ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 5, No.17 , September 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Detecting Infected Botnet Machines by Using the Traffic Behavior Analysis

Fahimeh Hasani[1], Ebrahim Mahdipour[2]

[1]Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
f.hasani77@gmail.com

[2]Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
e.mahdipour@gmail.com

Abstract

Despite the increase in attacks and other security challenges in cyberspace, we require new methods of detection and to develop new techniques for the new generations of attacks. One of these new threats are botnets. This article presents the means for identifying infected machines with botnets by using a behavioral analysis method. Work with botnets as a tool intended to carry out criminal activities has increased with large area in computer networks against large targets. The pattern of behavior By frequent studying on the nods and the visualization of traffic with FroceAtlas2 and Page Rank algorithms have been presented by analyzing the data traffic, as a result, the nodes that have the most interaction structure on bot in the network, have been identified as the machines infected with botnets.

Keywords: *Botnets, traffic analysis, network traffic visualization, infected machines, data visualization*

1. Introduction

Using of botnets has increased as a tool for criminal activities with a large area in computer networks against large targets. A botnet is a distributed environment that the victim`s system via Trojans and emails that contains malicious codes at the end of them and are used for targeting attacks by using a control center bot. We explore the interaction of botnets by using a behavioral analysis of network traffic such as, Internet Protocol addresses, source and destination, source and destination ports, protocol of transmission and traffic information and we can present a mechanism for detecting and eliminating the bots by using of behavioral patterns in traffic network Many ways of detecting of botnets Have been presented such as detecting botnets from the features of group activity[5], periodic behavior in relation to the control center of the Bot [2], using honey network [6], inactive listening and behavior visualization methods of Bot [9], for example; quick response time, size of small instructions, quick execution of instructions, parallel coordinates that had presented.

The proposed algorithm increases network traffic by approximately 8% with speed of detection accuracy and optimizing graph for visual diagnostic from overview. By default, there are three values for optimizing speed:
Parameter 0.1 under 50000 Node, 1 to 50000 Nodes and 10 more than 50000 Nodes.
Optimization is a description of a work for some nets. Measuring the quality and comparing different algorithms are going to allow us to construct spatialization scheme at each repetition.

2. Background and Target

Infected machines in the network form part of an army which is under the control of the bot's control center. It is very important to have a suitable method for stopping the bots from botnet's army to protect the network against their attacks. This study is going to be explored by two algorithms froceatlas2 and page rank as well as, surveying the behavioral of network by analyzing network and examining network traffic visualization techniques with Graph and Mathematics .The weaker nodes connected to stronger nodes by using repulsive and directed-force and energy model to some common features of the networks which come from the existence of many nets by a degree of nodes gather around a greater node.
The energy line model applies for creating tighter cluster with strong impact on diagram. We create a graph of the network traffic data by the use of algorithms force-directed and FroecAtlas2[8], that cover the analyzing node's link by applying page rank which can calculate the weight of each edge or node for the purpose of controlling the actions. Both aspects must be considered. How much is it going to take for the edge to pass from source to the destination host and how much is the rank of ultimate destination of traffic? Imagine, there is a relationship between bots with the same features within group activities. We cluster the nodes. The greater degree nodes apply in center of pattern. Clustering is that the

high nodes connected are grouped and the low nodes connected, are separated.

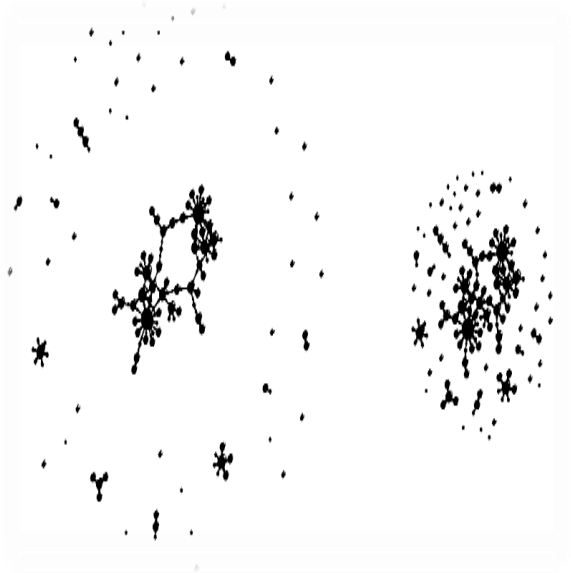

Fig.1 ForceAtlas2

Visualization is a sequence of repeated hypothesis, testing and exploring. This sequence helps the analyst to explore and innovate by his ability and experimentation.
Another visualization algorithms like byfanho algorithm [11], and Frochert man rain gold [10], have been presented also to simulate data. But forceatlas2 shows repetition and better quality at less duration.

2.1 Data visualization

Data visualization [7], is the conversion of information to a form of optical that enables the viewers to observe, review and make a sense and percept the information and also use it as a secure method of analysis
Data visualization has produced varieties of data, such as: image data, large-scale database, information of a video data matter which save Visual knowledge as one of the analysis. Therefore, data visualization is a plan with a specific goal. Until recently, data visualization has focused more on the three-dimensional visualization techniques of networks. Some of these data visualizations consist of, chernoff, coordinates star, multiple notions, and some other techniques. Displaying the star are grouped by numerical division on little multiple in a star plot. Tree mapping is another form of visualization that was first described by Johnson. The idea was that, a hierarchy is shown as a two-dimensional design, they presented relative size of hierarchy and the size on monitor screen by filling the space rectangular which has been nested. Detecting by forceatlas2and page range algorithms rank is a plan of a direct-force which deals with a simulation of physical system. The forces at the edge- absorption of node at the junction of nodes create a convergent movement which comes to a balanced position. It has a distinctive characteristic of forces help to commentary of data.
The algorithm is a resultant of a force and helps visual networks to be read. It has a distinct characteristic of an energy model. The way of optimization of speed is a network. It has a repulsive force that depends on degree. So, it establishes protection of optimal balance between speed and accuracy to node and it gives rapid convergence which occurs between network standards.
Force-direct comes to an approximate result of algorithm by setting a function of node with other nodes and the force between nodes.

2.2 Energy model

Each direct-force algorithm is based on a special formula for the gravitational force and simulates with a real inspiration [1], the formula of repulsion and absorption is also used for the repulsion of electrically charged particles. Graph spinalization algorithm has a major role in determining distance. The directing algorithm of a physical system works with gravity and repulsion force, this means that, there are forces with enough distance between coordinated subjects or nodes. The closer nodes have less adsorption and more desorption than the future nodes or vice versa. The distance between the forces can be linear, exponential or logarithmic.
The idea is that, the repulsion degree leads to the connecting of poor nodes to nodes that are well attached. The force Fr appropriates with product (degree+ one) from two nodes. Factor KR has defined by the following structure.

$$F_r(n_1, n_2) = K_r \frac{(\deg(n_1)+1)(\deg(n_2)+1)}{d(n_1, n_2)} \quad (1)$$

Modified nodes do not overlap with repulsion nodes. The idea is that, both forces repulsion and absorption used for calculating the size of nodes- in size of (n) node into distance calculation $d(n_1, n_2)$.
$\acute{d}(n_1, n_2) = d(n_1, n_2) - \text{size}(n_1) - size(n_2)$Is the distance of prevention the board to board of overlap.
If, $\acute{d}(n_1, n_2) < 0$ to be without overlap thus, instead of d from'd we have the following formula for calculating the force: $Fa(n_1, n_2) = \acute{d}(n_1, n_2)$

$$F_r(n_1, n_2) = K_r \frac{(\deg(n_1)+1)(\deg(n_2)+1)}{d(n_1, n_2)}$$

If, $\acute{d}(n_1, n_2) > 0$ contain overlap, we will have repulsion but without absorption.$Fa(n_1, n_2) = 0$

$$F_r(n_1, n_2) = K_r(\deg(n_1)+1)(\deg(n_2)+1)$$

If, we use the formula $\acute{d}(n_1, n_2) = 0$, we have neither absorption nor repulsion.

3. Page ranking algorithm

The Page Rank algorithm [4], is an analysis link node, both aspects must be considered, in observing the issue, time of the edge that goes from the source to the hostage destination and how much is the ration of final destination traffic. Suppose that the communication between bots have similar property in activities. We cluster the nodes in process of clustering. The nodes which are greater sets in center of scheme. Clustering is that fully connected nodes are grouped while less connected nodes are separated. The rank algorithm can be used as a set of matrix calculations. The scores of rank of link increased node evaluated by weight for exploring botnets. See the following equation:

$$P_t(i) = (1-d)\sum_{k=1}^{n} W(k) + d \sum_{(j,i)\in E} \frac{Pt-1(j)}{Oj} \quad (2)$$

At this study, we survey and explore the traffic data by presented algorithm. Aimed at developing effective visual traffic of the network which is then studied to detect suspicious patterns that might be infected by a botnet. We will study the features of botnets and their life duration. At the end, we survey a video surveillance system to detect the permeation of the botnet with visualization techniques.

3.1 How to analyze the data

Silk (System for Internet-Level Knowledge) [3], tool is an order- line contains a set of order- line tools that are used for analyzing and tracking cyber-attacks. Which has been applied for efficient-set, saving and analyzing current data. Silk tool creates a record, called flow record, in a system in case that records events. The tools of the set of silk's software start analyzing network problems after analyzing the records. The most important tool of this set is rwfilter. We store scans of network data which have been analyzed by silk, then it divides the data into smaller sets. The standard of choosing contains date, time, Internet Protocol address of origin and destination, packet length and the type of protocol, so we separate each list of divided data by clustering afterwards that we scheme their graph and study it by forceatlas2 algorithm.

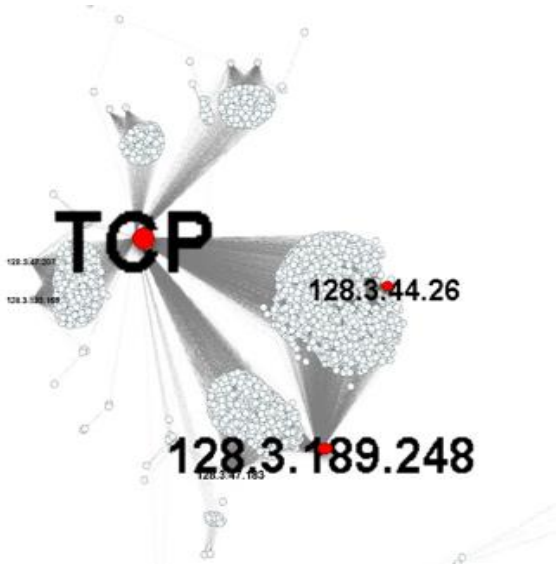

Fig.2 graph Force Atlas2 algorithm

The address of explored ip with the most degree. Ip addresses which have the most interaction with other addresses display in the graph force atlas 2 algorithm.

4. Proposed graph algorithm

We plot clustered data by traffic visualization software so we can determine the average grade, average degree weight, network thickness, density graph, power of play, eccentricity, modular class by repeating algorithm performance.

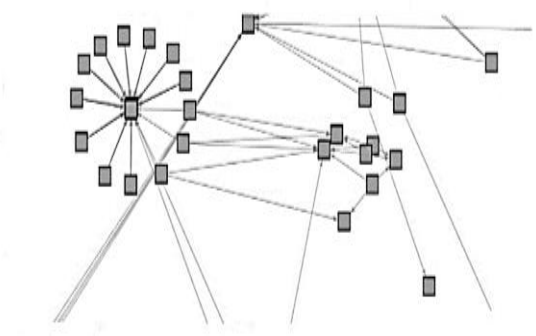

Fig.3 repeated algorithm graph

In this figure, finally, the node which has the most interaction with other nodes is situated in the center of sample by repetition of algorithm and the others are located around of it.

In the created graph, by default, we consider the value of rank at a minimum of 5 and a maximum of 31 to identify rank of nodes in ranking according to entrance degree. We can also distinct graph nodes by coloring the weight of edge and disorder ranked nodes for visualization and visual graph nodes. According to the algorithm applied to the graph, we find the node which has the most weight and highest degree in the table.

Data Table

Nodes | Edges | Configuration | Add node | Add edge | Search/Replace | Import Spreadsheet | Export table | More a

Nodes	Id	Label	In-Degree	Degree	Weighte...	Weighted In...	PageRank	Authority
128.3.44.26	128.3.44.26	128.3.44.26	2309	2309	2,309	2,309	0.087	0.189
128.3.189.248	128.3.189.248	128.3.189.248	3358	3358	3,358	3,358	0.126	0.276
128.3.44.112	128.3.44.112	128.3.44.112	44	44	44	44	0.002	0.004
128.3.47.183	128.3.47.183	128.3.47.183	1057	1057	1,057	1,057	0.04	0.087
128.3.100.81	128.3.100.81	128.3.100.81	172	172	172	172	0.007	0.014
128.3.46.252	128.3.46.252	128.3.46.252	2	2	2	2	0	0
128.3.95.149	128.3.95.149	128.3.95.149	2	2	2	2	0	0
128.3.45.164	128.3.45.164	128.3.45.164	274	274	274	274	0.01	0.023
128.3.97.204	128.3.97.204	128.3.97.204	12	12	12	12	0.001	0.001
128.3.45.225	128.3.45.225	128.3.45.225	14	14	14	14	0.001	0.001
128.3.47.255	128.3.47.255	128.3.47.255	1	1	1	1	0	0
128.3.45.10	128.3.45.10	128.3.45.10	62	62	62	62	0.002	0.005
128.3.47.207	128.3.47.207	128.3.47.207	36	36	36	36	0.001	0.003
128.3.212.208	128.3.212.208	128.3.212.208	14	14	14	14	0.001	0.001
128.3.44.94	128.3.44.94	128.3.44.94	2	2	2	2	0	0
128.3.100.204	128.3.100.204	128.3.100.204	18	18	18	18	0.001	0.002
128.3.47.161	128.3.47.161	128.3.47.161	7	7	7	7	0	0.001
128.3.164.194	128.3.164.194	128.3.164.194	17	17	17	17	0.001	0.001
128.3.47.46	128.3.47.46	128.3.47.46	8	8	8	8	0	0.001
128.3.193.169	128.3.193.169	128.3.193.169	7	7	7	7	0	0.001

Fig.4 table of algorithm data

We will explore a series of internet address protocols by studying the overall data and visual inspection, according to algorithm action, those which have the highest page ranking among other nodes. The ideal Internet Protocol addresses have more power to distribute -as well as weight and input degree than other nodes of graph at set of data. Therefore, the address is not reliable. We are going to penetrate to structure of a bot by studying an address, for instance, according to the figure2, the load of protocol traffic Tcp as well as the load of node traffic at 128.3.189.248 is high. Hence, we can penetrate to suspicious node. It has the most input and weight. Among the cars feature -which are under attack of bot is broad band and development of traffic. All the nodes that are connected to the ideal node are considered as infected machines.
Some utilized data and samples of suspicious addresses to bot considered at following table.

Table1: sample of explored bot

Number of data for one week 807589			
Number of data by the most weight 98546			
128.3.189.248	128.3.26.57	128.3.47.204	128.3.45.30
128.3.45.164	128.3.212.208	128.3.44.112	163.27.195.211
218.131.115.53	118.139.50.187	128.3.44.210	128.3.100.204

Table1, shows that the number of collected data during one week is 807589 and after applying the algorithm, 98546 data from that number of data have the highest weight and rank. Some samples of suspicious addresses to bot (bot infected machines) from data of the algorithm have been presented in table.

4.1 Architectural analysis of the proposed algorithm

The structure of algorithm for detecting bot infected machines.

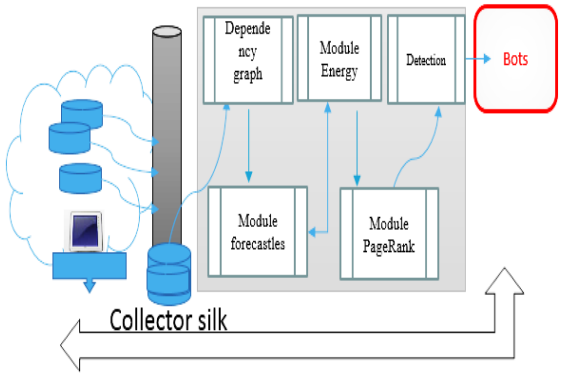

Fig. 5 architecture of proposed algorithm

Architecture algorithm is that: first we collect the data by silk and drawing the graph, then we apply force atlas2 algorithm on ideal graph and at last, we can discover nodes with the most weigh and rank that are the machines infected with bots by model of energy and Page Rank

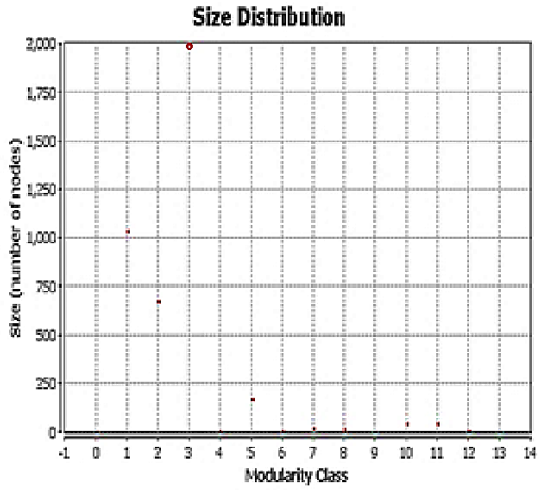

Fig.6 modular size

Dispersed scheme visualization is another method to consider bot behavior. Each point in diagram represents the modular class size based on the size of node. The distance is represented by color. Size, ranking and layer size of the data determine whether a node acts in a network as a bot. The size of network traffic distribution displays the load of traffic in the network, the load of bar observes by using the size of each node in modular class. The size of a node is greater so the weight is heavier. Figure 6 illustrates that the height of each point shows the size of nodes in modular class within one week and the size of nod in modular class has the most height from one to three so, it shows that there is something wrong with the activity of the modular class.

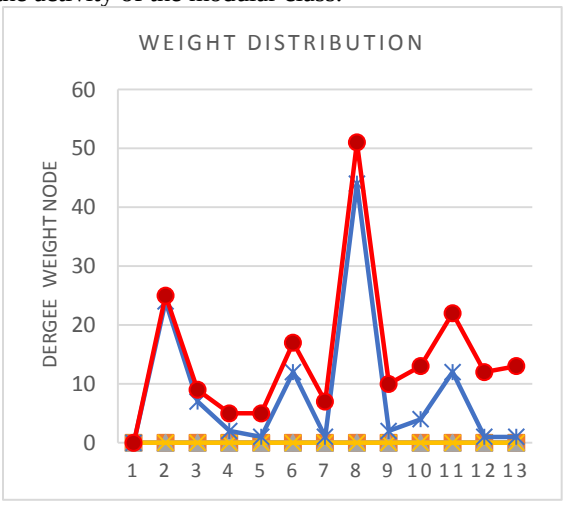

Fig.7 degree of nod's weight in graph

In figure 7, the weight of nods in graph is considered by calculating rank of nod after acting of directed force algorithm and the greatest weight represents suspicious activity. Presented algorithm almost increases rapid detection accuracy and optimized graph for better optical detection from general network traffic than other available techniques for visualization of a bot's behavior [9], such as, rapid response time, measure of small instructions, instant performing instructions, parallel coordinate which has been expressed.

5. Conclusions

This study reviewed the problem of detecting infected machines into botnets by using the analysis of traffic behavior. The nodes that have the most structural interaction such as a bot in a network were explored by analyzing, data traffic, recording the graphs, data analytic diagramming, frequent study of nodes and then took a specific pattern of them that was considered by comparing the behavior of the bots in the network and the node according to the pattern of the most weight and ranking.

References

[1] A. Noack, “Energy models for graph clustering” J. Graph Algorithms Appl., vol. 11, no. 2, pp. 453–480, 2007.

[2] B. AsSadhan, J. M. F. Moura and D. Lapsley, "Periodic behavior in botnet command and control channels traffic", in Proceedings of the 28th IEEE conference on Global telecommunications, Honolulu, Hawaii, USA, pp:2157-2162, 2009.

[3] Ch. Sanders،J.Smith،dJ.Bianco," Applied Network Security Collection, Detection, and Analysis Monitoring" CHAPTER4، 2014.

[4] J. Francois, S. Wang, R. State, T. Eng، "BotTrack: Tracking Botnets using Net Flow and PageRank" Submitted on 5 Aug 2011.61.

[5] K. Wang, C-Y. Huang, Shang-Jyh Lin, Ying-Dar Lin (2011). A fuzzy attern-based filtering algorithm for botnet detection. Computer Networks, 2011 Elsevier.

[6] KnowyourEnemy:TrackingBotnets,http://www.honeynet.org/Papers/bots/, 2005.

[7] M. Bastian, S. Heymann, and M. Jacomy, “Gephi: an open source software for exploring and manipulating networks,” in International AAAI Conference on Weblogs and Social Media. Association for the Advancement of Artificial Intelligence, 2009.

[8] M. M. Masud, T. Al-khateeb, L. Khan, B. Thuraisingham, K. W.Hamlen, “ Flow-based identification of botnet traffic by mining” multiple log file,” Proc. Int. Conf. Distributed Frameworks &Applications (DFmA 2008), 2008, pp.200-206.

[9] M.M. Jacomy, M. Heymann، M. Bastian “ForceAtlas2, A Continuous raph Layout Algorithm for Handy Network Visualization”, August 1, 2012.

[10] T. M. J. Fruchterman and E. M. Reingold, “Graph drawing by force-directed Placement,” Softw: Pract. Exper, vol. 21, no. 11, pp. 1129–1104, Nov.1991.

[11] Y. F. Hu, “Efficient and high quality force-directed graph drawing,” The Mathematica Journal, vol. 1۰, pp. 32–21, 2005.