ManTech International

Computer Forensics and Intrusion Analysis

Fuzzy Hashing

Jesse Kornblum

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

Overview

- A Reasonable Scenario
- MD5 Explained
- Where MD5 falls down
- Similarity
- Fuzzy Hashing
- Examples
- Questions

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

A Reasonable Scenario

- Disgruntled employee leaves BigCompany
  - Joins new startup in the same field
- New company introduces similar product
  - Using detailed information from Big Company
  - One particular document
  - Not publicly available
- BigCompany sues for trade secrets violation
- New company has eight employees

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

A Reasonable Scenario

- 100GB per employee hard drive, plus 200GB on server
- 1TB of data total
  - 536,000 reams of paper
  - About 29 semi-trucks

Photo courtesy sumsinnow via Flickr

MD5 Explained

How MD5 (roughly) works:

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

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

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?

?

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

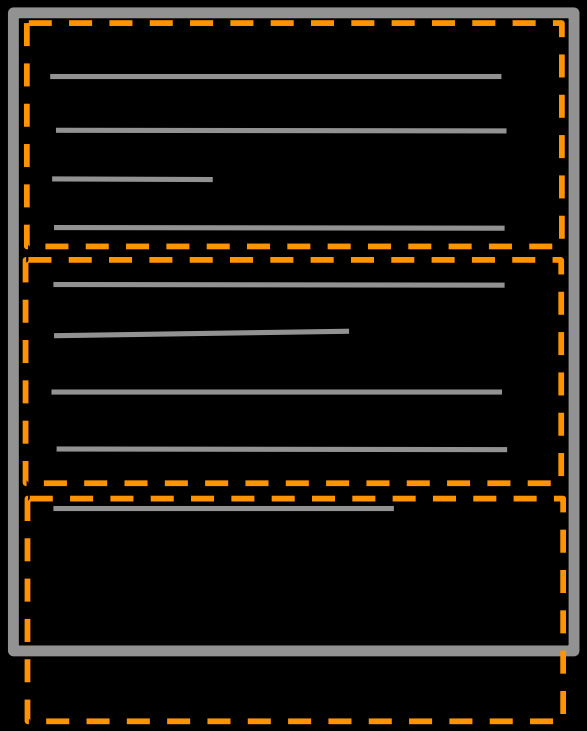
Similarity

- Different levels of similarity
  - Identical
  - Ones and zeros
  - Displays the same
  - Behaves the same
  - Thematically similar
  - Not similar

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

Piecewise Hashing

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



3b152e0baa367a8038373f6df

40c39f174a8756a2c266849b

fdb05977978a8bc69ecc46ec

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

Rolling Hash

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

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

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
- User report that rsync confuses similar Word documents

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

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

```
F o u r   s c o r e   ->  83,742,221

F o u r   s c o r e   ->       5

F o u r   s c o r e   ->    90,281
```

Rolling Hash

To update state (c,x,y,z,window) for a byte d:

```
y = y − x
y = y + size * d
x = x + d
x = x − window[c mod size]
window[c mod size] = d
c = c + 1
z = z  << 5
z = z  XOR d
return (x + y + z)
```

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

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

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

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

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

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

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

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

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

Rolling Hash

Deep into the darkness peering, long I stood 28163

there, wondering, fearing Doubting, dreaming dreams no mortals
ever dared to dream before 491522

; But the silence was unbroken, and the stillness gave no token,

And the only word there spoken was the whispered word,Lenore 57

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

145410213

!" Merely this, and nothing more

738210

Rolling Hash

Deep into the darkness peering, long I stood 28163

there, wondering, fearing Doubting, dreaming dreams no mortals ever dared to dream before 491522

; But the silence was unbroken, and the stillness gave no token,

And the only word there spoken was the whispered word,Lenore 57

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

145410213

!" Merely this, and nothing more

738210

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

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

Rolling Hash

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

Rolling Hash

Deep into the darkness peering, long I stood — 28163

there, wondering, fearing I AM THE LIZARD KING! Doubting,
dreaming dreams no mortals ever dared to dream before — 82910

; But the silence was unbroken, and the stillness gave no token,

And the only word there spoken was the whispered word,Lenore — 57

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

145410213

!" Merely this, and nothing more

738210

Rolling Hash

Deep into the darkness peering, long I stood — 28163

there, wondering, fearing I AM THE LIZARD KING! Doubting,
dreaming dreams no mortals ever dared to dream before — 82910

; But the silence was unbroken, and the stillness gave no token,

And the only word there spoken was the whispered word,Lenore — 57

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

145410213

!" Merely this, and nothing more

738210

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**

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

Demonstration

LAW ENFORCEMENT SENSITIVE

DO NOT DUPLICATE

WARNING:
EXPLICIT IMAGERY

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

Demonstration

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

Demonstration

Corrupted File

Known kitty porn

MATCH

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

Demonstration

Different File

Known kitty porn

No match

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

Demonstration

File Header

Known kitty porn

MATCH

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

Demonstration

File Footer

Known kitty porn

MATCH

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

Demonstration

File Footer (attached to header)

Known kitty porn

MATCH

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

Issues

- Does not work for similar looking graphics
- Unable to handle cropping, resizing, and other edits
- Confused by many small changes throughout input
- Computationally intensive
  - 7-10 times slower than MD5
- No way to sort signatures
  - Must compare each input to all known signatures

Fuzzy Hashing

- Matches similar but not identical bitstreams
  - Great for corrupted or partial documents
  - Also great for source code reuse
- Free Software
  - http://ssdeep.sf.net/
  - Windows, GUI, *nix, and OS X
  - Paper in *Digital Investigation*
    - http://dfrws.org/2006/proceedings/12-Kornblum.pdf

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

Future Research

"That algorithm is our last hope."

"No, there is another."

MANTECH SMA
COMPUTER FORENSICS & INTRUSION ANALYSIS

Questions

Jesse Kornblum

jesse.kornblum@mantech.com