Phương pháp phát hiện xâm nhập hiệu quả dựa trên học tổng hợp kháng mẫu biến thể trốn tránh

Trần Thị Minh Thư[1,2]

[1] Trường ĐH Công nghệ Thông tin [2] Đại học Quốc gia, Tp. Hồ Chí Minh

What ?

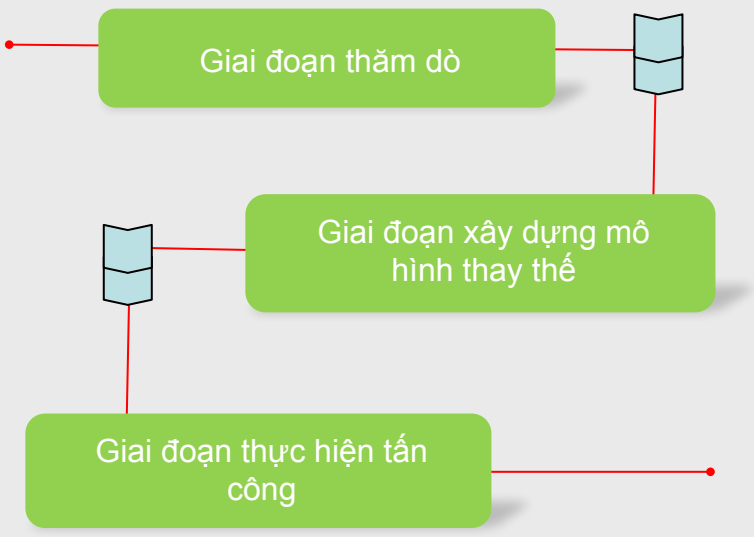
Giới thiệu về framework GPMT có các chức năng như sau:

- Tạo lưu lượng đối kháng thực tế, giữ được tính độc hại ban đầu và có khả năng xảy ra trong ngữ cảnh thực tế.
- Giải quyết vấn đề không ổn định của mô hình GAN.
- Giải quyết vấn đề sửa đổi feature để ánh xạ vào traffic-space..
- Được đánh giá mức độ hiệu quả bởi các mô hình hộp đen NIDS.

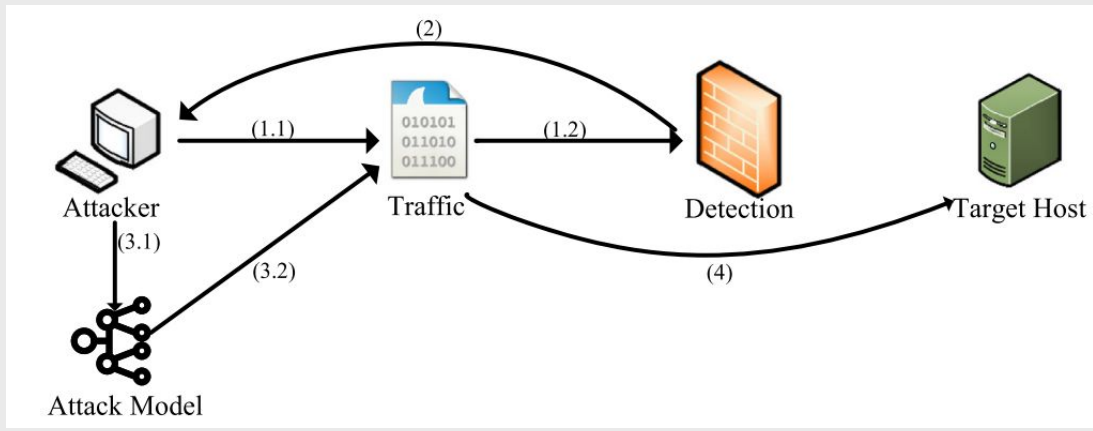
Why ?

- Thực hiện các cuộc tấn công đối kháng để đánh giá khả năng của các mô hình NIDS. Nhưng các phương pháp trước đây chỉ có kết quả tốt trong lý thuyết, không thể sử dụng trong ngữ cảnh thực tế và vi phạm quy tắc giao thức truyền thông hoặc không giữ được tính độc hại ban đầu.
- Nhiều nghiên cứu mô phỏng cuộc tấn công đối kháng vào các mô hình hộp đen trong ngữ cảnh thực tế bằng cách sử dụng kiến trúc GAN nhưng GAN không có tính ổn định trong quá trình tạo lưu lượng đối kháng.

Overview



Input: lưu lượng độc hại ban đầu.
Output: lưu lượng đối kháng trốn tránh mô hình hộp đen NIDS.



Description

1. Giai đoạn thăm dò

- Gửi lưu lượng độc hại vào mô hình hộp đen để nhận được các nhãn độc hại từ phản hồi của mô hình.
- Chọn 7 mô hình ML-based chính thống làm mô hình hộp đen.

2. Giai đoạn xây dựng mô hình thay thế

- Framework GPMT dựa trên WGAN để giải quyết tính không ổn định của GAN.
- Sử dụng Discriminator của WGAN làm mô hình thay thế cục bộ để có thể đào tạo lại nhiều lần.
- Mô hình thay thế có đầu vào là lưu lượng độc hại ban đầu (cụ thể là các tính năng temporal-spatial từ Time Series và Length Series) và nhãn thu được từ giai đoạn 1.
- Đầu ra của mô hình thay thế là phân loại nhị phân bắt chước mô hình hộp đen.

3. Giai đoạn thực hiện tấn công

- Generator của WGAN thêm nhiễu vào các tính năng temporal-spatial để tạo tính năng đối kháng.
- Đưa vào Discriminator để tiến hành phân loại mẫu giả từ Generator hay mẫu thật từ tập dữ liệu ban đầu.
- Nếu Discriminator phân loại đúng, đưa trở lại Generator để chỉnh sửa tạo mẫu đối kháng mạnh mẽ hơn.
- Quá trình là vòng lặp cho đến khi có thể đánh lừa Discriminator.
- Các tính năng temporal-spatial được sử dụng phổ biến trong các mô hình NIDS, dễ ánh xạ vào traffic-space nên có thể tạo lưu lượng đối kháng.
- Sửa đổi bằng cách trì hoãn thời gian gửi các gói và tăng chiều dài gói tin.
- Hàm mất mát (Wasserstein loss) được thiết kế để cải thiện khả năng của Discriminator và Generator.

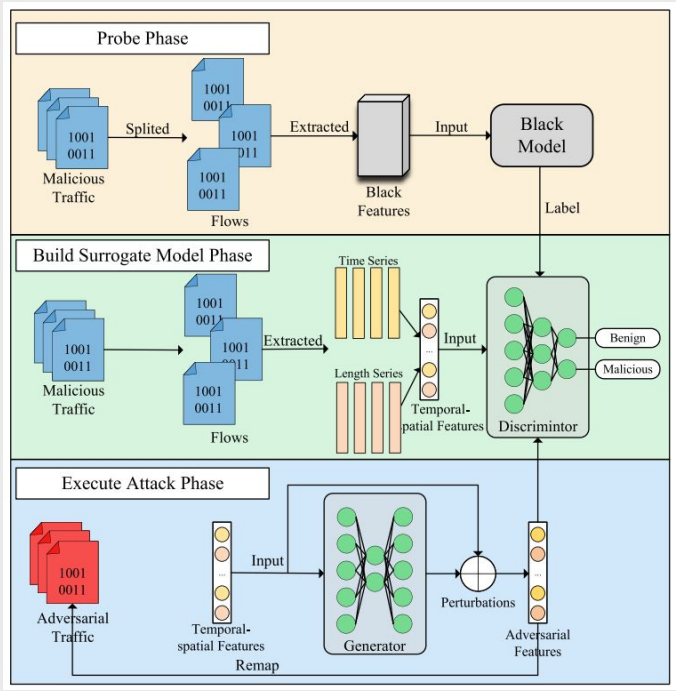


Figure 3. Kiến trúc framework GPMT dựa trên WGAN

4. Kết quả dự kiến

- Các mô hình hộp đen có hiệu suất phân loại tuyệt vời trên tập dữ liệu đào tạo.
- GPMT có hiệu quả trong việc tạo lưu lượng đối kháng thực tế mà vẫn giữ được tính độc hại ban đầu.

NII

Trần Thị Minh Thư – Trường Đại học Công nghệ Thông tin - ĐHQG, TPHCM

TEL : * Email : ***