Practical use of Approximate Hash Based Matching in digital investigations

Petter Christian Bjelland, Katrin Franke, André Årnes

2014.05.07
About me

- MSc. student at Gjøvik University College, Norway
  - Information security, specialization in digital forensics
  - Currently writing my master thesis
- Software developer
Agenda

- Motivation and summary of contributions
- Approximate matching (Fuzzy hashing) - What is it and why do we care?
- What is similarity?
- Three modes of Approximate Hash Based Matching (AHBM)
- Practical scenarios with AHBM
- Open research questions
Motivation

- What we can use approximate matching for?
Summary of contributions (1/4)

- **Paper:** Exploration of modus operandi for AHBM in digital investigations.
Summary of contributions (2/4)

- **Tool:** sddiff
  - Visualize similarity between two files.
  - Based on the feature selection algorithm in sdhash.
  - Find positions of fragments from the smaller file within the larger file.
  - https://github.com/pcbje/sddiff
Summary of contributions (3/4)

- **Tool**: Autopsy AHBM
  - The Sleuth Kit Autopsy 3 module for doing approximate matching.
  - 2nd place in Basistech 2013 Autopsy module development contest
    - (Yes, out of two participants)
  - [http://github.com/pcbje/autopsy-ahbm](http://github.com/pcbje/autopsy-ahbm)
  - Video presentation: [http://youtu.be/GBmZRufH_3o](http://youtu.be/GBmZRufH_3o)
Summary of contributions (4/4)

- **Tool:** makecluster
  - Split network into disjunct clusters. Handy when dealing with large, sparse networks.
  - [https://github.com/pcbje/makecluster](https://github.com/pcbje/makecluster)
Approximate Matching
Approximate Matching

- Techniques for the identification of similar data
  - Updated documents
  - Fragments of files in memory or hard drives
  - Pictures
  - Videos
  - +++

- Degree of similarity

- Useful when cryptographic hashes aren’t well suited
What is similarity?
What is similarity?

- With documents, there are two types of similarity we can measure:
  - Semantic similarity
  - Syntactic similarity
Semantic similarity

- Similarity from the perspective of humans
- Two documents are semantically identical if they communicate the same meaning
  - Different formats of the same document
Semantic similarity

Merzouga, Morocco
Syntactic similarity

- Similarity from the perspective of computers.
- Two documents are syntactically identical if have the same binary representation.
- Generally not suited for matching media files like pictures and videos.
  - Exception: Fragments of deleted pictures on a hard drive.
Syntactic similarity

- The brown **fox** jumped over the lazy dog.
- The brown **cat** jumped over the lazy dog.
Modes of Approximate Hash Based Matching
Modes of Approximate Hash Based Matching

- Searching
- Streaming
- Clustering
Mode: Searching

- Small input, large search space.
- Typically a query you perform once when you need it.
Mode: Streaming

- Large input, small search space.
- Typically a continuous query.
Mode: Clustering

- Large input, large search space.
- Organize and find patterns in unknown data.

![Diagram](image-url)
Practical scenarios with AHBM
Practical scenarios with AHBM

- Alternative versions detection
- File transfer detection
- Organization of documents
Alternative version detection

- We want to identify documents that are similar to one we have found.
- E.g. revisions of a PDF document or an email thread.
File transfer detection

- Detect when traces of some specified data is transferred over the network.
- E.g. download of specific software packages.

<table>
<thead>
<tr>
<th>Role</th>
<th>Filetype</th>
<th>Similarity score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Installer</td>
<td>Executable</td>
<td>31</td>
</tr>
<tr>
<td>Installed tool</td>
<td>Executable</td>
<td>1</td>
</tr>
<tr>
<td>User guide</td>
<td>PDF</td>
<td>2</td>
</tr>
<tr>
<td>Tool formatter</td>
<td>Executable</td>
<td>2</td>
</tr>
</tbody>
</table>
Organization of documents

- We are faced with a large, unknown corpus of data.
- We want to group together documents based on similarity.
Organization of malicious PDFs with sdhash

<table>
<thead>
<tr>
<th>Document 1</th>
<th>Document 2</th>
<th>Similarity score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>027</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>010</td>
</tr>
<tr>
<td>B</td>
<td>C</td>
<td>001</td>
</tr>
</tbody>
</table>
Organization of malicious PDFs with sdhash

<table>
<thead>
<tr>
<th>Document 1</th>
<th>Document 2</th>
<th>Similarity score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>027</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>010</td>
</tr>
<tr>
<td>B</td>
<td>C</td>
<td>001</td>
</tr>
</tbody>
</table>

Position of fragments of B in A:
Organization of malicious PDFs with sdhash

<table>
<thead>
<tr>
<th>Document 1</th>
<th>Document 2</th>
<th>Similarity score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>027</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>010</td>
</tr>
<tr>
<td>B</td>
<td>C</td>
<td>001</td>
</tr>
</tbody>
</table>

Position of fragments of C in A:
Organization of malicious PDFs with sdhash

<table>
<thead>
<tr>
<th>Document 1</th>
<th>Document 2</th>
<th>Similarity score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>027</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>010</td>
</tr>
<tr>
<td>B</td>
<td>C</td>
<td>001</td>
</tr>
</tbody>
</table>

Position of fragments of C in B:

```
[ ] [ ] [ ]
```
Open research questions
Open research questions (1/2)

- How can we improve approximate matching efficiency?
Open research questions (2/2)

- Can approximate matching be used to detect variations of known malware?
Conclusions and resources

- Three modes of approximate matching
  - Searching (ad-hoc query)
  - Streaming (continuous query)
  - Clustering (organization of data)

- Open source resources:
  - [http://sdhash.org](http://sdhash.org) (syntactic matching tool)
  - [http://phash.org](http://phash.org) (semantic matching tool)
  - [http://gephi.org](http://gephi.org) (graph visualization tool)
  - [https://github.com/pcbje](https://github.com/pcbje) (stuff)

- Contact:
  - petter.bjelland@hig.no