RI:Small: Consequences of Network Effects on Graph Learning Methods

Project Summary

Overview: Network effects refer to the fact that nodes, edges and paths scale at different rates: n, n^2 and 2^n, respectively. This proposal will study important connections between these scaling properties and graph learning methods such as graph neural networks (GNNs), random walks and large language models (LLMs such as BERT and ChatGPT). Graph learning has applications in web search (Page Rank), Product Search (Amazon), Biology, Finance and Traffic Analysis for military intelligence. Since much of this data is sensitive, this proposal will focus on applications in academic search where Semantic Scholar (S2) provides bulk download access to a large graph (with more than 200 million academic papers and 2 billion citations). S2 updates this data weekly, which is important since the literature is growing exponentially, doubling every 9 years (or perhaps 19 years). A number of preliminary experiments were performed on G_t, the citation graph at time t. Figure 1 compares two methods, ProNE (spectral clustering of the citation graph) and Specter (a BERT-like encoding of abstracts). One method is better than the other on large graphs ($t \to \infty$).

Keywords: Artificial Intelligence; Network Effects; Graph Neural Networks (GNNs); Spectral Clustering; Citation Graphs

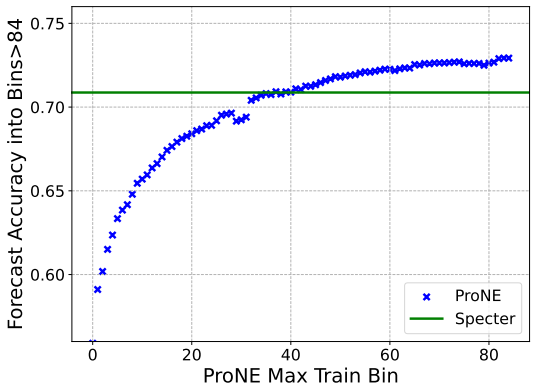


Figure 1: Unreasonable effectiveness of spectral clustering on large graphs.

Intellectual Merit: Large graphs are not like small graphs because of network effects. Much of the work on graph learning is based on benchmarks such as Open Graph Benchmark (OGB) and SciRepEval, much smaller than the crossover point in Figure 1. Just as research on algorithms values asymptotic performance on large problems, so too, performance on large problems should become a priority for research on graph learning. Large graphs are important for academic search, given the scale and growth.

Multiple embeddings create opportunities for robustness. Specter suffers from corner cases involving abstracts (missing abstracts, duplicate documents and abstracts in non-English languages). ProNE suffers from other corner cases such as papers with few (if any) links in the citation graph. Other methods such as GNNs (graph neural networks) assume most papers have both abstracts and links in graph, but Semantic Scholar (S2) is not like that: 70% of S2 papers are missing abstracts and/or links. Methods will be proposed for imputing missing values.

Broader Impacts: Improvements in recommender systems will make it easier for readers to find papers they should read, and authors to find papers they should cite, and program committees and funding agencies to assign submissions to reviewers that are well-informed and sympathetic to the topic area. Better recommendations will improve reviews (and the literature). Improvements on graph learning should transfer from academic search to recommendation systems based on more sensitive data such as product search, finance and traffic analysis.

Tools/Community/Education: In prior work at JSALT-2023, we created a search engine that encourages users to experiment with different embeddings and recommendation APIs. We hope to evolve the search engine into a hub like HuggingFace, where the community is encouraged to contribute embeddings and APIs, creating educational opportunities.

Project Description

Introduction

Intellectual Merit: Small graphs are not representative of large graphs because nodes, edges and paths scale at different rates: n, n^2 and 2^n. Much of the graph learning literature is based on benchmarks such as the Open Graph Benchmark (OGB) [1, 2] and SciRepEval [3] that measure a single data point (a single graph) well to the left of the crossover point in Figure 1. Just as research on algorithms focuses on asymptotic performance, performance on large graphs should become a priority for research on graph learning. Large graphs are important for academic search because the literature is already large, and will become exponentially larger, as the literature doubles between 9 years [4] and 19 years [5]. This project will develop, analyze and evaluate graph learning methods for G_t, citation graphs at time t, as $t \to \infty$.

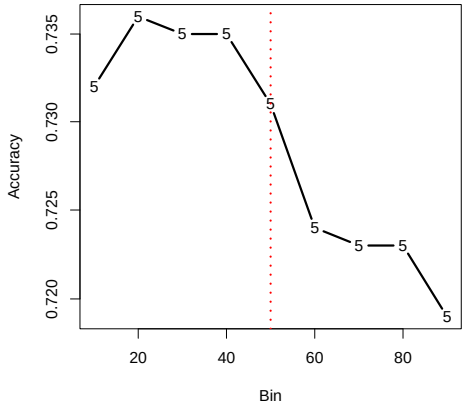


Figure 2: Accuracy degrades with h.

This project will also study forecasting and time invariance of graph learning methods. At JSALT-2023, our Better Together Team [6, 7] compared node-based embeddings and edge-based embeddings of 200 million papers in Semantic Scholar (S2) [8, 9]. We performed a number of preliminary experiments on G_t (labeled *bin*). Figure 2 uses a link prediction task [10, 11, 12] to test an edge-based model, ProNE [13], trained on $t = 50$ bins (denoted by the dashed red line). Accuracy on bin $t + h$ degrades with h, the forecasting horizon.

The rate of degradation indicates how often ProNE embeddings need to be updated. Some embeddings are time invariant and some are not. Embeddings based on edges (ProNE) improve with time, t, as papers receive citations years after publication. In contrast, embeddings based on nodes (Specter [14]) are time invariant because abstracts do not change after publication.

Graph Neural Networks (GNNs) [15, 16, 17, 18, 19] are a popular method for combining node and edge features into a single end-to-end model. On the other hand, there are advantages to a modular design with separate embeddings for nodes and edges because of (1) network effects, (2) computational requirements, (3) time invariance and (4) robustness to corner cases and missing values.

Broader Impacts (Tools/Community/Education):
In addition to experiments mentioned above, JSALT-2023 produced an academic search engine [20] shown in Figure 3. This UX encourages users to experiment with different embeddings and recommendation APIs. There are many academic search engines based on a single embedding/API, but we have found it useful to compare alternatives to learn what is good for what. The drop down menu in Figure 3 currently supports 8 choices [13, 14, 21, 22], and more will be added soon such as a recommen-

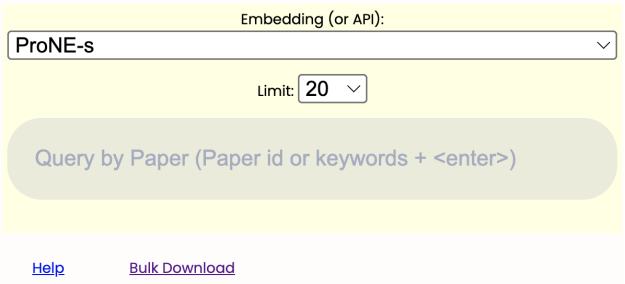


Figure 3: User Interface (UX)

Query	ProNE	S2 API	Description
[25]	[26, 27, 28, 29, 30]	[31, 32, 33, 34, 35]	Recommender systems
[36]	[37, 38, 39, 40, 41]	[42, 43, 44, 45, 46, 47]	Who should review what?
[48]	[49, 50, 51, 52]	[53, 54, 55, 56, 57]	Citation Recommendation
[3]	NA	[58, 59, 60, 61, 62]	Specter Benchmark
[63]	NA	[64, 42, 65, 66, 67]	A new benchmark for who should review what
[68]	NA	[69, 70, 71, 72, 73]	Who is who?
[74]	[75, 76, 77, 78, 79]	[80, 81, 82, 83, 84]	RAG

Table 1: Recommendations from the JSALT-2023 academic search engine.

dation API from NLM (National Library of Medicine) [23]. We envision the web site evolving into a hub like HuggingFace [24], where the community is encouraged to contribute embeddings and APIs, creating educational opportunities to learn about both the tools as well as the literature.

Improvements in recommender systems will make it easier for readers to find papers they should read, and authors to find papers they should cite, and program committees and funding agencies to assign submissions to reviewers that are well-informed and sympathetic to the topic area. Better recommendations will improve reviews (and the literature). Improvements on graph learning should transfer from academic search to recommendation systems based on more sensitive data such as product search, finance and traffic analysis.

Related Work

This project will extend much of the work cited below in three directions: (1) scale/network effects, (2) incremental updates/time invariance and (3) robustness to corner cases and missing values. One of the motivations for developing the JSALT-2023 academic search engine was to help authors write sections like this section on related work. When writing survey articles and sections on related work, there are four subtasks:

1. create a list of papers to discuss,
2. organize the list appropriately (by topic and/or time),
3. summarize papers, and
4. explain how those papers are relevant to the present discussion

Table 1 shows some output from the JSALT-2023 academic search engine. We have found this tool to be helpful for creating a list of papers to discuss. If one can find some good queries (first column in Table 1), then the tool finds many promising candidates. ProNE and S2 produce different candidates. Both are useful but in different ways. Recommendations from ProNE tend to have more citations and recommendations from Specter tend to be more recent.

It is more challenging to organize the recommendations, summarize the papers, and explain how they are relevant to the present discussion. It might be possible for RAG [74] to write a first draft, though that possibility raises obvious ethical questions. The JSALT-2023 search engine provides options to use chatbots and RAG to summarize papers one at a time, compare and contrast pairs of papers. The chatbot output is impressive, at least on first impression, but after a while, the output feels uninspired, superficial, repetitive and long-winded. S2 offers tl;dr summaries that are more useful, though tl;dr summaries do not attempt to compare and contrast pairs of papers.

There are a couple of missing values in Table 1 because ProNE is a transductive model, meaning it generates embeddings only for documents in the training set, but not for other documents.

Section 2.10.1 introduces the centroid approximation to generalize to documents beyond the training set, and Section 2.11 discusses methods to update ProNE incrementally.

In addition to the papers in Table 1, there is a considerable body of work on topics related to this proposal including: network embeddings [85], graph neural networks [15, 16, 17, 18, 19], temporal embeddings [86] and recommender systems [87]. The literature on citations is not as large, though still considerable:

- Early work: [88, 89, 90, 91, 92]
- Benchmarks: OGB, SciRepEval and others [93]
- Evaluations on benchmarks: [14, 94, 2]
- Training models on benchmarks: [95, 21]
- Survey papers: [96, 97, 57, 98, 99]

There are a number of use cases that are related to citation/link prediction:

1. For authors: what should I cite? [48]
2. For readers: what should I read? [25, 100]
3. For funding agencies and program committees: who should review what? [101, 102, 103, 42, 104, 63, 105]
4. Finding experts [106, 107, 108]: "Who knows" [109]

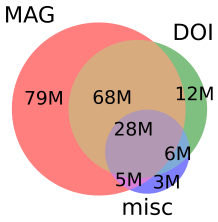


Figure 4: Semantic Scholar Sources

Given that there is so much related work, and so many interesting directions to pursue, we need to focus this project on a short list of paths forward. As mentioned above, we want to build on the references above, with an emphasis on: (1) scale, (2) computational requirements, (3) incremental updates and (4) robustness.

Materials

The 200M papers in S2 are from 7 sources: MAG (Microsoft Academic Graph) [110, 111], DOI [112], PubMed [113], PubMedCentral [114], DBLP [115], arXiv [116] and ACL [117]. There has been quite a bit of work in our field on the ACL Anthology and arXiv, but these sources are relatively small compared to MAG and DOI, as illustrated in Figure 4 where the 5 smaller sources, labeled misc, contribute only 3M papers beyond MAG ∪ DOI.

It has been suggested that most of the papers in S2 are in CS, but actually, S2 covers a diverse set of fields of study (fos): Medicine (45M), Chemistry (13M), Computer Science (13M), Biology (13M), Materials Science (10M), Engineering (8M), Physics (7M), Psychology (7M), Mathematics (5M), Political Science (4M), Business (4M), Sociology (3M), Geography (3M), Economics (3M), Environmental Science (3M), Geology (3M), History (2M), Art (2M), Philosophy (1M).

Some Examples

Figure 5 show two examples of the JSALT-2023 search engine. The query (top line) is followed by a number of recommendations. Each recommendation ends with four cosine scores, comparing the recommendation with the query, using four different embeddings [13, 21, 14, 19]. By construction, the scores are 1.0 for the top line. Since not all papers are in all embeddings, some scores can be missing (such as the GNN score for the second row in Figure 5 (top)).

	Citations	**Year**
ProNE	**13**	2013
S2	9	**2014**

Table 2: Medians

score	citationCount	Paper	Authors	year	More like this	Compare & Contrast	ProNE-s	SciNCL	Specter	GNN
	865	Personalizing Search via Automated Analysis of Interests and Activities	J. Teevan, S. Dumais, E. Horvitz	**2005**	Similar to this	Compare & Contrast	**1.0**	**1.0**	**1.0**	**1.0**
20	107	Personalizing search via automated analysis of interests and activities	J. Teevan, S. Dumais, E. Horvitz	2005	Similar to this	Compare & Contrast	0.992	1.0	1.0	
19	6	Personalizing Search via Automated Analysis of Interests and Activities	TeevanJaime, T. DumaisSusan, HorvitzEric	2018	Similar to this	Compare & Contrast	0.989	0.976	0.981	0.964
18	228	Personalizing web search using long term browsing history	N. Matthijs, Filip Radlinski	2011	Similar to this	Compare & Contrast	0.983	0.957	0.971	0.97
17	131	Enhancing personalized search by mining and modeling task behavior	Ryen W. White, Wei Chu, ..., Hongning Wang	2013	Similar to this	Compare & Contrast	0.945	0.957	0.967	0.952

score	citationCount	Paper	Authors	year	More like this	Compare & Contrast	ProNE-s	SciNCL	Specter	GNN
	865	Personalizing Search via Automated Analysis of Interests and Activities	J. Teevan, S. Dumais, E. Horvitz	**2005**	Similar to this	Compare & Contrast	**1.0**	**1.0**	**1.0**	**1.0**
0.997	564	Implicit user modeling for personalized search	Xuehua Shen, Bin Tan, ChengXiang Zhai	2005	Similar to this	Compare & Contrast	0.997	0.938	0.95	0.979
0.997	84	Beyond the Commons: Investigating the Value of Personalizing Web Search	J. Teevan, S. Dumais, E. Horvitz	2005	Similar to this	Compare & Contrast	0.997	0.919	0.933	0.99
0.996	203	Interest-based personalized search	Zhongmin Ma, Gautam Pant, O. Sheng	2007	Similar to this	Compare & Contrast	0.996	0.923	0.942	0.937

Figure 5: The query (top line) and recommendations from S2 API [118] (top) and ProNE (bottom).

The User Interface includes a number of simple, but surprisingly useful features such as a button that generates bibtex entries for all of the papers on the SRP (search results page). Most academic search engines make it easy to generate bibtex entries for one page at a time, but we have found it useful to generate bibtex entries for many papers with a single click.

Figure 5 show two buttons for each row (candidate). The "similar to this" button performs a search, using the candidate as the query. The "compare & contrast" button uses a chatbot to compare and contrast the candidate with the query.

The search engine uses approximate nearest neighbors (ANN) [119, 120, 121] to find recommendations that are near the input query, q. The two figures use the same query (top line) but different embeddings. Figure 5 (top) uses an API from S2, based on Specter Embeddings of abstracts, whereas Figure 5 (bottom) is based on ProNE embeddings of citation graphs. The two figures call out some interesting differences between the two embeddings.

1. Corner Cases and Missing Values: Since all embeddings are subject to corner cases, multiple representations create opportunities for robustness, as will be discussed in the next section. In particular, duplicate documents are more challenging for Specter than ProNE since duplicates have similar abstracts but different citations.
2. Priors: Priors played an important role in older machine learning methods such as Naive Bayes, but priors still play a role in more modern methods, though the role is less likely to be stated as explicitly. The S2 API tends to recommend more recent papers, and ProNE tends to recommend papers with more citations. These differences are shown in Table 2, which summarizes recommendations by the two methods over a set of 5273 SIGIR papers.

For both Specter and ProNE, we start with an embedding, $M \in \mathbb{R}^{N \times K}$, where N is the number of papers in the embedding and K is the number of hidden dimensions. M can be referred to as a vector database. In both cases, cosines of vectors can be interpreted as similarities of papers, though the two similarities are somewhat different. Specter similarities indicate the two abstracts use similar words, whereas ProNE similarities indicate the two papers are near one another in terms of random walks on the citation graph.

The two embeddings are computed in very different ways. For Specter, M is created by applying SciBERT [122], a BERT-like model to N abstracts. For ProNE, M is created by applying spectral clustering to the citation graph, using the nodevectors package [123]. GPUs with GBs of RAM work well for Specter, whereas CPUs with TBs of RAM are more effective for ProNE because spectral clustering is limited more by memory than computational cycles.

Outline

As mentioned above, this project will emphasize:

1. Network effects: how does performance scale the size of the graph,
2. Computational requirements: how does space and time scale with the size of the graph,
3. Time invariance: abstracts are time invariant, but citations are not
4. Robustness to corner cases and missing values.

The next section will discuss robustness. After that, we will introduce binning. Binning will be used to discuss network effects, computational requirements and time invariance. After binning, there will be a discussion of training (and computational requirements). Then we will discuss evaluation, with an emphasis on network effects. Finally, we will discuss how to impute missing values and incremental updates. These topics are relevant to robustness and computational requirements.

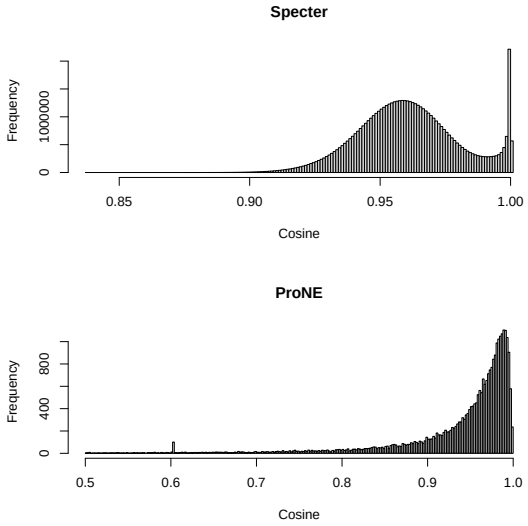


Figure 6: Histogram of $cos(q, cand_1)$ for 58M Specter recommendations (top). The spike near 1 is dominated by corner cases. ProNE is not bimodal (bottom).

Robustness To Corner Cases

Multiple redundant estimates of similarity create opportunities for robustness. Duplicates are one of many corner cases. Corner cases can often be detected by looking for large differences between Specter cosines and ProNE cosines. For example, duplicates will receive large Specter cosines because the abstracts are nearly the same, but much smaller ProNE cosines because one of the duplicates tends to have more citations than the other.

How many corner cases are there? The spike near 1 in Figure 6 (top) provides a rough estimate. This figure shows $cos(q, cand_1)$ for 58M queries, q, and the top candidate recommendation, $cand_1$. About 10% of Specter vectors in S2 have a nearby neighbor with a cosine of 0.99 or better. ProNE is very different, with a single mode and no second spike near 1, as shown in Figure 6 (bottom). Spot checks of the Specter spike near 1 suggest the spike is dominated by corner cases such as duplicates, missing/incorrect abstracts, and abstracts in languages other than English.

The last case is perhaps the most common case. Specter was designed for English. When Specter is applied to Chinese text, for example, the tokenizer returns many unknown tokens. As a result, two unrelated papers in Chinese (and many other languages) can appear to be more similar to one another than they should be. ProNE avoids these corner cases, because ProNE does not depend on abstracts, though ProNE runs into different corner cases involving papers with few (if any) links in the citation graph.

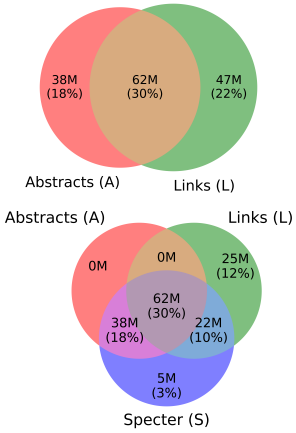


Figure 7: $|A \cap L| \approx 30\%$

Figure 7 (top) shows missing values are a common corner case in S2. Some papers have abstracts (A) and some have links in citation graph (L), but only 30% have both. Missing values are less common in benchmarks such as OGB and SciRepEval. Methods such as GNNs assume that most papers have both abstracts and links in the citation graph, but S2 is not like that.

We have run ProNE on L, and a number of models from HuggingFace on A. Semantic Scholar has more abstracts than A, but since those abstracts cannot be distributed, they graciously shared their Specter vectors with us. Figure 7 (bottom) compares those vectors (S) with A and L. Note that missing values are very common: $|S \cup L| \gg |S \cap L|$. Ensembles of abstract-based methods and link-based methods increase coverage from 30% to at least 70%. Nevertheless, coverage remains well below 100% because many papers are missing both abstracts and links.

Binning: Graph Partitioning

We construct a citation graph, $G = (V, E)$, based on data from S2. Each paper (vertex) has a primary document id. Document ids are often associated with a number of useful fields including: title, abstract, publication date, authors and references. Note that values can be missing (and incorrect). Publication dates are used to split G into a series of 100 subgraphs, G_t, the citation graph at time t. The 200M papers are sorted by publication date and assigned 100 bins (numbered 0-99) with about 2M papers per bin. We then construct 100 subgraphs: $G_{t_i} = (V_i, E_i)$ where $V_i \subset V$ and $E_i \subset E$. V_i contains the papers in bins 0 through bin i and E_i contains edges in E between papers in V_i.

Papers are removed from graphs if they are missing publication dates and/or edges. One might expect there to be 200M papers in V_{99}, but there are only 114M because nearly half of the papers are missing edges.

Figure 8 shows the growth of papers ($|V|$), citations ($|E|$) and $|E|/|V|$ over bins. There is considerable growth in all three panels in Figure 8 because of network effects.

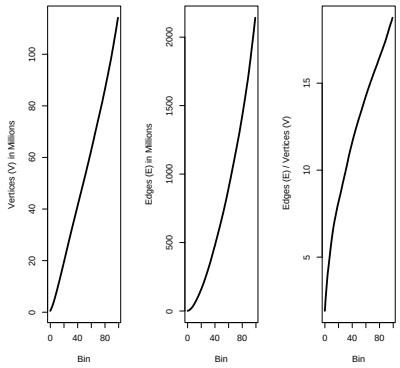


Figure 8: The 3 panels show growth in $|V|$, $|E|$ and $|E|/|V|$.

Training

We train 100 ProNE embeddings for each of the 100 G_t. Since ProNE requires large amounts of memory, we train the larger ProNE models on CPUs with TBs of memory. The bottleneck for ProNE is the initial prefactorization step, which computes an SVD on the input graph, using the sklearn function randomized_svd. Note that the input graph, G_t, is sparse ($|E_t| \ll |V_t|^2$), but SVD outputs a dense matrix, $U \in \mathbb{R}^{V \times K}$, where K is the number of hidden dimensions. We have been using $K = 280$ hidden dimensions, and $|E|/|V| < 20$ (see right panel of Figure 8). Since $K \gg |E_t|/|V_t|$, U consumes considerably more memory than G_t. Even though the output embedding is larger than the input graph, we find embeddings to be more convenient for certain computations such as approximate nearest neighbors (ANN).

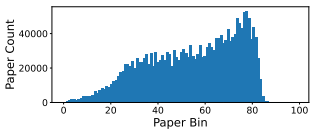


Figure 9: Specter training set peaks in 2017.

Time and space for prefactorization is shown in Figure 10. The two red lines in Figure 10 were fit to the observations (circles) using linear regression: hours $\sim$ poly(bin, 2) and GBs $\sim$ poly(bin, 2). The red regression lines make it clear that costs grow quickly with graph sizes. Experiments

such as Figure 1 show that larger graphs are better in terms of accuracy, but Figure 10 shows that larger graphs are more expensive in terms of space and time. Polynomial time and space may be unavoidable, though there may be ways to avoid SVD such as random projections [124], GEE [125, 126, 127] and incremental updates (to be discussed in subsection 2.11).

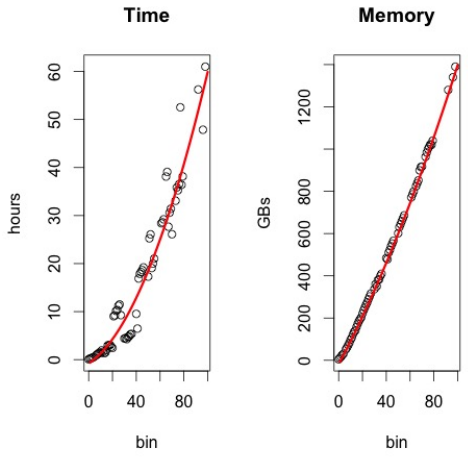


Figure 10: ProNE Prefactorization time and space (dominated by SVD)

GNNs merge node and edge features into a single embedding. As mentioned above, one of the motivations for separating these involves computational requirements. We use TBs of RAM to train ProNE on large graphs, whereas Specter uses GPUs and mini-batches in the standard way.

GPUs and TBs of RAM are about equally expensive. Depending on the workload, it may be preferable to invest more in one resource or the other. Spectral clustering (ProNE) is memory-bound, unlike deep networks, where computational cycles are the bottleneck.

Specter starts with a BERT-like model [128], and fine-tunes that model on a few million papers: $< query, pos, neg >$, where the *query* paper cites *pos* (in one or two hops), and *neg* are random negatives. Figure 9 shows the distribution of *query* by bin. Specter was trained on a single graph (not 100 graphs), a snapshot from a few years ago with a peak in 2017. Embeddings based on old data are likely to degrade over time. More seriously, because of network effects, small graphs (millions of edges) are not representative of large graphs (billions of edges).

As for time and space complexity, mini-batch training of Specter and GNNs can be done in $O(E)$ time and $O(1)$ space, which is clearly better than ProNE, where prefactorization requires considerably more memory. We suspect, though, that mini-batch approximations may be introducing larger errors for larger graphs. A work item for this proposal is to investigate this hypothesis.

Evaluation: Citation Prediction Task

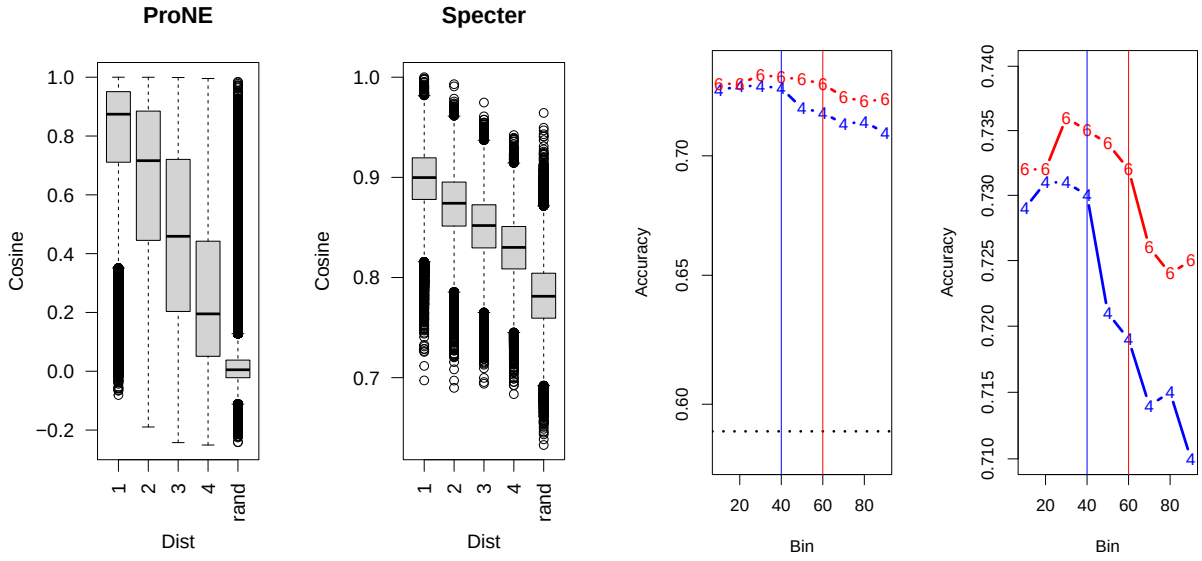
Figure 1 used the Citation Prediction Task described below to illustrate the unreasonable effectiveness of spectral clustering. That figure shows that ProNE (citations/edges) is better than Specter (abstracts/nodes) for larger t, with a crossover point at bin 41. Table 3 shows that this crossover point is considerably larger than ogbl-citation2, a link prediction task in OGB. Small graphs are not representative of larger graphs because of network effects.

Table 3: OGB is too small to see network effects

	Vertices	**Edges**
ogbl-citation2	3M	31M
crossover (bin 41)	42M	499M

Our Citation Prediction Task is simple: predict whether paper v_k and v_l cite one another i.e. $(v_k, v_l) \in E$. We define the distance $d(v_k, v_l)$ as the length of the shortest path between vertices v_k and v_l in the citation graph. Many link prediction tasks use random negatives, which are relatively easy to distinguish from true positives. To make the task harder, we sample relatively challenging negatives, where v_k and v_l are 2-4 hops from one another, i.e., $2 \leq d(v_k, v_l) \leq 4$.

Thus, the Citation Prediction task is to distinguish 1-hop (positive) from 2-4 hops (negative). Figure 11a shows cosines decline with distance (hops) for both ProNE and Specter.



(a) Cosines decline with distance.

(b) Accuracy degrades with h.

Figure 11: : Citation Prediction Task

The last bar, labeled *rand*, are random controls. These controls are not used in the citation prediction task, but they were added to Figure 11a to demonstrate that the negatives (2-4 hops) are more challenging than random negatives used in many link prediction experiments. These random controls also show that ProNE embeddings are centered so ProNE cosines are near zero for unrelated papers. Specter cosines have less dynamic range, and they are not centered at zero.

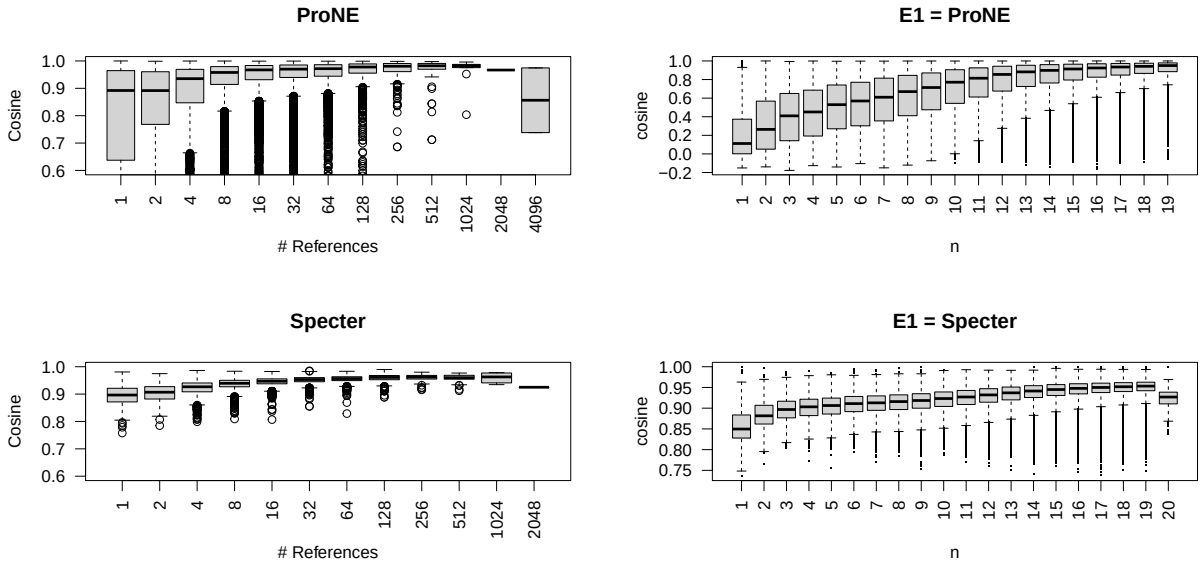
By construction, the Citation Prediction task is a binary classification task, with 28.9% positive labels (1-hop). Figure 11b shows accuracy_scores (from sklearn). The two panels of Figure 11b show the same accuracy scores, but on different y-axes. The panel on the left shows the two lines of interest (red and blue) are well above a random baseline (dashed line). This baseline at 59% is computed from the prior of 28.9%.

Figure 11b is like Figure 2, though Figure 2 was trained for $t = 50$ bins (denoted with the character 5), whereas Figure 11b was trained with $t = 40$ bins and $t = 60$ bins (denoted with the characters 4 and 6, respectively). Accuracy is reported for bins at $t + h$, where h (forecasting horizon) is an offset from the blue vertical line at $t = 40$ and the red vertical line at $t = 60$.

In general, accuracy increases with t and decreases with h:

- Increase with t (training data): The red line is consistently above the blue line because it is better to train on more (60) bins than fewer (40) bins.
- Decrease with h (forecasting horizon): Both lines decline after passing their respective vertical line at bin t, because forecasting into the future is harder (positive h) than predicting the past (negative h). Actually, the decline starts slightly before the vertical line (slightly negative h) because it takes time for recently published papers to be cited.

We conclude from these experiments that recommendation systems should be trained on as much citation data as possible, and they should be kept up to date. The degradation with h (forecasting horizon) can be interpreted as a penalty for failing to update often enough.



(a) The centroid approximation

(b) Better together approximation

Figure 12: Two approximations for imputing missing values.

Imputing Missing Values

A common criticism of spectral clustering methods such as ProNE is generalization beyond the training data. Such models are called *transductive*, meaning they generate vectors for the papers in the training set, but not for other papers such as submissions that have not been published yet. A similar criticism applies to Specter for papers with missing abstracts. The next two sections will introduce two methods for imputing missing vectors.

1. The centroid approximation: infer a missing vector from its references
2. The better-together approximation: infer a missing vector from papers that are nearby in another embedding

It is important to develop methods to impute missing vectors since missing vectors are common in ProNE, Specter and other embeddings.

The Centroid Approximation

The centroid approximation infers a vector from its references:

$$\widehat{vec}_{cen}(v_i, E) \approx \sum_{v_j \in refs(v_i)} vec(v_j, E)$$

where $vec(v, E)$ is the vector for paper v in embedding E and $refs(v)$ is a set of references in paper v. Figure 12a shows uses cosines to compare $vec(v_i)$ with the centroid approximation $\widehat{vec}(v_i, E)$ as a function of $|refs(v_i)|$. The centroid approximation is intended to impute missing values, but in Figure 12a, we compare approximations, $\widehat{vec}(v, E)$, with observations, $vec(v, E)$, when the observations are available (and not missing). Figure 12a shows the centroid approximation is reasonably

effective for both ProNE and Specter, though the approximation improves with $|refs(v_i)|$ (unless $|refs(v_i)|$ is extremely large).

There are many practical use cases for imputing missing values with the centroid approximation. Consider a use case where a paper is submitted for review. Who should review the submission? By construction, most unpublished submissions are not in S2, but the centroid approximation can be used to estimate vectors for submissions that cite papers in S2.

The Better-Together (BT) Approximation

The better together approximation infers vectors in one embedding, $vec(v, E_1)$, based on vectors in another embedding, $vec(v, E_2)$, where $vec(v, E)$ denotes the vector for paper v in embedding E. Figure 12b uses two embeddings, Specter and ProNE. In the top panel, E_1 is ProNE and E_2 is Specter. In the bottom panel, E_1 is Specter and E_2 is ProNE. The bt approximation is:

$$\widehat{vec}_{bt}(v_i, E_1) \approx \sum_{v_j \in near_k(v_i, E_2) \ \& \ v_j \in E_1} vec(v_j, E_1)$$

where $near_k(v, E)$ uses approximate nearest neighbors to find $k = 20$ papers near v in Embedding, E. Figure 12b show that $cos(v_i, \widehat{vec}_{bt}(v_i, E_1))$ improves with n, the number of vectors in the summation.

Incremental Updates And Time Invariance

As mentioned above, embeddings based on edges (ProNE) improve with time as papers receive more and more citations, whereas embeddings based on nodes (Specter [14]) are time invariant because abstracts do not change after publication. Since ProNE embeddings degrade with h, it is important to update ProNE embeddings frequently. Thus far, we have been computing ProNE embeddings from scratch, though a work item for this proposal is to come up with approximations to ProNE that support incremental updates. Since time and space complexity for ProNE grows super-linearly with graph size, we expect incremental updates to introduce approximation errors. We will develop and test approximations to ProNE that support incremental updates. A number of candidate approximations are: the centroid approximation, the better together approximation, random projections citeli2006very, and GEE (Graph Encoder Embedding) [125, 126, 127],

Sampling Embeddings And The Common Ground Approximation

The web site in Figure 3 provides a number of recommendation methods including an API from S2. The S2 API is limited to papers in computer science, about 10% of their collection, because the API is based on an approximate nearest neighbor method, FAISS [120], which is too expensive to index 100% of S2. This section will describe a sampling method that makes it possible to index large vector databases using well understood inverted files.

We assume we have a vector database, P, that is too large to fit in physical memory. To downsample P, we create a random sample, $S \subset P$, where $P \in \mathbb{R}^{n \times d}$ and $S \in \mathbb{R}^{n_s \times d}$. n is the number of rows in the large database, P, and n_s is the number of rows in the more manageable sample, S.

For each row, $p \in P$, we find k approximate nearest neighbors of p in S. We will refer to the set of k neighbors of p in S as a neighborhood, $neighborhood(p)$. The key observation is that neighborhoods can be used to approximate cosine similarity. That is, if two rows have a large cosine, then they are likely to share many of the same neighbors in S. We will refer to

$|neighborhood(p_i) \cap neighborhood(p_j)|$ as *common ground*, $cg(p_i, p_j)$. Thus, the observation above can be expressed as the common ground approximation: $cos(p_i, p_j) \sim cg(p_i, p_j)$.

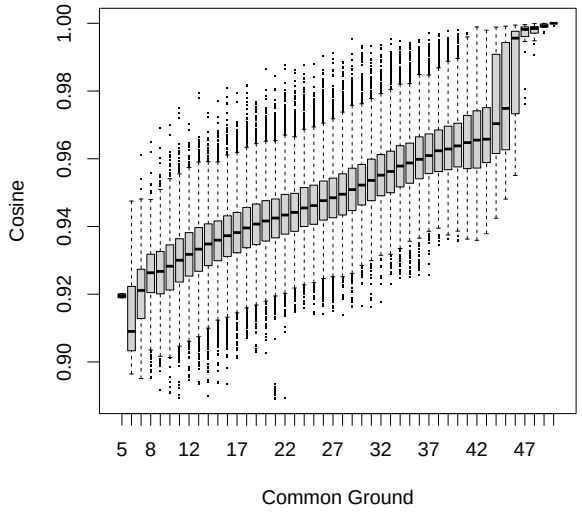


Figure 13: Common Ground Approximation

The intuition behind the common ground approximation is: "You shall know a word by the company it keeps" [129]. That is, it should be possible to infer p from its neighborhood, especially if k and n_s are large enough. That is, if we give you enough nearby landmarks, then you should be able to figure out where you are. Moreover, if we tell you that two vectors are near many of the same landmarks, then it is very likely that the two vectors are near one another.

Figure 13 provides some empirical evidence to support this intuition. The boxplots are computed over a set of 6894 SIGIR queries. For each query, q, we retrieve 19 candidates, c. For each of the 19 candidates, c, and each of the 6894 queries, q, we computed $cos(q, c)$ and $cg(q, c)$. Figure 13 shows there is a strong relationship between cg (common ground) and cos over these 19×6894 observations.

This assumption can be used to compute approximate nearest neighbors. At indexing time, we compute $neighborhood(v)$, the k nearest landmarks to paper v for all $v \in P$. For all those landmarks, l, we compute $postings(l)$, a list of papers near l.

At query time, we are given a query q. We use an off-the-shelf ANN method such as FAISS to find $neighborhood(q)$, the k nearest landmarks to q. For each of these k landmarks, we fetch $postings(l)$ from external memory with k disk seeks. We then compute common grounds by counting duplicates in these postings.

To Be Continued...

Results From Prior NSF Support

Neither PI Church nor Co-PI Chandrasekar have received NSF funding in the past five years, since they have been in industry. Church was a Co-PI for NSF #1040114 when he was at JHU in 2010.

Deliverables

1. Create and disseminate tools that make it easier for researchers to make better use of the scientific literature.
2. Resources: embeddings, models and code. We will share large files such as embeddings of 10^8 papers on Globus [130]. Code will be posted on GitHub. Models and benchmarks will be posted on HuggingFace. Some code and embeddings have been posted already on [7].
3. Timeliness: Support incremental updates to resources.
4. Use cases: (1) recommendations, (2) finding experts, (3) routing submissions to reviewers and (4) summarization.
5. Evaluation: Better benchmarks, as well as better numbers on established benchmarks.
6. Establish that it is important to experiment with large (and growing) graphs to appreciate network effects (Metcalfe's Law). Also, establish the value of multiple perspectives in theory and practice. We will generalize results from linear algebra to deep nets, as mentioned above. From a practical perspective, combinations of text and links create opportunities to deal with missing values and bad data.

Schedule/GANTT Chart

Figure 14 presents a GANTT chart for the work items in the proposal.

Accomplishments Thus Far, Risks And Contingencies

PI Church led a team on related topics at the 6-week 2023 Jelinek Summer Workshop on Speech and Language Technology [131]; some code and resources have already been posted on [7]. The deliverables in the previous section go well beyond what has already been done, and what will be accomplished during the Jelinek Summer Workshop. The long list of deliverables is a stretch goal. There is a risk that it may become necessary to make adjustments.

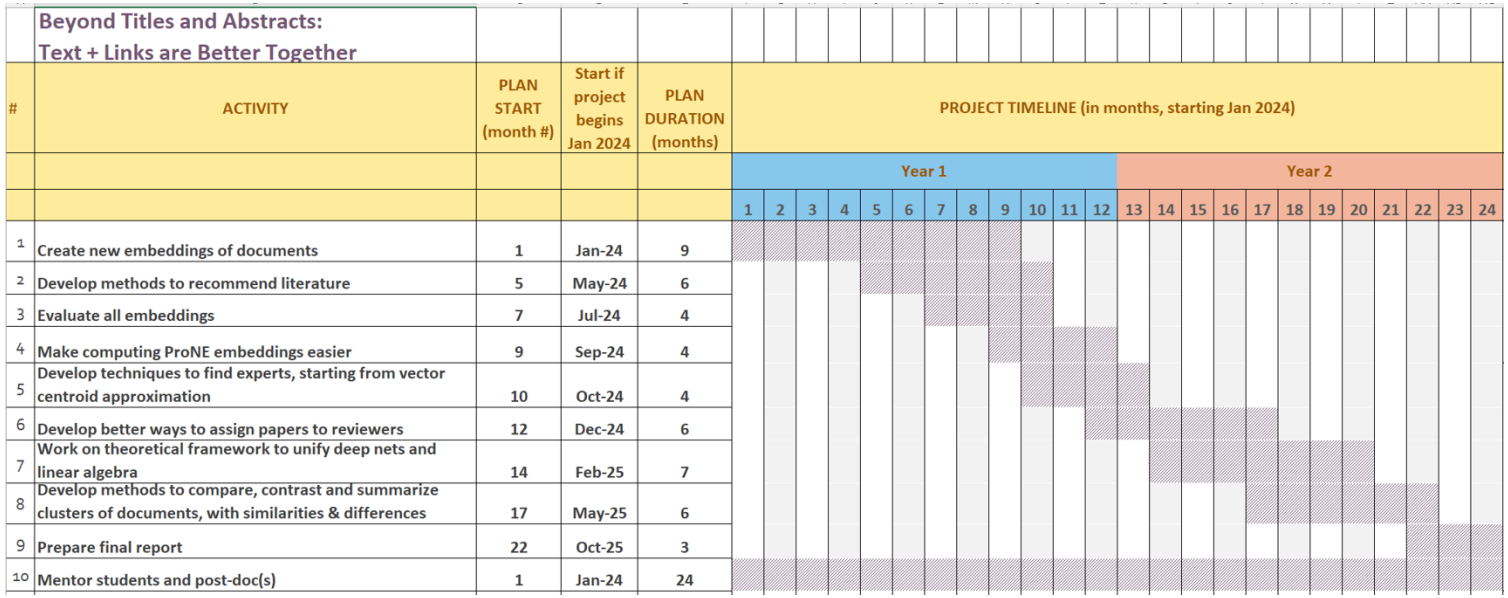
Broader Impacts

There are many opportunities for applications of the deliverables in academia and industry, including web search, product recommendations and traffic analysis (data mining on telephone call detail and internet packet headers). The proposed work will make it it easier for the community to make better use of the scientific literature.

Improving access to expertise will produce significant benefits to people in resource-poor countries and environments. It will be easier for researchers in diverse settings to contribute to the literature in meaningful ways. Science will advance in more productive ways when the right people can more easily see how they can contribute to critical projects.

There are opportunities to trial the tools that we will develop with students at Northeastern University. Northeastern has large numbers of excellent students. Many are the first in their family to attend college. Many are members of protected classes.

Northeastern has many campuses in many locations including Mills College in Oakland, California and the Roux Institute in Portland, Maine. Both locations offer opportunities to make the scientific literature more accessible to more diverse communities.



Beyond Titles and Abstracts: Text + Links are Better Together

#	ACTIVITY	PLAN START (month #)	Start if project begins Jan 2024	PLAN DURATION (months)
1	Create new embeddings of documents	1	Jan-24	9
2	Develop methods to recommend literature	5	May-24	6
3	Evaluate all embeddings	7	Jul-24	4
4	Make computing ProNE embeddings easier	9	Sep-24	4
5	Develop techniques to find experts, starting from vector centroid approximation	10	Oct-24	4
6	Develop better ways to assign papers to reviewers	12	Dec-24	6
7	Work on theoretical framework to unify deep nets and linear algebra	14	Feb-25	7
8	Develop methods to compare, contrast and summarize clusters of documents, with similarities & differences	17	May-25	6
9	Prepare final report	22	Oct-25	3
10	Mentor students and post-doc(s)	1	Jan-24	24

PROJECT TIMELINE (in months, starting Jan 2024)

Year 1 | Year 2

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

Figure 14: GANTT Chart for proposal

It would be interesting to see how effective the proposed tools are with users from more diverse backgrounds. We have already begun experimenting with these tools with users from a variety of research fields. So far, we have discovered opportunities for improving coverage in fields that go beyond STEM such as journalism and the law.

PI Qualifications

Church and Chandrasekar have considerable experience in computational linguistics and related fields, as indicated by their h-indexes on Google Scholar [132, 133]. They both worked at Microsoft on Bing, and have considerable experience with behavioral signals such as web search logs [134, 135]. They have also worked with large databases of telephone call detail at AT&T Bell Labs and elsewhere [136]. Call detail is often used for traffic analysis in intelligence applications where much can be inferred from the graph of who is communicating with whom. Data mining on call detail is similar to what we propose to do with citation graphs.

Church was an early advocate in 1990s of the revival of empirical methods and corpus-based lexicography. His most cited paper introduced what is now known as PMI (point-wise mutual information) [137]. Levy and Goldberg [138] established a connection between PMI and Word2Vec. There are clear connections between [138] and more recent methods such as deep nets (BERT) [122]. Qiu et al. [139] use methods like [138] to unify GNNs and spectral clustering. Church was also a founder of EMNLP and served as president from 1993 to 2011 of the ACL special interest group (SIGDAT) that runs EMNLP. Church was president of the ACL in 2012. He was an AT&T Fellow in 2001, ACL Fellow in 2015, Baidu Fellow in 2018 and ACM Fellow in 2023. Baidu is the largest web search company in China.

References Cited

[1] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec, "Open graph benchmark: Datasets for machine learning on graphs," *Advances in neural information processing systems*, vol. 33, pp. 22 118–22 133, 2020.

[2] W. Hu, M. Fey, H. Ren, M. Nakata, Y. Dong, and J. Leskovec, "OGB-LSC: A large-scale challenge for machine learning on graphs," *arXiv preprint arXiv:2103.09430*, 2021.

[3] A. Singh, M. D'Arcy, A. Cohan, D. Downey, and S. Feldman, "SciRepEval: A multi-format benchmark for scientific document representations," *ArXiv*, vol. abs/2211.13308, 2022.

[4] R. Van Noorden, "Global scientific output doubles every nine years," *Nature news blog*, 2014.

[5] L. Bornmann, R. Haunschild, and R. Mutz, "Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases," *Humanities and Social Sciences Communications*, vol. 8, no. 1, pp. 1–15, 2021.

[6] https://jsalt2023.univ-lemans.fr/en/better-together-text-context.html, 2023.

[7] "Better together team github," https://github.com/kwchurch/JSALT_Better_Together, 2023.

[8] A. D. Wade, "The semantic scholar academic graph (S2AG)," *Companion Proceedings of the Web Conference 2022*, 2022.

[9] R. M. Kinney, C. Anastasiades, R. Authur, I. Beltagy, J. Bragg, A. Buraczynski, I. Cachola, S. Candra, Y. Chandrasekhar, A. Cohan, M. Crawford, D. Downey, J. Dunkelberger, O. Etzioni, R. Evans, S. Feldman, J. Gorney, D. W. Graham, F. Hu, R. Huff, D. King, S. Kohlmeier, B. Kuehl, M. Langan, D. Lin, H. Liu, K. Lo, J. Lochner, K. MacMillan, T. Murray, C. Newell, S. R. Rao, S. Rohatgi, P. L. Sayre, Z. Shen, A. Singh, L. Soldaini, S. Subramanian, A. Tanaka, A. D. Wade, L. M. Wagner, L. L. Wang, C. Wilhelm, C. Wu, J. Yang, A. Zamarron, M. van Zuylen, and D. S. Weld, "The semantic scholar open data platform," *ArXiv*, vol. abs/2301.10140, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256194545

[10] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *Proceedings of the twelfth international conference on Information and knowledge management*, 2003, pp. 556–559.

[11] C. Wang, V. Satuluri, and S. Parthasarathy, "Local probabilistic models for link prediction," *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pp. 322–331, 2007. [Online]. Available: https://api.semanticscholar.org/CorpusID:8835297

[12] T. Tylenda, R. Angelova, and S. J. Bedathur, "Towards time-aware link prediction in evolving social networks," in *SNA-KDD '09*, 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:14033079

[13] J. Zhang, Y. Dong, Y. Wang, J. Tang, and M. Ding, "ProNE: Fast and scalable network representation learning." in *IJCAI*, vol. 19, 2019, pp. 4278–4284.

[14] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. Weld, "SPECTER: Document-level representation learning using citation-informed transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 2270–2282. [Online]. Available: https://aclanthology.org/2020.acl-main.207

[15] F. Scarselli, M. Gori, A. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, pp. 61–80, 2009.

[16] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, vol. 2, 2005, pp. 729–734 vol. 2.

[17] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.

[18] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, pp. 4–24, 2019.

[19] anonymous, "Graph neural networks," https://snap-stanford.github.io/cs224w-notes/machine-learning-with-networks/graph-neural-networks, 2020.

[20] https://www.semanticscholar.org/api-gallery/better-together, 2023.

[21] M. Ostendorff, N. Rethmeier, I. Augenstein, B. Gipp, and G. Rehm, "Neighborhood contrastive learning for scientific document representations with citation embeddings," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 11 670–11 688. [Online]. Available: https://aclanthology.org/2022.emnlp-main.802

[22] https://api.semanticscholar.org/recommendations/v1/papers/forpaper/21321bad706a9f9dbb502588b0bb393cf15fa052?from=all-cs&fields=title,externalIds,citationCount.

[23] https://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&db=pubmed&id=10491450&cmd=neighbor_score.

[24] https://huggingface.co/.

[25] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: a literature survey," *International Journal on Digital Libraries*, vol. 17, pp. 305 – 338, 2015.

[26] C. Nascimento, A. H. F. Laender, A. D. Silva, and M. A. Gonçalves, "A source independent framework for research paper recommendation," pp. 297–306, 2011.

[27] K. Sugiyama and M.-Y. Kan, "Scholarly paper recommendation via user's recent research interests," pp. 29–38, 2010.

[28] J. Sun, J. Ma, Z. Liu, and Y. Miao, "Leveraging content and connections for scientific article recommendation in social computing contexts," *Comput. J.*, vol. 57, pp. 1331–1342, 2014.

[29] C.-L. Huang, "Bayesian recommender system for social information sharing: Incorporating tag-based personalized interest and social relationships," *Intell. Data Anal.*, vol. 23, pp. 623–639, 2019.

[30] X. Bai, B. B. Cambazoglu, F. Gullo, A. Mantrach, and F. Silvestri, "Exploiting search history of users for news personalization," *Inf. Sci.*, vol. 385, pp. 125–137, 2017.

[31] B. Maake, S. Ojo, and T. Zuva, "Information processing in research paper recommender system classes," *Advances in Library and Information Science*, 2019.

[32] J. Beel and S. Langer, "A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems," pp. 153–168, 2015.

[33] J. Beel, S. Langer, M. Genzmehr, B. Gipp, C. Breitinger, and A. Nürnberger, "Research paper recommender system evaluation: a quantitative literature survey," in *RepSys '13*, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:4411601

[34] J. Beel, M. Genzmehr, S. Langer, A. Nürnberger, and B. Gipp, "A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation," pp. 7–14, 2013.

[35] D. M. S. C. V. K. Saini and J. R. Singh, "Recommendation system," *International Journal for Modern Trends in Science and Technology*, 2021.

[36] D. Mimno and A. McCallum, "Expertise modeling for matching papers with reviewers," in *Knowledge Discovery and Data Mining*, 2007.

[37] L. M. de Campos, J. M. Fernández-Luna, J. Huete, and L. Redondo-Expósito, "Lda-based term profiles for expert finding in a political setting," *Journal of Intelligent Information Systems*, vol. 56, pp. 529 – 559, 2021.

[38] J. Jin, Q. Geng, Q. Zhao, and L. Zhang, "Integrating the trend of research interest for reviewer assignment," *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017.

[39] H. Yin, B. Cui, H. Lu, Y. Huang, and J. Yao, "A unified model for stable and temporal topic detection from social media data," *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pp. 661–672, 2013.

[40] Y. Zhang, D. Ji, Y. Su, and P. Hu, "Topic analysis for online reviews with an author-experience-object-topic model," pp. 303–314, 2011.

[41] A. Daud, "Using time topic modeling for semantics-based dynamic research interest finding," *Knowl. Based Syst.*, vol. 26, pp. 154–163, 2012.

[42] Y. Zhang, Y.-J. Shen, X. Chen, B. Jin, and J. Han, ""why should i review this paper?" unifying semantic, topic, and citation factors for paper-reviewer matching," *ArXiv*, vol. abs/2310.14483, 2023.

[43] O. Anjum, A. V. Kamatar, T. Liang, J. Xiong, and W. mei W. Hwu, "Submission-aware reviewer profiling for reviewer recommender system," *ArXiv*, vol. abs/2211.04194, 2022.

[44] K. Balog and M. de Rijke, "Associating people and documents," pp. 296–308, 2008.

[45] V. Mangaravite and R. L. T. Santos, "On information-theoretic document-person associations for expert search in academia," *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016.

[46] O. Anjum, H. Gong, S. Bhat, W. mei W. Hwu, and J. Xiong, "Pare: A paper-reviewer matching approach using a common topic space," *ArXiv*, vol. abs/1909.11258, 2019.

[47] Y. Tu, N. Johri, D. Roth, and J. Hockenmaier, "Citation author topic model in expert search," pp. 1265–1273, 2010.

[48] M. Färber and A. Jatowt, "Citation recommendation: approaches and datasets," *International Journal on Digital Libraries*, vol. 21, pp. 375 – 405, 2020.

[49] A. Khadka and P. Knoth, "Using citation-context to reduce topic drifting on pure citation-based recommendation," *Proceedings of the 12th ACM Conference on Recommender Systems*, 2018.

[50] H. Jia and E. Saule, "Local is good: A fast citation recommendation approach," pp. 758–764, 2018.

[51] H. Chen, "A new citation recommendation strategy based on term functions in related studies section," *Journal of Data and Information Science*, vol. 6, pp. 75 – 98, 2020.

[52] R. Etemadi, M. Zihayat, K. Feng, J. Adelman, and E. Bagheri, "Collaborative experts discovery in social coding platforms," *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021.

[53] Z. Ali, I. Ullah, A. Khan, A. U. Jan, and K. Muhammad, "An overview and evaluation of citation recommendation models," *Scientometrics*, vol. 126, pp. 4083 – 4119, 2021.

[54] H. Jia and E. Saule, *An Analysis of Citation Recommender Systems: Beyond the Obvious*, 2017.

[55] C. Jebari, E. Herrera-Viedma, and M. Cobo, "Context-aware citation recommendation of scientific papers: comparative study, gaps and trends," *Scientometrics*, vol. 128, pp. 4243 – 4268, 2023.

[56] Z. Medic and J. Šnajder, "A survey of citation recommendation tasks and methods," *J. Comput. Inf. Technol.*, vol. 28, pp. 183–205, 2021.

[57] S. Ma, C. Zhang, and X. Liu, "A review of citation recommendation: from textual content to enriched context," *Scientometrics*, vol. 122, pp. 1445–1472, 2020.

[58] A. Razdaibiedina and A. Brechalov, "Miread: Simple method for learning high-quality representations from scientific documents," *ArXiv*, vol. abs/2305.04177, 2023.

[59] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld, "Specter: Document-level representation learning using citation-informed transformers," *ArXiv*, vol. abs/2004.07180, 2020.

[60] R. Seoh, H.-S. Chang, and A. McCallum, "Encoding multi-domain scientific papers by ensembling multiple cls tokens," *ArXiv*, vol. abs/2309.04333, 2023.

[61] M. Parisot and J. Zavrel, "Multi-objective representation learning for scientific document retrieval," pp. 80–88, 2022.

[62] R. Xu, Y. Yu, J. Ho, and C. Yang, *Weakly-Supervised Scientific Document Classification via Retrieval-Augmented Multi-Stage Training*, 2023.

[63] I. Stelmakh, J. Wieting, G. Neubig, and N. B. Shah, "A gold standard dataset for the reviewer assignment problem," *ArXiv*, vol. abs/2303.16750, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257805004

[64] C. K. Kreutz and R. Schenkel, "Revaside: Evaluation of assignments of suitable reviewer sets," *J. Data Intell.*, vol. 4, pp. 101–133, 2023.

[65] ——, *RevASIDE: Assignment of Suitable Reviewer Sets for Publications from Fixed Candidate Pools*, 2021.

[66] A. N. Medakene, K. Bouanane, and M. A. Eddoud, "A new approach for computing the matching degree in the paper-to-reviewer assignment problem," in *2019 International Conference on Theoretical and Applicative Aspects of Computer Science (ICTAACS)*, vol. 1, 2019, pp. 1–8.

[67] K. Leyton-Brown, Mausam, Y. Nandwani, H. Zarkoob, C. Cameron, N. Newman, and D. Raghu, "Matching papers and reviewers at large conferences," *ArXiv*, vol. abs/2202.12273, 2022.

[68] B. Chen, J. Zhang, F. Zhang, T. Han, Y. Cheng, X. Li, Y. Dong, and J. Tang, *Web-Scale Academic Name Disambiguation: The WhoIsWho Benchmark, Leaderboard, and Toolkit*, 2023.

[69] Y. Zhang, F. Zhang, P. Yao, and J. Tang, "Name disambiguation in AMiner: Clustering, maintenance, and human in the loop." *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018.

[70] S. Subramanian, D. King, D. Downey, and S. Feldman, "S2AND: A benchmark and evaluation system for author name disambiguation," *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 170–179, 2021.

[71] J. Zhang and J. Tang, "Name disambiguation in AMiner," *Science China Information Sciences*, vol. 64, 2020.

[72] Z. Boukhers and N. Bahubali, "Whois? deep author name disambiguation using bibliographic data," pp. 201–215, 2022.

[73] L. Zhang, W. Lu, and J. Yang, "Lagos-and: A large gold standard dataset for scholarly author name disambiguation," *Journal of the Association for Information Science and Technology*, vol. 74, pp. 168 – 185, 2021.

[74] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *ArXiv*, vol. abs/2005.11401, 2020.

[75] M. Lewis, M. Ghazvininejad, G. Ghosh, A. Aghajanyan, S. I. Wang, and L. Zettlemoyer, "Pre-training via paraphrasing," *ArXiv*, vol. abs/2006.15020, 2020.

[76] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "Realm: Retrieval-augmented language model pre-training," *ArXiv*, vol. abs/2002.08909, 2020.

[77] A. Lazaridou, E. Gribovskaya, W. Stokowiec, and N. Grigorev, "Internet-augmented language models through few-shot prompting for open-domain question answering," *ArXiv*, vol. abs/2203.05115, 2022.

[78] C. Alberti, D. Andor, E. Pitler, J. Devlin, and M. Collins, "Synthetic qa corpora generation with roundtrip consistency," *ArXiv*, vol. abs/1906.05416, 2019.

[79] Y. Tay, M. Dehghani, V. Q. Tran, X. García, D. Bahri, T. Schuster, H. Zheng, N. Houlsby, and D. Metzler, "Unifying language learning paradigms," *ArXiv*, vol. abs/2205.05131, 2022.

[80] C.-H. Tan, J.-C. Gu, C. Tao, Z.-H. Ling, C. Xu, H. Hu, X. Geng, and D. Jiang, "Tegtok: Augmenting text generation via task-specific and open-world knowledge," *ArXiv*, vol. abs/2203.08517, 2022.

[81] W. Yu, "Retrieval-augmented generation across heterogeneous knowledge," pp. 52–58, 2022.

[82] A. Asai, M. Gardner, and H. Hajishirzi, "Evidentiality-guided generation for knowledge-intensive nlp tasks," pp. 2226–2243, 2021.

[83] Z. Sun, X. Wang, Y. Tay, Y. Yang, and D. Zhou, "Recitation-augmented language models," *ArXiv*, vol. abs/2210.01296, 2022.

[84] H. Ivison and M. E. Peters, "Hyperdecoders: Instance-specific decoders for multi-task nlp," *ArXiv*, vol. abs/2203.08304, 2022.

[85] P. Cui, X. Wang, J. Pei, and W. Zhu, "A survey on network embedding," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, pp. 833–852, 2017.

[86] A. Li, S. P. Cornelius, Y.-Y. Liu, L. Wang, and A.-L. Barabási, "The fundamental advantages of temporal networks," *Science*, vol. 358, no. 6366, pp. 1042–1046, 2017.

[87] P. Resnick and H. R. Varian, "Recommender systems," *Communications of the ACM*, vol. 40, no. 3, pp. 56–58, 1997.

[88] T. Strohman, W. B. Croft, and D. D. Jensen, "Recommending citations for academic papers," in *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007.

[89] S. Bethard and D. Jurafsky, "Who should i cite: learning literature search models from citation behavior," *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010.

[90] Q. He, D. Kifer, J. Pei, P. Mitra, and C. L. Giles, "Citation recommendation without author supervision," in *Web Search and Data Mining*, 2011.

[91] C. Caragea, A. Silvescu, P. Mitra, and C. L. Giles, "Can't see the forest for the trees? a citation recommendation system," in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, 2013, pp. 111–114.

[92] K. Sugiyama and M.-Y. Kan, "Exploiting potential citation papers in scholarly paper recommendation," pp. 153–162, 2013.

[93] D. Roy, "An improved test collection and baselines for bibliographic citation recommendation," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 2271–2274.

[94] A. Singh, M. D. . Arcy, A. Cohan, D. Downey, and S. Feldman, "Scirepeval: A multi-format benchmark for scientific document representations," 11 2022. [Online]. Available: https://arxiv.org/abs/2211.13308v3

[95] M. Yasunaga, J. Leskovec, and P. Liang, "LinkBERT: Pretraining language models with document links," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8003–8016. [Online]. Available: https://aclanthology.org/2022.acl-long.551

[96] M. Färber and A. Jatowt, "Citation recommendation: approaches and datasets," *International Journal on Digital Libraries*, vol. 21, pp. 375–405, 2020. [Online]. Available: https://doi.org/10.1007/s00799-020-00288-2

[97] Z. Ali, P. Kefalas, K. Muhammad, B. Ali, and M. Imran, "Deep learning in citation recommendation models survey," *Expert Syst. Appl.*, vol. 162, p. 113790, 2020.

[98] R. S. Pillai and L. Deepthi, "A survey on citation recommendation system," in *2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)*. IEEE, 2022, pp. 423–429.

[99] Y. Liang and L.-K. Lee, "A systematic review of citation recommendation over the past two decades," *International Journal on Semantic Web and Information Systems*, 2023.

[100] L. Steinert, "Beyond similarity and accuracy - a new take on automating scientific paper recommendations," 2017, pp. 1–161.

[101] S. Dumais and J. Nielsen, "Automating the assignment of submitted manuscripts to reviewers," pp. 233–244, 1992.

[102] D. Yarowsky and R. Florian, "Taking the load off the conference chairs-towards a digital paper-routing assistant," in *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999. [Online]. Available: https://aclanthology.org/W99-0627

[103] D. Mimno and A. McCallum, "Expertise modeling for matching papers with reviewers," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 500–509.

[104] "ACL reviewer matching code," https://github.com/acl-org/reviewer-paper-matching.

[105] "Conference peer review with the semantic scholar api," https://blog.allenai.org/conference-peer-review-with-the-semantic-scholar-api-24ab9fce2324.

[106] D. Yimam-Seid and A. Kobsa, "Expert-finding systems for organizations: Problem and domain analysis and the demoir approach," *Journal of Organizational Computing and Electronic Commerce*, vol. 13, no. 1, pp. 1–24, 2003.

[107] M. T. Maybury, "Expert finding systems," https://www.mitre.org/sites/default/files/pdf/06_1115.pdf, 2006.

[108] Y. Tu, N. Johri, D. Roth, and J. Hockenmaier, "Citation author topic model in expert search," in *Coling 2010: Posters*. Beijing, China: Coling 2010 Organizing Committee, Aug. 2010, pp. 1265–1273. [Online]. Available: https://aclanthology.org/C10-2145

[109] L. A. Streeter and K. E. Lochbaum, "Who knows: A system based on automatic representation of semantic structure," in *User-Orient@articlemetcalfe2013metcalfe, title=Metcalfe's law after 40 years of Ethernet, author=Metcalfe, Bob, journal=Computer, volume=46, number=12, pages=26–31, year=2013, publisher=IEEE ed Content-Based Text and Image Handling*, 1988, pp. 380–388.

[110] K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, and A. Kanakia, "Microsoft academic graph: When experts are not enough," *Quantitative Science Studies*, vol. 1, no. 1, pp. 396–413, 2020.

[111] B.-J. Hsu, I. Shen, D. Eide, A. Chen, and R. Rogahn, "Microsoft academic graph," https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/.

[112] "DOI foundation," https://www.doi.org/.

[113] "Pubmed," https://pubmed.ncbi.nlm.nih.gov/.

[114] "Pubmed central," https://www.ncbi.nlm.nih.gov/pmc/.

[115] "DBLP," https://dblp.org/.

[116] "arXiv," https://arxiv.org/.

[117] "ACL anthology," https://aclanthology.org/.

[118] https://api.semanticscholar.org/recommendations/v1/papers/forpaper/21321bad706a9f9dbb502588b0bb393cf15fa052?from=all-cs&fields=title,externalIds,citationCount, 2024.

[119] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, 1998, pp. 604–613.

[120] Facebook Research, "Faiss," https://github.com/facebookresearch/faiss, 2019.

[121] S. Jayaram Subramanya, F. Devvrit, H. V. Simhadri, R. Krishnawamy, and R. Kadekodi, "DiskANN: Fast accurate billion-point nearest neighbor search on a single node," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[122] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[123] "Nodevectors," https://github.com/VHRanger/nodevectors.

[124] P. Li, T. J. Hastie, and K. W. Church, "Very sparse random projections," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 287–296.

[125] C. Shen, Q. Wang, and C. E. Priebe, "One-hot graph encoder embedding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[126] C. Shen, J. Larson, H. Trinh, X. Qin, Y. Park, and C. E. Priebe, "Discovering communication pattern shifts in large-scale labeled networks using encoder embedding and vertex dynamics," *IEEE Transactions on Network Science and Engineering*, vol. 11, no. 2, pp. 2100–2109, 2024.

[127] C. Shen, C. E. Priebe, J. Larson, and H. Trinh, "Synergistic graph fusion via encoder embedding," *arXiv preprint arXiv:2303.18051*, 2023.

[128] I. Beltagy, K. Lo, and A. Cohan, "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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3615–3620. [Online]. Available: https://aclanthology.org/D19-1371

[129] J. R. Firth, "A synopsis of linguistic theory, 1930-1955," *Studies in linguistic analysis*, 1957.

[130] "Globus," https://www.globus.org/.

[131] https://jsalt2023.univ-lemans.fr/en/better-together-text-context.html, 2023.

[132] https://scholar.google.com/citations?user=E6aqGvYAAAAJ&hl=en.

[133] https://scholar.google.com/citations?user=zZhCPGkAAAAJ&hl=en.

[134] Q. Mei and K. Church, "Entropy of search logs: how hard is search? with personalization? with backoff?" in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 2008, pp. 45–54.

[135] A. S. Bacioiu, D. M. Sauntry, J. S. Boyle, L. C. W. Wong, P. F. Leonard, and R. Chandrasekar, "Method and apparatus for analysis and decomposition of classifier data anomalies," Sep. 16 2008, uS Patent 7,426,497.

[136] D. Belanger, K. Church, and A. Hume, "Virtual data warehousing, data publishing and call detail," in *Databases in Telecommunications: International Workshop, Co-located with VLDB-99, Edinburgh, Scotland, UK, September 6th, 1999. Proceedings 1.* Springer, 2000, pp. 106–117.

[137] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational Linguistics*, vol. 16, no. 1, pp. 22–29, 1990. [Online]. Available: https://www.aclweb.org/anthology/J90-1003

[138] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *NIPS*, 2014.

[139] J. Qiu, Y. Dong, H. Ma, J. Li, K. Wang, and J. Tang, "Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec," *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2017.

[140] https://research.northeastern.edu/about/institutes-centers/.

[141] Semantic-Scholar, "Semantic scholar academic graph api: Providing a reliable source of scholarly data for developers," https://www.semanticscholar.org/product/api, 2017.

[142] J. Priem, H. Piwowar, and R. Orr, "OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts," 2022. [Online]. Available: https://arxiv.org/abs/2205.01833

Bio Sketches

NSF BIOGRAPHICAL SKETCH

Provide the following information for the Senior personnel.
Follow this format for each person. **DO NOT EXCEED 3 PAGES.**

IDENTIFYING INFORMATION:

NAME: Church, Kenneth

POSITION TITLE: Professor of Practice

ORGANIZATION AND LOCATION: Computer Science, San Jose, CA, USA

Professional Preparation:

ORGANIZATION AND LOCATION	DEGREE (if applicable)	DATE RECEIVED	FIELD OF STUDY
MIT, Cambridge, MA, USA	PHD	05/1983	Computer Science
MIT, Cambridge, MA, USA	MS	05/1980	Computer Science
MIT, Cambridge, MA, USA	BS	05/1978	Computer Science

Appointments and Positions

2022 - present	Professor of Practice, Computer Science, San Jose, CA, USA
2018 - 2022	Fellow, Baidu, Sunnyvale, CA, United States
2011 - 2018	Research Staff Member, IBM, Yorktown Heights, NY, USA
2009 - 2011	Chief Scientist, Johns Hopkins University, Human Language Technology Center of Excellence, Baltimore, MD, USA
2003 - 2009	Research Staff Member, Microsoft, Redmond, WA, USA
1983 - 2003	Department Head, AT&T Bell Laboratories, Murray Hill, NJ, USA

Products

Products Most Closely Related to the Proposed Project

1. Church K, Hanks P. Word association norms, mutual information, and lexicography. Computational linguistics. 1990; 16(1):22-29.
2. Li P, Hastie T, Church K. Very sparse random projections. ; c2006.
3. Church K. Last words: Reviewing the reviewers. Computational Linguistics. 2005; 31(4):575-578.
4. Church K, Kordoni V. Emerging trends: Sota-chasing. Natural Language Engineering. 2022; 28(2):249-269.
5. Church K. Emerging trends: Deep nets thrive on scale. Natural Language Engineering. 2022; 28(5):673-682.

Other Significant Products, Whether or Not Related to the Proposed Project

1. Mei Q, Zhou D, Church K. Query suggestion using hitting time. ; c2008.
2. Church K. One term or two?. ; c1995.
3. Mei Q, Church K. Entropy of search logs: how hard is search? with personalization? with

backoff?. ; c2008.
4. Li P, Church K. Using sketches to estimate associations. ; c2005.
5. Church K. A pendulum swung too far. Linguistic Issues in Language Technology. 2011; 6.

Synergistic Activities

1. President ACL (2012)
2. A founder of EMNLP (and president of SIGDAT from 1993 to 2011)
3. Author of Church & Hanks (1990), paper that introduced what is now known as PMI (pointwise mutual information). PMI, Word2vec and BERT are all based on collocations in lexicography, following Firth's famous quote, "You shall know a word by the company it keeps."
4. Active in many NLP conferences. Area Chair (AC) for ACL-2023 and Senior Area Chair (SAC) for IJCAI-2023.
5. Team leader for JSALT (2023 Jelinek Summer Workshop on Speech and Language Technology), an 8 week program in France.

Certification:

When the individual signs the certification on behalf of themselves, they are certifying that the information is current, accurate, and complete. This includes, but is not limited to, information related to domestic and foreign appointments and positions. Misrepresentations and/or omissions may be subject to prosecution and liability pursuant to, but not limited to, 18 U.S.C. §§ 287, 1001, 1031 and 31 U.S.C. §§ 3729-3733 and 3802.

Certified by Church, Kenneth in SciENcv on 2023-04-14 17:36:00

NSF BIOGRAPHICAL SKETCH

Provide the following information for the Senior personnel.
Follow this format for each person. **DO NOT EXCEED 3 PAGES.**

IDENTIFYING INFORMATION:

NAME: Chandrasekar, Raman

ORCID: 0000-0002-2755-0745

POSITION TITLE: Senior Principal Research Scientist

ORGANIZATION AND LOCATION: Institute for Experiential AI, Northeastern University, Seattle, WA, United States

Professional Preparation:

ORGANIZATION AND LOCATION	DEGREE (if applicable)	DATE RECEIVED	FIELD OF STUDY
Tata Institute of Fundamental Research/Univ of Bombay, Mumbai, Maharashtra, India	PHD	09/1994	Computer Science
Indian Institute of Technology Delhi, New Delhi, New Delhi, India	MS	05/1982	Physics

Appointments and Positions

2022 - present	Senior Principal Research Scientist, Institute for Experiential AI, Northeastern University, Seattle, WA, United States
2015 - present	Co-owner, KJ Consulting, Seattle, WA, United States
2019 - 2022	Clinical Professor, Northeastern University, Seattle, WA, United States
2017 - 2019	Part-time Lecturer, Northeastern University, Seattle, WA, United States
2012 - 2015	Senior Software Engineering Manager, ProQuest, Seatle, WA, United States
2010 - 2012	Applied Scientist, Evri.com, Seattle, WA, United States
1998 - 2010	Researcher, Microsoft/Microsoft Research, Redmond, WA, United States
1982 - 1998	(Various, culminating in) Research Scientist, National Center for Software Technology/Tata Institute of Fundamental Research, Mumbai, Not Applicable, N/A, India

Products

Products Most Closely Related to the Proposed Project

1. Sondhi P, Chandrasekar R. Domain-specific entity and relationship extraction from query logs. Proceedings of the American Society for Information Science and Technology. 2010 November; 47(1):1-2. Available from: https://onlinelibrary.wiley.com/doi/10.1002/meet.14504701458 DOI: 10.1002/meet.14504701458
2. Vydiswaran V, van den Eijkhof J, Chandrasekar R, Paradiso A, St. George J. News Sync: Enabling scenario-based news exploration. Proceedings of the American Society for

Information Science and Technology. 2011; 48(1):1-10. Available from: https://onlinelibrary.wiley.com/doi/10.1002/meet.2011.14504801078 DOI: 10.1002/meet.2011.14504801078

3. Diamond Ted, Price Susan, Chandrasekar Raman. Actions Speak Louder than Words: Analyzing large-scale query logs to improve the research experience. Code4Lib Journal. Retrieved from http://journal. code4lib. org/articles/8693. 2013.
4. Chandrasekar Raman, Slawson Dean A, Forney Michael K, Surendran Arungunram C, Choudhury Piali, Renshaw Erin. Presenting information related to topics extracted from event classes. 2010 March.

Other Significant Products, Whether or Not Related to the Proposed Project

1. Ma H, Chandrasekar R, Quirk C, Gupta A. Improving search engines using human computation games. Proceedings of the 18th ACM conference on Information and knowledge management. CIKM '09: Conference on Information and Knowledge Management; 02 1 09; Hong Kong China. New York, NY, USA: ACM; c2009. Available from: https://dl.acm.org/doi/10.1145/1645953.1645990 DOI: 10.1145/1645953.1645990
2. Chandrasekar R., Chen H., Corston-Oliver S., Brill E.. Subwebs for specialized search. Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; 2004; c2004. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-8644244082&partnerID=MN8TOARS eid: 2-s2.0-8644244082
3. Chandrasekar Raman, Doran Christine, Srinivas Bangalore. Motivations and methods for text simplification. Proceedings of the 16th conference on Computational linguistics-Volume 2; 1996; c1996.
4. Dziadosz S., Chandrasekar R.. Do thumbnail previews help users make better relevance decisions about web search results?. SIGIR Forum (ACM Special Interest Group on Information Retrieval). 2002; :365-366. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-0036989519&partnerID=MN8TOARS eid: 2-s2.0-0036989519

Synergistic Activities

1. Was Senior Manager, ProQuest, during 2012-2015, in charge of Development, Program Management, and Testing for ProQuest's academic search engine Summon serving several hundred University libraries, with a collection of over 2 billion documents.
2. Started Human Computation Workshop (now AAAI Human Computation Conference)
3. Did seminal work on Automatic Text Simplification.
4. 20+ issued patents in the areas of search, spelling, and related areas mostly while at Microsoft/Microsoft Research.
5. Handled large volumes of document, query, news, and other log data in jobs at Microsoft, Evri.com, ProQuest, and consulting jobs.

Certification:

When the individual signs the certification on behalf of themselves, they are certifying that the

information is current, accurate, and complete. This includes, but is not limited to, information related to domestic and foreign appointments and positions. Misrepresentations and/or omissions may be subject to prosecution and liability pursuant to, but not limited to, 18 U.S.C. §§ 287, 1001, 1031 and 31 U.S.C. §§ 3729-3733 and 3802.

Certified by Chandrasekar, Raman in SciENcv on 2023-04-17 19:46:22

Budget

Budget Justification

Current & Pending Support

We have to enter this using ScienCV, just like we enter our bio.
See C&P: https://www.nsf.gov/bfa/dias/policy/cps.jsp
Just need a 2 line summary, no in-kind

Facilities, Equipment, & Other Resources

We now briefly describe the facilities, equipment, and scientific environment available to the project team at Northeastern University.

Facilities The core facilities used will be Northeastern University's (NU) Institute for Experiential AI (EAI) offices, provided by the university. Collaborative work and user studies may also be conducted in the NU's campuses in San Jose, Seattle, Portland and Boston campuses. All offices are equipped with both wired and wireless networking. Project personnel may also work from their home offices.

Scientific Environment The Institute for Experiential AI (EAI) has close links with Northeastern University's Khoury College of Computer Science which is recognized as a leading computer science research department. NU is a leading research institution in the emerging field of Network Science. The University's many institutes [140] including the Roux Institute in Portland, Maine, and the Network Science Institute in Boston have scholars from the disparate fields of computer science, political science, biology, physics, etc. The project will benefit from academic contact and collaboration with faculty and researchers from these diverse departments.

Equipment The proposed project primarily involves data collection, memory and CPU/GPU-intensive computations, software implementation to compute embeddings, software development to power a number of web-applications, and extensive analysis. The bulk of the project personnel's computing will be done on Northeastern University's **Discovery cluster** (described below), provided by the University's ITS Research Computing team.

Research Computing The Northeastern ITS Research Computing (RC) team provides high-end research computing resources to all Northeastern University affiliated faculty, researchers, and students. The team also manages NU's partnership with the Massachusetts Green High Performance Computing Center (MGHPCC). Resources available to the NU community include a centralized high performance computing (HPC) cluster, storage, software, high-level technical and scientific consultations, education, documentation, and training. All of these resources are available to all faculty, research staff, and students, with RC staff available to assist researchers through consultations on how to leverage hardware and software for scientific applications and workflows.

As of January 2023, the Discovery cluster provides access to over 45,000 CPU cores and over 400 GPUs to all Northeastern faculty and students free of charge. Hardware currently available for research consists of a combination of Intel Xeon (Cascadelake, Skylake_avx512, Broadwell, Haswell, Sandybridge and Ivybridge) and AMD (Zen, Zen2) CPU microarchitectures. Additionally, a selection of NVIDIA Pascal (P100), Volta (V100), Turing (T4), and Ampere (A100) GPUs is available. Discovery is connected to the university network over 10 Gbps Ethernet (GbE) for high-speed data transfer, and Discovery provides 6 PB of available storage on a high-performance file system. Compute nodes are connected with either 10 GbE or high data rate InfiniBand (200 Gbps or 100 Gbps), supporting all types and scales of computational workloads.

A dedicated team of PhD scientists and staff manage the RC environment and support researchers in their use of the Discovery cluster resources. The RC team updates computational resources available through Discovery with the newest technologies on a yearly cycle to support the cutting-edge research being performed by Northeastern faculty and students.

Research groups who require access to dedicated computational resources can request to be part of a "buy-in" option, integrating their hardware into the Discovery cluster to provide unified access to both private and shared compute nodes for their research group members. Faculty-owned hardware that is part of the Discovery cluster is fully managed and maintained by the RC staff at no charge.

Alternatively, researchers can transfer grant dollars to RC to enable a 'Co-op' or paid allocations model. Principal Investigators (PIs) can purchase CPU or GPU hours, or storage capacity, tailored to the specific needs of a research project and/or grant. This alleviates the need for PIs to purchase capital equipment directly.

General Office/Computing Description Faculty, staff, and postdoctoral researchers are provided with furnished offices/shared office spaces/workstations, desktop and/or laptop computers and telephone services. Each department and research center has a common administrative area with a copy and fax machine and other office services and supplies.

Northeastern University computing facilities are connected via a layer-3 switched Ethernet-10/100/1000BT network. Wireless network is available throughout the Northeastern University campus. The facilities are constantly being upgraded to keep pace with developments in the computer industry. The college provides computing resources for both undergraduate and graduate courses, for individual student projects, and for faculty research.

The NEU Information Services technical support is available to faculty, students, and staff through the University's Computer Help Desk, which is a service of the University's centralized Information Services Department. The Computer Help Line/Call Center and IS Customer Services are dedicated to increasing the productivity and satisfaction of faculty, staff, and students using information technology at Northeastern University. They provide faculty and staff with a central point of contact for various resources, thereby providing an efficient and effective means to answer questions and solve problems.

Data Resources

One major source of data is the Allen Institute of AI, in Seattle, WA via their **Semantic Scholar** project. The Semantic Scholar API [141] supports a number of useful features. As discussed in Figure 4 in Section 2, the API provides access to more than 200 million papers from seven sources (MAG, DOI, PubMed, DBLP, PubMedCentral, arXiv and ACL), spanning a variety of disciplines (with an emphasis on STEM). The API also provides access to more than 2B citations, and Specter embeddings for about 125 million papers. As discussed above, the collection is growing rapidly. Semantic Scholar releases data updates frequently, making it freely available; we will continue to download updates for use in this project.

In addition to Semantic Scholar, much of this data is also available from **OpenAlex** project [142], also described in Section 2. The OpenAlex project was launched in early 2022 as a replacement for the Microsoft Academic Graph (MAG), which was being retired. The OpenAlex dataset consists of five types of scholarly entities – venues, concepts, works, authors, and institutions – and links between these. Data may be freely obtained from OpenAlex as a full dump (updated every 2 weeks), a REST API, or a GUI built on their API. In September 2022, OpenAlex contained about 209M works (papers), 213M authors, 124,000 venues (conferences, journals, repositories), 109,000 institutions and 65,000 concepts. As OpenAlex continues to crawl the web, these numbers will continue to increase.

Data Management Plan

The project team is committed to public access for the products of this research. The primary output of the proposed research will include (a) several kinds of embeddings computed over huge datasets. and (b) tools (programs and/or APIs) for ranked retrieval, routing, and recommending papers to read and/or cite. In this section, we describe our plans for collecting, organizing, managing, and releasing these products. Some code and data have already been released as part of our participating in the 2023 Jelinek Summer Workshop on Speech and Language Technology [7, 131].

Data Retention

We will retain the raw data, summarized and reported data, as well as all publications and software for at least three years. All software and data are stored on Northeastern servers that are automatically backed-up and replicated for redundancy. Northeastern offers a variety of solutions including their Discovery cluster as well as the University's Digital Repository Service (NU-DRS).

NU-DRS is a long-term digital asset management system developed and maintained by the Northeastern University Library. The NU-DRS provides the following services: 1) deposition of all file types, many of which are natively supported by the system, 2) provisioning and maintenance of a permanent identifier and URL for both the project space and individual data files via the library's handle server, 3) discovery, access and editorial control using the Shibboleth single sign-on identity management framework, and 4) data storage and backup services, provided jointly by the library and university information systems. The Research Data Management librarian and Digital Production Services staff will work with project staff to determine appropriate metadata models and deposition schedules. Project staff will ensure all relevant files are submitted either to available community resources or to Northeastern's repository system.

The raw data will be stored on Northeastern University's Discovery Cluster (see section on Facilities, Equipment, & Other Resources) research machines, which are backed-up regularly. We will store all data and code in on GitHub hosted on University servers. This data will also be regularly backed up and replicated. Portions of the code and large files (of the order of Terabytes) such as files of embeddings will be stored on Discovery Cluster storage rented for this purpose and made available via the University of Chicago's Globus not-for-profit research cyber-infrastructure. Globus supports relevant discovery and user authentication features, along with efficient network transfer of large files.

Data Format and Analysis Tools

All released data will include documentation describing how the data was collected and the format that the data is represented in. We will use open-source-compatible formats (ASCII, CSV, JSON, etc.) for the research data for this project. All code, scripts, and metadata that we use as a part of conducting the proposed research will be released with the data itself. This will allow other research groups to replicate our results themselves (using either our released data, or data sets that they collect on their own using our tools) and to extend our research.

Data Dissemination

At planned points in the project, updates of project code and data will be made freely available. No legal/ethical restrictions are required on distribution of the code or data, since the data that is sourced is free of such issues. This data will be made available for a estimated period of three years after the end of the project. The software and publicly released data sets will be served

from Northeastern's servers. Any inventions and discoveries that arise from the proposed research will be disclosed to Northeastern University's Technology Transfer Office, and will be released in accordance with the office's policies and procedures. The policies and provisions for re-use, re-distribution, and production of derivative forms will closely follow similar rules for Semantic Scholar code and data.

Access to raw, unprocessed data will be provided via the investigators. All code and data will be available for access, use, and sharing as soon it is reasonably possible after development/processing and will be preserved for a minimum of three years beyond the award period, as required by NSF guidelines.

Human Subjects Limitations

Not applicable.

Postdoctoral Researcher Mentoring Plan

Our postdoctoral researcher will embed within technical teams across the Institute for Experiential AI (EAI) and Khoury College of Computer Science, Northeastern University. This researcher will conduct foundational research in various aspects of natural language processing (NLP) and information retrieval (IR), using breaking work and emerging tools in these areas. The mentoring plan will be implemented through a combination of formal and informal activities. The following mentoring activities/components will help the postdoctoral fellow to contribute to the project and promote and expand their role as an interdisciplinary expert.

Training

The Project PI and Co-PI will provide mentorship to the postdoctoral researchers through regular meetings to discuss research progress, career development, and professional goals. The PI and Co-PI will also provide guidance on grant and project management, grant writing and manuscript preparation, and will facilitate opportunities for the postdoctoral researchers to present their research at national conferences and workshops.

Postdoctoral fellows will have access to and will participate in a variety of training activities at Northeastern's EAI and Khoury College. They will be invited to lectures both at EAI and Khoury College, by visiting speakers, faculty, and students from a wide range of disciplines. They will also have access to professional development workshops on grant writing, public engagement, and mentorship, and research projects.

Multi-Person Mentoring Team

Postdoctoral fellows will be assigned a mentoring team. The team will consist of at least one member of the project leadership team and one member of Khoury College. This mentoring team will help to ensure synergy between the postdoctoral fellow's project activities and their professional growth and development. The mentoring team will advocate for the postdoctoral fellow, help them navigate the difficulties of engaging in multidisciplinary research, and help them identify opportunities for growth that might not be apparent to a single mentor from a single discipline. Postdoctoral fellows will meet with their primary mentor weekly to discuss progress on assigned projects, identify areas where more background research is needed and how to attain it, discuss time management and prioritization, etc. Meetings with other mentors from the team will occur on a less frequent but regular basis.

Collaborative Engagement

Postdoctoral fellows will be provided opportunities to collaborate with other researchers on the team on co-authored projects and presentations. The mentoring plan will also include activities designed to enhance diversity, equity, and inclusion in science. The postdoctoral researcher will participate in workshops and seminars on topics such as implicit bias, cultural competence, and effective communication.

In summary, this mentoring plan is designed to provide a supportive and inclusive environment for the postdoctoral researcher as well as the whole project team, and to promote the development of the next generation of scientists.

Project Personnel And Partner Organizations

List of Project Personnel and Partner Organizations

1. Kenneth Church ; Northeastern University; PI
2. Raman Chandrasekar ; Northeastern University; Co-PI
3. John E. Ortega ; Northeastern University; Postdoctoral Researcher
4. Daniel Weld ; The Allen Institute of AI; unpaid collaborator
5. Gregory L. Shomo ; Research Computing, Northeastern University; unpaid collaborator

Support Letters

ON ORGANIZATIONAL LETTERHEAD

DATE

If the proposal submitted by Dr. Kenneth Church entitled "RI:Small: Beyond Titles and Abstracts: Text + Links are Better Together" is selected for funding by NSF, it is my intent to collaborate and/or commit resources as detailed in the Project Description or the Facilities, Equipment and Other Resources section of the proposal.

SIGNATURE NAME, TITLE