End-to-End Learning from Complex Multigraphs with Latent Graph Convolutional Networks

Floris Hermsen  
University of Amsterdam & Owlin  
Fabian Jansen  
ING WBAA

Peter Bloem  
VU Amsterdam  
Wolf Vos  
ING

ABSTRACT
We study the problem of end-to-end learning from complex multigraphs with potentially very large numbers of edges between two vertices, each edge labeled with rich information. Examples of such graphs include financial transactions, communication networks, or flights between airports. We propose Latent-Graph Convolutional Networks (L-GCNs), which can successfully propagate information from these edge labels to a latent adjacency tensor, after which further propagation and downstream tasks can be performed, such as node classification. We evaluate the performance of several variations of the model on two synthetic datasets simulating fraud in financial transaction networks, to ensure that the model must make use of edge labels in order to achieve good classification performance. We find that allowing for nonlinear interactions on a per-neighbor basis enhances performance significantly, while also showing promising results in an inductive setting.

KEYWORDS
Deep learning, Graph Convolutional Networks, end-to-end learning, multigraphs

1 INTRODUCTION
Much of human interaction can be viewed from the perspective of large, complex network structures. Examples vary from transportation networks, social media interactions and knowledge bases to financial transaction networks. A lot of valuable information resides in these networks, and detecting the right (complex) patterns may have implications ranging from more efficient logistics and improved targeted advertising, all the way to crime prevention and fraud detection. This last example of uncovering fraud in financial transaction networks will serve as the context of this study.

The complex nature of these types of data sets makes them ideally suited for deep learning techniques, and in recent years much progress has been made through the development of Graph Convolutional Networks (GCNs) [7]. These architectures and their spin-off siblings, such as Relational Graph Convolutional Networks (R-GCNs) [10] and Graph Attention Networks (GATs) [12], have shown great performance in entity classification and link prediction tasks, improving state-of-the-art results. Until now, however, these networks have been of a relatively straightforward structure, meaning that relations are binary in nature, such as scientific citation networks and simple social networks.

A financial transaction network, on the other hand, comes with many additional degrees of complexity. Relations between entities are not described by a binary encoding, not by a single weight, not even by a set of attributes: they are represented by entire sequences of varying length, composed of transactions and their individual characteristics (multi-edges). In the domain of node classification, labels related to the downstream task are often not available at the edge level. Additionally, one often does not even know which lower-level patterns one is looking for. All of this favors the development of neural networks that can be trained in an inductive setting. Devising an architecture that can take transaction-level information as well as network structure into account at the same time, is no straightforward task, and has yet to be demonstrated on data sets of this complexity within the domain of GCNs.

In this study, we investigate if we can perform end-to-end learning on multigraphs in the form of synthetic financial transaction networks. We propose the concept of Latent-Graph Convolutional Networks (L-GCNs), in which a learning mechanism transforms multi-edge populations (transaction sequences) into latent representations, which can then serve as input to a GCN-like architecture for further propagation in the form of an adjacency tensor. We hypothesize that such an architecture can exploit information unavailable to either non-GCN architectures relying on local neighborhood aggregation or standard GCN architectures that do not have access to transaction-level information. We use synthetic datasets we generate ourselves in order to be certain that important information, relevant to the downstream task, resides at the transaction level, and hence can only be extracted by means of successful end-to-end learning1.

We first demonstrate our prototypes by performing node classification tasks in a transductive setting, as is common within the realm of graph convolutional networks. This entails that the model is given access to the entire graph during training, yet with only a small subset of the node labels provided. In transductive learning, we do not expect the model to generalize to data that has not been seen during training. We then conclude by exploring how well this approach translates to the inductive case by applying our trained models to newly generated, unseen graphs.

2 RELATED WORK
2.1 GCN
Recently, a new field in machine learning has emerged in which neural network architectures are proposed that allow information to explicitly propagate along graphs. An important example of this are graph convolutional networks [2–4, 7]. In a publication by Kipf and Welling [7], spectral graph theory is leveraged to motivate a

1Our synthetic data sets and implementations can be found on GitHub: github.com/florishermsen/L-GCN
neural network architecture with the following propagation rule:

\[
H^{(l+1)} = \sigma(I + \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)}) W^{(l)},
\]

(1)

instructing how to compute a next layer of node embeddings \(H^{(l+1)}\) from the previous layer \(H^{(l)}\). Given a graph \(G\) with vertices \(v_i \in V\), \(I_V\) is the identity matrix of size \(|V| \times |V|\) and \(\tilde{A}\) is the adjacency matrix of shape \(|V| \times |V|\) with added self-connections\(^2\) and \(\tilde{D} = \sum_j A_{ij}\) represents a normalization matrix obtained through row and column-wise summation of \(\tilde{A}\). By convention, \(H^{(0)} = X\), which is the original \(|V| \times |F|\) matrix representing the graph vertices \(\mathcal{V}\) and their associated features \(\mathcal{F}\). Finally, \(\sigma\) is a nonlinear activation and \(W^{(l)} \in \mathbb{R}^{|F| \times |F|}\) is the layer-specific weight matrix, which determines the size of the output embedding. The final embedding layer \(H^{(M)}\) can serve as an output layer of appropriate size, depending on the downstream task at hand, for instance node classification.

Classic GCN architectures essentially facilitate diffusion of information over graphs. The more GCN layers an architecture contains, the more heavily the information is smoothed. It is therefore important to recognize that the performance of a traditional GCN relies heavily on a high degree of intraclass clustering \([7]\). This is a result of the convolutional kernel being determined by the adjacency matrix \(A\), which is inherently fixed. Also note that all vertices are either binary encoded or represented by a single weight. This is a limiting aspect when it comes to graphs containing more detailed information regarding relations between vertices.

2.2 Message Passing Frameworks

The propagation rule of the standard GCN (as displayed in Equation 1) can be reformulated as a special case of a message-passing framework (MPF) \([6]\). The main difference in notation is that an MPF defines propagation rules at the node level. This is more in line with the actual implementation of the architectures proposed in this study (see Section 4.5) and will therefore be used for the remainder of this document.

In the context of an MPF and a binary adjacency matrix, Equation 1 takes the following shape:

\[
h_i^{(l+1)} = \sigma\left(\frac{1}{c_i} h_i^{(l)} + \sum_{j \in N_i} h_j^{(l)} \right) W^{(l)},
\]

(2)

where \(h_i^{(l)}\) refers to the embedding of node \(v_i\) in layer \(l\), \(N_i\) denotes the set of direct neighbors of node \(v_i\), and \(c_i\) is a node-specific normalization constant, which typically takes a value of \(|N_i| + 1\) (the +1 being a remnant of added self-connections), in case of an undirected network.

In case the graph has edge weights \(w_{ij}\), Equation 2 can be expanded to reflect this as follows:

\[
h_i^{(l+1)} = \sigma\left(\frac{1}{c_i} w_{ij} h_i^{(l)} + \sum_{j \in N_i} w_{ij} h_j^{(l)} \right) W^{(l)},
\]

(3)

with the normalization constant now taking the form of:

\[
c_i = w_s + \sum_{j \in N_i} w_{ij}.
\]

(4)

Note that a self-weight \(w_s\) has been introduced, which can be set to be a constant as well as turned into a trainable parameter, the latter allowing the network to determine the strength of self-connections during training.

2.3 R-GCN

In situations often encountered in the realm of social networks, for instance, graphs contain edges representing different types of relations. In such a case, an edge can be represented by a one-hot encoded vector with a length corresponding to the cardinality of the set of all relations \(\mathcal{R}\). Schlichtkrull et al. \([10]\) use these vectors to replace the original adjacency matrix \(A\) from Equation 1 with an adjacency tensor of shape \(|V| \times |V| \times |\mathcal{R}|\). Formulated in the context of message passing frameworks, this leads to the following propagation rule:

\[
h_i^{(l+1)} = \sigma \left( \frac{1}{c_i} \sum_{r \in \mathcal{R}} \sum_{j \in N_i} w_r^{(l)} h_j^{(l)} W_r^{(l)} \right),
\]

(5)

where \(W_r \in \mathbb{R}^{|F| \times |F|\times|\mathcal{R}|\} \) now are relation-specific weight matrices, \(w^s\) denotes the weight matrix associated with self-connections (both trainable), and \(c_i\) are relation-specific normalization constants, typically of value \(|N_i\)\(^r\). In order to facilitate the directional nature of some interactions, the set of relations \(\mathcal{R}\) was expanded to contain a version in both edge directions (incoming and outgoing), by means of both canonical and inverse variations of each relation.

3 LATENT GRAPHS

3.1 Pseudo-relations

We can extend the multi-relational approach of R-GCNs to the case of edges being represented by a vector \(w_{ij}\) containing multiple weights across different edge attributes. These attributes can effectively be regarded as non-binary pseudo-relations \(\mathcal{R}\). The propagation rule now takes a form akin to Equation 3:

\[
h_i^{(l+1)} = \sigma \left( \frac{1}{c_i} \sum_{r \in \mathcal{R}} \sum_{j \in N_i} w_r^{(l)} h_j^{(l)} W_r^{(l)} \right),
\]

(6)

with \(c_i\) now taking the form of:

\[
c_i = \sum_{r \in \mathcal{R}} \left( \sum_{j \in N_i} w_r^{(l)} \right).
\]

(7)

Note that \(w_r^{(l)}\) from Equation 5 has been absorbed into \(W_r^{(l)}\). The weight of self-connections can now be learned by means of a self-weight vector \(w_s\). In essence, this method obtains an updated node representation for every pseudo-relation separately, after which a weighted summation takes place in order to arrive at the final embedding. The original publication by Schlichtkrull et al. \([10]\) alludes to the possibility of replacing the weighted sum with a more elaborate learning mechanism, such as a fully connected layer, but this was left for future work.
3.2 L-GCN

With L-GCN (Latent-Graph Convolutional Networks), we propose turning the edge weights \( w_{ij} \) from Equation 3 into trainable parameters in an end-to-end fashion, using the output of a learning mechanism which operates on either a vector describing multiple edge attributes or a sequence \( S_{ij} \) of such vectors (multi-edges):

\[
w_{ij} = \Gamma(S_{ij}),
\]

(8)

where \( \Gamma(\cdot) \) represents an arbitrary, differentiable learning function that is best suited for the data at hand. The information residing in the edge is thus transformed into a single weight, representing a latent relation between nodes \( u_j \) and \( v_j \). These weights can then serve as input for the GCN propagation rule as displayed in Equation 3.

Similarly, a learning mechanism can be chosen that outputs a latent relation embedding vector \( w_{ij} \in \mathbb{R}^L \) of arbitrary length \( L \):

\[
w_{ij} = \Gamma(S_{ij}),
\]

(9)

allowing for a more complex latent relation to be embedded. These learned representations can then similarly serve as input for the propagation rule from Equation 6 at the position of \( w_{ij} \).

3.3 Bidirectionality

We introduce propagation along both edge directions in a similar fashion as in the original publication by Schlichtkrull et al. [10], but adapted to the form of Equation 6. Therefore, the actual multi-edge embedding size is \( 2L \), with edges going in the original direction being represented by

\[
w_{ij}^c = [w_{ij}^1, w_{ij}^2, w_{ij}^3, w_{ij}^4, 0, 0, 0, 0]^{\top},
\]

(10)

and a carbon copy of the edge going in the other direction being represented by

\[
w_{ji}^c = [0, 0, 0, 0, w_{ij}^1, w_{ij}^2, w_{ij}^3, w_{ij}^4]^{\top},
\]

(11)

in the case of \( L = 4 \). The subsequent layers in the network are then able to process incoming and outgoing sets of transactions in different ways and independently. This allows for a bidirectional propagation of information across the graph. Note that the learning mechanism \( \Gamma(\cdot) \) needs to process each set of transactions \( S_{ij} \) only once, to generate both vectors.

3.4 L-GCN+

Equation 6 can be reformulated to first aggregate over all pseudo-relations \( R \) and only then aggregate over all neighbours \( N_i \) as follows:

\[
h_i^{(l+1)} = \sigma \left( \frac{1}{c_i} \sum_{r \in R} w_r^c h_i^{(l)} W_r^{(l)} + \frac{1}{c_i} \sum_{j \in N_i} \sum_{r \in R} w_{ij}^c h_j^{(l)} W_r^{(l)} \right).
\]

(12)

This is mathematically equivalent and more in line with the actual implementation. We can now also replace all relation-specific weight matrices \( W_r^{(l)} \in \mathbb{R}^{[F(l)] \times [F(l+1)]} \) by a single weight matrix \( W \in \mathbb{R}^{L \times [F(l)] \times [F(l+1)]} \), representing a fully connected layer, as follows:\(^3\)

\[
h_i^{(l+1)} = \sigma \left( \frac{1}{c_i} w_i h_i^{(l)} W^{(l)} + \frac{1}{c_i} \sum_{j \in N_i} w_{ij} h_j^{(l)} W^{(l)} \right).
\]

(13)

This enables a straightforward introduction of nonlinear interactions between edge embeddings and node attributes on a per-neighbor basis. This can be achieved by introducing an intermediate layer of size \( 2 \cdot |F(l+1)| \), splitting the original transformation into two matrices \( W_1 \in \mathbb{R}^{L \times |F(l)| \times 2 \cdot |F(l+1)|} \) and \( W_2 \in \mathbb{R}^{2 \cdot |F(l)+1| \times |F(l+1)|} \), allowing for intermediate nonlinearity:

\[
h_i^{(l+1)} = \sigma \left( \frac{1}{c_i} w_i h_i^{(l)} W_1^{(l)} + \sum_{j \in N_i} \frac{1}{c_i} w_{ij} h_j^{(l)} W_2^{(l)} \right),
\]

(14)

with \( f(\cdot) \) a simple, two-layer MLP with its own nonlinearity:

\[
f^{(l)}(x) = \sigma \left( x W_1^{(l)} W_2^{(l)} \right)
\]

(15)

We conjecture that this modification should enable the network to learn more complex patterns within the interactions between the node features and the latent relation attributes.

Note that for sake of brevity, we have omitted biases from the equations. In our implementations, weight matrices \( W \in \mathbb{R}^{AC \times B} \) are accompanied by bias vectors \( b \in \mathbb{R}^B \).

4 METHODS

4.1 Data Set

The synthetic data sets we use in our experiments simulate fraud in financial transaction networks. Vertices \( v_i \in V \) represent actors in the network, and (directed) multi-edges \( e_{ij} \in E \) represent the sequences of transactions \( S_{ij} \) from \( v_i \) to \( v_j \). Labels are only provided at the node level (\( C = 2 \) classes: fraud \( F \) and normal \( N \)), but relevant patterns are hidden in the transaction sets. We have generated two data sets (1-hop and 2-hop), which differ in how the information relevant to the classification task is distributed over the graph.

All actors (nodes \( v_i \)) are described by a set of attributes \( \mathcal{F} \) and all edges are represented by sets of transaction vectors \( t_k \in S_{ij} \). Transactions between the same vertices but in opposite direction are treated as two distinct edge populations \( S_{ji} \) and \( S_{ij} \). Individual transactions come with two properties \( t_{k1} \) and \( t_{k2} \) (time delta \( \Delta t \) and amount). For a more detailed explanation of the generated data sets, see Section 4.4.

4.2 Architecture Components

4.2.1 Multi-Edge Embeddings

We now introduce the learning function \( \Gamma(\cdot) \) in Equation 9 that operates on the sets of transactions \( S_{ij} \), which we will refer to as the MEE mechanism (Multi-Edge Embedding). In case of our prototypes, this submodel consists of a 1D convolutional operation with \( K \) kernels (size 3, stride 1) and \( Z \) channels, that slide over the sequences of transactions \( S_{ij} \). \( Z \) corresponds with the number of attributes for each transaction, which is \( 2 \) in case of our synthetic data sets (\( t_{k1} \) and \( t_{k2} \)). We perform a 1D max pooling operation across the output of each kernel (of shape \( K \times (|S_{ij}| - 2) \)) followed by an activation function in the form of ReLU():

\[
\text{ReLU}(x) = \max(0, x)
\]

The resulting layer of size \( K \) is then followed by two fully connected layers of sizes \( 2L \) and \( L \) in order to obtain our desired output size. Activation functions (also ReLU) are employed after every fully connected layer, and dropout (\( p = 0.2 \)) is applied to the second-to-last layer (of size \( 2L \)). The resulting vector \( \textbf{w}_{ij} \) can be viewed as a
The minimum size of such representations depends on data complexity, as well as the complexity of the network before and after the bottleneck. We therefore leave this to hyperparameter tuning, through which we incidentally also found a latent representation of size $L = 2C = 4$ to work favorably.

$$x^i = \frac{1}{\sqrt{|N_i^\text{in}|}} \sum_{j \in N_i^\text{in}} x_j + x_i^\text{out}$$

$$x^i' = \frac{1}{\sqrt{|N_i^\text{out}|}} \sum_{j' \in N_i^\text{out}} x_{j'} + x_i^\text{in}$$

### 4.2.1: vectors

If we recall Equations 13 and 14, the latent representation $w_{ij}$ obtained via the MEE mechanism are included by means of their tensor product with the previous node embeddings. We suspect that this way, the information residing in these latent representations is integrated into the new embeddings only in an implicit way, posing a challenge for the subsequent layers to retrieve informative components in relation to the training task at hand. We therefore conjecture that combining the concepts of L-GCN and DVE should result in increased performance, by first offering the latent representations of the relevant transaction sets directly on the node embeddings, and then proceeding with a GCN-like architecture.

### 4.3 Architectures

We will now briefly outline the different baseline and prototype architectures that we will include in our training.

**GCN** The classic GCN architecture (Equation 2) will be used as a baseline. To allow for a two-way flow of information across the graph, the original (directed) adjacency matrices are transformed into their undirected counterparts via $A' = A + A^T$. Note that an edge now only indicates that there exist transactions between the relevant nodes. Transaction-specific information is not taken into account.

**MEE-DVE** A non-GCN baseline architecture based on a simple local neighborhood aggregation that only makes use of the DVE mechanisms. After the expansion of the node features, two fully connected layers follow, with input and output sizes corresponding to the size of the node embeddings in the upcoming L-GCN architecture layers.

**L1-GCN** The first GCN-like architecture employing the MEE mechanism, which provides a single latent relation weight $w_{ij}$ ($L = 1$), combined with the GCN propagation rule from Equation 3.

**L4-GCN** The architecture in which the MEE mechanism provides embeddings in accordance with what was discussed in Section 4.2.1: vectors $w_{ij} \in \mathbb{R}^L$ with $L = 4$. Further propagation takes place according to the rule from Equation 13. A schematic overview of this architecture can be seen in Figure 3.

**L4-GCN+** A carbon copy of the L4-GCN architecture, with the additional introduction of nonlinear interaction on a per-neighbor basis, as described in Section 3.4. Propagation takes place according to the rule from Equation 14.
We simulate the task of fraud detection in a network of financial transactions between corporate entities. The challenge is to separate the “normal” actors from the fraudulent. Class labels are only available at the node level, but relevant information also resides in the raw transaction data. This scenario can be viewed as a specific case of a communication network, and there is a multitude of real-world data sets that have a similar shape, some of which were mentioned in the introduction.

An important shared characteristic among all of these examples is that the edge populations are sequential in nature, which can be leveraged by whichever learning mechanism (Γ in Equation 9) is applied. It should be trivial to generalize our experiments to also cover data sets with non-sequential edge populations by choosing a different function Γ(·) accordingly.

4.4.1 Fraud Detection

We simulate the task of fraud detection in a network of financial transactions between corporate entities. The challenge is to separate the “normal” actors from the fraudulent. Class labels are only available at the node level, but relevant information also resides in the raw transaction data. This scenario can be viewed as a specific case of a communication network, and there is a multitude of real-world data sets that have a similar shape, some of which were mentioned in the introduction.

An important shared characteristic among all of these examples is that the edge populations are sequential in nature, which can be leveraged by whichever learning mechanism (Γ in Equation 9) is applied. It should be trivial to generalize our experiments to also cover data sets with non-sequential edge populations by choosing a different function Γ(·) accordingly.

4.4 Data Generation

As mentioned before, the data sets used in our experiments are synthetic in nature. The main reason for this, is that this allows us full control over the patterns we generate in the transaction sequences. This way, we are absolutely certain that relevant information, important to the classification task, resides in the multi-edge populations, and can only be extracted by means of successful end-to-end learning. There are some additional benefits to working with synthetic data. Because of this full control over the hidden patterns, we can more easily perform model introspection after training, as we can generate edge-level labels. Finally, even though real-world networks with these structures are abundant, there is a lack of good quality public data sets (well-labeled data in general), in part due to privacy-related constraints. We leave the application of our models to real-world data to future work.

4.4.2 Network Generation

We first generate a set of 50,000 vertices \( v_i \in V \), randomly allocated to one of two classes, N (normal) or F (fraud), with a class ratio of 9:1. Next, we generate 125,000 edges \( e_{ij} \in E \) between said vertices to create our graph. The order of magnitude of these numbers and the vertex-to-edge ratio are similar to those of data sets commonly used for testing GCN-like architectures, such as Citeseer, Cora and Pubmed (scientific citation networks) [13].

Since most real-world networks have a degree distributions that follow power laws [9], we employ the concept of preferential attachment [1] in order to mimic this. In the original version of this model, one starts with a small, connected network, after which vertices are added one-by-one. The probability of a new vertex being connected to one of the existing nodes \( v_i \) scales directly with the number of edges the latter already has (its node degree \( k_i \)):

\[
p(e_{ij}) = \frac{k_j}{\sum_{k'} k'}.
\]

Since all nodes already have been generated, we need to transform this expression into probabilities for each node \( v_i \) to be either the source or the target node of a newly introduced edge:

\[
p(v_i) \propto (k_i + 1),
\]

Figure 3: Schematic overview of one layer of the L4-GCN architecture (out of two in total). Outputs \( w_{ij} \) from the MEE learning mechanism \( (L=4) \) are used to encode two similar but different latent relations in both directions. The R-GCN-like propagation rule from Equation 13 generates an intermediate tensor representation \( H^{(l)} \), which is flattened and transformed into the desired new node embeddings \( H^{(l+1)} \) by means of a fully connected layer applied to each intermediate node embedding \( h_i^{(l+1)} \).
We can now add sets of transactions $S_{ij}$ to the previously sampled edges. All transactions take place within the same time span of a year and have two attributes: time and amount. An edge can receive one of three types of transaction contracts:

1. weekly payments of fixed amount,
2. monthly payments of fixed amount,
3. payments with random intervals and random amount.

Types 1 and 2 receive additional small, randomized offsets on their time attributes in order to introduce a degree of noise. Depending on the classes of the source and target nodes, these transaction sets $S_{ij}$ are modified. In case that $v_i \in F$ and $v_j \in N$ we introduce fraud type A, having the following effects (in order of the previously discussed transaction types)$^4$:

1. some weekly payments are missing,
2. some monthly payments are missing,
3. some payments have a decreased amount by a factor 10.

In case $v_i \in N$ and $v_j \in F$ we introduce fraud type B, having similar but opposite effects:

1. some weekly payments occur twice,
2. some monthly payments occur twice,
3. some payments have an increased amount by a factor 5.

4.4.6 Data Summary

An overview of the two data sets can be seen in Table 1. Values for $|V|$ differ from their initial values due to the removal of zero-degree nodes (see Section 4.4.2). Values for $|E|$ also differ slightly because of the removal of generated self-connections.

Before the data sets are passed to the architectures, all node features are normalized: within each attribute, data points are transformed linearly in order to fit inside a $[0,1]$ range. In the transaction sets, individual times are transformed into relative time deltas within each sequence, disregarding all first transactions. Both the time deltas and transaction amounts are then transformed into their

---

$^4$All of these modifications take place with a per-transaction probability of $1/3$. 
We therefore split the transaction populations into appropriate with ratios 5/5/90. The validation sets were used during architecture and hyperparameter optimization. The test sets were held out entirely, until final settings and architectures were decided on. Both data sets, together with the used train/validation/test split schemes, are made publicly available for further experimentation\(^5\).

### 4.5.1 Hyperparameters

All GCN-like architectures consist of two of the GCN layers of their individual characteristics. Adding more layers does not provide any noticeable benefits (this may be related to excessive information diffusion). In all ANN architectures, we apply dropout \((p = 0.5)\) to the output of the first GCN-layer for regularization purposes. Note that each of the two GCN layers of every L-GCN-like architecture contain their own instance of the MEE learning mechanism, allowing for different embeddings to be learned in each stage of the propagation.

Hyperparameter sweeps were performed to find good node embedding sizes for \(H^{(1)}\) at layer 1, learning rate and weight decay. In case of architectures employing MEE, a number of convolutional kernels \(K = 20\) was found to be optimal with respect to other options explored, together with a learning rate of \(5 \cdot 10^{-4}\), weight decay of \(5 \cdot 10^{-4}\) and an intermediate embedding size of 20 for \(H^{(1)}\). We run all training sessions for 2000 epochs, after which training settles into an equilibrium.

## 5 RESULTS

All models were subjected to 10 training sessions with independent, random weight initialization and the settings described in Section 4.5.1. The resulting accuracy scores on the test sets are averaged and displayed in Table 2, accompanied by their standard error. Note that the resulting variance is only due to a different initialization of weights, as the training/validation/test splits do not vary. Because of the severe class imbalance, accuracy scores tend to be misleading. We therefore also display the AUC (area under the receiver operating characteristic curve), obtained from the original, decimal log values, with an appropriate normalization constant (again to produce data points approximately within a \([0,1]\) range).

### 4.5 Implementation

All architectures are built on top of PyTorch Geometric (PyG, version 1.2.0), an extension library for PyTorch centered around GCN-like architectures [5]. Training is performed using CUDA on Nvidia V100 Tensor Core GPUs.

For training speed optimization purposes, all transactions are stored in a single tensor. Transaction sets with fewer transactions than \(|S_{ij}|_{\text{max}}\) are padded with zeros. This has no impact on back-propagation (aside from edge cases) since the convolutional operation in our MEE learning mechanism is followed by a max-pool operation and an activation function (see Section 4.2.1). This does however, dramatically increase the memory allocation on the GPU. We therefore split the transaction populations into appropriate batches based on their number of transactions \(|S_{ij}|\). The net effect of this batching is a memory reduction of \(-6\times\) and a training speed increase of \(-2\times\). Note that training still takes place in a full-batch fashion; this splitting is solely employed in order to reduce the required amount of padding.

We use a weighted version of binary cross-entropy loss for all of our architectures, since this is most appropriate for binary classification problem with unbalanced classes. All training sessions employ the Adam optimizer.

For both data sets we use a fixed train/validation/test split scheme with ratios 5/5/90. The validation sets were used during architecture and hyperparameter optimization. The test sets were held out entirely, until final settings and architectures were decided on. Both data sets, together with the used train/validation/test split schemes, are made publicly available for further experimentation\(^5\).

### Table 1: The synthetic data sets and their characteristics: number of vertices and edges, sizes of classes F, N and M, and total number of transactions. Note that \(M \subseteq N\), with no labels for \(M\) in the final data set.

| Data Set | \(|V|\) | \(|E|\) | \(|N|\) | \(|F|\) | \(|M|\) | \(\Sigma|\mathcal{S}|\) |
|----------|--------|--------|--------|--------|--------|-------------|
| 1-Hop    | 41792  | 124996 | 37574  | 4218   | -      | 6643964     |
| 2-Hop    | 41758  | 124995 | 37585  | 4173   | 11392  | 6618483     |

![1-hop and 2-hop structures](image)

**Figure 4:** Schematic examples of a 1-hop structure (left) versus a 2-hop structure (right). In the 2-hop structure, fraudulent activity takes place at least once removed from the fraudulent actor in the network.

\(^5\)Our synthetic data sets and implementations can be found on GitHub: github.com/florishermsen/L-GCN
Table 2: Model performance on the 1-hop and 2-hop synthetic data sets. Accuracy and AUC (area under the receiver operator characteristics curve) are averaged over 10 runs and accompanied by their standard error. Also displayed are the duration of the training sessions and required GPU RAM allocation. Training was performed on Nvidia V100 Tensor Core GPUs.

| Architecture          | 1-Hop                      | 2-Hop                      | Time  | GPU RAM | N \_params |
|-----------------------|----------------------------|----------------------------|-------|---------|------------|
|                       | Accuracy | AUC    | Accuracy | AUC    |           |           |
| Baselines             |          |       |          |       |           |           |
| Random Guess          | 50.06 ± 0.08 | 0.500 ± 0.002 | 49.90 ± 0.09 | 0.497 ± 0.001 |           |           |
| Majority Class        | 89.91    | 0.500 | 90.01    | 0.500 |           |           |
| GCN                   | 49.62 ± 0.27 | 0.501 ± 0.001 | 56.50 ± 0.44 | 0.513 ± 0.001 | ~ 17s | 1.82 GB | 322 |
| MEE-DVE               | 97.41 ± 0.15 | 0.969 ± 0.007 | 80.22 ± 0.74 | 0.889 ± 0.012 | ~ 5m  | 4.64 GB | 746 |
| Prototypes            |          |       |          |       |           |           |
| L1-GCN                | 56.49 ± 2.83 | 0.514 ± 0.013 | 61.85 ± 1.76 | 0.556 ± 0.028 | ~ 9m  | 7.58 GB | 994 |
| L2-GCN                | 86.94 ± 5.29 | 0.862 ± 0.061 | 80.13 ± 2.41 | 0.862 ± 0.045 | ~ 10m | 7.96 GB | 1694 |
| L2-GCN+               | 95.38 ± 1.13 | 0.970 ± 0.006 | 85.50 ± 1.46 | 0.917 ± 0.034 | ~ 10m | 8.11 GB | 3746 |
| L4-GCN                | 97.10 ± 0.29 | 0.972 ± 0.004 | 89.65 ± 0.34 | 0.952 ± 0.004 | ~ 11m | 8.52 GB | 3118 |
| L4-GCN+               | 97.30 ± 0.41 | 0.983 ± 0.002 | 87.34 ± 1.26 | 0.951 ± 0.004 | ~ 11m | 8.47 GB | 6370 |
| L4-GCN+ & DVE         | 97.91 ± 0.21 | 0.987 ± 0.002 | 89.60 ± 0.97 | 0.962 ± 0.006 | ~ 14m | 11.5 GB | 8930 |

5.2 2-Hop Data Set

On the second data set, the baseline GCN architecture performs similarly. This is no surprise as both the class ratios and the distributions the node attributes are drawn from are identical. Whereas the MEE-DVE baseline architecture performed reasonably well on the 1-hop data set, its AUC and accuracy score are now significantly lower. This is in line with expectations, since the a lot of relevant information about individual nodes can no longer be found within the direct, local neighborhood.

The L1-GCN architecture still performs marginally better than random guessing for reasons outlined above. The performance of the other L-GCN prototypes still increases with each added level of complexity in a similar fashion, albeit at slightly lower performances (both accuracy and AUC). The 2-hop structure must therefore pose a slightly more difficult classification challenge to our models, as relevant information is hidden at a deeper level in the data set. The L4-GCN architectures, however, still perform surprisingly well with accuracy scores close to 90%, and the L4-GCN+ & DVE variation obtaining an average AUC of over 0.962. A box plot of the AUC scores of the L-GCN prototypes on both the 1-hop and 2-hop data sets can be seen in Figure 5.

5.3 Inductive Setting

Our investigation of model performance so far has taken place in a transductive setting, entailing that inference is performed on the same graph that was used during training (albeit with a small subset of the labels provided), as is common in the realm of GCNs. It is an interesting exercise, however to examine how well our trained models generalize (inductive learning). To this end, we generate similar but different versions of the 1 and 2-hop data sets (both graph structure, transaction sets and node attributes) and apply the same instances of our trained architectures. The results can be seen in Table 3 and Figure 6.

The first thing to notice is that all models show much lower levels of performance. Remarkably, the MEE-DVE baseline architecture now even performs on par with random guessing, which confirms the suspicion that the little performance shown in the transductive setting is
5.4 MEE Inspection

To verify that the trained architectures really extract information from the transactions themselves, we inspect the convolutional filters learned by the MEE mechanism. We extract the learned parameters in the MEE module from an instance of the best performing architecture (L4-GCN+ & DVE) trained on the 1-hop data set, and initialize a stand-alone version of the mechanism. We generate a new set of transactions from the same distributions and feed them to the model in order to retrieve the latent representations. Next, we subject these embeddings to dimensionality reduction employing the t-SNE algorithm [8], the results of which can be seen in Figure 7.

It is clear that the learning mechanism is able to offer latent representations based on which a distinction between the different transaction set and fraud types can be made. Similarly to other transaction set types, the majority of those of type 1 (weekly transactions) with fraud type B (represented by the red crosses) appear in the same latent cluster. Interestingly, the embeddings offer a distinction between transaction type combinations beyond just the type of fraud, even though this was not required for the downstream task.

We can delve deeper into this example by inspecting the MEE learning mechanism at the level of the convolutional kernels, in order to identify the relevant filters. This can be seen in Figure 8. Displayed are an example of a transaction set of type 1 and fraud type B (top), convolutional kernels related to such patterns (bottom) and their response to the data (middle). Transaction sets of type 1 and fraud type B present themselves with sudden, one-off drops in the values in channel 1 (log \( \Delta t \)), corresponding with occasional, chance-based double transactions (see Section 4.4.4). Some of the 20 kernels, 4 of which we show in Figure 8, have learned

\[
\begin{array}{|c|c|c|}
\hline
\text{Architecture} & \text{1-Hop AUC} & \text{2-Hop AUC} \\
\hline
\text{Baselines} & & \\
GCN & 0.500 \pm 0.000 & 0.500 \pm 0.000 \\
MEE-DVE & 0.499 \pm 0.001 & 0.500 \pm 0.000 \\
\hline
\text{Prototypes} & & \\
L1-GCN & 0.500 \pm 0.000 & 0.500 \pm 0.000 \\
L2-GCN & 0.581 \pm 0.041 & 0.639 \pm 0.046 \\
L2-GCN+ & 0.859 \pm 0.055 & \textbf{0.816} \pm 0.041 \\
L4-GCN & 0.653 \pm 0.051 & 0.575 \pm 0.024 \\
L4-GCN+ & \textbf{0.907} \pm 0.051 & 0.784 \pm 0.049 \\
L4-GCN+ & \textbf{0.860} \pm 0.054 & 0.802 \pm 0.046 \\
\hline
\end{array}
\]

Table 3: Model performance (AUC) in an inductive setting. AUC values are averaged over the same 10 runs from Table 2 and accompanied by their standard error.

indeed related to relying on graph-specific statistical fluctuations in the node attributes.

Other L-GCN architectures do generalize to some extent, however, the L4-GCN+ even obtaining an average AUC score over 0.90 on the 1-hop data set, with the majority of the measurements over 0.95 (see Figure 6). Introducing the "plus" variation of the architectures seems to be a crucial factor in this regard, with both L2-GCN+ and L4-GCN+ outperforming their "non-plus" counterparts by a wide margin. The lower score of the L4-GCN+ & DVE architecture should be attributed to the DVE mechanism being prone to overfitting on graph-specific patterns, as indicated by the drop in performance of the MEE-DVE architecture. Figure 6 Why the L-GCN+ variations are more prone to learning generalized patterns remains an open question for the time being. Note that this is especially surprising since the MEE learning mechanism is the same across all architectures.
Figure 8: An example transaction set of type 1 and fraud type B (top), convolutional kernels related to such patterns (bottom) and their response to the data (middle). Channel 2 responses are omitted since the input is constant. Kernel parameter values have been corrected for the kernel bias and are all scaled to within the same domain, since weights further down the architecture determine their influence.

6 DISCUSSION

In essence, a classical GCN is an information diffusion mechanism through which data residing in node attributes can be smoothed over local graph neighborhoods. Because this diffusion is essentially performed indiscriminately, its success relies heavily on the existence of an edge $e_{ij}$ being heavily correlated with whether vertices $v_i$ and $v_j$ belong to the same class. Kipf and Welling [7] showed that on data sets that fit that description, such as Citeseer, Cora and Pubmed, GCNs were indeed able to improve on node classification accuracy scores with respect to state-of-the-art results at the time.

In data sets that do not exhibit this correlation, a classic GCN is not likely to provide better classification accuracy results than random guessing. That is, unless there is additional information residing in edge attributes or multi-edge populations. We have shown that L-GCN-like architectures are able to exploit this by introducing trainable, discriminative information diffusion. A latent edge representation of size $L = 1$, however, does not offer enough degrees of freedom for the effective integration of information residing in the edges into the node embeddings and can only indicate the relative importance that needs to be assigned to a neighboring node.

Both the introduction of DVE and nonlinear interactions on a per-neighbor basis (L-GCN+) seem to have a significant effect on model performance. We conjecture that the improvement due to DVE relates to the output of the MEE mechanism influencing downstream layers in multiple ways, which we hypothesize leads to a more informative gradient. This also seems to make the models more robust with respect to model initialization, yet further experimentation should determine if this holds in every scenario.

The nonlinear interactions on a per-neighbor basis in L-GCN+ seem to be an essential ingredient in case one wishes to use L-GCNs in an inductive setting. This is one of the more interesting findings of this study, as the L-GCN+ architectures are more complex than their normal variations, containing over twice as many model parameters (see Table 2). This makes the fact that they are more inclined to generalize to unseen networks somewhat counter-intuitive. We conjecture that allowing nonlinear interactions on a per-neighbor basis favors the learning of generalized patterns but leave a more in-depth investigation of this phenomenon to future work.

Further modification of the MEE learning mechanism, most likely simplification, should allow all of our models to generalize better, if so desired. This most likely presents a trade-off between performance in the transductive setting versus the inductive setting. We leave further investigation of this to future work as well.

The way the MEE learning mechanism is configured in our prototypes makes use of the sequential nature of the transaction data. It is conceivable that this may not be a trait of all data sets of a similar structure. Fortunately, there are many different options that can process unordered edge populations, ranging from trainable aggregation schemes to attention mechanisms. Also, the current architectures cannot directly take multi-agent interactions into account, such as conversations between more than two persons in communication networks. Future work should investigate this further.

Working with synthetic data sets has enabled us to successfully demonstrate end-to-end learning on multi-edge graphs with GCN-like architectures. It has also provided the opportunity to show, in a controlled setting, the power of GCN-like architectures in scenarios where relevant information is hidden deeper in the network structure. These are all ideal circumstances when developing a prototype, however, and we do recognize that the demonstrated architectures will need to prove their value through application on real-world data sets. For now, this is left to future work.
7 CONCLUSIONS

In this study, we have shown that we can perform end-to-end learning on complex multi-edge graphs such as financial transaction data with graph convolutional networks. We have employed a learning mechanism that transforms multi-edge populations into latent relations, serving as input for R-GCN-like further propagation, introducing the concept of Latent-Graph Convolutional Networks (L-GCN).

We have successfully implemented the bidirectional propagation of information along directed edges inside the architecture itself, allowing for optimal flow of information. We have shown that our prototypes perform well on two synthetic data sets with relevant information hidden at different depths inside the network structure. Especially the data set with 2-hop correlations, our best prototypes significantly outperform the non-GCN baseline architecture.

Architectures that allow for additional interaction within latent relations display significantly increased performance. We have also shown that transferring latent edge representations directly to their adjacent nodes by means of a local neighborhood aggregation can yield added benefit in the transductive setting, and most notably seems to decrease model initialization sensitivity.

A brief examination of model performance in the inductive setting yields promising results, with a tentative conclusion that allowing for nonlinear interactions on a per-neighbor basis favors the learning of generalized patterns.

Further experimentation should clarify whether these observations also hold for different data sets and scenarios. The next logical step would be to assess the architectures in the context of real-world data sets. For now, we leave this to future work.

REFERENCES

[1] Réka Albert and Albert-László Barabási. 2002. Statistical mechanics of complex networks. Rev. Mod. Phys. 74 (Jan 2002), 47–97. Issue 1. https://doi.org/10.1103/RevModPhys.74.47

[2] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2014. Spectral Networks and Locally Connected Networks on Graphs. In 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings. http://arxiv.org/abs/1312.6203

[3] Michael Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. CoRR abs/1606.09375 (2016). arXiv:1606.09375 http://arxiv.org/abs/1606.09375

[4] David K. Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. 2015. Convolutional Networks on Graphs for Learning Molecular Fingerprints. CoRR abs/1509.09292 (2015). arXiv:1509.09292 http://arxiv.org/abs/1509.09292

[5] Matthias Fey and Jan E. Lenssen. 2019. Fast Graph Representation Learning with PyTorch Geometric. In ICLR Workshop on Representation Learning on Graphs and Manifolds.

[6] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural Message Passing for Quantum Chemistry. CoRR abs/1704.01212 (2017). arXiv:1704.01212 http://arxiv.org/abs/1704.01212

[7] Thomas N. Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. CoRR abs/1609.02907 (2016). arXiv:1609.02907 http://arxiv.org/abs/1609.02907

[8] L.J.P.D. Maaten and GE Hinton. 2008. Visualizing High-Dimensional Data using t-SNE. Journal of Machine Learning Research 9 (01 2008), 2579–2605.

[9] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. 2002. Hierarchical organization of modularity in metabolic networks. science 297, 5586 (2002), 1551–1555.

[10] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2017. Modeling Relational Data with Graph Convolutional Networks. In The Semantic Web - Springer International Publishing, 593–607.

[11] Kent A. Spackman. 1989. Signal Detect Theory: Valuable Tools for Evaluating Inductive Learning. In Proceedings of the Sixth International Workshop on Machine Learning. Morgan Kaufmann, San Francisco (CA), 160 – 163. https://doi.org/10.1016/B978-1-55860-036-2.50047-3

[12] Petar Veličković, Guillaume Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2017. Graph Attention Networks. arXiv e-prints, Article arXiv:1710.10903 (Oct 2017), arXiv:1710.10903 pages. arXiv:stat.ML/1710.10903

[13] Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. 2016. Revisiting Semi-Supervised Learning with Graph Embeddings. (2016). arXiv.cs.LG/1603.08861
A DATA SETS: SAMPLING DISTRIBUTIONS

A.1 Transaction Data

A.1.1 Type 1
The same amount for all transactions within the same sequence:

\[ P(\text{am}) \propto \begin{cases} e^{-\frac{(\text{am}-30)^2}{50^2}} & \text{if } \text{am} > 0 \\ 0, & \text{otherwise} \end{cases} \]

Different time deltas for individual transactions:

\[ \Delta t = 7 \text{ days} + A \text{ minutes} \]

\[ P(A) \propto e^{-\frac{A^2}{200^2}} \]

A.1.2 Type 2
The same amount for all transactions within the same sequence:

\[ P(\text{am}) \propto \begin{cases} e^{-\frac{(\text{am}-200)^2}{150^2}} & \text{if } \text{am} > 0 \\ 0, & \text{otherwise} \end{cases} \]

Different time deltas for individual transactions:

\[ \Delta t = 30 \text{ days} + B \text{ minutes} \]

\[ P(B) \propto e^{-\frac{B^2}{300^2}} \]

A.1.3 Type 3
Same maximum amount for all transactions within the same sequence:

\[ P(\text{max}) \propto \begin{cases} |e^{-\frac{(\text{max}-200)^2}{100^2}}| & \text{if } T \in \mathbb{Z} \\ 0, & \text{otherwise} \end{cases} \]

Different amounts for individual transactions:

\[ P(\text{am}) \propto \begin{cases} e^{-\frac{\text{am}}{3000}} & \text{if } 10 \leq \text{am} \leq \text{max} \\ 0, & \text{otherwise} \end{cases} \]

The same time delta base \( T \) for all transactions within the same sequence:

\[ P(T) \propto \begin{cases} |e^{-\frac{(T-10)^2}{200^2}}| & \text{if } T \in \mathbb{Z} \\ 0, & \text{otherwise} \end{cases} \]

Different time deltas for individual transactions:

\[ \Delta t = T \text{ days} + C \text{ days} + D \text{ hours} + E \text{ minutes} \]

\[ P(C) \propto e^{-\frac{C^2}{20^2}} \]

\[ P(D) \propto \begin{cases} 1, & \text{if } D \in \{1, 2, \ldots, 24\} \\ 0, & \text{otherwise} \end{cases} \]

\[ P(E) \propto \begin{cases} 1, & \text{if } E \in \{1, 2, \ldots, 60\} \\ 0, & \text{otherwise} \end{cases} \]

A.2 Node Features

A.2.1 Number of Employees

\[ P(x) \propto \begin{cases} e^{-0.005x}, & \text{if } 10 \leq x \leq 1500 \\ 0, & \text{otherwise} \end{cases} \]

A.2.2 Turnover

\[ P(x_2) \propto \begin{cases} e^{-0.00005x_2}, & \text{if } 1 \times 10^4 \leq x_2 \leq 1 \times 10^7 \\ 0, & \text{otherwise} \end{cases} \]

A.2.3 Profit

\[ x_3 = x_2 - x_3' \]

\[ P(x_3') \propto e^{-\frac{(x_3')^2}{200^2}} \]

A.2.4 Equity

\[ P(x_4) \propto \begin{cases} e^{-0.00003x_4}, & \text{if } 1 \times 10^5 \leq x_4 \leq 1 \times 10^7 \\ 0, & \text{otherwise} \end{cases} \]

A.2.5 Sector

\[ P(S) \propto \begin{cases} 2 + \sin(S)^2, & \text{if } S \in \{0, 1, 2, 3\} \\ 0, & \text{otherwise} \end{cases} \]

A.2.6 Region

\[ P(R) \propto \begin{cases} 3 + \sin(R+1)^2, & \text{if } R \in \{0, 1, 2, 3, 4\} \\ 0, & \text{otherwise} \end{cases} \]
Figure 9: Node degree distribution (in + out) for both synthetic data sets (left). Distribution of transaction set (edge populations) sizes for the 1-hop data set (right). The distribution for the 2-hop data set is similar.