STUDIA INFORMATICA 2011
Volume 32 Number 3A (98)

Tomasz BILSKI[1]
Politechnika Poznańska, Instytut Automatyki i Inżynierii Informatycznej

ANALIZA RUCHU NA PODSTAWIE WIELKOŚCI PAKIETÓW IP

Streszczenie. W artykule przedstawiono koncepcję wykorzystania długości pakietów IP w analizie ruchu, w sieciach komputerowych. W pierwszej części artykułu dokonano przeglądu ograniczeń związanych z klasyfikacją ruchu na podstawie numerów portów. Główna część artykułu poświęcona jest koncepcji analizy i klasyfikowania ruchu sieciowego na podstawie długości pojedynczych pakietów oraz rozkładu długości w zbiorach pakietów. W ostatniej części przedstawiono przykładowe zastosowanie w systemach wykrywania kradzieży usług.

Słowa kluczowe: ochrona danych, zapora sieciowa, IDS, pakiet IP

TRAFFIC ANALYSIS BASED ON IP PACKET SIZE

Summary. The paper discusses traffic analysis/classification problem. First part deals with some problems and limitations of analysis with a use of port numbers. Main part of the paper presents a concept of packet size based traffic classification. Exemplary application is provided in the last part of the paper.

Keywords: data security, firewall, IDS, IP packet

1. Wprowadzenie

Jednym z podstawowych elementów zarządzania sieciami oraz ochrony danych są analiza i klasyfikacja ruchu w sieci. Celem takiej analizy jest identyfikacja protokołu, usługi sieciowej bądź aplikacji wysyłającej/odbierającej pakiet lub strumień pakietów. W procesie identyfikacji usługi sieciowej, powiązanej z określoną jednostką transmisyjną (np. pakietem IP), wykorzystać można wiele informacji dostępnych w poszczególnych warstwach stosu TCP/IP, ze szczególnym uwzględnieniem warstw międzysieciowej i transportowej.

[1] Praca naukowa finansowana ze środków na naukę w latach 2010 – 2013, jako projekt badawczy.

W pracach teoretycznych [np. 2, 4, 5] do klasyfikowania ruchu sieciowego proponuje się, między innymi:

- czas trwania połączenia TCP,
- statystyki dotyczące odstępów między kolejnymi pakietami (ang. *interarrival time*),
- liczbę przesyłanych bajtów i pakietów,
- przepustowość,
- symetrię lub asymetrię ruchu.

W praktyce analiza jest najczęściej oparta na numerach portów oraz informacjach z warstwy zastosowań. Rozwiązania takie są powszechnie stosowane, niemniej mają swoje ograniczenia. Analiza na poziomie warstwy zastosowań w znacznym stopniu obciąża węzeł analizujący, a tym samym negatywnie wpływa na przepustowość sieci. Klasyfikacją ruchu w sieci zajmują się między innymi zapory sieciowe, systemy IDS, IPS i DLP [1].

W głównej części artykułu rozważa się metody analizy ruchu na podstawie długości pakietów oraz rozkładu tych długości w strumieniu pakietów. Pierwsza z metod jest względnie prosta i efektywna, a ponadto pozwala rozwiązać niektóre problemy, związane z analizą przeprowadzaną z użyciem numerów portów warstwy transportowej.

2. Ograniczenia związane z analizą na podstawie numerów portów

2.1. Wprowadzenie

Oczywistym sposobem określenia aplikacji wysyłającej/odbierającej jest analiza numerów portów warstwy transportowej. Numer portu źródłowego jest identyfikatorem procesu wysyłającego dane, a numer portu docelowego – procesu odbierającego dane. Istnieje duża liczba tzw. predefiniowanych numerów portów, przypisanych na stałe do określonych usług sieciowych i protokołów komunikacyjnych[2]. Przykładowo port 80 jest przypisany do protokołu HTTP. Analiza ruchu na podstawie numerów portów warstwy transportowej jest stosunkowo prostym i wydajnym rozwiązaniem. Jest stosowana na przykład do filtrowania pakietów w zaporach sieciowych. Analiza taka charakteryzuje się pewnymi ograniczeniami, a jej wyniki mogą być błędne lub niedokładne. Numer portu może być niedostępny do analizy lub może być użyty w nietypowy sposób. Wśród problemów można wyróżnić:

- niestandardowe wykorzystywanie portów predefiniowanych,
- dynamiczne ustalanie numerów portów w niektórych aplikacjach,
- fałszowanie numerów portów,

[2] Zajmuje się tym IETF.

- translację adresów NAT z użyciem portów,
- szyfrowanie pakietów.

2.2. Niestandardowe wykorzystanie portów predefiniowanych

Przypisanie określonego portu do danego protokołu/usługi nie jest jednoznaczne z zagwarantowaniem, że dana usługa w każdym przypadku będzie realizowana z użyciem danego portu i że dany port będzie powiązany tylko i wyłącznie z daną usługą lub protokołem komunikacyjnym. Przykładowo, port 25 przypisany przez IETF do protokołu SMTP jest używany przez dziesiątki koni trojańskich, w tym[3]: Ajan, Antigen, EPS, Gip, Gris, Happy99, Kuang2, MBT, Naebi, Shtirlitz, Stealth, Tapiras, Terminator, WinPC, WinSpy.

2.3. Dynamiczne ustalanie numerów portów

W niektórych aplikacjach sieciowych numer portu nie jest z góry znany, lecz ustalany w trakcie nawiązywania połączenia. Przykładem są systemy telefonii IP (a także inne usługi czasu rzeczywistego, bazujące na protokołach RTP i RTCP), wykorzystujące protokół sygnalizacyjny SIP lub H.323. Numery portów dla transmisji strumieni audio w trakcie rozmowy są ustalane w procesie sygnalizacji przed rozpoczęciem właściwej rozmowy.

Dodatkowym problemem może być to, że niektóre usługi korzystają z dużej liczby portów. Przykładowo jedna sesja telefonii IP może używać do 10 portów, a sesja związana z Web 2.0 i technologią AJAX korzysta z jeszcze większej liczby portów.

2.4. Translacja adresów IP

Powszechnie stosowana metoda translacji NAT (Network Address Translation) adresów IPv4 w nietypowy sposób wykorzystuje numery portów. W momencie dokonywania translacji adresu pakietu wychodzącego z sieci lokalnej numer portu źródłowego przestaje pełnić funkcję identyfikującą konkretny proces hosta-nadawcy, a staje się identyfikatorem samego hosta. Informacja umożliwiająca powiązanie danego pakietu (i ewentualnie pakietów zwrotnych) z konkretnym procesem jest dostępna w pamięci rutera dokonującego translacji. W takim przypadku poprawna klasyfikacja pakietu (powiązanie pakietu z usługą) na podstawie numeru portu staje się niemożliwa bez dostępu do tablicy w ruterze NAT.

[3] http://www.sans.org/security-resources/idfaq/oddports.php

2.5. Szyfrowanie pakietów

Numery portów stają się niedostępne dla analizy, jeżeli pakiet IP jest przesyłany w postaci zaszyfrowanej. W przypadku użycia protokołu IPSec, zarówno w trybie tunelowym, jak i transportowym, pole danych pakietu (a tym samym nagłówek TCP lub UDP z numerami portów) ma postać zaszyfrowaną.

3. Analiza długości pakietów IP

3.1. Uwagi ogólne

Długość pakietu IP jest zależna od wielu czynników, w tym: wielkości komunikatu warstwy zastosowań, protokołu warstwy transportowej, długości pola opcji w nagłówku IPv4 lub długości nagłówków opcjonalnych w IPv6.

Pakiet IP składa się z nagłówka i pola danych. Pole danych zawiera jednostkę transmisyjną protokołu warstwy transportowej, która również może zawierać nagłówek i pole danych, z jednostką transmisyjną warstwy aplikacji. Tak, więc wielkość pola danych pakietu IP jest zależna od wielkości nagłówków warstwy transportowej i aplikacji oraz wielkości pola danych w warstwie aplikacji. O ile wielkości nagłówków warstwy transportowej są stałe (TCP – 20 bajtów[4], UDP – 8 bajtów), o tyle wielkości nagłówków wielu protokołów warstwy aplikacji są zmienne, np. DHCP, SNMP, DNS, NTP, HTTP, SIP, RTP, RTCP. Niemniej w typowych zastosowaniach długości nagłówków komunikatów mieszczą się w określonych granicach, są przewidywalne a niekiedy są ściśle określone. Przykładem jest RTP w telefonii IP – jeżeli strumień audio jest przesyłany między dwoma węzłami, to nagłówek RTP składa się dokładnie z 12 bajtów. W takim przypadku jedynym elementem zmienny pozostaje długość pola danych komunikatu RTP (por. 3.3.1).

Zaletą klasyfikacji ruchu, bazującej na długości pakietów jest możliwość przeprowadzania analizy nawet w przypadku, gdy numer portu na to nie pozwala lub gdy pakiety są zaszyfrowane. Szyfrowanie w oczywisty sposób ukrywa zawartość pola danych pakietu, w tym także informacje umieszczone w nagłówkach warstw wyższych, transportowej i aplikacji. Pole długości pakietu, umieszczone w nagłówku IP, może zawierać fałszywą wartość, podobnie jak sfałszowany może być adres IP czy numer portu nadawcy. Węzeł analizujący pakiety nie ma prostej możliwości wykrycia fałszerstwa adresu IP czy numeru portu. Natomiast określenie faktycznej długości pakietu w przypadku sfałszowania nagłówka jest możliwe na podstawie informacji

[4] Nagłówek TCP może zawierać opcje o zmiennej długości, niemniej są one stosowane rzadko i zwykle tylko na etapie nawiązywania połączenia.

z warstwy łącza danych – w najprostszym przypadku długość pakietu IP jest równa aktualnej długości pola danych ramki ethernetowej.

Teoretycznie maksymalna długość pakietów IP jest ograniczona wielkością pola długości w nagłówku. W wersji IPv4 pole długości ma 16 bitów, w związku z tym maksymalna wielkość pakietu (w bajtach) wynosi 65 535[5]. Podobnie jest w IPv6, z tym, że w tej wersji możliwe jest wysyłanie także tzw. pakietów *jumbo* o długości do 4 GiB – długość pakietu jest podawana nie w standardowym 16-bitowym polu długości, lecz w nagłówku opcjonalnym. W praktyce wielkość pakietu jest ograniczona głównie przez właściwości protokołów warstwy łącza danych. W wersji IPv4 zakłada się, że każdy węzeł sieci jest w stanie przesyłać pakiety o długości co najmniej 576 bajtów – nie ma gwarancji, że pakiet o większej długości zostanie przesłany bez fragmentacji. W IPv6 limit ten został zwiększony do 1280 bajtów. W tym miejscu należy zauważyć, że powszechnie używane protokoły warstwy łącza danych spełniają ten warunek (tab. 1).

Tabela 1

Maksymalna wielkość pola danych w wybranych protokołach warstwy łącza danych

Protokół	Maksymalna wielkość pola danych [B]	Uwagi
Ethernet, ramka standardowa	1500	
Ethernet, *jumboframe*	9216	Brak powszechnej implementacji
IEEE 802.11	2312	
Fibre Channel	2112	

W większości przypadków rzeczywista długość pakietu jest znacznie mniejsza aniżeli dopuszczalna w danej sieci. Badania statystyczne ruchu w Internecie [3] wskazują, że około 40% wszystkich transmitowanych pakietów nie przekracza długości 40 bajtów, a średnia długość pakietu IP wynosi około 355 bajtów.

3.2. Analiza rozkładu długości pakietów

Do analizy ruchu wykorzystać można informacje na temat rozkładu długości w określonym zbiorze/strumieniu pakietów. Klasyfikacja ruchu sieciowego na podstawie rozkładu długości pakietów została zaproponowana w pracy [6]. Przedstawiana tam metoda może być stosowana pod warunkiem, że oprócz długości poszczególnych pakietów dostępne są także inne informacje, przy czym autorzy nie rozważają problemów, przedstawionych w rozdziale 2 niniejszego artykułu. Ponadto, konieczność analizowania zbiorów pakietów wyklucza możliwość pracy w trybie *on-line* i natychmiastowego blokowania transmisji po wykryciu zagrożenia.

[5] W tym miejscu warto przypomnieć atak typu Ping of Death, bazujący na sztucznie preparowanych pakietach o długości powyżej tej granicy.

Niektóre z aplikacji sieciowych charakteryzują się tym, że przesyłane, w ramach danej usługi, pakiety nie mają identycznej długości, niemniej rozkład długości w zbiorze pakietów jest określony. W takim przypadku, ze względu na procesy multipleksowania pakietów z różnych strumieni, w procesie analizy konieczne jest wstępne wyodrębnienie zbioru pakietów, który ma podlegać analizie. Można to zrobić na podstawie adresów IP i numerów portów – analizie rozkładu powinien być poddany zbiór pakietów o identycznych adresach i numerach portów.

3.3. Długości pakietów w wybranych zastosowaniach

3.3.1. Telefonia IP

W telefonii IP używane są dwa rodzaje protokołów: sygnalizacyjne i czasu rzeczywistego. W niektórych systemach obie funkcje realizowane są przez jeden protokół (przykładem jest protokół IAX). W większości przypadków używa się oddzielnych protokołów do sygnalizacji i transmisji mowy. Wielkości pakietów transmitowanych w ramach procesów sygnalizacji są zależne od protokołu sygnalizacyjnego (kilkadziesiąt bajtów w protokole H.323 i kilkaset bajtów w SIP) oraz ilości przesyłanych, w ramach negocjacji, parametrów. Natomiast wielkość pojedynczego pakietu przesyłanego w ramach rozmowy telefonicznej jest uzależniona przede wszystkim od zastosowanej metody kodowania i kompresji mowy (inaczej mówiąc od tego, czy użyty został kodek falowy czy predykcyjny).

W przypadku użycia najprostszego kodeka falowego, np. PCM, głos jest przesyłany w postaci strumienia pakietów o identycznych wielkościach[6]. Wielkość ta zależna jest od długości czasu próbkowania, reprezentowanego przez pojedynczy pakiet. Przykładowo, w typowej aplikacji VoIP, pojedynczy pakiet IP zawiera próbki sygnału analogowego, odpowiadającego 20 ms dźwięku[7]. W takim przypadku pole danych tego pakietu zawiera łącznie 180 bajtów: 160 bajtów próbek + 8 bajtów nagłówka UDP + 12 bajtów nagłówka RTP.

Kodeki liniowo-predykcyjne (ang. *linear-predictive*), takie jak G.728, G.729, charakteryzują się znacznie mniejszymi wymaganiami, związanymi z przepustowością kanału komunikacyjnego, a tym samym krótszymi pakietami. Zależnie od wersji kodeka, pojedynczy pakiet z zakodowanym dźwiękiem zawiera od 50 do 80 bajtów w polu danych. Długość pakietów przy transmisji mowy nie zmienia się, niemniej stosowanie systemu wykrywania aktywności rozmówcy VAD (*Voice Activity Detection*) powoduje, że w strumieniu pakietów mogą pojawić się pakiety o znacznie mniejszej długości, zawierające zamiast kilkudziesięciu bajtów zakodowanego głosu tylko 2 bajty danych dla generatora CNG (*Comfort Noise Generator*).

[6] Sporadycznie przesyłane są pakiety o innej długości, zawierające komunikaty kontrolne RTCP.

[7] To, że kolejne pakiety są transmitowane co 20 ms, też można wykorzystać do analizy realizowanej usługi, niemniej występowanie *jittera* powoduje, że odstępy między kolejnymi pakietami w węźle analizującym mogą być zmienne.

3.3.2. Gry komputerowe

Na podstawie długości pojedynczego pakietu trudno jest odpowiedzieć na pytanie, czy dany pakiet jest transmitowany w ramach komputerowej gry internetowej. Niemniej, jak wykazano w [3, 6], sesje komunikacyjne gier komputerowych mają dające się zidentyfikować charakterystyki strumieni pakietów. Dla różnych gier określić można oddzielnie rozkłady długości pakietów dla transmisji w obu kierunkach (od klienta do serwera i od serwera do klienta). Wartość średnia długości pakietu w transmisji do serwera wynosi zwykle kilkadziesiąt bajtów, w transmisji od serwera jest większa.

3.3.3. Systemy typu peer-to-peer

Inną grupę usług dających się zidentyfikować na podstawie wielkości transmitowanych pakietów stanowią systemy wymiany plików typu *peer-to-peer*. Dla konkretnych systemów tego typu (np. eMule, Bittorent) znane są rozkłady wielkości pakietów oraz średnie długości pakietów [6].

3.3.4. Inne przykłady

Znane są długości pakietów przesyłanych podczas niektórych ataków sieciowych. Przykładem jest popularny *Ping of Death*, w którym pole danych pakietu IP ma dokładnie 56 bajtów.

3.4. Podsumowanie

Cechą charakterystyczną niektórych usług sieciowych jest wielkość jednostki transmisyjnej, a tym samym wielkość pakietu IP. Podczas realizacji danej usługi wielkość ta może być stała, może się charakteryzować znanym rozkładem długości lub może się zawierać w określonych granicach (tab. 2).

Tabela 2

Wielkości pola danych pakietów IP w przykładowych zastosowaniach

Zastosowanie	Wielkość pola danych pakietu IP [B]	Uwagi
Gry internetowe	40-110	Na podstawie [3]. Długość zależna od gry, kierunku transmisji
VoIP z kodekiem typu LP np. G.728. G.729	50-80	Ilość danych zależna od konkretnego kodeka
Ping (także *Ping of Death*)	56	Komunikat ICMP typu *query*
World of Worldcraft	74	Dominująca wielkość pakietu, wg [6]
Skype	84	Dominująca wielkość pakietu, wg [6]
VoIP z kodekiem PCM	180, 260	
Bittorrent	377	Dominująca wielkość pakietu, wg [6]
eMule	1180	Dominująca wielkość pakietu, wg [6]

Analiza długości transmitowanych pakietów (a także rozkładu tych długości) może być wykorzystana jako dodatkowy czynnik klasyfikowania i ewentualnie blokowania transmisji pojedynczych pakietów lub ich strumieni.

4. Przykładowe zastosowanie

Generalnie analiza ruchu w sieci może być użyta do wykrywania różnorodnych naruszeń bezpieczeństwa. W tym, do wykrywania: ataków z zewnątrz, działania programów szkodliwych, awarii sprzętu i błędów oprogramowania, wewnętrznych i zewnętrznych kradzieży usług systemu informatycznego.

Pewnego rodzaju nadużycie zasobów ma miejsce wtedy, gdy uprawniony do pracy w systemie komputerowym użytkownik wykorzystuje ten system do celów prywatnych – w takim przypadku mamy do czynienia z kradzieżą usług (czasu pracy komputera, pojemności pamięci dyskowej, pasma transmisyjnego sieci lokalnej i rozległej). Kradzież usług (ang. *cyberslacking*) jest zjawiskiem powszechnie występującym. Nadużyć takich dopuszczają się wszystkie grupy pracowników, a straty są szacowane na miliardy dolarów rocznie w skali globalnej. Kradzież usług może być dokonana zarówno przez legalnego użytkownika danego systemu informatycznego, jak i przez intruza z zewnątrz. Z kradzieżą usług możemy mieć do czynienia w następujących, przykładowych sytuacjach:

- korzystanie z gier komputerowych (w tym gier dostępnych w Internecie),
- przesyłanie listów elektronicznych o charakterze prywatnym oraz rozmowy telefoniczne (w tym z użyciem VoIP) na tematy prywatne,
- korzystanie z usług typu *peer-to-peer* do pobierania plików audio, wideo, graficznych oraz obrazów płyt CD/DVD/BD,
- korzystanie z tzw. mediów strumieniowych, w tym radia internetowego i telewizji internetowej,
- przeglądanie stron/portali internetowych, np. społecznościowych, które nie mają związku ze służbowymi obowiązkami pracownika.

W celu zmniejszenia skali problemu stosuje się takie rozwiązania, jak oprogramowanie monitorujące czynności wykonywane przez użytkowników oraz serwery proxy, blokujące dostęp do określonych zasobów Internetu (w tym np. portali społecznościowych, serwerów gier *online*). Należy dodać, że wewnętrzne regulacje w zakresie dozwolonego użycia zasobów sprzętowo-programowych systemu informatycznego firmy powinny być szczegółowo opisane w polityce bezpieczeństwa.

W typowym przypadku kradzież usług wiąże się z wykorzystaniem innych protokołów i aplikacji sieciowych aniżeli tych, które są charakterystyczne dla normalnej pracy. Tak, więc wykrywanie zdarzeń tego typu jest możliwe także z wykorzystaniem technik opartych na analizie ruchu sieciowego. Długości, a także rozkłady długości pakietów transmitowanych w ramach realizacji usług sieciowych, zgodnych z polityką bezpieczeństwa są inne, aniżeli rozkłady wielkości pakietów transmitowanych w przypadku, gdy użytkownik nie przestrzega tej polityki. Wykrycie podczas analizy rozkładów pakietów o długościach charakterystycznych dla gier internetowych, systemów typu *peer-to-peer* może świadczyć o naruszeniu polityki bezpieczeństwa przez danego użytkownika.

5. Podsumowanie

Proponowana w artykule koncepcja analizy i klasyfikowania ruchu (identyfikacji używanego protokołu, aplikacji lub usługi sieciowej) na podstawie wielkości pakietu IP rozwiązuje pewne problemy klasyfikacji na podstawie numerów portów, w tym problemy związane z niestandardowym wykorzystaniem portów predefiniowanych, translacją NAT, szyfrowaniem zawartości pakietów. Wśród przykładowych zastosowań proponowanej metody wymienić można wykrywanie przypadków kradzieży usług, a więc wykorzystywania usług sieciowych niezgodnych z przyjętą w danej instytucji polityką bezpieczeństwa.

BIBLIOGRAFIA

1. Bace R. G.: Intrusion Detection, Macmillan Technical Publishing, Indianapolis, 2000.
2. Huang N.-F., Jai G.-Y., Chao H.-C.: A high accurate machine-learning algorithm for identifying application traffic in early stage, [w:] Proceedings of the IEEE ICC, 2008.
3. Joyce S.: Traffic on the Internet – Report, http://wand.cs.waikato.ac.nz/old/wand/publications/sarah-420.pdf, 2000.
4. McGregor A., Hall M., Lorier P., Brunskill J.: Flow clustering using machine learning techniques. W: Proceedings of the fifth passive and active measurement workshop (PAM 2004), March 2004.
5. Moore A., Zuev D.: Internet traffic classification using Bayesian analysis techniques, [w:] Proceedings of ACM international conference on measurement and modeling of computer systems (SIGMETERICS), 2005.

6. Ying-Dar L. i inni: Application classification using packet size distribution and port association, Journal of Network and Computer Applications, Vol. 32, Issue 5, September 2009, s. 1023÷1030.

Recenzenci: Dr hab. inż. Andrzej Kwiecień, prof. Pol. Śląskiej
Dr inż. Mirosław Skrzewski

Wpłynęło do Redakcji 7 marca 2011 r.

Abstract

The paper discusses traffic analysis problem. First part deals with some problems and limitations of analysis with a use of port numbers. Main part of the paper presents a concept of IP packet size based traffic classification. Two concepts are discussed. First one is based on packet sizes distribution in a given set of packets. The second one is based on a single packet size. Last part of the paper presents some remarks on application of the concept for cyberslacking detection.

Adres

Tomasz BILSKI: Politechnika Poznańska, Instytut Automatyki i Inżynierii Informatycznej, pl. Skłodowskiej-Curie 5, 60-965 Poznań, Polska, tomasz.bilski@put.poznan.pl