# Software Fingerprinting for Supply Chain Security

Ted Habeck Jiyong Jang Dhilung Kirat Ian Molloy JR Rao Doug Schales IBM Research Security



# Supply Chain Attacks

- SolarWinds (2019-2021) est. cost > \$100B
  - Malicious code (backdoor) pushed out through updates
- Dependency confusion (Feb 2021)
  - Private vs public packages (npm, PyPi, RubyGems)
- Codecov (Apr 2021)
  - DevOps tool. Vulnerability in CI. Bash uploader modified
- Kaseya (Jul 2021) ransom \$70M
  - IT solutions, including VSA (remote monitoring and management software) to deliver REvil ransomware
- Protestware (Mar 2022)
  - Popular NPM package wiped files in Russia and Belarus



# Codecov breach impacted 'hundreds' of customer networks: report



Updated: Reports suggest the initial hack may have led to a more extensive supply chain attack.

CISA-FBI Guidance for MSPs and their Customers Affected by the Kaseya VSA Supply-Chain Ransomware Attack

### Supply Chain Security Industry approach to protecting CI/CD pipelines



## Proposal: Software Fingerprints for Software Assurance

- Hashing (e.g., SHA2) verifies *exact* matches in files
  - One-bit changes result in large differences
- Signatures verify integrity assuming root of trust
  - Requires full trust in signer
- Fuzzy Hashing (e.g., **ssdeep**) provides partial matches to known files
  - Handles minor syntactic differences but lacks file semantics

### Ken Thompson's *Reflections on Trusting Trust*

- There is a need to verify code and binary integrity and identity
  - Sourced from multiple distributions, compilers, and optimizations
- Identify and verify legacy software in deployments
- Verify software version or backported patches, or code functionality



**Robustness for legacy** 



### **Stability across versions**

### Supply Chain Security Open security issues and residual risks



## Code Genome: Semantically meaningful fingerprint

int f4(int a){

- Across multiple architectures (x86, ARM, ...)
- Across multiple compilers (gcc, clang, ...)
- Across multiple optimization levels
- Handling obfuscation



Same Genome

x100000d10: 55

x100000d11: 48 89 e5

push rbp

mov rbp, rsp

ptr [rbp - 8], edi

# Key Idea: Code Genome construction



Genome can be constructed from closed-source/legacy code where source code is not easily available.

### Use Case 1: Finding Log4j Legacy Software Discovery

- Software deployments are a turducken
  - **zip**, **tar**, container image, **jar**, etc.
  - Dependencies often wrapped up
- No good provenance or CMDB
- Can't rely on standard directories, filenames, or hashes
- Code can be repackaged
- Applies across the board: CICD, DevSecOps and Legacy



# Use Case 2: SBOM Verification

- Problem
  - Each vendor creates SBOM of their own software including open-source and closed-source components. How can we verify its correctness (containing incorrect library mistakenly/maliciously) and completeness (missing library)?

– Value

- Given software, we can verify (generate) SBOM
- Support closed-source and legacy software without requiring source code access
- Help developers generate correct SBOM
- Vet software before integrating/deploying into a product



### Trust but Verify SBOM: Metadata vs. Code



#### CycloneDX SBOM Standard

CycloneDX is a modern standard for the software supply chain. SBOM, SaaSBOM, OBOM, Advisories, VEX, and more. CycloneDX is a OWASP Flagship Project.

⊙ OWASP & https://cyclonedx.org/ ♥ @CycloneDX\_Spec (Verified)

"Unfortunately, some images – such as the official node image on Docker Hub – incorrectly report the version of OpenSSL that's used by the Node.js runtime."

<pre># Syft detects the image contains OpenSSL 1.1.1 (not vulnerable): \$ syft node   grep openssl</pre>
openssl 1.1.1n-0+deb11u3 deb
<pre># But the node runtime actually uses the vulnerable OpenSSL 3.0.5! \$ docker run node -e "console.log(process.versions)"   grep openssl openssl: '3.0.5+quic',</pre>
# The Chainguard Image for node correctly reports the version of OpenSSL it uses, # due to its build-time SBOM.
\$ cosign download sbomplatform=linux/amd64 cgr.dev/chainguard/node   grep libssl libssl3-3.0.5-r3
# Syft doesn't have to detect what's in the image.
<pre>\$ syft cgr.dev/chainguard/node   grep libssl</pre>
libssl3 3.0.5-r3 apk

https://www.chainguard.dev/unchained/mitigating-critical-openssl-vulnerability-with-chainguard

#### Pinned

Specification Public Software Bill of Material (SBOM) standard designed for use in application security contexts and supply chain component analysis

● XSLT ☆ 168 % 34

#### Cyclonedx-python Public

Creates CycloneDX Software Bill of Materials (SBOM) from Python projects and environments.

● Pvthon ☆ 91 ♀ 35

Cyclonedx-node-module Public Creates CycloneDX Software Bill of Materials (SBOM) from Node.js projects ● JavaScript ☆ 86 % 67

Cyclonedx-maven-plugin Public Creates CycloneDX Software Bill of Materials (SBOM) from Maven projects

Creates CycloneDX Software Bill of Materials (SBOM) from .NET

● Java ☆ 125 ♀ 37

● C# ☆ 82 ¥ 41

Projects

Cvclonedx-dotnet Public

#### Cyclonedx-cli Public

CycloneDX CLI tool for SBOM analysis, merging, diffs and format conversions. ●C# ☆85 ¥21

This project provides a runnable Python-based application for generating CycloneDX bill-of-material documents from either:

- Your current Python Environment
- Your project's manifest (e.g. Pipfile.lock, poetry.lock or requirements.txt)
- Conda as a Package Manager

### 7 "bom-ref": "pkg:wget@1.20 "type": "library", "name": "wget", "version": "1.20.3-1ubun" "licenses": [

\$ sbom generation tools

Dockerfile > ...

2

3

4 5 6

### { "license": { "name": "UNKNOWN"

kerfile >	Dockerfile >	
FROM <u>ubuntu</u> :focal	1 FROM <u>ubuntu</u> :focal	
RUN apt-get update RUN apt-get install -y wget	3 RUN apt-get update 4 RUN apt-get install -y wget 5	
RUN apt-get update	6 RUN mv /var/lib/dpkg/status /var/l. 7 RUN touch /var/lib/dpkg/status 8	ib/dpkg/status.bak
of": "pkg:uget@1_20_3_1ubuptu2"	9 RUN apt-get update 10	
<pre>"" "pkg:wget@1.20.3-lubuntu2", " "library", " "wget", on": "1.20.3-lubuntu2", ses": [</pre>	<pre>sbom/docker &gt; grep wget sb sbom/docker &gt;</pre>	om.dpkg.json

# Demo: Verifying wget

	GNU Operating System Q Supported by the Free Software Foundation	
ABOUT GN	NU PHILOSOPHY LICENSES EDUCATION SOFTWARE >>	
GNU V	Wget	
GNU Wget is widely used Ir	s a <u>free software</u> package for retrieving files using HTTP, HTTPS, FTP and FTPS, the most Internet protocols. It is a non-interactive commandline tool, so it may easily be called from	
scripts, cron	jobs, terminals without X-Windows support, etc.	
GNU Wget ha including:	as many features to make retrieving large files or mirroring entire web or FTP sites easy,	
Can re	esume aborted downloads, using REST and RANGE	
<ul> <li>Can us</li> </ul>	se filename wild cards and recursively mirror directories	
<ul> <li>NLS-b</li> </ul>	based message files for many different languages	
<ul> <li>Option docum</li> </ul>	nally converts absolute links in downloaded documents to relative, so that downloaded nents may link to each other locally	
Runs o	on most UNIX-like operating systems as well as Microsoft Windows	
<ul> <li>Support</li> </ul>	orts HTTP proxies	
<ul> <li>Support</li> </ul>	orts HTTP cookies	
<ul> <li>Support</li> </ul>	orts persistent HTTP connections	
0 0 0	🕍 wget-1.19.4 — ./configure && make — ./configure — gcc < configure — 106×16	
[ imolloy@i checking f checking w checking f	ians-mbp > >-/Desktop/ssdeep/wget/wget-1.19.4 > ./configure && make 1656 + for a BSD-compatible install /usr/bin/install -c whether build environment is sane yes for a race-free mkdir -p/build-aux/install-sh -c -d	17:02:
checking f	for gawk no	
checking i	for nawk no	
	for awk awk	
checking w	whether make sets \$(MAKE) yes	
checking w	whether make supports nested variables yes	
checking b	build system type aarch64-apple-darwin21.6.0	
checking h	host system type aarch64-apple-darwin21.6.0	
checking w	whether make supports the include directive yes (GNU style)	
checking f	for gcc gcc	
checking w	whether the c compiler works yes	



## UI of POC

calhost:3001					
Unload files					
Max file size is 500mb.					
			= corinte		O Search
	Recents			<u> </u>	Gearch
	🔂 habeck	Today	Size	Kind	Date Added
	Applicatio	package.zip	9 KB	ZIP archive	Today at 11:04 F
	Desktop	раскаде			Today at 11:04 P
	Documents     Dovuploads	Previous 7 Days			
	Downloads	genrandomzip.sh	886 bytes	Shell Script	Oct 27, 2022 at
	Cloud				
	📑 Shared				
	⊖ pgAdm ≜				
	Red				
	Orange				Cancel Oper

🚲 React Redux App 🛛 🗙 🚽	÷		
C () localhost:3001/job/3ffff387-3c6	d-4584-9bb4-b5bea38bf391		۵ ۵ 🖈 🖿 🚨
l Scantool			::: 8
Total Tasks 176	Avg Task Time (seconds) O	File size (bytes) Total analysis time in seconds 7.3k 1	
uploads/3ffff387-3c6d-4584-9bb4	-b5bea38bf391/package.zip 7,318	package/479/14551/31208.zip 569 package/14303.zip 566	
	uploads/31111387-3c6d-4584-96b4-	package/2637.zip 563	
	ooseaaan Jurgaacage oo	package/30357.zip 571	

@ ☆ ☆ ♪ □ ≗

### 

IBM Scantool

#### History Upload

						ŵ
Date	$\downarrow$	Job ID		Statu	s	
10/27/2022, 4:06:29 PM		3ffff387-3c6d-4584-9bb4-b5bea38bf391		SUCC	EEDED	
10/27/2022, 3:57:08 PM		84916ee6-5029-4907-84b9-a348fc8326	0c	SUCC	EEDED	
10/27/2022, 3:43:35 PM		1c8f0caa-e016-4d32-b6aa-565f971ea7ef		SUCC	EEDED	
10/27/2022, 3:19:47 PM		1ed86fbc-7972-42fa-b945-0e40005a23e	e	SUCC	EEDED	
10/27/2022, 1:47:25 PM		fea4a97a-f831-4df1-aae8-78e3f6dad322		SUCC	EEDED	
10/27/2022, 1:37:18 PM		d4b4d451-c206-49f0-bb5b-1f42828d91b	4	SUCC	EEDED	
10/27/2022, 1:33:33 PM		9943ec9d-6231-44b7-a468-d0e2ac214d	7c	SUCC	EEDED	
10/27/2022, 1:18:28 PM		793ab921-8aff-4eae-ac46-f360432423dc		SUCC	EEDED	
10/27/2022, 12:26:19 PM		ab43df75-c1f0-422f-89b9-c0f768dc206b		SUCC	EEDED	
10/27/2022, 12:10:02 PM		3934eafc-e851-4c9e-8bbb-6efa94f8f0a0		SUCC	EEDED	
10/27/2022, 11:53:10 AM		676fccab-166e-4a59-969a-00e0436e7c8		SUCC	EEDED	
10/27/2022, 11:35:53 AM		57e5d1c6-293d-4033-9d08-81666fcb96e	e5	SUCC	EEDED	
Invalid Date		deef5c9b-77be-4274-be7f-797ef047d4a0		SUCC	EEDED	
ltems per page: 20 ∨	1 - 1	3 of 13 items	1∨ of	1 page		

Secure Software Supply Chain © 2022 IBM Corporation

### Status and Roadmap

### **Current Status**

- Several techniques for computing genes
- Support for binaries, packages, bytecode
- Cloud native application / processing engine
- Currently performing large-scale evaluation
- Additional features and capabilities in development

### Plans for Release:

- A version of genome generation
- Service demonstrating technology
- Utilities for handling genomes and querying service

### **Requests:**

- Welcome feedback, support and collaboration
- Insights on capabilities and use cases
- How to complement existing OpenSSF projects