

# Spatial-Temporal Graph ODE Networks for Traffic Flow Forecasting

Zheng Fang\*

Key Laboratory of Machine Perception (Ministry of Education), Peking University  
fang\_z@pku.edu.cn

Guojie Song<sup>‡</sup>

Key Laboratory of Machine Perception (Ministry of Education), Peking University  
gjsong@pku.edu.cn

Qingqing Long\*<sup>†</sup>

Alibaba Group  
lantu.lqq@alibaba-inc.com

Kunqing Xie

Key Laboratory of Machine Perception (Ministry of Education), Peking University  
kunqing@cis.pku.edu.cn

## ABSTRACT

Spatial-temporal forecasting has attracted tremendous attention in a wide range of applications, and traffic flow prediction is a canonical and typical example. The complex and long-range spatial-temporal correlations of traffic flow bring it to a most intractable challenge. Existing works typically utilize shallow graph convolution networks (GNNs) and temporal extracting modules to model spatial and temporal dependencies respectively. However, the representation ability of such models is limited due to: (1) shallow GNNs are incapable to capture long-range spatial correlations, (2) only spatial connections are considered and a mass of semantic connections are ignored, which are of great importance for a comprehensive understanding of traffic networks. To this end, we propose Spatial-Temporal Graph Ordinary Differential Equation Networks (STGODE).<sup>1</sup> Specifically, we capture spatial-temporal dynamics through a tensor-based ordinary differential equation (ODE), as a result, deeper networks can be constructed and spatial-temporal features are utilized synchronously. To understand the network more comprehensively, semantical adjacency matrix is considered in our model, and a well-design temporal dilated convolution structure is used to capture long term temporal dependencies. We evaluate our model on multiple real-world traffic datasets and superior performance is achieved over state-of-the-art baselines.

## CCS CONCEPTS

• **Information systems** → **Spatial-temporal systems**; • **Networks** → *Network structure*.

\*These authors contributed equally to the work.

<sup>†</sup>Work performed as a student of Peking University.

<sup>‡</sup>Corresponding Author.

<sup>1</sup>Codes are available at <https://github.com/square-coder/STGODE>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*KDD '21, August 14–18, 2021, Virtual Event, Singapore*

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8332-5/21/08...\$15.00

<https://doi.org/10.1145/3447548.3467430>

## KEYWORDS

Spatial Temporal Forecasting; Graph Neural Network; Neural ODE

### ACM Reference Format:

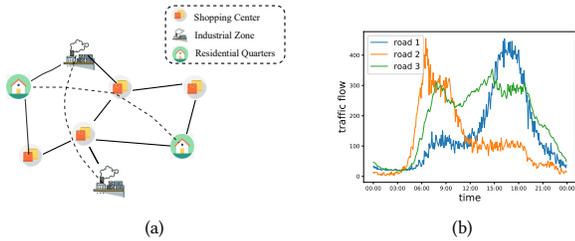
Zheng Fang, Qingqing Long, Guojie Song, and Kunqing Xie. 2021. Spatial-Temporal Graph ODE Networks for Traffic Flow Forecasting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3447548.3467430>

## 1 INTRODUCTION

Spatial-temporal forecasting has been widely studied in recent years. It has large scale applications in our daily life, such as traffic flow forecasting [8, 10], climate forecasting [3, 15], urban monitoring system analysis [22] and so on. For this reason, accurate spatial-temporal forecasting plays a significant role in improving the service quality of these applications. In this paper, we study one of the most representative in spatial-temporal forecasting, traffic flow forecasting, which is an indispensable component in Intelligent Transportation System (ITS). Traffic flow forecasting attempts to predict the future traffic flow given historical traffic conditions and underlying road networks.

This task is challenging principally due to the complex and long-range spatial-temporal dependencies in traffic networks. As an intrinsic phenomenon of traffic, the travel distances of different people vary a lot [25], which means that nearby and distant spatial dependencies largely exist at the same time. As Fig 1(a) shows, a node is not only connected to its geographical neighbors but also distant relevant nodes. Furthermore, traffic flow series exhibit diversified temporal pattern for their distinct behavior attributes as Fig 1(b) shows. Moreover, when the spatial attributes and temporal patterns are united, the complex interactions in between leading to an intractable problem for traffic flow forecasting.

Graph Neural Networks (GNNs) for traffic forecasting have attracted tremendous attention in recent years. Owing to its strong ability to deal with graph-structured data, GNN enables to update node representations by aggregating representations from their neighbors, whereby GNN yields effective and efficient performance in various tasks like node classification and graph classification [11, 16, 19, 21]. A large number of works have been proposed to utilize GNNs to extract spatial features in traffic networks, STGCN



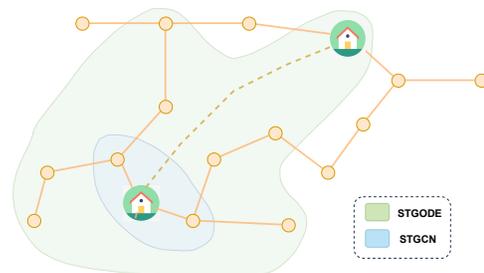
**Figure 1: (a) shows the geographical and semantic connections of nodes. (b) shows examples of traffic flow with diverse patterns, like morning peak, evening peak and relatively steady patterns.**

[32] and DCRNN [18] are the representative. Most of them combine GNNs with RNNs to obtain spatial representations and temporal representation respectively [24, 35], and multiple works improve recurrent structure with convolution structure for better training stability and efficiency [9, 34].

However, there are two problems that have been persistently neglected. On the one hand, most methods model spatial patterns and temporal patterns separately without considering their interactions, which restricts the representation ability of the models a lot. On the other hand, neural networks generally perform better with the stack of more layers, while GNNs benefit little from the depth. On the contrary, the best results are achieved when two-layer graph neural networks are cascaded, and more layers may lead to inferior performance in practice [17, 36]. Ordinary GNNs have been proved to suffer from the over-smoothing problem, i.e. all node representations will converge to the same value with deeper layers. Such drawbacks severely limit the depth of GNNs and make it hardly possible to obtain deeper and richer spatial features. However, to the best of our knowledge, there are few works considering network depth in spatial-temporal forecasting, which is of great importance for capturing long-range dependencies.

In our Spatial-Temporal Graph Ordinary Differential Equation Network (STGODE), several components are elaborately designed to tackle the aforementioned problems. First, in order to depict spatial correlations from both geographical and semantic views, we construct two types of adjacency matrices, i.e. spatial adjacency matrix and semantic adjacency matrix, based on spatial connectivity and semantical similarity of traffic flow respectively. Second, motivated by residual networks [12], residual connections are added between layers to alleviate the over-smoothing problem. Furthermore, it is proved that the discrete layers with residual connections can be viewed as a discretization of an Ordinary Differential Equation (ODE) [5], and so a continuous graph neural network (CGNN) is derived [31]. Here in this paper, a continuous GNN with residual connections is introduced to avoid the over-smoothing problem and hence be able to model long-range spatial-temporal dependencies. Last but not least, a spatial-temporal tensor is constructed to consider spatial and temporal patterns simultaneously and model complex spatial-temporal interactions. We present the superiority of our model with a toy example. As Fig 2 shows, compared with STGCN, STGODE possesses a wider receptive field and thus can

adjust outputs according to shifting circumstances to achieve better performance.



**Figure 2: A performance schematic of STGODE**

Our main contributions are summarized as follows,

- We propose a novel continuous representation of GNNs in tensor form for traffic flow forecasting, which breaks through the limit of network depth and improves the capacity of extracting longer-range spatial-temporal correlations, and a theoretical analysis is given in detail.
- We utilize both spatial neighbors and semantical neighbors of road nodes to consider spatial correlations comprehensively.
- Extensive experiments are conducted on real-world traffic datasets, and the results show that our model outperforms existing baseline models.

## 2 RELATED WORK

### 2.1 Traffic Flow Forecasting

In recent years a large body of research has been conducted on traffic flow forecasting, which has always been a critical problem in intelligent transportation systems(ITS)[23]. Traffic flow forecasting can be viewed as a spatial-temporal forecasting task leveraging spatial-temporal data collected by various sensors to predict future traffic conditions. Classic methods, including autoregressive integrated moving average (ARIMA), k-nearest neighbors algorithm (kNN), and support vector machine (SVM), can only take temporal information into account, without considering spatial features.[14, 28, 29]. Due to the limitation of modeling complex spatial-temporal relationships with classical methods, deep neural network models are proposed, which have been widely used in various challenging traffic prediction tasks. Specifically, FC-LSTM combines CNN and LSTM to model spatial and temporal relations through an extended fully-connected LSTM with embedded convolutional layers [26]. ST-ResNet utilizes a deep residual CNN network to predict citywide crowd flow [33], where the strong power of the residual network is exhibited. Despite impressive results that have been achieved, all above-mentioned methods are designed for grid data, thus not suitable for the traffic scene with graph-structured data.

### 2.2 Graph Neural Networks

GNN is an effective framework for the representation learning of graphs. GNNs follow a neighborhood aggregation scheme, where

the computation of node representation is carried out by sampling and aggregating features of neighboring nodes [11, 16, 20]. Strenuous efforts have been made to utilize graph convolution methods in traffic forecasting considering that traffic data is a classic kind of non-Euclidean structured graph data. For example, DCRNN [18] view the traffic flow as a diffusion process and captures the spatial dependency with bidirectional random walks on a directed graph. STGCN [32] builds a model with complete convolutional structures on both spatial and temporal view, which enables faster training speed with fewer parameters. ASTGCN [10] introduces attention mechanism to capture dynamics of spatial dependencies and temporal correlations. All these methods use two separate components to capture temporal and spatial dependencies respectively instead of simultaneously, thus STSGCN [27] makes attempts to incorporate spatial and temporal blocks altogether through an elaborately designed spatial-temporal synchronous modeling mechanism.

Long-range spatial-temporal relationship, as a common-sense in traffic circumstances, is expected to be explored with deeper neural networks. However, the over-smoothing phenomenon of deep GNNs, which has been proved in a great number of studies [17, 36], will lead to similar node representations. Thus the depth of GNNs is restricted, and the long-range dependencies between nodes are largely ignored.

### 2.3 Continuous GNNs

Neural Ordinary Differential Equation(ODE) [5] models a continuous dynamic system based on parameterizing the derivative of the hidden state using a neural network, instead of specifying discrete sequences of hidden layers. CGNN [31] first extends this method to graph-structured data, which develops a continuous message-passing layer through defining the derivatives as combined representations of current and initial nodes. The key factor for alleviating the over-smoothing effect is the use of restart distribution, which motivates us in this paper. With proving simple GCN as a discretization of a kind of ODE, they characterize the continuous dynamics of node representations and enable deeper networks. To the best of our knowledge, there are no works about graph ODE in spatial-temporal forecasting.

## 3 PRELIMINARIES

**Definition 1.** (Traffic network  $\mathcal{G}$ ) We represent the road network as a graph  $\mathcal{G} = (V, E, A)$ , where  $V$  is a set of  $N$  nodes;  $E$  is a set of edges;  $A \in \mathbb{R}^{N \times N}$  is an adjacency matrix. Here in this paper, two kinds of adjacency matrix are adopted, spatial adjacency matrix  $A^{sp}$  and semantic adjacency matrix  $A^{se}$ .

**Definition 2.** (Graph signal tensor  $\mathcal{X}$ ) We use  $\mathbf{x}_t^i \in \mathbb{R}^F$  to denote the observation of node  $i$  at time  $t$ , and  $F$  is the length of an observation vector.  $X_t = (\mathbf{x}_t^1, \mathbf{x}_t^2, \dots, \mathbf{x}_t^N) \in \mathbb{R}^{N \times F}$  denotes the observations of all nodes at time  $t$ .  $\mathcal{X} = (X_1, X_2, \dots, X_T) \in \mathbb{R}^{T \times N \times F}$  denotes the observations of all nodes at all time.

### 3.1 Problem Formulation

Given the tensor  $\mathcal{X}$  observed on a traffic network  $\mathcal{G}$ , the goal of traffic forecasting is to learn a mapping function  $f$  from the historical

$T$  observations to predict future  $T'$  traffic observations,

$$[X_{t-T+1}, X_{t-T+2}, \dots, X_t; \mathcal{G}] \xrightarrow{f} [X_{t+1}, X_{t+2}, \dots, X_{t+T'}].$$

### 3.2 Regularized adjacency matrix

Given an adjacency matrix  $A \in \mathbb{R}^{N \times N}$ , we typically normalize it as  $\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ , where  $D$  is the degree matrix of  $A$ .  $\tilde{A}$  has an eigenvalue decomposition [6] and the eigenvalues are in the interval  $[-1, 1]$ . Negative eigenvalues can lead to unstable training process, thus a self-loop is commonly added to avoid it. The regularized form [16] of  $\tilde{A}$  is adopted in this paper:

$$\hat{A} = \frac{\alpha}{2} \left( I + D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \right), \quad (1)$$

where  $\alpha \in (0, 1)$  is a hyperparameter, as a result the eigenvalues of  $\hat{A}$  are in the interval  $[0, \alpha]$ .

### 3.3 Neural ODE

We consider a continuous-time(depth) model,

$$\mathbf{x}(t) = \mathbf{x}(0) + \int_0^t \frac{d\mathbf{x}}{d\tau} d\tau = \mathbf{x}(0) + \int_0^t f(\mathbf{x}(\tau), \tau) d\tau, \quad (2)$$

where  $f(\mathbf{x}(\tau), \tau)$  will be parameterised by a neural network to model the hidden dynamic. We can backpropagate the process through an ODE solver without any internal operations [5], which allows us to build it just as a block for the whole neural network.

### 3.4 Tensor Calculation

A tensor  $\mathcal{T}$  can be viewed as a multidimensional array, and a tensor-matrix multiplication is defined on some mode fiber, for example,

$$(\mathcal{T} \times_2 M)_{ilk} = \sum_{j=1}^{n_2} \mathcal{T}_{ijk} \cdot M_{jl}, \quad (3)$$

where  $\mathcal{T} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ ,  $M \in \mathbb{R}^{n_2 \times n'_2}$ ,  $\mathcal{T} \times_2 M \in \mathbb{R}^{n_1 \times n'_2 \times n_3}$ ,  $\times_2$  denotes that the tensor-matrix multiplication is conducted on mode-2, i.e. the second subscript. There are some mathematical properties about tensor-matrix multiplication which will be used in this paper,

- $\mathcal{T} \times_i M_1 \times_i M_2 = \mathcal{T} \times_i (M_1 M_2)$
- $\mathcal{T} \times_i M_1 \times_j M_2 = \mathcal{T} \times_j M_2 \times_i M_1$  ( $i \neq j$ ).

Above properties can be easily proved with Eq 3 through the multiplication rule.

## 4 MODEL

Figure 3(a) shows the overall framework of our proposed model, i.e. Spatial-Temporal Graph ODE. It mainly consists of three components, two Spatial-Temporal Graph ODE (STGODE) layers composed of multiple STGODE blocks, a max-pooling layer, and an output layer. A STGODE block consists of two temporal dilation convolution (TCN) blocks and a tensor-based ODE solver in between, which is applied to capture complex and long-range spatial-temporal relationships simultaneously. The spatial adjacency matrix and the semantic adjacency matrix will be fed into the solver separately to obtain features from different levels. The details of the model will be described in the following section.

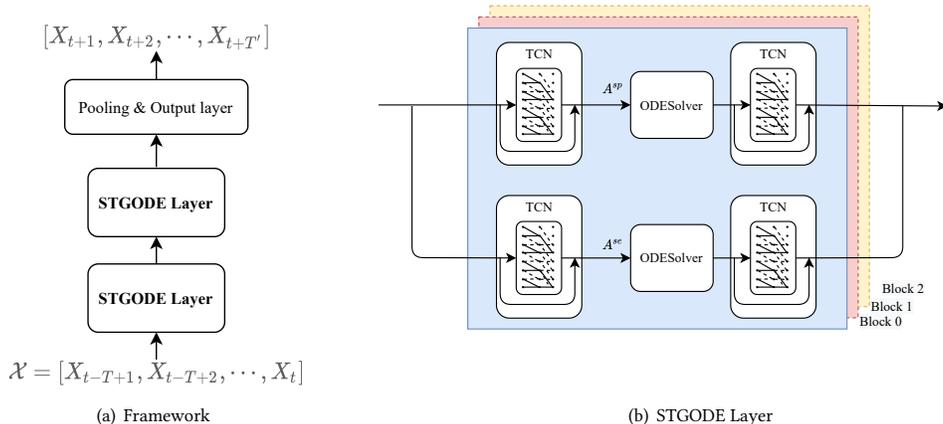


Figure 3: (a) is the framework of the STGODE network. Several STGODE blocks in parallel constitute a STGODE layer, and two STGODE layers are cascaded to extract higher-order features. (b) is the detail of STGODE blocks, where an ODE solver is sandwiched between two TCNs with residual connections and two kinds of adjacency matrices are utilized for more comprehensive characterization.

#### 4.1 Adjacency Matrix Construction

Two kinds of adjacency matrix are leveraged in our model. Following STGCN [32], the spatial adjacency matrix is defined as

$$A_{ij}^{sp} = \begin{cases} \exp\left(-\frac{d_{ij}^2}{\sigma^2}\right), & \text{if } \exp\left(-\frac{d_{ij}^2}{\sigma^2}\right) \geq \epsilon, \\ 0, & \text{otherwise} \end{cases}, \quad (4)$$

where  $d_{ij}$  is the distance between node  $i$  and node  $j$ .  $\sigma^2$  and  $\epsilon$  are thresholds to control sparsity of matrix  $A^{sp}$ .

Besides, contextual similarities between nodes provide a wealth of information and should be taken into consideration. For example, similar traffic patterns are shared among roads near commercial areas regardless of the remote geographical distance, while such correlations cannot be revealed in spatial graph. To capture above mentioned semantic correlations, the Dynamic Time Warping (DTW) algorithm is applied to calculate the similarity of two time series [1], which is superior to other metric methods on account of its sensitivity to shape similarity rather than point-wise similarity. As shown in Fig 4, the point  $a$  of series  $X$  will be related to the point  $c$  but not  $b$  of series  $Y$  by the DTW algorithm. Specifically, given two time series  $X = (x_1, x_2, \dots, x_m)$  and  $Y = (y_1, y_2, \dots, y_n)$ , DTW is a dynamic programming algorithm defined as

$$D(i, j) = \text{dist}(x_i, y_j) + \min(D(i-1, j), D(i, j-1), D(i-1, j-1)), \quad (5)$$

where  $D(i, j)$  represents the shortest distance between subseries  $X = (x_1, x_2, \dots, x_i)$  and  $Y = (y_1, y_2, \dots, y_j)$ , and  $\text{dist}(x_i, y_j)$  is some distance metric like absolute distance. As a result,  $\text{DTW}(X, Y) = D(m, n)$  is set as the final distance between  $X$  and  $Y$ , which better reflects the similarity of two time series compared to the Euclidean distance.

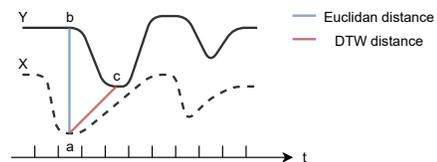


Figure 4: An example of the difference between the Euclidean distance and the DTW distance.

Accordingly, we define the semantic adjacency matrix through the DTW distance as following,

$$A_{ij}^{SE} = \begin{cases} 1, & \text{DTW}(X^i, X^j) < \epsilon \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

where  $X^i$  denotes time series of node  $i$ , and  $\epsilon$  determine the sparsity of the adjacency matrix.

#### 4.2 Tensor-based Spatial-Temporal Graph ODE

GNNs update embeddings of nodes through aggregating features of their own and neighbors with a graph convolution operation. The classic form of convolution operation can be formulated as:

$$H_{l+1} = \text{GCN}(H_l) = \sigma(\hat{A}H_lW), \quad (7)$$

where  $H_l \in \mathbb{R}^{N \times C}$  denotes the input of the  $l$ -th graph convolutional layer,  $\hat{A} \in \mathbb{R}^{N \times N}$  is the normalized adjacency matrix, and  $W \in \mathbb{R}^{C \times C'}$  is a learnable parameter matrix, which models the interaction among different features. However, such GNNs have been proved to suffer from over-smoothing problem when networks go deeper [17, 36], which largely limits the capacity of modeling long-range dependencies. For this reason, our STGODE block is proposed.

A discrete version is first shown as:

$$\mathcal{H}_{l+1} = \mathcal{H}_l \times_1 \hat{A} \times_2 U \times_3 W + \mathcal{H}_0, \quad (8)$$

where  $\mathcal{H}_l \in \mathbb{R}^{N \times T \times F}$  is a spatial-temporal tensor representing nodes' hidden embedding of the  $l$ -th layer,  $\times_i$  denotes the tensor-matrix multiplication on mode  $i$ ,  $\hat{A}$  is the regularized adjacency matrix,  $U$  is the temporal transform matrix,  $W$  is the feature transform matrix, and  $\mathcal{H}_0$  denotes the initial input of GNN, which can be acquired through another neural network. Different from existing works, we treat the spatial-temporal tensor as input and hence enable to handle spatial information and temporal information simultaneously. The intricate spatial-temporal correlation is coupled through tensor multiplication on each mode. Motivated by CGNN [31], a restart distribution  $\mathcal{H}_0$  is involved to alleviate the over-smoothing problem.

Specifically, the expansion of Eq 8 is shown as below,

$$\mathcal{H}_l = \sum_{i=0}^l \left( \mathcal{H}_0 \times_1 \hat{A}^i \times_2 U^i \times_3 W^i \right), \quad (9)$$

where we can see clearly that the output representation  $\mathcal{H}_l$  aggregates information from all layers, that's to say, the final outputs collect information from all no more than  $l$ -order neighbors without losing initial features. To show the necessity of the restart distribution, let's suppose another version without  $\mathcal{H}_0$  like

$$\mathcal{H}_{l+1} = \mathcal{H}_l \times_1 \hat{A} \times_2 U \times_3 W,$$

where the final output will be

$$\mathcal{H}_n = \mathcal{H}_0 \times_1 \hat{A}^n \times_2 U^n \times_3 W^n.$$

Take  $\hat{A}$  as a simple example, assuming  $\hat{A}$  has an eigenvalue decomposition as  $\hat{A} = P\Lambda P^T$ , where  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m)$  is a diagonal matrix. Obviously,

$$\begin{aligned} \hat{A}^n &= P \text{diag}(\lambda_1^n, \lambda_2^n, \dots, \lambda_m^n) P^T \\ &= \lambda_1^n P \text{diag}\left(1, \left(\frac{\lambda_2}{\lambda_1}\right)^n, \dots, \left(\frac{\lambda_m}{\lambda_1}\right)^n\right) P^T \\ &\rightarrow \lambda_1^n P \text{diag}(1, 0, \dots, 0) P^T \end{aligned} \quad (10)$$

when  $n$  goes to infinity with  $\lambda_1 > \lambda_2 > \dots > \lambda_m$ . The diagonal elements converge to zero except the largest one, which causes much loss of information.

Such residual structure as Eq 8 is powerful but tough to train due to its enormous amount of parameters, thus we aim to extend the discrete formulation to a continuous expression. Intuitively, we replace  $n$  with a continuous variable  $t$ , and view the expansion equation as a Riemann sum from 0 to  $n$  on  $i$ , which is,

$$\begin{aligned} \mathcal{H}_n &= \sum_{i=0}^n \left( \mathcal{H}_0 \times_1 \hat{A}^i \times_2 U^i \times_3 W^i \right) \\ &= \sum_{i=1}^{n+1} \left( \mathcal{H}_0 \times_1 \hat{A}^{(i-1) \times \Delta t} \times_2 U^{(i-1) \times \Delta t} \times_3 W^{(i-1) \times \Delta t} \right) \end{aligned} \quad (11)$$

where  $\Delta t = \frac{t+1}{n+1}$  with  $t = n$ . When  $n$  goes to  $\infty$ , we can formulate the following integral:

$$\mathcal{H}(t) = \int_0^{t+1} \mathcal{H}_0 \times_1 \hat{A}^\tau \times_2 U^\tau \times_3 W^\tau d\tau, \quad (12)$$

The critical point here is to transform the residual structure to an ODE structure, obviously we already have an ordinary differential

equation given by

$$\frac{d\mathcal{H}(t)}{dt} = \mathcal{H}_0 \times_1 \hat{A}^{t+1} \times_2 U^{t+1} \times_3 W^{t+1}, \quad (13)$$

but  $\hat{A}^{t+1}, U^{t+1}, W^{t+1}$  are intractable to compute especially when  $t$  is a non-integer. Motivated by the work in [31], we have the following corollary.

**Corollary 1.** The discrete update in Eq 8 is a discretization of following ODE:

$$\frac{d\mathcal{H}(t)}{dt} = \mathcal{H}(t) \times_1 \ln \hat{A} + \mathcal{H}(t) \times_2 \ln U + \mathcal{H}(t) \times_3 \ln W + \mathcal{H}_0, \quad (14)$$

where  $\mathcal{H}_0 = f(\mathcal{X})$  is the output of upstream networks.

**PROOF.** Starting from Eq 13, we consider the second-derivative of  $\mathcal{H}(t)$  through derivative rules,

$$\frac{d^2\mathcal{H}(t)}{dt^2} = \frac{d\mathcal{H}(t)}{dt} \times_1 \ln \hat{A} + \frac{d\mathcal{H}(t)}{dt} \times_2 \ln U + \frac{d\mathcal{H}(t)}{dt} \times_3 \ln W \quad (15)$$

Then integrate over  $t$  in both sides of the above equation, we can get:

$$\frac{d\mathcal{H}(t)}{dt} = \mathcal{H}(t) \times_1 \ln \hat{A} + \mathcal{H}(t) \times_2 \ln U + \mathcal{H}(t) \times_3 \ln W + \text{const} \quad (16)$$

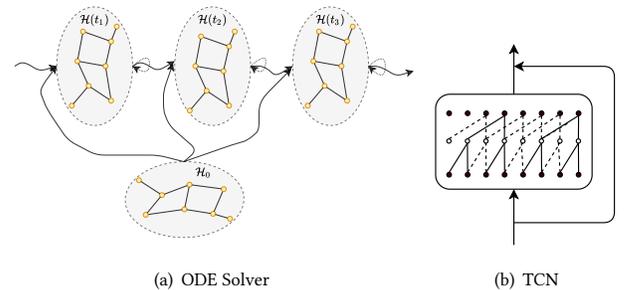
To solve the *const*, we can put Eq 13 and Eq 16 together, thus we have:

$$\begin{aligned} \text{const} &= \mathcal{H}_0 \times_1 \hat{A}^{t+1} \times_2 U^{t+1} \times_3 W^{t+1} \\ &\quad - \left( \mathcal{H}(t) \times_1 \ln \hat{A} + \mathcal{H}(t) \times_2 \ln U + \mathcal{H}(t) \times_3 \ln W \right). \end{aligned} \quad (17)$$

By letting  $t \rightarrow -1$  mathematically, we can easily get  $\text{const} = \mathcal{H}_0$ . So the corollary is proved.  $\square$

In practice, we can approximate the logarithm operation with its first order of Taylor expansion, i.e.  $\ln M \approx M - I$ . As a result, a simpler form is obtained,

$$\frac{d\mathcal{H}(t)}{dt} = \mathcal{H}(t) \times_1 (\hat{A} - I) + \mathcal{H}(t) \times_2 (U - I) + \mathcal{H}(t) \times_3 (W - I) + \mathcal{H}_0 \quad (18)$$



**Figure 5: (a) is the illustration of an ODE solver, which shows that the derivation of the hidden states is a function of current states and initial states. (b) represents the structure of TCN, which consists of a dilated convolution and a residual connection.**

The above ODE we deduced can be analytically solved as the following corollary.

**Corollary 2.** The analytic solution of the Eq 18 is given by

$$\begin{aligned} \mathcal{H}(t) = & \mathcal{H}_0 \times_1 e^{(\hat{A}-I)t} \times_2 e^{(U-I)t} \times_3 e^{(W-I)t} \\ & + \int_0^t \mathcal{H}_0 \times_1 e^{(\hat{A}-I)(t-s)} \times_2 e^{(U-I)(t-s)} \times_3 e^{(W-I)(t-s)} ds \end{aligned} \quad (19)$$

PROOF. Suppose

$$\mathcal{H}^*(t) = \mathcal{H}(t) \times_1 e^{(\hat{A}-I)t} \times_2 e^{(U-I)t} \times_3 e^{(W-I)t}, \quad (20)$$

then we have

$$\frac{d\mathcal{H}^*(t)}{dt} = \mathcal{H}_0 \times_1 e^{(\hat{A}-I)t} \times_2 e^{(U-I)t} \times_3 e^{(W-I)t}, \quad (21)$$

and this is derived from Eq 18. Integrate Eq 21 on both sides, and we can get the following result,

$$\mathcal{H}^*(t) = \mathcal{H}_0^* + \int_0^t \mathcal{H}_0 \times_1 e^{(\hat{A}-I)\tau} \times_2 e^{(U-I)\tau} \times_3 e^{(W-I)\tau} d\tau, \quad (22)$$

and hence  $\mathcal{H}(t)$  can be formulated as

$$\begin{aligned} \mathcal{H}(t) = & \mathcal{H}_0 \times_1 e^{(\hat{A}-I)t} \times_2 e^{(U-I)t} \times_3 e^{(W-I)t} \\ & + \int_0^t \mathcal{H}_0 \times_1 e^{(\hat{A}-I)(t-\tau)} \times_2 e^{(U-I)(t-\tau)} \times_3 e^{(W-I)(t-\tau)} d\tau \end{aligned} \quad (23)$$

□

In fact, the last integration can be solved further, but limited by space, we put it in the supplementary. Notice that the eigenvalues of  $\hat{A} - I$  is in the interval  $[-1, 0)$ , as a result,  $e^{(\hat{A}-I)t}$  will go to zero when  $t$  goes to  $\infty$ . However, unrestricted  $U$  and  $W$  will lead to divergent integrations as  $t$  goes to  $\infty$ . To enforce  $U$  and  $W$  to be a diagonalizable matrix with all the eigenvalues less than 1, we follow previous work [7] to parameterise  $U$  and  $W$  as  $U = P \text{diag}(\lambda) P^T$  and  $W = Q \text{diag}(\mu) Q^T$  respectively, where  $P$  and  $Q$  are learnable orthogonal matrices,  $\lambda$  and  $\mu$  are learnable vectors whose elements will be clamped to the interval  $(0, 1)$ .

So far, we have proved a continuous form of tensor-based hidden representation theoretically. Motivated by Neural ODE [5], we propose an STGODE learning framework,

$$\mathcal{H}(t) = \text{ODESolve} \left( \frac{d\mathcal{H}(t)}{dt}, \mathcal{H}_0, t \right), \quad (24)$$

where

$$\frac{d\mathcal{H}(t)}{dt} = \mathcal{H}(t) \times_1 (\hat{A} - I) + \mathcal{H}(t) \times_2 (U - I) + \mathcal{H}(t) \times_3 (W - I) + \mathcal{H}_0,$$

$\mathcal{H}_0$  denotes the initial value, which comes from the upstream network and the ODESolver is chosen as the Euler solver in our model.

### 4.3 Temporal Convolutional Blocks

Besides spatial correlations among different nodes, the long-term temporal correlations of the nodes themselves also matter. Although RNN-based models, like LSTM and GRU, are widely applied in time-series analysis, recurrent networks still suffer from some intrinsic drawbacks like time-consuming iterations, unstable gradients, and delayed responses to dynamic changes.

To enhance the performance of extracting long term temporal dependencies, a 1-D dilated temporal convolutional network along the time axis is adopted here.

$$H_{tcn}^l = \begin{cases} X & , l = 0 \\ \sigma(W^l *_d H_{tcn}^{l-1}) & , l = 1, 2, \dots, L \end{cases} \quad (25)$$

where  $X \in \mathbb{R}^{N \times T \times F}$  is the input of TCN,  $H_{tcn}^l \in \mathbb{R}^{N \times T \times F}$  is the output of the  $l$ -th layer of TCN, and  $W^l$  denotes the  $l$ -th convolution kernel. To expand the receptive field, an exponential dilation rate  $d^l = 2^{l-1}$  is adopted in temporal convolution. In the process, zero-padding strategy is utilized to keep time sequence length unchanged. What's more, a residual structure [12] is added to strengthen convolution performance as shown in Fig 5(b).

### 4.4 STGODE Layer

In this part, the overall STGODE layer is presented in detail. As illustrated in Fig 3(b), the "sandwich" structure is adopted which consist of two TCN blocks and a STGODE solver. Such structure enables flexible and sensible spatial-temporal information flows, and all-convolution structures have the superiority of fast training and parallelization. Stacked "sandwiches" further extend the model's ability to discover complex correlations.

In the construction of the model, we deploy two kinds of STGODE blocks, which accept different adjacency matrices, i.e. the spatial adjacency matrix and the semantic adjacency matrix. Two kinds of adjacency matrices are utilized to combine local dynamics and semantical relationships altogether, which greatly enhance the representation ability. Multiple blocks are deployed in parallel so that more complicated and multi-level correlations can be captured.

### 4.5 Others

After the STGODE layers, a max-pooling operation is carried out to aggregate information from different blocks selectively. Finally, a two-layer MLP is designed as the output layer to transform the output of the max-pooling layer to the final prediction.

Huber loss is selected as the loss function since it is less sensitive to outliers than the squared error loss [13],

$$L(Y, \hat{Y}) = \begin{cases} \frac{1}{2}(Y - \hat{Y})^2 & , |Y - \hat{Y}| \leq \delta \\ \delta|Y - \hat{Y}| - \frac{1}{2}\delta^2 & , \text{otherwise} \end{cases} \quad (26)$$

where  $\delta$  is a hyperparameter which controls the sensitivity to outliers.

## 5 EXPERIMENTS

### 5.1 Datasets

We verify the performance of STGODE on six real-world traffic datasets, PeMSD7(M), PeMSD7(L), PeMS03, PeMS04, PeMS07, and PeMS08, which are collected by the Caltrans Performance Measurement System(PeMS) in real time every 30 seconds[4]. The traffic data are aggregated into 5-minutes intervals, which means there are 288 time steps in the traffic flow for one day. The system has more than 39,000 detectors deployed on the highway in the major metropolitan areas in California. There are three kinds of traffic

measurements contained in the raw data, including traffic flow, average speed, and average occupancy.

Specifically, PeMSD3 has 358 sensors, and the time span of it is from September to November in 2018, including 91 days in total. PeMSD7(M) and PeMSD7(L) are two datasets selected from District 7 of California, which contains 288 and 1,026 sensors respectively. The time range of PeMSD7 is in the weekdays of May and June of 2012. And PeMSD8 is collected from July to August in 2016, which contains 170 sensors. The detail of datasets is listed in Table 1. Z-score normalization is applied to the input data, i.e. removing the mean and scaling to unit variance.

Datasets	#Sensors	#Edges	Time Steps
PeMSD7(M)	228	1132	12672
PeMSD7(L)	1026	10150	12672
PeMS03	358	547	26208
PeMS04	307	340	16992
PeMS07	883	866	28224
PeMS08	170	295	17856

**Table 1: Datasets description**

## 5.2 Baselines

We compare STODE with following baseline models:

- **ARIMA** [2]: Auto-Regressive Integrated Moving Average model, which is a well-known statistical model of time series analysis.
- **STGCN** [32]: Spatio-Temporal Graph Convolution Network, which utilizes graph convolution and 1D convolution to capture spatial dependencies and temporal correlations respectively.
- **DCRNN** [18]: Diffusion Convolution Recurrent Neural Network, which integrates graph convolution into an encoder-decoder gated recurrent unit.
- **GraphWaveNet** [30]: Graph WaveNet, which combines adaptive graph convolution with dilated casual convolution to capture spatial-temporal dependencies.
- **ASTGCN(r)** [10]: Attention based Spatial Temporal Graph Convolutional Networks, which utilize spatial and temporal attention mechanisms to model spatial-temporal dynamics respectively. In order to keep the fairness of comparison, only recent components of modeling periodicity are taken.
- **STSGCN** [27]: Spatial-Temporal Graph Synchronous Graph Convolutional Networks, which utilize multiple localized spatial-temporal subgraph modules to synchronously capture the localized spatial-temporal correlations directly.

## 5.3 Experimental Settings

We split all datasets with a ratio 6: 2: 2 into training sets, validation sets, and test sets. One hour of historical data is used to predict traffic conditions in the next 60 minutes.

All experiments are conducted on a Linux server(CPU: Intel(R) Xeon(R) CPU E5-2682 v4 @ 2.50GHz, GPU: NVIDIA TESLA V100 16GB). The hidden dimensions of TCN blocks are set to 64, 32, 64, and 3 STGODE blocks are contained in each layer. The regularized

hyperparameter  $\alpha$  is set to 0.8, the thresholds  $\sigma$  and  $\epsilon$  of the spatial adjacency matrix are set to 10 and 0.5 respectively, and the threshold  $\epsilon$  of the semantic adjacency matrix is set to 0.6.

We train our model using Adam optimizer with a learning rate of 0.01. The batch size is 32 and the training epoch is 200. Three kinds of evaluation metrics are adopted, including root mean squared errors(RMSE), mean absolute errors(MAE), and mean absolute percentage errors(MAPE).

## 5.4 Experimental Results and Analysis

Table 2 shows the results of our and competitive models for traffic flow forecasting. Our STGODE model is obviously superior to the baselines. Specifically, deep learning methods achieve better results than traditional statistical methods, as traditional methods like ARIMA only take temporal correlations into consideration and ignore spatial dependencies, whereas deep learning models can take advantage of spatial-temporal information. Among the deep learning baselines, all except STSGCN utilize two modules to model spatial dependencies and temporal correlations respectively, which overlook complex interactions between spatial information and temporal information, and STSGCN hence surpasses other models. But STSGCN only concentrates on localized spatial-temporal correlations, and turns turtle in global dependencies.

Our model yields the best performance regarding all the metrics for all datasets, which suggests the effectiveness of our spatial-temporal dependency modeling. The result can be attributed to three aspects:

- (1) We utilize a tensor-based ODE framework to extract longer-range spatial-temporal dependencies;
- (2) The semantical neighbors are introduced to establish global and comprehensive spatial relationships;
- (3) Temporal dilated convolution networks with residual connections help to capture long term temporal dependencies.

## 5.5 Case Study

Here we select two nodes from the road network to carry out a case study. As Fig 6 shows, the prediction results of STGODE are remarkably closer to the ground truth than STGCN [32]. In normal circumstances, the model generates a smooth prediction ignoring small oscillations to fight against noise. But when an abrupt change arises, our model enables a rapid response to it. This is because STGODE is able to utilize feature information from longer range geographical neighbors and semantic neighbors, which helps to accurately capture real-time dynamics and filter invalid information, while STGCN as a shallow network, is susceptible to few nearby neighbors and thus performs unstably.

## 5.6 Model Analysis

**5.6.1 Ablation Experiments.** To verify the effectiveness of different modules of STGODE, we conduct the following ablation experiments on PeMS04 dataset, and four variants of STGODE are designed.

- **STGCN\***: The ODE solver is replaced with a graph convolution layer to verify the effectiveness of ODE structures for extracting long-range dependencies.

Dataset	Metric	ARIMA	STGCN	DCRNN	ASTGCN(r)	GraphWaveNet	STSGCN	STODE
PeMSD7(M)	RMSE	13.20	7.55	7.18	6.87	6.24	5.93	<b>5.66</b>
	MAE	7.27	4.01	3.83	3.61	3.19	3.01	<b>2.97</b>
	MAPE	10.38	9.67	9.81	8.84	8.02	7.55	<b>7.36</b>
PeMSD7(L)	RMSE	12.39	8.28	8.33	7.64	7.09	6.88	<b>5.98</b>
	MAE	7.51	4.84	4.33	4.09	3.75	3.61	<b>3.22</b>
	MAPE	15.83	11.76	11.41	10.25	9.41	9.13	<b>7.94</b>
PeMS03	RMSE	47.59	30.42	30.31	29.56	32.77	29.21	<b>27.84</b>
	MAE	35.41	17.55	17.99	17.34	19.12	17.48	<b>16.50</b>
	MAPE	33.78	17.43	18.34	17.21	18.89	16.78	<b>16.69</b>
PeMS04	RMSE	48.80	36.01	37.65	35.22	39.66	33.65	<b>32.82</b>
	MAE	33.73	22.66	24.63	22.94	24.89	21.19	<b>20.84</b>
	MAPE	24.18	14.34	17.01	16.43	17.29	13.90	<b>13.77</b>
PeMS07	RMSE	59.27	39.34	38.61	37.87	41.50	39.03	<b>37.54</b>
	MAE	38.17	25.33	25.22	24.01	26.39	24.26	<b>22.99</b>
	MAPE	19.46	11.21	11.82	10.73	11.97	10.21	<b>10.14</b>
PeMS08	RMSE	44.32	27.88	27.83	26.22	30.04	26.80	<b>25.97</b>
	MAE	31.09	18.11	17.46	16.64	18.28	17.13	<b>16.81</b>
	MAPE	22.73	11.34	11.39	10.6	12.15	10.96	<b>10.62</b>

Table 2: Performance comparison of baseline models and STGODE on PeMS datasets.

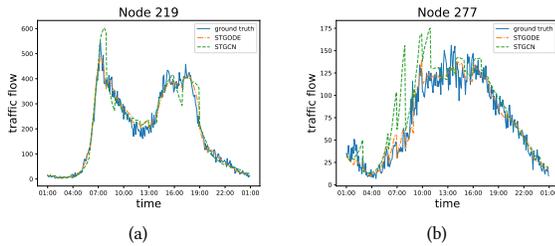


Figure 6: The comparison of prediction results between our model and STGCN.

- STGODE only spatial: This model does not consider semantic neighbors to verify the necessity of introducing a semantic adjacency matrix.
- STGODE-no-h0: The initial state is removed in the derivation of hidden states (Eq 14).
- STGODE-matrix-based: Reformulate the tensor-based ODE (Eq 14) to a matrix-based version as following,

$$\frac{dH(t)}{dt} = \ln \hat{A}H(t) + H(t) \ln W + H_0 \quad (27)$$

which means that the input tensor will be viewed as multiple matrices separately without considering temporal feature transform in ODE blocks.

The results are presented in Fig 7. Here we put STGCN and our STGCN\* together on account of their similar sandwich structures and the same way of convolution. The result shows that our STGCN\* performs much better than previous STGCN, which is contributed to our novel temporal convolution and the introduction of

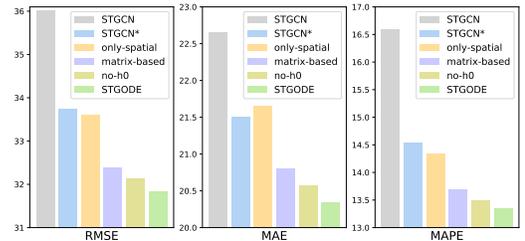
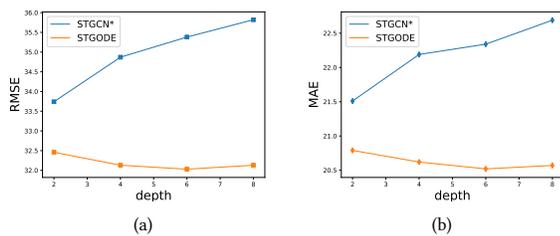


Figure 7: Ablation experiments of STGODE

semantical neighbors, and the poor result of STGODE with only spatial neighbors reinforces the latter point. The performance of the matrix-based version is also inferior to the tensor-based one, as it is incapable to consider spatial-temporal dependency simultaneously. And the result of STGODE without  $\mathcal{H}_0$  shows the importance of connecting the initial state.

5.6.2 *Parameter Analysis.* One major advantage of our STGODE model over other existing methods is that is robust to the over-smoothing problem and thus capable to construct deeper network structures. Here in Fig 8, we represent the performance of STGODE and STGCN\* under different depths, i.e. the input time length of STGODE solver and the number of convolution layers in STGCN\*. It is easy to see that, as the network depth increases, the performance of STGCN\* drops dramatically while the performance of our model is stable, which clearly shows the strong robustness of our model to extract longer-range dependencies.



**Figure 8: The performance of STGODE and STGCN when the network depth increasing.**

## 6 CONCLUSION

A tremendous number of works have been proposed to tackle the complex spatial-temporal problems, but few of them focus on how to extract long-range dependencies without being affected by the over-smoothing problem. In this paper, we present a novel tensor-based spatial-temporal forecasting model named STGODE. To the best of our knowledge, this is the first attempt to bridge continuous differential equations to the node representations of road networks in the area of traffic, which enables to construct deeper networks and leverage wider-range dependencies. Furthermore, the participation of semantic neighbors largely enhances the performance of the model. Extensive experiments prove the effectiveness of STGODE over many existing methods.

## ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 61876006 and No. 61572041).

## REFERENCES

- [1] Donald J Berndt and James Clifford. 1994. Using dynamic time warping to find patterns in time series.. In *KDD workshop*, Vol. 10. Seattle, WA, USA., 359–370.
- [2] George EP Box and David A Pierce. 1970. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association* 65, 332 (1970), 1509–1526.
- [3] James Buizer, Katharine Jacobs, and David Cash. 2016. Making short-term climate forecasts useful: Linking science and action. *Proceedings of the National Academy of Sciences* 113, 17 (2016), 4597–4602.
- [4] Chao Chen, Karl Petty, Alexander Skabardonis, Pravin Varaiya, and Zhanfeng Jia. 2001. Freeway performance measurement system: mining loop detector data. *Transportation Research Record* 1748, 1 (2001), 96–102.
- [5] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. 2018. Neural ordinary differential equations. *arXiv preprint arXiv:1806.07366* (2018).
- [6] Fan RK Chung and Fan Chung Graham. 1997. *Spectral graph theory*. Number 92. American Mathematical Soc.
- [7] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. 2017. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*. PMLR, 854–863.
- [8] Shengdong Du, Tianrui Li, Xun Gong, Yan Yang, and Shi Jinn Horng. 2017. Traffic flow forecasting based on hybrid deep learning framework. In *2017 12th international conference on intelligent systems and knowledge engineering (ISKE)*. IEEE, 1–6.
- [9] Shen Fang, Qi Zhang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. 2019. GSTNet: Global Spatial-Temporal Network for Traffic Flow Prediction.. In *IJCAI*. 2286–2293.
- [10] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. 2019. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 922–929.
- [11] William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *arXiv preprint arXiv:1706.02216* (2017).

- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [13] Peter J Huber. 1992. Robust estimation of a location parameter. In *Breakthroughs in statistics*. Springer, 492–518.
- [14] Young-Seon Jeong, Young-Ji Byon, Manoel Mendonca Castro-Neto, and Said M Easa. 2013. Supervised weighting-online learning algorithm for short-term traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems* 14, 4 (2013), 1700–1707.
- [15] Nicola Jones. 2017. How machine learning could help to improve climate forecasts. *Nature News* 548, 7668 (2017), 379.
- [16] Thomas Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference of Learning Representations*.
- [17] Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [18] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2018. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *International Conference on Learning Representations*.
- [19] Qingqing Long, Yilun Jin, Guojie Song, Yi Li, and Wei Lin. 2020. Graph Structural-topic Neural Network. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1065–1073.
- [20] Qingqing Long, Yilun Jin, Yi Wu, and Guojie Song. 2021. Theoretically Improving Graph Neural Networks via Anonymous Walk Graph Kernels. *arXiv preprint arXiv:2104.02995* (2021).
- [21] Qingqing Long, Yiming Wang, Lun Du, Guojie Song, Yilun Jin, and Wei Lin. 2019. Hierarchical Community Structure Preserving Network Embedding: A Subspace Approach. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 409–418.
- [22] Antonella Longo, Marco Zappatore, Mario Bochicchio, and Shamkant B Navathe. 2017. Crowd-sourced data collection for urban monitoring via mobile sensors. *ACM Transactions on Internet Technology (TOIT)* 18, 1 (2017), 1–21.
- [23] Attila M Nagy and Vilmos Simon. 2018. Survey on traffic prediction in smart cities. *Pervasive and Mobile Computing* 50 (2018), 148–163.
- [24] Zheyi Pan, Yuxuan Liang, Weifeng Wang, Yong Yu, Yu Zheng, and Junbo Zhang. 2019. Urban traffic prediction from spatio-temporal data using deep meta learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1720–1730.
- [25] Patrick Plötz, Niklas Jakobsson, and Frances Sprei. 2017. On the distribution of individual daily driving distances. *Transportation research part B: methodological* 101 (2017), 213–227.
- [26] Xingjian Shi, Zhoung Chen, Hao Wang, Dit Yan Yeung, Wai Kin Wong, and Wangchun Woo. 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Advances in Neural Information Processing Systems* (2015).
- [27] Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. 2020. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 914–921.
- [28] JWC Van Lint and CPIJ Van Hinsbergen. 2012. Short-term traffic and travel time prediction models. *Artificial Intelligence Applications to Critical Transportation Issues* 22, 1 (2012), 22–41.
- [29] Billy M Williams and Lester A Hoel. 2003. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. *Journal of transportation engineering* 129, 6 (2003), 664–672.
- [30] Z Wu, S Pan, G Long, J Jiang, and C Zhang. 2019. Graph WaveNet for Deep Spatial-Temporal Graph Modeling. In *The 28th International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization.
- [31] Louis-Pascal Xhonneux, Meng Qu, and Jian Tang. 2020. Continuous graph neural networks. In *International Conference on Machine Learning*. PMLR, 10432–10441.
- [32] Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2018. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 3634–3640.
- [33] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.
- [34] Qi Zhang, Jianlong Chang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. 2020. Spatio-temporal graph structure learning for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1177–1185.
- [35] Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. 2019. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems* 21, 9 (2019), 3848–3858.
- [36] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2018. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434* (2018).

## 7 APPENDIX

### 7.1 The calculation of the integration in Eq 19

PROOF. Suppose  $\hat{A} - I, U - I, W - I$  have eigenvalue decompositions  $P_1 \Lambda_1 P_1^{-1}, P_2 \Lambda_2 P_2^{-1}, P_3 \Lambda_3 P_3^{-1}$  respectively, then we have

$$\begin{aligned}
& \int_0^t \mathcal{H}_0 \times_1 e^{(\hat{A}-I)(t-\tau)} \times_2 e^{(U-I)(t-\tau)} \times_3 e^{(W-I)(t-\tau)} d\tau \\
&= \int_0^t \mathcal{H}_0 \times_1 P_1 e^{\Lambda_1(t-\tau)} P_1^{-1} \times_2 P_2 e^{\Lambda_2(t-\tau)} P_2^{-1} \times_3 P_3 e^{\Lambda_3(t-\tau)} P_3^{-1} d\tau \\
&= \int_0^t \mathcal{H}_0 \times_1 P_1 \times_2 P_2 \times_3 P_3 \times_1 e^{\Lambda_1(t-\tau)} \times_2 e^{\Lambda_2(t-\tau)} \\
&\quad \times_3 P_3 e^{\Lambda_3(t-\tau)} \times_1 P_1^{-1} \times_2 P_2^{-1} \times_3 P_3^{-1} d\tau, \\
&\quad \text{denote } \tilde{\mathcal{H}}_0 = \mathcal{H}_0 \times_1 P_1 \times_2 P_2 \times_3 P_3, \\
&\quad \int_0^t \mathcal{H}_0 \times_1 e^{(\hat{A}-I)(t-\tau)} \times_2 e^{(U-I)(t-\tau)} \times_3 e^{(W-I)(t-\tau)} d\tau \\
&= \int_0^t \tilde{\mathcal{H}}_0 \times_1 e^{\Lambda_1(t-\tau)} \times_2 e^{\Lambda_2(t-\tau)} \times_3 e^{\Lambda_3(t-\tau)} d\tau \times_1 P_1^{-1} \times_2 P_2^{-1} \times_3 P_3^{-1},
\end{aligned}$$

consider the integral element-wise, then we have,

$$\begin{aligned}
& \left( \int_0^t \tilde{\mathcal{H}}_0 \times_1 e^{\Lambda_1(t-\tau)} \times_2 e^{\Lambda_2(t-\tau)} \times_3 e^{\Lambda_3(t-\tau)} d\tau \right)_{ijk} \\
&= \int_0^t \tilde{\mathcal{H}}_{0ijk} \times_1 e^{\Lambda_{1ii}(t-\tau)} \times_2 e^{\Lambda_{2jj}(t-\tau)} \times_3 e^{\Lambda_{3kk}(t-\tau)} d\tau \\
&= - \frac{1}{\Lambda_{1ii} + \Lambda_{2jj} + \Lambda_{3kk}} \tilde{\mathcal{H}}_{0ijk} \times_1 e^{\Lambda_{1ii}(t-\tau)} \times_2 e^{\Lambda_{2jj}(t-\tau)} \times_3 e^{\Lambda_{3kk}(t-\tau)} \Big|_0^t \\
&= \frac{\tilde{\mathcal{H}}_{0ijk}}{\Lambda_{1ii} + \Lambda_{2jj} + \Lambda_{3kk}} \times_1 e^{\Lambda_{1ii}t} \times_2 e^{\Lambda_{2jj}t} \times_3 e^{\Lambda_{3kk}t} - \frac{\tilde{\mathcal{H}}_{0ijk}}{\Lambda_{1ii} + \Lambda_{2jj} + \Lambda_{3kk}} \\
&\text{thus, the result of the integration is as the following,} \\
&\quad \int_0^t \mathcal{H}_0 \times_1 e^{(\hat{A}-I)(t-\tau)} \times_2 e^{(U-I)(t-\tau)} \times_3 e^{(W-I)(t-\tau)} d\tau \\
&= \left( \frac{\tilde{\mathcal{H}}_{0ijk}}{\Lambda_{1ii} + \Lambda_{2jj} + \Lambda_{3kk}} \times_1 e^{\Lambda_{1ii}t} \times_2 e^{\Lambda_{2jj}t} \times_3 e^{\Lambda_{3kk}t} - \frac{\tilde{\mathcal{H}}_{0ijk}}{\Lambda_{1ii} + \Lambda_{2jj} + \Lambda_{3kk}} \right) \\
&\quad \times_1 P_1^{-1} \times_2 P_2^{-1} \times_3 P_3^{-1}.
\end{aligned}$$

□