Malware Analysis

A Survey of Visualization Systems for Malware Analysis

GVHD: TS. Trương Tuấn Anh
Học viên:
Đinh Thanh Phong
Trần Đại Vệ

Khảo sát các hệ thống trực quan cho phân tích mã độc

M. Wagner, F. Fischer , R. Luh , A. Haberson , A. Rind , D. A. Keim, and W. Aigner1
St. Poelten University of Applied Sciences, Austria
Vienna University of Technology, Austria
University of Konstanz, Germany

Mục tiêu bài báo cáo:

- Do mối đe dọa ngày càng tăng từ phần mềm độc hại (malware).
- Việc phân tích thực tế số lượng mẫu đáng ngờ ngày càng tăng và tốn nhiều thời gian.
- Việc sử dụng các hệ thống phân tích trực quan và trực quan có tính tương tác cao có thể giúp hỗ trợ quy trình phân tích này đối với việc điều tra, so sánh và tổng hợp các mẫu phần mềm độc hại.

- Bài báo cung cấp một cái nhìn tổng quan có hệ thống và phân loại các hệ thống trực quan hóa phần mềm độc hại từ quan điểm của phân tích trực quan.
- Xác định và đánh giá các nhà cung cấp dữ liệu và công cụ thương mại cho các hệ thống trực quan hóa phần mềm độc hại.

Giới thiệu đề tài

- Do số lượng mẫu quá nhiều và thực tế là phân tích thủ công của các chuyên gia là khá quá tải.
- Các phương pháp phân tích dữ liệu tự động đang rất cần thiết.
- Tuy nhiên, quá trình này không thể được tự động hóa hoàn toàn vì các chuyên gia cần phải tham gia để xác định, sửa chữa và phân biệt các kết quả trung gian, yêu cầu sự phân tích dữ liệu phức tạp và sự kết hợp giữa phân tích dữ liệu tự động với lý luận phân tích của các chuyên gia

- Bài báo cho thấy rằng việc sử dụng trực quan hóa tăng tốc đáng kể quá trình phát hiện phần mềm độc hại.

Các công cụ và thuật ngữ liên quan

- Data providers
- Analysis environments
- Base data

Các công cụ và thuật ngữ liên quan

Data providers

- Là bộ công cụ sử dụng các phương pháp phân tích tĩnh hoặc động (đôi khi cả hai) để thu thập thông tin về một khả năng phần mềm độc hại.
- Static analysis.
- Dynamic analysis (cần Analysis environments).

Môi trường phân tích mã độc
Analysis environments

Analysis environments:

- là môi trường để triển khai hệ thống phân tích phần mềm độc hại tương ứng.
- Tùy thuộc vào khả năng và yêu cầu của nhà cung cấp dữ liệu, các môi trường này có thể là máy vật lý, máy ảo hoặc hệ thống mô phỏng.

Môi trường phân tích mã độc
Analysis environments

Analysis environments:

- Máy Thật (vật lý): máy tính thật thực hiện một lấy mẫu trực tiếp trong hệ điều hành (HĐH).

- Máy vật lý thì thuận lợi cho phần mềm độc hại thực thi, nhưng mẫu độc tiềm ẩn có thể truy cập trực tiếp vào phần cứng mà nó đang chạy và nguy hại đến hệ điều hành.

- Nó cũng quan trọng để hãy nhớ rằng cài đặt lại/đặt lại máy vật lý tốn nhiều thời gian hơn so với việc đặt lại môi trường ảo hóa hoặc mô phỏng.

Môi trường phân tích mã độc
Analysis environments

Analysis environments:

- Emulated systems: là hệ thống không có chung tính chất với máy chủ.
- CPU and memory đều được giả lập.
- Emulated systems có thể bị virus phát hiện và hoạt động khác đi.

Môi trường phân tích mã độc
Analysis environments

Base data:

- Base data: mô tả loại dữ liệu được giám sát và đăng nhập bởi một nhà cung cấp. Có vô số thông tin được thu thập từ phân tích tĩnh và động, mỗi thông tin cung cấp cái nhìn sâu sắc cụ thể về bản chất và chức năng của một chương trình độc hại.
- Virus definition
- Packer information
- File and header information
- Library and function imports
- CPU instructions
- ...

So sánh các Providers

Process Monitor v3.93

Generic disassembler and generic debugger

So sánh các Providers

	Anubis	Cuckoo	CWSandbox	FireEye MAS	Joe Sandbox	ProcMon	APIMon	Generic disassembler	Generic debugger
Analysis mode and environment									
Static analysis support		✓✓		✓	✓✓			✓✓	
Dynamic analysis support	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓		✓✓
Native analysis environment			✓✓		✓✓	✓✓	✓✓	✓✓	✓✓
Virtual machine environment		✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓
Emulation environment	✓✓					✓✓	✓✓	✓✓	✓✓
(Simulated) Internet access	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓		
(Simulated) LAN services	✓✓		✓✓			✓✓	✓✓		
Interface									
Command line interface	✓✓	✓✓		✓✓	✓				
Graphical (web) interface (GUI)	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓
Sample input									
Single file submission	✓✓	✓✓	✓✓	✓✓	✓✓		✓✓	✓✓	✓✓
Folder submission	(✓)	(✓)		✓✓	✓✓				
URL/URI	✓✓	✓✓	✓✓	✓✓	✓✓				
Batch processing	(✓)	(✓)	(✓)	✓✓	(✓)	(✓)	(✓)	(✓)	(✓)
Interactive on-demand analysis	✓					✓✓	✓✓	✓✓	✓✓
Supported input file formats									
Windows executables (.exe)	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓
Windows libraries (.dll)	✓✓	✓✓	✓✓	✓✓	✓✓		✓	✓✓	✓
Microsoft Office files	✓✓	✓✓	✓✓	✓✓	✓✓				
Portable document format (.pdf)	✓✓	✓✓	✓✓	✓✓	✓✓				
Malicious URL scan	✓✓	✓✓	✓✓	✓✓	✓✓				
PHP files (.php)					✓✓				
Java file (.jar)	✓✓	✓✓	✓✓	✓✓	✓✓			✓✓	
Visual Basic scripts (.vbs)					✓✓				
Image files (.jpg, .png,...)		✓✓		✓✓					
Video files (.wmv, .flv,...)	✓	✓		✓✓	✓				
ZIP archive (.zip)	✓✓	✓✓		✓✓	✓✓			✓✓	

	Anubis	Cuckoo	CWSandbox	FireEye MAS	Joe Sandbox	ProcMon	APIMon	Generic disassembler	Generic debugger
Base data									
Virus definition/Malware name	✓✓	✓✓	✓✓	✓✓					
Behavior classification	✓				✓✓				
Packer information					✓			✓	✓
File information/File header		✓✓						✓✓	✓
Library imports/loads	✓✓	✓✓	✓✓		✓✓		✓	✓✓	✓
CPU instructions/assembly					✓✓			✓✓	✓✓
API calls	✓	✓✓	✓	✓	✓	✓	✓✓	✓✓	✓✓
System calls	✓	✓✓	✓	✓	✓	✓	✓✓	✓✓	✓✓
File system operations	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓	✓
Registry operations	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓	✓
Process/thread information	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓	✓
Network activity	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓	✓	✓
Report output									
PDF report	✓✓		✓✓						
HTML report	✓✓	✓✓	✓✓		✓✓				
XML report	✓✓		✓✓	✓✓	✓✓	✓✓			
TXT report	✓✓			✓✓			(✓)		✓✓
CSV report				✓✓		✓✓			
Native/Proprietary format						✓✓	✓✓	✓✓	✓✓
PCAP network dump	✓✓	✓✓	✓✓	✓✓	✓✓				
JSON report			✓✓	✓✓	✓✓				
Memory dumps		✓✓		✓✓	✓✓				
String dumps		✓✓			✓✓				
Screenshots		✓✓			✓✓				

Phương pháp nghiên cứu đề tài

- Để có 1 cái nhìn tổng thể về phương pháp trực quan hoá phân tích mã độc, bài báo tham khảo đến các thư viện điện tử (IEEE Xplore, ACM digital library, Google Scholar, and Academic Research Microsoft).
- Mỗi công cụ tìm kiếm có điểm mạnh và điểm yếu riêng.
- Đề tài được tổng hợp từ 25 bài báo phù hợp với chủ đề hệ thống trực quan hoá phân tích mã độc.

Hệ thống trực quan cho phân tích mã độc

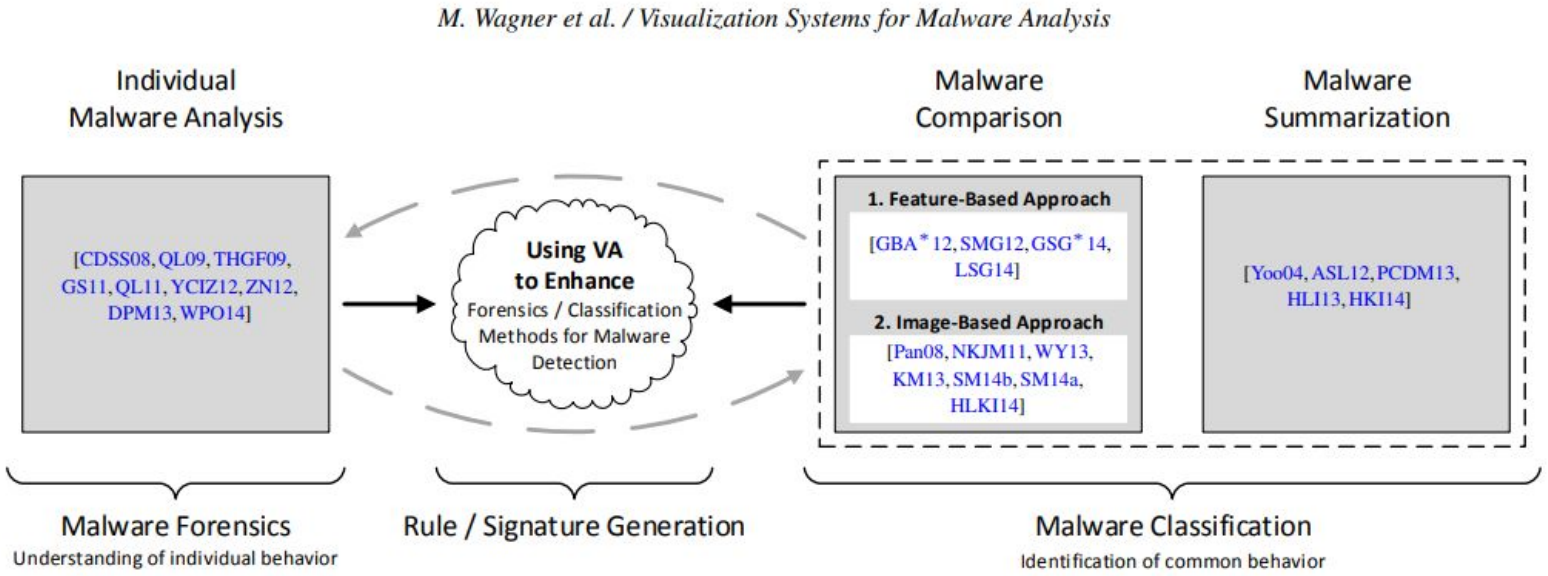


Figure 2: **Malware Visualization Taxonomy** – Categorization of malware visualization systems into three categories, namely (1) Individual Malware Analysis, (2) Malware Comparison, and (3) Malware Summarization. All systems have the ultimate goal to generate rules and signatures for fully-automated malware detection systems. While the first category tackles the problem of understanding the behavior of an individual malware sample for forensics, the latter two focus on the identification of common behavior for malware classification.

Hệ thống trực quan cho phân tích mã độc

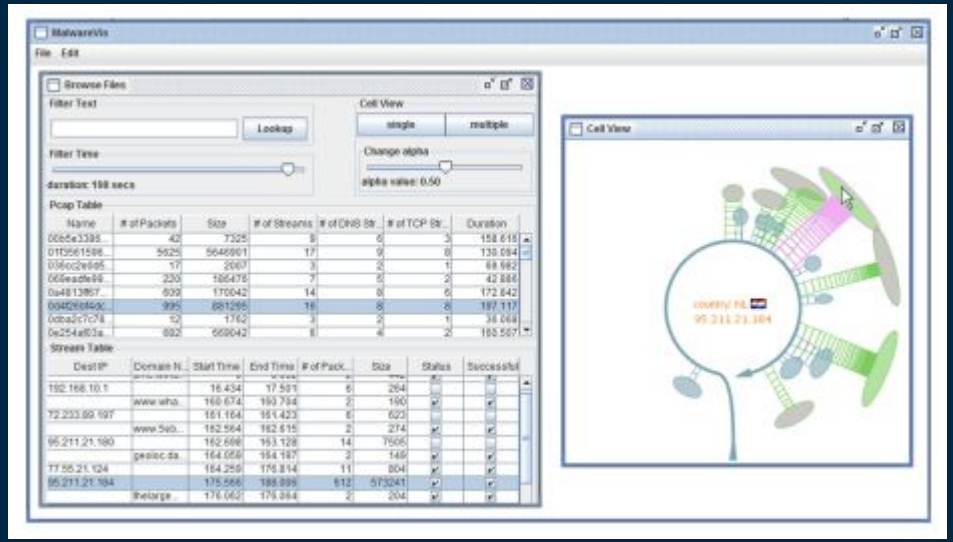
<table>
<tr><td colspan="2">Malware Forensics
Phân tích mã độc
Hiểu hành vi của phần mềm độc hại</td><td colspan="2">Malware Classification
Phân loại mã độc
Gán một phần mềm độc hại không xác định cho một nhóm các loại phần mềm độc hại đã biết</td></tr>
<tr><td colspan="2">Individual Malware Analysis</td><td>Malware Comparison</td><td>Malware Summarization</td></tr>
</table>

Hệ thống trực quan cho phân tích mã độc Individual Malware Analysis

Individual Malware Analysis:

Chứa các hệ thống trực quan hướng đến việc phân tích sâu rộng các mẫu phần mềm độc hại riêng lẻ.

Ví dụ: chỉ tập trung vào một loại hành vi cụ thể của phần mềm độc hại – hoạt động mạng. sau đó được hiển thị trực quan bằng biểu đồ giống như nét vẽ như có thể thấy trong hình. Điều này khám phá rất chi tiết hành vi đang thực hiện.



Hệ thống trực quan cho phân tích mã độc Individual Malware Analysis

Individual Malware Analysis:

Các công cụ khác có thể xem xét các tính năng khác nhau cùng một lúc, nhưng vẫn tập trung vào phân tích riêng lẻ của phần mềm độc hại mẫu.

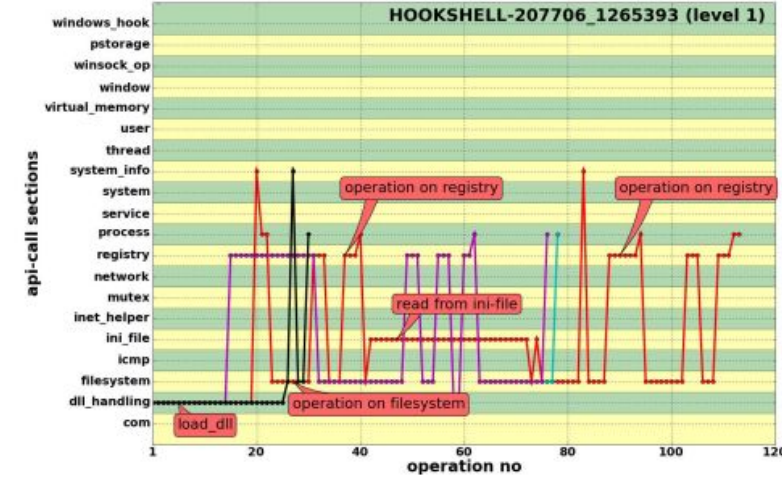


Figure 4: **Individual Malware Analysis** – Visual representation of system calls issued over time by an individual malware sample. *Image © 2009 IEEE. Reprinted, with permission, from [THGF09].*

Hệ thống trực quan cho phân tích mã độc
Visualization Support for Malware Comparison

Feature-Based Approach:

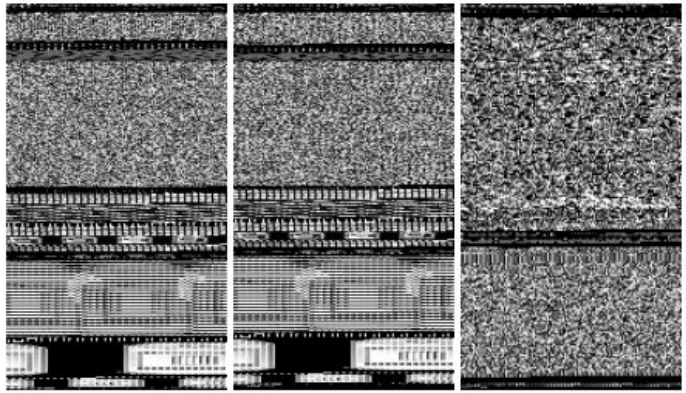
- Phương pháp tiếp cận dựa trên tính năng sử dụng các kỹ thuật phân tích trực quan để cho phép người dùng lọc, tìm kiếm, so sánh và khám phá nhiều loại thuộc tính được chiết xuất trong quá trình phân tích. Các hệ thống này cung cấp phương tiện để so sánh các mẫu phần mềm độc hại dựa trên sự tương đồng về tính năng của chúng.

Figure 5: **Comparison of Malware Characteristics** – Identifying similar malware samples to a focus sample by comparing them along different sets of characteristics (e.g., capabilities) [GSG*14]. *Image courtesy of Robert Gove.*

Hệ thống trực quan cho phân tích mã độc
Visualization Support for Malware Comparison

Image-Based Approach :

- Các phương pháp tiếp cận dựa trên hình ảnh có điểm chung là sử dụng trực quan để hiển thị hình ảnh cho từng mẫu phần mềm độc hại.



(a) FakeRean.D (b) FakeRean.E (c) Mebroot

Figure 6: **Comparison of Malware Images** – Visualizing malware executables as grayscale images is a common technique to visually identify similarities with low computation costs. *Image by the authors.*

Hệ thống trực quan cho phân tích mã độc
Visualization Support for Malware Summarization

Visualization Support for Malware Summarization :

- Thể hiện đa dạng hơn nhưng tất cả các công cụ liên quan đều chủ yếu cung cấp một số loại khả năng tóm tắt cho một số lượng lớn các mẫu phần mềm độc hại trong quá trình trực quan hóa.
- Một số sử dụng sử dụng bản đồ nhiệt để biểu diễn trực quan các hạt nhân được sử dụng cho trình phân loại Virus Code hỗ trợ để tóm tắt và cuối cùng phân loại các mẫu phần mềm độc hại.

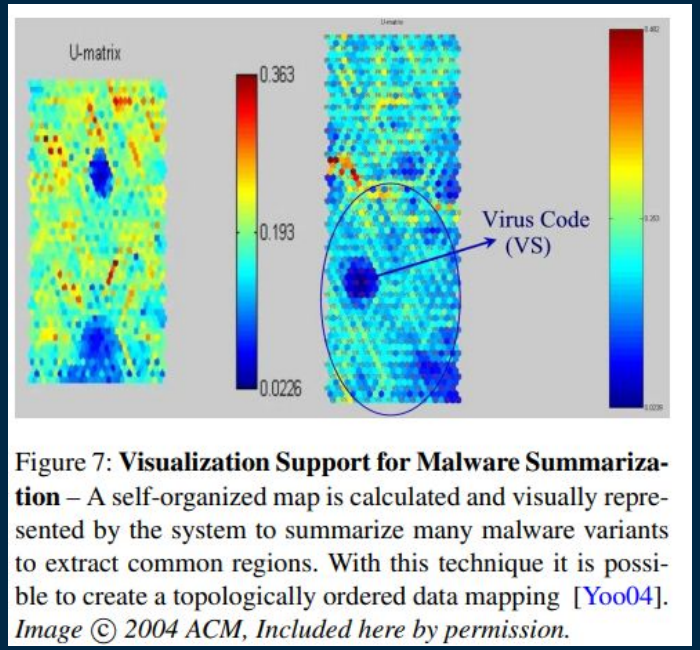
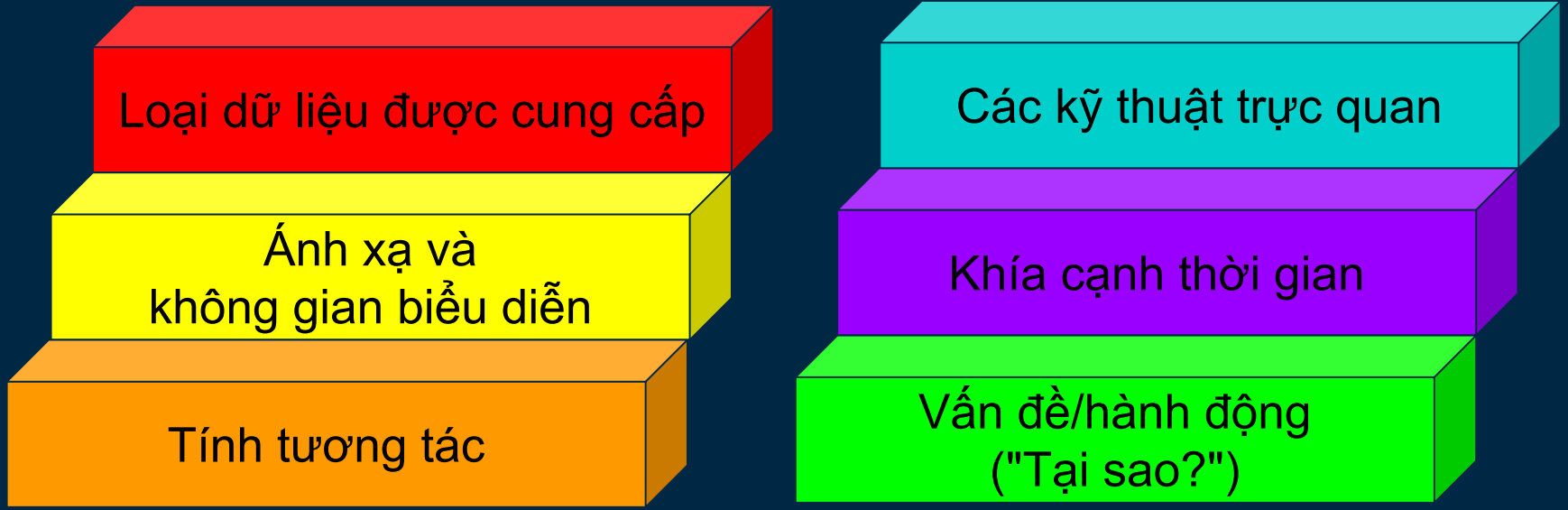


Figure 7: **Visualization Support for Malware Summarization** – A self-organized map is calculated and visually represented by the system to summarize many malware variants to extract common regions. With this technique it is possible to create a topologically ordered data mapping [Yoo04]. *Image © 2004 ACM, Included here by permission.*

Phân loại và so sánh

Để cung cấp một cái nhìn tổng quan, chúng tôi quyết định phân loại nhất quán tất cả các công cụ theo loại dữ liệu được cung cấp => tất cả các phân loại được sử dụng đều dựa trên các nguyên tắc phân loại được thiết lập trong cộng đồng trực quan hóa

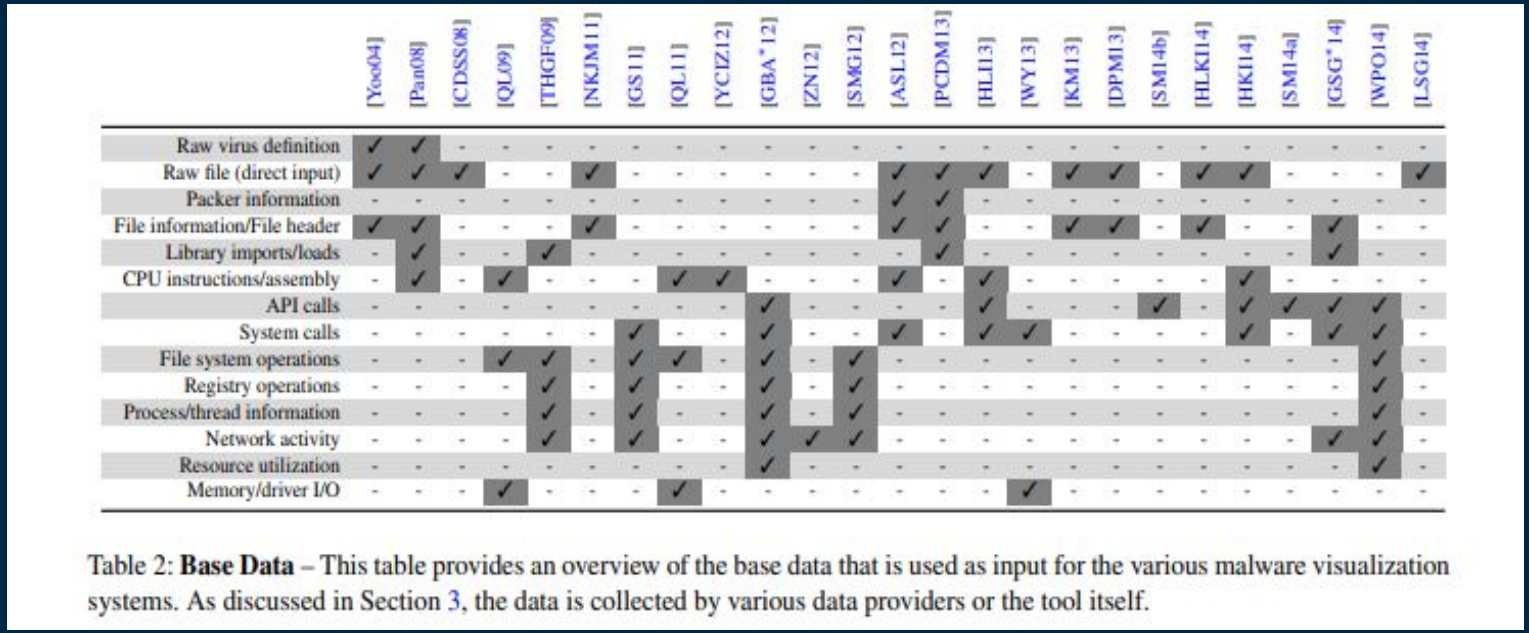


Data providers trong Visual Analytics

- Các yêu cầu đầu vào của mọi công cụ trực quan tương ứng với các định dạng đầu ra báo cáo được sử dụng bởi các nhà cung cấp dữ liệu.
- Dữ liệu cơ sở mô tả loại thông tin thực tế thu thập được từ các phân tích phần mềm độc hại – về cơ bản, dữ liệu này xác định loại hoạt động cụ thể của hệ thống được giám sát hoặc mã chương trình sẽ được trực quan hóa sau đó.

Data providers trong Visual Analytics

Bảng 2 hiển thị dữ liệu cơ sở được trực quan hóa bằng các giải pháp khác nhau của từng công cụ

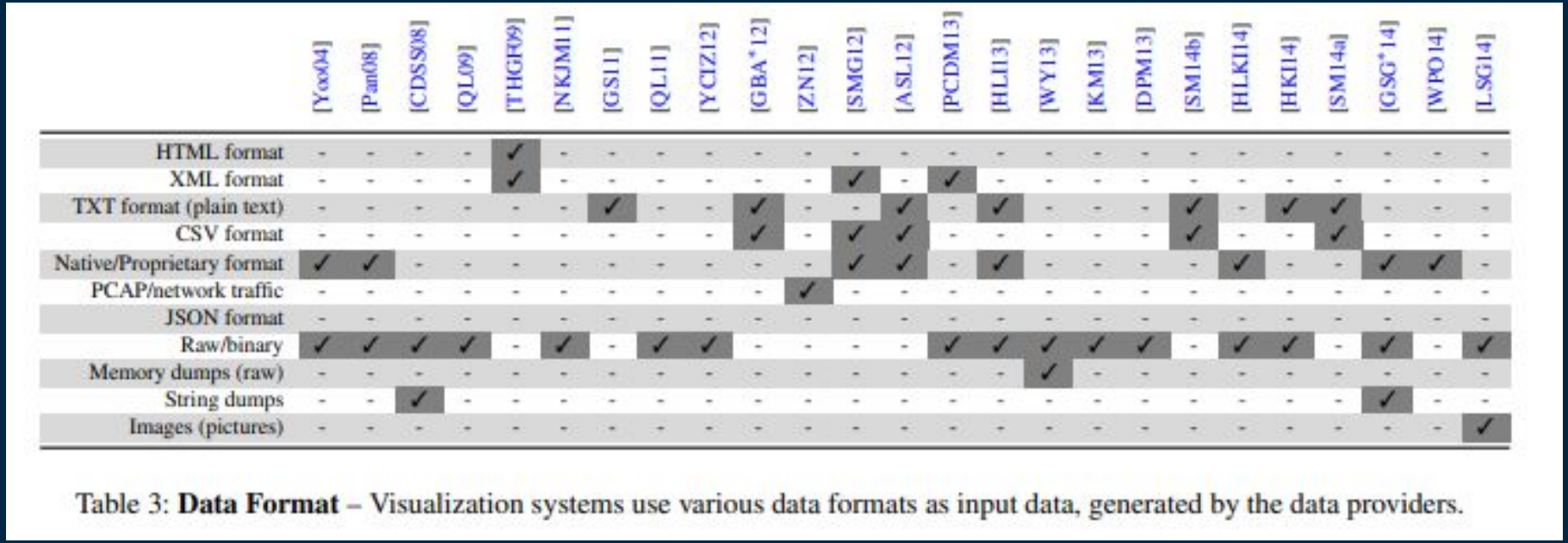


	[Yoo04]	[Pan08]	[CDSS08]	[QL09]	[THGF09]	[NKJM11]	[GS11]	[QL11]	[YCIZ12]	[GBA+12]	[ZN12]	[SMG12]	[ASL12]	[PCDM13]	[HLI13]	[WY13]	[KM13]	[DPM13]	[SM14b]	[HLKI14]	[HKI14]	[SM14a]	[GSG+14]	[WPO14]	[LSG14]
Raw virus definition	✓	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Raw file (direct input)	✓	✓	✓	-	-	✓	-	-	-	-	-	-	✓	✓	✓	-	✓	✓	-	✓	✓	-	-	-	✓
Packer information	-	-	-	-	-	-	-	-	-	-	-	-	✓	✓	-	-	-	-	-	-	-	-	-	-	-
File information/File header	✓	✓	-	-	-	✓	-	-	-	-	-	-	✓	✓	-	-	✓	✓	-	✓	-	-	✓	-	-
Library imports/loads	-	✓	-	-	✓	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	✓	-	-
CPU instructions/assembly	-	✓	-	✓	-	-	-	✓	✓	-	-	-	✓	-	✓	-	-	-	-	-	✓	-	-	-	-
API calls	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	✓	-	-	-	✓	-	✓	✓	✓	✓	-
System calls	-	-	-	-	-	-	✓	-	-	✓	-	-	✓	-	✓	✓	-	-	-	-	✓	-	✓	✓	-
File system operations	-	-	-	✓	✓	-	✓	✓	-	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-	✓	-
Registry operations	-	-	-	-	✓	-	✓	-	-	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-	✓	-
Process/thread information	-	-	-	-	✓	-	✓	-	-	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-	✓	-
Network activity	-	-	-	-	✓	-	✓	-	-	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	✓	✓	-
Resource utilization	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Memory/driver I/O	-	-	-	✓	-	-	-	✓	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-

Table 2: **Base Data** – This table provides an overview of the base data that is used as input for the various malware visualization systems. As discussed in Section 3, the data is collected by various data providers or the tool itself.

Data providers trong Visual Analytics

Bảng 3 liệt kê các định dạng xử lý dữ liệu tương ứng (định dạng đầu ra của nhà cung cấp) của từng công cụ



	[Yoo04]	[Pan08]	[CDSS08]	[QL09]	[THGF09]	[NKJM11]	[GS11]	[QL11]	[YCIZ12]	[GBA*12]	[ZN12]	[SMG12]	[ASL12]	[PCDM13]	[HLI13]	[WY13]	[KM13]	[DPM13]	[SM14b]	[HLKI14]	[HKI14]	[SM14a]	[GSG*14]	[WPO14]	[LSG14]
HTML format	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XML format	-	-	-	-	✓	-	-	-	-	-	-	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-
TXT format (plain text)	-	-	-	-	-	-	✓	-	-	✓	-	-	✓	-	✓	-	-	-	✓	-	✓	✓	-	-	-
CSV format	-	-	-	-	-	-	-	-	-	✓	-	✓	✓	-	-	-	-	-	✓	-	-	✓	-	-	-
Native/Proprietary format	✓	✓	-	-	-	-	-	-	-	-	-	✓	✓	-	✓	-	-	-	-	✓	-	-	✓	✓	-
PCAP/network traffic	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-
JSON format	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Raw/binary	✓	✓	✓	✓	-	✓	-	✓	✓	-	-	-	-	✓	✓	✓	✓	✓	-	✓	✓	-	✓	-	✓
Memory dumps (raw)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-
String dumps	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-
Images (pictures)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓

Table 3: **Data Format** – Visualization systems use various data formats as input data, generated by the data providers.

Data providers trong Visual Analytics

Sử dụng thông tin này, một nhà phân tích có thể chỉ cần chọn loại và định dạng mong muốn và chọn một dữ liệu phù hợp

Các bảng có thể được sử dụng làm tài liệu tham khảo cho các khả năng của công cụ và cách tiếp cận chung

Kỹ thuật trực quan hóa

Để phân loại các kỹ thuật trực quan hóa khác nhau, chúng tôi đã sử dụng "Trực quan hóa thông tin và dữ liệu khai thác"

Hiển thị 2D/3D tiêu chuẩn

Bao gồm các kỹ thuật trực quan hóa như biểu đồ x-y (x-y-z)

Ví dụ: biểu đồ phân tán, biểu đồ thanh và biểu đồ đường

Hiển thị biến đổi hình học

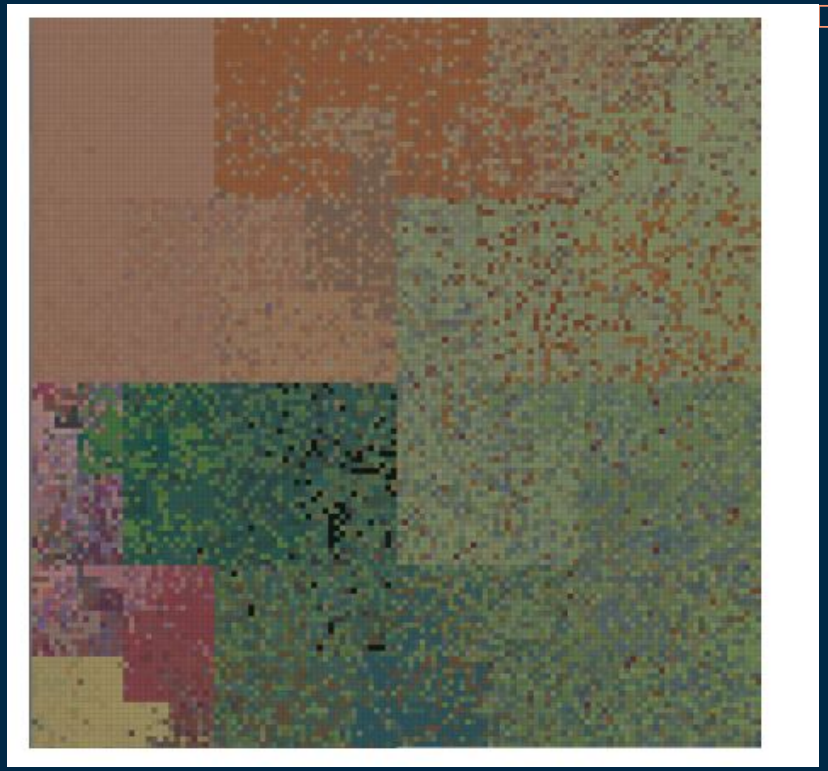
Nhằm mục đích trực quan hóa các phép biến đổi thú vị của bộ dữ liệu đa chiều
Ví dụ: ma trận biểu đồ phân tán, sơ đồ liên kết nút, tọa độ song song, stardinates

Hiển thị mang tính biểu tượng

Các thuộc tính của dữ liệu đa chiều được ánh xạ vào các tính năng của một biểu tượng để biểu diễn
Ví dụ: mặt chernoff, biểu tượng kim, biểu tượng ngôi sao, biểu tượng hình que, biểu tượng màu và ô xếp thanh

Hiển thị mật độ điểm ảnh

Mỗi điểm dữ liệu được ánh xạ tới một pixel màu và chúng được nhóm thành các vùng liền kề đại diện cho kích thước dữ liệu cá nhân.
Ví dụ: trực quan hóa ma trận.
Mỗi hình dạng pixel đại diện cho một mẫu phần mềm độc hại. Các mẫu phần mềm độc hại tương tự được sắp xếp cạnh nhau và được gán các giá trị màu tương tự để trực quan hóa các điểm chung



Hiển thị xếp chồng

- Biểu diễn cho dữ liệu phân cấp
  Ví dụ: xếp chồng theo thứ bậc, sơ đồ cây, vùng lân cận sơ đồ cây còn được gọi là Nmaps
- Biểu diễn cho bố cục phân cấp cho dữ liệu đa chiều
  Ví dụ: chiều xếp chồng

Kỹ thuật trực quan hóa

Hiển thị xếp chồng và hiển thị mang tính biểu tượng không thường được sử dụng trong lĩnh vực này

	[Yoo04]	[Pan08]	[CDSS08]	[QL09]	[THGF09]	[NKJM11]	[GS11]	[QL11]	[YCIZ12]	[GBA*12]	[ZN12]	[SMG12]	[ASL12]	[PCDM13]	[HL13]	[WY13]	[KM13]	[DPM13]	[SM14b]	[HLK14]	[HK14]	[SM14a]	[GSG*14]	[WPO14]	[LSG14]
Standard 2D Display	-	-	-	✓	✓	-	✓	✓	-	✓	-	-	-	-	-	✓	-	✓	✓	✓	-	✓	✓	✓	✓
Standard 3D Display	-	✓	-	✓	-	-	-	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Geometrically-transformed Display	✓	✓	-	✓	-	-	✓	✓	✓	-	✓	-	-	✓	-	-	-	-	-	-	-	-	-	✓	✓
Iconic Display	-	-	-	-	-	-	-	-	-	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	✓	-	-
Dense Pixel Display	✓	-	✓	-	-	✓	-	-	-	-	-	✓	✓	-	✓	-	✓	✓	✓	✓	✓	✓	✓	-	-
Stacked Display	-	-	-	-	✓	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	✓	-	-

Table 4: **Visualization Techniques** – A general overview of the most frequently used types of visualization techniques based on the taxonomy of Keim [Kei02].

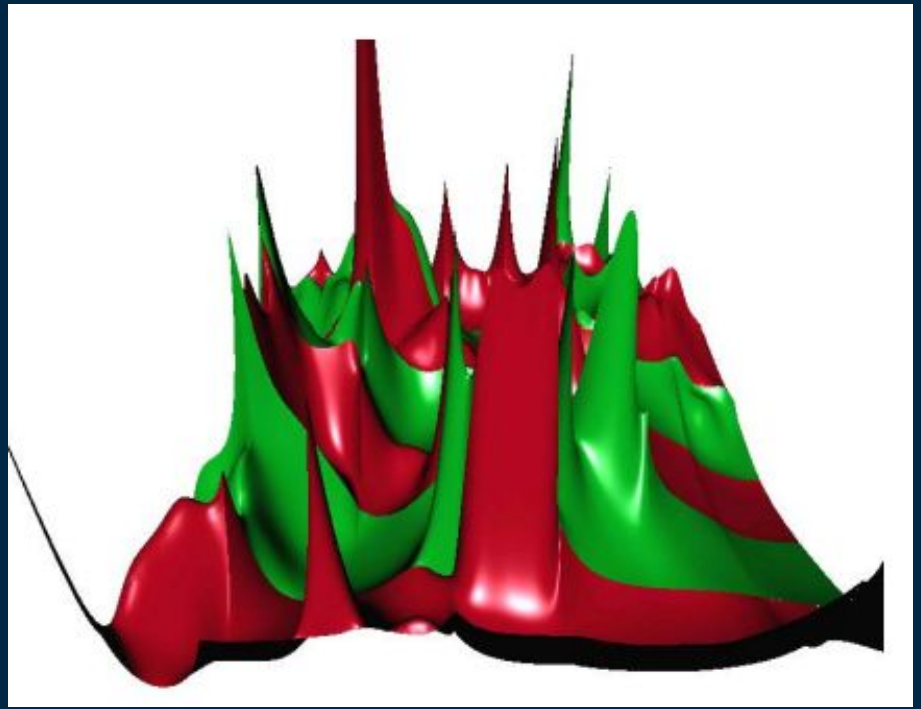
Không gian biểu diễn & Ánh xạ theo thời gian

Trong khi có rất nhiều cách biểu diễn trực quan để phân tích dữ liệu phần mềm độc hại, chúng ta có thể phân loại chúng theo hai sự chia đôi cơ bản: không gian biểu diễn và thời gian vật lý

- Nói chung, không gian biểu diễn của một hình ảnh trực quan có thể là 2D hoặc 3D. Chưa có sự đồng nhất trong cộng đồng
- Ánh xạ theo thời gian: thêm chiều của thời gian vào được sử dụng như một phần của trình chiếu hoặc hoạt ảnh. Cách tiếp cận động phù hợp với ánh xạ theo thời gian có thể được chiếu vào các biến trực quan khác nhau. Ngoài ra, phát trực tuyến nguồn dữ liệu hoặc phân tích tự động được cập nhật dần dần có thể dẫn đến ánh xạ động
    - Tĩnh: Dữ liệu được ánh xạ để hiển thị không gian và các biến trực quan khác không thay đổi theo thời gian. Ánh xạ tĩnh không loại trừ tính tương tác
    - Động: Dữ liệu hiển thị thay đổi theo thời gian vật lý mà không cần sự tương tác của người dùng

Không gian biểu diễn & Ánh xạ theo thời gian

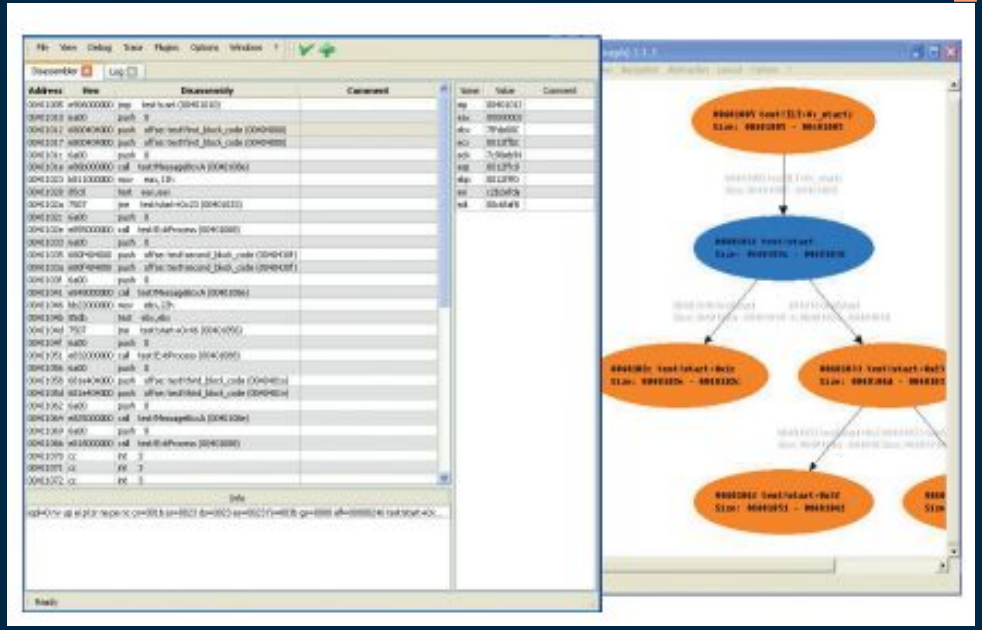
Panas đề xuất hình ảnh 3D để tạo chữ ký trực quan cụ thể giúp xác định sự bất thường trong một họ phần mềm độc hại



Không gian biểu diễn & Ánh xạ theo thời gian

Ánh xạ động

Trình gỡ lỗi trực quan để phân tích phần mềm độc hại bằng cách sử dụng sơ đồ liên kết nút có khả năng phát lại để hiển thị luồng thực thi của mẫu phần mềm độc hại trên thời gian



Không gian biểu diễn & Ánh xạ theo thời gian

Hầu hết các công cụ tập trung vào ánh xạ tĩnh. Lý do cho điều này có thể là nhiều công cụ sử dụng dữ liệu cơ sở không xem xét trình tự thời gian. Một lý do khác có thể là một đại diện năng động làm cho nó khó khăn hơn cho các nhà phân tích

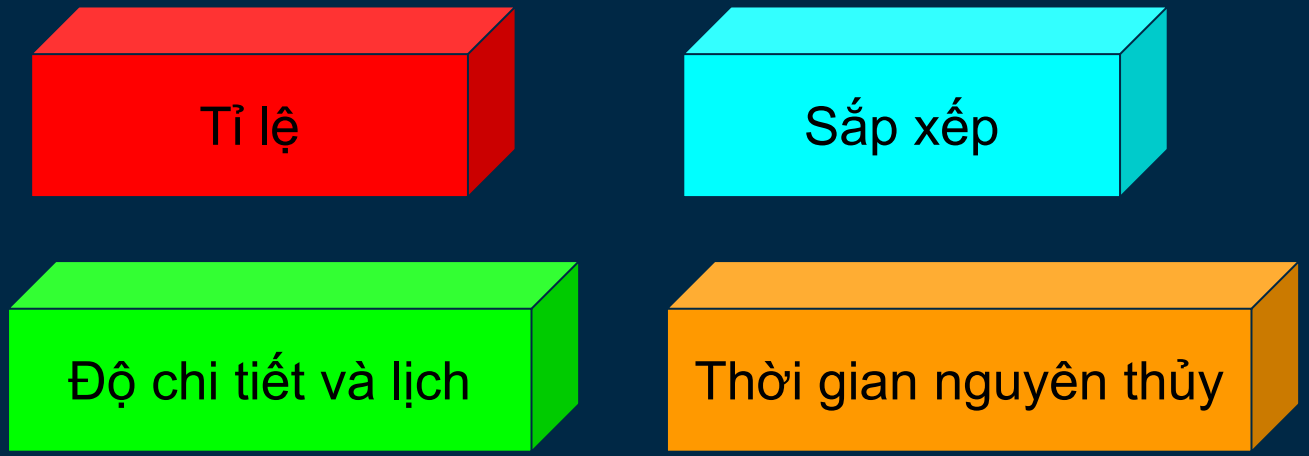
	[Yoo04]	[Pan08]	[CDSS08]	[QL09]	[THGF09]	[NKJM11]	[GSI1]	[QL11]	[YCIZ12]	[GBA*12]	[ZN12]	[SMG12]	[ASL12]	[PCDM13]	[HLI13]	[WY13]	[KM13]	[DPM13]	[SM14b]	[HLKI14]	[HKI14]	[SM14a]	[GSG*14]	[WPO14]	[LSG14]
Mapping ► Static	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mapping ► Dynamic	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Dimensionality ► 2D	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Dimensionality ► 3D	-	✓	-	✓	-	-	-	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 5: **Mapping & Representation Space** – An overview of used representation spaces in visualization systems. Almost all tools focus on static mappings.

Khía cạnh tạm thời

Dữ liệu định hướng theo thời gian đóng một vai trò quan trọng trong phân tích phần mềm độc hại. Thời gian có thể được mô hình hóa theo những cách khác nhau tùy thuộc vào phân tích mục tiêu

Ví dụ: thứ tự thực hiện của hệ thống hoặc lệnh gọi API có liên quan để xác định các mẫu hành vi nhất định (chẳng hạn như tạo và xóa tập tin sau đó)



Tỉ lệ

Thứ tự	Miền thời gian thứ tự chỉ thể hiện quan hệ giữa các khía cạnh thời gian Ví dụ: trước, sau
Rời rạc	Miền thời gian rời rạc biểu diễn bằng ánh xạ các giá trị thời gian tới một tập hợp các số nguyên
Liên tục	Miền thời gian liên tục, ánh xạ các giá trị thời gian tới một tập hợp các số thực.

Sắp xếp

Tuyến tính

Mỗi phần tử có một phần tử tiền nhiệm và một phần tử kế tiếp
Ví dụ: từ quá khứ đến tương lai

Chu kỳ

Nếu dữ liệu bao gồm một tập hợp thời gian lặp lại các giá trị, chúng ta đang nói về sự sắp xếp theo chu kỳ
Ví dụ: 4 mùa trong năm

Độ chi tiết và lịch

Không

Nếu các giá trị thời gian không được ánh xạ theo bất kỳ loại nào về mức độ chi tiết (ví dụ: năm, quý, tháng, v.v.) thì hệ thống sẽ không có độ chi tiết

Một

Độ chi tiết đơn mô tả ánh xạ các giá trị thời gian chỉ cho 1 loại đơn vị chi tiết (ví dụ: năm hoặc tháng)

Nhiều

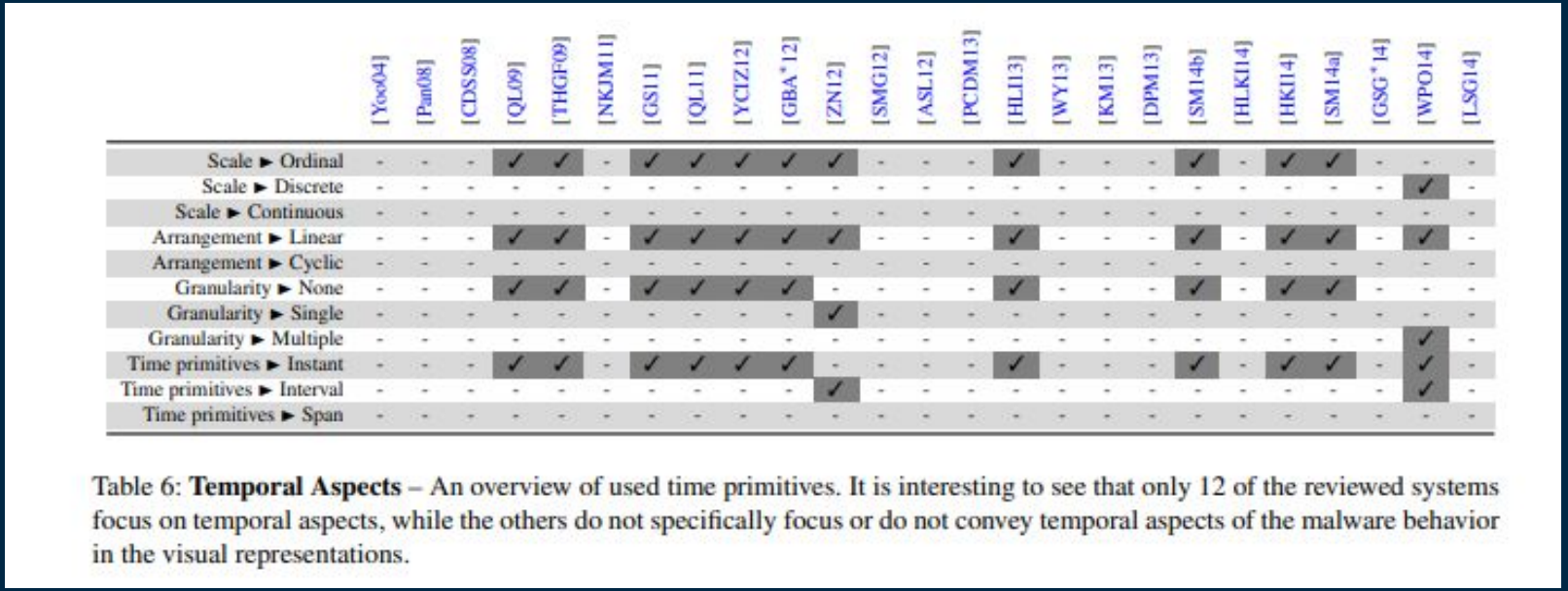
Với việc ánh xạ tới nhiều mức độ chi tiết, nó có thể chia các giá trị thời gian thành năm, quý, tháng. Một ánh xạ như vậy được gọi là lịch

Thời gian nguyên thủy

Tức thời	Một thời điểm duy nhất được gọi là tức thời Ví dụ: ngày 15 tháng 4 năm 2023
Khoảng	Khoảng là một phần thời gian giữa hai thời điểm Ví dụ: bắt đầu và kết thúc.
Quãng	Quãng là một giá trị nguyên thủy chưa được neo, đại diện cho khoảng thời gian được chỉ định (ví dụ: 4 giờ)

Khía cạnh tạm thời

Điều thú vị là chỉ có 12 trong số 25 công cụ sử dụng các khía cạnh thời gian. Tất cả 12 công cụ này đều có tuyến tính sắp xếp



	[Yoo04]	[Pan08]	[CDSS08]	[QL09]	[THGF09]	[NKJM11]	[GS11]	[QL11]	[YCIZ12]	[GBA+12]	[ZN12]	[SMGi12]	[ASL12]	[PCDM13]	[HLI13]	[WY13]	[KM13]	[DPM13]	[SM14b]	[HLKI14]	[HKI14]	[SM14a]	[GSG+14]	[WPO14]	[LSG14]
Scale ► Ordinal	-	-	-	✓	✓	-	✓	✓	✓	✓	✓	-	-	-	✓	-	-	-	✓	-	✓	✓	-	-	-
Scale ► Discrete	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Scale ► Continuous	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Arrangement ► Linear	-	-	-	✓	✓	-	✓	✓	✓	✓	✓	-	-	-	✓	-	-	-	✓	-	✓	✓	-	✓	-
Arrangement ► Cyclic	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Granularity ► None	-	-	-	✓	✓	-	✓	✓	✓	✓	-	-	-	-	✓	-	-	-	✓	-	✓	✓	-	-	-
Granularity ► Single	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Granularity ► Multiple	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Time primitives ► Instant	-	-	-	✓	✓	-	✓	✓	✓	✓	-	-	-	-	✓	-	-	-	✓	-	✓	✓	-	✓	-
Time primitives ► Interval	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Time primitives ► Span	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 6: **Temporal Aspects** – An overview of used time primitives. It is interesting to see that only 12 of the reviewed systems focus on temporal aspects, while the others do not specifically focus or do not convey temporal aspects of the malware behavior in the visual representations.

Tương tác

Để phân loại các khả năng tương tác của hệ thống, chúng tôi đã khám phá xem liệu các kỹ thuật tương tác như thu phóng, lọc, quét, chi tiết theo yêu cầu hoặc chải/liên kết có sẵn. Ngoài ra, chúng tôi đã cố gắng tìm hiểu xem có thể thay đổi động giữa các dữ liệu biểu diễn trực quan khác nhau

Hầu hết, các công cụ chỉ được gọi là tương tác nói chung mà không đưa ra lời giải thích chi tiết hơn. Do đó, chúng tôi quyết định giới hạn việc phân loại liệu hệ thống có hỗ trợ bất kỳ loại tương tác nào mà không đi vào chi tiết

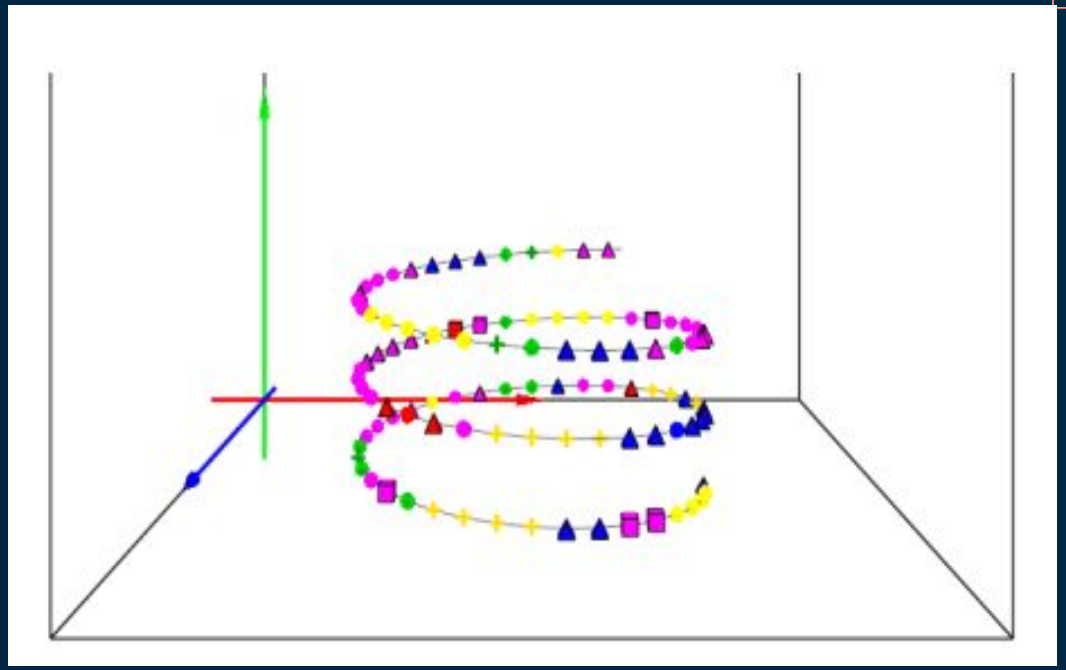
	[Yoo04]	[Pan08]	[CDSS08]	[QL09]	[THGF09]	[NKJM11]	[GS11]	[QL11]	[YCIZI12]	[GBA*12]	[ZN12]	[SMG12]	[ASL12]	[PCDM13]	[HLI13]	[WYI13]	[KM13]	[DPM13]	[SM14b]	[HLKI14]	[HKI14]	[SM14a]	[GSG*14]	[WPO14]	[LSG14]
Interaction	-	-	✓	✓	-	-	✓	✓	✓	✓	✓	✓	-	-	-	✓	-	✓	-	-	-	-	✓	✓	✓
No Interaction	✓	✓	-	-	✓	✓	-	-	-	-	-	-	✓	✓	✓	-	✓	-	✓	✓	✓	✓	-	-	-

Table 7: **Interactivity** – An overview of the level of interactivity in the visualizations used by the tools.

Tương tác

Tương tác

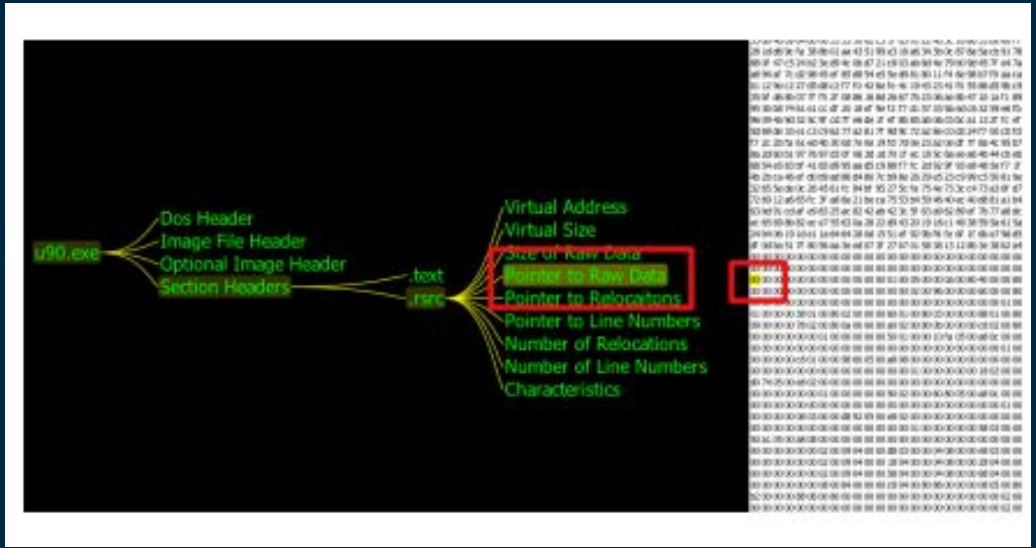
Công cụ này đại diện cho một chuỗi các hành động độc hại được sắp xếp theo thứ tự bằng cách sử dụng biểu tượng mang tính biểu tượng ở dạng xoắn ốc. Đối với việc khám phá dữ liệu, có thể phóng to và thu nhỏ, xoay, nghiêng, chọn các lát hành vi khác nhau, xem nhật ký văn bản và so sánh nó với dữ liệu hành vi có sẵn khác



Tương tác

Tương tác

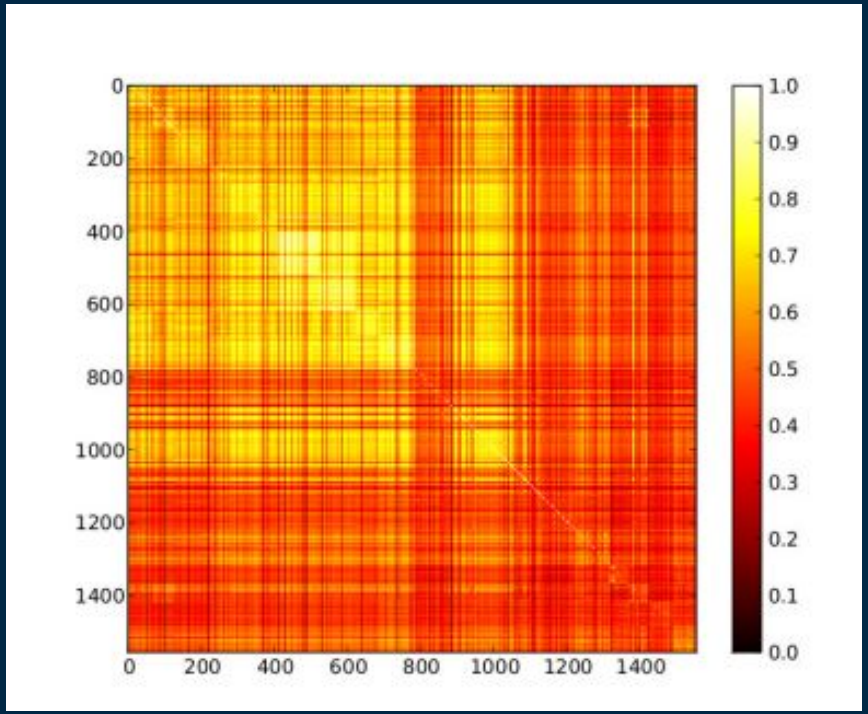
Liên kết trình soạn thảo hex thích hợp đối với biểu diễn dạng cây tương tác giúp nâng cao khả năng điều hướng và hiểu dữ liệu tiêu đề của phần mềm độc hại



Tương tác

Không tương tác

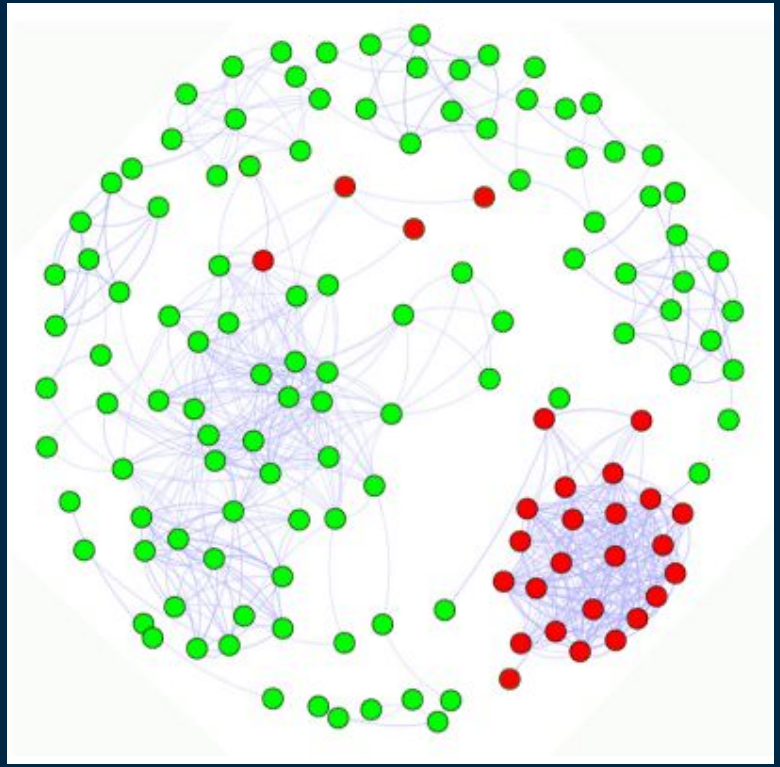
Trực quan hóa mật độ pixel cho thấy sự tương đồng giữa 780 mẫu phần mềm độc hại (trên cùng bên trái) và 776 mẫu lành tính (dưới cùng phải)



Tương tác

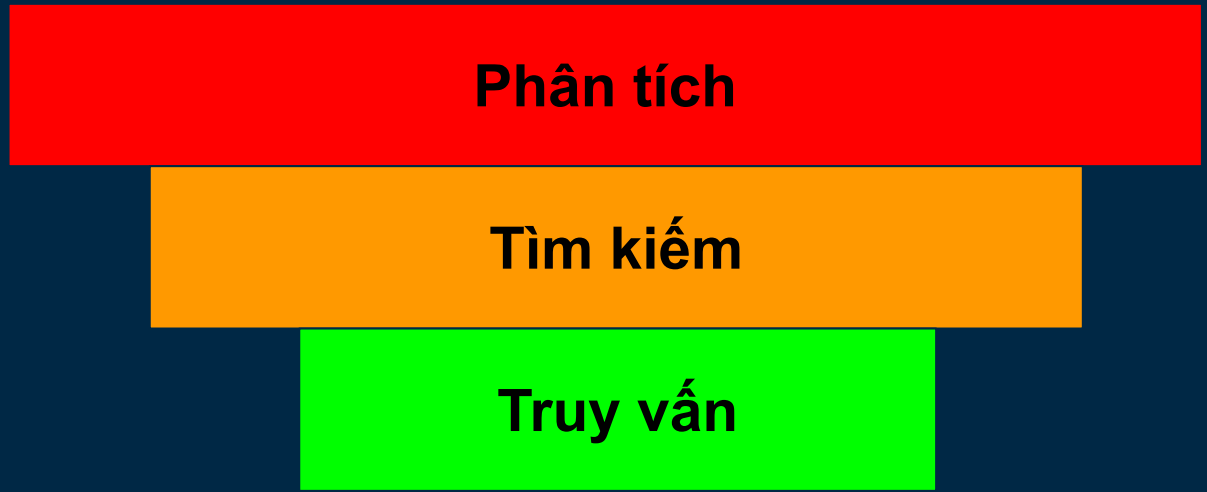
Không tương tác

Các mẫu phần mềm độc hại được sắp xếp trong sơ đồ liên kết nút với trọng số cạnh dựa trên cách nhiều hệ thống chống vi-rút gắn nhãn chúng trong cùng một danh mục. Các nút màu đỏ thuộc họ phần mềm độc hại đã biết. Vị trí của các nút được tính toán tự động và người dùng không thể tương tác với chúng

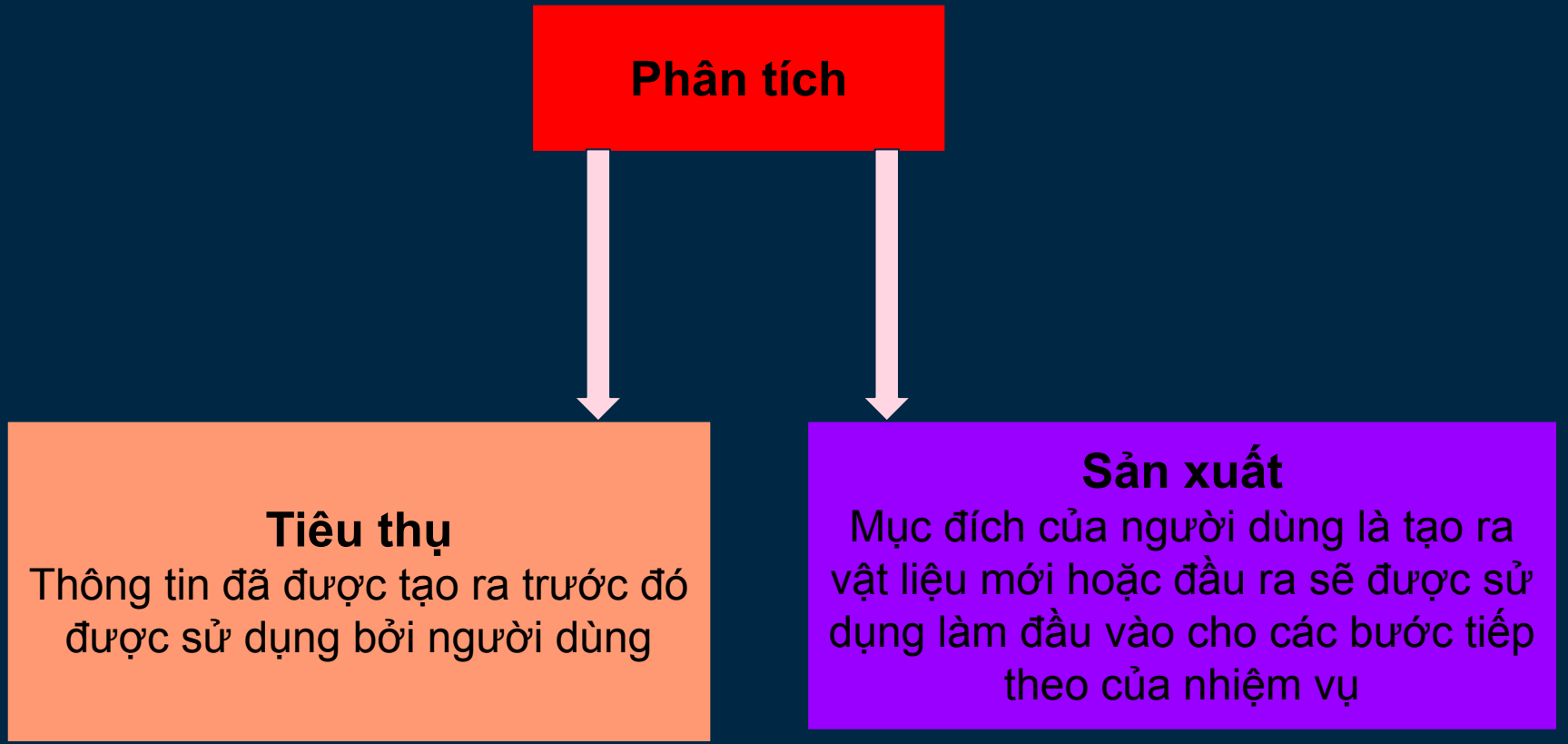


Vấn đề/Hành động (“Tại sao?”)

Munzner đã định nghĩa ba cấp độ hành động để mô tả mục tiêu của người dùng

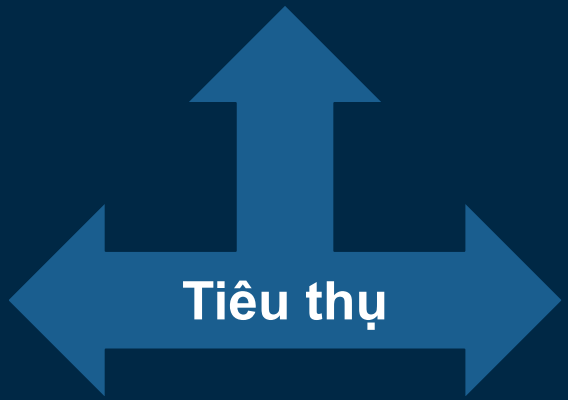


Phân tích


Phân tích
Tiêu thụ
Thông tin đã được tạo ra trước đó được sử dụng bởi người dùng
Sản xuất
Mục đích của người dùng là tạo ra vật liệu mới hoặc đầu ra sẽ được sử dụng làm đầu vào cho các bước tiếp theo của nhiệm vụ

Phân tích

Khám phá
Mô tả việc tạo ra và xác minh các giả thuyết bằng cách sử dụng khám phá trực quan cũng như thu được kiến thức mới về dữ liệu được trình bày



Thưởng thức
Đề cập đến việc sử dụng thông thường hoặc thưởng thức trực quan thuần túy mà không có mục tiêu hoặc nhu cầu cụ thể

Hiện diện
Truyền đạt thông tin sắc bén hoặc kể chuyện dựa trên dữ liệu trực quan

Phân tích

Chú thích
Chú thích bằng văn bản hoặc đồ họa cho trực quan hóa. Những chú thích này thường được thực hiện bằng tay

Giữ lại
Chụp hoặc lưu các yếu tố trực quan đã chọn.

Tạo ra
Tạo ra các phần tử dữ liệu mới dựa trên các phần tử dữ liệu hiện có. Như vậy, có thể rút ra thuộc tính mới từ thông tin hiện có hoặc để chuyển đổi một
kiểu dữ liệu sang kiểu dữ liệu khác

Tìm kiếm

Munzner chia khu vực phân tích thành bốn loại theo đó mã định danh và vị trí của các phần tử đích có được thể biết hoặc không. Trong cuộc khảo sát này, chúng tôi đã sử dụng mẫu phần mềm độc hại làm mục tiêu trong khi đặc điểm phần mềm độc hại đã được sử dụng làm vị trí.

Tra cứu
Mục tiêu và vị trí được biết đến

Định vị
Không rõ vị trí nhưng mục tiêu đã được biết đến

Duyệt
Biết vị trí nhưng không rõ mục tiêu

Khám phá
Mục tiêu và vị trí không rõ

Truy vấn

Khi mục tiêu cho việc tìm kiếm được xác định, bổ sung thông tin sẽ được truy vấn như một phần của mục tiêu cấp thấp nhất này. Munzner đặt tên cho ba loại truy vấn khác nhau

Xác định
Đề cập đến một mục tiêu duy nhất =>truy vấn xác định trả về các đặc điểm của mục tiêu

So sánh
Hoạt động truy vấn so sánh sử dụng các kỹ thuật phức tạp hơn

Tóm tắt
Cái nhìn toàn diện về mọi thứ cho tất cả các mục tiêu

Thảo luận và những thách thức trong tương lai

Cầu nối giữa các hạng mục

Trong Phần 5, chúng tôi đã xác định ba loại hệ thống trực quan hóa phần mềm độc hại giải quyết các vấn đề phụ khác nhau của điều tra phần mềm độc hại và phân loại ở cấp độ của các mẫu phần mềm độc hại riêng lẻ, so sánh các mẫu phần mềm độc hại và các tính năng phổ biến được tóm tắt từ các họ phần mềm độc hại

Vì có một mục tiêu chung là tạo ra các quy tắc hoặc chữ ký, nên có thể giả định rằng người dùng mục tiêu tiềm năng của cả ba loại hệ thống trực quan trùng lặp. Do đó, các hệ thống trực quan hóa phần mềm độc hại trong tương lai nên kiểm tra các thiết kế toàn diện

Thảo luận và những thách thức trong tương lai

Tích hợp các nguồn dữ liệu khác nhau

Phân tích phần mềm độc hại là dựa trên một loạt các dữ liệu cơ sở được thu thập bởi dữ liệu các nhà cung cấp theo các chế độ phân tích khác nhau (Phần 3). Vì phần mềm độc hại trở nên tinh vi hơn trong việc phát hiện và tránh phân tích, ngày càng có nhiều nhu cầu kết hợp các dữ liệu nhà cung cấp khác nhau

Ví dụ để kết hợp phân tích tĩnh và động

Thảo luận và những thách thức trong tương lai

Đặc tính hóa vấn đề và trừu tượng hóa cho phù hợp trực quan hóa

Nhiều hệ thống chỉ sử dụng trực quan hóa một cách hời hợt và dựa vào các tiêu chuẩn hiển thị. Tuy nhiên, những các phương pháp biểu diễn trực quan bị hạn chế về khả năng mở rộng trực quan của chúng. Tuy nhiên, có một tiềm năng cho các phương pháp đại diện mới hoặc được điều chỉnh để đáp ứng các nhu cầu đặc biệt của phân tích phần mềm độc hại. Nghiên cứu trực quan hóa theo hướng vấn đề phát triển mạnh từ sự hợp tác liên ngành với các chuyên gia trong lĩnh vực nhưng cần bắt đầu từ một đặc tính và trừu tượng hóa vấn đề vững chắc làm cơ sở cho thiết kế và đánh giá

Thảo luận và những thách thức trong tương lai

Thu hút kiến thức chuyên môn thông qua tương tác

Sự phát triển đa dạng của các dòng phần mềm độc hại, các nhà phân tích phần mềm độc hại cần liên tục điều chỉnh các cài đặt của hệ thống trực quan hóa của họ. Tính tương tác là điểm mạnh chính của các hệ thống trực quan hóa, cho phép các chuyên gia các phản hồi ngay lập tức.

Thảo luận và những thách thức trong tương lai

Đan xen các phương pháp phân tích với trực quan

Hiện nay hầu hết các hệ thống đều xây dựng các ẩn dụ trực quan của chúng trực tiếp trên đầu ra của các nhà cung cấp dữ liệu. Các phương pháp phân tích phải được xem xét cùng với các phương pháp biểu diễn trực quan để các giải pháp trực quan hóa phù hợp với vấn đề và có thể mở rộng

Kết luận

Trong cuộc khảo sát này, chúng tôi đã trình bày các nhà cung cấp dữ liệu hiện đang được sử dụng cũng như đánh giá có hệ thống các hệ thống trực quan hóa cho phân tích phần mềm độc hại

Trong bước đầu tiên, chúng tôi phân loại các nhà cung cấp dữ liệu liên quan đến phương pháp phân tích, môi trường cũng như định dạng dữ liệu đầu vào và đầu ra của họ.

Mỗi hệ thống phân tích tùy thuộc vào cách tiếp cận chung để xử lý và trực quan hóa, các khía cạnh thời gian nhất định, khả năng tương tác của chúng...

Chúng tôi cũng phân loại các hệ thống này theo các tệp và định dạng đầu vào của chúng, các kỹ thuật trực quan hóa được sử dụng.

Nhiều hệ thống được khảo sát thu thập dữ liệu phân tích nội bộ và phân tích dựa trên mẫu của mã nhị phân. Những hệ thống khác sử dụng các nhà cung cấp dữ liệu bên ngoài để truy xuất tính chất cụ thể

Kết luận

Về mặt kỹ thuật trực quan

- Hiển thị xếp chồng lên nhau và hiển thị mang tính biểu tượng không thường được sử dụng trong miền phần mềm độc hại
- Hầu hết các công cụ sử dụng hiển thị 2D tĩnh để hỗ trợ nhà phân tích
- Về không gian biểu diễn được sử dụng và ánh xạ theo thời gian, hầu hết các hệ thống đều sử dụng ánh xạ tĩnh
- Chỉ có 12 trong số 25 hệ thống phân tích xem xét thời gian
- Chỉ có 13 hệ thống được khảo sát hỗ trợ tương tác
- Trong số các hành động người dùng có sẵn, khám phá, trình bày và so sánh hoạt động là phổ biến nhất
- Việc xác định các mẫu phần mềm độc hại cụ thể thường không phải là ưu tiên

Cảm ơn thầy và các bạn.