

Graph Convolutional Networks for traffic anomaly

Yue Hu
Vanderbilt University
Nashville, Tennessee, USA
yue.hu@vanderbilt.edu

Ao Qu
Vanderbilt University
Nashville, Tennessee, USA
ao.qu@vanderbilt.edu

Dan Work
Vanderbilt University
Nashville, Tennessee, USA
dan.work@vanderbilt.edu

ABSTRACT

Event detection has been an important task in transportation, whose task is to detect points in time when large events disrupts a large portion of the urban traffic network. Travel information Origin-Destination (OD) matrix data by map service vendors has large potential to give us insights to discover historic patterns and distinguish anomalies. However, to fully capture the spatial and temporal traffic patterns remains a challenge, yet serves a crucial role for effective anomaly detection. Meanwhile, existing anomaly detection methods have not well-addressed the extreme data sparsity and high-dimension challenges, which are common in OD matrix datasets. To tackle these challenges, we formulate the problem in a novel way, as detecting anomalies in a set of directed weighted graphs representing the traffic conditions at each time interval. We further propose *Context augmented Graph Autoencoder (Con-GAE)*, that leverages graph embedding and context embedding techniques to capture the spatial traffic network patterns while working around the data sparsity and high-dimensionality issue. Con-GAE adopts an autoencoder framework and detect anomalies via semi-supervised learning. Extensive experiments show that our method can achieve up to 0.4 improvements of the area under the curve (AUC) score over state-of-art anomaly detection baselines, when applied on several real-world large scale OD matrix datasets.

1 INTRODUCTION

Event detection has been a long standing task in transportation [24, 43]. Usually, when large events take place, a large portion of the urban traffic network is disrupted, resulting in anomalous traffic conditions. Accurate and timely detection of such events can help with real-time resource allocation and congestion mitigation, enabling informed urban traffic management. Therefore, the objective of our work is to detect extreme traffic events, which are anomalies that manifest on large portions of the network at once.

Today, abundant transportation data facilitates real time traffic monitoring and anomaly detection, with the aid of machine learning techniques. One promising area is in urban mobility, where large fleets of instrumented vehicles (e.g., ride sharing services and taxis) collect abundant traffic information, which is then processed into *origin-destination* (OD) travel time data and made publicly available to assist urban mobility managers. Examples include data products from Uber Movement [35], taxi datasets in Chicago [5] and New York City [34], and micro-mobility data from bike share and scooter operators (e.g., [4]).

Despite the growing availability of such OD data, it is often challenging for urban mobility managers to directly use the data for performance monitoring and event detection. This is because the

data is highly structured in time and space and also high dimensional. It is also sparse with many origin-destination pairs lacking data at any given moment in time. The spatio-temporal structure, high dimensionality, and sparsity of the data inhibit standard anomaly detection approaches from achieving performance necessary for widespread adoption in current traffic management centers.

To address these challenges, we propose a *Context augmented graph autoencoder (Con-GAE)* that leverages graph embedding and context embedding techniques to capture the spatio-temporal traffic network patterns. The autoencoder uses graph convolutional layers modified to account for asymmetry in travel times to capture spatial patterns in the data, even when many of the travel time entries are missing. It also uses context embedding to capture daily and weekly periodicity in the data, to further enhance performance. Finally, Con-GAE adopts a hierarchical structure to aggregate the spatial and temporal embeddings of the traffic network at each time step into a single comprehensive graph embedding, to detect time contextualized anomaly graphs as a whole. Through comparisons with existing state of the art anomaly detection methods and through an ablation study, we show that Con-GAE can achieve performance improvements to the *area under the curve* (AUC) of 0.1 to 0.4 when detecting large scale anomalies in OD travel time data.

We give a high-level overview of the structure that enables Con-GAE to fully capture the spatial correlations between city zones. Intuitively, Zone A and Zone B can have similar travel times to other parts of the network if the zones are physically close together, or close in travel time. We consider physical closeness by using geographic information as input node features, and consider the travel-time connection via edge weights. Then Con-GAE learns the vector representation of each node given both node and weight information, such that Zone A and Zone B have similar embeddings when they are physically close or close in travel time. The node embeddings are further used to decode the edge weights, so two origin zones with similar embeddings will lead to similar travel times to other nodes in the network.

Our work is inspired by two threads of recent development in deep learning, namely graph convolutional networks and autoencoder-based anomaly detectors. The graph convolutional layers in Con-GAE are a modification of the *graph convolutional network* (GCN) [19]. Given a graph, GCN learns an embedding of the nodes in the graph that encodes structural information about the graph via aggregation of the embeddings of neighboring nodes on the graph. While modified GCN layers can capture the geometric structure of the network, the dynamic nature of traffic requires further consideration of the temporal dimension. Existing works on dynamic graphs [13, 49] mainly focus on modeling the short-term dependencies of consecutive graph instances. In comparison, we focus on exploiting the temporal periodicity of traffic graphs over longer time horizons.

Daily or weekly periodicity in the data can better support traffic anomaly detection than short term fluctuations. We further emphasize that while some approaches detect anomalous nodes and edges in dynamic graph [45, 47], our task of extreme event detection requires detecting anomalous graph as a whole, rather than a part of a graph as anomaly.

Con-GAE determines anomalies via an *autoencoder* (AE) framework [28] for semi-supervised anomaly detection. An AE consists of an encoder to compress the high-dimensional input into a low dimensional space, and a decoder to map the low dimensional representation back to the original input space minimizing the reconstruction error. When trained only on normal data, AE is assumed to encode and decode normal data well, but not anomalous data. Thus during testing, samples with large reconstruction errors are marked as anomalies. AE-based anomaly detectors have shown success in various high-dimensional anomaly detection tasks [6, 41, 48]. Here we specialize the AE network architecture to exploit spatio-temporal structure in mobility data.

We show the effectiveness of our model on large-scale real-world dataset. We apply Con-GAE to an Uber movement OD travel time dataset, to detect two kinds of synthesized anomalies, i.e., large spatial anomalies impacting the network, and traffic conditions which are otherwise normal but not for the given time. Such temporal anomalies occur, e.g., due to holidays. We vary the magnitude of the anomalies and the number of anomalous instances, and show that our model outperforms several state-of-the-art baseline methods for high-dimensional or spatio-temporal anomaly detection. Finally, we show how the model reveals the influence of a real large scale event contained in the dataset.

In summary, our contributions are as follows:

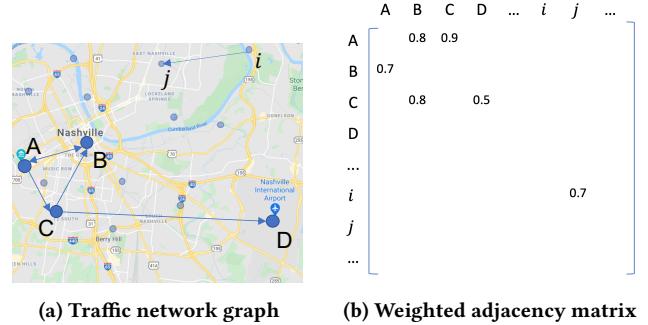
- We formulate the problem of anomaly detection in mobility OD datasets in a novel way, as detecting anomalies in a set of time dependent directed weighted graphs. Each graph has geographic information contained in the node features and travel time information in the edge weights.
- We propose Con-GAE, which is an autoencoder-based anomaly detector. Con-GAE uses GCN layers to generate node embeddings that capture spatial relationships, and time embeddings to capture the temporal periodicity.
- Experiments on an OD mobility dataset demonstrating Con-GAE outperforms several state-of-the-art anomaly detection methods.

2 METHODOLOGY

In this section, we first model the traffic network as a graph and formulate the anomaly detection problem. Then we address the graph encoder and decoder structures. Finally we address the loss function and regularization strategies used.

2.1 Modeling mobility data as a graph

We introduce the graph model of the OD mobility data, and then formulate the anomaly detection problem on the graph. The mobility data can be modeled as a set of time dependent directed weighted graphs. At discrete time $t \in \{1, 2, \dots, T\} = \mathcal{T}$, the graph is denoted as $G(t) = (\mathcal{V}, \mathcal{E}(t), \mathbf{W}(t))$. The node set \mathcal{V} is time invariant, representing the $N = |\mathcal{V}|$ distinct geographical zones in the city. There



(a) Traffic network graph

(b) Weighted adjacency matrix

Figure 1: Illustration of OD mobility graph (left) and corresponding weighted adjacency matrix (right) corresponding to scaled inverse travel times between points on the graph. The weighted adjacency matrix can be highly sparse due to lack of data on all OD pairs.

is an edge $e_{ij} \in \mathcal{E}(t)$ if the travel time information is available from node $v_i \in \mathcal{V}$ to $v_j \in \mathcal{V}$ at time t . The weighted adjacency matrix is denoted $\mathbf{W}(t) \in \mathbb{R}^{N \times N}$, where each element $w_{ij}(t)$ contains the scaled inverse travel time from node v_i to v_j at time t , such that larger values indicate shorter travel times. The geographic information for each zone is summarized in the node feature matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$, which could include, e.g., location, size, or land use. An OD graph is shown in Fig 1(a), with the corresponding adjacency matrix $\mathbf{W}(t)$ illustrated in Fig 1(b).

Our goal is the following. Given a set of graphs $\{G(t)|t \in \mathcal{T}\}$, find the anomalous times $t_a \in \mathcal{T}$ corresponding to large scale events disrupting the transportation network. We do so by training a context-augmented graph autoencoder that encodes each graph as a low-dimensional embedding conditioned on the time information, and decodes the embedding so as to minimize the average reconstruction error of the weighted adjacency matrix $\mathbf{W}(t)$. The autoencoder is trained only on normal data, so it produces high reconstruction errors when events occur.

2.2 Encoder

In an overview, the encoding process consists of three steps, as in Fig 2. We first apply GCN layers to encode the graph structure and node feature information, and learn an embedding vector for each node in the graph. We also use a distinct time embedding for each hour in the day and each day in the week to capture temporal information in the network. The node embeddings and the time embeddings are stacked in a vector, which is passed through a fully connected layer to arrive at a graph embedding of reduced dimension.

In the first step, we use a modified GraphSAGE [14] implementation of GCN to learn the node embedding, by aggregating the adjacent node information in a layer-by layer manner. The number of GraphSAGE layers stacked decides how many hops of neighborhood information is passed to each node [9]. Let $\mathcal{N}_i(t)$ denote the neighbors of node i . GraphSAGE propagates the node v_i embedding $h_i^l(t) \in \mathbb{R}^{d_l}$ in layer l of the network to the embedding a

$h_i^{l+1}(t) \in \mathbb{R}^{d_{l+1}}$ as:

$$\hat{h}_i^{l+1}(t) = \text{ReLU}\left(U^l \text{Concat}(h_i^l(t), \text{Agg}_{v_j \in N_i(t)} h_j^l(t))\right), \quad (1)$$

where $\hat{h}_i^{l+1}(t)$ is the non-normalized node embedding, Concat is the concatenation operator, and Agg is an aggregation operator described subsequently. The normalized node v_i embedding at layer $l + 1$ is obtained by projecting onto the unit ball: $h_i^{l+1}(t) = \hat{h}_i^{l+1}(t)/\|\hat{h}_i^{l+1}(t)\|_2$.

The standard GraphSAGE [14] aggregation operators for binary undirected edges aggregate information from the embeddings of the neighbors of v_i via mean, max-pooling, or LSTM operations. To account for the time varying weighted directed edges, we apply a weighted mean aggregation function. We take $N_i(t)$ as the set of nodes j for which a weighted edge e_{ji} is present at time t . Then the weighted mean aggregation operator reads:

$$\text{Agg}_{v_j \in N_i(t)} h_j^l(t) := \frac{\sum_{j \in N_i(t)} w_{ji}(t) h_j^l(t)}{\sum_{j \in N_i(t)} w_{ji}(t)}. \quad (2)$$

In the second step, we introduce context embeddings that capture the weekly and hourly periodicity of the mobility data. We define context as additional information not contained in the OD matrix that the encoder and decoder is conditioned on. In our case we use temporal information as context. Context is implemented by concatenating two fixed length vectors: one is the embedding of the day of week information, the other is the embedding of the hour of the day. The embeddings capture our prior that traffic conditions on a specific time of day or a specific day of the week should be similar. The hour embedding is denoted as $h_{\text{hour}} \in \mathbb{R}^{d_{\text{hour}}}$, with one embedding per hour of day; similarly the week embedding is denoted as $h_{\text{week}} \in \mathbb{R}^{d_{\text{week}}}$, with one embedding per day of the week. Given the graph at time t (e.g., 10am Tuesday Sept. 1), we use the corresponding h_{hour} (i.e., 10 am) and h_{week} (i.e., Tuesday) embedding. The entries of each embedding vector are initiated from i.i.d. normal distribution, then learned during training time.

In the last step, we combine the node and time embeddings into a single low-dimensional embedding through a fully connected layer:

$$\begin{aligned} \tilde{h}_G(t) &= \text{Concat}(h_1^L(t), h_2^L(t), \dots, h_N^L(t), h_{\text{hour}}(t), h_{\text{week}}(t)), \\ h_G(t) &= \text{ReLU}(U_G \tilde{h}_G(t)), \end{aligned} \quad (3)$$

where $h_G(t) \in \mathbb{R}^{d_g}$ is the final graph embedding at time t , and $U_G \in \mathbb{R}^{d_g \times (Nd_L + d_{\text{week}} + d_{\text{hour}})}$ is weight matrix.

2.3 Decoder

The decoder works in the opposite way as the encoder, as shown in Fig 3. We first calculate the node embedding from the graph embedding. Then, given each pair of node embeddings corresponding v_i and v_j at time t , we calculate the $w_{ij}(t)$ entry in the weighted adjacency matrix.

We explicitly condition the graph decoding on time by concatenating the graph embedding $h_G(t)$, with the corresponding time embeddings. Then a fully connected layer is used to recover a vector

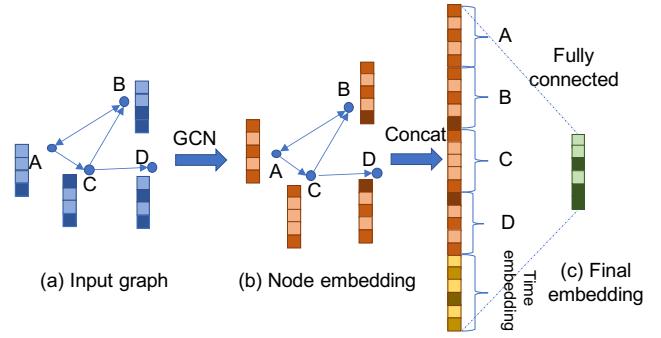


Figure 2: Illustration of encoder. We first apply GCN layers to aggregate the node features and edge weight information and calculate node embedding (b). Then we concatenate the node embedding and time embedding, and further shrink the embedding dimension, to calculate the embedding of the graph as a whole as in (c).

containing all node embeddings:

$$\begin{aligned} \tilde{h}'_G(t) &= \text{ReLU}(U'_G \text{Concat}(h_G(t), h_{\text{hour}}(t), h_{\text{week}}(t))), \\ \{h_1(t), h_2(t), \dots, h_N(t)\} &= \text{Unstack}(\tilde{h}'_G(t)), \end{aligned} \quad (4)$$

which is subsequently unstacked to recover the individual node embeddings $h_i(t) \in \mathbb{R}^{d_L}$. $U'_G \in \mathbb{R}^{Nd_L \times (d + d_{\text{week}} + d_{\text{hour}})}$ is a weight matrix to be trained.

Given a pair of nodes (v_i, v_j) and the corresponding recovered node embeddings $(h_i(t), h_j(t))$, we use an edge weight predictor to recover the edge weight $w_{ij}(t)$. Note that the predominantly used GAE edge predictor $w_{ij} = \sigma(h_i^T h_j)$ [18] cannot be adopted here, because it produces symmetric results for w_{ij} and w_{ji} . Because the graph corresponding to the mobility data is directed, forcing w_{ij}, w_{ji} to be the same would introduce bias. A bilinear predictor $w_{ij} = h_i^T Q h_j$ with Q to be learned is also popular for asymmetric decoding [42], but is slightly slower and performs slightly worse in our experiments. Instead, we use a parametrized MLP as a weight predictor, by first concatenating the two node embeddings and then applying fully connected layers:

$$\begin{aligned} \tilde{w}_{ij}(t) &= \text{ReLU}(U_{\text{dec}}^1 \text{Concat}(h_i(t), h_j(t))), \\ w'_{ij}(t) &= \sigma(U_{\text{dec}}^2 \tilde{w}_{ij}(t)), \end{aligned} \quad (5)$$

where $U_{\text{dec}}^1 \in \mathbb{R}^{d_e \times 2d_L}$ and $U_{\text{dec}}^2 \in \mathbb{R}^{1 \times d_e}$ are weight matrices. The final sigmoid layer σ ensures $w'_{ij}(t) \in [0, 1]$.

2.4 Loss function

We use the *mean squared error* (MSE) between the entries of the original and recovered weighted adjacency matrices as the loss function. For graph instance $G(t)$ at time t , the loss function reads:

$$\mathcal{L}(t) = \frac{1}{|\mathcal{E}(t)|} \sum_{e_{ij} \in \mathcal{E}(t)} (w_{ij}(t) - w'_{ij}(t))^2. \quad (6)$$

We use the Adam optimizer [17] to minimize the loss function (6) for training instances in mini-batches. During training, the values of $U^1, \dots, U^L, U_G, U'_G, U_{\text{dec}}^1, U_{\text{dec}}^2$, the 24 h_{hour} , and the 7 h_{week} time

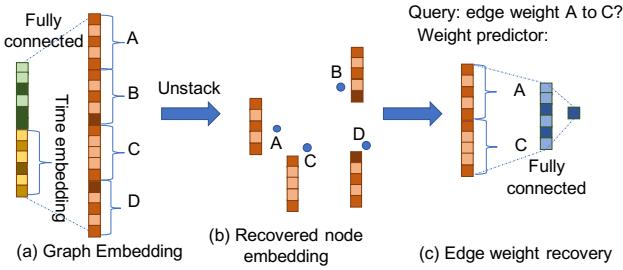


Figure 3: Illustration of decoder. We first calculate the node embeddings from graph encoding conditioned on time as in (a)(b). Then for each query of edge weight from one node to another, we use a weight predictor to calculate from the two corresponding node embeddings (c).

embeddings are updated. During testing, the loss function (6) for each testing instance is used as its anomaly score.

2.5 Regularization

For regularization during training, we conduct edge dropout before feeding the graph data into Con-GAE, by randomly dropping edges from the edge set $\mathcal{E}(t)$ with probability p_{e_drop} . Edge dropout amounts to masking a portion of edges and letting Con-GAE to predict edge weights without seeing the values. This can help improve the generalization ability, and force Con-GAE to learn the relations between nodes for correct prediction. We also apply dropout [30] with probability p_{drop} on the output of the GraphSAGE layers as well as on the time embeddings. Dropout prevents over-reliance on specific features that lead to locally optimal reconstruction solutions, such as the time embedding.

3 EXPERIMENTS

We use a large-scale real-world transportation dataset to evaluate Con-GAE. The experiments include: 1) a performance comparison of Con-GAE with other methods for spatial and temporal anomaly detection; 2) an ablation study to check the contribution of each component of Con-GAE; 3) a sensitivity study of the model hyper-parameters; and 4) a demonstration of the real-world performance of Con-GAE.

3.1 Dataset

Uber Movement dataset [35] provides multiple publicly available mobility datasets including origin-destination travel times at an hourly basis. We use data from the first two quarters of 2019 (Jan. - June) in Nashville, TN for analysis, which is the latest data available. The Q2 data is used for training, with the notable exclusion of data from April 15 - May 1, which is excluded because of a known large event, namely the NFL draft, occurred in Nashville on April 25-27. The Q1 data is used for testing, and the three-week period containing the NFL draft is used for a qualitative case study.

New York City (NYC) taxi dataset [34] is an open source dataset recording pick-up and drop-off time and locations for each taxi trip in NYC. We aggregate the trip information into origin-destination travel time matrices for taxi zones at an hourly basis. The data from

Table 1: Dataset statistics

Dataset	Uber Movement	NYC taxi	Chicago taxi
# graph samples	4331	4344	2880
avg travel time (min)	5.3	14.8	13.4
avg # edges per graph	1471	1482	440

Jan.1 - Mar.31, 2019 is used for training, and Apr.1 - June.30 2019 is used for testing.

Chicago taxi dataset [5] records the taxi trips in the City of Chicago, and is processed in the same way as NYC taxi dataset to result in origin-destination travel time matrices at an hourly basis. The data from Jan.1 - Mar.31 2019 is used for training, and Apr.1 - Apr.30, 2019 is used for testing.

Table 1 shows the statistics of the three datasets. The plot for travel time distribution can be found in appendix. For all three datasets, we restrict our analysis to the top 50 zones, covering at least 75% of all data. because no ground truth anomalies are available, synthetic events are added to testing sets, following a similar anomaly injection approach of [1, 45],

3.2 Baselines

We compare our method against both traditional and state-of-the-art models for anomaly detection spanning multiple detection strategies. The baselines are: *i*) A naive *historical average* (HA) approach which calculates an anomaly score as the mean squared error between all observed OD pairs with the historical mean value at the corresponding day of week and hour of day, thereby capturing periodicity of the data; *ii*) A *Robust tensor recovery and completion* (RTC) [16] method that uses a low-rank tensor decomposition to capture spatial-temporal correlations and detect outliers; *iii*) An autoencoder [15] using MLP for the encoder and decoder; *iv*) the *Long Short Term Memory Networks based Encoder-Decoder scheme for Anomaly Detection* (EncDec-AD) [25]; *v*) a *Deep autoencoding Gaussian mixture model* (DAGMM) [51]; *vi*) A RNN-based *deep structured energy based model* REBM [46]; and two graph-based baselines: *vii*) GCN [19]; *viii*) GraphSAGE [14]. The graph-based methods GCN and GraphSAGE have an encoder-decoder structure to encode each node in a low dimensional space, thus can serve as an autoencoder for anomaly detection. A description of each method can be found in supplementary material.

3.3 Experiment setup

We consider two kinds of anomalies: *i*) spatial anomalies, where the traffic conditions deviate from normal spatial patterns, and *ii*) temporal anomalies, where traffic conditions follow a correct spatial pattern, but is not the typical condition for the corresponding time of day.

Spatial anomalies are injected by first randomly selecting a fraction γ of time slices to pollute. For each polluted time slice, we randomly choose a fraction α of the OD pairs, and perturb the corresponding travel time by a factor drawn from an uniform distribution $\mathcal{U}(-\beta, \beta)$. Temporal anomalies are introduced by randomly selecting a fraction γ of time slices, and shifting the time associated with the data by 12 hours (e.g., 8pm becomes 8am and vice versa).

In the experiments, the pollution ratio and magnitudes are set to reflect the assumption that large events are infrequent, but they pollute relatively large portions of the network. We also vary the pollution ratios and magnitudes α , β , and γ to check the anomaly detection performance under various cases. For each case, we randomly generate the test set five times and calculate the average model performance.

The *area under the receiver operating characteristic curve* (AUC) is used as the metric to compare the methods. For the deep learning methods (i.e., Con-GAE, AE, DAGMM, EncDec-AD, REBM, GCN and GraphSAGE), the output of a given sample is an anomaly score. Using the anomaly scores of the test samples, the AUC is computed. For RTC, the output includes a sparse tensor containing anomalies. A anomaly score is calculated using the l_2 norm of the vector corresponding to each sample.

3.4 Model Configurations

The model settings for Con-GAE is as follows. The node feature matrix $X \in \mathbb{R}^{N \times d}$ is constructed with $d = 4$ corresponding to the minimum and maximum latitude and longitude extents of the zone corresponding to the node. The encoder uses $L = 2$ layers of GCN to learn the node-level embedding. The model is trained for 150 epochs with a batch size of 10. Out of the training set, 10% is kept out as validation set for early stopping. Other settings of Con-GAE for different datasets are as follows.

For Uber data, the two layers of node embedding $h_i^1(t) \in \mathbb{R}^{300}$ and $h_i^2(t) \in \mathbb{R}^{150}$ respectively. The dimension of the hour and week embeddings are $d_{\text{week}} = d_{\text{hour}} = 100$, and the graph embedding dimension is $d_g = 150$. The dropout rates $p_{\text{e_drop}}$ and p_{drop} are both set at 0.2. The initial learning rate of the Adam optimizer is 5×10^{-5} , and we decay the learning rate by 0.5 every 50 epochs.

For NYC data, the two layers of node embedding $h_i^1(t) \in \mathbb{R}^{150}$ and $h_i^2(t) \in \mathbb{R}^{50}$ respectively. The dimension of the hour and week embeddings are $d_{\text{week}} = d_{\text{hour}} = 100$, and the graph embedding dimension is $d_g = 50$. The dropout rates $p_{\text{e_drop}}$ and p_{drop} are both set at 0.2. The initial learning rate is 1×10^{-3} , and we decay the learning rate by 0.5 every 20 epochs.

For Chicago data, the two layers of node embedding $h_i^1(t) \in \mathbb{R}^{300}$ and $h_i^2(t) \in \mathbb{R}^{25}$ respectively. The dimension of the hour and week embeddings are $d_{\text{week}} = d_{\text{hour}} = 200$, and the graph embedding dimension is $d_g = 25$. The dropout rates $p_{\text{e_drop}}$ and p_{drop} are both set at 0.1. The initial learning rate is 1×10^{-3} , and we decay the learning rate by 0.5 every 20 epochs.

The AE, DAGMM, EncDec-AD, and REBM are implemented based on code in [12]. the respective hyper-parameters for each model are tuned to archive the best performance. GCN and GraphSAGE have the same encoding node dimension as our model. Detailed configurations are available in the supplementary materials.

3.5 Result and analysis

3.5.1 Model comparison. First, we compare Con-GAE with baseline methods for both both spatial and temporal anomaly detection, for different datasets. We fix the fraction of the time slices chosen to be polluted $\gamma = 10\%$, pollution magnitude $\alpha = 50\%$ and $\beta = 10\%$. The results shown in Table 2. We can see that Con-GAE constantly

Table 2: The AUC score for spatial and temporal anomaly detection of different datasets.

anomaly type	spatial anomaly			temporal anomaly		
dataset	Uber	NYC	Chicago	Uber	NYC	Chicago
HA	0.687	0.498	0.570	0.661	0.677	0.658
RTC	0.765	0.795	0.601	0.522	0.247	0.562
AE	0.812	0.671	0.727	0.517	0.592	0.460
EncDec-AD	0.582	0.782	0.715	0.549	0.659	0.671
REBM	0.859	0.759	0.739	0.482	0.805	0.482
DAGMM	0.546	0.603	0.669	0.466	0.397	0.414
GraphSAGE	0.840	0.741	0.674	0.542	0.464	0.426
GCN	0.717	0.593	0.765	0.548	0.416	0.421
Con-GAE	0.908	0.837	0.912	0.726	0.895	0.753

Table 3: The AUC score for spatial and temporal anomaly detection, under different anomaly rates. We vary the fraction γ of the time slices chosen to be polluted.

anomaly type	spatial anomaly			temporal anomaly		
anomaly rate γ	5%	10%	20%	5%	10%	20%
HA	0.728	0.687	0.711	0.658	0.661	0.659
RTC	0.736	0.765	0.790	0.579	0.522	0.509
AE	0.813	0.812	0.806	0.497	0.517	0.518
EncDec-AD	0.584	0.582	0.582	0.564	0.549	0.532
REBM	0.844	0.859	0.833	0.468	0.482	0.501
DAGMM	0.550	0.546	0.507	0.439	0.466	0.477
GraphSAGE	0.842	0.840	0.860	0.570	0.542	0.526
GCN	0.708	0.717	0.744	0.560	0.548	0.526
Con-GAE	0.903	0.908	0.913	0.752	0.726	0.693

outperforms the other methods, with an improvement in AUC score between 0.1 and 0.4. For spatial anomalies, Con-GAE is the only method to achieve an AUC above 0.9 for Uber data and Chicago data, and the only method above 0.8 for NYC data. For temporal anomalies, most other methods have AUC score around 0.5 except HA, meaning they are not well suited for detecting temporal anomalies, while Con-GAE achieves an AUC score of 0.7 or higher. These experiments highlight that when dealing with large scale mobility data, taking the graph structure into account can substantially boost the performance. Meanwhile, directly applying GCN and GraphSAGE model results in poor performance, showing the importance of designing hierarchical structure for graph-scale embedding beyond individual node embeddings when detecting network-wide anomalies. The temporal experiments shows that for data with periodicity, it is more effective to consider the long-term periodicity as in Con-GAE and HA, than to consider the short-term dependencies as in LSTM-ED and REBM, or no temporal dependencies as in DAGMM, AE, GCN and GraphSAGE.

Then, we compare the anomaly detection methods under different anomaly rates. We vary the fraction γ of the time slices chosen

Table 4: The AUC score for spatial anomaly detection, under different magnitudes of anomalies. Results are shown under different α deciding the OD pairs polluted, and different β corresponding to the perturbation magnitude.

spatial anom- aly rate α	25%			50%			
	anomaly magnitude β	5%	10%	20%	5%	10%	20%
HA	0.405	0.533	0.804	0.455	0.687	0.934	
RTC	0.626	0.699	0.863	0.648	0.765	0.942	
AE	0.294	0.572	0.936	0.405	0.812	0.994	
EncDec-AD	0.410	0.483	0.727	0.452	0.582	0.896	
REBM	0.389	0.633	0.958	0.491	0.859	0.997	
DAGMM	0.511	0.527	0.567	0.525	0.546	0.639	
GraphSAGE	0.381	0.627	0.963	0.491	0.840	1.000	
GCN	0.443	0.564	0.844	0.498	0.717	0.966	
Con-GAE	0.482	0.755	0.985	0.610	0.908	1.000	

to be polluted, fixing $\alpha = 50\%$ and $\beta = 10\%$. The result is shown in Table 3. We can see that the advantage of our method holds under different anomaly rates.

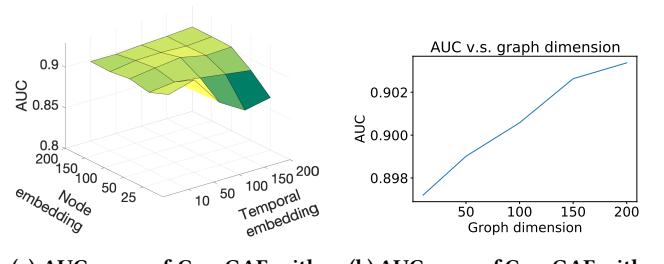
Next, we investigate the influence of anomaly magnitude. We focus on spatial anomaly detection, and vary the fraction α of OD pairs polluted and the uniform distribution $\mathcal{U}(-\beta, \beta)$ deciding the range of pollution ratio. We fix the time slices polluted at $\gamma = 10\%$. The result is shown in Table 4. No method obtains the best performance over all scenarios. All methods have higher AUC score with larger α and β , and perform less well when the pollution is more nuanced. Con-GAE has significant advantage over the other methods for $\beta = 10\%$ or larger. When β is lower than 10%, only the linear method RTC performs adequately well. Since our goal is to detect large events with significant disturbance to road networks, we conclude that Con-GAE is the most competitive method overall.

3.5.2 Ablation study. We compare the several variants of Con-GAE, to see how each component of Con-GAE helps with anomaly detection. We consider the following variants: *i) Con-GAE-sp*: We remove all temporal information the autoencoder, and only use the spatial graph data for training and detection. *ii) Con-GAE-t*: we remove the GCN layers, and only use the temporal information for training and detection. *iii) Con-GAE-fc*: we replace the GCN layers in the autoencoder with fully connected layers, while temporal information is retained. *iv) Con-GAE-NonContextDec*: we omit the context embedding and only use the graph embedding as input to the decoder. *v) Con-GAE-NonWeightedEnc*: We use the standard GraphSAGE implementation (1) with mean aggregator as first step of encoder, instead of the modified weighted mean aggregator.

The result is shown in Table 5. We can see that the modified GCN layers play a crucial role in anomaly detection. The AUC score is 0.2 lower if we remove the GCN layers (Con-GAE-t); and is 0.1 lower if we replace GCN layers with fully connected layers (Con-GAE-fc),

Table 5: Ablation study.

anomaly type anomaly rate γ	spatial anomaly			temporal anomaly		
	5%	10%	20%	5%	10%	20%
Con-GAE	0.903	0.908	0.913	0.752	0.726	0.693
Con-GAE-t	0.754	0.722	0.729	0.630	0.630	0.621
Con-GAE-sp	0.888	0.893	0.905	0.572	0.549	0.529
Con-GAE-fc	0.846	0.844	0.829	0.693	0.681	0.683
Con-GAE-NonContextDec	0.829	0.816	0.818	0.544	0.541	0.538
Con-GAE-NonWeightedEnc	0.835	0.813	0.819	0.606	0.591	0.576



(a) AUC score of Con-GAE with varying dimension of node and time embedding. (b) AUC score of Con-GAE with varying graph embedding dimension.

Figure 4: Sensitivity analysis of Con-GAE

or with standard GraphSAGE layers (Con-GAE-NonWeightedEnc). We observe that temporal information provides marginal benefit for spatial anomaly detection, but is crucial for temporal anomaly detection. Without temporal information, it is impossible for Con-GAE-sp to detect temporal anomalies, with an AUC score constantly around 0.5. The drop of performance in Con-GAE-NonContextDec shows that explicitly conditioning the decoder on time context helps the decoder to better reconstruct the graph, even though the temporal information might be encoded by the encoder.

3.5.3 Sensitivity analysis. Next, we explore how the performance of Con-GAE changes as a function of the node embedding dimension, the hour and week embedding dimension, the low dimensional final graph embedding, and the model depth. We fix $\alpha = 50\%$, $\beta = 10\%$ and $\gamma = 10\%$.

First, we explore the sensitivity to spatial and temporal encoding dimensions. We vary the node embedding dimension between 25 and 200, and vary the week and hour embedding dimension between 10 and 200. We fix the other parameters the same as in Model Configurations section. The result is shown in Fig 4a. We can see that Con-GAE is not too sensitive to spatial encoding dimension and temporal embedding dimension, with the AUC all in the range of 0.84–0.94. Most combinations work well except the case when spatial encoding dimension is too low (below 50) and the temporal embedding dimension is too high (above 100).

We then fix node embedding and temporal embedding dimensions, and explore the influence of the final graph embedding dimension d_g . We vary graph dimension from 10 to 150. Fig. 4b shows the result. We can see that AUC score is insensitive with scores slightly increasing but all around 0.9.

Table 6: Sensitivity to model depth

# GCN layers	1	2	3	4	5	6
AUC score	0.896	0.908	0.901	0.896	0.886	0.889

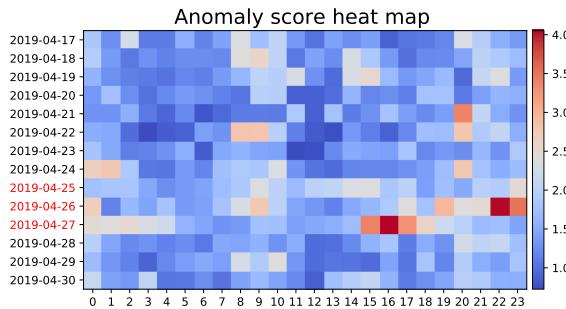


Figure 5: Anomaly scores for each hour. The NFL days are marked red. The score denotes the reconstruction error (6).

Lastly, we investigate the sensitivity to model depth. We vary the number of GCN layers from 1 to 6, fixing $h_i^1(t) \in \mathbb{R}^{300}$ and $h_i^l(t) \in \mathbb{R}^{150}$ for layers $l > 1$. The result is shown in Table 6. We can see that Con-GAE is not sensitive to depth, with AUC score all between 0.88 and 0.91.

3.6 Visualization of real-world traffic anomaly

In this section, we demonstrate the real-world application of Con-GAE. We use the test set 4/27-5/01 of 2019 Q2, which is not seen during training and is known to contain a large event (the NFL draft). The NFL draft brought several hundred thousand visitors to the city, and nightly events resulted in roadway closures in the downtown Nashville area.

Figure 5 shows the anomaly score (the reconstruction loss) for each hour of the studied period. The days under which the major activities of the draft take place are highlighted in red on the y-axis. We can see that the highest anomaly scores occur during the draft, specifically 4/26 19:00 until midnight, as well as 4/27 15:00–18:00. These are the time when the draft, post draft and post concert & fireworks take place [27].

Figure 6 plots the OD travel time data during two periods with high anomaly scores. For clarity, only a subset of travel time pairs are plotted, and only deviations in excess of 5 min from the historical average are shown. Fig 6(a) shows that the travel times from the region containing the draft to other parts of the city is significantly longer than usual. The third largest anomaly period occurs in the consecutive days of 4/21 to 4/24 at 20:00. This matches the record of *Tennessee Department of Transportation* [32] that there are alternating nightly maintenance road closures for the major freeway connecting downtown to the airport, resulting in substantially longer travel times (Fig 6(b)).

4 RELATED WORK

We review graph network embedding and autoencoder-based anomaly detection techniques that inspire Con-GAE.

4.1 Graph network embedding

In recent years, there has been rapid development in graph embedding-based learning algorithms. In particular, GCN has gained popularity because of its scalability and good performance [9]. Some benchmark variations of GCN include [14, 19, 36], which have become the building block of many complex graph network models [40]. Our work focuses on weighted directed graphs appropriate for mobility datasets, compared with binary undirected graphs in the original work of [14, 19].

A *graph autoencoder* (GAE) is an unsupervised learning approach that encodes the graph in a latent space, and reconstructs the graph structure from the encoding [40]. Since the introduction of GAE in [18, 33], it has been widely used, for example in link prediction [29, 39, 42], matrix completion [2], and graph clustering [37]. While existing GAE [18] learn node embeddings during encoding, our model further compresses the node embeddings of all nodes in a graph to a single graph embedding, and conditions it on temporal information to add additional context important for interpreting mobility data.

Given the intrinsic graph structure of transportation networks, recent research has demonstrated advantages in applying graph embedding on mobility data, for forecasting[7, 44], passenger demand prediction [38], multi-modal transportation recommendation [23], and human trajectory prediction [26]. While graph embedding for traffic anomaly detection is less studied, our present work work we show the ability of graph embedding to learn patterns in OD mobility data for anomaly detection.

4.2 Autoencoder-based anomaly detection

Deep autoencoder based methods have been widely used for anomaly detection [3, 6, 25, 28, 41, 48]. The encoder and decoder can have various structures based on different input data types. Besides the widely used MLP based encoder and decoders [3, 28, 41], *convolutional neural networks* (CNN) are applied on grid data [6], and *recurrent neural networks* (RNN) are applied on time series data [25, 31]. Recently, GCN-based autoencoders have been used for anomaly detection on graph data [8, 10, 20–22, 50], with an important emphasis node-level anomalies. In comparison, our work aims at graph-level anomaly detection, since each graph corresponds to the network wide traffic condition at a specific time instant that may be disturbed by a large scale event.

5 CONCLUSION

In this work, we address the problem of detecting large events in large origin destination mobility datasets. We formulate the problem of detecting these events as a problem of detecting anomalies in a set of time dependent directed graphs containing mobility data on the network. We introduced Con-GAEwhich is an autoencoder based anomaly detector that uses GCN layers and temporal embeddings to determine anomalies. Extensive experiments on synthetic test sets generated from real-world mobility data shows that Con-GAE outperforms several state-of-art anomaly detection methods in both spatial anomaly and temporal detection. In future work, we are interested to explore scalability of Con-GAE to networks with significantly more nodes, and for anomaly detection on other network datasets.

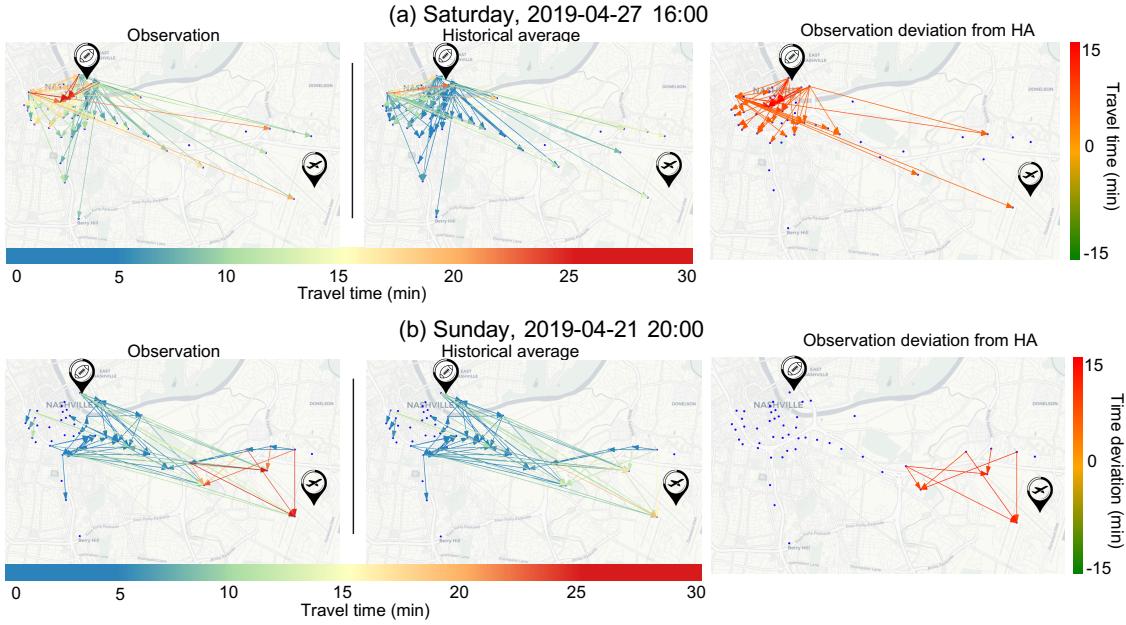


Figure 6: Traffic condition for the detected events. The OD pair travel time and comparison to historical conditions are plotted.

REFERENCES

- [1] Leman Akoglu, Hanghang Tong, and Danai Koutra. 2015. Graph based anomaly detection and description: a survey. *Data mining and knowledge discovery* 29, 3 (2015), 626–688.
- [2] Rianne van den Berg, Thomas N Kipf, and Max Welling. 2018. Graph convolutional matrix completion. *KDD’18 Deep Learning Day* (2018).
- [3] Andrea Borghesi, Andrea Bartolini, Michele Lombardi, Michela Milano, and Luca Benini. 2019. Anomaly detection using autoencoders in high performance computing systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9428–9433.
- [4] Chicago. 2020. Divvy Trips. <https://data.cityofchicago.org/Transportation/Divvy-Trips/fg6s-gzvg>.
- [5] Chicago. 2020. Taxi Trips. <https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew>.
- [6] Yong Shean Chong and Yong Haur Tay. 2017. Abnormal event detection in videos using spatiotemporal autoencoder. In *International Symposium on Neural Networks*. Springer, 189–196.
- [7] Zulong Diau, Xin Wang, Dafang Zhang, Yingru Liu, Kun Xie, and Shaoyao He. 2019. Dynamic spatial-temporal graph convolutional neural networks for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 890–897.
- [8] Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. 2019. Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM, 594–602.
- [9] Vijay Prakash Dwivedi, Chaitanya K Joshi, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. 2020. Benchmarking graph neural networks. *arXiv preprint arXiv:2003.00982* (2020).
- [10] Haoyi Fan, Fengbin Zhang, and Zuoyong Li. 2020. AnomalyDAE: Dual autoencoder for anomaly detection on attributed networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5685–5689.
- [11] Matthias Fey and Jan E. Lenssen. 2019. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.
- [12] Maxi Fischer, Willi Gierke, Thomas Kellermeier, Ajay Kesar, Axel Stebner, and Daniel Thevesen. 2019. Anomaly Detection on Time Series: An Evaluation of Deep Learning Methods. <https://github.com/KDD-OpenSource/DeepADoTS>. Accessed: 2020 Jul.
- [13] Palash Goyal, Nitin Kamra, Xinran He, and Yan Liu. 2018. Dyngem: Deep embedding method for dynamic graphs. *arXiv preprint arXiv:1805.11273* (2018).
- [14] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, 1024–1034.
- [15] Simon Hawkins, Hongxing He, Graham Williams, and Rohan Baxter. 2002. Outlier detection using replicator neural networks. In *International Conference on Data Warehousing and Knowledge Discovery*. Springer, 170–180.
- [16] Yue Hu and Dan Work. 2019. Robust Tensor Recovery with Fiber Outliers for Traffic Events. *arXiv preprint arXiv:1908.10198* (2019).
- [17] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR 2015* (2015).
- [18] Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* (2016).
- [19] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*.
- [20] Prakhyat G Kulkarni, SY Praneet, RB Raghav, and Bhaskarjyoti Das. 2020. Deep Detection of Anomalies in Static Attributed Graph. In *International Conference on Machine Learning, Image Processing, Network Security and Data Sciences*. Springer, 627–640.
- [21] Atsutoshi Kumagai, Tomoharu Iwata, and Yasuhiro Fujiwara. 2020. Semi-supervised Anomaly Detection on Attributed Graphs. *arXiv preprint arXiv:2002.12011* (2020).
- [22] Yuening Li, Xiao Huang, Jundong Li, Mengnan Du, and Na Zou. 2019. SpecAE: Spectral AutoEncoder for Anomaly Detection in Attributed Networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2233–2236.
- [23] Hao Liu, Ting Li, Renjun Hu, Yanjie Fu, Jingjing Gu, and Hui Xiong. 2019. Joint representation learning for multi-modal transportation recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1036–1043.
- [24] Wei Liu, Yu Zheng, Sanjay Chawla, Jing Yuan, and Xie Xing. 2011. Discovering spatio-temporal causal interactions in traffic data streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1010–1018.
- [25] Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. 2016. LSTM-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv:1607.00148* (2016).
- [26] Abdulla Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. 2020. Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14424–14432.
- [27] Nashville. 2019. 2019 NFL Draft Event Schedule. <https://www.visitmusiccity.com/nfldraft/nfl-draft-event-schedule>
- [28] Mayu Sakurada and Takehisa Yairi. 2014. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*. 4–11.

- [29] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*. Springer, 593–607.
- [30] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.
- [31] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. 2019. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2828–2837.
- [32] TDOT. 2019. TDOT Weekly Construction Report for Middle Tennessee. <https://preprod.tn.gov/tdot/news/2019/4/17/tdot-weekly-construction-report-for-middle-tennessee--april-18-24--2019.html>
- [33] Fei Tian, Bin Gao, Qing Cui, Enhong Chen, and Tie-Yan Liu. 2014. Learning deep representations for graph clustering. In *AAAI*, Vol. 14. 1293–1299.
- [34] NYC TLC. 2020. TLC Trip Record Data. <https://www1.nyc.gov/site/tlc/index.page>.
- [35] Uber. 2020. Uber Movement. <https://movement.uber.com/>.
- [36] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).
- [37] Chun Wang, Shirui Pan, Guodong Long, Xingquan Zhu, and Jing Jiang. 2017. Mgae: Marginalized graph autoencoder for graph clustering. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 889–898.
- [38] Yuandong Wang, Hongzhi Yin, Hongxu Chen, Tianyu Wo, Jie Xu, and Kai Zheng. 2019. Origin-destination matrix prediction via graph convolution: a new perspective of passenger demand modeling. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1227–1235.
- [39] Zihan Wang, Zhaochun Ren, Chunyu He, Peng Zhang, and Yue Hu. 2019. Robust Embedding with Multi-Level Structures for Link Prediction.. In *IJCAI*. 5240–5246.
- [40] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems* (2020).
- [41] Haowen Xu, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, Dan Pei, Yang Feng, et al. 2018. Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications. In *Proceedings of the 2018 World Wide Web Conference*. 187–196.
- [42] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575* (2014).
- [43] Shiming Yang, Konstantinos Kalpakis, and Alain Biem. 2014. Detecting road traffic events by coupling multiple timeseries with a nonparametric Bayesian method. *IEEE Transactions on Intelligent Transportation Systems* 15, 5 (2014), 1936–1946.
- [44] Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2017. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (2017).
- [45] Wenchao Yu, Wei Cheng, Charu C Aggarwal, Kai Zhang, Haifeng Chen, and Wei Wang. 2018. Netwalk: A flexible deep embedding approach for anomaly detection in dynamic networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2672–2681.
- [46] Shuangfei Zhai, Yu Cheng, Weineng Lu, and Zhongfei Zhang. 2016. Deep structured energy based models for anomaly detection. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*. 1100–1109.
- [47] Li Zheng, Zhenpeng Li, Jian Li, Zhao Li, and Jun Gao. 2019. AddGraph: Anomaly Detection in Dynamic Graph Using Attention-based Temporal GCN.. In *IJCAI*. 4419–4425.
- [48] Chong Zhou and Randy C Paffenroth. 2017. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 665–674.
- [49] Le-kui Zhou, Yang Yang, Xiang Ren, Fei Wu, and Yueteng Zhuang. 2018. Dynamic Network Embedding by Modeling Triadic Closure Process.. In *AAAI*. 571–578.
- [50] Dali Zhu, Yuchen Ma, and Yinlong Liu. 2020. DeepAD: A Joint Embedding Approach for Anomaly Detection on Attributed Networks. In *International Conference on Computational Science*. Springer, 294–307.
- [51] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. 2018. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*.

A SUPPLEMENTARY MATERIAL

A.1 Experiment data preparation and simulation

In this section, we detail the data reprocessing and simulation steps. We also show the data distribution of travel times for the three datasets used in the paper.

As described in the main text, the dataset is split into training and testing sets. The training set is assumed to have no major events (anomalies). To compare methods in a controlled setting, we create a test set with synthetic labeled anomalies which are injected into cleaned traffic conditions. We synthetically generate the test set from real data due to the unavailability of city scale traffic datasets pre-labeled with anomalies.

To generate cleaned traffic conditions in the test set which are later polluted with anomalies, we assume the travel time of each OD-pair in each hour of day in each day of week follows a Gaussian distribution. We calculate the mean and variance of the OD pair travel time of corresponding hours from the raw test data from the first quarter of 2019. We then resample the travel times based on the calculated mean and variance to arrive at a complete clean test dataset. Next we inject spatial and temporal anomalies respectively as described in the main text to arrive at the labeled test set with anomalies.

In the Uber Movement dataset corresponding to Nashville, TN, the region is divided into 2219 zones. However many of the OD pairs never record travel times. To learn as much useful information as possible and ensure learning efficiency, we select the top 50 most connected zones, so that the selected zone has one or more trip to at least one third of the total 2219 zones. This results in 41% missing data rate in the training set, and a 30% missing data rate in the test set. The minimum and maximum latitude and longitude extents of each zone are used as the node features, which are scaled to 0-1 before feeding into Con-GAE.

The data distribution of travel times for the three datasets used in the paper are shown in Fig. 7. We can see that Uber Movement dataset contains trips of shorter duration than NYC or Chicago taxi data. Meanwhile, Uber data and Chicago taxi data has more concentrated travel time distribution than NYC taxi data.

A.2 Baseline methods

In this section, we give a brief introduction to each method, and document the model configurations.

- RTC [16]. Robust tensor recovery and completion for group anomaly detection (RTC) is a low-rank tensor decomposition based method that exploits spatio-temporal correlation. This is a non-deep learning baseline method.

RTC has one hyper parameter λ , which controls the tradeoff between the rank of the normal tensor and the size and magnitude of the outlier tensor. We set $\lambda = 0.278$ for Uber data, set $\lambda = 0.390$ for NYC data, and set $\lambda = 0.227$ for Chicago data. The value is chosen empirically following the settings described in the article.

- AE [15]. Autoencoder (AE) is a widely-used neural network model for anomaly detection. It uses a fully connected neural network to build an encoder and a decoder to compress and

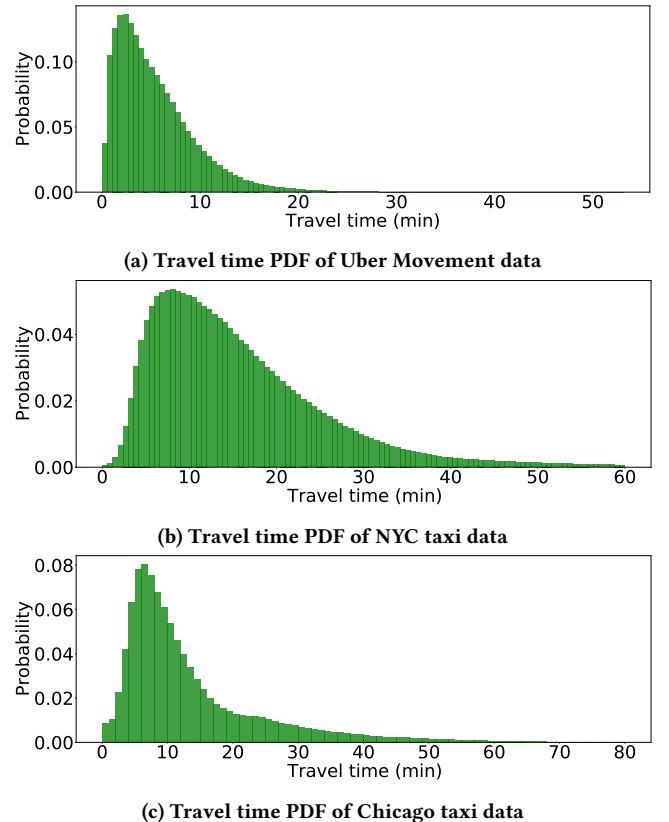


Figure 7: Travel time distribution for three datasets.

reconstruct the data. It uses the reconstruction error as the anomaly score.

We set the hidden size (i.e., the dimension of the low dimensional encoded state) of AE as the same setting used in Con-GAE. Namely, the hidden size is 150 for Uber data, 50 for NYC, and 25 for Chicago data. The learning rate is 0.0005, and the number of epochs is 30. The values were determined to maximize performance using grid search. The sequential length is set as 1, i.e., each input sample contains the traffic conditions of one time step. All other parameters are set as the default values, with random seed set at 0.

- EncDec-AD [25]. EncDec-AD is an LSTM-based encoder-decoder model for time series anomaly detection. EncDec-AD learns the normal pattern of a multi-dimensional time series and detects anomalous times that deviate from the normal pattern via the reconstruction error.

We set the hidden size (i.e., the dimension of the low dimensional encoded state) of AE as the same setting used in Con-GAE. Namely, the hidden size is 150 for Uber data, 50 for NYC, and 25 for Chicago data. The learning rate and number of epochs are 0.0001 and 120 for Uber data, 0.0005 and 30 for NYC, and 0.001 and 70 for Chicago data. EncDec-AD takes time series windows as input, and we set the sequential length at 24 for Uber and Chicago data, and 72 for NYC data, which are also determined via grid search. All other

- parameters are set as the default values, with random seed set at 0.
- DAGMM [51]. *Deep autoencoding Gaussian mixture model* (DAGMM) conducts unsupervised anomaly detection by finding low-dimensional data representations via a deep autoencoder, and modeling the data distribution in the low-dimensional space via a Gaussian mixture model. Instead of setting the hidden size of DAGMM at 150 like the Con-GAE, a grid search revealed the best performance is achieved at a hidden size of 10 for Uber data, 50 for NYC data, and 5 for Chicago data. While the other methods (Con-GAE, AE, EncDec-AD) are reconstruction-based, DAGMM relies on clustering in a low-dimensional space. An encoding dimension too low might not capture the original information, while high dimensional clustering can encounter a curse of dimensionality. The learning rate and number of epochs are 0.0001 and 20 for Uber data, 0.0005 and 50 for NYC data, and 0.0001 and 20 for Chicago data, which yields the best result after parameter grid search. The sequential length is set as 1, i.e., each input sample contains the traffic conditions of one time step. The rest parameters are set at the default values, with the random seed at 0.
 - REBM [46]. REBM is the RNN-based version of *deep structured energy based models* (EBM) for time series anomaly detection. REBM models the data distribution with neural networks, and detects the anomaly by the energy (negative log probability) and reconstruction error. We set the hidden size (i.e., the dimension of the low dimensional encoded state) of AE as the same setting used in Con-GAE. Namely, the hidden size is 150 for Uber data, 50 for NYC, and 25 for Chicago data. The learning rate and number of epochs are 0.001 and 100 for Uber data, 0.0001 and 100 for NYC, and 0.001 and 50 for Chicago data. All other parameters are set as the default values, with random seed set at 0.
 - GCN [19]. As a baseline formulation of Graph neural network, GCN updates node features by averaging over the features of neighbor nodes. We use the final node feature matrix of all nodes as the encoding for the traffic graph, and use the same decoder as our Con-GAE model to decode the edge weights from the node encoding. The reconstruction error across all edges is used as anomaly score. For GCN we use the same structure as Con-GAEmodel. Namely, we use two layers of GCN, the node embeddings are set at 300 and 150 for Uber data; 150 and 50 for NYC data, and 300 and 25 for Chicago data. The dropout rate is set at 0.2. The learning rate is set at 0.0001 for Uber data, 0.001 for NYC data, and 0.01 for Chicago data.
 - GraphSAGE [14] improves from GCN by explicitly concatenating the feature of the node itself after aggregating the neighboring node features. Its encoding-decoding structure is the same as our GCN comparison method. It is akin to our Con-GAE model without the aggregating layer from node embedding to graph embedding and without the time embedding. For GraphSAGE we use the same structure as Con-GAEmodel. Namely, we use two layers of GCN, the node embeddings

for the two layers are set at 300 and 150 for Uber data; 150 and 50 for NYC data, and 300 and 25 for Chicago data. The dropout rate is set at 0.2. The learning rate is set at 0.0001 for Uber data, 0.0005 for NYC data, and 0.05 for Chicago data.

For the methods that do not directly accommodate missing data, namely HA, AE, DAGMM, EncDec-AD and REBM, we impute the missing data via temporal interpolation (i.e., via linear interpolation when values are present to interpolate, or via a mean value when interpolation cannot be performed).

A.3 Reproducibility

In this section, we specify the computing hardware and software used for experiments, as well as the data availability.

The GPU used for the experiments is a GTX 1080 (12GB) operated on Ubuntu 18.04.4 LTS. We use the python package Pytorch Geometric [11] version 1.4.3 to implement the GCN.

The Uber Movement time series data is from Nashville, TN, which is not yet publicly available on the Uber Movement portal [35]. Uber Movement is planning to release the data on the portal, timed with the publication of this work. Uber Movement data is available under a Creative Commons, Attribution Non-Commercial license.¹.

¹<https://movement.uber.com/faqs?lang=en-US>