Model for Network Traffic Identification at the Application Level

Samih. M. Jammoul[1], Ark. M. Andreev[2], Vladimir V. Suzev[3], Vitaly E. Chulkov[4]
Bauman Moscow State Technical University
Moscow, Russia
[1]samihj@gmail.com, [2]arkandreev@gmail.com, [3]v.suzev@bmstu.ru, [4]vechulkov@bmstu.ru

***Abstract*— Network tunnel applications are one of the most known methods to avoid monitoring or filtering in the firewall, these applications usually are used to avoid monitoring or to run prohibited applications. Recognizing such these types of tunnels is one of the most challenging issues in network traffic identification. In this paper, a new model of network traffic identification at application level is proposed. The model is able to identify some of the most known tunneling applications.**

Keywords*— *packet switching networks; network traffic identification; network tunneling; traffic encapsulation

I. INTRODUCTION

Network tunneling is a kind of encapsulation the real traffic within another legal protocol and forwarding it to its real destination through a third connection party. Network tunneling is one of the most known method to escape monitoring or to use prohibited internet application behind the firewall; most of the used tunneling application encrypt the contents as well. The tunneling applications are divided into two types according to the level of tunnel protocol.

- Tunnel application at IP level, where the tunnel application forwards traffic for all client traffic at the same time, the most known application of this type of tunnels is IPsec.
- Tunnel application at TCP level, where this type of tunnel is used to forward traffic of one application, like tunnels for web browsing, or file transfer. The most known TCP tunnel applications are SSH (Secure Shell) tunnels [4] and TOR (The Onion Router) [5].

II. RELATED WORK

This report presents a new model of network traffic identification at the application level; the presented model can be used as well to identify tunnel application at TCP level. However, in the following, we present briefly some of the most important work in the tunnel identification domain.

In the works [1,2] is proposed using machine learning technics to identify TOR traffic. In [1], a comparison of using three of machine learning is performed, the used feature set consists of twenty-five features, and most of them are based on the flow time statistics. The results of the experiments show a high detection precision for TOR and Non-Tor traffic, as well as types of traffic within the TOR tunnels. In fact, the disadvantages of this method are, first, the big number of features affects the performance of the method, so it is not suitable to use the model in wideband networks or at ISP level. Second, the proposed method does not work in the real time, so it is not suitable for working in IPS or firewall.

In the work [3], authors proposed a new method to identify the encrypted applications within the SSH tunnels. The authors used two vectors of features, packet size and time between successive packets. The main disadvantages of this method are, first, the method considers that, just one TCP session (for the carried application) is transmitted over the tunnel at same time. This assumption is not correct for all type of applications, for example, the web application may generate several TCP session between client and server at the same time. Second, the new versions of SSH protocol add a random number of bytes to packet size, so the usage of packet size does not lead to accurate results anymore.

III. NEW NETWORK TRAFFIC IDENTIFICATION MODEL

In the current work, we present a new model for network traffic identification, which identifies the traffic at the application level. The proposed model can identify different type of applications including TCP tunnel applications, like TOR and SSH application.

The identification model of one chosen application is based on hidden Markov model – HMM [6], where its Parameters are

$$\theta = \{\pi, A, B\}$$

Where the initial state distribution π is the probability of the being in state i in the initial time

$$\pi_i = P(z_i = s_N)$$

Transition probability matrix between states

$$A = \{a_{ij}\}, 1 \le i, j \le N$$

The work is being supported by the Russian Ministry of Education and Science (Project #2.7782.2017/BC dated 10/3/2017).

978-1-5386-5824-6/18/$31.00 ©2018 IEEE

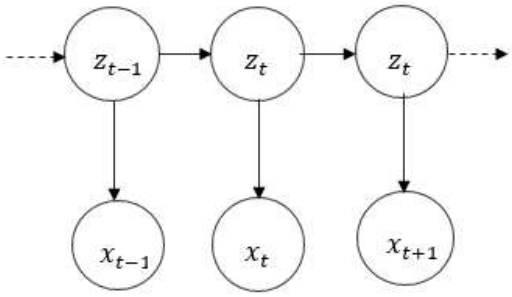

Fig. 1. Hidden Markov model

The probability of moving from the state i to state j,

$$a_{ij} = P(z_{t+1} = s_j / z_t = s_i)$$

The emission matrix

$$B = \{b_{ij}\}, 1 \le j \le M, 1 \le i \le N$$

The probability of appearing the observation *j* in the state *i*.

$$b_{ij} = P(x_t = o_j / z_t = s_i)$$

The observation are the only known values in HMM, the other model parameters are calculated basing on the observations, and the main task of the model is to find the unknown (hidden) parameters

$$\theta = \{\pi, A, B\}$$

Baum-Welch algorithm, which is a special case of the expectation maximization algorithm [7], is used to find model parameters [8]. This algorithm aims to find the parameters, which maximize the probability of appearing the training set in the model:

$$\theta^* = ARG(Max_\theta(P(X \mid \theta))) \quad (1)$$

In the current network traffic identification model, two observation vectors are used, packet size and time between successive packets. In order to increase the performance of the model, it is suggested to minimize the cardinal of observation sets, so for the packet size is calculated as the following:

$$l_{model} = \frac{l_{real}}{30} \quad (2)$$

The used unit to measure the time between successive packets is 0.1 microsecond, the maximum idle time for TCP connection is 10 minutes. To reduce this range of this parameter we propose using:

$$t_{model} = 10\log_{10}(t_{real}) \quad (3)$$

To facilitate the calculation in the proposed model, we consider that, the two observation parameters are independents, so the probability of appearing an observation (size, time) is calculated as follows:

$$P(O_l, O_t \mid \theta_l, \theta_t) = P(O_l \mid \theta_l) P(O_t \mid \theta_t) \quad (4)$$

Based on the previous assumption, two separated hidden Markov models are built to identify each of the considered network applications; the first one uses the packet size as observation parameter

$$\theta_l = \{\pi_l, A_l, B_l\}$$

The second uses the time between the successive packets

$$\theta_t = \{\pi_t, A_t, B_t\}.$$

In the previous works [9,10], it is suggested to use random values for model initial parameters values in Baum-Welch algorithm. The method of using random values to initiate model parameters has mainly two disadvantages, first, the final model parameter are changed by changing the initial values, this means, repeating same experiments in same conditions leads to different results!. Second, Baum-Welch algorithm does not converge all the time when using random initial values, so the random initial values method does not offer stable results.

In the current proposed model, it's suggested to estimate the parameters initial values based on Gaussian mixture model (GMM), this assumption means that the distribution of the observation parameters (packet size and time) is approximated to Gaussian mixture distribution

$$P(x) = \sum_{i=1}^{k} \varphi_i N(\mu_i, \sigma_i)$$

Each component in Gaussian mixture distribution is considered equivalent to a state emission distribution as the following:

$$b_{l,ij} = N_l(j, \mu_i, \sigma_i) \quad (5)$$

$$b_{t,ij} = N_t(j, \mu_i, \sigma_i) \quad (6)$$

To estimate the initial state distribution, we calculate the distribution of the first packet of each flow as the following:

$$\pi_{l,i} = \sum_{j=-50}^{50} P(l, j) B_{l,ij} \quad (7)$$

$$\pi_{t,i} = \sum_{j=1}^{100} P(t, j) B_{t,ij} \quad (8)$$

where the value *P(l, j)* represents the number of the flows of the studied application, which have size packet equals to *j*, divided by the number of all application flows. The same for the value *P(t, j)*, which represents the number of the flows of the studied application, which have time between first abd second packets equals to *j*, divided by the number of all application flows.

Fig. 2. Shows the proposed model of network traffic identification for a set of chosen applications. In the training phase, the model parameters are calculated for each of the chosen applications by using its training dataset. In the testing phase, when a new flow is showing up, the probabilities of appearing this flow are calculated for all considered application, the maximum probability specifies the application with taking into consideration that its value should be upper than a specified level.

IV. EXPERIMENTS AND RESULTS

To test the proposed model, we use two group of datasets, first one is gathered in Bauman – Moscow State Technical University, computer networks department, and we got the second dataset for testing from University of New Brunswick–Cyber Security Lab (Canada). We did some tests using the following application: Web application (HTTP and HTTPS), Email (IMAPS), chat application (WhatsApp), tunnel applications TOR and SSH. The used application within the application tunnels (TOR and SSH) are Web application (HTTP and HTTPS). We use the precision to measure the effectiveness of the proposed model:

$$precision = \frac{TP}{TP + FP} \quad (9)$$

The results in Fig. 3. Show that 9 packets are sufficient to identify all the chosen application with precision > 95%. The needed number of states for each of the chosen models is between 4–6 states for each model depending on the type of application.

By using this model, we calculated as well the percentage of the identified flows of the application, which is defined by the recall function as the following:

$$recall = \frac{TP}{TP + FP} \quad (10)$$

Fig. 4. Shows that most of the application flows are identified with a very high percentage > 90% when using over than 5 packets.

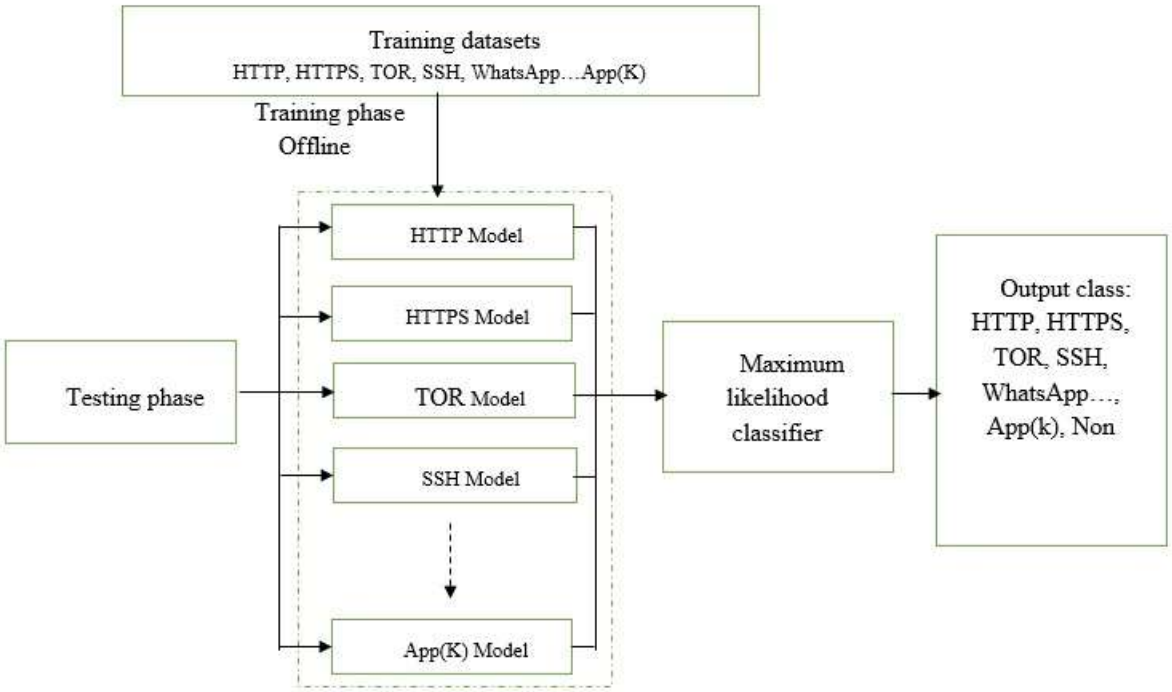

Fig. 2. Network traffic identification model

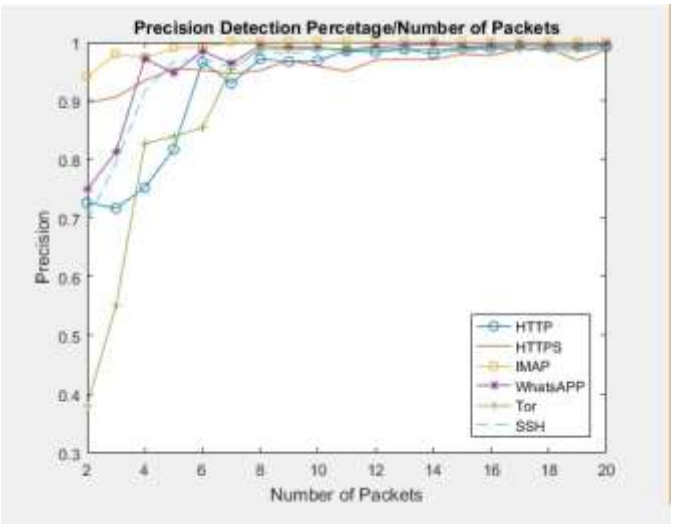

Fig. 3. The precision of identification model regarding the number of flow packets

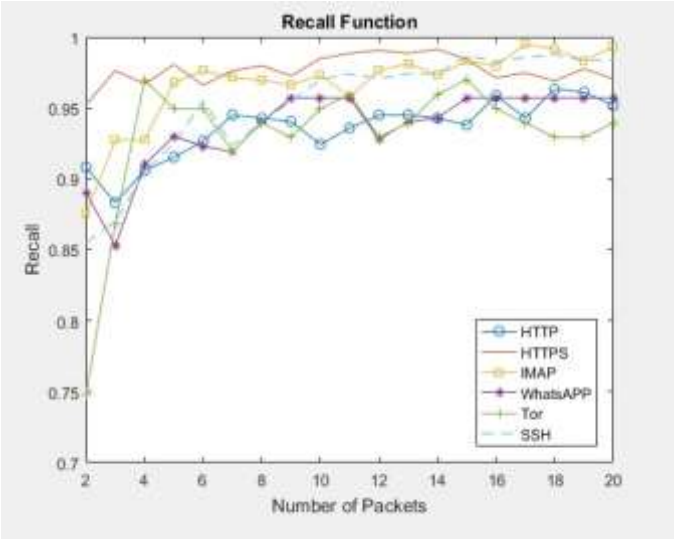

Fig. 4. Recall of identification model regarding the number of flow packets

V. Conclusion

In current work, we presented a new model for network traffic identification based on hidden Markov model. The proposed model as well can identify the application, which generates the flow in real time, where it is needed less than 10 packets to identify the application with high accuracy. The current model identifies the tunnel application as well, which is used to transfer type of one application, identification of the tunnel application at IP level will be discussed in our feature work.

Acknowledgment

The authors would like to thank the Canadian Institute of Cyber Security, University of Brunswick for providing us with the dataset to carry out this work.

References

[1] Arash Habibi Lashkari, Gerard Draper-Gil, Mohammad Saiful Islam Mamun and Ali A. Ghorbani, "Characterization of Tor Traffic Using Time Based Features", In the proceeding of the 3rd International Conference on Information System Security and Privacy, SCITEPRESS, Porto, Portugal, 2017.

[2] Hodo E, Bellekens X, Iorkyase E, Hamilton A, Tachtatzis C, Atkinson R (2017) Machine learning approach for detection of nontor traffic. ARES'17, August 2017, Reggio Callabria, ITALY.

[3] Maurizio Dusi, Manuel Crotti, Francesco Gringoli, Luca Salgarelli, "Detection of Encrypted Tunnels across Network Boundaries", 2008 IEEE International Conference on Communications.

[4] The Secure Shell (SSH) Transport Layer Protocol[Электронный ресурс] / 2018. – Режим доступа https://tools.ietf.org/html/rfc4253

[5] The Onion Router, Wikipedia, [OnLine] / 2018. – Режим доступа https://en.wikipedia.org/wiki/Tor_(anonymity_network)

[6] L.R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition / L.R. Rabiner // Proceedings of the IEEE. 1989. Volume 77, № 2. С. 257–285.

[7] Baum-Welch Algorithm [Online] / Wikipedia. 2017. – Available at: https://ru.wikipedia.org/wiki/Алгоритм_Баума_–_Велша

[8] J.A. Bilmes, A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, university of Berkeley, CA, Technical Report ICSI-TR-97-021, 1998.

[9] Dainotti, A., De Donato W., Pescape A., Rossi P.S., (2008) Classification of network traffic via packet-level hidden markov models. IEEE Global Telecommunications Conference (GLOBECOM) 2008, New Orleans, LA, USA.

[10] Wright, C.V., Monrose, F., Masson, G.M., HMM profiles for network traffic classification (extended abstract), in Proc. ACM Workshop on Visualization and Data Mining for Computer Security, pp. 9–15, Oct. 2004.