ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

HÀ MẠNH KIÊN

KỸ THUẬT PHÂN LỚP DỮ LIỆU VÀ ỨNG DỤNG TRONG PHÁT HIỆN MÃ ĐỘC

***Chuyên ngành*: Khoa học máy tính**
Mã số:60.48.01.01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

***NGƯỜI HƯỚNG DẪN KHOA HỌC*:**
TS.Lương Thế Dũng

THÁI NGUYÊN - 2015

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình của riêng tôi.

Các số liệu, kết quả nêu trong luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Qua đây tôi xin chân thành cảm ơn toàn thể các thầy cô trong khoa đào tạo sau đại học Trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên, những người đã trực tiếp giảng dạy, truyền đạt cho tôi kiến thức chuyên môn và phương pháp làm việc khoa học.

Đặc biệt, tôi xin chân thành cảm ơn **TS. Lương Thế Dũng** ,đã tận tình hướng dẫn để tôi có thể hoàn thành luận văn này.

Tôi cũng xin gửi lời cảm ơn tới gia đình, bạn bè, đồng nghiệp đã giúp đỡ, động viên và tạo điều kiện cho tôi trong quá trình làm luận văn.

Tác giả luận văn

Hà Mạnh Kiên

MỤC LỤC

DANH MỤC BẢNG

DẠNH MỤC HÌNH

ĐẶT VẤN ĐỀ

Khi nhu cầu về việc sử dụng internet của con người ngày càng tăng thì cũng là lúc mối đe dọa xuất hiện ngày càng nhiều, nổi bật là đe dọa của mã độc hại. Mã độc là một loại phần mềm hệ thống do các tin tặc hay các kẻ nghịch ngợm tạo ra nhằm gây hại cho máy tính. Tùy theo cách thức mà tin tặc dung, sự nguy hại của các loại phần mềm khác nhau từ chỗ chỉ hiển thị các cửa sổ thông báo cho đến việc tấn công chiếm máy và lây lan sang máy khác như virut. Xuất hiện bất kỳ đâu trên môi trường của các thiết bị điện tử như các dĩa mềm, usb, đến môi trường Internet trong các webside, trong các tin nhắn, trong hòm thư điện tử của người dùng, trong các phần mềm tiện ích……Khi mã độc hại đã nhiễm vào một máy tính nào đó thì nó sẽ lây lan sang máy tính khác là khá nhanh và khó lường trước được.

Công nghệ thông tin liên tục phát triển và thay đổi, nhiều phần mềm mới ra đời mang đến cho con người nhiều tiện ích hơn. Do vậy để chống lại các loại mã độc hại người ta thường sử dụng các chương trình phát hiện và loại bỏ mã độc hại. Tuy nhiên việc phát hiện mã độc hại của các chương trình hiện nay thường dựa trên các thuật toán đối sánh mẫu và quan trọng là một cơ sở dữ liệu đầy đủ và cập nhật thường xuyên những mẫu mới. Để có một cơ sở dữ liệu như đã nêu cần một chương trình quản lý một cách hiệu quả và tốt rất nhiều công sức để tạo ra các mẫu mã độc hại. Một phương pháp mới hiện nay là dựa trên các mô hình toán học để phát hiện ra các mã độc hại mới mà không sử dụng các cơ sở dữ liệu mẫu, trong đó khai phá dữ liệu là một phương pháp quan trọng và đang được nhiều người quan tâm. Chính vì vậy luận văn này tiến hành nghiên cứu, tìm hiểu các kỹ thuật phân lớp dữ liệu và ứng dụng trong phát hiện mã độc. Nhằm xấy dựng ra các mô hình, thuật toán để phát hiện và đánh giá các mô hình đó.

CHƯƠNG 1
TỔNG QUAN VỀ MÃ ĐỘC HẠI

1.1. Các loại mã độc

1.1.1. Virus

Virus là một loại mã độc hại (Maliciuos code) có khả năng tự nhân bản và lây nhiễm chính nó vào các file, chương trình hoặc máy tính. Như vậy virus máy tính phải luôn luôn bám vào một vật chủ (đó là file dữ liệu hoặc file ứng dụng) để lây lan. Các chương trình diệt virus dựa vào đặc tính này để thực thi việc phòng chống và diệt virus, để quét các file trên thiết bị lưu, quét các file trước khi lưu xuống ổ cứng… Điều này cũng giải thích vì sao đôi khi các phần mềm diệt virus tại PC đưa ra thông báo "phát hiện ra virus nhưng không diệt được" khi thấy có dấu hiệu hoạt động của virus trên PC, bởi vì "vật mang virus" lại nằm ở máy khác nên không thể thực thi việc xoá đoạn mã độc hại đó.

Compiled Virus là virus mà mã thực thi của nó đã được dịch hoàn chỉnh bởi một trình biên dịch để nó có thể thực thi trực tiếp từ hệ điều hành. Các loại boot virus như (Michelangelo và Stoned), file virus (như Jerusalem) rất phổ biến trong những năm 80 là virus thuộc nhóm này, compiled virus cũng có thể là pha trộn bởi cả boot virus và file virus trong cùng một phiên bản.

Interpreted Virus là một tổ hợp của mã nguồn mã chỉ thực thi được dưới sự hỗ trợ của một ứng dụng cụ thể hoặc một dịch vụ cụ thể trong hệ thống. Một cách đơn giản, virus kiểu này chỉ là một tập lệnh, cho đến khi ứng dụng gọi thì nó mới được thực thi. Macro virus, scripting virus là các virus nằm trong dạng này. Macro virus rất phổ biến trong các ứng dụng Microsoft Office khi tận dụng khả năng kiểm soát việc tạo và mở file để thực thi và lây nhiễm. Sự khác nhau giữa macro virus và scripting virus là: Macro virus là

tập lệnh thực thi bởi một ứng dụng cụ thể, còn scripting virus là tập lệnh chạy bằng một service của hệ điều hành. Melisa là một ví dụ xuất sắc về Macro virus, Love Stages là ví dụ cho scripting virus.

1.1.2. Worm

Worm cũng là một chương trình có khả năng tự nhân bản và tự lây nhiễm trong hệ thống tuy nhiên nó có khả năng "tự đóng gói", điều đó có nghĩa là Worm không cần phải có "file chủ" để mang nó khi nhiễm vào hệ thống. Như vậy, có thể thấy rằng chỉ dùng các chương trình quét file sẽ không diệt được Worm trong hệ thống vì Worm không "bám" vào một file hoặc một vùng nào đó trên đĩa cứng. Mục tiêu của Worm bao gồm cả làm lãng phí nguồn lực băng thông của mạng và phá hoại hệ thống như xoá file, tạo backdoor, thả keylogger,… Tấn công của Worm có đặc trưng là lan rộng cực kỳ nhanh chóng do không cần tác động của con người (như khởi động máy, copy file hay đóng/mở file). Worm có thể chia làm 2 loại:

Network Service Worm lan truyền bằng cách lợi dụng các lỗ hổng bảo mật của mạng, của hệ điều hành hoặc của ứng dụng. Sasser là ví dụ cho loại sâu này.

Mass Mailing Worm là một dạng tấn công qua dịch vụ mail, tuy nhiên nó tự đóng gói để tấn công và lây nhiễm chứ không bám vào vật chủ là email. Khi sâu này lây nhiễm vào hệ thống, nó thường cố gắng tìm kiếm sổ địa chỉ và tự gửi bản thân nó đến các địa chỉ thu nhặt được. Việc gửi đồng thời cho toàn bộ các địa chỉ thường gây quá tải cho mạng hoặc cho máy chủ mail. Netsky, Mydoom là ví dụ cho thể loại này.

1.1.3. Trojan Horse

Trojan Horse là loại mã độc hại được đặt theo sự tích "Ngựa thành Troy". Trojan horse không có khả năng tự nhân bản tuy nhiên nó lây vào hệ thống với biểu hiện rất bình thường nhưng thực chất bên trong có ẩn chứa các đoạn mã với mục đích gây hại. Trojan có thể gây hại theo ba cách sau:

Tiếp tục thực thi các chức năng của chương trình mà nó bám vào, bên cạnh đó thực thi các hoạt động gây hại một cách riêng biệt (ví dụ như gửi một trò chơi dụ cho người dùng sử dụng, bên cạnh đó là một chương trình đánh cắp password).

Tiếp tục thực thi các chức năng của chương trình mà nó bám vào, nhưng sửa đổi một số chức năng để gây tổn hại (ví dụ như một trojan giả lập một cửa sổ login để lấy password) hoặc che dấu các hành động phá hoại khác (ví dụ như trojan che dấu cho các tiến trình độc hại khác bằng cách tắt các hiển thị của hệ thống).

Thực thi luôn một chương trình gây hại bằng cách núp dưới danh một chương trình không có hại (ví dụ như một trojan được giới thiệu như là một chò chơi hoặc một tool trên mạng, người dùng chỉ cần kích hoạt file này là lập tức dữ liệu trên PC sẽ bị xoá hết).

Có 7 loại trojan chính:

Trojan truy cập từ xa: Được thiết kế để cho kẻ tấn công có khả năng từ xa chiếm quyền điều khiển của máy bị hại. Các trojan này thường dấu vào các trò chơi và các chương trình nhỏ làm cho người dùng mất cảnh giác.

Trojan gửi dữ liệu: Nó thực hiện việc lấy và gửi dữ liệu nhạy cảm như mật khẩu, thông tin thẻ tín dụng, các tệp nhật ký, địa chỉ email… cho kẻ tấn công. Trojan này có thể tìm kiếm cụ thể thông tin hoặc cài phần mềm đọc trộm bàn phím và gửi toàn bộ các phím bấm về cho kẻ tấn công.

Trojan hủy hoại: Thực hiện việc xóa các tệp tin. Loại trojan này giống với virus và thường có thể bị phát hiện bởi các chương trình diệt virus.

Trojan kiểu proxy: Sử dụng máy tính bị hại làm proxy, qua đó có thể sử dụng máy bị hại để thực hiện các hành vi lừa gạt hay đánh phá các máy tính khác.

Trojan FTP: Được thiết kế để mở cổng 21 và cho phép tin tặc kết nối vào máy bị hại sử dụng FTP.