Security of Infocommunication Systems and Internet of Things

2026 Vol 4, No 1

https://doi.org/10.31861/sisiot2026.1.01023

Received 05 June 2026; revised 20 June 2026; accepted 29 June 2026; published 30 June 2026

Limitations of Graph Neural Networks in the Task of Detecting Insider Threats

Ihor Dobrynin* **and Vadym Pantelieiev**

V.V. Popovskyy Department of Infocommunication Engineering, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

*Corresponding author (E-mail: ihor.dobrynin@nure.ua)

ABSTRACT The article investigates a fundamental limitation of graph neural networks in insider threat detection: oversmoothing of node representations, which is most pronounced for privileged users. To make the value of the graph explicit, insiders are assigned not only a weak local signal but also a relational one – an anomalous pattern of links to sensitive resources: feature-only detection yields an area under the ROC curve of about 0.53, whereas one or two aggregation steps raise it to 0.93 – 0.96. Privilege is modelled as a role label that correlates with, but is distinct from, connectivity (Spearman rank correlation 0.43). Verification is performed using a trained graph convolutional network, averaged across stratified splits. As depth increases, detection quality declines; for privileged hubs, the area under the ROC curve falls to the random-guessing level at eight layers and below it by sixteen, whereas for ordinary nodes it remains markedly higher. Two degradation mechanisms are distinguished: label heterophily (early quality loss) and oversmoothing proper (deep collapse). The mitigations are presented and quantitatively compared for both mechanisms: the baseline network is virtually helpless on the privileged segment (AUC 0.42), whereas PairNorm raises this to 0.84, GPR-GNN to 0.88, and H2GCN and GCNII to 0.93 and 0.94, respectively, almost closing the gap between privileged and ordinary nodes; GPR-GNN proves practically depth-invariant. The results are of practical value for designing insider threat detection systems that are resilient to quality degradation within the most high-risk user segment. The robustness of these conclusions to the parameters of graph generation is confirmed by a dedicated analysis that varies the graph size, the number of communities, the mean degree, and the privileged fraction.

KEYWORDS insider threats, graph neural networks, heterophily, privileged users, anomaly detection.

I. INTRODUCTION

The shift from signature analysis to behavioural modelling has changed the very logic of countering internal threats. Whereas an external attacker must breach the perimeter, an insider already resides within the information system, holds legitimate credentials, and acts within granted permissions, so that their activity is indistinguishable from routine staff work. Researchers studying information security risk management tools point directly to the exhaustion of rules oriented to known attack patterns and to the need for behavioural models capable of dynamically assessing risk [1]. The same limitation of the signature-based approach is noted by developers of multi-agent intrusion-detection systems that combine analysis of known threats with machine-learning-based behavioural modelling [2].

Graph neural networks have become a natural tool for tasks in which what matters is not so much individual actions as the relations between subjects and resources. A problem that remains insufficiently addressed is the hidden conflict between the structural role of privileged users and the operating mechanism of deep graph architectures. Administrators and managers, whose compromise causes the greatest damage, become high-degree hubs in the behaviour graph. It is precisely for them that repeated feature aggregation erases the individuality of the representation most quickly, so that the model risks perceiving the behaviour of an insider-administrator as the norm in their own environment. The quantitative substantiation and empirical verification of this effect in a trained model, the separation of its causes, and the experimental evaluation of means to overcome it constitute the scientific problem addressed in this article.

Systematic reviews of graph neural networks for anomaly detection confirm the benefit of accounting for the environment's topology, where feature-based methods reach their limits [3]. For cybersecurity applications, heterogeneous graph architectures that distinguish node and relation types and reconcile the semantics of heterogeneous logs are advantageous [4]. The empirical basis of most works is the CERT Insider Threat dataset: models that combine graph convolutional networks with recurrent components achieve high areas under the ROC curve on versions r5.2 and r6.2, yet also register a drop in recall on more difficult subsets with few insiders [5]. Generalised models for forecasting cybersecurity anomalies based on ensembles of neural components achieve high accuracy mainly at the aggregate level, without guarantees of uniform quality for individual user categories [6]. Related solutions are developed within hybrid defence strategies [7] and comparative studies of neural-network algorithms for network anomaly detection [8, 9].

A separate line comprises works devoted to the fundamental limitation of deep graph networks – oversmoothing. Surveys define it as the degradation of representations in which nodes become almost indistinguishable, and associate it with a sharp quality drop

after only a few layers [10]. A theoretical analysis of graph convolution via energy functionals shows that linear graph convolutions behave as low-pass filters and monotonically reduce the Dirichlet energy [11], while multi-scale energy-enhancing approaches enable the preservation of distinguishability in deeper networks [12]. It has also been shown that energy metrics are reliable only for very deep networks, whereas the rank of the representation matrix is a more robust alternative [13]. The influence of structure on anomaly-detection quality is studied separately: it has been established that extremely high-degree nodes systematically distort feature aggregation and worsen discrimination in the most connected parts of the graph [14]. Alongside oversmoothing, degradation is caused by two further phenomena: label heterophily, in which aggregation mixes the representations of nodes of opposite classes [15], and oversquashing, which limits signal transmission through structural bottlenecks [16]. Among architectural mitigations, the best known are residual connections to the initial representation [17], PairNorm normalisation of pairs of representations [18], and DropEdge edge dropping [19]. Applied issues of class imbalance and decision transparency are addressed in works combining graph models with generative data augmentation [20] and interpretability mechanisms [21]. What has remained outside the scope of existing work is the combination of both lines: a targeted analysis of how oversmoothing affects the detection of privileged insiders, specifically, why the effect is stronger for them, how to separate it from heterophily, and by what means it can be mitigated.

The aim of the article is the quantitative substantiation and experimental verification of the impact of oversmoothing of graph neural networks on the quality of insider-threat detection for privileged users, the separation of the mechanisms of representation degradation, and the identification and quantitative comparison of the means of overcoming it that are suitable for the most high-risk segment of accounts. To achieve this aim, the following tasks are set:

- to formalise a graph model of user behaviour and explain the structural exceptionality of privileged nodes, separating privilege from connectivity;
- to reveal the mathematical nature of oversmoothing and justify the choice of diagnostic metrics with a correct definition of the Laplacian;
- to conduct a computational experiment with a relational insider signal on a trained graph convolutional network and measure the dynamics of smoothing and of detection quality for privileged and ordinary nodes with statistical averaging;
- to separate the contributions of heterophily and oversmoothing and to experimentally evaluate the means of increasing model robustness.

II. STUDING METODOLOGY

Logs of logons, file accesses, e-mail exchange, removable-media use, and network connections describe not isolated events but a web of interactions, in which what becomes suspicious is not the event itself but its position in the structure of links. The appeal of a graph representation of behaviour is due to several interrelated properties: 'user–resource' and 'user–user' relations are encoded directly, so that the context of an action becomes part of the representation; the message-passing mechanism aggregates the features of neighbours and detects an anomaly by the mismatch of behaviour with its environment even for normal individual features; and the combination of structural and attribute information increases robustness to the sparsity and incompleteness of logs [5].

The structure of the behaviour graph model underlying the analysis is summarised in Table 1. Nodes correspond to subjects and resources, edges to interactions recorded in the logs, and attributes to aggregated behavioural characteristics over a reporting period. Note that, for the analysis of oversmoothing, a homogeneous projection of this graph is used below; full type-dependent aggregation is considered as a mitigation.

TABLE 1. Structure of the user-behaviour graph model based on CERT-type logs.

Graph element	Semantics	Attributes and features	Approx. scale
"User" node	Employee account or per-day profile	Role, department, action frequencies, deviations from baseline	thousands of nodes
"Resource" node	Workstation, file resource, mailbox, network segment	Type, sensitivity level, access mode	hundreds – thousands
"Logon" edge	Authentication event on a device	Time, duration, own/foreign PC flag	tens of thousands
"Access" edge	File access or e-mail interaction	Direction, volume, off-hours	tens of thousands
Privileged node	Administrator, manager, holder of broad rights	High connectivity, access to many resources	single-digit %

In this study, a “privileged user” is defined as a system account with extensive privileges, the compromise of which poses a disproportionately high threat to the system (for example, system administrators, IT security specialists, senior executives, as well as service or automation accounts with a high level of access). Such users differ from ordinary users in two ways: they have elevated access rights to many resources, including the most sensitive ones; structurally, in the behavioural graph, this is reflected in a node’s high degree of connectivity and centrality, whereas a regular user remains a node with a low degree of connectivity and is largely confined to their own community. The first parameter defines the privilege from a security perspective; the second represents it in the

graph model, and it is precisely this node structure that determines the behaviour of excessive smoothing, which is analyzed below.

Privileged users interact with many resources and colleagues, thereby acquiring a high degree of connectivity and becoming structural hubs. Importantly, privilege and degree, although correlated, are not identical: some administrators are low-activity, and some highly connected nodes have no elevated rights. Therefore, in the model, privilege is set as a separate role label, only moderately correlated with degree (discussed below). Seizing privileged access can nullify the remaining lines of defence, so detecting anomalies in the behaviour of holders of broad permissions is of priority importance [9].

Graph convolutional networks are built on the repeated aggregation of neighbour features. The basic operation is multiplication by the symmetrically normalized adjacency matrix with self-loops, defined in formula (1):

$$\hat{A} = \tilde{D}^{-1/2}(A+I)\tilde{D}^{-1/2}, \quad (1)$$

where A is the adjacency matrix of the graph, I is the identity matrix that adds self-loops, and $\tilde{D}$ is the diagonal degree matrix of the graph with self-loops (of the matrix $A+I$). The propagation of features over L layers in the linear approximation is described by formula (2):

$$H^{(L)} = \hat{A}^{L} \cdot H^{(0)}, \quad (2)$$

where $H^{(0)}$ is the matrix of initial node features and $H^{(L)}$ their representations after L aggregation steps. Repeated multiplication by the matrix $\hat{A}$ drives the representations toward the dominant eigen-subspace of the propagation operator; the corresponding eigenvector has components proportional to the square root of the node degree with self-loop, so that with each layer the representations are increasingly determined by structural position rather than by individual behavioural features [10]. A quantitative measure of smoothness is the Dirichlet energy, which assesses the average divergence of the representations of adjacent nodes, in the normalized form (3):

$$\begin{aligned} &E(H) - tr(H^T \cdot \hat{L} \cdot H) / tr(H^T \cdot H), \\ &\hat{L} = I - D^{-1/2} \cdot A \cdot D^{-1/2} \end{aligned}, \quad (3)$$

where $\hat{L}$ is the standard symmetrically normalized graph Laplacian without self-loops. We stress that the propagation operator $\hat{A}$ in formula (1) (with self-loops) and the Laplacian $\hat{L}$ in formula (3) (without self-loops) are different matrices: the energy is correctly measured precisely with respect to the classical Laplacian, whereas the self-loop augmentation concerns aggregation only. A decay of the Dirichlet energy indicates the loss of the high-frequency components of the representations, i.e., the erasure of differences between nodes [11]; energy metrics should be complemented by rank-based indicators sensitive to smoothing even before the critical energy drop [13].

III. EXPERIMENTS

A. Synthetic graph generation. To test the hypotheses put forward, a synthetic behaviour graph of 1500 user nodes with a heavy-tailed degree distribution (mean degree 12.7) was constructed, generated by a modified degree-corrected block model: eight communities set a homogeneous feature structure, while the degree correction produces hubs. Each node is assigned a 32-dimensional feature vector, obtained by averaging the community profile with noise. Eight percent of the nodes are assigned to the insider category regardless of privilege, which excludes trivial recognition by degree alone.

The principal difference from simplified formulations lies in the nature of the insider signal. In addition to a weak local feature deviation, insiders are assigned a relational signal: an anomalous pattern of links – additional edges to a fixed cluster of sensitive resources, atypical for an ordinary user. As a result, an insider's own features remain almost normal within their community, while the anomaly manifests only after aggregating neighbourhood features. As a result, feature-only detection (equivalent to zero depth) yields an area under the ROC curve of about 0.53, whereas already one or two propagation steps raise it to 0.93 – 0.96. This removes the logical contradiction of simplified experiments in which the graph provided no gain, and restores the correct logic 'the graph is useful, but excessive depth destroys it'.

Privilege is set as a role label (ten percent of the nodes) correlated with, but distinct from, degree: the mean degree of privileged nodes is 37.1 versus 10.0 for ordinary ones, yet the Spearman rank correlation between privilege and degree equals only 0.43.

The figure of 10% was adopted as a rough estimate of the order of quantity, rather than as a precisely calibrated metric: in corporate environments, accounts with administrative or managerial privileges typically constitute a small minority (usually significantly less than approximately 10 – 15%, assuming administration follows the “least privilege” principle). The exact proportion varies depending on the organization and is not fixed in the CERT dataset on internal threats. Therefore, in this study, this value is treated as an experimental parameter, and its impact is examined in detail in the robustness analysis below (Table 4), where the proportion of accounts with privileges varies between 5% and 15% without altering the qualitative conclusions.

B. Baseline model and training setup. The base detector is represented by a standard graph convolutional network, in which each layer performs normalized neighbourhood aggregation according to Equation (1), followed by a training linear transformation and a ReLU nonlinearity; the input linear layer maps the 32-dimensional node features to a hidden width of 48, and the output linear layer generates logits for two classes. The depth of the network L (the number of graph convolution layers) is an independent study variable and ranges from 1 to 16, and in the stability analysis – up to 32. The model is trained transductively for node classification: nodes are split into 60% training and 40% test subsets, stratified by the “insider” label, and each configuration is repeated across 10 independent stratified splits, with means and standard deviations reported. Training minimizes the class-weighted cross-entropy loss (the “insider” class is weighted by the “negative-to-positive” ratio to compensate for an 8% class imbalance) using the Adam optimizer with a learning rate of 0.01, weight decay of $5 \cdot 10^{-4}$, and a

dropout rate of 0.5 over 110 epochs; the operating threshold used for the F1 score is set to the baseline score for insiders. All architectures designed to reduce risk and resist heterophilia, presented below (residual/GCNII, PairNorm, DropEdge, GPR-GNN, H2GCN), are trained using this identical protocol; therefore, the observed differences reflect architectural features rather than optimization choices. To diagnose smoothing, we additionally calculated, separately, the normalized Dirichlet energy, the cosine similarity of representations with neighbours, and the local difference for privileged and ordinary nodes under the assumption of linear propagation for depths ranging from zero to sixteen layers.

The dynamics of smoothing are illustrated in Fig. 1. The normalized Dirichlet energy decays exponentially and reaches a low level already after six to eight layers (a). The cosine similarity of representations to neighbours for privileged nodes grows ahead of schedule: it exceeds the corresponding indicator of ordinary nodes at all depths and approaches unity by four layers (b). The accelerated smoothing of hubs has a structural cause: a high-degree node averages the features of a large neighbourhood, so its own behavioural specificity dissolves in fewer propagation steps, while a larger projection onto the dominant eigenvector of the operator accelerates the approach to the limiting regime [14].

The direct consequence of smoothing is reflected in Table 2 and Fig. 2. The area under the ROC curve, which at shallow depth ranges from 0.93 to 0.96 and significantly exceeds the feature-only baseline (0.53), decreases as depth increases. Most importantly, the degradation is uneven across categories: for ordinary nodes, the AUC at eight layers is still 0.77, whereas for privileged hubs it falls to about 0.49 (the random-guessing level) and to 0.44 at sixteen layers. Thus, precisely on the most high-risk segment of users, the deep model effectively loses its detection ability.

The F1 score gives an analogous picture (Fig. 3): for privileged insiders, it declines more steeply, since the individual behavioural signal of hubs is erased earlier. The obtained result is consistent with observations of a recall drop on difficult CERT subsets as model depth grows [5].

Separation of heterophily and oversmoothing. Degradation already begins at shallow depth (between the first and second layers), and the AUC decreases from 0.96 to 0.94, long before the energy collapse. This points to a second mechanism, distinct from oversmoothing – label heterophily. In the generated graph the structural homophily is high (the share of edges between nodes of the same community is 0.82), yet with respect to the 'insider' label the environment is heterophilous: among the neighbours of an ordinary insider, insiders constitute only about 8.5 percent, i.e., at the base-rate level. Therefore, even one or two rounds of aggregation inevitably mix an insider's representation with those of normal neighbours [15], which explains the early drop.

By contrast, the collapse at eight or more layers is energetic in nature (Fig. 1) and corresponds to classical oversmoothing [10]. It is practically important that these mechanisms require different mitigations: heterophily is alleviated by heterophily-resistant architectures and type-dependent aggregation, whereas oversmoothing is alleviated by preserving the energy of representations.

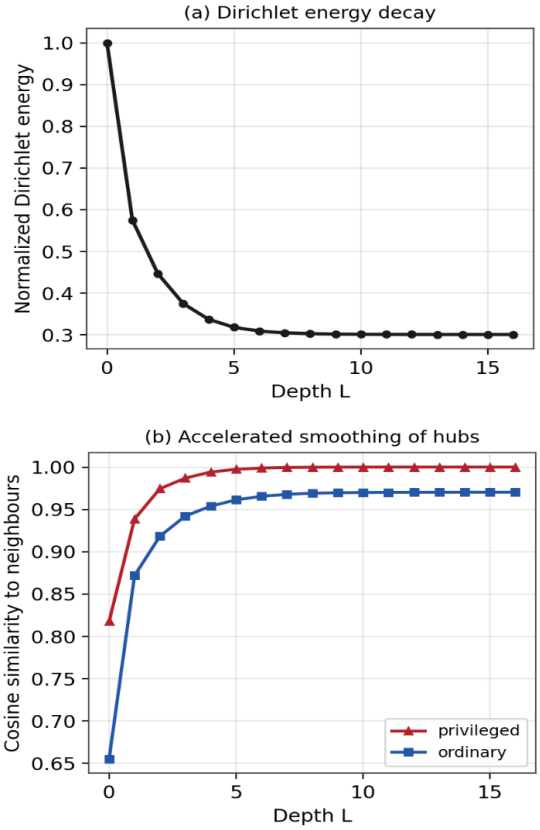

FIG. 1. Decay of the normalized Dirichlet energy (a) and the accelerated growth of the cosine similarity of hub representations (b) with increasing network depth.

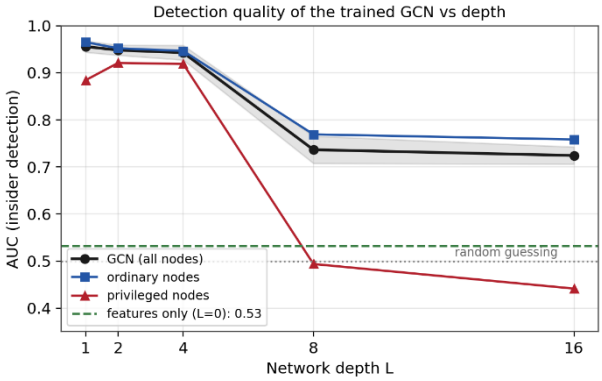

FIG. 2. Degradation of the detection quality of the trained GCN with increasing depth: for privileged nodes, the AUC falls to the random-guessing level at eight layers and below it at sixteen.

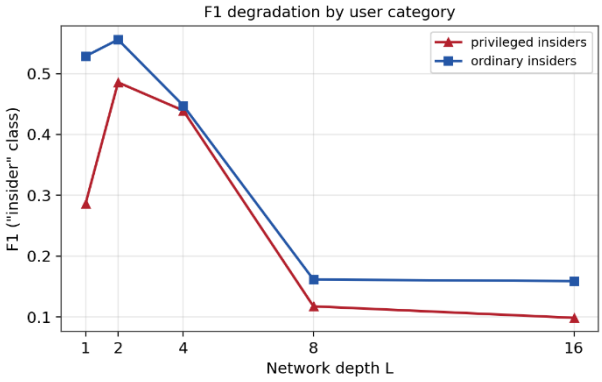

FIG. 3. Degradation of the F1 score for the "insider" class by user category.

TABLE 2. Smoothing and detection-quality indicators of the trained GCN as a function of depth (mean ± standard deviation over 10 splits).

Depth L	Dirichlet energy (norm.)	Similarity priv./ord.	AUC (all)	AUC priv.	AUC ord.
1	0.575	0.94 / 0.87	0.956 ± 0.012	0.884	0.965
2	0.447	0.97 / 0.92	0.948 ± 0.011	0.921	0.952
4	0.337	0.99 / 0.95	0.943 ± 0.016	0.919	0.947
8	0.303	1.00 / 0.97	0.737 ± 0.029	0.494	0.769
16	0.301	1.00 / 0.97	0.724 ± 0.018	0.442	0.758
1	0.575	0.94 / 0.87	0.956 ± 0.012	0.884	0.965
2	0.447	0.97 / 0.92	0.948 ± 0.011	0.921	0.952
4	0.337	0.99 / 0.95	0.943 ± 0.016	0.919	0.947
8	0.303	1.00 / 0.97	0.737 ± 0.029	0.494	0.769
16	0.301	1.00 / 0.97	0.724 ± 0.018	0.442	0.758

The proposed directions for overcoming the limitation were not confined to a list but tested experimentally on the same testbed. The baseline GCN was compared with three mitigations: residual connections to the initial representation in the spirit of GCNII [19], PairNorm normalization [18], and DropEdge edge dropping [19]. The results are shown in Fig. 4.

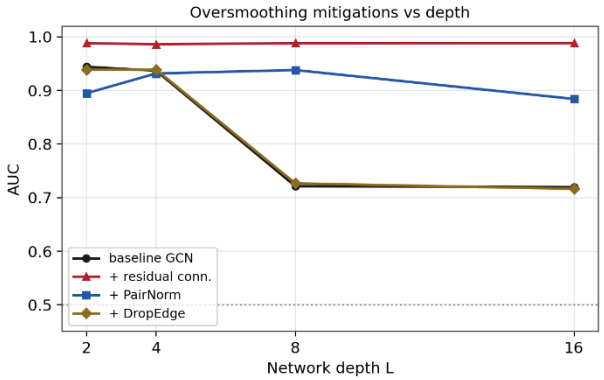

FIG. 4. Detection quality for the baseline GCN and three oversmoothing mitigations as a function of depth.

Residual connections proved the most effective: they not only prevent collapse but also allow a deep network (sixteen layers) to maintain an AUC of 0.99, because the initial representation is constantly 'mixed in' and prevents the approach to the limiting regime. PairNorm also largely preserves distinguishability (AUC about 0.89–0.94 at depths 8–16). By contrast, edge dropping proved insufficient in this regime: under strong self-loop smoothing, it only slows but does not prevent collapse (the AUC at eight layers remains at the baseline level, about 0.72). This result demonstrates that the mitigations are not interchangeable, and their choice must take into account the dominant degradation mechanism.

Since two different degradation mechanisms were identified, the comparison was extended to architectures that address each of them. Against oversmoothing, GCNII with tuning of the initial-residual parameter α and identity mapping in the weight matrices was studied [17]; against label heterophily, two heterophily-resistant architectures were studied: GPR-GNN with learnable generalized-PageRank weights that may be negative and therefore preserve high-frequency components [22], and a simplified H2GCN with separation of node and neighbour representations, aggregation of higher-order neighbourhoods, and combination of intermediate representations (jumping knowledge) [15].

C. Descriptions of the architectures. GCNII [17] combines two mechanisms against oversmoothing. The initial residual mixes the original representation into each layer with weight α, while the identity mapping preserves part of the representation, bypassing the weight matrix with a coefficient $\beta_\ell = \ln(\lambda/\ell + 1)$ that decays with depth. Each layer is computed as $H^{(\ell+1)} = \sigma(((1-\alpha)\hat{A}H^{(\ell)} + \alpha H^{(0)})((1-\beta_\ell)I + \beta_\ell W^{(\ell)}))$. The parameter α controls the strength of the 'anchoring' to the input features, so its tuning is key.

GPR-GNN [22] first transforms the features with a multilayer perceptron and then forms the final representation as a signed sum of propagations $Z = \Sigma\, \gamma_k \hat{A}^k H$, where the weights γ_k are learnable. The possibility of negative weights allows the model to 'subtract' the contribution of a heterophilous neighbourhood, which makes it robust to both smoothing and label heterophily.

H2GCN [15] implements three design ideas for heterophilous graphs: separation of a node's own representation from those of its neighbours (aggregation without self-loops), separate accounting for higher-order neighbourhoods (one- and two-hop), and combining the intermediate representations across all rounds (jumping knowledge). As a result, a node's individual signal is neither mixed with its environment nor erased, so the architecture is almost independent of depth.

The tuning of α for GCNII (Fig. 5) shows that the detection of privileged nodes is restored for any value of the parameter (the area under the ROC curve for this group rises monotonically from 0.90 at $\alpha = 0.1$ to 0.97 at $\alpha = 0.5$, whereas for the baseline network it fell to 0.42), while the best overall quality (0.988) is given by $\alpha = 0.5$; the value $\alpha = 0.3$ already yields 0.98 overall and 0.94 on the privileged segment and is used in the comparison. This confirms that the constant 'mixing in' of the initial representation eliminates the collapse and removes the dependence of quality on depth over a wide range of α.

The summary comparison (Table 3, Fig. 6) establishes a clear hierarchy by demonstrating the ability to restore detection of the privileged segment, where the baseline network is effectively helpless (0.42, the random-guessing level).

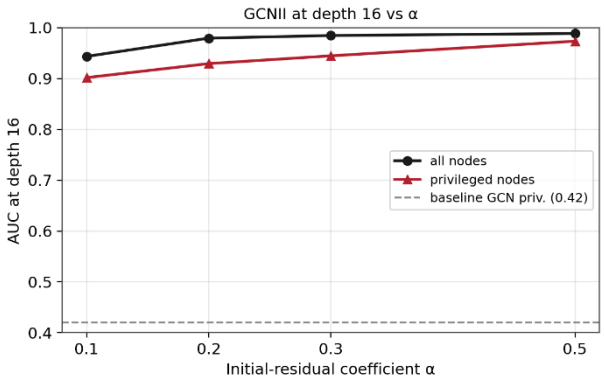

FIG. 5. Dependence of GCNII quality at depth 16 on the initial-residual coefficient α.

TABLE 3. Comparison of architectures at depth 16 (mean ± standard deviation over 8 splits).

Architecture	AUC (all)	AUC privileged	AUC ordinary
Baseline GCN	0.722 ± 0.013	0.420	0.758
PairNorm [18]	0.908 ± 0.019	0.837	0.914
GPR-GNN [22]	0.955 ± 0.019	0.878	0.962
H2GCN [15]	0.965 ± 0.012	0.930	0.968
GCNII, α=0.3 [17]	0.984 ± 0.005	0.944	0.988

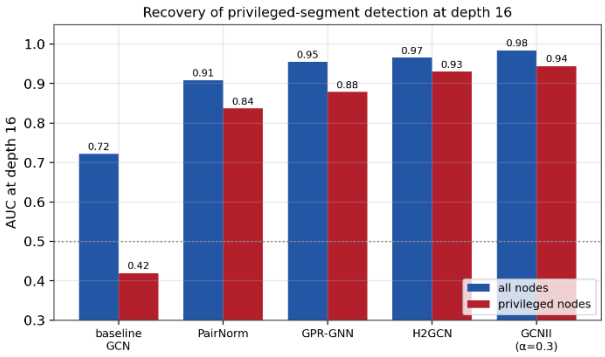

FIG. 6. Detection quality for all nodes and for privileged nodes at depth 16: recovery of privileged-segment detection by different architectures.

PairNorm raises this indicator to 0.84, GPR-GNN to 0.88, and the heterophily-resistant H2GCN and the energy-preserving GCNII to 0.93 and 0.94, respectively, essentially co-leading and almost closing the gap between privileged and ordinary nodes. The strong result of H2GCN is natural: since the insider label is heterophilous, separating a node's own representation from the aggregated neighbourhood and using higher orders preserves precisely the high-frequency signal that distinguishes an insider from their normal environment; GCNII attains a comparable level because, for this predominantly structural relational signal, constant mixing-in of the input representation is equally effective.

Robustness to depth (Fig. 7) reveals an additional difference. The baseline network collapses: for privileged nodes the area under the ROC curve falls from 0.88 at two layers to 0.38 at thirty-two. GCNII maintains quality up to sixteen layers and only slowly declines at extreme depth (0.91 at thirty-two layers). GPR-GNN proves practically invariant to depth: owing to learnable layer weights that automatically reduce the contribution of over-smoothed deep components, and its quality for privileged nodes remains at 0.85–0.87 up to thirty-two layers. This result is consistent with the theoretical property of GPR-GNNs, which avoid oversmoothing without requiring a shallow network [22].

Thus, the mitigations are not merely listed but quantitatively compared for both degradation mechanisms. For practical insider-detection systems, this yields a concrete recommendation: in the presence of a pronouncedly heterophilous label (rare insiders among predominantly normal neighbours), priority should be given to heterophily-resistant architectures (H2GCN, GPR-GNN), complementing them with energy-preserving means (GCNII residual connections) for depth robustness, with mandatory quality control on the privileged segment.

To ensure that the conclusions are not an artefact of a single configuration, a sensitivity analysis with respect to the strength of the relational signal was carried out (Fig. 8). The qualitative shape of the dependence is preserved at all signal levels: shallow depth yields high quality, and large depth yields collapse. At the same time, the weaker the signal, the deeper the collapse: for the weakest relational signal the AUC of the deep network drops to 0.56, i.e., almost to the random-guessing level. This confirms the robustness of the main conclusion and reconciles it with the most severe scenarios described in the literature.

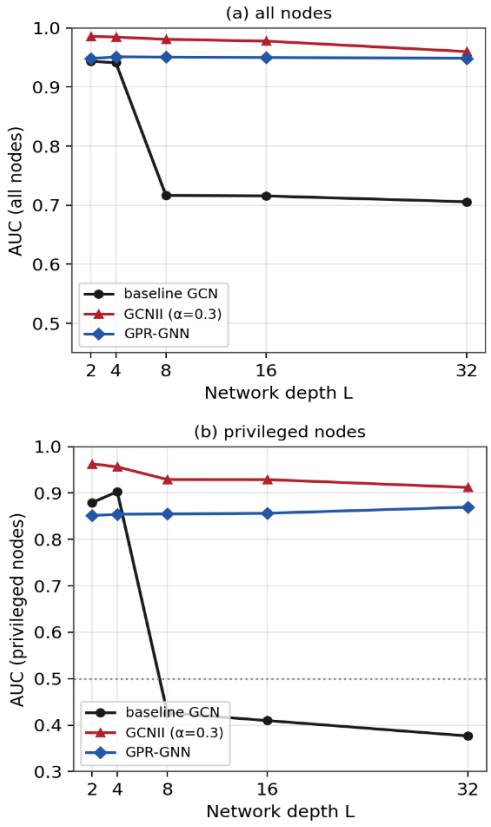

FIG. 7. Robustness of architectures to increasing depth for all nodes (a) and for privileged nodes (b).

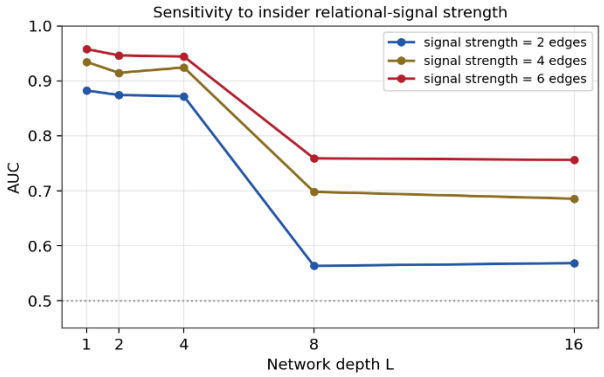

FIG. 8. Sensitivity of detection quality to the strength of the insider relational signal.

In addition to signal strength, the robustness of the findings with respect to the graph-generation process parameters was tested. Starting from the baseline configuration, we sequentially varied the number of nodes (1,000–2,500), the number of communities (5–12), the average degree of connectivity (8–18), and the proportion of privileged nodes (5–15%); for each configuration, three metrics were measured: AUC based solely on features, AUC for privileged nodes of the baseline GCN at depth 8 (collapse indicator), and AUC for privileged nodes of GCNII at depth 16 (recovery indicator). The results are presented in Table 4. A consistent pattern emerges across all configurations: detection based solely on features remains at a level close to random (0.50–0.57), confirming the necessity of the graph; the baseline network loses the privileged segment at depth (0.21–0.65, i.e., at or below the random level for this group); while the energy-saving architecture restores it to 0.87–0.97. Thus, the conclusions do not depend on the specific structural parameters of the selected configuration but reflect the general interaction between the node structure and the network depth.

TABLE 4. Robustness of the main conclusions to graph-generation parameters (privileged-node AUC).

Configuration	Feature-only AUC	Baseline GCN, priv. (L=8)	GCNII, priv. (L=16)
Reference (N=1500, K=8, ⟨d⟩=12, 10 %)	0.53	0.39	0.91
Nodes N = 1000	0.57	0.51	0.97
Nodes N = 2500	0.53	0.55	0.90
Communities K = 5	0.51	0.65	0.88
Communities K = 12	0.53	0.60	0.87
Mean degree = 8	0.53	0.43	0.94
Mean degree = 18	0.53	0.37	0.91
Privileged fraction 5 %	0.53	0.21	0.90
Privileged fraction 15 %	0.53	0.48	0.94

IV. LIMITATIONS OF THE STUDY

The results were obtained on a synthetic graph whose structure is aligned with the statistics of CERT-type logs, but not on real corporate logs; transferring the conclusions requires separate verification. The analysis of smoothing was performed on a homogeneous projection of the graph and on a static snapshot, whereas real insider-detection systems rely on heterogeneous and temporal architectures [3-5] capable of partially compensating for the loss of the local signal. Finally, three of the possible mitigations were tested; a broader comparison (in particular, of heterophily-resistant architectures and means against oversquashing [16]) remains a subject for further research.

V. CONCLUSIONS

Oversmoothing is not a marginal technical inconvenience but a systemic limitation of graph models in the insider-detection task, which is most pronounced for privileged users. The experiment with a relational signal showed that the graph is indeed useful – one or two aggregation steps raise the area under the ROC curve from 0.53 to 0.93–0.96 – yet further increasing depth destroys quality. In a trained graph convolutional network, it was established that for privileged hubs this indicator falls to the random-guessing level at eight layers and below it by sixteen, whereas for ordinary nodes it remains markedly higher; the F1 score for privileged insiders declines faster. Privilege was separated from connectivity (Spearman rank correlation 0.43), and the degradation was decomposed into two mechanisms: an early quality loss due to label heterophily and a deep collapse caused by proper oversmoothing.

The mitigations are not merely listed; they are quantitatively compared for both degradation mechanisms. On the privileged segment, the baseline network is helpless (AUC 0.42), whereas PairNorm, GPR-GNN, H2GCN, and GCNII restore its detection to 0.84, 0.88, 0.93, and 0.94, respectively; H2GCN, designed for heterophilous graphs, and GCNII, which preserves the input representation, essentially co-lead and almost eliminate the gap between privileged and ordinary nodes, while GPR-GNN, owing to learnable layer weights, remains practically depth-invariant. Hence, a concrete practical recommendation follows: instead of mechanically increasing the depth of the baseline architecture to protect the privileged segment, it is advisable to use heterophily-resistant architectures (H2GCN, GPR-GNN) complemented by energy-preserving means (GCNII residual connections), with mandatory separate quality control for privileged accounts.

For the first time, it has been established and quantitatively assessed that oversmoothing causes a disproportionately faster degradation of detection specifically for privileged hub users (for them the area under the ROC curve falls to the random-guessing level at eight layers and below it by sixteen, whereas for ordinary nodes it remains markedly higher), and this effect is for the first time separated from the trivial influence of connectivity – privilege is set as a distinct role label with a rank correlation with degree of only 0.43; the decline of detection quality is decomposed into two independent mechanisms (label heterophily (early quality loss) and oversmoothing proper (deep collapse)) which have not previously been distinguished in the context of insider detection.

The methodology has been improved: instead of a linear propagation operator, a trained graph convolutional network with statistical averaging over stratified splits is used, and the discriminative insider signal is supplemented with a relational component, so that the usefulness of the

graph representation becomes explicit (the area under the ROC curve grows from 0.53 for features alone to 0.93–0.96 after one or two aggregation steps) and the 'best zero depth' artefact inherent in simplified formulations is eliminated.

Further development has been obtained by the methods of overcoming oversmoothing and heterophily, which are for the first time compared with one another precisely by their ability to restore detection on the privileged segment; this made it possible to build a quantitative hierarchy of means (baseline network 0.42 → PairNorm 0.84 → GPR-GNN 0.88 → H2GCN 0.93 → GCNII 0.94) and to formulate a practical recommendation on the choice of architecture depending on the dominant degradation mechanism.

A prospect for further research is to verify the proposed approach on real logs from corporate environments, as well as to extend the analysis to heterogeneous and temporal graph architectures.

AUTHOR CONTRIBUTIONS

I.D. – conceptualisation, supervision; V.P. – methodology, software, validation, investigation, formal analysis, investigation, writing-original draft preparation.

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

[1] I. Dobrynin, T. Radivilova, N. Maltseva, and D. Ageyev, "Use of approaches to the methodology of factor analysis of information risks for the quantitative assessment of information risks based on the formation of cause-and-effect links," in *Proc. 2018 Int. Sci.-Pract. Conf. Problems of Infocommunications. Science and Technology (PIC S&T)*, Kharkiv, Ukraine, 2018, pp. 229–232. doi: 10.1109/INFOCOMMST.2018.8632022.

[2] T. Radivilova, K. Lyudmyla, O. Lemeshko, D. Ageyev, M. Tawalbeh, and A. Ilkov, "Analysis of approaches of monitoring, intrusion detection and identification of network attacks," in *Proc. 2020 IEEE Int. Conf. Problems of Infocommunications. Science and Technology (PIC S&T)*, 2020, pp. 819–822. doi: 10.1109/PICST51311.2020.9467973.

[3] F. Ares-Robledo, H. Rifà-Pous, and R. Clarisó, "Graph neural networks for anomaly detection: A systematic review of dynamic temporal approaches," *Artif. Intell. Rev.*, vol. 59, Art. no. 129, 2026. doi: 10.1007/s10462-026-11532-7.

[4] S. Jiang, "A survey of heterogeneous graph neural networks for cybersecurity anomaly detection," arXiv:2510.26307, 2025. [Online]. Available: https://arxiv.org/abs/2510.26307

[5] R. Yumlembam, B. Issac, S. M. Jacob, L. Yang, and D. Krishnan, "Insider threat detection using GCN and Bi-LSTM with explicit and implicit graph representations," *IEEE Trans. Artif. Intell.*, 2025. [Online]. Available: https://arxiv.org/abs/2512.18483

[6] O. Neretin and V. Kharchenko, "Information technology for assessing and ensuring cybersecurity of large language models," *Secure Inf. Syst. IoT (SISIOT)*, vol. 3, no. 2, p. 02020, Dec. 2025. doi: 10.31861/sisiot2025.2.02020.

[7] D. Ageyev, L. Kirichenko, T. Radivilova, M. Tawalbeh, and O. Baranovskyi, "Method of self-similar load balancing in network intrusion detection system," in *Proc. 2018 28th Int. Conf. Radioelektronika*, Prague, Czech Republic, 2018, pp. 1–4. doi: 10.1109/RADIOELEK.2018.8376406.

[8] T. Radivilova, L. Kirichenko, A. S. Alghawli, D. Ageyev, O. Mulesa, O. Baranovskyi, A. Ilkov, V. Kulbachnyi, and O. Bondarenko, "Statistical and signature analysis methods of intrusion detection," in *Information Security Technologies in the Decentralized Distributed Networks*, Lecture Notes on Data Engineering and Communications Technologies, vol. 115, R. Oliynykov, O. Kuznetsov, O. Lemeshko, and T. Radivilova, Eds. Cham, Switzerland: Springer, 2022. doi: 10.1007/978-3-030-95161-0_5.

[9] T. Radivilova, L. Kirichenko, D. Ageyev, M. Tawalbeh, V. Bulakh, and P. Zinchenko, "Intrusion detection based on machine learning using fractal properties of traffic realizations," in *Proc. 2019 IEEE Int. Conf. Advanced Trends in Information Theory (ATIT)*, Kyiv, Ukraine, 2019, pp. 218–221. doi: 10.1109/ATIT49449.2019.9030452.

[10] T. K. Rusch, M. M. Bronstein, and S. Mishra, "A survey on oversmoothing in graph neural networks," arXiv:2303.10993. [Online]. Available: https://arxiv.org/abs/2303.10993

[11] F. Di Giovanni, J. Rowbottom, B. P. Chamberlain, T. Markovich, and M. M. Bronstein, "Understanding convolution on graphs via energies," arXiv:2206.10991. [Online]. Available: https://arxiv.org/abs/2206.10991

[12] J. Chen, Y. Wang, C. Bodnar, R. Ying, P. Liò, and Y. G. Wang, "Dirichlet energy enhancement of graph neural networks by framelet augmentation," arXiv:2311.05767. doi: 10.48550/arXiv.2311.05767.

[13] K. Zhang, P. Deidda, D. J. Higham, and F. Tudisco, "Are we measuring oversmoothing in graph neural networks correctly?," arXiv:2502.04591, 2025. doi: 10.48550/arXiv.2502.04591.

[14] X. Sun et al., "Understanding the influence of extremely high-degree nodes on graph anomaly detection," in *Pattern Recognition (ICPR 2025)*. Cham, Switzerland: Springer, 2025. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-78183-4_2

[15] J. Zhu, Y. Yan, L. Zhao, M. Heimann, L. Akoglu, and D. Koutra, "Beyond homophily in graph neural networks: Current limitations and effective designs," arXiv:2006.11468, 2020. [Online]. Available: https://arxiv.org/abs/2006.11468

[16] U. Alon and E. Yahav, "On the bottleneck of graph neural networks and its practical implications," arXiv:2006.05205, 2020. [Online]. Available: https://arxiv.org/abs/2006.05205

[17] M. Chen, Z. Wei, Z. Huang, B. Ding, and Y. Li, "Simple and deep graph convolutional networks," arXiv:2007.02133, 2020. [Online]. Available: https://arxiv.org/abs/2007.02133

[18] L. Zhao and L. Akoglu, "PairNorm: Tackling oversmoothing in GNNs," arXiv:1909.12223, 2020. [Online]. Available: https://arxiv.org/abs/1909.12223

[19] Y. Rong, W. Huang, T. Xu, and J. Huang, "DropEdge: Towards deep graph convolutional networks on node classification," arXiv:1907.10903, 2020. [Online]. Available: https://arxiv.org/abs/1907.10903

[20] M. Yang and C. Liu, "XGA-E: an explainability-enhanced graph neural network for network traffic anomaly detection," *Cybersecurity*, vol. 9, Art. no. 99, 2026. doi: 10.1186/s42400-025-00487-x.

[21] P. Lavanya, H. A. Glory, M. Aggarwal, M. et al., "Unmasking insider threats using a robust hybrid optimized generative pretrained neural network approach," *Sci. Rep.*, vol. 15, Art. no. 26718, 2025. doi: 10.1038/s41598-025-12127-y.

[22] E. Chien, J. Peng, P. Li, and O. Milenkovic, "Adaptive universal generalized PageRank graph neural network," arXiv:2006.07988, 2021. [Online]. Available: https://arxiv.org/abs/2006.07988

Ihor Dobrynin

Research activity: research supervisor of winners of national competitions for students' scientific works in the field of Information Security; expert auditor of information security management systems. Scientific directions: process approaches to the audit of information security management systems, assessment and minimization of information risks; Project Academic Advocate of ISACA.

ORCID ID: 0000-0001-8910-2609

Vadym Pantelieiev

Research Interests: Research and refinement of methods for predicting domestic incidents through the analysis of social network structures and interactions

Research Publications: 3 research publications.

ORCID ID: 0009-0008-7824-8782

Обмеження графових нейронних мереж у задачі виявлення інсайдерських загроз

Ігор Добринін*, Вадим Пантєлєєв

Кафедра інфокомунікаційної інженерії ім.В.В. Поповського, Харківський національний університет радіоелектроніки, Харків, Україна

*Автор-кореспондент (Електронна адреса: ihor.dobrynin@nure.ua)

АНОТАЦІЯ У статті досліджується фундаментальне обмеження графових нейронних мереж у виявленні внутрішніх загроз: надмірне згладжування представлень вузлів, яке найбільш виражено для користувачів із привілеями. Щоб зробити значення графа очевидним, інсайдерам присвоюється не лише слабкий локальний сигнал, а й реляційний – аномальний шаблон зв'язків із чутливими ресурсами: виявлення лише за ознаками дає площу під кривою ROC приблизно 0,53, тоді як один або два етапи агрегації підвищують її до 0,93–0,96. Привілей моделюється як мітка ролі, яка корелює з зв'язністю, але відрізняється від неї (рангова кореляція Спірмена 0,43). Верифікація виконується за допомогою навченої графової конволюційної мережі, усередненої за стратифікованими поділами. Зі збільшенням глибини якість виявлення знижується; для привілейованих хабів площа під кривою ROC падає до рівня випадкового вгадування на восьми рівнях і нижче нього на шістнадцятому, тоді як для звичайних вузлів вона залишається помітно вищою. Виокремлюються два механізми погіршення якості: гетерофілія міток (рання втрата якості) та власне надмірне згладжування (глибокий колапс). Для обох механізмів представлено та кількісно порівняно заходи щодо пом'якшення наслідків: базова мережа практично безсила щодо привілейованого сегмента (AUC 0,42), тоді як PairNorm підвищує цей показник до 0,84, GPR-GNN – до 0,88, а H2GCN та GCNII – до 0,93 та 0,94 відповідно, майже повністю усуваючи розрив між привілейованими та звичайними вузлами; GPR-GNN виявляється практично незалежним від глибини. Отримані результати мають практичне значення для розробки систем виявлення внутрішніх загроз, стійких до погіршення якості в сегменті користувачів з найвищим рівнем ризику. Стійкість цих висновків до параметрів генерації графів підтверджується спеціальним аналізом, у якому варіюються розмір графа, кількість спільнот, середній ступінь зв'язаності та частка привілейованих вузлів.

КЛЮЧОВІ СЛОВА внутрішні загрози, графові нейронні мережі, гетерофілія, користувачі з привілеями, виявлення аномалій.