Adversarial Label-Flipping Attack and Defense for Anomaly Detection in Spatial Crowdsourcing UAV Services

Junaid Akram, Ali Anaissi, Awais Akram, Rajkumar Singh Rathore, Rutvij H. Jhaveri

***Abstract*—The rapid expansion of Graph Neural Networks (GNNs) in consumer electronics and Vehicular Edge Computing (VEC) enhanced Internet of Drone Things (IoDT) services highlights the need for strong defenses against cyber attacks. One significant but overlooked threat is adversarial label-flipping, where attackers slightly change training labels to disrupt the system. This issue is critical in spatial crowdsourcing UAV networks that use potentially insecure labels. Our study investigates these attacks on GNNs, emphasizing a serious security problem. We introduce UAVGuard, an innovative attack model that uses continuous approximations for complex objectives and a simplified GNN structure for effective gradient-based attacks. Our analysis shows that GNNs' vulnerability mainly comes from overfitting to these manipulated labels. To counter this, we offer a defensive framework that uses a community-preserving self-supervised task as a regularization method. Tests on three real-world datasets, including various IBRL modalities, demonstrate UAVGuard's effectiveness and our defense architecture's resilience to label-flipping attacks. This research enhances our understanding of these threats to GNNs and provides practical defenses, improving the security of UAV services in spatial crowdsourcing within VEC-enhanced IoDT systems.**

***Index Terms*—Graph Neural Networks, Label-Flipping Attacks, Spatial Crowdsourcing, UAV Networks, Bi-Level Optimization, Data Security.**

I. Introduction

The integration of Graph Neural Networks (GNNs) into consumer electronics and Vehicular Edge Computing (VEC) has revolutionized spatial crowdsourcing and Unmanned Aerial Vehicle (UAV) services. VEC, by bringing computational resources closer to data sources, significantly enhances UAV networks' capabilities, enabling rapid processing and decision-making [1]–[4]. This advancement is crucial for real-time applications such as traffic monitoring, environmental surveillance, and public safety [5]–[7], where swift analytics are essential. Within this framework, GNNs excel at modeling complex interactions within UAV networks, facilitating breakthroughs in anomaly detection, route optimization, and service allocation. However, the adoption of GNNs in these advanced contexts introduces significant vulnerabilities, particularly to adversarial attacks like poisoning [8], [9], where training data manipulation can severely compromise model integrity.

A specific but underexplored threat in UAV services is adversarial label-flipping attacks [10], [11], where a minor fraction of training labels are maliciously altered. These attacks, while investigated in various domains [12], have critical implications in UAV networks that rely on real-time, externally sourced data. The key question arises: "How do adversarial label-flipping attacks affect the reliability of GNNs in UAV networks, and what defense mechanisms can be implemented to mitigate these attacks?" Consider a scenario where a GNN model [13] is deployed within a UAV network. The manipulation of even a small number of labels, such as inaccurately altering a UAV's status, can significantly reduce the model's predictive accuracy, undermining operational decisions. This situation illustrates how susceptible GNNs are to targeted label-flipping attacks, highlighting the urgent need for robust defense mechanisms to maintain data integrity and decision-making accuracy in UAV services.

To address these challenges, we introduce **UAVGuard**, an innovative adversarial label-flipping attack model specifically designed for GNNs in UAV networks. UAVGuard addresses the dual challenges of bi-level optimization and non-differentiability in label-flipping attacks by using GNN linearization and continuous surrogates, which are smooth, differentiable approximations of non-differentiable functions, for gradient-based optimization [14]–[19]. This approach, along with an approximate closed form of GNNs, significantly enhances the execution of effective, gradient-based attacks on GNNs, establishing UAVGuard as a novel contribution to the field.

Recognizing the critical need for data integrity in UAV networks, we propose a self-supervised defense framework aimed at bolstering GNN resilience against label-flipping attacks. This framework leverages the natural community structures within UAV networks, employing community detection to generate robust training signals. By incorporating a community-preserving self-supervised task as a novel regularization technique, our defense approach effectively prevents overfitting to tampered labels, ensuring the reliability of GNN-based decision processes in UAV services.

The implications of our research extend beyond UAV networks to broader consumer electronics and technologies. UAVGuard enhances the security and robustness of systems

Junaid Akram and Ali Anaissi are with School of Computer Science, The University of Sydney, Camperdown, NSW 2008 Australia (e-mail: jakr7229@sydney.edu.au, ali.anaissi@sydney.edu.au).

Awais Akram is with COMSATS University Islamabad, Vehari, Pakistan (e-mail: awais.akram.1212@gmail.com).

Rajkumar Singh Rathore is with Cardiff School of Technologies, Cardiff Metropolitan University, United Kingdom. (e-mail: rsrathore@cardiffmet.ac.uk)

Rutvij H. Jhaveri is with Department of Computer Science and Engineering, School of Technology, Pandit Deendayal Energy University, India (e-mail: rutvij.jhaveri@sot.pdpu.ac.in).

Corresponding Authors: Rutvij H. Jhaveri, Junaid Akram

relying on GNNs, providing critical insights into safeguarding advanced computational models in consumer applications. By addressing existing limitations in defense mechanisms against adversarial label-flipping attacks, our work contributes to developing more secure and reliable deployments of GNNs in various consumer technology contexts.

In summary, UAVGuard represents a significant advancement in cybersecurity for consumer electronics, particularly in UAV services. Our approach effectively addresses the vulnerabilities of GNNs to adversarial attacks, providing a robust defense mechanism that enhances data integrity and operational reliability in critical real-time applications. This research not only improves the security of UAV networks but also offers valuable strategies for protecting GNNs in a wide range of consumer electronics applications. Our key contributions are as follows:

- We investigate the robustness of GNNs against label-flipping attacks in the newly emerging area of VEC combined with UAV services. We present **UAVGuard**, an advanced adversarial label-flipping attack model tailored for GNNs. This model efficiently executes gradient-based attacks on GNNs by leveraging an approximate closed form of GNNs and addressing the challenges posed by non-differentiable components.
- We propose a pioneering defense mechanism designed to counteract label-flipping attacks on GNNs within the VEC-UAV framework. Our defense strategy employs a novel community-preserving self-supervised learning approach. This innovative method significantly reduces the risk of overfitting to manipulated labels, thereby bolstering the robustness of GNNs to such adversarial threats.
- Through extensive empirical analysis on three distinct real-world datasets from the IBRL dataset, encompassing temperature, humidity, and light modalities, we demonstrate the critical vulnerability of GNNs to adversarial label-flipping attacks in VEC-enhanced UAV service environments. Our results validate the effectiveness of the UAVGuard attack model. Furthermore, we show that our defense framework significantly improves the resilience of traditional GNNs and their variants against label-flipping attacks, highlighting its efficacy and adaptability in safeguarding VEC-driven UAV services.

The paper is organized as follows: Section II reviews related work on GNNs, anomaly detection in graphs, learning with noisy data, and adversarial attacks, including label-flipping attacks and defenses. Section III presents the problem formulation, detailing our GCN model, adversarial attack strategy, attack objective, and defense strategy for anomaly detection. Section IV evaluates the effectiveness of UAVGuard and our defense framework. Finally, Section V concludes by summarizing our contributions and suggesting future research directions for securing GNNs in UAV services.

II. Related Work

GNNs have become a cornerstone in processing graph-structured data, with applications ranging from node classification to link prediction and graph clustering [20]–[24]. Among the subclasses of GNNs, Graph Convolutional Networks (GCNs) stand out for their ability to leverage graph topologies through convolution operations, driving research towards enhancements such as simplification techniques, the addition of connectivity layers, and attention mechanisms [25]. This work specifically targets GCNs and their derivatives to explore vulnerabilities to adversarial attacks. The realm of anomaly detection within graph data, particularly in dynamic graph structures, has witnessed significant advancements [26], [27]. While initial methods were grounded in traditional machine learning [28], recent approaches harness deep learning to achieve superior anomaly detection performance [20], [29]–[31]. These methodologies, however, often treat structural and temporal data independently and struggle with the challenge of noisy labels, underscoring a need for refined strategies that address these limitations.

The challenge of learning with noisy data, especially in non-IID graph data, has spurred various strategies aimed at enhancing the robustness of GNNs against label noise [32], [33]. Despite the plethora of research in other domains, the application of these strategies in graph data is not as widespread, partly due to the unique challenges posed by the graph structure [34]. Adversarial attacks on GNNs have been thoroughly investigated, revealing vulnerabilities to manipulations such as edge rewiring or node feature alterations [35]–[37]. Particularly, poisoning attacks that tamper with training data present significant threats due to the complex nature of bi-level optimization problems associated with these attacks [8]. This work contributes to this domain by directly addressing these challenges, presenting a novel perspective on adversarial label-flipping attacks and defenses.

Adversarial label-flipping attacks have gained attention for their potential to severely degrade the performance of learning models, including GNNs, by exploiting gradient information to induce misclassification [38]. Despite the existence of strategies against such attacks in other learning paradigms, the adaptation of these strategies to GNNs has been limited, with few works like Liu et al. [14] providing a foundation for this study. Common defense strategies often involve data sanitization [39], yet there is a scarcity of comprehensive defense mechanisms tailored to GNNs. This paper aims to fill this gap by focusing on both adversarial label-flipping attack models and defense strategies within the context of GNNs.

III. System Model

A. Problem Formulation

The Internet of Drone Things (IoDT) systems has emerged as a pivotal technology for environmental monitoring, utilizing a dense network of UAV nodes to gather comprehensive spatial and temporal data. These systems capitalize on the strong correlation between the data collected by a UAV node and its geospatially adjacent nodes. For example, when a UAV node identifies a bushfire, it is likely that neighboring nodes will register related changes in their environmental readings, such as elevated temperatures or variations in air quality parameters like carbon dioxide levels and humidity.

This study leverages these spatiotemporal correlations for enhanced anomaly detection in IoDT systems, approaching the

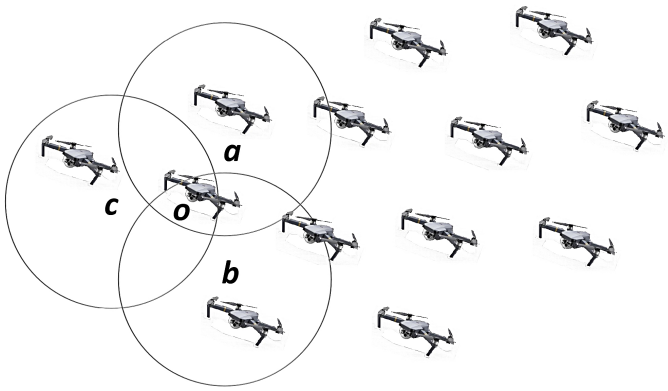


Fig. 1. Illustration of synchronized UAV monitoring capturing a bushfire event, highlighting the spatial correlation among data collected by geospatially adjacent UAV nodes.

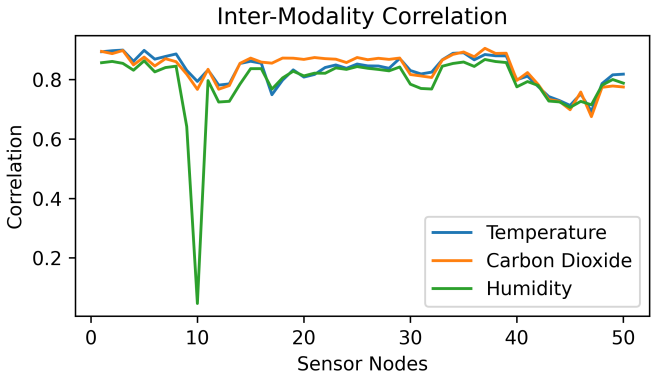


Fig. 2. Demonstration of inter-modality correlation in UAV-collected data during a bushfire event, showcasing the temporal correlation among temperature, carbon dioxide levels, and humidity.

challenge through a regression problem based on time series forecasting. For a given time series dataset $\{X_t\}_{t\in T}$ and a window W, we propose a GNN model for prediction:

$$X_{t+W} = f\left(X_t, X_{t-1}, \ldots, X_{t+W-1}|\theta\right), \tag{1}$$

where f signifies the mapping function, and θ embodies the model parameters. By simplifying f into a neural network with a single hidden layer, we adapt the prediction equation as follows:

$$X_{t+W} = \theta_0 + \sum_{j=1}^{D} \theta_j g\left(\theta_{0j} + \sum_{i=1}^{W} \theta_{ij} X_{t+i-1}\right), \tag{2}$$

with D denoting the number of nodes in the hidden layer, g as the activation function, and θ_0, θ_j, θ_{0j}, θ_{ij} as the trainable parameters. The anomaly detection model f is tasked with forecasting data for forthcoming timestamps, with anomalies identified when the prediction error surpasses a predefined threshold η, typically set based on the highest inference score obtained on a validation dataset.

$$\text{Score}\left(\tilde{X}_{t+W}, X_{t+W}\right) > \eta, \tag{3}$$

This methodology introduces an innovative GNN model tailored for IoDT systems, exploiting the systems' topology and intrinsic spatiotemporal data correlations for efficient anomaly detection. By integrating diverse data across UAV nodes, modalities, and temporal instances, our model aims to accurately identify anomalies within the IoDT data flow.

B. UAV Attribute Framework

Let us envisage a framework tailored for the IoDT system, represented as $G = (A, X)$. Within this framework, $A \in R^{M\times M}$ signifies a matrix that encapsulates the intricate interconnections binding the UAVs, with M depicting the aggregate tally of UAVs incorporated in the network. Concurrently, the attribute matrix for the M UAVs is represented as $X \in R^{M\times D}$. The attributes of each individual UAV can be succinctly articulated through a D-dimensional vector. By coalescing these D-dimensional vectors from the entirety of the M UAVs, we obtain the holistic attribute matrix X, which serves as the foundational pillar of our UAV framework.

C. Graph Convolutional Network (GCN) Model

The Graph Convolutional Network (GCN) model is designed to decipher the complex spatial and temporal correlations embedded in the data ecosystem of the Internet of Drone Things (IoDT) system. It operates through several layers, each meticulously processing graph-structured data to distill pertinent features. The model bifurcates into two core components: the extraction of spatial features via graph convolutions and the extraction of temporal features through a time series forecasting model. Spatial feature extraction within the GCN is facilitated by a sequence of graph convolution layers. These layers apply a convolution operation to the graph data, enabling the model to discern spatial relationships among nodes (UAVs). The mathematical expression for the graph convolution operation at any given layer is delineated as follows:

$$H^{(l+1)} = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right), \tag{4}$$

where:

- At the l-th layer, the node feature matrix is represented by $H^{(l)}$ while the input feature matrix at time t is represented by $H^{(0)} = X_t$.
- $\hat{A} = A + I_M$ denotes the adjacency matrix of the graph G augmented with self-connections, I_M being the identity matrix.
- $\hat{D}$ is the diagonal degree matrix of $\hat{A}$, where $\hat{D}_{ii} = \sum_j \hat{A}_{ij}$.
- The weight matrix for the l-th graph convolution layer is represented as $W^{(l)}$.
- σ signifies a non-linear activation function, such as the rectified linear unit (ReLU).

Information propagation through the network is governed by the following layer-wise update rule:

$$H^{(l+1)} = \sigma\left(\text{GraphConv}(H^{(l)}, A)\right), \tag{5}$$

with GraphConv$(\cdot)$ encapsulating the graph convolution operation as previously defined.

Subsequent to the extraction of spatial features, the GCN integrates a temporal dimension to capture the evolving dynamics of the data. The temporal dynamics are encapsulated using a linear time series forecasting model, articulated as:

$$X_{t+W} = \Theta_0 + \sum_{i=1}^{W} \Theta_i H_{t+i-1}, \tag{6}$$

where:

- W delineates the window length for the time series forecasting.
- Θ_0, Θ_i are the trainable parameters of the model.
- H_{t+i-1} reflects the output of the GCN at timestamp $t+i-1$.

The overarching architecture of the GCN model is summarized as follows:

1) Input Layer: Receives the multi-modal data X_t from the IoDT system.
2) Graph Convolution Layers: A series of layers executing the graph convolution operation to elicit spatial features.
3) Temporal Forecasting Layer: A linear model delineating the temporal dynamics of the data.
4) Output Layer: Generates the forecasted data $\tilde{X}_{t+W}$ for anomaly detection purposes.

The construction of the GCN model is based on several key assumptions about the UAV networks' graph structure. It assumes that the graph is static and well-defined, with all node connections accurately represented. This might not always hold true in dynamic UAV networks, where node connections can frequently change due to UAV mobility or varying environmental conditions. Additionally, the model assumes that the features of connected nodes are correlated, which might not always be the case in heterogeneous networks with diverse UAV functionalities. A limitation of this approach is its reliance on accurate graph structure representation. Inaccurate or incomplete graph structures can significantly degrade the model's performance. Moreover, the model's complexity increases with the graph size, which can pose scalability issues for large-scale UAV networks. Ensuring the robustness of the model in such scenarios necessitates further research and optimization.

The initialization and optimization of model parameters are critical steps in ensuring the effective performance of the GCN. Parameters are initialized using Xavier initialization to ensure that the weights are set to values that maintain the variance of the activations constant across layers. This helps in mitigating issues related to vanishing or exploding gradients. During the optimization phase, the Adam optimizer is employed due to its adaptive learning rate capabilities, which are particularly useful in handling the sparse and diverse nature of graph data. The learning rate, number of epochs, and dropout rates are tuned using cross-validation on the validation set to prevent overfitting and ensure generalizability.

D. Anomaly Detection

Anomaly detection within IoDT systems, facilitated by the Graph Convolutional Network (GCN) model, is predicated on discerning deviations between forecasted data and actual observations. This crucial process encompasses the determination of anomaly scores and thresholds for anomaly identification, leveraging the GCN model's ability to capture the intricate spatial and temporal correlations present in IoDT data.

The anomaly score, serving as a quantitative indicator of prediction deviations, is calculated at each timestamp to pinpoint potential anomalies. It is mathematically expressed as the root mean square error (RMSE) between the GCN model's predicted values and the actual observed data:

$$\text{Score}(\tilde{X}_{t+W}, X_{t+W}) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left\|\tilde{X}_{t+W}^{(m)} - X_{t+W}^{(m)}\right\|^2}, \tag{7}$$

where $\tilde{X}_{t+W}$ denotes the matrix of predicted values for the forthcoming W timestamps, X_{t+W} represents the matrix of actual observed values, and M signifies the number of UAV nodes within the IoDT framework.

A pronounced anomaly score suggests a significant deviation in the model's predictions from the observed values, thereby flagging potential anomalies. For effective anomaly detection, a threshold η is established. Exceeding this threshold with the anomaly score indicates an anomaly, prompting further investigation:

$$\text{if Score}(\tilde{X}_{t+W}, X_{t+W}) > \eta, \text{ then flag as anomaly.} \tag{8}$$

The threshold η is determined from the distribution of anomaly scores on a validation dataset, typically set at a level to capture the upper quantile of the score distribution to ensure that only significant deviations are flagged as anomalies. This approach allows for a balanced sensitivity and specificity in anomaly detection, providing a robust framework for identifying genuine anomalies within normal operating conditions.

The anomaly detection process unfolds in a structured manner, beginning with the computation of anomaly scores using the predicted and actual values. These scores are then compared against the η threshold to flag timestamps where the scores indicate anomalies. Subsequent steps involve a thorough investigation of flagged timestamps to confirm anomalies and implement appropriate measures.

E. Formulation of Adversarial Attack Strategy

In this section, we unveil our innovative adversarial attack strategy, named Label-flipping Adversarial Framework for Anomaly Detection in Graph-based Networks (UAVGuard), devised for Graph Convolutional Networks (GCNs) in spatial crowdsourcing UAV services, building on the concepts discussed in [40]. UAVGuard's main aim is to conduct covert label-flipping attacks that are both effective and hard to detect. We first describe the attack goal, which presents an intricate, non-differentiable bi-level optimization problem, rendering conventional gradient-based techniques useless. To counter this complexity, we adopt a dual-strategy approach:

1) **GCN Approximation**: We present an approximate closed-form solution for the GCN. This is essential

for breaking down the difficult bi-level optimization problem into a single-level optimization problem that is easier to handle.

2) **Continuous Surrogate Design**: We construct continuous surrogate elements for the non-differentiable components of the attack objective. These stand-ins are designed to enable the use of gradient-based optimization while preserving the fundamental functionality of the original components.

With these strategies, the optimization of the attack objective becomes viable using gradient-based techniques.

1) Defining the Attack Objective: Aligned with existing research [14]–[16], our study focuses on binary classification in spatial crowdsourcing UAV networks. Node labels are represented as $\mathbf{y} \in \{-1, +1\}^N$. The labels for unlabeled nodes (V_U) and labeled nodes (V_L) are given as $\mathbf{y}_U \in \{-1, +1\}^{N_U}$ and $\mathbf{y}_L \in \{-1, +1\}^{N_L}$, respectively. Our attack strategy discreetly alters these labels to induce misclassifications, challenging the UAV network's anomaly detection capabilities.

Optimizing parameters for a two-layer Graph Convolutional Network (GCN) in spatial crowdsourcing UAV services is crucial. This is done semi-supervisedly, aiming to minimize classification loss for the labeled nodes (V_L). The loss function is defined as:

$$L\left(\boldsymbol{\theta}; \mathbf{A}, \mathbf{X}, \mathbf{y}_L\right) = \sum_{i}^{N_L} \ell\left(f_{\boldsymbol{\theta}}(\mathbf{A}, \mathbf{X})_L[i], \mathbf{y}_L[i]\right) + \lambda \|\boldsymbol{\theta}\|_2^2, \tag{9}$$

where the point-wise loss function $\ell(\cdot, \cdot)$ is specific to node classification. The GCN's output, $f_\theta(\mathbf{A}, \mathbf{X})$, ranges within $[-1, +1]^N$, reflecting the binary classification aspect of our anomaly detection framework.

The GCN model's effectiveness is judged by its test accuracy, measured as the 0-1 test error:

$$L_{0-1}\left(\boldsymbol{\theta}; \mathbf{A}, \mathbf{X}, \mathbf{y}_U\right) = \frac{1}{N_U} \sum_{i}^{N_U} \mathbb{I}\left[\operatorname{sign}\left(f_{\boldsymbol{\theta}}(\mathbf{A}, \mathbf{X})_U\right)[i] \neq \mathbf{y}_U[i]\right], \tag{10}$$

$\mathbb{I}[\cdot]$ is the non-differentiable indicator function, and $\operatorname{sign}(\cdot)$ is the Heaviside step function. This step function is essential for categorizing nodes' predictions into discrete classes, a critical component of anomaly detection in UAV networks.

Baseline Models: The selection of baseline models is crucial for a robust comparative context. We benchmark UAVGuard against several established models:

RGCN (Robust Graph Convolutional Network): RGCN is selected due to its design to counteract adversarial topology and feature attacks using Gaussian distributions for node hidden representations and a variance-based attention mechanism. This model represents state-of-the-art robustness strategies against adversarial perturbations in graph data.

Sanitation: This method employs k-Nearest Neighbors (k-NN) to identify and correct potentially tainted training labels, directly addressing the challenge of label noise.

To enrich the comparative context, we include additional baseline models:

GCN with Data Augmentation: This model uses data augmentation techniques to enhance GCN robustness by introducing synthetic variations of the training data, improving the model's ability to generalize and resist adversarial attacks.

Graph Attention Network (GAT): GAT is included to evaluate how attention mechanisms in GNNs impact resilience to label-flipping attacks, leveraging its ability to weigh the importance of different node features and edges.

Graph Isomorphism Network (GIN): GIN, known for its powerful representation capabilities, often outperforms other GNNs in various tasks. Including GIN helps assess the robustness of our attack strategy against a highly expressive GNN model.

By comparing UAVGuard with these diverse baseline models, we can thoroughly evaluate its performance and highlight its strengths. This comprehensive comparison ensures that our findings are robust and applicable to a wide range of scenarios in spatial crowdsourcing UAV networks.

Evaluation Metrics: The choice of evaluation metrics is crucial for assessing the performance and robustness of our models. We employ accuracy, robustness score, and anomaly detection consistency. Accuracy measures overall predictive capability, robustness score evaluates resilience to label-flipping attacks, and anomaly detection consistency assesses the model's ability to reliably identify anomalies under adversarial conditions. These metrics provide a comprehensive evaluation framework that captures both effectiveness and robustness against adversarial threats.

F. Objectives of the Label-Flipping Adversarial Attack

The UAVGuard strategy in spatial crowdsourcing UAV networks aims to discreetly modify a subset of training labels $\mathbf{y}_L$. This manipulation entails flipping certain labels to significantly reduce the test accuracy of a retrained Graph Convolutional Network (GCN). Label flipping is done by $(\delta \odot \mathbf{y}_L)$, where the flipping vector is $\delta \in \{+1, -1\}^{N_L}$, the Hadamard product is represented by $\odot$, the flipping label $\mathbf{y}_L[i]$ is indicated by $\delta[i] = -1$, and no flip is implied by $\delta[i] = 1$. UAVGuard's formal objective is expressed as:

$$\begin{aligned} &\min_{\boldsymbol{\delta}} -L_{0-1}\left(\boldsymbol{\theta}^*; \mathbf{A}, \mathbf{X}, \mathbf{y}_U\right) \\ &\text{s.t.} \quad \boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} L\left(\boldsymbol{\theta}; \mathbf{A}, \mathbf{X}, \boldsymbol{\delta} \odot \mathbf{y}_L\right), \quad \|\boldsymbol{\delta} - \mathbf{1}\|_0 \leq \epsilon N_L, \end{aligned} \tag{11}$$

Here, ϵ denotes a small flipping ratio, essential for the attack's subtlety and undetectability. The $L_{0-1}\left(\boldsymbol{\theta}^*; \mathbf{A}, \mathbf{X}, \mathbf{y}_U\right)$ outer optimization on the $0-1$ test error is inherently bi-level since the inner optimization on classification loss $L\left(\boldsymbol{\theta}; \mathbf{A}, \mathbf{X}, \boldsymbol{\delta} \odot \mathbf{y}_L\right)$ serves as a limitation.

Note that θ^* are the GCN's 'optimal' parameters for a given δ.

The complexity of this attack objective stems from its non-differentiable nature, typical of bi-level optimization problems. As noted before, this bi-level optimization is quite difficult in the absence of a closed form for the inner model. The complexity of the optimization process is increased by the addition of non-differentiable components such as the flipping vector

δ, the Heaviside step function $\mathrm{sign}(\cdot)$, and the characteristic function $\mathbb{I}[\cdot]$.

1) Approximation and Simplification of GCN for Anomaly Detection: In the field of spatial crowdsourcing UAV services, tackling the bi-level optimization challenge in GCN-based anomaly detection requires approximating the GCN to a more manageable form. This transformation into a single-level problem is achieved through two primary steps:

1) **GCN Linearization:** The complexity of the Graph Convolutional Network (GCN) can be a hindrance in deriving a closed form. However, Wu et al. [25] suggest that this complexity might be unnecessary for specific applications, like anomaly detection in UAV networks. To simplify the model while retaining the essence of graph convolutions, we linearize the GCN as follows:

$$\mathrm{Softmax}\left(\hat{\mathbf{A}}\left(\hat{\mathbf{A}}\mathbf{X}\mathbf{W}^{(1)}\right)\mathbf{W}^{(2)}\right) = \mathrm{Softmax}\left(\hat{\mathbf{A}}^2\mathbf{X}\mathbf{W}\right), \tag{12}$$

In this instance, integrating weights $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ into a single matrix reduces the complexity of the model. This leads to $\boldsymbol{\theta} = \mathbf{W}$, where $\boldsymbol{\theta} \in \mathbb{R}^d$.

2) **Simplification of Inner Classification Loss:** To enable the use of an ordinary least squares (OLS) estimator in its closed form, we propose replacing the inner classification loss with a regression loss. This change aligns with the OLS's strengths and simplifies the optimization within the GCN framework.

These modifications effectively streamline the GCN model, making it more suitable for anomaly detection tasks in spatial crowdsourcing UAV services. This approach addresses adversarial label-flipping attacks, ensuring robust and efficient anomaly detection.

The linearization of the Graph Convolutional Network (GCN) as outlined in Eq. (12) simplifies the model for spatial crowdsourcing UAV services, converting the GCN into a basic feature propagation step ($\hat{\mathbf{A}}^2\mathbf{X}$) followed by logistic regression. However, logistic regression has limitations in terms of closed-form efficiency, particularly with limited samples [41]. To address this, we propose shifting to a linear regression model, benefiting from the Ordinary Least Squares (OLS) estimator's closed-form existence. We do this by substituting a regression loss (squared loss) for the inner classification loss ℓ. This yields the GCN approximation (also called GCN (appr)) that is shown below:

$$\boldsymbol{\theta}^* = \frac{1}{N_L}\arg\min_{\boldsymbol{\theta}} \left\| \left(\hat{\mathbf{A}}^2\mathbf{X}\boldsymbol{\theta}\right)_L - \mathbf{y}_L \right\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2. \tag{13}$$

Here, θ^* denotes the optimal parameter set for GCN (appr). The OLS estimator is applied to derive the closed-form of θ^* as follows:

$$\begin{aligned}\boldsymbol{\theta}^* &= \left(\left(\hat{\mathbf{A}}^2\mathbf{X}\right)_L^T\left(\hat{\mathbf{A}}^2\mathbf{X}\right)_L + \lambda\mathbf{I}\right)^{-1}\left(\hat{\mathbf{A}}^2\mathbf{X}\right)_L^T(\boldsymbol{\delta}\odot\mathbf{y}_L)\\ &= \mathbf{P}\left(\boldsymbol{\delta}\odot\mathbf{y}_L\right),\end{aligned} \tag{14}$$

where $\mathbf{P}$ can be precomputed for a given graph structure. This simplifies parameter optimization in the GCN, making it more feasible for anomaly detection under adversarial conditions like label-flipping attacks.

Empirical results show that our GCN (appr) matches the standard GCN's performance across various label-flipping ratios, crucial for robustness in spatial crowdsourcing UAV services. We effectively reduce the complex bi-level problem to a single-level problem using the approximated GCN, which improves robustness against label-flipping attacks. The optimization problem is reformulated as:

$$\begin{aligned}&\min_{\boldsymbol{\delta}} -L_{0-1}\left(\mathbf{P}\left(\boldsymbol{\delta}\odot\mathbf{y}_L\right);\mathbf{A},\mathbf{X},\mathbf{y}_U\right)\\ &\text{s.t.}\quad \|\boldsymbol{\delta}-\mathbf{1}\|_0 \le \epsilon N_L.\end{aligned} \tag{15}$$

Here, the goal is to minimize the negative 0-1 test error, constrained by the sparsity of $\boldsymbol{\delta}$, controlled by ϵ, the maximum proportion of label flips allowed. This transformation into a single-level problem simplifies computation and aligns with the operational needs of UAV networks in spatial crowdsourcing, where efficiency is key.

2) Refinement of Non-Differentiable Elements for Enhanced Model Robustness: For spatial crowdsourcing UAV services, ensuring the robustness of anomaly detection models against adversarial attacks is critical. The challenge with optimizing the attack objective in Eq. (15) lies in its non-differentiable elements. We offer continuous surrogates for these components to strengthen the model's defense against hostile label-flipping attacks.

Smooth Approximation of the Heaviside Step Function: The conventional 0-1 error, L_{0-1}, is inherently non-differentiable. To circumvent this, we substitute a smooth approximation $\tilde{h}(x) = \tanh(\tau x)$ for the Heaviside step function $\mathrm{sign}(\cdot)$. The parameter τ acts as a smoothness coefficient, allowing for fine-tuning of the approximation's accuracy. A higher τ value leads to a closer resemblance to the original step function, thereby enhancing the flexibility and robustness of the model. An in-depth investigation of the influence of τ reveals that as τ increases, the model's sensitivity to small perturbations decreases, which helps in mitigating the impact of adversarial attacks. However, if τ is set too high, the approximation may lose the benefits of the original step function, thus finding a balance is crucial.

Approximated Predictions of Test Nodes: The approximated predictions for test nodes, $\tilde{\mathbf{y}}_U$, are obtained using label flips $\delta\odot\mathbf{y}_L$ as follows:

$$\tilde{\mathbf{y}}_U = \tilde{h}\left(\hat{\mathbf{A}}^2\mathbf{X}\boldsymbol{\theta}^*\left(\delta\odot\mathbf{y}_L\right)\right)_U, \tag{16}$$

where θ^* is derived from Eq. (14). This step is critical in preserving the model's prediction accuracy amid adversarial manipulation.

Continuous Differentiable 0-1 Error: Using the approximated test node predictions, $\tilde{\mathbf{y}}_U$, in the original 0-1 error formula (Eq. (10)), results in a continuous and differentiable version of the 0-1 error:

$$L_{0-1}\left(\boldsymbol{\theta}^{*}; \mathbf{A}, \mathbf{X}, \mathbf{y}_{U}\right) \stackrel{\text{def}}{=} \frac{1}{N_{U}} \sum_{i}^{N_{U}} \left(\tilde{\mathbf{y}}_{U}[i] \odot \mathbf{y}_{U}[i]\right). \quad (17)$$

This reformulation significantly enhances the model's ability to detect and counteract adversarial label-flipping attacks, safeguarding the integrity of anomaly detection.

In defending against adversarial attacks, optimizing the label-flipping strategy is essential. The discrete nature of the label-flipping operation vector δ poses a challenge. To mitigate this, we adopt a probabilistic approach to model the flipping decisions.

Modeling Label-Flipping as Bernoulli Random Variables: The components of δ are represented as mutually independent Bernoulli random variables α, as per [14]. The probability of the i-th node flipping is $P(\delta[i] = -1) = \alpha[i]$. This probabilistic form makes label flipping more adaptable.

Sampling Strategy and Reparameterization Trick: From $\mathcal{B}(1, \alpha)$, the flipping vector δ is obtained, where $\delta = 2\mathrm{p} - 1$ and $\mathrm{p} \sim \mathcal{B}(1, \alpha)$. This approach takes advantage of the "reparameterization trick" [42] to simplify the optimization of the objective function when the parameters are stochastic.

Differentiable Attack Objective: These adjustments transform the attack objective in Eq. (11) into a differentiable and manageable form:

$$\mathcal{L}(\alpha) := - \mathop{\mathbb{E}}_{\substack{\mathbf{p} \sim \mathcal{B}(1, \boldsymbol{\alpha}) \\ \delta = 2\mathbf{p} - 1}} \left[\frac{1}{N_{U}} \sum_{i}^{N_{U}} \tilde{\mathbf{y}}_{U}[i] \odot \mathbf{y}_{U}[i] \right]. \quad (18)$$

This reformulation enables the application of gradient-based optimizers to minimize the objective function, determining the optimal 'flipping' probability α. By systematically varying α and analyzing the resulting model performance, we can fine-tune the parameters to achieve an optimal balance between robustness and flexibility. This approach is crucial for improving the resilience of anomaly detection systems against sophisticated adversarial label-flipping attacks in spatial crowdsourcing UAV services.

3) Optimization Strategy for Enhanced Defense: In spatial crowdsourcing UAV services, optimizing defense mechanisms against adversarial label-flipping attacks is imperative. Our model, UAVGuard, detailed in Algorithm 1, is designed to bolster the resilience of anomaly detection systems in UAV services against such attacks.

Algorithm Overview: UAVGuard begins by preprocessing the matrix $\mathbf{P}$ (Eq. (14)), which requires solving an inverse matrix with a time complexity of $O\left(d^{3}\right)$. This is computationally feasible given the typically small feature dimension d. The essence of the algorithm is the iterative training of the probability vector α, optimized using stochastic gradient descent. The final step implements a strategic label-flipping based on the largest ϵN_L elements in α, following the ϵ-greedy strategy [14].

Efficiency and Performance: UAVGuard's efficiency is highlighted by its average time for generating perturbations across datasets like Polblogs, Cora, Citeseer, and Pubmed, typically under 50 seconds. This rapid performance is crucial for real-time anomaly detection in spatial crowdsourcing UAV services.

Algorithm 1 UAVGuard: Enhanced Defense against Label-flipping Attack for Anomaly Detection in UAV Services

Input: Graph consisting of $y_L, \mathrm{A}, \mathrm{X}$, and y_U (alternatively, GCN predictions $\hat{y}_U$), flipping ratio ϵ, smoothing factor τ, and number of iterations T.
Output: The modified training labels y'_L for robust anomaly detection.
1: $y'_L = y_L$
2: Pre-process $\mathbf{P}$ in Eq. (14).
3: **for** each t within T **do**
4: Generate sample flips with $p \sim \mathcal{B}(1, \alpha)$ and set $\delta = 2p - 1$;
5: Compute $\mathcal{L}(\alpha)$ as per Eq. (18);
6: Update α by $\frac{\partial \mathcal{L}(\alpha)}{\partial \alpha}$;
7: **end for**
8: Modify the labels in y'_L by choosing the highest ϵN_L values in α for optimal defense.

G. Enhancing GCN Robustness through Community Structure Integration

Graph Convolutional Networks (GCNs) face significant vulnerability to label-flipping attacks in spatial crowdsourcing UAV services, especially in anomaly detection tasks. We explore the learning dynamics of GCNs under such adversarial conditions and observe that while training accuracy remains stable, the discrepancy between training and validation accuracies increases. This trend indicates a tendency for GCNs to overfit on manipulated labels.

To counteract this overfitting issue, we propose leveraging the inherent community structures present in graph data as a form of signal regularization. Spatial interaction graphs in UAV networks often exhibit distinct community structures, like clusters of UAVs operating within similar geographical areas or sharing task characteristics. These community structures can provide valuable insights for node classification tasks, helping guide the classification process in GCNs.

Integrating community-based signals into the GCN framework, we aim to improve the model's resilience against label-flipping attacks and ensure robust anomaly detection in spatial crowdsourcing UAV services. Our defense strategy comprises two main stages:

First, community label derivation involves utilizing graph embedding techniques such as DeepWalk and adversarially regularized graph autoencoders to transform nodes into a lower-dimensional space, capturing the inherent structural properties of the UAV network. Clustering algorithms like K-means are then applied to these embeddings, forming distinct groups. These clusters provide pseudo community labels $\mathbf{Y}^{c}$ for the nodes. The selection of these methods is justified by their proven effectiveness in capturing community structures and their scalability to large graph datasets, which is crucial for UAV networks.

Second, community-aware GCN training incorporates a self-supervised task into the GCN training process using the derived community labels. This task involves predicting a K-class softmax distribution corresponding to the community labels using the features from the penultimate layer of the GCN, $\mathbf{H}^{(L-1)}$. This auxiliary task enhances community-aware learning and improves the GCN's resistance to label-flipping attacks. The specific features that provide the proposed defense mechanism with its unique capacity to mitigate label-flipping attacks include:

1) **Community-Based Regularization:** By using community detection, we introduce a regularization mechanism that leverages the natural community structure within UAV networks. This helps in mitigating the overfitting of the model to manipulated labels by providing additional, reliable training signals that reflect the true underlying structure of the data.
2) **Multi-Task Learning:** The inclusion of a self-supervised task that predicts community labels alongside the main classification task helps to regularize the training process. This multi-task learning approach distributes the learning objective across multiple related tasks, reducing the risk of the model focusing too narrowly on potentially corrupted labels.

The community-based softmax output is represented by:

$$L_c\left(\boldsymbol{\theta}^{(L-1)}, \mathbf{W}^c; \mathbf{A}, \mathbf{X}, \mathbf{Y}^c\right) = -\sum_{v \in V_L^c} \sum_{i=1}^{K} \mathbf{Y}_{v,i}^c \log \mathbf{Z}_{v,i}^c, \tag{19}$$

where $\mathbf{Z}_{v,i}^c = \operatorname{Softmax}\left(\hat{\mathbf{A}}\mathbf{H}^{(L-1)}\mathbf{W}^c\right)$. Separate task-specific output layers $\mathbf{W}^{(L)}$ and $\mathbf{W}^c$ are used, and the lower levels $\boldsymbol{\theta}^{(L-1)}$ share parameters with both the major and auxiliary tasks. This architecture penalizes overfitting to misleading signals in flipped nodes by using community-level information as a regularizer. The comprehensive loss function L_{MT} for the training process is defined as:

$$\begin{aligned} L_{MT}\left(\mathbf{A}, \mathbf{X}, \mathbf{Y}, \mathbf{Y}^c\right) &= L\left(\boldsymbol{\theta}^{(L-1)}, \mathbf{W}^{(L)}; \mathbf{A}, \mathbf{X}, \mathbf{Y}\right) \\ &+ \lambda_c L_c\left(\boldsymbol{\theta}^{(L-1)}, \mathbf{W}^c; \mathbf{A}, \mathbf{X}, \mathbf{Y}^c\right). \end{aligned} \tag{20}$$

Our entire loss returns to the original GCN loss function when $\lambda_c = 0$. We set $\lambda_c = 1$ by default for this investigation. Note that the self-supervised part of the loss does not require the ground truth training labels $\mathbf{Y}$ as input.

Thus, the integration of community structures into the GCN framework not only mitigates the risk of overfitting due to additional training tasks but also enhances the robustness of the model against label-flipping attacks. This approach leverages the inherent structural properties of UAV networks, providing a robust defense mechanism that is both practical and effective.

IV. Analysis and Discussion

In spatial crowdsourcing UAV services, the robustness of Graph Convolutional Networks (GCNs) against adversarial label-flipping attacks is a critical concern. To thoroughly evaluate our proposed attack and defense strategies, we utilize the Intel Berkeley Research Lab Wireless Sensor Network (WSN) dataset [43], which offers a diverse range of real-world sensor data, making it ideal for this study.

A. Experimental Setup

Our experimental framework adheres to established standards from prior research [13], [44], using the comprehensive IBRL WSN dataset, which includes sensor readings collected from 54 sensors between February 28th and April 5th, 2004. The dataset spans three key modalities:

1) **Temperature**: Sensor readings reflecting ambient temperature, essential for environmental analysis.
2) **Humidity**: Measurements of air moisture content, crucial for atmospheric condition-sensitive applications.
3) **Light**: Readings on light intensity, important for applications like daylight harvesting and energy management.

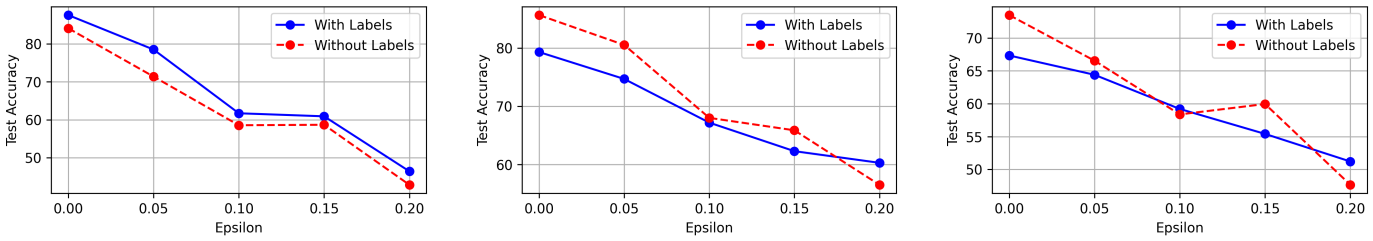
Each modality is treated as a distinct dataset in our experiments, allowing for a comprehensive assessment of GCNs' robustness against label-flipping attacks across various types of sensor data.

We begin with data preprocessing, handling missing values and normalizing sensor readings. The dataset is then divided into training, validation, and test sets. The training set is used to train the GCN model with a fixed number of labeled data points. Hyperparameters such as learning rate, number of epochs, and dropout rates are tuned using the validation set to optimize performance and prevent overfitting. Our model evaluation employs several metrics: accuracy to measure overall predictive performance, robustness score to assess resilience to label-flipping attacks, and anomaly detection consistency to evaluate the model's capability to detect anomalies under adversarial conditions. We compare our methods against various baselines, including traditional GCNs without defense and state-of-the-art defense strategies in graph-based learning for sensor networks.

Despite the IBRL WSN dataset being relatively old, its comprehensive and varied sensor data make it a valuable resource for evaluating our methods. To ensure the scalability and generalizability of our results, we also discuss the applicability of our methods to more recent and larger datasets. Our approach is designed to adapt to different scales and types of data, crucial for real-world applications in UAV networks and beyond. Our models are not confined to GCNs; they can be generalized to other GNN architectures such as Graph Attention Networks (GATs) and Graph Isomorphism Networks (GINs). This adaptability ensures that our models can be effectively employed in various graph-based learning scenarios, providing robust defense mechanisms against adversarial attacks across different applications.

B. Attack in Spatial Crowdsourcing UAV Services

This section evaluates the effectiveness and transferability of UAVGuard in spatial crowdsourcing UAV services, focusing on anomaly detection with the IBRL dataset. We consider scenarios where the attacker has access to true test labels and scenarios where such labels are unknown.



(a) Temperature. (b) Humidity. (c) Light.

Fig. 3. Performance of UAVGuard attack model without access to true test labels across temperature, humidity, and light modalities. This demonstrates the robustness of the attack strategy using model predictions instead of actual labels, underscoring the need for strong defense mechanisms in UAV networks.

Effectiveness of attack model. We assess UAVGuard's effectiveness on the temperature, humidity, and light modalities of the IBRL dataset, varying the flipped nodes ratio $\epsilon \in [0.05, 0.1, 0.15, 0.2]$. We use five random splits for each modality, running the model five times per split and averaging results over 25 runs. The attacks target semi-supervised binary classification tasks in anomaly detection.

UAVGuard is benchmarked against these baselines:

1) Random-flipping attacks (**Rnd**), flipping labels randomly.
2) Degree-based flipping attacks (**Deg**), targeting highly connected nodes.
3) Label-flipping attacks against label propagation (**LPattack**) [14].

Two variants of UAVGuard are also evaluated:

1) **UAVGuard**$_{\mathrm{LP}}$: Uses label propagation for approximating predictions of unlabeled data $\tilde{\mathbf{y}}_h$.
2) **UAVGuard**$_{\mathrm{MSE}}$: Uses mean squared error instead of 0-1 error.

Each model's attack loss function is used to optimize it.

Test accuracy for the IBRL dataset modalities under various flip ratios ϵ is shown in Table I. Key observations include:

- UAVGuard significantly outperforms Random-flipping (Rnd) across all metrics, indicating that UAV service anomaly detection models are vulnerable to sophisticated adversarial label-flipping attacks. As ϵ increases, performance under UAVGuard consistently worsens, demonstrating its ability to exploit learning algorithm weaknesses.
- UAVGuard shows superior performance compared to other label-flipping strategies, underlining its tailored effectiveness for GCNs, unlike Degree-based (Deg) and LPattack methods.
- Comparing UAVGuard$_{\mathrm{LP}}$ and UAVGuard$_{\mathrm{MSE}}$ with the standard UAVGuard, it's clear that the latter outperforms both. This demonstrates how well UAVGuard's continuous surrogates and approximated closed version of GCN work for spatial crowdsourcing UAV services in anomaly detection.

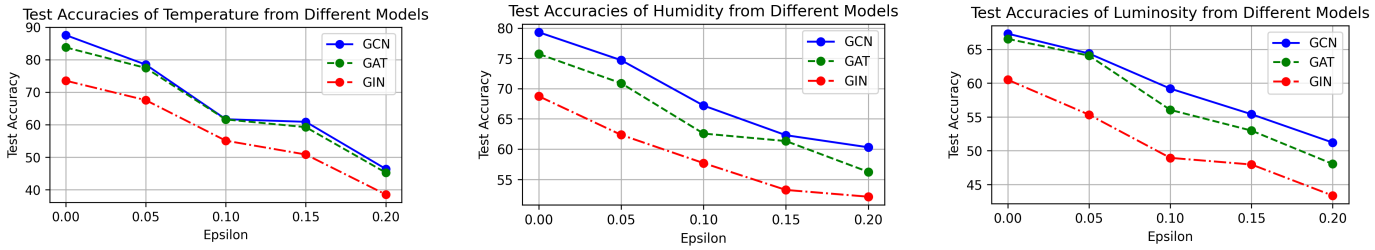
Our study reveals that sophisticated adversarial attacks, like UAVGuard, can significantly impact the performance of anomaly detection models in spatial crowdsourcing UAV services. These findings underscore the need for robust and adaptive defense mechanisms in such applications.

Transferability of attacks. A key consideration in practical attack scenarios is the generalizability of attacks across different models, particularly when attackers are unaware of the specific model in use. We examine the transferability of our UAVGuard model in spatial crowdsourcing UAV services for anomaly detection. We apply perturbations optimized for a GCN model to other graph neural networks, including Graph Isomorphism Networks (GIN) [45] and Graph Attention Networks (GAT) [46], using default hyper-parameter settings. Results in Fig. 4 indicate that UAVGuard's perturbations are not only transferable but occasionally even more effective on various models within the IBRL dataset. This highlights UAVGuard's versatility and potential threat across different model architectures, emphasizing the need for robust defenses in anomaly detection within spatial crowdsourcing UAV services.

Analysis of the number of labeled nodes per class. To assess the performance of the UAVGuard model in hostile situation, we examine the impact of increasing the number of labeled nodes per class (L/C) from 20 to 80. Figure 5 illustrates these findings. The figure shows that for each IBRL dataset, the performance trends for different L/C values follow a similar pattern, with a decline in performance as the flip ratio ϵ increases. Higher L/C values result in better performance, both on clean data and under adversarial conditions. This trend can be attributed to:

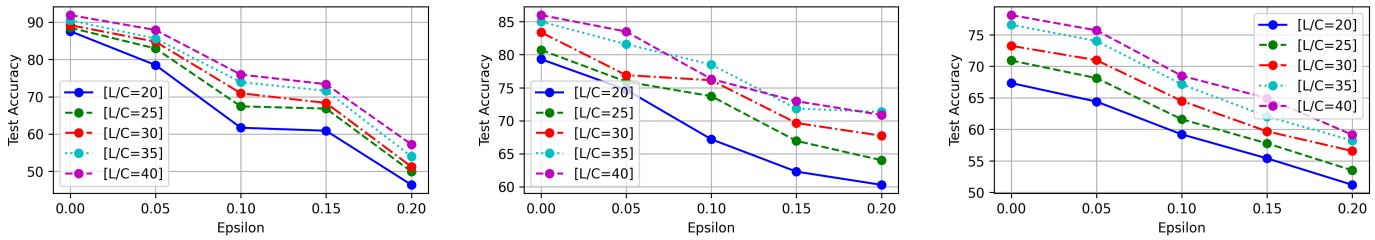
1) Enhanced propagation of label information with more labeled nodes, leading to improved classification confidence and reduced model vulnerability.
2) Greater resistance to overfitting to flipped labels due to a larger training dataset.

Evaluating the Attack Model Without True Test Labels. In this section, a realistic scenario—common in applications like as spatial crowdsourcing UAV services—where attackers do not have access to genuine test labels $\mathbf{y}_U$ is examined. Alternatively, attackers replace this with GCN model predictions on test nodes $\hat{\mathbf{y}}_U$. To illustrate this, we utilize the temperature, humidity, and light modalities from the IBRL dataset. Fig. 3 shows that the absence of true test labels does not significantly affect the effectiveness of UAVGuard. This is relevant for spatial crowdsourcing UAV services, where ground-truth labels may not be available or are hidden for security reasons. The experiment shows that attackers can use estimated labels $\hat{\mathbf{y}}_U$ to effectively compromise GCN models. This underscores the



(a) Temperature. (b) Humidity. (c) Light.

Fig. 4. UAVGuard attack model's transferability



(a) Temperature. (b) Humidity. (c) Light.

Fig. 5. Analyzing the label rate for our attack model UAVGuard. L/C represents the number of labeled nodes per class for the training set.

need for robust defense mechanisms in spatial crowdsourcing UAV services, highlighting the importance of strategies that are inherently resilient to adversarial manipulations, beyond just the secrecy of test labels.

C. Defense Evaluation in Spatial Crowdsourcing UAV Services

We evaluate our defense framework tailored for spatial crowdsourcing UAV services using the IBRL dataset, focusing on temperature, humidity, and light modalities. Our community-preserving self-supervised defense framework is benchmarked against:

1) **RGCN**: A robust Graph Convolutional Network model counteracting adversarial topology and feature attacks, using Gaussian distributions for node hidden representations and a variance-based attention mechanism.
2) **Sanitation**: A defense method against label-flipping attacks that makes use of k-Nearest Neighbors (k-NN) to locate and fix possibly tainted training labels.

Both baselines are initialized according to their original publications, with parameter tuning for optimized performance. Our proposed framework uses the same hyper-parameters as the standard GCN model [13]. The evaluation aims to assess the effectiveness and generalizability of our defense mechanism in spatial crowdsourcing UAV services under diverse environmental data conditions.

The choice of evaluation metrics in our experiments is crucial for comprehensively assessing the performance and robustness of our proposed models. We employ several key metrics, including accuracy, robustness score, and anomaly detection consistency. Accuracy is a fundamental metric that measures the overall predictive capability of the model, providing a clear indication of its performance on the labeled data. The robustness score is specifically designed to evaluate the model's resilience to label-flipping attacks, which is central to our study's focus. This metric helps in understanding how well the model can withstand and function under adversarial conditions. Anomaly detection consistency is critical in applications involving UAV services, as it assesses the model's ability to reliably identify anomalies even when faced with adversarial perturbations. This metric ensures that the model maintains its operational effectiveness and reliability in real-world scenarios where adversarial attacks are prevalent. Together, these metrics provide a comprehensive evaluation framework that captures both the effectiveness and robustness of our models against adversarial threats.

For a thorough evaluation, each defense method is executed five times across different training/validation/testing splits, with the average results reported. This process aims to demonstrate the resilience of our defense framework in spatial crowdsourcing UAV services, especially for critical applications like environmental monitoring anomaly detection.

Effectiveness of Defense Framework in UAV Services Context. We use the IBRL dataset to assess our proposed defense system against label-flipping assaults, with particular attention to the modalities of light, humidity, and temperature. Table II summarizes the results, and our key observations include:

- Our defense framework substantially improves the robustness of the GCN model across all modalities. This highlights the effectiveness of the community-preserving self-supervised task in counteracting misinformation from flipped nodes in spatial crowdsourcing scenarios.
- Traditional methods, such as RGCN and Sanitization, show less efficacy in this context. RGCN, designed for feature/topology attacks, fails to address label-flipping

TABLE I
ATTACK MODEL RESULTS UNDER VARIOUS ϵ FLIP RATIOS. BETTER PERFORMANCE IS INDICATED BY A LOWER TEST ACCURACY (%). IN RELATION TO THE CLEAN MODEL, WE ALSO PROVIDE THE PERFORMANCE DECREASE RATE. BETTER ATTACK PERFORMANCE IS INDICATED BY A HIGHER DROP RATE.

Modality	ϵ	Rnd	Deg	LPattack	UAVGuard$_{LP}$	UAVGuard$_{\text{MSE}}$	UAVGuard
Temperature	0.05	84.7(−2.16%)	84.8(−2.04%)	81.2(−7.14%)	79.5(−8.56%)	80.8(−7.08%)	78.9(−9.33%)
	0.10	84.8(−2.04%)	81.3(−6.62%)	68.5(−20.47%)	67.0(−21.85%)	73.2(−13.66%)	62.4(−28.28%)
	0.15	84.6(−2.39%)	73.8(−15.95%)	60.2(−28.47%)	60.6(−27.94%)	65.4(−22.81%)	61.7(−27.94%)
	0.20	82.1(−5.24%)	57.8(−31.38%)	52.3(−35.62%)	52.5(−35.39%)	49.4(−39.92%)	47.2(−41.49%)
Humidity	0.05	78.3(−1.01%)	78.2(−1.14%)	75.5(−4.64%)	75.0(−5.27%)	74.2(−6.28%)	74.7(−5.65%)
	0.10	78.3(−1.01%)	76.3(−3.89%)	70.7(−10.76%)	69.0(−12.90%)	67.8(−14.42%)	67.2(−15.17%)
	0.15	75.9(−3.51%)	73.5(−6.93%)	65.7(−15.24%)	65.9(−14.98%)	62.5(−19.27%)	62.3(−19.52%)
	0.20	75.1(−4.52%)	71.4(−7.59%)	62.7(−18.91%)	61.9(−19.92%)	60.3(−21.94%)	60.3(−21.94%)
Light	0.05	66.4(−1.45%)	65.9(−2.19%)	66.0(−2.04%)	64.9(−3.65%)	65.2(−3.21%)	64.4(−4.39%)
	0.10	66.4(−1.45%)	62.6(−6.98%)	63.0(−6.41%)	62.4(−7.28%)	62.3(−7.42%)	59.2(−11.97%)
	0.15	65.7(−2.48%)	59.4(−11.18%)	59.8(−10.59%)	57.9(−13.38%)	59.3(−11.45%)	55.4(−17.07%)
	0.20	58.3(−12.03%)	58.1(−12.32%)	58.0(−12.47%)	56.7(−14.38%)	54.7(−17.31%)	51.2(−22.44%)

TABLE II
RESULTS OF DEFENSE FRAMEWORK FOR THE IBRL DATASET WITH THREE MODALITIES: TEMPERATURE, HUMIDITY, AND LIGHT. BETTER PERFORMANCE IS INDICATED BY HIGHER TEST ACCURACY (%).

Modality	ϵ	**GCN**	**RGCN**	**Sanitization**	**Ours$_{\text{GCN}}$**
Temperature	0.05	78.7	78.1	78.6	**84.2**
	0.10	61.9	60.8	61.7	**70.1**
	0.15	61.1	61.8	58.7	**63.2**
	0.20	46.6	47.8	46.5	**56.4**
Humidity	0.05	78.8	79.4	78.7	**80.0**
	0.10	73.9	75.2	73.9	**77.6**
	0.15	66.6	67.6	66.5	**72.2**
	0.20	61.4	62.1	61.3	**64.5**
Light	0.05	64.7	64.4	64.6	**69.3**
	0.10	58.9	59.2	58.9	**62.9**
	0.15	56.4	57.5	56.3	**57.7**
	0.20	55.6	56.1	55.5	**57.5**

adequately. Sanitization, dependent on extensive training data, is less effective in the semi-supervised learning environment typical in UAV-based monitoring systems.

- The performance improvement in the light modality dataset is less pronounced, suggesting a weaker community structure in this modality. This indicates the importance of strong community structures in enhancing resilience against adversarial attacks in spatial crowdsourcing applications.

These findings highlight how, in the complex environment of spatial crowdsourcing UAV services, our community-preserving self-supervised defensive architecture may be used to strengthen GCN models' robustness against label-flipping attacks.

V. CONCLUSION AND FUTURE WORK

In this study, we pioneered the investigation into adversarial label-flipping attacks within the domain of Graph Neural Networks (GNNs), particularly focusing on anomaly detection in spatial crowdsourcing UAV services. We developed a unique label-flipping attack model called UAVGuard, which employs a novel method for handling non-differentiable objectives and an approximate closed form of GNNs. This approach adeptly addresses non-differentiability issues and bi-level optimization difficulties. Our thorough analyses using the IBRL dataset, featuring three unique modalities—light, humidity, and temperature—demonstrate that even small changes in training labels can significantly impact GNNs' classification performance. This vulnerability is particularly critical in the context of spatial crowdsourcing for UAV services, where data integrity is paramount. To combat this, we developed a unique defense mechanism based on community-preserving self-supervised learning. By reducing the tendency to overfit manipulated nodes, this technique significantly strengthens GNNs' defense against label-flipping attacks. Our empirical evaluations across various datasets, including the specialized UAV service context, affirm the robustness and effectiveness of our defense strategy.

The practical implications of our work are significant for real-world UAV networks, emphasizing the need for robust defense mechanisms to ensure data integrity and accurate decision-making in critical applications like environmental monitoring, public safety, and traffic management. Implementing UAVGuard and our defense framework can greatly enhance UAV service reliability, mitigating adversarial attack disruptions.

Future research should explore clean-label poisoning attacks, expand our defense framework to other adversarial attacks, and test its applicability across different GNN architectures and real-world scenarios. Enhancing the self-supervised learning framework with advanced community detection algorithms, real-time defense mechanisms, and adaptive learning strategies will be crucial for advancing secure and resilient GNN applications in UAV services and other critical domains.

REFERENCES

[1] M. Umair, R. H. Jhaveri, M. N. Riaz, H. Chi, and S. Malebary, "Chained-drones: Blockchain-based privacy-preserving framework for secure and intelligent service provisioning in internet of drone things," *Computers and Electrical Engineering*, vol. 110, p. 108772, 2023.

[2] R. H. Jhaveri, M. Alazab, and H. Chi, "Bc-iodt: blockchain-based framework for authentication in internet of drone things," in *Proceedings of the 5th International ACM Mobicom Workshop on Drone Assisted Wireless Communications for 5G and Beyond*, pp. 115–120, 2022.

[3] J. Akram, M. Aamir, R. Raut, A. Anaissi, R. H. Jhaveri, and A. Akram, "Ai-generated content-as-a-service in iomt-based smart homes: Personalizing patient care with human digital twins," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 1939–1950, 2024.

[4] H. S. Munawar, Z. Gharineiat, and S. Imran Khan, "A framework for burnt area mapping and evacuation problem using aerial imagery analysis," *Fire*, vol. 5, no. 4, p. 122, 2022.

[5] A. Tahir, A. Akram, A. Z. Kouzani, and M. P. Mahmud, "Cloud-and fog-integrated smart grid model for efficient resource utilisation," *Sensors*, vol. 21, no. 23, p. 7846, 2021.

[6] H. S. Munawar, A. Z. Kouzani, and M. P. Mahmud, "Using adaptive sensors for optimised target coverage in wireless sensor networks," *Sensors*, vol. 22, no. 3, p. 1083, 2022.

[7] H. S. Munawar, J. Akram, S. I. Khan, F. Ullah, and B. J. Choi, "Drone-as-a-service (daas) for covid-19 self-testing kits delivery in smart healthcare setups: A technological perspective," *ICT Express*, vol. 9, no. 4, pp. 748–753, 2023.

[8] L. Sun, Y. Dou, C. Yang, K. Zhang, J. Wang, S. Y. Philip, L. He, and B. Li, "Adversarial attack and defense on graph data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 8, pp. 7693–7711, 2022.

[9] J. Akram and A. Anaissi, "Decentralized pki framework for data integrity in spatial crowdsourcing drone services," in *2024 IEEE International Conference on Web Services (ICWS)*, 2024.

[10] C. Sandeepa, B. Siniarski, S. Wang, and M. Liyanage, "Rec-def: A recommendation-based defence mechanism for privacy preservation in federated learning systems," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2716–2728, 2024.

[11] Y. Liu, K.-F. Tsang, C. K. Wu, Y. Wei, H. Wang, and H. Zhu, "Ieee p2668-compliant multi-layer iot-ddos defense system using deep reinforcement learning," *IEEE Transactions on Consumer Electronics*, vol. 69, no. 1, pp. 49–64, 2023.

[12] S. Shrivastava, S. John, A. Rajesh, and P. K. Bora, "Collision penalty-based defense against collusion attacks in cognitive radio enabled smart devices," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 3963–3976, 2024.

[13] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[14] X. Liu, S. Si, X. Zhu, Y. Li, and C.-J. Hsieh, "A unified framework for data poisoning attack to graph-based semi-supervised learning," *arXiv preprint arXiv:1910.14147*, 2019.

[15] P. W. Koh, J. Steinhardt, and P. Liang, "Stronger Data Poisoning Attacks Break Data Sanitization Defenses," 2021. arXiv:1811.00741 [cs, stat].

[16] A. Paudice, L. Muñoz-González, and E. C. Lupu, "Label sanitization against label flipping poisoning attacks," in *ECML PKDD 2018 Workshops: Nemesis 2018, UrbReas 2018, SoGood 2018, IWAISe 2018, and Green Data Mining 2018, Dublin, Ireland, September 10-14, 2018, Proceedings 18*, pp. 5–15, Springer, 2019.

[17] J. Akram, A. Anaissi, W. Othman, A. Alabdulatif, and A. Akram, "Dronessl: Self-supervised multimodal anomaly detection in internet of drone things," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 4287–4298, 2024.

[18] J. Akram, A. Anaissi, R. S. Rathore, R. H. Jhaveri, and A. Akram, "Galtrust: Generative adverserial learning-based framework for trust management in spatial crowdsourcing drone services," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2285–2296, 2024.

[19] M. Mehmood, N. Javaid, S. H. Abbasi, A. Rahman, and F. Saeed, "Efficient resource distribution in cloud and fog computing," in *Advances in Network-Based Information Systems: The 21st International Conference on Network-Based Information Systems (NBiS-2018)*, pp. 209–221, Springer, 2019.

[20] X. Wu, H. Wang, S. Li, J. Dai, and Z. Ren, "Prior indicator guided anchor learning for multi-view subspace clustering," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 144–154, 2024.

[21] Y. Sun, Z. Ren, Z. Cui, and X. Shen, "Feature weighted multi-view graph clustering," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 401–413, 2024.

[22] Y. Mei, Z. Ren, B. Wu, T. Yang, and Y. Shao, "Multi-view comprehensive graph clustering," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 3279–3288, 2024.

[23] T. K. Behera, S. Bakshi, M. A. Khan, and H. M. Albarakati, "A lightweight multiscale-multiobject deep segmentation architecture for uav-based consumer applications," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 3740–3753, 2024.

[24] J. Akram and A. Anaissi, "Privacy-first crowdsourcing: Blockchain and local differential privacy in crowdsourced drone services," in *2024 IEEE International Conference on Web Services (ICWS)*, 2024.

[25] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger, "Simplifying graph convolutional networks," in *International conference on machine learning*, pp. 6861–6871, PMLR, 2019.

[26] J. Akram, A. Anaissi, R. S. Rathore, R. H. Jhaveri, and A. Akram, "Digital twin-driven trust management in open ran-based spatial crowdsourcing drone services," *IEEE Transactions on Green Communications and Networking*, vol. 8, no. 2, pp. 841–855, 2024.

[27] M. Ali, P. Scandurra, F. Moretti, and H. H. R. Sherazi, "Anomaly detection in public street lighting data using unsupervised clustering," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 4524–4535, 2024.

[28] H. Zhao, M. Liu, S. Qiu, and X. Cao, "Satellite unsupervised anomaly detection based on deconvolution-reconstructed temporal convolutional autoencoder," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2989–2998, 2024.

[29] C. Xiao, X. Xu, Y. Lei, K. Zhang, S. Liu, and F. Zhou, "Counterfactual graph learning for anomaly detection on attributed networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 10, pp. 10540–10553, 2023.

[30] D.-J. Kim, N. G. B. Amma, and V. Sarveshwaran, "A novel split learning-based consumer electronics network traffic anomaly detection framework for smart city environment," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 4197–4204, 2024.

[31] A. Mukherjee, D. De, N. Dey, R. G. Crespo, and E. Herrera-Viedma, "Disastdrone: A disaster aware consumer internet of drone things system in ultra-low latent 6g network," *IEEE Transactions on Consumer Electronics*, vol. 69, no. 1, pp. 38–48, 2023.

[32] J. Li, J. Xie, L. Qian, L. Zhu, S. Tang, F. Wu, Y. Yang, Y. Zhuang, and X. E. Wang, "Compositional temporal grounding with structured variational cross-graph correspondence learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3032–3041, 2022.

[33] J. Li, S. Tang, L. Zhu, W. Zhang, Y. Yang, T.-S. Chua, F. Wu, and Y. Zhuang, "Variational cross-graph reasoning and adaptive structured semantics learning for compositional temporal grounding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12601–12617, 2023.

[34] E. Dai, C. Aggarwal, and S. Wang, "Nrgnn: Learning a label noise resistant graph neural network on sparsely and noisily labeled graphs," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 227–236, 2021.

[35] H. Chang, Y. Rong, T. Xu, W. Huang, H. Zhang, P. Cui, W. Zhu, and J. Huang, "A restricted black-box adversarial framework towards attacking graph embedding models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 3389–3396, 2020.

[36] J. Akram and A. Anaissi, "Ddrm: Distributed drone reputation management for trust and reliability in crowdsourced drone services," in *2024 IEEE International Conference on Web Services (ICWS)*, 2024.

[37] S. K. Dwivedi, M. Abdussami, R. Amin, and M. K. Khan, "D³apts: Design of ecc-based authentication protocol and data storage for tactile internet enabled iod system with blockchain," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 4239–4248, 2024.

[38] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1944–1952, 2017.

[39] R. Taheri, R. Javidan, M. Shojafar, Z. Pooranian, A. Miri, and M. Conti, "On defending against label flipping attacks on malware detection systems," *Neural Computing and Applications*, vol. 32, pp. 14781–14800, 2020.

[40] M. Zhang, L. Hu, C. Shi, and X. Wang, "Adversarial label-flipping attack and defense for graph neural networks," in *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 791–800, 2020.

[41] E. Yang, A. C. Lozano, and P. K. Ravikumar, "Closed-form estimators for high-dimensional generalized linear models," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[42] M. Figurnov, S. Mohamed, and A. Mnih, "Implicit reparameterization gradients," *Advances in neural information processing systems*, vol. 31, 2018.

[43] S. Madden, "Intel Berkeley Research lab." http://db.csail.mit.edu/labdata/labdata.html, 2004.

[44] D. Zhu, Z. Zhang, P. Cui, and W. Zhu, "Robust graph convolutional networks against adversarial attacks," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1399–1407, 2019.

[45] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," *arXiv preprint arXiv:1810.00826*, 2018.

[46] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, *et al.*, "Graph attention networks," *stat*, vol. 1050, no. 20, pp. 10–48550, 2017.