Open Source Vulnerability Metrics

Dr. Tate

Seth Goodwin, Michael Follari, Jaron Dunham,
Gabe Wilmoth, Rohit Gade

UNC GREENSBORO

Overview

Software Heritage Graph (SHGD):

- Largest public archive of source code and development history
- More than one billion commits from over eight million projects
- Over one TB of data

National Vulnerability (NVD):

- Over 120,000 distinct security vulnerabilities
- Each entry identifying software affected, severity, and etc..

Goals

Cross-reference known software vulnerabilities found on the *National Vulnerability Dataset* with commits found in the *Software Heritage Graph Dataset*

- Is there a relationship between project activity and vulnerability severity?
- How long is there between when a software vulnerability is discovered and when it's patched?
- How long is there between a fix and a new software release?

Task Breakdown

Seth Goodwin

- Looking through and cleaning revision.csv

Michael Follari

- Looking through and cleaning release.csv

Jaron Dunham

- Parsing NVD data into useful and relevant data

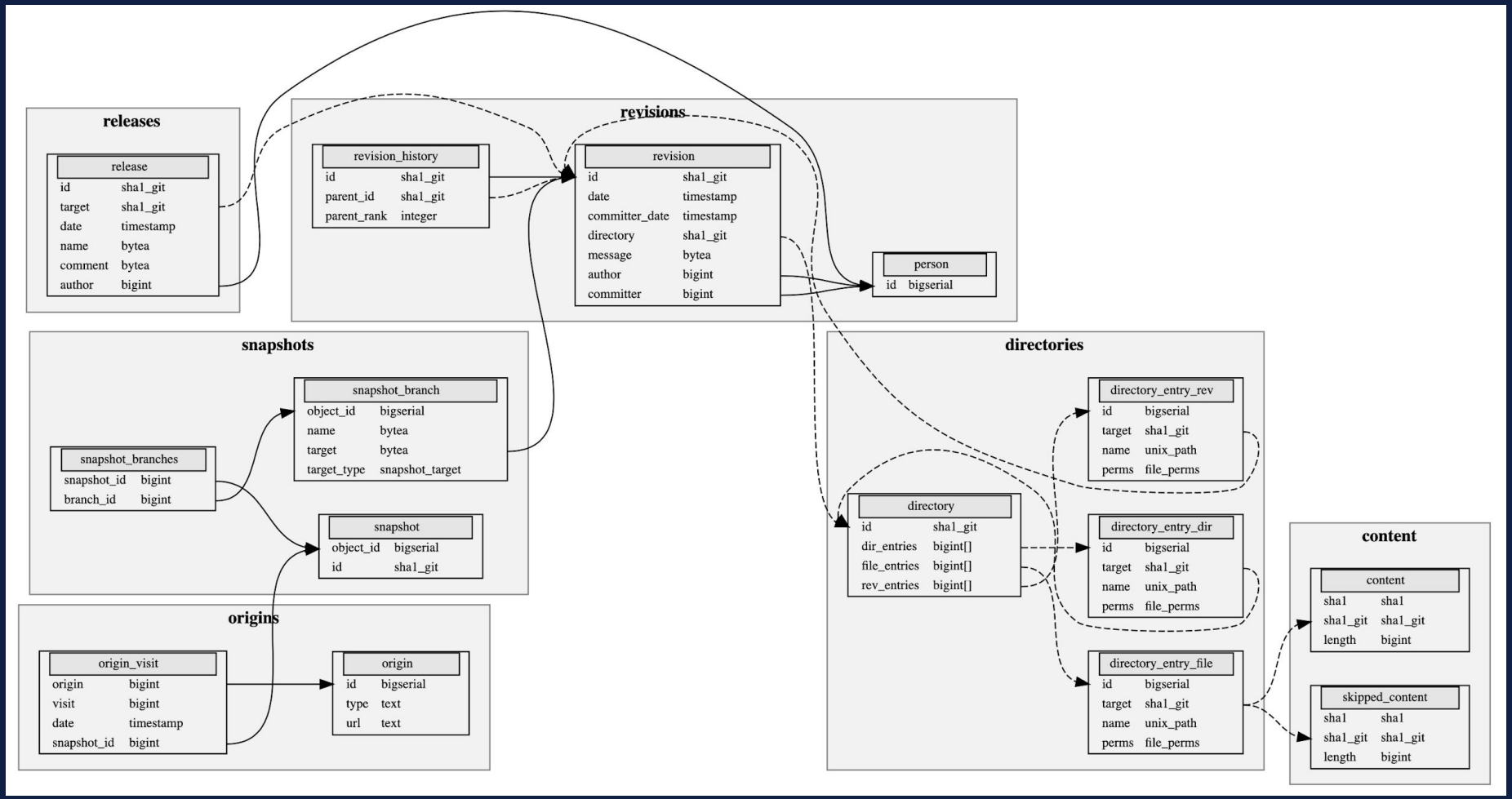
Gabe Wilmoth

- Trying to connect NVD with SWHGD

Rohit Gade

- Extracting GitHub hash id from links found on the NVD

SHGD - Simplified Schema



UNC GREENSBORO

Data Dictionary (SHGD)

- revision: contains the revisions stored in the archive.
  - id (bytes): the intrinsic identifier of the revision, recursively computed with the Git SHA-1 algorithm. For Git repositories, this corresponds to the revision hash.
  - date (timestamp): the date the revision was authored
  - committer_date (timestamp): the date the revision was committed
  - author (integer): the author of the revision
  - committer (integer): the committer of the revision
  - message (bytes): the revision message
  - directory (bytes): the Git SHA-1 of the directory the revision points to. Every revision points to the root directory of the project source tree to which it corresponds.
- release: contains the releases stored in the archive.
  - id (bytes)
  - target (bytes): the Git SHA-1 of the object the release points to.
  - name (bytes): the release name
  - date (timestamp), author (integer), message (bytes)

Data Dictionary (NVD)

- id (string): unique identifier for each reported vulnerability marked by year and numeric increment (CVE-yyyy-iiiii)
- reference (string): hyperlink reference to vulnerability and/or patch
- severity:
  - confidentiality (string): marker scoring how vulnerable said vulnerability left the software to unauthorized disclosure of information
  - integrity (string): marker scoring how vulnerable said vulnerability left the software to unauthorized modification of information
  - availability (string): marker scoring how vulnerable said vulnerability left the software to being maliciously taken down
    - marker values: (none, low, partial, high)

Data Statistic

Software Heritage Graph Dataset (SHGD):

- 1.2 TB of data
- Sample dataset of 3052 unique popular repositories
  - 1000 most popular GitHub projects in python + 131 from Gitlab
  - 1000 most popular PyPi projects
  - 1000 most popular Debian packages

National Vulnerability Dataset (NVD):

- 123,029 unique vulnerability entries
  - Spanning 1988 - Present

Data Extraction

	id	date	message
0	01714ff5fd94a846f7dc3456a52e6f2dcd36ca0b	2015-04-02 01:11:50	Implement dask.array.take\n\nIn principle, we ...
1	01d76a45b817be788eab3e27a93e41c74a6d8957	2010-08-14 17:15:31	Convert to spaces.\n
2	03ac0bf5d03af97bc7dc7c5aa67d8ee346f8cd1c	2013-09-27 17:02:55	Merge pull request #4887 from cpcloud/groupby-...
3	05c9110ba2615d42af41a77138bc62dd18278320	2011-10-25 20:11:52	Merge pull request #162 from gabrielhurley/use...
4	06de8d8e88d5b6311ea3feae369c85d157c9dfe3	2017-10-31 20:29:21	Update patch set 1\n\nPatch Set 1: Presubmit-V...
...	...	...	...
5188989	fb5183dd25cb0bde1f8a1da20d07b940883f8f17	2012-10-16 02:18:13	Fixes for unary and indexing operations.\n
5188990	fc05c476f53b6aa6188070601fada55a6677ef01	2017-10-23 20:49:07	Revert "Update CONTRIBUTING.md"\n\nThis revert...
5188991	fcd97879e3cd57a15f91db81ff88da7c6c114b98	2013-08-27 14:12:35	Improve SEO tools CSS across themes\n\nbzr rev...
5188992	fd71164472400e8f373a1d9de7d6e92a6aa8be07	2017-05-23 21:18:40	Update CONTRIBUTING.md\n\nFixing broken issues...
5188993	fe181f4848a8c774155b8d853c2f53f7e7679872	2010-06-18 23:10:20	Only run CoalesceExtSubRegs when we can expect...

SHGD: 5,188,994 total rows

Data Extraction

NVD Dataframe head: 4695 total rows

	cveid	year	hash	desc	link	publishedDate
0	654	2019	fd37bd8fb2b9c306079db505e0e3fe79a096c31c	phpIPAM version 1.3.2 and earlier contains a C...	http://github.com/phpipam/phpipam/commit/fd37b...	2019-02-04T21:29Z
1	659	2019	33e2692a37b5b6340cf5bec1a84e541460983c03	Chamilo Chamilo-lms version 1.11.8 and earlier...	http://github.com/chamilo/chamilo-lms/commit/3...	2019-02-04T21:29Z
2	660	2019	b97a4b658814b2de8b9f2a3bce491c002d34de31	FFMPEG version 4.1 contains a CWE-129: Improper...	http://github.com/FFmpeg/FFmpeg/commit/b97a4b6...	2019-02-04T21:29Z
3	661	2019	33e2692a37b5b6340cf5bec1a84e541460983c03	Chamilo Chamilo-lms version 1.11.8 and earlier...	http://github.com/chamilo/chamilo-lms/commit/3...	2019-02-04T21:29Z
4	826	2019	81a4b8620188e89f7e4fc985f3c89b58d4bcc86b	utils/find-opencv.js in node-opencv (aka OpenC...	http://github.com/peterbraden/node-opencv/comm...	2019-03-26T01:29Z

Related Work