
SBA
Research

PeekaTorrent

Leveraging P2P Hash Values for Digital Forensics

Sebastian Neuner, *Martin Schmiedecker*,
Edgar Weippl

SBA Research

Problem Description

We are drowning in data:

- processes and best-practices do not scale well
- 12TB hard drives recently presented
- sector hashing, unziping and unpacking, ...

SBA Research

Problem Description

SBA Research

Problem Description

We'd like to ignore:

SBA Research

Problem Description

Hashing is prevalent:

- DHTs, P2P file-sharing (SHA-1)
- Dropbox (4MB, SHA-256)
- file whitelisting (NSRL):
  - full file (SHA256, SHA1 & MD5)
  - fuzzy files (ssdeep, sdhash)
  - blocks (MD5b4096, MD5b8192)

SBA Research

Problem Description

Hashing is prevalent:

- DHTs, **P2P file-sharing (SHA-1)**
- Dropbox (4MB, SHA-256)
- file whitelisting (NSRL):
  - full file (SHA256, SHA1 & MD5)
  - fuzzy files (ssdeep, sdhash)
  - blocks (MD5b4096, MD5b8192)

SBA Research

Problem Description

Couldn’t we add to this:

- exclude commonly found files
- mostly totally irrelevant for investigation
- even before looking at files manually

SBA Research

peekaTorrent

SBA Research

peekaTorrent

peekaTorrent

General idea:

- leverage publicly shared hash values
- more granular than files, but less than sectors

- it's all in the *.torrent*
- copyright-free!

SBA Research

peekaTorrent

BitTorrent uses chunking:

- all files are concatenated
- then split in chunks (=*pieces*)
- most often 256kb, (observed 16kb-16mb)
- depending on implementation and user preference

SBA Research

peekaTorrent

Instead of hashing sectors, or files:

- variable hash windows (2^n)
- iterate over each sector
- build on *bulk_extractor*

Then pipe it all into *hashdb*, see what drops out

SBA Research

peekaTorrent

Benefits:

- also deleted & partially overwritten files
- fast!
- less false-positives
- hashdb files can be easily shared

SBA Research

peekaTorrent

Use cases:

- file whitelisting (torrents)
- file blacklisting
- custom hashsets: source code, email attachements, sharepoint, ...

SBA Research

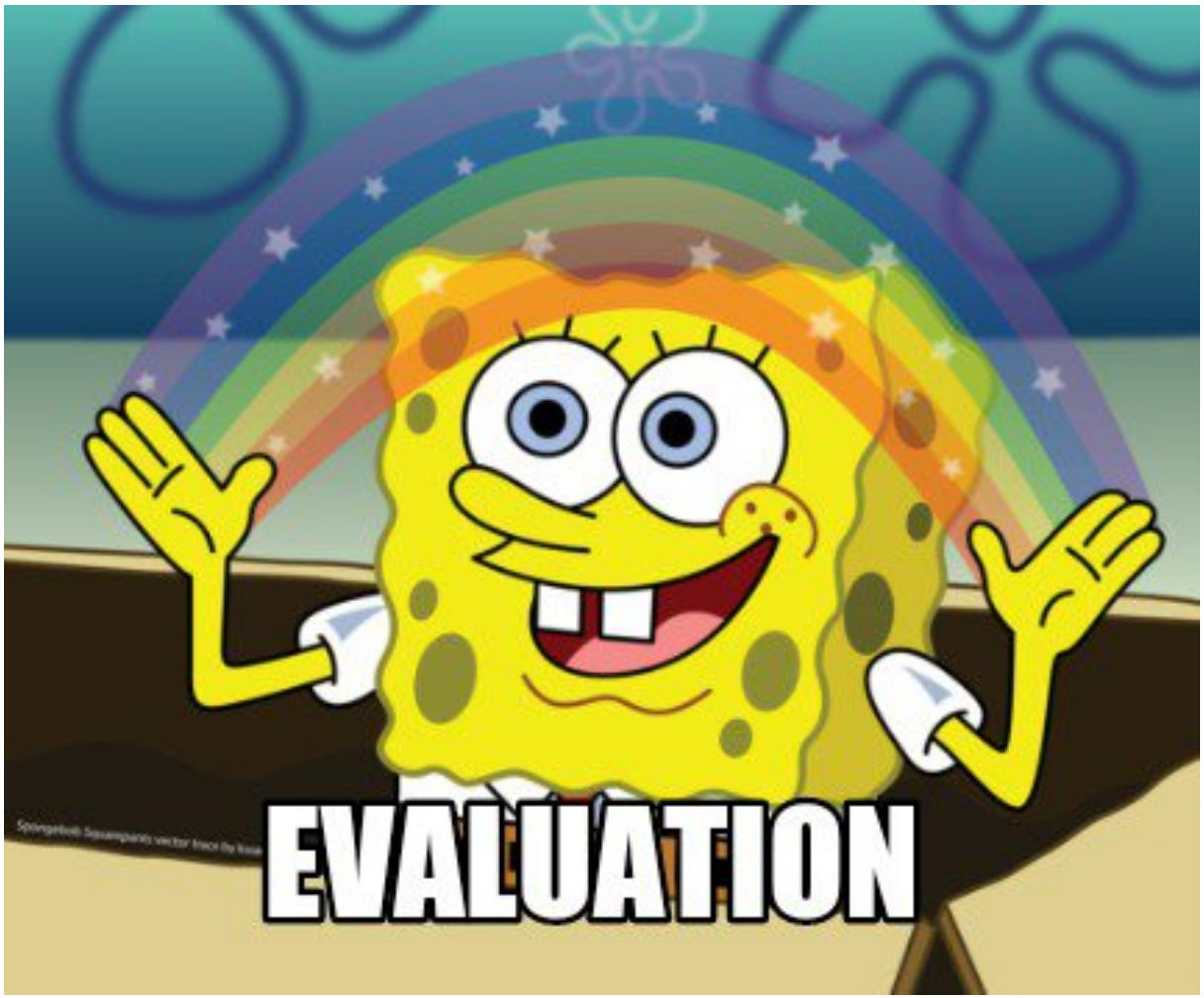
peekaTorrent

Simplistic use:

- create torrent with files of interest
- don’t publish/announce it
- pipe into hashdb, done

SBA Research

Evaluation



SBA Research

Evaluation

Collected data:

- in total: 2.65 million torrent files
- crawling Piratebay & KAT
- multiple data dumps

- 3.3 billion unique chunk hashes
- up to 2.6 PB of data

SBA Research

Evaluation

Some numbers:

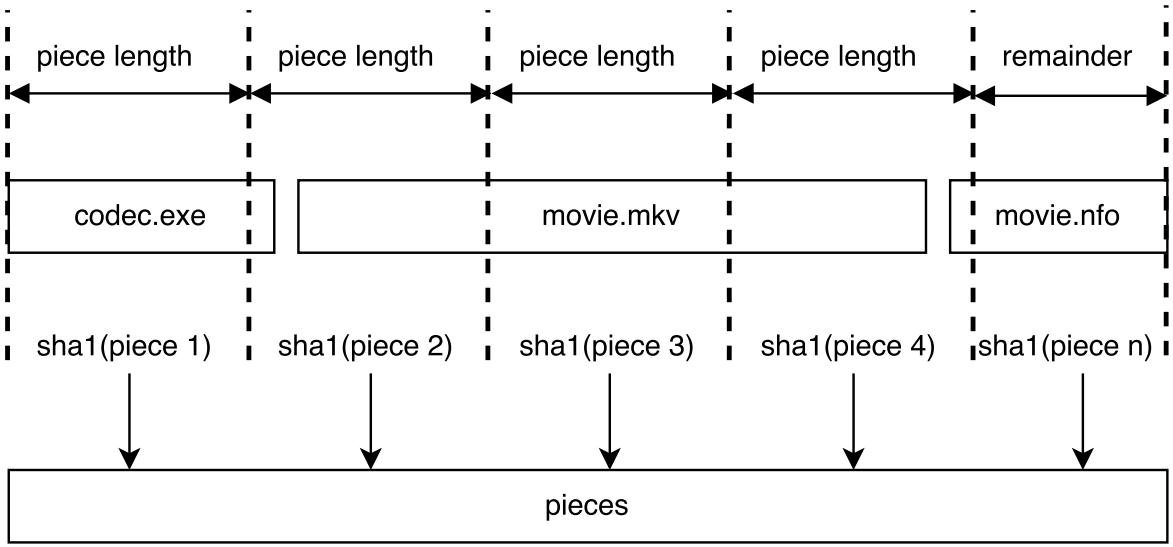
- 1 GB filesystem, Ubuntu Desktop = 1158 chunks
- running bulk_extractor: 220s (Notebook), 23s (Server)
- running hashdb: few seconds

SBA Research

Limitations

Non-usable data:

- chunks consisting of two files
- fragementation on disk



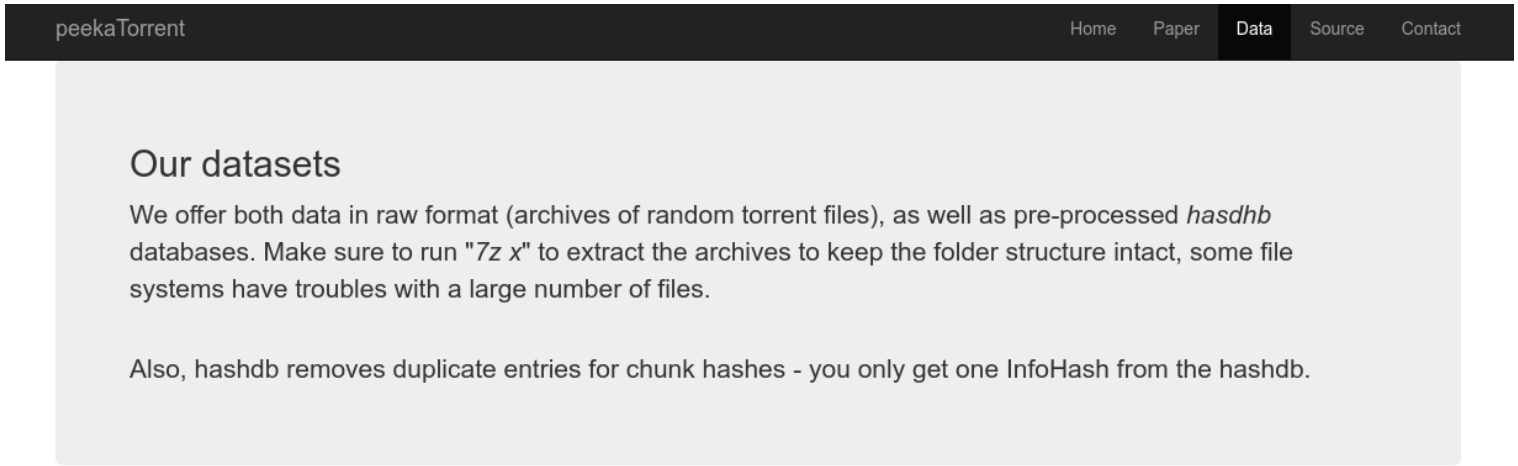
Future Work

What's needed:

- more .torrents!
- more data
- investigate data set more closely (duplicates)
- get feedback

SBA Research

Sharing is Caring



peekaTorrent

Home Paper Data Source Contact

Our datasets

We offer both data in raw format (archives of random torrent files), as well as pre-processed *hasdhb* databases. Make sure to run "*7z x*" to extract the archives to keep the folder structure intact, some file systems have troubles with a large number of files.

Also, hashdb removes duplicate entries for chunk hashes - you only get one InfoHash from the hashdb.

Torrent Archives

2.65 million torrent files 66 GB compressed, 84 GB raw

85,000 torrent files piece_length 16k, 1 GB compressed, 3 GB raw
100,000 torrent files piece_length 32k, 2.2 GB compressed, 3.7 GB raw
345,000 torrent files piece_length 64k, 3.9 GB compressed, 5.4 GB raw
213,000 torrent files piece_length 128k, 4.4 GB compressed, 5.4 GB raw
730,000 torrent files piece_length 256k, 25 GB compressed, 31 GB raw
318,000 torrent files piece_length 512k, 8.2 GB compressed, 9.5 GB raw
332,000 torrent files piece_length 1024k, 7.3 GB compressed, 8.8 GB raw
189,000 torrent files piece_length 2048k, 4.3 GB compressed, 5.3 GB raw
171,000 torrent files piece_length 4096k, 4.7 GB compressed, 6.1 GB raw

hashdb Datasets with InfoHash

50 mio hashes, piece_length 16k, 2.2 GB compressed, 5 GB raw
111 mio hashes, piece_length 32k, 4.7 GB compressed, 10 GB raw
192 mio hashes, piece_length 64k, 8.1 GB compressed, 15 GB raw
220 mio hashes, piece_length 128k, 9.1 GB compressed, 17 GB raw
1.2 bil hashes, piece_length 256k, 49 GB compressed, 81 GB raw
405 mio hashes, piece_length 512k, 17 GB compressed, 30 GB raw
350 mio hashes, piece_length 1024k, 15 GB compressed, 26 GB raw
205 mio hashes, piece_length 2048k, 8.6 GB compressed, 16 GB raw
226 mio hashes, piece_length 4096k, 9.4 GB compressed, 18 GB raw

SBA Research

Questions?

Thank you for your attention

URL: https://peekatorrent.org

Email: mschmiedecker@sba-research.org

Twitter: @fr333k

SBA Research