**Preprint – IJCAI 2025**: This is the accepted version made available for conference attendees.
**Do not cite**. The final version will appear in the IJCAI 2025 proceedings.

# Detecting Illicit Massage Businesses by Leveraging Graph Machine Learning

**Vasuki Garg**[1] , **Osman Y. Özaltın**[1] , **Maria E. Mayorga**[1] and **Sherrie Bosisto**[2]
[1]Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695
[2]Global Emancipation Network, Clermont, Florida 34715
{vgarg5, oyozalti, memayorg}@ncsu.edu, sherrie@globalemancipation.ngo

## Abstract

Thousands of Illicit Massage Businesses (IMBs) are estimated to be operating in the United States by disguising themselves as legitimate establishments while exploiting trafficked workers, harming both the victims and the massage industry. The increasing digital presence of these illicit businesses presents an opportunity for detection, a crucial task for law enforcement and social service agencies aiming to disrupt their operations. Our research leverages user-generated business reviews from Yelp.com, enriched with data from multiple sources, including RubMaps.ch, U.S. Census records, GIS data, and licensing information. We present a feasibility study of developing a graph convolutional network (GCN) for a novel application and exploring its benefits and drawbacks in identifying IMBs. The novelty of our approach lies in its ability to link and analyze businesses, reviews, and reviewers within a heterogeneous network and employ a relational GCN to capture their complex relationships.

## 1 Introduction

The U.S. Department of State defines human trafficking as the use of force, fraud, or coercion to exploit individuals for commercial sex or labor services [Department of State, 2025]. Illicit Massage Businesses (IMBs) are businesses that pose as legitimate establishments but engage in a hybrid form of human trafficking, which includes sex and labor trafficking. According to the Human Trafficking Institute, IMBs thrive on coercion and deception, where victims are lured by false employment advertisements, trapped in debt bondage, psychologically manipulated through fear and shame, and economically exploited by being forced to work long hours for minimal pay [Janis, 2020]. It is estimated that more than 15,000 IMBs operate in the U.S. [The Network, 2024]. Also, this hybrid form of exploitation generates an estimated annual revenue of 12.8 billion USD [Bouche and Crotty, 2017]. Therefore, developing effective methods to detect these establishments will help law enforcement and social service organizations to disrupt their pervasive illicit activities.

Many methodologies have been developed to identify factors associated with IMBs. The evolution of the landscape of the sexually oriented businesses out of the traditional red-light districts with the change in policing approaches and access to better transport and internet services led to comprehensive geospatial studies [Aalbers and Sabat, 2012; Lasker, 2001; Murphy and Venkatesh, 2006]. Analysis of data from massage review board websites like RubMaps.ch, combined with foot traffic data from camera footage, has also been performed, which generated predictions for the total annual demand of IMBs and their spatial clustering based on demographic characteristics in the Houston area [Crotty and Bouché, 2018]. These methodologies have been further expanded to other parts of the U.S., such as Los Angeles County and New York City [Chin *et al.*, 2019] and across the nation [White *et al.*, 2021], establishing common census features for the IMB locations, such as socio-demographic and household characteristics. Further, a study of the geospatial distribution of sex trafficking offenses led to a conclusion of a direct relationship between the closeness to highways, cheap hotels, motels, and the number of such offenses [Mletzko *et al.*, 2018], reaffirmed by an IMB prevalence study [de Vries and Radford, 2021].

The abundant digital presence of IMBs presents both a challenge and an opportunity to detect them. This has given rise to studies aiming to detect IMBs by mining the information on business review websites like Yelp.com [Diaz and Panangadan, 2020]. Approaches to ensemble multiple sentiment analysis methods to understand reviewers' perspectives have been developed [Mensikova and Mattmann, 2018] as well as lexicon and word embedding-based text classification models [Li *et al.*, 2023]. Unlike our study, these approaches have mainly focused on analyzing the review text.

The closest study to ours is that of Tobey et al. [Tobey *et al.*, 2022], which also aims to identify IMBs. They use risk score and decision tree models to detect IMBs by focusing on Yelp reviews enhanced with multi-faceted features [Tobey *et al.*, 2022]. Our work uses a similar data collection and feature creation approach. However, the methods proposed in Tobey et al. [Tobey *et al.*, 2022] treat each business as a standalone, isolated entity and do not take advantage of the information presented by the potential network linking businesses, their reviews, and reviewers. As IMBs operate covertly, analyzing these links enables the aggregation of features from multiple sources to gain deeper insights into illicit operations and potentially improve detection efficacy.

## 2 Related Work and Our Contributions

This section discusses the previous work on illicit node detection and our contributions.

### 2.1 GNN-based Illicit Detection

Graph Neural Networks (GNNs) extend the traditional Neural Networks by learning over data with a graph structure [Scarselli *et al.*, 2009]. Given GNN's ability to explore the underlying interactions between graph components efficiently, they have been established as a leading framework to interpret and learn features from tabular data [Li *et al.*, 2024]. Many different GNN methods, such as Attention-based GNN, Convolutional-based GNN, Meta-path-based GNN, and transformer-based GNN, have been applied to fraud detection [Motie and Raahemi, 2024].

Among the more traditional non-graph-based neural networks, Convolutional Neural Networks have the innate ability to model spatial data effectively. Therefore, they have been utilized to analyze graph data through Graph Convolutional Networks (GCNs), making these one of the most widely adopted types of GNNs [Bhatti *et al.*, 2023]. Most of the applications concerning node classification using a GCN are in the area of financial fraud detection, where each node corresponds to a transaction, and an edge between two transactions represents congruency at a particular feature, such as Message Authentication Code (MAC) ID and/or time interval [Liu *et al.*, 2021]. Other works using homogeneous graphs (graphs with a single type of nodes and edges) with transactional data have considered nodes as the transaction's source and destination while the edge as the transaction itself [Zou and Cheng, 2024].

An extension to this line of work entails using a multi-relational heterogeneous graph or a Relational Graph Convolutional Network (RGCN), which involves different types of nodes representing merchants, cards, and transactions, as well as the edges representing the connections between a card and a transaction and between a transaction and a merchant [Harish *et al.*, 2024]. RGCNs have also been used to detect hostile posts using token embeddings of posts as nodes and edges representing syntactic relationships between these tokens [Sarthak *et al.*, 2021]. An augmentation of RGCN with hierarchical graph contrastive learning has been applied to fake review detection as well [Yao *et al.*, 2024].

One of the most comparable works to our study methodologically uses Hack Forums (a common platform for underground markets) to build a heterogeneous graph for illicitly traded product identification using five types of nodes: buyers, vendors, products, comments, post topics, and six types of edges. That work incorporates buyer and vendor attributes as well as product, comment, and post features to build links between illicit buyers and vendors to detect illicitly traded products [Fan *et al.*, 2020]. Our work is similar in that we incorporate multiple data sources focused on different entities of a business review system to build a heterogeneous graph. The network model built over this heterogeneous graph, described in more detail in Section 3.1, allows us to explore the illicit characteristics of businesses based on their shared reviewers.

**Contributions of Our Work**

- We propose an RGCN model to create a heterogeneous network of massage businesses' user-generated data for business-level classification.
- We conduct extensive experiments to perform a baseline comparison of the proposed RGCN with other state-of-the-art models to showcase the model's competitive performance and feasibility.

## 3 Methodology

Our approach consists of four major components (see Figure 1), and this section describes each.

### 3.1 Graph Construction

A heterogeneous graph can integrate different node types—massage businesses, reviews, and reviewers, enabling the representation of their interactions while preserving unique features for each. Our primary focus is on business nodes. Given labeled businesses, our inference task is a node classification problem, predicting whether a business is *illicit* or *non-illicit*. To the best of our knowledge, no prior work has constructed a network of massage businesses based on reviews and reviewers for IMB detection.

**Heterogeneous Graph Definition**

We denote our heterogeneous undirected graph by $G = (V, E, R)$, where:

- $V$ denotes the set of nodes: `businesses`, `reviews`, and `reviewers`. Each node type has a specific set of features.
- $E$ denotes the set of edges between the nodes: `business` – `review`, and `review` – `reviewer`.
- $R$ denotes the set of relations between the nodes: `has review` (for `business - review` edge) and `written by` (for `review - reviewer` edge).

Figure 2 shows a subgraph of the network for a business node, illustrating the underlying connections between different node types in the heterogeneous graph.

### 3.2 Message Passing and Aggregation in GNNs

GNNs work on the principle that nodes are connected to other nodes of the same type, also known as graph homophily [Luan *et al.*, 2024]. To take advantage of the graph structure and the interaction between nodes, GNNs operate on a three-step process; first, a message-passing step is undertaken, which propagates the information between nodes. This message passing is performed over a set of nodes in the neighborhood of a node, also called the receptive field of the node [Valsesia *et al.*, 2023]. The next step aggregates all the propagated information through a weighted sum. The last step combines the features of the node and the propagated messages from its receptive field.

b , that is, the set of nodes connected to $i$ through any of the edge types. $W_0^{(l+1)}$ and $W_1^{(l+1)}$ are the self-loop and the

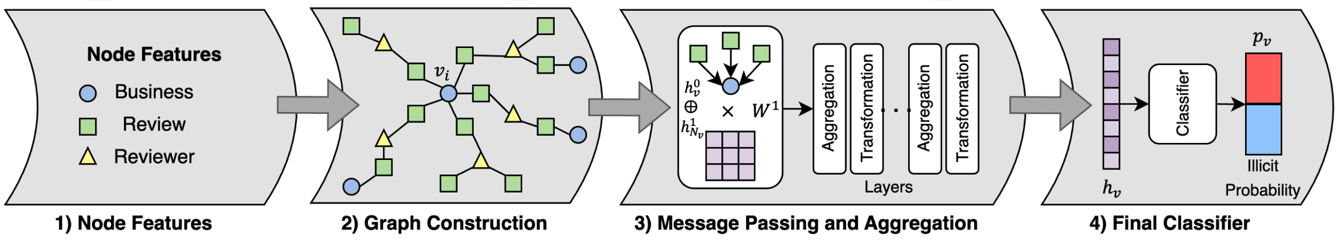


Figure 1: GNN architecture

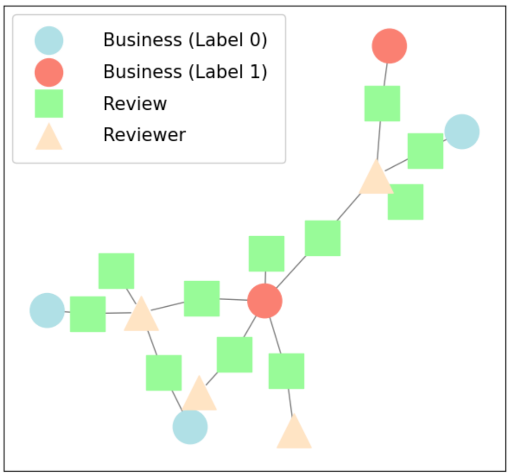


Figure 2: Heterogeneous subgraph for one business node

neighborhood weight aggregation matrices for the $(l+1)^{th}$ layer, and $\sigma$ is the non-linear activation function.

### 3.3 Heterogeneous RGCNs

Zhou et al. [Zhou *et al.*, 2020] present an overview of the evolution of GNNs to GCNs. Further, to capture the information from different types of relations and edges, heterogeneous RGCNs extend the neighborhood aggregation of the GNNs through a separate message-passing process for each type of relation, leading to relation-specific weight matrices. These matrices allow weighted aggregation of all nodes pertaining to each relation, which are then summed across all types. The final step aggregates the source node's features and the propagated messages. The feature representation of node $i$ for the $(l+1)^{th}$ RGCN layer can be formulated as:

$$h_i^{(l+1)} = \sigma \Bigg( W_0^{(l+1)} h_i^{(l)} + \sum_{r \in R} \sum_{j \in N_i(r)} \frac{1}{|N_i(r)|} W_r^{(l+1)} h_j^{(l)} \Bigg), \quad (1)$$

where $R$ denotes the set of all relations in the heterogeneous graph and $N_i(r)$ denotes the neighborhood nodes for the source node $i$ under relation $r$, $W_0^{(l+1)}$ and $W_r^{(l+1)}$ are the learnable weight matrices in the $(l+1)^{th}$ layer, and $\sigma$ is the non-linear activation function.

### 3.4 Classifier

The final business node embeddings generated from the RGCN model are passed through a logistic regression classifier, which applies a linear transformation followed by a sigmoid activation function to compute classification probabilities. The classification threshold is set to 0.5, as it serves as a neutral decision boundary where the model is equally uncertain between the two classes. We denote the model with this classifier as RGCN_LR, and the classifier formulation is given below:

$$z_i = W_{\text{out}} \cdot e_i + b, \quad (2)$$

$$p_i = \sigma(z_i), \quad (3)$$

$$class_i = \begin{cases} 1, & p_i \geq 0.5, \\ 0, & p_i < 0.5, \end{cases} \quad (4)$$

where $W_{\text{out}}$ is the output weight matrix, $e_i$ is the extracted embeddings after $L^{th}$ layers, $b$ is the bias, $\sigma$ is the sigmoid function, $p_i$ is the output probability, and $class_i$ is the predicted class of the business as *illicit* or *non-illicit*.

## 4 Experiments

### 4.1 Dataset Description

We integrate multi-source data to extract business, review, and reviewer information and generate an information-rich heterogeneous graph of massage businesses in Colorado (CO), Florida (FL), and Texas (TX). These states are chosen for the analysis as Florida and Texas are considered hotspots for IMBs [Janis, 2020] and due to Colorado's local law enforcement agency partnerships with our collaborator Global Emancipation Network (GEN). GEN is a non-profit that uses data analytics and technology to fight Human Trafficking [Global Emancipation Network, 2024], which has provided access to the datasets including:

- **Yelp reviews:** business features: name, address, phone number, service category, and price range; review features: text, author, date, and rating.
- **RubMaps reviews:** business features: name, address, and phone number; review features: text, username, date, amount paid, tip paid, and worker demographics.
- **GIS (Geographic Information System):** locations of truck stops, military bases, highways, police stations, and public schools in the considered state.
- **NLCD (National Land Cover Database):** locations of different land cover types.

| Node Type | CO | FL | TX |
|---|---|---|---|
| Total | 13610 | 23704 | 36562 |
| Business | 425 | 785 | 1230 |
| Review | 7662 | 13584 | 21824 |
| Reviewer | 5523 | 9335 | 13508 |

Table 1: RGCN node count across undersampled datasets

| Edge Type | CO | FL | TX |
|---|---|---|---|
| Total | 15324 | 27168 | 43648 |
| Business - Review | 7662 | 13584 | 21824 |
| Review - Reviewer | 7662 | 13584 | 21824 |

Table 2: RGCN edge count across undersampled datasets

- **U.S. census at the census tract level:** demographics: % non-white, % foreign-born, and % ages 20 to 29; socioeconomic status: median household income, % over 25 with bachelor's degree, and % over 25 with master's degree; housing & household composition: % housing vacant, % housing rented, % non-family households, % households with children, and average household size; employment & industry features: % employed in manufacturing, and % employed in education, health care, and social assistance;
- **Business license records:** business features: name, address, phone number, license number, license status, and administrative orders from regulatory institutions.

## 4.2 Data Pre-processing

This section discusses the data pre-processing pipeline for creating the heterogeneous graph. Tables 1 & 2 show the node and edge statistics of the constructed graphs for the datasets of each state.

### Business Features

**Geocoding and GIS Analysis.** Businesses from the Yelp dataset are geocoded to get distances to truck stops, highways, military bases, police stations, and schools. These places are potential factors influencing the location of IMBs, based on crime opportunity theory [de Vries, 2023] and previous stakeholder interviews [Tobey *et al.*, 2022].

**Business Labeling.** The businesses are labeled according to the criteria described in Tobey et al. [Tobey *et al.*, 2022], where we use the features from the RubMaps dataset (review count, last review date, and specific keywords) and the Business license records dataset (status = revoked, surrendered, suspended). The businesses corresponding to the label = 0 are *non-illicit*, while those corresponding to the label = 1 are *illicit*.

**Categorical Feature Creation.** To improve the interpretability of the results, the continuous features are converted into binary features using low, medium, and high quantiles, and categorical features are one-hot encoded.

**Feature Selection.** Univariate logistic regression is performed on each feature to select the statistically significant features. *Table 3: Selected Data Features* in Tobey et al. [Tobey *et al.*, 2022] shows a complete list of business features.

### Review Features

**Review Text Processing and Embedding Creation.** We process each review through standard NLP techniques such as stopword removal, tokenization, and lemmatization. Embeddings are fixed-length numerical vectors representing a given text while capturing semantic meaning. We use a Doc2Vec model pre-trained on massage business reviews with 600-dimensional vector representations [Li *et al.*, 2023]. This high-dimensional vector is reduced to a lower dimension using Principal Component Analysis (PCA).

**Sentiment Analysis.** Sentiment Analysis quantifies a review or an opinion into three categories: positive, negative, and neutral. We use a RoBERTa model, an optimized version of BERT (Bidirectional Encoder Representations from Transformers), which was pre-trained on tweets and fine-tuned for sentiment analysis [Barbieri *et al.*, 2020].

### Reviewer Features

**Gender Identification.** The gender feature is created using the reviewer's username and the *gender guesser* Python package (version 0.4.0), followed by one-hot encoding. This is driven by the observation that IMBs predominantly serve male customers [Crotty and Bouché, 2018].

## 4.3 Baseline Classifiers

This section establishes the baseline classifiers, which include logistic regression, logistic regression with weight balancing, and random forest, which are used to benchmark the performance of the proposed RGCN approach. Logistic regression is chosen for its simplicity as a linear model, its effectiveness for binary classification, and a version that addresses class imbalance. Random Forest is chosen because it can effectively learn non-linear relations. As the aim of the study is to evaluate the effectiveness of the network structure in business classification by using RGCN methodology with business, review, and reviewer links, the baseline models use only the business features.

## 4.4 Experimental Setup

**Class Imbalance.** We tackle the class imbalance in the datasets (Table 3) by undersampling. Specifically, our undersampling strategy ensures that the number of non-illicit businesses sampled is four times that of illicit businesses. Additionally, we prioritize selecting non-illicit businesses with the highest number of reviews for under-sampling to ensure a more connected network.

**Loss Function.** We utilize a Negative Log Likelihood loss function, which is defined as follows:

$$L_{\text{NLL}}(x, y) = -\log p(y|x) \quad (5)$$

where $x$ denotes the vector representation of each business obtained from the RGCN model, $y$ is the label, and $p(y|x)$ is the predicted probability for the true label $y$.

| Business Type | CO | FL | TX |
|---|---|---|---|
| Total | 1926 | 4774 | 4699 |
| Label = 1 (illicit) | 85 | 157 | 246 |
| Label = 0 (non-illicit) | 1841 | 4617 | 4453 |
| Imbalance ratio | 0.046 | 0.034 | 0.055 |

Table 3: Business count

**Performance Metrics.** In order to address the bias due to class imbalance [Luque *et al.*, 2019], we use the area under the ROC curve (AUC) [Zou and Cheng, 2024] to evaluate our results along with Recall and F1-Score.

**Hyperparameters.** For the RGCN_LR model, the search space for the hyperparameters is: epochs in $\{50, 100, 150\}$, dropout in $\{0.1, 0.3, 0.5\}$, number of layers in $\{3, 4, 5\}$, and number of neighbors sampled in $\{1, 5, 10, \text{all}\}$. In order to maintain the brevity and conclusiveness of our results, we do not report results across all of these configurations. However, we showcase hyperparameter sensitivity across two important parameters for the network: the number of layers/hops and the number of neighbors sampled at each layer/hop in Section 5.2. The final hyperparameters chosen for the model are shown in Table 4.

**Implementation.** The training and testing sets are created as stratified partitions of 80% and 20% of the undersampled data set, respectively (see Table 5). We perform 10-fold stratified cross-validation using the training data across two undersampled datasets: Colorado (the dataset with the lowest labeled illicit businesses) and Texas (with the highest labeled illicit businesses). From the top ten models with respect to AUC values across both datasets, we identify the best-performing common hyperparameter configuration and report its results on the test sets. We used test datasets from different states (CO, FL, and TX) in order to validate the generalizability of our approach. We demonstrate our results using three states due to the substantial effort and time required for manual labeling. The analysis is implemented with Python 3.11.11, Pytorch 2.4.0, and DGL 2.4.0, and a seed is set at 42 for both the DGL [Wang *et al.*, 2019] and the Pytorch packages for reproducibility. The model is trained using Google Colab on a virtualized environment with an Intel Xeon CPU @ 2.20GHz (two cores, four threads).[1]

## 5 Numerical Results

### 5.1 Classification Performance

Table 6 showcases the performance of the RGCN_LR model and the baseline models across the three test datasets. The RGCN_LR gives a notable improvement of 0.18 in Recall and 0.10 in F1-Score compared to logistic regression for the CO dataset, the state with the least training data, thereby highlighting its ability to leverage network-informed learning in the case of scarce data, which is typically the case for real-world illicit business detection. It also maintains competitive AUC values across all other states.

[1]The synthetic dataset and the code are available at: https://github.com/Vasuki-Garg/rgcn-imb-detection

| Hyperparameter | Value |
|---|---|
| #Layers | 5 |
| Batch Size | 128 |
| Hidden Dimensions | 64 |
| Dropout | 0.5 |
| Epochs | 100 |
| Early Stopping Criteria | 20 |
| #Neighbors | 5 |
| Batch Normalization | TRUE |
| Self Loop | TRUE |

Table 4: Hyperparameter configuration

| Business | CO | | FL | | TX | |
|---|---|---|---|---|---|---|
| Type | Train | Test | Train | Test | Train | Test |
| Total | 340 | 85 | 630 | 155 | 985 | 245 |
| Label = 1 | 68 | 17 | 126 | 31 | 197 | 49 |
| Label = 0 | 272 | 68 | 504 | 124 | 788 | 196 |
| Ratio | 0.25 | 0.25 | 0.25 | 0.25 | 0.25 | 0.25 |

Table 5: Train and test data split statistics for CO, FL, and TX

### 5.2 Parameter Sensitivity

We analyze the sensitivity of the RGCN_LR model on two hyperparameters using 10-fold stratified cross-validation on the training data.

**Number of Layers/Hops**

The model's performance was evaluated by varying the number of layers in $\{3, 4, 5\}$ while fixing other parameters, as in Table 4. This exploration helped in understanding the impact of the network's depth on classification performance. We can infer from Figure 3 that the model shows minimal sensitivity to the number of layers for FL and TX, while the F1-Score and Recall for CO improve by 0.08 and 0.07 as the number of layers increases from three to five, reinforcing the importance of multi-hop information aggregation in the network in the case of scarce data. Since the model with five layers

| State | Model | Recall | F1-Score | AUC |
|---|---|---|---|---|
| CO | Logistic Reg. | 0.2941 | 0.4348 | 0.8279 |
| | Logistic Reg. (bal) | 0.5294 | 0.5000 | 0.8209 |
| | Random Forest | 0.2353 | 0.3478 | 0.8183 |
| | RGCN_LR | 0.4706 | 0.5333 | 0.7889 |
| FL | Logistic Reg. | 0.8065 | 0.8621 | 0.9119 |
| | Logistic Reg. (bal) | 0.8065 | 0.8333 | 0.9112 |
| | Random Forest | 0.8387 | 0.8814 | 0.9192 |
| | RGCN_LR | 0.8065 | 0.8475 | 0.9099 |
| TX | Logistic Reg. | 0.7755 | 0.8352 | 0.9860 |
| | Logistic Reg. (bal) | 0.8980 | 0.8544 | 0.9853 |
| | Random Forest | 0.7143 | 0.8235 | 0.9743 |
| | RGCN_LR | 0.7755 | 0.8261 | 0.9826 |

Table 6: Model performance across states

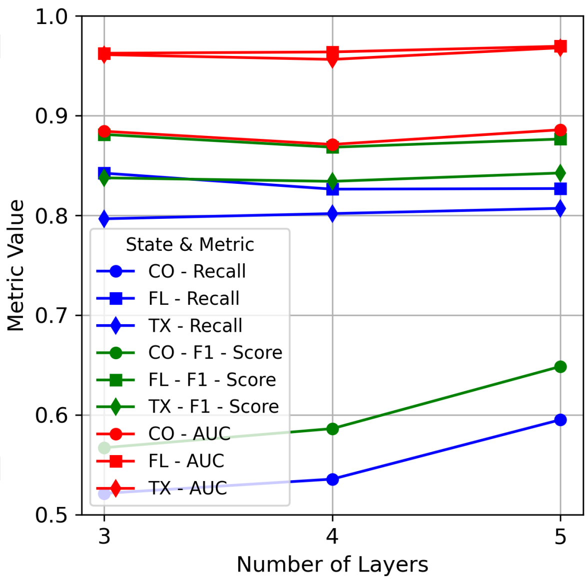


Figure 3: Model performance with different numbers of layers

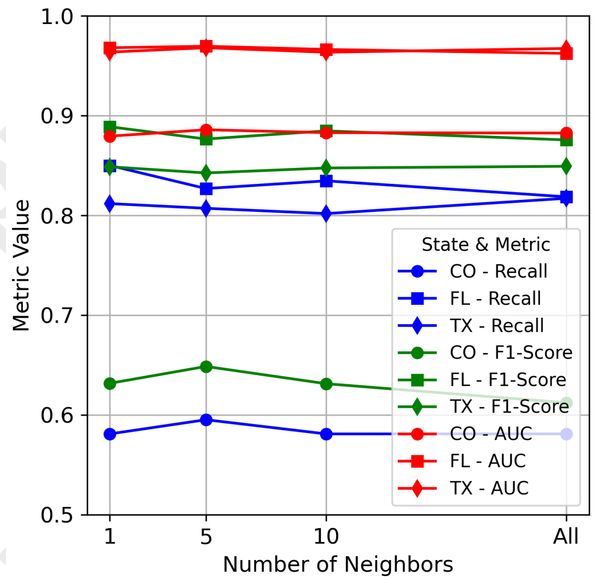


Figure 4: Model performance with different numbers of neighbors

performed well across both CO and TX (the hyperparameter tuning datasets), this value was chosen for the final configuration.

**Neighborhood Sampling**
Further tuning was performed with respect to the number of neighbors sampled. This analysis aimed to determine the optimal number of neighbors to sample during the graph convolution process. Figure 4 shows that the RGCN_LR model exhibits marginal sensitivity, and a single neighbor sampled at each layer performs equally well. However, as the model with five neighbors compared to a single neighbor sampled performed moderately better with 0.02 and 0.01 improvement in F1-Score and Recall for CO, this value was chosen for the final configuration.

### 5.3 Ablation Studies

We validate the importance of key components of our RGCN_LR model with the following two studies using 10-fold stratified cross-validation on the training data:

**Review and Reviewer Node Features**
Since the focus of the problem is on illicit business detection, in this study, we seek to understand the contribution of the review and reviewer node features. We show the model's performance in two cases, with review and reviewer node features (this model is denoted as RGCN_LR) and without these features (denoted as RGCN_LR_wo). We can infer from Table 7 that RGCN_LR shows an improvement in F1-Score (0.05 & 0.02) for CO and TX compared to RGCN_LR_wo. However, by analyzing the AUC trends, we can also deduce that the addition of features shows marginal improvement toward the discriminative ability of the model and that the model primarily focuses on business features to make predictions.

| State | Model | Recall | F1-Score | AUC |
|---|---|---|---|---|
| CO | RGCN_LR_wo | 0.5524 | 0.6035 | 0.8915 |
| | RGCN_LR | 0.5952 | 0.6486 | 0.8858 |
| | RGCN_MLP | 0.7571 | 0.6083 | 0.8862 |
| FL | RGCN_LR_wo | 0.8340 | 0.8806 | 0.9620 |
| | RGCN_LR | 0.8269 | 0.8765 | 0.9695 |
| | RGCN_MLP | 0.8583 | 0.8718 | 0.9672 |
| TX | RGCN_LR_wo | 0.7913 | 0.8267 | 0.9547 |
| | RGCN_LR | 0.8071 | 0.8425 | 0.9680 |
| | RGCN_MLP | 0.8374 | 0.7961 | 0.9578 |

Table 7: Model performance w/wo review & reviewer features

**Explainability.** To make a visual comparison between RGCN_LR and RGCN_LR_wo models' abilities to discriminate between illicit and non-illicit businesses, we present t-SNE plots [van der Maaten and Hinton, 2008] for TX. Specifically, we map the embeddings of both models. The red and the blue nodes in Figures 5 & 6 correspond to non-illicit and illicit businesses. Comparing these, we see that adding review and reviewer features leads to better model separation and discriminative ability, which can also be concluded from the AUC values for TX in Table 7.

**Final Classifier**
In order to assess the discriminative ability of the final classifier, which is the logistic regression in RGCN_LR, we compare it with an RGCN model with Multi-Layer Perceptron (denoted as RGCN_MLP). Table 7 also shows that RGCN_LR gives higher F1-Scores (0.04 and 0.05) and comparable AUC across CO and TX. However, RGCN_MLP shows consistently higher Recall (0.16, 0.03, and 0.03) across all states

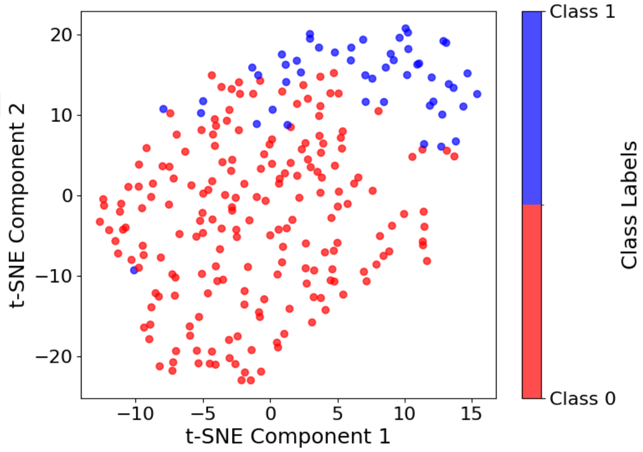


Figure 5: t-SNE plot for RGCN_LR

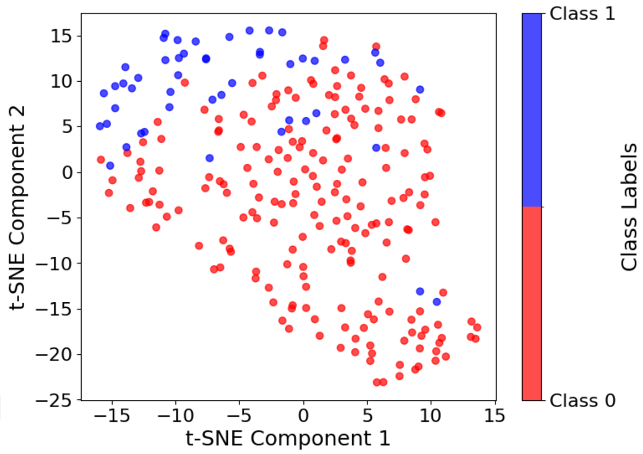


Figure 6: t-SNE plot for RGCN_LR_wo

(CO, FL, and TX), which is preferable in illicit business detection, given the high costs associated with false negatives.

## 6 Real-world Implementation

The model's performance and its ability to generate meaningful results across the three datasets, as summarized in Table 7, demonstrate that our approach can be applied to any business-review dataset containing features related to businesses, reviews, and their reviewers. Such a dataset should also include the necessary identifier information to establish associations between businesses and reviews, as well as between reviews and reviewers, to enable the generation of informative links for effective classification. Since the classification performance in our approach relies on the integration of neighborhood information, it is sensitive to the number of hops/layers and the number of neighbors in each layer of the RGCN. The sensitivity of these parameters to performance can be inferred from Section 5.2

To evaluate performance, we used Recall, F1-score, and AUC. However, the choice of an appropriate performance metric depends on the use case. For instance, some investigative agencies may prioritize Precision to avoid the consequence of labeling most businesses as illicit, leading to operational wastage of resources, while others may emphasize Recall to ensure that as many illicit businesses as possible are identified. While our approach alone will not directly lead to the disruption of illicit businesses or arrests, it is intended as a decision-support framework to help investigative agencies prioritize their limited resources more effectively. The insights from this work can inform advocacy for stricter regulations, improved labor protections, and greater transparency within the massage industry, which is vulnerable to illicit activity.

## 7 Conclusions

In this work, we perform a feasibility study of identifying illicit massage businesses on a review platform using a relational graph convolutional network approach that creates a heterogeneous network by linking businesses, reviews, and reviewers. The comprehensive experiments showcase comparable performance to other state-of-the-art baseline models, with the largest improvements seen in the smallest dataset. This establishes the importance of combining data from multiple sources to detect illicit massage businesses and, in turn, disrupt human trafficking activities.

The sensitivity analysis with respect to the number of neighbors and the ablation study without the review and reviewer nodes reveal marginal informational gains from the network. These results direct us towards the need to create a denser network. As we employ only two edge and relation types, future work will explore further avenues for connecting businesses, such as license and financial records. Our work showcases how business-review datasets can be analyzed from a graph machine learning perspective, primarily focusing on RGCNs. This opens up opportunities for the approach to be extended to other GNN architectures, such as heterogeneous graph transformers, however, these are more data-intensive and would require access to larger datasets for effective training. The proposed network methodology can also be extended to incorporate link prediction across businesses, which is crucial for law enforcement agencies in building human trafficking cases.

## 8 Collaborations

In this project, we collaborated with Global Emancipation Network (GEN), a nonprofit committed to countering human trafficking. Furthermore, CINA facilitated interactions with Homeland Security Investigations (HSI) and other Department of Homeland Security (DHS) stakeholders. These collaborations provided subject matter expertise and context-specific insights; such as the fact that illicit massage businesses are interconnected and do not operate individually. Our collaborators provided access to data and made the labor-intensive labeling process possible. Their domain knowledge also informed our feature engineering process and helped identify relevant parameters and variables that guided the construction of our model.

## Acknowledgements

This work was supported by the Criminal Investigations Network Analysis Center (CINA).

## References

[Aalbers and Sabat, 2012] Manuel B. Aalbers and Magdalena Sabat. Re-making a Landscape of Prostitution: the Amsterdam Red Light District. *City*, 16(1-2):112–128, 2012.

[Barbieri *et al.*, 2020] Francesco Barbieri, José Camacho-Collados, Leonardo Neves, and Luis E. Anke. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *CoRR*, abs/2010.12421, 2020.

[Bhatti *et al.*, 2023] Uzair A. Bhatti, Hao Tang, Guilu Wu, Shah Marjan, and Aamir Hussain. Deep Learning with Graph Convolutional Networks: An Overview and Latest Applications in Computational Intelligence. *International Journal of Intelligent Systems*, 2023(1):8342104, 2023.

[Bouche and Crotty, 2017] Vanessa Bouche and Sean M. Crotty. Estimating demand for illicit massage businesses in Houston, Texas. *Journal of Human Trafficking*, 4(4):279–297, 2017.

[Chin *et al.*, 2019] John J. Chin, Lois M. Takahashi, and Douglas J. Wiebe. Where and Why Do Illicit Businesses Cluster? Comparing Sexually Oriented Massage Parlors in Los Angeles County and New York City. *Journal of Planning Education and Research*, 43(1):106–121, 2019.

[Crotty and Bouché, 2018] Sean M. Crotty and Vanessa Bouché. The Red-Light Network: Exploring the Locational Strategies of Illicit Massage Businesses in Houston, Texas. *Papers in Applied Geography*, 4(2):205–227, 2018.

[de Vries and Radford, 2021] Ieke de Vries and Jason Radford. Identifying online risk markers of hard-to-observe crimes through semi-inductive triangulation: The case of human trafficking in the United States. *The British Journal of Criminology*, 62(3):639–658, 2021.

[de Vries, 2023] Ieke de Vries. Examining the geography of illicit massage businesses hosting commercial sex and sex trafficking in the united states: The role of census tract and city-level factors. *Crime & Delinquency*, 69(11):2218–2242, 2023.

[Department of State, 2025] U.S. Department of State. Understanding Human Trafficking. https://www.state.gov/what-is-trafficking-in-persons/, 2025. [Accessed 05-02-2025].

[Diaz and Panangadan, 2020] Maria Diaz and Anand Panangadan. Natural Language-based Integration of Online Review Datasets for Identification of Sex Trafficking Businesses. In *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 259–264, 2020.

[Fan *et al.*, 2020] Yujie Fan, Yanfang Ye, Qian Peng, Jianfei Zhang, Yiming Zhang, Xusheng Xiao, Chuan Shi, Qi Xiong, Fudong Shao, and Liang Zhao. Metagraph Aggregated Heterogeneous Graph Neural Network for Illicit Traded Product Identification in Underground Market. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 132–141, 2020.

[Global Emancipation Network, 2024] Global Emancipation Network. What We Do. https://www.globalemancipation.ngo/whatwedo/, 2024. [Accessed 06-02-2025].

[Harish *et al.*, 2024] Sunisha Harish, Chirag Lakhanpal, and Amir H. Jafari. Leveraging graph-based learning for credit card fraud detection: a comparative study of classical, deep learning and graph-based approaches. *Neural Computing and Applications*, 36:21873–21883, 2024.

[Janis, 2020] Elizabeth Ranade Janis. Unmasking Trafficking in Illicit Massage Businesses Across the United States - Human Trafficking Institute — traffickinginstitute.org. https://traffickinginstitute.org/illicit-massage-businesses/, 2020. [Accessed 11-06-2024].

[Lasker, 2001] Stephanie Lasker. Sex and the city: zoning pornography peddlers and live nude shows. *UCLA Law Review*, 49:1139–1185, 2001.

[Li *et al.*, 2023] Ruoting Li, Margaret Tobey, Maria E. Mayorga, Sherrie Caltagirone, and Osman Y. Özaltın. Detecting Human Trafficking: Automated Classification of Online Customer Reviews of Massage Businesses. *Manufacturing & Service Operations Management*, 25(3):1051–1065, 2023.

[Li *et al.*, 2024] Cheng-Te Li, Yu-Che Tsai, Chih-Yao Chen, and Jay C. Liao. Graph Neural Networks for Tabular Data Learning: A Survey with Taxonomy and Directions. *arXiv*, 2401.02143, 2024.

[Liu *et al.*, 2021] GuanJun Liu, Jing Tang, Yue Tian, and Jiacun Wang. Graph Neural Network for Credit Card Fraud Detection. In *2021 International Conference on Cyber-Physical Social Intelligence (ICCSI)*, pages 1–6, 2021.

[Luan *et al.*, 2024] Sitao Luan, Chenqing Hua, Minkai Xu, Qincheng Lu, Jiaqi Zhu, Xiao-Wen Chang, Jie Fu, Jure Leskovec, and Doina Precup. When Do Graph Neural Networks Help with Node Classification? Investigating the Impact of Homophily Principle on Node Distinguishability. *arXiv*, 2304.14274, 2024.

[Luque *et al.*, 2019] Amalia Luque, Alejandro Carrasco, Alejandro Martín, and Ana de las Heras. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91:216–231, 2019.

[Mensikova and Mattmann, 2018] Anastasija Mensikova and Chris A. Mattmann. Ensemble Sentiment Analysis to Identify Human Trafficking in Web Data. In *Workshop on Graph Techniques for Adversarial Activity Analytics (GTA 2018), Marina Del Rey, CA, USA*, pages 5–9, 2018.

[Mletzko *et al.*, 2018] Deborah Mletzko, Lucia Summers, and Ashley N. Arnio. Spatial patterns of urban sex trafficking. *Journal of Criminal Justice*, 58:87–96, 2018.

[Motie and Raahemi, 2024] Soroor Motie and Bijan Raahemi. Financial fraud detection using graph neural networks: A systematic review. *Expert Systems with Applications*, 240:122156, 2024.

[Murphy and Venkatesh, 2006] Alexandra K. Murphy and Sudhir A. Venkatesh. Vice Careers: The Changing Contours of Sex Work in New York City. *Qualitative Sociology*, 29(2):129–154, 2006.

[Sarthak *et al.*, 2021] Sarthak, Shikhar Shukla, and Karm V. Arya. Detecting Hostile Posts using Relational Graph Convolutional Network. *CoRR*, abs/2101.03485, 2021.

[Scarselli *et al.*, 2009] Franco Scarselli, Marco Gori, Ah C. Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[The Network, 2024] The Network. What is the Illicit Massage Industry? — The Network — thenetworkteam.org. https://www.thenetworkteam.org/research/what-is-the-illicit-massage-industry, 2024. [Accessed 08-02-2025].

[Tobey *et al.*, 2022] Margaret Tobey, Ruoting Li, Osman Y. Özaltın, Maria E. Mayorga, and Sherrie Caltagirone. Interpretable models for the automated detection of human trafficking in illicit massage businesses. *IISE Transactions*, 56(3):311–324, 2022.

[Valsesia *et al.*, 2023] Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Ran-gnns: Breaking the Capacity Limits of Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):4610–4619, 2023.

[van der Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[Wang *et al.*, 2019] Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*, 2019.

[White *et al.*, 2021] Anna White, Seth Guikema, and Bridgette Carr. Why are You Here? Modeling Illicit Massage Business Location Characteristics with Machine Learning. *Journal of Human Trafficking*, 10(1):20–40, 2021.

[Yao *et al.*, 2024] Jianrong Yao, Ling Jiang, Chenglong Shi, and Surong Yan. Fake review detection with label-consistent and hierarchical-relation-aware graph contrastive learning. *Knowledge-Based Systems*, 302:112385, 2024.

[Zhou *et al.*, 2020] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.

[Zou and Cheng, 2024] Yao Zou and Dawei Cheng. Effective High-order Graph Representation Learning for Credit Card Fraud Detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 7581–7589, 2024.