ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Nguyễn Quang Huy

PHÁT HIỆN CÁC BOTNET NGANG HÀNG SỬ DỤNG CÁC ĐẶC ĐIỂM CẤU TRÚC ĐỒ THỊ

KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY
Ngành: Công nghệ thông tin

HÀ NỘI - 2023

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Nguyễn Quang Huy

PHÁT HIỆN CÁC BOTNET NGANG HÀNG SỬ DỤNG CÁC ĐẶC ĐIỂM CẤU TRÚC ĐỒ THỊ

KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY
Ngành: Công nghệ thông tin

Cán bộ hướng dẫn: TS. Nguyễn Đại Thọ

HÀ NỘI - 2023

VIETNAM NATIONAL UNIVERSITY, HANOI
UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Nguyen Quang Huy

PEER-TO-PEER BOTNET DETECTION USING GRAPH STRUCTURE FEATURES

BACHELOR'S THESIS
Major: Information Technology

Supervisor: Dr. Nguyen Dai Tho

HANOI - 2023

Tóm Tắt

Việc gia tăng số lượng các thiết bị kết nối Internet đã và đang đặt ra nhiều thách thức về bảo mật trên không gian mạng. Trong đó, các mạng botnet ngang hàng đang trở thành một trong những mối đe dọa chính đối với an ninh mạng vì đóng vai trò nền tảng cho rất nhiều loại tội phạm mạng khác nhau. Đã có nhiều nghiên cứu về vấn đề này và nhiều giải pháp phát hiện botnet ngang hàng đã được đề xuất, tuy nhiên, nhiều đặc trưng quan trọng vẫn chưa được đề cập dẫn đến kết quả đạt được chưa cao. Một giải pháp phát hiện botnet ngang hàng mới được đề xuất trong khóa luận dựa trên việc phân tích hành vi của botnet thông qua lưu lượng mạng và đồ thị giao tiếp. Bên cạnh những đặc trưng thông thường của botnet như tính đa dạng đích, tính liên hệ chung, đa dạng kích thước gói tin hay tần suất liên hệ dày đặc, khóa luận đề xuất ba độ đo mới dựa trên phân tích các đặc điểm cấu trúc đồ thị, áp dụng trong việc phân tách các cộng đồng hợp pháp và cộng đồng chứa lưu lượng độc hại trên đồ thị liên hệ chung. Một bộ dữ liệu mới được xây dựng bằng cách kết hợp các lưu lượng hợp pháp và botnet phổ biến với lưu lượng ứng dụng ngang hàng được thu thập mới để tăng thêm tính thách thức cho thực nghiệm. Kết quả cho thấy hệ thống đề xuất của khóa luận có thể phát hiện botnet hiệu quả với độ chính xác và độ nhạy cao trong nhiều bộ thông số khác nhau.

***Từ khóa**: P2P botnet, phát hiện P2P botnet, an ninh mạng, đặc điểm cấu trúc đồ thị.*

Abstract

The increasing number of Internet-connected devices is posing many challenges to network security. Among them, peer-to-peer botnets have become one of the main threats to cybersecurity, as they serve as a platform for various types of cybercrime. Many studies have been conducted on this issue and many solutions for detecting peer-to-peer botnets have been proposed. However, a lot of important features have not been addressed, leading to suboptimal results. A new solution for detecting peer-to-peer botnets has been proposed in this thesis based on the analysis of their behavior through network traffic and communication graphs. In addition to the common features of botnets such as diverse targets, shared connections, diverse packet sizes, and frequent dense communication, the thesis proposes three new metrics based on the analysis of the structural characteristics of communication graphs. These metrics are applied to separate legitimate communities from those containing malicious traffic on a shared communication graph. A new dataset has been constructed by combining legitimate and popular botnet traffic with the newly collected peer-to-peer application traffic to increase the challenge for experimentation. The results show that the proposed system in the thesis can effectively detect botnets with high accuracy and sensitivity in various parameter settings.

***Key words**: P2P botnet, P2P botnet detection, cybersecurity, network security, graph structure features.*

Lời cảm ơn

Đầu tiên, tôi muốn gửi lời cảm ơn chân thành đến TS. Nguyễn Đại Thọ, người đã trực tiếp hướng dẫn chỉ bảo và định hướng tôi để tôi có thể hoàn thiện khóa luận này.

Tiếp theo, cho phép tôi được gửi lời cảm ơn tới Khoa Công Nghệ Thông Tin - Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội và các thầy cô trong Khoa đã tạo điều kiện thuận lợi cho tôi được học tập, nghiên cứu và thực hiện đề tài tốt nghiệp này.

Tôi xin được cảm ơn đến tập thể thành viên Phòng thí nghiệm An toàn thông tin - Trường Đại học Công Nghệ đã đồng hành cùng với tôi trong quá trình nghiên cứu cũng như xây dựng khóa luận này.

Tôi cũng xin được cảm ơn tập thể lớp K64-C-CLC đã chỉ bảo và giúp đỡ trong những lúc khó khăn cũng như đã tạo ra một môi trường lý tưởng cho tôi phát triển trong suốt bốn năm đại học.

Cuối cùng tôi xin dành lời cảm ơn đến gia đình, những người đã yêu thương, tạo động lực để tôi cố gắng rèn luyện, hoàn thiện mình hơn và tới bạn Nguyễn Thị Thư, người đã động viên cũng như giúp đỡ tôi vượt qua những lúc khó khăn để hoàn thiện khóa luận này.

Tôi xin chân thành cảm ơn!

Lời cam đoan

Tôi xin cam đoan toàn bộ khóa luận về đề tài "Phát hiện các botnet ngang hàng sử dụng các đặc điểm cấu trúc đồ thị" là của tôi, tất cả đề xuất và kết quả đều do tôi thực hiện dưới sự hướng dẫn của TS. Nguyễn Đại Thọ. Những kiến thức, phương pháp nghiên cứu liên quan đến khóa luận này đều đã được trích dẫn và ghi chú rõ ràng trong danh sách những tài liệu tham khảo. Tôi xin chấp nhận tất cả những truy cứu về trách nhiệm theo quy định của Trường Đại học Công nghệ - ĐHQG Hà Nội nếu có những hành vi không trung thực.

Hà Nội, ngày 24 tháng 05 năm 2023
Sinh viên

Nguyễn Quang Huy

Mục lục

Bảng thuật ngữ

Thuật ngữ tiếng Anh	**Từ viết tắt**	**Ý nghĩa tiếng Việt**
Command and Control	C&C	Chỉ huy và kiểm soát
Peer-to-Peer	P2P	Ngang hàng
Internet of Things	IoT	Vạn vật kết nối Internet
Distributed Denial of Service	DDoS	Tấn công từ chối dịch vụ phân tán
Internet Relay Chat	IRC	Giao thức giao tiếp trực tuyến qua Internet
Hypertext Transfer Protocol Secure	HTTPS	Giao thức truyền tải siêu văn bản an toàn
Intrusion Detection System	IDS	Hệ thống phát hiện xâm nhập
Internet Protocol	IP	Giao thức mạng
Destination Diversity Ratio	DDR	Tỉ lệ đa dạng địa chỉ đích
Mutual Contact Ratio	MCR	Tỉ lệ liên hệ chung
Flow Loss-response Rate	FLR	Tỉ lệ luồng không phản hồi

Danh sách hình vẽ

Danh sách bảng

Chương 1

Mở đầu

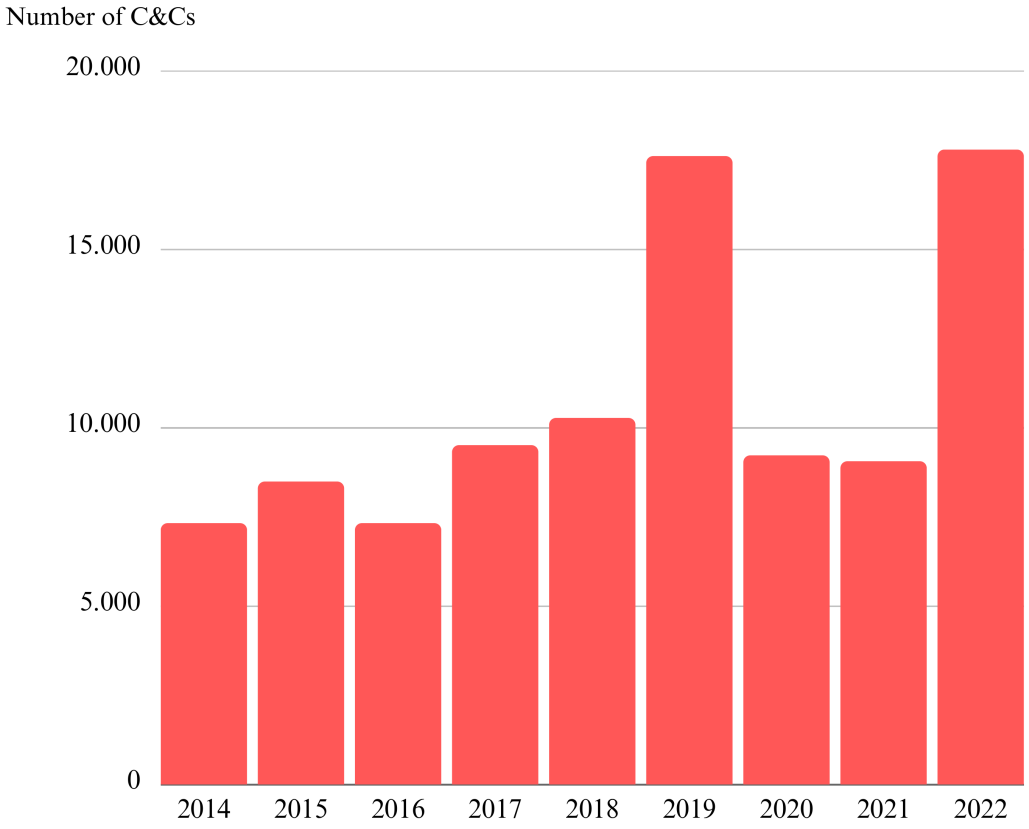
1.1 Bối cảnh

Ngày nay, trong thời đại công nghệ phát triển, hàng tỷ thiết bị điện tử như điện thoại di động, máy tính xách tay, máy tính bảng hay các thiết bị thông minh trong gia đình đều được kết nối với Internet. Số lượng các thiết bị này đang ngày càng tăng lên với tốc độ nhanh chóng, mang lại nhiều lợi ích cho con người trong cuộc sống. Theo thống kê của Statista [1], chỉ tính riêng trong tháng 1 năm 2023 đã có khoảng hơn năm tỉ người dùng Internet trên toàn thế giới, chiếm khoảng 64% dân số thế giới.

Việc gia tăng số lượng các thiết bị kết nối Internet mang lại nhiều cơ hội và lợi ích cho con người, nhưng cũng đặt ra nhiều thách thức về an ninh mạng. Một trong những hình thức đe dọa đáng quan tâm nhất hiện này là sự hình thành của các botnet từ các thiết bị trong mạng lưới Internet, từ đó làm cơ sở cho việc thực hiện các hành vi tấn công mạng.

Botnet là một mạng lưới các máy bị xâm nhập và nhiễm mã độc, được điều khiển bởi một trung tâm điều khiển từ xa, còn gọi là máy chủ C&C (Command and Control server) hay botmaster [2]. Kẻ tấn công sẽ khai thác các thiết bị này để thực hiện các hoạt động bất hợp pháp như phát tán thư rác, đánh cắp dữ liệu, lây truyền phần mềm độc hại hay tấn công từ chối dịch vụ phân tán (DDoS). Theo thống kê hàng năm của Spamhus, số lượng máy chủ C&C được phát hiện vào năm 2019 đã tăng gần gấp ba lần so với năm 2014, và nó cũng tiếp tục tăng mạnh từ cuối năm 2021 cho đến nay (hình 1.1).

Botnet truyền thống thường được tổ chức theo kiến trúc tập trung, trong đó tất cả các bot trong một mạng botnet sẽ giao tiếp và nhận lệnh từ một hoặc một vài máy chủ đóng vai trò như là trung tâm điều khiển từ xa [2]. Thiết kế này làm cho mạng botnet dễ



Hình 1.1: Thống kê của Spamhus về số lượng máy chủ C&C được phát hiện qua các năm

bị phát hiện và ngăn chặn hơn do có một số điểm tập trung để kiểm soát toàn bộ mạng. Nếu các nút điều khiển này bị phát hiện hoặc tấn công, botnet có thể bị dừng hoạt động. Ngoài ra, botnet truyền thống cũng thường giới hạn về quy mô do khó mở rộng hoặc quản lý các bot một cách hiệu quả chỉ với vài máy chủ tập trung.

Để tránh những nhược điểm của kiến trúc tập trung, các mạng botnet ngày nay thường sử dụng kiến trúc mạng ngang hàng (P2P) [3]. Các mạng P2P không có nút tập trung để ra lệnh và điều khiển, khi đó mỗi máy sẽ hoạt động với vai trò của cả máy khách và máy chủ. Vì vậy, ngay cả khi một nút ngoại tuyến hoặc bị phát hiện, các nút khác trong mạng cũng sẽ không bị ảnh hưởng nhiều.

Các botnet P2P đã và đang trở thành một trong những mối đe dọa chính đối với Internet. Storm được biết đến như một trong những botnet ngang hàng đầu tiên, được phát hiện vào tháng 1 năm 2007, và nó đã nhanh chóng trở nên khổng lồ với khoảng gần một triệu máy chủ bị nhiễm mỗi tháng [4]. Vào tháng 3 năm 2010, Microsoft đã phát hiện và gỡ bỏ các máy chủ của mạng botnet Waledac, trước đó, nó đã lây nhiễm cho hằng trăm ngàn máy tính và có khả năng gửi từ một đến hai tỉ tin nhắn rác mỗi ngày [5].

Các cuộc tấn công của botnet P2P được ghi nhận đã gây thiệt hại lớn cho các cá nhân, tổ chức, hay đặc biệt là các ngân hàng.

1.2 Đặt vấn đề

Botnet ngang hàng đã và đang tạo ra nhiều thách thức cho các nhà nghiên cứu an ninh mạng. Việc thiết kế một hệ thống phát hiện botnet P2P hiệu quả là rất khó khăn. Thứ nhất, các mạng botnet có xu hướng hoạt động lén lút và dành phần lớn thời gian trong giai đoạn chờ đợi trước khi thực hiện bất kỳ hoạt động độc hại nào. Thứ hai, các botnet có xu hướng mã hóa kênh truyền C&C khiến các phương pháp dựa trên kiểm tra gói tin như dựa vào chữ ký không hiệu quả. Thứ ba, vai trò của một bot có thể thay đổi linh hoạt tùy thuộc vào cấu trúc hiện tại của mạng botnet.

Các phương pháp dựa trên đặc điểm bất thường được đánh giá là phù hợp nhất trong việc phát hiện botnet ngang hàng. Có nhiều kỹ thuật đã được đề xuất để xác định P2P botnet dựa trên sự bất thường. Một số trong đó dựa vào học máy và học sâu như [6], [7]. Lợi thế của phương pháp này là tính linh hoạt, khả năng thích nghi và phát hiện các botnet mới. Tuy nhiên, điểm yếu của chúng là yêu cầu cao về tài nguyên phần cứng cũng như dữ liệu học tập phải đủ tốt. Thực tế cho thấy rằng các phương pháp dựa trên học máy hiện nay có hiệu quả chưa được cao.

Một cách khác để phát hiện P2P botnet là dựa trên việc phân tích lưu lượng mạng giữa các máy, từ đó tìm ra những điểm khác biệt về hành vi và khuôn dạng giao tiếp đặc trưng như lưu lượng truy cập lớn, độ trễ mạng cao, lưu lượng trên các cổng bất thường [8]. Từ đó có thể phát hiện được P2P botnet nhanh chóng và hiệu quả với những đặc tính vốn có. Các đề xuất theo hướng tiếp cận trên có thể kể đến như [9], [10], [11], [12]. Các giải pháp này thường quan sát một số đặc điểm hành vi từ lưu lượng mạng của P2P botnet, từ đó xây dựng các hệ thống phân cụm, các đồ thị liên hệ hoặc sử dụng các ngưỡng để phân tách. Tuy nhiên, có một điểm chung là hầu hết các nghiên cứu thường chỉ đánh giá được một số khía cạnh, không bao quát được tất cả hành vi đặc trưng quan trọng của botnet. Một số khác lại có những cách xây dựng thuật toán chưa hiệu quả dẫn đến tỉ lệ phát hiện P2P botnet không tốt, cũng như số lượng ứng dụng hợp pháp bị phát hiện nhầm là botnet tương đối cao.

1.3 Mục tiêu và đóng góp của khóa luận

Mục tiêu của khóa luận hướng đến cải thiện độ chính xác trong việc phát hiện botnet P2P thông qua việc sử dụng phương pháp dựa trên phân tích hành vi cộng đồng. Từ các nghiên cứu trước đây của tôi và cộng sự dưới sự hướng dẫn của TS. Nguyễn Đại Thọ, một bài báo liên quan đã được công bố tại hội nghị ICIIT 2023 (2023 8th International Conference on Intelligent Information Technology) [13] và trình bày tại hội nghị sinh viên nghiên cứu khoa học Khoa Công nghệ thông tin - trường Đại học Công Nghệ - Đại học Quốc gia Hà Nội. Đóng góp của chúng tôi trong nghiên cứu đó bao gồm:

- Đề xuất một tính chất mới của botnet là sự thay đổi một khoảng nhỏ và liên tục trong độ dài gói tin, từ đó kết hợp với đặc tính của mạng P2P để phát hiện các lưu lượng ngang hàng.
- Sử dụng đặc tính về việc duy trì các kết nối giữa các máy trong mạng botnet P2P, áp dụng vào việc xây dựng đồ thị liên hệ lẫn nhau cấp luồng mạng dựa vào tần suất liên hệ, tạo ra hệ thống mới phát hiện botnet với tỉ lệ dương tính giả thấp hơn.
- Thu thập các nguồn dữ liệu và xây dựng bộ dữ liệu mới thử thách hơn, từ đó đánh giá tốt hơn hoạt động của mô hình.

Trong khóa luận này, tôi tiếp tục đề xuất những giải pháp để cải tiến phương pháp phát hiện các mạng botnet P2P dựa trên việc phân tích hành vi hội thoại giữa các máy kết hợp với các đặc điểm cấu trúc đồ thị. Khóa luận kết hợp một số đặc điểm quan trọng của P2P botnet đã được sử dụng trong các nghiên cứu trước đây, từ đó cải tiến để có một phương pháp mới hiệu quả và chính xác hơn. Những đóng góp chính của khóa luận như sau:

- Phân tích một số đặc điểm cấu trúc của đồ thị đã được áp dụng trong việc phát hiện P2P botnet trên đồ thị giao tiếp đơn giản, từ đó đề xuất một số giải pháp áp dụng trong đồ thị liên hệ lẫn nhau cấp luồng mạng để phân tách lưu lượng độc hại và hợp pháp.
- Thiết lập môi trường để thu thập dữ liệu P2P hợp pháp và kết hợp với bộ dữ liệu sử dụng trong nghiên cứu trước đây để loại bỏ những thiên vị và tạo ra tập dữ liệu mới mang tính thách thức hơn.

- Hệ thống đề xuất của khóa luận đạt được tỉ lệ phát hiện botnet 100% và không có dương tính giả nào trên tập dữ liệu thực nghiệm. Hệ thống hoàn toàn khả thi để triển khai thực tế trong việc phát hiện botnet ngang hàng tại các tổ chức.

1.4 Cấu trúc của khóa luận

Khóa luận bao gồm 4 chương và phần Kết luận:

Chương 1: Giới thiệu về bài toán phát hiện mạng botnet ngang hàng và những đóng góp chính của khóa luận.

Chương 2: Giới thiệu tổng quan về phát hiện botnet ngang hàng và các nghiên cứu liên quan.

Chương 3: Mô tả chi tiết hệ thống đề xuất phát hiện botnet ngang hàng dựa trên hành vi cộng đồng của botnet và các đặc tính cấu trúc trên đồ thị cấp luồng mạng.

Chương 4: Mô tả phương pháp thực nghiệm, cách thức xây dựng dữ liệu và đánh giá hiệu quả của phương pháp.

Kết luận: Tổng kết về những kết quả nghiên cứu đã đạt được trong khóa luận, những hạn chế còn tồn tại và đề xuất những công việc sẽ thực hiện trong tương lai.

Chương 2

Tổng quan về phát hiện botnet ngang hàng

2.1 Tổng quan về botnet

Botnet là một mạng lưới các thiết bị điện tử đã bị xâm nhập bởi một phần mềm độc hại, được điều khiển từ xa bởi kẻ tấn công thông qua một máy chủ chỉ huy và kiểm soát [2]. Botnet thường bao gồm các máy tính, điện thoại di động, thiết bị IoT (Internet of Things) và các thiết bị kết nối mạng khác. Các bot trong mạng botnet thường chạy các chương trình bot mà kẻ tấn công có thể sử dụng để thực hiện các hoạt động tấn công mạng, bao gồm tấn công DDoS, lừa đảo trực tuyến, spam, phát tán phần mềm độc hại và trộm thông tin cá nhân. Một mạng botnet trung bình có thể bao gồm từ mười nghìn đến một triệu máy bị xâm nhập.

Kẻ tấn công thường kiểm soát botnet của mình thông qua một trong hai kiến trúc: mô hình tập trung với việc giao tiếp trực tiếp giữa người điều khiển bot và các máy trong mạng, và hệ thống phân tán với nhiều liên kết giữa tất cả các thiết bị bot bị nhiễm.

Kiến trúc tập trung

Thế hệ botnet đầu tiên thường hoạt động trên kiến trúc máy khách - máy chủ, trong đó các bot được kết nối với một hoặc một vài nút điều khiển và chỉ huy, được gọi là các máy chủ command-and-control (C&C servers) [2]. Các máy chủ này có toàn quyền kiểm soát hoạt động của các bot và có thể gửi lệnh cho chúng thông qua các kênh giao tiếp bảo mật, chẳng hạn như IRC hoặc HTTPS.

Với mô hình tập trung, kẻ tấn công chỉ cần điều khiển một vài trung tâm duy nhất

để quản lý toàn bộ botnet. Điều này giúp họ dễ dàng giám sát và kiểm soát các hoạt động của botnet, cũng như tập trung các tài nguyên và năng lực để tấn công mục tiêu. Botnet tập trung cũng có khả năng dễ dàng thay đổi và cập nhật các chức năng và tính năng mới mà không cần thay đổi tệp thực thi độc hại trên từng thiết bị nhiễm. Điều này giúp botnet tập trung đáp ứng được các yêu cầu và mục tiêu tấn công mới một cách nhanh chóng và hiệu quả.

Tuy nhiên, loại kiến trúc này lại có những nhược điểm nghiêm trọng, như sự dễ bị phát hiện và đánh sập do có một số điểm tập trung duy nhất để tấn công. Một khi các điểm tập trung này bị phát hiện, toàn bộ mạng botnet sẽ bị phá hủy. Kiến trúc máy chủ - máy khách tạo ra một lưu lượng truy cập tập trung lớn trên một số địa chỉ IP, dễ dàng bị phát hiện bởi các công cụ giám sát và phân tích lưu lượng mạng. Các chuyên gia an ninh mạng có thể sử dụng các kỹ thuật giám sát mạng để phát hiện các mẫu hoạt động của botnet, bao gồm thời gian liên lạc, tần suất liên lạc, các địa chỉ IP sử dụng, các giao thức mạng và mật độ truy cập. Nếu lưu lượng truy cập tập trung vào một số máy chủ cụ thể, có thể dự đoán rằng một botnet tập trung đang hoạt động ở đó. Ngoài ra, botnet tập trung cũng có thể bị giới hạn về quy mô do số lượng bot được điều khiển từ một điểm tập trung là có hạn. Do đó, để tăng tính ổn định và khả năng tồn tại lâu dài, nhiều botnet ngày nay đã chuyển sang sử dụng kiến trúc phân tán với cách thức trao đổi thông tin phức tạp hơn.

Kiến trúc phân tán

Trong kiến trúc phân tán, các bot được kết nối với nhau và có thể gửi và nhận lệnh từ những bot khác trong mạng [3]. Trong loại kiến trúc này, không có máy chủ trung tâm. Tất cả các yêu cầu được xử lý bởi các đồng nghiệp khác nhau trong mạng. Trong các kiến trúc hoàn toàn phi tập trung, các hành vi ngang hàng giữa các máy sẽ tương đối giống nhau. Các đồng nghiệp đóng vai trò là máy chủ khi xử lý truy vấn tìm kiếm tệp và là máy khách khi yêu cầu tệp. Tuy nhiên, trong các kiến trúc phi tập trung một phần, các đồng nghiệp có khả năng tính toán và băng thông mạng tốt hơn sẽ có cơ hội trở thành "siêu ngang hàng", đóng vai trò quan trọng hơn trong mạng.

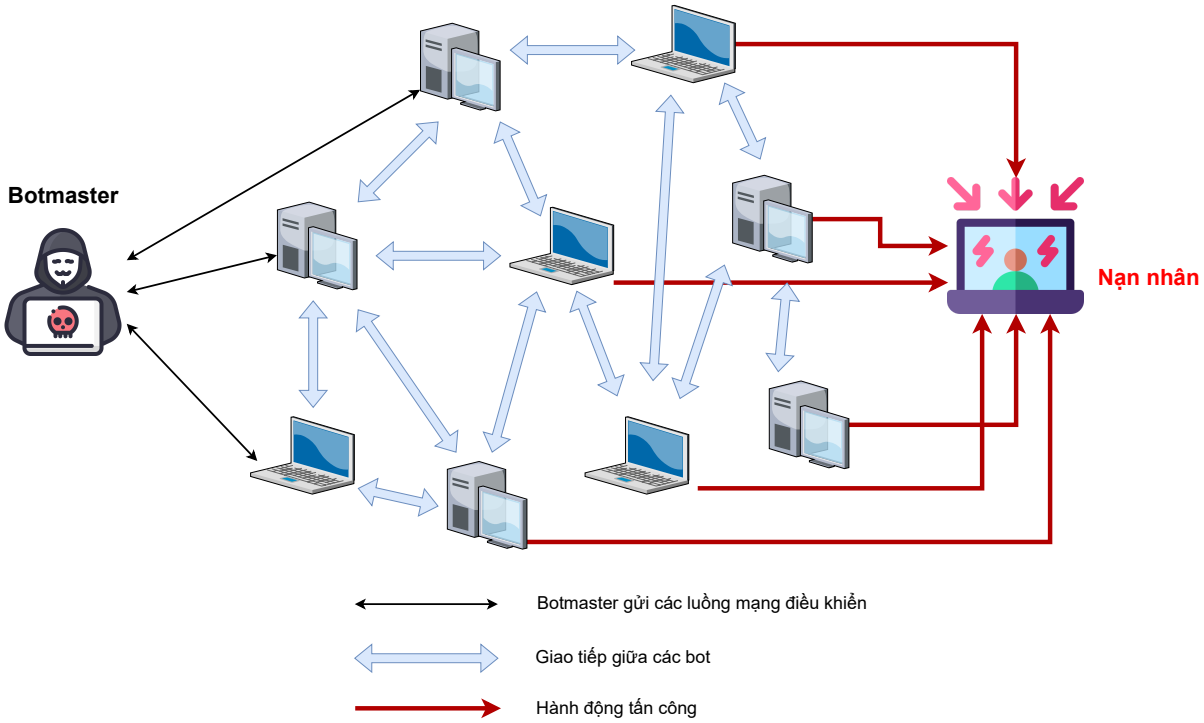
Kiến trúc phân tán của botnet có nhiều ưu điểm hơn so với kiến trúc tập trung. Trong kiến trúc này, các bot được phân tán trên nhiều máy chủ khác nhau, khiến cho việc phát hiện và ngăn chặn botnet trở nên khó khăn. Hơn nữa, do không có một máy chủ điều khiển chính nào, việc tấn công vào botnet cũng trở nên phức tạp hơn. Bên cạnh đó, kiến trúc phân tán còn giúp botnet trở nên ổn định, do sự chia sẻ công việc và trách nhiệm giữa các bot. Nếu một bot bị tấn công hay bị ngưng hoạt động, các bot khác có thể tiếp tục hoạt động mà không bị ảnh hưởng nhiều. Tuy nhiên, kiến trúc phân tán cũng

có một số hạn chế, ví dụ như khó khăn trong việc quản lý và điều khiển các bot, cũng như khó khăn trong việc thực hiện các hoạt động đồng bộ và phân tán thông tin giữa các bot.

Khóa luận này sẽ tập trung vào các giải pháp phát hiện botnet ngang hàng do những ưu điểm và thách thức mà nó tạo ra. Phần tiếp theo của chương sẽ trình bày về vòng đời phát triển của mạng botnet ngang hàng và các phương pháp để phát hiện chúng.

2.2 Vòng đời phát triển của mạng botnet ngang hàng

Vòng đời của một botnet ngang hàng có nhiều giai đoạn, bao gồm lây nhiễm và lan truyền, giao tiếp và kiểm soát, thực hiện tấn công, cập nhật và bảo trì [14]. Hình 2.1 thể hiện hoạt động của một mạng botnet P2P sau khi mạng được hình thành.



Hình 2.1: Hoạt động của mạng botnet ngang hàng

Giai đoạn lây nhiễm và lan truyền: Ở giai đoạn này, kẻ tấn công sẽ xâm nhập nhiều máy tính trên Internet và tìm cách để điều khiển chúng từ xa thông qua các loại mã độc ẩn chứa trong các phần mềm độc hại được phát tán qua các kênh khác nhau như các trang web, trình duyệt, email hay các nền tảng mạng xã hội. Các máy tính bị xâm nhập sẽ có chứa sẵn một danh sách các máy ngang hàng và nó sẽ tìm cách liên hệ với chúng để cập nhật thêm danh sách ngang hàng lân cận cũng như thông báo tới các ngang hàng khác. Ngoài ra, botnet cũng có thể được hình thành trên cơ sở của một mạng P2P có sẵn

khác, lưu lượng của chúng khi đó có thể được trộn lẫn với lưu lượng P2P hợp pháp, khiến chúng khó bị phát hiện hơn.

Giai đoạn giao tiếp và kiểm soát: Các gói tin C&C được trao đổi giữa các máy trong mạng botnet [3]. Giai đoạn này chiếm phần lớn thời gian trong vòng đời của một mạng botnet. Cơ chế C&C là phần chính và vô cùng quan trọng trong thiết kế của mạng botnet nói chung và mạng P2P botnet nói riêng, ảnh hưởng đến sức mạnh và khả năng ẩn mình của mạng botnet đó. Các cơ chế C&C của P2P botnet dựa trên tính chất của mạng ngang hàng, vì vậy mỗi máy sẽ đóng hai vai trò là máy chủ phân phối lệnh và máy khách nhận lệnh.

Các cơ chế C&C bao gồm cơ chế kéo và cơ chế đẩy (pull and push) [3]. Cơ chế kéo đề cập đến việc các bot trong mạng P2P chủ động truy xuất các lệnh từ một nơi được định trước mà kẻ tấn công sẽ phân phát lệnh. Ngược lại, cơ chế đẩy nghĩa là là các bot sẽ thụ động chờ lệnh được gửi đến và sau đó sẽ chuyển tiếp lệnh đến các ngang hàng khác. Các mạng botnet P2P ngày nay thường sử dụng linh hoạt cả hai cơ chế này. Khi một bot nhận được một lệnh, nó sẽ cố gắng chuyển tiếp lệnh đó tới các ngang hàng (cơ chế đẩy), đồng thời những máy khác cũng sẽ định kỳ liên hệ với các máy ngang hàng của chúng để cố gắng truy xuất các lệnh mới (cơ chế kéo).

Việc liên hệ định kỳ với các ngang hàng cũng góp phần giúp duy trì cấu trúc mạng cho mạng botnet P2P. Mỗi máy sẽ lưu trữ một danh sách ngang hàng và tiến hành trao đổi danh sách ngang hàng mỗi khi giao tiếp với máy khác trong mạng, đồng thời cập nhật lại nếu cần thiết.

Giai đoạn thực hiện tấn công: Các bot trong mạng sẽ nhận lệnh từ botmaster và thực hiện các hoạt động độc hại trong giai đoạn tấn công. Các cuộc tấn công này có thể kể đến như tấn công từ chối dịch vụ phân tán DDoS, gửi thư rác, lừa đảo, gian lận nhấp chuột, lan truyền vi-rút và một số hoạt động khác.

Giai đoạn cập nhật và bảo trì: Sau khi tấn công, để tăng cường cho chiến lược phòng ngừa, nếu một số bot bị phát hiện, việc xóa những máy này lưu trong tệp nhị phân (danh sách ngang hàng) của các peer khác được thực hiện. Nếu có nhiều bot bị phát hiện, kẻ tấn công sẽ lên kế hoạch để xây dựng mạng botnet mới.

2.3 Các phương pháp phát hiện botnet ngang hàng

Phát hiện botnet ngang hàng là một nhiệm vụ quan trọng để bảo vệ hệ thống và dữ liệu của các tổ chức. Các phương pháp phát hiện botnet ngang hàng được chia thành hai

loại chính: dựa trên honeypot và dựa trên hệ thống phát hiện xâm nhập (IDS) [14].

2.3.1 Phát hiện dựa trên honeypot

Phát hiện botnet ngang hàng dựa trên honeypot là một phương pháp phổ biến để xác định sự tồn tại và hoạt động của botnet [14]. Honeypot là một hệ thống giả lập, được thiết lập nhằm mô phỏng các thiết bị hoặc ứng dụng để thu hút các kẻ tấn công. Khi botnet tấn công vào honeypot, nó sẽ để lại dấu vết và cung cấp thông tin về hành vi và tính năng của botnet.

Việc phát hiện botnet dựa trên honeypot giúp các chuyên gia an ninh mạng có thể tìm hiểu và nghiên cứu botnet một cách an toàn, mà không gây tác động đến hệ thống hoặc thiết bị thật. Các chuyên gia có thể thu thập thông tin về địa chỉ IP, cổng mở, giao thức sử dụng và các lệnh điều khiển được gửi đến các bot để phân tích và giải mã các hành động của botnet. Ngoài ra, honeypot còn có thể được sử dụng để phát hiện các mối đe dọa mới và không rõ nguồn gốc, cũng như giúp cho các chuyên gia có thể phát hiện các lỗ hổng trong hệ thống mạng của mình để nâng cao độ an toàn.

Tuy nhiên, phương pháp phát hiện botnet ngang hàng dựa trên honeypot hiện nay có nhiều hạn chế. Kẻ tấn công có thể sử dụng các kỹ thuật giả mạo thông tin hoặc ẩn danh để che giấu hành vi của botnet, khiến cho honeypot không thể phát hiện được botnet một cách chính xác. Hơn nữa, honeypot có thể trở thành một mục tiêu tấn công cho kẻ xấu, khiến cho hệ thống mạng của người sử dụng dễ bị tấn công. Do đó, các chuyên gia an ninh mạng cần phải sử dụng phương pháp này một cách cẩn thận và kết hợp với các phương pháp khác để đảm bảo tính hiệu quả và an toàn của quá trình phát hiện botnet ngang hàng.

2.3.2 Phát hiện dựa trên hệ thống phát hiện xâm nhập (IDS)

Rất nhiều nhà nghiên cứu đã phân tích các mạng botnet P2P gần đây để xác định các tính năng đặc trưng của chúng và đề xuất những hệ thống phát hiện xâm nhập. Có hai hướng tiếp cận chính là dựa vào chữ ký và dựa trên sự bất thường của mạng botnet [14].

Các phương pháp dựa vào chữ ký sử dụng một cơ sở dữ liệu chữ ký đã biết trước đó, so sánh với các tập tin hoạt động trên mạng để xác định xem chúng có chữ ký giống với của botnet hay không. Đây là phương pháp phổ biến nhất và đơn giản nhất trong các phương pháp phát hiện botnet. Tuy nhiên điều này đang dần trở nên bất khả thi do việc mã hóa gói tin của các mạng botnet gây cản trở cho việc kiểm tra chữ ký cũng như không

thể phát hiện các botnet mới.

Việc phát hiện botnet dựa trên phân tích sự bất thường đang được các nhà nghiên cứu ưa chuộng hơn để có thể ứng phó với các biến thể botnet mới đang ngày càng tinh vi. Trong phần này, khóa luận sẽ trình bày về một số kỹ thuật phát hiện botnet ngang hàng dựa trên sự bất thường.

2.3.2.1 Phát hiện dựa trên hành vi

Phát hiện dựa trên hành vi là một phương pháp phát hiện P2P botnet dựa trên việc phân tích và so sánh hành vi của các hoạt động mạng để phát hiện các hoạt động không bình thường hoặc đáng ngờ của các bot trong mạng botnet. Các đặc điểm hành vi của botnet hình thành từ kiến trúc cơ bản và cơ chế hoạt động của nó. Các bot liên tục thực hiện các giao tiếp để duy trì cấu trúc mạng, phản hồi các bot lân cận cũng như nhận lệnh từ botmaster. Các hành vi của các bot trong mạng khi đó sẽ biểu hiện không giống với hành vi của các ứng dụng hợp pháp, đặc biệt là các ứng dụng được sử dụng trực tiếp bởi con người. Các kỹ thuật phát hiện dựa trên lưu lượng và dựa trên đồ thị phần lớn sẽ phân tích đặc điểm hành vi của mạng.

2.3.2.2 Phát hiện dựa trên lưu lượng truy cập

Phương pháp phát hiện botnet dựa trên lưu lượng mạng là một trong những phương pháp hiệu quả để phát hiện botnet. Khi bot P2P đóng vai trò là một đồng đẳng trong mạng botnet P2P, nó cần thiết lập kết nối với càng nhiều đồng nghiệp lân cận càng tốt để xây dựng mạng bot. Do đó, bot P2P sẽ tạo ra nhiều lưu lượng mạng bất thường.

Một trong những phương pháp phát hiện botnet dựa trên lưu lượng mạng phổ biến nhất là phương pháp dựa trên mô hình luồng nhiều pha [14]. Mô hình luồng nhiều pha sẽ xử lý một lượng lớn dữ liệu mạng để thực hiện phân tích sự trao đổi thông tin giữa các ngang hàng và các đồng nghiệp hàng xóm được liên kết. Phương pháp này sẽ xác định các luồng lưu lượng mạng giống với mô hình mạng của botnet. Các mô hình này bao gồm các trạng thái và sự chuyển đổi giữa các trạng thái, ví dụ như các trạng thái của bot khi thiết lập kết nối với đồng nghiệp hoặc khi truyền dữ liệu.

Phương pháp dựa trên lưu lượng truy cập có thể phát hiện được các loại botnet khác nhau, bao gồm cả các botnet mới được phát triển với độ chính xác cao. Tuy nhiên, nhược điểm của phương pháp này là có thể gây ra nhiễu đối với các lưu lượng mạng bình thường dẫn đến tỉ lệ dương tính giả cao nếu không phân tích và áp dụng chính xác.

Ngoài ra, phương pháp này đòi hỏi nhiều dữ liệu lưu lượng mạng và phải thường xuyên cập nhật để có thể phát hiện được các botnet mới. Thêm vào đó, việc xây dựng các mô hình luồng nhiều pha phức tạp và đòi hỏi kỹ thuật chuyên môn cao để triển khai.

2.3.2.3 Phát hiện dựa trên đồ thị

Phương pháp dựa trên đồ thị là một phương pháp phát hiện botnet ngang hàng dựa trên sự phân tích đồ thị các kết nối giữa các máy tính. Mục tiêu của phương pháp này là phát hiện các kết nối bất thường và các hoạt động của botnet trên đồ thị. Cấu trúc đồ thị là một tính năng vốn có của mạng botnet và rất hữu ích để hiểu cách các mạng botnet, đặc biệt là mạng botnet ngang hàng giao tiếp nội bộ. Trong mạng P2P, đồ thị truyền thông tạo bởi các kết nối C&C thể hiện các kết cấu liên kết đặc trưng và rất hữu ích để phân loại lưu lượng truy cập và phát hiện mạng botnet.

Phương pháp này có khả năng phát hiện botnet một cách chính xác và hiệu quả. Nó giúp phát hiện được các hành vi bất thường của botnet thông qua việc phân tích đồ thị. Phương pháp này cũng có thể phát hiện được các botnet mới và chưa được biết đến trước đó.

Nhược điểm của phương pháp này cần có độ phức tạp cao, đặc biệt là khi phải xử lý các đồ thị lớn và phức tạp. Cũng giống như phương pháp dựa trên lưu lượng mạng, kỹ thuật này cần có dữ liệu đầu vào chính xác, đầy đủ và được cập nhật thường xuyên để hoạt động hiệu quả.

2.3.2.4 Phát hiện dựa trên các thuật toán học máy

Phương pháp phát hiện botnet dựa trên học máy (machine learning) là một phương pháp tiên tiến để phát hiện botnet trong mạng máy tính. Các thuật toán học máy được sử dụng để xây dựng mô hình phát hiện botnet dựa trên các tính năng của lưu lượng mạng, giao tiếp trong mạng và các đặc điểm hành vi. Sau khi được huấn luyện trên các tập dữ liệu được gán nhãn, mô hình này có thể phân loại gói tin mạng mới và xác định liệu chúng có phải là botnet hay không.

Ưu điểm của phương pháp phát hiện botnet dựa trên học máy là tính linh hoạt, tính tự động hóa cũng như khả năng phát hiện các botnet mới tương đối tốt. Hệ thống được thiết kế để tự động học và tìm ra các đặc trưng quan trọng để phát hiện botnet. Điều này giảm thiểu sự phụ thuộc vào kinh nghiệm của nhà nghiên cứu và giúp tăng độ chính xác của phương pháp.

Tuy nhiên, phương pháp này đòi hỏi yêu cầu cao về tài nguyên phần cứng do cần hoạt động liên tục, cũng như dữ liệu học tập phải đủ tốt. Thêm vào đó, độ chính xác sẽ phụ thuộc nhiều vào thuật toán học máy được sử dụng, dẫn đến hiệu quả của nhiều phương pháp hiện nay chưa được cao.

2.4 Các nghiên cứu liên quan

Phần này trình bày về một số nghiên cứu liên quan và cơ sở kiến thức sử dụng trong các nghiên cứu đó. Các giải pháp này sử dụng các đặc tính về hành vi của botnet thông qua việc phân tích lưu lượng mạng và đồ thị giao tiếp giữa các máy để phát hiện các botnet ngang hàng. Một số tính chất được sử dụng trong các nghiên cứu được áp dụng trong khóa luận.

2.4.1 Enhanced PeerHunter: Phương pháp phát hiện các botnet ngang hàng thông qua hành vi cộng đồng ở cấp độ luồng mạng

2.4.1.1 Tổng quan về phương pháp

Enhanced PeerHunter [9] đề xuất một phương pháp phát hiện botnet ngang hàng hiệu quả dựa trên đặc tính hành vi cộng đồng. Giải thuật bắt đầu quá trình phát hiện P2P botnet bằng cách lọc những luồng mạng P2P. Một máy P2P thường liên lạc với những máy ngang hàng phân bố ở nhiều mạng khác nhau, vậy nên có thể phát hiện luồng mạng P2P dựa trên chỉ số đa dạng đích đến. Tiếp theo, khái niệm "liên hệ chung" (mutual contact) được sử dụng để xây dựng đồ thị. Các cặp bot trong cùng một botnet thường có xu hướng có một lượng lớn liên hệ chung và nó thường cao hơn rất nhiều so với những cặp máy không phải botnet hoặc những cặp máy thuộc hai botnet khác nhau. Bên cạnh những đặc trưng về giao tiếp như đã đề cập, Enhanced PeerHunter còn khai thác hai đặc trưng khác của cộng đồng P2P botnet có thể sử dụng để phân biệt giữa cộng đồng botnet với những máy thông thường trên đồ thị liên hệ chung: đặc trưng về lưu lượng (lưu lượng mạng được nhóm bởi kích thước gói tin vào và ra), và đặc trưng cấu trúc cộng đồng (với những kết nối qua liên hệ chung, các bot trong cùng một botnet có xu hướng tạo thành các bè phái). Enhanced PeerHunter có tỉ lệ phát hiện botnet thành công tuyệt đối và không có dương tính giả trong tập dữ liệu của họ.

2.4.1.2 Lý thuyết liên quan

Đặc điểm của mạng P2P

Do tính chất đặc trưng của mạng ngang hàng, lưu lượng mạng được tạo bởi các máy chủ trong mạng P2P cũng cho thấy những điểm riêng biệt. Ví dụ, một số lượng lớn peer ngang hàng sẽ thường xuyên tham gia và rời khỏi mạng (peer churn). Đây là một trong những đặc trưng rất phổ biến trong mạng P2P [15]. Điều này tạo ra một tỉ lệ tương đối cao về số kết nối thất bại trong mạng. Ngoài ra, các ứng dụng P2P thường lưu trữ sẵn và liên hệ với địa chỉ IP của các đồng nghiệp khác mà không cần đến các truy vấn DNS [16].

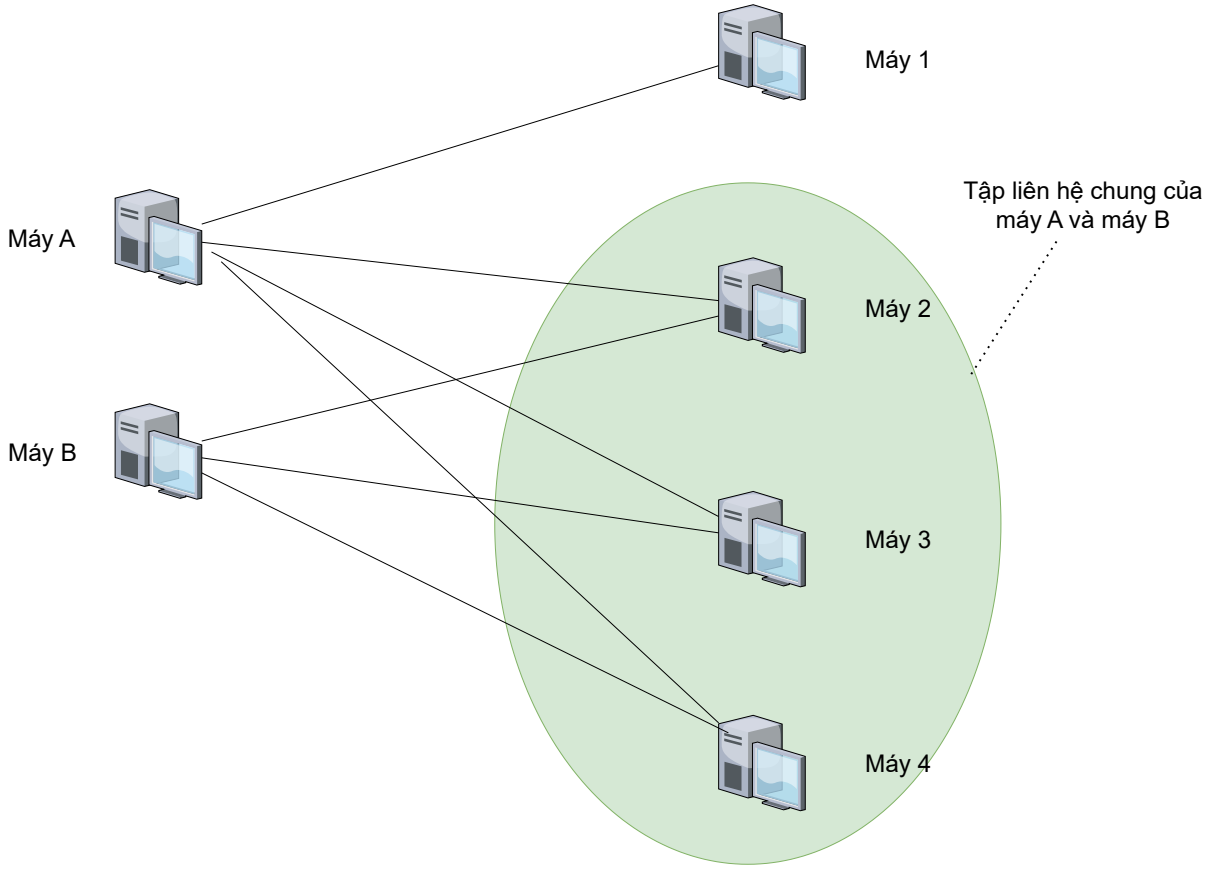
Một tính chất phổ biến khác của mạng ngang hàng là các máy thường nằm rải rác trong nhiều mạng thực tế khác nhau, dẫn đến mỗi máy chủ P2P thường xuyên liên hệ với một số lượng lớn các mạng vật lý [17]. Nghiên cứu của Zhang và các cộng sự [9] cũng chỉ ra rằng tính chất này tỏ ra hiệu quả trong việc phát hiện luồng mạng P2P hơn hai tính chất kể trên.

Liên hệ chung

Liên hệ chung (mutual contact) giữa hai máy là tập hợp các máy mà hai máy này đều có luồng mạng kết nối tới [18]. Hình 2.2 mô tả một ví dụ về liên hệ chung giữa hai máy A và B. Máy A có kết nối tới bốn máy {1, 2, 3, 4}, máy B có kết nối với ba máy {2, 3, 4}. Khi đó {2, 3, 4} là tập liên hệ chung của A và B.

Các liên hệ chung là một đặc điểm tự nhiên của mạng botnet P2P. Điều này dựa trên một tính chất cơ sở trong hoạt động của mạng botnet P2P là các bot sẽ thường xuyên giao tiếp với các đồng nghiệp để nhận lệnh, cập nhật và duy trì cấu trúc mạng. Mặc dù các bot khác nhau có thể sẽ giao tiếp với các đồng nghiệp khác nhau do các bot chọn ngẫu nhiên các peer ngang hàng, tuy nhiên có một thực tế rằng hầu như luôn có một vài liên hệ chung giữa hai máy trong mạng, nhất là khi hai máy đó cùng thuộc một mạng vật lý. Ngoài ra, các máy trong mạng botnet thường cùng nhau thực hiện cuộc tấn công. Những nạn nhân khi đó rất có thể là liên hệ chung của các bot trong mạng.

Đặc tính này cũng có thể có trong các ứng dụng P2P hợp pháp. Một cặp máy chủ hợp pháp có thể có một nhóm nhỏ các liên hệ chung nếu chúng giao tiếp với cùng một nhóm máy ngang hàng một cách tình cờ hoặc cùng giao tiếp với một số máy chủ phổ biến như Google, Facebook. Tuy nhiên, trên thực tế, các máy chủ P2P hợp pháp hoạt động với các mục đích khác nhau thường không kết nối tới cùng một nhóm máy ngang hàng.



Hình 2.2: Ví dụ về liên hệ chung giữa hai máy

Hành vi cộng đồng

Các bot trong cùng một mạng botnet ngang hàng luôn hoạt động cùng nhau như một cộng đồng. Một cộng đồng (hay một cụm) trong đồ thị là các đỉnh có xác suất được kết nối với nhau cao hơn so với các thành viên của các nhóm khác [19]. Mỗi cộng đồng khi đó sẽ thể hiện đặc điểm hành vi không giống nhau phụ thuộc vào các bot. Việc phát hiện các cộng đồng là đơn giản hơn việc phát hiện riêng lẻ từng bot. Ba đặc trưng về hành vi cộng đồng sau đây đã được sử dụng trong Enhanced PeerHunter [9]:

Đặc trưng về lưu lượng

Các phương pháp phát hiện botnet sử dụng các đặc trưng lưu lượng mạng đã được thảo luận rộng rãi trong nhiều nghiên cứu trước đây. Trong Enhanced PeerHunter [9], đặc trưng về kích thước gói của các lưu lượng mạng đã được sử dụng. Đối với các ứng dụng P2P, kích thước gói tin phụ thuộc vào nhu cầu của người dùng. Tuy vậy, các luồng mạng với mục đích quản lý thường cố định về kích thước gói. Các ứng dụng P2P và các botnet P2P, hay giữa các P2P hợp pháp khác nhau và P2P botnet khác nhau sẽ có những dải kích thước gói tin không giống nhau.

Đặc trưng về giao tiếp

Hai tính chất về đa dạng đích và liên hệ chung được sử dụng để thể hiện hành vi của cộng đồng botnet.

- *Tỉ lệ đa dạng đích trung bình:* Tỉ lệ này thể hiện hành vi P2P của mạng botnet ngang hàng. Do tính chất phi tập trung của mạng P2P, luồng mạng P2P có xu hướng có tỉ lệ đa dạng đích cao hơn luồng mạng không ngang hàng. Tỉ lệ đa dạng đích của mỗi luồng mạng ở đây được tính tương đối bằng tỉ lệ giữa số lượng 16 tiền tố IP đích khác nhau và tổng số lượng IP đích của luồng mạng đó. Trong mạng botnet ngang hàng, tỉ lệ đa dạng đích cũng có phần lớn hơn do các bot thường sẽ cố gắng lưu trữ ít nhất các địa chỉ của cùng mạng vật lý [20]. Ngược lại, đích đến của luồng mạng P2P hợp pháp thường phụ thuộc vào người dùng, điều này dẫn đến tỉ lệ đa dạng địa chỉ đích của chúng thay đổi rất nhiều tùy vào mục đích sử dụng.
- *Tỉ lệ liên hệ chung trung bình:* Tỉ lệ này thể hiện hành vi đặc trưng của các mạng botnet. Tỷ lệ liên hệ lẫn nhau (MCR) giữa một cặp máy chủ bằng số lượng liên hệ chung giữa chúng chia cho tổng số liên hệ riêng biệt của cả hai. MCR của một cặp bot trong cùng một mạng botnet thường cao hơn nhiều so với MCR của một cặp ứng dụng hợp pháp hoặc một cặp bot từ các mạng botnet khác nhau. Thêm vào đó, mỗi cặp bot trong cùng một mạng botnet thường có MCR tương tự nhau.

Đặc trưng về cấu trúc cộng đồng

Mỗi cặp bot trong cùng một botnet có xu hướng có một số lượng các liên hệ lẫn nhau. Nếu chúng ta coi mỗi máy là một đỉnh và tồn tại một cạnh giữa một cặp máy khi chúng có liên hệ chung, thì các bot trong cùng một mạng botnet có xu hướng hình thành các nhóm mà ở đó bất kỳ hai đỉnh nào cũng có liên kết đến nhau. Ngược lại, các địa chỉ liên hệ của các máy chủ hợp pháp khác nhau có thể không tồn tại điểm chung, do phụ thuộc lớn vào người dùng. Vì vậy, khả năng các cộng đồng hợp pháp hình thành các cụm nhóm như định nghĩa ở trên là tương đối thấp.

2.4.2 BotCluster: Hệ thống phân cụm botnet ngang hàng dựa trên phiên

2.4.2.1 Tổng quan về phương pháp

BotCluster [10] là một hệ thống phân cụm botnet ngang hàng dựa trên các phiên. Đầu tiên, các bản ghi lưu lượng mạng được gộp thành các phiên. Tiếp theo, lưu lượng

không phải P2P và lành tính sẽ được lọc ra bằng cách tham chiếu đến danh sách trắng và sử dụng tham số tỷ lệ luồng không có phản hồi (FLR). Sau đó những phiên có khả năng cao là có những hành vi của botnet ngang hàng được lọc ra và gom thành những nhóm dựa trên bốn thuộc tính cơ bản của P2P botnet. BotCluster đạt được độ chính xác trung bình là 97.58%.

2.4.2.2 Lý thuyết liên quan

Tỉ lề luồng không phản hồi (FLR)

Tỉ lệ luồng không phản hồi (FLR) là một độ đo trong phát hiện lưu lượng ngang hàng dựa trên mô hình luồng nhiều pha. Nó được tính bằng tỉ lệ giữa số lượng yêu cầu đến (incoming request) không nhận được phản hồi và tổng số lượng yêu cầu đến trong một khoảng thời gian nhất định. FLR tăng tương đương với việc có một số lượng lớn các gói tin mạng bị mất hoặc bị thất lạc trong quá trình truyền tải. Điều này có thể là do các botnet cố gắng chặn các luồng mạng hoặc làm giảm tốc độ truyền tải để che giấu hoạt động của chúng.

Mạng ngang hàng nói chung cũng có thể có tỉ lệ mất gói tin cao do hiện tượng "peer churn" [15]. Thêm vào đó, một số mạng hợp pháp có thể tăng tốc độ truyền tải bằng cách tăng tần số truyền hoặc giảm độ trễ, điều này có thể dẫn đến mất gói tin. Một số khác có thể được cấu hình để sử dụng các giao thức không đồng bộ, cũng là một nguyên nhân dẫn đến mất gói tin. Tuy nhiên, trong mạng ngang hàng hợp pháp, tỉ lệ mất gói tin thường thấp hơn so với các mạng ngang hàng bất hợp pháp được điều khiển bởi botnet.

Một số đặc trưng của P2P botnet

Trong BotCluster [10], bốn thuộc tính cơ bản sau của botnet ngang hàng đã được sử dụng:

- *Các phiên P2P của botnet có phân phối dày đặc trong các véc tơ đặc trưng của chúng và có thể được lọc qua phân cụm.* Ngược lại, các phiên P2P của ứng dụng hợp pháp có phân phối không đồng đều trong véc tơ đặc trưng do chúng không có chung các hành vi bất thường giống như các máy trong mạng botnet. Vectơ đặc trưng ở đây được hiểu là một biểu diễn toán học của các thuộc tính của phiên, chẳng hạn như kích thước gói tin, thời lượng và địa chỉ IP nguồn/đích.

- *Các bot ngang hàng thường xuyên liên hệ với các đồng nghiệp của nó và cập nhật*

thông tin. Đây là một đặc trưng cơ bản của mạng botnet ngang hàng, đã được đề cập trong phần 2.2. Việc liên lạc giữa các bot cũng cho phép botnet thực hiện các hoạt động phức tạp hơn mà có thể đòi hỏi sự đồng bộ hóa giữa các bot để thực hiện một cách hiệu quả, chẳng hạn như tấn công phân tán hoặc tấn công từ chối dịch vụ (DDoS). Các bot cũng có thể cập nhật thông tin về hệ thống, chẳng hạn như phiên bản hệ điều hành, trình duyệt web và các ứng dụng đang chạy, để giúp các hacker điều khiển botnet tìm kiếm các lỗ hổng bảo mật và khai thác chúng.

- *Các bot ngang hàng liên hệ với nhiều đồng nghiệp khác nhau trong mỗi lần thực hiện giao tiếp.* Mỗi máy trong mạng botnet ngang hàng sẽ có chứa một tệp nhị phân lưu danh sách các hàng xóm ngang hàng của chúng [20]. Danh sách này thường cố định về số bản ghi, và được cập nhật liên tục mỗi khi có máy mới tham gia hay biến mất khỏi mạng.

- *Các bot trong cùng một mạng botnet P2P có xu hướng thể hiện các kiểu hành vi tương tự mặc dù đã cố gắng làm xáo trộn hoặc che giấu các hoạt động của chúng.* Điều này là do các botnet P2P được thiết kế để hoạt động theo cách phi tập trung, với các bot giao tiếp với nhau để chia sẻ thông tin và nhận lệnh từ botmaster. Do thiết kế này, các bot trong cùng một mạng botnet P2P cần phải có một mức độ nhất quán nhất định trong các mẫu hành vi của chúng để đảm bảo rằng chúng có thể giao tiếp hiệu quả với nhau và thực hiện các nhiệm vụ của mình.

2.4.3 Xác định các cộng đồng botnet ngang hàng qua lưu lượng mạng sử dụng các độ đo cấu trúc cộng đồng

2.4.3.1 Tổng quan về phương pháp

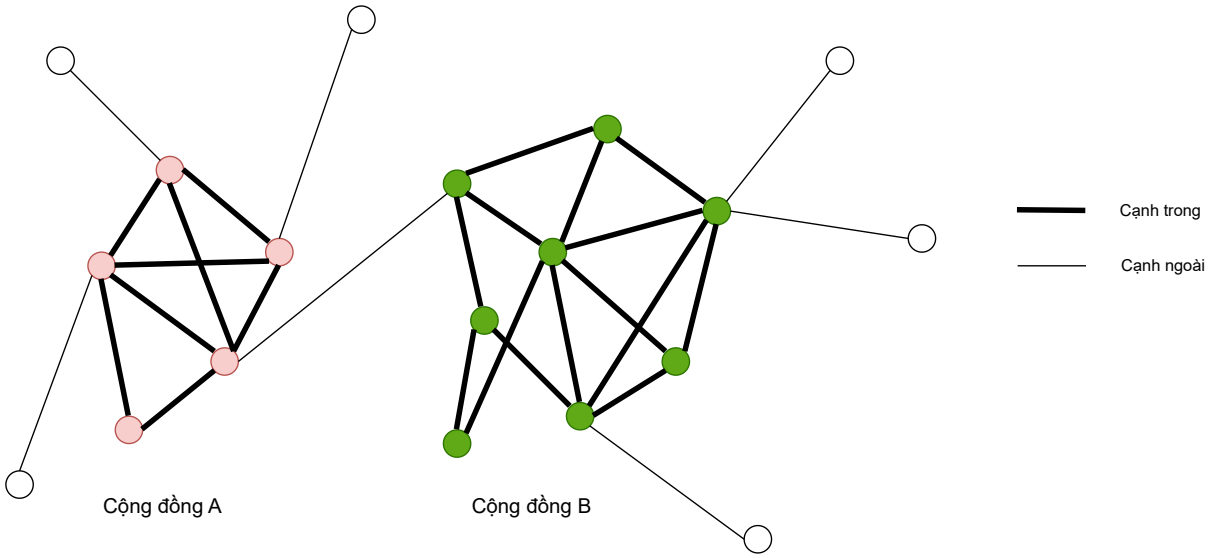
Joshi và các cộng sự [12] đã xây dựng đồ thị giao tiếp giữa các máy, sau đó sử dụng thuật toán Louvain để xác định các cộng đồng trong đồ thị. Joshi mô tả các đặc điểm hành vi giao tiếp của các mạng botnet P2P sử dụng một số độ đo cấu trúc cộng đồng trong đồ thị. Các độ đo này sau đó thông qua phân tích và thực nghiệm cho thấy một số trong đó như mức độ bậc nội bộ trung bình hay hệ số tương quan rất phù hợp để xác định các cộng đồng tương ứng trong biểu đồ giao tiếp là botnet hay ứng dụng hợp pháp. Phương pháp học không giám sát sau đó được áp dụng giúp tăng hiệu quả của mô hình, kết quả thực nghiệm cho thấy các mạng botnet ngang hàng lớn được phát hiện ngay cả với lưu lượng được giám sát chỉ 10%.

2.4.3.2 Lý thuyết liên quan

Phần này trình bày một số độ đo cấu trúc đồ thị được phân tích và sử dụng trong nghiên cứu của Joshi [12] để phân tách các ứng dụng và mạng botnet ngang hàng trong đồ thị liên hệ.

Bậc nội bộ trong cộng đồng

Với một cộng đồng C trong một đồ thị, bậc nội bộ (internal degree) của nó bằng hai lần số lượng cạnh nằm trong cộng đồng. Một cạnh được coi là thuộc một cộng đồng nếu hai đỉnh mà cạnh đó nối đều thuộc cộng đồng. Nếu một cạnh chỉ có một đầu mút thuộc cộng đồng, cạnh đó được coi là cạnh ngoài của cộng đồng [19]. Hình 2.3 mô tả một ví dụ về các cộng đồng trong đồ thị cùng với cạnh trong và cạnh ngoài của chúng. Cộng đồng A và B tương ứng sẽ có tổng bậc nội bộ lần lượt là 16 và 28.



Hình 2.3: Ví dụ về cộng đồng trong đồ thị

Các đồng nghiệp trong mạng ngang hàng thường liên hệ định kỳ với một số lượng lớn đồng nghiệp khác, điều này đặc biệt đúng với mạng botnet như đã chỉ ra ở phần 2.2. Nếu coi mỗi máy chủ là một đỉnh của đồ thị và giữa hai đỉnh có cạnh nối nếu hai máy đó trên thực tế có kết nối với nhau, khi đó peer trong một mạng ngang hàng sẽ tạo thành một cộng đồng, và các cộng đồng này có xu hướng có nhiều cạnh trong, dẫn đến bậc nội bộ trung bình của nó cao hơn các cộng đồng khác trong đồ thị.

Độ dẫn

Độ dẫn (conductance) là một độ đo được sử dụng để đánh giá độ tách biệt giữa các tập đỉnh trong một đồ thị [21]. Giá trị của nó càng thấp tức là tỉ lệ số lượng cạnh đi ra

ngoài tập đó so với số lượng cạnh nối giữa các đỉnh trong tập càng nhỏ, hay các đỉnh trong tập đó liên kết chặt chẽ với nhau và ít liên kết với các đỉnh khác bên ngoài. Độ dẫn được sử dụng trong các thuật toán phân cụm đồ thị, giúp tìm cách tách một đồ thị thành các tập đỉnh có tính chất liên kết cao bên trong mỗi tập và tách biệt với phần còn lại.

Một cộng đồng các nút phần lớn tiếp xúc với các nút khác trong cộng đồng so với phần còn lại của mạng, vì vậy nó có khả năng có độ dẫn thấp. Thêm vào đó, các cộng đồng P2P được cho là có độ dẫn thấp hơn cả do những tính chất cơ bản của mạng ngang hàng. Các cộng đồng botnet P2P cũng có khả năng sẽ có độ dẫn thấp hơn các cộng đồng P2P hợp pháp do nó thường thể hiện tính cộng đồng rõ rệt hơn.

Đường kính

Đường kính (diameter) trong đồ thị là độ dài dài nhất trong các đường đi ngắn nhất giữa các đỉnh [12]. Với một cộng đồng các nút, khái niệm đường kính được hiểu tương tự. Trong các giao tiếp thông thường trên mạng, thông tin được trao đổi trực tiếp giữa hai điểm cuối, vì vậy đường kính trong trường hợp này sẽ tương đối nhỏ. Ngược lại, các cộng đồng P2P dự kiến sẽ có đường kính lớn hơn các cộng đồng khác do đặc tính lan truyền thông tin trong mạng. Các cộng đồng botnet P2P cũng có khả năng sẽ có độ dài đường kính lớn hơn so với cộng đồng P2P hợp pháp do sự chuyển tiếp thông điệp trong mạng botnet được thực hiện một cách mạnh mẽ hơn.

Độ tương đồng

Một đồ thị được coi là thể hiện sự tương đồng nếu các đỉnh có xu hướng kết nối nhiều hơn với các đỉnh khác có cùng một số thuộc tính tương tự. Ngược lại, nếu các đỉnh thường kết nối với các đỉnh khác không giống nhau về thuộc tính, đồ thị đó không thể hiện sự tương đồng [22].

Bậc là một thuộc tính cơ bản của một đỉnh trong đồ thị và có thể sử dụng để đo sự tương đồng. Trong mô hình giao tiếp máy chủ - máy khách (server - client), phần lớn các máy kết nối với một số máy chủ như Google, Facebook dẫn đến một sự không tương đồng về bậc. Mặt khác, việc liên hệ với một số lượng đồng nghiệp nhất định trong mạng P2P, đặc biệt là mạng botnet ngang hàng, dẫn đến sự tương đồng tương đối cao về bậc trong đồ thị giao tiếp của chúng.

Công thức để tính độ tương đồng dựa trên bậc của các đỉnh trong đồ thị vô hướng lần đầu tiên được đề xuất trong nghiên cứu của Newman và các cộng sự [22]:

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2} \tag{2.1}$$

Trong đó, j, k là bậc thặng dư (excess degree) của đỉnh, có giá trị bằng bậc của các đỉnh đó trừ đi một, e_{jk} là tỉ lệ các cạnh nối hai đỉnh có bậc thặng dư là j và k, q_k là giá trị phân phối xác suất của bậc thặng dư k và σ_q là độ lệch chuẩn của phân phối q_k. q_k có thể được tính dựa vào giá trị phân phối bậc p_k và bậc trung bình z của đồ thị theo công thức sau:

$$q_k = \frac{(k+1)p_{k+1}}{z} \tag{2.2}$$

Giá trị của độ tương đồng r nằm trong khoảng $[-1, 1]$, trong đó, với $r = 1$ có nghĩa là các đỉnh có kết nối với nhau trong đồ thị đó hoàn toàn tương đồng. Ngược lại khi $r = -1$, không có bất kỳ sự tương đồng nào giữa các đỉnh được kết nối trong đồ thị.

2.4.4 Bàn luận về các phương pháp

Trong ba phương pháp được đề cập, Enhanced PeerHunter có hiệu quả tốt nhất mặc dù Enhanced PeerHunter không sử dụng nhiều đặc trưng của P2P botnet bằng BotCluster, đặc biệt là tính chất tần suất liên hệ giữa hai máy trong mạng botnet tương đối cao. Việc phát hiện các botnet trong BotCluster sử dụng sự giống nhau của các vectơ đặc trưng của phiên (một phiên là một nhóm các luồng mạng có cùng một số tính chất). Các luồng mạng trong cùng một phiên có thể thuộc về các mạng botnet khác nhau hoặc các ứng dụng khác nhau. Trong khi đó phương pháp dựa trên mức lưu lượng mạng trong Enhanced PeerHunter sẽ phân tách từng mạng botnet và ứng dụng hợp pháp ngay cả khi chúng hoạt động trên cùng một máy chủ. Các định nghĩa về sự đa dạng của điểm đến và liên hệ lẫn nhau cũng đại diện cho các hành vi độc đáo của botnet và có thể phân tách lưu lượng hợp pháp và độc hại một cách khá rõ ràng.

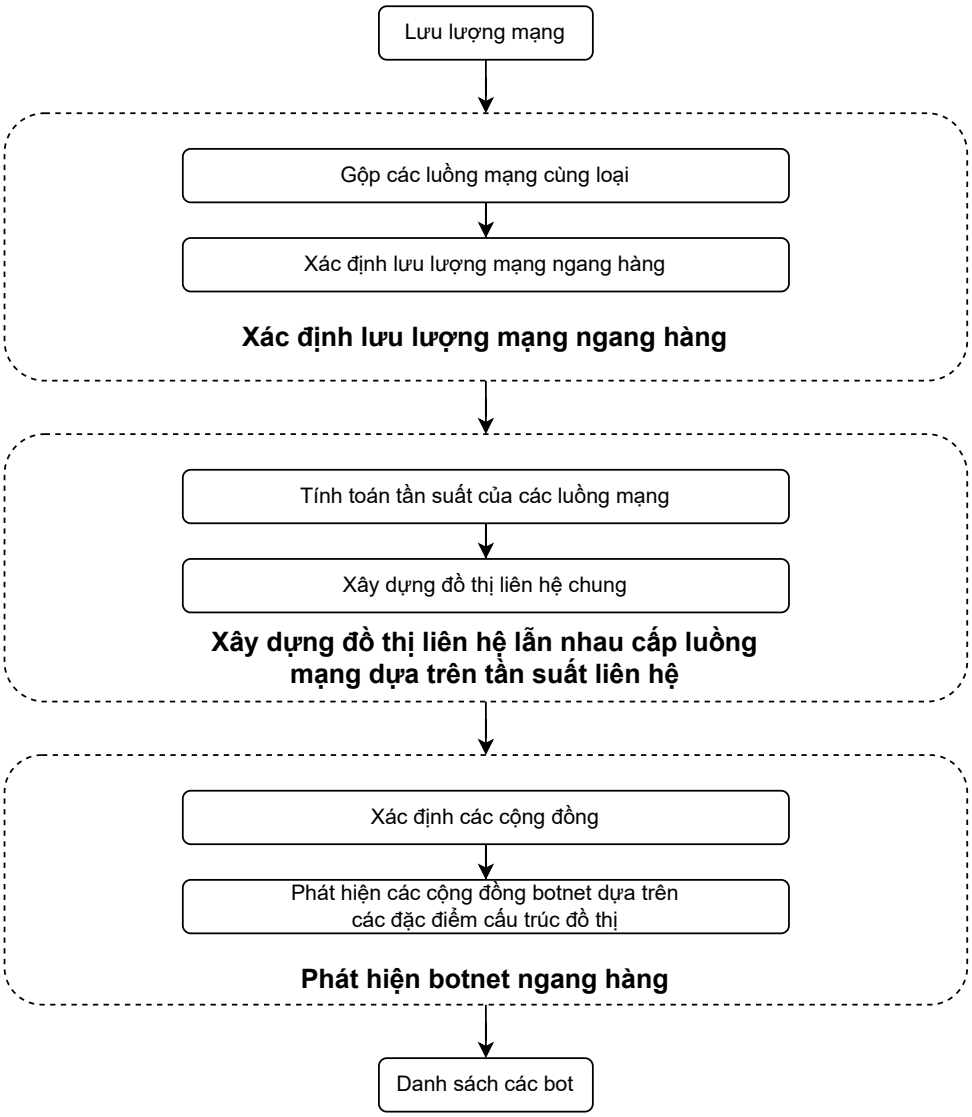
Tuy nhiên, việc Enhanced PeerHunter có khả năng phát hiện botnet với độ chính xác lên tới 100% một phần là do việc đánh giá hiệu năng của Enhanced PeerHunter được thực hiện trên một tập dữ liệu ít phức tạp. Bộ dữ liệu này chứa cả dữ liệu của bot P2P và ứng dụng P2P, nhưng không có ứng dụng hợp pháp nào có các đặc điểm đa dạng điểm đến và tiếp xúc lẫn nhau, dẫn đến việc thuật toán của họ hoạt động rất hiệu quả và không có kết quả dương tính giả. Trong nghiên cứu đã công bố trước đây của tôi và cộng sự [13], chúng tôi đã thực hiện kiểm nghiệm Enhanced PeerHunter với bộ dữ liệu khác khó hơn, kết quả là một số dương tính giả đã được nhìn thấy. Vì vậy, trong nghiên cứu đó, chúng tôi đã xây dựng một hệ thống mới dựa trên hệ thống cũ của Enhanced PeerHunter kết hợp với đặc tính tần suất liên hệ được chỉ ra trong BotCluster, cùng với đó là đặc tính kích thước gói tin liên tục thay đổi. Kết quả thực nghiệm trên hệ thống mới cho thấy nó hiệu quả hơn hẳn Enhanced PeerHunter về độ chính xác. Tuy vậy, hệ thống đã xây

dựng này vẫn còn những nhược điểm và có thể không hiệu quả trong một số bộ dữ liệu khó hơn nữa. Qua thu thập dữ liệu thực tế từ một ứng dụng ngang hàng hợp pháp với tính thách thức cao, việc phân tách lưu lượng ngang hàng hợp pháp và độc hại xuất hiện những sai số. Vì vậy, trong khóa luận này, tôi tiếp tục đề xuất những ý tưởng mới để cải tiến nó.

Một trong số đó dựa trên các đặc điểm cấu trúc đồ thị vì hệ thống đã đề xuất của chúng tôi phát hiện các botnet ngang hàng bằng cách xây dựng đồ thị liên hệ chung và phân tích hành vi botnet trên đồ thị này. Thêm vào đó, đã có một số nghiên cứu khác phát hiện botnet ngang hàng dựa trên các đặc tính cấu trúc trên đồ thị giao tiếp và đạt được những kết quả tốt, điển hình là đề xuất của Joshi và các cộng sự [12].

Chương 3

Hệ thống đề xuất của khóa luận



Hình 3.1: Tổng quan hệ thống

3.1 Tổng quan hệ thống

Hệ thống đề xuất của khóa luận được xây dựng dựa trên việc phân tích hành vi của botnet thông qua lưu lượng giao tiếp của botnet trong mạng. Hệ thống bao gồm ba mô-đun chính như trong hình 3.1. Ở mô-đun đầu tiên, tính chất đa dạng kích thước gói tin và đa dạng đích được sử dụng để xác định các lưu lương ngang hàng. Tiếp theo đó, ở thành phần thứ hai, tần suất liên hệ cao giữa các máy trong mạng botnet là tiền đề cho việc xây dựng đồ thị liên hệ chung cấp luồng mạng. Tại thành phần cuối cùng, giải thuật kết hợp đặc trưng về đa dạng đích và tính liên hệ chung với việc phân tích hành vi botnet qua các đặc điểm cấu trúc cộng đồng để phân tách các lưu lượng độc hại và hợp pháp.

Phần tiếp theo của chương này sẽ lần lượt trình bày chi tiết từng giải thuật, trong đó, giải thuật thứ ba (mô-đun phát hiện botnet) được chú trọng và là đề xuất chính của khóa luận này. Hai thành phần đầu tiên được sử dụng lại từ công trình nghiên cứu trước đó của tôi [13].

3.2 Xác định luồng mạng P2P

Mục tiêu của phần này là xác định những lưu lượng mạng nào là do các mạng ngang hàng tạo ra. Đầu vào cho mô-đun này là bản ghi các lưu lượng mạng và hai tham số ngưỡng, đầu ra là các lưu lượng mạng còn lại sau khi loại bỏ lưu lượng không phải của ứng dụng P2P.

3.2.1 Phân tích

Sự thay đổi kích thước gói tin

Các bot trong một mạng botnet thường xuyên giao tiếp với nhau để nhận lệnh, cập nhật và duy trì cấu trúc mạng ngang hàng. Ở một số mạng botnet ngang hàng, kích thước các gói tin được gửi đi trong những lần khác nhau thường thay đổi và đôi khi chỉ chênh nhau một khoảng nhỏ. Kỹ thuật này được sử dụng để tránh sự nghi ngờ và làm giảm khả năng bị phát hiện cho mạng botnet. Sality là một ví dụ điển hình cho tính chất này [23]. Qua quan sát thực nghiệm từ bộ dữ liệu, khóa luận cũng chỉ ra rằng sự thay đổi này là rất nhỏ giữa các kết nối cùng loại, do các mạng botnet đó đệm thêm một số byte ngẫu nhiên bất kỳ vào gói tin gốc. Áp dụng tính chất này, các luồng mạng đầu vào sẽ được nhóm lại và coi là cùng loại nếu chúng có cùng những thông tin cơ bản như địa chỉ IP nguồn, địa chỉ IP đích và giao thức, đồng thời có kích thước trung bình gói nằm trong cùng một khoảng cho trước.

Các luồng mạng sau khi nhóm sẽ sử dụng tính chất đa dạng đích, đã được đề cập trong phần 2.4.1.2 để phân tách lưu lượng P2P và không phải P2P. Bằng cách sử dụng một tham số ngưỡng, các luồng mạng với mức độ phân tán tới các mạng vật lý khác nhau tương đối cao sẽ được giữ lại. Nghiên cứu của Zhang [9] cũng đã cho thấy sự hiệu quả trong việc áp dụng đặc điểm này.

3.2.2 Chi tiết giải thuật

Chi tiết thuật toán được thể hiện trong giải thuật 1.

Nhóm luồng mạng cùng loại

Đầu vào từ tập dữ liệu là một tập các luồng mạng F dưới dạng các bản ghi, trong đó, mỗi luồng f_i đã được xử lý trước và bao gồm bộ năm phần tử đại diện cho địa chỉ IP nguồn ip_{src}, địa chỉ IP đích ip_{dst}, giao thức truyền tin $proto$, kích thước gói tin gửi đi và nhận về trung bình (bpp_{out} và bpp_{in}).

ip_{src}	ip_{dst}	proto	bpp_{out}	bpp_{in}
180.217.196.80	223.10.147.156	udp	82	76
180.217.196.80	223.10.147.156	udp	83	74
180.217.196.80	223.10.147.156	udp	88	78
180.217.196.80	223.10.147.156	udp	70	60
180.217.196.80	223.10.147.156	udp	74	67
180.217.196.80	36.78.177.141	udp	68	61
180.217.196.80	36.78.177.141	udp	65	87
180.217.196.80	36.78.177.141	udp	69	80
180.217.196.80	36.78.177.141	udp	63	88
180.217.196.80	78.190.127.96	udp	86	79
180.217.196.80	78.190.127.96	udp	87	70

ip_{src}	ip_{dst}	proto	bpp_{out}	bpp_{in}
180.217.196.80	223.10.147.156	udp	8	7
180.217.196.80	223.10.147.156	udp	8	7
180.217.196.80	223.10.147.156	udp	8	7
180.217.196.80	223.10.147.156	udp	7	6
180.217.196.80	223.10.147.156	udp	7	6
180.217.196.80	36.78.177.141	udp	6	6
180.217.196.80	36.78.177.141	udp	6	8
180.217.196.80	36.78.177.141	udp	6	8
180.217.196.80	36.78.177.141	udp	6	8
180.217.196.80	78.190.127.96	udp	8	7
180.217.196.80	78.190.127.96	udp	8	7

Hình 3.2: Ví dụ về nhóm các luồng mạng khi $\theta_f = 10$. Các bản ghi cùng màu tượng trưng cho các luồng mạng thuộc cùng một loại sau khi thực hiện xử lý dữ liệu

Đầu tiên, tham số ngưỡng kích thước gói tin Θ_f được sử dụng để tính toán lại giá trị độ dài gói tin trung bình gửi đi bpp_{out} và độ dài gói tin trung bình truyền tới bpp_{in} cho mỗi luồng mạng f_i. Giá trị của cả hai biến này được tiến hành chia cho ngưỡng Θ_f, sau đó giữ lại phần nguyên. Bằng cách này, hai luồng mạng có cùng hai điểm đầu cuối và giao thức giống nhau, kích thước trung bình mỗi gói tin ra và vào cùng trong khoảng $[a, a + \Theta_f)$, $[b, b + \Theta_f)$ (trong đó a, b là bội số của Θ_f) sẽ trở thành những luồng mạng giống hệt nhau. Hình 3.2 là một ví dụ minh họa của điều này.

Giải thuật này được tích hợp sử dụng khung phần mềm MapReduce [24] để giảm đáng kể thời gian tính toán khi tập dữ liệu đầu vào lớn. Mô-đun MAP thực hiện việc thay

đổi và nhóm các luồng mạng thành các bộ $(key, value)$ trong đó key bao gồm bộ bốn phần tử $[ip_{src}, proto, bpp_{out}, bpp_{in}]$ và $value$ là tập các ip_{dst} tương ứng, như được trình bày ở giải thuật 1.

Algorithm 1: Xác định luồng mạng P2P

Data: F - tập các lưu lượng mạng
Θ_f - ngưỡng kích thước gói tin
Θ_{dd} - ngưỡng đa dạng đích
Result: F^* - tập các lưu lượng mạng ngang hàng
Function MAP ($f_i = [ip_{src}, ip_{dst}, proto, bpp_{out}, bpp_{in}]$) **:**
  $newbpp_{out} \leftarrow \left[\frac{bpp_{out}}{\Theta_f}\right]$
  $newbpp_{in} \leftarrow \left[\frac{bpp_{in}}{\Theta_f}\right]$
  Key là $[ip_{src}, proto, newbpp_{out}, newbpp_{in}]$
  $Value$ là ip_{dst}
  emit $(Key, Value)$
Function REDUCE ($Key, Value[]$) **:**
  $dd_{key} = \varnothing$
  for *địa chỉ ip đích* $v \in Value[]$ **do**
    v_{16} là 16 bit đầu của địa chỉ ip v
    $dd_{key} \leftarrow dd_{key} \cup \{v_{16}\}$
  end
  if $\left\|dd_{key}\right\|$ *lớn hơn hoặc bằng* Θ_{dd} **then**
    for *địa chỉ ip đích* $v \in Value[]$ **do**
      emit (Key, v)
    end
  end

Xác định các lưu lượng ngang hàng

Thành phần này là hàm Reduce trong Giải thuật 1. Trong mô-đun này, số lượng tiền tố /16 khác nhau của địa chỉ IP đích với mỗi bộ bốn phần tử $[ip_{src}, proto, bpp_{out}, bpp_{in}]$ được tính toán. Nếu nó vượt quá ngưỡng đa dạng đích đến Θ_{dd}, những luồng mạng liên quan được coi là lưu lượng ngang hàng. Những luồng mạng còn lại được xác định là không phải của ứng dụng P2P và được loại bỏ. Dữ liệu được truyền tới mô-đun tiếp theo sẽ nhỏ hơn đáng kể.

3.3 Xây dựng đồ thị liên hệ lẫn nhau cấp luồng mạng dựa trên tần suất liên hệ

Mục tiêu của thành phần này là xây dựng một đồ thị liên hệ lẫn nhau cấp luồng mạng dựa trên tần suất liên hệ. Đầu vào cho mô-đun này là các bản ghi lưu lượng ngang hàng và một tham số ngưỡng tần suất, đầu ra là một đồ thị được xây dựng.

3.3.1 Phân tích

Đã có rất nhiều nghiên cứu trước đây xây dựng đồ thị giao tiếp làm cơ sở để phát hiện các mạng botnet ngang hàng như [25], [12] trong đó mỗi đỉnh của đồ thị sẽ ứng với một máy chủ, giữa hai đỉnh tồn tại cạnh nối nếu hai máy đó có kết nối với nhau. Cách xây dựng đồ thị như trên vẫn còn đơn giản và chưa tận dụng được các đặc trưng vốn có của botnet. Trong Enhanced PeerHunter [9], Zhang và cộng sự đã đề xuất một cách xây dựng đồ thị liên hệ mới sử dụng tính chất liên hệ chung (phần 2.4.1.2) và đặc trưng về kích thước gói (phần 2.4.1.2). Khi đó, mỗi đỉnh của đồ thị này sẽ là một cụm luồng mạng, được xác định bởi bộ bốn phần tử địa chỉ nguồn, giao thức, kích thước gói tin trung bình gửi đi và nhận vào. Giữa hai đỉnh sẽ có cạnh nối nếu như hai cụm luồng mạng này có liên hệ chung và có chung các đặc điểm về giao thức truyền cũng như kích thước gói tin. Liên hệ chung giữa hai cụm luồng mạng được hiểu tương tự như liên hệ chung giữa hai máy chủ. Việc xác định các đỉnh của đồ thị theo cách này giúp phân tách tương đối lưu lượng ứng dụng P2P hợp pháp và lưu lượng botnet P2P ngay cả khi chúng cùng hoạt động trên một máy tính vật lý do sự khác nhau về mặt kích thước gói tin giữa các loại luồng mạng ngang hàng. Vì vậy, cách xây dựng đồ thị như trên được áp dụng trong giải thuật của khóa luận.

Một điểm nữa, dựa theo tính chất đã trình bày trong phần 2.2, một botnet sẽ liên tục gửi lại các gói tin với tần suất cao để nhận lệnh, cập nhật thông tin và duy trì cấu trúc mạng. Vì vậy, những luồng mạng có tần suất liên lạc lại thấp sẽ có khả năng cao không phải là luồng mạng của botnet. Những luồng mạng như vậy sẽ bị loại bỏ khi xây dựng đồ thị. Chi tiết về đặc điểm này được giải thích rõ ràng hơn trong khóa luận [26].

3.3.2 Chi tiết giải thuật

Chi tiết thuật toán được thể hiện trong giải thuật 2.

Tính toán tần suất của các luồng mạng

Cho một tập các luồng mạng ngang hàng F, giải thuật thực hiện tính tần suất của mỗi luồng mạng khác nhau. Cấu trúc dữ liệu Map được sử dụng để lưu kết quả tính toán. Cấu trúc dữ liệu Map là một cấu trúc dữ liệu trong đó dữ liệu được lưu trữ dưới dạng các cặp chìa khóa - giá trị (key - value). Hình 3.3 thể hiện ví dụ về tính tần suất của các luồng mạng.

ip_{src}	ip_{dst}	proto	bpp_{out}	bpp_{in}
180.217.196.80	223.10.147.156	udp	8	7
180.217.196.80	223.10.147.156	udp	8	7
180.217.196.80	223.10.147.156	udp	8	7
180.217.196.80	223.10.147.156	udp	7	6
180.217.196.80	223.10.147.156	udp	7	6
180.217.196.80	36.78.177.141	udp	6	6
180.217.196.80	36.78.177.141	udp	6	8
180.217.196.80	36.78.177.141	udp	6	8
180.217.196.80	36.78.177.141	udp	6	8
180.217.196.80	78.190.127.96	udp	8	7
180.217.196.80	78.190.127.96	udp	8	7

ip_{src}	ip_{dst}	proto	bpp_{out}	bpp_{in}	**Tần suất**
180.217.196.80	223.10.147.156	udp	8	7	**3**
180.217.196.80	223.10.147.156	udp	7	6	**2**
180.217.196.80	36.78.177.141	udp	6	6	**1**
180.217.196.80	36.78.177.141	udp	6	8	**3**
180.217.196.80	78.190.127.96	udp	8	7	**2**

Hình 3.3: Ví dụ tính tần suất của các luồng mạng

Xây dựng đồ thị liên hệ chung

Mỗi cụm luồng mạng FB được đặc trưng bởi bộ bốn phần tử $[ip_{src}, proto, bpp_{out}, bpp_{in}]$. Các luồng mạng giống nhau về bốn phần tử này sẽ thuộc cùng một cụm luồng mạng.

Đầu tiên, giải thuật thực hiện tính toán tập địa chỉ IP đích tương ứng với mỗi cụm luồng mạng FB bằng cách duyệt qua từng luồng mạng và hợp các giá trị IP đích của các luồng mạng thuộc cùng một cụm.

Sau đó, đồ thị liên hệ chung được xây dựng với mỗi đỉnh là một cụm luồng mạng. Mỗi đỉnh này cũng được gán trọng số bằng tỉ lệ đa dạng đích trung bình của cụm luồng mạng đó, để làm tiền đề cho bước xác định botnet ngang hàng. Trọng số ddr_v của đỉnh

được tính qua công thức 3.1, trong đó FB_v là tập địa chỉ IP đích của cụm luồng mạng v, $DD(FB_v)$ là tập giá trị 16 tiền tố khác nhau trong FB_v.

$$ddr_v = \frac{\|DD(FB_v)\|}{\|FB_v\|} \tag{3.1}$$

Tập cạnh của đồ thị được xây dựng như sau:

- Với mỗi cặp đỉnh i, j tương ứng là các cụm luồng mạng, nếu hại cụm luồng mạng đó khác nhau về bộ ba phần tử $[proto, bpp_{out}, bpp_{in}]$, khi đó sẽ không tồn tại cạnh nối hai đỉnh i và j.
- Nếu hai cụm luồng mạng i và j giống nhau về bộ ba phần tử $[proto, bpp_{out}, bpp_{in}]$, bước tiếp theo của giải thuật thực hiện tính $total\_ip$ là số lượng phần tử của phép hợp từ hai tập địa chỉ IP đích của hai cụm luồng mạng đó, $same\_ip$ là số lượng phần tử của phép giao từ hai tập địa chỉ IP đích nhưng chỉ tính đến những địa chỉ đích có tần suất liên hệ với mỗi cụm luồng mạng lớn hơn hoặc bằng giá trị ngưỡng Θ_{fre} cho trước.
- Thực hiện tính toán giá trị $mrc_{i,j}$ như thể hiện trong công thức 3.2. Nếu giá trị này này lớn hơn không, đồng nghĩa với việc tồn tại liên hệ chung giữa hai cụm luồng mạng, khi đó sẽ tồn tại cạnh nối hai cụm luồng mạng i, j có trọng số chính bằng $mcr_{i,j}$.

$$mcr_{i,j} = \frac{same\_ip}{total\_ip} \tag{3.2}$$

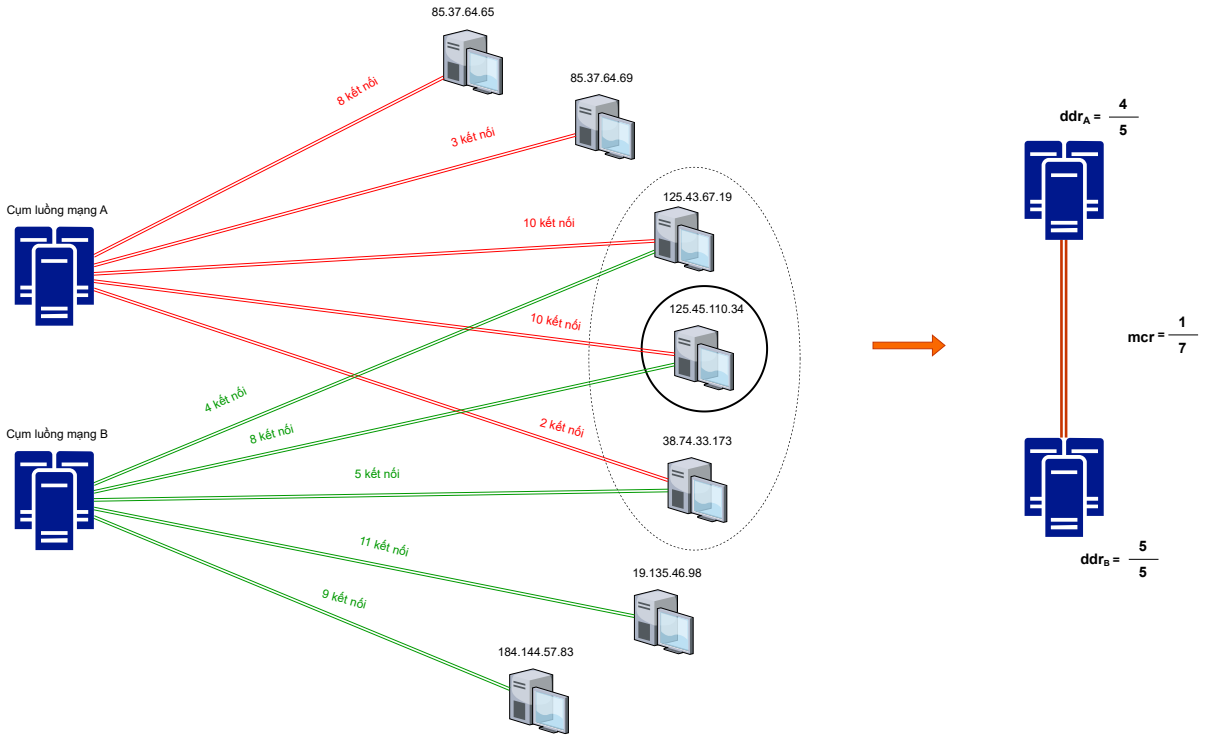
Hình 3.4 mô tả một ví dụ về xây dựng cạnh giữa hai cụm luồng mạng. Phần bên trong hình e-líp nét đứt là các liên hệ chung, phần e-líp nét liền là liên hệ chung thỏa mãn tần suất với mỗi cụm luồng mạng lớn hơn hoặc bằng Θ_{fre}.

3.4 Phát hiện botnet ngang hàng dựa trên các độ đo cấu trúc đồ thị

Mục tiêu của thành phần này là xác định cụ thể các bot ngang hàng dựa trên các đặc điểm cấu trúc đồ thị. Đầu vào cho mô-đun này là đồ thị liên hệ chung của các cụm luồng mạng và các tham số ngưỡng, đầu ra là danh sách các bot ngang hàng được phát hiện.

Algorithm 2: Xây dựng đồ thị liên hệ lẫn nhau cấp luồng mạng

Data: F - tập các lưu lượng mạng ngang hàng
Θ_{fre} - ngưỡng tần suất
Result: $G_{mc} = (V, E)$ - đồ thị liên hệ chung
fre là một CTDL map với khóa là luồng mạng và giá trị là tần suất
for *luồng mạng* $f \in F$ **do**
  if f *tồn tại trong tập khóa của* fre **then**
    $fre_f \leftarrow fre_f + 1$
  else
    Thêm khóa f cho fre
    $fre_f \leftarrow 1$
  end
end
FB là một CTDL map với khóa là cụm luồng mạng và dữ liệu là tập địa chỉ IP đích
for *luồng mạng* $f \in F$ **do**
  key là $[ip_{src}, proto, bpp_{out}, bpp_{in}]$ của f
  if key *tồn tại trong tập khóa của* FB **then**
    $FB_{key} \leftarrow FB_{key} \cup \{ip_{dst}\}$
  else
    Thêm khóa key cho FB
    $FB_{key} = \{ip_{dst}\}$
  end
end
$V = \varnothing, E = \varnothing$
for *cụm luồng mạng* $v \in$ *tập khóa của* FB **do**
  $ddr_v \leftarrow \frac{\|DD(FB_v)\|}{\|FB_v\|}$
  $V \leftarrow V \cup \{v\}$
end
for *cụm luồng mạng* $i, j \in$ *tập khóa của* FB *và* $i \neq j$ **do**
  `/* (i,j) và (j,i) là giống nhau và chỉ được tính một lần */`
  if i *và* j *có chung* $[proto, bpp_{out}, bpp_{in}]$ **then**
    $total\_ip \leftarrow \|FB_i \cup FB_j\|$
    $same\_ip \leftarrow 0$
    for *địa chỉ IP* $tmp \in FB_i \cap FB_j$ **do**
      $key1$ là luồng mạng kết hợp bởi cụm luồng mạng i với $ip_{dst} = tmp$
      $key2$ là luồng mạng kết hợp bởi cụm luồng mạng j với $ip_{dst} = tmp$
      if *cả* fre_{key1} *và* fre_{key2} *đều lớn hơn hoặc bằng* Θ_{fre} **then**
        $same\_ip \leftarrow same\_ip + 1$
      end
    end
    $mcr_{i,j} \leftarrow \frac{same\_ip}{total\_ip}$
    if $mcr_{i,j}$ *lớn hơn 0* **then**
      e là cạnh nối hai cụm luồng mạng i và j và có trọng số bằng $mcr_{i,j}$
      $E \leftarrow E \cup \{e\}$
    end
  end
end
return $G_{mc} = (V, E)$



Hình 3.4: Ví dụ xây dựng cạnh nối hai cụm luồng mạng khi $\Theta_{fre} = 8$

3.4.1 Phân tích

Mô-đun này bao gồm hai bước chính: phát hiện các cộng đồng trong đồ thị và xác định các cộng đồng là cộng đồng botnet.

Phát hiện các cộng đồng trong đồ thị

Tính năng phát hiện cộng đồng nhằm mục đích nhóm cùng một loại cụm luồng mạng ngang hàng vào cùng một cộng đồng cụm luồng mạng. Các máy trong cùng một mạng ngang hàng thường sinh ra các luồng mạng có tính chất tương đối giống nhau, vì vậy, các cụm luồng mạng thuộc cùng một cộng đồng khả năng cao thuộc cùng một mạng ngang hàng. Mỗi cụm luồng mạng sẽ thuộc về một cộng đồng cụm luồng mạng duy nhất, nhưng mỗi máy chủ có thể thuộc về nhiều cộng đồng khác nhau do có thể tồn tại nhiều ứng dụng ngang hàng hoạt động trên một máy. Ngoài ra, mỗi mạng botnet có thể chứa một số cộng đồng cụm luồng mạng khác nhau.

Để xác định các cộng đồng, giải thuật phát hiện cộng đồng của Louvain [27] đã được sử dụng. Giải thuật của Louvain phân tách các cộng đồng dựa vào mật độ giữa các cạnh nối các đỉnh trong đồ thị (mật độ cao của các cạnh trong cộng đồng và mật độ thấp của các cạnh giữa các cộng đồng khác nhau). Điều này là hoàn toàn phù hợp cho vấn đề phát hiện cộng đồng mạng botnet P2P. Ngoài ra, giải thuật này vượt trội so với nhiều

phương pháp khác về thời gian tính toán [27] và độ phức tạp tính toán tốt (phân tích một mạng thông thường gồm hai triệu nút mất khoảng hai phút [27]).

Xác định các cộng đồng là của botnet dựa trên các đặc điểm cấu trúc đồ thị

Có rất nhiều đặc điểm cấu trúc của đồ thị có thể được sử dụng để phân tách các cộng đồng botnet và hợp pháp trong đồ thị liên hệ chung cấp luồng mạng, hai trong số đó dựa vào tính chất đa dạng đích và liên hệ chung, đã được áp dụng hiệu quả bởi Zhang và các cộng sự [9]. Dựa vào cơ sở lý thuyết đã trình bày trong phần 2.4.1.2, các cộng đồng botnet ngang hàng cấp luồng mạng thường có tỉ lệ đa dạng đích trung bình và tỉ lệ liên hệ chung trung bình cao hơn các cộng đồng khác. Trong đồ thị được xây dựng như trình bày ở phần 3.3, hai độ đo trên chính là giá trị trung bình của trọng số đỉnh và giá trị trung bình của trọng số cạnh. Khóa luận này cũng sẽ áp dụng hai tính chất trên.

Thêm vào đó, phần 2.4.3.2 đề cập đến bốn độ đo cấu trúc đồ thị khác: bậc nội bộ trong cộng đồng, độ dẫn, đường kính và độ tương đồng giữa các đỉnh được kết nối. Joshi trong luận án tiến sĩ của mình [12] đã thử nghiệm bốn độ đo này trong việc xác định các cộng đồng botnet trên đồ thị giao tiếp. Kết quả thực nghiệm cho thấy chỉ có hai độ đo là bậc nội bộ trong cộng đồng và độ tương đồng tỏ ra hiệu quả. Mặt khác, từ đặc điểm của các độ đo, dễ dàng nhận thấy tính chất cho hai độ đo độ dẫn và đường kính không có nhiều ý nghĩa trên đồ thị liên hệ chung cấp luồng mạng. Vì vậy, khóa luận sẽ chỉ quan tâm đến hai độ đo còn lại, thực hiện phân tích để ứng dụng trong ngữ cảnh của đồ thị liên hệ lẫn nhau thay vì đồ thị giao tiếp.

Bậc nội bộ trong cộng đồng

Trong đồ thị giao tiếp, bậc nội bộ của các cộng đồng botnet có xu hướng cao hơn các cộng đồng khác do sự liên hệ định kỳ với một số lượng lớn đồng nghiệp. Trong ngữ cảnh của đồ thị liên hệ chung cấp luồng mạng, điều này vẫn đúng do trong mạng botnet, các lệnh, thông tin cập nhật hay duy trì cấu trúc mạng được chuyển tiếp sang một lượng lớn các peer khác, và các gói tin này có chung các đặc điểm về giao thức hay kích thước gói tin do có cùng mục đích sử dụng, từ đó sẽ hình thành lên những kết nối dày đặc trong đồ thị liên hệ chung cấp luồng mạng.

Vì các cộng đồng có thể có số lượng thành viên khác nhau, do đó để có tính tương quan giữa các cộng đồng, khóa luận định nghĩa bậc nội bộ trung bình trong cộng đồng theo công thức sau, trong đó, V, E tương ứng là tập đỉnh và tập cạnh của đồ thị:

$$AVGID = \frac{2 * \|E\|}{\|V\| * (\|V\| - 1)} \tag{3.3}$$

Độ tương đồng giữa các đỉnh

Một đồ thị được coi là thể hiện sự tương đồng nếu các đỉnh có xu hướng kết nối nhiều hơn với các đỉnh khác có cùng một số thuộc tính tương tự, chẳng hạn như là bậc của đỉnh. Nếu xét trong ngữ cảnh đồ thị giao tiếp của riêng các mạng ngang hàng nói chung, mạng botnet P2P nói riêng, giá trị sự tương đồng về bậc tương đối cao theo phần 2.4.3.2.

Trong ngữ cảnh của đồ thị liên hệ chung, các cụm luồng mạng botnet có cùng kích thước và giao thức truyền tin và cùng có kết nối đến một địa chỉ khác cũng có khả năng cao tương đồng về bậc. Ví dụ, khi một máy A trong mạng botnet muốn gửi thông điệp nhận được từ quản trị viên bot đến các máy khác, nó sẽ thực hiện trao đổi các gói tin về cơ bản là giống nhau tới các hàng xóm ngang hàng được lưu trữ sẵn. Khi đó, cụm luồng mạng khi xét tại máy chủ của các hàng xóm đó sẽ đều có cạnh nối trong đồ thị liên hệ chung, tạo nên một sự tương đồng về bậc. Ngoài ra, nhóm này có xác suất cao sẽ không tồn tại các bậc ngoại lai riêng lẻ do các cụm luồng mạng khác tuy cũng có khả năng kết nối tới máy A, nhưng vì đặc thù khác nhau trong mục đích giao tiếp, kích thước gói tin và giao thức sẽ không tương đồng, dẫn tới không có cạnh nối trên đồ thị liên hệ chung cấp luồng mạng.

Công thức 2.1 được sử dụng để tính độ tương đồng về bậc giữa các đỉnh được kết nối của một đồ thị. Để áp dụng được trong ngữ cảnh giải thuật của khóa luận, cần phải tính được độ tương đồng ở trong một cộng đồng nhất định, nhưng vẫn xét đến những kết nối từ cộng đồng đó ra ngoài. Để thực hiện điều này, ta sẽ tìm cách tính độ tương đồng một cách riêng lẻ trên từng đỉnh của đồ thị, khi đó độ tương đồng của một cộng đồng khi xét trong ngữ cảnh đồ thị lớn có thể coi bằng tổng độ tương đồng của các đỉnh thuộc nó. Piraveenan và các cộng sự [28] đã đề xuất công thức tính độ tương đồng trên từng đỉnh như sau:

$$r_v = \frac{j(j+1)(\bar{k} - \mu_q)}{2M\sigma_q^2} \tag{3.4}$$

Trong đó, j là bậc thặng dư (excess degree), $\bar{k}$ là giá trị trung bình bậc thặng dư của các đỉnh hàng xóm của đỉnh tương ứng, q_k là giá trị phân phối xác suất của bậc thặng dư k, μ_q và σ_q là kỳ vọng và độ lệch chuẩn của phân phối q_k.

Tuy vậy, nếu tính độ tương đồng của một cộng đồng nhất định trong ngữ cảnh của đồ thị lớn, kết quả tính toán có thể không phản ánh chính xác, do một cộng đồng chỉ là một phần rất nhỏ của đồ thị to, và sẽ bị ảnh hưởng rất nhiều bởi phần còn lại. Vì vậy,

khóa luận đề xuất tính toán độ tương đồng của cộng đồng trong ngữ cảnh thành phần liên thông chứa cộng đồng đó. Một thành phần liên thông của một đồ thị vô hướng là một đồ thị con trong đó giữa bất kì hai đỉnh nào đều tồn tại một đường đi trực tiếp hoặc qua các đỉnh khác đến nhau.

Hệ số biến thiên đa dạng đích

Ngoài độ đo tương đồng dựa trên bậc, khóa luận đề xuất một độ đo mới dựa trên sự tương đồng về số lượng địa chỉ đích của các luồng mạng. Dễ dàng nhận thấy bậc trong đồ thị giao tiếp, khi xét trong ngữ cảnh đồ thị liên hệ chung, chính bằng số địa chỉ IP khác nhau mà cụm luồng mạng kết nối tới. Thêm vào đó, mỗi máy trong mạng botnet ngang hàng lưu trữ một danh sách hàng xóm ngang hàng với số lượng phần tử cố định. Ví dụ, mỗi bot GameOver Zeus có một danh sách gồm 60 người hàng xóm và liên lạc với những người hàng xóm đó sau mỗi 30 phút, trong khi đó, Sality có danh sách ngang hàng lớn hơn bao gồm 1000 bản ghi [20]. Mặt khác, các ứng dụng ngang hàng không có quy chuẩn nào về số lượng IP sẽ liên hệ của mỗi máy trong mạng, vì vậy những mạng này khả năng sẽ có sự tương đồng không cao.

Để tính toán độ đo này, khóa luận sử dụng công thức hệ số biến thiên xung quanh giá trị trung bình của một quần thể (coefficient of variation) [29] áp dụng trên tập số lượng các tiền tố /16 khác nhau trong tập địa chỉ IP đích của mỗi cụm luồng mạng trong mỗi cộng đồng theo công thức:

$$c_v = \frac{\sigma}{\mu} \tag{3.5}$$

Trong đó, μ và σ tương ứng là giá trị trung bình và độ lệch chuẩn của quần thể. Giá trị c_v càng gần 0 khi sự tương đồng càng thể hiện rõ.

Việc sử dụng tập 16 tiền tố khác nhau thay vì cả tập địa chỉ IP đích vì các bot thường sẽ cố gắng liên lạc với ít máy nhất trong cùng một mạng, và nếu không thể kết nối tới một máy, nó nhiều khả năng sẽ tìm một máy khác trong mạng cũng là đồng nghiệp ngang hàng để thay thế trong danh sách hàng xóm. Điều này dẫn đến số lượng 16 tiền tố khác nhau trong các địa chỉ IP đích của các máy trong cùng mạng botnet sẽ ít biến động hơn ngay cả khi xảy ra hiện tượng "peer churn".

3.4.2 Chi tiết giải thuật

Chi tiết thuật toán được thể hiện trong giải thuật 3.

Xác định các cộng đồng

Đầu vào cho thành phần này là đồ thị liên hệ chung cấp luồng mạng G_{mc} được xây dựng trong mô-đun 3.3. Thuật toán Louvain được sử dụng để xác định các cộng đồng trên đồ thị này.

Phát hiện các cộng đồng botnet dựa trên các đặc điểm cấu trúc đồ thị

Đầu tiên, với mỗi thành phần liên thông trong đồ thị G_{mc}, giải thuật thực hiện tính toán chỉ số tương đồng về bậc trên từng đỉnh của thành phần liên thông đó.

Sau đó, với mỗi cộng đồng com_i, các chỉ số tỉ lệ đa dạng đích trung bình $avgddr_i$, tỉ lệ liên hệ chung trung bình $avgmcr_i$, bậc nội bộ trung bình $avgid_i$, độ tương đồng bậc lad_i, hệ số biến thiên đa dạng đích cv_i được tính toán như trong giải thuật 3. Trong đó ddr_j là tỉ lệ đa dạng đích hay trọng số của đỉnh j, $mcr_{j,k}$ là tỉ lệ liên hệ chung hay trọng số cạnh giữa hai đỉnh j và k, LA_j là giá trị độ tương đồng tính trên đỉnh j, $\mu_{\|DD_{V_{com_i}}\|}$ và $\sigma_{\|DD_{V_{com_i}}\|}$ lần lượt là giá trị trung bình và độ lệch chuẩn của số lượng các tiền tố /16 khác nhau trong tập địa chỉ IP đích của mỗi cụm luồng mạng trong cộng đồng.

Các cộng đồng có tỉ lệ đa dạng đích trung bình, tỉ lệ liên hệ chung trung bình, bậc nội bộ trung bình, độ tương đồng bậc lớn hơn các ngưỡng tương ứng và hệ số biến thiên đa dạng đích nhỏ hơn ngưỡng tương đồng đa dạng đích được coi là cộng đồng botnet ngang hàng.

Ở bước tiếp theo, các cụm luồng mạng trong các cộng đồng này được duyệt để xác định chính xác các bot.

Algorithm 3: Xác định P2P botnet

Data: $G_{mc} = (V, E)$ - đồ thị liên hệ chung
Θ_{avgddr} - ngưỡng đa dạng đích trung bình
Θ_{avgmcr} - ngưỡng liên hệ chung trung bình
Θ_{avgid} - ngưỡng bậc nội bộ trung bình
Θ_{lad} - ngưỡng độ tương đồng bậc
Θ_{cv} - ngưỡng độ tương đồng đa dạng đích
Result: S_{bot} - tập các bot ngang hàng
$Com \leftarrow$ Louvain(G_{mc})
LA là CTDL map với khóa là tập V và giá trị là độ tương đồng
for *mỗi thành phần liên thông* $ConCom$ *trong* G_{mc} **do**
  $LA \leftarrow LA \cup$ LocalAssortativity$(ConCom)$
end
$S_{botCom} = \varnothing$
for *cộng đồng* $com_i \in Com$ **do**
  $avgddr_i \leftarrow \frac{\sum_{j \in V_{com_i}} ddr_j}{\left\|V_{com_i}\right\|}$
  $avgmcr_i \leftarrow \frac{2 * \sum_{j,k \in E_{com_i}} mcr_{j,k}}{\left\|V_{com_i}\right\| * \left(\left\|V_{com_i}\right\| - 1\right)}$
  $avgid_i \leftarrow \frac{2 * \left\|E_{com_i}\right\|}{\left\|V_{com_i}\right\| * \left(\left\|V_{com_i}\right\| - 1\right)}$
  $lad_i \leftarrow \sum_{j \in V_{com_i}} LA_j$
  $cv_i \leftarrow \frac{\sigma_{\left\|DD_{V_{com_i}}\right\|}}{\mu_{\left\|DD_{V_{com_i}}\right\|}}$
  if $avgddr_i$ *lớn hơn* Θ_{avgddr}
   && $avgmcr_i$ *lớn hơn* Θ_{avgmcr}
   && $avgid_i$ *lớn hơn* Θ_{avgid}
   && lad_i *lớn hơn* Θ_{lad}
   && cv_i *nhỏ hơn* Θ_{cv} **then**
    $S_{botCom} \leftarrow S_{botCom} \cup \{com_i\}$
  end
end
$S_{bot} = \varnothing$
for *cộng đồng* $com_i \in S_{botCom}$ **do**
  for *cụm luồng mạng* $fb \in S_{botCom}$ **do**
    $S_{bot} \leftarrow S_{bot} \cup \{ip_{src}\}$
  end
end
return S_{bot}

Chương 4

Thực nghiệm

Chương 5 của khóa luận sẽ trình bày về những thực nghiệm của hệ thống đề xuất và đánh giá về những kết quả thực nghiệm đó. Mã nguồn chương trình và tập dữ liệu sử dụng cho thực nghiệm được công khai tại đường dẫn https://github.com/ginsama01/Improved-PeerCatcher-Graduation-Thesis.

4.1 Môi trường thực nghiệm

Bảng 4.1: Cấu hình máy tính

Cấu hình máy tính thực nghiệm	
CPU	Intel Core i5 8265U
RAM	12GB
Ổ cứng	256 GB SSD M.2 SATA 3
GPU	NVIDIA Geforce MX150
OS	Ubuntu 22.04 64bit

Bảng 4.2: Bảng các công cụ và môi trường

STT	Tên công cụ	Chú thích
1	Java JDK 1.8	Môi trường phát triển
2	Python 3.10.6	Môi trường phát triển
3	IntelliJ IDEA Ultimate	Môi trường lập trình
4	Visual Studio Code	Môi trường lập trình
5	Tcpdump	Công cụ dòng lệnh giúp phân tích và ghi lại lưu lượng mạng
6	CICFlowMeter	Công cụ mã nguồn mở giúp chuyển đổi dữ liệu pcap thành csv
7	VMware Work Station	Môi trường máy ảo

4.2 Xây dựng tập dữ liệu thực nghiệm

Tập dữ liệu sử dụng trong khóa luận này được thu thập từ nhiều nguồn khác nhau nhằm tăng tính công bằng và thách thức cho việc thực nghiệm.

4.2.1 Dữ liệu từ nguồn công khai

Lưu lượng mạng P2P hợp pháp

Khóa luận sử dụng lưu lượng mạng từ Đại học Georgia [17], gồm lưu lượng mạng của 5 ứng dụng P2P hợp pháp phổ biến: eMule, FrostWire, uTorrent, Vuze, Skype, được thu thập trong khoảng ba tuần.

Lưu lượng mạng P2P botnet

Một phần lưu lượng mạng botnet sử dụng là từ Đại học Georgia [17], chứa lưu lượng mạng 24 giờ của 13 máy chủ Storm và 3 máy chủ Waledac. Phần còn lại là từ Đại học SouthFlorida [30], chứa lưu lượng mạng trong 24 giờ của 5 bot Sality, 8 bot Kelihos và 8 bot ZeroAccess.

Lưu lượng mạng nền

Lưu lượng mạng nền sử dụng có nguồn gốc từ kho lưu trữ lưu lượng truy cập của MAWI [31], bao gồm dấu vết mạng 24 giờ của các máy tính trên đường trục chính WIDE xuyên Thái Bình Dương vào ngày 10/12/2014. Bộ dữ liệu chứa xấp xỉ 407.523.221 luồng mạng và 48.607.304 địa chỉ IP khác nhau. Tuy nhiên, chỉ 10000 địa chỉ IP nguồn và các luồng mạng tương ứng sẽ được sử dụng trong khóa luận này.

4.2.2 Dữ liệu tự thu thập

Ngoài các lưu lượng mạng từ các nguồn công khai, tôi đã thiết lập môi trường để ghi lại lưu lượng mạng hợp pháp của một ứng dụng tải tệp ngang hàng có tên là Deluge. Quá trình thu thập được thực hiện như sau:

- Đầu tiên, ba máy ảo Ubuntu 16.04 được thiết lập trên môi trường của VMWare Work Station. Ba máy ảo này đều được cài phần mềm tải tệp ngang hàng Deluge. Để thuận tiện cho việc ghi lại các lưu lượng mạng, tôi tiến hành cấu hình để Deluge chỉ hoạt động trên cổng 7000 và 7001.

- Sau đó, công cụ dòng lệnh *tcpdump* được sử dụng trên máy tính thật, ghi lại thông tin về các luồng mạng đi qua các máy ảo trên cổng 7000 và 7001.

- Quá trình thu thập được chạy liên tục trong ba ngày. Ngày đầu tiên, ứng dụng Deluge được sử dụng để tải các file giống hệt nhau ở cả ba máy. Ngày thứ hai, các file được tải xuống ở các máy có một phần giống nhau. Ngày thứ ba, mỗi máy sẽ tải những file riêng biệt không giống nhau. Việc thực hiện như vậy nhằm tăng tính khách quan cho bộ dữ liệu.

- Dữ liệu sau đó được *tcpdump* lưu vào một tệp định dạng *pcap*, sau đó được tiền xử lý trước khi tiến hành sử dụng cho thực nghiệm.

4.2.3 Tiền xử lý dữ liệu

Các lưu lượng mạng đều được lưu dưới định dạng tệp *pcap*, do đó, để có thể tiến hành thực nghiệm, dữ liệu cần được chuyển về dạng tệp văn bản. Sau đó, chỉ những trường dữ liệu cần thiết cho quá trình thực nghiệm sẽ được giữ lại để tối ưu hóa bộ nhớ lưu trữ.

Một phần trong số các dữ liệu đầu vào, bao gồm lưu lượng mạng 24 giờ của 4 ứng dụng ngang hàng eMule, FrostWire, uTorrent, Vuze, lưu lượng mạng của 5 P2P botnet Kelihos, Waledac, Storm, Sality, ZeroAccess và lưu lượng mạng nền từ MAWI đã được xử lý và công khai bởi Zhang và các cộng sự [9].

Phần còn lại, gồm lưu lượng mạng trong 24 giờ của 9 máy chủ chạy Skype và lưu lượng 72 giờ của 3 máy chủ chạy Deluge sẽ được đưa qua công cụ mã nguồn mở *CICFlowMeter* để chuyển từ định dạng *pcap* thành định dạng *csv*. Sau đó, một đoạn mã nguồn python được tôi triển khai để chuyển dữ liệu thành định dạng văn bản và chỉ giữ

lại các trường cần thiết là địa chỉ nguồn, địa chỉ đích, giao thức, kích thước gói tin gửi đi trung bình, kích thước gói tin nhận trung bình. Ngoài ra, do tất cả các lưu lượng mạng, ngoại trừ của ứng dụng Deluge đều là trong 24 giờ, vì vậy, lưu lượng 72 giờ của 3 máy chủ Deluge sẽ được phân tách theo từng ngày, tạo ra 9 máy chủ Deluge hoạt động trong 24 giờ.

Hệ thống đề xuất của khóa luận trên thực tế sẽ được triển khai tại các thiết bị giám sát mạng, các cổng chuyển mạng switch, route hay hệ thống tường lửa trong hệ thống mạng của tổ chức. Với việc triển khai như vậy, lưu lượng mạng sẽ tạo thành cấu trúc hai bên và chỉ những kết nối từ bên trong mạng nội bộ ra ngoài hoặc ngược lại được ghi nhận. Vì vậy, để đảm bảo tính thực tế, các kết nối giữa các máy của cùng loại ứng dụng hợp pháp hay botnet sẽ bị loại bỏ, vì chúng được thu thập khi cùng chạy trong một mạng nội bộ. Ngoài ra, như đã đề cập ở phần trên, chỉ có 10000 địa chỉ IP trong dấu vết mạng nền được sử dụng trong khóa luận, tất cả các lưu lượng kết nối giữa 10000 máy này với nhau cũng được loại bỏ.

Để tăng thêm tính thách thức cho bộ dữ liệu, địa chỉ IP của các máy chủ chạy ứng dụng P2P hợp pháp và botnet sẽ được ánh xạ ngẫu nhiên vào các giá trị trong tập 10000 địa chỉ IP của dấu vết mạng nền. Tất cả các công việc trên đều được tôi tự thực hiện sử dụng ngôn ngữ Python.

Bảng 4.3 thể hiện tổng số máy chủ của từng loại ứng dụng và botnet ngang hàng cũng như tổng số bản ghi lưu lượng tương ứng.

Bảng 4.3: Thống kê về bộ dữ liệu

Loại ứng dụng	Số máy chủ	Số bản ghi luồng mạng
eMule	16	3581543
FrostWire	16	3287937
uTorrent	14	7523902
Vuze	14	6189647
Skype	9	2088185
Deluge	9	888956
Kelihos	8	122182
Waledac	3	1109508
Storm	13	8603399
Sality	5	5599435
ZeroAccess	8	709299

4.3 Kết quả thực nghiệm và đánh giá

4.3.1 Kết quả thực nghiệm và đánh giá về xác định lưu lượng ngang hàng

Phần này sẽ đánh giá mức độ hiệu quả của việc lọc các lưu lượng ngang hàng. Mục tiêu chính của việc lọc lưu lượng P2P trong mô-đun 3.2 là giảm thiểu số máy chủ của tập dữ liệu và các luồng mạng trong khi vẫn đảm bảo các máy chạy P2P được gán nhãn cùng phần lớn lưu lượng mạng của chúng vẫn được giữ lại. Điều này sẽ giúp giảm phần lớn thời gian tính toán cho các mô đun sau của giải thuật. Việc giảm tỉ lệ dương tính giả trong phát hiện các máy chủ chạy ứng dụng ngang hàng không phải là yếu tố quan trọng nhất của thành phần này, vì chúng khả năng cao cũng sẽ bị lọc ở các mô-đun sau đó.

Khóa luận tiến hành thực nghiệm đánh giá cho thành phần xác định lưu lượng P2P qua nhiều thông số ngưỡng Θ_{dd} khác nhau với thông số ngưỡng kích thước gói tin Θ_f được đặt cố định bằng 10. Quan sát từ tập dữ liệu cũng như thực tế đã được kiểm nghiệm từ nghiên cứu đã công bố trước đó [13] cho thấy rằng ngưỡng $\Theta_f = 10$ là phù hợp và mang lại kết quả tốt nhất.

Bảng 4.4: Kết quả thực nghiệm đánh giá mô đun xác định lưu lượng P2P với $\Theta_f = 10$

Θ_{dd}	Số lượng máy chạy ứng dụng P2P đã gán nhãn được phát hiện (/115)	Số lượng máy được xác định có chạy ứng dụng P2P (/10000)	Tỉ lệ phần trăm lưu lượng mạng còn lại của các máy chạy ứng dụng P2P được gán nhãn (%)
5	115	470	99.50
10	115	274	98.84
20	115	178	97.87
30	115	158	97.07
50	115	142	95.95
70	115	128	95.11
100	113	121	94.06
200	104	104	91.31

Bảng 4.4 cho thấy kết quả thực nghiệm của việc xác định các lưu lượng ngang hàng với giá trị $\Theta_f = 10$. Với các giá trị Θ_{dd} trong khoảng $[5, 70]$, tất cả các máy chủ được gán nhãn là tồn tại lưu lượng ngang hàng đều được phát hiện với tỉ lệ phần trăm lưu lượng mạng của các máy đó được giữ lại tương đối cao (> 95%). Vì trong dữ liệu bản ghi các luồng mạng của ứng dụng P2P, một số trong đó không phải là lưu lượng ngang hàng do

trong quá trình thu thập dữ liệu, vẫn có khả năng một lượng nhỏ các ứng dụng client-server vẫn hoạt động ngầm và lưu lượng của chúng được ghi nhận vào. Vì vậy, con số 95% lưu lượng mạng của các máy chạy ứng dụng P2P được giữ lại là chấp nhận được.

Khi giá trị Θ_{dd} lớn, có một số máy chạy ứng dụng P2P đã không được phát hiện, ví dụ như khi $\Theta_{dd} = 200$, có tới 11 máy chủ như vậy. Việc xác định thiếu làm ảnh hưởng lớn đến quá trình phát hiện P2P botnet do có thể chúng đã bị lọc ngay từ bước này.

Một thông số đáng lưu tâm khác là số lượng máy chủ được xác định có chạy ứng dụng P2P trên tổng số 10000 máy chủ của tập dữ liệu. Khi $\Theta_{dd} < 20$, số lượng máy như vậy là tương đối cao. Trên thực tế, trong lưu lượng mạng nền được sử dụng, khả năng tồn tại những máy chủ có chứa lưu lượng ngang hàng là có thể xảy ra nhưng phần lớn sẽ không thể hiện đầy đủ tính chất của ứng dụng ngang hàng do dấu vết mạng nền này được thu thập trên một đường trục mạng xuyên Thái Bình Dương trong khi tồn tại rất nhiều đường trục liên kết như vậy.

Từ những phân tích trên, các giá trị Θ_{dd} trong khoảng $[20, 70]$ được đánh giá là phù hợp nhất để thành phần xác định lưu lượng mạng ngang hàng hoạt động tốt. Giá trị $\Theta_{dd} = 30$ sẽ được sử dụng trong các thử nghiệm tiếp theo của khóa luận.

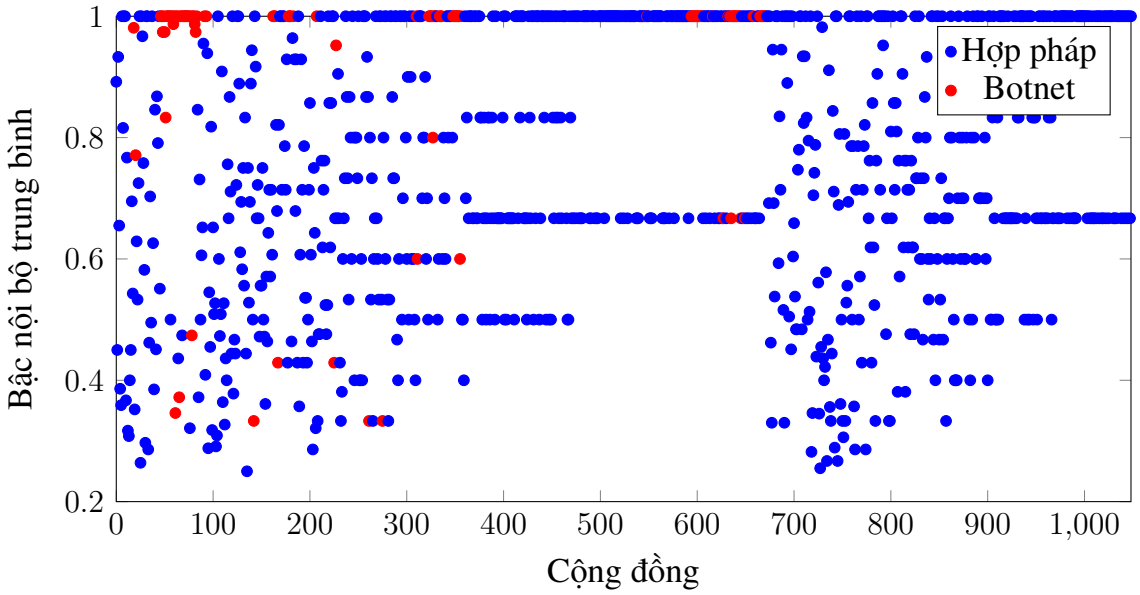
4.3.2 Kết quả thực nghiệm và đánh giá về phát hiện botnet

4.3.2.1 Đánh giá khả năng phát hiện các cộng đồng botnet của ba độ đo được đề xuất

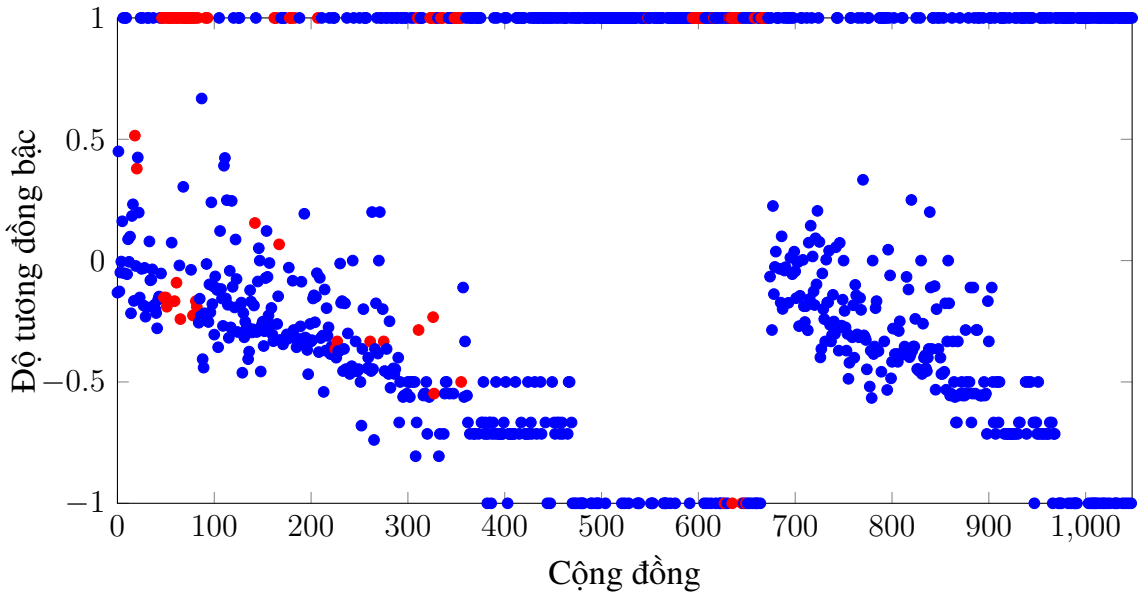
Phần này sẽ trình bày những thực nghiệm và đánh giá tính hiệu quả trong việc xác định các cộng đồng là botnet hay hợp pháp của ba độ đo cấu trúc được sử dụng trong mô đun 3.4. Các thực nghiệm sẽ sử dụng ngưỡng tần suất $\Theta_{fre} = 4$ dựa trên cơ sở của nghiên cứu đã công bố trước đó [13] vì bộ dữ liệu trong khóa luận này là mở rộng của bộ dữ liệu sử dụng trong nghiên cứu. Thêm vào đó, bảng 4.7 cũng cho thấy rằng giá trị $\Theta_{fre} = 4$ là phù hợp nhất trong việc xây dựng đồ thị giao tiếp cấp luồng mạng và phát hiện các cộng đồng botnet trên đồ thị.

Hình 4.1, 4.2, 4.3 thể hiện phân phối giá trị của ba độ đo: bậc nội bộ trung bình, độ tương đồng bậc, hệ số biến thiên đa dạng đích của các cộng đồng trên đồ thị liên hệ chung cấp luồng mạng. Các điểm màu xanh là các cộng đồng tạo thành từ các lưu lượng hợp pháp. Các điểm màu đỏ là các cộng đồng có tồn tại ít nhất một đỉnh trong đó là cụm luồng mạng của một máy chủ nhiễm botnet, nhưng vẫn có khả năng không phải là cộng đồng botnet.

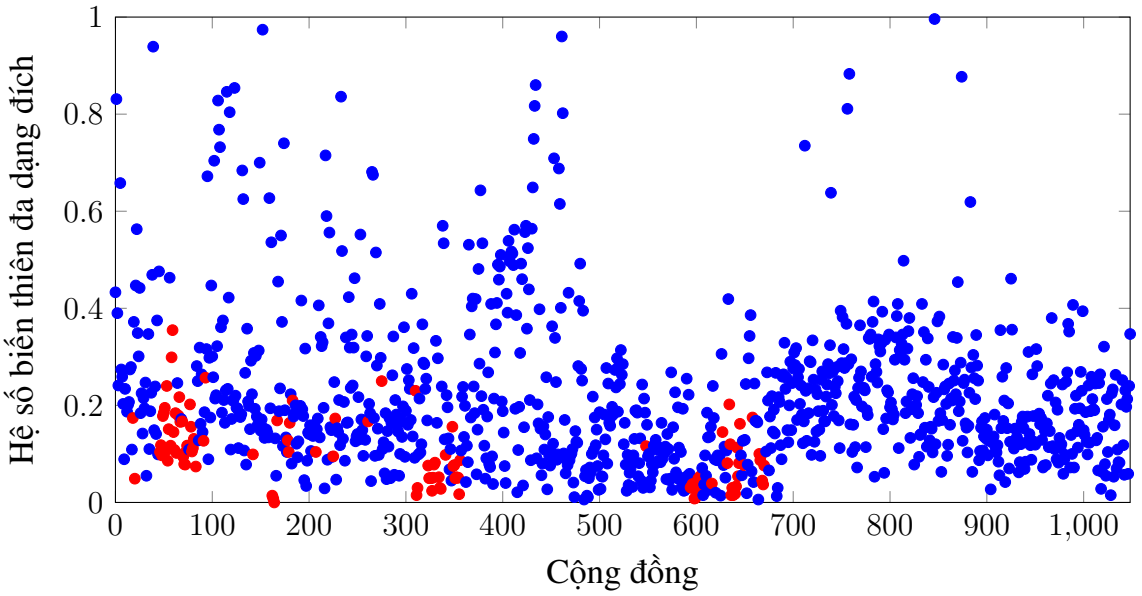
Hình 4.1: Biểu đồ phân phối giá trị bậc nội bộ trung bình của các cộng đồng



Hình 4.2: Biểu đồ phân phối giá trị độ tương đồng bậc của các cộng đồng



Hình 4.3: Biểu đồ phân phối giá trị hệ số biến thiên đa dạng đích của các cộng đồng



Đánh giá độ đo bậc nội bộ trung bình của cộng đồng

Hình 4.1 cho thấy phần lớn các cộng đồng nghi là của botnet đều có giá trị bậc nội bộ trung bình đạt tuyệt đối, tức là giữa hai đỉnh bất kỳ luôn tồn tại cạnh nối. Một số ít khác có giá trị của độ đo này không cao, có thể lý giải là do các cộng đồng này không phải của các cụm luồng mạng đặc trưng của botnet (các luồng truyền thông tin, duy trì cấu trúc mạng) hoặc không hình thành hoàn toàn từ lưu lượng botnet.

Tuy nhiên, vẫn có một lượng lớn các cộng đồng hợp pháp cũng đạt giá trị tuyệt đối với độ đo này. Trên thực tế, để tạo thành một cộng đồng, giữa các đỉnh phải có sự liên kết dày đặc nhất là đối với những cộng đồng có số lượng đỉnh tương đối ít. Bộ dữ liệu dùng để thực nghiệm không có nhiều máy chủ nếu xét riêng với mỗi loại ứng dụng P2P, vì vậy độ đo bậc nội bộ trung bình sẽ không quá hiệu quả trong việc phân tách các cộng đồng độc hại và hợp pháp.

Xét về tổng thể, giá trị độ đo này đối với các cộng đồng botnet lớn hơn hoặc bằng giá trị của nó trên các cộng đồng hợp pháp. Do đó, độ đo vẫn sẽ hiệu quả trong nhiều trường hợp nếu sử dụng chung với những giải thuật hay độ đo khác.

Đánh giá độ đo độ tương đồng bậc

Như thể hiện trong hình 4.2, các cộng đồng botnet chủ yếu có giá trị độ đo độ tương đồng về bậc lớn hơn hoặc bằng các cộng đồng khác. Tương tự như độ đo bậc nội bộ trung bình, khi các cộng đồng bao gồm lượng nhỏ thành viên, nếu giữa hai đỉnh bất kỳ đều có cạnh nối, đồng thời không có thêm cạnh nối ra một cụm luồng mạng nào khác

(do với mỗi ứng dụng P2P tồn tại lượng nhỏ các peer nên khả năng xảy ra càng thấp), khi đó giá trị độ đo này sẽ rất cao. Vì vậy, độ đo này không hiệu quả trong bộ dữ liệu thực nghiệm của khóa luận nhưng vẫn có thể sử dụng cùng với những phép đo khác để tăng độ chính xác của việc phát hiện cộng đồng botnet.

Đánh giá độ đo hệ số biến thiên đa dạng đích của các cộng đồng

Hình 4.3 thể hiện phân phối giá trị hệ số biến thiên đa dạng đích của các cộng đồng. Giá trị này đối với các cộng đồng botnet phân phối chủ yếu trong khoảng [0, 0.2], tương đối thấp so với các cộng đồng khác. Đối với công thức hệ số biến thiên trong quần thể được sử dụng để tính toán cho độ đo, giá trị càng thấp càng thể hiện sự tương đồng cao. Các cộng đồng hợp pháp phân phối giá trị trong một khoảng rất rộng do các mạng ngang hàng này thường không có quy chuẩn về số lượng máy chủ liên hệ hay bảng thông tin hàng xóm giống như mạng botnet. Mặc dù vẫn có một lượng nhỏ có giá trị độ đo tương đương với botnet, xét về tổng thể, độ đo này có thể phân tách giữa cộng đồng hợp pháp và botnet với tỉ lệ dương tính giả thấp và nếu được kết hợp với một số giải thuật khác sẽ có thể tạo ra hệ thống có độ chính xác cao.

4.3.2.2 Đánh giá giải thuật phát hiện botnet của hệ thống đề xuất

Phần này sẽ đánh giá tính hiệu quả trong việc xác định botnet của hệ thống đề xuất qua việc kết hợp các độ đo.

Để thuận tiện cho việc đánh giá, khóa luận sử dụng ba phép đo phổ biến sau đây:

- *Độ chính xác*:

$$Precision = \frac{TP}{TP + FP} \tag{4.1}$$

  Trong đó, TP là số lượng các bot được phát hiện chính xác, FP là số lượng dương tính giả hay máy bị nhận nhầm là nhiễm botnet. Độ chính xác càng cao tức là số lượng máy bị xác định nhầm là botnet càng thấp.

- *Độ nhạy*:

$$Recall = \frac{TP}{TP + FN} \tag{4.2}$$

  Trong đó, TP là số lượng các bot được phát hiện chính xác, FN là số lượng máy bị nhiễm botnet nhưng không được phát hiện. Độ nhạy càng cao chứng tỏ số lượng bot bị bỏ xót càng ít.

Bảng 4.5: Kết quả thực nghiệm hệ thống với những bộ Θ_{cv}, Θ_{avgid}, Θ_{lad} khác nhau và $\Theta_{fre} = 4$, $\Theta_{avgmcr} = 0.1$, $\Theta_{avgddr} = 0.3$.

Θ_{cv}	Θ_{avgid}	Θ_{lad}	FP	Precision	Recall	F-score
0.06	0.6	0.0	0	100.00%	64.86%	78.69%
		0.8	0	100.00%	64.86%	78.69%
	1.0	0.0	0	100.00%	64.86%	78.69%
		0.8	0	100.00%	64.86%	78.69%
0.08	0.6	0.0	0	100.00%	100.00%	100.00%
		0.8	0	100.00%	100.00%	100.00%
	1.0	0.0	0	100.00%	100.00%	100.00%
		0.8	0	100.00%	100.00%	100.00%
0.10	0.6	0.0	3	92.50%	100.00%	96.10%
		0.8	3	92.50%	100.00%	96.10%
	1.0	0.0	3	92.50%	100.00%	96.10%
		0.8	3	92.50%	100.00%	96.10%
0.15	0.6	0.0	7	84.09%	100.00%	91.36%
		0.8	7	84.09%	100.00%	91.36%
	1.0	0.0	7	84.09%	100.00%	91.36%
		0.8	7	84.09%	100.00%	91.36%

- *Điểm F1 (điểm F)*:

$$F - score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{4.3}$$

Điểm F1 là giá trị trung bình điều hòa của độ chính xác và độ nhạy.

Bảng 4.5 là kết quả thực nghiệm của hệ thống đề xuất với các bộ giá trị ngưỡng Θ_{cv}, Θ_{avgid}, Θ_{lad} khác nhau và các giá trị ngưỡng còn lại cố định ở một mức phù hợp. Kết quả cho thấy rằng hệ thống hoạt động tốt với nhiều bộ thông số khác nhau và phụ thuộc chủ yếu vào ngưỡng của độ đo hệ số biến thiên đa dạng đích. Như đánh giá ở phần 4.3.2.1, hai độ đo bậc nội bộ trung bình và sự tương đồng về bậc tỏ ra không thực sự hiệu quả trong bộ dữ liệu thực nghiệm, tuy nhiên vẫn có thể sử dụng kết hợp để tăng độ chính xác.

Với giá trị ngưỡng $\Theta_{cv} <= 0.08$, hệ thống có độ chính xác 100% với không kết quả dương tính giả nào. Ngược lại, khi $\Theta_{cv} >= 0.08$, độ nhạy đạt giá trị tuyệt đối, độ chính xác vẫn tương đối cao khi Θ_{cv} trong khoảng $[0.08, 0, 15]$. Ngưỡng $\Theta_{cv} = 0.08$ là phù hợp nhất cho việc phân tách giữa các cộng đồng botnet và hợp pháp.

Khi $\Theta_{cv} = 0.06$, độ nhạy của hệ thống giảm do một số cộng đồng botnet không được phát hiện. Các cụm luồng mạng trong các cộng đồng này chủ yếu là của các bot Storm. Loại botnet này có bảng hàng xóm ngang hàng rất lớn, dẫn đến các máy trong mạng liên hệ với rất nhiều ngang hàng khác nhau. Trong mạng ngang hàng nói chung, hiện tượng mất gói tin rất dễ xảy ra đặc biệt là trong những mạng lớn do hiện tượng các máy thường xuyên được thêm vào hoặc loại bỏ khỏi mạng. Khi gói tin không có phản hồi, các bot trong mạng botnet sẽ loại bỏ địa chỉ IP đích khỏi bảng hàng xóm ngang hàng và thay thế bằng một ngang hàng khác. Với việc lưu nhiều địa chỉ của các ngang hàng khác, tỉ lệ các thay đổi rất dễ xảy ra, dẫn đến sự không đồng đều trong tổng số lượng IP mà máy đó đã liên hệ.

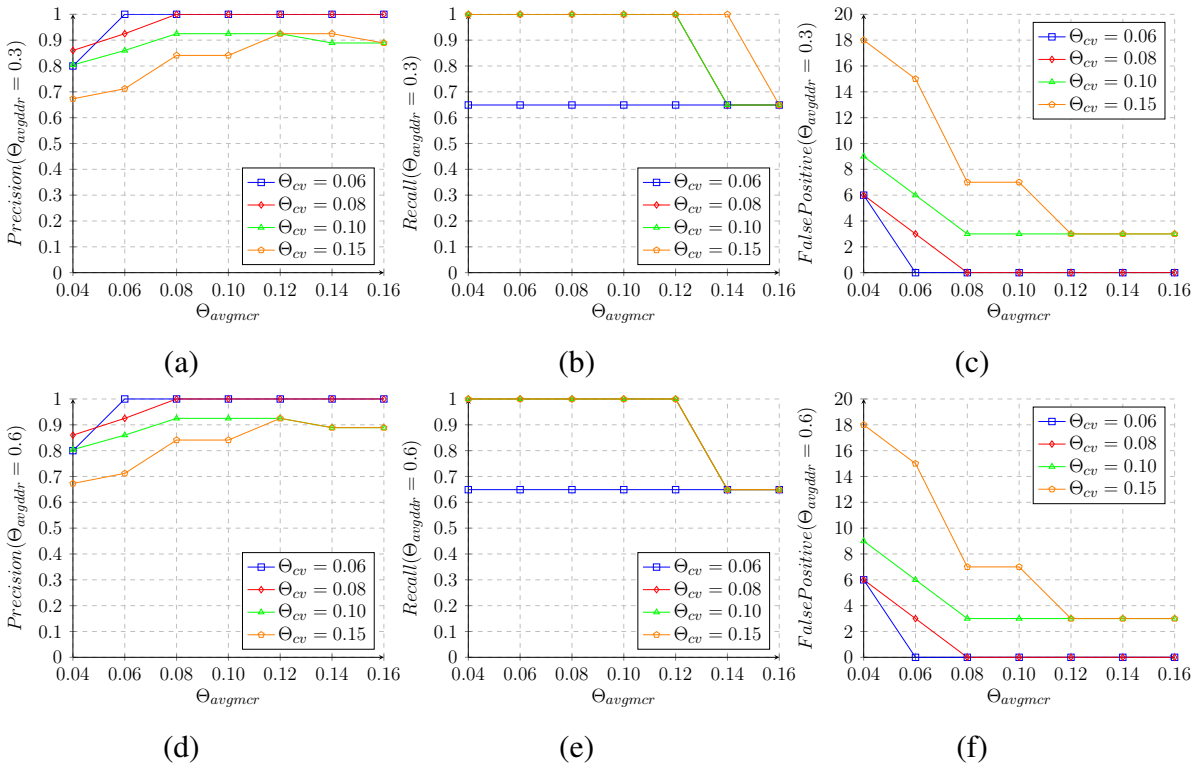
Khi $\Theta_{cv} = 0.1$, một số máy chỉ chứa lưu lượng hợp pháp bị phát hiện nhầm là botnet. Một số mạng ngang hàng bình thường cũng có hệ số tương quan đa dạng đích tương đối thấp, đặc biệt là các mạng có cấu trúc xây dựng dựa trên nền tảng liên lạc của một số máy nhất định. Trong trường hợp này, ba máy tính chạy ứng dụng tải tệp ngang hàng Deluge cùng tải một số tệp giống hệt nhau từ cùng một nguồn, dẫn đến giá trị độ đo trên thấp hơn so với một số cộng đồng hợp pháp khác và bị phát hiện nhầm là botnet với ngưỡng $\Theta_{cv} = 0.1$.

Khóa luận cũng tiến hành thực nghiệm với các thông số khác nhau của Θ_{cv}, Θ_{avgddr}, Θ_{avgmcr}, các thông số còn lại được đặt ở một ngưỡng thích hợp, để đánh giá thêm về tính hiệu quả khi ứng dụng độ đo hệ số biến thiên số lượng địa chỉ đích kết hợp với các giải thuật khác. Các biểu đồ trên hình 4.4 cho thấy ngưỡng $\Theta_{cv} = 0.08$ hoạt động tốt khi kết hợp với nhiều thông số ngưỡng tỉ lệ đa dạng đích trung bình và tỉ lệ liên hệ chung trung bình. Hệ thống đạt tỉ lệ chính xác và độ nhạy tuyệt đối hay không có kết quả dương tính giả nào khi $\Theta_{avgddr} \in [0.3, 0.6]$, $\Theta_{avgmcr} \in [0.08, 0.12]$ và $\Theta_{cv} = 0.08$.

4.3.3 Đánh giá tổng thể hệ thống đề xuất

4.3.3.1 Đánh giá tính hiệu quả

Với bộ dữ liệu đầu vào được sử dụng trong khóa luận, hệ thống đề xuất cho kết quả thực nghiệm rất tốt. Hệ thống đã phát hiện đầy đủ 37 máy chủ bị nhiễm botnet mà không có kết quả dương tính giả nào khi cài đặt cấu hình thông số tương ứng là $\Theta_f = 10$, $\Theta_{dd} = 30$, $\Theta_{fre} = 4$, $\Theta_{avgddr} \in [0.3, 0.6]$, $\Theta_{avgmcr} \in [0.08, 0.12]$, $\Theta_{cv} = 0.08$, $\Theta_{avgid} \in [0.6, 1]$, $\Theta_{lad} \in [0, 0.8]$.



Hình 4.4: Độ chính các, độ nhạy và số dương tính giả với những bộ giá trị khác nhau của Θ_{cv}, Θ_{avgddr}, Θ_{avgmcr}

4.3.3.2 Đánh giá khả năng áp dụng thực tế

Hệ thống được xây dựng để có thể áp dụng tại các thiết bị giám sát mạng, các cổng chuyển mạng switch, route hay tường lửa trong hệ thống mạng của tổ chức. Để đánh giá tính khả thi, tôi đã tính toán thời gian chạy trên bộ dữ liệu thực nghiệm gồm 10000 máy chủ (tương ứng với một tổ chức có 10000 máy tính). Kết quả cho thấy hệ thống chạy mất khoảng 8 phút với cấu hình máy tính trong bảng 4.1.

Hệ thống có thể hoạt động tốt với dữ liệu đầu vào lớn do khả năng tính toán song song của một số thành phần, sử dụng khung phần mềm MapReduce, đặc biệt là thành phần xác định lưu lượng ngang hàng. Bảng 4.4 cho thấy rằng số lượng máy chủ còn lại sau bước phát hiện lưu lượng ngang hàng là tương đối nhỏ. Ví dụ với thông số $\Theta_{dd} = 30$ được sử dụng trong các thực nghiệm của hệ thống, chỉ có 158/10000 máy chủ được giữ lại cho thành phần tiếp theo. Có thể nói Θ_{dd} là một tham số quan trọng ảnh hưởng đến khả năng ứng dụng thực tế của hệ thống. Khi Θ_{dd} tăng, thời gian xử lý sẽ giảm do số lượng máy chủ còn lại sau quá trình phát hiện lưu lượng P2P cũng giảm theo. Do đó, với việc điều chỉnh Θ_{dd} hợp lý, hệ thống đề xuất hoàn toàn có thể mở rộng để xử lý dữ liệu mạng trong thế giới thực.

Bảng 4.6: Kết quả thực nghiệm với những thông số tốt nhất của hệ thống Enhanced PeerHunter

Θ_{avgddr}	Θ_{avgmcr}	FP	Precision	Recall	F-score
0.2	0.05	28	56.92%	100.00%	72.55%
	0.15	13	74.00%	100.00%	85.06%
	0.25	13	74.00%	100.00%	85.06%
	0.35	13	74.00%	100.00%	85.06%
	0.45	12	74.47%	94.59%	83.33%
0.4	0.05	28	56.92%	100.00%	72.55%
	0.15	13	74.00%	100.00%	85.06%
	0.25	13	74.00%	100.00%	85.06%
	0.35	13	74.00%	100.00%	85.06%
	0.45	12	74.47%	100.00%	85.37%
0.6	0.05	17	68.52%	100.00%	81.32%
	0.15	13	74.00%	100.00%	85.06%
	0.25	13	74.00%	100.00%	85.06%
	0.35	13	74.00%	100.00%	85.06%
	0.45	12	74.47%	100.00%	85.37%

Ngoài ra, nhờ việc xây dựng đồ thị với các đỉnh là cụm luồng mạng đặc trưng bởi kích thước gói tin, hệ thống có thể hoạt động tốt khi lưu lượng mạng trên một máy chủ bao gồm cả lưu lượng ngang hàng hợp pháp và độc hại. Điều này là rất phù hợp vì trên thực tế, một máy tính thường sử dụng đan xen các loại ứng dụng khác nhau và lưu lượng của P2P botnet sẽ bị trộn lẫn trong đó.

4.3.4 So sánh với các nghiên cứu khác

4.3.4.1 So sánh với hệ thống Enhanced PeerHunter

Hệ thống đề xuất trong khóa luận sử dụng nhiều tính chất và đặc điểm hành vi của botnet đã được chỉ ra trong Enhanced PeerHunter [9], vì vậy, có nhiều điểm chung giữa hai hệ thống. Thứ nhất, cả hai hệ thống đều tận dụng các biểu đồ liên hệ lẫn nhau ở cấp độ luồng mạng. Thứ hai, thành phần phát hiện luồng mạng P2P đều dựa trên tính đa dạng đích đến. Thứ ba, các đặc điểm hành vi cộng đồng được sử dụng để xác định các nhóm mạng botnet.

Điểm khác biệt giữa hai hệ thống là một số đặc điểm khác của botnet đã được áp dụng để giảm tỉ lệ phát hiện nhầm trên những tập dữ liệu thách thức cao. Thứ nhất, trước

khi thực hiện phát hiện luồng mạng P2P, hệ thống đề xuất có thành phần tổng hợp các luồng mạng theo byte trên mỗi gói. Thứ hai, giải thuật thực hiện tính toán trọng số liên hệ lẫn nhau giữa các gói luồng mạng bằng cách sử dụng một ý tưởng mới gọi là tần suất của luồng mạng. Hai cải tiến này đã được áp dụng trong công trình trước đó của tôi [13]. Trong khóa luận này, ba độ đo cấu trúc đồ thị được thêm vào trong quá trình lọc các cộng đồng trên đồ thị liên hệ chung và đã hoạt động rất tốt.

Khóa luận đã tiến hành thực nghiệm hệ thống của Enhanced PeerHunter trên tập dữ liệu của khóa luận với những bộ thông số tốt nhất để tiến hành so sánh. Bảng 4.6 cho thấy Enhanced PeerHunter có độ nhạy rất cao, phát hiện được hầu hết các loại botnet trong nhiều cài đặt cấu hình khác nhau. Tuy nhiên, độ chính xác của Enhanced PeerHunter lại tương đối thấp do việc tồn tại những mạng ngang hàng hợp pháp cũng có sự liên kết cao như botnet, dẫn đến bị phát hiện nhầm là botnet.

4.3.4.2 So sánh với nghiên cứu đã công bố

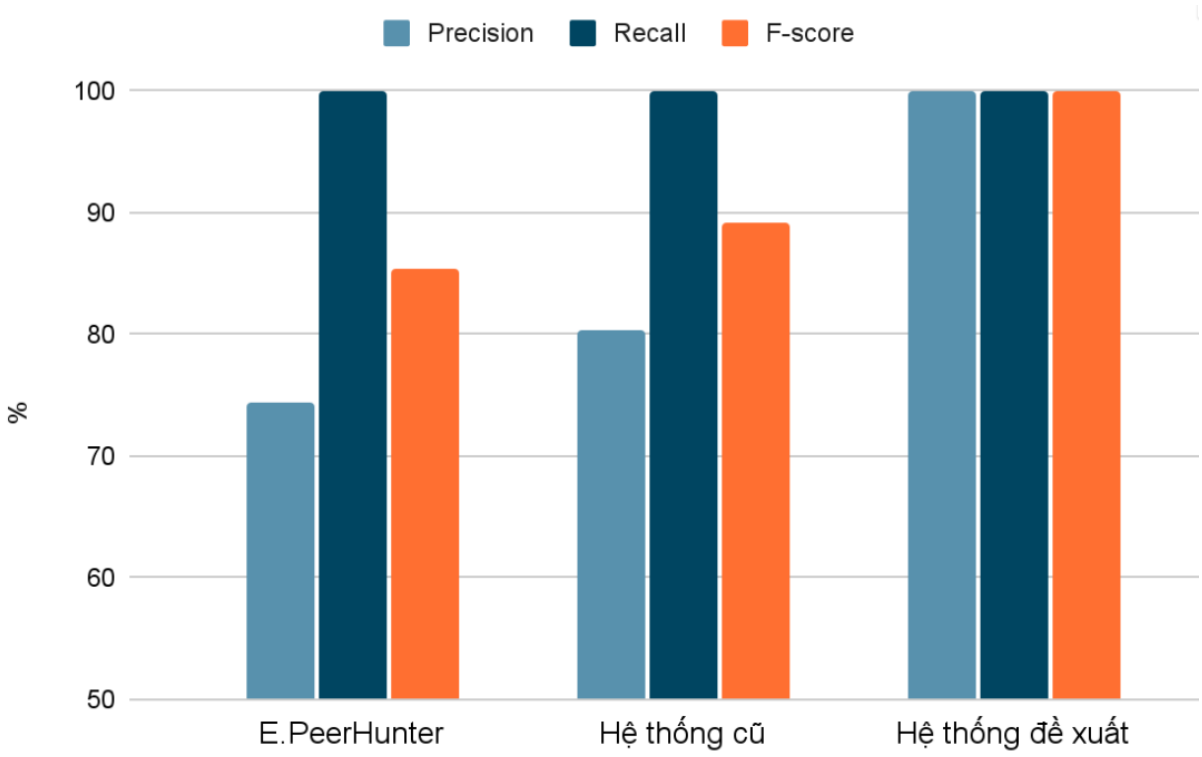
Hệ thống được đề xuất trong khóa luận là một phiên bản cải tiến của công trình nghiên cứu đã được tôi và các cộng sự công bố [13]. Những đề xuất trong khóa luận chủ yếu hướng tới việc tăng độ chính xác cho thành phần xác định cộng đồng nào là của botnet trên đồ thị liên hệ chung. Ba độ đo mới đã được xây dựng dựa trên các độ đo cấu trúc đồ thị bao gồm độ đo bậc nội bộ trung bình, độ đo tính tương đồng về bậc và độ đo hệ số biến thiên về tính đa dạng đích. Thành phần phát hiện các cụm liên kết cao sau quá trình phát hiện cộng đồng botnet cũng được loại bỏ do độ đo bậc nội bộ trung bình đã bao quát, thể hiện chính xác hơn cho giải thuật này. Cùng với đó, một bộ dữ liệu mới được xây dựng dựa trên việc thu thập các lưu lượng thực tế để tăng tính thực tế và khó khăn cho việc phát hiện botnet.

Khóa luận đã tiến hành thực nghiệm bộ dữ liệu mới trên hệ thống đã công bố trước đó với một loạt các thông số khác nhau. Kết quả được thể hiện trong bảng 4.7. Trong các kết quả có độ nhạy cao nhất, tức là không có máy bị nhiễm botnet nào không được phát hiện, độ chính xác trong các trường hợp đó không đạt tuyệt đối do xuất hiện những dương tính giả. Các máy chủ hợp pháp bị nhận nhầm là nhiễm botnet này chủ yếu là các máy chủ chạy Deluge. Các máy này được thiết lập để thu thập với tính thách thức vô cùng cao do những cách triển khai khác nhau trong từng ngày thu thập.

Hình 4.5 thể hiện tương quan về độ chính xác, độ nhạy và điểm F1 của 3 hệ thống được so sánh khi thực nghiệm trên bộ dữ liệu đề xuất của khóa luận. Hệ thống đề xuất hoạt động tốt nhất với độ chính xác và độ nhạy đều đạt 100%.

Bảng 4.7: Kết quả thực nghiệm trên hệ thống cũ với các giá trị khác nhau của Θ_{fre}, Θ_{avgmcr} và Θ_{avgddr}.

Θ_{fre}	Θ_{avgmcr}	Θ_{avgddr}	FP	Precision	Recall	F-score
1	0.05	0.3	105	26.05%	100%	41.34%
		0.6	74	33.34%	100%	50%
		0.9	73	33.64%	100%	50.34%
	0.1	0.3	63	37%	100%	54.02%
		0.6	63	37%	100%	54.02%
		0.9	61	37.76%	100%	54.81%
	0.15	0.3	58	38.95%	100%	56.06%
		0.6	58	38.95%	100%	56.06%
		0.9	41	47.43%	100%	64.35%
2	0.05	0.3	70	34.58%	100%	51.38%
		0.6	68	35.24%	100%	52.11%
		0.9	64	36.63%	100%	53.62%
	0.1	0.3	26	58.73%	100%	74%
		0.6	26	58.73%	100%	74%
		0.9	23	59.65%	91.89%	72.34%
	0.15	0.3	13	74%	100%	85.06%
		0.6	13	74%	100%	85.06%
		0.9	3	87.50%	56.76%	68.85%
4	0.05	0.3	31	54.41%	100%	70.48%
		0.6	31	54.41%	100%	70.48%
		0.9	27	43.75%	56.75%	49.41%
	0.1	0.3	9	80.43%	100%	89.15%
		0.6	9	80.43%	100%	89.15%
		0.9	3	86.36%	51.35%	64.41%
	0.15	0.3	9	80.43%	100%	89.16%
		0.6	9	78.05%	86.49%	82.05%
		0.9	0	100%	43.24%	60.38%
6	0.05	0.3	20	64.91%	100%	78.72%
		0.6	20	64.91%	100%	78.72%
		0.9	14	60%	56.75%	58.34%
	0.1	0.3	9	78.05%	86.49%	82.05%
		0.6	9	78.05%	86.49%	82.05%
		0.9	3	84.21%	43.24%	57.14%
	0.15	0.3	9	67.85%	51.35%	58.46%
		0.6	9	67.85%	51.35%	58.46%
		0.9	0	100%	43.24%	60.38%



Hình 4.5: So sánh kết quả thực nghiệm của các hệ thống trên tập dữ liệu đề xuất.

Kết luận

Khóa luận đã đề xuất một phương pháp mới hiệu quả hơn trong việc phát hiện các mạng botnet ngang hàng dựa trên việc phân tích các hành vi của botnet qua các đặc điểm cấu trúc đồ thị kết hợp với các đặc tính đặc trưng khác. Đầu tiên, đặc tính đa dạng kích thước gói tin và đa dạng địa chỉ đích được sử dụng để phát hiện các lưu lượng ngang hàng. Sau đó, đồ thị liên hệ chung cấp luồng mạng được xây dựng thông qua tần suất liên hệ của các luồng mạng. Cuối cùng, các độ đo mới được đề xuất trong khóa luận dựa trên đặc điểm cấu trúc đồ thị bao gồm bậc nội bộ trung bình của cộng đồng, độ tương đồng về bậc, hệ số biến thiên số lượng địa chỉ IP đích, kết hợp với đặc tính liên hệ chung và đa dạng đích, được sử dụng để phân tích hành vi của mạng botnet và phân tách chúng với các cộng đồng mạng hợp pháp.

Giải pháp đề xuất đã được thực nghiệm trên một tập dữ liệu có tính thực tế và thách thức cao, trong đó kết hợp các lưu lượng hợp pháp và botnet phổ biến với lưu lượng ứng dụng ngang hàng được thu thập mới. Kết quả chỉ ra rằng phương pháp đề xuất đạt hiệu suất rất cao với độ chính xác và độ nhạy đều đạt 100% trên bộ dữ liệu thực nghiệm. Kết quả thực nghiệm cũng cho thấy độ đo hệ số biến thiên về số lượng địa chỉ đích hoạt động rất tốt trong việc phân tách lưu lượng hợp pháp và độc hại. Hai độ đo còn lại gồm bậc nội bộ trung bình và độ tương đồng về bậc, tuy không hiệu quả khi sử dụng độc lập nhưng vẫn có thể áp dụng chung với các giải thuật khác để tăng độ chính xác.

Khóa luận cũng đưa ra những thực nghiệm trên các hệ thống liên quan và cho thấy rằng hệ thống đề xuất hoạt động hiệu quả hơn khi không có trường hợp nào bị nhận nhầm là nhiễm botnet. Kết quả này là tiền đề để có thể ứng dụng hệ thống trong các mạng thực tế. Các phân tích cũng chỉ ra giải pháp đề xuất có khả năng ứng dụng thực tế rất cao do có thời gian chạy hợp lý và có thể phân tách nhiều loại lưu lượng khác nhau hoạt động trên cùng một máy chủ.

Định hướng mở rộng nghiên cứu trong tương lai

Trong phạm vi của một khóa luận tốt nghiệp, do những hạn chế về thời gian, kiến

thức cũng như hạn chế về mặt dữ liệu, môi trường thực nghiệm, mô hình này có thể chưa được đánh giá chính xác và vẫn cần phải được cải thiện nhiều để có thể hoàn thiện hơn nữa. Thêm vào đó, các mạng botnet mới ngày càng phát triển và có khả năng thích nghi dần và có các biện pháp hạn chế sự hiệu quả của các phương pháp phát hiện nếu không có sự thay đổi thường xuyên. Vì vậy, khóa luận đề xuất một số hướng nghiên cứu trong tương lai như sau:

- Phát hiện các thuộc tính botnet P2P ẩn chưa được biết đến cũng như phân tích các mạng botnet mới, áp dụng để cải tiến hệ thống trong tương lai thích nghi với nhiều loại botnet.
- Áp dụng các phương pháp học máy, học sâu để tự động chọn những cấu hình tham số tốt nhất cho hệ thống, đồng thời kết hợp với nhiều kỹ thuật khác để tăng tính hiệu quả cho mô hình.
- Thu thập và cập nhật bộ dữ liệu đặc biệt là lưu lượng của các loại botnet mới và nguy hiểm để tập dữ liệu mang tính thực tế, thách thức hơn, giúp cho việc đánh giá các hệ thống được chính xác.

Danh mục bài báo đã xuất bản

[Pub 1] **Quang Huy Nguyen**, Trung Kien Dang, Dai Tho Nguyen. "A More Efficient System for Peer-to-Peer Botnet Detection." In *The 8th International Conference on Intelligent Information Technology ICIIT 2023*, 2023, (In Press).

Tài liệu tham khảo

[1] Statista, "Number of internet and social media users worldwide as of january 2023." [Online]. Available: https://www.statista.com/statistics/617136/digital-populatio-worldwide/

[2] S. S. Silva, R. M. Silva, R. C. Pinto, and R. M. Salles, "Botnets: A survey," *Computer Networks*, vol. 57, no. 2, pp. 378–403, 2013, botnet Activity: Analysis, Detection and Shutdown. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128612003568

[3] P. Wang, B. Aslam, and C. C. Zou, *Peer-to-Peer Botnets*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 335–350. [Online]. Available: https://doi.org/10.1007/978-3-642-04117-4_18

[4] Cisco and IronPort, "2008 internet malware trends: Storm and the future of social engineering," 2008. [Online]. Available: https://www.pmi.it/app/uploads/2018/02/000346-K7fvYO.pdf

[5] L. Whitney, "With legal nod, microsoft ambushes waledac botnet," 2010. [Online]. Available: https://www.cnet.com/news/privacy/with-legal-nod-microsoft-ambushes-waledac-botnet/

[6] K. Singh, S. C. Guntuku, A. Thakur, and C. Hota, "Big data analytics framework for peer-to-peer botnet detection using random forests," *Information Sciences*, vol. 278, pp. 488–497, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025514003570

[7] M. Alauthman, N. Aslam, M. Al-kasassbeh, S. Khan, A. Al-Qerem, and K.-K. Raymond Choo, "An efficient reinforcement learning-based botnet detection approach," *Journal of Network and Computer Applications*, vol. 150, p. 102479, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S108480451930339X

[8] M. Elhalabi, S. Manickam, L. Bani Melhim, M. Anbar, and H. Alhalabi, "A review of peer-to-peer botnet detection techniques," *Journal of Computer Science*, vol. 10, p. 169, 01 2013.

[9] D. Zhuang and J. M. Chang, "Enhanced peerhunter: Detecting peer-to-peer botnets through network-flow level community behavior analysis," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 6, pp. 1485–1500, 2019.

[10] C.-Y. Wang, C.-L. Ou, Y.-E. Zhang, F.-M. Cho, P.-H. Chen, J.-B. Chang, and C.-K. Shieh, "Botcluster: A session-based p2p botnet clustering system on netflow," *Computer Networks*, vol. 145, 09 2018.

[11] H. Hang, X. Wei, M. Faloutsos, and T. Eliassi-Rad, "Entelecheia: Detecting p2p botnets in their waiting stage," in *2013 IFIP Networking Conference*, 2013, pp. 1–9.

[12] H. P. Joshi and R. Dutta, "Identifying p2p communities in network traffic using measures of community connections : Ieee cns 20 poster," in *2020 IEEE Conference on Communications and Network Security (CNS)*, 2020, pp. 1–2.

[13] N. Quang Huy, D. Trung Kien, and N. Dai Tho, "A more efficient system for peer-to-peer botnet detection," in *2023 8th International Conference on Intelligent Information Technology*, ser. ICIIT 2023.

[14] M. A. Al-Shareeda, S. Manickam, M. A. Saare, S. Karuppayah, and M. A. Alazzawi, "Detection mechanisms for peer-to-peer botnets: A comparative study," in *2022 8th International Conference on Contemporary Information Technology and Mathematics (ICCITM)*, 2022, pp. 267–272.

[15] D. Stutzbach and R. Rejaie, "Understanding churn in peer-to-peer networks," 10 2006, pp. 189–202.

[16] H.-S. Wu, N.-F. Huang, and G.-H. Lin, "Identifying the use of data/voice/video-based p2p traffic by dns-query behavior," 06 2009, pp. 1–5.

[17] B. Rahbarinia, R. Perdisci, A. Lanzi, and K. Li, "Peerrush: Mining for unwanted p2p traffic," *Journal of Information Security and Applications*, vol. 19, no. 3, pp. 194–208, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214212614000143

[18] B. Coskun, S. Dietrich, and N. Memon, “Friends of an enemy: Identifying local members of peer-to-peer botnets using mutual contacts,” 12 2010, pp. 131–140.

[19] S. Fortunato and D. Hric, “Community detection in networks: A user guide,” *Physics Reports*, vol. 659, pp. 1–44, 2016, community detection in networks: A user guide. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0370157316302964

[20] S. Karuppayah, *Advanced Monitoring in P2P Botnets: A Dual Perspective*, 01 2018.

[21] Wikipedia, “Conductance (graph).” [Online]. Available: https://en.wikipedia.org/wiki/Conductance_(graph)

[22] M. E. J. Newman, “Mixing patterns in networks,” *Physical Review E*, vol. 67, no. 2, feb 2003. [Online]. Available: https://doi.org/10.11032Fphysreve.67.026126

[23] C. Rossow and C. J. Dietrich, “Provex: Detecting botnets with encrypted command and control channels,” in *Detection of Intrusions and Malware, and Vulnerability Assessment*, K. Rieck, P. Stewin, and J.-P. Seifert, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 21–40.

[24] J. Dean and S. Ghemawat, “Mapreduce: Simplified data processing on large clusters,” *Commun. ACM*, vol. 51, no. 1, p. 107–113, jan 2008. [Online]. Available: https://doi.org/10.1145/1327452.1327492

[25] S. Nagaraja, P. Mittal, C.-Y. Hong, M. C. Caesar, and N. Borisov, “Botgrep: Finding p2p bots with structured graph analysis,” in *USENIX Security Symposium*, 2010.

[26] Đặng Trung Kiên, “Phát hiện các botnet ngang hàng thông qua phân tích các biểu hiện cộng đồng,” 2023.

[27] V. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics Theory and Experiment*, vol. 2008, 04 2008.

[28] M. Piraveenan, M. Prokopenko, and A. Zomaya, “Local assortativeness in scale-free networks,” *http://dx.doi.org/10.1209/0295-5075/89/49901*, vol. 84, 10 2008.

[29] Wikipedia, “Coefficient of variation.” [Online]. Available: https://en.wikipedia.org/wiki/Coefficient_of_variation

[30] “Free malware sample sources for researchers,” Mar 2021. [Online]. Available: https://zeltser.com/malware-sample-sources/

[31] “Mawi working group traffic archive,” 2008. [Online]. Available: http://mawi.wide.ad.jp/mawi/ditl/ditl201412/