Hidden Node Recognition Utilizing Wireless LAN Sensing Data in Indoor Environments

Hayato Mukasa* and Takeo Fujii*
* Advanced Wireless and Communication Research Center (AWCC), The University of Electro-Communications
1–5–1 Chofugaoka, Chofu, Tokyo 182–8585, Japan
Emails: {mukasa, fujii}@awcc.uec.ac.jp

***Abstract*—The Internet of Things (IoT) is attracting attention for monitoring the status of equipment in indoor environments such as factories. In wireless LANs used in the IoT, carrier sense multiple access with collision avoidance (CSMA/CA) achieves autonomous distributed control. However, it is known that the communication efficiency in CSMA/CA is greatly degraded due to the hidden node problem. Therefore, to detect the occurrence of hidden nodes, this paper proposes a method for discriminate the cause of packet errors and a method for discriminate whether two nodes are hidden nodes. By learning the information from the packets using logistic regression, it is possible to discriminate the cause of packet errors and whether the two nodes are hidden nodes. Through computer simulations, we show that the method can discriminate the cause of packet errors with information that can be collected more easily than conventional methods. We also show that when the offered load is 9 [Mbps], it is possible to discriminate whether two node are hidden nodes with 80% probability using training data and data collected with a different node placement pattern.**

***Index Terms*—Indoor Environment, Hidden Node, Spectrum Database, CSMA/CA, Logistic Regression**

I. Introduction

The rapid development of wireless communication technology has resulted in large and complex networks with increasing communication traffic. One example is the Internet of Things (IoT), where everything is connected to the Internet. The number of IoT devices is increasing every year, and it is predicted to exceed 50 billion by 2028 [1].

Wireless local area networks (LAN) used in the IoT employ carrier sense multiple access with collision avoidance (CSMA/CA) as the access method [2]. In CSMA/CA, frequency sharing is achieved by carrier sense, a method of interference avoidance. However, if there are nodes that cannot be detected by carrier sense, mutual interference will occur due to packet collision. This is known as the hidden node problem [3]. The hidden node problem degrades the communication quality and increases the communication traffic. In indoor environments, the hidden node problem is expected to be more severe than in outdoor environments because of the more complex and densely populated obstacles.

In indoor wireless environments such as factories, IoT is used to monitor equipment status and automate control [4]. Since the radio environment changes fluidly with time and location, proactive control is required by predicting the communication quality in advance. Generally, in outdoor environments, the received signal strength indicator (RSSI) is used as an indicator to predict communication quality. However, in indoor environments, interference from the hidden node problem described above can cause a decrease in throughput even with a good RSSI value. Therefore, it is important to recognize the mutual interference between nodes to predict the communication quality in indoor environments.

In this study, we investigate a solution to the hidden node problem by recognizing the wireless environment using packet information that can be easily obtained from nodes and access points (APs) in a wireless LAN. We collect packet information such as RSSI, retransmission counts, and the number of failed packets from nodes and APs, and compile it into a spectrum database [5]. Then, the information stored in the database is learned using logistic regression [6] to discriminate the causes of packet errors and hidden node relationships.

A method is proposed to discriminate whether propagation loss or packet collisions are the dominant cause of packet errors [7]. RSSI, retransmission rate, and packet error rate (PER) from the node and AP are collected in a spectrum database, and the discrimination is made by learning using logistic regression. However, this method requires the acquisition of PER, which requires complex statistical processing, making it difficult to perform real-time environment recognition. Therefore, the proposed method uses RSSI, the number of transmission failure packets, and the retransmissions count as input parameters for logistic regression and discrimination cause of packet error. Since these indicators require less time for statistical processing than the aforementioned PER, they can perform environment recognition in real time. In addition, the method for determining the cause of packet errors cannot detect which node is the hidden node among the communicating nodes. Therefore, we also propose a method to detect whether two nodes are hidden nodes. In addition to collecting RSSI, the number of transmission failure packets, and the retransmissions count from the node and AP, the proposed method also collects RSSI observed by other APs in a database. Based on the collected information, the proposed method uses logistic regression to detect whether two nodes are in a hidden node or not.

The remainder of this paper is organized as follows. Section II describes the system model, and Section III explains the packet error cause determination method using PER. Section IV describes the proposed method, and Section V describes simulations to evaluate the proposed method. Finally, we

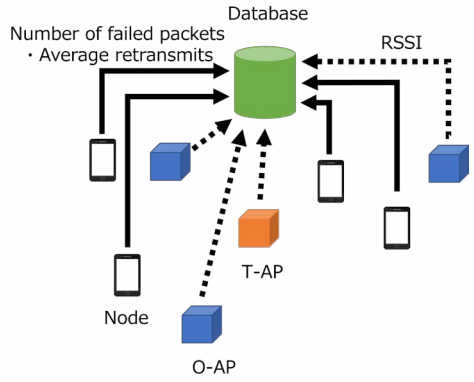


Fig. 1. System Model.

conclude this paper in Section VI.

II. System Model

We assume IEEE 802.11 g as a communication system. An AP that transmits packets from a node is defined as a transmitting AP (T-AP). Three other APs are placed around the T-AP. These APs are defined as O-APs. As shown in Fig. 1, the data collected by the T-AP, O-AP, and each node are transmitted to the database every second. The number of nodes is assumed to be four, and each node sends a packet to one T-AP.

A. Spectrum Database

To predict the radio environment with a high accuracy, the use of spectrum databases has been proposed [5] [9]. The database is constructed by accumulating packet information and RSSI easily obtained from nodes. The database will be utilized to predict the radio environment and improve the communication quality.

In this study, data such as RSSI from APs and the number of transmission failure packets and retransmissions count from nodes are collected, and then spectrum database is constructed. The information necessary for learning is obtained from the constructed database, and learning is performed by logistic regression and discrimination of hidden nodes is performed using the learning results.

B. Logistic Regression

Logistic regression is a generalized linear model for binary classification using a binomial distribution and a logistic sigmoid function [6]. The logistic sigmoid function takes a real number y as input and outputs a real number between 0 and 1, so the probability can be predicted and it is used for classification. The logistic sigmoid function is expressed by

$$\sigma(y) = \frac{1}{1+\exp(-y)}. \tag{1}$$

The posterior probability of classification class C_1 in logistic regression is given by the following equation,

$$P(C_1|\phi) = y(\phi) = \sigma(w^{\mathrm{T}}\phi), \tag{2}$$

where ϕ is the feature vector and w is the weight vector. Thus, the posterior probability of the other classification class C_2 is expressed by

$$P(C_2|\phi) = 1 - P(C_1|\phi). \tag{3}$$

To determine the weights of the logistic regression model by the maximum likelihood method, the derivative of the logistic sigmoid function is used, which is expressed as

$$\frac{d\sigma}{dy} = \sigma(1-\sigma). \tag{4}$$

The data set is $\{\phi_n, t_n\}$. The likelihood function of $n = \{1, \cdots, N\}$ at $t_n \in \{0, 1\}$, $\phi = \phi(x_n)$ is given by

$$P(t|w) = \prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}, \tag{5}$$

where $t = (t_1, \cdots, t_n)^{\mathrm{T}}$, $y_n = p(C_1|\phi)$. The error function expressed in Eq. (6) is the logarithm of the likelihood function multiplied by a minus sign, and is called the cross-entropy error function,

$$E(w) = -\ln P(t|w) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}, \tag{6}$$

where $y_n = \sigma(a_n)a_n = w^{\mathrm{T}}\phi_n$. The gradient $\nabla E(w)$ of the error function for weight w is given by

$$\nabla E(w) = \sum_{n=1}^{N}(y_n - t_n)\phi_n. \tag{7}$$

Learning w by performing sequential updates using $\nabla E(w)$.

III. Logistic Regression Based Packet Error Discrimination Method Using PER (PED-PER)

To detect the occurrence of hidden nodes, a method to estimate the collision status of packets using a spectrum database is proposed [7]. Based on the information collected from nodes and T-APs, it discriminates whether propagation loss or packet collision is the predominant cause of packet errors. This method is defined as "PED-PER". The following procedure is used to make the decision.

First, each node communicates with the T-AP. As shown in Fig. 1, the node sends packets to the T-AP for communication. Packets are generated according to the Poisson distribution shown in the following equation

$$f(k) = \frac{\lambda^k \exp(-\lambda)}{k!}, \tag{8}$$

where λ represents the expected value of the number of packets generated per unit time and k represents the number of packets generated per unit time.

The offered load G, which is the amount of traffic requested by a user, is expressed as

$$G = \frac{\lambda \times l \times b}{t}, \tag{9}$$

where t is the time [s], l is the packet length [bits], and b is the number of nodes connected to the channel.

As shown in Fig. 1, the values observed at the T-AP, O-AP, and each node are sent to the database every second. Learning by logistic regression is performed based on the information accumulated in the database. In learning by logistic regression, it is necessary to assign labels to the training data in advance.

By using the cause of packet error as the output, it is possible to discriminate whether propagation loss or packet collision due to hidden nodes is the predominant cause of packet error for any packet.

IV. Proposed Method

A. Discriminating the Cause of Packet Errors Using Instantaneous Data

PED-PER uses PER to discriminate the cause of packet errors. PER is the ratio of the number of transmission failure packets to the total number of transmission packets. PED-PER used simulation to evaluate the method. Note that packet capture is required to apply this method in actual measurements. When conducting packet capture, statistical processing such as recognizing packets during retransmission takes time.

It is assumed that the method of discriminating the cause of packet errors will be applied as an indicator for channel switching. A channel-switching method using a discriminative model of packet error causes has also been proposed in [7]. Since adaptive channel switching must be performed using the discrimination results, the emphasis is on real-time environment awareness. PED-PER that require time-consuming statistical processing are considered to be suboptimal because it is difficult to perform real-time environment recognition. Therefore, it is necessary to use parameters as inputs that obtain data easily in a real environment and perform statistical processing easily.

In addition, from the definition of PER, If the total number of transmission packets is a few, a small change in the number of transmission failure packets may cause a large change in the PER. If the offered load is low, the effect of fluctuations in PER is expected to be large because the probability of packet generation is reduced. Therefore, instead of using the PER, it is better to use the number of failed packets. The number of transmission failure packets is an ”instantaneous” value that can be easily processed statistically and obtained in real time. Based on the above, we propose to use RSSI, the average number of retransmissions, and the number of failed packets after standardization of the input data. These data are ”instantaneous data” that can be easily obtained in real-time when performing packet capture in a real environment compared to PER.

B. Discriminating Hidden Node Relationships Using Logistic Regression

The packet error cause discrimination method can discriminate whether propagation loss or packet collision is the dominant cause of packet errors. If packet errors are caused by packet collisions, packet collisions are frequently caused by hidden nodes. It is possible to detect that a hidden node problem is occurring. However, this method cannot discriminate which nodes are hidden nodes. Therefore, this section discusses a method to discriminate whether two nodes are hidden nodes.

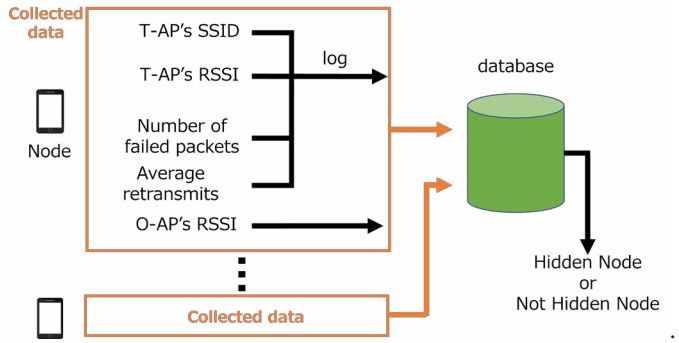


Fig. 2. Hidden node detection

Fig. 2 shows a schematic diagram of the proposed method. The following procedure is used to make a decision.

First, the nodes communicate with the T-AP. As shown in Fig. 1, the nodes send a packet to the T-AP for communication. Packets are generated according to the Poisson distribution shown in Eq. (8).

As shown in Fig. 1, the values observed by the T-AP, O-AP, and each node are transmitted from the node to the database every second. At this time, the T-AP and O-APs send the RSSI, and each node sends the number of transmission failure packets and the average number of retransmissions to the database.

Using the information accumulated in the database, logistic regression is used to discriminate whether two nodes are hidden nodes. The input parameters are RSSI, the average number of retransmissions, the number of transmission failure packets, and RSSIs at three O-APs. These parameters are standardized to mean 0 and variance 1 as input parameters and trained by logistic regression.

Since carrier sense is not possible between nodes that are hidden nodes, the output of the learning process is the carrier sense availability or unavailability. The RSSI is measured between the two target nodes in advance, and if it is less than the carrier sense threshold ϵ_{cs} [dB], it is determined that carrier sense is not possible, and if it is greater than ϵ_{cs} [dB], it is determined that carrier sense is possible. When carrier sense is not possible, a hidden node label is assigned to the training data, and when carrier sense is possible, a non-hidden node label is assigned to the training data. Utilizing the training model, discriminate whether two nodes are hidden nodes.

V. Simulation Results

A. Discriminating the Cause of Packet Errors

In this section, we evaluate a method for discriminating the cause of packet errors with instantaneous values by simulation.

1) Simulation Conditions: Training data was prepared to discriminate the cause of packet errors using logistic regression. The environment for acquiring training data was set up as follows.

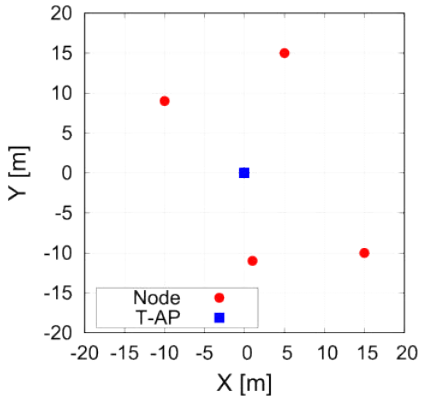


Fig. 3. Placement of nodes and T-AP in determining the cause of packet errors.

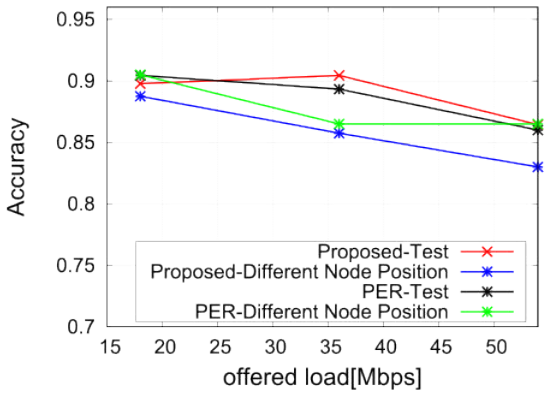


Fig. 4. Accuracy of packet error cause ddiscrimination for each offered load.

TABLE I
SIMULATION PARAMETERS.

Node Placement Area	40 [m] × 40 [m]
Frequency f	2.4 [GHz]
Number of Nodes	4
Number of T-AP	1
Number of O-AP	3
Payload Length	1500 [Bytes]
Data Rate	54 [Mbps]
DIFS	34 [μs]
SIFS	16 [μs]
ACK Data rate	24 [Mbps]
ACK Length	14 [bytes]
CW Minimum	15
CW Maximum	1023
Maximum Number of Retransmissions	6
Transmission Power P_{Tx}	10 [mW]
Noise Level	-95 [dBm]
Received Power Threshold	-65 [dBm]
Carrier Sense Thresholds ϵ_{cs}	-62 [dB]
SINR Threshold	24.6 [dB]
ACK Received Power Threshold	-74 [dBm]
ACK SINR Threshold	17 [dB]
Pathloss Coefficient n	3
Fading Model	Rayleigh Fading

TABLE II
NUMBER OF DATA: PACKET ERROR CAUSES DISCRIMINATION.

Number of dataset	1800
Number of training data	1350
Number of test data	450
Number of other 100 positions dataset	400

Fig. 3 shows an example of nodes and T-AP placement. This time, the coordinates of the T-AP communicating with each node were fixed as (0,0). For the four nodes, the coordinates were randomly changed, and data measurements were taken in 30 different configurations. Data were collected every second for 15 seconds, creating a total of 450 datasets. In addition to the training data, data were collected in 100 random patterns for 1 second each, creating a separate dataset (defined as "other 100 positions data"). Other simulation parameters are shown in Table I.

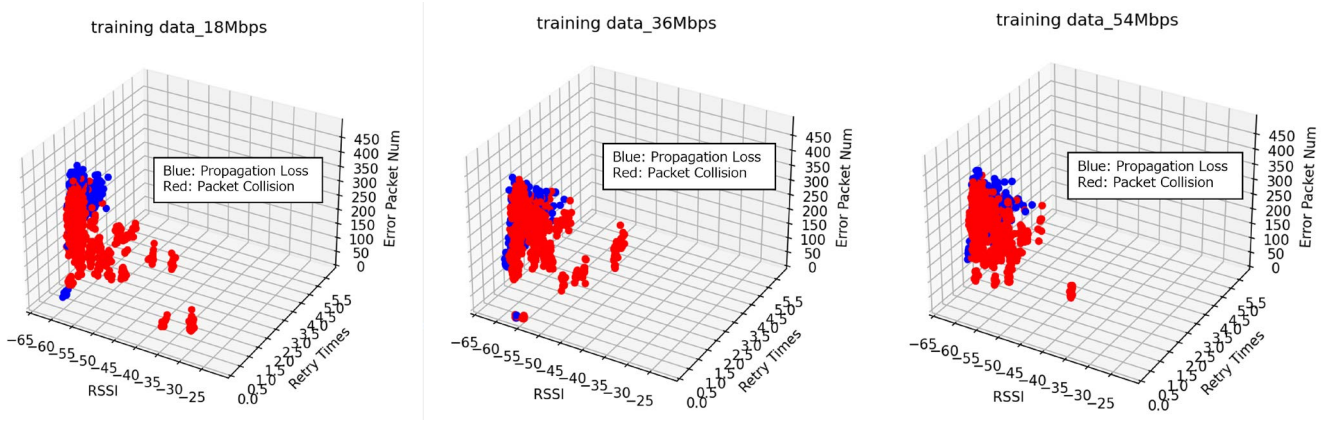
Table II shows the number of data used for training. The dataset was divided into training and test data at a ratio of 3:1, and the training data was used to construct the learning model. In addition to the test data, another 100 positions data set was used to evaluate the accuracy of the training model. This allows us to evaluate the accuracy of the learning model when it is applied to a different node arrangement than the training data used to construct the learning model.

2) Simulation Results: To verify the usefulness of the proposed method, we evaluated the accuracy of the learning model for the proposed method and PED-PER. The proposed method uses RSSI, the average number of retransmissions, and the number of transmission failure packets as input parameters, while the PED-PER uses RSSI, retransmission rate, and PER. Fig. 4 shows the accuracy when the offered load is 18, 36, and 54 [Mbps]. The accuracy is the ratio of the total number of predicted data to the number of data whose predictions match.

In Fig. 4, Proposed-Test represents the accuracy when the proposed method is used on test data. Proposed-Different Node Position represents the accuracy when the proposed method is used on other 100 positions datasets collected at different node positions than the training data. PER-Test represents the accuracy when PED-PER is used on the test data, and PER-Different Node Position represents the accuracy when PED-PER is used on other 100 positions datasets.

On the test data, the proposed method achieves almost the same estimation accuracy as PED-PER. On the other 100 positions data, the accuracy degraded by about 2%, but the proposed method was able to estimate about 85% for all offered loads. As mentioned in Section III, PED-PER uses PER as an input parameter, which requires time-consuming statistical processing. The proposed method uses data that does not require time-consuming statistical processing, such as the number of transmission failure packets, as an input parameter, and thus can perform environment recognition in real-time better than PED-PER. The proposed method achieves relatively the same level of estimation accuracy as PED-PER, indicating the superiority of the proposed method in real-time environment recognition.

Next, we discuss the difference in accuracy for different offered loads. Fig. 5 shows the actual observed data labels when the offered load is 18, 36, and 54 [Mbps], respectively. Fig. 4 shows that when the offered load is the largest at 54 [Mbps], both the proposed method and the PED-PER have lower accuracy rates. Fig. 5 shows that the higher the offered load, the smaller the scatter of features. When the offered load is 54 [Mbps], the feature variation is smaller, which may



(a) 18[Mbps] (b) 36[Mbps] (c) 54[Mbps]

Fig. 5. Actual data labels for each offered load.

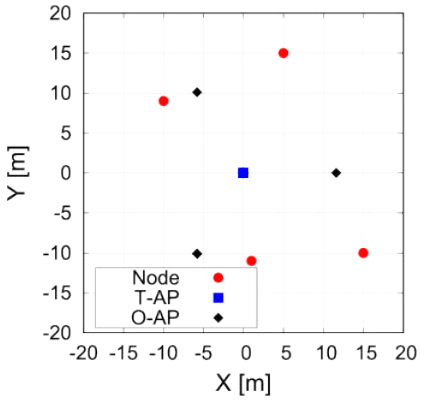


Fig. 6. Placement of nodes, T-APs, and O-APs in detecting hidden nodes.

TABLE III
COORDINATES OF T-AP AND O-APS IN DETECTING HIDDEN NODES

AP Types	Coordinate (x, y)
T-AP	(0, 0)
O-AP1	(11.6, 0)
O-AP2	(-5.8, 10.1)
O-AP3	(-5.8, -10.1)

indicate that the classification accuracy by logistic regression has decreased.

B. Hidden Node Detection

In this section, we evaluate a method for detecting whether two nodes are hidden nodes using simulations.

1) Simulation Conditions: Training data was prepared to detect whether two nodes are hidden nodes using logistic regression. The environment for acquiring training data was set up as follows.

Fig. 6 shows an example of nodes, T-AP, and O-APs placement. The coordinates of the T-AP communicating with each node were fixed as (0,0). The placement of the O-APs was determined by calculating the limit of carrier sense d_{limit} from the position of the T-AP based on Eq. (10) and placing them evenly around the circumference of the T-AP.

$$P_{\mathrm{rx}} = P_{\mathrm{tx}} - 20\log_{10}\left(\frac{4\pi df}{c}\right) - 10n\log_{10}(d), \qquad (10)$$

where P_{rx} is received power, P_{tx} is transmitted power, d is the distance between two points, f is frequency, n is pathloss coefficent, and $c = 3.0 \cdot 10^8$ [m] is the speed of light. The coordinates of the O-APs were shown in Table III.

For the four nodes, the coordinates were randomly changed, and data measurements were taken in 300 different configurations. Data were collected every second for 15 seconds, creating a total of 4500 datasets.

Table IV shows the number of data used for training. The dataset was divided into training and test data at a ratio of 3:1, and the training data constructed the learning model. To verify the change in the accuracy depending on the number of training data, we trained 10 different patterns of node placement: $30, 60, \cdots, 270, 300$. In addition to the training data, data were collected in 100 random patterns per second, creating a separate dataset (defined as "other 100 positions data"). Other simulation parameters are shown in Table I.

2) Simulation Results: To verify the usefulness of the proposed method, we evaluated the accuracy of the learning model in the proposed method.

The accuracy for each offered load on the test data is shown in Fig. 7. The horizontal axis indicates the number of node placement patterns used for training. According to Fig. 7, although there is some variation, the accuracy does not increase as the number of node placement patterns increases and is generally constant.

The accuracy for each offered load on the other 100

TABLE IV
NUMBER OF DATA: HIDDEN NODE DETECTION.

Number of node placement patterns	30	60	90	120	150	180	210	240	270	300
Number of dataset	2700	5400	8100	10800	13500	16200	18900	21600	24300	27000
Number of training data	2025	4050	6075	8100	10125	12150	14175	16200	18225	20250
Number of test data	675	1350	2025	2700	3375	4050	4725	5400	6075	6750
Number of other 100 positions data	100	100	100	100	100	100	100	100	100	100

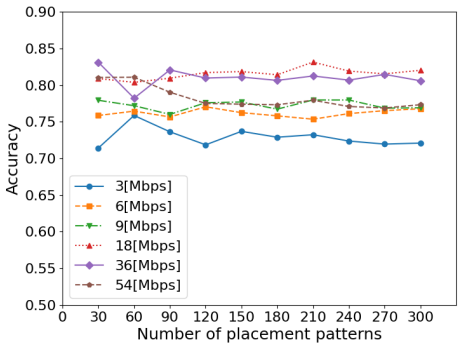


Fig. 7. Accuracy of hidden node detection for each offered load: test data.

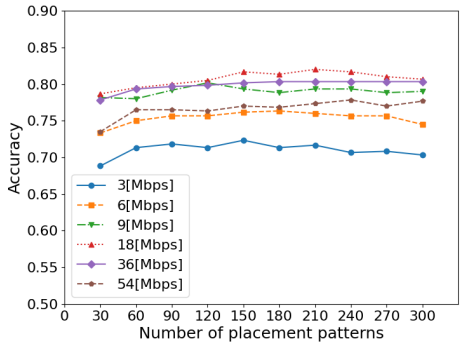


Fig. 8. Accuracy of hidden node detection for each offered load: other 100 positions data.

positions data is shown in Fig. 8. Fig. 8 shows that for any offered load, the accuracy tends to increase as the number of node placement patterns increases, but converges at around 150 patterns. As the number of node placement patterns increases, the number of measurement points increases and measurement costs increase, so it is desirable to reduce the number of node placement patterns used to build the learning model. Therefore, it is recommended that the number of node placement patterns is set at 150 in this method.

Fig. 8 shows that the accuracy is generally around 80% when the offered load is 9, 18, and 36 [Mbps]. As the offered load decreases to 6 [Mbps] and 3 [Mbps], the accuracy tends to decrease. From Eq. (8) and (9), the probability of packet generation decreases as the offered load decreases, and thus the frequency of packet collisions decreases even if the nodes are hidden terminals. Therefore, the difference feature between the bidden node case and non-hidden node case is smaller, and the accuracy may have decreased due to the smaller variation in the feature. On the other hand, the accuracy also decreased when the offered load was 54 [Mbps], $i.e.$, full buffer. The packet generation probability is higher, which increases the probability of packet collisions when nodes are not hidden nodes. Therefore, the difference feature between the bidden node case and non-hidden node case. is smaller, and the accuracy rate may have decreased due to the smaller variation in the feature values.

VI. CONCLUSION

This paper proposed a method to discriminate the cause of packet errors and whether two nodes are hidden nodes by learning packet information using logistic regression. Simulation results show that the proposed method can discriminate the cause of packet errors at a similar level using packet information that can be obtained more easily than conventional methods. In addition, when the offered load was 9 [Mbps], it was possible to detect whether two nodes were hidden terminals with a probability of 80% with the training data and the data collected with a different node placement pattern.

REFERENCES

[1] Ericsson, "Ericsson Mobility Report November 2022," Nov. 2022, https://www.ericsson.com/4ae28d/assets/local/reports-papers/mobility-report/documents/2022/ericsson-mobility-report-november-2022.pdf, Accessed: May 24, 2023.

[2] L. S. C. of the IEEE Computer Society, "Wireless LAN medium access control (MAC) and physical layer (PHY) specification," In *ANSI/IEEE Std 802.11*, pp.1–512, Sept. 1999.

[3] A. Rahman and P. Gburzynski, "Hidden problems with the hidden node problem," In *Proc. 23rd Biennial Symp. on Commun., 2006*, pp.270–273, May 2006.

[4] T. -H. Wang et al., "The management control system for plant factory that uses the IoT technology in combination with Augmented Reality technology," *2020 International Symposium on Computer, Consumer, and Control (IS3C)*, Taichung City, Taiwan, 2020, pp. 1-4.

[5] K. Sato, M. Kitamura, K. Inage, and T. Fujii, "Measurement-based spectrum database for flexible spectrum management," *IEICE Trans. on Commun.*, vol.E98.B, pp.2004-2013 Oct. 2015.

[6] C. M. Bishop, " Pattern Recognition and Machine Learning," *Springer*, pp.113–206, Feb. 2006.

[7] F. Aizawa and T. Fujii, "A Study of Hidden Node Discrimination Method using Wireless LAN Packets," *Proc. SmartCom 2019*, Nov. 2019.

[8] J. Mitola and G.Q. Maguire, "Cognitive radio: making software radios more personal," *IEEE Personal Commun.*, vol.6, no. 4, pp.13–18, Aug. 1999.

[9] G. Ding, J. Wang, Q. hui Wu, Y. Yao, F. Song, and T. Tsiftsis, "Cellular-base-station-assisted device-to-device in TV White space," *IEEE J. on Selected Areas in Commun.*, vol.34, no.1, pp.107–121, Jan. 2015.