Journal of Computing and Information Technology - CIT 20, 2012, 1, 33–39
doi:10.2498/cit.1001840

Predicting Number of Zombies in DDoS Attacks Using Pace Regression Model

B. B. Gupta

[1]University of New Brunswick, Canada
[2]Department of Electronics and Computer Engineering, Indian Institute of Technology Roorkee, India

A DDoS attacker attempts to disrupt a target, by flooding it with illegitimate packets which are generated from a large number of zombies, usurping its bandwidth and overtaxing it to prevent legitimate inquiries from getting through. This paper reports the evaluation results of proposed approach that is used to predict number of zombies using Pace Regression Model. A relationship is established between number of zombies and observed deviation in sample entropy. Various statistical performance measures, such as R^2, *CC*, *SSE*, *MSE*, *RMSE*, *NMSE*, η, *MAE* are used to measure the performance of the regression model. Network topologies similar to Internet used for simulation are generated using Transit-Stub model of GT-ITM topology generator. NS-2 network simulator on Linux platform is used as simulation test bed for launching DDoS attacks with varied number of zombies. The simulation results are promising as we are able to predict number of zombies efficiently using Pace Regression Model with considerably less error rate.

Keywords: DDoS attack, intrusion detection, pace regression, zombies, entropy

1. Introduction

The Internet has become a popular medium of commercial activity and this has raised the risks, both, for attackers and security personnel. DDoS attacks compromise availability of the information system through various means [1,2]. One of the major challenges in defending against DDoS attacks is to accurately detect their occurrences in the first place. Anomaly based DDoS detection systems construct profile of the traffic normally seen in the network, and identify anomalies whenever traffic deviate from normal profile beyond a threshold [3]. This extension of deviation is normally not utilized. We use Pace regression [4] based approach that utilize this extension of deviation from detection threshold, to predict number of zombies. A real time estimation of the number of zombies in DDoS scenario is helpful to suppress the effect of attack by choosing predicted number of most suspicious attack sources for either filtering or rate limiting. We have assumed that zombies have not spoof header information of out going packets. Moore et. al [5] have already made a similar kind of attempt, in which they used backscatter analysis to estimate number of spoofed addresses involved in DDoS attack. This is an offline analysis based on unsolicited responses.

Our objective is to find the relationship between number of zombies involved in a flooding DDoS attack and deviation in sample entropy. In order to predict number of zombies, Pace Regression Model is used. To measure the performance of the proposed approach, we have calculated various statistical performance measures. Internet type topologies used for simulation are generated using Transit-Stub model of GT-ITM topology generator [6]. NS-2 network simulator [7] on Linux platform is used as simulation test bed for launching DDoS attacks with varied number of zombies. In our simulation experiments, attack traffic rate is fixed to 25 Mbps in total; therefore, the mean attack rate per zombie is varied from 0.25 Mbps to 2.5 Mbps and total zombie machines range between 10 and 100 to generate attack traffic.

The remainder of the paper is organized as follows. Section 2 contains overview of Pace

Regression Model. Section 3 presents various statistical performance measures. Detection scheme is described in Section 4. Section 5 describes experimental setup and performance analysis in details. Model development is presented in Section 6. Section 7 contains simulation results and discussion. Finally, Section 8 concludes the paper.

2. Pace Regression Model

In its simplest form, regression analysis [8,9] involves finding the best straight line relationship to explain how the variation in an outcome variable, Y depends on the variation in a predictor variable, X. Hence, regression analysis is a statistical tool for the investigation of relationships between variables. Variables which are used to explain, other variables are called explanatory variables. Variable which is explained is called response variable. A response variable is also called a dependent variable, and an explanatory variable is sometime called an independent variable, or a predictor, or regressor. When there is only one explanatory variable, the regression model is called a simple regression, whereas if there are more than one explanatory variable, the regression model is called multiple regression.

Pace regression [4] is a form of linear regression analysis that has shown to outperform other types of linear model-fitting methods, particularly when the number of features is large and some of them are mutually dependent. Pace regression includes a form of feature selection, therefore not all features are actually used in the resulting models.

Input and Output: In Pace Regression Model, a relationship is developed between number of zombies Y (output) and observed deviation in sample entropy X as input. Here X is equal to $(H_c - H_n)$. Our proposed regression based approach utilizes this deviation in sample entropy X to predict number of zombies.

3. Statistical Performance Measures

The different statistical parameters are adjusted during calibration to get the best statistical agreement between observed and simulated variables. For this purpose, various performance measures, such as Coefficient of Determination (R^2), Coefficient of Correlation (CC), Standard Error of Estimate (SSE), Mean Square Error (MSE), Root Mean Square Error ($RMSE$), Normalized Mean Square Error ($NMSE$), Nash–Sutcliffe Efficiency Index (η) and Mean Absolute Error (MAE) are used to measure the performance of the proposed regression model. These measures are defined below.

i) Coefficient of Determination (R^2):
Coefficient of determination (R^2) is a descriptive measure of the strength of the regression relationship, a measure how well the regression line fit to the data. R^2 is the proportion of variance in dependent variable which can be predicted from independent variable.

$$R^2 = \frac{\left(\sum_{i=1}^{N}(Y_o - \bar{Y}_o)(Y_c - \bar{Y}_c)\right)^2}{\left[\sum_{i=1}^{N}(Y_o - \bar{Y}_o)^2 \cdot \sum_{i=1}^{N}(Y_c - \bar{Y}_c)^2\right]} \quad (1)$$

ii) Coefficient of Correlation (CC):
The Coefficient of Correlation (CC) can be defined as:

$$CC = \frac{\sum_{i=1}^{N}(Y_o - \bar{Y}_o)(Y_c - \bar{Y}_c)}{\left[\sum_{i=1}^{N}(Y_o - \bar{Y}_o)^2 \cdot \sum_{i=1}^{N}(Y_c - \bar{Y}_c)^2\right]^{1/2}} \quad (2)$$

iii) Sum of Squared Errors (SSE):
The Sum of Squared Errors (SSE) can be defined as:

$$SSE = \sum_{i=1}^{N}(Y_o - Y_c)^2 \quad (3)$$

iv) Mean Square Error (MSE):
The Mean Square Error (MSE) between observed and computed outputs can be defined as:

$$MSE = \frac{\sum_{i=1}^{N}(Y_c - Y_o)^2}{N} \quad (4)$$

v) Root Mean Square Error ($RMSE$):
The Root Mean Square Error ($RMSE$) between observed and computed outputs can be defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Y_c - Y_o)^2}{N}} \quad (5)$$

vi) Normalized Mean Square Error ($NMSE$):
The Normalized Mean Square Error ($NMSE$) between observed and computed outputs can be defined as:

$$NMSE = \frac{\frac{1}{N}\sum_{i=1}^{N} (Y_c - Y_o)^2}{\sigma_{obs}^2} \quad (6)$$

vii) Nash–Sutcliffe Efficiency Index (η):
The Nash–Sutcliffe Efficiency Index (η) can be defined as:

$$\eta = 1 - \frac{\sum_{i=1}^{N} (Y_c - Y_o)^2}{\sum_{i=1}^{N} (Y_o - \bar{Y}_o)^2} \quad (7)$$

viii) Mean Absolute Error (MAE):
Mean Absolute Error (MAE) can be defined as follows:

$$MAE = 1 - \frac{\sum_{i=1}^{N} |Y_c - Y_o|}{\sum_{i=1}^{N} |Y_o - \bar{Y}_o|} \quad (8)$$

where N represents the number of feature vectors prepared, Y_o and Y_c denote the observed and the simulated values of dependent variable respectively, $\bar{Y}_o$ and σ_{obs}^2 are the mean and the standard deviation of the observed dependent variable respectively.

4. Detection of Attacks

Here, we will discuss proposed detection system that is part of access router or can belong to separate unit that interacts with access router to detect attack traffic. Entropy [10] based DDoS scheme is used to construct profile of the traffic normally seen in the network, and identify anomalies whenever traffic goes out of profile. A metric that captures the degree of dispersal or concentration of a distribution is sample entropy. Sample entropy $H(X)$ is

$$H(X) = -\sum_{i=1}^{N} p_i \log_2(p_i) \quad (9)$$

where p_i is n_i/S. Here n_i represent total number of bytes arrivals for a flow i in $\{t-\Delta, t\}$ and $S = \sum_{i=1}^{N} n_i$, $i = 1, 2, \ldots N$. The value of sample entropy lies in the range $0 - \log_2 N$.

To detect the attack, the value of $H_c(X)$ is being calculated in time window Δ continuously; whenever there is appreciable deviation from $H_n(X)$, various types of DDoS attacks are detected. $H_c(X)$, and $H_n(X)$ give entropy value at the time of detection of attack and entropy value for normal profile respectively.

5. Experimental Setup And Performance Analysis

In this section, we evaluate our proposed scheme using simulations. The simulations are carried out using NS2 network simulator [7]. We show that false positives and false negatives (or various error rates) triggered by our scheme are considerably less. This implies that profiles built are reasonably stable and are able to predict number of zombies correctly.

5.1. Simulation Environment

Real-world Internet type topologies generated using Transit-Stub model of GT-ITM topology generator [6] are used to test our proposed scheme, where transit domains are treated as different Internet Service Provider (ISP) networks i.e. Autonomous Systems (AS). For simulations, we use ISP level topology, which contains four transit domains, with each domain containing twelve transit nodes i.e. transit routers. All the four transit domains have two peer links at transit nodes with adjacent transit domains. Remaining ten transit nodes are connected to ten stub domains, one stub domain per transit node. Stub domains are used to connect transit domains with customer domains, as each

stub domain contains a customer domain with ten legitimate client machines. So a total of four hundred legitimate client machines are used to generate background traffic. To generate attack traffic, total zombie machines range between 10 and 100. Transit domain four contains the server machine to be attacked by zombie machines. A short scale simulation topology is shown in Figure 1.

Currently, the majority of the DDoS attacks are flooding, so we will consider detection of a wide range of flooding attacks in this section. The legitimate clients are TCP agents. The attackers are modeled by UDP agents. A UDP connection is used instead of a TCP one because in a practical attack flow, the attacker would normally never follow the basic rules of TCP, i.e. waiting for ACK packets before the next window of outstanding packets can be sent, etc. The attack traffic rate is fixed to 25 Mbps in total; therefore, the mean attack rate per zombie is varied from 0.25 Mbps to 2.5 Mbps. In our experiments, the monitoring time window was set to 200 ms, as the typical domestic Internet RTT is around 100 ms and the average global Internet RTT is 140 ms [11]. Total false positive alarms are minimum with high detection rate using this value of monitoring window. The simulations are repeated and different attack scenarios are compared by varying total number of zombie machines and at fixed attack strengths.

6. Model Development

In order to predict number of zombies $(\hat{Y})$ from deviation $(H_C - H_n)$ in entropy value, simulation experiments are done at the same attack strength 25 Mbps in total and varying number of

Actual Number of Zombies (Y)	Deviation in Entropy (X) $(H_c - H_n)$
10	0.045
15	0.046
20	0.048
25	0.050
30	0.068
35	0.087
40	0.099
45	0.111
50	0.121
55	0.130
60	0.139
65	0.148
70	0.157
75	0.163
80	0.170
85	0.176
90	0.182
95	0.189
100	0.192

Table 1. Deviation in entropy with actual number of zombies.

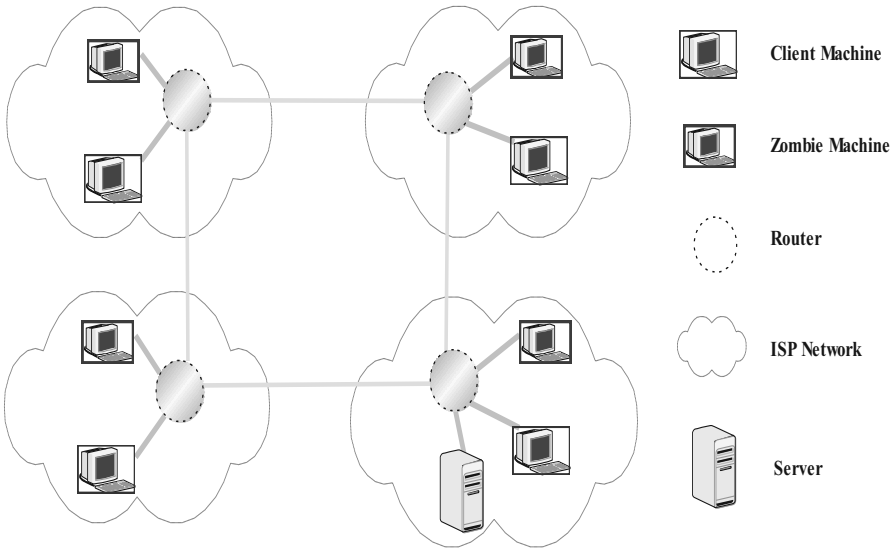

Figure 1. A short scale simulation topology.

zombies from 10-100 with increment of 5 zombies i.e. the mean attack rate per zombie from 0.25 Mbps–2.5 Mbps. Table 1 represents deviation in entropy with actual number of zombies.

Pace Regression Model is developed using the number of zombies (Y) and deviation $(H_C - H_n)$ in entropy value as discussed in Table 1 to fit the regression equation.

7. Results and Discussion

We have developed Pace Regression Model as discussed in Section 6. Various performance measures are used to check the accuracy of this model.

The number of zombies can be computed and compared with actual number of zombies using proposed regression model. The comparison between actual number of zombies and predicted number of zombies using Pace Regression Model is depicted in Figure 2.

To represent false positive (falsely predicted normal clients as zombies) and false negative (zombies are identified as normal client), we plot residual error. Positive cycle of residual error curve represents false positive, while negative cycle represents false negative. Table 2 shows residual error for Pace Regression Model. Figure 3 represents residual error for Pace Regression Model.

Table 3 shows values of various performance measures. It can be inferred from Table 3 that for Pace Regression Model values of R^2, *CC*, *SSE*, *MSE*, *RMSE*, *NMSE*, η, *MAE* are 0.98, 0.99, 368.15, 19.38, 4.40, 0.69, 0.97 and 0.84 respectively. Hence number of zombies predicted by this model is close to the observed number of the zombies.

(X) Entropy Variation	(Y) Number of Zombies	Residual error
0.045	10	3.16
0.046	15	−1.20
0.048	20	−5.31
0.050	25	−9.19
0.068	30	−4.30
0.087	35	1.17
0.099	40	2.84
0.111	45	4.77
0.121	50	5.34
0.130	55	5.24
0.139	60	5.36
0.148	65	4.94
0.157	70	5.40
0.163	75	3.72
0.170	80	2.16
0.176	85	0.92
0.182	90	−1.20
0.189	95	−2.39
0.192	100	−5.36

Table 2. Summary of residual error for Pace Regression Model.

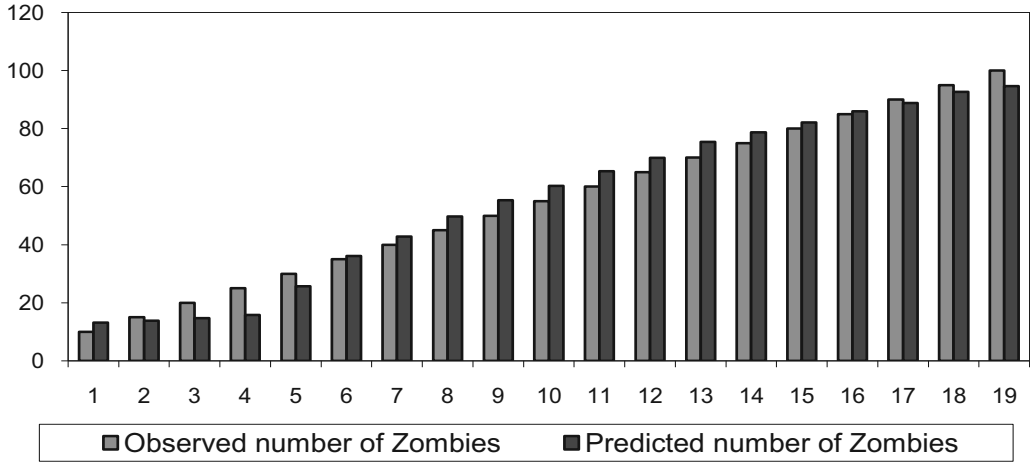

Figure 2. Comparison between actual number of zombies and predicted number of zombies using Pace Regression Model.

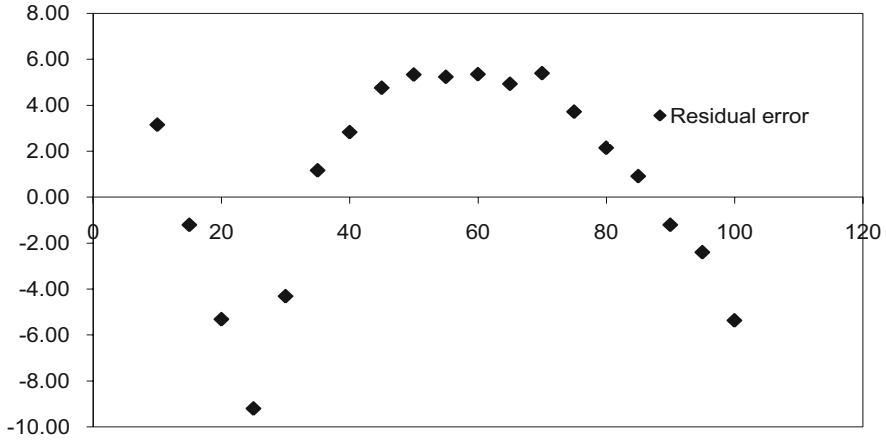

Figure 3. Residual error in Pace Regression Model.

R^2	0.98
CC	0.99
SSE	368.15
MSE	19.38
RMSE	4.40
NMSE	0.69
η	0.97
MAE	0.84

Table 3. Values of various performance measures.

8. Conclusion And Future Work

Recently, a number of highly publicized incidents of DDoS make clear that it is a complex and difficult problem. Several schemes have been proposed on how to defend against these attacks, but they suffer from a range of problems, some of them being impractical and others not being effective against these attacks. This paper investigates suitability of Pace Regression Model to predict number of zombies involved in a flooding DDoS attack from deviation $(H_c(X) - X_n(X))$ in sample entropy. We have calculated various statistical performance measures i.e. R^2, *CC*, *SSE*, *MSE*, *RMSE*, *NMSE*, η, *MAE* and residual error and their values are 0.98, 0.99, 368.15, 19.38, 4.40, 0.69, 0.97 and 0.84 respectively. Therefore, total number of predicted zombies using Pace Regression Model is very close to observe/actual number of zombies. However, simulation results are promising as we are able to predict number of zombies efficiently. Experimental study using a real time test bed can strongly validate our claim.

Acknowledgment

The authors gratefully acknowledge the financial support of the Ministry of Human Resource Development (MHRD), Government of India for partial work reported in the paper.

References

[1] B. B. GUPTA, M. MISRA, R. C. JOSHI, An ISP level Solution to Combat DDoS attacks using Combined Statistical Based Approach, in *International Journal of Information Assurance and Security (JIAS)*, vol. 3, issue 2, Dynamic Publishers Inc., USA, pp. 102–110, 2008.

[2] B. B. GUPTA, R. C. JOSHI, M. MISRA, Defending against Distributed Denial of Service Attacks: Issues and Challenges, in *Information Security Journal: A Global Perspective*, vol. 18, number 5, Taylor & Francis Group, UK, pp. 224-247, 2009.

[3] B. B. GUPTA, R. C. JOSHI, M. MISRA, ANN Based Scheme to Predict Number of Zombies involved in a DDoS Attack, *International Journal of Network Security (IJNS)*, vol. 14, no. 1, ISSN 1816-3548, pp. 36–45, 2012.

[4] Y. WANG & I. H. WITTEN, *Pace Regression*, Technical Report 99/12, Department of Computer Science, University of Waikato, September 1999.

[5] D. MOORE, C. SHANNON, D. J. BROWN, G. VOELKER, S. SAVAGE, Inferring Internet Denial-of-Service Activity, *ACM Transactions on Computer Systems*, 24 (2), 115–139, (2006).

[6] GT-ITM Traffic Generator Documentation and Tool. Available at: `http://www.cc.gatech.edu/fac/EllenLegura/ graphs.html`.

[7] NS Documentation. Available at: `http://www.isi.edu/nsnam/ns`.

[8] D. V. LINDLEY, (1987) Regression and Correlation Analysis, *New Palgrave: A Dictionary of Economics*, vol. 4, pp. 120–23.

[9] DAVID A. FREEDMAN, *Statistical Models: Theory and Practice*, Cambridge University Press, 2005.

[10] C. E. SHANNON, A Mathematical Theory of Communication, *ACM SIGMOBILE Mobile Computing and Communication Review*, vol. 5, pp. 3–55, 2001.

[11] B. GIBSON, TCP Limitations on File Transfer Performance Hamper the Global Internet, *White paper*, Sept. 2006. Available at: `http://www.niwotnetworks.com/gbx/TCPLimitsFastFileTransfer.htm`.

Received: June, 2010
Accepted: May, 2012

Contact address:
B. B. Gupta
University of New Brunswick, Canada
e-mail: `gupta.brij@gmail.com`

B. B. GUPTA received the Bachelor of Engg. degree in information technology from the Rajasthan University, Jaipur, India. He was a topper in the graduate studies and was awarded with the Institute Fellowship for his excellent performance. He got his PhD degree from Dept. of Electronics and Computer Engg, Indian Institute of Technology Roorkee, India. In 2009, he was selected for the prestigious Canadian Commonwealth Scholarship and awarded with Government of Canada Award. He spent more than 6 months at University of Saskatchewan (UofS), Canada to complete a portion of his research work. Dr. Gupta has published more than 35 research papers in international journals and conferences of high repute. He has visited more than 6 countries to presented his research papers. His biography is selected to be published in the 30th edition of the prestigious Marquis Who's Who in the World, 2012. He has also served as Technical Program Committee (TPC) member of more than 15 international conferences worldwide. Dr. Gupta is member of IEEE, SIGCOMM, The Society of Digital Information and Wireless Communications (SDIWC), Internet Society, Institute of Nanotechnology, Life Member, International Association of Engineers (IAENG), life member of the International Association of Computer Science and Information Technology (IACSIT). His research areas include information security, cyber security, intrusion detection, network performance evaluation. Currently, he is doing post doctoral research at the Faculty of Computer Science, University of New Brunswick, Canada.