Recommender System based on Empirical Study of Geolocated Clustering and Prediction Services for Botnets Cyber-Intelligence in Malaysia

Nazri Ahmad Zamani[1], Aswami Fadillah Mohd Ariffin[2], Siti Norul Huda Sheikh Abdullah[3]
[1,2]CyberSecurity Malaysia, Level 5, Sapura Mines Seri Kembangan, Malaysia
[3]Cyber Security Faculty of Information Science and Technology Universiti Kebangsaan Bangi, Malaysia

***Abstract*—A recommender system is becoming a popular platform that predicts the ratings or preferences in studying human behaviors and habits. The predictive system is widely used especially in marketing, retailing and product development. The system responds to users preferences in goods and services and gives recommendations via Machine Learning algorithms deployed catered specifically for such services. The same recommender system can be built for predicting botnets attack. Via our Integrated Cyber-Evidence (ICE) Big Data system, we build a recommender system based on collected data on telemetric Botnets networks traffics. The recommender system is trained periodically on cyber-threats enriched data from Coordinated Malware Eradication & Remedial Platform system (CMERP), specifically the geolocations and the timestamp of the attacks. The machine learning is based on K-Means and DBSCAN clustering. The result is a recommendation of top potential attacks based on ranks from a given geolocations coordinates. The recommendation also includes alerts on locations with high density of certain botnets types.**

Keywords*— *Botnets; recommender system; predictive analytics; Big Data; cyber-threat intelligence; K-Means; DBSCAN

I. INTRODUCTION

Botnets are growing threats at global scale in recent years. This cyber phenomenon is growing exponentially with the increased of broadband penetration onto the global population. Furthermore, this trend is thought be directly related with mobile devices and computers are getting cheaper and more powerful over time. Botnets are malicious software (or malware) that ceded control of users devices and computers. These malware are responsible for being mediums for DDOS attacks, ransomware, and data mining [8]. According to SpamHaus Project site [9], the company malware division identified that there are 9500 botnets C&C servers detected on 1122 different networks worldwide. Malaysia is no stranger to the threat- the country is listed in rank 21 in the world botnet threat list with 96049 listed incidents detected in just 2017 alone [10, 11]. Adopting cyber security strategy framework provides an insight into the government's approach for protection of cyber space in the country [17].

Curbing the botnets incidents is quite an impossible feat as botnets infect machines via strength-in-numbers strategy. The more machine the botnets are able to infect, the better. Looking on the bright side, botnets are always in communication with their bot herders (or Command and Control) via their respective communication protocols. The exhibit communication IP addresses of these botnets can be traced of their geolocations. This information can be analyzed via Machine Learning to predict their next attacks patterns via geolocation vs. time information. Through this predictive analysis, IT experts are able to take precautions steps in mitigating the risks before the predicted botnets attack are taking place. Whether the predicted attacks are precise or the otherwise, considering the volume and the velocity of botnets attacks any good preparation could save an organization from any damage that they might incurred.

A recommender system is one of the popular applications that is built on top of a Machine Learning predictive analytics algorithms. It is widely used by companies such as Amazon, Target Corporation, and Netflix to drive sales. A recommender system predicts consumers' interests and provides recommendation to them through Machine Learning. The same concept can be applied for botnets patterns. This paper demonstrates our implementation of such recommendation system that is developed from acquired telemetric botnets sensors data. The predictive analytic is used to learn and analyze botnets source IPs and the target IPs. The results from such learning can be used to provide either a warning on the top-10 botnets attack or a warning on certain botnets attack based on user input geolocations.

The feasibility of such system opens up a novel defense mechanism that is scalable to the volume, velocity and the veracity of botnets activities nationwide versus timeline. Such scalability is an enabling feature to analyze more data and to come up with more accurate results in this era of ubiquitous computing. An accurate recommendation system is essential in not just as monitor-alert system but as to provide better consequential planning and actions for remedial or triage.

II. BACKGROUND

Botnet signals the bot herders from internet-connected devices IP addresses. From the signals the location of the source IPs and the Command & Control (C&C) IPs can be located and logged via sensors. The number of recorded signals can be in huge volume and in high velocity in daily basis. Analyzing the logs can cost a lot of computing resources and therefore only a Big Data set up can handles such magnitude.

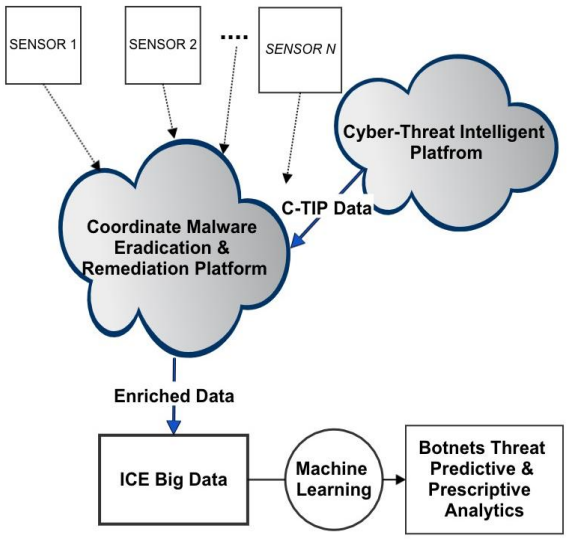

Fig. 1. ICE Recommendation System for Botnets Attack Prediction Services Schematic.

CyberSecurity Malaysia has developed Integrated Cyber-Evidence (ICE), a Big Data platform to analyze huge volume of digital evidence. The system comprises of nine processing nodes in which are all managed via Apache Metron [14].

For the recommendation system, ICE gets the input from malware cyber-intelligence enriched data from Coordinated Malware & Remediation Platform (CMERP). CMERP is another Big Data platform that CyberSecurity Malaysia developed to analyze in real-time data-in-motion from sensors deployed and from the Microsoft Cyber-intelligent Platform (or C-TIP). C-TIP here is a cloud service platform that Microsoft developed to monitor and fight botnets and the threat actors that run the malicious codes [13]. The enriched data that the CMERP tabulated is a precursor to the ICE system, in which the predictive analytics that leads to recommendation system can take place while the data is at rest. The following Fig.1 shows the schematic of how ICE botnets recommendation system works.

In building a recommender system, it is essential to firstly determine the type of application that is required and how it can be useful from the data ingested. The first step is to do the Exploratory Data Analysis (or EDA). EDA is performed in order to understand the data through statistical and visualization means in order to knife-out strategy to build a recommendation system. The botnets cardinality and IP addresses geolocations are drawn to determine the right algorithms that could cluster the geo-patterns. K-Means is one of the common used clustering algorithms in partitioning observations on the data-space into Voronoi cells. Another clustering algorithm under consideration is Density-based spatial clustering of applications with noise (or DBSCAN). DBSCAN is known for its density-based clustering- it groups together points that are closely packed together. With similar motivation of [18,19] study, projecting these Botnets into these data-space clusters may fit the predictive requirements for the geo-patterns of the signals.

III. Related Works

There are several ways used in various cyber security researches [16] to detect malware either in anti-virus software or end point protection such as signature based and behavioral based. Unlike signature-based, behavioral based able to detect malware that uses obfuscation technique even though it is time consuming with considerable false positive. This paper is inspired by the works of Coreia [12] and Casey [15]. Casey proposed a recommendation-verification system for predicting the Zeus malware infections via the Signaling Game methodology. Coreia on the other hand is using statistical characterization of botnets and their respective Command & Control traffics. His studies are focusing on the characterization of Denial of Service (DoS) attacks, spamming or phishing activities. Wei Xu [3] build a system on top various data feeds that predicts malware via malicious DNS domain. The system is leveraging on the knowledge of the life cycle of malicious domains, as well as the observation of resource re-use across different attacks. Another similar work by Truong & Cheng, where they proposed a method to detect Zeus and Conficker that utilizes domain fluxing by analyzing the extracted the DNS traffic length and expected value that can distinguish between a domain name, by a human or botnets [5]. Nguyen & Tran on the other hand, modeled user behaviors and applying heuristic analysis approach to mobile logs generated during device operation process [7]. For the task, they proposed a lightweight semantic formalization in the form of physical and logical taxonomy for classifying collected raw log data. There is also work done on honeypot dataset presented by Dowling & Seamus in [4]. In this paper, Seamus presented their analysis on honeypot dataset to establish attack types and corresponding temporal patterns. Their analysis shows the calculation of the probability of each attack type occurring at a particular time of day. Then they test these probabilities with a random sample from the honeypot dataset to see the geo-distributed patterns of the attacks. These attacks can take many forms and can come from different geographical sources. Another work that involves the applications of honeypot and sandboxing is by Mariconti, where he uses classification technique to learn about the different network behavior patterns demonstrated by target malware and generic malware [1]. They set up a sandbox and infected virtual machines with malware, recording all resulting malware activities on the network and then extracted meaningful features for classification. Gupta in [6] used Probabilistics Data Structure, specifically Bloom filter for setting membership on data the number of hits on suspicious nodes per unit time in network traffic data for their IDS (Intrusion Detection System). Lastly, Mezzour used empirical test alternative hypotheses on factors variations in the number of malware encounters [2]. Their analysis is focusing on regression analysis to test for the effect of computing and monetary resources, web behavior, computer piracy, cyber-security expertise, apart from the number of malware encounters. Another improvement of anomaly intrusion detection system has been proposed based on hybrid approach namely Fuzzy-ART and k-means clustering algorithm [20]. They swap the usage of both of the methods for getting the initial stated value for K-means, using Fuzzy-ART whereas

Fuzzy-ART uses K-means to reassign the data instances within clusters.

IV. Proposed System

The proposed system comprises of two sub-systems. The first sub-system is the training on the geolocation and the latter is the recommendation system. The training is the process of learning the enriched data, in which consists of preprocessing, clustering, in-hulling and merging geolocation data from cyber-threats enriched intelligence data. A prerequisite for the training phase to initiate is the EDA in getting the data overview in term of statistics, data visualization.

The recommender system is the process that responds to the geolocation input by user and return location alerts on types of botnets threat in the enquired vicinity. The system would also check at which cluster hulls that the requested geolocations inputs are closely correlate. It will return a warning for a user if the location is indeed a hotbed for botnets infections.

A. Exploratory Data Analysis

The main task of EDA is the hypothesis testing via statistics and data visualization. The first step is to seek the magnitude by cardinality of botnets activities in Malaysia. The following Fig.2 shows the word clouds visualization of the botnets infection in Malaysia in the first quarter of 2016 alone.

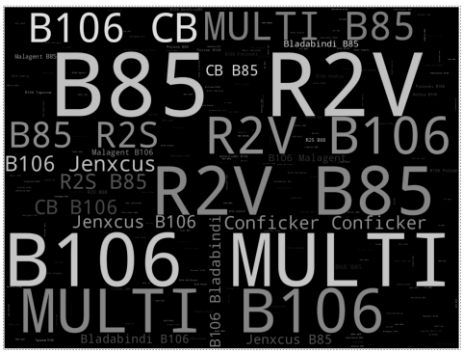

Fig. 2. The Word Clouds that Depict the Magnitude of Botnets Infection in Term of Cardinality.

From the word cloud, we can assume the top-3 botnets infections (with their own descriptions and threat risk) as tabulated in Table 1.

TABLE I. The Rank of Botnets Infections Analyzed from the Data

Rank	Threat code	Description	Threat Risk
1	B85-R2V	Family of worms that can steal usernames and passwords. Some can use PC for DoS attacks.	**Severe**
2	B106-MULTI	PC infected with multiple types of Botnets malware.	**Severe**
3	B106-CB	PC that is infected with both B106-Bladabindi & B106-Jenxcus	**Severe**

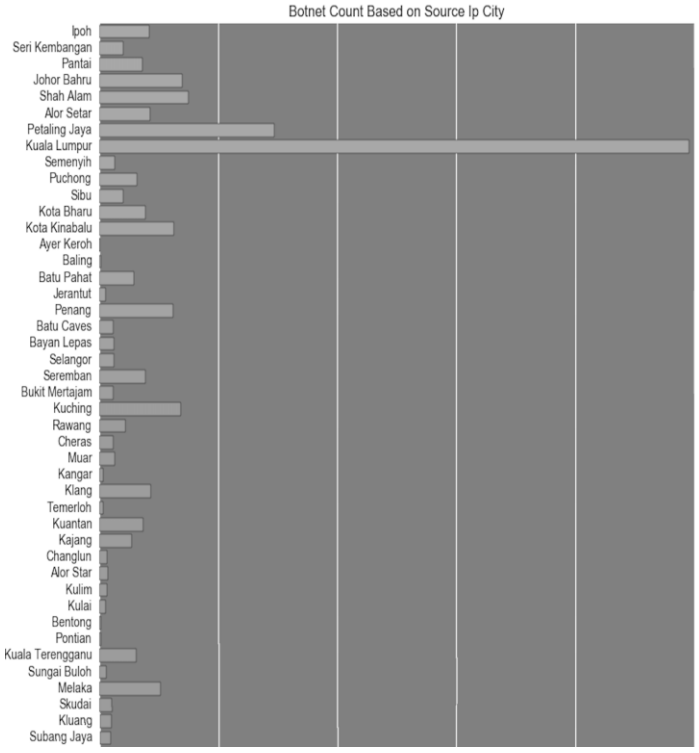

Fig. 3. The Histogram of Botnets Infections Versus Source IP Cities.

The next analysis is to find the distribution of infections in each of the city observed. Fig.3 shows the histogram of the number of infections over cities source IPs. The two cities which have the highest number of botnets infections depicted from the histogram are Petaling Jaya and Kuala Lumpur. Kuala Lumpur has the highest record of infection: almost reached a ten millions of records at just in the first quarter of 2016. Petaling Jaya on the other hand is lagged behind in number by less than a million. The numbers of infections of these two adjacent cities can easily balance out the number of other cities combined. In fact 70% of the infections are focused on regions adjacent to these two cities, in which are hubs for government offices, financials, manufacturing and commercials. Therefore, the predictive analytics for the botnet infections can be given top priority on these cities. Furthermore, the threats presented in Table 1 are commonly found in these areas.

B. Training Process

The Training process is a phase of selected machine learning algorithm learning from the ingested enriched data. The data is further pre-processed to discriminate columns required for the analysis. The data is then trained in two-tiers Machine Learning – K-Means and DBSCAN. K-Means is the first-tier clustering that clustered and partitioned the botnets IP address geolocation based on the number of botnets observations. The resulting Voronoi partitions are then taken on to the second-tier clustering, in which DBSCAN is to mark outlier points that lie alone in low-density regions (whose nearest neighbors are wide apart). The final cluster results are then merged for the next Recommendation process. Fig.4 shows the explained the proposed Training process.

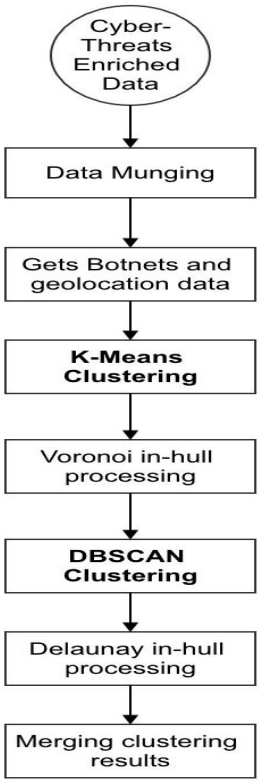

Fig. 4. The Proposed Training Process Flow.

One of the challenges in the training phase is to determine the optimal value of *k* in calculating the K-Means for such huge dataset. We solved the problem by estimating the value via the Elbow method. The idea of the elbow method is to run K-Means clustering on the data for a range of values of *k* (for this setup 1 to 15), and for each value of *k* we calculate the sum of squared errors (SSE). The goal is to choose a small value of *k* that is still has a low SSE, and the elbow of the curve is usually represents of where the SSE value have diminishing returns by the increasing *k*. Fig. 5 shows the line plot of SSE versus the increasing *k* values. Here, the optimal value *k* at the elbow of the curve is *k*=5. This value is the one applied to the K-means clustering for the dataset.

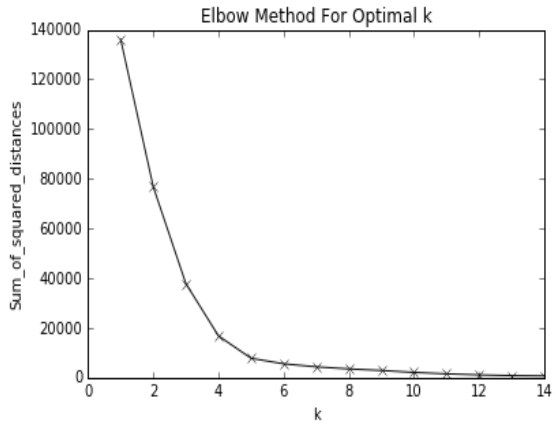

Fig. 5. The Elbow Method for Determining the Optimal *K* Value for K-Means.

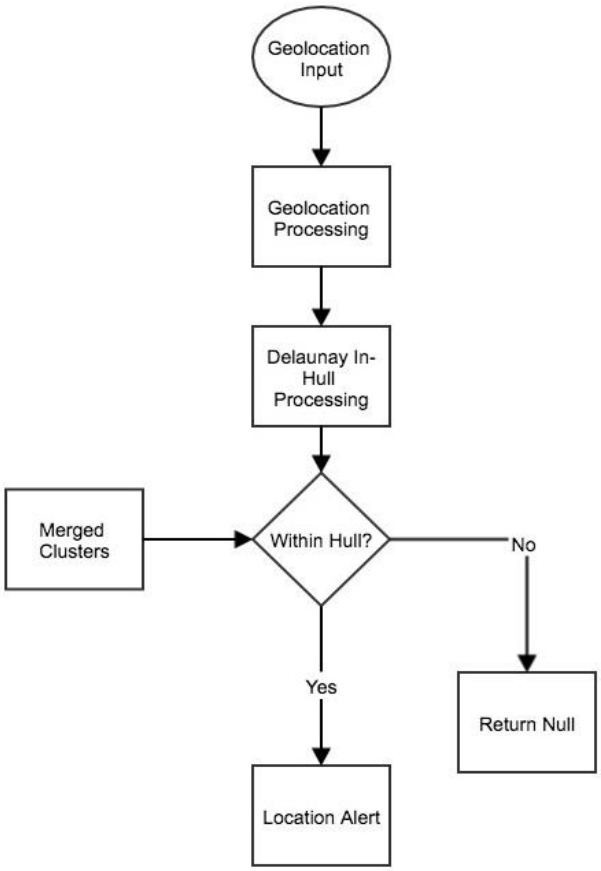

Fig. 6. The Recommendation Process Flow.

C. Recommendation Process

The recommendation process is where the system gives possible predicted botnets attacks based on a geolocation input. The input is processed with in-hull Voronoi and Delaunay processing as to find whether the distance correlation between the enquired geolocation points to the clusters of trained botnets threats is within hull. If the points are within hull, the location alert is issued with suggestion of related botnets types that may exist within the vicinity of the location. Fig.6 shows the Recommendation process flow.

V. Method of Experiments

The experiment is conducted on the enriched Cyber-Threats Intelligence data from the mentioned first quarter of year 2016. From the EDA performed on the data, we had decided to focus the predictive analysis on Kuala Lumpur, Petaling Jaya and other connected cities around Klang Valley. The recommendation system will show the top-3 botnets within the enclosure of the Klang Valley. In the second part of the recommendation system, an alert is given by the system on the threats that may exist from a given geolocation.

VI. Analysis and Results

Our analysis shows that from the millions of log on the botnets infections, the K-Means clusters formed the Voronoi as depicted in the Fig.7. Each of the Voronoi cell represents malware cluster label.

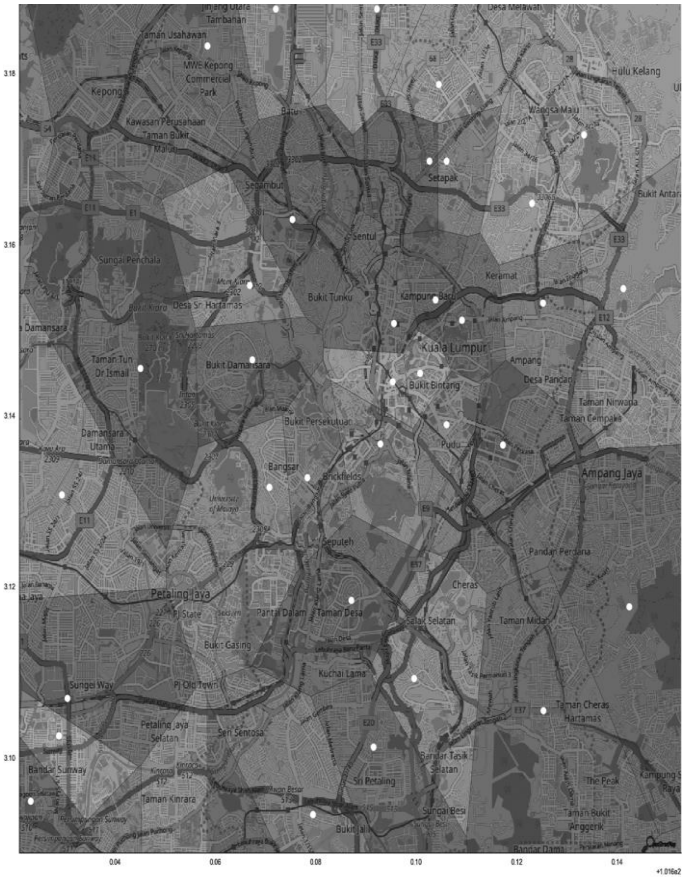

Fig. 7. The Voronoi Cells of K-means Clustering of Botnet Threats Around the Klang Valley.

From a selected geolocation point around the city center of Klang Valley, we can see the top 3 botnets threats, and the suggested locations of each suggested botnet within radius from the requested point. The Fig.8 shows the example of top three predicted botnets and the suggested dots are areas of which botnets infections are found to be active within the point parameters.

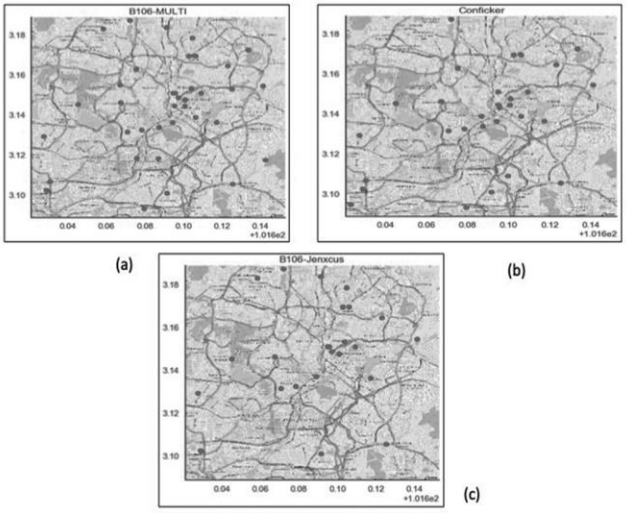

Fig. 8. The Example of Suggested top-3 Botnet Threats from a Requested Geolocations Point–(a) B106-MULTI, (b) Conficker, and (d) B106-Jenxcus.

Fig. 9. The Alert of Botnet Threats around the given Geolocations Point, (*101.67, 3.139003)*.

In the second part of the recommendation system, a geolocation is chosen randomly at (*101.67, 3.139003)*. As shown in the Fig. 9, the system gives a recommendation of types of botnets threats existed. There are 73 warnings issued on the threats existed within the parameter of the requested geolocations, and is dot-labeled to differentiate each of the threat.

VII. Conclusions

Recommendation system is crucial in informing the public on the predicted botnet threat landscape based on their local area. The system is capable of providing predicted botnets threats that users have to be aware of in an area. The recommendation system works via Machine Learning algorithms inside ICE Big Data environment. ICE system learns the botnets threats IP geolocations through K-Means and DBSCAN clustering and partitioning the threats geographically. As a result the system will provide the top-three botnets in the alert feeds along with other suggested targeted areas on the map for awareness.

Acknowledgment

This Botnet Recommendation system is a project that is a part of the DSTIN program on Integrated Cyber-Evidence platform development, granted by Ministry of Energy, Science, Technology, Environment and Climate Change (MESTECC) to CyberSecurity Malaysia. A research grant UKM-AP2017-005/2 under Universiti Kebangsaan Malaysia has been a collaborator and partner to CyberSecurity Malaysia in research, academic and outreach programs especially in the field of digital forensics.

REFERENCES

[1] Mariconti, Enrico, et al. "What's your major threat? On the differences between the network behavior of targeted and commodity malware." Availability, Reliability and Security (ARES), 2016 11th International Conference on. IEEE, 2016.

[2] Mezzour, Ghita, Kathleen M. Carley, and L. Richard Carley. "An empirical study of global malware encounters." *Proceedings of the 2015 Symposium and Bootcamp on the Science of Security*. ACM, 2015.

[3] Xu, Wei, Kyle Sanders, and Yanxin Zhang. "We know it before you do: predicting malicious domains." *Virus Bulletin Conference September*. 2014.

[4] Dowling, Seamus, Michael Schukat, and Hugh Melvin. "Using analysis of temporal variances within a honeypot dataset to better predict attack type probability." Internet Technology and Secured Transactions (ICITST), 2017 12th International Conference for. IEEE, 2017.

[5] Truong, D.-T. Cheng, G. "Detecting domain-flux botnet based on DNS traffic features in managed network". Security and Communication Networks. Issues 14 Volume. 2016.

[6] Gupta, D, Garg, S Singh, A Batra, S Kumar, N Obaidat, M S. "ProIDS: Probabilistic Data Structures Based Intrusion Detection System for Network Traffic Monitoring". GLOBECOM 2017 - 2017 IEEE Global Communications Conference. 2017.

[7] Nguyen, G, B M Nguyen, D Tran, and L Hluchy. "A Heuristics Approach to Mine Behavioural Data Logs in Mobile Malware Detection System." Data and Knowledge Engineering 115: 129–51, 2018.

[8] Tara Seals. "Bad Botnet Growth Skyrockets in 2017". Infosecurity Magazine. https://www.infosecurity-magazine.com/news/bad-botnet-growth-skyrockets-in. 2018.

[9] SpamHaus. The SpamHaus Project. https://www.spamhaus.org/statistics/botnet-cc. 2018.

[10] Composite Block Listing. "CBL Breakdown by Country, Highest by Count". Composite Block Listing, A Division of SPAMHAUS. https://www.abuseat.org/public/country.html. 2018.

[11] SeyedAliReza Vaziri. "Botnets: As We See Them in 2017". Ripe NCC. https://labs.ripe.net/Members/alireza_vaziri/botnet. 2018.

[12] Correia, Pedro, Eduardo Rocha, António Nogueira, and Paulo Salvador. "Statistical Characterization of the Botnets C&C Traffic." Procedia Technology 1: 158–66. 2012.

[13] Microsoft. Microsoft Secure. https://www.microsoft.com/en-us/security/intelligence. 2018.

[14] Apache Metron. http://metron.apache.org. 2018.

[15] William Casey, Evan Wright, Jose Andre Morales, Michael Appel, Jeff Gennari, Bud Mishra. "Agent-based trace learning in a recommendation-verification system for cybersecurity". 2014 9th International Conference on Malicious and Unwanted Software: The Americas (MALWARE). IEEE. 2015.

[16] Rami Sihwail, Khairuddin Omar, Khairul Akram Zainol Ariffin. A Survey on Malware Analysis Techniques: Static, Dynamic, Hybrid and Memory Analysis, 8 (4-2) : 1662-1671, 2018

[17] Salamzada, Khosraw and Zarina Shukur, and Marini Abu Bakar. "A framework for cybersecurity strategy for developing countries: case study of Afghanistan". Asia-Pacific Journal of Information Technology and Multimedia, 4 (1). pp. 1-10. ISSN 2289-2192. 2015.

[18] Mohammed Ariff Abdullah, S.N.H.S. Abdullah, Md Jan Nordin. "Additional Feet-on-the-Street Deployment Method for Indexed Crime Prevention Initiative", Jurnal Pengurusan 53(2018) 20 pages ,2018.

[19] Kamarul Ismail, and Nasir Nayan, and Siti Naielah Ibrahim. "Improving the tool for analyzing Malaysia's demographic change: data standardization analysis to form geo-demographics classification profiles using k-means algorithms". Geografia : Malaysian Journal of Society and Space, 12 (6). pp. 34-42. ISSN 2180-2491. 2016.

[20] Zulaiha Ali Othman, and Afaf Muftah Adabashi, and Suhaila Zainudin, and Saadat M. Al Hashmi. "Improvement anomaly intrusion detection using Fuzzy-ART based on K-means based on SNC Labeling". Jurnal Teknologi Maklumat dan Multimedia, 10 . pp. 1-11. ISSN 1823-0113 Item availablity restricted. 2011.