Fuzzy Hashing

Jesse Kornblum
Introduction

- Interactive Presentation
  - The answer is always “it depends”
  - Ask me anything at any time

- Computer Forensics Engineer
  - Link between academics and practitioners

- Former AFOSI agent

- Tool Developer
  - foremost, md5deep, FRED
Outline

- Introduction
- The Joy of Hashing
- MD5 Completely Explained [Abridged]
- Piecewise Hashing
- Rolling Hash
- Fuzzy Hashing
- Matching
- Demonstration
- Issues
- Future Research
The Joy of Hashing

- Hashing reduces any input to a fixed size output
- Lots of different hashing algorithms
  - Most are not suited for forensics
- We use cryptographic hashing algorithms
  - MD5, SHA-1, Whirlpool, Tiger
  - These are great for finding identical things
  - Can’t be used to find similar things
What is “Similar”

- This is a hard question
- “How are you” vs. “How are y0u”
- Blade Runner vs. The Fifth Element
  - Plot
  - Themes
  - Motifs (e.g. Messages at lunch counter)
  - Streams of ones and zeros
Streams of Ones and Zeros

- For our purposes, searching for file homologies
  - Files with large sequences of identical ones and zeros
- Two Microsoft Word Documents with different metadata

```
C:\>md5deep -b *.doc
63bb20b0df871ae390af2af0b0a33248  BusinessPlan.doc
618e7bf2232d7f29406246c692e4dd10  StolenPlan.doc
```
Disclaimer

- I didn’t invent this math
- Originally Dr. Andrew Tridgell
  - Samba
  - rsync was part of his thesis
  - Modified slightly for spamsum
    - Spam detector in his “junk code” folder
- Discovered by G.
  - User report that rsync confuses similar Word documents
MD5 Explained

How MD5 (roughly) works:

3. Start with an initial state
4. Look at fixed size block of input
   - Do mathy stuff with current state and block
   - Get new state
5. Advance to next block of input
6. Repeat steps 2 and 3 until out of input blocks
7. Ending state is the hash
MD5 Explained

- If you change one bit in the middle, you change the next state
- Which ends up changing the end result
- Is this a good thing or a bad thing?
**Piecewise Hashing**

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

![Diagram showing hash values]

- `3b152e0bba367a8038373f6df`
- `40c39f174a8756a2c266849b`
- `fdb05977978a8bc69ecc46ec`
Rolling Hash

- It would be nice to set boundaries such that
  - Insertions and deletions are contained within a block
Rolling Hash

- A different kind of hash function
- Produces a pseudorandom output for every position in a file
  - Depends only on last few bytes
  - Lots of academic work on these
  - Just mathy tricks

Four score -> 83,742,221
Four score -> 5
Four score -> 90,281
To update state \((c, x, y, z, window)\) for a byte \(d\):

\[
\begin{align*}
y &= y - x \\
y &= y + \text{size} \times d \\
x &= x + d \\
x &= x - \text{window}[c \mod \text{size}] \\
\text{window}[c \mod \text{size}] &= d \\
c &= c + 1 \\
z &= z \ll 5 \\
z &= z \text{ XOR } d \\
\text{return } (x + y + z)
\end{align*}
\]
Rolling Hash

- We use the rolling hash to generate block boundaries
- Select some values as trigger points
- When we hit a trigger point, end the block

Example
- Excerpt from "The Raven" by Edgar Allan Poe
- Triggers on ood and ore
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
Rolling Hash

- How do we choose the triggers?
  - Chosen randomly, before reading the file
  - Based on the size of the input file
  - Really just a set of numbers
  - Has nothing to do with type of input data
Fuzzy Hashing

- Combine Rolling Hash with a Traditional Hash
- Use Fowler/Noll/Vo (FNV) hash
  - That’s what Tridgell did
  - Faster and less complex than MD5
  - We’re only using a small part of the result

- Start reading file, compute Rolling and Traditional Hashes
- When Rolling Hash triggers
  - Record LSB of Traditional Hash value
- When finished, combine LSBs to make signature
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 I AM THE LIZARD KING! 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
Matching

Signature 1: 3 2 7 3 0
Signature 2: 3 0 7 3 0

- **Edit Distance**
  - Number of insertions, modifications and deletions to turn Signature 1 into Signature 2.
  - For the example above, the edit distance is one.

- **Signatures (and thus files) match when the ratio of the edit distance to the length is small**
Demonstration

- Word Documents
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
WARNING: EXPLICIT IMAGERY
Demonstration
Demonstration

- Needle in a haystack

Known kitty porn MATCH
Demonstration

- Does not match similar looking images

Known kitty porn

no match
(00000380.JPG)
Demonstration

- File header

Known kitty porn  MATCH
Demonstration

- File footer

Known kitty porn    MATCH
Demonstration

- Source code reuse
Issues

- Does not work for similar looking graphics
  - For that use imgSeek
    http://www.imgseek.net/
- 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
Issues

- Matching via Fuzzy Hashing is not proof!
  - Similar does not mean identical
- An excellent basis for further examination
  - Probable cause

“I ran a program called “ssdeep” that uses Context Triggered Piecewise Hashing (aka “Fuzzy Hashing”) on SUBJECT’s hard drive and a set of signatures of known child pornography. The program identified 417 files on SUBJECT’s hard drive that matched the known images of child pornography. The output of the program, demonstrates that a significant majority if not the entirety of each matching file is identical to the known child pornography image that it is said to match.

Based on the presence of child pornography on SUBJECT’s hard drive, the Government requests to search SUBJECT’s primary residence...”
Future Research

- Need hashes of known files
  - Hash sets like NSRL or Hashkeeper
  - How much information to record?
    - File size, content?

- File Footer Reconstruction
  - Record headers when making signatures
  - Append recovered footers
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.
- Find blocks that are "JPEGy" or "GIFy"
  - Lots of academic research
  - No practical tools
- What makes things similar?
  - Lots of research to date
  - Most focuses on plagiarism by CS students
Fuzzy Hashing allows examiners to find documents that are similar but not quite identical.

- ssdeep is Free and Open Source
  - [http://ssdeep.sf.net/](http://ssdeep.sf.net/)
  - Windows executable and source code
  - Graphical Front End now available!

Questions? Write to me at jesse.kornblum@mantech.com.  
- I am always available for LE
Questions

Jesse Kornblum
jesse.kornblum@mantech.com