Graph Neural Networks for Trust Evaluation: Criteria, State-of-the-Art, and Future Directions

Tingxi Luo*, Jie Wang*, *Graduate Student Member, IEEE,* Zheng Yan, *Fellow, IEEE,* and Erol Gelenbe, *Life Fellow, IEEE*

***Abstract*—The process of quantifying trust considers the factors that affect it, which can be applied to identify malicious behavior, reduce uncertainty, and facilitate decision-making. Traditional trust evaluation methods based on statistics and reasoning, rely heavily on domain knowledge, which limits their practical applications. Graph Neural Networks (GNNs) are a new Machine Learning (ML) paradigm that can revolutionize the evaluation of trust, by modeling relationships as graphs to simplify relevant data and automating end-to-end evaluation. Thus, a variety of GNN-based trust evaluation models have been developed for different applications. However, there is still a gap in the literature regarding a review on these advances with discussion about remaining challenges. To bridge this gap, we conduct the first review on GNN-based trust evaluation. We first propose a set of criteria in terms of trust-related attributes, correctness, functionality, and overhead. Then, we propose a taxonomy of existing GNN-based trust evaluation models, followed by a review using the proposed criteria to analyze their pros and cons. A quantitative analysis of the recent cutting-edge models is also provided. Based on the review and experimental results, we identify key challenges and suggest future research directions.**

***Index Terms*—Graph neural networks, trust evaluation, trust assessment, trust prediction, trust.**

I. Introduction

Trust evaluation plays a crucial role in cybersecurity. It offers a valuable approach to quantify trust by considering the factors that affect trust. It has been widely applied into various cybersystems, such as social networks, Cyber-Physical Systems (CPS), communication networks, and fog/cloud computing systems. In these systems, trust evaluation can assist in intrusion detection, service selection, task assignment, access control, data fusion, trustworthy routing, recommendation, incentives, system optimization, and so on [1].

Traditional trust evaluation models can be classified into statistical models, reasoning models, and Machine Learning (ML) models [2]. The statistical models use extensive interaction data and simple statistical methods to assess trust. The effectiveness of these models relies on weight selection and data availability, making it difficult to evaluate the trustworthiness of new users or those with few interactions. The reasoning models infer trust using carefully-designed rules such as Subjective Logic [1]. However, such models tend to be overly idealized and lack adaptability into complex and dynamic real-world scenarios. In contrast, the ML models can learn intricate patterns from big data to make them applicable across different contexts, showing significant potential.

Graph Neural Networks (GNNs) are a relatively new type of ML model designed to handle graph-structured data, offering outstanding advantages in trust evaluation [2]. First, trust relationships between entities in various cybersystems can be naturally represented as graph data, where nodes represent different types of entities and edges depict their trust relationships. Second, the message-passing mechanism of GNNs enables trust propagation and aggregation, thus satisfying basic trust properties, such as conditional transferability. Moreover, GNNs provide an end-to-end evaluation manner, allowing raw graph data to be directly fed into GNN models, thereby simplifying the process of evaluation.

We can find other reviews on trust evaluation or GNNs in the literature. For instance, Wang *et al.* [1] reviewed ML methods for trust evaluation, but did not discuss GNNs. While some reviews have addressed GNNs and their variants, their use in trust evaluation has not yet been surveyed. The absence of a dedicated literature review on GNN-based trust evaluation, along with a lack of comprehensive evaluation criteria and technical classifications, hinders the understanding of advances offered by GNNs for the open problems in this field.

To this end, in this paper, we review GNN models for trust evaluation. We begin by proposing a set of criteria that GNN-based trust evaluation models should follow, concerning trust-related attributes, correctness, functionality, and overhead. Then, we introduce a taxonomy of these models and conduct a thorough review by employing our proposed criteria as a measure to justify their pros and cons. Additionally, we perform experimental analyses on cutting-edge GNN-based trust evaluation models to uncover in-depth insights. Finally, according to our review and experimental findings, we identify key challenges and propose future research directions to advance the research on GNN-based trust evaluation.

II. Background Knowledge

This section introduces the fundamental knowledge of trust and trust evaluation, as well as the basics of GNNs and a workflow of GNN-based trust evaluation.

* indicates equal contribution.

T. X. Luo, J. Wang, Z. Yan (corresponding author), are with the State Key Lab of ISN, School of Cyber Engineering, Xidian University, Xi'an, Shaanxi, 710026 China (email: luotingxi930@foxmail.com, jwang1997@stu.xidian.edu.cn, zyan@xidian.edu.cn).

Erol Gelenbe is with the Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Gliwice, Poland, also with King's College London, and CNRS I3S Université Côte d'Azur, Nice, France (email: seg@iitis.pl).

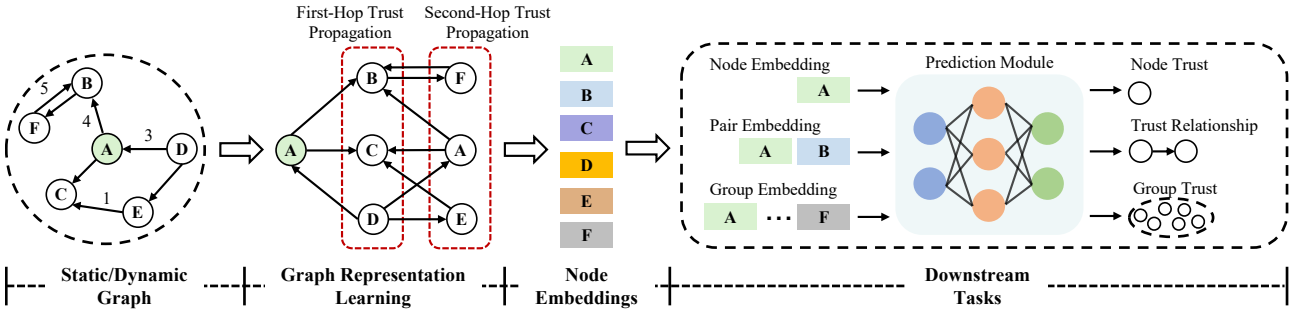


Fig. 1: Workflow of GNN-based trust evaluation.

A. Trust and Trust Evaluation

Trust can be defined as the confidence or belief of one entity in another regarding a specific context. It can mitigate potential risks inherent in social communications and interactions. Trust holds several key properties, including subjectivity, dynamicity, context-awareness, asymmetry, conditional transferability, and composability [2], explained in Section III-A1.

Trust evaluation or trust assessment is the process of quantifying trust of a trustor in a trustee by taking the factors that affect trust into consideration, including security, usability, maintainability, reliability, etc. Herein, the "trustor" refers to an entity placing trust, while the "trustee" is an entity being trusted. Using ML for trust evaluation is also known as trust prediction, as ML typically uses past interaction data to predict future trust relationships. Trust evaluation is one of the important means for identifying internal and external attackers, mitigating uncertainty, and facilitating decision-making [2].

B. Graph Neural Networks

GNNs are a class of ML models specifically designed for graph-structured data that consist of nodes and edges. In such graphs, nodes can represent users, devices, and atoms in networks like social networks, industrial CPS, Internet of Things (IoT), and molecules, while edges represent connections or interactions between these nodes [3, 4]. The core of GNNs is their message-passing mechanism: each node receives messages (including node attributes and edge features) from its neighbors, aggregates them, and updates its own representation (i.e., embedding) based on the aggregated messages [5]. This process is performed iteratively over multiple hops, enabling GNNs to capture complex patterns and dependencies within the graph.

C. GNN for Trust Evaluation

GNNs have shown impressive potential across diverse applications, such as social networking analysis, traffic analysis, and resource allocation [3]. This success has attracted researchers to revolutionize trust evaluation using GNNs. Fig. 1 presents a workflow of GNN-based trust evaluation. First, trust relationships between nodes are modeled as a static or dynamic graph, where edge weights represent varying levels of trust. Then, this graph is fed into a graph representation learning module. Leveraging the message-passing mechanism, trust information is propagated and aggregated based on the graph structure, resulting in the generation of node embeddings. Finally, a prediction module utilizes different combinations of these embeddings to predict/quantify node trust, trust relationships between nodes, or group trust.

GNN for Trust vs. Trustworthy GNNs: GNN-based trust evaluation models are designed to predict trust levels, whereas the trustworthiness of GNNs focus on such aspects as robustness, privacy, fairness, and explainability of GNN models. Although both are related to "trust", they differ in research objectives and application scenarios. This paper focuses on the applications of GNNs in trust evaluation.

III. Evaluation Criteria

This section proposes a set of criteria in terms of trust-related attributes, correctness, functionality, and overhead, according to which we can comprehensively analyze the strengths and weaknesses of existing GNN-based trust evaluation models in order to identify key challenges. The taxonomy of these criteria is illustrated in Fig. 2.

A. Trust-Related Attributes

Trust-related attributes are integral elements that impact the accuracy of trust evaluation models. These attributes encompass trust properties, node characteristics, and heterogeneity. Models considering these attributes can well represent real-world trust relationships and have potential to achieve high accuracy.

1) Trust Properties: Trust has some basic properties that should be incorporated into GNN-based trust evaluation models, as listed below. Although subjectivity is one intrinsic nature of trust, we exclude it herein for two reasons: i) GNNs aim to assess trust using objectively available data (e.g., interactions), and ii) interaction data, such as ratings given by a trustor to a trustee, inherently reflect the trustor's subjective perspective.

- Dynamicity: Trust changes with new interactions and tends to decay over time. For instance, recent interactions are typically more important than historical ones [6]. Hence, a trust evaluation model should take into account trust dynamicity and discriminate the significance of trust relationships formed over time.
- Context-awareness: The level of trust between a trustor and a trustee varies across different contexts. For instance, a person may trust someone's mathematical ability but

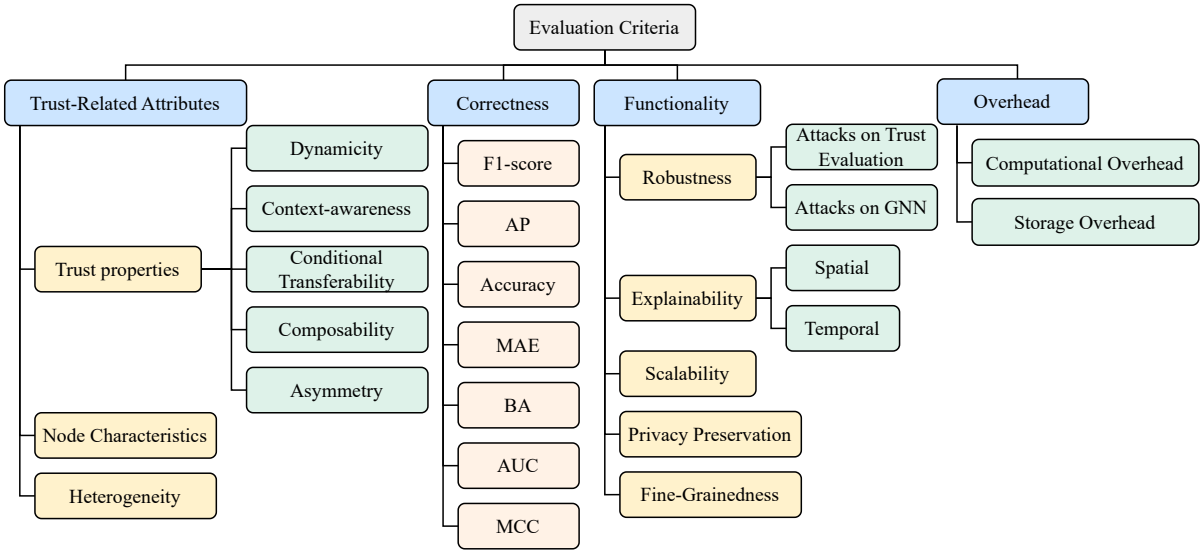


Fig. 2: Taxonomy of evaluation criteria.

not his athletic skills. As such, considering contextual information is crucial for fine-grained trust evaluation.

- Conditional transferability: Trust is propagative or conditionally transitive, suggesting that one might extend trust in another entity recommended by a trusted intermediary. Recognizing this nuanced property is vital for accurately modeling how trust extends across connections, leading to the formation of trust chains.
- Composability: Due to conditional transferability, a trustor may establish multiple trust chains towards a trustee. In this case, the trustor needs to combine trust information received from different chains to derive an overall trust level [5]. A trust evaluation model that considers composability can effectively extract and leverage sufficient information towards precise evaluation.
- Asymmetry: It implies that the level of trust one individual places in another may not be equally reciprocated. This property stems from differences in individual subjective perceptions. It also emphasizes the distinct roles in trust evaluation, where a node can act either as a trustor or a trustee. A trust evaluation model should account for these two possible roles to comprehensively capture the asymmetry of trust relationships.

2) Node Characteristics: Node characteristics encompass the inherent attributes of the nodes themselves. Taking social networks as an example, these attributes include age, occupation, and hobbies of user nodes. This information is crucial for evaluating social trust, as trust establishment is often related to user characteristics [7]. Thus, involving node characteristics is beneficial in developing effective GNN-based trust evaluation models.

3) Heterogeneity: Heterogeneity refers to graphs with more than two types of nodes or edges [8]. Compared to homogeneous graphs, heterogeneous graphs excel at modeling complex real-world networks [4]. For example, social networks show heterogeneity through diverse user-to-user and user-to-item interactions, both of which are valuable for trust evaluation. Consequently, GNN models should support heterogeneity to embrace rich information to gain high evaluation accuracy and practical applicability.

B. Correctness

Correctness quantifies the effectiveness of trust evaluation models using metrics such as F1-score, Average Precision (AP), Accuracy, Mean Absolute Error (MAE), Balanced Accuracy (BA), the Area Under the ROC Curve (AUC), and Matthews Correlation Coefficient (MCC). Each metric provides unique insights into model effectiveness and is applicable in different scenarios. For instance, BA and MCC remain effective when a dataset is imbalanced, while MAE is suitable for regression tasks where trust values are continuous. Additional details on these metrics can be found in references [2, 9, 10].

C. Functionality

Functionality refers to the functions that a GNN-based trust evaluation model supports. We consider five types of functions: robustness, explainability, scalability, privacy preservation, and fine-grainedness. The more functions a model supports, the more applicable it is across various contexts.

1) Robustness: Robustness refers to the capability of a GNN-based trust evaluation model to resist attacks, no matter they target the trust evaluation or the GNN model itself.

a) Attacks on trust evaluation: There are numerous attacks on trust evaluation [1], and we introduce five common types. Bad-mouthing and good-mouthing attacks occur when malicious nodes manipulate ratings to undermine the trustworthiness of a well-behaved node or boost the trustworthiness of a malicious node. On-off attacks (i.e., conflict behavior attacks) involve attackers alternately exhibiting honest and dishonest behaviors, allowing them to maintain a certain level of trust within a system, making them still possible to achieve destructive goals. Collusion attacks occur when multiple attackers conspire to manipulate the result of trust evaluation. In Sybil attacks, a malicious node has multiple

identities to make it possible to influence the evaluation on other nodes' trustworthiness.

b) Attacks on GNNs: Apart from trust-related attacks, inherent vulnerabilities within the GNN model can lead to adversarial attacks during both training and testing phases [2]. In the training phase, attackers can inject false or harmful samples into the training data to degrade the model's performance, known as poisoning attacks. In the testing phase, attackers manipulate input data to cause the model to produce incorrect evaluation results, known as evasion attacks.

2) Explainability: Explainability refers to how easily humans can understand the trust evaluation results or operations of a GNN model. GNNs function like a black box, making their results difficult to follow, reducing user trust and hindering their widespread adoption [2]. Therefore, studying explainability from both spatial and temporal aspects is crucial to enhance the model's trustworthiness and user acceptance.

3) Scalability: Scalability refers to the capability of a model to remain effective without a significant increase in overhead as the graph size increases. Real-world networks often exhibit very large scales. Therefore, GNN models need to address the challenge of scaling to large graph structures while maintaining high performance.

4) Privacy Preservation: Privacy Preservation refers to safeguarding the privacy of individual users' information during trust evaluation, preventing the disclosure or misuse of sensitive data. Trust evaluation often necessitates examining sensitive interactions between users. Thus, GNN models should incorporate privacy preservation measures to align with user expectations or government regulations.

5) Fine-Grainedness: Fine-grainedness describes the ability of a trust evaluation model to predict various levels of trust, which is critical for two reasons. First, by modeling trust at multiple levels or with a continuous digital value, the model can capture meaningful node representations. Second, this ability facilitates nuanced decision-making, making the model broadly applicable. Thus, fine-grained trust evaluation is highly preferred, as widely supported by traditional trust evaluation methods.

D. Overhead

Overhead represents the costs of training and using a GNN-based trust evaluation model, mostly including the costs spent on computation and storage. Simple yet effective models are preferred in both the ML community and practical applications. For example, IoT devices are inherently resource-constrained, making it infeasible to deploy heavy trust evaluation models. In this paper, we focus on examining the computational and storage overhead of these models using big O notation, while restricting our analysis to one-hop neighbors for simplicity.

IV. GNN-Based Trust Evaluation Models

In this section, we first present a taxonomy of existing GNN-based trust evaluation models, followed by a detailed review on them by employing the proposed criteria as an evaluation measure. Table I summarizes and compares the reviewed models regarding our proposed criteria except correctness. Furthermore, we perform an quantitative analysis on the correctness of cutting-edge evaluation models, as shown in Table II.

A. Taxonomy

Based on whether time is incorporated, we classify the existing models into two primary categories: static model and dynamic model, refer to Fig. 3. We further divide the static model into three categories: graph convolution, graph attention, and chain-based models, depending on the way of trust propagation. The dynamic model can be classified into discrete-time and continuous-time models according to temporal partitioning methods. In what follows, we introduce each category in detail.

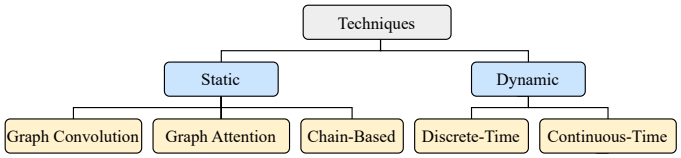


Fig. 3: Taxonomy of GNN-based trust evaluation models.

1) Static Model: The static model focuses on static graphs (or snapshots) captured at a specific time instance and learns node representations exclusively based on structural information (e.g., trust relationships). Depending on trust propagation methods, static models can be divided into three categories:

a) Graph convolution model: In the graph convolution model [9, 11–13], trust information is propagated with equal weights. Specifically, GNNs employ a convolutional operation to update node representations, leveraging features of individual nodes and their local neighborhoods. This is the most straightforward and efficient manner of trust propagation.

b) Graph attention model: The graph attention model [7, 8, 14] propagates trust information with varying weights by employing an attention mechanism. It assigns coefficients to adjacent nodes based on their significance to target nodes [7, 8], which allows the model to focus on critical information, such as node importance or special trust relationships, within a network.

c) Chain-based model: Essentially, the above two types of models propagate trust using the message-passing mechanism of GNNs. Instead, the chain-based model [5] explicitly propagates trust through trust chains, which are sequences of nodes with trust relationship extensions. By defining trust chains from one node to another, some basic trust properties can be effectively supported and enhanced. For instance, trust chains are inherently directional, making them compatible with trust asymmetry.

2) Dynamic Model: The dynamic model is capable of handling dynamic graphs, where trust relationships or node features change over time. It incorporates time-series information through carefully-designed time modeling methods, enhancing its ability to capture the temporal evolution of trust relationships. Depending on the method employed for time modeling, the dynamic models can be categorized into two categories:

a) Discrete-time model: The discrete-time model [2, 10] represents dynamic graphs using a sequence of time-ordered snapshots observed at different timeslots. It first learns spatial features within each snapshot and then captures temporal patterns across different snapshots using recurrent methods [10] or position-aware attention mechanisms [2]. The discrete-time models are efficient and simple but may lose structural dependencies across snapshots.

b) Continuous-time model: The continuous-time model [6] represents dynamic graphs using streaming interactions that update in real time. It employs a time encoding function to map a continuous-time domain into a vector space, fully utilizing the timestamps of interactions. While the continuous-time models excel at capturing subtle changes in trust relationships over time, the process of encoding temporal features could incur expensive overhead.

B. Review on Static Models

1) Graph Convolution Models: Lin *et al.* [9] proposed the first GNN-based trust evaluation model called Guardian in Online Social Networks (OSNs). They defined popularity trust (in-degree) and engagement trust (out-degree) to model trust asymmetry. By stacking multiple trust convolutional layers, OSN users are able to receive trust propagated from their multi-hop neighbors, thus satisfying conditional transferability and composability. Leveraging localized graph convolutions, Guardian accelerates trust evaluation by up to 2,827× with comparable accuracy compared to a neural network-based method, demonstrating its high scalability for large-scale networks. Additionally, it can predict trust relationships with four levels, owning potential to support fine-grained evaluation. However, Guardian studies a homogeneous network, ignoring the heterogeneous nature of real-world networks.

Building upon the architecture of Guardian [9], researchers have adapted it for various scenarios. Zhan *et al.* [11] developed a Trust Reinforcement Evaluation Framework (TREF) to assist worker recruitment in mobile crowdsourcing. This framework, based on Guardian, incorporates expert knowledge to fully address conditionally transitive and composable natures of trust. It introduces a trust benefit layer to select a worker group with mutual trust, which is particularly important for the fulfillment of crowdsourcing tasks through multiple-worker collaboration. Since this framework is rooted in Guardian, it meets the same criteria as Guardian. However, its scalability is not empirically validated, and the introduction of expert knowledge may limit its generality.

In addition to this, Jiang *et al.* [12] proposed a trust-based fraud detection model in Social Internet of Things (SIoT). This model uses a Guardian [9]-like architecture to derive user embeddings, and thus can also support conditional transferability and composability. Differently, it investigates trust in multi-relation scenarios and introduces a trust-aware neighbor difference aggregation method to amplify the differences between normal users and fraudsters. By assigning varying importance to user embeddings learned under different relations, this model addresses graph heterogeneity. Nevertheless, its scalability remains untested. Fine-grainedness is not available as the focus of this work is on fraud detection rather than trust evaluation.

Existing studies isolate users' preferences from their social relationships, while Wang *et al.* [13] argued that they influence each other. As such, they proposed JoRTGNN by considering this mutual influence. JoRTGNN learns users' trust embeddings using a user-user trust network based on Balance Theory. Meanwhile, it learns user interaction embeddings and item embeddings using a user-item interaction network through a graph convolution mechanism. By designing a joint loss for trust prediction and item recommendation, JoRTGNN facilitates their mutual influence. It achieves conditional transferability and composability by stacking multiple graph convolution layers. By incorporating both user-user and user-item relations, JoRTGNN supports heterogeneity. However, the inclusion of these diverse relations also limits its scalability. Moreover, JoRTGNN does not model two roles in trust evaluation, failing to satisfy asymmetry. Only trust and distrust relationships can be predicted, lacking fine-grainedness.

Additional Remarks: The above four models effectively capture intricate relationships between nodes through graph convolution mechanisms, surpassing traditional trust evaluation methods. However, they assign uniform weights to different neighbors during trust propagation and aggregation, which fails to accurately reflect reality. Additionally, they simplify network modeling by focusing on a static snapshot, which not only ignores the dynamic nature of trust but may also lead to the mistaken use of later interactions to predict past trust relationships. Contextual information and node characteristics, which are crucial for accurate evaluation, are not addressed at all. Moreover, the lack of considerations on robustness, explainability, and privacy preservation limits their practical significance. The computational and storage complexities of these models are shown in Table I.

2) Graph Attention Models: To address the issue of graph convolution models overlooking the importance of different neighboring nodes, Jiang *et al.* [7] proposed GATrust based on graph attention mechanisms in OSNs. GATrust considers multi-aspect properties of users, including user features (e.g., personal hobbies), topological structure, and known social relationships. Thus, node characteristics are well incorporated. Similar to Guardian [9], GATrust defines popularity trust and engagement trust to address asymmetry. By incorporating a graph attention layer, GATrust fuses these properties with varying importance, enabling conditional transferability and composability. It also supports fine-grained evaluation by modeling trust at different levels. However, the integration of multi-faceted user properties increases computational overhead and limits scalability. Moreover, heterogeneity is not supported.

Except for node characteristics, Badr *et al.* [14] argued that incorporating edge features can also improve evaluation accuracy. They proposed GBTrust in Peer-to-Peer (P2P) networks by taking into account edge direction and features, such as transaction value/frequency and local trust. GBTrust initially divides a graph into two subgraphs based on node roles in trust relationships, supporting asymmetry. It then employs an attention mechanism to differentiate the influence of edge directions and multi-dimensional features, providing a nuanced

TABLE I: Summary and comparison of GNN-based trust evaluation models.

Techniques		References	Scenarios	Trust-Related Attributes: Trust properties: Dyn	CA	CT	Com	Asy	NC	Het	Functionality: Robustness: ATE	AG	Explainability: Spa	Tem	Sca	PP	FG	Overhead: Computational Overhead	Storage Overhead
Static	Graph Convolution	Guardian [9]	OSN	×	×	√	√	√	×	×	×	×	×	-	√	×	√	$O(NF_2^2+EF_2)$	$O(E+F_2^2+NF_2)$
		TREF [11]	MC	×	×	√	√	√	×	×	×	×	×	-	-	×	√	$O(NF_2^2+EF_2)$	$O(E+F_2^2+NF_2)$
		T-FrauDet [12]	SIoT	×	×	√	√	√	×	√	×	×	×	-	-	×	-	$O(NF_2^2+EF_2+PF_2)$	$O(E+F_2^2+NF_2+PF_2)$
		JoRTGNN [13]	OSN	×	×	√	√	×	×	√	×	×	×	-	-	×	×	$O(NF_1^2+EF_1+PF_1)$	$O(E+F_1^2+NF_1+PF_1)$
	Graph Attention	GATrust [7]	OSN	×	×	√	√	√	√	×	×	×	×	-	-	×	√	$O(NF_3^2+EF_3)$	$O(E+F_3^2+NF_3)$
		GBTrust [14]	P2P	×	×	√	√	√	√	×	×	×	×	-	-	×	√	$O(NF_4^2+EF_4)$	$O(E+F_4^2+NF_4)$
		KGTrust [8]	SIoT	×	×	√	√	√	√	√	×	×	×	-	-	×	×	$O(NF_1^2+EF_1+PF_1)$	$O(E+F_1^2+NF_1+PF_1)$
	Chain-Based	TrustGNN [5]	OSN	×	×	√	√	√	×	×	×	×	√	-	-	×	√	$O(NF_2^2+EF_2)$	$O(E+F_2^2+NF_2)$
Dynamic	Discrete-Time	DTrust [10]	OSN	√	×	√	√	√	×	×	×	×	×	×	-	×	√	$O(NF_2^2+EF_2+TNH(F_2+H))$	$O(F_2^2+EF_2+TNH+(F_2+H)H)$
		MATA [15]	OSN	√	×	√	√	√	×	×	×	×	×	×	-	×	√	$O(NF_2^2+EF_2+TNH(F_2+H))$	$O(F_2^2+EF_2+TNH+(F_2+H)H)$
		TrustGuard [2]	Generic	√	×	√	√	√	×	×	√	×	√	√	√	×	√	$O(TNF_2^2+EF_2)$	$O(TNF_2^2+T^2+NF_2+F_2^2)$
	Continuous-Time	Medley [6]	OSN	√	×	√	√	×	×	×	×	×	×	×	-	×	√	$O(NF_3^2+EF_3)$	$O(E+F_3^2+NF_3)$

1. OSN: Online Social Network; SIoT: Social Internet of Things; MC: Mobile Crowdsourcing; P2P: Peer-to-Peer Network.
2. Dyn: Dynamicity; CA: Context-Awareness; CT: Conditional Transferability; Com: Composability; Asy: Asymmetry; NC: Node Characteristics; Het: Heterogeneity; ATE: Attacks on Trust Evaluation; AG: Attacks on GNNs; Spa: Spatial; Tem: Temporal; Sca: Scalability; PP: Privacy Preservation; FG: Fine-Grainedness.
3. N: Number of nodes; E: Number of edges/interaction types; P: Number of interaction types; T: Number of snapshots; H: The hidden state dimension of GRU; $F_1/F_2/F_3/F_4$: Embedding dimensions when considering one/two/three/four features.
4. √: supported/resisted; ×: not supported/resisted; -: not mentioned/related.

understanding of node relationships. By stacking multiple attention layers, GBTrust meets conditional transferability and composability. It also considers varying levels of trust, thus able to deliver a fine-grained evaluation result. However, its reliance on node and edge features specific to P2P networks hinders its generality. Heterogeneity is not considered.

Most models study homogeneous user-to-user trust networks, missing complex interactions in real-world heterogeneous networks. In response, Yu *et al.* [8] introduced KGTrust, which captures different semantic meanings of user and object nodes in SIoT. KGTrust consists of three layers. First, a embedding layer generates user embeddings from comments reflecting their preferences, as well as object embeddings from external knowledge. Thus, this layer fully considers nodes' inherent characteristics. Next, a heterogeneous convolutional layer defines trustor and trustee roles to support asymmetry, and applies a discriminative attention mechanism to handle various node types and interactions, addressing heterogeneity. By stacking multiple heterogeneous convolutional layers, KGTrust supports conditional transferability and composability. Finally, a prediction layer predicts trust relationships between any user pairs. However, KGTrust does not account for diverse trust levels, failing to support fine-grainedness.

Additional Remarks: The above three models employ graph attention mechanisms, enabling selective aggregation of trust information from different nodes, which provides a more realistic representation compared to graph convolution models. However, their overhead (refer to Table I) increases accordingly due to the need to compute node importance. Thus, how to enhance their scalability remains challenging. Similar to graph convolution models, these attention-based models fail to satisfy two basic trust properties: dynamicity and context-awareness, leaving a lot of room for improving evaluation accuracy. In addition, they lack support on robustness, explainability, and privacy preservation, which require further investigation to enhance their practical applicability.

3) Chain-Based Model: To explicitly consider transferability and composability, Huo *et al.* [5] proposed TrustGNN in OSNs. TrustGNN first introduces trust chains to define directional propagation patterns, modeling conditional transferability and asymmetry. Then, an attention mechanism is adopted to discriminately aggregate information across trust chains, supporting composability. Additionally, TrustGNN provides spatial explainability by visualizing the importance of each chain. Experimental results show that TrustGNN can handle networks with different levels of trust, thus satisfying fine-grainedness. However, other criteria like dynamicity and context-awareness are not discussed in this work. Its computational and storage complexities are presented in Table I.

C. Review on Dynamic Models

1) Discrete-Time Models: Static models focus on specific snapshots, failing to capture trust dynamicity. In response, Wen *et al.* [10] proposed DTrust in OSNs by splitting a dynamic graph into time-ordered snapshots. DTrust comprises three units. A static feature aggregation unit captures spatial features for users within each snapshot using a mechanism similar to Guardian, thus some basic trust properties can be supported as well. A dynamic feature unit takes a sequence of spatial features derived from different snapshots as its input, and learns the evolution of trust relationships via Gated Recurrent Unit (GRUs). A prediction unit receives two user embeddings containing both spatial and temporal features and then outputs their current or future trust relationship. DTrust offers fine-grained evaluation by evaluating trust with different levels. However, it lacks support on robustness, explainability, and scalability.

To enhance robustness, Jafarian *et al.* [15] proposed MATA, an extension of the DTrust [10] architecture that incorporates an attention layer and a reputation assessment module. Specifically, the attention layer helps identify on-off attackers who try to conceal their behavioral fluctuations over a long period by assigning varying weights to different snapshots. The reputation assessment module based on clustering can detect suspicious nodes. However, some parameters are required to be set in this module, which limits generality. MATA satisfies the same criteria as DTrust while showing additional resilience against bad/good-mouthing attacks and on-off attacks.

Wang *et al.* [2] proposed TrustGuard to simultaneously address dynamicity, robustness, and explainability for the first time. Unlike DTrust [10] and MATA [15], TrustGuard employs a position-aware attention mechanism to learn temporal patterns across time-ordered snapshots, which is more efficient than GRUs. Additionally, it designs a robust aggregator based

TABLE II: Correctness of representative GNN-based trust evaluation models.

Datasets		Metrics	Static Models		Dynamic Models	
			Graph Convolution	Graph Attention	Discrete-Time	Continuous-Time
			Guardian [9]	GATrust [7]	TrustGuard [2]	Medley [6]
Static Datasets	Advogato	MCC	0.595±0.006	0.599±0.001	**0.601±0.004**	-
		AUC	0.892±0.003	0.891±0.002	**0.896±0.003**	-
		F1-macro	0.697±0.005	**0.700±0.002**	0.699±0.005	-
	PGP	MCC	**0.739±0.002**	0.733±0.007	0.729±0.004	-
		AUC	0.887±0.005	0.890±0.004	**0.894±0.011**	-
		F1-macro	0.649±0.035	0.666±0.007	**0.682±0.038**	-
Dynamic Datasets	Bitcoin-OTC	MCC	0.224±0.071	0.192±0.076	**0.381±0.014**	0.364±0.006
		AUC	0.660±0.023	0.634±0.044	**0.727±0.007**	0.688±0.020
		F1-macro	0.559±0.064	0.541±0.059	**0.682±0.007**	0.632±0.007
	Bitcoin-Alpha	MCC	0.159±0.013	0.141±0.035	0.188±0.007	**0.203±0.017**
		AUC	0.530±0.071	0.516±0.033	**0.642±0.026**	0.599±0.016
		F1-macro	0.504±0.011	0.505±0.014	**0.589±0.004**	0.539±0.005

The best result is marked in **bold**; "-" indicates that the model is not applicable to the corresponding dataset.

on network homophily, propagating and aggregating trust by considering user similarity to counter malicious attacks. Beyond meeting the criteria satisfied by DTrust, TrustGuard demonstrates resilience to typical trust-related attacks as well as scalability to large-scale networks. To offer explainability, it visualizes user similarities to identify the importance of each neighbor and attention scores to illustrate the learned temporal patterns, offering insights into spatial and temporal dimensions.

Additional Remarks: DTrust [10] and TrustGuard [2] effectively capture temporal patterns using GRUs and a position-aware attention mechanism, respectively, while MATA [15] integrates both techniques to deal with monotonic assumption of GRUs. In comparison to the static models, the dynamic models consider an additional time dimension, which allows for accurate modeling of trust relationships by accounting for evolving interactions and behaviors over time. Regarding robustness, TrustGuard makes the first attempt to counter bad/good-mouthing attacks and on-off attacks by designing a robust aggregator. It also demonstrates that its spatial-temporal architecture can naturally counter on-off attacks. Differently, MATA introduces an attention layer to deal with inconsistent behaviors across snapshots and a reputation assessment module to identify suspicious nodes. However, there is still a significant gap in developing robust GNN-based trust evaluation models that can defend against various types of attacks on trust evaluation, as well as attacks on GNN model itself. Moreover, all these models do not consider context-awareness, a basic nature of trust. They focus solely on user-to-user interactions, overlooking heterogeneity and node characteristics. Essential privacy is not protected during trust evaluation, either. Refer to Table I, their computational and storage complexities are higher than static models due to temporal information modeling.

2) Continuous-Time Models: In cases of high event frequency or short time intervals, discrete-time models may miss precise temporal information. In response, Lin *et al.* [6] proposed Medley, which uses continuous-time representations in OSNs. Medley incorporates three types of embeddings: user, time, and interaction embeddings. User embeddings are generated based on users' social ties. Functional time encoding maps timestamps into high-dimensional vectors, fully leveraging each timestamp. Interaction embeddings showing different trust levels are modeled to support fine-grained evaluation. All embeddings are propagated and aggregated through attention-based layers, satisfying conditional transferability and composability while identifying the importance of interactions formed over time. However, Medley does not explicitly model the two distinct roles of a trust relationship, thus cannot fully support asymmetry. It also struggles with high overhead when interactions are frequently updated, making it unsuitable for large-scale networks. Other criteria like context-awareness and heterogeneity are not investigated in this work. Its computational and storage complexities are presented in Table I.

D. Experimental Analysis of Cutting-edge Evaluation Models

In this subsection, we quantitatively evaluate four representative **GNN-based trust evaluation models**: Guardian [9], GATrust [7], TrustGuard [2], and Medley [6]. The evaluation focuses on their correctness in predicting static or dynamic trust relationships between nodes, referred to as an **edge classification task**. We utilize four **datasets** collected from different scenarios. Advogato and PGP are static datasets [9] that come from online social networks for open-source software developers and public certification networks, respectively. Both datasets include four levels of trust. Bitcoin-OTC and Bitcoin-Alpha are dynamic datasets [6] collected from an open market where users can make transactions using Bitcoins. These datasets have two levels of trust. For **evaluation metrics**, we employ Matthews Correlation Coefficient (MCC), the Area Under the ROC Curve (AUC), and F1-macro that are well-suited for handling imbalanced datasets, as suggested in [2]. For **implementation details**, our experiments are conducted on a machine equipped with an Intel(R) Core(TM) i7-12700K CPU and an NVIDIA GeForce RTX 3060 GPU. The GNN models are implemented using the Pytorch framework (version 1.8.1 + cu111). In addition, for dynamic datasets, trust relationships are chronologically split into 80% for training and 20% for testing. For static datasets, the trust relationships are randomly split in the same 80%-20% ratio. We report the average results obtained from 5 runs for each experiment.

As shown in Table II, we observe that Guardian, GATrust, and TrusrGuard are comparable on static datasets, with TrustGuard slightly ahead. This can be attributed to the simplicity of the static datasets, which limit the performance of dynamic models. When it comes to dynamic datasets, we find that dynamic models consistently outperform static ones. This superiority is due to dynamic models' ability to effectively capture temporal dependencies and adapt to evolving trust relationships over time. Additionally, Guardian exhibits slightly better performance than GATrust, likely because GATrust's high complexity reduces its generalizability across different datasets. While TrustGuard achieves the best overall performance, it is worth noting that determining an appropriate observation frequency (which affects the number of snapshots) for discrete-time models is a challenging task. This issue would somehow affect the practicality of these models.

V. Key Challenges and Future Research Directions

Through the above serious review, we identify key challenges faced by current research and suggest future research directions accordingly in this section.

A. Limited Support on Context-Awareness

None of the existing models support context-awareness, leading to coarse-grained evaluation results. Since trust relationships vary across contexts, it is essential to incorporate specific contextual information and identify context-aware trust. However, the lack of sufficient contextual information and fine-grained labels in existing datasets make investigating context-aware trust evaluation particularly challenging. Future research could explore heterogeneous networks enriched with contextual information. For example, in networks like Epinions and Ciao [8], both user nodes and item nodes exist, and item categories can serve as contexts. In this case, researchers could first make full use of item categories to predict context-aware trust, and then link these predictions with labeled overall trust, thus avoiding the reliance on fine-grained labels. In addition, Large Language Models (LLMs), with their extensive knowledge base and strong reasoning abilities, show great potential to generate context-specific trust labels on datasets and produce high-quality embeddings based on limited contextual information.

B. Inadequate Consideration of Robustness

Model robustness against poisoning or evasion attacks is rarely studied in the current literature, with only TrustGuard [2] and MATA [15] making efforts to defend against attacks on trust evaluation. However, the vulnerabilities of GNN itself deserve significant attention since adversarial attacks are more disruptive than trust-related attacks. Some defense strategies from the image domain could be borrowed for enhancing model robustness. For instance, adversarial training is effective against evasion attacks. Additionally, graph preprocessing techniques based on Jaccard or cosine similarities, as well as low-rank based defenses are useful for filtering adversarial edges. Incorporating the properties of clean graphs as constraints within the loss function can also facilitate the training of a robust GNN model. Last but not least, LLMs can capture the inherent features of nodes using their textual attributes to generate node embeddings. When integrated with these LLM-derived embeddings, GNNs may become resilient to adversarial attacks since the rich semantic information offered by LLM within the embeddings strengthens their ability to distinguish between normal and malicious node behaviors.

C. Limited Explainable Evaluation Results

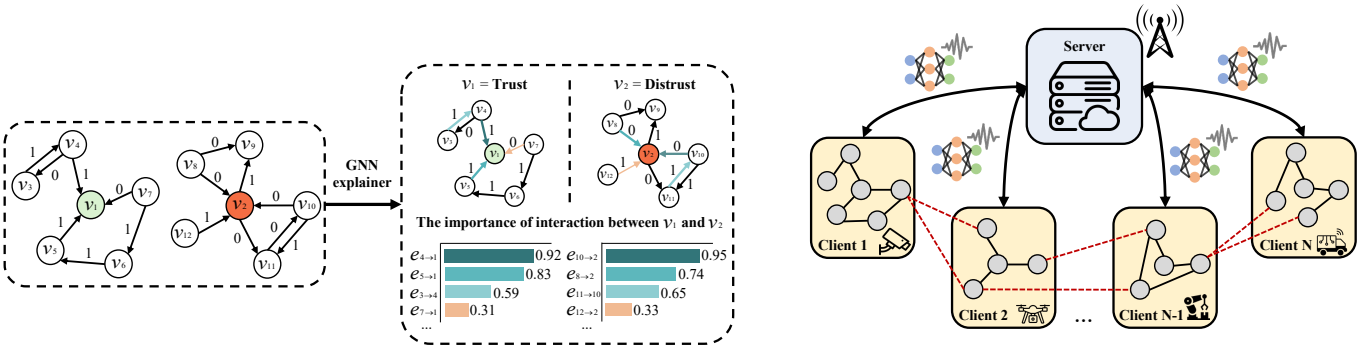
Existing models make significant efforts for improving the accuracy of trust evaluation but neglect the explainability of evaluation results from a human perspective. This oversight impacts user acceptance on these models. Future research could incorporate explainability tools to address this issue. For instance, visualizing the parameters learned by a GNN model can help users understand how an evaluation result is derived [2, 5]. Additionally, identifying small yet representative subgraphs via specific explanation methods is also beneficial. Fig. 4(a) illustrates an example of explainable evaluation results. By using a GNN explainer, users can clearly understand how each evaluation result is derived and identify the interactions that have the greatest influence on the final result. For example, v_1 is deemed trustworthy because it receives two trusted relationships with high weights, while a distrusted relationship is assigned a low weight. This visualization not only aids users in assessing the rationality of the evaluation results but also enhances user trust in the GNN model.

D. Lack of Privacy Preservation

None of the existing models support privacy preservation, which is highly expected by users and required by governmental policies. Federated Learning (FL) is a distributed ML framework that offers a degree of privacy preservation. Applying such an FL framework with differential privacy techniques is a promising approach to achieve privacy-preserving trust evaluation. Considering a CPS, as shown in Fig. 4(b), where devices serve as FL clients, while a base station acts as an FL server. Each device constructs an interaction subgraph with its connected peers to evaluate their trustworthiness. However, due to resource constraints, devices are unable to collect a complete graph, resulting in low accuracy of local models. Additionally, privacy regulations prohibit the direct upload of raw data to the server. In this context, although FL-enabled GNNs present an effective solution by facilitating collaborative learning without compromising privacy, new challenges arise when FL meets GNNs, e.g., how to deal with missing links across local subgraphs, which requires special investigation.

E. Inefficiency of Dynamic Models

Existing models suffer from scalability issues due to high complexity to capture the dynamic nature of trust. Continuous-time models consider all timestamps, leading to high costs, while discrete-time models reduce these costs by simplifying



(a) An example of explainable evaluation results

(b) FL-enabled GNNs for privacy-preserving trust evaluation

Fig. 4: Examples of explainable and privacy-preserving GNN-based trust evaluation.

dynamic graph representations, yet still face challenges when being applied to large-scale graphs. From a model perspective, it is promising to design simple architectures or consider non-learnable time encoding techniques. From a data perspective, neighbor sampling can be adopted to select limited yet important neighbors for information propagation and aggregation. Improvements in both aspects show great potential to enhance the scalability of existing models.

F. Limited Application Scenarios

Research on GNN-based trust evaluation is still in its infancy, primarily focusing on social networks. However, other promising application scenarios [3], such as 6G heterogeneous networks, computing power networks, and industrial CPS, require further investigation. The reasons are as follows: (i) Real-world networks can be naturally modeled as graphs for effective GNN analysis. (ii) There is a pressing need to establish trust within these networks. For example, it is challenging to ensure that all nodes across different network domains in 6G networks are trustworthy. In such cases, trust evaluation becomes crucial for understanding the dynamically changed trust status of the network, facilitating trustworthy networking, as outlined by ITU Recommendation Y.3053.

VI. Conclusion

In this paper, we review existing trust evaluation models that employ GNNs. We first propose both qualitative and quantitative criteria to evaluate trust evaluation performance. Then, we provided a technical taxonomy of existing models, and employed our proposed criteria to conduct an in-depth review and analysis of existing GNN models for trust evaluation following this taxonomy. Moreover, we compared the effectiveness and validity of some cutting-edge models with regard to trust using commonly used datasets. As a result, we identified several key challenges and proposed future directions that advance the research on GNN-based trust evaluation.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grants U23A20300; in part by the Key Research Project of Shaanxi Natural Science Foundation under Grant 2023-JC-ZD-35; in part by the Concept Verification Funding of Hangzhou Institute of Technology of Xidian University under Grant GNYZ2024XX007; in part by the 111 Project under Grant B16037; in part by the EU Horizon DOSS Project under Grant Agreement No. 101120270; in part by the Fundamental Research Funds for the Central Universities; and in part by the Innovation Fund of Xidian University.

References

[1] J. Wang, X. Jing, Z. Yan, Y. Fu, W. Pedrycz, and L. T. Yang, "A survey on trust evaluation based on machine learning," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–36, 2020.

[2] J. Wang, Z. Yan, J. Lan, E. Bertino, and W. Pedrycz, "Trustguard: Gnn-based robust and explainable trust evaluation with dynamicity support," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 5, pp. 4433–4450, 2024.

[3] J. Suárez-Varela, P. Almasan, M. Ferriol-Galmés, K. Rusek, F. Geyer, X. Cheng, X. Shi, S. Xiao, F. Scarselli, A. Cabellos-Aparicio, and P. Barlet-Ros, "Graph neural networks for communication networks: Context, use cases and opportunities," *IEEE Network*, vol. 37, no. 3, pp. 146–153, 2022.

[4] J. Li, R. Zheng, H. Feng, M. Li, and X. Zhuang, "Permutation equivariant graph framelets for heterophilous graph learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 9, pp. 11 634–11 648, 2024.

[5] C. Huo, D. He, C. Liang, D. Jin, T. Qiu, and L. Wu, "Trustgnn: Graph neural network-based trust evaluation via learnable propagative and composable nature," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 10, pp. 14 205–14 217, 2024.

[6] W. Lin and B. Li, "Medley: Predicting social trust in time-varying online social networks," in *Proceedings of IEEE Conference on Computer Communications*, 2021, pp. 1–10.

[7] N. Jiang, W. Jie, J. Li, X. Liu, and D. Jin, "Gatrust: A multi-aspect graph attention network model for trust assessment in osns," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 6, pp. 5865–5878, 2023.

[8] Z. Yu, D. Jin, C. Huo, Z. Wang, X. Liu, H. Qi, J. Wu, and L. Wu, "Kgtrust: Evaluating trustworthiness of siot via knowledge enhanced graph neural networks," in *Proceedings of the ACM Web Conference*, 2023, pp. 727–736.

[9] W. Lin, Z. Gao, and B. Li, "Guardian: Evaluating trust in online social networks with graph convolutional networks," in *Proceedings of IEEE Conference on Computer Communications*, 2020, pp. 914–923.

[10] J. Wen, N. Jiang, J. Li, X. Liu, H. Chen, Y. Ren, Z. Yuan, and Z. Tu, "Dtrust: Toward dynamic trust levels assessment in

time-varying online social networks," in *Proceedings of IEEE Conference on Computer Communications*, 2023, pp. 1–10.
[11] Z. Zhan, Y. Wang, P. Duan, A. M. V. V. Sai, Z. Liu, C. Xiang, X. Tong, W. Wang, and Z. Cai, "Enhancing worker recruitment in collaborative mobile crowdsourcing: A graph neural network trust evaluation approach," *IEEE Transactions on Mobile Computing*, vol. 23, no. 10, pp. 10 093–10 110, 2024.
[12] N. Jiang, W. Gu, L. Li, F. Zhou, S. Qiu, T. Zhou, and H. Chen, "Tfd: Trust-based fraud detection in siot with graph convolutional networks," *IEEE Transactions on Consumer Electronics*, pp. 1–12, 2024.
[13] G. Wang, H. Wang, J. Gong, and J. Ma, "Joint item recommendation and trust prediction with graph neural networks," *Knowledge-Based Systems*, vol. 285, 2023, Art. no. 111340.
[14] B. Bellaj, A. Ouaddah, A. Mezrioui, N. Crespi, and E. Bertin, "Gbtrust: Leveraging edge attention in graph neural networks for trust management in p2p networks," in *Proceedings of IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, 2023, pp. 1272–1278.
[15] B. Jafarian, N. Yazdani, and M. S. Haghighi, "Using attentive temporal gnn for dynamic trust assessment in the presence of malicious entities," *Expert Systems with Applications*, vol. 260, 2025, Art. no. 125391.

Tingxi Luo received the B.S. degree in information security from Tianjin University of Technology in 2023. He is currently pursuing the master's degree with the School of Cyber Engineering, Xidian University. His research interests include graph neural networks, trust evaluation, and trustworthy artificial intelligence.

Jie Wang received the B.S. degree in network engineering from Xidian University in 2020, where he is currently pursuing the Ph.D. degree with the School of Cyber Engineering. His research interests include graph neural networks, trust evaluation, and trustworthy artificial intelligence.

Zheng Yan (Fellow, IEEE) is currently a Full Professor with Xidian University, China. Her research interests are in trust, security, privacy, and data analytics. She authored over 410 peer-reviewed publications and solely authored two books. She is an inventor of 212 patents, more than 150 of which have been adopted by industry with wide applications. She was invited to delivery more than 30 keynotes and talks in international conferences. She served or is serving as the Editor-in-Chief of Information Sciences and an area/associate/guest editor for over 60 journals. She is also a fellow of IET, AAIA, and AIIA. She served as the general or program chair for over 40 international conferences. She initiated IEEE international conference on Blockchain as the Steering Committee Co-Chair. Recently, she achieved the Distinguished Inventor Award of Nokia, the N²Women: Stars in Computer Networking and Communications, the IEEE TCSC Award for Excellence, the IEEE TEMS Distinguished Leadership Award, the ELEC Impact Award, the IEEE ComSoc TCBD Best Journal Paper Award, and several best paper awards.

Erol Gelenbe (Life Fellow, IEEE) is a Professor in the Institute of Theoretical and Applied Informatics of the Polish Academy of Sciences (IITIS-PAN), Poland, a Visiting Professor at King's College, London, UK, and a Researcher at the I3S CNRS Laboratory, Université Côte d'Azur. Also a Fellow of ACM, he previously held the Dennis Gabor Chair at Imperial College London. His research addresses quantitative methods for the design and evaluation of computer systems and networks, including Cybersecurity, Sustainability and Quality of Service. Recipient of several prizes including the ACM-SIGMETRICS Life-Time Achievement Award (2008), the Grand Prix France-Télécom of the French Academy of Sciences (1996), the UK IET Oliver Lodge Medal for Innovation (2010), and the Mustafa Prize (2017), he is an elected Fellow of Academia Europaea, the French National Academy of Technologies, the Turkish Science Academy, the Royal Academy of Belgium, the Polish Academy of Sciences, and several other academies and professional societies.