

Report project



**Trích xuất các thuộc tính tập tin PE
phục vụ phát hiện mã độc Windows**

**A study on extracting PE file features
for Windows malware detection**

NT114.021.ANTN



Nguyễn Vũ Anh Duy Nguyễn Văn Khang Kim

CONTENT

01.

Tổng quan

02.

Phương pháp nghiên cứu

03.

Kết quả

04.

Hướng nghiên cứu tiếp theo

05.

Kết luận

TỔNG QUAN

01

Ngữ cảnh

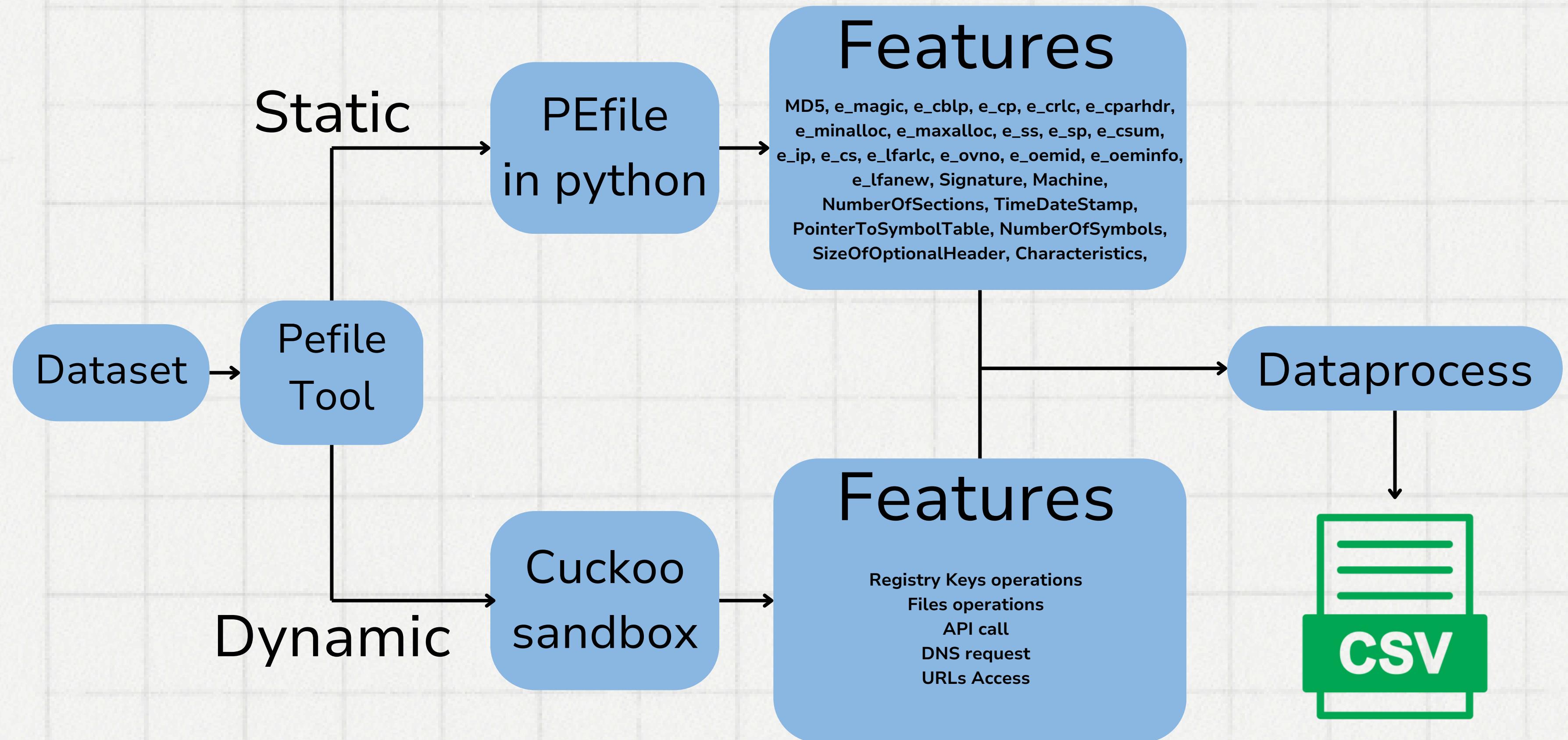
- Trong vài năm trở lại đây, mã độc windows đang ngày càng phổ biến và nguy hiểm hơn.
- Để phát hiện và ngăn chặn mã độc các phương pháp học máy đang đem lại kết quả tốt.
- Để phương pháp học máy có thể phát triển và chính xác nhất, cần phải có một tập dữ liệu đủ lớn và đầy đủ.
- Bài nghiên cứu này sẽ tìm hiểu về việc trích xuất thuộc tính các tập tin PE (Portable Executable) trong Windows để tạo ra một tập dữ liệu đa dạng về thuộc tính bao gồm cả thuộc tính tĩnh và động.

Mục tiêu

- Trích xuất thuộc tính tĩnh và động tập tin PE.

02

PHƯƠNG PHÁP NGHIÊN CỨU



Trích xuất thuộc tính tĩnh

- Khi phân tích mã độc bằng Machine Learning, các thuộc tính tĩnh đóng vai trò quan trọng trong việc xác định đặc điểm và hành vi của mã độc mà không cần thực thi mã.
- Các thuộc tính tĩnh:
 - *MD5, e_magic, e_cblp, e_cp, e_crlc, e_cparhdr, e_minalloc, e_maxalloc, e_ss, e_sp, e_csum, e_ip, e_cs, e_lfarlc, e_ovno, e_oemid, e_oeminfo, e_lfanew, Signature, Machine, NumberOfSections, TimeStamp, PointerToSymbolTable, NumberOfSymbols, SizeOfOptionalHeader, Characteristics, Magic, MajorLinkerVersion, MinorLinkerVersion, SizeOfCode, SizeOfInitializedData, SizeOfUninitializedData, AddressOfEntryPoint, BaseOfCode, BaseOfData, ImageBase, SectionAlignment, FileAlignment, MajorOperatingSystemVersion, MinorOperatingSystemVersion, MajorImageVersion, Rawsize, SectionsMeanVirtualsize, SectionsMinVirtualsize, ...*

Trích xuất thuộc tính tĩnh

- Thư viện pefile trong Python là một công cụ mạnh mẽ để phân tích các tập tin PE trên hệ điều hành Windows.
- Thư viện pefile cung cấp các phương pháp và chức năng để trích xuất và phân tích thông tin từ các tập tin PE, giúp các nhà nghiên cứu bảo mật và nhà phát triển phần mềm có thể phân tích mã độc và các tập tin thực thi khác một cách hiệu quả.
- Threading: Một tập dữ liệu sẽ có từ vài ngàn đến mấy chục ngàn tệp thực thi, để tăng tốc cho quá trình trích xuất ta có thể sử dụng phân luồng để tận dụng tối đa hiệu suất của máy.

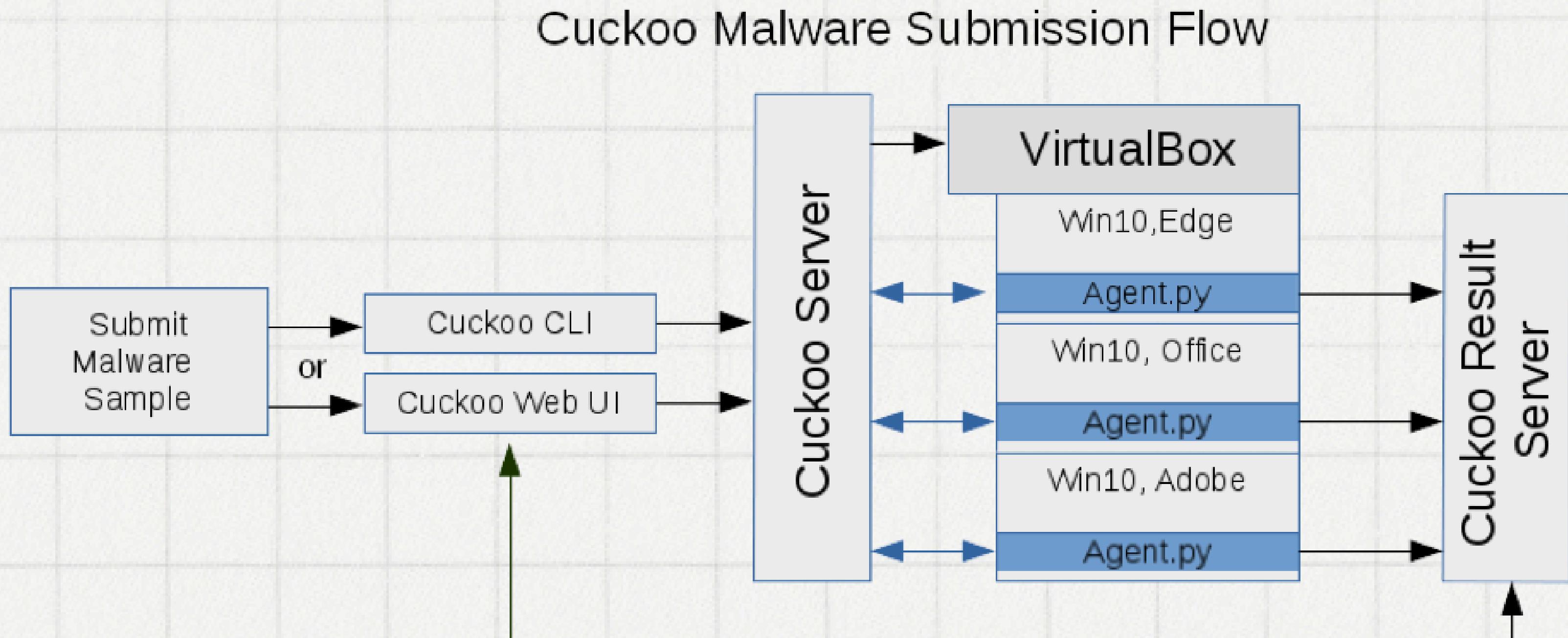
Trích xuất thuộc tính động

- Thuộc tính động của mã độc là các đặc điểm được thu thập khi mã độc được thực thi trong một môi trường giám sát, chẳng hạn như sandbox. Những thuộc tính này giúp xác định hành vi thực tế của mã độc và cung cấp thông tin chi tiết về cách mã độc tương tác với hệ thống và mạng lưới.
- Các thuộc tính động được trích xuất:
 - Registry Keys operations: Chứa thông tin về thay đổi hay truy cập sổ đăng ký như việc ghi sổ đăng ký, xóa sổ đăng ký, mở và đọc.
 - Files operations: Chứa thông tin về tệp đã tạo, tệp đã sửa đổi, tệp đã xóa và số lượng tệp bị lỗi.
 - API calls: API là một tập hợp các lệnh gọi chương trình con hoặc hàm được sử dụng để liên lạc giữa hai thành phần phần mềm hoặc liên lạc giữa các thành phần phần mềm và phần cứng.
 - DNS requests: Số lượng truy vấn DNS được trích xuất dưới dạng các tính năng trong phần thông tin tóm tắt.
 - URLs Access: Số lượng các URL được truy cập.

Cuckoo sandbox

- Hộp cát (sandbox) là một môi trường cô lập và an toàn được tạo ra để thực thi các ứng dụng, chương trình, hoặc tệp tin mà không ảnh hưởng đến hệ thống chủ. Mục đích chính của sandbox là cung cấp một không gian thử nghiệm an toàn để kiểm tra và phân tích các ứng dụng mà không cần lo lắng về các tác động tiềm ẩn đến hệ thống hoặc dữ liệu quan trọng.
- Cuckoo Sandbox là một hệ thống phân tích mã độc mã nguồn mở mạnh mẽ, được thiết kế để tự động phân tích các tập tin đáng ngờ trong các môi trường ảo hóa. Cuckoo Sandbox cho phép người dùng thực thi và theo dõi hành vi của mã độc trong một môi trường cô lập, thu thập thông tin chi tiết về hành vi của mã độc và các tác động của nó đến hệ thống

Cuckoo sandbox



Cuckoo API

- Cuckoo API là một phần mở rộng của Cuckoo Sandbox cho phép người dùng tương tác và điều khiển Cuckoo Sandbox từ các ứng dụng bên ngoài thông qua giao diện lập trình ứng dụng (API). API cung cấp các phương thức và endpoints để thực hiện các tác vụ như gửi tệp tin để phân tích, lấy kết quả phân tích, quản lý mẫu mã độc và cấu hình môi trường phân tích.
- Một vài API thường dùng.
 - /tasks/create/file: Thêm một tập tin vào danh sách các nhiệm vụ đang chờ xử lý. Trả về ID của tác vụ mới được tạo.
 - /tasks/list: Trả về danh sách các nhiệm vụ.
 - /tasks/sample: Trả về danh sách nhiệm vụ cho mẫu.
 - /tasks/report: Trả về báo cáo được liên kết với ID tác vụ đã chỉ định.
 - /tasks/delete: Xóa tác vụ đã cho khỏi cơ sở dữ liệu và xóa kết quả.

Performance

- Không giống như phân tích tĩnh, hiệu suất phụ thuộc vào tốc độ xử lý của máy. Trong phân tích động, hiệu suất phụ thuộc vào khả năng phân tích mẫu của cuckoo sandbox.
- Theo tìm hiểu, tốc độ phân tích của cuckoo sandbox chủ yếu phụ thuộc vào 2 yếu tố, là số máy ảo cuckoo và khả năng tạo báo cáo của cuckoo server.
- Để nâng cao hiệu suất, cách đơn giản nhất chính là tạo nhiều hơn 1 máy ảo.

03

KẾT QUẢ NGHIÊN CỨU

Dataset

- Dataset (10690)
 - Benign (1000)
 - Benign Test (300)
 - Benign Train (700)
 - Virus (9690)
 - Virus Test (2691)
 - Locker(99), Mediyes(435), Winwebsec(1320) Zbot(630)
 - Zeroaccess(207)
 - Virus Train (6999)
 - Locker(231) Mediyes(1015) Winwebsec(3080) Zbot(1470)
 - Zeroaccess(483)

Thuộc tính tĩnh

- Chạy chương trình trích xuất thuộc tính tĩnh với tập Benign Test.

```
khangkim@khangkimvm:~/Desktop$ python static_v1.py -s Dataset/Benign/Benign_test/ -d ./ -bov 0  
khangkim@khangkimvm:~/Desktop$
```

Thuộc tính tĩnh

• Kết quả

Thuộc tính động

- Khởi động cuckoo api.

```
(cuckoo) cuckoo@khangkimvm:~/cuckoo$ cuckoo api --host 127.0.0.1 --port 8888
/home/cuckoo/.virtualenvs/cuckoo/lib/python2.7/site-packages/sflock/decode/office.py:12: CryptographyDeprecationWarning: Python 2 is no longer supported by the Python core team. Support for it is now deprecated in cryptography, and will be removed in the next release.
    from cryptography.hazmat.backends import default_backend
2024-07-04 21:17:30,041 [werkzeug] INFO: * Running on http://127.0.0.1:8888/ (Press CTRL+C to quit)
2024-07-04 21:19:31,818 [werkzeug] INFO: 127.0.0.1 - - [04/Jul/2024 21:19:31] "GET /machines/list HTTP/1.1" 401 -
2024-07-04 21:32:13,155 [werkzeug] INFO: 127.0.0.1 - - [04/Jul/2024 21:32:13] "GET /machines/list HTTP/1.1" 200 -
```

Thuộc tính động

- Kiểm tra các máy ảo.

```
cuckoo@khangkimvm:/home/khangkim/Desktop$ curl -H "Authorization: Bearer BHjsFvgfjQN7hiYXnVJjEg" http://localhost:8888/machines/list
{
  "machines": [
    {
      "id": 1,
      "interface": "vboxnet0",
      "ip": "192.168.56.102",
      "label": "cuckoo21",
      "locked": false,
      "locked_changed_on": "2024-07-05 00:34:12",
      "name": "cuckoo21",
      "options": [],
      "platform": "windows",
      "rcparams": {},
      "resultserver_ip": "192.168.56.1",
      "resultserver_port": 2042,
      "snapshot": null,
      "status": "poweroff",
      "status_changed_on": "2024-07-05 00:34:12",
      "tags": []
    },
    {
      "id": 2,
      "interface": "vboxnet0",
      "ip": "192.168.56.104",
      "label": "cuckoo41",
      "locked": false,
      "locked_changed_on": "2024-07-05 00:33:59",
      "name": "cuckoo41",
      "options": [],
      "platform": "windows",
      "rcparams": {},
      "resultserver_ip": "192.168.56.1",
      "resultserver_port": 2042,
      "snapshot": null,
      "status": "poweroff",
      "status_changed_on": "2024-07-05 00:33:59",
      "tags": []
    }
  ]
}
cuckoo@khangkimvm:/home/khangkim/Desktop$
```

Thuộc tính động

- Chạy chương trình dynamic.py

```
khangkim@khangkimvm:~/Desktop$ python dynamic_v3.py -s Dataset/Benign/Benign_test/
tasks: 300
tasks: 300
tasks: 300
tasks: 300
cvtres.exe: reported
tasks: 299
deinterlace.exe: reported
tasks: 298
tasks: 298
fc-scan.exe: reported
tasks: 297
tail.exe: reported
tasks: 296
toast.exe: reported
tasks: 295
sysprep.exe: reported
tasks: 294
tasks: 294
LockAppHost.exe: reported
tasks: 4
tasks: 4
TDEnvCleanup.exe: reported
tasks: 3
u2d.exe: reported
tasks: 2
octave-cli.exe: reported
tasks: 1
semi-flatten.exe: reported
tasks: 0
khangkim@khangkimvm:~/Desktop$
```

Thuộc tính động

Thuộc tính động

| | A | B | C | D | E | F | G | H | I | J | K |
|----|--|---------------|-------------|----------------|----------------|---|---|---|---|---|---|
| 1 | name | regkey_opened | regkey_read | regkey_written | regkey_deleted | | | | | | |
| 2 | cvtres.exe | 0 | 0 | 0 | 0 | | | | | | |
| 3 | deinterlace.exe | 0 | 0 | 0 | 0 | | | | | | |
| 4 | fc-scan.exe | 0 | 0 | 0 | 0 | | | | | | |
| 5 | tail.exe | 0 | 0 | 0 | 0 | | | | | | |
| 6 | toast.exe | 0 | 6 | 0 | 0 | | | | | | |
| 7 | sysprep.exe | 0 | 0 | 0 | 0 | | | | | | |
| 8 | color-to-alpha.exe | 0 | 0 | 0 | 0 | | | | | | |
| 9 | tr.exe | 0 | 0 | 0 | 0 | | | | | | |
| 10 | echo.exe | 0 | 0 | 0 | 0 | | | | | | |
| 11 | pmsort.exe | 0 | 0 | 0 | 0 | | | | | | |
| 12 | malias.exe | 0 | 0 | 0 | 0 | | | | | | |
| 13 | php-win.exe | 0 | 0 | 0 | 0 | | | | | | |
| 14 | fc-validate.exe | 0 | 0 | 0 | 0 | | | | | | |
| 15 | iconv.exe | 0 | 0 | 0 | 0 | | | | | | |
| 16 | SCANPST.EXE | 0 | 0 | 0 | 0 | | | | | | |
| 17 | sleep.exe | 0 | 0 | 0 | 0 | | | | | | |
| 18 | proquota.exe | 0 | 0 | 0 | 0 | | | | | | |
| 19 | max-rgb.exe | 0 | 0 | 0 | 0 | | | | | | |
| 20 | win7appid.exe | 0 | 0 | 0 | 0 | | | | | | |
| 21 | perl.exe | 0 | 0 | 0 | 0 | | | | | | |
| 22 | datacopy.exe | 0 | 0 | 0 | 0 | | | | | | |
| 23 | ThumbnailExtractionHost.exe | 0 | 16 | 0 | 0 | | | | | | |
| 24 | qconvex.exe | 0 | 0 | 0 | 0 | | | | | | |
| 25 | TsWpfWrp.exe | 76 | 330 | 4 | 0 | | | | | | |
| 26 | twain.exe | 0 | 0 | 0 | 0 | | | | | | |
| 27 | Common.DBConnection.exe | 12 | 6 | 0 | 0 | | | | | | |
| 28 | dbus-update-activation-environment.exe | 0 | 0 | 0 | 0 | | | | | | |
| 29 | UpgradeResultsUI.exe | 0 | 0 | 0 | 0 | | | | | | |
| 30 | editbin.exe | 0 | 0 | 0 | 0 | | | | | | |
| 31 | SETLANG.EXE | 0 | 0 | 0 | 0 | | | | | | |
| 32 | newmail.exe | 0 | 0 | 0 | 0 | | | | | | |
| 33 | d2u.exe | 0 | 0 | 0 | 0 | | | | | | |
| 34 | gsl-randist.exe | 0 | 0 | 0 | 0 | | | | | | |
| 35 | fc-cache.exe | 0 | 0 | 0 | 0 | | | | | | |
| 36 | vshost.exe | 12 | 6 | 0 | 0 | | | | | | |
| 37 | qdelauay.exe | 0 | 0 | 0 | 0 | | | | | | |

Thuộc tính động

Thuộc tính động

| A1 | fx Σ = name | B |
|----|--|--|
| 1 | A | B |
| 1 | name | dns |
| 2 | cvtres.exe | dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 3 | deinterlace.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 4 | fc-scan.exe | dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 5 | tail.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 6 | toast.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 7 | sysprep.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 8 | color-to-alpha.exe | dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 9 | tr.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 10 | echo.exe | dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 11 | pmsort.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 12 | malias.exe | dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 13 | php-win.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 14 | fc-validate.exe | dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 15 | iconv.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 16 | SCANPST.EXE | dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 17 | sleep.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 18 | proquota.exe | dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 19 | max-rgb.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 20 | win7appid.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 21 | perl.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 22 | datacopy.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 23 | ThumbnailExtractionHost.exe | dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 24 | qconvex.exe | dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 25 | TsWpfWrp.exe | www.msftncsi.com.crl.microsoft.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 26 | twain.exe | dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 27 | Common.DBConnection.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 28 | dbus-update-activation-environment.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 29 | UpgradeResultsUI.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 30 | editbin.exe | dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 31 | SETLANG.EXE | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 32 | newmail.exe | dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 33 | d2u.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 34 | gsl-randist.exe | dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 35 | fc-cache.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 36 | vhost.exe | dns.msftncsi.com.teredo.ipv6.microsoft.com, |
| 37 | qdelauay.exe | www.msftncsi.com.dns.msftncsi.com.teredo.ipv6.microsoft.com, |

04

HƯỚNG NGHIÊN CỨU TIẾP THEO

Hướng nghiên cứu tiếp theo

- Thực hiện tìm hiểu và trích xuất các thuộc tính mới.
- Xây dựng tập dataset hoàn chỉnh, chất lượng cao.
- Dùng phương pháp học máy, học sâu đã biết để đánh giá bộ dataset.

05

KẾT LUẬN

Kết luận

- Một dataset chất lượng cao đóng vai trò quan trọng trong việc cải thiện hiệu suất và độ chính xác của các mô hình học máy.
- Dataset chất lượng đảm bảo rằng dữ liệu được cung cấp cho mô hình là chính xác, đầy đủ và đa dạng, giúp mô hình học được các đặc trưng thực sự quan trọng và tránh việc học các mẫu nhiễu hoặc không liên quan.
- Khi dữ liệu huấn luyện phản ánh chính xác các tình huống thực tế mà mô hình sẽ gặp phải trong trường hợp này là các đặc, hành vi của mã độc, mô hình có thể đưa ra các dự đoán chính xác hơn và tổng quát hóa tốt hơn cho các dữ liệu mới.
- Một dataset kém chất lượng có thể dẫn đến mô hình bị sai lệch, kém hiệu quả và dễ bị overfitting hoặc underfitting, do đó làm giảm khả năng áp dụng của mô hình trong thực tế.
- Vì vậy, việc xây dựng và duy trì một dataset chất lượng, đầy đủ, là một yếu tố then chốt trong việc phát triển các giải pháp học máy hiệu quả và tin cậy trong việc phát hiện mã độc Windows.

**Thank you
very much!**