



**PTIT- PTITMALWARE**

**Giảng viên:** Đinh Trường Duy

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA AN TOÀN THÔNG TIN**

***Hà Nội - 2024***

# **1. Nội dung và hướng dẫn thực hiện bài thực hành**

## **1.1. Mục đích**

Bài thực hành này nhằm hướng dẫn sinh viên biết và hiểu về các bước phân tích tĩnh một tập tin PDF nghi ngờ có mã độc. Sinh viên sẽ sử dụng bộ công cụ Pdfid, Peepdf và Pdf-parser để phân tích một tập tin PDF nhằm xác định một số thuộc tính cơ bản và đoạn mã độc ẩn giấu bên trong.

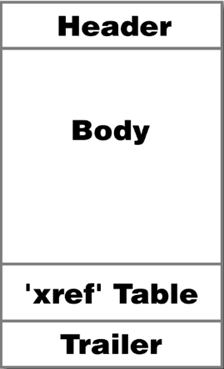
## **1.2. Yêu cầu đối với sinh viên**

Có kiến thức cơ bản về hệ điều hành Linux, mã độc, cấu trúc của 1 tập tin PDF.

Cấu trúc định dạng của 1 tập tin PDF:

PDF là định dạng tài liệu di động có thể được sử dụng để trình bày các tài liệu bao gồm văn bản, hình ảnh, thành phần đa phương tiện, liên kết trang web và hơn thế nữa. Định dạng này được thiết kế để có thể hoạt động trên tất cả các công cụ PDF trên tất cả các Hệ điều hành.

PDF có nhiều chức năng hơn là chỉ văn bản: nó có thể bao gồm hình ảnh và các thành phần đa phương tiện khác, được bảo vệ bằng mật khẩu, thực thi JavaScript, v.v. Cấu trúc cơ bản của một tệp PDF được trình bày trong hình dưới đây:



Hình 1. Cấu trúc tập tin PDF

**PDF File Header**

Tệp PDF đều bắt đầu bằng tiêu đề chứa mã định danh duy nhất cho PDF và phiên bản có định dạng, chẳng hạn như %PDF-1.x trong đó x nằm trong khoảng từ 1-7.

**File Body**

Trong phần nội dung của tài liệu PDF, có các đối tượng thường bao gồm luồng văn bản, hình ảnh, các thành phần đa phương tiện khác, v.v. Phần Nội dung được sử dụng để chứa tất cả dữ liệu của tài liệu được hiển thị cho người dùng.

**Cross-Reference Table**

Bảng tham chiếu chéo chứa thông tin cho phép truy cập ngẫu nhiên vào các đối tượng gián tiếp trong tệp để không cần phải đọc toàn bộ tệp để định vị bất kỳ đối tượng nào.

**File Trailer**

Đoạn giới thiệu của tệp PDF cho phép người đọc tuân thủ nhanh chóng tìm thấy bảng tham chiếu chéo và một số đối tượng đặc biệt. Người đọc phù hợp nên đọc tệp PDF từ cuối.

## **1.3. Nội dung thực hành**

* *Link Github:* [*https://github.com/chumaii/ptit-pdfmalware*](https://github.com/chumaii/ptit-pdfmalware)
* Khởi động bài lab:

Vào terminal, gõ: *labtainer ptit-pdfmalware*

(Chú ý: sinh viên sử dụng mã sinh viên của mình để nhập thông tin người thực hiện bài lab khi có yêu cầu, để sử dụng khi chấm điểm)

Sau khi khởi động xong một terminal ảo sẽ xuất hiện, đại diện cho máy ptit-pdfmalware

Trong bài lab này sẽ có sẵn tập tin PDF có tên secret.pdf và các tập tin pdfid.py, peepdf.py, pdf-parser.py là bộ ba công cụ để thực hiện phân tích.

* Sinh viên làm các nhiệm vụ dưới đây để hoàn thành bài lab:
* Nhiệm vụ 1: Xác định các đối tượng đáng ngờ bằng công cụ pdfid.
* Đối tượng là những gì chúng ta đang theo đuổi. Nó có thể là một hình ảnh, javascript, biểu mẫu, nội dung văn bản, v.v.
* Đầu tiên sinh viên sử dụng câu lệnh: *python pdfid.py secret.pdf*

để kiểm tra các luồng của tập tin PDF.

Ý nghĩa các luồng xuất hiện:

**/Encrypt** – Biểu thị rằng PDF được mã hóa và cần có khóa để giải mã nó

**/Objstm** - Điều này được sử dụng để xác định luồng đối tượng có thể ẩn các đối tượng cụ thể khác có thể là văn bản, thành phần đồ họa hoặc tập lệnh. Sẽ là nơi tốt để ẩn đối tượng độc hại

**/JS** hoặc **/JavaScript** - JavaScript được nhúng trong tài liệu. Bất kỳ Javascript độc hại nào cũng có thể được tìm thấy từ đây

**/AA** - Điều này xác định Hành động tự động được nhúng trong tài liệu và sẽ tự động kích hoạt khi người dùng mở tài liệu.

**/Acroform** hoặc **/XFA** - Điều này cho biết các biểu mẫu Adobe có được sử dụng trong tài liệu PDF hay không

**/Launch** - Thao tác này khởi chạy một chương trình

* Sau khi pdfid đưa ra kết quả, hãy tìm các luồng được đề cập ở trên có giá trị **Khác 0** để điều tra thêm
* Nhiệm vụ 2: Xác định số tham chiếu hoặc id đối tượng (các số trỏ đến đối tượng quan tâm) bằng công cụ peepdf.
* Khi các đối tượng đáng ngờ được tìm thấy từ pdfid, đã đến lúc đi sâu vào các đối tượng bằng cách sử dụng peedpf và trình phân tích cú pháp pdf
* Peepdf được sử dụng riêng để xác định Id đối tượng hoặc số tham chiếu của các đối tượng đáng ngờ
* Sinh viên gõ lệnh: *./peepdf/peepdf.py secret.pdf*
* Quan sát kết quả trả về của peepdf để biết các luồng cần điều tra có bao nhiêu object và là object số mấy.
* Nhiệm vụ 3: Truy xuất dữ liệu thô từ các đối tượng đáng ngờ bằng công cụ pdf-parser.
* Pdf-parser là công cụ giúp phân tích cấu trúc tập tin PDF, tìm kiếm và trích xuất thông tin của các đối tượng trong tập tin PDF.
* Sau khi đã biết được id của các đối tượng đáng ngờ ở phần trước, sinh viên gõ lệnh sau để thực hiện truy xuất nội dung của đối tượng cụ thể:

*python pdf-parser.py secret.pdf –o <id\_đối tượng>*

* Quan sát kết quả trả về cho biết ý nghĩa nội dung của các đối tượng đáng ngờ.
* Nhiệm vụ 4: Đọc tập tin secret.pdf để xem tên tập tin cụ thể được chèn vào
* Sinh viên chạy lệnh: *cat secret.pdf*

và cho biết title của tập tin HTML được nhúng và là gì.

* Giải thích cách mã độc hoạt động thông qua đoạn mã Javascript.
* In ra màn hình đường dẫn URL mà mã độc kết nối đến.
* Kết thúc bài lab:

Trên terminal đầu tiên sử dụng câu lệnh sau để kết thúc bài lab:

*stoplab ptit-pdfmalware*

Khi bài lab kết thúc, một tệp zip lưu kết quả được tạo và lưu vào một vị trí được hiển thị bên dưới stoplab.

* Khởi động lại bài lab:

Trong quá trình làm bài sinh viên cần thực hiện lại bài lab, dùng câu lệnh:

*labtainer -r pdfmalware*