Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/081L02

Implementation of Certified list for Botnet Detection

Aneel Rahim[1,*], *Khizar Hayat*[2] *and Tai-hoon Kim* [3,*]

[1] Telecommunication Software and System Group, WIT, Waterford, Ireland
[2] International Islamic University, Islamabad, Pakistan
[3] Department of Convergence Security, Sungshin Women's University, Seoul, South Korea

Received: 8 Apr. 2013, Revised: 3 Aug. 2013, Accepted: 4 Aug. 2013
Published online: 1 Apr. 2014

Abstract: Bots are compromised computers and combine together form a network called Botnet. Bots perform the task whatever their master ordered them. Communication between Bots and their master is done through different ways but most common way is Internet Relay Chat (IRC). Using IRC bots communicate with their master, master also sends commands to bots with IRC. Encryption is also done to secure communication between master and bots. Bots perform malicious activities. Different mechanism exists to detect the botnet. One detection method is also IP Blacklist but it also has some problems. If the IP spoofing is used by the attacker then some legitimate users become the part of IP blacklist. In this paper we have developed a certified list. This certified list is used to eliminate legitimate users from IP Blacklist to solve its deficiency. We implemented the proposed model in Java and experimental results are showing the effectiveness, correctness, reliability and usability of the proposed model. We calculated qualities of blacklists in term of responsiveness, completeness and percentage of completeness of IP blacklist.

Keywords: IP Blacklist, Bots, Botnet, malicious,Botnet defense

1 Introduction

Botnet works according to the commands which their master order them[1]. They enable attacker to launch different attacks for instance (spamming [2,3], Information Leakage [4,2], DDoS attacks,Click Fraud, Identity Theft[4] etc.). There are many detection schemes but more noticeable schemes are i.e., Anomaly-based Detection [5], Mining-based Detection [6], Signature-based Detection [7] and DNS-based Detection [8]. Blacklists are also used for the detection and blockage of the botnets attacks. For example Spamhaus list and the Bleeding Snort rule [9]. In this paper mathematical model is developed for calculating the certified list. Certified list contains the legitimate users which were blocked in IP blacklist. Then we subtract this certified list from IP blacklist and now at this point IP blacklist possibly will not contain legitimate users. We measured the qualities of IP blacklist in term of responsiveness, completeness and percentage of completeness.
This paper is organized as follows: in Section 2, we will discuss the Literature Survey of botnet detection techniques. IP blacklist is used to detect the botnet and certain parameters are proposed to detect the quality and effectiveness of IP blacklist. In Section 3, we enhance the mathematical model for IP blacklist to detect botnet. In Section 4 we will discuss implementation of proposed mathematical model, in Section 5 we will present the experimental results and lastly in Section 6 conclusion is given.

2 Related Work

Botnet Detection is a very difficult task. Botnet detection only can be done when they communicate at large scale [14]. For example, in DDoS attack information is required from variety of different data sources after having enough information from different sources then we can detect a Botnet malicious activity[10]. There are some botnet detection techniques given bellow.

2.1 Botnet Detection Using IP Blacklist

David et al [11] proposed the idea to detect the botnet attacks with help of IP blacklist and develop the

* Corresponding author e-mail: arahim@tssg.org, taihoonn@paran.com

framework to measure the quality of blacklist. In order to determine the effectiveness of the framework, responsiveness, completeness and percentage of completeness of IP blacklist are required to be measured first. Responsiveness is the average reported latencies for the IP blacklist.

$$\gamma(Bi) = \frac{\sum_{j=0}^{|B_i|} L_{ij}}{|I|} \tag{1}$$

Bi is the subset of blacklist containing group of IP addresses.
I is the set of infected IP addresses.
L is the set of latencies and Lij shows the infected address IPj in blacklist Bi.
Completeness determines how many of the infected addresses (set I) were reported in each blacklist.

$$\psi(Bi) = \left| \cup_{j=0}^{B_i} L_{ij} \neq [log(T_{out})] \right| \tag{2}$$

where Tout is the predefined timeout interval. Usually it is equal to 5 days.
Percentage of Completeness is

$$\chi(Bi) = \frac{\psi(Bi)}{|I|} \tag{3}$$

2.2 Enhanced IP Blacklist for Botnet Detection

IP blacklist obtained from multiple sources are used to block the users that are the part of botnet. But the problem in this case is that sometimes legitimate users are also blocked. This problem occurs when selfish nodes used spoofed IPs and do some malicious activities like dos, spamming, etc and server will put his IP in blacklist. Rahim et al.[12] proposed certified list to remove the legitimate user from blacklist and identify the botnet attacks more accurately. We also measure the responsiveness and completeness of blacklist by adding the certified list.

$$Blacklist(B_L) = \{B_{L1}, B_{L2}, B_{L3}, . \quad . \quad . \quad . \quad , B_{Ln}\}$$

$$B_{L1} = \{I_{P1}, I_{P2}, I_{P3}, . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad , I_{Pm}\}$$

$$B_{L2} = \{I_{P1}, I_{P2}, I_{P3}, . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad , I_{Pp}\}$$

. . . . . .
. . . . . .
. . . . . .
. . . . . .

$$B_{Ln} = \{I_{P1}, I_{P2}, I_{P3}, . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad , I_{Pq}\}$$

$$Certifiedlist(C_L) = \{C_{L1}, C_{L2}, C_{L3}, . \quad . \quad . \quad . \quad , C_{Lw}\}$$

$$C_{L1} = \{I_{P1}, I_{P2}, I_{P3}, . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad , I_{Px}\}$$

$$C_{L2} = \{I_{P1}, I_{P2}, I_{P3}, . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad , I_{Py}\}$$

. . . . . .
. . . . . .
. . . . . .
. . . . . .

$$C_{Ln} = \{I_{P1}, I_{P2}, I_{P3}, . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad , I_{Pz}\}$$

B_L is the set of blacklist of blocked IP addresses and B_{Li} BLi is the subset of B_L. C_L is the set of certified list of IP addresses and C_{Li} is the subset of C_L.

$$\gamma(B_{Li} - C_{Li}) = \frac{\sum_{j=0}^{|B_{Li}-C_{Li}|} L_{ij}}{|I|} \tag{4}$$

$$K_i = B_{Li} - C_{Li}$$

$$\gamma(K_i) = \frac{\sum_{j=0}^{|K_i|} L_{ij}}{|I|} \tag{5}$$

$$x = K_i$$

$$y = |I|$$

$$\gamma(K_i) = \frac{\sum_{j=0}^{x} L_{ij}}{y} \tag{6}$$

$$\psi(K_i) = \left| \cup_{j=0}^{x} L_{ij} \neq [log(T_{out})] \right| \tag{7}$$

$$\chi(K_i) = \frac{\psi(K_i)}{y} \tag{8}$$

3 Proposed Solution

The major objective of the research is to minimize the blockage of legitimate users in the IP Blacklist[15]. For

this purpose we proposed a solution which is given below in detail.

3.1 Development of Certified List

IP Blacklist is enhanced by the mathematical model so that the blockage of legitimate user will be minimized in IP blacklist. For this purpose certified list is proposed to remove the legitimate user from blacklist and identify the botnet attacks more accurately. Certified list is developed by calculating the following parameters against each IP address in IP Blacklist.

3.1.1 Type of Attack

In each IP Blacklist, against each IP the type of attack is identified and every attack is assigned a value with respect to its type and nature. There are some famous attacks which botnet masters launch with the help of their botnets. We assign them values according to their relevance as shown in Table 3.1.

Table 1: Values of Different Types of Attacks

Types of Attacks	Values
Crawl	3
Spam Yield	4
Spam	5
Dictionary Attack	7
Bad Event	8

a) Crawl: This type of attack is consisted of a set of user profiles that are used as rogue agent for the injection into the clickstream navigation data and would change the future behavior of the system.
b) Spam Yield: This is defined as the number of spam messages causing from web harvesting behavior. The greater the spam yield, the extra damaging and intimidating the harvester is deliberated to be.c) Spam: Spam means that flooding the Internet with several replicas of the same message and send to the many people they are willing to receive or not. Spam is used for the purpose of commercial advertising, habitually for uncertain products, get-rich-quick schemes.
d) Dictionary Attack: A dictionary attack means that beating a cipher or authentication mechanism by trying to define its decryption key by examining likely opportunities. This type of attack uses a directed procedure of continually trying all the words in a whole list called a dictionary. A dictionary attack cracks only those options which are most possible to succeed, usually resulting from a list of words for instance a dictionary or a bible etc. Mostly, dictionary attacks succeed because various people have a trend to pick passwords which are small single words found in dictionaries and easily anticipated differences on words, such as adding a digit.
e) Bad Events: Remote File Injections, DDOS Attacks, Malicious Website, Malware, Fake Software, Fake Websites, Criticizing Peoples Philosophies, Religions, Advancement of Obvious Sexual Material and Websites.

3.1.2 Occurrence Time

In IP Blacklist, against each IP the Occurrence Time is calculated. A default time duration is fixed and if a certain IP is blocked before the default time duration then it will be assigned a value by 0 and otherwise value will be non 0. The predefined fix time is year 2006. The occurrence time after the year 2006 will be more than 0 and before year 2006 it will be 0. The Time duration is taken from 2006 to 2011.Other assumption values are shown in Table 3.2 below.

Table 2: Values of Different Occurrence Times

Occurrence Time	Values
2006	2
2007	3
2008	4
2009	5
2010	7
2011	9

3.1.3 Frequency of attack (FA)

Frequency of attack means that how many times an attack is launch by a certain IP address. We assign values 0 to 10 to attacks according to their frequency as in Table 3.3.

3.1.4 Priority (P)

Priority is calculated by adding the above parameters Type of Attack (TA), Occurrence Time (OC) and Frequency of attack (FA).
Priority (P) = (OC+ FA+TA)/3
High priority means IP should be Blacklisted and low priority means IP should reported in certified list. Now by applying the proposed model, we can calculate certified list and certified list will contain users that was not actually attacker but because of their IP Spoofed by the

Table 3: Values for Different Frequencies of Attacks

Frequency of Attacks	Values
1 to 1000	2
1000 to 5000	5
5000 to 10000	7
More than 10000	9

attacker they added into IP blacklist. Now finally by subtracting the certified list from IP blacklist, the resulting new Blacklist will contain possibly only the attackers. This way legitimate user will not blocked and will not suffer. This will improve the performance of network and also makes the IP blacklist more accurate and easily manageable.

3.2 Measure the Qualities of Blacklist

Finally the qualities of IP Blacklists are calculated in term of Responsiveness, Completeness and Percentage of Completeness to show that which IP Blacklist is more accurate and best by applying the proposed model.
Responsiveness is the average reported latencies for an IP Blacklist. Low responsiveness means that IP Blacklist is more accurate and best because it can detect malicious activities in less time. The latency is the period of time between when an infection or disinfection of a computer is completed and when it finally comes to be informed (listed/delisted respectively) in a botnet IP blacklist.
Completeness is calculated as a factor between the number of infected IP addresses which are reported by a particular blacklist within the predefined timeout Tout and the total amount of infected hosts reported in the baseline. More Completeness means that an IP Blacklist is more accurate and best because it can detect more malicious activities.
The Percentage of Completeness x is calculated to evaluate which one of the existing IP blacklist providers is the best when reporting IP addresses of bots which are the participants of a malicious financial botnet. Percentage of Completeness calculated as a factor between the number of infected IP addresses which are reported by a particular blacklist within the predefined timeout Tout and the total amount of infected hosts reported in the baseline. By dividing the reported IPs (Completeness) with the total number of infected IPs which are reported in baseline and by multiplying 100 we get the Percentage of Completeness.

4 Implementation and Results

The proposed model is implemented in JAVA JDK 1.6 and finally result validation is performed with the existing model. IP Blacklists are used for experiments and purposed model is applied on that IP Blacklists to show that how many legitimate users are suffering in these IP Blacklist and qualities of these IP Blacklist are also calculated in term of Responsiveness, Completeness and Percentage of Completeness with the both models purposed and existing model. Finally the result validation is performed with existing model in term of responsiveness, completeness, percentage of completeness and in term of legitimate users blockage. Four different IP Blacklists have been taken from the project honeypot [13] for two countries Pakistan and India. These IP Blacklist contains total 640 IPs addresses form Pakistan and India which are blacklisted. After getting data the Infection set I is defined. In the experiments the infection set I is equal to 60 and it contains 60 IP addresses. This infection set is defined for both countries Pakistan and India separately.

4.1 Responsiveness of IP Blacklists for Pakistan and India without Certified List

Low responsiveness means that IP Blacklist is more accurate and best because it can detect malicious activities in less time as in Table 4.1. In case of Pakistan

Table 4: Responsiveness of IP Blacklists for Pakistan and India

Blacklists	Pakistan	India
A	97	158
B	159	146
C	115	124
D	124	119

we found that IP Blacklist A is best and B is worst on basis of time taken to detect malicious activities. Where as in case of India A is worse and D is best.

4.2 Completeness of IP Blacklists for Pakistan and India without Certified List

Completeness of each IP Blacklist is calculated for Pakistan and India. Completeness is calculated as a factor between the number of infected IP addresses which are reported by a particular blacklist within the predefined timeout Tout and the total amount of infected hosts reported in the baseline. We found that in case of Pakistan IP Blacklist B as given in Table 4.3 is best from all IP Blacklist because it can detect more malicious activities within the predefined timeout Tout and IP Blacklist A is worst in this scenario. Where as in case of India A is best and D is worse.

Table 5: Completeness of IP Blacklists for Pakistan and India

Blacklists	Pakistan	India
A	24	37
B	38	35
C	28	31
D	30	30

4.3 Percentage of Completeness of IP Blacklists for Pakistan and India without Certified List

Percentage of Completeness is calculated for both countries Pakistan and India as mentioned in Table 4.4. Percentage of Completeness calculated as a factor between the number of infected IP addresses which are reported by a particular blacklist within the predefined timeout Tout and the total amount of infected hosts reported in the baseline. By dividing the reported IPs

Table 6: Percentage of Completeness of IP Blacklists for Pakistan and India

Blacklists	Pakistan	India
A	24	37
B	38	35
C	28	31
D	30	30

(Completeness) with the total number of infected IPs which are reported in baseline and by multiplying 100 we get the Percentage of Completeness. We found that for Pakistan IP Blacklist B is best from all IP Blacklist because it can detect more percentage of malicious activities within the predefined timeout Tout. IP Blacklist A is worst in this scenario because it detected less percentage of malicious activities. Whereas for India IP Blacklist A is best and IP Blacklist D is the worst.

4.4 Blockage of Legitimate Users

We checked the blockage of legitimate users in four IP Blacklist for the Pakistan and India as in Table 4.5. We found that IP Blacklist A is best from all IP Blacklist in term of legitimate users blockage because it blocked less number of legitimate users and IP Blacklist B is worst because it blocked more legitimate users. Whereas in case of India IP Blacklist A is best and IP Blacklist C is the worst.

Table 7: Legtimate users blockage for Pakistan and India

Blacklists	Pakistan	India
A	3	4
B	16	6
C	10	8
D	10	5

4.5 Responsiveness of IP Blacklists for Pakistan and India with Certified List

In this case certified lists are developed for both countries Pakistan and India as in Table 4.6. Certified list is developed by calculating the parameters, type of attack, occurrence time, frequency of attack and priority. Priority is calculated by adding the above stated parameters. High

Table 8: Responsiveness of IP Blacklists for Pakistan and India

Blacklists	Pakistan	India
A	83	104
B	93	121
C	74	93
D	84	98

priority means IP should be Blacklisted and low priority means IP should reported in certified list. If the priority value is equal to or less than 5 then this IP should not be blacklisted and if the priority value is more than five then this IP should be reported in IP Blacklist. In case of Pakistan we found that IP Blacklist C is best from all IP Blacklist because it can detect malicious activities in less time. IP Blacklist B is the worst.

4.6 Completeness of IP Blacklists for Pakistan and India with Certified List

In this case completeness of each IP Blacklist is calculated for Pakistan and India with certified list as in Table 4.7. We found for Pakistan that IP Blacklist B is best from all IP Blacklist because it can detect more malicious activities within the predefined timeout Tout. IP Blacklist C is worst. Whereas for India IP Blacklist A is best and C is the worst.

4.7 Percentage of Completeness of IP Blacklists for Pakistan and India with Certified List

In this case Percentage of Completeness is calculated for both countries Pakistan and India with certified list as in

Table 9: Completeness of IP Blacklists for Pakistan and India

Blacklists	Pakistan	India
A	21	33
B	22	29
C	18	23
D	20	25

Table 10: Percentage of Completeness of IP Blacklists for Pakistan and India

Blacklists	Pakistan	India
A	35	55
B	36.6	48.3
C	30	38.3
D	33.3	41.6

Table 4.8. For Pakistan we found that IP Blacklist B is best from all IP Blacklists because it can detect more percentage of malicious activities within the predefined timeout Tout. IP Blacklist C is worst in this scenario. Whereas for India A is best from all IP Blacklists because it can detect more percentage of malicious activities and IP Blacklist C is worst.

5 Comparison

The comparison of both models is done through responsiveness, completeness, percentage of completeness.

5.1 Comparison of Responsiveness With and Without Certified List

Responsiveness for both models is compared for Pakistan and India as mentioned in Fig. 5.1(a, b). Comparison results are showing that which IP Blacklist is best in term of responsiveness. Results are showing that proposed model is best in term of responsiveness from existing model. Proposed model can detect malicious activities in less time than the existing model and it detects correct malicious activities than the existing model. The existing model detects some surplus activities as malicious activities. The IP Blacklist C for the case of Pakistan is best in proposed model it can detect malicious activities in less time. But for the same case in existing model IP Blacklist A is best in term of Responsiveness but it is not correct because exiting model detected some legitimate users as malicious activities. The IP Blacklist C for the case of India is best in proposed model it can detect malicious activities in less time. But for the same case in existing model IP Blacklist D is best in term of Responsiveness but it is not correct because exiting model detected some legitimate users as malicious activities.

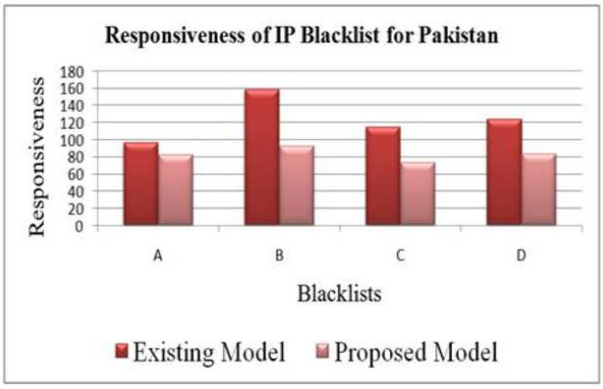

Fig. 1: Responsiveness for Pakistan

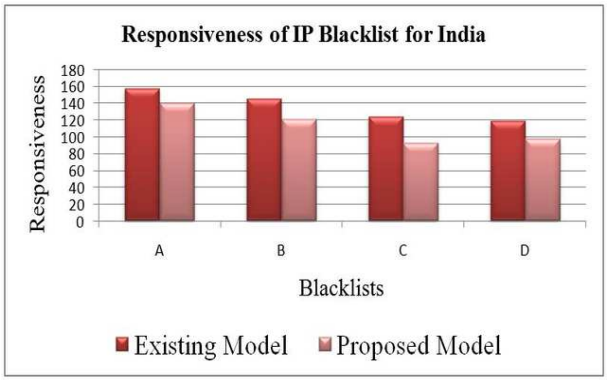

Fig. 2: Responsiveness for India

5.2 Comparison of Completeness With and Without Certified List

Completeness is compared for Pakistan and India as mentioned in Fig. 5.2(a, b). Results are showing that proposed model is best in term of completeness from existing model. Proposed model can detect correct malicious than the existing model. The existing model detects some surplus activities as malicious activities.

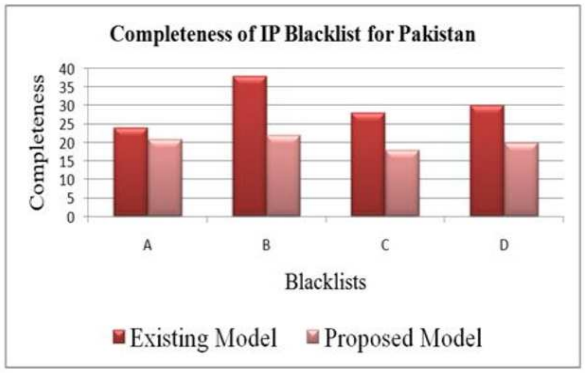

Fig. 3: Completeness for Pakistan

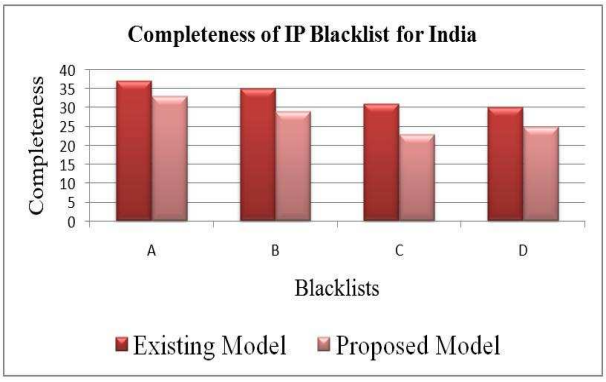

Fig. 4: Completeness for India

5.3 Comparison of Percentage of Completeness With and Without Certified List

Percentage of completeness for both models is compared for Pakistan and India as mentioned in Fig. 5.1(a, b). Results are showing that proposed model is best in term of percentage of completeness from existing model. Proposed model can detect correct malicious activities than the existing model. The IP Blacklist B for Pakistan is best in proposed model it can detect more and correct percentage of malicious activities within predefined timeout Tout. But for the same case in existing model IP Blacklist B is best in term of percentage of completeness but it is not correct because exiting model detected some legitimate users as malicious activities. The IP Blacklist A for India is best in proposed model it can detect correct malicious activities within predefined timeout Tout. But for the same case in existing model IP Blacklist A is best in term of completeness but there are included some surplus malicious activities because exiting model detected some legitimate users as malicious activities.

The IP Blacklist B for the case of Pakistan is best in proposed model it can detect more and correct malicious activities within predefined timeout Tout. But for the same case in existing model IP Blacklist B is best in term of completeness but it is not correct because exiting model detected some legitimate users as malicious activities. The IP Blacklist A for the case of India is best in proposed model it can detect correct malicious activities within predefined timeout Tout. But for the same case in existing model IP Blacklist A is best in term of completeness but there are included some surplus malicious activities because exiting model detected some legitimate users as malicious activities.

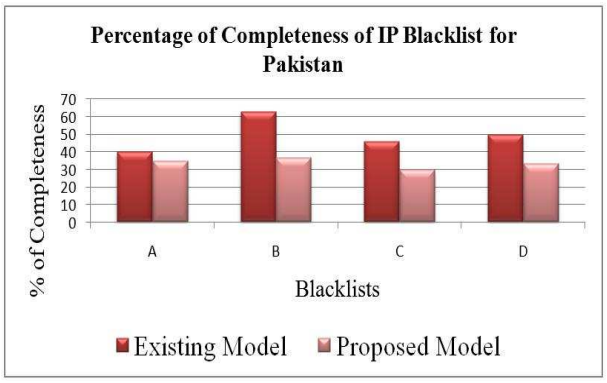

Fig. 5: Percentage of Completeness for Pakistan

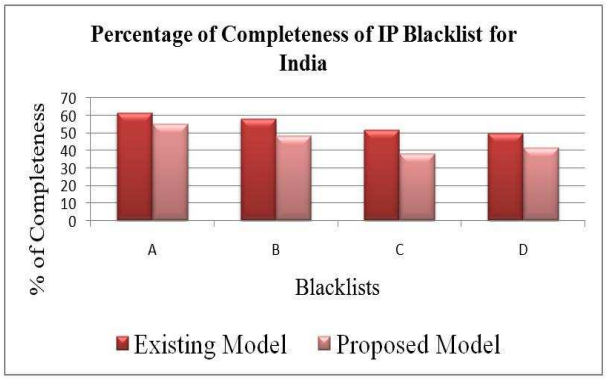

Fig. 6: Percentage of Completeness for India

6 Conclusion

The existing model do not deals with legitimate users but the proposed Certified list deals with legitimate users and this model try to minimize the blockage of these legitimate users as much as possible. Results are proving that fact. Proposed model minimized 3.75% legitimate users from IP Blacklist A, 20% from B, 12.5% from C and also 12.5% from D in the case of Pakistan. Proposed model minimized 5% legitimate users from IP Blacklist A, 7.5% from B, 10% from C and also 6.25% from D in the case of India.

References

[1] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, C. Kruegel, G. Vigna, In: Proceedings of the 16th ACM conference on Computer and communications security, 635 (2009).
[2] K. Pappas, Communications News, **45**, 12 (2008).
[3] P. Sroufe, S. Phithakkitnukoon, R. Dantu, J. Cangussu, In: Proceedings of the 6th IEEE Conference on Consumer Communications and Networking Conference, 1074 (2009).
[4] B. T. Holz, M. Kotter, G. Wicherski, Online Available, http://www.honeynet.org/papers/bots, (2012).
[5] W. Tylman, In: Proceedings of the Third International Conference on Dependability of Computer Systems, **211**, (2008).
[6] C. Dartigue, H. Jang, W. Zeng, In: Proceedings of the Seventh Annual Communication Networks and Services Research Conference, 372 (2009).
[7] S. Neelakantan, S. Rao, In: Proceedings of the Third International Conference on Internet Monitoring and Protection, 80 (2008).
[8] D. A. L. Romana, S. Kubota, K. Sugitani, Y. Musashi, In: Proceedings of the 2008 First International Conference on Intelligent Networks and Intelligent Systems, 205 (2008).
[9] M. P. Collins, T. J. Shimeall, S. Faber, J. Janies, R. Weaver, M. D.Shon, J. Kadane, In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, 93 (2007).
[10] J. Govil, J. Govil, In: IEEE Electro/Information Technology Conference, **215**, (2007)
[11] D. Oro, J. Luna, T. Felguera, M. Vilanova and J. Serna, In: Sixth International Conference on Information Assurance and Security (IAS), (2010).
[12] A. Rahim, K. Hayat, M. Sher,T.-h Kim, In: Information Journal, **14**, 3335 (2011).
[13] http://www.projecthoneypot.org/, (2012).
[14] A. Rahim, F. B. Muhaya, In: FGIT-SecTech/DRBC 231, (2010).
[15] K. Hayat Master thesis, Supervisor M. Sher, International Islamic University, (2011).

Aneel Rahim is Postdoctoral Researcher at Telecommunication Software and System Group, WIT, Waterford, Ireland. He is editor of Computer Science Journal and Guest editor of Multimedia Tools and Applications Springer, Telecommunication System Journal Springer, Information international journal and Journal of Internet Technology.

Khizar Hayat has done Bachelor in Computer science from COMSATS Institute of Information Technology Islamabad and Master degree from International Islamic University Pakistan.He has published several journal publications and also reviewer of international journals and conferences.His main area of research is adhoc networks,BOTNETs etc.

Tai-hoon Kim is Professor at Department of Convergence Security, Sungshin Women's University, Seoul, South Korea. His research area includes adhoc networks, security, botnet detection, sensor network. etc.