, we use the widely accepted, standard formulations of novelty and diversity (Castells et al., 2021; Smyth & McClave, 2001) as presented in Equations 1 and 2 below.

For a set of k recommendations r provided for some query q , we denote the novelty of the set of recommendations as

$$\text{recommendation novelty} = \frac{1}{k} \sum_{i \in r} \frac{x_q \cdot x_i}{\|x_q\|_2 \|x_i\|_2} \quad (1)$$

which is the mean cosine distance between the representation x_i of a recommended item i and that of the query. Similarly, the diversity of the set of k recommendations can be defined as

$$\text{recommendation diversity} = \frac{1}{k} \sum_{i,j \in r} \frac{x_i \cdot x_j}{\|x_i\|_2 \|x_j\|_2} \quad (2)$$

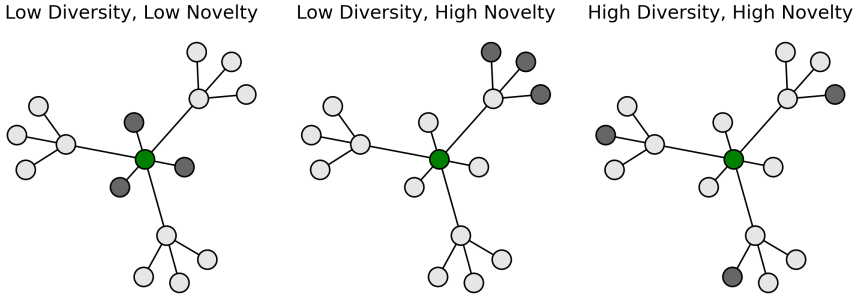


Fig. 2 A toy example designed to offer a visual demonstration of the network perspective of novelty and diversity in recommendation. The central node represents a query and the shaded nodes show examples of recommendations with different diversity and novelty scores, as measured using equations 1 and 2, given *DeepWalk* embeddings as descriptions of nodes.

which is the mean pairwise cosine distance between the representations of the recommended items.

We consider diversity and novelty from two perspectives: the citation network perspective, and the article content perspective, thus giving four metrics. Specifically, when a system is posed with a query and provides a set of recommendations, we measure: (i) citation network recommendation *novelty* as the mean cosine distance between the *DeepWalk* representation of the query, and the *DeepWalk* representation of the recommendations, such that recommendations are considered ‘novel’ if they are positioned far away from the query in the citation network; (ii) citation network recommendation *diversity* as the mean cosine distance between the *DeepWalk* representations of the recommendations, such that a set of recommendations are considered ‘diverse’ if they are positioned far away from each other in the citation network; (iii) article content recommendation *novelty* as the mean cosine distance between the *SciBERT* representation of the query, and the *SciBERT* representation of the recommendations, such that recommendations are considered ‘novel’ if they have content that differs from the query; (iv) article content recommendation *diversity* as the mean cosine distance between the *SciBERT* representations of the recommendations, such that recommendations are considered ‘diverse’ if they have content that differs from each other.

Metrics (i) and (ii) above (diverse and novel recommendations according the citation graph perspective) are illustrated in the toy example provided in Figure 2. The citation graph perspective represents a “relational space of research articles” as described by Kozłowski et al. (2021), and as citation networks have been shown to map the disciplinary structure of research areas (Asatani, Mori, Ochi, & Sakata, 2018; Boyack & Klavans, 2010; Fortunato et al., 2018), distance in the citation network is indicative of a disciplinary distance between research papers.

Similarly, the article content perspective considers distances between the *SciBERT* embeddings of the title and abstract of the research papers when

measuring diversity and novelty of recommendations. As discussed in Section 2.3, the *SciBERT* embeddings that we use to represent the content of research papers correspond to the final hidden state representations of the [CLS] token in the large pre-trained BERT model, *SciBERT*. Therefore, the article content perspective represents a “*semantic* space of research articles” as described by (Kozłowski et al., 2021), and therefore distances measured according to SciBERT offer an alternative and complementary measure of the disciplinary distance between research articles. It follows that recommender systems which maximise these definitions of novelty and diversity can help to counter the filter bubble effect and promote interdisciplinary knowledge transfer.

Of course, the utility of novel and diverse recommendations is largely contingent on their relevance. As such, these metrics should be used in tandem with traditional measures of article relevance, such as those outlined in Section 3.1. Further, we also suggest calculating and reporting the novelty and diversity of the subset of the recommended papers that are *relevant* to the query. We note that in some cases, the diversity of the set of recommended and relevant papers may not be a particularly meaningful statistic, as it is often calculated from a set of papers that is substantially smaller than k . However, we include it here for completeness.

4 Experiment Design

In this section we describe the design of an experiment to demonstrate our proposed evaluation methods and to investigate the potential of RP-Rec-Sys to facilitate interdisciplinary research and knowledge transfer. Specifically, we evaluate different recommenders – each employing a different representation method – to explore the effect of upstream research paper embedding methods on the disciplinary distance (novelty and diversity) of downstream recommendations in an RP-Rec-Sys context. We implement four different approaches to research paper embedding: one baseline, two state-of-the-art, and a recent method yet to be implemented in an RP-Rec-Sys context. These implementations are detailed in Section 4.2. We generate recommendations according to these embeddings using a simple, representation-agnostic recommender system which is described in Section 4.3. Subsequently, we evaluate each set of recommendations according to the framework described in Section 3. Section 4.4 outlines the evaluation metrics we employ, and how these metrics can relate embedding methods to their most appropriate recommendation use-case. In addition to the primary use case central to this work – reducing filter bubbles and promoting interdisciplinary knowledge transfer – we also consider more traditional use cases, such as information retrieval and literature review.

4.1 Data

To conduct our experiment, we require a densely-connected citation network that includes research from a variety scientific disciplines. We constructed a novel citation graph using Semantic Scholar citation information for 58,513

research papers. To ensure that this graph contained regions of interdisciplinary research, we collected the set of research articles according to the following process:

- i Select a sample of 8 topics from the All Science Journal Categorisation (ASJC): ‘Computer Science’, ‘Mathematics’, ‘Chemistry’, ‘Medicine’, ‘Social Sciences’, ‘Neuroscience’, ‘Engineering’, and ‘Biochemistry, Genetics and Molecular Biology’. The purpose of this sample is to seed the search process and ensure disciplinary diversity across the final citation network;
- 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 citation relationship (‘cited by’ or ‘citing’) with at least 10 articles in the seed set. The choice of 10 articles was made heuristically: a lower threshold would result in a larger, but significantly sparser citation network, while a higher threshold would lead to smaller network that may lack sufficient density at the margins between the disciplines. While expanding the network, we do not limit our search to the original ASJC sample topics;
- iv Filter the resulting network by removing any articles for which abstract text is unavailable or in a language other than English, and retain only the largest fully connected component in the network.

This process produced a densely-connected, multidisciplinary citation graph with title and abstract text for each article, and included 836,857 citations among the 58,513 research papers. The earliest paper was published in 1893 and the latest papers were published in 2022, but more than 96% of the papers were published in the period 2000–2022. We used a language model that was pre-trained on a large scientific corpus to represent the text or articles as dense document vectors. Specifically, for each article, we encoded its concatenated title and abstract using the pre-trained SciBERT model (Beltagy et al., 2019) and used the final representation of the [CLS] token to represent each document. The dataset is available at <https://doi.org/10.7910/DVN/P62Q5V>.

4.2 Representations for Recommendation

We implement and evaluate four methods for research paper embedding:

1. *TF-IDF*, an established baseline for most document representation tasks;
2. *GraphSAGE* (Hamilton, Ying, & Leskovec, 2017), a scalable GNN architecture, commonly used for research paper representation learning and recommendation (Gao et al., 2023; Jiang et al., 2023; Kozłowski et al., 2021);
3. *SPECTER* (Cohan et al., 2020), an NLP-based approach also applied in downstream recommendation tasks (Ali et al., 2022; Church et al., 2024; Singh et al., 2023);
4. *ComBSAGE* (Cunningham & Greene, 2023), a novel approach to paper embedding based on the *GraphSAGE* architecture.

For completeness, we now briefly outline the recent Community-Based Sample and Aggregation (ComBSAGE) (Cunningham & Greene, 2023) 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. This generates scientific document embeddings that are more appropriate for 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 is related to the *over-smoothing* phenomenon, a known weakness in many existing GNN architectures, where node representations converge and become indistinguishable from their neighbours (D. Chen et al., 2020). It also connects to the *over-squashing* problem, where information is lost in nodes that act as bottlenecks in the network (Topping, Giovanni, Chamberlain, Dong, & Bronstein, 2022). The proposed method of message aggregation splits messages from distinct communities, such that they can be aggregated separately before being combined (see Fig 3). 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, such that it may be recommended to or retrieved by authors in those secondary disciplines.

In the next section, we will outline the downstream RP-Rec-Sys framework that we apply to the chosen embedding methods, in an effort to characterise the differences in the recommendations that they provide.

4.3 Generating Recommendations

To isolate the effect of the representation learning phase on downstream recommendations, we implement a simple recommendation framework (similar to (Bhagavatula, Feldman, Power, & Ammar, 2018)), that can be applied independently of the upstream embedding method. Given a query paper q from a set of research papers P , we seek to recommend the top k relevant papers according to the paper descriptions or feature space H . That is, learn some function $g(h_q, h_i) = p(\text{recommend paper } i \mid \text{query paper } q)$, which computes the probability of recommending paper i , given query q (or the probability that paper i is relevant to query paper q). Thus, a set of recommendations is realised as the top k papers which maximise g for a given query.

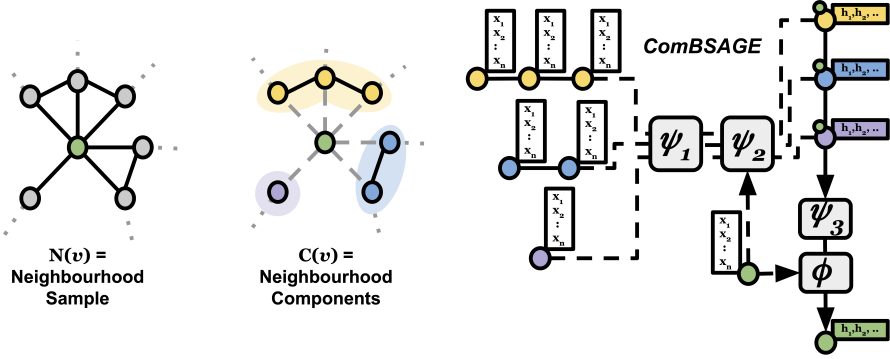


Fig. 3 One layer of the ComBSAGE MPGNN framework. Unlike traditional MPGNNs, messages are not aggregated uniformly across all neighbours. The neighbours of node v are grouped by the neighbourhood component function C . Messages from each component in $C(v)$ are aggregated separately by ψ_1, ψ_2 , before combination (ψ_3) and update (ϕ). For full details on the architecture, see the workshop proceedings (Cunningham & Greene, 2023).

To evaluate the recommendations produced by different systems, we divide the citation graph $G = (V, E)$ into discrete time steps. For a given year y , $G_y = (V_y, E_y)$ contains all papers published up to y , $V_{y-2} \subseteq V_{y-1} \subseteq V_y$, similarly $E_{y-2} \subseteq E_{y-1} \subseteq E_y$. Thus, the citation graph for 2016 contains all papers published up to and including 2016, and any citations between those papers: $G_{2014} \subseteq G_{2015} \subseteq G_{2016}$. The paper metadata/content, as described by X , remains constant after publication. According to each paper embedding method, we train an embedding model f on all data up to and including 2016. We then infer paper embeddings for all papers in that period: $f(G_{2016}, X) = H$, and train a multi-layer perceptron (MLP) recommendation function g , on any co-citing pairs where the co-citing paper is published no later than 2017. The MLP g is trained to compute the probability of a co-citation occurring between a paper i and query paper q , based on their respective embeddings h_i and h_q . Accordingly, the probability of recommending a paper i , given query q is given by

$$p(\text{recommend paper } i \mid \text{query paper } q) = \sigma(W_2 \cdot a + b_2) \quad (3)$$

where $a = \sigma(W_1 \cdot \text{CONCAT}(h_q, h_i) + b_1)$, W_1, b_1 and W_2, b_2 are the weights and biases of layers 1 and 2 of the MLP, and σ is a non-linear activation function. The use of small (two-layer) feed-forward neural networks as the recommendation layer in larger RP-Rec-Sys is well established in the literature (Bhagavatula et al., 2018; Huang, Wu, Liang, Mitra, & Giles, 2015; Li, Chen, Pettit, & Rijke, 2019). Finally, we evaluate each approach to embedding using f and g to make recommendations for every paper published in 2017 (6211 papers). A recommendation is considered to be *relevant* if the query paper and recommended paper are co-cited any time after 2017 (i.e., during 2017–2023 in our data). Any co-citation between two papers according to Semantic Scholar is recorded, so we do not limit this relationship to only the co-citations which

occur in the citation network we have collected. If there is a direct citation between the query paper and some co-cited paper, then we exclude that co-cited paper from the list of papers relevant to the query paper. This is done to avoid rewarding a system that makes recommendations directly from a query paper’s bibliography or citations. We choose 2017 as the year for evaluation so that the citation network is sufficiently dense for training purposes, while also allowing the query papers to have sufficient time to accumulate citations, and thus co-citations/relevant papers.

4.4 Evaluating Recommendations

Using co-citation as a ground truth, we adopt two common metrics for assessing the relevance of a set of recommendations: *precision* and *recall*. For top- k recommendations, $\text{precision}@k$ refers the proportion of recommended articles that are relevant, while $\text{recall}@k$ is proportion of the relevant articles that are recommended. In our later experiments, we report precision and recall at $k \in \{10, 20, 30\}$, where the lower end is chosen to represent the type of recommendations that might be regularly offered to a researcher/user (e.g. a short-list of recommendations in a weekly email), while the upper end could represent the case of an author conducting a literature review. Precision and recall scores are presented in Table 1. Every paper published in 2017 is used a query paper, and we use bootstrapping to report means and standard errors. In practice, we consider recommendation relevance to be paramount across all RP-Rec-Sys use-cases. In particular, in a retrieval use-case, recommendation precision should be valued highly, while in a literature review setting, recommendation precision and recall would be the most important metrics. We also evaluate the overall ordering of recommendations using the area under the ROC curve (AUC), normalised Discounted Cumulative Gain (nDCG), and Mean Reciprocal Rank (MRR). The formula for computing AUC for a query q given a list of relevant items R is offered in Equation 4 and can be interpreted as the likelihood that a random relevant item is recommended above a random irrelevant item.

$$\text{AUC}(R)_q = \frac{1}{|R|(n - |R|)} \sum_{r \in R} \sum_{r' \notin R} \delta(r < r') \quad (4)$$

Here $\delta(r < r')$ has the value 1 if the relevant item r is ranked above the irrelevant item r' and 0 otherwise. Similar to AUC, nDCG evaluates the ordering of the recommended items, but it includes an additional weighting to penalise cases where irrelevant items are ranked very highly. Equation 5 describes the calculation of Discounted Cumulative gain with a binary definition of relevance.

$$\text{DCG}(R)_q = \sum_{i=1}^{i \leq k} \frac{\delta(i)}{\log_2(i+1)} \quad (5)$$

In this case $\delta(i)$ takes the value 1 if i is a relevant item (i.e., $i \in R$) and 0 otherwise. To normalise DCG to yield a score in the range of $[0,1]$, we divide by

the ideal DCG score for the set of recommendations. MRR simply reports the reciprocal of the average position/rank of the first relevant item. For example, an MRR of 1.0 would indicate that the top recommendation was relevant, while a score 0.1 would indicate that the 10th recommended item was relevant. AUC, nDCG, and MRR scores are presented in Table 2.

In addition to recommendation relevance, we consider the recommendation novelty and recommendation diversity of the result sets. Tables 3, 4, 5, and 6 report the novelty and diversity of recommendations papers, according to both the network and content perspectives as discussed in Section 3. Recommendation novelty and recommendation diversity may have broad appeal across many RP-Rec-Sys applications. For instance, they are valuable in minimising redundant recommendations, as discussed in Section 2.2. Yet, they are uniquely important to the use-case central to this work: promoting interdisciplinary research, reading and citation patterns using RP-Rec-Sys. The results of these experiments are presented in Section 5 below. Code for the experiments is available at https://github.com/eoghancunn/facilitating_idr.

5 Results and Discussion

Table 1 presents the precision and recall results for the content-based recommendation evaluation described in Section 3. The GNN-based methods (*GraphSAGE* and *ComBSAGE*) outperform *SPECTER* in recommendation precision. While some of the differences are significant ($p < 0.05$), they are modest. *GraphSAGE* has the highest recommendation recall, with no significant difference between *SPECTER* and *ComBSAGE* (at %95 confidence). According to these results, we conclude that the state-of-the-art methods perform with similar precision, and given the improved recall achieved by the *GraphSAGE* method, it should be preferred for use-cases like literature review. Since the TF-IDF baseline is not competitive with the other state-of-the-art methods, we will not consider the TF-IDF results further. It would be misleading to compare the recommendation diversity and novelty of two different sets of recommendations that have very different relevance scores, as the utility of diverse or novel recommendations is contingent on their relevance.

Table 2 reports the AUC scores for recommendations offered by each embedding method. These scores show that throughout the entire ordering of recommended papers, the *ComBSAGE* method is competitive with the state-of-the-art embedding methods when recommending relevant papers ahead of irrelevant papers. The nDCG scores confirm this result, although this metric shows that the *GraphSAGE* method may be less likely to include irrelevant recommendations in the highest recommendations. Finally, the MRR scores measure the average position of the first relevant recommendation. According to these values, *GraphSAGE*, *ComBSAGE* and *SPECTER* offer their first relevant recommendation at positions 3.3, 3.8 and 4.3 respectively.

	Precision ($\mu \pm 2\sigma$)			Recall ($\mu \pm 2\sigma$)		
	@10	@20	@30	@10	@20	@30
TF-IDF	0.037 $\pm$ 0.002	0.039 $\pm$ 0.002	0.040 $\pm$ 0.002	0.002 $\pm$ 0.000	0.003 $\pm$ 0.000	0.005 $\pm$ 0.001
GraphSAGE	0.134 $\pm$ 0.005	0.129 $\pm$ 0.005	0.125 $\pm$ 0.004	0.018 $\pm$ 0.002	0.029 $\pm$ 0.002	0.041 $\pm$ 0.003
SPECTER	0.126 $\pm$ 0.005	0.117 $\pm$ 0.004	0.113 $\pm$ 0.004	0.014 $\pm$ 0.002	0.023 $\pm$ 0.002	0.031 $\pm$ 0.002
CombSAGE	0.138 $\pm$ 0.005	0.133 $\pm$ 0.005	0.130 $\pm$ 0.005	0.012 $\pm$ 0.001	0.022 $\pm$ 0.002	0.031 $\pm$ 0.002

Table 1 Precision and recall scores for 4 different paper embedding methods. Each method is used to produce the content representations for research papers in a content-based recommendation experiment. Every paper published in 2017 is given as a query to the system, and the model recommends the top 10, 20 and 30 papers. A recommendation is considered relevant if the query paper and recommended paper are co-cited at some point in the future (up to 2023). We compute precision and recall scores for all papers and use bootstrapping to report mean and standard deviation.

	AUC	nDCG	MRR
GraphSage	0.838	0.450	0.300
SPECTER	0.822	0.427	0.233
ComBSAGE	0.840	0.440	0.264

Table 2 Area under the ROC curve (AUC), normalised Discounted Cumulative Gain (nDCG) and Mean Reciprocal Rank (MRR) for 3 different paper embedding methods. Each method is used to produce the content representations for research papers in a content-based recommendation experiment. Every paper published in 2017 is given as a query to the system, and all possible recommendations are ranked. A recommendation is considered relevant if the query paper and recommended paper are co-cited at some point in the future (up to 2023). We compute scores for all queries and report the mean.

Tables 3 and 4 report recommendation novelty scores. As discussed in Section 3.2, recommendation novelty is measured as the average distance/dissimilarity between a query paper and a recommended paper, and is measured from two perspectives. Table 3 reports the *article content* recommendation novelty scores measured using distances between *SciBERT* embeddings, and Table 4 reports *citation network* recommendation novelty scores measured using distances between *DeepWalk* embeddings. According to Table 3, the textual content of recommendations made by the *ComBSAGE* model are more distant from the query than those made by the *GraphSAGE* or *SPECTER* models. A similar pattern is also apparent in Table 4. That is, *ComBSAGE* recommends research papers that are further from the query in the citation graph; these are more novel recommendations in the sense of Equation 1. We also compute the novelty of the subset of the recommended papers that are *relevant* to the query. Again we find that the *ComBSAGE* model provides the most distant recommendations even when we consider only the relevant papers.

To help contextualise the benefits of the proposed *ComBSAGE* embedding method, and to verify the utility of our proposed metrics, consider the following: the average shortest path distance between a pair of papers that received a co-citation in 2017 is 2.3 citations. At $k = 10$, the average *relevant* paper

	Novelty (Title+Abstract Distance)					
	Recommended			(Relevant Subset)		
	@10	@20	@30	@10	@20	@30
GraphSAGE	0.28	0.28	0.28	0.20	0.20	0.20
SPECTER	0.28	0.28	0.28	0.22	0.22	0.22
CombSAGE	0.31	0.30	0.30	0.25	0.25	0.25

Table 3 Novelty of recommendations according to *SciBERT* embeddings of query and recommended papers. At $k = 10$, the increase in novelty of the *relevant* recommendations offered by *CombSAGE* (according to the Cohen’s d effect size), is ‘*moderate*’ when compared to the next most precise method, *GraphSAGE* ($d = 0.42$), and ‘*small*’ when compared to *SPECTER* ($d = 0.2$).

	Novelty (Graph Distance)					
	Recommended			(Relevant Subset)		
	@10	@20	@30	@10	@20	@30
GraphSAGE	0.55	0.55	0.55	0.37	0.37	0.37
SPECTER	0.58	0.58	0.58	0.42	0.42	0.42
CombSAGE	0.59	0.58	0.58	0.46	0.45	0.45

Table 4 Novelty of recommendations according to *DeepWalk* embeddings of query and recommended papers. At $k = 10$, the increase in novelty of the *relevant* recommendations offered by *CombSAGE* (according to the Cohen’s d effect size), is ‘*moderate*’ when compared to the next most precise method, *GraphSAGE* ($d = 0.43$), and ‘*small*’ when compared to *SPECTER* ($d = 0.07$).

recommended by the GraphSAGE model has a shortest path distance of 2.3 citations from the query. According to these numbers, the recommendations provided by the *GraphSAGE* model occur at a distance similar to the connections and combinations identified by authors in that same year. This distance (2.3 citations) is consistent with numbers claimed by Fortunato et al. (2018): “*new links fall between things only one or two steps away from each other, implying that when scientists choose new topics, they prefer things directly related to their current expertise*”. Meanwhile, the average *relevant* paper recommended by the model employing *CombSAGE* has a shortest path distance of 2.6 citations from the query, confirming that models that maximise our proposed novelty metrics, provide more distant recommendations than models with lower scores. Further, we find that the connections recommended by the models with the greatest diversity and novelty are more distant on average than the connections made by authors in the same year, indicating the potential for RP-Rec-Sys to introduce authors to relevant works that are *further* from their primary focus. According to Cohen’s d effect sizes, the increases in novelty (or ‘unexpectedness’) achieved by the *CombSAGE* model range from *small* to *moderate* (see table captions). Crucially, these increases come at no cost to precision.

Finally, Tables 5 and 6 report the recommendation diversity, measured as the average pairwise distance/dissimilarity between recommended papers,

given the *SciBERT* embeddings and *DeepWalk* embeddings for papers respectively. According to Table 5 (*article content* recommendation diversity), the set of recommendations made by the *CombSAGE* model are more semantically diverse than those made by either the *GraphSAGE* or *SPECTER* models. Similarly, *CombSAGE* recommends research papers that are distributed more widely across the full citation graph (see Table 6, *citation network* recommendation diversity). Although we calculate and include the diversity scores for the subset of recommendations that are *relevant* for completeness, those scores are calculated on small sets of papers and thus may not be informative. Again, we find that effects according to Cohen’s *d* (see table captions) range from *small* to *moderate*.

	Diversity (Title+Abstract Distance)					
	Recommended			(Relevant Subset)		
	@10	@20	@30	@10	@20	@30
GraphSAGE	0.26	0.27	0.27	0.10	0.12	0.14
SPECTER	0.25	0.26	0.27	0.10	0.12	0.13
CombSAGE	0.27	0.28	0.29	0.11	0.13	0.15

Table 5 Diversity of recommendations according to *SciBERT* embeddings of recommended papers. At $k = 10$, the increase in diversity of the recommendations offered by *CombSAGE* (according to the Cohen’s *d* effect size), is ‘*small*’ when compared to the next most precise method, *GraphSAGE* ($d = 0.27$), and ‘*moderate*’ when compared to *SPECTER* ($d = 0.5$).

	Diversity (Graph Distance)					
	Recommended			(Relevant Subset)		
	@10	@20	@30	@10	@20	@30
GraphSAGE	0.51	0.53	0.55	0.18	0.22	0.25
SPECTER	0.52	0.56	0.57	0.19	0.23	0.26
CombSAGE	0.55	0.58	0.59	0.21	0.25	0.28

Table 6 Diversity of recommendations according to *DeepWalk* embeddings of recommended papers. At $k = 10$, the increase in diversity of the recommendations offered by *CombSAGE* (according to the Cohen’s *d* effect size), is ‘*small–moderate*’ when compared to the next most precise method, *GraphSAGE* ($d = 0.31$), and ‘*small*’ when compared to *SPECTER* ($d = 0.09$).

6 Conclusions

As researchers become increasingly reliant on automated systems for research paper recommendation and retrieval, the risks of the filter-bubble effect leading to research isolation become greater. Despite the attention and value

ascribed to *diversity* and *novelty* metrics in the broader recommender systems literature (Castells et al., 2021; Smyth & McClave, 2001), they have been largely overlooked in RP-Rec-Sys contexts (Ali et al., 2021; Kreutz & Schenkel, 2022). Diverse and novel recommendations are desirable for many reasons, (see Section 2), but for RP-Rec-Sys systems, they are uniquely important. Research paper recommendations that are novel and/or diverse can bridge different disciplines, helping to mitigate filter-bubble effects and siloed reading habits.

In this work, we have presented new network science and NLP-based metrics for research paper recommendation diversity and novelty. While remaining consistent with the consensus in the broader Rec-Sys literature, our methods are tailored to the case of RP-Rec-Sys and are based on measures of disciplinary distance between papers. Recommendations that maximise these metrics, while remaining relevant to the user and their queries, can actively promote interdisciplinary research, by facilitating the discovery of latent connections between otherwise distant topics. In pursuit of this goal, we have evaluated four different research paper representation methods in an RP-Rec-Sys, and we have demonstrated how a recent approach (*ComBSAGE*), can produce more more diverse and novel recommendations without compromising precision, when compared to other state-of-the-art methods. Thus we show the potential for RP-Rec-Sys to facilitate interdisciplinary knowledge transfer and prevent the perpetuation of information silos.

As a secondary contribution, we can characterise the nature of recommendations provided by alternative methods, such that they might be prescribed to the use-case for which they are best suited. For example, with the highest recall, particularly at larger recommendation sizes ($k = 30$), the *GraphSAGE* method may be the most appropriate choice of embedding method for scholars conducting literature review. In this scenario, a high recall on the relevant documents is likely to be paramount, whereas diverse and unexpected recommendations are potentially less useful. In this way, we recognise the measurable impact of document embedding methods on the quality and nature of downstream recommendations. Thus, we have demonstrated that the choice of representational approach within RP-Rec-Sys is important and should be grounded in the system’s intended use-case or target audience.

Going forward, widespread adoption of our proposed metrics will encourage greater interdisciplinary and disciplinary diversity in RP-Rec-Sys. There are many avenues for future work in the pursuit of interdisciplinary knowledge transfer. First of all, other components within RP-Rec-Sys could be optimised towards greater recommendation diversity and novelty. For example, candidate reordering approaches have been proposed for movie recommendation to induce diversity in the recommendation set (Boim, Milo, & Novgorodov, 2011; M. Zhang & Hurley, 2008). Further, specific tools could be developed to help users to tackle the language-barrier associated with reading research from other disciplines. For example, literature mapping tools such as the VOSViewer (Van Eck & Waltman, 2010) could be employed to compare the

usage of key words across different recommended research areas from the specific perspective of the user and their query.

References

- Achakulvisut, T., Acuna, D.E., Ruangrong, T., Kording, K. (2016). Science concierge: A fast content-based recommendation system for scientific publications. *PLOS ONE*, 11(7), e0158423.
- Ali, Z., Qi, G., Kefalas, P., Khusro, S., Khan, I., Muhammad, K. (2022, nov). SPR-SMN: Scientific paper recommendation employing specter with memory network. *Scientometrics*, 127(11), 6763–6785.
- Ali, Z., Ullah, I., Khan, A., Ullah Jan, A., Muhammad, K. (2021). An overview and evaluation of citation recommendation models. *Scientometrics*, 126, 4083–4119.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17), 3389–3402.
- Amami, M., Pasi, G., Stella, F., Faiz, R. (2016). An LDA-based approach to scientific paper recommendation. *Natural language processing and information systems: 21st international conference on applications of natural language to information systems* (pp. 200–210).
- Asatani, K., Mori, J., Ochi, M., Sakata, I. (2018). Detecting trends in academic research from a citation network using network representation learning. *PloS ONE*, 13(5), e0197260.
- Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., Xia, F. (2019). Scientific paper recommendation: A survey. *IEEE Access*, 7, 9324–9339.
- Beltagy, I., Cohan, A., Lo, K. (2019). SciBERT: Pretrained contextualized embeddings for scientific text. *arXiv preprint: 1903.10676*.
- Bhagavatula, C., Feldman, S., Power, R., Ammar, W. (2018). Content-based citation recommendation. *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human*

- language technologies, volume 1 (long papers) (pp. 238–251).
- Boim, R., Milo, T., Novgorodov, S. (2011). Diversification and refinement in collaborative filtering recommender. *Proceedings of the 20th ACM international conference on information and knowledge management* (pp. 739–744).
- Bornmann, L., Tekles, A., Zhang, H.H., Fred, Y.Y. (2019). Do we measure novelty when we analyze unusual combinations of cited references? a validation study of bibliometric novelty indicators based on F1000Prime data. *Journal of Informetrics*, 13(4), 100979.
- 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.
- Castells, P., Hurley, N., Vargas, S. (2021). Novelty and diversity in recommender systems. *Recommender systems handbook* (pp. 603–646). New York, NY: Springer.
- Chaudhuri, A., Sinhababu, N., Sarma, M., Samanta, D. (2021). Hidden features identification for designing an efficient research article recommendation system. *International Journal on Digital Libraries*, 22(2), 233–249.
- Chen, C., McCain, K., White, H., Lin, X. (2002). Mapping scientometrics (1981–2001). *Proceedings of the American Society for Information Science and Technology*, 39(1), 25–34.
- Chen, D., Lin, Y., Li, W., Li, P., Zhou, J., Sun, X. (2020). Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, pp. 3438–3445).
- Chen, S., Song, Y., Shu, F., Larivière, V. (2022). Interdisciplinarity and impact: the effects of the citation time window. *Scientometrics*, 1–22.
- Church, K., Alonso, O., Vickers, P., Sun, J., Ebrahimi, A., Chandrasekar, R. (2024). Academic article recommendation using multiple perspectives. *arXiv preprint 2407.05836*.

- Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D.S. (2020). Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint 2004.07180*.
- Cunningham, E., & Greene, D. (2023). Graph embedding for mapping interdisciplinary research networks. *Companion proceedings of the ACM Web Conference 2023 (WWW'23)* (p. 784–789). New York, NY, USA: Association for Computing Machinery.
- Dai, T., Zhu, L., Wang, Y., Zhang, H., Cai, X., Zheng, Y. (2018). Joint model feature regression and topic learning for global citation recommendation. *IEEE Access*, 7, 1706–1720.
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint 1810.04805*.
- Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., ... others (2018). Science of science. *Science*, 359(6379), eaao0185.
- Gao, C., Zheng, Y., Li, N., Li, Y., Qin, Y., Piao, J., ... others (2023). A survey of graph neural networks for recommender systems: Challenges, methods, and directions. *ACM Transactions on Recommender Systems*, 1(1), 1–51.
- Ghorab, M.R., Zhou, D., O'connor, A., Wade, V. (2013). Personalised information retrieval: survey and classification. *User Modeling and User-Adapted Interaction*, 23, 381–443.
- Gmür, M. (2003). Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics*, 57(1), 27–57.
- Hamilton, W.L., Ying, R., Leskovec, J. (2017). Inductive representation learning on large graphs. *Proceedings of the 31st international conference on neural information processing systems* (pp. 1025–1035).
- Hua, S., Chen, W., Li, Z., Zhao, P., Zhao, L. (2020). Path-based academic paper recommendation. *Proceedings of the 21st international conference*

on web information systems engineering (WISE 2020), Part II (pp. 343–356).

- Huang, W., Wu, Z., Liang, C., Mitra, P., Giles, C. (2015). A neural probabilistic model for context based citation recommendation. *Proceedings of the AAAI conference on artificial intelligence* (Vol. 29).
- Hui, S., Wei, M., XiaoLiang, Z., JunYan, J., YanBing, L., ShuJuan, C. (2020). A hybrid paper recommendation method by using heterogeneous graph and metadata. *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8).
- Jiang, C., Ma, X., Zeng, J., Zhang, Y., Yang, T., Deng, Q. (2023). Taprec: time-aware paper recommendation via the modeling of researchers' dynamic preferences. *Scientometrics*, 1–19.
- Jo, W.S., Liu, L., Wang, D. (2022). See further upon the giants: Quantifying intellectual lineage in science. *Quantitative Science Studies*, 3(2), 319–330.
- Kaminskas, M., & Bridge, D. (2016). Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1), 1–42.
- Koren, Y., Rendle, S., Bell, R. (2021). Advances in collaborative filtering. *Recommender Systems Handbook*, 91–142.
- Kozłowski, D., Dusdal, J., Pang, J., Zilian, A. (2021). Semantic and relational spaces in science of science: deep learning models for article vectorisation. *Scientometrics*, 1–30.
- Kreutz, C.K., & Schenkel, R. (2022). Scientific paper recommendation systems: a literature review of recent publications. *International Journal on Digital Libraries*, 23(4), 335–369.
- Larivière, V., Haustein, S., Börner, K. (2015). Long-distance interdisciplinarity leads to higher scientific impact. *PLOS ONE*, 10(3), e0122565.

- Li, X., Chen, Y., Pettit, B., Rijke, M.D. (2019, mar). Personalised reranking of paper recommendations using paper content and user behavior. *ACM Transactions on Information Systems (TOIS)*, 37(3).
- Liu, J., Kong, X., Zhou, X., Wang, L., Zhang, D., Lee, I., ... Xia, F. (2019). Data mining and information retrieval in the 21st century: A bibliographic review. *Computer Science Review*, 34, 100193.
- Lops, P., De Gemmis, M., Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. *Recommender Systems Handbook*, 73–105.
- Markus, T., Hillebrand, H., Hornidge, A.-K., Krause, G., Schlüter, A. (2018). Disciplinary diversity in marine sciences: the urgent case for an integration of research. *ICES Journal of Marine Science*, 75(2), 502–509.
- Mei, X., Cai, X., Xu, S., Li, W., Pan, S., Yang, L. (2022). Mutually reinforced network embedding: An integrated approach to research paper recommendation. *Expert Systems with Applications*, 204, 117616.
- Nakatsuji, M., Fujiwara, Y., Tanaka, A., Uchiyama, T., Fujimura, K., Ishida, T. (2010). Classical music for rock fans? novel recommendations for expanding user interests. *Proceedings of the 19th ACM international conference on information and knowledge management* (pp. 949–958).
- Nishioka, C., Hauke, J., Scherp, A. (2019). Research paper recommender system with serendipity using tweets vs. diversification. *Proceedings of digital libraries at the crossroads of digital information for the future: 21st international conference on asia-pacific digital libraries (ICADL 2019)* (pp. 63–70).
- Nogueira, R., Jiang, Z., Cho, K., Lin, J. (2020). Navigation-based candidate expansion and pretrained language models for citation recommendation. *Scientometrics*, 125, 3001–3016.
- Okamura, K. (2019). Interdisciplinarity revisited: evidence for research impact and dynamism. *Palgrave Communications*, 5(1), 1–9.
- Portenoy, J., Radensky, M., West, J.D., Horvitz, E., Weld, D.S., Hope, T. (2022). Bursting scientific filter bubbles: Boosting innovation via novel

- author discovery. *Proceedings of the 2022 CHI conference on human factors in computing systems* (pp. 1–13).
- Porter, A., Cohen, A., David Roessner, J., Perreault, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics*, 72(1), 117–147.
- Qi, F., Zhou, H., Sun, B., Huang, Y., Zhang, L. (2024). Facilitating interdisciplinarity: the contributions of boundary-crossing activities among disciplines. *Scientometrics*, 1–19.
- Ricci, F., Rokach, L., Shapira, B. (2015). Recommender systems: introduction and challenges. *Recommender Systems Handbook*, 1–34.
- Rodriguez-Prieto, O., Araujo, L., Martinez-Romo, J. (2019). Discovering related scientific literature beyond semantic similarity: a new co-citation approach. *Scientometrics*, 120, 105–127.
- Sharma, B., Willis, V.C., Huettner, C.S., Beaty, K., Snowdon, J.L., Xue, S., . . . Micheleni, V. (2020). Predictive article recommendation using natural language processing and machine learning to support evidence updates in domain-specific knowledge graphs. *JAMIA Open*, 3(3), 332–337.
- Shi, F., & Evans, J. (2023). Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *Nature Communications*, 14(1), 1641.
- Shi, Y., Larson, M., Hanjalic, A. (2014). Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)*, 47(1), 1–45.
- Shin, H., Kim, K., Kogler, D.F. (2022, 07). Scientific collaboration, research funding, and novelty in scientific knowledge. *PLOS ONE*, 17(7), 1–22.
- Singh, A., D’Arcy, M., Cohan, A., Downey, D., Feldman, S. (2023). Scirepeval: A multi-format benchmark for scientific document representations. *Conference on empirical methods in natural language processing (EMNLP 2023)* (pp. 5548–5566).

- 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.
- Smyth, B. (2007). Case-based recommendation. *The adaptive web: methods and strategies of web personalization* (pp. 342–376). Springer.
- Smyth, B., & McClave, P. (2001). Similarity vs. diversity. D.W. Aha & I. Watson (Eds.), *Case-based reasoning research and development* (pp. 347–361). Springer Berlin Heidelberg.
- Topping, J., Giovanni, F.D., Chamberlain, B.P., Dong, X., Bronstein, M.M. (2022). Understanding over-squashing and bottlenecks on graphs via curvature. *Proceedings of the 10th international conference on learning representations (ICLR)*.
- Trujillo, C.M., & Long, T.M. (2018). Document co-citation analysis to enhance transdisciplinary research. *Science advances*, 4(1), e1701130.
- Uzzi, B., Mukherjee, S., Stringer, M., Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342(6157), 468–472.
- Van Eck, N., & Waltman, L. (2010). Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538.
- Vargas, S., & Castells, P. (2011). Rank and relevance in novelty and diversity metrics for recommender systems. *Proceedings of the 5th ACM conference on recommender systems* (pp. 109–116).
- Yu, C., Wang, C., Zhang, T., Bu, Y., Xu, J. (2023). Analyzing research diversity of scholars based on multi-dimensional calculation of knowledge entities. *Scientometrics*, 1–30.
- Zhang, J., & Zhu, L. (2022). Citation recommendation using semantic representation of cited papers' relations and content. *Expert Systems with Applications*, 187, 115826.
- Zhang, L. (2013). The definition of novelty in recommendation system. *Journal of Engineering Science & Technology Review*, 6(3).

Zhang, M., & Hurley, N. (2008). Avoiding monotony: improving the diversity of recommendation lists. *Proceedings of the 2008 ACM conference on recommender systems* (pp. 123–130).

Acknowledgments

This research was supported by Research Ireland under Grant Number SFI/12/RC/2289_P2. This work builds substantially on a short paper previously published in workshop proceedings ([Cunningham & Greene, 2023](#)).

Author Contributions

EC: Conceptualisation, Data Curation, Formal Analysis, Investigation, Methodology, Visualisation, Writing – Original Draft

BS: Conceptualisation, Validation, Writing – Review & Editing

DG: Conceptualisation, Funding Acquisition, Validation, Writing – Review & Editing

Competing Interests

The authors declare no competing interests.