Revealing Influenced Selected Feature for P2P Botnet Detection

Wan Ahmad Ramzi W.Y[1], Faizal M. A[2], Rudy Fadhlee M. D[3] and Nur Hidayah M. S[4]

[1] Masjid Tanah Community College, Ministry Of Higher Education Malaysia, Paya Rumput, 78300 Masjid Tanah , Melaka.
[2,3,4] Information Security, Digital Forensic, and Computer Networking (INSFORNET), Department of Computer System and Communication, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia.

Abstract: P2P botnet has become a serious security threat for computer networking systems. Botnet attack causes a great financial loss and badly impact the information and communication technology (ICT) system. Current botnet detection mechanisms have limitations and flaws to deal with P2P botnets which famously known for their complexity and scalable attack. Studies show that botnets behavior can be detected based on several detection features. However, some of the feature parameters may not represent botnet behavior and may lead to higher false alarm detection rate. In this paper, we reveal selected feature that influences P2P botnets detection. The result obtained by selecting features shows detection attack rate of 99.74%.

Keywords**:** P2P Botnet, Botnet Detection, Feature Selection, Malware, Flow Analysis, Regression

1. Introduction

Nowadays communication networks are widely used in various fields such as business, education, banking and building communication networks around the world. All this activity takes place every day and certainly the security is the factor that necessary to ensure they are safe from any cyber-criminal activity. To protect user data, research on security features should be review and understand in order to prevent any dangerous threat to Internet users. The challenge of dealing vast complexity and massive network, in line with the fast increasing of threat such as malicious software(malware) or botnets. Botnets are remotely controlled networks of hijacked computers by botmasters [1]. According to researcher [31] the word Bot is derived from the word ROBOT which will infected host machine to become a ZOMBIE machine. The intention is to do the distribution of spam emails, coordination of distributed denial-of-service attacks, and automated identity theft [2]. It is a necessity to understanding traffic behavior for each network activity as a measure of protection from the attack. Botnet uses revision numbers updates techniques through the internet over the time which lead to the characteristic of a high degree of anonymity, make the botnet is very difficult to recognize.

Botnet attacks in Malaysia rising from year after year. Furthermore, it is the most advanced cyber threats nowadays. The example of botnet attack includes e-mail spamming, sniffing network traffic, malware distribution, fraud, Distributed Denial of Service (DDoS), and more. Botnets can result in huge losses to a nation and can be used to hack into the confidential information either banking or defense system in the country. According to the Incidents Report of General Incident Classification Statistic 2015 reported by Malaysia Cyber Security, botnet agent is leading with the big number of cases especially fraud and spam ranked top 2 in the total cases reported to the Cyber Security from January to October year 2015 [3].

The report is supported by the ‘Kosmo’ Online news stated that Malaysia being a target of more than a million botnet attacks aimed at invading the database in the local computer system during the first 10 months the year 2015 [4]. Hence, botnet attacks must be prevented with the need for more researchers doing research on bot-net detection.

Recently, more study focuses on P2P Botnet because it is considered a serious threat to P2P application users.

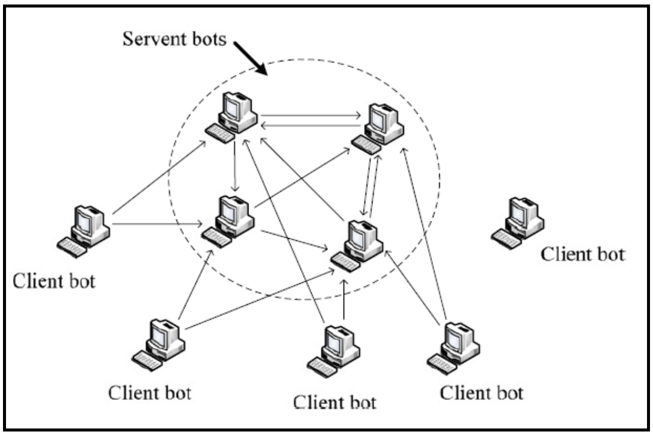

Figure 1. Architecture of P2P Botnet [5]

P2P botnet is a decentralized network topology causes the detection even harder for security researchers to trace the communication source compare to previous bot-net topology. Fig. 1 shows the Command and Control architecture of the P2P botnet. Each bots initiating a connection from one into another with its own peer list which only involved servant bots’ IP addresses and connects to all the bots through his own peer list to form a P2P botnet [5]. P2P botnet has made companies’ and bank’s web-sites as their victim, using the DDoS attack. Among victims involved are WordPress, PayPal, MasterCard and Yahoo mail [6]. Zero Access rank as the largest P2P botnet size estimated 1.9 million zombies, profitable to US$60,000 up to US$120,000 a year for a basic package that has been sold as a service on various underground hacker forums [7].

Therefore, it is important for the security mechanism to design a good system to prevent any suspicious activity from accessing the system or data resources. This will avoid the significant losses suffered by the company as claimed. Designing a good botnet detection system requires the system to detect efficiently in term of detection accuracy and low false detection alarm. One of the ways to reduce false alarm is by exploring the best detection features for more effective intrusion detection system. It is supported by Amrita and P. Ahmed [8] which stated that the effect of feature se-lection capable to shorten the training and testing time, guaranteed on high detection rates and makes IDS suitable for real time and on-line detection of attacks.

Therefore, this paper is aim to reveal selected feature that influences P2P Botnet detection. This paper consists of following section: section 2 discussing on related work of P2P Botnet Detection using detection feature. The methodology is briefly discussed in section 3 and followed by result and discussion in section 4. Conclusion and future work in section 5. Lastly, the acknowledgment and references at the end of the paper.

2. Related Work

In this section, a study related to the topic is conducted to support the importance of the research. The definition of network flow, feature selection, and statistical approach will be discussed in this section.

2.1. Network Flow Analysis

Network flow analysis is defined as traffic stream consists of same source IP, destination IP, network protocol, source port and destination port that traveling between two computers and it came with a common set of identifiers [9]. The devices such as routers and switches in the computer network can generate traffic flow depends on the stream of traffic through it. Traffic data will then be sent to a flow collector that will generate statistical reports from flow updates.

Flow analysis can be used to determine traffic statistic in overall and also a good approach to understand the traffic traversing the network which capable to tracks the fields such as Source interface, Source and destination IP, layer 4 protocol, types of service value and much more. In addition, many research used flow analysis on botnet detection to reveal the message contents between server and clients by monitoring of network traffic [10]. The advantages of using network traffic characteristic for detection is immune to the encryption algorithm and computationally cheaper than other approaches.

2.2. Feature Selection

According to Features selection for Network Intrusion Detection System (NIDS) is one of the fundamental steps in detecting botnets as it is used as classification in datasets entity and one of basic step in preprocessing of data mining. Features selection are widely used as a technique to eliminate the redundant and unnecessary features which also describe as a process of selecting a subset of related features that contributing in element of NIDS and having a false data and redundant data may result in false correlations which will interfere in the learning process of the classifiers [11]. According to authors [12], features selection technique is applied by many practitioners for reducing dimensionality by focusing on small subset of relevant features from the original based on a specific assessment of the relevant criteria, which usual-ly leads to better learning performance such as higher learning accuracy for classifi-cation which is lower computational cost, and better model interpretability.

Based on evaluation criteria, features selection can be categorized into three categories known as filter model, wrapper model, and hybrid model. Filter model depends on general attributes of the training dataset independently from classifier feedback to select best features. Meanwhile wrapper model optimises classifiers as part of features selection[12]. Since Wrapper models obtain better predictive accura-cy estimates than filter models [13] and aim to select features that maximize the quality, therefore the Wrapper model is chosen with flavors Forward feature selection technique would be applied for this research. This is supported by [14] stated that the preferred approach for feature selection is wrapper approach as it can handle large dimensional data while filter approach has less computational complexity and it uses independent subset evaluation criteria for subset evaluation, so wrapper approach is suitable for feature selection.

2.3. Statistical Approach

According to authors [15] Statistical method refers to a range of techniques and procedures for analyzing data, interpreting data, displaying data, and making decisions based on data. The authors [29] noted that Statistical approach is a solution for encrypted traffics classification. Meanwhile, researchers [16] noted statistical methods is the science of learning from data which is a set of principles or procedures that use by the scientist in their pursuit of knowledge.

Many of the researchers use statistical techniques to find the best way to discovering, identify their chosen methodology with thought this approach is the most appropriate, using the estimated values of the parameters. This is supported by researchers [17] which use a statistical approach to producing a novel technique to formalization methodology in identifying a critical malicious pattern among malware families. The authors present basic blocks of the malware control graph which classify them into their corresponding malware family by computing the Frequency Distribution Ratio for each basic block within each malware family. The author found their novel approach is more consistent compared to related works. Authors [18] proposed a random effect logistic regression model to predicting anomaly detection. The research-based on a sample of 49,427 random observations for 42 variables of the KDD-99 dataset containing ‘normal’ and ‘anomaly ‘ connections. Six selected features with five input variable selection are performed consists of discrete, continuous and Binary data type categories. Although the proposed model has high classification accuracy and a high percentage of data set validation, there is a weakness in this approach which does not accommodate the situation where the system exhibit different attack probabilities under the same condition.

The Authors [30] performing a combination between statistical method and machine learning method to solve the dependencies problem on the signature network packets. The idea is to see the simples’ way to measure an elements in the developed system to identify network interruptions. Seven statistical features are chosen to studies its intrusive behavior from network traffic which contributing to the effectiveness of Intrusion Detection System (IDSs) against multi types of network attacks. Researchers [19] performing logistic regression for Malware signature based detection. This research produces a solution to the limitations on signature-based and anomaly based. Two theory generated which make this methodology is better than the previous signature based detection and even better compared to anomaly detection method with the less false positive rate. A sample of the Slammer worm is used to demonstrate slammer model test between three IDS model which are Slammer Logistic Regression Signature, Signature Detection McAfee and Signature Detection Norton. The proposed model had a 100% detection rate with no false positive.

Logistic regression is one of the tools for applied statistics and commonly used in the discrete analysis. The advantages of logistic regression are ease of use, flexibility, and the ability to apply logistic regression to many subject areas [20]. According to author [21], Logistic regression is one of the regression analysis approaches which are used to predict an outcome when the dependent variable is categorical (binary variable). Meanwhile, researchers [22] using logistic regression as one of the testing models to binary based detection approach and had the best accuracies and scores.

Most of the researchers' only focus on the detection technique and not to disclose the influence of the selected features. The exploration of the features still needs to be revealed to produce a robust botnet detection techniques that can be built to identify properties of the features. Besides, there is no researcher mentioning or using proper technique in identifying the static threshold to differentiate normal traffic activities and abnormal traffic activities. Therefore, logistic regression is the suitable approach to detect P2P botnet to identify a suitable threshold value for botnet detection.

By revealing the influence of the selected features, it will filter and find the most significant features from the total of selected feature from Feature Selection Module and it may increase a level of confidence to the contribution of each feature before it can be used in the detection module. This technique is based on previous researcher [23] which is focusing on time-based traffic feature or derived feature and using Bro to distinguish normal and abnormal connection in the network traffic. The output from the timed based module is used for IP comparison and lastly went through logistic regression model to access the feature influence inside the detection model [24].

3. Methodology

3.1. Data Collection

The data collection is been done by performing test lab in a P2P environment which setup at Security Laboratory of Universiti Teknikal Malaysia Melaka (UTeM). The test lab consists of four computers and one switch that connected to the external network. The malicious files are provided by Cyber security Malaysia and it will be used by releasing it into four client computers and let it on with open connection with P2P software applications is running for one week period of time. The purpose of test lab is to see the interaction between malicious files and C&C server or outside peer that would be contacted each other and recognize as abnormal traffic.

Each dataset has a different amount of network flow examples and classes. The dataset is named according to its botnets variants to avoid any dataset confusion. Table 1 describes the information dataset names on a number of examples available and the number of classes given to differentiate normal and abnormal traffic.

Table 1. P2P Botnet Dataset

CSV NAMES	P2P BOTNET	SIZE	EXAMPLE S
Cryptowall.csv	Cryptowall	1.692 MB	5088
Neris.csv	Neris	8.72 MB	26592
Kelihos.csv	Kelihos	13.92 MB	42450
Rbot.csv	Rbot	1.75 MB	5237
T.bot.csv	T.bot	1.73 MB	5179
Zbot.csv	Zbot	3.96 MB	11644
Zeroaccess.csv	Zeroaccess	1.34 MB	5131

3.2. The Pre-Processing: Principles of Traffic Network Analysis

In preprocessing, data will be going thru data cleaning, which missing attributes will be filled, reducing the noise inside the dataset, data integration, where multiple data will be tested, data transformation and data discretization will be held.

During the process of data cleaning, variant file in .csv format has some noise or imperfection of data that must be overcome. From the 89 attributes that extracted based on TCP traffic, only 48 attributes were selected. There are outliers and discrepancies in codes or names that must be solved. Therefore, the data must be going through a cleaning process before it can be run in classification technique. Inside the value of each attribute, there are attributes contains a missing value such as “NA” which means ‘Not Available’ and for symbol ‘?’ bring the inconsistent with the definition of the variable. The software Rapid Miner deducts that there are missing values. Any attributes or features that contained these value will be eliminated according to it line or records. It is because the classification techniques cannot run with the data that contain a missing value or else it will become noise to the accuracy rate. In this datasets, unnecessary attributes have been identifying and data normalization has been conducted. Out of 89 attributes, only 48 attributes has been filtered and tested.

3.3. Implementing Feature Selection Module

Feature selection method is to choose a subset of the variable in the training set and used the chosen variable as predicted features [28]. In this phase, Feature selection will be held to select the best features out of 52 features that were tested. In the pre-processing, Feature selection is categorized in Dimensional Reduction which to reduce of patterns in the patterns and can be done by heuristic methods such as step-wise forward selection, step-wise backward elimination and decision-tree induction [25]. Feature selection methods can further be broadly categorized into filter models, wrapper models and embedded models. The filter model relies on measures of the general characteristics such as distance, consistency, dependency, information, and correlation. While wrapper model uses the predictive accuracy of a predetermined learning algorithm to determine the quality of selected features [12].

Features selection module functionality is capable to eliminate the redundant and unnecessary features which also describe as a process of selecting a subset of related features that contributing in the element of NIDS and having a false data and redundant data. It may result in false correlations which will interfere in the learning process of the classifiers [11]. In this phase, the Wrapper model is used by applying for-ward operator in the RapidMiner analyzer tools as it optimizes classifiers and able to handle large dimensional data.

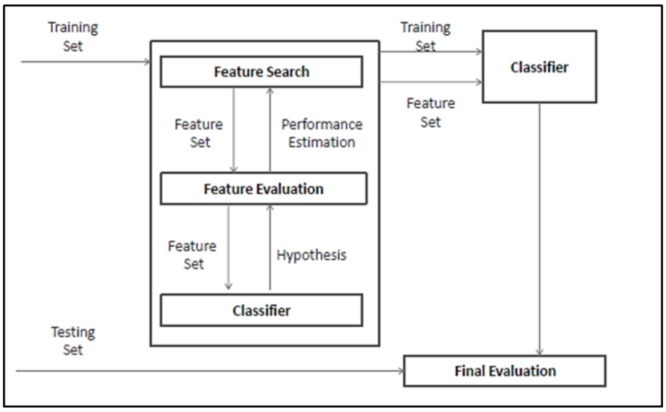

Figure 2. Framework for Wrapper Methods of Feature Selection Classification. [12]

Wrapper method typically performs a three-step process which are:

Step 1: Searching a subset of features
Step 2: Assess the option subset of features using a classifier performance
Step 3: Step 1 and step 2 is repeated to achieve the desired results.

Fig. 2 illustrates a general framework for wrapper methods that were used at this stage to find the selected features out of 52 features. Three components are involved according to the figure which are Feature search, Feature evaluation, and Classification.

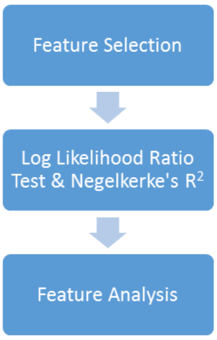

Figure 3. Feature Influence Process Flow

In wrapper model, the first components of Feature searching will generate the set of feature which then will evaluate the estimate performance by the classifier in the feature Evaluation component. The estimation result will be used again by the searching agent for the next iteration process to find the highest estimated value of the features selected for learning classifier as the final results.

Forward selection is used as classifier performance performing greedy searching strategies which start searching at the empty set of features and then gradually joined into a larger subset. The purpose is to find the best evaluation value of the selected feature out of total features that are tested.

To explore the feature influence, researcher applied Log Likelihood Ratio test and Negelkerke's R2 test to reveal the influence of the feature from the selected features generated by feature selection process. The result of the model then is used to identify significant contribution among the feature selected. Figure 3 shows the process flow in exploring the feature influence in detecting P2P botnet attack.

3.3.1. Log-likelihood Ratio Test

Log-likelihood Ratio Test is a hypothesis test used in statistics. The test purpose is to help the researcher make the decision to choose the best model by comparing two models. In this paper, log-likelihood ratio test is implemented to find the best detection model for P2P Botnet. The formula for the test is:

$$x^2 = 2\,Log\ Likelihood\big(New\ (with\ predictor)\big) - Log\ Likelihood\big(Baseline(without\ predictor)\big)$$

3.3.2. Nagelkerke's R2 Test

Nagelkerke's R2 test is one of the coefficients of determination (R^2) test on the regression model. Nagelkerke's R^2 test is an adjustment version of Cox and Snell's R^2 that adjust the value scale ranging from 0 to 1. The test is used to find the best fit model for P2P Botnet Detection. The formula for the test is:

$$R^2 = \frac{1 - \left\{\dfrac{L(M_{Intercept})}{L(M_{Full})}\right\}^{2/N}}{1 - L(M_{Intercept})^{2/N}}$$

52 features were tested with wrapper feature selection from all P2P botnets variance represented by Botnet's attribute class named Mybotnet. Only six features were selected as shown in Table 2. The detailed description of the features produced by feature selection module is described as in Table 2. This selected feature would go through validation for evaluation performance. The next subsection discussed the process involved in the feature influence module.

Table 2. Selected Feature Description

Feature Name	**Description**
pushed data pkts_a2b	The count of all the packets seen with the PUSH bit set in the TCP header.
pushed data pkts_b2a	The count of all the packets seen with the PUSH bit set in the TCP header.
Max_Win_ Adv_a2b	The largest window advertisement that was sent from the destination to the source
Max_Win_ Adv_b2a	The largest window advertisement that was sent from the destination to the source
pure acks sent	The total number of ack packets sent between the hosts that were not piggy-backed with data (just the TCP header and no TCP data payload) and did not have any of the SYN/FIN/RST set.
throughput	The average throughput calculated as the unique bytes sent divided by the elapsed time i.e., the value reported in the unique bytes sent field divided by the elapsed time (the time difference between the capture of the first and last packets in the direction).

3.4. Performance Evaluation

The outcome of the analysis of the performance of method will be done upon the analysis of performance evaluation practices. Empirical observation of the efficiency on IDS is done by utilizing the confusion matrix [26]. True Positive (TP) represents attack examples correctly classified as 1(attack); True Negative (TN) represents non-malicious examples correctly classified as 0 (normal); False Positive (FP), which represents non-malicious examples misclassified as 1 and False Negative (FN) represents attack examples misclassified as 0. Using the parameters of the confusion matrix the performance measures are stated as follows:

- Detection Rate (DR), which is the probability of malicious examples correctly classified as malicious among all the malicious examples. DR = TP/ (FN+TP)
- False Positive Rate (FPR), which is the probability of non-malicious examples misclassified as malicious among all the frames. FPR = FP/ (FP+TN)
- False Negative Rate (FNR), which is the probability of malicious examples misclassified as normal among all the malicious frames. FNR = FN/(FN+TP)
- Overall Success Rate (OSR), or Accuracy, which is the probability of any examples correctly classified. OSR= (TN+TP)/(TP+FP+TN+FN)

4. Result and Discussion

There are six features selected after went through feature selection process which are: *pushed data pkts_a2b, pushed data pkts_b2a, Max_Win_Adv_a2b, Max_Win_Adv_b2a, Pure_acks_sent_b2a, throughput*. All these features are tested and analyzed using binomial logistic regression in SPSS Statistics. The purpose to run those features using statistical approach is to explore an influence features and might be useful to improve detection system especially P2P botnet attack.

Table 3. Determining Influence Features

		B	Wald	df	Sig.	95% C.I.for EXP(B)	
						Lower	Upper
Step 1[a]	pushed_data_pkts_a2b	-.839	4516.820	1	.000	.422	.443
	Constant	4.077	14159.092	1	.000		
Step 2[b]	pushed_data_pkts_a2b	-1.216	4318.137	1	.000	.286	.307
	pushed_data_pkts_b2a	.794	936.013	1	.000	2.103	2.329
	Constant	4.339	12828.634	1	.000		
Step 3[c]	**pure_acks_sent_b2a**	.931	597.473	1	.000	2.356	2.735
	pushed_data_pkts_a2b	-1.335	3724.092	1	.000	.252	.275
	pushed_data_pkts_b2a	.269	101.459	1	.000	1.242	1.379
	Constant	3.763	5771.159	1	.000		

a. Variable(s) entered on step 1: pushed_data_pkts_a2b.
b. Variable(s) entered on step 2: pushed_data_pkts_b2a.
c. Variable(s) entered on step 3: pure_acks_sent_b2a.

Table variable in the equation found that the insignificant features are not present in the table because it is automatically excluded in the equation as we apply improvements method (Forward LR) which only present the significant features. As the quality of the logistic regression improves. Only the significant features are included in the logistic regression equation. The selected feature is calculated using regression function *(3.763 + 0.931*x1 – 1.335*x2 + .269*x3),* which determine the significant features according to coefficient value. The positive coefficient value implies the possibilities of what is happening are more than the baseline and negative coefficient value implies the possibilities of what is happening is lower than the baseline. In Table 3, there are only two influence features that were chosen for the next process as *pushed_data_pkts_a2b* has a negative coefficient value and it is considered as an insignificant feature. The two influence features are *pushed_data_pkts_b2a* and *pure_act_pkts_a2b*.

4.1. Exploring Feature Influence for Feature Pushed_data_pkts_b2a.

Table 4 shows the value of the likelihood ratio statistic after the feature is included in the model. If only the constant was included, -2LL = 11961.505 and this value has been reduced to 9758.518. This reduction means that the features have a significant influenced at predicting the outcome (botnet).

Table 4 also shows the Nagelkerke's value for the new model which is 0.525. The result shows that was closed to one which means that the feature selected gave a good influence to the model in predicting the outcome.

Moreover, chi-square (x2) test also can be used to verify whether the feature gives a significant contribution to the model. Table 4 shows the value of chi-square is 2202.986 and the p value is highly significant at 0.05 and 0.001 levels. Thus, Pushed_data_pkts_b2a gives a good effect to the model in predicting the outcome. The summary of the influence of *Pushed_data_pkts_b2a* is shown in Table 4.

Table 4. Model Summary of the Influence of Feature for *Pushed_data_pkts_b2a*

Feature	-2 Log Likelihood	Nagelkerke R^2	Chi-square	Df	Sig.	Wald
Pushed_data_pkts_b2a	9758.518	0.525	2202.986	1	0.000	936.013

4.2. Exploring Feature Influence for Feature pure_act_pkts_a2b.

The result from analysis shows that *pure_act_pkts_a2b* gave a significant influence on the model in predicting botnet detection. Table 5 shows the value of the likelihood ratio statistic after the feature is included in the model. If only the constant was included, -2LL = 9758.518 and this value has been reduced to 8201.782. This reduction means that the features have a significant influenced at predicting the outcome (botnet).

Table 5 also shows the Nagelkerke's value for the new model which is 0.608. The result shows that was closed to one which means that the feature selected gave a good influence to the model in predicting the outcome.

Moreover, chi-square (x^2) test also can be used to verify whether the feature gives a significant contribution to the model. Table 5 shows the value of chi-square is 1556.736 and the p value is highly significant at 0.05 and 0.001 levels. Thus, *pure_act_pkts_a2b* gives a good effect to the model in predicting the outcome. The summary of the influence feature *pure_act_pkts_a2b* is shown in Table 5.

Table 5. Model Summary of the Influence of Feature for *pure_act_pkts_a2b*

Feature	-2 Log Likelihood	Nagelkerke R^2	Chi-square	Df	Sig.	Wald
pure_act_pkts_a2b	8201.782	0.608	1556.736	1	0.000	597.473

4.3. Assessing the Model based on Classification Table for Overall Feature

The classification table can be used to assess the fit of the model. The fit of the model can be assessed by using the classification table. Table 6 and Table 7 show the result of the classification table of the null model and full model.

Table 6. Classification of Null Model

Observed		Predicted	
		Class	
		Normal	Botnet
Class	Normal	0	2326
	Botnet	0	46165

From the Table 6, we can conclude:
Detection Attack Rate = 100%
False Positive (FP) = 47.9%
Detection Normal Rate = 0%
False Negative (FN) = 0%
Overall Detection Rate = 95.2%

Table 6 indicates the detection for normal traffic labeled as Normal while for abnormal traffic is labeled as Botnet. From this table, the classification of Null Model shows, the detection normal rate of the model is 0% correct in classifying the normal while the detection of attack rate is 100%. False positive rate indicates 47.9% and false negative is 0%. Overall accuracy for Null classification is 95.2% which is high enough in detecting intrusion or P2P botnet in network traffic. The highest percentage of the detection attack rate resulted in 100% meaning that the network tested already in full risk for an attacker activities. This is because this set of tables describes the baseline model which is a model that does not include our explanatory variables, the predictions of this baseline model are made purely on whichever category occurred most often in our dataset. In this case, the model always guesses ‘yes’ because it shows 100% detection of a botnet (46165 compared to 2326). The overall percentage row tells that this approach to prediction is correct 95.2% of botnet detection.

Table 7. Classification of Full Model

Observed		Predicted	
		Class	
		Normal	Botnet
Class	Normal	1637	689
	Botnet	117	46048

From the Table 7, we can conclude:
Detection Attack Rate = 99.74%
False Positive (FP) = 1.47%
Detection Normal Rate = 70.4%
False Negative (FN) = 6.67%
Overall Detection Rate = 98.3%

Based on Table 7, the classification of the full model means that the predictor was included inside the model and will generate a different result with Null Model especially the improvement in overall accuracy of the detection. Detection attack rate produces still in high rate, 99.74% correct in classifying the attack and false positive indicate 1.47%. The false negative was increased to 6.67% from the full model. Overall Detection Rate was only 98.3% but it is an acceptable value according to authors [27] which stated that the capabilities of current botnet detection system is enough at 80% for botnet detection and may generate a better prediction, thus had capabilities to distinguish the classification between normal and abnormal traffic.

Referring to the overall percentage, Null Model generates an Overall accuracy at 95.2%, while Full Model which applied logistic regression model to the data and shows an increment up to 98.3%. Thus, the improvement of Overall accuracy indicated that the model is fitted and may significantly contribute to the intrusion detection system.

5. Conclusion and Future Work

In conclusion, this paper has revealed several influenced features for P2P Botnet Detection. It was found that the features model fitting produced the best fit for the data. The real traffic test proving that P2P botnet detection is determined by *pushed_data_pkts_b2a* and *pure_act_pkts_a2b*. The analysis result shows the feature generated giving a good contribution in P2P botnet detection with higher detection rate, thus fulfilled the objectives of the research. Based on the accuracy of detection produce in this research, a further study is needed in order to identify the appropriate threshold value. Good evaluation on detection rate does not mean it can be used in any techniques. It still needs an improvement in developing better technique to identify threshold value for P2P Botnet detection itself to increase the detection rate and may possibly apply in a different technique. Since this study only focusing on TCP protocol, the intention to apply other protocol such UDP is also recommended. Besides that, the future research also aims to look at IDS log, to determine suitable threshold in the traffic which may contribute to the detection accuracy of IDS in order to differentiate normal and abnormal activities in the network.

6. Acknowledgement

This work has been supported under Universiti Teknikal Malaysia Melaka research grant GLUAR/CSM/2016/FTMK-CACT/I00013 and Ministry of Higher Education Malaysia (HLP). The authors would like to thank INSFORNET Research Group for their incredible support by providing facilities to conduct the study.

References

[1] A. Karim, R. Bin Salleh, M. Shiraz, S. A. A. Shah, I. Awan, and N. B. Anuar, “Botnet detection techniques: review, future trends, and issues,” *J. Zhejiang Univ. Sci. C*, vol. 15, no. 11, pp. 943–983, 2014.

[2] D. Plohmann, E. Gerhards-Padilla, and F. Leder, “Botnets: Detection, Measurement, Disinfection & Defence,” *Inf. Secur.*, p. 153, 2011.

[3] M. C. E. R. T. MYCERT, “Incidents Report of General Incident Classification Statistic 2015,” 2015. [Online]. Available: https://www.mycert.org.my/assets/graph/pdf/2015-1.pdf. [Accessed: 15-Jan-2016].

[4] S. A. M. Yusof, “Serangan siber ancam negara,” *Kosmo Online News*, 2015. [Online].Available:http://www.kosmo.com.my/kosm

o/content.asp?y=2015&dt=1119&pub=Kosmo&sec=Negara&pg=ne_10.htm. [Accessed: 20-Nov-2015].

[5] Q. Han, W. Yu, Y. Zhang, and Z. Zhao, "Modeling and evaluating of typical advanced peer-to-peer botnet," *Perform. Eval.*, vol. 72, pp. 1–15, 2014.

[6] M. J. Elhalabi, S. Manickam, L. B. Melhim, M. Anbar, and H. Alhalabi, "A review of peer-to-peer botnet detection techniques," *J. Comput. Sci.*, vol. 10, no. 1, pp. 169–177, 2013.

[7] A. Neville and R. Gibb, "ZeroAccess Indepth," 2013.

[8] Amrita and P. Ahmed, "A Study of Feature Selection Methods in Intrusion Detection System : a Survey," *Int. J. Comput. Sci. Eng. Inf. Technol. Res.*, vol. 2, no. 3, pp. 1–25, 2012.

[9] NetFort Technologies Limited, "Flow Analysis Versus Packet Analysis . What Should You Choose ?," 2014. Retrieved from https://www.netfort.com/wp-content/uploads/PDF/WhitePapers/NetFlow-Vs-Packet-Analysis-What-Should-You-Choose.pdf [Accessed on March 8, 2017].

[10] D. Zhao, I. Traore, A. Ghorbani, B. Sayed, S. Saad, and W. Lu, "Peer to peer botnet detection based on flow intervals," *IFIP Adv. Inf. Commun. Technol.*, vol. 376 AICT, no. 1, pp. 87–102, 2012.

[11] V. Bolon-Canedo, N. Sanchez-Marono, and A. Alonso-Betanzos, "On the behavior of feature selection methods dealing with noise and relevance over synthetic scenarios," *Proc. Int. Jt. Conf. Neural Networks*, pp. 1530–1537, 2011.

[12] J. Tang, S. Alelyani, and H. Liu, "Feature Selection for Classification: A Review," *Data Classif. Algorithms Appl.*, pp. 37–64, 2014.

[13] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1–2, pp. 273–324, 1997.

[14] V. Kumar, "Feature Selection: A literature Review," *Smart Comput. Rev.*, vol. 4, no. 3, 2014.

[15] K. Yang and J. Trewn, *Multivariate statistical methods in quality management*. McGraw Hill Professional, 2004.

[16] R. L. Ott and M. T. Longnecker, *An introduction to statistical methods and data analysis*. Cengage Learning, 2008.

[17] V. Ghanaei, C. S. Iliopoulos, and R. E. Overill, "A Statistical Approach for Discovering Critical Malicious Patterns in Malware Families," Seventh Int. Conf. Pervasive Patterns Appl. 2015", IARIA, no. c, pp. 21–26, 2015.

[18] M. S. Mok, S. Y. Sohn, and Y. H. Ju, "Random effects logistic regression model for anomaly detection," Expert Syst. Appl., vol. 37, no. 10, pp. 7162–7166, 2010.

[19] K. Hughes and Y. Qu, "A Theoretical Model: Using Logistic Regression for Malware Signature based Detection," in The 10th International Conference on Dependable, Autonomic, and Secure Computing (DASC-2012), 2012.

[20] D. W. Hosmer and S. Lemeshow, "Applied logistic regression", vol. 2nd, no. 1. 2000.

[21] A. K. I. El-Koka, "Regularization Parameter Optimization in Logistic Regression Analysis and Support Vector Machines," Dongseo University, 2013.

[22] V. Nivargi, M. Bhaowal, and T. Lee, "Machine Learning Based Botnet Detection," CS 229 Final Proj. Rep., 2006.

[23] M. A. et al. Faizal, "Statistical Approach for Validating Static Threshold in Fast Attack Detection," vol. 4, no. 1, pp. 53–72,2010.

[24] S. et al. Faizal, M.A. Shahrin, "FEATURE SELECTION FOR DETECTING FAST ATTACK IN IDS," vol. 2, no. 2, pp. 39–56, 2008.

[25] N. H. Son, "Data cleaning and Data preprocessing." 2006.

[26] D. Jurafsky and J. H. Martin, "Speech and Language Processing," 2016. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/. [Accessed: 20-Apr-2017].

[27] M. Eslahi, "Bots and Botnets : An Overview of Characteristics, Detection and Challenges," 2012.

[28] Nur Hidayah Mohd Saudi et al. "*Revealing the* Feature Influence in HTTP Botnet Detection" International Journal of Communication Networks and Information Security (IJCNIS), Vol. 9, No. 2, 274-281, August 2017

[29] Aun Yichiet, Selvakumar Manickam, Shankar Karuppayah," A Review on Features' Robustness in High Diversity Mobile Traffic Classifications" International Journal of Communication Networks and Information Security (IJCNIS), Vol 9, No 2,294-304,August 2017

[30] Rastegari, S., Lam, C.P. and Hingston, P., "A Statistical Rule Learning Approach to Network Intrusion Detection," In IT Convergence and Security (ICITCS), 2015 5th International Conference on Kuala Lumpur, Malaysia, pp.1-5, IEEE, August, 2015.

[31] Kritika Govind, S. Selvakumar,"Auto-Pattern Programmable Kernel Filter (Auto-PPKF) for Suppression of Bot Generated Traffic", IJCNIS, vol.6, no.1, pp.48-54,2014.