Open Source Vulnerability Metrics

Dr. Tate

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

UNC GREENSBORO

Data Statistics

Software Heritage Graph Dataset (SHGD)

- Largest public archive of open source software code
- 1+ billion commits from 8+ million projects
- 1.2 TB of data
- Teaser Datasets
    - popular-4k (23 GB)
    - popular-3k-python (4.7 GB)

National Vulnerability (NVD)

- 123,029 security vulnerabilities
    - Vulnerability description, affected software, severity, and impact

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 (2.19 GB)

Michael Follari

- Explore/Clean release.csv, Explore Time Correlation

Jaron Dunham

- Parsing NVD data into useful and relevant data

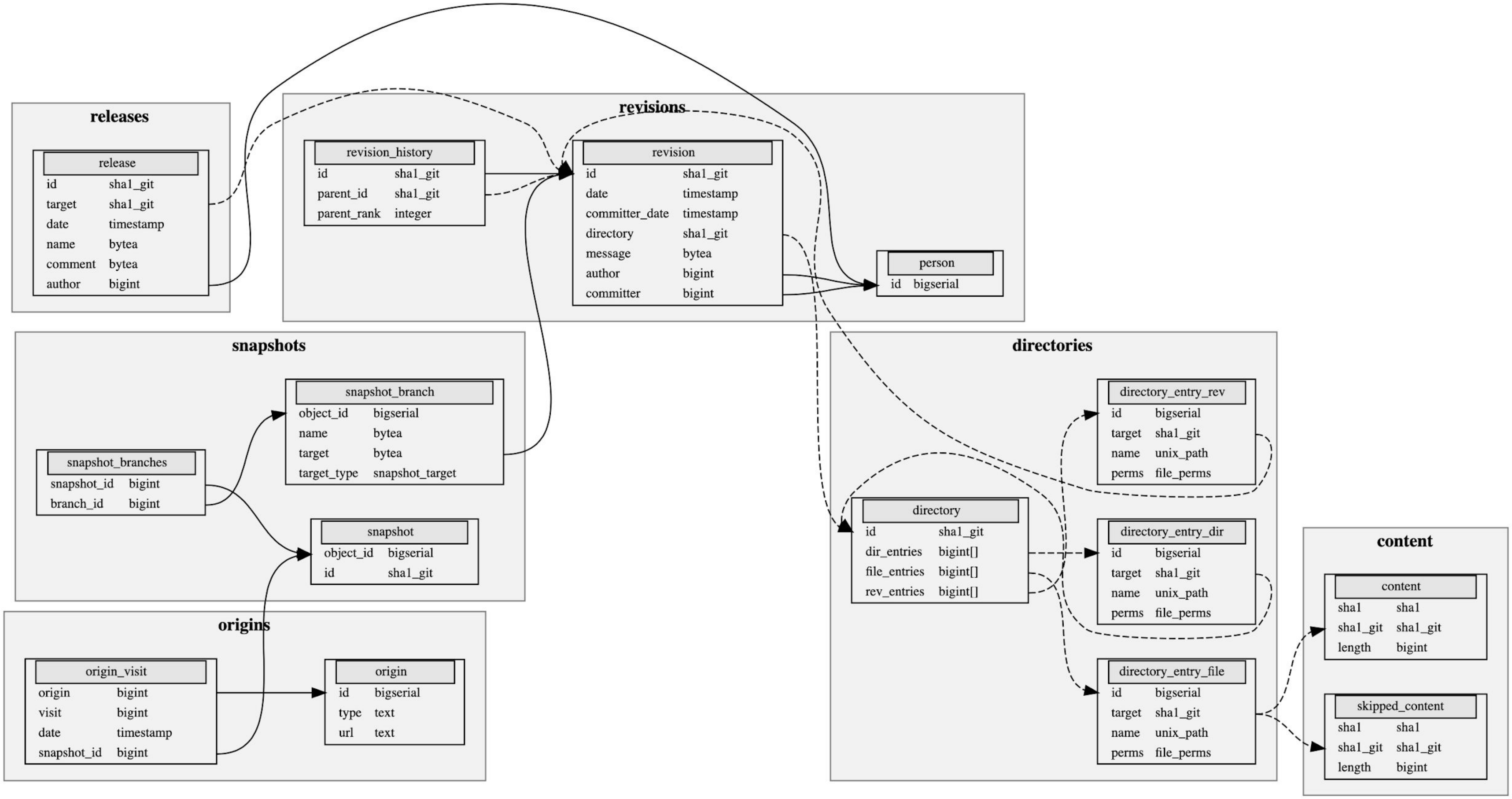
Gabe Wilmoth

- Trying to connect NVD with SHGD

Rohit Gade

- Extracting GitHub hash id from links found on the NVD

SHGD - Schema



releases

release	
id	sha1_git
target	sha1_git
date	timestamp
name	bytea
comment	bytea
author	bigint

revisions

revision_history	
id	sha1_git
parent_id	sha1_git
parent_rank	integer

revision	
id	sha1_git
date	timestamp
committer_date	timestamp
directory	sha1_git
message	bytea
author	bigint
committer	bigint

person	
id	bigserial

snapshots

snapshot_branches	
snapshot_id	bigint
branch_id	bigint

snapshot_branch	
object_id	bigserial
name	bytea
target	bytea
target_type	snapshot_target

snapshot	
object_id	bigserial
id	sha1_git

origins

origin_visit	
origin	bigint
visit	bigint
date	timestamp
snapshot_id	bigint

origin	
id	bigserial
type	text
url	text

directories

directory	
id	sha1_git
dir_entries	bigint[]
file_entries	bigint[]
rev_entries	bigint[]

directory_entry_rev	
id	bigserial
target	sha1_git
name	unix_path
perms	file_perms

directory_entry_dir	
id	bigserial
target	sha1_git
name	unix_path
perms	file_perms

directory_entry_file	
id	bigserial
target	sha1_git
name	unix_path
perms	file_perms

content

content	
sha1	sha1
sha1_git	sha1_git
length	bigint

skipped_content	
sha1	sha1
sha1_git	sha1_git
length	bigint

https://docs.softwareheritage.org/devel/swh-dataset/graph/schema.html

Time Correlation Scouting

Explore viability of time correlation

- Spike in releases after vulnerability discovery

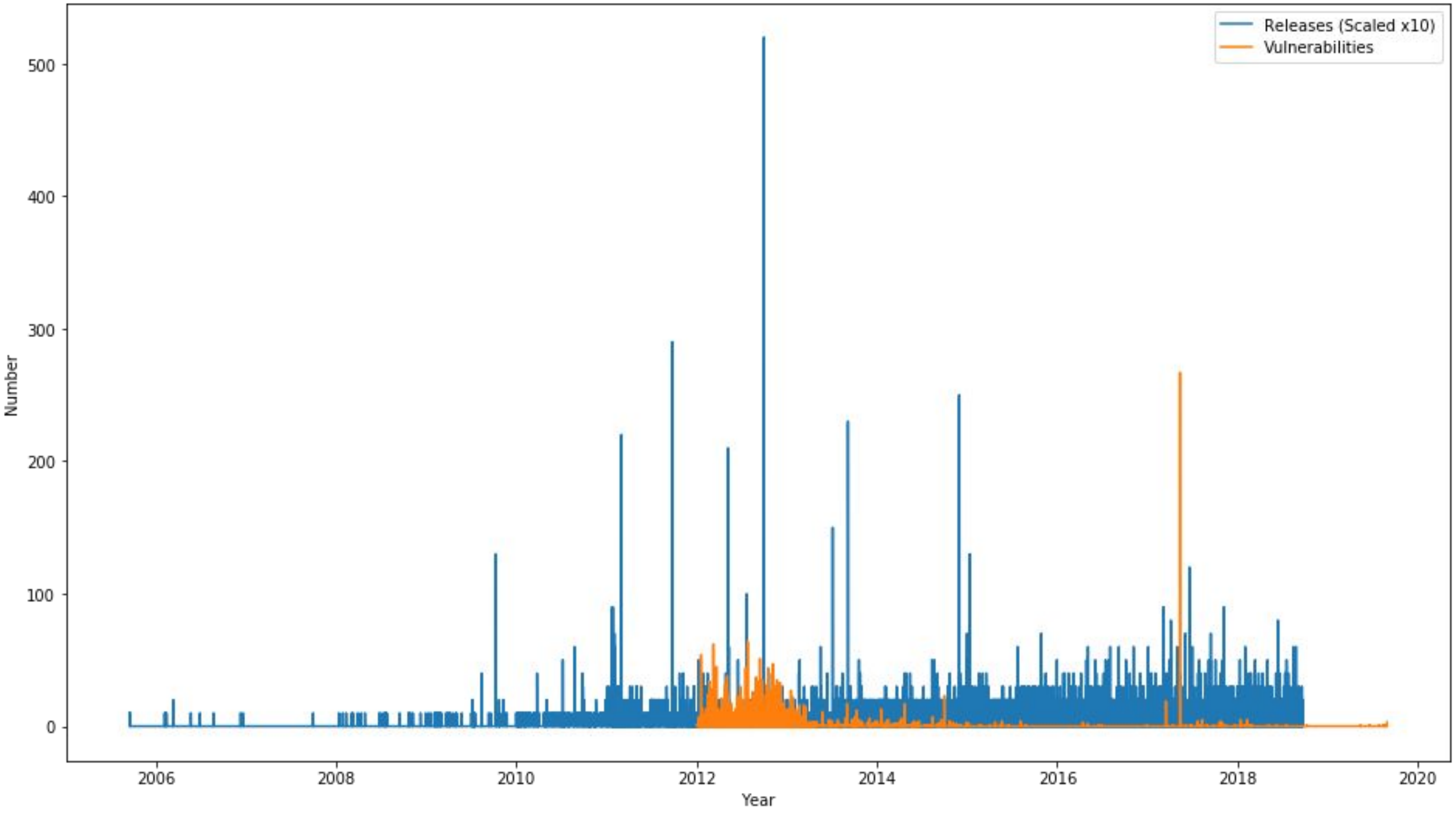
Data

- Software “releases”
  - When software is updated with an official release
  - Small dataset (1.4GB), python-3k teaser (6MB)
- NVD
  - Reference date of vulnerability discovery

Goals

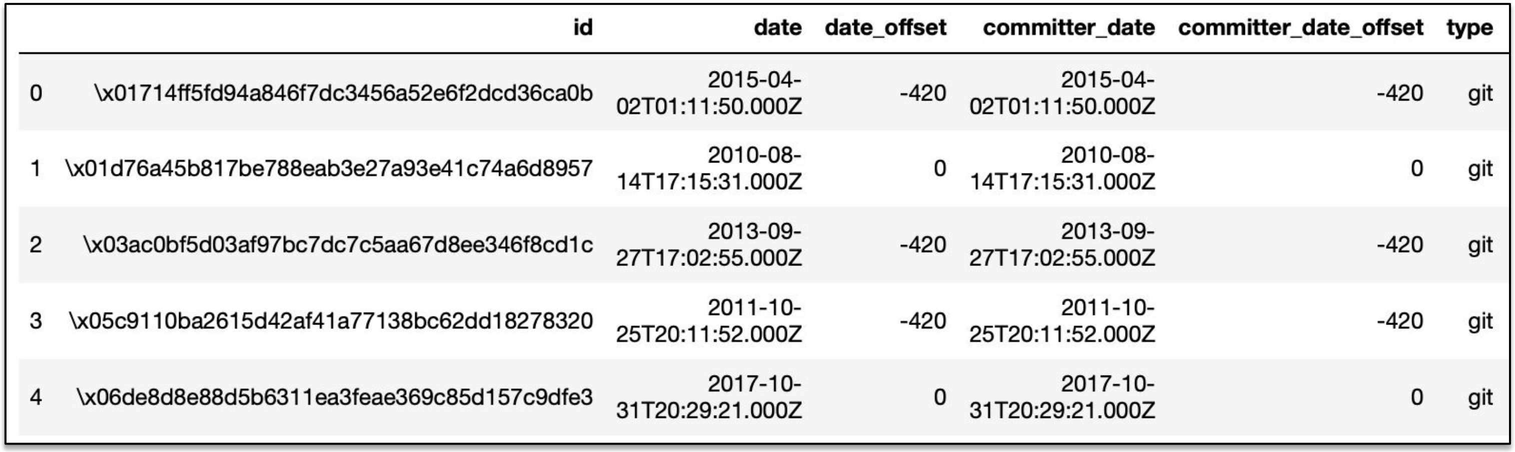
- Delay between vulnerability and fix
- Relationship between severity and delay

Time Correlation?

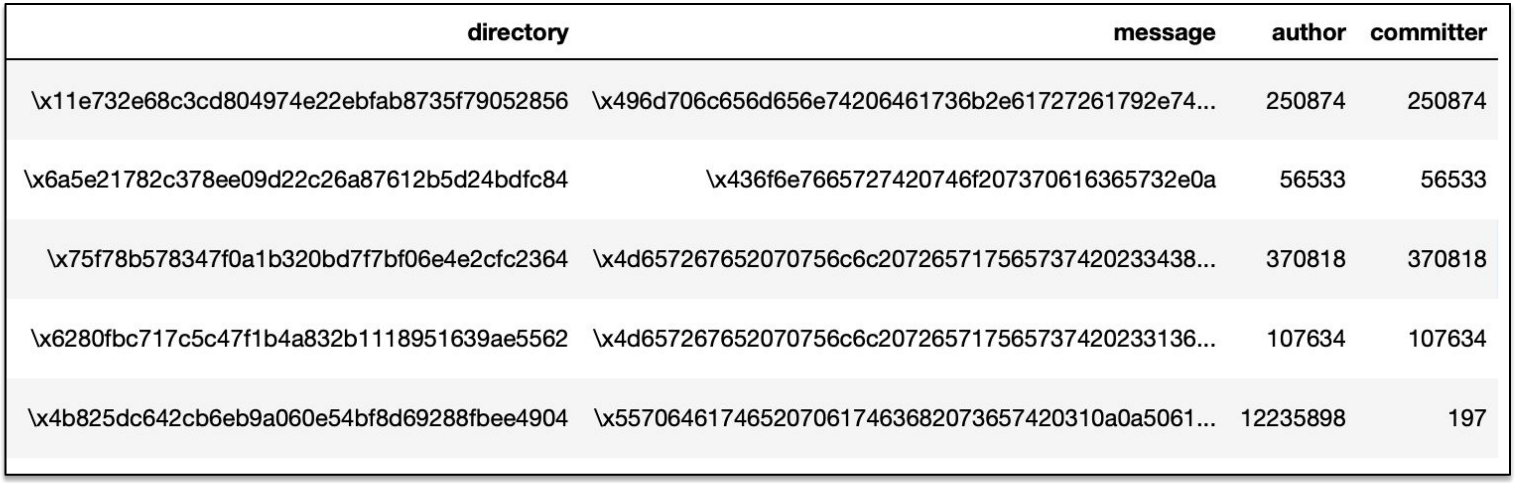


Python-3k software releases plotted with vulnerabilities from 2012 onward

SHGD - Raw Data Extraction



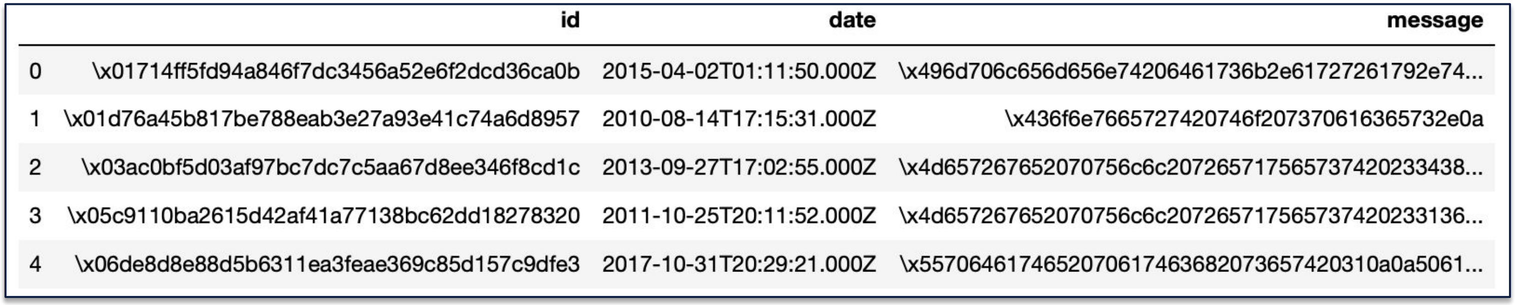
	id	date	date_offset	committer_date	committer_date_offset	type
0	\x01714ff5fd94a846f7dc3456a52e6f2dcd36ca0b	2015-04-02T01:11:50.000Z	-420	2015-04-02T01:11:50.000Z	-420	git
1	\x01d76a45b817be788eab3e27a93e41c74a6d8957	2010-08-14T17:15:31.000Z	0	2010-08-14T17:15:31.000Z	0	git
2	\x03ac0bf5d03af97bc7dc7c5aa67d8ee346f8cd1c	2013-09-27T17:02:55.000Z	-420	2013-09-27T17:02:55.000Z	-420	git
3	\x05c9110ba2615d42af41a77138bc62dd18278320	2011-10-25T20:11:52.000Z	-420	2011-10-25T20:11:52.000Z	-420	git
4	\x06de8d8e88d5b6311ea3feae369c85d157c9dfe3	2017-10-31T20:29:21.000Z	0	2017-10-31T20:29:21.000Z	0	git



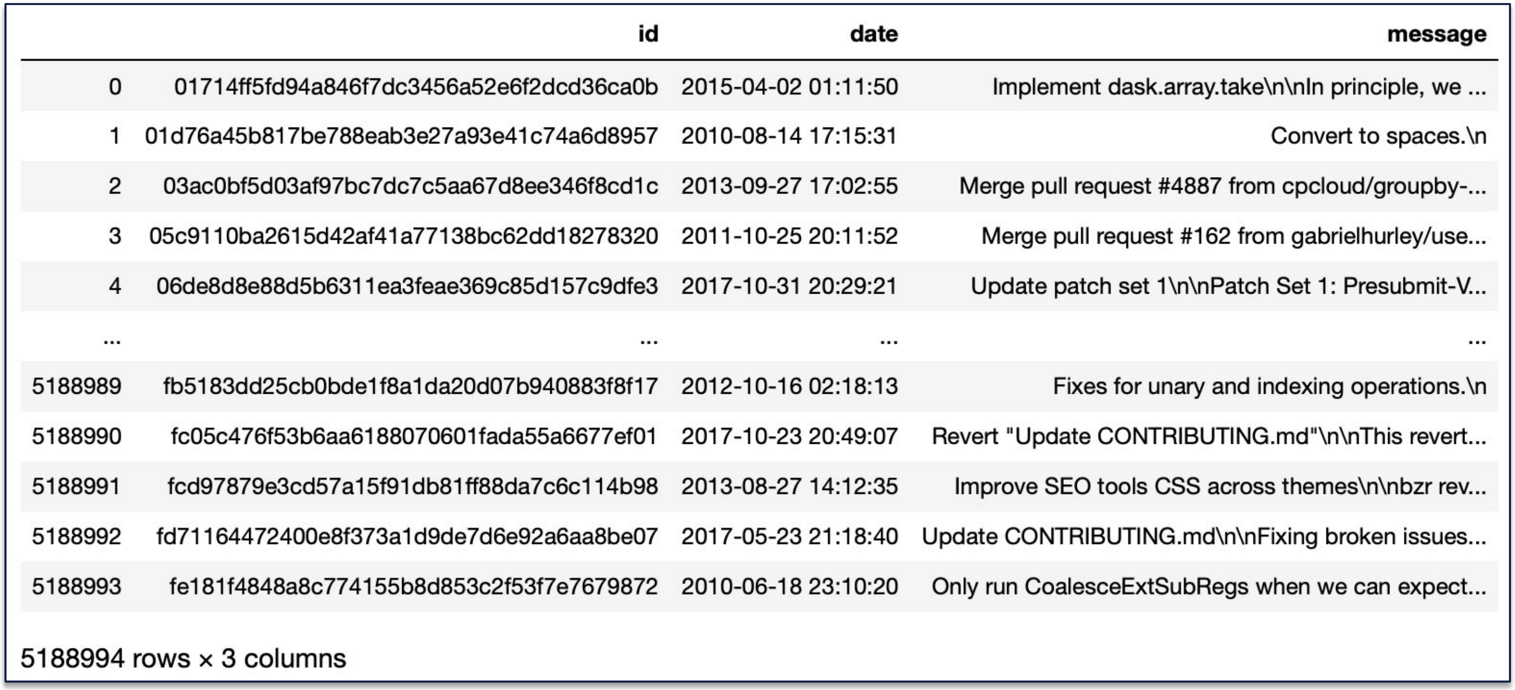
directory	message	author	committer
\x11e732e68c3cd804974e22ebfab8735f79052856	\x496d706c656d656e74206461736b2e61727261792e74...	250874	250874
\x6a5e21782c378ee09d22c26a87612b5d24bdfc84	\x436f6e7665727420746f207370616365732e0a	56533	56533
\x75f78b578347f0a1b320bd7f7bf06e4e2cfc2364	\x4d657267652070756c6c207265717565737420233438...	370818	370818
\x6280fbc717c5c47f1b4a832b1118951639ae5562	\x4d657267652070756c6c207265717565737420233136...	107634	107634
\x4b825dc642cb6eb9a060e54bf8d69288fbee4904	\x557064617465207061746368207365742031 0a0a5061...	12235898	197

revision.csv

SHGD - Data Cleaning



	id	date	message
0	\x01714ff5fd94a846f7dc3456a52e6f2dcd36ca0b	2015-04-02T01:11:50.000Z	\x496d706c656d656e74206461736b2e61727261792e74...
1	\x01d76a45b817be788eab3e27a93e41c74a6d8957	2010-08-14T17:15:31.000Z	\x436f6e7665727420746f207370616365732e0a
2	\x03ac0bf5d03af97bc7dc7c5aa67d8ee346f8cd1c	2013-09-27T17:02:55.000Z	\x4d657267652070756c6c207265717565737420233438...
3	\x05c9110ba2615d42af41a77138bc62dd18278320	2011-10-25T20:11:52.000Z	\x4d657267652070756c6c207265717565737420233136...
4	\x06de8d8e88d5b6311ea3feae369c85d157c9dfe3	2017-10-31T20:29:21.000Z	\x5570646174652070617463682073657420310a0a5061...

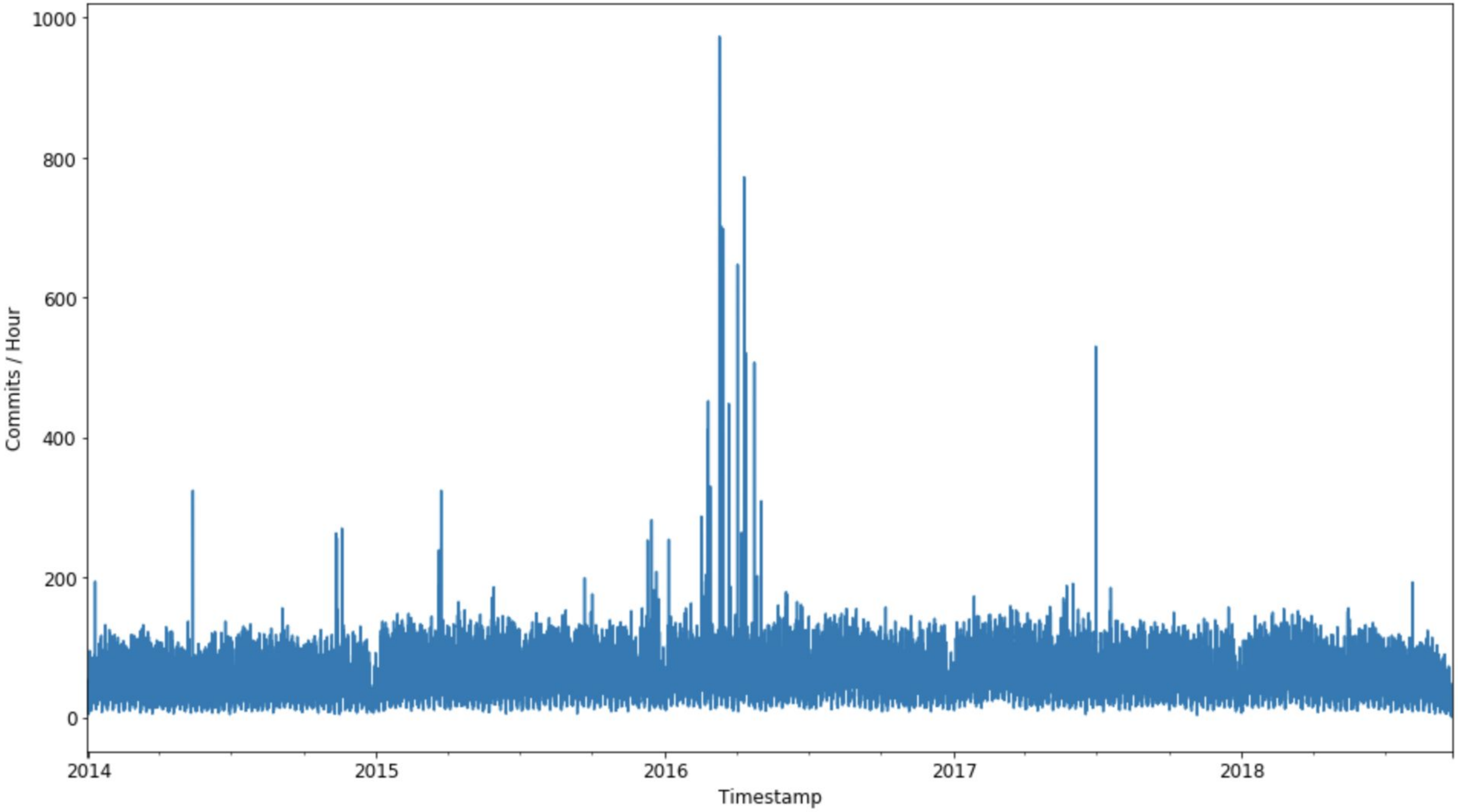


	id	date	message
0	01714ff5fd94a846f7dc3456a52e6f2dcd36ca0b	2015-04-02 01:11:50	Implement dask.array.take\n\nIn principle, we ...
1	01d76a45b817be788eab3e27a93e41c74a6d8957	2010-08-14 17:15:31	Convert to spaces.\n
2	03ac0bf5d03af97bc7dc7c5aa67d8ee346f8cd1c	2013-09-27 17:02:55	Merge pull request #4887 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...

5188994 rows × 3 columns

revision.csv

GitHub Activity

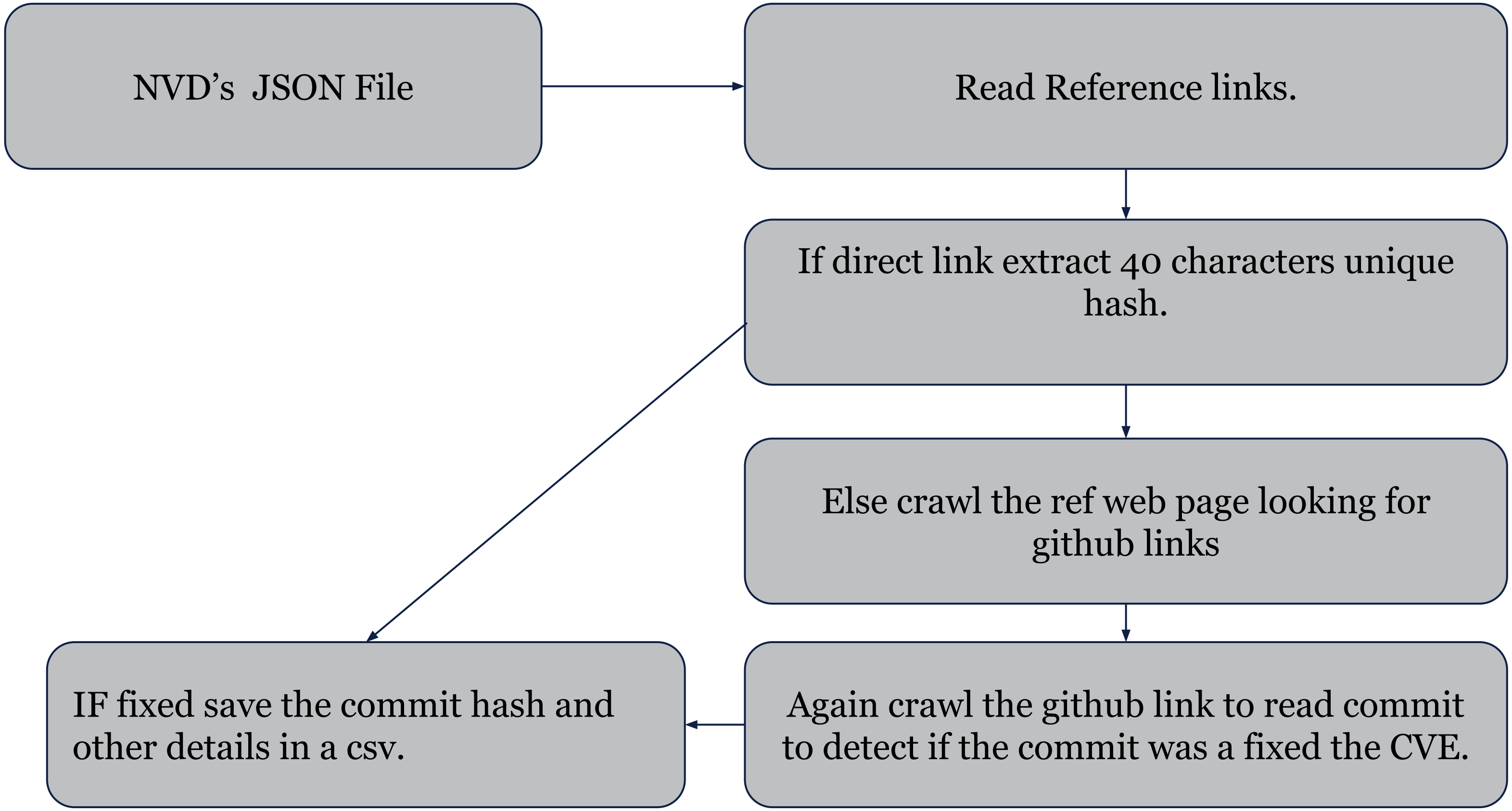


Commits / hour during the last 5 years

Parsing NVD data into useful and relevant data

- Taking the NVD json files and parsing the data into more easily readable information
- Makes it more convenient to cross reference with SHGD
- Once cross referenced with the SHGD, any parts included in the referenced section can be freely accessed

Extracting commit hashes



NVD’s JSON File

Read Reference links.

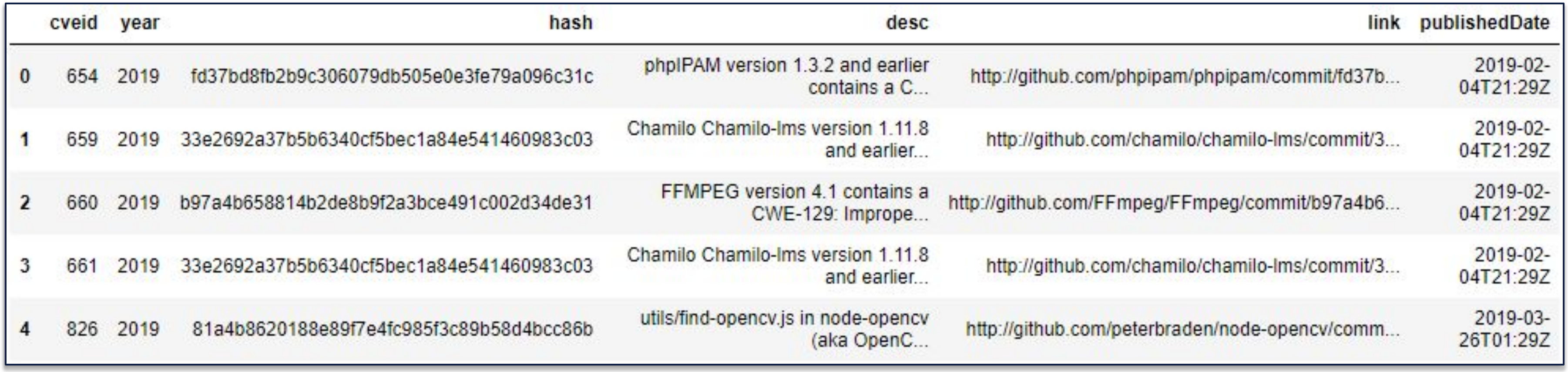
If direct link extract 40 characters unique hash.

Else crawl the ref web page looking for github links

Again crawl the github link to read commit to detect if the commit was a fixed the CVE.

IF fixed save the commit hash and other details in a csv.

NVD Data Extraction

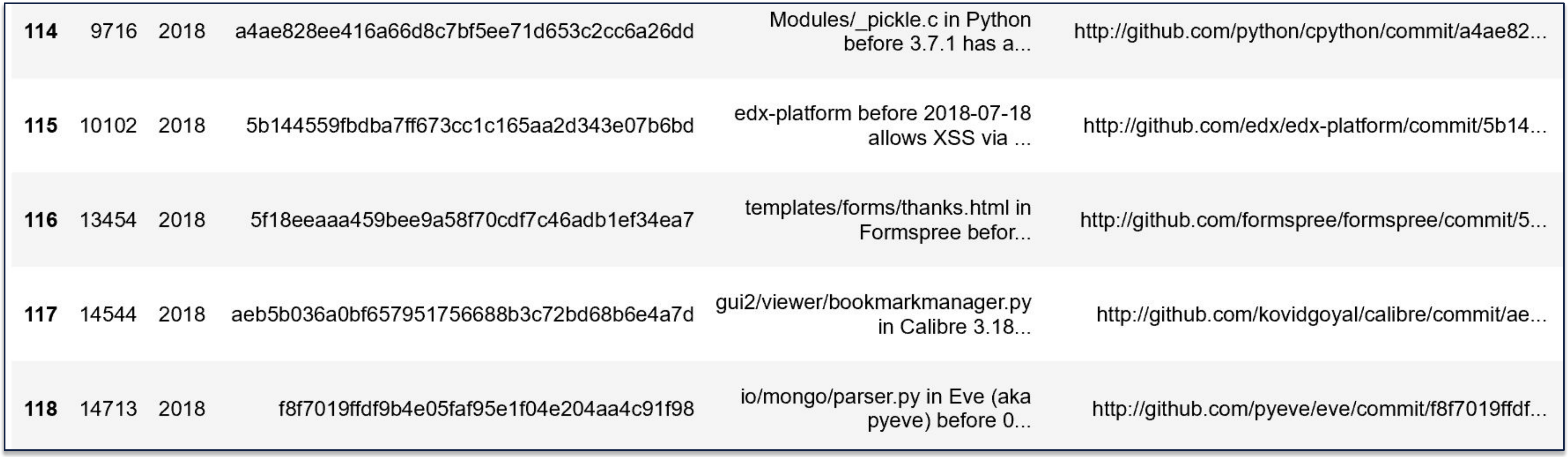


	cveid	year	hash	desc	link	publishedDate
0	654	2019	fd37bd8fb2b9c306079db505e0e3fe79a096c31c	phpIPAM version 1.3.2 and earlier contains a C...	http://github.com/phpipam/phpipam/commit/fd37b...	2019-02-04T21:29Z
1	659	2019	33e2692a37b5b6340cf5bec1a84e541460983c03	Chamilo Chamilo-lms version 1.11.8 and 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: Imprope...	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

2019 NVD Dataframe head: 4695 total rows

Cross Referencing NVD & SHGD Hashes

- 119 Total rows in common between NVD teaser and SHGD
- Are hashes a viable way to link the two datasets?



114	9716	2018	a4ae828ee416a66d8c7bf5ee71d653c2cc6a26dd	Modules/_pickle.c in Python before 3.7.1 has a...	http://github.com/python/cpython/commit/a4ae82...
115	10102	2018	5b144559fbdba7ff673cc1c165aa2d343e07b6bd	edx-platform before 2018-07-18 allows XSS via ...	http://github.com/edx/edx-platform/commit/5b14...
116	13454	2018	5f18eeaaa459bee9a58f70cdf7c46adb1ef34ea7	templates/forms/thanks.html in Formspree befor...	http://github.com/formspree/formspree/commit/5...
117	14544	2018	aeb5b036a0bf657951756688b3c72bd68b6e4a7d	gui2/viewer/bookmarkmanager.py in Calibre 3.18...	http://github.com/kovidgoyal/calibre/commit/ae...
118	14713	2018	f8f7019ffdf9b4e05faf95e1f04e204aa4c91f98	io/mongo/parser.py in Eve (aka pyeve) before 0...	http://github.com/pyeve/eve/commit/f8f7019ffdf...

Cross Referencing NVD & SHGD Hashes cont…

- 4,695 parsed hashes from NVD
- 5,188,994 hashes in SHGD teaser
- 2.5% of NVD data has been matched to SHGD teaser
- Exploring new ways to match data and larger dataset
  - Exploring full 106GB dataset
    - Looking into Hashes, Dates, Version, and Names

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): GitHub commit hash
  - 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)

Questions?