CINES: Explore Citation Network and Event Sequences for Citation Forecasting

Fang He
Penn State University
University Park, PA, USA
fxh35@psu.edu

Wang-Chien Lee
Penn State University
University Park, PA, USA
wlee@cse.psu.edu

Tao-Yang Fu
Penn State University
University Park, PA, USA
txf225@psu.edu

Zhen Lei
Penn State University
University Park, PA, USA
zxl26@psu.edu

ABSTRACT

Citations of scientific papers and patents reveal the knowledge flow and usually serve as the metric for evaluating their novelty and impacts in the field. Citation Forecasting thus has various applications in the real world. Existing works on citation forecasting typically exploit the sequential properties of citation events, without exploring the citation network. In this paper, we propose to explore both the citation network and the related citation event sequences which provide valuable information for future citation forecasting. We propose a novel *Citation Network and Event Sequence (CINES)* Model to encode signals in the citation network and related citation event sequences into various types of embeddings for decoding to the arrivals of future citations. Moreover, we propose a *temporal network attention* and three alternative designs of *bidirectional feature propagation* to aggregate the retrospective and prospective aspects of publications in the citation network, coupled with the citation event sequence embeddings learned by a *two-level attention mechanism* for the citation forecasting. We evaluate our models and baselines on both a U.S. patent dataset and a DBLP dataset. Experimental results show that our models outperform the state-of-the-art methods, i.e., RMTPP, CYAN-RNN, Intensity-RNN, and PC-RNN, reducing the forecasting error by 37.76% - 75.32%.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

Neural networks; Citation forecasting; Citation network

ACM Reference Format:
Fang He, Wang-Chien Lee, Tao-Yang Fu, and Zhen Lei. 2021. CINES: Explore Citation Network and Event Sequences for Citation Forecasting . In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3404835.3462903

1 INTRODUCTION

Scientific papers and patents reveal scientific discoveries and technological innovation. In addition to revealing the knowledge flows, citations of publications[1] have been widely used as evaluation metrics for assessing discoveries and innovations' novelty and impact [2, 8, 18], researcher competence [7], journal impacts [4], performance of research institutes [10], emerging new research and technologies [12, 13], and technology cycles, among others. Due to the wide applications of citations, researchers have seen growing interests in the task of *citation forecasting* on scientific publications to assess their long-term potential.

One approach to citation forecasting is to treat the arrivals of citations to a publication as a dynamic process of citation events affected by a number of factors and model the occurrence of these events mathematically.[2] As such, future citation events are predicted using the developed mathematical models. For example, Wang, Song, and Barabasi propose a WSB model by factoring in the total number of citations a paper received, its age and fitness (i.e., paper quality) in a manually tailored equation [19]. He et al. improve the WSB model to better predict atypical articles whose citations arrives in atypical patterns (e.g., articles that sleep with no citations for a while and then awaken to receive many citations) [6]. Some other studies model the series of citation events as a temporal point process using some pre-selected conditional intensity functions, e.g., the reinforced poison process [17, 21] and Hawkes process [14, 23, 26]. However, the aforementioned models do not achieve good performance due to the limited power of identified factors or strong assumptions unfit with the real-world scenarios.

Another more recent approach to citation forecasting involves deep learning. In particular, *recurrent neural network* (RNN) has been employed to emulate the temporal point process for citation forecasting, exploring the chained neural cells in RNN to capture the conditional intensity of previous citations [3, 9, 20, 22]. However, this approach results in numerical instability in training and a computational bottleneck in inference due to the complexity of density functions adopted [9]. Recently, Ji. et al. propose a sequence-to-sequence model, called *PC-RNN*, for patent citation forecast of a given *focal patent*, which employs an attention mechanism to capture the dependencies in time sequences of citations to the focal patent, citations to other patents of its inventors, and citations to other patents owned its assignees [9]. While some "indirect" citations are explored as additional signals to supplement direct citation events to the focal patent, this work does not consider different characteristics of individual patents, leading to significant information loss. Additionally, it does not exploit the rich information in the *publication citation network (PCN)*, that contains not only information and traits of a publication, but also "knowledge flow" and "impact" revealed by edges in the citation network.

[1] The term "publication" refers to the papers or patents in this paper.
[2] We refer to the happening of a citation in terms of its timing as *citation event*.

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462903

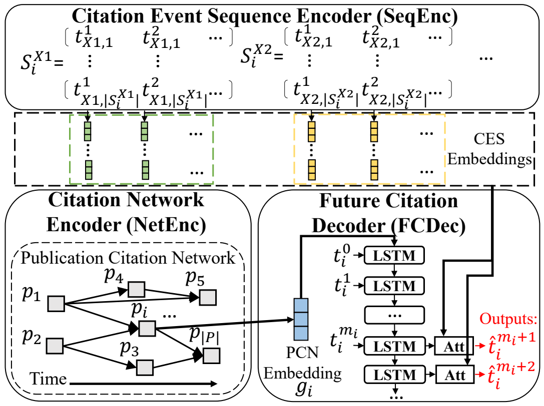

Figure 1: The CINES Model

In this work, we aim to exploit signals embedded in PCN and various sequences of citation events for citation forecasting. We argue that information in both PCN and sequences of citation events are useful for citation forecasting. The PCN, along with the content and metadata describing publications, provides the linkage among publications via citations, which can be viewed in *retrospective* and *prospective* aspects to capture knowledge flow and impact. In other words, a citation serves two roles with different implications. On one hand, the retrospective aspect highlights that the *references* documented in a focal publication signals the knowledge propagation from the references to it. On the other hand, the prospective aspect stresses that the *forward citations* of a publication signal its influence on subsequent innovations and in certain fields (i.e., its future impacts). In addition, PCN as a whole is a rich data repository that embeds temporally evolving technological trends and development. The position of the focal publication in such a network and its surrounding network structural information may bring useful signals for forecasting future citations. Thus, in this work, we aim to capture various information from PCN for each publication.

As the arrivals of future citations to a publication can be seen as a sequence of citation events (also called *citation event sequences*), it is intuitive that the past citation event sequences of a focal publication are useful for its future citation forecasting. Furthermore, we argue that in addition to the past citation event sequence associated with the focal publication, the citation event sequences associated with its *related publications*, including its references, similar publications of its authors (or its inventors) and so on, are valuable supplements. Thus, it's a good idea to complement the information captured in PCN with signals in *related citation event sequences* for future citation forecasting.

Designing a citation forecast framework to incorporate both information from PCN and related citation event sequences is challenging due to their distinctive representations, i.e., PCN is a graph while related citation event sequences are time-variant sequences. To meet the challenge, we propose a new neural network model, namely *Citation Network and Event Sequences (CINES)*, based on the encoder-decoder paradigm to forecast the arrivals of future citations for a focal publication. As shown in Figure 1, CINES consists of three components: i) a *Citation Network Encoder (NetEnc)*, ii) a *Citation Event Sequence Encoder (SeqEnc)*, and iii) a *Future Citation Decoder (FCDec)*. NetEnc is a graph neural network that encodes content, traits and network structural information in PCN into PCN embeddings for publications in PCN. The novelty in our design of NetEnc lies in the exploration of *bidirectional feature propagation* to capture both prospective and retrospective aspects of citations and the proposal of a *temporal network attention (TNA) mechanism* to capture the factor of the *temporal distance* in citations. On the other hand, to encode existing citation event sequences of related publications to the focal publication, SeqEnc exploits the *Long Short Term Memory (LSTM)*, a variant of recurrent neural networks known for processing sequential data, to encode various citation event sequences into corresponding sequences of *citation event sequence (CES)* embeddings. Note that to handle different categories of citation event sequences, SeqEnc learns multiple LSTMs, one for each category. Finally, FCDec takes the PCN embedding, various CES embedding sequences and the citation event sequence of the focal publication as the inputs to forecast the arrival of its future citations. As future citations arrive as a temporal sequence, FCDec exploits an LSTM for forecasting. Specifically, at each time step, FCDec explores a *two-level event/publication aware attention* mechanism to learn the attention weights for aggregation of various CES embedding sequences into one aggregated CES embedding, which represents the overall influence from related citation event sequences at that time step. Finally, both the aggregated CES embedding and the LSTM hidden state at that time step (which incorporates the PCN embedding and the historical citation event sequence of the focal publication) are fed into a fully-connected layer to predict the arrival time of the next future citation.

We conduct extensive experiments on two large-scale real-world publication datasets (i.e., U.S. patents and DBLP papers) to evaluate the proposed CINES model in comparison with several state-of-the-art models. The results show that CINES outperforms the baselines by reducing the Mean Absolute Error (MAE) by 37.76% and 75.32% on the patent and paper citation forecasting, respectively.

The major contributions made in this work are as follows.

- We propose to explore both the PCN and the related citation event sequences, which provide valuable information inherent to the dynamic pattern of citation events in complementary forms, for future citation forecasting.
- We propose *CINES*, a new end-to-end neural network model, for publication citation forecasting. The design of CINES is tailored for exploiting unique characteristics of information embedded in a large-scale temporal network and a variety of citation event sequences.
- We conduct extensive experiments on U.S. patents and DBLP papers to evaluate the proposed CINES against the state-of-the-arts in research on citation forecast. Experimental results show that CINES significantly outperforms the baselines by reducing the forecasting error (in terms of MAE) up to 75.32%. Ablation study finds that signals from PCN and related citation events sequences are complementary for citation forecasting, while the former is generally more important.

2 RELATED WORKS

The citation forecasting task on publications aims to forecast the arrival of future citations to a focal publication. Existing works generally fall into two categories: i) mathematically modeling citations to a focal publication as a dynamic process of citing events; and ii) learning machine learning models using historical citation data.

2.1 Mathematical Modeling

Early studies on citation forecasting model the arrival of citations as a phenomenon of citation events dominated by certain factors. They analyze the factors, manually formulate a mathematical model, and fit citation events to the model [6, 19, 25]. For example, by considering the total number of citations a publication received, the age of the publication (which affects the long-term decay of its novelty), and the fitness (i.e., how different it is from other publications), Wang, Song and Barabasi propose the WSB model to predict future citations to the publication [19]. He et al. extend the WSB model to handle two atypical types of publications, i.e., awakened articles (first sleep and then gradually gain the citations) and second-act articles (have two peaks of the citation arrivals) [6]. Alternatively, some studies model the arrivals of citing events as a *temporal point process* [3, 8, 14, 17, 21, 23, 26]. This line of studies differ in the conditional intensity functions used in the temporal point process (e.g., reinforced poison process in [17, 21] and Hawkes process [5, 14, 23, 26]) and priors used in the models. However, these mathematical models suffer from poor performance due to the limits of identified factors or the strong assumptions behind used conditional intensity functions, unfitting with real-world scenarios.

2.2 Machine Learning

Instead of manually formulating mathematical models, recent studies on citation forecasting explore machine learning methods to learn predictive models from the historical citation data [1, 3, 15, 20, 22, 24]. Some works extract various types of features, e.g., author-wise attributes, paper-specific and venue-centric features, from publications to train regression models for citation forecasting [1, 24].

Owing to the limits caused by manually designed conditional intensity functions in temporal point processes, *recurrent neural networks (RNNs)* are exploited to learn flexible exponential functions with its cells [3, 20, 22]. However, these RNN-based methods face numerical instability in training due to their designs of the exponential functions [9]. In addition, those works face data sparsity issues. For new publications, which have limited or no citations, it is hard for these models to predict future citations if the models overfit on historical citing events. To address these issues in patent data, a sequence-to-sequence model, called *PC-RNN*, is proposed to capture dependencies among three types of citing event sequences, including existing citations to the focal patent, citations to the patents of its authors and citations to the patents of its assignee [9].

In this work, we propose to exploit publication citation networks and citation event sequences, which have not been explored together in previous works for citation forecasting. Moreover, the designs of components in our model, CINES, are new and tailored for unique characteristics of information embedded the large-scale publication citation networks and various citation event sequences.

3 PROBLEM FORMULATION

In this section, we formally define the notions of *publication citation network (PCN)*, *citation event sequences* and some related terms. Accordingly, we formulate the *publication citation forecasting* problem.

Definition 1. Publication Citation Network (PCN). A *publication citation network* is a directed graph $D = (P, E, \phi)$, where P is the node set and E is the edge set. Each node $p_i \in P$ denotes a publication; each edge $e_{ij} \in E$ denotes a citation, i.e., p_i cites p_j; and $\phi : P \rightarrow \mathcal{F}$ is a feature extraction function that maps each node to $\mathcal{F}$, the set of its features, such as publication date and title.

Definition 2. Reference Set and **Forward Citation Set**. Given a PCN $D = (P, E, \phi)$ and a publication p_i, we refer to the set of references cited by p_i as the *reference set* of p_i, i.e., $R(p_i) = \{p_j | p_j \in P, e_{ij} \in E\}$. For simplicity, we use R_i to denote $R(p_i)$. Similarly, we refer to the set of publications that cite p_i as the *forward citation set* (or *citation set* in short) of p_i, i.e., $C(p_i) = \{p_j | p_j \in P, e_{ji} \in E\}$ and use C_i to denote $C(p_i)$.

The publications in the citation set C_i of a publication p_i contain the existing citations to p_i, and thus may provide useful information for the arrival of its future citations. We are particularly interested in exploiting the temporal signal associated with the happenings of those citations, referred as *citation events*, as defined below.

Definition 3. Citation Event. Given a publication p_j which cites p_i, i.e., $p_j \in C_i$, the *citation event* of the citation e_{ji}) is the time distance between the publishing of p_i and p_j, i.e., $|Date(p_j) - Date(p_i)|$ where $Date(p)$ is the publishing date of p.

By sorting the citation events of forward citations to a publication p_i in ascending order, we have the *citation event sequence* of p_i, as defined below.

Definition 4. Citation Event Sequence. Given the citation set C_i, the *citation event sequence* of p_i is the sorted set of existing citation events $S_i = \{t_i^n | 1 \leq n \leq m$ where t_i^n is the n^{th} citation event to p_i and m is the total number of existing citations to p_i.$\}$.

In addition to the signal of direct citations to the focal publication, we argue that citation event sequences to some *related publications*, especially those that exhibit correlated signals, may be useful for citation forecasting. In this work, we identify several categories of related publications, depending on the nature of publications, i.e., patents or papers. For patents, the related publications of a focal patent p_i include i) patents referenced by p_i, ii) patents citing p_i, iii) similar patents by the inventors of p_i, iv) similar patents owned by the assignee of p_i. Similarly, for papers, we explore related publications in (i) and (ii) above but replace inventors in (iii) and assignee in (iv) by authors and affiliations, respectively. Accordingly, for p_i, there are multiple related publications in each category X, and each related publication has a corresponding *related citation event sequence* as defined below.

Definition 5. Related Citation Event Sequences. Let X_i denote the set of related publications to p_i in category X, e.g., when $X = R$, the set of related publications R_i is the reference set of p_i. Given the kth related publication in X_i, its *related citation event sequence* is represented as $\{t_{X,k}^1, t_{X,k}^2, ..., t_{X,k}^{m_{X,k}}\}$, where $m_{X,k}$ is the total number of publications citing this kth related publication. Thus, all the related citation event sequences corresponding to category X form a set S_i^X as follows.

$$S_i^X = \{\{t_{X,1}^1, t_{X,1}^2, ..., t_{X,1}^{m_{X,1}}\}, ..., \{t_{X,|S_i^X|}^1, t_{X,|S_i^X|}^2, ..., t_{X,|S_i^X|}^{m_{X,|S_i^X|}}\}\} \quad (1)$$

Notice that the content and metadata of publications, the citation network, the citation event sequence of publications, and related citation event sequences of publications in various categories can all be extracted from a dataset of scientific publications. Thus, we formulate the problem of *learning for citation forecasting* as follows.

Definition 6. Learning for Citation Forecasting Problem. Given a dataset of scientific publications, the goal is to learn a citation forecasting model M that maps i) a focal publication p_i, along with ii) its existing citation events $S_i = \{t_i^1, t_i^2, ..., t_i^{m_i}\}$, iii) the citation network $D = (P, E, \phi)$, and iv) various related citation event sequences $\{S_i^X | X \text{ is a category of related publication.}\}$ to the future citation events of p_i, i.e., $\{t_i^{m_i+1}, t_i^{m_i+2}, ...\}$.

The problem of learning for citation forecasting faces two challenges: i) data sparsity due to the limited number of existing citation events to the focal publication; and ii) rich yet complex signals hidden in various sequences of citation events as well as content and relationships among publications form the citation network. To address the data sparsity issue, we explore both the citation network and related citation event sequences to supplement the existing citation events of the focal publication for citation forecasting. To distill the complex signals, we analyze the characteristics of the citation network and citation event sequences, aiming to design an effective model for citation forecasting. As the citation network has a temporal graph structure in which citations (i.e., directional edges in the graph) have both retrospective and prospective aspects on how publications are influencing or influenced by their neighbors. Thus, an idea is to explore *bidirectional feature propagation* in a temporal graph attention network to capture the aforementioned roles of citations in the citation network. On the other hand, to capture the dependencies in various related citation event sequences, our idea is to exploit recurrent neural networks to encode signals in related citation event sequences by a two-level attention mechanism, which at the *event* level first aggregates weighted signals amongst citation events for each related publication, and then at the *publication* level further aggregates the aggregated signals of related publications in the same category according to the similarity between the related and focal publications. The proposed CINES model is developed based on these ideas.

4 DESIGN OF THE CINES MODEL

In this section, we first introduce the model architecture and then detail our design on components of the CINES model.

4.1 Model Architecture

To explore the supplementary signals from i) the *publication citation network (PCN)* and ii) *related citation event sequences* of publications, we follow the encoder-decoder paradigm of neural network architecture in the design of CINES, which consists of three components (see Figure 1): 1) *Citation Network Encoder (NetEnc)* to encode information in PCN, i.e., content and traits of publications and their surrounding network structures; 2) *Citation Event Sequence Encoder (SeqEnc)* to encode temporal dependencies in citation event sequences of various related publications; and 3) *Future Citation Decoder (FCDec)* to predict future citation events of a given focal publication by taking its existing citation events and decode the encoding (in form of embeddings) learned from NetEnc and SeqEnd.

Specifically, taking the focal publication p_i as an example, NetEnc encodes the content and traits of p_i and the network structural information of its surroundings into its *PCN embedding*, denoted by g_i. Meanwhile, SeqEnc encodes related citation event sequences of p_i, corresponding to various categories (denoted by $X1, X2, ...$) into a series of *CES embeddings*, which consists of multiple *embedding sequences* for each category (shown in different colors in Figure 1), while each embedding sequence is corresponding to a related publication and each individual CES embedding is corresponding to a citation event. Finally, FCDec takes the existing (direct) citation events of p_i, i.e., $S_i = \{t_i^1, ..., t_i^{m_i}\}$ where m_i is the number of forward citations to p_i received so far, as input, while decoding the PCN embedding g_i and all related CES embeddings to predict the future citation events of p_i (i.e., $\{t_i^{m_i+1}, t_i^{m_i+2}, etc.\}$).

4.2 Citation Network Encoder

In this section, we detail our design of the Citation Network Encoder (NetEnc) which aims to encode content and traits and the surrounding structural information of publications in PCN. Note that the topics covered in a publication, its location in the PCN, and indicators of knowledge flow and future impact are all potential factors affecting its value which in turn may be related to its future citations. Thus, an effective encoder should be able to capture those signals for publications in PCN.

Due to the nature of PCN, i.e., it is a graph, we design NetEnc as a graph neural network (GNN) model to exploit its underlying ideas of feature transformation and propagation in the network. However, as mentioned previously, there exist retrospective and prospective aspects in citations due to their roles and implications in knowledge propagation and potential impact among publications. Additionally, PCN is a temporal network where the time distance between cited and citing publications obviously contributes its uniqueness. Conventional GNN models, based on our best knowledge, do not take the special characteristics of PCN into consideration and thus may not work well. To address these issues, our design of NetEnc explores three different ideas, namely, *Merge, Alternate* and *Cross*, to facilitate *bidirectional feature propagation* in order to accommodate both prospective and retrospective aspects of citations, while proposing a temporal attention mechanism, called *temporal network attention (TNA)*, to address the temporal property of citations. In the following, we first present TNA and then discuss alternative strategies to extend TNA for bidirectional feature propagation.

4.2.1 **Temporal Network Attention.** We discuss the proposed *temporal network attention (TNA)* under the context of "forward" feature propagation (i.e., the knowledge diffusion) via citation edges in PCN, as the idea also applies for "backward" feature propagation. Our design has multiple layers, each of which models the feature transition from a node's 1-hop neighbors (i.e., its references) to the node. As such, in each layer, each node collects the features within the neighborhood of its 1-hop references. With L layers, a node/publication collects features propagated within its L-hop neighborhood of references. Intuitively, the references of a publication may have different influences on it, and thus we propose to model the different impacts of the references in our design. We argue that feature transitions along citation edges are different under the following factors: i) the contents and traits being transmitted, i.e., the transitions from two references with different topics are different; and ii) its temporal distance, i.e., the transitions from a reference published a long time ago and another reference published recently are different. Thus, they should be treated differently. Accordingly, for a node to aggregate information from its references at each layer, TNA performs feature transformation based on temporal distances of reference edges to carry out a weighted aggregation by a graph attention.

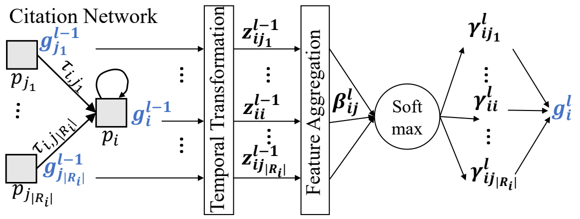

Figure 2: Layer l of the Temporal Network Attention

Figure 2 shows the l-th layer of TNA, which models the feature propagation process from references of a publication p_i to generate the layer-l embedding g_i^l of p_i.[3] As shown, the l-th layer transforms the layer-(l-1) embedding of p_i and those of its references into intermediate embeddings and aggregates them to generate the layer-l embedding of p_i. More specifically, for each reference p_j (j = j_1, j_2, ... $j_{|R_i|}$), where the temporal distance between p_i and p_j is $\tau_{i,j}$, an intermediate embedding $z_{ij}^{l-1} = W^l_{f_{step}(\tau_{i,j})} \cdot g_j^{l-1}$ (where $W^l_{f_{step}(\tau_{i,j})}$ is a transformation matrix tailored by the temporal distance τ) is generated. Note that the transformation matrix W is predetermined by a step function f_{step} which designates a number of temporal distance ranges by some time steps, e.g., W_1 for 0-1 year, W_2 for 1-2 year, and so on. Additionally, we use a threshold δ_τ to designate "long" temporal distance and use a special W_{δ_τ} for such citations.

Next, we aggregate the intermediate embeddings z_{ij}^{l-1} (where $p_j \in R_i \cup \{p_i\}$) based on the *importance* of p_j to p_i as follows.

$$\beta_{i,j}^l = LeakyReLU(\vec{\alpha} \cdot (z_{ij}^{l-1} || z_{ii}^{l-1})) \tag{2}$$

where || denotes the concatenation of the two embeddings and $\vec{\alpha}$ is a learned latent vector to help calculate the unnormalized attention score (i.e., importance).

Finally, we normalize the attention scores with a softmax function as follows.

$$\gamma_{i,j}^l = \frac{exp(\beta_{i,j}^l)}{\sum_{p_j \in R_i \cup \{p_i\}} exp(\beta_{i,j}^l)} \tag{3}$$

The normalized scores serve as the weights to aggregate the layer-(l-1) embeddings of the references and p_i to generate the layer-l embedding of p_i as follows.

$$g_i^l = \sigma\Big(\sum_{p_j \in R_i \cup \{p_i\}} \gamma_{i,j}^l W^l_{f_{step}(\tau_{i,j})} \cdot g_j^{l-1}\Big) \tag{4}$$

where σ is a non-linear activation function.

4.2.2 **Bidirectional Feature Propagation.** Owing to the retrospective and prospective aspects of citations, the PCN can be logically treated as two PCNs with opposite directions of feature propagation. As TNA discussed earlier works for only one direction, we extend it for bidirectional feature propagation with three alternative designs, namely, *Merge, Alternate* and *Cross*.

Merge: As mentioned, we may treat PCN as two separate networks (where the citation edges have opposite directions) to capture the knowledge flow from references and future impacts from forward citations independently. With this simple design, the PCN embedding of p_i, i.e., g_i, can be obtained by concatenating the two embeddings g_i^L and g'^L_i (learned from backward feature propagation).

[3]Note that the initial embedding of p_i, denoted by g_i^0, is pre-learned on its textual content such as publication title and description by a representation learning model, e.g., Word2Vec or Doc2Vec[11, 16].

Alternate: A potential weakness in the design of Merge is that the embeddings from PCN and Reverse PCN are obtained *independently* and simply concatenated, which does not reflect the potential interplay between retrospective and prospective roles of citations. Thus, this design aims to fuse both aspects in one embedding by taking turns to facilitate feature propagation in different directions. More specifically, let the embedding of p_i generated in the $(2l-1)$-th layer (odd layer) be g_i^{2l-1} and that generated in $2l$-th layer (even layer) be g_i^{2l}. As feature propagation in a layer is based on the embeddings obtained in the previous layer (in which feature propagation is proceeded in opposite directions of the current layer), the information in both retrospective and prospective aspects is fused into the embedding through multiple iterations of the alternative layers.

Cross: This design integrates both ideas of Merge and Alternate to learns two embeddings for each publication. Instead of learning them independently like Merge, they are learned dependently like Alternate. Like Merge, we use two networks, i.e., PCN and Reverse PCN, to carry out forward feature propagation and backward feature propagation, respectively. However, after the embeddings g_i and g_i' are generated, they are *swapped* before being used in the next layer. Therefore, the final embeddings are generated exactly like being learned by Alternate, except that their initial directions of feature flow are opposite.

Later in Section 5, we experimentally compare these three extensions on TNA to decide the NetEnc in our CINES model.

4.3 Citation Event Sequence Encoder

In this section, we introduce our design of *Citation Event Sequence Encoder (SeqEnc)*, which captures the dependencies among the existing citation event sequences of various publications related to the focal publication p_i. As discussed earlier, a related publication belongs to some category X.[4] Moreover, for p_i, there are multiple related publications in each category and correspondingly each related publication has a related citation event sequence.

For a publication p_i, we group its related citation event sequences of category X into a set S_i^X. For the k-th citation event sequence in S_i^X, let its number of citation events be $m_{X,k}$, we denote the s-th citation event in the sequence as $t^s_{X,k}$, which is the temporal distance between the publication corresponding to the s-th citation event and p_i. Figure 3 illustrates our design of SeqEnc. To encode the

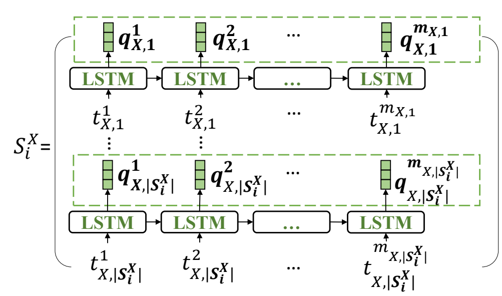

Figure 3: Design of the Citation Event Sequence Encoder

related citation event sequences, we exploit the Long Short Term Memory (LSTM) model, to embed each citation event sequence into a series of embeddings. Instead of using a general LSTM for

[4]Please refer to Section 3 for categories of related publications considered.

all citation event sequences, *we use one LSTM for each category of citation event sequences to better differentiate their relevance.* More specifically, for the k-th sequence in S_i^X, i.e., $\{t^1_{X,k}, t^2_{X,k}, ..., t^{m_{X,k}}_{X,k}\}$, we feed the citation events one by one to the LSTM model for X which is used for encode all the citation event sequences of category X. At each time step, a citation event is fed to the LSTM, which encodes all the citation events (and their conditional dependency) before the current time step into an embedding. For example (as shown in Figure 3), we generate $q^1_{X,k}$ as the embedding for $t^1_{X,k}$, $q^2_{X,k}$ as the embedding for $\{t^1_{X,k}, t^2_{X,k}\}$, and so on. Finally, the LSTM outputs a series of *citation event sequence (CES) embeddings*, i.e., $\{q^s_{X,k} | 1 \le k \le |S_i^X|, 1 \le s \le m_{X,k}\}$, as the representation of the citation event sequence. Note that we encode a citation event sequence to a series of CES embeddings instead of only one single embedding for the whole sequence in order to better capture the different and increasing impacts brought by occurrence of individual citation events. As such, the CES embeddings for all categories of related publications, i.e., $\bigcup_X \{q^s_{X,k} | 1 \le k \le |S_i^X|, 1 \le s \le m_{X,k}$ for $p_i\}$, are passed to the Future Citation Decoder for decoding and exploited for citation forecasting.

4.4 Future Citation Decoder

Here we discuss the design of *Future Citation Encoder (FCDec)*, which takes the existing citation event sequence of a focal publication, its PCN embedding and the CES embeddings of its related citation event sequences as input to predict its future citation events.

Owing to the sequential nature of the existing and future citation events, FCDec exploits an LSTM model to capture the dependencies among the previous citations and the upcoming citations for forecasting. To exploit signals from PCN, FCDec intuitively takes the PCN embedding of the focal publication as the initial hidden state to the LSTM. On the other hand, FCDec aims to exploit the many CES embeddings captured in the steps of citation event sequences and thus needs to find a way to effectively comb the signals.

Notice that the overall CES embeddings generated by SeqEnc consist of groups of CES embedding sequences which in turn consist of CES embeddings corresponding to time steps in their sequences. In other words, the CES embeddings contains signals related to citation events in various categories of related publications. To decode useful information from them, we propose a two-level *event/publication-aware attention* mechanism, which systematically aggregates signals captured in event and publication levels. At the event level, we capture the signals in different time steps of citation events by learning attention weights for citation events to aggregate the individual CES embeddings into embeddings of their corresponding citation event sequences. At the publication level, we further capture the aggregated signals in related citation event sequences of the same category X into a *categorical CES embedding* for X by learning attention weights for these sequences.

Figure 4 illustrates our design of FCDec, where g_i, i.e., the PCN embedding of the focal publication p_i, is fed as the initial hidden state to the LSTM. At time step n, the LSTM cell takes the hidden state from cell $n-1$ and the citation event t_i^n to output an intermediate embedding h_i^n which summarizes all previous citation events and information in the PCN embedding g_i. The embedding h_i^n is fed into a fully connected (FC) network, along with the concatenated

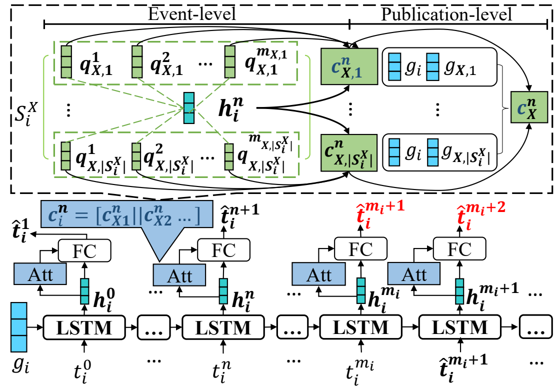

Figure 4: Design of the Future Citation Decoder

categorical CES embeddings obtained by the two-level attention mechanism to predict the next citation event $\hat{t}_i^{n+1}$. As such, after FCDec scans through existing citation events $t_i^0, t_i^1, ...t_i^{m_i}$, it starts to forecast future citation events $t_i^{m_i+1}, t_i^{m_i+2}$, and so on.

As illustrated in the dashed box on top of Figure 4, the categorical CES embedding of category X, denoted as c_X^n, is obtained by the proposed event/publication attention mechanism. For the event-level attention which aggregates the CES embeddings generated from a citation event sequence by SeqEnc, e.g., $q^1_{X,k}, q^2_{X,k}, ..., q^{m_{X,k}}_{X,k}$, the attention score representing the importance of the s-th citation event $q^s_{X,k}$ to the time step n of forecasting is defined as follows.

$$\beta^{n,s}_{X,k} = f^X_{evt}(q^s_{X,k}, h_i^n) = V^X_{evt} tanh(W^X_{evt}(q^s_{X,k} || h_i^n)) \quad (5)$$

where f^X_{evt} is the attention function in the event-level for category X with a learnable vector V^X_{evt} and a learnable matrix W^X_{evt}.

Then the attention scores are normalized with a *softmax* function as follows.

$$\gamma^{n,s}_{X,k} = \frac{exp(\beta^{n,s}_{X,k})}{\sum_{s=1,2,...,m_{X,k}} exp(\beta^{n,s}_{X,k})} \quad (6)$$

The normalized attention scores are used as weights to aggregate their corresponding CES embeddings in the citation event sequence of the k-th related publication, resulting in $c^n_{X,k}$ as follows.

$$c^n_{X,k} = \sum_{s=1,2,...,m_{X,k}} \gamma^{n,s}_{X,k} q^s_{X,k} \quad (7)$$

In turn, the publication-level attention further aggregates the embeddings of different related publications in category X generated by the event-level attention, i.e., $c^n_{X,1}, ..., c^n_{X,|S_i^X|}$. Considering that the content and traits of a related publication may be correlated to the citation event sequence of the focal publication (i.e., related publications with content similar with the focal publication may have more similar citation event sequence with that of the focal publication), we use the publication-level attention scores to weigh the importance of the k-th related publication at the time step n by multiplying the attention score derived from the CES embedding $c^n_{X,k}$ and the PCN embedding similarity derived from $g_{X,k}$ and g_i. Following this idea, we exploit a function f_g to capture the importance of the k-th related publication by estimating the PCN embedding similarity between the related publication (i.e.,

$g_{X,k}$) and the focal publication (i.e., g_i) as follows.

$$f_g(g_{X,k}, g_i) = V^g tanh(W^g(g_{X,k}||g_i)) \tag{8}$$

where V^g and W^g are learnbale parameters. Similarly, we exploit a function f^X_{pub} to capture the importance brought from the citation event sequence based on the CES embedding as follows.

$$f^X_{pub}(c^n_{X,k}, h^n_i) = V^X_{pub} tanh(W^X_{pub}(c^n_{X,k}||h^n_i)) \tag{9}$$

Then we multiply f_g and f^X_{pub} to to generate the attention score, $\beta^{n,k}_X$, which represents the importance of the k-th related publication in category X at the time point n of forecasting as follows.

$$\begin{aligned}\beta^{n,k}_X =& f_g(g_{X,k}, g_i) \cdot f^X_{pub}(c^n_{X,k}, h^n_i)\\ =& V^g tanh(W^g(g_{X,k}||g_i)) \cdot V^X_{pub} tanh(W^X_{pub}(c^n_{X,k}||h^n_i))\end{aligned} \tag{10}$$

Finally, we normalize the attention scores and exploit them to weight the CES embeddings of related publications to generate the categorical CES embedding of category X, i.e., c^n_X, as follow,

$$\gamma^{n,k}_X = \frac{exp(\beta^{n,k}_X)}{\sum_{k=1,2,...,|S^X_i|} exp(\beta^{n,k}_X)}, \quad c^n_X = \sum_{k=1,2,...,S^X_i} \gamma^{n,k}_X c^n_{X,k} \tag{11}$$

Here c^n_X represents the total impacts of all citation event sequences in the category X at time step n of forecasting. The categorical CES embeddings of all categories, i.e., c^n_{X1}, c^n_{X2}, ..., etc., are concatenated as c^n_i to fed to a fully connected (FC) network, together with the embedding h^n_i to predict $\hat{t}^{n+1}_i$. The loss function $\mathcal{L}$ of the CINES model is the Mean Absolute Error between the predicted arrivals and the ground truth, i.e., $\mathcal{L} = \sum_{i=1,...,|P|} \sum_{n=1,...,m_i} |\hat{t}^n_i - t^n_i|$.

5 EXPERIMENTS

In this section, we empirically evaluate the CINES model using two real-world publication datasets against several baseline methods.

5.1 Experiment Setup

We collect two datasets for evaluation: i) a U.S. Patent dataset publically accessible from United States Publication and Trademark Office (USPTO) website, and ii) a DBLP paper dataset collected from the DBLP website. For each dataset, we build a citation network using the publications between a starting year (denoted as *Start*) and the year of forecasting (denoted as *Forecast*), to predict the future citation events till the ending year (denoted as *End*). We show the statistics of the datasets in Table 1. For each patent, we collect its id, title, description, publication date, the list of inventors, the assignee, the U.S. patent class and a list of its references. For each patent p_i, we represent the publication date (denoted by year/month/day) and U.S. patent class in one-hot embedding, apply Word2vec to transform the title and description into Word embedding, and finally concatenate them to serve as the initial embedding g^0_i in NetEnc. In addition, we prepare the related citation event sequences for each patent of different categories.[5] As the number of related publications in some categories (i.e., patents by the inventors of p_i and patents owned by the assignee of p_i) may be very large, we select at most 10 publications that have the most similar content with the focal publication, measured by the distance between their Word embeddings. The DBLP paper dataset is processed similarly for the categories of related publications by the authors and affiliation of the focal paper.

[5]Please refer to Section 3 for the different categories for papers and patents respectively.

Table 1: Statistic of the Datasets

Dataset	No. of Publications	*Start*	*Forecast*	*End*
U.S. Patents	980,849	1980	1999	2011
DBLP Papers	4,107,325	1995	2010	2019

For evaluation, we randomly split each dataset into three sets: 80% as the training set, 10% as the validation set and 10% as the testing set. To train the CINES model, we feed the citation network (by the end of the *Forecast* year) and the related citation event sequences of each focal publication in the training set, and train the model by minimizing the errors in predicting existing citation events one by one, i.e., predict $\hat{t}^n_i$ based on $t^0_i, t^1_i, ..., t^{n-1}_i$ for $n = 1, ..., m_i$. For training, we set batch size as 8, and exploit *Adam* optimizer with initial learning rate set as 0.0001. In the testing phase, we evaluate the model by measuring the forecasting error of the future citations, i.e., $t^{m_i+1}_i, t^{m_i+2}_i, ...$, till the *End* year. We measure the model performance with the *Mean Absolute Value (MAE)* between the predicted future citations and the ground truth.

5.2 Baselines for Comparison

The following are the baselines compared in our evaluation.

RMTPP [3] models a citation event sequence as a point process to predict both the next citation event and the type of point in the citation event sequence. Since RMTPP only models one sequence, in the experiments we only feed the existing citation events of the focal publication as the input to RMTPP for prediction.

CYAN-RNN [20] is an intensity-based RNN model which forecasts the time and user of the next resharing behavior in social media datasets. By forcing the k-th output to be the $(k+1)$-th input, this model can be regarded as a sequence generator to be used for publication citation prediction. In the experiment, we feed the historical citation event sequence to CYAN-RNN for prediction.

Intensity-RNN [22] takes an event arrival sequence and an evenly distributed background time series as inputs to predict the arrival of the next event. In our experiment, we feed the historical citation event sequence of the focal publication with a simulated evenly distributed background time series (as [22] did) for prediction.

PC-RNN [9] employs an attention-of-attention mechanism to capture the dependencies among existing citations, inventor's citations and assignee's citations for citation forecasting. In our experiment, the three sequences generated based on the proposed pre-processing method [9] are fed to PC-RNN for prediction.

5.3 Parameter Tuning

We tune the parameters in CINES (with different bidirectional feature propagation designs, i.e., Merge, Alternate, and Cross) and show the results on the patent dataset in Figure 5.[6] By increasing *the layer number L* in NetEnc from 1 to 5, Figure 5(a) shows that better performance is achieved with more layers and it converges when L= 3 (thus chosen as default). For *the dimensionality of the PCN embedding* d_g, Figure 5(b) shows that 128 is the best (i.e., a higher dimensionality does not help) and thus chosen as the default of d_g. For *the temporal distance unit* u_τ,[7] Figure 5(c) shows that

[6]Parameter tuning for papers show similar results but not shown due to space limit.
[7]u_τ refers to the unit range of temporal distance in f_{step}, i.e., the same transformation matrix is used for temporal distance within certain ranges specified by u_τ.

CINES performs the best with u_τ= 1 year, and it gets worse with a larger τ_u, which leads to fewer ranges of temporal distance and thus fewer transformation matrices used, i.e., the model does not work as well when the same matrix is used on a range over 1 year. Finally, for the temporal distance threshold τ_δ in TNA, which specifies *long* temporal distances, Figure 5(d) shows that CINES performs better with a larger threshold. We set δ_τ as 15 years.

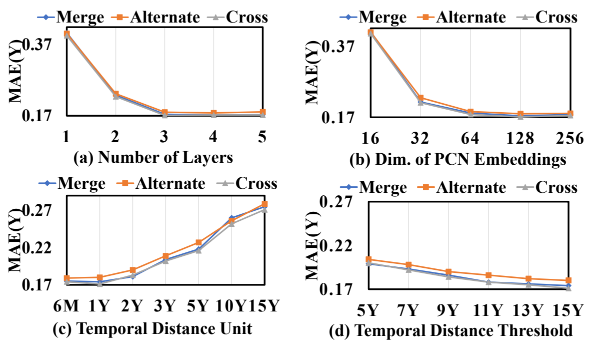

Figure 5: Parameter Tuning in CINES

Notice that the transformation matrices W's in NetEnc could be shared in different layers. We compare three strategies: i) W-Unique - different transformation matrices for different layers; ii) W-Shared - same transformation matrix for all layers; and iii) W-Direct - different transformation matrix for direct references and indirect references (i.e., references of references), respectively. Results in Table 2 show that W-Unique works the best, suggesting the feature transformation at different layers should be treated differently.

Table 2: Evaluation of the Transformation Matrices Settings

Model	W-Unique	W-Direct	W-Shared
CINES-Merge	0.174	0.180	0.185
CINES-Alternate	0.179	0.189	0.201
CINES-Cross	0.171	0.174	0.182

5.4 Ablation Study

In this section, we perform an ablation study to assess the impacts of various information (signals) in CINES on its performance. We consider the following variants of the CINES model by reducing information or replacing PCN and CES embeddings used in FCDec for forecasting. i) `Complete`: the complete version of CINES; ii) `NoNet`: not taking PCN embedding as input to FCDec; iii), `NoRef`: one-directional feature propagation in NetEnc, not from referenced publications to the citing publications; iv) `NoCit`, one-directional feature propagation in NetEnc, not from citing publications to referenced publications; v) `NoSeq`: not taking CES embeddings as input to FCDec; vi) `RawCon`: taking pre-trained Word embedding instead of PCN embedding as input to FCDec; vi) `NoPubAt`: replacing the publication-level attention in FCDec by averaging the embeddings of related publications as the categorical CES embedding. (i.e., remove f_g from Eq (10); and viii) `NoSplit`: use one aggregated citation event sequence for all related publications in a category instead of multiple sequences, i.e., one for each related publications, for each category as input to FCDec.

The blue bars in Figure 6 show the result of ablation study on patents. We observe that `Complete` outperforms all variants, reducing the MAE on patents by 6.55% (0.171 v.s. 0.183) to 34.73%

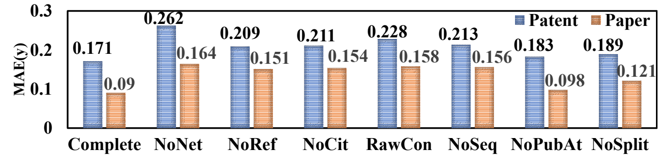

Figure 6: Ablation Study of CINES on Patents and Papers

(0.171 v.s. 0.262). Between `NoNet` and `NoSeq`, we observe the former has more performance deterioration against `Complete` than the latter, indicating that the PCN embedding is more important than CES embeddings in citation forecasting. By comparing `NoNet` and `RawCon`, we observe that capturing the content and traits of the focal publication into a Word embedding, while not as good as the PCN embedding, is much better than not using it. In addition, we observe that `NoRef` has 22.2% more MAE than `Complete` and `NoCit` has MAE by 23.4%, validates our idea for exploring bidirectional feature transmission to capture both retrospective and prospective aspects of information flow on citations.

Between `Complete` and `NoSplit`, the former reduces 9.52% in MAE, suggesting that exploring signals in citation event sequences for individual related publications in the same category works better than using one mixed long citation event sequence aggregated from all related publications in the category, as the latter may inject noisy and misleading signals. Finally, `Complete` outperforms `NoPubAt` by 6.55% in MAE, validating our idea that measuring the patent similarity in the publication-level attention helps CINES to capture signals from different related patents and aggregate the CES embeddings better.

The results of the ablation study on papers dataset are also shown in Figure 6 as the orange bars. Comparing the blue bars and orange bars, we observe that for the patent citation forecasting, `NoNet` results in 53.22% more MAE than the `Complete` while `NoSeq` results in 24.56% more MAE than the `Complete`, which suggests that the citation network plays a more important role in patent citation forecasting. As for the paper citation forecasting, we observe that `NoNet` results in 82.22% more MAE than the `Complete`, while `NoSeq` results in 73.33% more MAE than the `Complete`, which suggests the same conclusion (but their importance are closer).

5.5 Comparison against Baselines

We evaluate CINES against several baselines. Figure 7 shows that the best of CINES variants in bidirectional feature transformation (denoted by Merge, Alternate and Cross) significantly reduces MAE from the best baseline by 40.21% for patents and 46.43% for papers. CINES-Cross performs the best among CINES variants. Between CINES-Cross and CINES-Merge, the former captures more information than the latter as CINES-Cross combines two CINE-Alternate models with initial feature propagation in opposite directions, while CINES-Merge combines two independent feature propagation.

Among the baselines, RMTPP and CYAN-RNN simply simulate the intensity function with the exponential function of the hidden

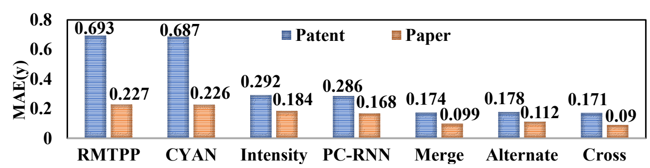

Figure 7: Performance Evaluation against Baselines

states in the RNN. Instead, Intensity-RNN learns the RNNs with a more flexible structure and higher capability, and thus outperforms those two models (e.g., 0.292 MAE by Intensity-RNN v.s. 0.693 MAE by RMTPP and 0.687 MAE by CYAN-RNN). Owing to the extra information extracted from related publications of inventors, assignee and the power of attention mechanism explored, PC-RNN outperforms Intensity-RNN slightly.

Again, CINES significantly outperforms PC-RNN (the winner amongst all the baselines). CINES's superior performance is due to the following reasons: i) CINES explores strong signals from PCN by designing a novel NetEnc to effectively capturing the content and traits of publications as well as the network structural information of the PCN; ii) CINES deals with related citation event sequences in an effective way by designing a well-tailored SeqEnc that distinguishes and captures temporal dependencies of citation events in individual related publications instead of mixing citation event sequences of related patents in the same category into one long sequence (as PC-RNN does); iii) CINES make effective citation forecasting by designing an FCDec that exploits the strength of an LSTM to decode signals from existing citation events and the PCN embedding and explore a two-level attention mechanism to distill important signals from the CES embeddings.

5.6 Sensitivity Tests

In this section, we conduct a number of sensitivity tests to observe how the CINES model performance is affected by various factors, including i) the number of references in the focal publication; ii) the number of forward citations to the focal publication; and iii) the publication year of the focal publication.

First, we plot in Figure 8 the MAE of citation forecasting on focal publications with the various number of references and forward citations, respectively. As shown, the MAE is quickly lowered when the number of references/citations in a publication increases from 0 to 4, and the performance improvement is reduced after then. This observation shows that both references and citations provide important information for citation forecasting.

In Figure 9, we plot MAE of citation forecasting on patents published in various years to evaluate the impact of publication year.[8] The blue curve, showing the forecasting errors made by the Complete CINES model on patents published from year 1980 to 1999, first drops and then rises. A hypothesized explanation is that the publications published in early years have fewer references than those published in later years (as references published before 1980 are cleared from the dataset due to the lack of content information) and similarly, the publications published in later years tend to be cited less often than those published earlier. To verify this hypothesis, we remove the signals of references and forward citations from CINES separately and plot their MAEs by orange and green curves, denoted by RemoveRef and RemoveCit, respectively. As shown, the MAE of RemoveRef increases as the number of citations tends to decreases for later publications, while the MAE of RemoveCit decreases as the number of references tends to increases for later publications. Moreover, the gap between RemoveRef and Complete increases over years while the gap between RemoveCit and Complete decreases, which supports our explanation.

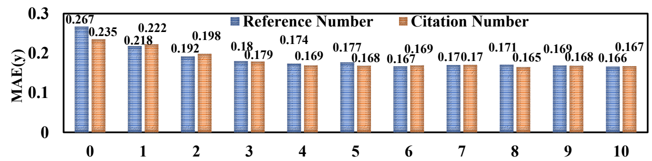

Figure 8: Evaluation of References and Citations Numbers

5.7 Summary of Experimental Findings

Here we summarize our main observations. First, capturing the content and traits of publications as well as the structural information of the citation network helps citation forecasting. Second, differentiating fine-grained temporal distances in Temporal Network Attention helps to distinguish the impacts of references corresponding to various temporal distances, which improves citation forecasting. Third, related citation event sequences, with an event/publication-aware attention to capture different influence of related citation events to the focal publication based on publication content and traits, are complementary to citation network in citation forecasting. Finally, we observe a trend of the forecasting errors (first drop and then arise) made by the CINES model for publications published over the years in our datasets. This is because that publications in early years are missing some references and those in later years are in lack of citations. Generally speaking the CINES model performs best for publications with rich references and citations.

6 CONCLUSION

We propose the CINES model for future citation forecasting by exploring both the publication citation network (PCN) and the citation event sequences, where the ideas of temporal network attention and bidirectional feature propagation are explored for the PCN and a two-level event/publication-aware attention mechanism is developed for citation event sequences. Empirically, we demonstrate the superiority of CINES to the current state of the arts in research on future citation forecasting.

In future works, we plan to analyze citation strategies and study the performance of CINES on short-term and long-term forecasting. Moreover, we plan to explore temporal attention in social networks.

7 ACKNOWLEDGEMENT

We thank reviewers for comments. This work is supported in part by the National Science Foundation under Grant No. IIS-1717084.

[8]Result of the evaluation on papers is similar but not shown due to the space limit.

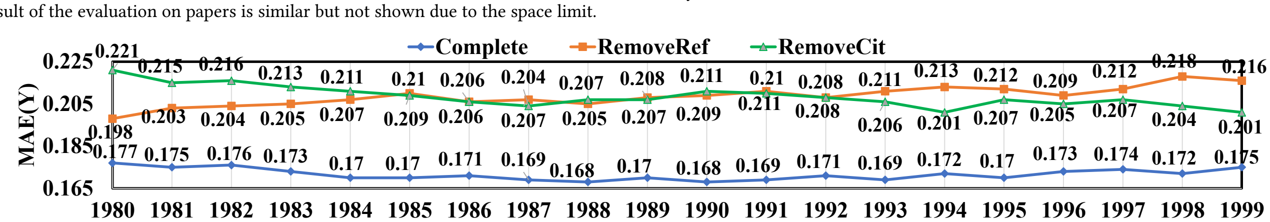

Figure 9: Forecasting MAE for Patents Published in Different Years

REFERENCES

[1] Tanmoy Chakraborty, Suhansanu Kumar, Pawan Goyal, Niloy Ganguly, and Animesh Mukherjee. 2014. Towards a stratified learning approach to predict future citation counts. In *IEEE/ACM Joint Conference on Digital Libraries*. IEEE, 351–360.

[2] Changwoo Choi, Seungkyum Kim, and Yongtae Park. 2007. A patent-based cross impact analysis for quantitative estimation of technological impact: The case of information and communication technology. *Technological Forecasting and Social Change* 74, 8 (2007), 1296–1314.

[3] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. 2016. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1555–1564.

[4] Eugene Garfield. 2006. The history and meaning of the journal impact factor. *jama* 295, 1 (2006), 90–93.

[5] Alan G Hawkes. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58, 1 (1971), 83–90.

[6] Zhongyang He, Zhen Lei, and Dashun Wang. 2018. Modeling citation dynamics of "atypical" articles. *Journal of the Association for Information Science and Technology* 69, 9 (2018), 1148–1160.

[7] Jorge E Hirsch. 2010. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics* 85, 3 (2010), 741–754.

[8] Hyun Jin Jang, Han-Gyun Woo, and Changyong Lee. 2017. Hawkes process-based technology impact analysis. *Journal of Informetrics* 11, 2 (2017), 511–529.

[9] Taoran Ji, Zhiqian Chen, Nathan Self, Kaiqun Fu, Chang-Tien Lu, and Naren Ramakrishnan. 2019. Patent citation dynamics modeling via multi-attention recurrent networks. *arXiv preprint arXiv:1905.10022* (2019).

[10] AL Kinney. 2007. National scientific facilities and their science impact on non-biomedical research. *Proceedings of the National Academy of Sciences* 104, 46 (2007), 17943–17947.

[11] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. 1188–1196.

[12] Changyong Lee, Ohjin Kwon, Myeongjung Kim, and Daeil Kwon. 2018. Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting and Social Change* 127 (2018), 291–303.

[13] Sungjoo Lee, Byungun Yoon, and Yongtae Park. 2009. An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation* 29, 6-7 (2009), 481–497.

[14] Xin Liu, Junchi Yan, Shuai Xiao, Xiangfeng Wang, Hongyuan Zha, and Stephen M Chu. 2017. On predictive patent valuation: Forecasting patent citations and their types. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[15] Cynthia Lokker, K Ann McKibbon, R James McKinlay, Nancy L Wilczynski, and R Brian Haynes. 2008. Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. *Bmj* 336, 7645 (2008), 655–657.

[16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[17] Hua-Wei Shen, Dashun Wang, Chaoming Song, and Albert-László Barabási. 2014. Modeling and predicting popularity dynamics via reinforced Poisson processes.. In *AAAI*, Vol. 14. 291–297.

[18] Dennis Verhoeven, Jurriën Bakker, and Reinhilde Veugelers. 2016. Measuring technological novelty with patent-based indicators. *Research Policy* 45, 3 (2016), 707–723.

[19] Dashun Wang, Chaoming Song, and Albert-László Barabási. 2013. Quantifying long-term scientific impact. *Science* 342, 6154 (2013), 127–132.

[20] Yongqing Wang, Huawei Shen, Shenghua Liu, Jinhua Gao, and Xueqi Cheng. 2017. Cascade Dynamics Modeling with Attention-based Recurrent Neural Network.. In *IJCAI*. 2985–2991.

[21] Shuai Xiao, Junchi Yan, Changsheng Li, Bo Jin, Xiangfeng Wang, Xiaokang Yang, Stephen M Chu, and Hongyuan Zha. 2016. On Modeling and Predicting Individual Paper Citation Count over Time.. In *IJCAI*. 2676–2682.

[22] Shuai Xiao, Junchi Yan, Xiaokang Yang, Hongyuan Zha, and Stephen M Chu. 2017. Modeling the intensity function of point process via recurrent neural networks. In *Thirty-first aaai conference on artificial intelligence*.

[23] Junchi Yan, Shuai Xiao, Changsheng Li, Bo Jin, Xiangfeng Wang, Bin Ke, Xiaokang Yang, and Hongyuan Zha. 2016. Modeling Contagious Merger and Acquisition via Point Processes with a Profile Regression Prior.. In *IJCAI*. 2690–2696.

[24] Rui Yan, Jie Tang, Xiaobing Liu, Dongdong Shan, and Xiaoming Li. 2011. Citation count prediction: learning to estimate future citations for literature. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. 1247–1252.

[25] Tian Yu, Guang Yu, Peng-Yu Li, and Liang Wang. 2014. Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics* 101, 2 (2014), 1233–1252.

[26] Ke Zhou, Hongyuan Zha, and Le Song. 2013. Learning triggering kernels for multi-dimensional hawkes processes. In *International Conference on Machine Learning*. 1301–1309.