**Tóm tắt bài báo**

***Tóm Tắt Nội Dung Chính:***

* **Vấn Đề Nghiên Cứu**: Nghiên cứu tập trung vào việc tìm kiếm các ví dụ đối kháng để lừa gạt các phân loại malware PDF dựa trên học máy (ML). Vấn đề này được cho là thách thức hơn so với các mô hình ML dành cho xử lý hình ảnh do cấu trúc phức tạp của tệp PDF và yêu cầu phải duy trì hành vi độc hại của PDF được tạo ra.
* **Giải Pháp Đề Xuất**: Đề xuất một biến thể của mạng Generative Adversarial Networks (GANs) để tạo ra các biến thể malware PDF lẩn trốn (không bị lỗi), có thể được phân loại là lành tính bởi các bộ phân loại hiện có nhưng vẫn giữ nguyên hành vi độc hại ban đầu.
* **Quá Trình Đánh Giá**: Kỹ thuật được đánh giá trên ba bộ phân loại malware PDF tiêu biểu và hiệu quả cũng được kiểm tra với các công cụ chống virus từ VirusTotal.

***Chi tiết kỹ thuật:***

* **Mô Hình Đề Xuất**: Mô hình GAN đề xuất sử dụng bộ phân loại mục tiêu như một bộ phân loại thứ hai để tạo ra các biến thể PDF lẩn trốn nhanh chóng, thông qua quy trình lựa chọn đặc điểm mới bao gồm các đặc điểm độc đáo được trích xuất từ các tệp PDF độc hại.
* **Thử Nghiệm và Hiệu Quả**: Mô hình tìm thấy các biến thể lẩn trốn một cách nhanh chóng cho tất cả các mẫu đã chọn, đặt ra mối quan tâm về an ninh nghiêm trọng trong tình huống có kẻ xâm nhập.

***Kết quả đánh giá:***

* Mô hình đạt hiệu suất cao trong việc tạo ra các biến thể lẩn trốn mà không bị lỗi, với số lượng tối đa là 12 thao tác thay đổi và hiệu quả cao hơn 13 lần so với các phương pháp trước đây.
* Các mẫu thử nghiệm được chọn từ kho lưu trữ malware Contagio và được kiểm tra với các bộ phân loại malware PDF cũng như các công cụ chống virus từ VirusTotal.

***Nguồn dữ liệu:***

1. **Thu thập dữ liệu từ VirusTotal**: Các mẫu malware PDF được thu thập từ VirusTotal. Tổng cộng có 10,673 tệp trong bộ dữ liệu này.
2. **Sử dụng dữ liệu từ Contagio:** Ngoài ra, bộ dữ liệu từ Contagio cũng được sử dụng, bao gồm 9,109 tệp PDF lành tính và 11,105 tệp PDF độc hại.
3. **Lựa chọn mẫu CVE từ Exploit-db**: Để mở rộng phạm vi nghiên cứu, các mẫu CVE được chọn từ Exploit-db, một nguồn cung cấp các mã và tệp chứng minh khái niệm (PoC) cho các lỗ hổng bảo mật đã biết.
4. **Phân tích động Cuckoo Sandbox**: Các tệp từ bộ dữ liệu Contagio được phân tích bằng hệ thống phân tích động Cuckoo Sandbox. Từ 6,105 tệp, 1,503 tệp cho thấy hoạt động mạng độc hại và cuối cùng, 500 mẫu được chọn ngẫu nhiên cho quá trình đánh giá.
5. **Xử lý và lọc dữ liệu**: Trong quá trình xử lý, các tệp được phân tích để xác nhận cấu trúc cây của chúng và lọc ra những tệp có cùng đặc điểm. Cuối cùng, từ 712 tệp độc đáo còn lại sau quá trình lọc, 709 tệp được xác nhận là độc hại bởi các bộ phân loại mục tiêu.

**Giai đoạn 1**: Tiền xử lý PDF (Pre-processing of PDF):

-Using PDFrw để phân tích cú pháp file PDF và biến chúng thành dạng biểu diễn cây.

**A diagram of a document

Description automatically generated**

- Từ biểu diễn cây, một tập đặc tính có thể được hình thành. Mỗi đường dẫn từ nút gốc đến nút lá và giá trị của nó được coi là một đặc điểm. Mọi giá trị trong kiểu chuỗi đều được chuyển đổi thành giá trị nguyên là 1 và một giá trị được cho dưới dạng mảng các giá trị sẽ được chuyển đổi thành giá trị trung bình của các giá trị trong một mảng.

A diagram of a computer language

Description automatically generated  
  
**Giai đoạn 2**: Huấn luyện PDF-GAN (PDF-GAN Training): Giai đoạn này tập trung vào việc huấn luyện mô hình GAN. Mô hình này bao gồm bốn phần: bộ tạo (generator), bộ phân biệt (discriminator), bộ phân loại nguyên mẫu (surrogate classifier) và bộ phân loại thù địch (adversarial classifier). Mục tiêu của bộ tạo là tạo ra các biến thể của file PDF gốc sao cho nó có thể qua mặt bộ phân loại mục tiêu, trong khi bộ phân biệt và bộ phân loại cố gắng phân biệt và phân loại chính xác các mẫu dữ liệu.  
  
**Giai đoạn 3**: Phát hiện (Detection): Giai đoạn cuối cùng liên quan đến việc kiểm tra và xác nhận rằng các mẫu PDF độc hại đã được tạo ra bởi PDF-GAN có thể qua mặt các bộ phân loại phần mềm độc hại và vẫn duy trì hành vi độc hại của chúng.