Identification of Information Security Threats Using Data Mining Approach in Campus Network

Norkhushaini Awang*[a], Ganthan Narayana Samy[a], Noor Hafizah Hassan[a], Nurazean Maarop[a], Pritheega Magalingam[a] and Norshaliza Kamaruddin[a]

[a] Razak Faculty of Technology and Informatics, Level 5, Menara Razak, Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia

E-mail: *shaini@tmsk.uitm.edu.my

Abstract. Comprehensive risk assessment implementation in an organization is crucial in order to safeguard valuable organization assets and to minimize information security threats. Thus, inadequate information security risk assessment may result in compromised confidentiality, integrity, and availability of the information system due to unauthorized access particularly in the education domain. Therefore, the objective of this paper is to identify several information security threat risks related to the University Information System. Hence, data from intrusion prevention system (IPS) has been collected from the selected university campus network. Moreover, under Python language, Anaconda is used as a machine learning environment to do the data analysis of the collected data. Basically, the analysis of the university campus network data identified various types of information security threats such as database-related attacks. The contribution of this research is to guide the network administrator to develop an appropriate incident response plan based on the identified threats from the risk assessment activity.

***Keywords*— university information system (UIS); intrusion prevention system (IPS); data mining; network threats.**

1. Introduction

Currently, information security is critical assets for an organization to protect their information in their business. Information security is defined when admin protecting the information, hardware, storage and network system from vulnerabilities and threats (1). There are challenges involves in implementing information security in an organization. Data inside the systems is valuable and critical to the business. As discussed by (2) stated that the core principles of information security are Confidentiality, Integrity, and Availability (CIA) that forms the basis of asset protection, authenticity, accountability, reliability, and non-repudiation. Authors also stated that the CIA is the three major principles of information protection. Compromising these principles leaves systems in risk. According to (3) emphasize that first among the targets in cyberspace are colleges and universities because they contain resources with open access. Therefore, it is needed to conduct a study in a university information system (UIS). As mention by (4) stated that organizing knowledge of systems behavior, network admin have a better chance of identifying the origins of risk events come whether from within the system or from its immediate environment. With the help of data science, which a multidisciplinary blend of data inference, algorithm development, and technology help researcher in order to solve analytically complex problems in understanding system behavior.

2. Related work

UIS is referred to as a collection of hardware, software, people, data, and information that provides managers with the tools for organizing, evaluating, and efficiently running their departments (5). The author also described that UIS components are student information system, library information system, faculty information system, and a finance system, as illustrated in Figure 1. As mention by Louisiana State University (2017), UIS is developed to enhance, maintain, and protect the academic, research, and administrative information resources of the university. One of the primary responsibilities of UIS to the information resources of the university is to protect and ensure the confidentiality of the information stored in the university's databases. Further discussion from (6) stated that university campuses are equipped with the most technologically advanced places in the world by providing facilities like extensive Wi-Fi support, online learning using lecture capture software, digital library, classroom virtualization and web conferencing. However, these advancement make universities' computing environment vulnerable to hacking targets because of large open networks.

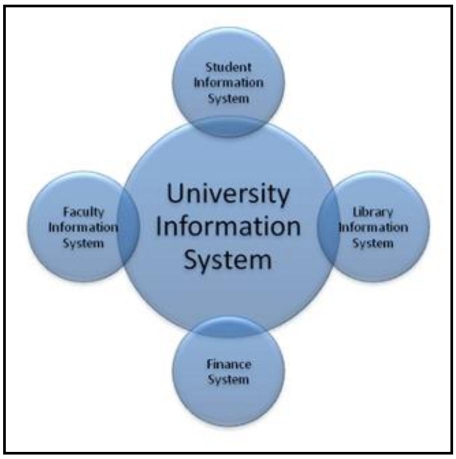
Student Information System
University Information System
Faculty Information System
Library Information System
Finance System

Fig. 1 Example of University Information System (5)

A study from (7) emphasized that the difficulty in implementing information security in higher education because lack of system integration and understanding across different departments and systems. Therefore, a systematic approach is needed to assess the risks in UIS, so that the risks can be mitigated in time in order to protect the assets. The research about university information security threats also been discussed by (8) in the report. The author emphasizes that many university users do not understand basic information security threats. Universities users neglecting about a compromised system that can be used to attack another system through a computer network. They also overlooked on password protection, which can lead to identity theft. As in university network, we have to increase campus users' awareness regarding the confidentiality of our systems and networks content. As discussed by (9) threat model can help network admin to understand what kind of rival they are securing it against. Security is not only about developing technical solutions, but also users have to understand which information is critical to campus network and what is important to protect.

Higher education institutions are susceptible to cyber-attacks. Elements such as open networks, large volumes of data and freedom of public access expose them to a variety of cyber threats and risks.

According to (8) mention that the Computer Emergency Response Team reported that the number of network incidents in the university campus in 2001 was 52,658, which jumped from 21,756 the previous year. The increasing of threats in today's universities because of the lead of technological advancement (10). The author also discussed access technology in the university campus results in a vulnerable computing environment with more security threats. University campuses are demonstrating some advancement technology by providing facilities like extensive Wi-Fi support, online learning using lecture capture software, digital library, classroom virtualization and web conferencing. All these advancements make University's computing environment particularly vulnerable because in contrast to hacking targets like banks, college and university computing environments are often large open networks. A discussion from (11) stated that to define and identify critical infrastructures at the national level, risk assessment can be conducted in order to manage their own infrastructures to plan a decision-making. Figure 2 shows data breaches that affected the university information system.

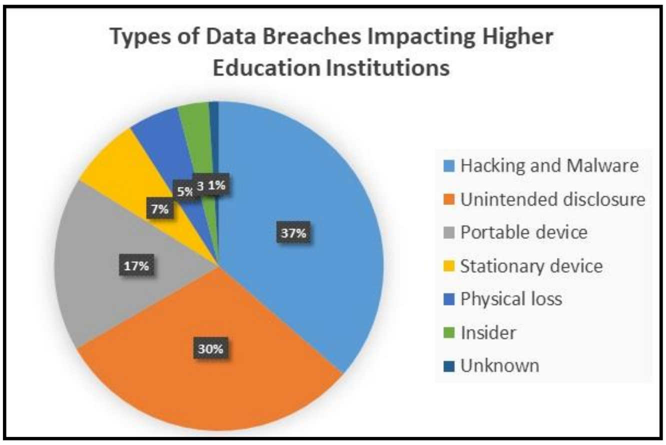

Fig. 2 Types of Data Breaches in UIS (12)

A discussion on UIS threats also been discussed by (10) where authors explained that campus network mainly suffers following security threats as groups below:

1. Phishing, ransomware, and malware: Cybercriminals uses emails or Web accounts that spoof official mailings for financial gain. University's young students are at most of being the victim of a phishing attack that results in malware or ransomware downloads.
2. Wi-Fi: University which provides Wi-Fi access on the University campus which is great in technology advancement view, but it can cause security problems.
3. Viruses Spreading through Social Media: Young adults of University are the most passionate users of social media like Facebook, Twitter and YouTube. This implies that in University's network malware can spread through social media sites.
4. Mobile Devices: Students are early adopters of technology, and new devices are frequently visible in campus; from iPads to new android phones, daily new launched devices are having upgraded versions of operating systems that can easily be infected by the smart attacker and also ready to infect University's network.
5. Embedded Devices: Embedded connectivity improves the risks for viruses and more threats to the network.

Therefore, it is important to conduct research on risk assessment in the university information system. It also significance to develop and implement proper security controls based on the results of their internal risk assessment and vulnerability assessment. Both approaches can be used to improve efficiency towards achieving desired security levels.

Information Security Risk Assessment is an on-going process of discovering, correcting and preventing security problems. The risk assessment is part of a risk management process designed to provide appropriate levels of security for information systems. Authors (13) stated that data is important in risk assessment process to do identification and implementation of decision making. Quantitative description of specific data using statistical models provides tools for translation and provide method in the conventional risk assessment paradigm. Other researchers (14) studied on assessing risk using data analytics also stated that data analytics is a promising way to turn information into outcomes, enhance decision-making, make data-driven discoveries, minimize risk, and have a valuable insights that would otherwise remain hidden. Adaptation of data analytics in business give a major impact in doing the risk assessment. As been discussed by (15), data analytics is a techniques to process relevant data, which are to be tied with suitable capability to deal with unexpected events and provide the right support to enable risk management. From their research shows that how data analytics is relevant in doing risk assessment as tools to make accurate decision for securing campus network. Authors from (16) claimed that network infrastructure is secure. However, rather than wondering if the infrastructure would be attacked from malicious actions, IT administrators shifted towards trying to understand when it will happen, and what the consequences will be to prevent before the malicious attacked. As mention from (17) in their study, they conducted from unified risk assessment to personalized risk prediction. The process move from fixed information to relationship between human factors. They found that the risk prediction helped them to get the high-accuracy model in conducting a risk assessment. As in our research, we implemented data mining approach to have the data analysis in assessing the risk in campus network. As mentioned by (18) data mining turns a large collection of data into knowledge. Process of data mining is when researchers sort massive data sets in identifying patterns and establish relationships to solve problems through data analysis. We capture raw data from IPS and analysis them into useful information to network admin.

Figure 3 explained about a general road map on pattern of data mining research. Mentioned by (18), most research studies mainly address three pattern mining aspects which are patterns mined, mining methodologies, and applications. Some studies, however, integrate multiple aspects; for example, different applications may need to mine different patterns, which naturally leads to the development of new mining methodologies. As in our research, we integrate more than one aspects in doing the data mining for this research.

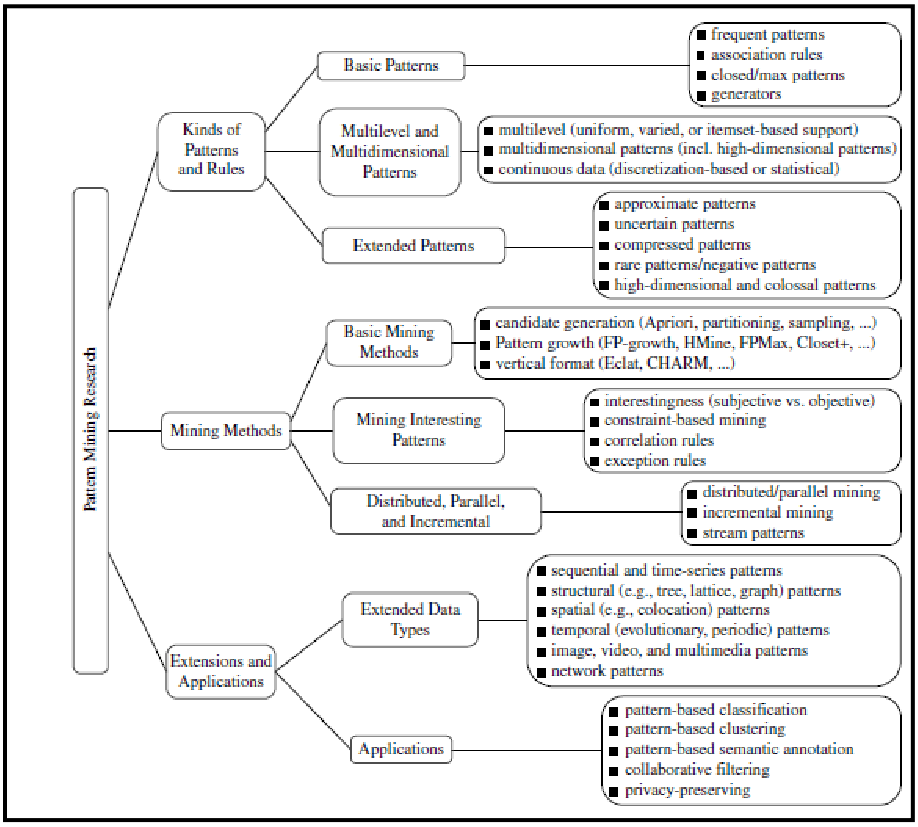
Pattern Mining Research
Kinds of Patterns and Rules
Basic Patterns
■ frequent patterns
■ association rules
■ closed/max patterns
■ generators
Multilevel and Multidimensional Patterns
■ multilevel (uniform, varied, or itemset-based support)
■ multidimensional patterns (incl. high-dimensional patterns)
■ continuous data (discretization-based or statistical)
Extended Patterns
■ approximate patterns
■ uncertain patterns
■ compressed patterns
■ rare patterns/negative patterns
■ high-dimensional and colossal patterns
Mining Methods
Basic Mining Methods
■ candidate generation (Apriori, partitioning, sampling, ...)
■ Pattern growth (FP-growth, HMine, FPMax, Closet+, ...)
■ vertical format (Eclat, CHARM, ...)
Mining Interesting Patterns
■ interestingness (subjective vs. objective)
■ constraint-based mining
■ correlation rules
■ exception rules
Distributed, Parallel, and Incremental
■ distributed/parallel mining
■ incremental mining
■ stream patterns
Extensions and Applications
Extended Data Types
■ sequential and time-series patterns
■ structural (e.g., tree, lattice, graph) patterns
■ spatial (e.g., colocation) patterns
■ temporal (evolutionary, periodic) patterns
■ image, video, and multimedia patterns
■ network patterns
Applications
■ pattern-based classification
■ pattern-based clustering
■ pattern-based semantic annotation
■ collaborative filtering
■ privacy-preserving

Fig. 3 Road Map on Pattern Mining Research (18)

According to (19) stated that data mining techniques have the capability to find hidden pattern from the secondary data in large databases. It leads in creating a prediction model for desired output. Data mining also helped in finding the accuracy of the algorithms. A research from (20) explained that they used three layers of analysis before decide to make any plan on their network. The methods they used are to develop structured firewall analysis, re-use in other firewall projects and presented for use by others with similar challenges. Benefit from this approach, they learned the attacks behaviour from analysis that they did and operation have been useful in both the perimeter and their secondary data centre firewalls projects.

3. Methodology

We have conducted an experiment from IPS data packet. The data is collected from IPS in local university network to analysis the behaviour of attack. There are about 5,000 threats log from January 2017 until December 2017. Python language is used as a tool to do the experiment. Anaconda is being used as a machine learning environment to do the data analysis. Anaconda is an open source distribution of the Python programming languages, one of data science platform and it is used in machine learning related application aiming at simplifying package management and deployment.

During the experiment, we used some of the anaconda libraries such as *pandas, matplotlib, seaborn, urllib.request, json,* and *socket.* In this experiment, we used tracking patterns data mining where this is the most basic techniques in data mining in learning to recognize patterns in data sets. This method is usually a recognition of some abnormality in data happened at regular intervals. Next discussion we presented the output from the data analysis that we conducted using the log from IPS. Figure 4 below explained research procedure conducted in this research.

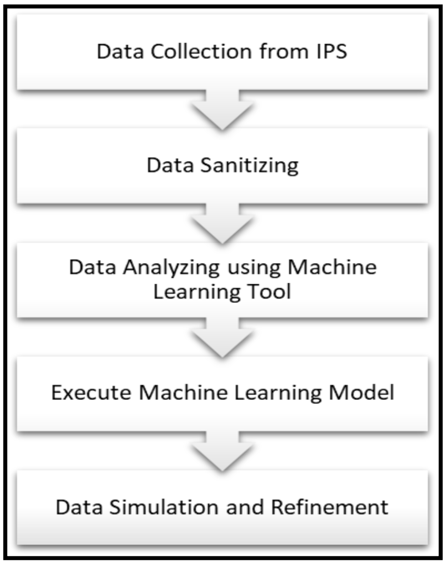
Data Collection from IPS
Data Sanitizing
Data Analyzing using Machine Learning Tool
Execute Machine Learning Model
Data Simulation and Refinement

Fig. 4 Research Procedure

3.1. Attack Analysis

From attacks capture in IPS, we have identified attacks that capture into the university network. In university network, there are some application from these systems which are financial portal, staff portal, leave portal, e-mail system, complaints system, facilities system, support system, student information management system, student's affair system, library system and learning management system. We have analysis packet form the network. Below is the list of top 15 attacks capture from university IPS. The highest number of attack is on Linux Kernel nfsd. This is a remote denial of service vulnerability exists in the Linux Kernel. The vulnerability of this is due to an implementation flaw which may result in a buffer overflow in the NFS subsystem of the Linux Kernel. The first 7 attacks from the graph are associated with database-related attacks, which targeted both MySQL and Oracle database. This local university network used MySQL and Oracle database for storing their users' data.

1. Linux: Linux Kernel nfsd CAP MKNOD Security Bypass Attack
2. ORACLE: XDB_PITRIG_PKG.PITRIG_DROPMETADATA Buffer Overflow II
3. RAT: GhostRat Traffic Detected
4. ORACLE: Server String Conversion Function Buffer Overflow
5. ORACLE: Application Server Printenv Information Disclosure
6. MySQL: MariaDB memcmp Function Security Bypass Vulnerability
7. MySQL: Password Brute Force
8. RDP: Microsoft Windows RDP Server Abnormal Termination
9. DCERPC: Suspicious DCERPC Call II
10. RDP: EsteemAudit Windows Remote Desktop Protocol Exploit

11. DCERPC: Suspicious DCERPC Call
12. ORACLE: Web Listener Batch File Vulnerability
13. Oracle: HTTP Server mod_access Restriction Bypass Vulnerability
14. HTTP: VMware Server Directory Traversal Vulnerability
15. ORACLE: Application Server Default Page SQL

3.2. Attack Category Analysis

From all attacks capture, they are in a high and medium severity to the university network. And most of the attack is done to exploit the network. An exploit takes advantage of a weakness in an operating system, application or any other software code, including application plug-ins or software libraries. Exploits are ultimately errors in the software development process that leave holes in the software's built-in security that cybercriminals can then use to access the software and, by extension, your entire computer. Exploits are commonly classified according to the type of vulnerability they exploit, such as zero-day, DoS, spoofing and Cross-Site Scripting (XXS). From the graph also shows how suspicious packet capture from IPS higher to the code-execution group. Code execution vulnerabilities occur where the output or content served from a Web application can be manipulated in such a way that it triggers server-side code execution. In some poorly written web applications that allow users to modify server-side files such as by posting to a message board or guestbook. It is sometimes possible to inject code in the scripting language of the application itself.

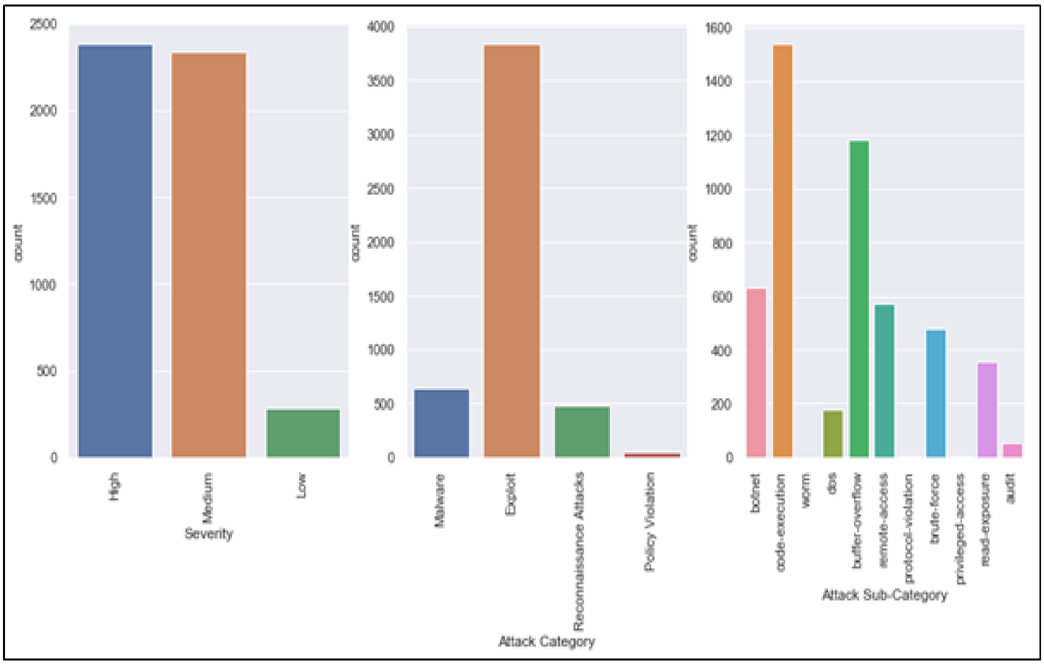

Fig. 5 Categories of Attack

In Figure 5 shows categories graphs in describing attacks in campus network. The discussion is observed from 3 group of attacks level, which are the severity, attack category and attack sub-category. The first graph show about level of severity of attack based on the policy setting up from the university environment. Second graph shows that exploit is the highest contribution of attack category in campus network and last graph shows other sub-category of attacks which code execution contribute the most in this campus network.

3.3. Network Port Analysis

We analysed network port attacker attempted to go inside the server. The attacks are attempted the server through port 443 and 3306. Port 443 is usually associated with TLS (https), and 3306 is usually with MySQL/MariaDB database. Campus network use secure sockets layer (SSL) and transport layer security (TLS) to encrypt their internet communications. The encryption protocols are utilized to ensure privacy and ensure data integrity. Unfortunately, the encryption protocols secure all application data, whether it is legitimate or malicious. As university aware about encrypting network traffic to protect data from potential attacks or exposure, attackers recover their Secure Sockets Layer/Transport Layer Security (SSL/TLS) information to hide their malicious activities. The port 3306 is a port associated with MySQL/MariaDB database. MySQL is the world’s most popular open source database system and MariaDB is the world’s fastest growing open source database system. A common attack at this port is the brute forcing of the root password for the MySQL database.

3.4. Time Analysis

In this topic, we discussed about time of attacks through a year 2017. Figure 6 shows the analysis of time of attacks for 24 hours’ time. From the graph, we can conclude that time of attack mostly happen from 12 midnight to 6 morning. The university network is not active at this time. Attackers prefer to do their activities at this time scale to prevent from being notice by the network administrator. Another time scale that attackers attack the university’s network is at 7 morning to 12 afternoon. At this time, insider users may become potential attackers in contributing this numbers of attacks.

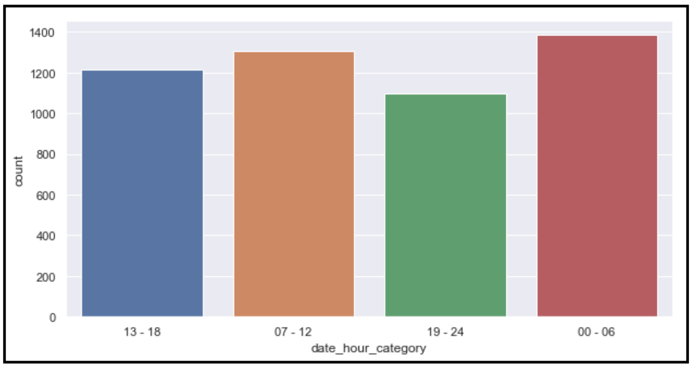

Fig. 6 Time in 24 hours Attack

A discussion continues to Figure 7 about activities throughout a year. This time scale is refer to a university’s calendar.

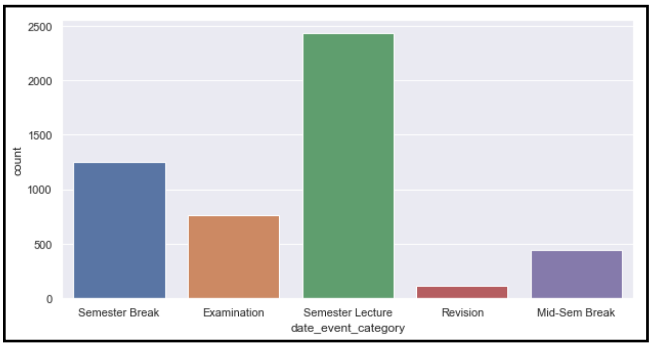

Fig. 7 Attacker Targeted Time

From this graph, we can see the attack happen during semester lecture period. One of the reasons this period is high frequency because of the insider users attack or spoof as legitimate campus users. An insider attack is a malicious attack committed on a network or computer system by a person with authorized system access. Insiders that perform attacks have a distinct advantage over external attackers because they have authorized system access and familiar with network architecture and system policies and procedures. In addition, less security against insider attacks because organizations focus on protection from external attacks.

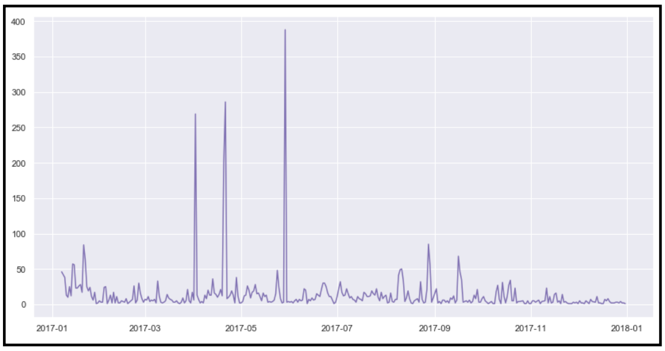

Fig. 8 Frequent Attack Time

From the graph in Figure 8 show that the most frequent date this university network detected attack are on 29 May 2017 about 388 attack hits, 21 April 2017 about 286 attack hits, 2 April 2017 about 269 attack hits and 20 April 2017 about 201 attack hits.

Among all top five attacks observed from those 4 different date, we analyzed that attacks associate with Oracle & MySQL database are the highest number of attempted from the attacker. We proposed to the administrator to update the software patches since attackers exploit the vulnerabilities from this type of databases. We also found an attack, which called EsteemAudit with highest hit. This attack targeted Windows Remote Desktop system on Windows XP and Windows Server 2003. This attack happen because of Microsoft stop releasing security updates to these two OSs.

4. Conclusions

As in University Information System, there are massive amount of personal information such as lecturer information, student information, staff information, exam information and the open nature of the higher education online learning environment also likely a target for a security breach. As a network administrator concern, developing a comprehensive incident response plan is important by conducting a risk assessment activity. It is important to understand our universities potential risks. In this paper, researchers proposed to use data from IPS as one of the steps in creating a library of risks. After narrow down threats list, we use machine learning software to take the next step of data mining and data analysis. By looking and the data from IPS, researchers used machine learning tool to track data patterns, and analyzing trends which lead to new or emerging risks.

The risk assessment is an important part of a risk management process to secure information systems. However, the protection of network system has become one of the challenges to the organizations especially with the increase of cybercrimes in university environment. Institution are expose to the cyberattack and this become a source of financial liability to universities. It is important to analyze our current security posture in campus network.

References

[1] Sengupta A, Manna A, Mazumdar C. 2013. A Graph-Based Approach for Managing Enterprise Information System Security. Cloud Ubiquitous Comput Emerg Technol (CUBE), 2013 Int Conf.;137–43.

[2] Asosheh A, Khodkari H, Hajinazari P. 2013. A Practical Implementation of ISMS. In. p. 1–17.

[3] Jones R, Stallings TJ. 2011 Challenges to network security on college campuses *.;37–42.

[4] Powell JH, Mustafee N, Chen AS, Hammond M. 2016. System-focused risk identification and assessment for disaster preparedness: Dynamic threat analysis. Eur J Oper Res [Internet].;254(2):550–64. Available from: http://dx.doi.org/10.1016/j.ejor.2016.04.037

[5] El-Ghareeb HA. 2009. E-Learning and Management Information Systems. eLearn [Internet]. 2009(9):8. Available from: http://portal.acm.org/citation.cfm?doid=1599450.1621693

[6] Joshi C, Singh UK. 2017. Information Security Risks Management Framework – A Step Towards Mitigating Security Risks In University Network. J Inf Secur Appl [Internet].;35:128–37. Available from: http://dx.doi.org/10.1016/j.jisa.2017.06.006

[7] Kam H. 2014. Information Security In Higher Education: A Neo-Institutional Perspective. J Inf Priv Secur. Volume 10(1):Pages 28-43.

[8] Roberts TL. 2013. Information Security in Higher Education: Threats & Response [Internet]. Global Information Assurance Certification Paper. SANS Institute;. p. 1–12. Available from: http://zma.es/Incident Handler/real-world-arp-spoofing/real-world-arp-spoofing_487.pdf

[9] Naagas MA, Palaoag TD. 2018. A Threat-Driven Approach to Modeling a Campus Network Security. In: International Conference on Communications and Broadband Networking.. p. 1–7.

[10] Joshi C. 2016 Quantitative Information Security Risk Assessment Model for University Computing Environment. In: International Conference on Information Technology.. p. 69–74.

[11] Cavallini S, Bisogni F, Bardoscia M, Bellotti R, Cavallini S, Bisogni F, et al. 2016. Assessing Potential Casualties in Critical Events. In: International Conference on Critical Infrastructure Protection. p. 231–42.

[12] Raman A, Kabir F, Hejazi S, Aggarwal K. 2016. Cybersecurity in higher education: the changing threat landscape [Internet]. EY technologies blog. p. 1–11. Available from: https://consulting.ey.com/cybersecurity-in-higher-education-the-changing-threat-landscape/

[13] Aylward LL. 2018. Integration of biomonitoring data into risk assessment. Curr Opin Toxicol [Internet].;9:14–20. Available from: https://doi.org/10.1016/j.cotox.2018.05.001

[14] Gürdür D, El-khoury J, Törngren M. 2019. Digitalizing Swedish industry: What is next?: Data analytics readiness assessment of Swedish industry, according to survey results. Comput Ind [Internet].;105:153–63. Available from: https://doi.org/10.1016/j.compind.2018.12.011

[15] Paltrinieri N, Comfort L, Reniers G. 2019. Learning about risk: Machine learning for risk assessment. Saf Sci [Internet].;118(June):475–86. Available from: https://doi.org/10.1016/j.ssci.2019.06.001

[16] Bilge L, Han Y, Dell ’Amico M. 2017. RiskTeller: Predicting the Risk of Cyber Incidents. ACM Conf Comput Commun Secur [Internet].;(1):1299–311. Available from: http://dl.acm.org/citation.cfm?doid=3133956.3134022%0Ahttps://acmccs.github.io/papers/p1299-bilgeA.pdf

[17] Wang Y, Xu W, Zhang Y, Qin Y, Zhang W, Wu X. 2017. Machine Learning Methods for Driving Risk Prediction. In: 3rd ACM SIGSPATIAL Workshop on Emergency Management. ACM

[18] Han J, Kamber M, Pei J. 2012. Data Mining Concepts and Techniques, 3rd Edition. 740 p.

[19] Muhamad W, Wan T, Laila N, Ghani A, Drus SM. 2019. Data Mining Techniques for Disease Risk Prediction Model : A Systematic Literature Review. In: Proceedings of the 3rd International Conference of Reliable Information and Communication Technology (IRICT 2018) [Internet]. Springer International Publishing; p. 40–6. Available from: http://dx.doi.org/10.1007/978-3-319-99007-1_4

[20] Cary K. 2019. Firewall Analysis and Operation Methods. SANS Institute; p. 1–37.