Evaluating Scholarly Impact: Towards Content-Aware Bibliometrics

Saurav Manchanda and **George Karypis**
University of Minnesota
Twin Cities, MN, USA
manch043,karypis@umn.edu

Abstract

Quantitatively measuring the impact-related aspects of scientific, engineering, and technological (SET) innovations is a fundamental problem with broad applications. Traditional citation-based measures for assessing the impact of innovations and related entities do not take into account the content of the publications. This limits their ability to provide rigorous quality-related metrics because they cannot account for the reasons that led to a citation. We present approaches to estimate content-aware bibliometrics to quantitatively measure the scholarly impact of a publication. Our approaches assess the impact of a cited publication by the extent to which the cited publication informs the citing publication. We introduce a new metric, called *Content Informed Index* (CII), that uses the content of the paper as a source of distant-supervision, to quantify how much the cited-node informs the citing-node. We evaluate the weights estimated by our approach on three manually annotated datasets, where the annotations quantify the extent of information in the citation. Particularly, we evaluate how well the ranking imposed by our approach associates with the ranking imposed by the manual annotations. CII achieves up to 103% improvement in performance as compared to the second-best performing approach.

1 Introduction

Scientific, engineering, and technological (SET) innovations have been the drivers behind many of the significant positive advances in our modern economy, society, and life. To measure various impact-related aspects of these innovations various quantitative metrics have been developed and deployed. These metrics play an important role as they are used to influence how resources are allocated, assess the performance of personnel, identify intellectual property (IP)-related takeover targets, value a company's intangible assets, and identify strategic and/or emerging competitors.

Citation networks of peered-reviewed scholarly publications (e.g., journal/conference articles and patents) have widely been used and studied in order to derive such metrics for the various entities involved (e.g., articles, researchers, institutions, companies, journals, conferences, countries, etc. (Aguinis et al., 2012)). However, most of these traditional metrics, such as citation counts and h-index, treat all citations and publications equally and do not take into account the content of the publications and the context in which a prior scholarly work was cited. Another related line of work, such as PageRank (Page et al., 1999) and HITS (Kleinberg, 1999) considers the node centrality (as a proxy for influence) but still operate in a content-agnostic manner.

Content-agnostic metrics fail to precisely size up the scholarly impact of an article as they do not differentiate between the possible reasons that a scholarly work is being cited. In addition, they can be easily manipulated by the presence of malicious entities, such as publication venues indulging in self-citations, which leads to high impact factor, or a group of scholars citing each others' work. For example, Journal Citation Reports (JCR)[1] routinely suppresses many journals that indulge in *citation stacking*, a practice where the reviewers and journal editors pressure authors to cite papers that either they wrote or that are published in their journal. Thus, there is a need to establish content-aware metrics to accurately measure various innovation-related aspects such as their significance, novelty, impact, and market value. Such metrics are essential for ensuring that SET-driven innovations will play an ever more significant role in the future.

A straightforward solution to develop content-aware metrics is to manually annotate the citations,

[1] http://help.incites.clarivate.com/incitesLiveJCR/JCRGroup/titleSuppressions.html

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6041–6053
November 7–11, 2021. ©2021 Association for Computational Linguistics

where the annotations describe the reasons for the citations. These annotations can then be used to train a machine-learning system that takes the content of the publications as input and predicts the reasons for the citation. Along this direction, there has been considerable effort to identify important citations (Valenzuela et al., 2015; Jurgens et al., 2018; Cohan et al., 2019). However, generating labeled data for such supervised approaches is difficult and time-consuming, especially when the meaning of the labels is user-defined.

In this work, we present approaches to estimate content-aware bibliometrics to quantitatively measure the scholarly impact of a publication. Our approaches are distant supervised, that require no manual annotation. The proposed approaches leverage the readily available content of the papers as a source of distant supervision. Our approaches assess the impact of a cited publication by the extent to which it informs the citing publication. They automatically estimate the weights of the edges in the citation network, such that higher-weighted edges correspond to higher-impact citations. We use these weights to introduce a new metric, called *Content Informed Index* (CII). We evaluate CII on three manually annotated datasets, where the annotations tell us the citation importance, thus, quantify the extent of information in the citation. Particularly, we evaluate how well the ranking imposed by CII associates with the ranking imposed by the manual annotations. The proposed approach achieves up to 103% improvement in performance as compared to the second-best performing approach.

2 Related Work

The research areas relevant to the work present in this paper belong to *citation indexing*, *citation recommendation*, *link prediction* approaches, and *distant-supervised credit attribution* approaches, and *citation-intent classification* approaches. We briefly discuss these below:

2.1 Citation indexing

A citation index indexes the links between publications that authors make when they cite other publications. Citation indexes aim to improve the dissemination and retrieval of scientific literature. CiteSeer (Giles et al., 1998; Li et al., 2006) is the first automated citation indexing system that works by downloading publications from the Web and converting them to text. It then parses the papers to extract the citations and the context in which the citations are made in the body of the paper, storing this information in a database. Other examples of popular citation indices include Google Scholar[2], Web of Science[3] by Clarivate Analytics, Scopus[4] by Elsevier and Semantic Scholar[5]. Some examples of subject-specific citation indices include INSPIRE-HEP[6] which covers high energy physics, PubMed[7], which covers life sciences and biomedical topics and Astrophysics Data System[8] which covers astronomy and physics.

2.2 Citation recommendation

Citation recommendation describes the task of recommending citations for a given text. It is an essential task, as all claims written by the authors need to be backed up to ensure reliability and truthfulness. The approaches developed for citation recommendation can be grouped into 4 groups as follows(Färber and Jatowt, 2020): hand-crafted feature-based, topic-modeling-based, machine-translation-based, and neural-network-based approaches. Hand-crafted feature-based approaches are based on features are manually engineered by the developers. For example, text similarity between the citation context and the candidate papers can be used as one of the text-based features. Examples of hand-crafted feature-based approaches include (Färber and Jatowt, 2020; He et al., 2011; LIU et al., 2016; Livne et al., 2014; Rokach et al., 1978). Topic modeling based approaches represent the candidate papers' text and the citation contexts using abstract topics and thereby exploiting the latent semantic structure of texts. Examples of topic modeling-based approaches include (He et al., 2010; Kataria et al., 2010). The machine-translation-based approaches apply the idea of translating the citation context into the cited document to find the candidate papers worth citing. Examples in this category include (He et al., 2012; Huang et al., 2012). Finally, the popular examples of neural-network-based models include (Ebesu and Fang, 2017; Han et al., 2018; Huang et al., 2015; Kobayashi et al., 2018; Tang et al., 2014; Yin and Li, 2017).

[2] https://scholar.google.com/

[3] http://www.webofknowledge.com/

[4] https://www.scopus.com/

[5] https://www.semanticscholar.org/

[6] https://inspirehep.net/

[7] https://pubmed.ncbi.nlm.nih.gov/

[8] http://ads.harvard.edu/

2.3 Link-prediction

Link-prediction is the problem of predicting the existence of a link (connection) between two nodes in a network. A good link-prediction model predicts the likelihood of a link between two nodes, thus, link-prediction can be a useful tool to find likely citations in a citation network. The citation recommendation task described previously can be thought of as a special case of link-prediction. Following the taxonomy described in (Martínez et al., 2016), link-prediction approaches can be broadly categorized into three categories: similarity-based approaches, probabilistic and statistical approaches, and algorithmic approaches. The similarity-based approaches assume that nodes tend to form links with other similar nodes and that two nodes are similar if they are connected to similar nodes or are near in the network according to a given similarity function. Examples of popular similarity functions include number of common neighbors (Liben-Nowell and Kleinberg, 2007), Adamic-Adar index (Adamic and Adar, 2003), etc. The probabilistic and statistical approaches assume that the network has a known structure. These approaches estimate the model parameters of the network structure using statistical methods and use these parameters to calculate the likelihood of the presence of a link between two nodes. Examples of probabilistic and statistical approaches include (Guimerà and Sales-Pardo, 2009; Huang, 2010; Wang et al., 2007). Algorithmic approaches directly use the link-prediction as supervision to build the model. For example, link-prediction task can be formulated as a binary classification task where the positive instances are the pair of nodes that are connected in the network, and negative instances are the unconnected nodes. Examples include (Menon and Elkan, 2011; Bliss et al., 2014). Unsupervised or self-supervised node embedding (such as DeepWalk (Perozzi et al., 2014), node2vec (Grover and Leskovec, 2016)), followed by training a binary classifier and Graph Neural network approaches such as GraphSage (Hamilton et al., 2017) belong to this category.

2.4 Distant-supervised credit-attribution

Various distant-supervised approaches have been developed for credit-attribution on text documents. A document may be associated with multiple labels but all the labels do not apply with equal specificity to the individual parts of the documents. *Credit attribution* problem refers to identifying the specificity of labels to different parts of the document. Various probabilistic and neural-network-based approaches have been developed for this problem, such as Labeled Latent Dirichlet Allocation (LLDA) (Ramage et al., 2009), Partially Labeled Dirichlet Allocation (PLDA) (Ramage et al., 2011), Multi-Label Topic Model (MLTM) (Soleimani and Miller, 2017), Segmentation with Refinement (SEG-REFINE) (Manchanda and Karypis, 2018), and Credit Attribution with Attention (CAWA) (Manchanda and Karypis, 2020).

Another line of work uses distant-supervised credit-attribution for query-understanding in product search. Examples include, (i) using the reformulation logs as a source of distant-supervision to estimate a weight for each term in the query that indicates the importance of the term towards expressing the query's product intent (Manchanda et al., 2019a,b); and (ii) annotating individual terms in a query with the corresponding intended product characteristics, using the characteristics of the engaged products as a source of distant-supervision (Manchanda et al., 2020).

2.5 Citation-intent classification

In general, these approaches treat citation-intent classification as a text classification problem and require the availability of training data with ground truth annotations. Representative examples include rule-based approaches (Pham and Hoffmann, 2003; Garzone and Mercer, 2000) as well as machine-learning driven approaches (Valenzuela et al., 2015; Jurgens et al., 2018; Cohan et al., 2019). Generating labeled data for these supervised approaches is difficult and time-consuming, especially when the meaning of the labels is user-defined. In contrast, our approaches require no manual annotation.

3 Content-Informed Index (CII)

To address the disadvantages of content-agnostic bibliometrics, we present approaches that use machine-learning to estimate content-aware bibliometrics to measure the scholarly impact of a publication. Our approaches are distant supervised, requiring no manual annotation. They automatically estimate the weights of the edges in the citation network, such that edges with higher weights correspond to higher-impact citations. We use these weights to come up with a new metric, called *Content Informed Index* (CII). Next, we discuss the assumptions behind CII and provide deeper details.

3.1 Assumptions and problem definition

In the absence of labels that define the *impact*, we assume that the extent to which a cited paper informs (contributes or is used by) the citing paper is an indication of the citation's impact. We assume that each paper P_i can be represented as a set of *concepts* C_i, a subset of which are the *historical* concepts that were already known prior to P_i. These historical set of concepts of the paper P_i are borrowed from the papers that P_i cites, and are denoted by H_i.

The contribution of a cited paper P_j towards the citing paper P_i is the set of concepts that P_i borrows from P_j, i.e., the set of concepts $C_j \cap H_i$. The task at hand is to quantitatively approximate the extent to which C_j contributes towards H_i, and hence contributes towards C_i. Next, we describe the framework that we employed to achieve this.

3.2 Representing the set of concepts associated with a paper

Figuring out the explicit-human-interpretable concepts associated with a paper is not trivial, and can be interpreted differently by different audiences. However, in our case, we are interested in getting a representation of the semantic meaning associated with the concepts, rather than the concepts themselves. One of the simple approaches to get the representation of the semantic meaning associated with the concepts is to use the pre-trained representation (embedding) of the text associated with the concepts themselves. Being trained on language-modeling tasks, such pre-trained representations easily capture semantic meanings of words/sequence of words. For simplicity, we use the representations pre-trained on scientific documents provided by ScispaCy (Neumann et al., 2019). In addition, we only use the representation of the abstract to get the representation of the concepts of a paper. The representation of C_i is denoted by $r(C_i)$.

Note that we can use more sophisticated representation techniques for this part, but limit ourselves to abstract representations provided by ScispaCy[9] for simplicity (further discussed in Section 7). Other potential improvements include: (i) using better pre-trained representations such as BERT (Devlin et al., 2019), ELMo (Peters et al., 2018), etc., and (ii) representation for a more representative summary of the paper than the abstract. Further, CII is not suitable for the class of papers for which our assumptions do not hold. A particular case is of the review papers, which tend to have a lot of content, and a limited-word abstract may not be a representative summary of the complete paper. Thus, the CII estimates that depend on these papers would not be reliable.

3.3 Representing the set of historical concepts H_i

As the set of historical concepts H_i is a union of the borrowed concepts from the cited papers (C_j), we simply represent the set of historical concepts as a weighted linear combination of the representation of the concepts of the cited papers, i.e.,

$$\begin{aligned} r(H_i) \quad &= \sum_{P_i \text{ cites } P_j} \tilde{w}_{ji} r(C_j) \\ \text{subject to} \quad &\sum_{P_i \text{ cites } P_j} \tilde{w}_{ji}^2 = 1 \\ &\tilde{w}_{ji} \geq 0; \forall (i,j). \end{aligned} \tag{1}$$

We have the constrained norm condition ($\sum_{P_i \text{ cites } P_j} \tilde{w}_{ji}^2 = 1$) to make the representation of $r(H_i)$ agnostic to the number of cited-papers (a paper can cite multiple papers to reference the same borrowed concepts)[10]. We model the weights $\tilde{w}_{ji}$ as a function of the concepts of the cited paper, and the concepts in the citation context. The approach to estimating these weights is described next.

3.4 Supervision task

Since CII does not depend upon the availability of explicit manual annotations, we need to address the challenge of finding an alternative task, with similar underlying principles as the task at hand. Recall that, CII assumes the extent to which a cited paper informs (or *explains*) the citing paper is an indication of the citation's impact. In this direction, we propose to minimize the *explanation* loss, where the *explanation* tries to explain the concepts C_i of the paper P_i using the historical concepts H_i i.e., the concepts of the cited papers (C_j). Thus, we formulate our problem as a distant-supervised

[9] The evaluation dataset contains papers from many scientific domains but ScispaCy is specific to biomedical/clinical texts, and performed better than the word2vec embeddings pre-trained on general web crawled text.

[10] In addition to using the $L2$-norm as constraint, we also experimented with $L1$-norm, but the setup with $L1$-norm constraint lead to sparse $\tilde{w}_{ji}$ and lower performance as compared to the setup with $L2$-norm constraint.

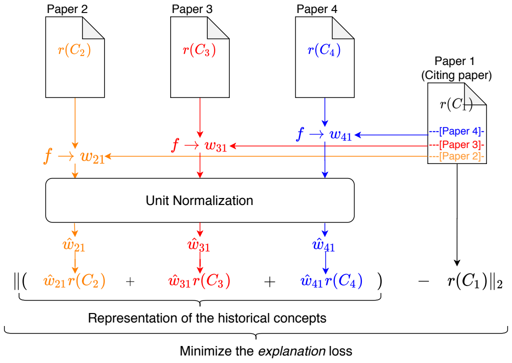

Figure 1: Overview of Content-Informed Index. Paper P_1 cites papers P_2, P_3 and P_4. The weights w_{21}, w_{31}, and w_{41} quantifies the extent to which P_2, P_3 and P_4 informs P_1, respectively. The function f is implemented as a Multilayer Perceptron.

problem, and the content of the papers acts as a source of distant-supervision. Combining it with the discussion in Sections 3.2 and 3.3, we formally describe our formulation as follows: We model the weights $\tilde{w}_{ji}$ in Equation 1 as the normalized similarity measure between the concepts of the cited paper, and the concepts in the citation context. Thus, to estimate $\tilde{w}_{ji}$, we first estimate unnormalized $\tilde{w}_{ji}$, denoted by w_{ji}, and then normalize w_{ji} so as to have unit norm. The unnormalized weight w_{ji} is precisely the extent to which C_j contributes towards H_i (and hence C_i), i.e., the weight that we wish to estimate in this paper. Specifically, the above discussion leads to the following mathematical formulation:

$$\begin{aligned}
\underset{f}{\text{minimize}} \quad & \sum_i \|r(C_i) - \sum_{P_i \text{ cites } P_j} \tilde{w}_{ji} r(C_j)\|^2 \\
\text{subject to} \quad & \tilde{w}_{ji} = \frac{w_{ji}}{\sqrt{\sum_{P_i \text{ cites } P_j} w_{ji}^2}}; \forall (i,j), \\
& w_{ji} = f(r(C_j), r(C_{ji})); \forall (i,j), \\
& w_{ji} \geq 0; \forall (i,j).
\end{aligned} \tag{2}$$

We estimate w_{ji} as a multilayer perceptron, that takes as input the representations of the concepts in the cited paper and the concepts in the citation context. Similar to $r(C_j)$, we use the ScispaCy vector representation for the citation context as the representation of the context and denote it by $r(C_{ji})$. To take care non-negativity constraint for the w_{ji}, the function $f(\cdot)$ can be implemented as a multilayer perceptron, with a single output node, and a non-negative mapping at the output node. Note that, if the set of weights w_{ji} minimize Equation (2), then so will any scalar multiplication of the weights w_{ji}. This can potentially lead to the estimated weights being incomparable across different citing papers. Empirically, we found that having an additional max-bound constraint on the estimated weights ($w_{ji} \leq b$) helps to avoid this pitfall[11], as it essentially limits the projection space of the weights w_{ji}. We do not need to explicitly set the max-bound b, but it is implicitly set by the $L2$ regularization of the weights of the function f. The $L2$ regularization parameter is treated as a hyperparameter. Figure 1 shows an overview of Content-Informed Index (CII).

4 Experimental methodology

4.1 Evaluation methodology and metrics

We need to evaluate how well the weights estimated by our proposed approach quantifies the extent to which a cited paper informs the citing paper. To this end, we leverage various manually annotated datasets (explained later in Section 4.3), where the annotations quantify the extent of information in the citation. The task inherently becomes an ordinal association, and we need to evaluate how well the ranking imposed by our proposed method associates with the ranking imposed by the manual annotations. As a measure of rank correlation, we use the non-parametric Somers' Delta (Somers, 1962) (denoted by Δ). Values of Δ range from $-1(100\%$ negative association, or perfect inversion) to $+1(100\%$ positive association, or perfect agreement).

4.2 Baselines

We choose representative baselines from diverse categories as discussed below:

4.2.1 Link-prediction approaches

The citation weights that we estimate in this paper can also be looked at from the link-prediction perspective, i.e., assigning a score to every citation (link) in the citation graph, that encodes the likelihood of the existence of a link. We compare against two link-prediction methods, one based on the classic network embedding approach, and the other belonging to Graph Neural Network approaches.

- DeepWalk (Perozzi et al., 2014) is a popular method to learn node embeddings. Once we

[11] In theory, the pitfall still remains, but we believe that the input representations lie in smooth space, so the output of the multilayer perceptron is also smooth enough. This factor, along with the max-bound avoids drastic differences between the estimated weights across different citing papers

have node embeddings as the output of DeepWalk, we train a binary classifier, with the positive instances as the pairs of nodes which are connected in the network, and negative instances are the unconnected nodes (generated using negative sampling). We provide results using two different classifiers: Logistic Regression (denoted by DeepWalk+LR) and Multilayer Perceptron (denoted by DeepWalk+MLP). Note that Deepwalk is a transductive model, and does not use the content of the papers to estimate the model.

- GraphSage (Hamilton et al., 2017) is a Graph Convolutional Network (GCN) based framework for inductive representation learning on graphs. GraphSage uses the link-prediction loss for training, so does not use a second step (as in DeepWalk) to train the classifier. Note that, GraphSage is an inductive model, so considers the content of the papers in addition to the network topology.

4.2.2 Text-similarity based baselines

We can think of the function f as a similarity measure between the cited paper and the citation context. Thus, we consider the following similarity measures as our baselines: We use the same pre-trained representations as we used as an input to CII, and cosine similarity as the similarity measure, which is a popular similarity measure for text data.

- Similarity-Abstract-Context: Similarity between the cited abstract and the citation context.
- Similarity-Context-Abstract: Similarity between the citing abstract and the citation context.
- Similarity-Abstract-Abstract: Similarity between the cited abstract and citing abstract.

To calculate each of the above similarity measures, we use the same pre-trained representations as we used as an input to CII, and cosine similarity as the similarity measure. The baselines belonging to this category can also be thought of as similarity-based link prediction approaches.

4.2.3 Reference Frequency based baselines

We also consider another simple baseline, referred to as *Reference Frequency*, where we assume that the more frequently the cited paper is referenced in the citing paper, the higher the chances of the cited paper informing the citing paper. This assumption has also been used as a feature in prior supervised approaches (Valenzuela et al., 2015). The absolute frequency of referencing a cited-paper may provide a good signal regarding the information borrowed from the cited paper when comparing with other papers being cited by the same citing paper. However, as the citation behavior differs between papers, the absolute frequency may not be comparable across different citing papers. Thus, we also provide results after doing normalization of the absolute frequency of the citation references for each citing paper. We provide results for mean, max, and min normalization. Specifically, given a citation and the corresponding citing paper, the information weight for a citation is calculated by dividing the number of references of that citation, by the mean, max, and min of references of all the citations in that citing paper, respectively.

4.3 Datasets

The Semantic Scholar Open Research Corpus (S2ORC): The S2ORC (Lo et al., 2020) dataset is a citation graph of 81.1 million academic publications and 380.5 million citation edges. We only consider the publications for which full-text is available and abstract contains at least 50 words. This leaves us with a total of $5,653,297$ papers, and $30,533,111$ edges (citations).

ACL-2015: The ACL-2015 (Valenzuela et al., 2015) dataset contains 465 citations gathered from the ACL anthology[12], represented as tuples of (cited paper, citing paper), with ordinal labels ranging from 0 to 3, in increasing order of importance. The citations were annotated by one expert, followed by annotation by another expert on a subset of the dataset, to verify the inter-annotator agreement. We only use the citations for which we have the inter-annotator agreement, and the citations are present in the S2ORC dataset we described before. The selected dataset contains 300 citations among 316 unique publications. The total number of unique citing publications are 283 and the total number of unique cited publications are 38.

ACL-ARC: The ACL-ARC (Jurgens et al., 2018) is a dataset of citation intents based on a sample of papers from the ACL Anthology Reference Corpus (Bird et al., 2008) and includes 1,941 citation instances from 186 papers and is annotated by domain experts. The dataset provides ACL IDs for the papers in the ACL corpus, but does not provide an identifier to the papers outside the ACL

[12] https://www.aclweb.org/anthology/

Table 1: Results on the Somers' Δ metric.

Model	ACL-2015	ACL-ARC	SciCite
Content-Informed Index (CII)	**0.428 ± 0.013**	**0.308 ± 0.010**	**0.296 ± 0.006**
Reference Frequency (Absolute)	0.325 ± 0.000	**0.308 ± 0.000**	0.144 ± 0.000
Reference Frequency (Mean-normalized)	0.351 ± 0.000	0.300 ± 0.000	0.120 ± 0.000
Reference Frequency (Min-normalized)	0.321 ± 0.000	0.298 ± 0.000	0.145 ± 0.000
Reference Frequency (Max-normalized)	0.270 ± 0.000	0.172 ± 0.000	0.035 ± 0.000
Similarity-Abstract-Abstract	−0.041 ± 0.000	0.091 ± 0.000	−0.003 ± 0.000
Similarity-Abstract-Context	−0.147 ± 0.000	0.090 ± 0.000	−0.125 ± 0.000
Similarity-Context-Abstract	0.013 ± 0.000	−0.062 ± 0.000	−0.202 ± 0.000
Deepwalk+Logistic-Regression	−0.071 ± 0.016	0.190 ± 0.006	−0.037 ± 0.018
Deepwalk+Multilayer-Perceptron	−0.026 ± 0.011	0.205 ± 0.024	−0.047 ± 0.015
GraphSage	0.023 ± 0.045	0.132 ± 0.024	0.049 ± 0.019

[1] **Boldfaced** entries correspond to the overall best-performing method.
[2] The ACL-ARC dataset is used as a validation dataset for parameter selection.
[3] The reported results are the mean±standard-deviation corresponding to five different runs with different seeds.
[4] The Reference-frequency and text-similarity based baselines are deterministic, thus the std-deviation of their results is zero.

corpus, making it difficult to map many citations to the S2ORC corpus. However, it provided the titles of those papers, and we used these titles to map these papers to the papers in the S2ORC dataset, if matching titles were found. The annotations in ACL-ARC are provided at individual citation-context level, leading to multiple annotations for some of the (cited paper, citing paper) pair. In such cases, we chose the highest-informing annotation for such (cited paper, citing paper) pairs. The selected dataset contains 460 citations among 547 unique publications. The total number of unique citing publications are 145 and the total number of unique cited publications are 413.

SciCite: SciCite (Cohan et al., 2019) is a dataset of citation intents based on a sample of papers from the Semantic Scholar corpus[13], consisting of papers in general computer science and medicine domains. Citation intent was labeled using crowdsourcing. The annotators were asked to identify the intent of a citation, and were directed to select among three citation intent options: Method, Result/Comparison and Background. This resulted in a total 9, 159 crowdsourced instances. We use the citations that are present in the S2ORC dataset we described before. Similar to ACL-ARC, the annotations are provided at individual citation-context level, leading to multiple annotations for some of the (cited paper, citing paper) pair. For such cases, we chose the highest-informing annotation for the (cited paper, citing paper) pairs. The selected dataset contains 352 citations among 704 unique publications. There is no repeated citing or cited publication in this dataset, thus, the total number of unique citing as well as unique cited publications are 352 each.

4.4 Parameter selection

We treat one of the evaluation datasets (ACL-ARC) as the validation set and chose the hyperparameters of our approaches and baselines concerning best performance on this dataset. For DeepWalk, we use the implementation provided here[14], with the default parameters, except the dimensionality of the estimated representations, which is set to 200 (for the sake of fairness, as the used 200 dimensional text representations for CII). For the models that require learning, i.e., the logistic regression part of Deepwalk, MLP part of Deepwalk, GraphSage, and CII, we used the ADAM (Kingma and Ba, 2015) optimizer, with an initial learning rate of 0.0001, and further use step learning rate scheduler, by exponentially decaying the learning rate by a factor of 0.2 every epoch. We use $L2$ regularization of 0.0001. The function f in CII was implemented as a multilayer perceptron, with three hidden layers, with 256, 64, and 8 neurons, respectively. We use the same network architecture for the MLP that we train on top of DeepWalk representations. We train the logistic regression and MLP parts of Deepwalk, GraphSage, and CII for a maximum of 50 epochs, and do early-stopping if the validation performance does not improve for 5 epochs. For GraphSage, we use the implementation provided by DGL[15]. We used a mini-batch size of 1024 for training.

[13] https://www.semanticscholar.org/

[14] https://github.com/xgfs/deepwalk-c

[15] https://github.com/dmlc/dgl/blob/master/examples/pytorch/graphsage

5 Results and discussion

5.1 Quantitative analysis

Table 1 shows the performance of the various approaches on the Somers' Delta (Δ) for each of the three evaluation datasets. For ACL-2015 and SciCite, CII outperforms the competing approaches; while for the ACL-ARC dataset, CII performs on par with the best performing approach. The improvement of CII over the second-best performing approach is 22% and 103%, on the ACL-2015 and SciCite datasets, respectively.

Interestingly, the simplest baseline, Reference-frequency, and its normalized forms are the second-best performing approaches. While Reference-frequency performs at par with the CII on the ACL-ARC dataset, it does not perform as well on the other two datasets. This can be attributed to the fact that the number of unique citing papers in the ACL-ARC dataset is relatively small. Thus, many citations in ACL-ARC are shared by the same citing paper, which is not the case with the other two datasets. Thus, as mentioned in Section 4.2, the absolute frequency of referencing a cited-paper may provide a good signal regarding the information borrowed from the cited paper, when comparing with other papers being cited by the same citing paper. Further, even the normalized forms of the Reference-frequency lead to only a marginal increase in performance for the ACL-2015 and SciCite datasets. Thus, the simple normalizations (such as mean, max, and min normalization used in this paper), are not sufficient to address the difference in citation behavior between different papers.

Furthermore, we observe that simple similarity-based approaches, such as cosine-similarity between pairs of various entities (each combination of citing abstract, citing abstract, and citation-context) perform close to random scoring (Δ value of close to zero). This validates that the simple similarity measures, like cosine similarity, are not sufficient to manifest the information that a cited-paper lends to the citing-paper; thus, showing the necessity of more expressive approaches, like CII.

In addition, the other learning-based link-prediction-based approaches perform considerably worse than the simple baseline reference-frequency. While on ACL-2015 and SciCite datasets, they perform close to random scoring, the performance on ACL-ARC dataset is better than the random baseline. For the link-prediction approaches to perform well, the basic assumption is that the majority of the edges (links) in the training set are indeed the informing citations. If such assumption holds, the link-prediction approaches can pick the majority signal (informing citations) and ignore the noise (non-informing citations) owing to the low-dimensional projections of the nodes (or edges). However, such assumption does not hold in the citation graphs, with only a fraction of citation being the informing citations. For example, it has been estimated that authors read only 20% of the works they cite (Simkin and Roychowdhury, 2002).

5.2 Qualitative analysis

To understand the patterns that the proposed approach CII learns, we look into the data instances with the highest and lowest predicted weights. As the function f takes as input both the abstract of the cited paper and the citation context, the learned patterns can be a complex function of the cited paper abstract and the citation context. Thus, for simplicity, we limit the discussion here to understand the linguistic patterns in the citation context, and their association with the predicted weights.

We repeat the same exercise for the citation-contexts with the lowest predicted weights. Figures 2 and 3 shows the wordclouds for the highest weighted citations and lowest weighted citations, respectively. These figures show clear discriminatory patterns between the highest-weighted and lowest-weighted citations, that relate well with the information carried by a citation. For example, the words such as 'used' and 'using' are very frequent in the citation contexts of the highest weighted citations. This is expected, as such verbs provide a strong signal that the cited work was indeed employed by the citing paper. Another interesting pattern in the highest weighted citations is the presence of words like 'fig', 'figure', and 'table'. Such words are usually present when the authors describe important concepts, such as methods and results. As such, citations in these important sections indicate that the cited work is used/extended in the citing paper, which signals importance.

On the other hand, the wordcloud for the least weighted citations (Figure 3) is dominated by weasel words such as 'may', 'many', 'however', etc. The words such as 'many' commonly occur in the related work section of the paper, where the paper presents some examples of other related works to emphasize the problem that the citing paper is solving. The words like 'may', 'however', 'but' etc

Table 2: Examples of lowest and highest CII-weighted citations.

Citing-paper title	Relavant citation-context	Score
AAV-mediated gene therapy for retinal disorders: from mouse to man	The number of regenerating axons per nerve was then calculated at each distance using a previously developed formula (Lim et al., 2016; Bei et al., 2016), with the total number of axons equal to πr^2.	$8.32\mathrm{e}{-1}$
	Most experimental therapies that stimulate RGC axon regeneration involve interventions at the time of injury or, in the case of many gene therapies, prior to injury (Buch et al., 2008). While such studies are valuable for identifying therapeutic targets and elucidating mechanisms of RGC axon regeneration, they are not readily translatable to human patients.	$3.46\mathrm{e}{-5}$
Hippocampal Memory Traces are Differentially Modulated by Experience, Time, and Adult Neurogenesis	Exploration in response to a novel open field (OF) was measured as previously described (Richardson-Jones et al., 2010).	$8.33\mathrm{e}{-1}$
	Many models have stressed the importance of the hippocampus (HPC) subregions in distinguishing similar patterns (pattern separation) and in completing partial patterns (pattern completion) (Bakker et al., 2008; Leutgeb et al., 2007; Marr, 1971; McHugh et al., 2007; O'Reilly and McClelland, 1994; Treves and Rolls, 1992)..	$1.55\mathrm{e}{-5}$

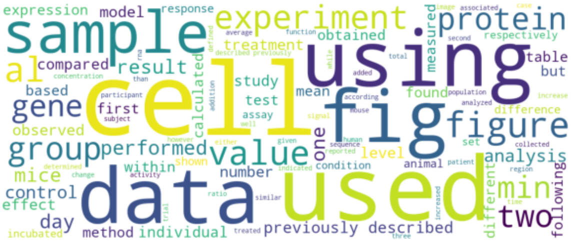

Figure 2: Word-cloud for the words that appear in the citation context of the citations with the highest predict importance weights.

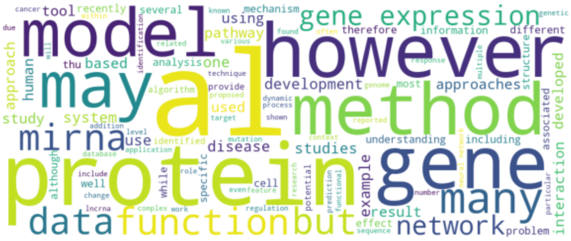

Figure 3: Word-cloud for the words that appear in the citation context of the citations with the least predict importance weights.

are commonly used to describe some limitations of the cited work. Such citations are expected to be incidental and carry less information.

We also look at some examples of individual citation contexts and the predicted weights for them. Table 2 shows two citing papers, with an example of a high weighted citation and an example of a low weighted citation for each of those papers. For these examples, we see that the high predicted weight corresponds to cited work indeed being employed by the citing paper. For example, the high weight citations for the papers titled 'AAV-mediated gene therapy for retinal disorders: from mouse to man' and 'Hippocampal Memory Traces are Differentially Modulated by Experience, Time, and Adult Neurogenesis' in Table 2 correspond to formulas employed by these papers, that were developed in the cited papers. Similarly, the lowest weighted citations correspond to cited papers that are not informative. For example, for the paper titled 'AAV-mediated gene therapy for retinal disorders: from mouse to man', the lower-weighted citation describes the limitation of the cited paper. Similarly, for the paper titled 'Hippocampal Memory Traces are Differentially Modulated by Experience, Time, and Adult Neurogenesis', the lower-weighted citation corresponds to background work, which is not an informing citation.

6 Discussion and Conclusions

In this paper, we presented approaches to estimate content-aware bibliometrics to accurately quantitatively measure the scholarly impact of a publication. Our distant-supervised approaches use the content of the publications to weight the edges of a citation network, where the weights quantify the extent to which the cited-publication informs the citing-publication. Experiments on the three manually annotated datasets show the advantage of using the proposed method on the competing approaches. The code is available on GitHub[16].

Our work makes a step towards developing content-aware bibliometrics, and envision that the proposed method will serve as a motivation to develop other rigorous quality-related metrics.

[16] https://github.com/gurdaspuriya/Evaluating-Scholarly-Impact/

7 Broader impact and ethics discussion

Quantitative metrics to measure the impact-related aspects of scientific, engineering, and technological (SET) innovations play an important role in the modern society. These metrics are used to influence how resources are allocated, assess the performance of personnel, identify intellectual property (IP)-related takeover targets, value a company's intangible assets (IP is such an asset), and identify strategic and/or emerging competitors. Thus, metrics that accurately and quantitatively the innovation-related aspects, are essential for ensuring that SET-driven innovations will play an ever more significant role in the future. This paper is a step in this direction.

While our discussion and evaluation focused on identifying informing citations, our approach is not restricted to this domain, and can be used to derive impact metrics for the various involved entities. For example, the content-aware weights estimated by the CII convert the original unweighted citation network to a weighted one. Consequently, this weighted network can be used to derive impact metrics for the various involved entities, like the publications, authors etc. For example, to find the impact of a publication, the sum of weights outgoing from its corresponding node can be used to quantify the impact of the publication, instead of using vanilla citation count. Further, the impact can be propagated through generations of citations (similar to CiteRank (Walker et al., 2005)), by simply doing a weighted pagerank on this weighted graph.

However, as there are benefits, there are also risks and concerns. Like other bibliometrics, CII is also prone to be manipulated by the bad actors. For example, the citation contexts can be constructed in a way (using particular keywords as shown in Figures 2 and 3) so as to fool CII. A way of mitigating these risks is to use more advanced information extraction approaches for the accurate assessment of the citation context. In this direction, we can leverage the extensive literature on concept and context extraction in NLP: from the highly specific ('does this cited paper really discuss the entity our approach found in the citing sentence?') to much more general ('is this mention positive or negative?') and much in between. Having said that, it is also important for an impact metric needs to be simple to be widely adopted, and added complexity can lead to issues of trust and acceptance by the user community. Thus, we encourage the research community and policy makers to come together to understand and evaluate the specific impacts and risks of using more expressive and relatively complex metrics. We envision that this paper will serve as a motivation to continue the discussion in the aforementioned directions.

Acknowledgements

This work was supported in part by NSF (1447788, 1704074, 1757916, 1834251), Army Research Office (W911NF1810344), and the Digital Technology Center at the University of Minnesota. Access to research and computing facilities was provided by the Digital Technology Center and the Minnesota Supercomputing Institute.

References

Lada A Adamic and Eytan Adar. 2003. Friends and neighbors on the web. *Social networks*, 25(3):211–230.

Herman Aguinis, Isabel Suárez-González, Gustavo Lannelongue, and Harry Joo. 2012. Scholarly impact revisited. *Academy of Management Perspectives*, 26(2):105–132.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Catherine A Bliss, Morgan R Frank, Christopher M Danforth, and Peter Sheridan Dodds. 2014. An evolutionary algorithm approach to link prediction in dynamic social networks. *Journal of Computational Science*, 5(5):750–764.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. 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)*, pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Travis Ebesu and Yi Fang. 2017. Neural citation network for context-aware citation recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 1093–1096. ACM.

Michael Färber and Adam Jatowt. 2020. Citation recommendation: Approaches and datasets. *ArXiv preprint*, abs/2002.06961.

Mark Garzone and Robert E Mercer. 2000. Towards an automated citation classifier. In *Conference of the canadian society for computational studies of intelligence*, pages 337–346. Springer.

C Lee Giles, Kurt D Bollacker, and Steve Lawrence. 1998. Citeseer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pages 89–98.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 855–864. ACM.

Roger Guimerà and Marta Sales-Pardo. 2009. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1024–1034.

Jialong Han, Yan Song, Wayne Xin Zhao, Shuming Shi, and Haisong Zhang. 2018. hyperdoc2vec: Distributed representations of hypertext documents. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2384–2394, Melbourne, Australia. Association for Computational Linguistics.

Jing He, Jian-Yun Nie, Yang Lu, and Wayne Xin Zhao. 2012. Position-aligned translation model for citation recommendation. In *International symposium on string processing and information retrieval*, pages 251–263. Springer.

Qi He, Daniel Kifer, Jian Pei, Prasenjit Mitra, and C. Lee Giles. 2011. Citation recommendation without author supervision. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, pages 755–764. ACM.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and C. Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 421–430. ACM.

Wenyi Huang, Saurabh Kataria, Cornelia Caragea, Prasenjit Mitra, C. Lee Giles, and Lior Rokach. 2012. Recommending citations: translating papers into references. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 1910–1914. ACM.

Wenyi Huang, Zhaohui Wu, Chen Liang, Prasenjit Mitra, and C. Lee Giles. 2015. A neural probabilistic model for context based citation recommendation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2404–2410. AAAI Press.

Zan Huang. 2010. Link prediction based on graph topology: The predictive value of generalized clustering coefficient. *Available at SSRN 1634014*.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Saurabh Kataria, Prasenjit Mitra, and Sumit Bhatia. 2010. Utilizing context in generative bayesian models for linked corpus. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*. AAAI Press.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jon M Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.

Yuta Kobayashi, Masashi Shimbo, and Yuji Matsumoto. 2018. Citation recommendation using distributed representation of discourse facets in scientific articles. In *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries*, pages 243–251.

Huajing Li, Isaac G. Councill, Wang-Chien Lee, and C. Lee Giles. 2006. Citeseerx: an architecture and web service design for an academic document search engine. In *Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23-26, 2006*, pages 883–884. ACM.

David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031.

Ya'ning LIU, Rui YAN, and Hongfei YAN. 2016. Personalized citation recommendation based on user's preference and language model. *Journal of Chinese Information Processing*, (2):18.

Avishay Livne, Vivek Gokuladas, Jaime Teevan, Susan T. Dumais, and Eytan Adar. 2014. Citesight: supporting contextual citation recommendation using differential search. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 807–816. ACM.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Saurav Manchanda and George Karypis. 2018. Text segmentation on multilabel documents: A distant-supervised approach. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1170–1175. IEEE.

Saurav Manchanda and George Karypis. 2020. Cawa: An attention-network for credit attribution. In *AAAI*, pages 8472–8479.

Saurav Manchanda, Mohit Sharma, and George Karypis. 2019a. Intent term selection and refinement in e-commerce queries. *ArXiv preprint*, abs/1908.08564.

Saurav Manchanda, Mohit Sharma, and George Karypis. 2019b. Intent term weighting in e-commerce queries. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 2345–2348. ACM.

Saurav Manchanda, Mohit Sharma, and George Karypis. 2020. Distant-supervised slot-filling for e-commerce queries. *ArXiv preprint*, abs/2012.08134.

Víctor Martínez, Fernando Berzal, and Juan-Carlos Cubero. 2016. A survey of link prediction in complex networks. *ACM computing surveys (CSUR)*, 49(4):1–33.

Aditya Krishna Menon and Charles Elkan. 2011. Link prediction via matrix factorization. In *Joint european conference on machine learning and knowledge discovery in databases*, pages 437–452. Springer.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: online learning of social representations. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 701–710. ACM.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Son Bao Pham and Achim Hoffmann. 2003. A new approach for scientific citation classification using cue phrases. In *Australasian Joint Conference on Artificial Intelligence*, pages 759–771. Springer.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, Singapore. Association for Computational Linguistics.

Daniel Ramage, Christopher D. Manning, and Susan T. Dumais. 2011. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, pages 457–465. ACM.

Lior Rokach, Prasenjit Mitra, Saurabh Kataria, Wenyi Huang, and Lee Giles. 1978. A supervised learning method for context-aware citation recommendation in a large corpus. *INVITED SPEAKER: Analyzing the Performance of Top-K Retrieval Algorithms*, page 1978.

Mikhail V Simkin and Vwani P Roychowdhury. 2002. Read before you cite! *arXiv preprint cond-mat/0212043*.

Hossein Soleimani and David J Miller. 2017. Semisupervised, multilabel, multi-instance learning for structured data. *Neural computation*, 29(4):1053–1102.

Robert H Somers. 1962. A new asymmetric measure of association for ordinal variables. *American sociological review*, pages 799–811.

Xuewei Tang, Xiaojun Wan, and Xun Zhang. 2014. Cross-language context-aware citation recommendation in scientific articles. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 817–826. ACM.

Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*.

D. Walker, H. Xie, Koon-Kiu Yan, and S. Maslov. 2005. Citerank: A google-inspired ranking algorithm for citation networks.

Chao Wang, Venu Satuluri, and Srinivasan Parthasarathy. 2007. Local probabilistic models for link prediction. In *Seventh IEEE international conference on data mining (ICDM 2007)*, pages 322–331. IEEE.

Jun Yin and Xiaoming Li. 2017. Personalized citation recommendation via convolutional neural networks. In *Asia-Pacific web (APWeb) and web-age information management (WAIM) joint conference on web and big data*, pages 285–293. Springer.