BỘ DỮ LIỆU DẠNG NETFLOW DÙNG TRONG PHÁT HIỆN XÂM NHẬP TRÁI PHÉP VÀ ỨNG DỤNG

Nguyễn Hoàng Giang*, Trần Quang Anh+
*Cục Công nghệ thông tin & Thống kê Hải quan
+ Học Viện Công Nghệ Bưu Chính Viễn Thông

Tóm tắt: Các bộ dữ liệu mẫu về xâm nhập trái phép trong mạng máy tính hiện đã và đang được ứng dụng rất rộng rãi trong việc nghiên cứu phát hiện xâm nhập mạng trái phép. Trên thế giới đã có nhiều bộ dữ liệu khác nhau, mỗi bộ dữ liệu có ưu, nhược điểm khác nhau. Bộ dữ liệu dạng Netflow có nhiều ưu điểm trong việc phát hiện xâm nhập trái phép, đặc biệt trong mạng có lưu lượng dữ liệu lớn. Hiện tại, bộ dữ liệu của DARPA vẫn đang được các nhà khoa học sử dụng trong nghiên cứu phát hiện xâm nhập trái phép, tuy nhiên bộ dữ liệu DARPA không ở dạng Netflow. Mục tiêu của bài báo này trình bày một phương thức xây dựng bộ dữ liệu dạng Netflow từ nguồn dữ liệu DARPA; và ứng dụng bộ dữ liệu này trong phát hiện xâm nhập trái phép bằng phương pháp học máy. Bộ dữ liệu này có thể được sử dụng rộng rãi trong nghiên cứu phát hiện xâm nhập trái phép dựa trên Netflow.

Từ khóa: Bộ dữ liệu (dataset), Naïve Bayes, Netflow, phát hiện xâm nhập trái phép (IDS).

I. GIỚI THIỆU

Ngày nay, mạng máy tính thường xuyên là các mục tiêu tấn công của tin tặc nhằm mục đích ăn cắp dữ liệu bí mật quan trọng của tổ chức hoặc làm dừng hệ thống cung cấp dịch vụ của tổ chức. Để phát hiện và ngăn chặn các cuộc tấn công này, có rất nhiều các giải pháp phần cứng cũng như phần mềm ra đời. Các giải pháp đó có thể là IDS (Intrusion Detection Systems), IPS (Intrusion Prevention Systems), IDP (Intrusion Detection Prevention Systems), Firewall, hoặc hệ thống giám sát. Để nghiên cứu, cho ra đời các giải pháp, công nghệ về IDS, IPS, IDP... rất cần thiết phải có các bộ dữ liệu mẫu về xâm nhập trái phép để thực hiện việc huấn luyện và kiểm thử.

Netflow là một giao thức do hãng Cisco phát triển vào những năm 1996, được phát triển thành một công nghệ giám sát lưu lượng mạng.

Hiện nay, Netflow đã được xây dựng thành tiêu chuẩn và sử dụng hầu hết trong các thiết bị mạng Router của Cisco, Juniper, Extreme, Habour... Netflow đã được phát triển qua nhiều phiên bản: version 1 đến version 10; trong đó thông dụng nhất hiện nay là version 5, version 7 và version 9. Netflow cho phép thực hiện giám sát, phân tích, tính toán lưu lượng gói. Một trong các ưu điểm của Netflow so với các giao thức khác là nó cho phép định danh và phân loại những loại tấn công như DoS, DDoS, Worm... theo thời gian thực dựa vào những sự hành vi thay đổi bất thường trong mạng, đặc biệt trong mạng có lưu lượng lớn. Do vậy, việc xây dựng một bộ dữ liệu Dataset dạng Netflow là cần thiết để có thể tận dụng được hết các ưu điểm của giao thức này.

Tác giả liên hệ: Nguyễn Hoàng Giang,
email: giangnh@customs.gov.vn.

Đến tòa soạn: 28/3/2016, chỉnh sửa: 08/5/2016, chấp nhận đăng: 30/5/2016.

Bảng I. Tổng hợp tập dữ liệu trong các nghiên cứu về IDS dựa trên thống kê

Tác giả	Năm công bố	Định dạng dữ liệu	Tập dữ liệu sử dụng	Phương pháp thực hiện
Eskin	2000	Packet-based	DARPA99	Probability Model
Manikopoulos and Papavassilou	2002	Packet-based	Real-life	Statistical model with neural network
Mahoney and Chan	2003	Packet-based	DARPA99	LERAD algorithm
Chan et al	2003	Packet-based	DARPA99	Learning rules
Wang and Stolfo	2004	Packet-based	DARPA99	Payload-based algorithm
Song et al	2007	Packet-based	KDDCUP99	Gaussian mixture model
Chhabra et al	2008	Packet-based	Real-time	FDR method
Lu and Ghorbani	2009	Packet-based & Flow-based	DARPA99	Wavelet analysis
Wattenberg et al	2011	Packet-based	Real-time	GLRT model
Yu	2012	Packet-based	Real-time	Adaptive CUSUM

Bảng II. Tổng hợp tập dữ liệu trong các nghiên cứu về IDS dựa trên phân loại

Tác giả	Năm công bố	Định dạng dữ liệu	Tập dữ liệu sử dụng	Phương pháp thực hiện
Tong et al	2005	Packet-based	DARPA99, TCPSTAT	KPCC model
Gaddam et al	2007	Packet-based	NAD, DED, MSD	K-means + ID3
Khan et al	2007	Packet-based	DARPA98	DGSOT + SVM
Das et al	2008	Packet-based	KDDCUP99	APD algorithm
Lu and Tong	2009	Packet-based	DARPA99	CUSUM – EM
Quadeer et al	2010	Packet-based	Real-time	Traffic statistics
Wagner et al	2011	Flow-based	Flow Traces	Kernel OCSVM
Muda et al	2011	Other	KDDCUP99	KMNB algorithm
Kang et al	2012	Packet-based	DARPA98	Differentiated SVĐ

Để xây dựng được một bộ dữ liệu phục vụ cho nghiên cứu đòi hỏi phải thực hiện rất nghiêm túc và tốn thời gian. Đó là phải thiết lập được môi trường mạng, cài đặt phần mềm, có hiểu biết và biết sử dụng các công cụ để thực hiện tấn công thực tế, bắt giữ và đánh nhãn gói tin trên mạng để hình thành bộ dữ liệu. Trên thế giới hiện nay tồn tại một số bộ dữ liệu nổi tiếng như DARPA, KDD-99, ISCX... Tuy vậy, các bộ dữ liệu này tồn tại ở dạng Tcpdump, không phải ở dạng Netflow nên không ứng dụng được trong nghiên cứu về IDS trên Netflow. Các bộ dữ liệu ở dạng Netflow rất ít, nếu có thì hoặc không đầy đủ (như bộ UT) hoặc chưa hoàn chỉnh (như bộ dữ liệu được công bố [8], chỉ xây dựng bộ dữ liệu Netflow cho một loại tấn công). Theo tổng hợp [11], các công trình nghiên cứu về IDS sử dụng phương pháp học máy (học máy dựa trên thống kê và học máy dựa trên phân loại được trình bày trong Bảng I và II) hiện nay phần lớn đều sử dụng định dạng dữ liệu là Packet-based. Điều này có nghĩa là hiện nay chưa có, hoặc có rất ít các bộ dữ liệu định dạng Netflow được công bố để phục vụ mục đích nghiên cứu về IDS.

Trên cơ sở những phân tích, lập luận trên, nhóm tác giả đã xác định mục tiêu của bài báo này là thực hiện xây dựng một bộ dữ liệu dạng Netflow hoàn chỉnh trên cơ sở bộ dữ liệu DARPA nổi tiếng và ứng dụng trong phát hiện xâm nhập trái phép.

Phần còn lại của bài báo được chia thành các mục sau: Mục II giới thiệu các bộ dữ liệu hiện có đã được công bố rộng rãi; Mục III trình bày về phương pháp và quá trình xây dựng bộ dữ liệu; Mục IV thực hiện mô tả về các bộ dữ liệu đã xây dựng được; Mục V trình bày về ứng dụng của bộ dữ liệu trong phương pháp

học máy để phát hiện xâm nhập trái phép đối với một loại xâm nhập; cuối cùng là phần kết luận và hướng nghiên cứu trong tương lai.

II. CÁC BỘ DỮ LIỆU DÙNG TRONG PHÁT HIỆN XÂM NHẬP TRÁI PHÉP

A. Dữ liệu DARPA

Bộ dữ liệu DARPA hình thành do Cục dự án nghiên cứu cao cấp Bộ quốc phòng Mỹ (Defense Advanced Research Project Agency) tài trợ đề tài xây dựng cơ sở dữ liệu mẫ xâm nhập trái phép tại Phòng thí nghiệm Lincoln, Đại học MIT [1]. Để xây dựng tập dữ liệu này, các nhà khoa học đã lấy dữ liệu của một mạng quân sự Mỹ khi hoạt động bình thường làm dữ liệu bình thường; sau đó đưa thêm các dữ liệu xâm nhập trái phép vào trong tập dữ liệu đó. Cách làm trên cho phép biết được chắc chắn đâu là dữ liệu bình thường, đâu là dữ liệu xâm nhập trái phép.

Mỗi dữ liệu của DARPA bao gồm dữ liệu mạng và dữ liệu máy chủ tương ứng. Dữ liệu mạng được thu thập và lưu trữ ở dạng Tcpdump. Dữ liệu máy chủ được lưu giữ ở dạng BSM (Basic Security Module). Tập dữ liệu bao gồm dữ liệu thu thập trong vòng 5 tuần. Đi kèm với dữ liệu là tài liệu mô tả dữ liệu khá chi tiết, bao gồm loại xâm nhập, thời gian bắt đầu, thời gian kết kết, địa chỉ máy tấn công, địa chỉ máy bị tấn công đối với mỗi sự kiện xâm nhập trái phép. Toàn bộ dữ liệu có kích thước khoảng 10Gb, trong đó gồm 54 loại xâm nhập được phân làm 4 nhóm: R2L (Remote to Local – là nhóm các xâm nhập cho phép kẻ tấn công từ xa lấy được quyền của người dung máy chủ), U2R (User to Root – là nhóm các xâm nhập cho phép người dùng bình thường trên máy chủ có thể đoạt quyền quản trị root), DoS (Denial of Service – là nhóm tấn công từ chối dịch vụ, phá hoạt tính sẵn sàng của hệ thống), Probe (là nhóm tấn công do thám, ảnh hưởng đến tính bảo mật của hệ thống, đồng thời cung cấp các thông tin cần thiết để tiến hành các bước tấn công tiếp theo. Các hình thức xâm nhập trái phép được thể hiện trong bảng sau:

Bảng III. Các nhóm xâm nhập trái phép trong dữ liệu DARPA

Nhóm	Tên loại tấn công
R2L	Dictionary, Ftpwrite, Guest, Httptunnel, Imap, Named, ncftp, netbus, netcat, Phf, ppmacro, Sendmail, sshtrojan, Xlock, Xsnoop
U2R	anypw, casesen, Eject, Ffbconfig, Fdformat, Loadmodule, ntfsdos, Perl, Ps, sechole, Xterm, yaga
DoS	Apache2, arppoison, Back, Crashiis, dosnuke, Land, Mailbomb, SYN Flood (Neptune), Ping of Death (POD), Process table, selfping, Smurf, sshprocesstable, Syslogd, tcpreset, Teardrop, UDPstorm
Probe	insidesniffer, Ipsweep, ls_domain, Mscan, NTinfoscan, Nmap, queso, resetscan, Saint, Satan

Nhược điểm lớn nhất của bộ dữ liệu DARPA là được thu thập và lưu giữ ở dạng Tcpdump, có kích thước lớn.

B. Dữ liệu KDD-99

Như đã đề cập ở Mục II.A, dữ liệu DARPA do lưu ở dạng Tcpdump. Nên để có thể sử dụng để đánh giá các phương pháp, thuật toán, dữ liệu này cần thông qua một quá trình xử lý ban đầu, bao gồm: Định nghĩa các sự kiện, lựa chọn đặc trưng của các sự kiện, sau đó trích rút đặc trưng và lưu các dự kiện dưới dạng các vector. Như vậy, các phương pháp xử lý ban đầu khác nhau có thể cho các định nghĩa khác nhau về sự kiện hay các đặc trưng khác nhau, từ đó dẫn đến khó khăn trong việc so sánh, phân tích các thuật toán xâm nhập trái phép. Vì thế, với sự tài trợ của DARPA, hội nghị về khai pháp dữ liệu và phát triển tri thức năm 1999 (Knowledge Discovery and Data Mining 1999 – viết tắt là KDD -99) đã thực hiện quá trình xử lý ban đầu đối với tập dữ liệu của Darpa và cho ra tập dữ liệu KDD-99 [2]. Dữ liệu KDD-99 đã định nghĩa sự kiện dựa trên nền tảng của kết nối TCP/IP: Mỗi sự kiện bao gồm các hoạt động mạng sinh ra khi một máy chủ kết nối với một máy chủ khác, và các hoạt động bên trong máy chủ bị kết nối đó trong thời gian kết nối.

Tập dữ liệu KDD-99 được phân thành hai tập dữ liệu: Tập dữ liệu huấn luyện và tập dữ liệu thử nghiệm. các nhóm dữ liệu trong tập dữ liệu KDD-99 giống như trong bảng I, ngoài ra còn thêm nhóm dữ liệu NORMAL là các dữ liệu

bình thường. Phân bố dữ liệu theo nhóm trong tập dữ liệu KDD-99 được trình bày trong bảng sau.

Bảng IV. Phân bố dữ liệu theo nhóm trong tập huấn luyện

Nhóm	Số lượng	Phần trăm (%)
R2L	1.126	0.023
U2R	52	0.001
DoS	3.883.370	79.278
Probe	4.102	0.839
NORMAL	972.781	19.859

Bảng V. Phân bố dữ liệu theo nhóm trong tập thử nghiệm

Nhóm	Số lượng	Phần trăm (%)
R2L	14.745	4.738
U2R	246	0.079
DoS	231.455	74.374
Probe	14.166	1.339
NORMAL	60.593	19.47

Theo các bảng nêu trên, chúng ta để ý thấy số lượng cũng như tỷ lệ % của nhóm xâm nhập DoS và Probe rất lớn. Điều này không có nghĩa là các nhóm DoS và Probe xảy ra nhiều mà là do KDD-99 định nghĩa sự kiện dựa trên kết nối TCP/IP. Thông thường mỗi đợt tấn công DoS và Probe thường sinh ra rất nhiều kết nối, vì vậy trong tập dữ liệu KDD-99 mỗi kết nối TCP/IP được xem như một sự kiện.

Mỗi dữ liệu trong KDD-99 được trích rút thành 41 đặc trưng, gồm 4 phần: Phần thứ nhất (từ đặc trưng 1 đến 9) là các đặc trưng cơ bản của kết nối TCP/IP; Phần thứ hai (từ đặc trưng 10 đến 22) là các đặc trưng của máy chủ bị kết nối; Phần thứ ba (từ đặc trưng 23 đến 31) là các đặc trưng về lưu lượng trong khoảng thời gian 2 giây; Phần thứ tư (từ đặc trưng 32 đến 41) là các đặc trưng về lưu lượng trong khoảng thời gian 256 giây.

C. Dữ liệu ISCX

Information Security Centre of Excellence (ISCX) là một trung tâm nghiên cứu về an toàn thông tin của trường đại học New Brunswick (UNB) – Canada. Xuất phát từ yêu cầu nghiên cứu hệ thống IDS đòi hỏi phải có một bộ Dataset chính xác, đầy đủ, ISCX đã xây dựng một mô hình mạng, mô phỏng các cuộc tấn công trong mạng dựa trên các giao thức HTTP, SMTP, SSH, IMAP, POP3 và FTP. Những luồng dữ liệu thông thường và bất thường được bắt giữ và được đánh dấu. Bộ dữ liệu này đã được giới thiệu bởi Ali Shiravi, Hadi Shiravi, Mahbod Tavallaee, Ali A. Ghorbani tại bài báo "Toward developing a systematic approach to generate benchmark datasets for intrusion detection, Computers & Security, Volume 31, Issue 3, May 2012, Pages 357 -374, ISSN 0167-4048, 10.1016/j.cose.2011.12.012.(http://www.sciencedirect.com/science/article/pii/S0167404811001672).

Bộ dữ liệu UNB ISCX 2012 IDS [9] bao gồm dữ liệu thu thập trong vòng 7 ngày, gồm cả dữ liệu thông thường và bất thường, cụ thể:

Bảng VI. Bộ dữ liệu ISCX

Thứ	Ngày	Mô tả	Kích thước dữ liệu (GB)
6	11/6/2010	Dữ liệu thông thường	16.1
7	12/6/2010	Dữ liệu thông thường	4.22
Chủ nhật	13/6/2010	Infiltrating the network from inside và dữ liệu thông thường	3.95
2	14/6/2010	HTTP Denial of Service và dữ liệu thông thường	6.85
3	15/6/2010	Distributed Denial of Service using an IRC Botnet	23.4
4	16/6/2010	Dữ liệu thông thường	17.6
5	17/6/2010	Brute Force SSH và dữ liệu thông thường	12.3

Bộ dữ liệu ISCX cũng ở dạng Tcpdump.

D. Dữ liệu UT

Tập dữ liệu UT là tập do nhóm nghiên cứu tại đại học Twente của Hà Lan xây dựng theo dạng Netflow [3]. Tập dữ liệu này được xây dựng bằng phương pháp thu thập dữ liệu xâm nhập trái phép thực tế trên nguyên lý Honeypot. Nhóm nghiên cứu đã xây dựng một Honeypot – một mạng máy tính không có người

sử dụng; như vậy nếu có lưu lượng mạng phát sinh thì đó chính là lưu lượng xâm nhập.

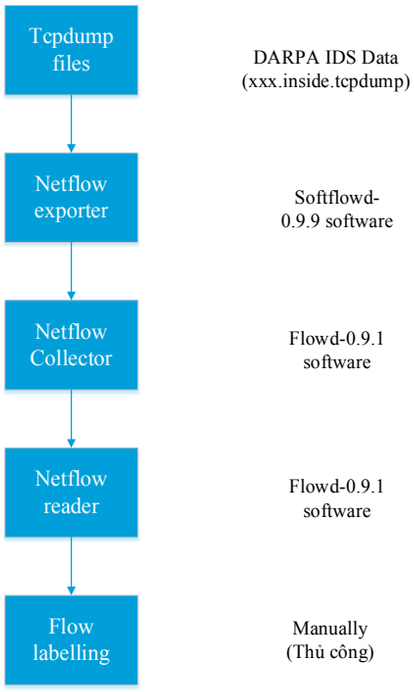
Đây là dữ liệu theo dạng Netflow, tuy nhiên khi so sánh với tập dữ liệu của DARPA và KDD-99, dữ liệu UT có một số vấn đề:

- Trong tập dữ liệu UT không có dữ liệu bình thường;
- Các dữ liệu xâm nhập trái phép được xây dựng tự động bởi Honeypot (trong khi các dữ liệu DARPA và KDD-99 được xây dựng bởi các chuyên gia về an ninh mạng);
- Các dữ liệu xâm nhập trái phép chỉ có một loại duy nhất là dữ liệu bất thường.

III. XÂY DỰNG BỘ DỮ LIỆU DẠNG NETFLOW DÙNG TRONG IDS

A. Phương pháp xây dựng

Nhóm tác giả đã thực hiện chuyển đổi dữ liệu DARPA thành dữ liệu dạng Netflow theo sơ đồ như Hình 1.

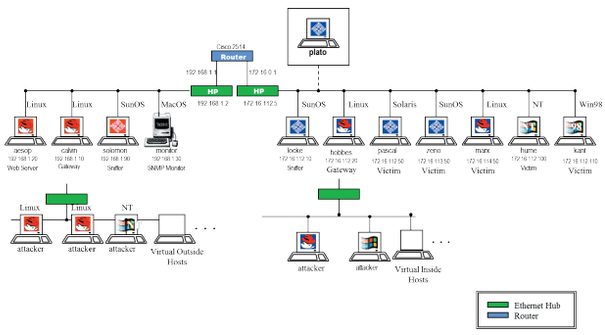


Hình 1. Sơ đồ chuyển đổi từ dữ liệu Tcpdump sang Netflow

Dữ liệu đầu vào của bộ chuyển đổi này là dữ liệu ở định dạng Tcpdump (bộ dữ liệu của DARPA). Dữ liệu này được truyền đến Module Netflow exporter. Module Netflow exporter thực hiện đọc dữ liệu Tcpdump, sau đó trích rút ra các flow, tạo ra các gói tin theo chuẩn Netflow (v5, v7, v9) và gửi đến Module Netflow collector. Module Netflow collector thu thập các gói tin Netflow và lưu dữ liệu Netflow này vào bộ nhớ (ổ cứng). Module Netflow reader sẽ đọc các dữ liệu Netflow từ bộ nhớ và hiển thị theo yêu cầu của người dùng. Sau đó, nhóm tác giả sẽ căn cứ vào tài liệu mô tả các cuộc tấn công trái phép của DARPA để tiến hành đánh nhãn bằng tay các flow xâm nhập trái phép và các flow bình thường. Sau quá trình này, chúng ta đã thu thập được bộ dữ liệu dạng Netflow đầy đủ từ bộ dữ liệu DARPA.

B. Quá trình thực hiện

Dữ liệu đầu vào của hệ thống chuyển đổi chính là các file dữ liệu Tcpdump (inside.Tcpdump) trong tập dữ liệu DARPA. Dữ liệu Tcpdump này được thu thập bằng cách bắt các gói tin trong mạng nội bộ (mạng mô phỏng xâm nhập trái phép DARPA). Về lý thuyết, nó bao gồm toàn bộ lưu lượng mạng đến và đi từ tất cả máy chủ bên trong mạng. Tuy nhiên, do có sự cố trong quá trình thực hiện nên không có dữ liệu Tcpdump của ngày Thứ 3 (Tuesday) của Tuần 4.



Hình 2. Sơ đồ mạng mô phỏng xâm nhập trái phép DARPA (Phần inside là phần phía tay phải, dải mạng 172.16.0.0) [1]

Module Netflow exporter được xây dựng dựa trên phần mềm nguồn mở Softflowd phiên bản 0.9.9 [4]. Sau khi cài đặt và chạy, Softflowd đọc file dữ liệu ở dạng Tcpdump, sau đó sinh ra các gói tin Netflow theo version thiết lập. Ở đây, chúng tôi sử dụng phiên bản version 9, đây là phiên bản mới

nhất hiện nay mà Cisco công bố. Khi chạy phần mềm softflowd, nảy sinh một vấn đề đó chính là thời gian bắt đầu và thời gian kết thúc của flow thu thập được lại chính là thời gian tham chiếu tới thời gian hiện tại của máy chủ cài đặt phần mềm softflowd, chứ không phải thời điểm năm 1999 khi dữ liệu Tcpdump được thu thập. Điều này dẫn đến sai số về timestamp khi thu thập các file Netflow, mà vấn đề thời gian là vấn đề rất quan trọng đối với phương pháp chuyển đổi này, vì từ nhãn thời gian mới có thể đánh nhãn thủ công các xâm nhập trái phép đã được công bố bởi DARPA. Để giải quyết vấn đề này, chúng tôi đã phải tham chiếu lại thời gian thực hiện của DARPA, thiết lập giờ của máy chủ về thời điểm năm 1999 gần thời điểm DARPA thực hiện thu thập dữ liệu. Tuy vậy, vẫn còn sai số trong thu thập thời gian bắt đầu, thời gian kết thúc. Việc này lại phải thực hiện hiệu chỉnh bằng tay, với độ chính xác đến từng giây.

Module Netflow collector được xây dựng dựa trên phần mềm mã nguồn mở Flowd phiên bản 0.9.1 [4]. Module này thu thập các flow và lưu trữ trong bộ nhớ (ổ cứng) để sử dụng cho các bước tiếp theo.

Module Netflow reader là một cấu phần nằm trong bộ phần mềm mã nguồn mở Flowd. Module này có nhiệm vụ đọc dữ liệu mà Module Flowd đã thu thập và lưu trữ trong bộ nhớ. Nó thực hiện đọc các trường trong Netflow. Mặc dù Netflow có rất nhiều trường, tuy nhiên, chúng ta quan tâm tới một số trường quan trọng, được sử dụng trong phát hiện xâm nhập trái phép. Bao gồm:

- Source IP, source port;
- Destination IP, destination port;
- Protocol, flag;
- Packets; Octets;
- Flow-start, flow-finish.

Vì bộ sniffer thu thập dữ liệu DARPA đặt trong cùng dải mạng có nhiệm vụ thu thập thông tin của các máy chủ victim bị tấn công, nên trong dữ liệu inside.Tcpdump thu thập được chứa đựng tất cả các luồng dữ liệu đến, đi các máy chủ victim. Để thuận tiện cho việc thao tác đối với dữ liệu của từng máy chủ victim, cũng như thuận tiện cho việc đánh nhãn sau này, chúng tôi thực hiện chỉnh sửa đoạn mã cấu hình trong file cấu hình của phần mềm Flowd để thực hiện thu thập dữ liệu Netflow cho từng máy chủ victim. Kết quả, chúng tôi đã thu thập được 4 bộ dữ liệu Netflow cho 4 máy chủ victim là pascal (172.16.112.50), zeno (172.16.113.50), marx (172.16.114.50) và hume (172.16.112.100).

Bước cuối cùng, chúng tôi đã thực hiện đánh dấu bằng phương pháp thủ công các flow xâm nhập trái phép dựa theo tài liệu công bố, mô tả của DARPA. Quá trình đánh dấu thực hiện dựa trên thời gian bắt đầu, thời gian kết thúc, địa chỉ IP nguồn, địa chỉ IP đích, cổng dịch vụ đích. Việc đánh dấu cho các luồng dữ liệu khá dễ dàng bằng việc sử dụng tài liệu mô tả của DARPA kết hợp các công cụ lọc (filter) theo từng thuộc tính (địa chỉ IP đích, cổng dịch vụ đích); hơn nữa các flow dữ liệu thuộc mỗi loại tấn công thường liên tục và có dấu hiệu tương đối giống nhau. Chính vì vậy, việc đánh dấu cho các luồng dữ liệu rất nhanh và có độ chính xác cao. Chỉ có một số rất ít trường hợp do nhiều lý do khách quan (sai lệch thời gian milisecond trong quá trình chuyển đổi) và chủ quan (do ghi nhận chưa chính xác trong tài liệu mô tả của DARPA), chúng tôi nhận thấy có một số chỗ không thống nhất về thời gian nhãn tấn công. Lưu ý: một cuộc tấn công có thể bao gồm nhiều flow, nhưng mỗi flow chỉ thuộc về một cuộc tấn công nhất định. Đó chính là cơ sở để có thể đánh dấu các flow là xâm nhập trái phép hay bình thường.

C. Kết quả chuyển đổi và so sánh với tập dữ liệu gốc

Bảng sau sẽ thực hiện so sánh một số thông số giữa bộ dữ liệu gốc dạng Tcpdump và bộ dữ liệu chuyển đổi Netflow:

Bảng VI. Bảng so sánh thông số của 02 bộ dữ liệu

Tuần	(Kích thước file (byte		Số lượng Packets trong Tcpdump	Số lượng Flow trong Netflow
	Tcpdump	Netflow		
Week1	1.929.080.092	160.344.163	7.810.861	342.837
Week2	1.613.234.838	193.322.991	7.199.540	394.623
Week3	2.215.279.595	165.561.281	8.912.974	316.613
Week4	1.571.862.354	134.873.898	7.655.034	310.053
Week5	3.413.554.375	299.771.599	14.299.343	511.289
Tổng cộng	10.743.011.254	953.873.932	45.877.752	1.875.415

Từ bảng so sánh một số thông số của hai bộ dữ liệu (Tcpdump và Netflow), chúng ta có thể nhận thấy:

- Kích thước bộ dữ liệu Netflow giảm đi rất nhiều lần so với kích thước bộ dữ liệu Tcpdump (khoảng 1/10 lần);
- Số lượng dữ liệu cần xử lý của bộ dữ liệu Netflow cũng giảm hơn rất nhiều lần so với bộ dữ liệu Tcpdump, cụ thể chỉ khoảng 1.875.415 flows so với 45.877.752 packets (tức là giảm còn khoảng 1/20 lần);
- Số lượng các trường dữ liệu trong một flow cũng ít hơn nhiều so với các trường dữ liệu trong một packet.

Do đó, việc thao tác, xử lý dữ liệu trên bộ dữ liệu Netflow này bằng phương pháp học máy sẽ dễ dàng, nhanh chóng hơn nhiều so với trên bộ dữ liệu Tcpdump. Chính vì thế, dữ liệu Netflow sẽ phù hợp hơn trong các mạng máy tính có lưu lượng lớn, đòi hỏi thời gian xử lý nhanh.

IV. MÔ TẢ CÁC TẬP DỮ LIỆU NETFLOW DARPA

Như đã đề cập ở Mục III.B, chúng tôi đã thu thập và phân tách được 4 bộ dữ liệu Netflow tương ứng với 4 máy chủ victim. Trong phạm vi bài báo này, chúng tôi trình bày thông số cơ bản của từng bộ dữ liệu Netflow của các máy chủ Pascal (172.16.112.50), zeno (172.16.113.50), marx (172.16.114.50) và hume (172.16.112.100), được trình bày chi tiết như sau:

A. Bộ dữ liệu cho máy chủ pascal

Bảng VII. Các thông số cơ bản của bộ dữ liệu Netflow máy chủ Pascal

Mô tả	Giá trị
Kích thước dữ liệu DARPA ở dạng Tcpdump	Xấp xỉ 10Gb
Số lượng flow đến máy chủ Pascal	170.153
Số lượng flow tấn công vào máy chủ Pascal	29.416
Số lượng flow bình thường vào máy chủ Pascal	140.737
Số lượng flow kết nối vào cổng dịch vụ ftp của máy chủ Pascal	649
Số lượng flow tấn công vào cổng dịch vụ ftp của máy chủ Pascal	70
Số lượng flow kết nối bình thường vào cổng dịch vụ ftp của máy chủ Pascal	579
Số lượng flow kết nối vào cổng dịch vụ 22 của máy chủ Pascal	763
Số lượng flow tấn công vào cổng dịch vụ 22 của máy chủ Pascal	239
Số lượng flow kết nối bình thường vào cổng dịch vụ 22 của máy chủ Pascal	3.176
Số lượng flow kết nối vào cổng dịch vụ 23 của máy chủ Pascal	3.246
Số lượng flow tấn công vào cổng dịch vụ 23 của máy chủ Pascal	70
Số lượng flow kết nối bình thường vào cổng dịch vụ 23 của máy chủ Pascal	3.176
Số lượng flow kết nối vào cổng dịch vụ 25 của máy chủ Pascal	3.145
Số lượng flow tấn công vào cổng dịch vụ 25 của máy chủ Pascal	1.176
Số lượng flow kết nối bình thường vào cổng dịch vụ 25 của máy chủ Pascal	1.969
Số lượng flow kết nối vào cổng dịch vụ khác của máy chủ Pascal	130.787
Số lượng flow tấn công vào cổng dịch vụ khác của máy chủ Pascal	2.065
Số lượng flow kết nối bình thường vào cổng dịch vụ khác của máy chủ Pascal	128.722

Bảng VIII. Số lượng tấn công từ từng máy chủ

Máy chủ	Số lượng tấn công
206.47.98.151	501
10.20.30.40	20.480
Mạng [209.X.Y.Z]	5.108
Mạng [172.16.X.Y]	539
Khác	2.788

Bảng IX. Số lượng tấn công theo các cổng đích

Cổng	Số lượng tấn công
0	1.997
25	1.176
22	524
23	70
20	36
21	34
80	28
53	24
110	21
Các cổng khác	25.506

B. Bộ dữ liệu cho máy chủ Marx

Bảng X. Các thông số cơ bản của bộ dữ liệu Netflow máy chủ Marx

Mô tả	Giá trị
Kích thước dữ liệu DARPA ở dạng Tcpdump	Xấp xỉ 10Gb
Số lượng flow đến máy chủ Marx	184.050
Số lượng flow tấn công vào máy chủ Marx	89.830
Số lượng flow bình thường vào máy chủ Marx	94.220
Số lượng flow kết nối vào cổng dịch vụ ftp của máy chủ Marx	561
Số lượng flow tấn công vào cổng dịch vụ ftp của máy chủ Marx	117
Số lượng flow kết nối bình thường vào cổng dịch vụ ftp của máy chủ Marx	444
Số lượng flow kết nối vào cổng dịch vụ 22 của máy chủ Marx	283
Số lượng flow tấn công vào cổng dịch vụ 22 của máy chủ Marx	56
Số lượng flow kết nối bình thường vào cổng dịch vụ 22 của máy chủ Marx	227
Số lượng flow kết nối vào cổng dịch vụ 23 của máy chủ Marx	1.626
Số lượng flow tấn công vào cổng dịch vụ 23 của máy chủ Marx	65
Số lượng flow kết nối bình thường vào cổng dịch vụ 23 của máy chủ Marx	1.561
Số lượng flow kết nối vào cổng dịch vụ 25 của máy chủ Marx	1.889
Số lượng flow tấn công vào cổng dịch vụ 25 của máy chủ Marx	561
Số lượng flow kết nối bình thường vào cổng dịch vụ 25 của máy chủ Marx	1.328
Số lượng flow kết nối vào cổng dịch vụ 80 của máy chủ Marx	32.379
Số lượng flow tấn công vào cổng dịch vụ 80 của máy chủ Marx	2.030
Số lượng flow kết nối bình thường vào cổng dịch vụ 80 của máy chủ Marx	30.349
Số lượng flow kết nối vào cổng dịch vụ khác của máy chủ Marx	146.715
Số lượng flow tấn công vào cổng dịch vụ khác của máy chủ Marx	86.814
Số lượng flow kết nối bình thường vào cổng dịch vụ khác của máy chủ Marx	59.901

Bảng XI. Số lượng tấn công từ từng máy chủ

Máy chủ	Số lượng tấn công
10.20.30.40	40.960
Mạng [172.16.X.Y]	12.238
Khác	36.633

Bảng XII. Số lượng tấn công theo các cổng đích

Cổng	Số lượng tấn công
0	433
25	561
22	56
23	65
20	59
21	58
80	2.030
53	61
110	55
Các cổng khác	86.452

C. Bộ dữ liệu cho máy chủ Zeno

Bảng XIII. Các thông số cơ bản của bộ dữ liệu Netflow máy chủ Zeno

Mô tả	Giá trị
Kích thước dữ liệu DARPA ở dạng Tcpdump	Xấp xỉ 10Gb
Số lượng flow đến máy chủ Zeno	37.923
Số lượng flow tấn công vào máy chủ Zeno	1.019
Số lượng flow bình thường vào máy chủ Zeno	36.904
Số lượng flow kết nối vào cổng dịch vụ ftp của máy chủ Zeno	1.061
Số lượng flow tấn công vào cổng dịch vụ ftp của máy chủ Zeno	478
Số lượng flow kết nối bình thường vào cổng dịch vụ ftp của máy chủ Zeno	583
Số lượng flow kết nối vào cổng dịch vụ 23 của máy chủ Zeno	1.463
Số lượng flow tấn công vào cổng dịch vụ 23 của máy chủ Zeno	13
Số lượng flow kết nối bình thường vào cổng dịch vụ 23 của máy chủ Zeno	1.450
Số lượng flow kết nối vào cổng dịch vụ 25 của máy chủ Zeno	2.208
Số lượng flow tấn công vào cổng dịch vụ 25 của máy chủ Zeno	496
Số lượng flow kết nối bình thường vào cổng dịch vụ 25 của máy chủ Zeno	1.712
Số lượng flow kết nối vào cổng dịch vụ khác của máy chủ Zeno	31.191
Số lượng flow tấn công vào cổng dịch vụ khác của máy chủ Zeno	104
Số lượng flow kết nối bình thường vào cổng dịch vụ khác của máy chủ Zeno	31.087

Bảng XIV. Số lượng tấn công từ từng máy chủ

Máy chủ	Số lượng tấn công
11.21.31.41	400
Mạng [172.16.X.Y]	6
Khác	613

Bảng XV. Số lượng tấn công theo các cổng đích

Cổng	Số lượng tấn công
0	3
25	496
22	1
23	1
20	1
21	477
80	6
Các cổng khác	76

E. Bộ dữ liệu cho máy chủ Hume

Bảng XVI. Các thông số cơ bản của bộ dữ liệu Netflow máy chủ Hume

Mô tả	Giá trị
Kích thước dữ liệu DARPA ở dạng Tcpdump	Xấp xỉ 10Gb
Số lượng flow đến máy chủ Hume	294.286
Số lượng flow tấn công vào máy chủ Hume	393
Số lượng flow bình thường vào máy chủ Hume	293.893
Số lượng flow kết nối vào cổng dịch vụ ftp của máy chủ Hume	13.326
Số lượng flow tấn công vào cổng dịch vụ ftp của máy chủ Hume	25
Số lượng flow kết nối bình thường vào cổng dịch vụ ftp của máy chủ Hume	13.301
Số lượng flow kết nối vào cổng dịch vụ 23 của máy chủ Hume	399
Số lượng flow tấn công vào cổng dịch vụ 23 của máy chủ Hume	31
Số lượng flow kết nối bình thường vào cổng dịch vụ 23 của máy chủ Hume	368
Số lượng flow kết nối vào cổng dịch vụ 25 của máy chủ Hume	17.935
Số lượng flow tấn công vào cổng dịch vụ 25 của máy chủ Hume	10
Số lượng flow kết nối bình thường vào cổng dịch vụ 25 của máy chủ Hume	17.925
Số lượng flow kết nối vào cổng dịch vụ 80 của máy chủ Hume	21.664
Số lượng flow tấn công vào cổng dịch vụ 80 của máy chủ Hume	58
Số lượng flow kết nối bình thường vào cổng dịch vụ 80 của máy chủ Hume	21.606
Số lượng flow kết nối vào cổng dịch vụ khác của máy chủ Hume	240.962
Số lượng flow tấn công vào cổng dịch vụ khác của máy chủ Hume	269
Số lượng flow kết nối bình thường vào cổng dịch vụ khác của máy chủ Hume	240.693

Bảng XVII. Số lượng tấn công từ từng máy chủ

Máy chủ	Số lượng tấn công
Mạng [1.12.120.X]	48
Mạng [172.16.X.Y]	15
Khác	330

Bảng XVIII. Số lượng tấn công theo các cổng đích

Cổng	Số lượng tấn công
0	253
25	10
22	1
23	31
20	16
21	9
80	58
Các cổng khác	15

V. ỨNG DỤNG BỘ DỮ LIỆU NETFLOW TRONG PHÁT HIỆN XÂM NHẬP TRÁI PHÉP

A. Trích chọn đặc trưng

Như đã đề cập ở trên, bộ dữ liệu Netflow gồm rất nhiều trường dữ teliệu khác nhau. Tuy nhiên, để ứng dụng trong phát hiện xâm nhập trái phép, chúng tôi lựa chọn sử dụng các đặt trưng như sau:

Bảng XIX. Các đặc trưng lựa chọn trong phát hiện xâm nhập trái phép

Tên của đặc trưng	Mô tả
Protocol	(Giao thức (TCP hoặc UDP
Packets	Số lượng gói tin (packet) trong một flow
Octets	Số lượng bytes trong một flow
Flags	Số dạng hexa biểu thị cờ của flow, được gán nhãn để xác định flow nào là bình thường, flow nào là bất thường

Các đặc trưng được trích chọn nêu trên đều ở dạng số (numeric) nên rất thuận lợi cho việc thử nghiệm phát hiện xâm nhập trái phép bằng phương pháp học máy, mô phỏng trên phần mềm Weka [5].

B. Lựa chọn thuật toán học máy

Các tiêu chí được sử dụng để đánh giá hiệu năng của hệ thống phát hiện xâm nhập trái phép [6]:

- Confusion Matrix:

Bảng XX. Confusion Matrix

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

- True Positives (TP): Số lượng các bất thường được phân loại đúng là bất thường;
- True Negatives (TN): Số lượng các bình thường được phân loại đúng là bình thường;
- False Positives (FP): Số lượng các bình thường được phân loại sai là bất thường;
- False Negatives (FN): Số lượng các bất thường được phân loại sai thành bình thường.
- True Positive Rate (TPR):

$$TPR = Recall = \frac{TP}{TP + FN} \quad (1)$$

- False Positive Rate (FPR):

$$FP = \frac{FP}{FP + TN} \quad (2)$$

- Precision (P): là thước đo một hệ thống có khả năng phát hiện bình thường hay bất thường

$$P = \frac{TP}{TP + FP} \quad (3)$$

- Accuracy (A): Độ chính xác

$$A = \frac{TP + TN}{TP + FN + FP + TN} \quad (4)$$

Chúng tôi sử dụng các thuật toán học máy SVM

(Support Vector Machines) và Naive Bayes để thực hiện thử nghiệm phân loại và phát hiện xâm nhập trái phép; đồng thời đánh giá hiệu năng của từng thuật toán học máy. Phần mềm được sử dụng để thực hiện các thuật toán học máy là phần mềm Weka.

Trong thuật toán học máy SVM, việc sử dụng các kiểu hàm nhân (kernel function) khác nhau có thể cho kết quả đánh giá hiệu năng khác nhau. Có 4 kiểu hàm nhân trong SVM:

- Hàm nhân tuyến tính (Linear kernel) có dạng:

$$K_{linear}(x_1, x_2) = x_1^T x_2 + c \quad (5)$$

- Hàm nhân đa thức (Polynomial kernel) có dạng:

$$K_{poly}(x_1, x_2) = (ax_1^T x_2 + c)^d \quad (6)$$

- Hàm nhân RBF (RBF kernel) có dạng:

$$K_{RBF}(x_1, x_2) = e^{\gamma \|x_1 - x_2\|^2} \quad (7)$$

- Hàm nhân đường xích-ma (sigmoid kernel) có dạng:

$$K_{sigmoid}(x_1, x_2) = \tanh(ax_1^T x_2 + c) \quad (8)$$

C. Dữ liệu huấn luyện và kiểm thử

Để tính toán hiệu năng tổng thể của các thuật toán học máy, chúng tôi sử dụng phương pháp đánh giá 10-fold cross-validation của Weka. Với phương pháp này, bộ dữ liệu Dataset sẽ được chia một cách ngẫu nhiên thành 10 tập con. Với bộ 10 tập con, 1 tập con sẽ được sử dụng cho mục đích kiểm thử, 9 tập con khác được sử dụng cho mục đích dữ liệu huấn luyện. Phương pháp 10-fold cross-validation của Weka sẽ thực hiện lặp đi lặp lại 10 lần với tập dữ liệu, mỗi lần với một tập con làm tập kiểm thử. Kết quả của 10 lần thực hiện sẽ được tính giá trị trung bình để xác định hiệu năng tổng thể của từng thuật toán học máy.

Trong 4 bộ dữ liệu Netflow của 4 máy chủ victim, chúng tôi thấy rằng máy chủ Pascal là máy chủ được thử nghiệm tấn công nhiều nhất. Đối với máy chủ Pascal, dịch vụ mail (cổng dịch vụ đích 25) là cổng dịch vụ bị tấn công nhiều nhất. Do đó, chúng tôi đã lựa chọn bộ dữ liệu Netflow của máy chủ Pascal, với dịch vụ mail để kiểm thử.

Bảng XXI. Đặc trưng dữ liệu thử nghiệm

Thuộc tính	Giá trị
Proto	(TCP) 6
Flags	1b; 1b:::mailbomb; 2.0; 2:::portsweep; 1:::portsweep; 2:::neptune; 16.0; 1b:::ps; 6:::queso; 12:::queso; 7:::queso; c6:::queso; 1.0; 8.0; 1b:::ffbconfig; 6.0; 17.0; 13.0
Octets	89369 – 46
Packets	70 – 1

Trước khi thực hiện thử nghiệm, dữ liệu trải qua giai đoạn chuẩn hóa sử dụng kỹ thuật Discretize của Weka nhằm tăng tính chính xác cho kết quả thử nghiệm. Kết quả kiểm thử đối với từng thuật toán đối với khả năng phát hiện tấn công mailbomb như sau:

Bảng XXII. Kết quả thử nghiệm với các thuật toán

	Naive Bayes	SVM linear) (kernel	SVM polynomial) (kernel	SVM RBF) (kernel	SVM sigmoid) (kernel
TP	0.994	0.994	0.994	0.994	0.994
FP	0.001	0.001	0.004	0.001	0.001
P	0.990	0.990	0.988	0.990	0.990
Recall	0.994	0.994	0.994	0.994	0.994

VI. KẾT LUẬN

Trong phạm vi của bài báo này, chúng tôi đã trình bày mục tiêu và ý nghĩa của việc phải xây dựng bộ dữ liệu dạng Netflow cho bộ dữ liệu xâm nhập trái phép DARPA. Bằng các công cụ mã nguồn mở, kết hợp với tài liệu mô tả về nhãn tấn công của DARPA và cách thức đánh nhãn thủ công, chúng tôi đã xây dựng thành công 4 bộ dữ liệu Netflow tương ứng với các máy chủ Victim.

Chúng tôi cũng đã sử dụng công cụ WEKA, với các thuật toán học máy SVM và Naive Bayes đi kèm để thực hiện thử nghiệm phát hiện xâm nhập trái phép trên bộ dữ liệu Netflow đã xây dựng.

DARPA và ISCX đã thực hiện xây dựng các bộ dữ liệu Tcpdump rất công phu và đồ sộ. Nội dung bài báo mới thực hiện xây dựng hoàn chỉnh bộ dữ liệu Netflow từ bộ dữ liệu DARPA inside. Trong tương lai, chúng tôi sẽ tiếp tục hoàn thiện trọn vẹn bộ dữ liệu DARPA và ISCX để phục vụ cho việc nghiên cứu, thử nghiệm.

TÀI LIỆU THAM KHẢO

[1]. DARPA Intrusion Detection Data Sets, https://www.ll.mit.edu/ideval/data/;

[2]. KDD Cup 1999 Data, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html;

[3]. UT Dataset, https://www.ietf.org/proceedings/78/slides/NMRG-2.pdf;

[4]. Softflowd, Flowd software, http://www.mindrot.org/projects;

[5]. Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/;

[6]. M. E. Elhamahmy, H. N. Elmahdy, I. A. Saroit, "A New Approach for Evaluating Intrusion Detection System" , International Journal of Artificial Intelligent Systems and Machine Learning, vol. 2, no. 11, Nov. 2010.

[7]. A. M. Riad, Ibrahim Elhenawy, Ahmed Hassan and Nancy Awadallah, "Visualize network anomaly detection by using k-means clustering algorithm", International Journal of Computer Networks & Communications (IJCNC), vol.5, no. 5, Sep. 2013

[8]. Q.A. Tran, F. Jiang, J. Hu, "A real-time Netflow-based intrusion detection system with improved BBNN and high-frequency field programmable gate arrays," Proceedings of the 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 2012, pp. 201-208, Liverpool, UK

[9]. UNB ISCX Intrusion Detection Evaluation Dataset, http://www.unb.ca/research/iscx/Dataset/iscx-IDS-Dataset.html;

[10]. Ali Shiravi, Hadi Shiravi, Mahbod Tavallaee, Ali A. Ghorbani, "Toward developing a systematic approach to generate benchmark Datasets for intrusion detection," Computers & Security, vol. 31, no. 3, pp. 357-374, May 2012

[11]. Monowar H. Bhuyan, D. K. Bhattachryya, J. K. Kalita, "Network Anomaly Detection: Methods, Systems and Tools," IEEE Communications Surveys & Tutorials, vol.16, no. 1, pp. 303-336, 2014

NETFLOW DATASET IN INTRUSION DETECTION SYSTEM AND APPLICATIONS

Abstract: Instrusion datasets in computer networks have been widely applied in the study of network intrusion detection system. There are many different datasets, each has advantages and disadvantages. Netflow dataset has several advantages in intrusion detection system, particularly in large traffic data network. Currently, DARPA dataset is still used in research to detect intrusions, but the dataset is not in the form of Netflow. The objective of this paper is to present a method of building a Netflow dataset from the DARPA dataset; and its applications in detecting intrusions by machine learning methods. This dataset can be used widely in research of Netflow-based intrusion detection.

Nguyễn Hoàng Giang nhận bằng kỹ sư ngành Công nghệ thông tin chương trình Đào tạo kỹ sư chất lượng cao PFIEV (Việt – Pháp) tại Đại học Bách Khoa Hà Nội năm 2004. Hiện tại anh đang học thạc sỹ chuyên ngành Hệ thống thông tin tại Học viện Công nghệ Bưu chính viễn thông. Hướng nghiên cứu hiện tại: phát hiện xâm nhập mạng trái phép; bảo mật mạng.

Trần Quang Anh nhận bằng tiến sĩ chuyên ngành Xử lý tín hiệu và thông tin tại Đại học Thanh Hoa, Trung Quốc năm 2003, hiện là giảng viên Học viện Công nghệ Bưu chính Viễn thông. Hướng nghiên cứu chính là Phát hiện xâm nhập trái phép, Lọc thư rác và tin nhắn rác, Máy vectơ hỗ trợ, Giải thuật tiến hóa, Ứng dụng FPGA trong an ninh mạng.