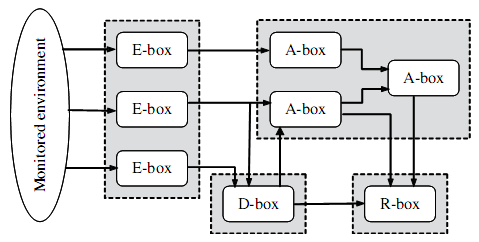
# Anomaly-based network intrusion detection: techniques, systems and challenges (2008)

Keywords: Network security, Threat, Intrusion detection, Anomaly Detection, IDS

Paper này giới thiệu các thong tin tổng quát về IDS, nội dung của paper là về các nền tảng chính của Anomaly- Network IDS (A-NIDS) và phân loại chúng. Thông qua đó xác định những vấn đề cần giải quyết đối với mỗi loại.

Định nghĩa: IDS là một công cụ về bảo mật có tác dụng chính là làm tăng khả năng bảo mật cho các hệ thống thong tin và hệ thống lien lạc.



Trên đây là kiến trúc tổng quát của IDS của group CIDF (Common Intrusion Detection Framework) đưa ra vào năm 1998.

* E-Box: chứa các network sensor để nhận các thong tin, sự kiện cho việc phân tích ở các box khác
* D-box: là các thành phần để lưu trữ thong tin từ E-box rồi cung cấp thong tin một cách tuần tự cho các box phía sau.
* A box: phân tích các sự kiện và xác định các hành vi nguy hiểm
* R-box: thực thi các tác vụ cần thiết khi phát hiện tấn công.

Một điểm khá quan trọng ở đây là cần phải chú ý về việc định nghĩa các giao thức và định dạng của message để tương tác giữa các thành phần.

Phân loại:

* Dựa trên nguồn thong tin:
  + Host-based: phân tích các thong tin dựa trên OS
  + Network-based
* Dựa trên cách phân tích thong tin:
  + Signature-based: phân tích thong tin dựa trên các pattern đã định nghĩa trước.
  + Anomaly-based: ước tính các tác động hiện tại và so sánh nó với tác động bình thường, nếu nó vượt quá một ngưỡng định trước thì đưa ra cảnh báo. Cũng có thể so sánh với các tác động bất thường được định nghĩa, nếu thấp hơn mức định sẵn thì cảnh báo.

Quá trình phân tích của kĩ thuật anomaly-based thong thường qua 3 giai đoạn:

* Parameterization: các dữ liệu thu được sẽ được chuyển về các dạng định trước
* Training: các tác động bình thường (hoặc bất thường) được xây dựng.
* Detection: dựa trên thong tin thu được và trained data, so sánh để xác định tấn công.

Kĩ thuật anomaly-based của Network IDS được phân thành 3 nhánh chính: để so sánh các hướng, có thể đánh giá dựa trên hiệu quả và chi phí (tài nguyên, tiền bạc) của hướng đó

* Statistical-based: các hoạt động diễn ra sẽ được lưu và được thống kê lại dựa trên các hệ thống đo định trước. Có thể chia làm 3 loại:
  + Univariate model: mỗi biến cho mỗi thuộc tính sẽ được tính riêng biệt nhau
  + Multivariate model: đánh giá có xem xét đến sự tương tác qua lại giữa các thuộc tính
  + Time series model: model này bao gồm một interval timer, bộ đếm sự kiện và thước đo giá trị cho resource. Nó sẽ kiểm tra về thứ tự, khoảng thời gian, giá trị của các sự kiện thu được, rồi thong qua các công thức xác suất để xác định tấn công.
* Knowledge-based: được sử dụng khá phổ biến hiện nay. Nó sẽ phân loại các thong tin dựa vào các luật định trước. Đầu tiên các thuộc tính sẽ được chỉ định trong quá trình training. Sau đó các luật, thong số và phương thức phân loại sẽ được chỉ định. Cuối cùng, thong tin sẽ được phân loại dựa trên các dữ liệu có trước đó.
* Machine learning-based: xác định các dữ liệu dung để train. Có khá nhiều mô hình được ứng dụng vào hướng này:
  + Bayesian networks: kết quả thu được giống với hệ thống statistical nhưng cần sử dụng nhiều nỗ lực tính toán hơn. Kết quả dựa nhiều vào các giả thiết về hệ thống vì thế dễ dẫn đến dự đoán sai.
  + Markov models: cũng giống như trên dựa quá nhiều vào giả thiết.
  + Neural networks:
  + Fuzzy logic techniques: kĩ thuật này khá tốn kém tài nguyên hệ thống
  + Genetic algorithm: có thể tìm được kết quả theo nhiều hướng dựa trên các kĩ thuật search mạnh và linh hoạt, mà không cần giả thiết về hệ thống. nhưng tốn kém tài nguyên
  + Clustering and outlier detection: các dữ liệu thu được sẽ được nhóm lại thành các cluster. Mỗi cluster sẽ được biểu diễn bởi các điểm. sẽ có các điểm không thuộc cluster nào sẽ là các outlier và được dung cho quá trình phân tích. Kĩ thuật này dung để xác định các tấn công từ các dữ liệu thô.

Ngoài các kĩ thuật trên, còn cần phải chú ý đến việc xử lý với lượng thong tin trong các tập dữ liệu. Có 2 kĩ thuật lien quan là:

* Principal component analysis(PCA): để giảm độ phức tạp của tập dữ liệu
* Association rules discovery: có mục đích đạt được thong tin từ các dữ liệu khác nhau từ trong tập dữ liệu training.

Kĩ thuật Data mining: bao gồm các giải thuật learning vào các khối lượng dữ liệu lớn để xác định các thong tin hữu ích. Kĩ thuật này cũng đã được ứng dụng vào việc correlate traffic instrances trong các cơ sở dữ liệu lien quan đến network và nó cũng là một hướng phát triển của IDS.

Trong paper còn liệt kê một vài hệ thống A-NIDS hiện có dựa trên các nền tảng và lĩnh vực hướng nghiên cứu.