Kompressionsclustern von Malware

Jedes Jahr werden immer mehr Malware-Angriffe registriert. Der Schaden durch Cyberkriminalität weltweit wird mittlerweile auf über 200 Milliarden Euro geschätzt. Ransomware allein ist für 20 % der Schadenssumme verantwortlich, wobei zuletzt eine Ransomware namens „StalinLocker“ für Aufsehen sorgte. Diese Ransomware fordert kein Lösegeld, sondern das Wissen über das Gründungsdatum der Sowjetunion.

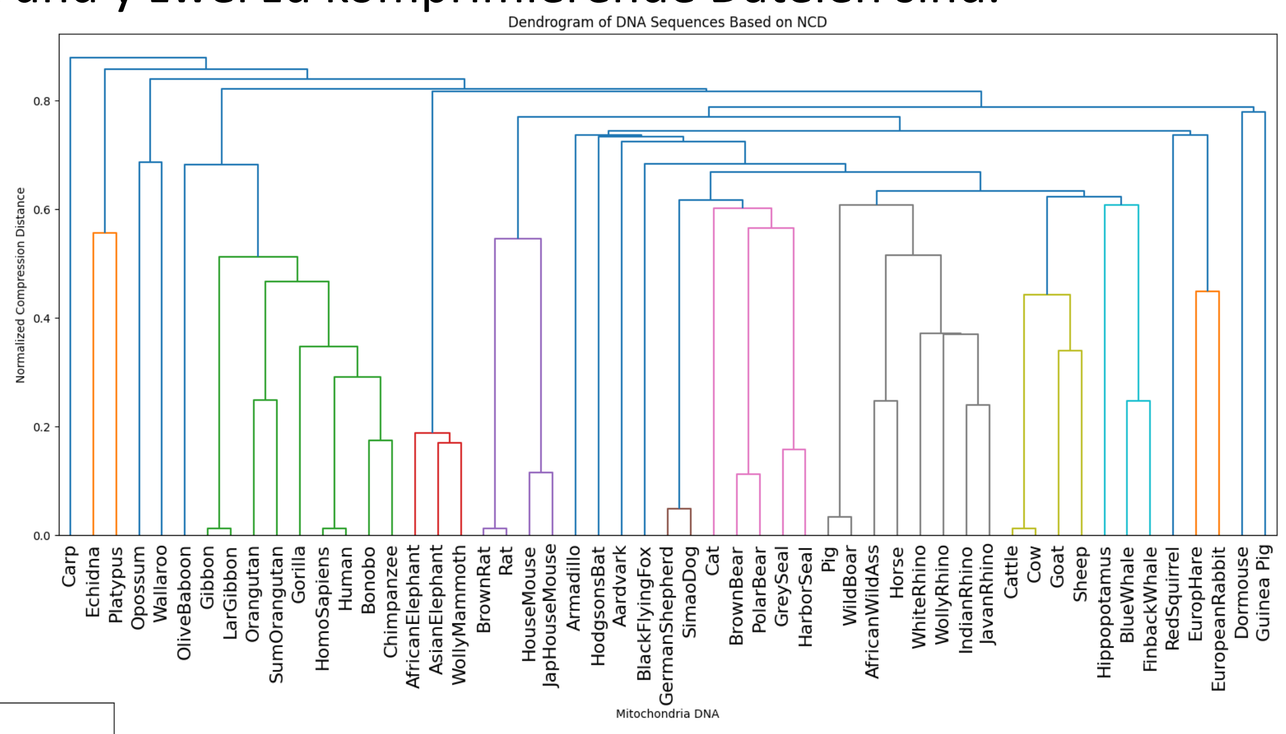
Wegen der steigenden Anzahl der Angriffe ist es besonders wichtig, neue Methoden zur Malwarebekämpfung zu entwickeln.

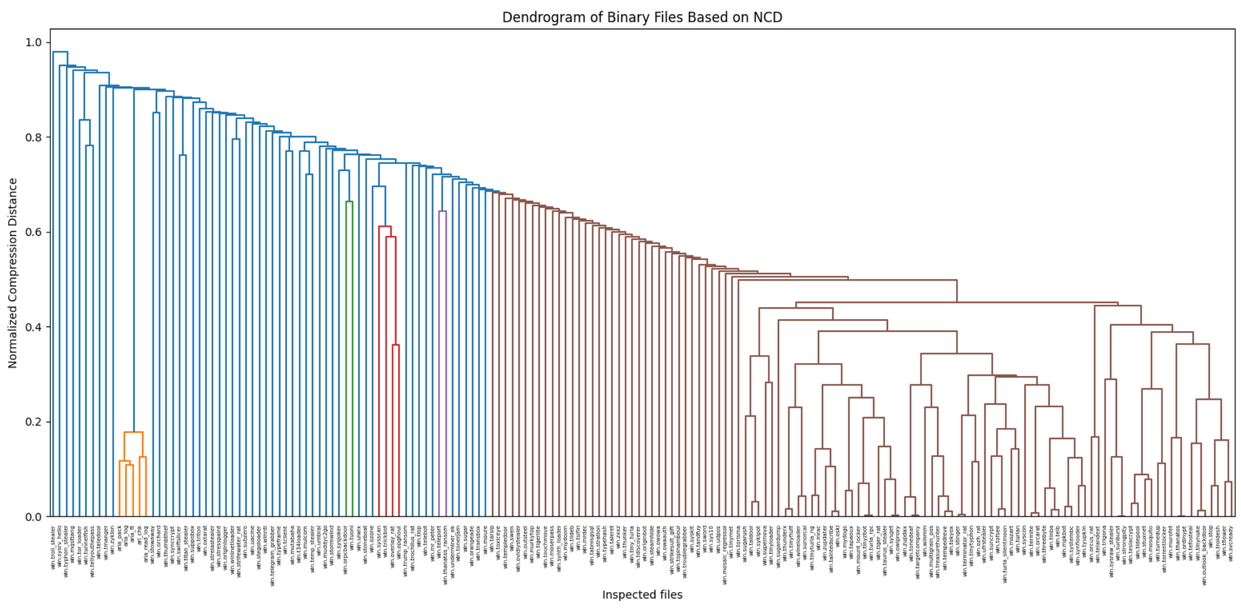
Unser Projekt hat das Ziel, ein Tool zu entwickeln, das zur Verienfachung der Analyse neuer Malware Ähnlichkeiten zu bereits bekannter Schadsoftware aufzeigen soll.

In unserem Projekt wird die Mustererkennung mithilfe von Kompressionsverfahren untersucht. Hierfür wurde ein Algorithmus entworfen, der mit LZMA-Kompression Daten liefert. Aus diesen Daten wird eine NCD-Matrix gebildet, und mit Agglomerative Clustering werden die Datensätze gruppiert.
Dabei steht NCD für „Normalized Compression Distance“, welche eine Ähnlichkeitsmetrik ist und lässt sich mit der unten stehenden Formel berechnen, wo x und y zwei zu komprimierende Dateien sind.

$$NCD(x,y) = \frac{C(x,y) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}},$$

Zuerst, zur Überprüfung, ob unser Algorithmus sinnvolle Ergebnisse liefert, wurde er auf Mitochondrien DNA-Sequenzen von Tieren getestet. Einen Ausschnitt aus einem solchen Verfahrenstest kann man rechts sehen.





	folder_name	file_path	original_size	compressed_size	compression_rate	ncd_win_originbot	ncd_js_cryptonight	ncd_win_zupdater
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	osx.kandykorn	/home/kali/Documents/test51/...	379824	85430	77.51	0.9988	0.9976	0.4991
2	elf.zerobot	/home/kali/Documents/test51/elf.zerobot/...	9727768	4526694	53.47	1.0	0.9998	0.9643
3	osx.kuiper	/home/kali/Documents/test51/osx.kuiper/...	2390096	807611	66.21	0.9998	0.9992	0.7996
4	elf.hiatus_rat	/home/kali/Documents/test51/elf.hiatus_ra...	1034032	419722	59.41	0.9997	0.9984	0.6147
5	win.outsteel	/home/kali/Documents/test51/win.outsteel...	112922	17680	84.34	0.9935	0.9951	0.499
6	win.ovidiystealer	/home/kali/Documents/test51/...	102400	26776	73.85	0.9678	0.9997	0.4963
7	win.wininetloader	/home/kali/Documents/test51/...	1305088	445599	65.86	0.9991	0.9996	0.6364
8	jar.crossrat	/home/kali/Documents/test51/jar.crossrat/...	222543	215690	3.08	0.9998	0.9999	0.5013
9	elf.kobalos	/home/kali/Documents/test51/elf.kobalos/...	36624	14242	61.11	0.9925	0.9984	0.5008
10	win.muddyc2go	/home/kali/Documents/test51/...	192000	122233	36.34	0.9869	0.9977	0.4948
11	jar.supremebot	/home/kali/Documents/test51/...	104889	95964	8.51	1.0	1.0004	0.5011

Die uns bereitgestellte Datenbank umfasst mehr als 2.000 Viren. Auf dem Bild links sind lediglich 10% davon zu sehen, und selbst dieser Ausschnitt erscheint überladen. Daher wird für die Analyse auch die von uns erstellte Datenbank mit SQL-Anfragen genutzt.
Zur Veranschaulichung der Funktionsweise des Dendrogramms haben wir Binärdateien von Linux Systemprogrammen hinzugefügt, die orange hervorgehoben sind.
Die Überladung der Dendogramme ist zugleich kein kein großes Hindernis, da es am Ende des Projekts darauf ankommt, neue Viren in die Datenbank einzuspeisen, um ihren Ähnlichkeitsgrad zu bereits bekannten Viren festzustellen.

Ausblick: Die NCD-Matrix scheint gute Ergebnisse für Viren zu liefern, deren Eigenkompression zwischen 40 % und 70 % liegt. Die Bildung der Matrix erfordert jedoch exponentiell viel Zeit. Daher bleibt ein großer Teil der Virendatenbank noch zu untersuchen. Daher bleibt ein großer Teil der Virendatenbank in einer erweiterten Analyse noch zu untersuchen.

h_da hochschule darmstadt informatik | member of EUT+ EUROPEAN UNIVERSITY OF TECHNOLOGY

Praxisprojekt 2
Informatik Dual
2025

Gruppe A