Design of Retransmission Mechanism for Decentralized Inference with Graph Neural Networks

Jiaying Zhang, Yutong Jiang, Xingyu Liu, Mengyuan Lee, Huiguo Gao, and Guanding Yu
College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China
E-mail: yuguanding@zju.edu.cn

***Abstract*—Graph neural network (GNN) is widely applied in various fields, especially for graph data. Moreover, it is an effective technique for decentralized inference tasks, where information exchange among neighbors relies on wireless communications. However, wireless channel impairments and noise decrease the accuracy of prediction. To remedy imperfect wireless transmission and enhance the prediction robustness, we propose a novel retransmission mechanism with adaptive modulation that could select the appropriate modulation order adaptively for each retransmission. Compared to the traditional method that determines the modulation order based on bit error ratio (BER), we bring in a new indicator called robust prediction to select an appropriate modulation order for each transmission. Under the requirement of prediction robustness, the error-tolerance of GNNs is exploited and a higher modulation order can be used in the proposed mechanism compared with the traditional method, thus reducing the communication overhead and improving the data rate. Meanwhile, we combine the signals of different retransmission with the soft-bit maximum ratio combine (SBMRC) technique. Simulation results verify the effectiveness of the proposed retransmission mechanism.**

I. Introduction

Nowadays, graph neural network (GNN) is a graph analysis tool widely applied in various fields, such as computer vision [4] and intelligent traffic [5]. Given that a wireless network can be naturally modeled as a graph model, GNN can be used to solve corresponding optimization problems [3]. Meanwhile, GNN is an effective technology for decentralized control and management in wireless networks since the inference stage of GNNs can be implemented in a decentralized manner where information exchange among neighbors relies on wireless communications. The main bottleneck, however, is wireless channel impairments that deteriorate the prediction robustness of GNN.

To overcome this obstacle, retransmission is an effective method. Recently, a retransmission mechanism based on the maximal-ratio combining (MRC) is proposed in [3] for the decentralized inference with GNNs. However, this work does not take adaptive modulation into account and assumes that all nodes transmit at the same data rate for simplicity. The proposed mechanism in [3] can be further improved by adjusting the modulation order for each transmission. Generally, the modulation order is determined based on the channel quality. Specifically, the 3rd Generation Partnership Project (3GPP) protocol claims that the highest modulation order with block error ratio (BLER) < 0.1 is generally selected. However, this method neglects the error-tolerance of GNNs and leads to unnecessary high demand on communication resources for inference tasks. Meanwhile, the authors in [8] propose an adaptive modulation scheme for wireless federated edge learning, which achieves the trade-off between the learning latency and the convergence rate caused by stochastic channel error. Inspired by above ideas, we propose a novel retransmission mechanism with adaptive modulation for the decentralized inference tasks of GNNs.

Specifically, we first model the wireless network as a graph model and use GNN binary classifier to solve its corresponding binary classification problem. Each node can get its predicted label by exchanging information among its neighbors through wireless channels. For each transmission, we propose an adaptive modulation algorithm to select the appropriate modulation order based on channel conditions and prediction requirements. Appropriate modulation order not only aims at reaching the prediction requirement, but also reaches a higher data rate compared with the traditional method. Meanwhile, the technique of soft-bit maximum ratio combine (SBMRC) [6] is adopted to combine the signals from different transmissions with different modulation orders, thus decreasing the bit error ratio (BER) and improving prediction accuracy. Simulation results show that the proposed method can successfully achieve prediction accuracy target with few retransmissions. Furthermore, a higher data rate can be reached by the proposed method than the traditional method while achieving the same transmission requirement.

The remainder of this paper is organized as follows. The problem formulation and the GNN binary classifier are introduced in Section II. Section III introduces the proposed retransmission mechanism with adaptive modulation for decentralized GNNs. The test results and performance analysis of the proposed method are presented in Section IV. Finally, we conclude this paper in Section V.

II. System Model and Problem Formulation

As depicted in Fig. 1, we consider a wireless communication system with the node set $\mathcal{V}$ and the edge set $\mathcal{E}$, which can be represented as a wireless graph $G(\mathcal{V}, \mathcal{E})$.

Specifically, each device is modeled as a node. Two nodes are neighbors if they can communicate with each other, and we use an edge to connect them.

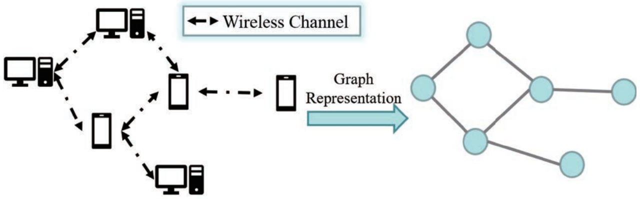

Fig. 1. A general wireless network with $N = 6$ nodes and its corresponding wireless graph model.

A. Binary Classification Problem

In this paper, we focus on the binary classification problem in wireless network for simplicity, which is widely found in wireless networks [3] and can be extended to more complicated problems. Specifically, each node $v \in \mathcal{V}$ can be classified into two classes and be labeled as 1 or -1. Meanwhile, each node $v \in \mathcal{V}$ holds an internal vector $\boldsymbol{x}_v$ which includes the needed characteristics for the classification task. For convenient storage and transmission, the vector is coded into a binary node feature vector $\boldsymbol{x}_v \in \{0,1\}^p$, where p is the feature dimension of the vector. Generally, we assume that each node knowns its neighbors in the network and can estimate the signal-to-noise ratios (SNRs) of the received signals.

B. Decentralized GNN Binary Classifier Model

The above binary classification problem over $G(\mathcal{V}, \mathcal{E})$ can be solved by the GNN binary classifier. The basic operations of GNN binary classifier are as follows. Generally speaking, by taking the graph adjacency matrix and node features as the input, GNN could learn the hidden state $\boldsymbol{h}_v$ of each node $v \in \mathcal{V}$. Then, the hidden state $\boldsymbol{h}_v$ can be put into a fully-connected layer for binary classification and the binary label could be predicted.

In this paper, we focus on the decentralized implementation of the GNN binary classifier. Firstly, we denote the set of neighbor of v as $N(v)$. Meanwhile, $\boldsymbol{X} \in \{0,1\}^{N \times p}$ is called node feature matrix consisting of feature vectors of all nodes, $\boldsymbol{\theta} \in \mathbb{R}^{p \times D}$ is the matrix of learning parameters, and $\hat{\boldsymbol{A}} \in \mathbb{R}^{N \times N}$ is called graph filter that defines how to integrate information of the neighborhood. Then, for each node $v \in \mathcal{V}$, its hidden state is updated as [3]

$$\hat{\boldsymbol{h}}_v = \hat{\boldsymbol{a}}_v \boldsymbol{X}_v \boldsymbol{\theta} + \hat{a}_v \boldsymbol{x}_v \boldsymbol{\theta}, \tag{1}$$

$$\boldsymbol{h}_v = \sigma^{(r)}(\hat{\boldsymbol{h}}_v), \tag{2}$$

where $\sigma^{(r)}(\cdot)$ is the ReLU function, $\boldsymbol{X}_v = \boldsymbol{X}_{\{N(v),:\}} \in \{0,1\}^{|N(v)| \times p}$ is defined as the sliced node feature matrix, $\hat{\boldsymbol{a}}_v = \hat{\boldsymbol{A}}[v,:]_{\{N(v)\}} \in \mathbb{R}^{1 \times |N(v)|}$ keeps the graph filter entries needed by node v to integrate information from nodes in $N(v)$. After getting the hidden state of each node v, we input it into a fully-connected layer for binary classification and the operation can be written as [3]

$$y_v = \sigma^{(s)}\left(\boldsymbol{h}_v \boldsymbol{\omega} + b\right), \tag{3}$$

where $y_v \in \mathbb{R}$ is the predicted output value of node v and $\boldsymbol{\omega} \in \mathbb{R}^{D \times 1}$ is the vector of learning weights. Finally, we can get the predicted binary label of node v and denote it as [3]

$$c_v = sgn(y_v - 1). \tag{4}$$

Therefore, the above operations in (1) - (4) at node v could be summarized as [3]

$$c_v = J\left([\hat{\boldsymbol{a}}_v, \hat{a}_v], [\boldsymbol{X}_v, \boldsymbol{x}_v]; \mathcal{W}\right). \tag{5}$$

From (5), each node v could get the predicted label c_v by using the GNN binary classifier in the decentralized manner. And the inference procedure of each node is as follows. First, the target node $v \in \mathcal{V}$ sends a request to its neighbors $u \in N(v)$. Then, each neighbor node u sends its internal vector $\boldsymbol{x}_u$ to the target node after receiving the request. Finally, node v can complete the classification task and get the label by following (5). Similarly, all nodes in the graph could predict their label in the decentralized manner.

However, the received signal $\boldsymbol{x}_u$ will be contaminated while transmitting through fading and noisy wireless channel, thus resulting in incorrect inference. Therefore, we need to remedy imperfect wireless transmission and enhance the prediction robustness by developing appropriate methods.

In order to solve the above problem, retransmission is needed. For each transmission, the traditional method usually sets a hard BER bound to choose the modulation order. But this method could result in high communication overhead and low data rate. This motivates us to propose a new mechanism by exploiting the error-tolerance of GNNs.

III. Retransmission Mechanism With Adaptive Modulation

In this section, we will introduce our proposed retransmission mechanism to enhance the prediction robustness. For simplicity, we further assume that the communication system is uncoded.

A. Algorithm Overview

To exploit the error-tolerance of GNNs, we adopt robustness requirement $p_v^{(t)}$ as the transmission metric to propose a novel retransmission mechanism with adaptive modulation. Specially, for each retransmission, we compare the robust probability $p_v^{(r)}$ of the current prediction with the pre-given robustness requirement $p_v^{(t)}$ of node v to determine whether the retransmission ends. By following

[3], the robustness probability $p_v^{(r)}$ of the current prediction of node $v \in \mathcal{V}$ is given by

$$p_v^{(r)} = \prod_{u \in N(v)} \left[\sum_{i=0}^{q_v^U} \binom{p}{i} (\varepsilon_{vu})^i (1-\varepsilon_{vu})^{p-i} \right], \quad (6)$$

where q_v^U is a solution to the optimization problem proposed in [3], which can be interpreted as the average maximum number of errors in each received signal of node v. Specifically, the optimization problem is formulated as

$$q_v^U \triangleq \max_q q, \quad (7)$$

subject to

$$z(\boldsymbol{q}_v^{'}, \hat{\boldsymbol{X}}_v) > 0, \forall q_{vu}^{'} \leq q, u \in N(v), \quad (8)$$

where $z(\cdot,\cdot)$ is defined as

$$z(\boldsymbol{q}_v, \hat{\boldsymbol{X}}_v) \triangleq \min_{\{\tilde{\boldsymbol{X}}_v, \boldsymbol{h}_v, \hat{\boldsymbol{h}}_v\}} \hat{c}_v(\boldsymbol{h}_v \boldsymbol{\omega} + b), \quad (9)$$

and the details can be found in [3]. Meanwhile, ε_{vu} is the BER of the signal transmitted from node u to node v, which can be estimated according to signal-noise ratio (SNR), modulation method and order.

While $p_v^{(r)} < p_v^{(t)}$, it means that the robustness requirement is not satisfied and the retransmission is needed. Specifically, the possible reasons are that the BER of some received signals is too large. Therefore, for the next retransmission, we need to find out these received signals with high BER and request the corresponding neighbors to retransmit these signals with appropriate modulation order, which will be discussed in details later. The retransmission ends until $p_v^{(r)} > p_v^{(t)}$. Moreover, the signals from different transmissions are combined by the SBMRC technique for a higher communication efficiency. Finally, the target node $v \in \mathcal{V}$ can get its robust predicted label 1 or -1. Compared to the traditional method, our mechanism can reach a higher data rate under the same prediction requirement.

B. Adaptive Modulation

As mentioned above, unqualified neighbors will retransmit the signal with appropriate modulation order when retransmission is needed. First, we need to identify unqualified neighbors. To do this, we need to find out the threshold of BER. In an average sense, we can transform equation (6) to

$$\sqrt[N(v)]{p_v^{(t)}} = \sum_{i=0}^{q_v^U} \binom{p}{i} \left(\varepsilon_{vu}^{'}\right)^i \left(1-\varepsilon_{vu}^{'}\right)^{p-i}, \quad (10)$$

where $\varepsilon_{vu}^{'} = \varepsilon_v^{'}, \forall u \in N(v)$. As $p_v^{(t)}$ is given, we can use numerical methods to calculate the value of $\varepsilon_v^{'}$ based on the monotonicity. If BER $> \varepsilon_v^{'}$, the corresponding neighbors need to retransmit.

Next, for each neighbor u who needs to retransmit, the retransmission modulation order should be selected. For the first transmission, the highest modulation order is adopted. For the subsequent retransmission, the modulation model can be computed based on q_v^U, BER, and $p_v^{(t)}$. Specifically, because of the limited choices of modulation order, the value of appropriate modulation order can be calculated by enumeration until the threshold of BER is achieved, which will be discussed in details later. At the same time, the value of q_v^U, BER and $p_v^{(r)}$ will be updated.

C. Signal Combiner

In our mechanism, for each node, signals are transmitted with different modulation orders for different transmissions. Therefore, the generally adopted combining technique, maximal-ratio combining (MRC) that can only combines signals with the same modulation level [9], [10], cannot be used in our mechanism. A technique named SBMRC that is a low complexity diversity combining scheme for signals with different modulation orders has been proposed in [6]. Meanwhile, SBMRC reduces complexity while exhibiting performance that is close to the optimal maximum likelihood detector (MLD). Specifically, taking MQAM as an example, the output SNR can be combined by [6]

$$\gamma_{\text{all}} = \sum_{i=0}^{L-1} d_{M_i}^2 \gamma_i, \quad (11)$$

where L is the total transmission times of a neighbor node, γ_i is the SNR of the i-th transmission, and M_i is the modulation order of the i-th transmission. Meanwhile, d_{M_i} is given by $d_{M_i} = \sqrt{\frac{3\log_2 M_i}{2(M_i-1)}}$. For example, $d_4 = 1, d_{16} = 0.6325$, and $d_{64} = 0.378$. Equation (11) suggests that the total SNR of the SBMRC is a weighted sum of the individual SNRs, and these weights depend on the modulation orders of the signals to be combined. In addition, the tight bounds on the average BER by averaging over the probability density functions (PDFs) of the SNRs is given as [6]

$$\tau Q(\sqrt{2\gamma_{\text{all}}}) < BER < Q(\sqrt{2\gamma_{\text{all}}}), \quad (12)$$

where τ is a constant related to the modulation levels of the signals to be combined and $Q(\cdot)$ is the Q-function defined as $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$. In this paper, we adopt the upper bound so that the value of τ can be neglected.

D. Choice of Modulation Order

There is a trade-off between BER and data rate while choosing the modulation order. In the proposed algorithm, the highest modulation order that makes the BER reach the threshold value is adopted. Specifically, we start with the modulation order M of the current transmission. Then γ_{all} can be calculated by (11) based on M, γ_M, and the historical transmission information. After that, the BER could be estimated by the upper bound in equation (12). Then the estimated BER is compared with the threshold value. If it is smaller than the threshold value, the modulation order of the current transmission is adopted. Otherwise, the modulation order is reduced by enumeration and the above checking process is repeated. The concrete adaptive modulation algorithm is summarized in Table I.

TABLE I
RETRANSMISSION MECHANISM WITH ADAPTIVE MODULATION

Algorithm 1 Retransmission Mechanism with Adaptive Modulation

for node v in $\mathcal{V}$ **do**
  initial transmission:
  Send transmission request to each neighbor $u \in N(v)$.
  Receive signals from neighbors:$\{\hat{\boldsymbol{x}}_u, \forall u \in N(v)\}$ (or say $\hat{\boldsymbol{X}}_v$) with $M_{u_0} = 64$.
  Compute current BERs for all received signals using graph neural network: $\{\varepsilon_{vu}, \forall u \in N(v)\}$.
  Use local GNN binary classifier to get the predicted label $\hat{c}_v$.
  Get q_v^U [3] and then calculate $p_v^{(r)}$ using (10).
  retran $= 1$.
  retransmission:
  while $p_v^{(r)} < p_v^{(t)}$ **do**
    Compute ideal BERs $\{\varepsilon'_{vu}, \forall u \in N(v)\}$ with robustness requirement $p_v^{(t)}$ using (6).
    $M_{u_{\text{retran}}}$=64.
    for neighbor u with $\varepsilon_{vu} > \varepsilon'_{vu}$ **do**
      adaptive modulation:
      Get γ_{all} using (11) and get BER ε_{vu} using (12).
      if $\varepsilon_{vu} < \varepsilon'_{vu}$ **then**
        break
      else
        $M_{u_{\text{retran}}} = M_{u_{\text{retran}}}/2$.
      end if
    end for
    Recompute BERs for all newly received signal.
    Recalculate $\hat{c}_v, q_v^U, p_v^{(r)}$ with newly received signals.
    retran $=$ retran $+ 1$.
  end while
end for

IV. TEST RESULTS

In this section, we will test the performance of the proposed mechanism. All codes are implemented by python3.8.

A. Simulation Setup

We conduct simulation on the synthetic random geometric graph data. Specifically, we consider a 2000 m by 2000 m two-dimensional square area with $N \in \{50, 100, 150, 200\}$ nodes distributed in the square area uniformly. Then, we assume that two nodes are neighbors if their distance is less than 500 m.

Moreover, each graph corresponds to a GNN binary classifier $J(\cdot, \cdot; \mathcal{W})$ for binary classification task. Following [7], each entry of the parameter in set $\mathcal{W}$ is generated independently from Gaussian distribution $\mathcal{N}\left(0, 10^2\right)$. Similarly, for all nodes in a graph, each entry of their node features $\{\boldsymbol{x}_u\}_{u \in N(v)}$ is generated independently following the Bernoulli distribution with a probability of success of 0.3. After that, the GNN binary classifier as shown in (5) is used to obtain the true labels of all nodes, which means that if the wireless transmission is perfect, the misclassification rate will be zero. For simplicity, we assume that all nodes have the same target robustness probability $p_v^{(t)}$ and use $64QAM$ for the initial transmission. And the modulation order candidates are $\{2, 4, 8, 16, 32, 64\}$. In this section, the results presented are the average performance over all nodes in the 50 testing graphs. Our simulation parameters are summarized in Table II.

TABLE II
SIMULATION PARAMETERS

Parameter	Value
Square Area	2000 m $\times$ 2000 m
Communication Radius	500 m
Bandwidth	10 MHz
Noise Spectral Density	-174 dBm/Hz
Path Loss Model	128.1+37.6log(d[km])
Shadowing Standard Deviation	8 dB
Target Robustness Probability, $p_v^{(t)}$	80%
Node Feature Dimension, p	32
Hidden State Dimension, D	32
Transmit Power, P	0.1 w

B. Simulation Results

We adopt the robust rate as a performance metric that indicates the percentages of nodes reaching transmission requirement at the initial transmission. The four lines in Fig. 2 represent scenarios with different numbers of nodes in one graph. The simulation result indicates that almost no node could reach transmission requirement at the first transmission with the increasing $p_v^{(t)}$. Therefore, retransmission always exists. Moreover, the higher the transmission requirement is, the lower the robust rate will be.

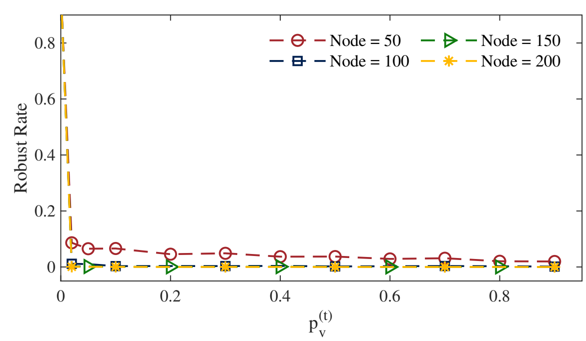

Fig. 2. Robust rate at first transmission.

In order to test the performance of the proposed mechanism, we compare the data rate $\mathcal{R}$ between the traditional method and the proposed retransmission mechanism with adaptive modulation. Specially, $\mathcal{R}$ is defined as the average transmission rate as

$$\mathcal{R} = \frac{np}{\sum_k \sum_b \frac{p}{\log_2 M_k^b}}, \tag{13}$$

where n is the number of neighbors of the node, p is the node feature dimension, and M_k^b means the modulation order of the k-th neighbor in the b-th transmission round.

Furthermore, the average data rate of all nodes in one graph is defined as the average data rate of this graph.

Fig. 3 shows that our proposed algorithm outperforms the traditional method for all $p_v^{(t)}$. To ensure a fair comparison, we set BER bound based on the value of the prediction requirement $p_v^{(t)}$. Specifically, after the first transmission, $\varepsilon_v^{'}$ can be estimated (about 10^{-4} order of magnitude) and the BER bound of the traditional method is set equally to the value of $\varepsilon_v^{'}$. We can see that under the same robustness requirement, the proposed method can find a higher modulation order by exploiting the error-tolerance of GNN, thus resulting in higher data rate and lower communication overhead.

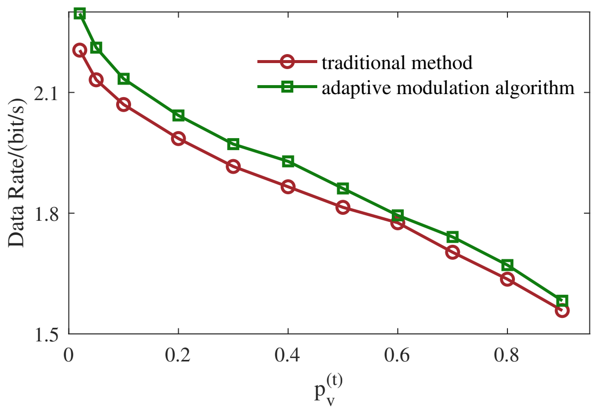

Fig. 3. Data rate with different method.

In addition, error rate is another performance metric that represents the percentages of misclassification nodes. And we prove that both the proposed method and the traditional method could keep the acceptable error rate. As shown in Fig. 4, the error rate decreases as $p_v^{(t)}$ increases. When $p_v^{(t)} > 0.6$, the error rate of adaptive modulation algorithm is almost consistent with traditional method, converging to zero. Both methods can keep the error rate to almost zero. While our method can achieve a higher data rate metioned above.

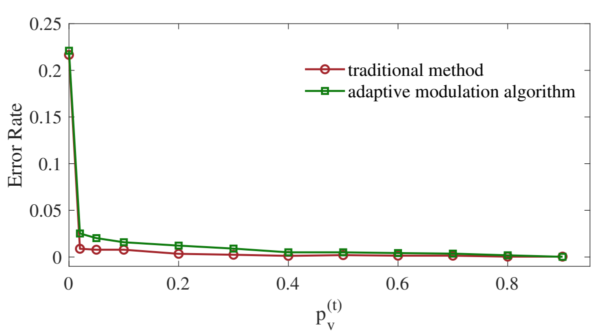

Fig. 4. Error rate with different method.

Furthermore, with the purpose of testing the scalability of the proposed method, we change the number of nodes in the graph (node density) while keeping other conditions unchanged. As shown in Fig. 5, we find that the error rate converges to zero for scenarios with different node densities. Moreover, as node density increases, the data rate tends to decrease on the contrary. It can be explained as follows. Higher node density leads to more neighbors for one node. Then, the robustness probability is smaller according to (6). Therefore, lower modulation order is reasonably adopted for higher prediction robustness.

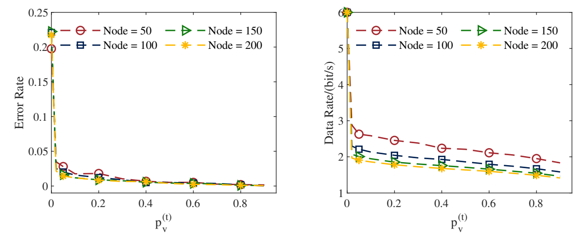

Fig. 5. Influence of node density with the proposed mechanism.

V. CONCLUSIONS

GNN can be used to effectively solve the decentralized inference tasks. However, information exchange among neighbors relies on imperfect wireless communications. This paper aims to remedy imperfect wireless transmission and enhance the prediction robustness with the proposed retransmission mechanism with adaptive modulation. The key idea is to select appropriate modulation order for each retransmission. To further enhance communication efficiency, the SBMRC technique is adopted for combination of signals. Then, we consider a decentralized GNN binary classifier model to verify the effectiveness of the proposed method. Simulation results show that the proposed methods can achieve robustness requirement after few retransmissions. Compared to the traditional method, it can reduce the overhead of communication and improve the data rate.

REFERENCES

[1] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks", *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4-24, Jan. 2021.

[2] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, and M. Sun, "Graph neural networks: A review of methods and applications", *arXiv preprint arXiv:1812.08434*, Jul. 2019.

[3] M. Lee, G. Yu, and H. Dai, "Decentralized inference with graph neural networks in wireless communication systems", *arXiv preprint arXiv:2104.09027*, Apr. 2021.

[4] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei,"Scene graph generation by iterative message passing", in *Proc. of CVPR*, 2017.

[5] H. Yao, F. Wu, J. Ke, X. Tang, Y. Jia, S. Lu, P. Gong, J. Ye, and Z. Li, "Deep multi-view spatial-temporal network for taxi demand prediction", in *Proc. of AAAI*, 2018.

[6] A. B. Sediq and H. Yanikomeroglu, "Performance analysis of soft-bit maximal ratio combining in cooperative relay network", *IEEE Trans. Wireless Commun.*, vol. 8, no. 10, pp. 4934-4939, Oct. 2009.

[7] S. Zhang, M. Wang, S. Liu, P.-Y. Chen, and J. Xiong, "Fast learning of graph neural networks with guaranteed generalizability: one-hidden-layer cas", in *ICML*, 2020.

[8] X. Xu, G. Yu, and S. Liu, "Adaptive modulation for wireless federated learning", in *PIMRC*, 2021.

[9] S. Hares, H. Yanikomeroglu, and B. Hashem, "Diversity- and AMC (adaptive modulation and coding)-aware routing in TDMA multihop networks", in *Proc. IEEE GLOBECOM*, 2003.

[10] B. Can, H. Yanikomeroglu, F. Onat, E. Carvalho, and H. Yomo, "Efficient cooperative diversity schemes and radio resource allocation for IEEE 802.16j", in *Proc. IEEE WCNC*, 2008.