ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH **TRƯỜNG ĐẠI HỌC**

**CÔNG NGHỆ THÔNG TIN**

**CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM Độc Lập - Tự Do - Hạnh Phúc**

**ĐĂNG KÝ ĐỀ TÀI KHÓA LUẬN TỐT NGHIỆP**

|  |
| --- |
| **TÊN ĐỀ TÀI: TRIỂN KHAI HONEYPOT KẾT HỢP VỚI DEEP LEARNING ĐỂ MÔ PHỎNG HÀNH VI MÁY CHỦ** |
| **TÊN ĐỀ TÀI TIẾNG ANH: HONEYPOT IMPLEMENTATION BASED ON DEEP LEARNING TO SIMULATE SERVER BEHAVIOR** |
| **Cán bộ hướng dẫn:** |
| **Thời gian thực hiện:** Từ ngày …/2024 đến ngày …/2024. |
| **Sinh viên thực hiện:**  Nguyễn Hoàng Thảo Quyên - 230202014 **Lớp:** CS2205.CH181  **Email:** quyennht.18@grad.uit.edu.vn **Điện thoại:** |
| **Nội dung đề tài:**  **Giới thiệu:**  Trong thế giới công nghệ ngày nay, với sự phát triển không ngừng của công nghệ internet, mối đe dọa về an ninh mạng ngày càng trở nên nghiêm trọng và phức tạp hơn bao giờ hết. Để bảo vệ mạng lưới của khỏi các cuộc tấn công mạng, các chuyên gia an ninh thông tin đang ngày càng quan tâm đến việc sử dụng honeypot - một công cụ mạng được thiết kế để thu hút và ghi lại các cuộc tấn công mạng.  Tuy nhiên, dù honeypot đã trở thành một phần không thể thiếu trong chiến lược an ninh mạng của các tổ chức, nhưng vẫn có vấn đề tồn tại: tính chân thực của honeypot. Các honeypot hiện tại thường không thể mô phỏng hành vi của máy chủ một cách chân thật. Điều này là một rủi ro tiềm ẩn, khi mà tin tặc có thể dựa vào một vài dấu hiệu để phát hiện và tránh qua các honeypot.  Để giải quyết vấn đề này, trong nghiên cứu này, tôi đề xuất một phương pháp mới để xây dựng honeypot thông minh, sử dụng dữ liệu thu thập được từ máy chủ thực tế và kỹ thuật học máy sâu để tạo ra một honeypot có khả năng mô phỏng hành vi máy chủ chân thực và đáng tin cậy từ cách giao tiếp đến phản ứng với các cuộc tấn công mạng. Cụ thể:  *Input:* các câu lệnh từ hệ thống.  *Output:* phản hồi giống như máy chủ thật, bao gồm việc mô phỏng các tiến trình, cấu hình hệ thống, ghi lại các sự kiện hệ thống…  **Mục tiêu:**   * Tìm hiểu các honeypot có khả năng mô phỏng hành vi của máy chủ thật một cách chân thực để tin tặc khó phân biệt được với máy chủ thực. * Sử dụng dữ liệu thực tế thu thập được từ máy chủ thật để huấn luyện mô hình máy học để tạo ra một honeypot có khả năng mô phỏng hành vi dựa trên dữ liệu thực tế, giúp tăng tính chân thực và hiệu quả của honeypot. * Sử dụng các thuật toán học máy sâu như mạng nơ-ron hồi quy (RNNs), và mạng tăng cường (GANs) để xây dựng một mô hình mô phỏng hành vi cho honeypot.   **Phạm vi:**   * Dữ liệu được thu thập từ máy chủ sẽ bao gồm cấu hình hệ thống, các tiến trình của các máy chủ thật, các dữ liệu từ syslog để biết được các phản hồi của máy chủ thực đến các câu lệnh nhận vào. * Máy chủ honeypot sẽ được giới hạn là một máy chủ web nên sẽ nhận thêm các dữ liệu về gói tin mạng như HTTP, HTTPS, DNS.   **Đối tượng:**   * Các thuật toán GAN và RNN sẽ được sử dụng trong đề tài này.   **Nội dung:**   * Nghiên cứu thuật toán GAN và RNNs, xây dựng mô hình học máy, huấn luyện mô hình để xác định độ đo của từng mô hình trên tập dữ liệu. * Chọn dùng các bộ dữ liệu như Syslog Dataset, HTTP Archive,… để huấn luyện cấu hình hệ thống và các bộ dữ liệu như Apache HTTP Server Logs, Nginx HTTP Server Logs, … để huấn luyện về các gói tin mạng. * Huấn luyện mô hình theo các bộ dữ liệu trên và đánh giá hệ thống. * Xây dựng honeypot và chạy thử trên môi trường thực tế.   **Phương pháp:**   * Dùng GAN để tự sinh và tự đánh giá các phản hồi có chân thực hay không, sau đó dùng mô hình RNN để tạo ra các biến thể của các phản hồi này. RNN có thể học và tái tạo lại các mẫu từ dữ liệu được tạo ra bởi GAN, giúp tạo ra các phản hồi mô phỏng chân thực hơn. * Tìm hiểu cách đánh giá mô hình GAN bằng độ đo Precision và Recall, đánh giá mô hình RNN bằng Perplexity. * Huấn luyện mô hình GAN và RNN, so sánh và đánh giá kết quả. * Dựng honeypot trong môi trường thử nghiệm và tìm cách áp dụng mô hình vừa huấn luyện vào.   **Kết quả dự kiến:**   * Một honeypot áp dụng mô hình máy học đã huấn luyện với các thông số đạt yêu cầu có thể phân tích các loại tấn công theo mẫu tấn công và phản hồi dựa theo kết quả phân tích. * Báo cáo các phương pháp và kỹ thuật của mô hình máy học đã phát triển được. Kết quả thực nghiệm, đánh giá, so sánh với các phương pháp khác.   **Tài liệu tham khảo:**  [1] Siniosoglou, I., Efstathopoulos, G., Pliatsios, D., Moscholios, I. D., Sarigiannidis, A., Sakellari, G., … Sarigiannidis, P. (2020). NeuralPot: An Industrial Honeypot Implementation Based On Deep Neural Networks. 2020 IEEE Symposium on Computers and Communications (ISCC). doi:10.1109/iscc50000.2020.9219712  [2] Fan, W., Du, Z., Smith-Creasey, M., & Fernandez, D. (2019). HoneyDOC: An Efficient Honeypot Architecture Enabling All-Round Design. IEEE Journal on Selected Areas in Communications, 1–1. doi:10.1109/jsac.2019.2894307  [3] Jiang, K., & Zheng, H. (2020). Design and Implementation of A Machine Learning Enhanced Web Honeypot System. 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). doi:10.1109/cisp-bmei51763.2020.9263640  [4] Mehta, V., Bahadur, P., Kapoor, M., Singh, P., & Rajpoot, S. (2015). Threat prediction using honeypot and machine learning. 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE). doi:10.1109/ablaze.2015.7155011  [5] Matin, I. M. M., & Rahardjo, B. (2020). The Use of Honeypot in Machine Learning Based on Malware Detection: A Review. 2020 8th International Conference on Cyber and IT Service Management (CITSM). doi:10.1109/citsm50537.2020.9268794 |
| **Kế hoạch thực hiện:**  + Tuần 1 - 6: Tìm hiểu các thuật toán GAN và RNN phù hợp, tìm hiểu và chọn dùng các bộ dữ liệu phù hợp với mô hình.  Kết quả dự kiến:   * Tài liệu chi tiết cấu trúc các mô hình GAN và RNN. * Tài liệu về độ đo Precision và Recall, độ đo Perplexity. * Các bộ dữ liệu được sử dụng để huấn luyện mô hình.   + Tuần 3 - 12: Huấn luyện thuật toán GAN và RNN, ghi chép lại kết quả kèm đánh giá và so sánh.  Kết quả dự kiến:   * Bảng kết quả đánh giá và theo dõi thực nghiệm của thuật toán GAN và RNN đã phát triển dựa trên các bộ dữ liệu đã chọn.   + Tuần 10 - 16: Tim hiểu các loại honeypot và lựa chọn một honeypot phù hợp để xây dựng chương trình demo trên môi trường thử nghiệm, đồng thời áp dụng các mô hình máy học đã huấn luyện vào chương trình này.  Kết quả dự kiến:   * Tài liệu chi tiết cấu trúc mô hình honeypot sử dụng trong đề tài. * Chương trình honeypot đang chạy đã áp dụng mô hình GAN và RNN vào. |

|  |  |
| --- | --- |
| **Xác nhận của CBHD**  (Ký tên và ghi rõ họ tên) | **TP. HCM, ngày….tháng …..năm 2024**  **Sinh viên**  (Ký tên và ghi rõ họ tên) |