Investigating Links between Illicit Massage Businesses through Natural Language Processing and Graph Machine Learning

Vasuki Garg[1], Osman Y. Özaltın[1], Maria E. Mayorga[1], Sherrie Bosisto[2]

[1]Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695
[2]Global Emancipation Network, Clermont, FL 34715
{vgarg5, oyozalti, memayorg}@ncsu.edu, sherrie@globalemancipation.ngo

Abstract

Human trafficking exploits vulnerable individuals through forced sex or labor. Illicit massage businesses offer a clandestine front to illicit activities by disguising themselves as legitimate businesses. This makes it challenging for law enforcement agencies and anti-trafficking organizations to detect these enterprises and their associated entities, disrupt the network, and save victims. We adopt a multi-stream data integration approach primarily focusing on consumer-generated business reviews on Yelp.com, enriched with features from contextual data sources, such as the U.S. Census and business license records. We propose a novel decision support framework that extends the traditional link prediction methods by defining a higher-order neighborhood to detect links between pairs of massage businesses and the exposure of businesses to illicit activities related to human trafficking. We achieve this by introducing a bespoke subgraph extraction strategy in GNNs where the node features are derived using NLP techniques. Comprehensive experimental results demonstrate the competitive performance of our approach over the baseline methods.

1 Introduction

The U.S. Department of State (2025) defines human trafficking as the use of force, fraud, or coercion to exploit an individual for commercial sex or labor. Human trafficking has infiltrated the massage industry due to weak regulations, governance, and laws (Organisation for Economic Co-operation and Development, 2016). Illicit Massage Businesses (IMBs) sell commercial sex while disguising themselves as legitimate businesses. They exploit the victims for both sex and labor while also harming the legitimate massage industry. The true extent of IMBs is unknown; however, The Network (2024) estimates that more than 15,000 IMBs operated in the U.S. in 2024. These entities often operate within a network that carries out various illicit activities, such as drug and arms trafficking, and money laundering (Miklaucic and Brewer, 2013). Given the scale of human suffering caused by their operations, several investigative agencies, both in the public and private sectors, are seeking to identify IMBs and their connections. However, identifying key indicators from the vast volume of data associated with these businesses poses a significant challenge in unearthing their discreet and intertwined operations.

Existing studies employed different approaches to identify IMBs and counter human trafficking. Aalbers and Sabat (2012); Lasker (2002); Murphy and Venkatesh (2006) applied geo-spatial analysis methods to study the spread of sexually-oriented businesses, driven by the proliferation of transportation and the internet. Mletzko et al. (2018); de Vries and Radford (2021) examined the distribution of sex-trafficking offenses to determine IMB hotspots, such as highways and motels. Lugo-Graulich (2024) employed linguistic analysis to identify the distinguishing characteristics between sex-trafficking and consensual work in escort advertisements. Davy (2022) studied the role of relationships between human trafficking victims and traffickers in victim recruitment, control, and exploitation, through qualitative analysis of structured input from people with lived experiences.

With an increasing online presence, other researchers (Crotty and Bouché, 2018; Chin et al., 2023; White et al., 2024) spatially clustered IMBs and predicted their demand by analyzing customer reviews on massage boards and foot traffic data from video surveillance in combination with demographic factors. Other studies developed classification models to detect IMBs based on features extracted from customer reviews on publicly available business review websites, such as Yelp.com, either by applying sentiment analysis (Mensikova and Mattmann, 2018), or by developing a lexicon

Findings of the Association for Computational Linguistics: ACL 2026, pages 34080–34096

vocabulary (Li et al., 2023). The closest study to our work (Garg et al., 2025) employed a Graph Convolutional Network (GCN) model to learn features based on the relationships between businesses, reviews, and reviewers, represented as nodes on a network. In this study, we use similar data collection and feature extraction methods. Unlike Garg et al. (2025), which classified individual massage businesses into illicit and non-illicit categories, this work focuses on the links between massage businesses.

2 Related Work and Our Contributions

2.1 Link Prediction

Link prediction aims to infer connections between entities within a domain, which may be of the same type (e.g., businesses) or different types (e.g., businesses and review writers). Early link prediction methods (Liben-Nowell and Kleinberg, 2007) leveraged similarity scores based on network measures, such as the number of common neighbors or the node degrees. The integration of embedding-based models, such as DeepWalk (Perozzi et al., 2014), Node2Vec (Grover and Leskovec, 2016), and Graph Machine Learning (GML), has provided analytical tools for processing vast amounts of data to infer links within a network. For example, GraphSAGE (Hamilton et al., 2017) aggregates neighborhood information into node embeddings and uses these embeddings for link prediction. Formulating link prediction as a classification problem, the SEAL method (Zhang and Chen, 2018) extracts the subgraph around the inferred link and uses a Graph Neural Network (GNN) for prediction. Traditional GNNs assume homophily. However, the business review network studied in our work is heterophilic, where connected nodes have different labels. Specifically, businesses are connected to reviews, which, in turn, are connected to the reviewers. Neighborhood aggregation can negatively impact the performance of GNNs in such graphs. Thus, we propose a new definition (Section 4.1) of neighborhood based on higher-order links.

2.2 Link Prediction to Combat Human Trafficking

Prior studies have mainly focused on analyzing online escort advertisements to uncover trafficking networks. Cockbain et al. (2011) applied social network analysis to identify key hub nodes whose interdiction could significantly disrupt the network. Szekely et al. (2015) developed a knowledge graph that includes nodes representing ads, people, locations, and contact data. Links between nodes are predicted using text and image similarity measures, as well as entity resolution methods. Chambers et al. (2019) proposed a neural network–based approach for extracting phone numbers to link escort advertisements. Li et al. (2022) combined classic rule-based and dictionary extractors with a contextualized language model to recognize entities with ambiguous names in escort ads, establishing links between these entities and the ads. Vajiac et al. (2023) proposed a micro clustering approach to identify links between escort ads. Saxena et al. (2023) predicted links between human trafficking vendors on the basis of authorship features in language patterns.

No prior work has examined business review data for link prediction between IMBs, nor has any study considered predicting different types of links. We address this gap in the literature by developing a GML–based prediction framework built on business review data. We demonstrate the applicability of the proposed framework by developing models to predict: (i) whether a pair of massage businesses are linked through their reviews and reviewers; (ii) whether the link between a pair of businesses is illicit, that is, it involves an illicit business; (iii) whether a massage business is illicit; and (iv) whether a massage business is illicit or linked (exposed) to any illicit business.

2.3 Main Contributions

Driven by investigative goals and domain expertise, the main contributions of this study include:

- Combining text embeddings extracted via Natural Language Processing (NLP) with network-based graph features and contextual features of nodes within GML to create a link prediction framework based on a business review dataset.
- Developing a subgraph extraction and labeling approach to infer higher-order links between nodes in heterophilic graphs.
- Extending traditional link prediction methods to consider different link types (i.e., illicit and non-illicit links).
- Expanding the link prediction task between two nodes to exposure detection, which aims to predict the neighborhood type of a node based on its own class and links to any other node of a certain class.

3 Methodology

3.1 Graph Construction

We represent the business review dataset as an undirected heterogeneous graph with three node types: `business`, `review`, and `reviewer`. Two types of edges capture the relationship between the nodes: `business-review` and `review-reviewer` edges. We refer to this graph as the *business review graph*. We focus on predicting the `links` between the businesses. A pair of businesses is linked if there is a path connecting them. Given the above definition of node and edge types, the minimum length of a path linking businesses `A` and `B` is 4 if the same reviewer `X` provided reviews for both businesses, i.e. `A-reviewA-X-reviewB-B`. Thus, two businesses can be linked by a path of $4, 8, 12 \ldots$ hops. To focus on more direct links and maximize the performance of the proposed approach, we consider predicting 4-hop and 8-hop links between business nodes in the business review graph (Section 4.1).

3.2 Subgraph Extraction

In GML, subgraphs are used to define the receptive field or neighborhood structure that influences inference (Valsesia et al., 2023). A small receptive field may provide insufficient information, whereas larger receptive fields can lead to over-smoothing (Hamilton, 2020). In this study, we extract subgraphs from the business review graph. For prediction tasks involving pairs of businesses (link-level), we extract the m-hop ego subgraph around each business node. For tasks involving a single business (node-level), we extract the m-hop ego subgraph centered on the business node. We choose the value of the parameter m on the basis of the link definition in our experiments (Section 4.1) as well as to balance information gain and over-smoothing.

To prevent label leakage in link prediction tasks involving business pairs, we remove all direct 4-hop links connecting the two target businesses when constructing training subgraphs. In addition, we remove 8-hop links by randomly selecting one endpoint of the link and deleting the corresponding 4-hop path connected to that endpoint. This procedure ensures that the extracted training subgraphs do not explicitly include the target links while preserving the surrounding structural context.

Each business, review, and reviewer node in the subgraph has tailored contextual features described in Section 4.1. To capture the positional encoding of each node with respect to the target business node(s), we append z_1 and z_2 distance scores to the node features. These scores are calculated using Equations 1 and 2.

$$z_1 = 1 + min(d_A, d_B), \quad d = d_A + d_B, \quad (1)$$

$$z_2 = \begin{cases} \lfloor d/2 \rfloor \left(\lfloor d/2 \rfloor + d\%2 - 1 \right) & \text{if } d < \infty, \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

where d_A and d_B are the shortest path distances from the target business nodes A and B in link prediction. For prediction tasks involving a single business node A, we only use $z_1 = 1 + d_A$.

3.3 Subgraph Encoding

We train a Graph Neural Network (GNN) to encode the structural and node features of a subgraph into an embedding vector. GNNs first propagate the information between nodes in a message-passing step. This message passing is performed over the neighborhood of each node. The next step aggregates the propagated information through a weighted sum. The final step combines the node's features with the propagated messages from its neighborhood.

We apply message passing and aggregation over multiple neighborhood layers around each node (Kipf and Welling, 2016a). The first neighborhood layer ($l = 1$) of a node consists of its immediate neighbors reachable in one hop. The second layer ($l = 2$) includes nodes reachable in two hops, and higher-order layers are defined analogously. Equation (3) demonstrates the calculation of the embedding $h_i^{(l+1)}$ for node i in layer $l + 1$.

$$h_i^{(l+1)} = \sigma \left(W_0^{(l+1)} h_i^{(l)} + \sum_{j \in N_i^{(l+1)}} W_1^{(l+1)} h_j^{(l)} \right), \quad (3)$$

where $N_i^{(l+1)}$ is the set of nodes in neighborhood layer $l + 1$ of node i, $W_0^{(l+1)}$ and $W_1^{(l+1)}$ are the self-loop and the neighborhood weight aggregation matrices, and σ is the nonlinear activation function. We set the dimension of the node embedding vector h_i^l and the number of layers $l = 1, \ldots, L$ using hyperparameter tuning. We maintain the same node embedding dimension in each layer, except for the last layer, where the dimension of h_i^L is set to 1. The final embedding vector (h_i) of node i is obtained by concatenating its embedding vectors in each layer, i.e., $h_i = [h_i^1, h_i^2, \ldots, h_i^L]$. We sort h_i's based on their last entry and populate the k

(node pool percentile) portion of them in a node embedding matrix. The value of k is set using hyperparameter tuning. We then process the node embedding matrix using a Convolutional Neural Network (CNN) to obtain an embedding vector for the entire subgraph. We pass the final subgraph embedding through a fully connected multi-layer perceptron for classification.

3.4 Benchmark Methods

We implement two benchmark methods for link prediction between a pair of businesses: Preferential Attachment (PA) and Logistic Regression (LR). For link class prediction or node-level prediction tasks, we use LR only. PA infers a link between a given pair of businesses based on the product of their node degrees. LR applies a linear transformation to business features, followed by a nonlinear activation function to produce class probabilities.

PA is a purely structure-based method that requires no training but does not incorporate node attributes. In contrast, LR leverages features of the target business node(s) but fails to capture structural information from the business review graph. Our proposed approach addresses these limitations by jointly modeling both subgraph structural features and node attributes.

4 Computational Experiments

This section describes the business review dataset, the feature construction process for the three node types (businesses, reviews, and reviewers), and the experimental design and implementation.

4.1 Data

We leverage multi-source data to construct business review graphs for massage businesses in Colorado (CO), Florida (FL), and Texas (TX). We chose these states for analysis because FL and TX are IMB hotspots (Janis, 2020) and our collaborator, Global Emancipation Network (GEN), has partnerships with a local law enforcement agency in CO. GEN is a nonprofit that uses data analytics and technology to fight human trafficking (Global Emancipation Network, 2024), which has provided the following datasets:

Yelp reviews: business name, address, phone number, service category, price range; review text, username, date, rating.

RubMaps reviews: business name, address, phone number; review text, username, date.

Business license records: business name, address, phone number, license number, license status, and administrative orders from regulatory agencies.

We have augmented these datasets with contextual features from publicly available data sources:

U.S. Census at the census tract level: demographic and socioeconomic variables, housing & household composition, employment & industry features.

GIS (Geographic Information System): locations of highways, truck stops, military bases, police stations, and public schools.

NLCD (National Land Cover Database): land cover types, e.g., developed, low/high intensity.

Business Features. Massage businesses from the Yelp dataset are geocoded to get distances to truck stops, highways, military bases, police stations, and schools. These places influence the location of IMBs based on crime opportunity theory (de Vries, 2023) and stakeholder interviews (Tobey et al., 2024). We adopt the business labeling procedure (non-illicit:0, illicit:1) from Tobey et al. (2024), which utilizes the review and business features obtained from RubMaps.ch, as well as the license records. Statistically significant features are selected using univariate logistic regression. *Table 3: Selected Data Features* in Tobey et al. (Tobey et al., 2024) shows a complete list of business features.

Review Features. After applying standard natural language processing (NLP) techniques, such as stop-word removal, lemmatization, and tokenization, we convert review text into 600-dimensional embedding vectors using a pretrained Doc2Vec model (Li et al., 2023). These vectors are mapped to a lower-dimensional space using Principal Component Analysis (PCA). To create informative features, we also employ the lexicon analysis from Li et al. (2023), where they develop a Yelp-specific lexicon for IMBs, comprising a vocabulary of 169 keywords selected based on their high frequency in illicit reviews and input from domain experts. An illicit review explicitly mentions or implies commercial sex or other indicators of human trafficking at a business. While commercial sex alone does not constitute human trafficking unless induced by force, fraud, or coercion (U.S. Department of State, 2025), evidence (Dank et al., 2014) suggests that a non-negligible share of massage business workers engaged in commercial sex are trafficking victims. They weight lexicon terms by strength (1 for potential signs for commercial sex and 2 for strong

indicators) and train a classifier using normalized lexicon scores. We include the score and the classifier's output in review features. We further perform sentiment analysis using a RoBERTa model (Barbieri et al., 2020) to classify reviews as positive, neutral, or negative, and also incorporate the review ratings (1-5) as a feature.

Reviewer Features. Driven by the observation that IMBs predominantly serve male customers (Crotty and Bouché, 2018), we use the *gender guesser* package to create a gender feature based on the reviewer's username.

GNNs show better performance at a lower hop neighborhood by avoiding over-smoothing (Hamilton, 2020) and at higher edge homophily (Luan et al., 2022) (where edge homophily defines the proportion of edges that join the nodes of the same class). We generalize this definition to hop homophily (i.e., the proportion of m-hop links that connect the same class nodes). Table 3 shows that 4 and 8 hops exhibit high hop homophily and also dense population (i.e., high total number of pairs for training, testing, and validation). Therefore, we select 4 and 8 hops for the link definition.

4.2 Design of Experiments

This section defines four experiments, which are motivated by investigative objectives aimed at identifying and disrupting illicit business networks. The first two experiments focus on relationships between pairs of businesses, whereas the last two experiments involve individual businesses. All experiments are formulated as binary classification tasks (with Classes 1 and 0), where Class 1 consistently represents businesses of investigative importance.

Experiment 1: Link Prediction. This analysis enables investigators to predict links between pairs of massage businesses. We define positive links as observed 4-hop and 8-hop connections between business node pairs (Class 1). To preserve class imbalance, we sample a comparable number of business node pairs that do not exhibit any connections within eight hops (Class 0).

Experiment 2: Link Classification. This analysis enables investigators to prioritize the examination of businesses involved in illicit connections. The experiment predicts whether a linked pair of businesses includes any illicit business. We define Class 1 links (illicit links) as observed 4-hop or 8-hop connections between business node pairs in which at least one business is illicit. We sample a comparable number of Class 0 links from business pairs in which both businesses are non-illicit (benign) and are connected by 4-hop or 8-hop paths.

Experiment 3: Business Classification. This experiment predicts whether a business is illicit (Class 1) or non-illicit (Class 0). Although the primary contribution of our work lies in link prediction and classification, we include this experiment to demonstrate the robustness of the proposed approach in learning effective embeddings for node classification.

Experiment 4: Illicit Exposure Detection. This analysis enables investigators to prioritize the examination of businesses that are illicit or involved in illicit connections. The experiment predicts whether a massage business is illicit or has connections to an illicit business through 4 hops (Class 1). Businesses in Class 0 are non-illicit and have no links to illicit businesses. This experiment does not classify the specific links associated with a given business. If a business is predicted to belong to Class 1, the analysis in Experiment 2 can be applied to further identify the specific illicit links.

4.3 Implementation

Data Preparation. We employ a three-step approach to generate data for model training, testing, and validation in the link-based experiments (1 and 2). First, we perform undersampling to address the class imbalance between illicit and non-illicit businesses. Since non-illicit businesses are abundant in the data, we select the ones with the most reviews in our sampling approach and maintain an imbalance ratio of 0.25 (Table 4). Then, we perform stratified splitting across the CO, FL, and TX datasets, dividing them into 80% model development and 20% held-out testing sets. We further stratify the development data into 80% training and 20% validation sets, with the validation set used for early stopping. This inductive splitting across businesses prevents data leakage and aligns with the purpose of our framework, which is to make inferences about unknown businesses and networks. Finally, we sample training, validation, and testing business pairs from their respective business-review graphs (Table 5) using the class definitions in each experiment. In particular, we sample up to 10,000 business pairs for Class 1 and up to 10,000 pairs for Class 0, considering all pairs if the set is smaller. The data preparation for Experiments 3 and 4 in-

volves only the stratified splitting of businesses into training, validation, and testing sets.

Each dataset (CO, FL, and TX) belongs to a different jurisdiction, which is governed by different policies and regulations, so we expect the business networks to differ. To analyze the contribution of these differences and diversity in the datasets, we report the distribution of statistically significant features in the training sets by categorizing them and measuring their prevalence, i.e., the mean proportion of businesses for which the binary business features within the given category are active (value = 1). We can infer from Table 1 that CO exhibits a Yelp-based features dominance, while FL and TX incorporate more Census-based features. We also see the highest prevalence for geographic features (GIS & NLCD) across the three datasets.

To analyze the network created by these datasets, we report their network measures and statistics. Table 6 illustrates that CO leads to the densest network while TX is the sparsest. Even though the average degree (Avg. deg) is similar, higher-order structure among the three varies, as bigger mean z_1 or ($\bar{z_1}$) and mean z_2 or ($\bar{z_2}$) scores indicate larger local neighborhoods for TX.

Hyperparameter Tuning. The key hyperparameters chosen for tuning along with their search space include: number of GNN layers $L \in \{2, 4, 6\}$, hidden dimension of the node embeddings for the GNN layers $|h_i^l| \in \{8, 16, 32\}$, node pool percentile $k \in \{0.3, 0.6, 0.9\}$, as well as the number of input and output channels of the convolutional layer $(c_1, c_2) \in \{(4, 8), (8, 16)\}$.

We use the FL dataset for hyperparameter optimization, as it is the medium-sized set in terms of number of businesses, reviews, and reviewers (Table 7). We perform hyperparameter tuning for each experiment based on the Area Under the Receiver Operating Characteristic Curve (AUC) value and use the finalized hyperparameters (as shown in Table 8) for a single-run testing. For Experiments 1 and 2, we sample 2,500 business pairs in Class 0 and Class 1 from the businesses in the training set. For Experiments 3 and 4, we consider the entire training set, as these experiments are based on businesses rather than business pairs. We further perform 5-fold stratified cross-validation across all experiments and report mean results across all folds with their sample standard deviations.

We report two performance metrics, AUC and Average Precision (Avg Prec), i.e., the area under the precision-recall curve. These threshold-agnostic metrics are particularly suitable for imbalanced datasets. While AUC is compared against the baseline of 0.5, Avg Prec is interpreted with respect to the positive class prevalence (Saito and Rehmsmeier, 2015) as illustrated in Table 9. We implement the framework on Google Colab using Python 3.11, equipped with an NVIDIA L4 GPU and an Intel Xeon 2.20 GHz CPU.[1]

4.4 Numerical Results

Table 2 presents the performance of the GML-based prediction model (GNN) as well as the performance of the LR and PA methods over the held-out testing sets for CO, FL, and TX.

Link-level Experiments. The GML-based methods consistently achieve the highest AUC and Avg Prec across CO and TX in Experiments 1 (Link Prediction) and 2 (Link Classification), showcasing robust performance even with the added difficulty of classification on top of structural link prediction. We see a notable improvement of 0.2692 and 0.2909 in the AUC and Avg Prec for CO (dataset with the lowest labeled businesses) compared to the LR model, highlighting the importance of high-order links and network-informed learning in a low-data regime. The LR model, which is trained on handcrafted business features, also outperforms PA, with the effect more pronounced in larger datasets (FL and TX), demonstrating the importance of features and model training over untrained network measures. Competitive LR results for FL across both experiments suggest that the subgraph structure is homogeneous for both classes.

Node-level Experiments. In Experiment 3 (Node Classification), we observe significant improvements in LR performance metrics, suggesting that direct business features are highly predictive. Even though LR dominates, the learned embeddings from the GNN model show competitive performance and do not degrade. GNN achieves an extremely high Avg Prec score of 0.9920 for TX in Experiment 4 (Illicit Exposure Detection) with substantial improvements over LR for CO and FL, establishing the criticality of neighborhood aggregation through network-informed learning in detecting exposure.

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

Category	CO		FL		TX	
	% Features	Prevalence	% Features	Prevalence	% Features	Prevalence
Yelp	**57.1%**	0.2040	35.7%	0.3041	35.7%	0.3043
Census	25.0%	0.3340	**37.5%**	0.3333	**37.5%**	0.3333
GIS & NLCD	14.3%	**0.3879**	12.5%	**0.4269**	12.5%	**0.4262**
Others	3.6%	0.0037	14.3%	0.1559	14.3%	0.2049

Table 1: Business features distribution and their mean prevalence across the three datasets.

Exp	Method	CO		FL		TX	
		AUC	Avg Prec	AUC	Avg Prec	AUC	Avg Prec
1	GNN	**0.7781**	**0.8076**	**0.6492**	**0.6984**	**0.7324**	**0.7876**
	LR	0.5089	0.5167	0.6468	0.6068	0.6015	0.5722
	PA*	0.4951	0.4888	0.5202	0.5225	0.4873	0.5020
2	GNN	**0.7885**	**0.7718**	0.7725	0.8056	**0.5485**	**0.5304**
	LR	0.7384	0.7666	**0.7733**	**0.8200**	0.5000	0.5000
3	GNN	0.6626	**0.5912**	0.8971	0.8265	0.8580	0.7726
	LR	**0.8097**	0.5093	**0.9242**	**0.8968**	**0.9596**	**0.9138**
4	GNN	**0.5950**	**0.9387**	**0.7414**	**0.9434**	**0.8534**	**0.9920**
	LR	0.5029	0.9014	0.6145	0.9136	0.6530	0.9780

Table 2: Prediction performance across four experiments. * PA is only used in Experiment 1 because it is suitable for link prediction, not for link classification.

Class Imbalance. We test the performance of the GNN, LR, and PA models at various imbalance ratios to ascertain generalizability for link-level experiments. In Figure 1, for CO, while LR is the least sensitive, GNN outperforms the former across all ratios for both Experiments 1 and 2.

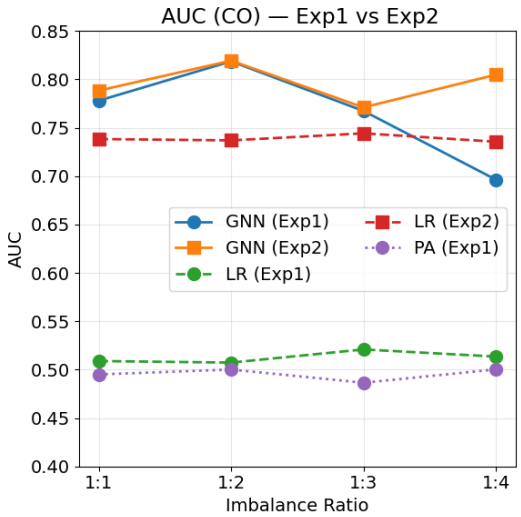

Figure 1: Model performance at different imbalance ratios for link-level experiments. Analogous results for FL and TX are presented in Figure 6 and Figure 7.

4.5 Parameter Sensitivity

Number of GNN Layers (L). We vary the number of GNN layers within $\{2, 4, 6\}$ while fixing other parameters to assess the implication of neighborhood layers. Figure 2 suggests that the model's AUC is highly sensitive in Experiment 2, showing an improvement of 0.2263 when the number of layers is increased from 2 to 6. This reinforces the impact of GNN's message passing and aggregation towards creating informative representations for predicting the link class. In contrast, for the simplistic task of link prediction in Experiment 1, we see a decrease in performance, which is consistent with the phenomenon of over-smoothing. Experiment 3 (Node Classification) shows the lowest sensitivity, while performance improves as L increases for Experiments 2 and 4 (Link Classification and Exposure Detection), due to their reliance on neighborhood context.

Node Pool Percentile (k). We examine the node pool size by varying k within $\{0.3, 0.6, 0.9\}$. A k value of 0.6 is interpreted as the 60^{th} percentile among all training subgraph sizes (number of nodes) to be chosen as the pooling size. In Figure 3, Experiments 1, 2, and 4 show peak AUC at $k = 0.6$, which prevents over-smoothing due to noise from weakly informative nodes. Experi-

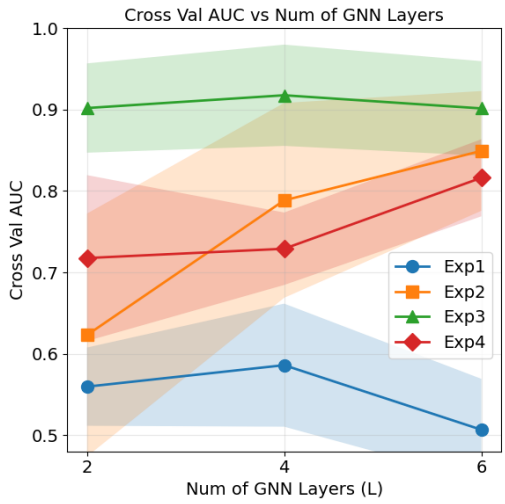

Figure 2: Model cross-validation AUC ± std dev values with different numbers of layers on the FL dataset.

ment 3 shows minimal sensitivity to k, indicating that business classification depends primarily on its own features.

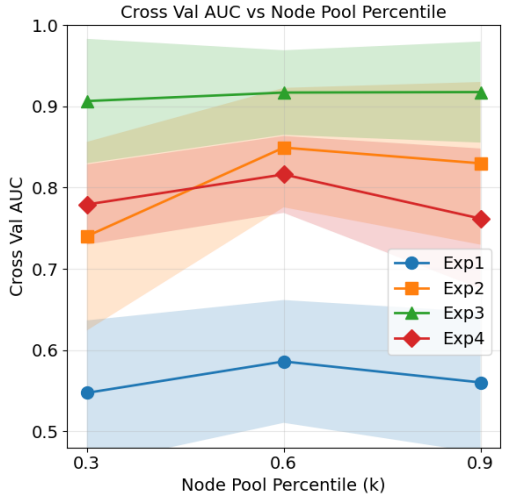

Figure 3: Model cross-validation AUC ± std dev values with different node pool percentile on the FL dataset.

4.6 Ablation Studies

Positional Encoding Scores (z_1, z_2). In this section, we analyze the contribution of the z_1 and z_2 positional encoding scores in the performance of the GNN model. To do so, we define a new model named GNN_woz, i.e., GNN without these scores. Table 10 shows that GNN outperforms GNN_woz at most instances of link-level experiments, establishing the value of network topology. For the node-based experiments, where node features dominate (as evidenced by the improved performance of LR), both models exhibit comparable performance.

Feature Pruning. To analyze the impact of feature quality, we remove the least informative features, ranked by integrated gradient scores (Section 5), 5 of 28 features in CO and 10 of 56 features in FL and TX. Since the primary contribution of this work is link prediction and classification, we emphasize Experiments 1 and 2 and tune hyperparameters (Tables 8 and 11) for each state. Table 12 summarizes the performance of GNN_woz (model with all business features) and GNN_sel (model with selected features). We observe the largest performance gaps in FL, where pruning improves performance in Experiment 2, owing to a reduction in feature noise towards link classification, while the performance degrades in Experiment 1, indicating that all features contribute to link prediction.

Dataset Splitting. We vary our dataset splitting strategy to a temporal approach. We sort businesses in ascending order by their 75$^{\text{th}}$ percentile review date, select the first 80% as the development data, and use the remaining for testing, while defining the threshold as the cutoff business's 75$^{\text{th}}$ percentile date. We then ensure the training data only contains reviews prior to the threshold, while testing businesses leverage the full time frame to emulate real-world deployment. Following this, we employ our regular pipeline of undersampling businesses and sampling link types. We compare our framework with two Variational Graph Auto-Encoders (VGAE) (Kipf and Welling, 2016b), VGAE_1 with a GCN (Kipf and Welling, 2016a) and VGAE_2 with a GraphSage (Hamilton et al., 2017) encoder. Results in Tables 13 and 14 showcase that GNN outperforms the VGAE-based models across all experiments, datasets, and splitting strategies, demonstrating the robustness of our approach. Temporal splitting also improves the performance for TX, highlighting that the dataset with the largest labeled businesses allows the model to learn patterns that generalize across time.

Test Set with Real-World Imbalance. We create a test set with a real-world imbalance to stress-test our approach. The Network (2024) estimates that 15,000 IMBs operated in the U.S. in 2024, and the IBISWorld (2025) reports 202,221 massage service businesses in the U.S. in 2024, resulting in an imbalance ratio of 0.08 (illicit to non-illicit). Table 15 illustrates that GNN achieves higher average precision scores than the baselines across all datasets, and therefore reinforces its generalizability in the real-world setting.

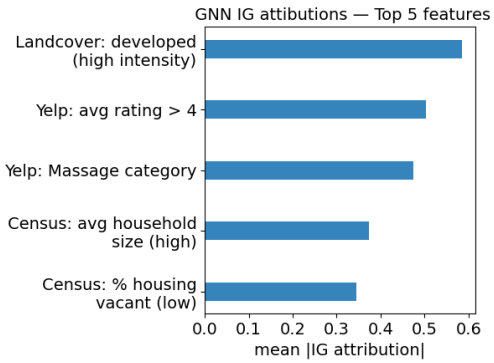

Figure 4: Mean IG attributions from the GNN Model.

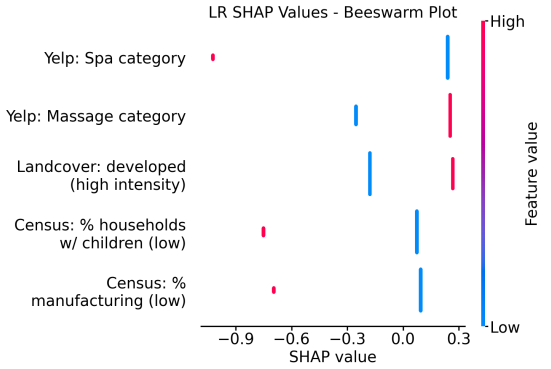

Figure 5: Beeswarm plot of the SHAP values.

5 Feature Importance

To assess the importance of each feature in the GNN model, we leverage a gradient-based method called Integrated Gradients (IG) (Sundararajan et al., 2017), which we implement using Captum (Kokhlikyan et al., 2020). IG evaluates the model at m different inputs created by linearly interpolating the original input features x and a neutral baseline x'. It then averages the gradients of the output with respect to those inputs to generate an attribution score IG_i, which captures the importance of the i^{th} feature, defined in Equation 4.

$$IG_i = (x_i - x_i') \cdot \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F\left(x' + \frac{k}{m}(x - x')\right)}{\partial x_i}, \tag{4}$$

We calculate the average of the absolute values of these scores across all nodes and subgraphs in the testing set to generate final feature-level attributions. For this, we use the GNN_woz model to focus on business features. Figure 4 depicts the top five most important business features in the test set for Experiment 2 (Link Classification) for CO. To generate attribution scores for a link, we can calculate the sum of the absolute values of the attribution scores of the endpoint nodes and consider this as a proxy for link importance. Figure 8 depicts a subgraph with its links highlighted on the basis of these attributions.

To compare the feature importance of GNN_woz with LR, we generate the SHAP values (Lundberg and Lee, 2017) of the LR features. Figure 5 presents a beeswarm plot highlighting the five most influential features based on SHAP values.

In Figures 4 and 5, two of the top five features (explained in Table 16) appear in both models, showcasing consistency in identifying salient signals. Since the GNN model leverages the network structure and message passing across business, review, and reviewer nodes, it identifies different business features of high importance for classifying a link.

6 Discussion and Conclusions

This work studies the network between illicit massage businesses using multifaceted business-review data and extends the link prediction models to classify links and detect exposure. The LR model utilizes handcrafted business features, leveraging Natural Language Processing and the domain expertise of collaborators, and builds on a logistic regression classifier without incorporating network information. To overcome this, we introduce a graph machine learning-based GNN model that learns latent structural relationships between business, review, and reviewer nodes in a business review dataset and infers high-order links between businesses. We introduce four experiments motivated by investigative goals. Two link-level experiments, which aim to predict and classify the link between two business nodes, and two node-level experiments, which classify a business node and detect its exposure to illicit business nodes. The GNN model outperforms LR and VGAE models in the former set of experiments, where the relative structure of the subgraph around the target business nodes is key, whereas LR dominates when individual business features are crucial. In this work, we focus on learning a network to infer links between business nodes; an extension of this work can incorporate predicting links between different node types. Additionally, as we observed improved performance across both the smallest and largest datasets, future work will aim to study graph-topology characteristics to elucidate this performance gain.

7 Ethical Considerations

The proposed framework can enhance the transparency of disciplinary actions and inform regulatory policies that protect vulnerable industries, such as the massage industry. Furthermore, with comprehensive experiments driven by investigative goals to identify illicit massage businesses and their connections, the models demonstrate a real-world use case that can serve as a decision-support tool for law enforcement agencies and counter-human trafficking organizations, facilitating the proper allocation of investigative resources to uncover a hidden network of illicit businesses. The classification models proposed in this work can generate false positive results. Therefore, the models should be used cautiously to inform decision-making and prioritize investigations. This work supports reproducibility without raising ethical concerns or posing risks to society. Additionally, the datasets were constructed in accordance with strict ethical guidelines, ensuring the anonymity of businesses and reviewers. The reviews used in this work are redacted to remove the names of individual people. However, it can contain offensive content about commercial sex. To protect data, we share it only with relevant researchers and investigators through a data use agreement.

8 Limitations

We employ the inductive splitting of businesses into training, validation, and testing sets to prevent data leakage; however, this results in the training graphs being larger in size than the validation and testing graphs. A nuanced splitting strategy to create disjoint business sets with graphs of similar characteristics can improve model performance. We also limit sampling to 2,500 business pairs for hyperparameter optimization and 10,000 business pairs for testing to control computational complexity. Multiple replications of hyperparameter tuning and model testing with additional samples will enhance the robustness of our results. We used three datasets containing Yelp reviews for massage businesses in CO, FL, and TX, which limits the generalizability of the results to other states with a significant domain shift; however, the proposed model is generalizable to other datasets. Finally, manual labeling of the data required domain knowledge and careful reasoning, which constrained the availability of annotated data.

References

Manuel B. Aalbers and Magdalena Sabat. 2012. Remaking a landscape of prostitution: the Amsterdam red light district. *City*, 16(1-2):112–128.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Nathanael Chambers, Timothy Forman, Catherine Griswold, Kevin Lu, Yogaish Khastgir, and Stephen Steckler. 2019. Character-based models for adversarial phone extraction: Preventing human sex trafficking. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

John J. Chin, Lois M. Takahashi, and Douglas J. Wiebe. 2023. 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.

Eleanor Cockbain, Helen Brayley, and Gloria Laycock. 2011. Exploring internal child sex trafficking networks using social network analysis. *Policing: A Journal of Policy and Practice*, 5(2):144–157.

Sean M. Crotty and Vanessa Bouché. 2018. The red-light network: Exploring the locational strategies of illicit massage businesses in Houston, Texas. *Papers in Applied Geography*, 4(2):205–227.

Meredith Dank, Bilal Khan, P. Mitchell Downey, Cybele Kotonias, Deborah Mayer, Colleen Owens, Laura Pacifici, and Lilly Yu. 2014. Estimating the size and structure of the underground commercial sex economy in eight major us cities. Research report, Urban Institute.

Deanna Davy. 2022. Trafficked by someone I know: A qualitative study of the relationships between trafficking victims and human traffickers in Albania. Research report, UNICEF Albania & IDRA.

Ieke de Vries. 2023. 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.

Ieke de Vries and Jason Radford. 2021. 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.

Vasuki Garg, Osman Y. Özaltın, Maria E. Mayorga, and Sherrie Bosisto. 2025. Detecting illicit massage businesses by leveraging graph machine learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 9647–9655. International Joint Conferences on Artificial Intelligence Organization. AI and Social Good.

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

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 855–864, New York, NY, USA. Association for Computing Machinery.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

William L. Hamilton. 2020. *Graph Representation Learning*, 1 edition. Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer Cham.

IBISWorld. 2025. Massage services in the US - number of businesses (2005–2031). https://www.ibisworld.com/united-states/number-of-businesses/massage-services/6028/. [Accessed 08-02-2025].

Elizabeth Ranade Janis. 2020. Illicit massage businesses: The pervasive, insidious form of trafficking happening across the United States. https://traffickinginstitute.org/illicit-massage-businesses/. [Accessed 11-06-2024].

Thomas N. Kipf and Max Welling. 2016a. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907.

Thomas N. Kipf and Max Welling. 2016b. Variational graph auto-encoders. *Preprint*, arXiv:1611.07308.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch. *CoRR*, abs/2009.07896.

Stephanie Lasker. 2002. Sex and the city: Zoning pornography peddlers and live nude shows. *UCLA Law Review*, 49(4):1139–1185.

Ruoting Li, Margaret Tobey, Maria E. Mayorga, Sherrie Caltagirone, and Osman Y. Özaltın. 2023. Detecting human trafficking: Automated classification of online customer reviews of massage businesses. *Manufacturing & Service Operations Management*, 25(3):1051–1065.

Yifei Li, Pratheeksha Nair, Kellin Pelrine, and Reihaneh Rabbany. 2022. Extracting person names from user generated text: Named-entity recognition for combating human trafficking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2854–2868, Dublin, Ireland. Association for Computational Linguistics.

David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031.

Sitao Luan, Chenqing Hua, Qincheng Lu, Jiaqi Zhu, Mingde Zhao, Shuyuan Zhang, Xiao-Wen Chang, and Doina Precup. 2022. Revisiting heterophily for graph neural networks. *Preprint*, arXiv:2210.07606.

Kristina Lugo-Graulich. 2024. *Indicators of Sex Trafficking in Online Escort Ads, 7 U.S. States, 2013–2020*. Inter-university Consortium for Political and Social Research, Ann Arbor, MI.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Anastasija Mensikova and Chris A. Mattmann. 2018. 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.

Michael Miklaucic and Jacqueline Brewer, editors. 2013. *Convergence: Illicit Networks and National Security in the Age of Globalization*. National Defense University Press, Washington, DC.

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

Alexandra K. Murphy and Sudhir A. Venkatesh. 2006. Vice careers: The changing contours of sex work in new york city. *Qualitative Sociology*, 29(2):129–154.

Organisation for Economic Co-operation and Development. 2016. *Trafficking in Persons and Corruption: Breaking the Chain*. OECD Publishing, Paris. Chapter 1: Trafficking in persons: Weak governance and growing profits.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 701–710, New York, NY, USA. Association for Computing Machinery.

Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):1–21.

Vageesh Saxena, Benjamin Ashpole, Gijs van Dijck, and Gerasimos Spanakis. 2023. IDTraffickers: An authorship attribution dataset to link and connect potential human-trafficking operations on text escort advertisements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8444–8464, Singapore. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Pedro Szekely, Craig A. Knoblock, Jason Slepicka, Andrew Philpot, Amandeep Singh, Chengye Yin, Dipsy Kapoor, Prem Natarajan, Daniel Marcu, Kevin Knight, David Stallard, Subessware S. Karunamoorthy, Rajagopal Bojanapalli, Steven Minton, Brian Amanatullah, Todd Hughes, Mike Tamayo, David Flynt, Rachel Artiss, and 4 others. 2015. Building and using a knowledge graph to combat human trafficking. In *The Semantic Web - ISWC 2015*, pages 205–221. Springer Cham.

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

Margaret Tobey, Ruoting Li, Osman Y. Özaltın, Maria E. Mayorga, and Sherrie Caltagirone. 2024. Interpretable models for the automated detection of human trafficking in illicit massage businesses. *IISE Transactions*, 56(3):311–324.

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

Catalina Vajiac, Meng-Chieh Lee, Aayushi Kulshrestha, Sacha Levy, Namyong Park, Andreas Olligschlaeger, Cara Jones, Reihaneh Rabbany, and Christos Faloutsos. 2023. Deltashield: Information theory for human- trafficking detection. *ACM Trans. Knowl. Discov. Data*, 17(2).

Diego Valsesia, Giulia Fracastoro, and Enrico Magli. 2023. Ran-gnns: Breaking the capacity limits of graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):4610–4619.

Anna White, Seth Guikema, and Bridgette Carr. 2024. Why are you here? modeling illicit massage business location characteristics with machine learning. *Journal of Human Trafficking*, 10(1):20–40.

Muhan Zhang and Yixin Chen. 2018. Link prediction based on graph neural networks. *CoRR*, abs/1802.09691.

A Appendix: Business Pairs and Homophily

Data	Hop	Connected Business Pairs	Hop Homophily
CO	4	3715	81.5%
	8	40326	79.2%
	12	37335	65.4%
	16	2837	23.6%
	20	42	90.5%
FL	4	9171	87.1%
	8	134336	85.2%
	12	122394	64.2%
	16	12386	21.3%
	20	1052	39.6%
TX	4	22121	84.4%
	8	413551	81.1%
	12	268967	55.0%
	16	14416	32.2%
	20	342	71.6%

Table 3: Number of connected business pairs and homophily across hops for CO, FL, and TX.

B Appendix: Undersampled Business Data Statistics

Data	Total	Illicit	Non-Illicit	Ratio
CO	425	85	340	0.25
FL	785	157	628	0.25
TX	1230	246	984	0.25

Table 4: Business statistics for CO, FL, and TX.

C Appendix: Class Distribution across the Datasets after Splitting

Exp	Data	Split	Class 1	Class 0
1	CO	Train	13311	23545
		Val	305	1973
		Test	765	2805
		Total	14381	28323
	FL	Train	44790	80961
		Val	975	6900
		Test	1792	10454
		Total	47557	98315
	TX	Train	141470	167821
		Val	3291	16015
		Test	6747	23388
		Total	151508	207224
2	CO	Train	2715	10596
		Val	83	222
		Test	214	551
		Total	3012	11369
	FL	Train	6614	38176
		Val	111	864
		Test	204	1588
		Total	6929	40628
	TX	Train	25483	115987
		Val	543	2748
		Test	1564	5183
		Total	27590	123918

Table 5: Total Class 1 and Class 0 links across splits for Experiments 1 and 2.

D Appendix: Network Measures

Data	Avg. deg	Density	$\bar{z}_1$	$\bar{z}_2$
CO	2.0394	0.0007	15	380
FL	2.0484	0.0004	19	306
TX	2.1009	0.0002	27	756

Table 6: Network statistics and measures in the training sets of CO, FL, and TX.

E Appendix: Node Types and Counts

Data	Business	Review	Reviewer
CO	425	7662	5523
FL	785	13584	9335
TX	1230	21824	13508

Table 7: Number of businesses, reviews, and reviewers in CO, FL, and TX.

F Appendix: Hyperparameters for FL across the Four Experiments

Param	Exp 1	Exp 2	Exp 3	Exp 4
L	4	6	4	6
$\|h_i^l\|$	8	32	32	32
k	0.6	0.6	0.9	0.6
(c_1, c_2)	(4,8)	(4,8)	(4,8)	(4,8)
m	6	6	6	6
Loss	BCE	BCE	BCE	BCE
Optimizer	Adam	Adam	Adam	Adam
Epochs	100	100	100	100

Table 8: Hyperparameters across experiments for FL. The first four parameters are tuned. BCE: Binary Cross Entropy.

G Appendix: Positive Class Prevalence

Data	Exp 1	Exp 2	Exp 3	Exp 4
CO	0.2143	0.2797	0.2000	0.2000
FL	0.1463	0.1138	0.1975	0.1975
TX	0.2239	0.2318	0.1992	0.1992

Table 9: Prevalence of the positive class in the three test sets for average precision (Avg Prec) baseline values.

H Appendix: Ablation Study - Positional Encoding Scores

Data	Exp	Method	AUC	Avg Prec
CO	1	GNN	**0.7781**	**0.8076**
		GNN_woz	0.7030	0.6921
	2	GNN	**0.7885**	**0.7718**
		GNN_woz	0.5022	0.5049
	3	GNN	**0.6626**	**0.5912**
		GNN_woz	0.6280	0.5260
	4	GNN	**0.5950**	0.9387
		GNN_woz	**0.5950**	**0.9404**
FL	1	GNN	**0.6492**	**0.6984**
		GNN_woz	0.6059	0.5886
	2	GNN	**0.7725**	**0.8056**
		GNN_woz	0.4963	0.4891
	3	GNN	0.8971	0.8265
		GNN_woz	**0.9188**	**0.8682**
	4	GNN	**0.7414**	**0.9434**
		GNN_woz	0.7091	0.9343
TX	1	GNN	**0.7324**	**0.7876**
		GNN_woz	0.6233	0.6111
	2	GNN	0.5485	0.5304
		GNN_woz	**0.5733**	**0.5748**
	3	GNN	**0.8580**	**0.7726**
		GNN_woz	0.8510	0.7129
	4	GNN	**0.8534**	**0.9920**
		GNN_woz	0.8197	0.9892

Table 10: Prediction performance of the GNN model with and without (z_1, z_2) positional encodings across datasets.

I Appendix: Key Hyperparameters Tuned for CO and TX

	CO		TX	
Param	Exp 1	Exp 2	Exp 1	Exp 2
L	4	6	4	6
$\|h_i^l\|$	16	32	8	32
k	0.9	0.3	0.9	0.3
(c_1, c_2)	(8,16)	(8,16)	(4,8)	(8,16)

Table 11: Tuned hyperparameters across Experiments 1 and 2 for CO and TX.

J Appendix: Ablation Study - Feature Pruning

Data	Exp	Method	AUC	Avg Prec
CO	1	GNN_woz	**0.7177**	**0.6846**
		GNN_sel	0.6860	0.6646
	2	GNN_woz	0.5512	**0.5639**
		GNN_sel	**0.5691**	0.5542
FL	1	GNN_woz	**0.6059**	**0.5886**
		GNN_sel	0.5485	0.5365
	2	GNN_woz	0.4963	0.4891
		GNN_sel	**0.6751**	**0.6587**
TX	1	GNN_woz	0.5728	0.5670
		GNN_sel	**0.5992**	**0.6014**
	2	GNN_woz	**0.5000**	**0.5000**
		GNN_sel	**0.5000**	**0.5000**

Table 12: Prediction performance of the models with IG-based feature pruning using fine-tuned hyperparameters for each dataset.

K Appendix: Ablation Study - Dataset Splitting

Data	Exp	Method	AUC	Avg Prec
CO	1	GNN	**0.7295**	**0.7820**
		VGAE_1	0.5455	0.5174
		VGAE_2	0.4952	0.4974
	2	GNN	**0.8384**	**0.8500**
		VGAE_1	0.2729	0.3825
		VGAE_2	0.6382	0.6235
FL	1	GNN	**0.6492**	**0.6984**
		VGAE_1	0.6465	0.5912
		VGAE_2	0.6105	0.5895
	2	GNN	**0.7725**	**0.8056**
		VGAE_1	0.4852	0.4944
		VGAE_2	0.6757	0.6789
TX	1	GNN	**0.7187**	**0.7804**
		VGAE_1	0.5707	0.5427
		VGAE_2	0.5670	0.5558
	2	GNN	**0.5000**	**0.5000**
		VGAE_1	0.1673	0.3350
		VGAE_2	0.1713	0.3352

Table 13: Prediction performance with inductive splitting using fine-tuned hyperparameters for each dataset.

Data	Exp	Method	AUC	Avg Prec
CO	1	GNN	**0.7192**	**0.7438**
		VGAE_1	0.6247	0.5841
		VGAE_2	0.5400	0.5310
	2	GNN	**0.7206**	**0.7897**
		VGAE_1	0.3970	0.4706
		VGAE_2	0.7054	0.7213
FL	1	GNN	**0.7900**	**0.8037**
		VGAE_1	0.5985	0.5694
		VGAE_2	0.6274	0.5781
	2	GNN	**0.6734**	**0.7186**
		VGAE_1	0.4588	0.4781
		VGAE_2	0.6237	0.6205
TX	1	GNN	**0.7879**	**0.8326**
		VGAE_1	0.6063	0.5719
		VGAE_2	0.6516	0.6144
	2	GNN	**0.8194**	**0.8130**
		VGAE_1	0.4398	0.4854
		VGAE_2	0.7975	0.7804

Table 14: Prediction performance with temporal splitting using fine-tuned hyperparameters for each dataset.

L Appendix: Ablation Study - Test Set with Real-World Imbalance

Data	Exp	Method	AUC	Avg Prec
CO	1	GNN	0.7510	**0.4925**
		VGAE_1	**0.7729**	0.2236
		VGAE_2	0.5287	0.1006
	2	GNN	**0.6730**	**0.3175**
		VGAE_1	0.4808	0.1661
		VGAE_2	0.6130	0.2670
FL	1	GNN	0.7345	**0.4884**
		VGAE_1	**0.7791**	0.2329
		VGAE_2	0.6155	0.1207
	2	GNN	**0.6132**	**0.0993**
		VGAE_1	0.4955	0.0649
		VGAE_2	0.5832	0.0921
TX	1	GNN	**0.8057**	**0.6508**
		VGAE_1	0.6850	0.2466
		VGAE_2	0.6113	0.1943
	2	GNN	**0.5000**	**0.1268**
		VGAE_1	0.3968	0.0958
		VGAE_2	0.2661	0.0802

Table 15: Prediction performance of the models on the test set with real-world imbalance using fine-tuned hyperparameters for each dataset.

M Appendix: Model Performance at different Imbalance ratios for link-level Experiments for FL and TX

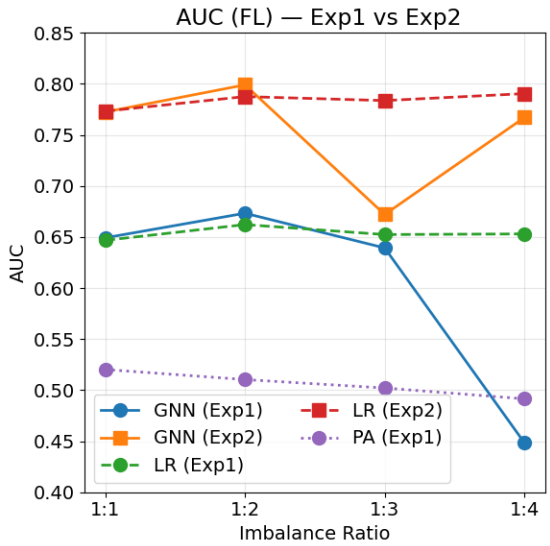

Figure 6: Model performance at different imbalance ratios for link-level experiments for FL.

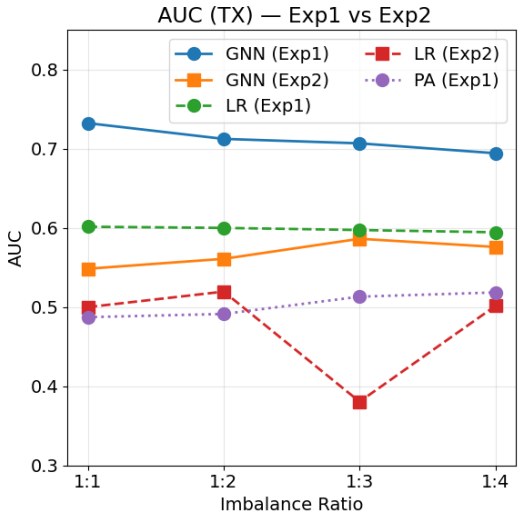

Figure 7: Model performance at different imbalance ratios for link-level experiments for TX.

N Appendix: Subgraph with Link Importance

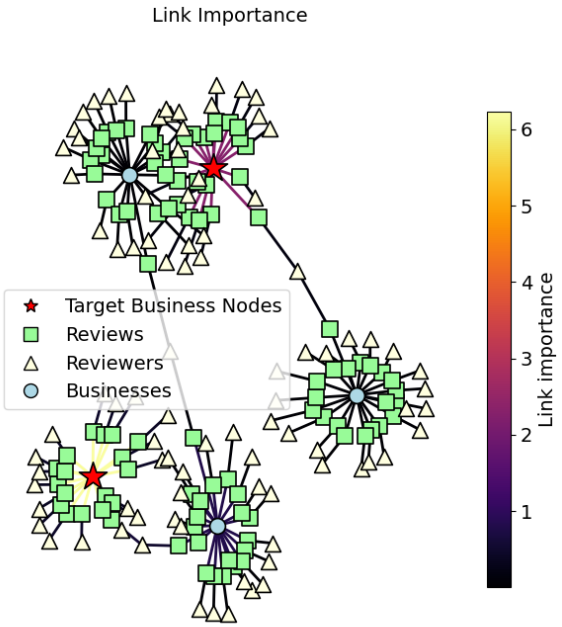

Figure 8: A Class 1 (Experiment 2) subgraph with highlighted link importance.

O Appendix: Description of Business Features

For the complete list, please refer to *Table 3: Selected Data Features* in (Tobey et al., 2024)

Features	Description
Yelp: Spa category	if business categorizes under Day Spas, Medical Spas, Saunas, Float-Spa, Beauty & Spas
Yelp: Massage category	if business categorizes under Massage Therapy, Massage, Reflexology, Reiki, Tui-Na & Massage Schools
Yelp: average rating > 4	if the average of all Yelp review ratings is greater than 4
Census: % households with children (low)	if the percentage of households with children in the zip code is low
Census: % manufacturing industry (low)	if the percentage of people employed in the zip code who are in the manufacturing industry is low
Census: % housing vacant (low)	if the percentage of vacant housing in the zip code is low
Census: average household size (high)	if the average household size in the zip code is low
Landcover: developed (high intensity)	if landcover type in the zip code is underdeveloped high intensity (NLCD)

Table 16: Description of binary business features.