Mencari Jerami di antara Tumpukan Jarum dengan Fuzzy Hash

Beberapa saat yang lalu, ketika sedang menunggu rekan-rekan praktisi forensik digital melakukan akuisisi pada beberapa perangkat komunikasi, mata penulis terpaku pada sebuah poster SANS forensic yang berisi teknik identifikasi *malware*. Dari beberapa teknik identifikasi tersebut, terdapat salah satu teknik untuk menghitung kesamaan dari dua buah *malware* yang berbeda. Penulis bertanya di dalam hati, mengapa teknik tersebut tidak digunakan untuk mencari dokumen elektronik yang memiliki kesamaan. Pertanyaan tersebut tidak menuai jawaban hari itu….karena penulis bertanya di dalam hati.

Pengantar

Sebagai praktisi forensik digital, terutama yang berfokus di e-discovery, kegiatan menemukan file sudah menjadi makanan sehari-hari. Namun adakalanya diperlukan juga bagi praktisi forensik digital untuk menemukan file yang serupa dengan file yang sudah ditemukan (file yang sudah di-*update,* di-*edit,* di-*save as* dsb). Fuzzy hash dapat digunakan sebagai salah satu alternative untuk menyelesaikan permasalahan tersebut.

Hash dan fuzzy hash

Hash function adalah fungsi matematis yang digunakan untuk "merangkum" data elektronik yang berjumlah berapa pun menjadi jumlah tetap yang disebut dengan *hash values*. Data elektronik yang berbeda kontennya, akan menghasilkan *hash value* yang berbeda. Konsep ini yang menjadi kunci pentingnya penggunaan *hash function* untuk menjamin keaslian dari sebuah data elektronik.

Sedikit berbeda dengan *hash function* tradisional, *fuzzy hash* "memotong-motong" bagian dari sebuah dokumen elektronik, dan kemudian menggunakan teknik *hash* tradisional untuk menghitung *hash values* dari masing-masing potongan untuk kemudian dibandingkan.

Bagi praktisi forensik digital yang berfokus di bidang *malware analysis,* konsep *fuzzy hash* tersebut sering digunakan untuk menemukan kekerabatan dari berbagai *malware*. *Malware* yang beredar di "alam liar", dan memiliki *source code* yang serupa, sehingga dengan menggunakan *fuzzy hash* dapat dipetakan *malware* yang mirip dan berkerabat. Sedangkan, di bidang e-discovery, ternyata teknik fuzzy hash seringkali dipakai di kasus *copyright infringement.*

Ssdeep

ssdeep merupakan implementasi dari konsep *fuzzy hash*. Untuk penggunaan dalam e-discovery, khususnya menemukan file-file yang memiliki kemiripan, dapat digunakan beberapa teknik ssdeep berikut:

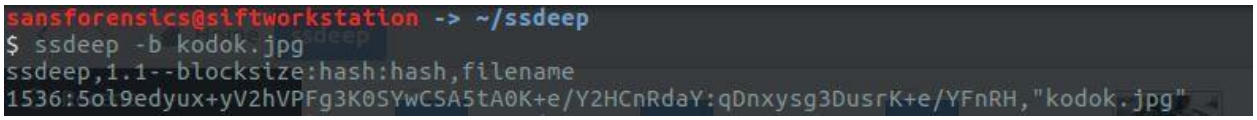
1. Menghitung ssdeep hash.

   untuk menghitung ssdeep hash, dapat digunakan perintah:

   ```
   ssdeep -b nama_file
   ```

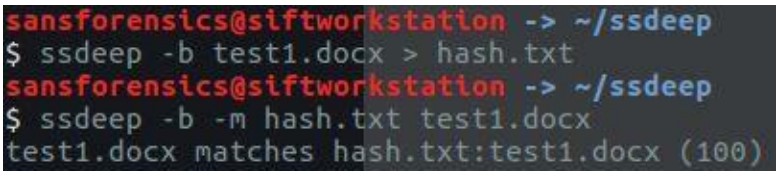
   atau apabila hasil hash tersebut akan disimpan ke dalam sebuah file:

   ```
   ssdeep -b nama_file > hash.txt
   ```



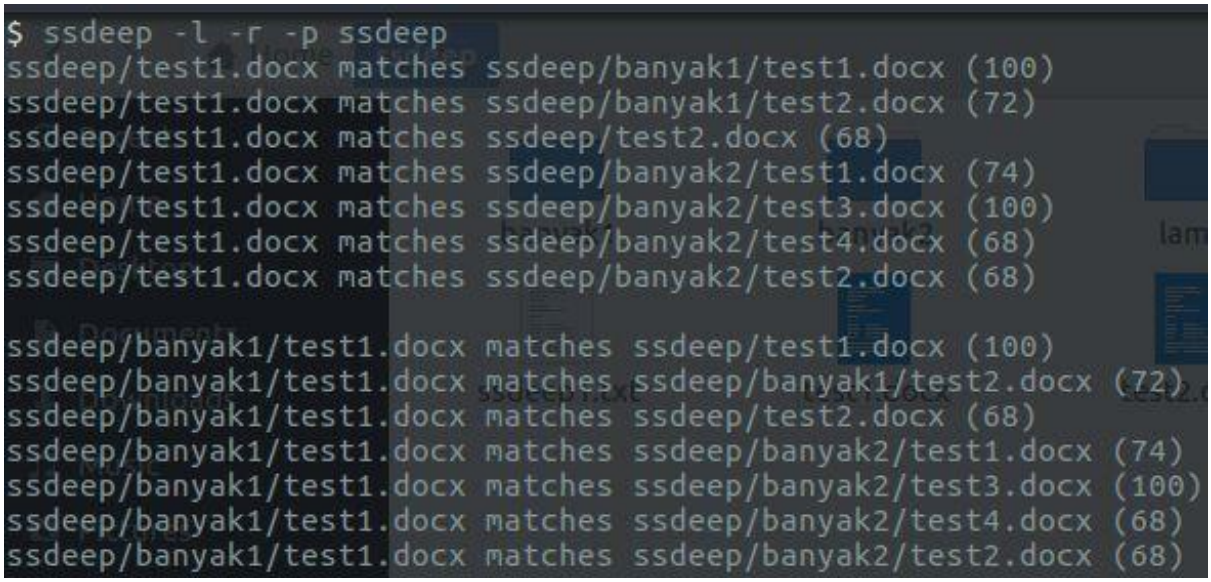
2. Membandingkan hash yang disimpan dalam file dengan:

   ```
   ssdeep -b -m hash.txt nama_file
   ```



3. Apabila anda pemalas kelas kakap, dapat dilakukan perbandingan langsung seluruh folder secara *recursive* dengan

   ```
   ssdeep -l -r -p nama_directory
   ```



keterangan:

1. angka di dalam kurung merupakan persentase kecocokan
2. flags:
   -b: tidak perlu tampilkan *path*
   -m: tandingkan dengan *hash*
   -l: gunakan *path relative*
   -r: *recursive*
   -p: tampilkan semua yang sama

Kekurangan

Selain itu presentasi kemiripan kedua belah file seringkali tidak mencerminkan kecocokan yang tepat, hal ini dikarenakan sehingga terdapat kemungkinan terdapat potongan yang memiliki *hash* yang sama, walaupun sesungguhnya kedua file asal *hash* tersebut tidak berkerabat dekat. Contohnya adalah *header* dari sebuah *filetype.*

To do

Dikarenakan cara penyimpanan data yang berbeda pada file-file media (gambar, film), terdapat beberapa kemungkinan pengembangan yang dapat dibahas di masa depan, antara lain menggunakan ssdeep untuk mencari file gambar yang mirip, ataupun media lain seperti film atau lagu.