Identifying Almost Identical Files Using Context Triggered Piecewise Hashing

By

Jesse Kornblum

Presented At

The Digital Forensic Research Conference

DFRWS 2006 USA  Lafayette, IN (Aug 14th - 16th)

DFRWS is dedicated to the sharing of knowledge and ideas about digital forensics research. Ever since it organized the first open workshop devoted to digital forensics in 2001, DFRWS continues to bring academics and practitioners together in an informal environment. As a non-profit, volunteer organization, DFRWS sponsors technical working groups, annual conferences and challenges to help drive the direction of research and development.

http://dfrws.org
ManTech SMA

Computer Forensics and Intrusion Analysis

Fuzzy Hashing

Jesse Kornblum
Too Many Pictures

- Child Pornography cases
  - Hundreds of thousands of images
  - MD5 not effective for carved files
Cryptographic Hashing

- Algorithms like MD5, SHA-1
- Generate single hash for entire input
- Any change greatly alters hash

```
e41b1427a018fbb264c8adf0a
7f48e4b990a2d637fc363efc8
```
**Piecewise Hashing**

- Developed for integrity during imaging
- Divide input into equal sized sections and hash
- Insert or delete changes all subsequent hashes

![Diagram]

- 3b152e0baa367a8038373f6df
- 40c39f174a8756a2c266849b
- fdb05977978a8bc69ecc46ec
Rolling Hash

- Function triggered by current context of input
Fuzzy Hashing

- Piecewise hashing with boundaries defined by when rolling hash triggers

Signature: 11776
Fuzzy Hashing

- Changes only affect one small part of signature

New Signature: 11726
Original: 11776
Rolling Hash

To update the hash for a byte $d$:

$y = y - x$

$y = y + \text{size} \times d$

$x = x + d$

$x = x - \text{window}[c \mod \text{size}]$

$\text{window}[c \mod \text{size}] = d$

$c = c + 1$

$z = z \ll 5$

$z = z \text{ XOR } d$

\text{return } (x + y + z)$
Rolling Hash

- Choose triggers such that
  - `rolling_hash(d) mod block_size = block_size – 1`
  - Depends only on previous seven bytes

- Example
  - Excerpt from "The Raven" by Edgar Allan Poe
  - Based on file size, triggers on `ood` and `ore`
Rolling Hash

Deep into the darkness peering, long I stood there, wondering, fearing
Doubting, dreaming dreams no mortals ever dared to dream before;
But the silence was unbroken, and the stillness gave no token,
And the only word there spoken was the whispered word,
Lenore?, This I whispered, and an echo murmured back the word,
"Lenore!" Merely this, and nothing more.
Deep into the darkness peering, long I stood there, wondering, fearing
Doubting, dreaming dreams no mortals ever dared to dream before;
But the silence was unbroken, and the stillness gave no token,
And the only word there spoken was the whispered word,
Lenore?, This I whispered, and an echo murmured back the word,
"Lenore!" Merely this, and nothing more.
Deep into the darkness peering, long I stood
there, wondering, fearing Doubting, dreaming dreams no mortals
ever dared to dream before

; But the silence was unbroken, and the stillness gave no token,
And the only word there spoken was the whispered word, Lenore

? This I whispered, and an echo murmured back the word, "Lenore"

!" Merely this, and nothing more.
Demonstration
Demonstration

- Needle in a haystack

Known kitty porn MATCH
Demonstration

- No false positives

Known kitty porn

no match (00000380.JPG)
Known kitty porn MATCH
Demonstration

- File footers

Known kitty porn  MATCH
Issues

- Not perfect
  - Confused by many small changes throughout input
  - Unable to handle cropping, resizing, and other edits
- Computationally intensive
  - 7-10 times slower than MD5
- No way to sort signatures
  - Must compare each input to all known signatures
Future Research

- Need Error Rate Computation
  - I am a practitioner, not math geek
- For court, need error rate
  - How similar is similar?
Future Research

- File Footer Reconstruction
  - Record headers when making signatures
  - Append recovered footers
- Need to parse known files
  - How much information to record?
  - Best storage method?
Future Research

- File footer Reconstruction

Known kitty porn  File header with footer appended
Future Research

- Finding footers and middles
  - Current carvers require true footer
  - Encase, iLook, Foremost, Scalpel, etc.
- The formatted drive scenario
- Find blocks that are "JPEGy" or "GIFy"
  - Lots of academic research
  - No practical tools
ssdeep to be published August 14th
- Free software!
- http://ssdeep.sf.net/
Questions

Jesse Kornblum - ManTech CFIA
jesse.kornblum@mantech.com - 410-312-5548