ANOMALY DETECTION ON USER BROWSING BEHAVIORS FOR PREVENTION APP_DDOS

Vidya Jadhav[1] and Prakash Devale[2]
[1]Student, Department of Information Technology, Bharti Vidyapeeth Deemed University, Pune, India
[2]Professor & Head, Department of Information Technology, Bharti Vidyapeeth Deemed University, Pune, India

ABSTRACT

Some of the hardest to mitigate distributed denial of service attacks (DDoS) are ones targeting the application layer. Over the time, researchers proposed many solutions to prevent denial of service attacks (DDoS) from IP and TCP layers instead of the application layer. New application Layer based DDoS attacks utilizing legitimate HTTP requests to overwhelm victim resources are more undetectable. This may be more serious when such attacks mimic or occur during the flash crowd event of the website. This paper present a new application layer anomaly detection and filtering based on Web user browsing behavior for create defense against Distributed Denial of Service Attack(DDoS). Based on hyperlink characteristics such as request sequences of web pages. This paper, uses a large scale Hidden Semi Markov Model (HsMM) to describe the web access behavior and online implementation of model based observation sequence on user browsing behavior fitting to the model measure of user's normality.

KEYWORDS: *Hidden Semi Markov Model, APP_DDOS, user's normality detection, browsing behavior.*

I. INTRODUCTION

In the last couple of years, attacks against the Web application layer have required increased attention from security professionals. The main APP_DDOS attack techniques that have been used, is utilizing the HTTP "/GET" request by requesting home page of victim server repeatedly. Without specifying URL of web page of victim website, attackers easily find out the domain name of the victim web site. Many statistical or dynamical techniques that have been used to create defense against distributed denial of service (DDOS) attack on web application.

Statistical detection detect Automated attacks using tools such as Nikto or Whisker or Nessus Attacks that check for server misconfiguration, HTML hidden field attacks (only if GET data –rare) Authentication brute-forcing attacks, Order ID brute-forcing attacks (possibly) – but if it is POST data, then order IDs cannot be seen .Static Detection fail to detect attacks that overflows various HTTP header field, Web Application attacks in a POST form. Statistical method can hardly distinguish the vicious HTTP request from the normal one [12].

To overcome these issues, anomaly detection system on web browsing behavior, this supports detection of new APP_DDOS attacks. This paper presents a model to capture the browsing patterns of web users using Hidden Semi Markov Model (HsMM) and to detect the APP_DDOS Attacks.

II. RELATED WORK

Most of current research has focus on network layer (TCP/IP) instead of application layer. To detect DDOS attack IP address, time to leave (TTL) values were used [1][2]. C. Douligeris and A. Mitrokotsa [3] classify DDOS defense mechanism depending on the activity deployed and location deployment. Cabrera [4] shown that Statistical Tests applied in the time series of MIB(Management

Information Base) traffic at the Target and the Attacker are effective in wextracting the correct variables for monitoring in the Attacker Machine.

To the best of my knowledge, a few existing work has been done on the detection of APP_DDOS attacks. S. Ranjan[5] deployed a counter mechanism which assign a suspicious measure to a session in proportion to legitimate behaviour and decide when whether the session is serviced using DDOS Scheduler. C. Kruegel introduced a novel approach to perform anomaly detection using HTTP queries parameter (e.g String length of an attribute value) [6].

The existing work for web user behavior can be summarized as the following ways 1) Based on probabilistic model, a double Pareto distribution for long normal distribution and link choice for the revisiting etc.[9]. 2) Based on click stream and web contents e.g. data mining [10] to capture web user's usage patterns from page content and click streams data set. 3) Based on Markov chain e.g. Markov chain to model the URL access patterns that are observed on the navigation logs based on the previous state[11] . 4) User behaviour to implement anomaly detection e.g. uses system call data sets generated by program to detect the anomaly access of UNIX system based on data mining [13]

Disadvantages with existing system

1) This system does not take into account the user's series of operation information e.g. which page will be requested next. They can not explain the browsing behavior of a user because the next page the user will browse is primarily determined by the current page he is browsing
2) The method omits dwell time that the user stays on a page while reading and they do not consider the cases that a user may not follow the hyperlink provided by the current page.
3) From the network perspective, protecting is considered in effective. attacks flows can still incur congestion along the attack path
4) It is very hard to identify DDoS attack flows at sources since the traffic is not so aggregate.

Thus a new system is designed that take into account the users series of operation information. There is an intensive computation for page content processing and data mining and hence they are very suitable for online detection. The dwell time that the user stays on a page while reading and we can find cases that a user may follow the hyperlinks provided by the current page.

III. APP_DDOS ATTACKS

APP_DDOS Attacks may exhausting the limited server resources such as CPU cycle ,network bandwidth, DRAM space, database, disk or specific protocol data structures, including service degradation or outage in computing infrastructures for the client [7]. System downtime resulting from DDOS attacks could lead to million dollars' loss. Thus, APP_DDOS attacks may cause more serious threats in high speed internet because increasing in computational complexity of internet application & larger network bandwidth those server resources may become bottleneck of that application.

First characteristics of APP_DDOS attacks is that attacker targeting at some popular Websites are increasing moving away from pure bandwidth flooding to more surreptitious attacks that hide in normal flash crowds of the website. Thus, such website become more & more demands of information broadcast and e-commerce, the challenges of network security are how to detect and respond to the APP_DDOS attacks if they occur during a flash crowd event.

Second characteristics of APP_DDOS attacks is that application layer request originating from compromised hosts on internet are indistinguishable from those generated by legitimate users. APP_DDOS attacks can be mounted with legitimate request from legitimately connected network computer. To launch the attacks, APP_DDOS attacks utilize the weakness enabled by the standard practice of opening service such as HTTP and HTTPS (TCP port 80) through most firewalls. Many protocol or applications, both legitimate and illegitimate, can use these openings to tunnel through firewalls by connecting over a standard TCP port 80. Legitimate users may request services to the website, but these clients are unable to complete their transactions, website will be put busy giving responses to the Zombie processes. In this paper, APP_DDOS attacks can be identified by using browsing behavior of user, the elements of browsing behaviour of user are HTTP request rate, page viewing time, page requesting sequence.

IV. PROBLEMS WITH APP_DDOS DETECTION

The main aim of a DDoS defense system is to `relieve victim's resources from high volume of counterfeit packets sent by attackers from distributed locations, so that these resources could be used to serve legitimate users. There are four approaches to combat with DDoS attack as proposed by Douligeris et al. [3]: Prevention, Detection and Characterization, Trace back, and Tolerance and Mitigation. Attack prevention aims to fix security holes, such as insecure protocols, weak authentication schemes and vulnerable computer systems, which can be used as stepping stones to launch a DoS attack. This approach aims to improve the global security level and is the best solution to DoS attacks in theory. Attack detection aims to detect DDoS attacks in the process of an attack and characterization helps to distinguish attack traffic from legitimate traffic. Trace back aims to locate the attack sources regardless of the spoofed source IP addresses in either process of attack (active) or after the attack (passive). Tolerance and mitigation aims to eliminate or curtail the effects of an attack and try to maximize the Quality of Services (QoS) under attack. Carl et al. Douligeris et al. and Mirkovic et al. have reviewed a lot of research schemes based on these approaches but still no comprehensive solution to tackle DDoS attacks exist. One of the main reasons behind it is lack of comprehensive knowledge about DDoS incidents. Furthermore the design and implementation of a comprehensive solution which can defend Internet from variety of APP_ DDOS attacks is hindered by following challenges:

1. Large number of unwitting participants.
2. No common characteristics of DDoS streams.
3. Use of legitimate traffic models by attackers.
4. No administrative domain cooperation.
5. Automated DDoS attack tools.
6. Hidden identity of participants because of source addresses spoofing.
7. Persistent security holes on the Internet.
8. Lack of attack information.
9. The APP_DDOS attacks utilize high layer protocol to pass through most of the current anomaly detection system designed for low layer & arrive at victim website.
10. Flooding is not the unique way for the APP_DDOS. There are many other forms, such as consuming the resources of the server, arranging the malicious traffic to mimic the average request rate of legitimate user or utilizing the large scale botnet to produce low rate attack flows.
11. APP_DDOS attacks usually depend on successful TCP connection, which makes the general defense schemes based on detection of spoofed IP address useless.

V. WEB BROWSING BEHAVIOR

The browsing behavior of web user is mainly influenced by the structure of website (e.g. hyperlink and the web documents) and the way users access web pages. Web user browsing behavior can be abstracted & profiled by user request sequences. User can access the web pages by two ways. First users click a hyperlink pointing to a page, the browser will send number of request for the page and it's in line objects. Then, user may follow series of hyperlink provided by the current browsing pages to complete his access. Second way, the user jump from one page to another by typing URLs in address bar, selecting from the favorites of the browser or using navigation tools.

Fig 1 shows web browsing model. Webpage clicked by a web user can uniquely represented by semi Markov state(S). State transition probability matrix A presents the hyperlink relation between different webpages. The duration of a state present the number of HTTP requests received by the webserver. The output sequences of each state throughout its duration present those requests of the clicked page which pass through all proxies and then arrive at webserver. Take a simple example to explain these relations by fig.1 The unseen page sequences is page1,page2,page3 .Except those responded by cashes or proxies, HTTP request sequences received by the webserver is(r1,r2,r3,r4,r5,r6,r7,r8,r9,r10,r11). When the observed request sequences inputted to the HsMM, the algorithm may group them into three clusters (r1,r2,r3,r4), (r5,r6,r7), (r8,r9,r10,r11) and denote them state sequence (1,2,3). The state transition probability a12 represent the probability that page2 may be

accessed after accessing current page1 by the user. The duration of the first state 1 is d=4, which means 4 HTTP requests of page1 arrived at the webserver.
Frequency of the clicking behavior of user for multiple page requests will be calculated by using HsMM

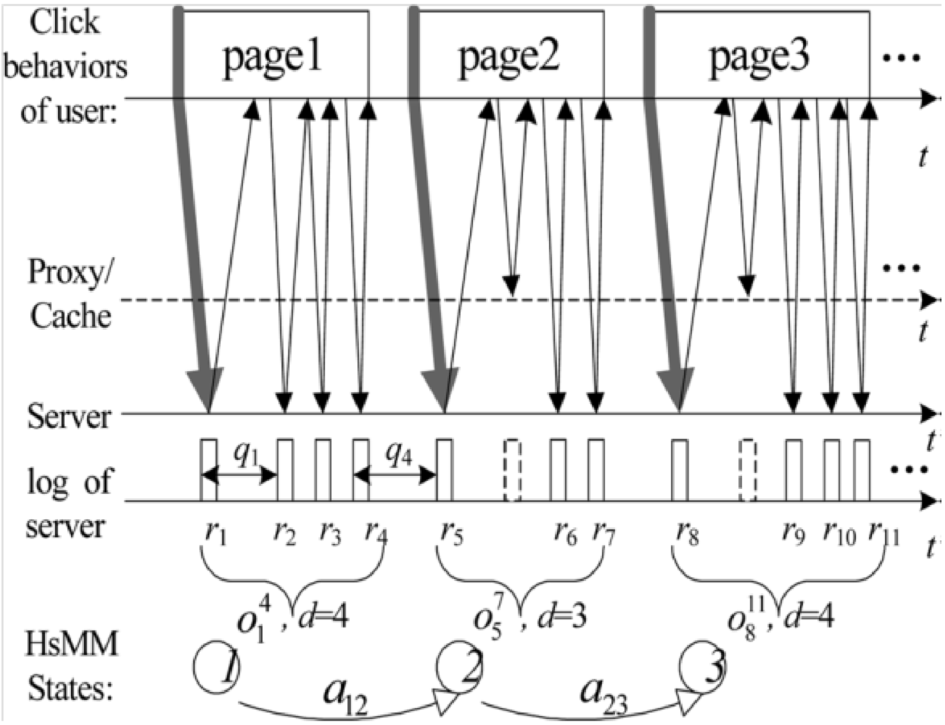

Figure 1: Web browsing behavior

VI. TECHNIQUE USED OR ALGORITHMS USED

To achieve early attack detection and filtering for the application-layer-based DDoS attack we use an extended hidden semi-Markov model is proposed to describe the browsing behaviors of web surfers. In order to reduce the computational amount introduced by the model's large state space, a novel forward algorithm is derived for the online implementation of the model based on M algorithm. Entropy of the user's HTTP sequence fitting to the model is used as a criterion to measure the user's normality.

6.1 Hidden Semi-Markov Model

HsMM is an extension of the hidden Markov Model with explicit state duration. It is a stochastic finite state machine, specified by (**S, π, A, P**) where:

1 **S** is a discrete set of hidden states with cardinality N, i.e. S = {1, N}.

2 ***π*** is the probability distribution for the initial state $\boldsymbol{\pi}_{\mathbf{m}} \equiv$ **Pr** [s1 = *m*], s_t denotes the state that the system takes at time and *m Є S.* The initial state probability distribution satisfies $\Sigma_m \pi_m = 1$;

3 **A** is the state transition matrix with probabilities: $a_{mn} \equiv \Pr[s_t = n \mid s_{t-1} = m]$, m, n Є S, and the state transition coefficients satisfy $\Sigma_n a_{mn} = 1$;

4 **P** is the state duration matrix with probabilities: *pm* (d) ≡ Pr[r_t = d | s_t = m], r_t denote the remaining (or residual) time of the current state s_t, m Є S, d Є {1,…,D}, D is the maximum interval between any two consecutive state transitions, and the state duration coefficients satisfy $\Sigma_d p_m (d) = 1$.

Consider a semi-Markov chain of M states, denoted s_1,s_2…….$S_{M,}$ with the probability of transition from state s_m to state s_n being denoted a_{mn}(m, n=1,2….M). The initial state probability distribution is given by {π_m } . Let o_t stands for the observable output at t and let qt denote the state of the semi-

Markov chain at time t, where t = 1,2,….T. The observable and the state are related through the conditional probability distribution $b_m\ (v_k) = Pr[o_t = v_k \mid q_t = s_m]$.where $\{v_k\}$ is a set of k distinct values that may assumed by observation $o_{t\ .}$ $b_m(o_{a|b}) = \pi_{t=a\ |\ b}\ b_m(o_t)$ when the "conditional independence" of outputs is assumed, where $o_{a|b} = \{o_t : a \le t \le b\}$ represent the observation sequences from time a toy time b. If the pair process $(q_{t,}\ r_t)$ takes on value (sm,d), the semi Markov chain will remain in the current state sm until time t+d-1 and transits to another state at time t+d, where $d \ge 1$. Let λ stands for the complete set of model parameters $\lambda = (\{a_{mn}\}, \{\pi_m\}, \{b_m(v_k)\}, \{p_m(d)\})$.

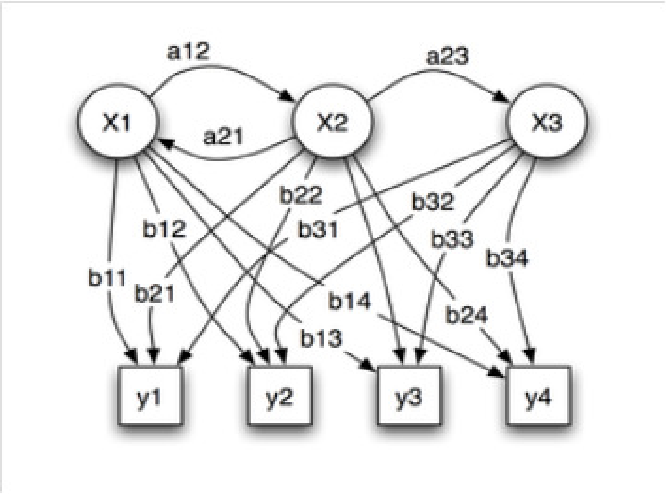

Figure 2: Markov Chain

We first define the forward and backward variable.

We define the forward variable by

$$\alpha_t(m,d) = Pr\ [o_{1|t}, (q_t,r_t) = (s_m,d)] \qquad (1)$$

A transition into state $(q_t,r_t) = (s_m,d)$ takes place either from $(q_{t-1},r_{t-1}) = (s_m, d+1)$ or from $(q_{t-1},r_{t-1}) = (s_n,1)$ for $n \neq m$. Therefore , we readily obtain the following forward recursion formula

$$\alpha_t(m, d) = \alpha_{t-1}(m, d + 1)\ b_m(o_t) + \left(\textstyle\sum_{n \neq m} \alpha t - 1(n, 1) amn\right) .b_m(o_t)p_m(d), \quad d \ge 1 \qquad (2)$$

for a given state sm and time t > 1, with the initial condition

$$\alpha_1(m, d) = \pi_m b_m(o_1)p_m(d). \qquad (3)$$

We define backward variable by

$$\beta t(m,d) = Pr[o_{t+1|T} | (q_t, r_t) = (s_m, d)]. \qquad (4)$$

By examining the possible states that follow $(q_t, r_t) = (s_m, d)$, we see that when d > 1 the next state must be $(q_{t+1}, r_{t+1}) = (s_m, d-1)$, and when d=1 it must be $(q_{t+1}, r_{t+1}) = (s_n, d')$ for some $n \neq m$ and $d' \ge 1$. We thus have the following recursion formula:

$$\beta t(m,d) = b_m(o_{t+1})\beta_{t+1}(m,d-1) \qquad \text{for } d > 1 \qquad (5)$$

and

$$\beta_t(m, 1) = \textstyle\sum_{n \neq m} a_{mn}\, b_n(o_{t+1}) \left(\sum_{d \ge 1} p_n(d)\ \beta_{t+1}(n,d)\right) \qquad (6)$$

for a given states sm and time t < T , with the initial condition (in the backward recursive steps)

$$\beta_T(m, d) = 1 \qquad d \ge 1 \qquad (7)$$

the algorithm of HsMM can be found in [15] & [16].

6.2 M-Algorithm for Normality Detection

The M-algorithm is being widely adopted in decoding digital communications because it requires far fewer computations than the Viterbi algorithm. The aim of the M-algorithm is to find a path with distortion or likelihood metrics as good as possible (i.e., minimize the distortion criterion between the symbols associated to the path and the input sequence).

M-Algorithm work as follow:

I. Select Only the best **M** paths ,at time t.
II. Each path associated with value called as path metric, which act as distortion measure of the path and is the accumulation of transition metric.
III. The transition metric is the distance between the symbol associated to a trellis transition and the input symbol.
IV. Path metric is criterion to select best M path.
V. To the next time instant t+1 by extending the M paths is has retained to generate N.M new paths.
VI. All terminal branches compared to input data to path metric and the (N-1).
VII. Deleted M poorest paths.
VIII. Until all the input sequences have been processed this process is repeated.

VII. ANOMALY DETECTION

Anomaly detection relies on detecting behaviors that are abnormal with respect to some normal standard. Many anomaly detection systems and approaches have been developed to detect the faint signs of DDoS attacks. Due to constraint in computing power, the detector and filter is unable to adapt its policy rapidly. Because the web access behavior is short term stable[14]. The filter policy must be fixed for only a short period of time. Define Td as a length of the request sequence for anomaly detection. For a given HTTP request sequences of the lth user, we calculate the average entropy from mean entropy of the model. If the deviation is larger than a predefined threshold the user is regarded as an abnormal one, and the request sequences will be described by the filter when the resources is scarce. Otherwise user's request can pass through the filter and arrive at the victim smoothly. Then , when given slot is time out , the model can implement the online update by the self adaptive algorithm proposed in [15].

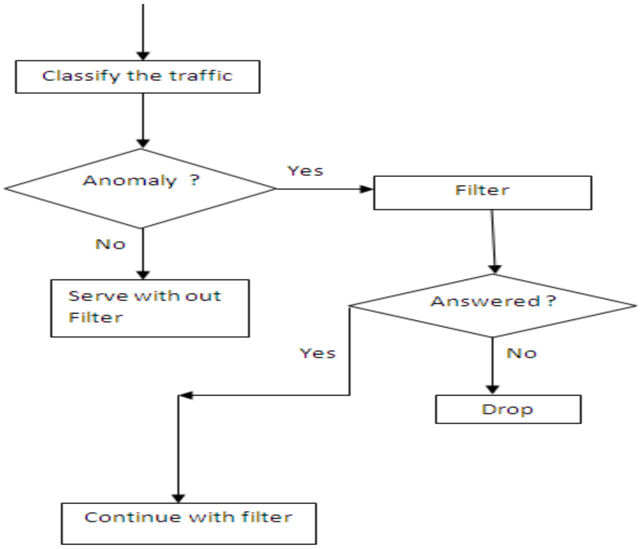

Figure 3: Algorithm for anomaly detection

VIII. PROPOSED SYSTEM

1) Monitor browsing behavior of web surfer.

2) HsMM will be used to calculate behavior of the system for abnormal user browsing, which will done by maintaining state transition.

3) Train system to distinguish between normal user browsing and abnormal user browsing, which can be done by Normality Detection and Filter policies. Detector and filter between internet and the victim will accept the HTTP request and decides whether to accept or not.

4) Make use of efficient algorithm to minimize the lot of computations for anomaly detection, so M-algorithm will be used to minimize these lots of computations.

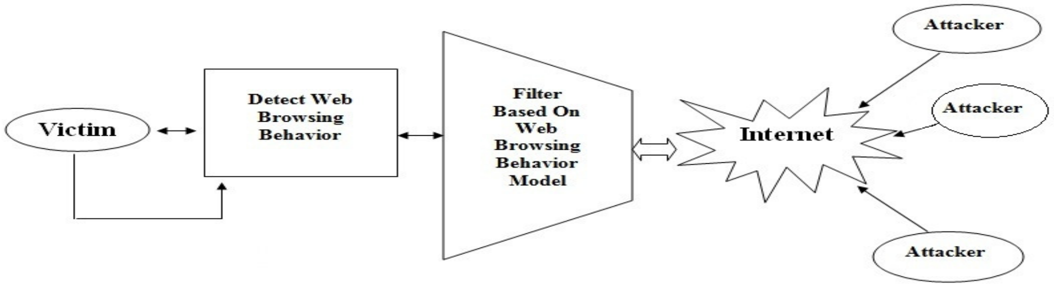
Attacker
Attacker
Attacker
Victim
Detect Web Browsing Behavior
Filter Based On Web Browsing Behavior Model
Internet

Figure 4: Anomaly detection based on behavior model

IX. RESULTS

We try to insert the APP_DDOS attack request into normal traffic shown in fig.5(a). In order to generate a stealthy attack which is not easily detected by the traditional methods, each attack node's output traffic to approximate the average request rate of normal user. The APP_DDOS attack aggregated from the low rate malicious traffic show in fig 5(b).

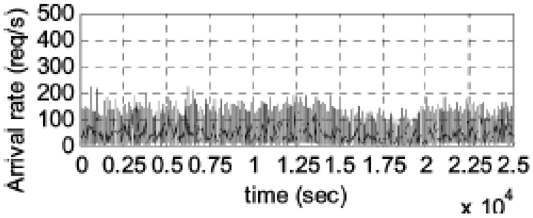

Figure 5(a) : Arrival rate Vs time of traffic without Attack

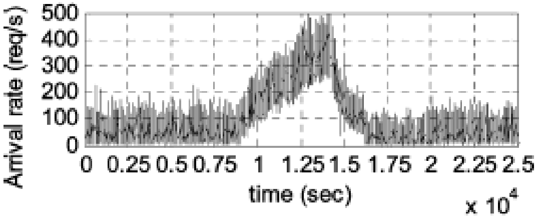

Figure 5(b) : Arrival time Vs time of traffic with Attack

X. CONCLUSION AND FUTURE SCOPE

This paper focuses on protecting Web servers from APP_DDOS attacks by using web browsing behaviour of user. We presented novel algorithm based on Large Hidden semi-Markov model that distinguish the normal and deviated behavior users. A set of real traffic data collected from an educational website and applied the M-algorithm to differentiate the normal and abnormal behaviors.

Several issues will need further research: 1) if all clients are getting service from one proxy and Zombie is behind that proxy among the legitimate clients, blocking the IP results the service annoy and service delays to the legitimate users also.2) applying this model for other schemes to detect the App.DDoS attacks, such as FTP attacks.

REFERENCES

[1]. C. Jin, H. Wang, and K. G. Shin, "Hop-count filtering: An effective defense against spoofed traffic," in *Proc. ACM Conf Computer and Communications Security*, 2003, pp. 30–41.

[2]. T. Peng, K. R. mohanarao, and C. Leckie, "Protection from distributed denial of service attacks using history-based IP filtering," in *Proc. IEEE Int. Conf. Communications*, May 2003, vol. 1, pp. 482–486.

[3]. C. Douligeris and A. Mitrokotsa, "DDoS attacks and defense mechanisms:Classification and state-of-the-art," *Computer Networks: The Int. J. Computer and Telecommunications Networking*, vol. 44, no. 5,pp. 643–666, Apr. 2004.

[4]. J. B. D. Cabrera *et al.*, "Proactive detection of distributed denial of service attacks using MIB traffic variables a feasibility study," in *Proc. IEEE/IFIP Int. Symp. Integrated Network Management*, May 2001, pp. 609–622.

[5]. S. Ranjan, R. Swaminathan, M. Uysal, and E. Knightly, "DDoS-resilient scheduling to counter application layer attacks under imperfect detection," in *Proc. IEEE INFOCOM*, Apr. 2006 [Online]. Available: http://www- ece.rice.edu/~networks/papers/dos-sched.pdf

[6]. C. Krugel & G. Vigna "Anomaly detection of Web-based attacks" in CCS'03,October 27-31,2003 washingtone, DC,USA.

[7]. S. Ranjan, R. Karrer, and Knightly, "Wide area redirection of dynamic content by Internet data centers," in *Proc. 23rd Ann. Joint Conf. IEEE Comput. Commun. Soc.*, Mar. 7–11, 2004, vol. 2, pp. 816–826.

[8]. S.-Z. Yu and H. Kobayashi, "An efficient forward-backward algorithm for an explicit duration hidden Markov model," *IEEE Signal Process. Lett.*, vol. 10, no. 1, pp. 11–14, Jan. 2003.

[9]. S. Z. Yu, Z. Liu, M. Squillante, C. Xia, and L. Zhang, "A hidden semi-Markov model for web workload self-similarity," in *Proc. 21st IEEE Int. Performance, Computing, and Communications Conf. (IPCCC 2002)*, Phoenix, AZ, Apr. 002, pp. 65–72.

[10]. S. Bürklen *et al.*, "User centric walk: An integrated approach for modeling the browsing behavior of users on the web," in *Proc. 38th Annu. Simulation Symp. (ANSS'05)*, Apr. 2005, pp. 149–159.

[11]. J. Velásquez, H. Yasuda, and T. Aoki, "Combining the web content and usage mining to understand the visitor behavior in a web site," in *Proc. 3rd IEEE Int. Conf. Data Mining (ICDM'03)*, Nov. 2003, pp. 669–672.

[12]. D. Dhyani, S. S. Bhowmick, and W.-K. Ng, "Modelling and predicting web page accesses using Markov processes," in *Proc. 14th Int. Workshop on the Database and Expert Systems Applications (DEXA'03)*, 2003, pp. 332–336.

[13]. J. Mirkovic, G. Prier, and P. L. Reiher, "Attacking DDoS at the source," in *Proc. 10th IEEE Int. Conf. Network Protocols*, Sep. 2002, pp. 312–321.

[14]. X. D. Hoang, J. Hu, and P. Bertok, "A multi-layer model for anomaly intrusion detection using program sequences of system calls," in *Proc. 11th IEEE Int. Conf. Networks*, Oct. 2003, pp. 531–536.

[15]. M. Kantardzic, *Data Mining Concepts, Models, Methods And Algorithm*. New York: IEEE Press, 2002.

[16]. X. Yi and Y. Shunzheng, "A dynamic anomaly detection model for web user behavior based on HsMM," in *Proc. 10th Int. Conf. Computer Supported Cooperative Work in Design (CSCWD 2006)*, Nanjing, China, May 2006, vol. 2, pp. 811–816.

[17]. S.-Z. Yu and H. Kobayashi, "An efficient forward-backward algorithm for an explicit duration hidden Markov model," *IEEE Signal Process. Lett.*, vol. 10, no. 1, pp. 11–14, Jan. 2003.

Biography:

Vidya Jadhav, PG Scholer in information Technology at bharatividya peeth deemed university, Pune. Her field of interst are computer networking, Operating syatem and anomaly detection.

Prakash Devale, presently working as a professor and Head department of Information Technology at bharati vidyapeeth deemed University College of Engineering, Pune. He received his ME from Bharati Vidyapeeth University and pursuing Ph.D degree in natural language processing.