An Attentional Multi-scale Co-evolving Model for Dynamic Link Prediction

Guozhen Zhang
Department of Electronic Engineering, Beijing National Research Center for Information Science and Technology, Tsinghua University
Beijing, China
zgz18@mails.tsinghua.edu.cn

Tian Ye
PBC School of Finance, Tsinghua University
Beijing, China
yet.22@pbcsf.tsinghua.edu.cn

Depeng Jin
Department of Electronic Engineering, Beijing National Research Center for Information Science and Technology, Tsinghua University
Beijing, China
jindp@tsinghua.edu.cn

Yong Li
Department of Electronic Engineering, Beijing National Research Center for Information Science and Technology, Tsinghua University
Beijing, China
liyong07@tsinghua.edu.cn

ABSTRACT

Dynamic link prediction is essential for a wide range of domains, including social networks, bioinformatics, knowledge bases, and recommender systems. Existing works have demonstrated that structural information and temporal information are two of the most important information for this problem. However, existing works either focus on modeling them independently or modeling the temporal dynamics of a single structural scale, neglecting the complex correlations among them. This paper proposes to model the inherent correlations among the evolving dynamics of different structural scales for dynamic link prediction. Following this idea, we propose an Attentional Multi-scale Co-evolving Network (AMCNet). Specifically, We model multi-scale structural information by a motif-based graph neural network with multi-scale pooling. Then, we design a hierarchical attention-based sequence-to-sequence model for learning the complex correlations among the evolution dynamics of different structural scales. Extensive experiments on four real-world datasets with different characteristics demonstrate that AMCNet significantly outperforms the state-of-the-art in both single-step and multi-step dynamic link prediction tasks.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Networks** → **Online social networks**.

KEYWORDS

dynamic link prediction, multi-scale information, multi-scale co-evolving model

WWW '23, May 1–5, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04.
https://doi.org/10.1145/3543507.3583396

ACM Reference Format:
Guozhen Zhang, Tian Ye, Depeng Jin, and Yong Li. 2023. An Attentional Multi-scale Co-evolving Model for Dynamic Link Prediction. In *Proceedings of the ACM Web Conference 2023 (WWW '23), May 1–5, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3543507.3583396

1 INTRODUCTION

Dynamic link prediction, i.e., estimating the likelihood of a future connection between two given nodes in a time-varying network based on observed data, is one of the most critical problems in the network science literature [31, 37]. There has been a burst of methods tackling the problem in the last few years and it has been the foundation of various applications, such as recommender systems [38, 40], social network analysis [11, 27], and anomaly detection [20].

Previous works have demonstrated two types of most useful information for the task. One is temporal information, which characterizes the evolving dynamics of the network [22, 32, 34]. The other is structural information that suggests that network topology reflects whether two nodes are more likely to form a link in the future [3, 15, 43].

As illustrated in Figure 1, these two types of information are deeply connected. Specifically, from a structural perspective, we can characterize a given network in three scales: microscopic level, which focuses on the state of each node and edge; mesoscopic level, which concerns the states of groups and communities; and macroscopic level, which characterizes the state of the whole network [2, 33], such as the degree distribution or the network growth rate. Our key observation is that the temporal dynamics of different structural scales complement one another and are coherent in the meantime. We can illustrate it from two perspectives. On the one hand, networks are constituted of individuals and their interconnections, and thus macroscopic temporal dynamics naturally originate from microscopic temporal dynamics of how each individual chooses to build connections with others. On the other hand, research has suggested that human behaviors are dynamically influenced by their social connections [19], and this influence affects a wide range of attributes and behaviors, such as political orientation, music tastes, and even how people choose new friends [4, 23].

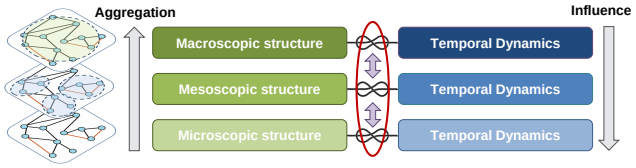

Figure 1: The illustration on the complex correlations among the temporal dynamics of different structural scales.

In other words, microscopic temporal dynamics are affected by mesoscopic temporal dynamics. Further, mesoscopic community formation is affected by how networks evolve at a macroscopic level [1].

Most of the existing works model the two types of information independently [7, 9, 44], and thereby result in poor prediction performance. Some recent advancements try to model the connections between the structural information and the temporal information [6, 22, 27, 29, 43]. However, these works either only focus on modeling the temporal dynamics of a single structural scale or model different structural scales independently, neglecting the complex correlations among the temporal dynamics of different structural scales.

To bridge the gaps in the literature, we present an Attentional Multi-scale Co-evolving Network (AMCNet) to model the coherence among the evolving dynamics of different structural scales for dynamic link prediction. Specifically, to model multi-scale structural and temporal information, we first design a multi-scale representation learning module, which learns node embeddings based on graph attentional networks and leverages a multi-scale pooling algorithm to obtain embeddings that capture the structural information of different scales. We further enhance the model's representation power at the mesoscopic level by introducing motifs [30], the basic building blocks of complex networks at a mesoscopic level, to construct motif graphs and obtain mesoscopic node embedding by fusing all the node embeddings learned from motif graphs. Second, we develop a multi-scale co-evolving model to learn the temporal dynamics of each structural scale. Third, to learn the inherent correlations among the temporal dynamics of different structural scales, we propose to leverage the higher-scale representations to guide the learning process of the lower-scale representations by a novel hierarchical attention-based model. Finally, we jointly leverage the embeddings of different scales to predict future links. Extensive experiments on four real-world datasets with different characteristics demonstrate the superior performance of our proposed model. We make our code and data publicly available[1].

We summarize our main contributions as follows:

- To the best of our knowledge, we are the first to study the inherent correlations among the evolving dynamics of different structural scales for dynamic link prediction.
- We propose a novel dynamic link prediction model, AMCNet, which jointly models multi-scale structural information and the complex correlations among the evolving dynamics of different structural scales.
- Extensive experiments on four real-world datasets with different characteristics demonstrate that AMCNet significantly outperforms the state-of-the-art in both single-step and multi-step dynamic link prediction tasks, and it achieves a performance gain of up to 7.2% in terms of AUC compared with the best baseline.

[1] https://github.com/tsinghua-fib-lab/AMCNet

2 PROBLEM FORMULATION

The goal of dynamic link prediction is to predict the future connections between two given nodes in a time-varying network based on historical data. Let $G = (\mathsf{V}, \mathsf{E})$ denotes a time-varying network, where V is the node set and E is the link set, with its adjecent matrix denoted as A. Following prior works, we split G into a series of equal-spaced snapshots $G = \{G_1, G_2, \ldots, G_N\}$, and we formally define dynamic graph prediction as follows: given a historical series of snapshots $G = \{G_1, G_2, \ldots, G_T\}$, we aim to predict the link structure at future time steps, i.e., $G = \{G_{T+1}, G_{T+2}, \ldots, G_{T+n}\}$.

3 ATTENTIONAL MULTI-SCALE CO-EVOLVING NETWORK (AMCNET)

To learn the inherent correlations among the evolving dynamics of different structural scales, we propose AMCNet and show its overall architecture in Figure 2. It contains two key components: a multi-scale representation learning module and a multi-scale evolving module. The main idea is to learn the representation of different scales with the multi-scale representation learning module first, and then build connections among the temporal dynamics of different scales and learn the complex co-evolving dynamics with the multi-scale evolving module. In the following sections, we elaborate on the details of the two key modules.

3.1 Multi-scale Representation Learning

Microscopic Representation. To obtain microscopic representations, i.e., node representation, of each snapshot, graph neural networks are common choices, which have a better representation power compared with traditional random-walk-based methods [14] and can handle graphs with or without features at the same time. We adopt Graph Attention Network [35] (GAT), one of the state-of-the-art graph neural network architectures, so that the model can focus on the most important information. For each time snapshots, It takes node features x_i^t as inputs and outputs the learned representations of H^t, which can be formulated as follows,

$$\begin{aligned} e_{ij}^t &= \sigma\left(a^T \left[Wx_i^t \| Wx_j^t\right]\right), \quad \forall (i,j) \in \mathsf{E}_t, \\ \alpha_{ij}^t &= \frac{\exp\left(e_{ij}^t\right)}{\sum_{k \in \mathsf{N}_i} \exp\left(e_{ik}^t\right)}, \quad \forall (i,j) \in \mathsf{E}_t, \\ H_i^t &= \sigma\left(\sum_{j \in \mathsf{N}_i} \alpha_{ij}^t W x_j^t\right) \quad \forall i \in \mathsf{V}_t, \end{aligned} \tag{1}$$

where N_i denotes the neighbors of node i and H_i^t denotes the node representation of node i at the snapshot t. $H^t = [H_1^t, H_2^t, \ldots, H_{N_v}^t]$, $N_v = |\mathsf{V}|$. a and W are learnable parameters. $\sigma(\cdot)$ is a non-linear activation function, and we adopt ReLU [26] in our implementation. $x_i^t \in \mathbb{R}^D, \forall i \in \mathsf{V}$ with D as the dimension of node features.

Mesoscopic Representation. As the connection between the microscopic graph structure and macroscopic graph structure, the

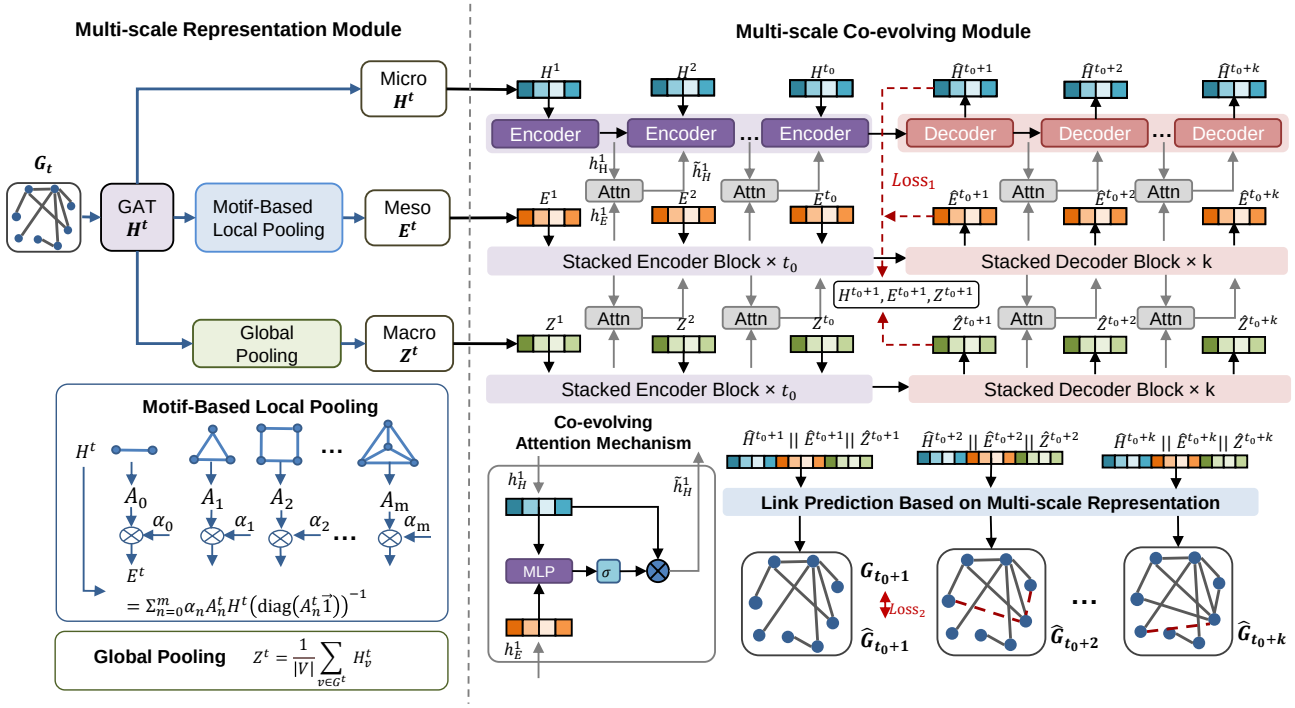

Figure 2: The architecture of AMCNet. It contains two modules: a multi-scale representation learning module, which can learn embeddings that capture the structural information of different scales, and a multi-scale co-evolving module with a novel hierarchical attention mechanism, which can capture the complex correlations among the evolution dynamics of different structural scales.

mesoscopic graph structure contains the richest information. Most existing works neglect this structural scale [6, 22] and thus are not able to learn the complex correlations between different structural scales.

At first, we follow prior works and obtain mesoscopic representation by performing a simple mean pooling on the microscopic representations [13]. However, in our experiments, we found that such an operation without learnable parameters is not enough to capture the rich information on the mesoscopic scale. To solve this challenge, we introduce motif, a special subgraph that is regarded as the basic building block of complex networks, and propose a motif-based local pooling method to enhance the mesoscopic representation module.

Specifically, given an undirected graph G and a set of motifs $\mathsf{M} = \{M_0, M_1, \ldots, M_m\}$ with m as the number of chosen motifs, we can construct a set of motif-based graphs from the original graph with their adjacent matrices denoted as $\mathsf{A}^t = A_0^t, A_1^t, \ldots, A_m^t$ by assigning an edge to two nodes if they are in the same motif in the original graph, which can be fomulated as follows,

$$A_{n\ i,j}^t = \begin{cases} 1 & i = j, \\ 1 & i, j \text{ are in the same motif } M_n, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

With the motif-based graphs, we can obtain mesoscopic representations E^t by performing a weighted pooling on the learned microscopic representations, which can be formulated as follows,

$$\begin{aligned} \alpha_n^t &= a_n^T H^t, \\ E^t &= \Sigma_{n=0}^m \alpha_n^t A_n^t H^t (\mathrm{diag}(A_n^t \vec{1}))^{-1}, \end{aligned} \tag{3}$$

where a_n is a learnable parameter and α_n is the learnable weight corresponding to the n-th motif. Note that we normalize the mesoscopic representations for each node by its degree. $\mathrm{diag}(\cdot)$ refers to the operation of constructing a diagonal matrix.

Macroscopic Representation. Following prior works, we use a global pooling to encode the whole graph information and derive its macroscopic representation Z^t, which can be formulated as follows,

$$Z^t = \frac{1}{|\mathsf{V}^t|} \tilde{\mathsf{O}}_{i \in \mathsf{V}^t} H_i^t. \tag{4}$$

Note that since the core idea of this paper is to model the complex relationship among the temporal dynamics of different structural scales, obtaining the representations of each structural scale is only its first step. To validate our idea, we aim to capture the information of different structural scales with a minimal design in the multi-scale representation learning module. These designs can be easily extended to more complicated ones, such as hyperbolic graph neural networks [5] and self-attention-based graph pooling [18] for better practical performance.

3.2 Multi-scale Co-evolving module

From the multi-scale representation learning module, we obtain the microscopic representation H^t, mesoscopic representation E^t,

and macroscopic representation Z^t of the graph at the snapshot t. To learn the inherent correlations among the evolving dynamics of different structural scales, we propose a multi-scale co-evolving module. Our main idea is that modeling the coherence among multi-scale structural-temporal dynamics is critical for dynamic link prediction, and we can explain why it works from two perspectives:

- First, from an information theory perspective, the amount of noise in the data and the level of information aggregation are typically inversely correlated. The higher the structural level, the less the information, and the smaller the noise. Thus, the temporal dynamics of the higher structural scale are more predictable than the lower structural scale. Therefore, higher-scale predictions help correct the potential systematic bias of the lower-scale predictions, which impose a constraint on the learned model to be scale-invariant.
- Second, the information of different structural scales captures different characteristics of the graph and thus complements one another [3, 22]. Jointly modeling the temporal dynamics of different scales enables the model to leverage information from a different range of contexts to make predictions.

To realize our idea, we designed a hierarchical attention-based model with two key designs. The first is to learn the evolving dynamics of different structural scales simultaneously. The second is to build connections between the learning process of different scales by a co-evolving attention mechanism, which uses higher-scale representations to guide the learning process of the lower-scale representations.

Sequence to Sequence Backbone. To model the temporal information of each structural scale, we leverage a sequence-to-sequence model (Seq2Seq) as the backbone. In this way, we can deal with single-step link prediction and multiple-step link prediction simultaneously. Specifically, we use a traditional encoder-decoder framework constructed with LSTM blocks and feed the representations of different structural scales into a different Seq2Seq model to learn their evolving dynamics simultaneously. Taking the microscopic scale as an example, we can formulate the process as follows,

$$\begin{aligned} h_H^t, c_H^t &= \text{LSTM}\ \ H^t, h_H^{t-1}, c_H^{t-1}\ \ , \\ h_{\hat{H}}^t, c_{\hat{H}}^t &= \text{LSTM}\ \ \hat{H}^{t-1}, h_{\hat{H}}^{t-1}, c_{\hat{H}}^{t-1}\ \ , \\ \hat{H}^t &= \tanh\ \ W h_H^t \end{aligned} \tag{5}$$

where h_H^t and $h_{\hat{H}}^t$ refer to the hidden state of the encoder and the decoder, respectively. c_H^t and $c_{\hat{H}}^t$ are the states of the memory cell of the encoder and the decoder, respectively. $\hat{H}^t$ is the predicted microscopic representations at time t. Similarly, we obtain the mesoscopic representations $\hat{E}^t$ and macroscopic representations $\hat{Z}^t$, as illustrated in Figure 2.

Co-evolving Attention Mechanism. To learn the coherence among the evolving dynamics of different scales, we leverage the higher-scale representations to guide the learning process of the lower-scale representations by a novel hierarchical attention mechanism. Specifically, as illustrated in Figure 2, we change the input of the lower level Seq2Seq model from the hidden states of the last time step to a state that learned from both the lower structural scale's hidden states and the higher structural scale's hidden states with an element-wise co-evolving attention mechanism. In other words, we let macroscopic hidden states h_Z^t guide the evolution learning process of mesoscopic hidden states h_E^t and mesoscopic hidden states h_E^t guide the evolution learning process of microscopic hidden states h_H^t. Taking the microscopic scale as an example, we formulate the attention mechanism as follows,

$$\begin{aligned} \beta^t &= \text{softmax}\ \ \sigma\ \ W_\beta\ \ h_H^t \| h_E^t\ \ , \\ \tilde{h}_H^t &= \beta^t \odot h_H^t \end{aligned} \tag{6}$$

where β^t is the attention vector, W_β is a learnable parameter, and $\odot$ refers to the element-wise product. We can further reformulate the sequence-to-sequence structure with a hierarchical co-evolving attention mechanism. For the encoder,

$$\begin{aligned} \tilde{h}_H^{t-1} &= \text{Attention}(h_H^{t-1}, h_E^{t-1}), \\ h_H^t, c_H^t &= \text{LSTM}\ \ H^t, \tilde{h}_H^{t-1}, c_H^{t-1}\ \ , \end{aligned} \tag{7}$$

For the decoder,

$$\begin{aligned} \tilde{h}_{\hat{H}}^{t-1} &= \text{Attention}(h_{\hat{H}}^{t-1}, h_{\hat{E}}^{t-1}), \\ h_{\hat{H}}^t, c_{\hat{H}}^t &= \text{LSTM}\ \ \hat{H}^t, \tilde{h}_{\hat{H}}^{t-1}, c_H^{t-1}\ \ , \\ \hat{H}^t &= \tanh\ \ W \tilde{h}_H^t\ \ . \end{aligned} \tag{8}$$

We can obtain the mesoscopic representations in a similar way. Note that in the computational process, the model must first complete the computation of the hidden states of the higher structural scale at time t (e.g., h_E^t). Only then can it compute the hidden states of the lower structural scale (e.g., h_H^t).

3.3 Link Prediction Based on Multi-scale Representations

Through the multi-scale co-evolving module, we obtain representations of different structural scales for each node. We jointly leverage them for link prediction by training a predictor network with two fully connected layers. For example, to predict whether there will be a link at time t between two nodes i and j, we can develop the formulation as follows,

$$\begin{aligned} \hat{Y}_i^t &= \text{Concat}\ \ \hat{H}_i^t, \hat{E}_i^t, \hat{Z}_i^t \qquad \forall i \in V, \\ P_t(i,j) &= \sigma\ \ W_2\ \ \sigma\ \ W_1\ \ \hat{Y}_i^t \| \hat{Y}_j^t\ \ + b_1\ \ + b_2, \end{aligned} \tag{9}$$

where $P_t(i,j) \in [0,1]$ represents the probability of forming a link between node i and node j in the future. W_1, W_2, b_1, b_2 are trainable parameters.

3.4 Training

We first train the GAT to get micro representation H_i^t of each node at each snapshot. We expect the node representations to learn the structural information of the graph sufficiently well so that it is able to perform well in link prediction tasks. Inspired by the work of Sankar et al. [29], we let the nodes co-occurring in a fix-length random walk to have similar representations by leveraging a binary

Datasets	Enron	UCI	Youtube	Foursquare
Node Feature	×	×	×	✓
# Nodes	150	1899	2993	2940
# Links	1526	13838	88587	6772
Timespan (days)	1137	193	165	665
Slice days	45	9	7	20
# snapshots	26	21	24	34
# Avg links per snapshot	616	11258	45810	4807
# Avg new links per snapshot	177	898	3691	199

Table 1: Summary statistics of the four datasets.

cross-entropy loss based on fixed-length random walks, which can be formulated as follows,

$$\mathcal{L} = \sum_{t=1}^{T} \sum_{v \in V} \Big(\sum_{u \in \mathcal{N}^t_{\text{walk}}(v)} -\log\ \sigma\left(H^t_u, H^t_v\right) - w_n \sum_{u' \in P^t_n(v)} \log\ 1 - \sigma\left(H^t_{u'}, H^t_v\right) \Big), \tag{10}$$

where $\mathcal{N}^t_{\text{walk}}(v)$ is the set of co-occurring nodes of node v in the fixed-length random walks at the time snapshot t. $P^t_n(v)$ is a negative sampling distribution correlated to degrees, and w_n is the negative sampling ratio.

Then, we train the multi-scale co-evolving module by optimizing a loss function that consists of two parts. The first part is the mean square error loss (MSE) of the Seq2Seq model, which characterizes how well the model captures the evolution dynamics and can be formulated as follows,

$$\mathcal{L}_1 = \frac{1}{k|V|} \sum_{t=t_0+1}^{t_0+k} \sum_{i \in V^t} \left\| \hat{Y}^t_i - Y^t_i \right\|_2^2, \tag{11}$$

where $\hat{Y}^t_i$ is the output of the decoder, and k is the required prediction steps. The second part is a binary cross-entropy loss for the link predictor that predicts whether there will be a link between two nodes, which can be formulated as follows,

$$\mathcal{L}_2 = -\sum_{t=t_0+1}^{t_0+k} \Big(\sum_{e_{ij} \in E^t} \ln\left(\sigma\left(P_t(i,j)\right)\right) + \sum_{e_{ij} \notin E^t} \ln\left(1 - \sigma\left(P_t(i,j)\right)\right) \Big). \tag{12}$$

We optimize the weighted sum of the two losses during training, which can be formulated as follows,

$$\mathcal{L}_{\text{evolve}} = \mathcal{L}_1 + \alpha \mathcal{L}_2, \tag{13}$$

where α is a tunable hyper-parameter that balances the two losses.

4 EXPERIMENTS

4.1 Datasets

We conduct experiments on four dynamic network datasets that differ in category, scale and density. The statistics of the datasets is shown in Table 1. Here, we briefly introduce them as follows:

Enron [28]: An email communication network where each edge represents an email interaction between two people. Enron is a small network with only 150 nodes and 1526 time-stamped edges spanning more than three years. It is denser than other networks where the average node degree equals 4.

UC Irvine messages (UCI) [17]: UCI is a network of online forums at the University of California, Irvine. If two students interact on the same forum post, they are connected. It has a total of 1899 nodes spanning in half a year.

Youtube [25]: Youtube is a popular video sharing website. We obtain the data between February 2007 and July 2007, including over 1.1 million users and 4.9 million edges, which denotes users' following relationships. Considering the computational efficiency, we randomly select 3,000 active users with their corresponding edges.

Foursquare [41]: This dataset includes check-in data collected from Foursquare on a global scale from April 2012 to January 2014, as well as two snapshots of users' social networks before and after the data collection period. In this work, we use the data collected from Tokyo and focus on the evolution of social networks. To obtain fine-grained time-stamped social networks, we first obtain new relationships by calculating the difference between the two social network snapshots. Then, we assume these new relationships are formed when two users post a check-in at the same place at the same time the first time.

4.2 Baselines and Experiment Settings

Baselines. We compare the performance of AMCNet with eleven state-of-the-art methods from three research lines, which we introduce as follows:

Heuristics methods:

- Common Neighbors(CN) [21]: A heuristics method based on a similarity score that measures how many mutual friends two have.
- Newton [36]: It takes inspiration from Newton's gravitational law and models the degree centrality as the mass of the nodes and the lengths of shortest paths between two nodes as distances.

Static Link Prediction methods:

- Node2vec [14]: A node embedding method based on biased random walk sampling.
- GCN [16]: An inductive node representation learning framework for the graph.
- GAT [35]: A variant of GCN that use self-attention to aggregate messages.
- HARP [7]: a hierarchical method for learning low dimensional embeddings of a graph's nodes which preserves higher-order structural features.

Dynamic Link Prediction methods:

- DySAT [29]: A dynamic graph neural network which computes node representations through joint self-attention along the two dimensions of the structural neighborhood and temporal dynamics.
- DynamicTriad [43]: A dynamic graph embedding technique that preserves both structural information and evolution patterns through modeling the triadic closure process.
- GC-LSTM [8]: GC-LSTM embeds the two-layer GCN in the LSTM to learn the spatio-temporal information for end-to-end dynamic link prediction.

Models	Enron		UCI		Youtube		Foursquare	
	AUC	MAP	AUC	MAP	AUC	MAP	AUC	MAP
CN [21]	0.686	0.678	0.528	0.524	0.570	0.562	0.504	0.504
Newton [36]	0.685	0.691	0.570	0.566	0.570	0.570	0.510	0.511
node2vec [14]	0.720	0.707	0.641	0.614	0.644	0.642	0.732	0.769
GCN [16]	0.529	0.535	0.647	0.647	0.622	0.683	0.562	0.583
GAT [35]	0.525	0.533	0.599	0.615	0.645	0.575	0.551	0.578
HARP [7]	0.534	0.504	0.670	0.682	0.635	0.613	0.580	0.577
DySAT [29]	0.600	0.606	0.651	0.626	0.654	0.613	0.633	0.617
DynamicTriad [43]	0.532	0.529	0.656	0.649	0.677	0.664	0.559	0.562
GC-LSTM [8]	0.574	0.556	0.687	0.687	0.623	0.621	0.616	0.605
TGAT [40]	0.583	0.567	0.654	0.659	0.637	0.618	0.599	0.593
CAW-N [39]	0.733	0.758	0.718	0.711	0.689	0.665	0.768	0.775
HTGN [42]	0.660	0.667	0.704	0.696	0.711	0.695	0.699	0.682
AMCNet	0.750	0.766	0.731	0.712	0.762	0.715	0.781	0.800

Table 2: The performance evaluation results on four different datasets for multi-step prediction.

- TGAT [40]: It proposes the temporal graph attention layer to capture the temporal-topological features.
- CAW-N [39]: CAW-N is a state-of-the-art dynamic graph representation learning method based on causal anonymous walks, which leverages a novel strategy to make the node identities anonymized.
- HTGN [42]: It leverages hyperbolic graph neural network and hyperbolic gated recurrent neural network to model the evolving dynamics of the graph.

Note that most of the existing methods cannot directly fit into the multi-step link prediction setting. Following prior works [29], we use the latest learned embeddings to predict multiple future time steps independently.

Experiment Settings. We test our model on both multi-scale link prediction and single-step link prediction tasks. For both tasks, we train our model on the historically observed graph snapshots $G = \{G_1, G_2, \ldots, G_T\}$ to derive the multi-scale representations corresponding to each time step and let it predict the connections in the future k time steps, where $k = 1$ for single-step link prediction and $k > 1$ for multi-step link prediction. In our experiment, we use the first 9 snapshots as the training samples, the next 3 snapshots as the validation samples, and the next 6 snapshots as the test samples. All the experiments are conducted in transductive settings. We focus on newly added links in future snapshots and regard them as positive samples while sampling an equal number of non-links as negative samples. We use the area under ROC curve (AUC) [14] and mean average precision (MAP) [12] to evaluate model performances.

Reproducibility. In the multi-scale representation learning module, We adopt a two-layer GAT network with 16 and 8 attention heads, respectively, and an embedding size of 128. When optimizing loss function Eqs.(10), we follow the strategy of DySAT[29], sampling 10 walks of length 20 per node, each with a context window size of 10. In order to avoid over-fitting, we apply l_2 regularization with $\lambda = 5e^{-4}$ and dropout rate of 0.5. We use mini-batch gradient descent with Adam for training, and perform a grid search on the learning rate in a range of {1e-4,5e-4,1e-3,5e-3,1e-2,5e-2}. In constructing the mesoscopic representation, the motif weights are learnable parameters with initialized value of {0.4,0.1,0.1,0.1,0.1,0.1,0.1,0.1}. In the multi-scale co-evolving module, the input and output steps of the sequence-to-sequence model are set to three. The model's learning rate has a grid search range of {0.0005, 0.001, 0.005, 0.01}. When optimizing the joint loss function 11 and 12, we determine the grid search range of the hyper-parameter α with regards to the magnitudes of L_1 and L_2. As a result, the grid search range for α is set as {1e-2,5e-2,1e-3,5e-3,1e-4,5e-4}. We also perform a grid search on other hyper-parameters, including the batch size and the l_2 regularization coefficient, to find the best hyper-parameters for AMCNet. For reproducibility, we make our implementation codes of AMCNet available (the link is presented in the introduction section).

Models	Enron	UCI	Youtube	Foursquare
CN [21]	0.673	0.589	0.561	0.508
Newton [36]	0.726	0.662	0.562	0.511
node2vec [14]	0.800	0.640	0.631	0.733
GCN [16]	0.505	0.581	0.608	0.569
GAT [35]	0.544	0.617	0.677	0.563
HARP [7]	0.408	0.680	0.648	0.585
DySAT [29]	0.607	0.616	0.657	0.623
DynamicTriad [43]	0.504	0.643	0.701	0.556
GC-LSTM [8]	0.530	0.736	0.642	0.646
TGAT [40]	0.608	0.671	0.644	0.623
CAW-N [39]	0.834	0.741	0.706	**0.773**
HTGN [42]	0.692	0.728	0.717	0.703
AMCNet	0.838	0.739	0.767	0.772

Table 3: The AUC of the single-step link prediction results.

4.3 Main Results

Multi-step Link Prediction. To examine the effectiveness of our model, we compare it with the state-of-the-art baselines from three groups of research lines for multi-step link prediction and show the results in Table 2. Overall, we have three key observations.

- First, AMCNet consistently outperforms all state-of-the-art baselines across all four real-world datasets. Specifically, it provides a relative performance gain of 2.3%, 1.8%, 7.2%, 1.7% in terms of AUC, and 1.1%, 1.4%, 2.9%, 3.2% in terms of mAP, on the Enron, UCI, Youtube, and Foursquare dataset, respectively, which demonstrates its effectiveness and robustness. The results also indicate that learning the complex correlations of the temporal dynamics of different structural scales is indeed important for the multi-step link prediction task.

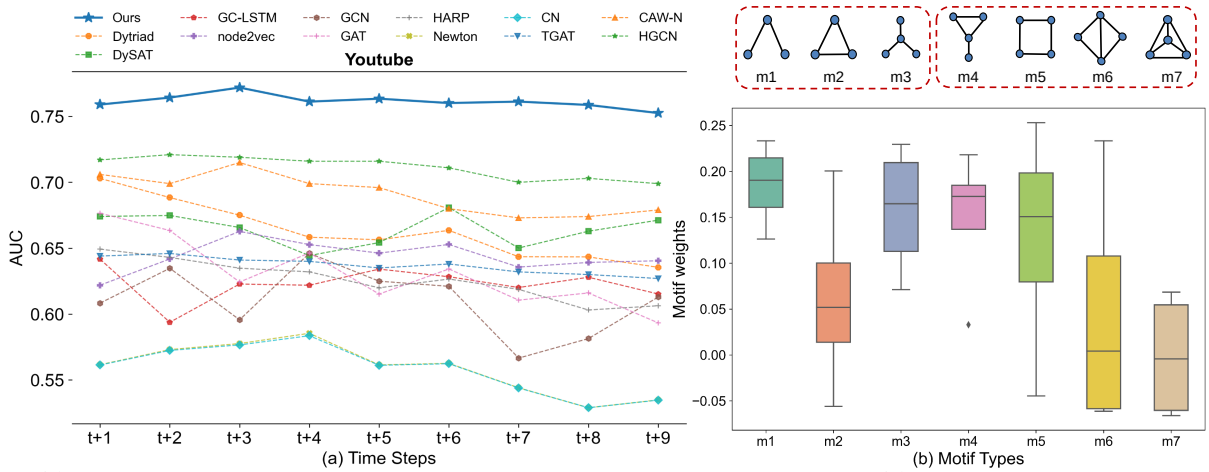

Figure 3: (a) The model's performance in a longer time steps on the Youtube dataset. (b) An illustration of motifs and the distribution of the learned motif weights.

- Second, among all baseline models, dynamic link prediction methods generally perform better than static ones since they incorporate temporal information into the model design. Heuristic methods, including Common Neighbor and Newton's method, perform well on small and dense datasets Enron but are close to random prediction on other large datasets.
- Third, on larger and more dynamic datasets, such as the Youtube dataset, our model performs consistently better than other models, which demonstrates its superior scalability.

Single-step Link Prediction. Since most of the past literature has focused on single-step prediction tasks for link prediction, we also compare the performance of the present model, AMCNet, on single-step prediction with the baseline methods. The evaluation metric is the single-step AUC based on new links, and the results are shown in Table 3. We can see that our model outperforms almost all state-of-the-art models on the single-step link prediction task with more significant improvements for the larger datasets with more nodes, which demonstrates its effectiveness.

Generalizability in longer step predictions. To further investigate our model's generalizability along the temporal dimension, we test the performances of AMCNet with longer time steps. Specifically, we conduct an experiment on the Youtube dataset to predict the next 9 time steps and visualize the result in Figure 3(a). We can see that our models' performance is stable as time goes by, which suggests the model has good generalizability along the temporal dimension.

4.4 Ablation Study

Our model consists of two main modules, the multi-scale representation learning module and the multi-scale co-evolving module. To further verify the effectiveness of the two modules, we set up a series of ablation study.

The Effectiveness of the Multi-scale Representation. The model has three scales of representation: microscopic, mesoscopic, and macroscopic. To verify their effectiveness, we remove the macroscopic and mesoscopic representations separately while retaining the sequence-to-sequence structure and attention mechanism. The experimental results show that both the introduction of mesoscopic and macroscopic representations improve the model's performance, which suggests that jointly modeling the temporal dynamics of different scales is effective. We further examine the effectiveness of the motif-based mesoscopic modeling by substituting it with attention-based pooling [18]. The results are shown in Table 4 (Without motifs). It shows that the model's performance drops significantly without our proposed motif-based designs, demonstrating its effectiveness.

The Effectiveness of the Sequence-to-Sequence Structure. We introduce a sequence-to-sequence structure to learn the evolutionary dynamics of the graph. To demonstrate the effectiveness of the design, we remove the sequence-to-sequence structure in the ablation study and directly leverage the multi-scale representations on the last time snapshot of the training set to make predictions following prior works [29]. Specifically, $Y^{t_0} = \text{Concat}\ H^{t_0}, E^{t_0}, Z^{t_0}$ is used to predict the probability of an future edge of node v from $t_0 + 1$ to $t_0 + k$. As shown in Table 4, the sequence-to-sequence structure learns the underlying graph evolution patterns and is able to improve the accuracy of future multi-step predictions.

The Effectiveness of the Co-evolving Attention Mechanism. To model the complex correlations among the temporal dynamics of different structural scales, the model is designed with learnable attention weights that explicitly characterize the micro-meso-macro relationship. In this ablation experiment, we remove the mechanism, i.e., the three scales of representation are learned separately. The experimental results show that the model's performance drops significantly without the attention mechanism, which supports our main motivation that modeling the complex correlations among the temporal dynamics of different structural scales is critical for dynamic link prediction.

4.5 The Role of Different Motifs

To better understand the role that different motifs play in the model, we carry out a more in-depth analysis of the importance of each motif. All motifs involved in the experiment are shown in Figure 3(b).

Models	Enron		UCI		Youtube		Foursquare	
	AUC	MAP	AUC	MAP	AUC	MAP	AUC	MAP
Only Micro	0.700	0.686	0.690	0.660	0.735	0.682	0.725	0.743
Only Micro and Meso	0.739	0.740	0.726	0.696	0.745	0.699	0.760	0.780
Without motifs	0.717	0.714	0.725	0.695	0.756	0.698	0.785	0.801
Without the Seq2Seq Structure	0.739	0.727	0.675	0.653	0.748	0.700	0.694	0.698
Without the Co-evolving Attention Mechanism	0.733	0.732	0.731	0.711	0.736	0.695	0.759	0.788
AMCNet	**0.750**	**0.766**	**0.731**	**0.712**	**0.762**	**0.715**	**0.794**	**0.816**

Table 4: The ablation study results.

Models	Enron	UCI	Youtube	Foursquare
Without motifs	0.717	0.725	0.756	0.785
Add three-node motifs	0.748	0.730	0.762	0.794
Add four-node motifs	0.750	0.731	0.762	0.794

Table 5: The AUC of AMCNet with two-node, three-node and four-node motifs on the four datasets.

We aim to answer the following questions: Can the introduction of motifs with a higher number of nodes improve the model's performance? How do different motifs contribute to the dynamic link prediction, and why are some motifs more informative?

As shown in Table 5, the introduction of three-node motifs boosts the AUC on all four datasets, but when we include the four-node motifs, there is a relatively small performance improvement on two datasets and no change on the other two. The results suggest that three-node motifs are the most informative, and the marginal effect of adding more motifs decreases as the number of nodes in the motif increases.

Figure 3(b) shows the distribution of the learned motif weights on the four datasets. As we can see, there are two important observations. First, m1/m3/m4 play a positive role, while m2/m6/m7 play a negative role. Second, m1/m3/m4/m7 have less variance on different datasets. A possible explanation for the phenomenon is that open triads with three nodes tend to be connected with each other, which is the most informative for the link prediction tasks. This phenomenon is also referred to as the triadic closure process [10]. Specifically, m1/m3/m4 contain open triangles. Thus, when we construct the motif-based adjacency matrix, the social theory is inherently integrated into the mesoscopic modeling to strengthen the connections between nodes with common neighbors, thus enhancing the model's effectiveness. In contrast, m2/m6/m7, which contain only closed triangles, focus only on nodes that are already closely connected. Thus, it contains redundant information similar to microscopic node representations and thereby has less effect on the model's prediction results.

5 RELATED WORK

Link prediction has been a long-standing problem in the network science literature. In the early days, researchers typically neglected the temporal characteristics of links and formulated it as a static prediction problem [21, 24]. Despite its simplicity, this formulation has limited downstream applications. With the maturity of recurrent neural networks and the attention mechanism, there has been a burst of methods for dynamic link prediction in the last few years [31, 37], which is also the focus of this work.

Existing works have demonstrated two types of most useful information for this problem: temporal information and structural information [3, 15, 22, 32, 34, 43]. Most of the existing works model them independently [7, 9, 44]. For example, to model the structural information better, Chen et al. [7] propose a hierarchical representation learning framework based on existing random walk based node embedding algorithms. In terms of capturing the temporal information, a representative strategy is *temporal smoothness* that imposes constraints to ensure the network embeddings do not change dramatically in consecutive time steps [9, 44]. The performance of these methods is generally unsatisfactory and unstable.

A few recent studies try to leverage both information simultaneously [6, 22, 27, 29, 43]. For example, Sankar et al. [29] propose DySAT, which jointly leverages structural and temporal information by applying self-attention to learn the network embeddings and the temporal dynamics. Wang et al [39] propose CAW-N that applies a novel anonymization strategy on temporal random walks to make the method inductive and capable of modeling temporal network motifs. Zhou et al. [43] explicitly model the triadic closure process in a temporal smoothness model. However, these works fail to capture the complex relationships between the temporal information and the structural information of different scales. To tackle this challenge, we present an attentional multi-scale co-evolving network that can learn the inherent correlation among the evolving dynamics of different structural scales for dynamic link prediction.

6 CONCLUSION

In this paper, we propose an attentional multi-scale co-evolving network, AMCNet, to model the inhernt correlations among the evolving dynamics of different structural scales. Extensive experiments on four real-world datasets with different characteristics demonstrate its superior performance on both single-step and multi-step dynamic link prediction tasks. Further ablation study shows the effectiveness of our designs. A meaningful future direction is to investigate how to model multi-scale co-evolving dynamics in continuous time rather than in snapshots.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China under 2022YFF0606904, and the National Natural Science Foundation of China under U21B2036, U20B2060, 62272260.

REFERENCES

[1] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. 2006. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 44–54.

[2] Mireia Bolíbar. 2016. Macro, meso, micro: Broadening the 'social'of social network analysis with a mixed methods approach. *Quality & Quantity* 50, 5 (2016), 2217–2236.

[3] Lei Cai and Shuiwang Ji. 2020. A multi-scale approach for graph link prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 3308–3315.

[4] Damon Centola. 2010. The spread of behavior in an online social network experiment. *science* 329, 5996 (2010), 1194–1197.

[5] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. 2019. Hyperbolic graph convolutional neural networks. *Advances in neural information processing systems* 32 (2019).

[6] Huiyuan Chen and Jing Li. 2018. Exploiting structural and temporal evolution in dynamic link prediction. In *Proceedings of the 27th ACM International conference on information and knowledge management*. 427–436.

[7] Haochen Chen, Bryan Perozzi, Yifan Hu, and Steven Skiena. 2018. Harp: Hierarchical representation learning for networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[8] Jinyin Chen, Xueke Wang, and Xuanheng Xu. 2021. GC-LSTM: graph convolution embedded LSTM for dynamic network link prediction. *Applied Intelligence* (2021), 1–16.

[9] Yun Chi, Xiaodan Song, Dengyong Zhou, Koji Hino, and Belle L Tseng. 2007. Evolutionary spectral clustering by incorporating temporal smoothness. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 153–162.

[10] James S Coleman. 1994. *Foundations of social theory*. Harvard university press.

[11] Nur Nasuha Daud, Siti Hafizah Ab Hamid, Muntadher Saadoon, Firdaus Sahran, and Nor Badrul Anuar. 2020. Applications of link prediction in social networks: A review. *Journal of Network and Computer Applications* 166 (2020), 102716.

[12] Vijay Prakash Dwivedi, Chaitanya K Joshi, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. 2020. Benchmarking graph neural networks. *arXiv preprint arXiv:2003.00982* (2020).

[13] Hongyang Gao and Shuiwang Ji. 2019. Graph u-nets. In *international conference on machine learning*. PMLR, 2083–2092.

[14] 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*. 855–864.

[15] Chengbin Hou, Han Zhang, Shan He, and Ke Tang. 2020. Glodyne: Global topology preserving dynamic network embedding. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[16] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[17] Jérôme Kunegis. 2013. Konect: the koblenz network collection. In *Proceedings of the 22nd international conference on world wide web*. 1343–1350.

[18] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. 2019. Self-attention graph pooling. *arXiv preprint arXiv:1904.08082* (2019).

[19] Kevin Lewis, Marco Gonzalez, and Jason Kaufman. 2012. Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences* 109, 1 (2012), 68–72.

[20] Jia Li, Zhichao Han, Hong Cheng, Jiao Su, Pengyun Wang, Jianfeng Zhang, and Lujia Pan. 2019. Predicting path failure in time-evolving graphs. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1279–1289.

[21] 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 (2007), 1019–1031.

[22] Yuanfu Lu, Xiao Wang, Chuan Shi, Philip S Yu, and Yanfang Ye. 2019. Temporal network embedding with micro-and macro-dynamics. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 469–478.

[23] Noah Mark. 1998. Birds of a feather sing together. *Social forces* 77, 2 (1998), 453–485.

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

[25] Alan E Mislove. 2009. *Online social networks: measurement, analysis, and applications to distributed information systems*. Rice University.

[26] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 807–814.

[27] Zhenyu Qiu, Jia Wu, Wenbin Hu, Bo Du, Guocai Yuan, and Philip Yu. 2021. Temporal Link Prediction with Motifs for Social Networks. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[28] Ryan Rossi and Nesreen Ahmed. 2015. The network data repository with interactive graph analytics and visualization. In *Twenty-ninth AAAI conference on artificial intelligence*.

[29] Aravind Sankar, Yanhong Wu, Liang Gou, Wei Zhang, and Hao Yang. 2020. Dysat: Deep neural representation learning on dynamic graphs via self-attention networks. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 519–527.

[30] Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. 2002. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics* 31, 1 (2002), 64–68.

[31] Joakim Skarding, Bogdan Gabrys, and Katarzyna Musial. 2021. Foundations and Modeling of Dynamic Networks Using Dynamic Graph Neural Networks: A Survey. *IEEE Access* 9 (2021), 79143–79168.

[32] Aynaz Taheri, Kevin Gimpel, and Tanya Berger-Wolf. 2019. Learning to represent the evolution of dynamic graphs with recurrent models. In *Companion proceedings of the 2019 world wide web conference*. 301–307.

[33] Gergely Tibely, Lauri Kovanen, Marton Karsai, Kimmo Kaski, Janos Kertesz, and Jari Saramäki. 2011. Communities and beyond: mesoscopic analysis of a large social network with complementary methods. *Physical Review E* 83, 5 (2011), 056125.

[34] Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. 2019. Dyrep: Learning representations over dynamic graphs. In *International conference on learning representations*.

[35] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[36] Akanda Wahid-Ul-Ashraf, Marcin Budka, and Katarzyna Musial-Gabrys. 2017. Newton's gravitational law for link prediction in social networks. In *International Conference on Complex Networks and their Applications*. Springer, 93–104.

[37] Meihong Wang, Linling Qiu, and Xiaoli Wang. 2021. A survey on knowledge graph embeddings for link prediction. *Symmetry* 13, 3 (2021), 485.

[38] Shoujin Wang, Liang Hu, Yan Wang, Xiangnan He, Quan Z Sheng, Mehmet A Orgun, Longbing Cao, Francesco Ricci, and Philip S Yu. 2021. Graph learning based recommender systems: A review. *arXiv preprint arXiv:2105.06339* (2021).

[39] Yanbang Wang, Yen-Yu Chang, Yunyu Liu, Jure Leskovec, and Pan Li. 2021. Inductive representation learning in temporal networks via causal anonymous walks. *arXiv preprint arXiv:2101.05974* (2021).

[40] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2020. Inductive representation learning on temporal graphs. *arXiv preprint arXiv:2002.07962* (2020).

[41] Dingqi Yang, Bingqing Qu, Jie Yang, and Philippe Cudre-Mauroux. 2019. Revisiting user mobility and social relationships in lbsns: a hypergraph embedding approach. In *The world wide web conference*. 2147–2157.

[42] Menglin Yang, Min Zhou, Marcus Kalander, Zengfeng Huang, and Irwin King. 2021. Discrete-time temporal network embedding via implicit hierarchical learning in hyperbolic space. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1975–1985.

[43] Lekui Zhou, Yang Yang, Xiang Ren, Fei Wu, and Yueting Zhuang. 2018. Dynamic network embedding by modeling triadic closure process. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[44] Linhong Zhu, Dong Guo, Junming Yin, Greg Ver Steeg, and Aram Galstyan. 2016. Scalable temporal latent space inference for link prediction in dynamic social networks. *IEEE Transactions on Knowledge and Data Engineering* 28, 10 (2016), 2765–2777.