Модель идентификации трафика сетей передачи данных на уровне приложения

С. М. Джаммул[1], А. М. Андреев[2], В. В. Сюзев[3], В. Е. Чулков[4]
МГТУ им. Н.Э. Баумана
[1]samihj@gmail.com, [2]arkandreev@gmail.com, [3]v.suzev@bmstu.ru, [4]vechulkov@bmstu.ru

***Аннотация*. TCP-туннель является одним из самых известных способов избежать фильтрации в брандмауэре, где пользователь может посетить запрещенные сайты или использовать запрещенные приложения. Распознавание зашифрованного TCP-туннеля в трафике сети передачи данных является одной из самых сложных задач идентификации сетевого трафика. В этой работе представлены некоторые типы TCP-туннеля, а также предложен новый метод обнаружения зашифрованного TCP-туннеля на основе скрытой марковской модели.**

Ключевые слова: сети передачи данных; идентификация трафика; анализ трафика; TCP-туннель; инкапсуляция трафика

I. Введение

Самый известный способ скрытия передаваемых данных через интернет является переносом этих данных с использованием другого протокола (носитель) после осуществления их шифрования. Этот метод называется туннельным протоколом или приложением. К типам туннельных протоколов можно отнести:

- Туннельный протокол на IP уровне, где весь трафик хоста проходит через этот туннель, самый известный туннельный протокол на IP уровне является IPSec.
- Туннельный протокол на уровне TCP-протокол, где этот тип туннеля используется для передачи трафика одного приложения. Самые известные туннельные протоколы на этом уровне являются SSH (Secure Shell protocol)[5] и TOP (The Onion Router) [6], где эти приложения используются в основном для просмотра веб-сайтов без мониторинга в брандмауэре.

II. Анализ существующих методов идентификации туннельных приложений

В этом материале представлена модель идентификации сетевого трафика, и ее использование для идентификации туннельных приложений на TCP уровне, это следует из анализа результатов некоторых самых важных работ в области идентификации туннельных приложений.

Работа выполнена при финансовой поддержке министерство образования и науки РФ, проект №2.7782.2017/ВС, от 10/3/2017

В работах [1, 2] предложено использовать метод машинного обучения для идентификации трафика TOP. В работе [1] используются три типа машинного обучения, а набор признаков составляется из 25 признаков на основе временных характеристик потока. При этом с высокой точностью осуществляется идентификация TOP части трафика TOP и не-TOP. Но на самом деле, этот подход обладает двумя недостатками, во-первых, большое число признаков не подходит к условиям идентификации трафика в сетях с широкой полосой пропускания, во-вторых, этот метод не работает в реальное время из-за выбора признаков на уровне потока, которые вычисляются после окончания потока.

В работе [3] представлен метод идентификации SSH туннели, и также тип протокола, который проходит через это туннель. В этой работе используются два вектора параметров, длина пакета и интервал времени между пакетами. Главные недостатки этого подхода являются, во-первых, в работе предполагается, что SSH туннель проводит только один TCP сеанс, и на основе этого предположения считаются входные параметры модели идентификации. Но это редко случается в реальном трафике, где одно приложение генерирует несколько TCP сеансов в одно то же время, как в случае веб-приложений. Во-вторых, новые версии SSH приложений (туннелей) добавляют случайное число байтов на перенесенных пакетах (избыточность), поэтому сигнатура приложения, вычисленная на основе длины перенесенных пакетов, не является адекватной в этом виде туннели.

III. Предложена модель идентификации сетевого трафика

В настоящей работе представляется модель идентификации сетевого трафика на основе скрытой Марковской модели (СММ), рис 1. Предложенная модель идентифицирует сетевой трафик на уровне приложения, в том числе туннельные приложения SSH и TOP.

Параметры скрытой Марковской модели

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

где вероятность находиться в состояниях в начальное время:

$$\pi = \{\pi_1, \pi_2, ..., \pi_N\}$$

$$\pi_i = P(z_1 = s_i)$$

Матрица перехода между состояниями

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

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

Матрица вероятностей появления наблюдения в состояния

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

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

В этой модели известны только значении наблюдений, а все остальные параметры неизвестны, и задача заключается в построении модели и нахождении ее неизвестных параметров[6].Чтобы найти параметры модели используются алгоритм Баума–Велша (Baum–Welch) [8] который является частным случаем EM алгоритма (Expectation Maximization algorithm) [7].

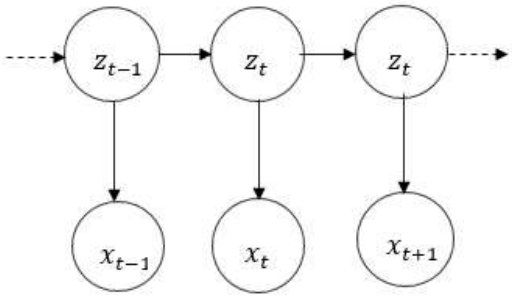

Рис. 1. Скрытая Марковская модель

Цель алгоритма Баума–Велша является оценивание параметров модели, при которых получаются максимальная вероятность выявления набора наблюдений (данные тренировки), т.е. алгоритм найдет:

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

В предложенной модели используются два наблюдаемых значения: длина пакета *l*, интервал времени между пакетами *t*. Также для повышения производительности предлагается уменьшить мощность пространства наблюдений следующим образом:

$$l_{Model} = [\frac{l_{Real}}{30}] \quad (2)$$

$$t_{Model} = [10\log_{10}(t_{Real})] \quad (3)$$

В данной модели идентификации сетевого трафика предполагается, что значения наблюдений *l* и *t* независимые друг от друга, поэтому справедливо:

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

Исходя из этого, построены две самостоятельные СММ для идентификации сетевых приложений с использованием параметров *l* и *t*:

- параметры модель СММ с использованием *l*

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

- параметры модель СММ с использованием *t*

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

Оценка начальных значений параметров в существующих моделях классификации сетевого трафика на основе СММ, как работы [9, 10], используют случайные значения, но этот метод обладает два недостатки, во-первых, при повторении экспериментов в одинаковых условиях результаты изменяются. Во-вторых, случайные значения параметров не постоянно ведет к сходимости алгоритма Баума–Велша. В настоящей модели для оценки начальных значений параметров используется модель гауссовых смесей МГС (Gaussian mixture model), где начальные распределения наблюдений в состояниях аппроксимируются гауссовым распределениям:

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

При этом мы считаем, что каждый компонент в смеси эквивалентен одному состоянию, поэтому значения распределений наблюдений в состояниях *B* в каждой модели СММ вычисляются следующим образом

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

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

Для вычисления вероятности нахождения в состояниях в начальное время, мы вычисляем распределение длины первого пакета каждого потока, а также распределение интервала времени между первым и вторым пакетами каждого потока. Вероятность каждого состояния в начальный момент времени вычисляется следующим образом:

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

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

где *P(l ,j)* число потоков приложения, где первый пакет имеет длину равно *j*, на число потоков, и *P(t,j)* число потоков приложения, где время между первым и вторым пакетами равно *j*, на число потоков.

Рис. 2 показывает структуру модели, где построена специальная модель идентификации для каждого приложения APP. На этапе обучения, вычисляются параметры каждого интересного приложения APP с использованием его набора данных обучения (наблюдений). На этапе тестирования, когда наступает новый поток, проводится вычисление вероятности наступления этого потока на каждой модели приложения. После сравнения вычисленных вероятностей, ожидаемое приложение определяется на основе максимального значения.

IV. Эксперименты и результаты

Для тестирования предложенную модель идентификации сетевого трафика генерирован набор трафика в локальной сети университета МГТУ им. Н.Э. Баумана, и также использован набор трафика из университета Нью-Брансуика (Канада). Изучемые приложения в экспрементах являются: веб приложения (HTTP и HTTPS), электронная почта (IMAPS), чат приложения (WhatsApp) и туннельные приложения SSH и TOP. Эксперименты проводятся с использованием туннельных приложений, чтобы проводить Веб приложение.

Эксперименты проводятся с использованием от 4 до 6 состояний для обеих частей модели (интервал времени между пакетами и длина пакетов) которые зависят от изучаемого приложения.

Рис. 2. Модель идентификации сетевого трафика на основе CMM

Рис. 3 показывает, что 9 пакетов достаточно для идентификации всех приложения с точностью > 95%, где показатель точности определяется следующим образом:

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

Мы вычисляли полноту идентификации потоков приложений, которая определяется следующим образом:

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

Рис. 4 показывает, что после 5 пакетов полнота идентификации всех изучаемых приложений являются выше чем 90%.

V. Заключение

В этой работе представлена модель идентификации сетевого трафика на основе CMM, эта модель идентифицирует разные типы приложения, в том числе, туннельные приложения на уровне TCP протокола с высокой точностью, и также эта модель идентифицирует сетевой трафик в реальном времени где достаточно меньше чем 10 пакетов для идентификации приложения.

Наша следующая задача в этой тематике идентификация типов приложений в туннельных приложениях на уровне IP протокол.

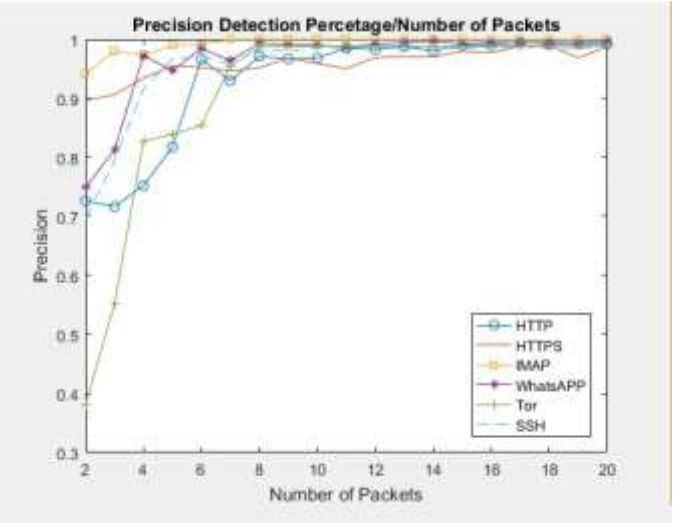

Рис. 3. Точность идентификации приложений на основе числа использованных пакетов.

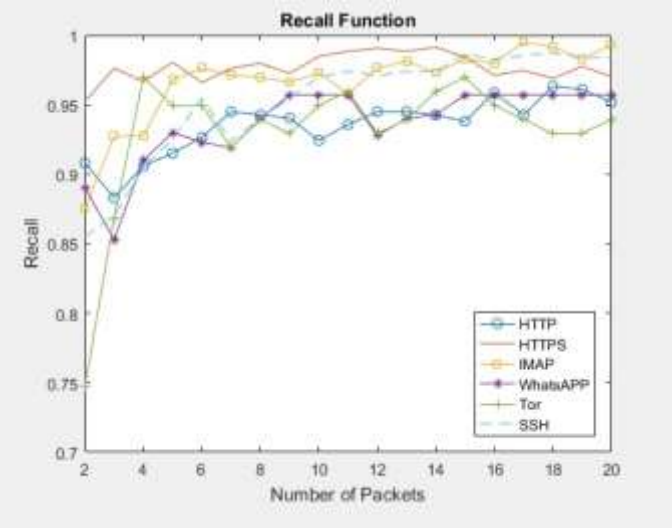

Рис. 4. Полнота модели идентификации приложений на основе числа пакетов

Список литературы

[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, [Электронный ресурс] / 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] Алгоритм Баума–Велша [Электронный ресурс] / Википедия. 2017. – Режим доступа: 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.