An Effective Double-layer Detection System Against Social Engineering Attacks

***Abstract*—In recent years, social engineering attacks that use phishing emails as the medium and target specific groups of people have occurred frequently. Current enterprise systems are difficult to detect social engineering attack events dominated by human factors and the detection methods are relatively independent. Therefore, we propose a double-layer detection framework based on deep learning technology in this paper. Firstly, a phishing email detection model based on Long Short-Term Memory (LSTM) and extreme gradient boosting tree (XGBoost) is designed from the perspective of individual security. Then, an insider threat detection model based on Bidirectional LSTM and Attention mechanism is designed from the perspective of group security. Finally, combined with the social engineering network attack simulation theory, a social engineering attack and defense simulation platform is established. It is used to evaluate the effectiveness of the phishing email detection model and the insider threat detection model. We establish user roles based on corporate real log data from the aspects of human subject attribute, group relationship and personality psychology, and simulate phishing email attacks and insider threat attacks to analyze the correlation between attack process and user threat in detail. The experimental results show that our proposed framework has the characteristics of early detection, timely detection and after-the-fact investigation, which can effectively detect the risks of phishing attacks and insider threats faced by enterprise systems.**

***Index Terms*—Social Engineering Attack, Double-layer Detection, Deep Learning, Simulation Verification**

I. Introduction

In recent years, the extensive development of the Internet and the popularization of artificial intelligence technology have gradually changed people's work habit and lifestyle. New forms of social interaction and office, such as short video broadcast, we-media, online education, remote office, have brought a lot of convenience and huge benefits to people. At the same time, network security and information disclosure problems have become increasingly serious. More and more security incidents have social engineering factors.

There are two main types of social engineering attacks. One is based on human attack, the attacker fully excavates the victim's information from the aspects of sociology, psychology, and interpersonal relationship, and establishes an association mapping between the fragmented information of the network and people, forming a complete and clear dynamic information topology structure. Another is based on technical means. The most common attack is phishing email attack, where the attacker uses malicious links or malicious attachments in the email to obtain the network credentials and personal data of the attacked person. The fourth quarter report of the Anti-Phishing Working Group in 2020 shows that the number of phishing attacks has doubled during 2020, compared with the same period in 2019. Financial institutions, online email and SaaS website categories are the most frequently victims of phishing attacks [1].

At the same time, insider threats to enterprise are more destructive than phishing attacks from the outside. Employees of an enterprise system have access to sensitive data and special operations, and they can bypass physical monitoring. Once an internal attack is implemented, it is difficult to be detected, and the subsequent consequences are more serious. For example, in 2018, SunTrust Bank discovered that a former employee may have stolen personal information including names, addresses, phone numbers, and account balances of more than 1.5 million customers, and sold it to a criminal organization. The 2020 Insider Threat Report [2] stated that 72% of organizations have observed that insider attacks have become more frequent in the past 12 months. Among the various types of data attacked by insiders, customer data (61%) is considered most vulnerable, followed by financial data (54%) and intellectual property (53%).

Phishing email attack is technical attack, and insider threat attack can be regarded as human-based attacks. The ultimate goal of an attacker launching a phishing attack is to launch internal attacks with the help of enterprise insiders to obtain private data or damage the system. When employees click on the disguised link in a phishing email or establishes a contact with an attacker due to their psychological weaknesses such as curiosity, greed, and reciprocity, they may trigger internal attacks. In addition, without the participation of external attackers, internal employees may still generate internal attacks due to factors such as work, family and society. It is necessary to explore the correlation characteristics of external attacks and insider threats, especially those caused by human factors, and establish a multi-layer attack defense mechanism from technical means to help individuals and enterprises to defend against social engineering attacks.

The main contributions of this paper are as follows:

- We design a multi-model fusion phishing email detection method, which fully combines the advantages of LSTM networks strong ability to extract deep semantic features from the subject and body of emails and the high running speed of XGBoost. The model generalization ability and the accuracy rate of detection is improved by adding a custom loss function in training process.
- We design an insider threat detection model based on Bidirectional LSTM and Attention mechanism. The Bidirectional LSTM network can avoid the problem of long sequence information being forgotten and Attention mechanism can get the key information in the sequence. The validity of the model for multi-domain time series and anomaly detection are verified by comparing with

different models and existing insider threat detection models.

- We design a double-layer detection framework for social engineering attacks and simulate phishing email attacks and insider threat attacks to dynamically analyze the threat correlation between external attacks and internal attacks.

The rest of this paper is organized as follows, Section II presents some related defense and detection technology of social engineering attacks. Section III introduces the double-layer detection framework and the principle of model implementation. In Section IV, we conduct experiments to evaluate the double-layer detection framework on our developed simulation verification platform and analyze the attack process in detail. Finally, we conclude the paper in Section V.

II. Related Work

Social engineering involves concepts such as sociology, psychology and information security. Attackers use the personality psychological weaknesses, interpersonal interaction and loopholes in the regulations to achieve the purpose of attack. In the process of daily social or active contact, the attacker establishes trust with the attacked by means of conversation, deception and transaction to obtains useful information to achieve the purpose of penetration.

A. Social Engineering Attack and Model

Social engineering attacks can be roughly divided into human-based attacks and technology-based attacks. Harley first proposed a classification method of social engineering attacks, which listed seven kinds of social engineering attacks [3], including camouflage, password theft, spam and other common attacks. Anthony *et al.* found out the gaps in network security related to social engineering and web phishing by investigating the solutions adopted by organizations against attack media. Bakhshi [4] studied the vulnerability of target enterprise users to social engineering attacks by simulating phishing attacks. The research abstractly described the privacy threats caused by social engineering attacks, but it lacked applicability. At present, most of the social engineering models are based on the attack cycle model, which divides the attack model into 4-8 stages, each stage completes a specific target task.

B. Defence Against Social Engineering Attacks

In the past few years, the concept and attack of social engineering have been widely mentioned, but the research on defense of social engineering attacks is still in its infancy. Email is an important medium for daily communication between enterprises and people, and it is also one of the most commonly used media for phishing attacks. As the internal and external networks of enterprises gradually become blurred, phishing attacks become more complex and advanced, and internal attacks dominated by human factors occur frequently. Researchers have also made some achievements in the research on phishing email detection and insider threat detection technology.

Detection methods based on machine learning and deep learning are currently the mainstream method for detecting phishing emails and insider threat. Fette *et al.* [5] and Cohen *et al.* [6] used random forest classifier to detect spam and phishing emails, but Fettes method has a high false positive rate for phishing email detection, and Cohens approach ignored the external features of email and was not connected to the internet, so the real-time performance is insufficient. Most of the detection methods based on deep learning are based on text feature. Egozi *et al.* [7] used NLP technology to extract 26 features from stop words, word counts, and punctuation for the email text content, and correctly identified more than 80% of phishing emails and 95% of normal emails. The THEMIS deep learning model proposed by Fang *et al.* [8] used an improved RCNN combined with Attention mechanism to detect email contents at both character and word level, and can detect phishing emails with a higher degree of disguise, and has better performance than LSTM.

Rashid *et al.* [9] proposed an insider threat detection method using the HMM model to model the normal behavior of users and simulate the deviation from normal behavior over time, but it is not suitable for long sequence detection. Chi *et al.* [10] used psychological characteristics as auxiliary features, combined with language analysis and K-means algorithm to analyze communication logs such as emails to determine whether employees meet certain personality criteria, and calculate the risk level of each employee. Tuor *et al.* [11] studied the insider threat detection method based on deep learning, extracted 414-dimensional feature vectors from system logs, and used fully connected neural network and LSTM to detect anomalies according to the statistical characteristics of user behavior. The model can extract contextual information from the user's operation sequence over time, but the detection performance deteriorates when the sequence is long.

The research on simulation verification of social engineering attacks can help to clearly explain all attack components and their relationships to defend against social engineering attacks. In 2014, Mouton *et al.* [12] proposed ten kinds of social engineering attack templates from real social engineering attack instances, verified the social engineering attack detection model by applying social engineering attack scenarios, and provided detailed attack processes and steps. In 2016, Wilcox *et al.* [13] proposed a social media policy framework SESM, which focuses on reducing the risk of enterprises from social engineering attacks through ICT security policy control. In 2019, Zheng *et al.* [14] proposed a social engineering framework based on conversation and dialogue, and analyzed the usability of the proposed framework. But it failed to evaluate and verify in the real network environment.

III. Our Proposed Scheme

A. Overview

Attackers usually adopt social engineering methods to collect target information, and formulate attack plans and escape strategies. If the attacker chooses the phishing email attack method, he can select the appropriate phishing email from the phishing template library according to the target information.

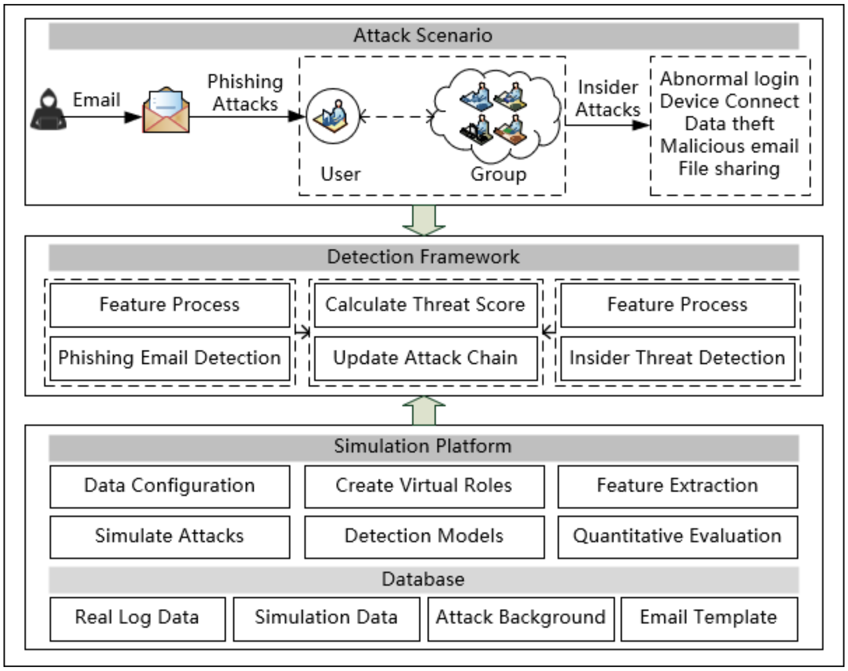


Fig. 1: Structure of Double-layer Detection Framework and Simulation Verification Platform.

The employee may view the email after receiving it and may click on the disguise links in the phishing email. The employee may be lured by the content of the web page to be infected and to generate internal attacks on the enterprise such as data interception and system damage. The structure of double-layer detection framework and simulation verification platform is shown in Fig. 1. The upper layer describes the attacker, attack media and attack scenarios of internal employees. The middle layer depicts the proposed framework which provides hierarchical detection of attacks and analyzes the threat association in each step. The bottom layer is the basic module of the simulation platform for verifying the efficacy of the framework. The main functions of the platform include multi-source heterogeneous real data and simulation data fusion, data configuration and model loading, social engineering virtual roles generation, multi-dimensional social engineering event information mining and feature extraction, social engineering attack simulation, and quantitative evaluation.

B. *Phishing Email Detection Model*

We design a multi-model fusion phishing email detection model called L-XGB (Phishing Email Detection Method Based on LSTM and XGBoost). The length of the email subject is generally no more than 100 words, which is a short text type, while the email body content is a long text type. After word segmentation and vectorization respectively, the LSTM network is used to extract the subject information and the Bi-LSTM network is used to extract the implicit information of the email body, which can avoid the information forgotten by LSTM due to the long sequence. After that, the subject and content features are spliced with the numeric features and input into the XGBoost classifier to participate in training evaluation. XGBoost [15] is an ensemble learning algorithm that belongs to the Boosting school, it realizes the generation of weak learners by optimizing the structured loss function. XGBoost algorithm does not use the search algorithm, but directly uses the first second derivatives of the loss function and improves the performance of the algorithm by pre-sorting and weighted quantile. XGBoost can artificially define the loss function, which can further increase the generalization ability of the model.

C. *Insider Threat Detection Model*

The second-layer detection mechanism is to protect group security. We design a MDFTS (Multi-Domain Feature Time Series) extraction algorithm for heterogeneous data sources, and use Bi-LSTM network combined with Attention mechanism to establish the BiLA-ITD (Insider Threat Detection Method Based on Bi-LSTM and Attention Mechanism) model. Different from most existing machine learning anomaly detection algorithms which are only suitable for detecting transient abnormal events, BiLA-ITD uses bidirectional LSTM network to extract the time series features of user behavior expressed in sentences to avoid forgetting certain information in the model due to the long input sequence, and refer to the user's past behavior sequence and future event information (such as a user leaving the company after making a lot of attacks) during anomaly detection. The Bi-LSTM output vector is used as the input of the attention layer. Attention mechanism can effectively capture the dynamic characteristics of the data and the key features of the sequence. Finally, the output of attention layer is connected with the numerical statistical features and passed to the next neural network layer for classification calculation, and the final classification result is calculated through the Sigmoid function.

IV. Experimental Evaluation

To verify the feasibility of our proposed double-layer detection framework, we conducted experiments in the Anaconda3 environment of Windows system, where the Python version is 3.7.7, and the Tensorflow version is 2.3.0. In the following, we will introduce the simulation verification platform, the data set used and the specific simulation verification process, and finally analyze the evaluation results and case results.

A. Dataset

The experimental data of simulation verification are extracted from multi-source heterogeneous data sources, and the experimental data of phishing email attack are real phishing email cases. The normal email dataset used for phishing email detection comes from the Enron email dataset. The corpus contains about 500,000 messages, which are stored in folders. The phishing email dataset used is from the monkey.org website. We randomly selected 2000 emails from the normal email dataset and the phishing email dataset as the training and testing datasets for phishing email detection model. We used the CMU-CERT r6.2 dataset for the insider threat detection model. The 6.2 version of CMU-CERT dataset contains 4000 user's behavior sequence and operation log during 18 months, it not only simulates the three main types of attack behavior data of insiders' carrying out system damage, information theft and internal fraud, but also contains a large amount of normal user background data, which is suitable for the study of user behavior sequences, relational asset models, and decoy psychology models.

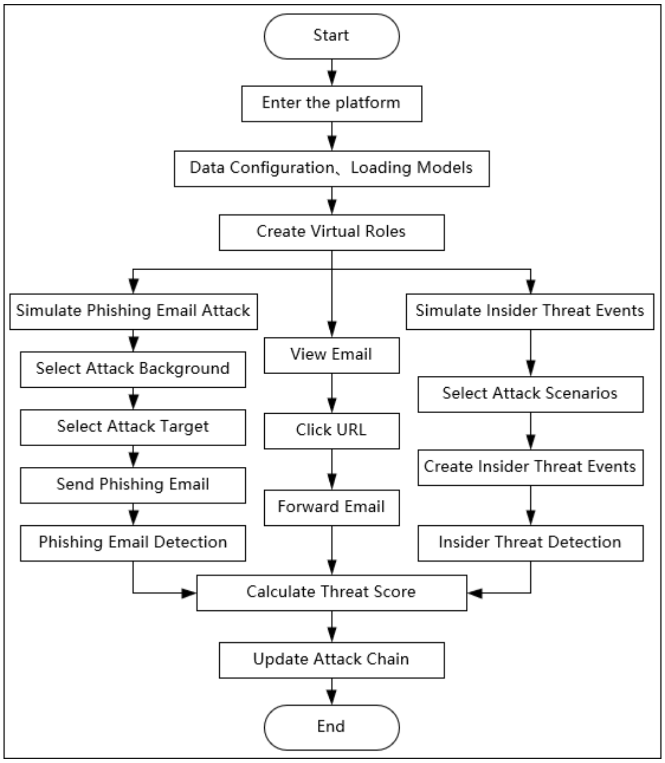


Fig. 2: Simulation Verification Process.

B. Simulation Verification Steps

The simulation verification process is based on the real case that 1.5 million customer data of SunTrust Bank in the United States were stolen and sold by internal employee in April 2018. We summarized typical phishing email cases in recent years and built the phishing email template library according to the content of the emails and the psychological characteristics of the victims. The internal attack data comes from the CERT-r6.2 dataset. The specific simulation verification process are shown in Fig. 2.

Step 1: Data Configuration and Load Models: After the platform is running, it is necessary to configure the data first and load the detection models. In addition, the platform supports custom addition of phishing email attack cases background, including the real background content, the attacker's real identity, the attacker's disguised identity, the way to obtain information, the identity of the target and case background preservation path.

Step 2: Virtual Role Modeling: The information of virtual roles are extracted from the configured multi-source heterogeneous log data. The newly created virtual role is added to the system. One can view user details, view the receive status and content of email, and delete user roles in system.

Step 3: Simulate Phishing Email Attack: After selecting the attack event background and the attack target, the attacker writes phishing email content according to the personality and psychology of the target user. Phishing email includes email subject, email content, phishing links, email attachments, and email format. The attacker can select the phishing email content from the phishing email template library for the attack target.

Step 4: Send Phishing Email: When the attacker has finished writing the content of the phishing email, he can send the phishing email to the target user through the system. During the sending process, the system automatically detects whether the email is a phishing email, and updates the recipient's mailbox status after the detection is completed.

Step 5: Simulate Insider Threat Attacks: Select attack scenarios to create insider threat events. The system automatically detects whether the target user has internal attack behavior, and updates the detection status and result charts after the detection is completed.

Step 6: Calculate Threat Score and Update Attack Chain: The threat score is determined by URL click rate, phishing email forwarding rate, internal attack scenarios and attack days.

Step 7: User Early Warning and Review: One can choose to trust the insider or review the insider based on the detection results.

C. Experimental Results and Case Analysis

1) Experimental Results: We trained and evaluated the L-XGB model and the BiLA-ITD model separately. The L-XGB model took 156.58ms in the process of detecting 4550 samples, and obtained 98.59% precision rate and 1.41% false positive rate on the test set. The precision rate of BiLA-ITD

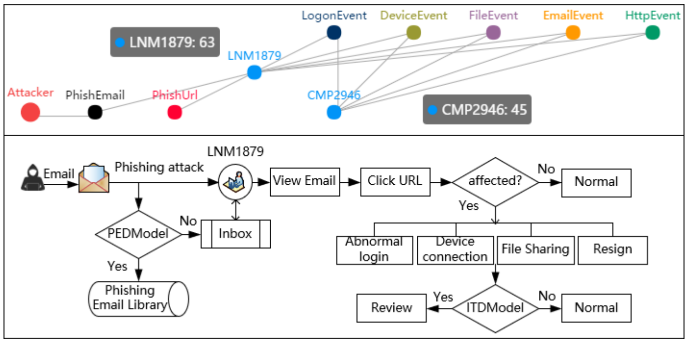


Fig. 3: Attack Chain and Detection Process.

TABLE I: Case Analysis of Attacks

Phishing Email Attack	
Attacker's real identity:	Hackers and third party organizations;
Attacker's disguised identity:	eBay official technician;
The way to obtain information:	Social engineering methods;
The identity of the target:	SunTrust Bank employee;
Social relations:	Technical support and transaction relations;
Exploit weakness:	The psychological characteristics of the victim such as dissatisfaction, impulse and easy trust;
Specific attack methods:	The target is induced to enter the phishing website, and the attacker reveals the transaction news.
Insider Threat Attack	
Attack motivation:	The failure of job promotion and the dissatisfaction of superiors and the temptation of outsiders;
Preparation period:	About a week;
Convenience of position:	The insider has access to database and sensitive data, and is familiar with physical monitoring;
Attack scenario:	Abnormal logon, device connection, file sharing, data theft;
Characteristics:	Suddenness, continuous, isolation;
After the attack:	Leave the company;
Consequences:	A large amount of customer data was stolen;

model in detecting 195,125 user behavior sequences on the test set is 96.4%, and the false positive rate is 0.035.

In the process of simulating phishing attack and insider threat attack, the attack chain chart is depicted in Fig. 3. The insider threat detection result is shown in Fig. 4. Next, we specifically analyze the content of each stage according to the simulation verification process.

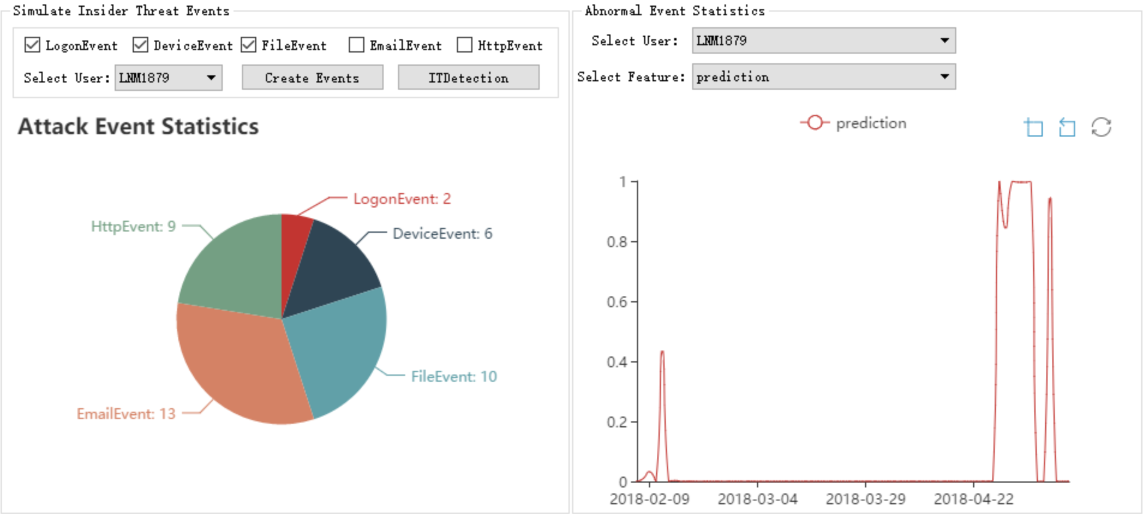
2) Case Analysis: Table I summarizes the analysis of phishing email attack and insider threat attack. We selected the real case that 1.5 million customer data of SunTrust Bank in the United States were stolen and sold by internal employee in April 2018 as the background of the simulated phishing email attack and chose LNM1879 as the experimental user. In order to unify the format of real user data and simulated data, we define each user role information in json format. For example, the user LNM1879 is defined as:

```
{
LNM1879:{IsThreat : true, IsInsider : true,
BasicInfo: {Name : “”, Sex : “” },
WorkInfo: {Role : “”, Department : “” ...},
SocialInfo: {Email : “”, Facebook : “” },
CharacterInfo: {O : “”, C : “”, E : “”, A : “”, N : “”},
PhishEmail: {IsNewEmail : true, IsPhishing : true},
InsiderThreat: {AttackDays : 0, AttackScenario : {}}}
}
```

Each user's information includes basic information, work information, social information, personality and psychological information, phishing attack information, and insider threat information. The platform counts the user's role information by histogram, and clusters the users according to the department and supervisor relationship. Moreover, the platform automatically analyzes the users' personality characteristics according to different users' personality psychological evaluation scores. For example, the user LNM1879 is a computer scientist in the system engineering department and his main work is commercial engineering research. The openness, conscientiousness, and agreeableness scores of LNM1879 are relatively low, but the extraversion and neuroticism scores are relatively high, indicating that he cannot cope with higher work pressure, and does things without considering the consequences. He may be self-centred sometimes but normally the mood is optimistic.

The attacker used LNM1879's personality characteristics combined with the collected information to send him a phishing email. The content of the phishing email is about the user's eBay account update verification. The phishing email informs LNM1879 that his eBay account needs to verify the validity and identity information of the personal account due to security updates, and requires LNM1879 to click the link in the email to login eBay account to fill in personal information. The attacker used the eBay official identity to notify LNM1879 that the eBay account could not be used within 5 days, and then “good faith” reminded LNM1879 that he would periodically receive site updates information. The email format is HTML format and the phishing link is displayed in the form of a visual button. The phishing email detection model detects the email as a normal email even though it is a phishing email. LNM1879 is an anxious person. After the attacker disguised as an “eBay official” reminder, LNM1879 may immediately click on the link in the email due to anxiety, trust and other complex psychology. When he clicked on the phishing link to enter the target website, we set the **IsThreat** value to **true**, which means that LNM1879 is threatening. The threat coefficient is determined by the URL click rate, email forwarding rate, whether to generate internal attacks and the number of internal attacks. The threat score of insider is the sum of the URL click threat score, the email forwarding threat score, the attack days threat score, and the attack scenario threat score.

Since the real internal attacks of SunTrust Bank employees cannot be obtained, our platform provides the function of simulating insider threat events. In the following week, the attacker increases contact with LNM1879. LNM1879 was a computer engineer in SunTrust Bank. He used his authority to bypass the company's physical monitoring and administrator monitoring during off-hours, logged on to the data server and used a mobile device to copy a large amount of customer data, and uploaded some confidential documents to the online



Simulate Insider Threat Events

☑ LogonEvent ☑ DeviceEvent ☑ FileEvent □ EmailEvent □ HttpEvent

Select User: LNM1879

Abnormal Event Statistics

Select User: LNM1879

Select Feature: prediction

Fig. 4: Insider Threat Attack Scenario and Detection Results

website specified by attacker. LNM1879 quickly went through the resignation procedures and left the company on April 19, 2018. On April 20, the company discovered that customer data was stolen and used BiLA-ITD model to detect the company's suspicious employees' behavior log data in the past week. According to the detection results, the company found that the employee LNM1879 who had left the company had stolen customer data.

V. Conclusion

Aiming at the threat of social engineering attack faced by enterprise systems and the lack of corresponding effective comprehensive prevention system, this paper has proposed a double-layer detection framework from the perspective of phishing attack and insider threat attack. We have designed a multi-model fusion phishing email detection model and insider threat detection model based on deep learning. In addition, we have developed a prototype system, which has the functions of multi-dimensional social engineering event information mining and feature extraction, social engineering virtual roles generation, social engineering attack simulation and quantitative evaluation. Then, we have simulated the phishing email attack and insider threat attack on the simulation platform according to the real attack case background. By establishing a social engineering attack chain, we have analyzed the threat correlation between the attacker, the attack media, the attacked target, and the internal attack in detail. Finally, the feasibility of the double-layer detection framework is objectively evaluated through the simulation process. Our work has provided theoretical and practical guidance for enterprise system to deal with social engineering attacks dominated by human factors.

Acknowledgment

This research is supported by the National Key R&D Program of China (2017YFB0802805). Daojing He is the corresponding author of this article.

References

[1] APWG, "Phishing activity trends report 4^{th} quarter 2020," Anti-Phishing Working Group, Tech. Rep., 2020. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q4_2020.pdf

[2] Cybersecurity, "2020 insider threat report," Cybersecurity Insiders, Tech. Rep., 2020.

[3] T. R. Peltier, "Social engineering: Concepts and solutions," *Information Security Journal*, vol. 15, no. 5, p. 13, 2006.

[4] T. Bakhshi, "Social engineering: revisiting end-user awareness and susceptibility to classic attack vectors," in *2017 13th International Conference on Emerging Technologies (ICET)*, Islamabad, Pakistan, Dec. 27-28, 2017, pp. 1–6.

[5] I. Fette, N. Sadeh, and A. Tomasic, "Learning to detect phishing emails," in *Proceedings of the 16th international conference on World Wide Web*, Banff, Alberta, Canada, 2007, pp. 649–656.

[6] A. Cohen, N. Nissim, and Y. Elovici, "Novel set of general descriptive features for enhanced detection of malicious emails using machine learning methods," *Expert Systems with Applications*, vol. 110, pp. 143–169, 2018.

[7] G. Egozi and R. Verma, "Phishing email detection using robust nlp techniques," in *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, Singapore, Singapore, Nov. 17-20, 2018, pp. 7–12.

[8] Y. Fang, C. Zhang, C. Huang, L. Liu, and Y. Yang, "Phishing email detection using improved rcnn model with multilevel vectors and attention mechanism," *IEEE Access*, vol. 7, pp. 56 329–56 340, 2019.

[9] T. Rashid, I. Agrafiotis, and J. R. Nurse, "A new take on detecting insider threats: exploring the use of hidden markov models," in *Proceedings of the 8th ACM CCS International workshop on managing insider security threats*, Vienna, Austria, Oct. 2016, pp. 47–56.

[10] H. Chi, C. Scarllet, Z. G. Prodanoff, and D. Hubbard, "Determining predisposition to insider threat activities by using text analysis," in *2016 Future Technologies Conference (FTC)*, San Francisco, CA, USA, Dec. 6-7, 2016, pp. 985–990.

[11] A. Tuor, R. Baerwolf, N. Knowles, B. Hutchinson, N. Nichols, and R. Jasper, "Recurrent neural network language models for open vocabulary event-level cyber anomaly detection," *arXiv preprint arXiv:1712.00557*, 2017.

[12] F. Mouton, L. Leenen, and H. S. Venter, "Social engineering attack examples, templates and scenarios," *Computers & Security*, vol. 59, pp. 186–209, 2016.

[13] H. Wilcox and M. Bhattacharya, "A framework to mitigate social engineering through social media within the enterprise," in *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, Hefei, China, Jun. 5-7, 2016, pp. 1039–1044.

[14] K. Zheng, T. Wu, X. Wang, B. Wu, and C. Wu, "A session and dialogue-based social engineering framework," *IEEE Access*, vol. 7, pp. 67 781–67 794, 2019.

[15] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, San Francisco, California, USA, Aug. 2016, pp. 785–794.