FbHash: A New Similarity Hashing Scheme for Digital Forensics

Timotej Knez
63..

Sebastian Mežnar
27192031

Jasmina Pegan
63170423

POVZETEK
nek povzetek

Kategorija in opis področja
E.3 [**Data encryption**]

Splošni izrazi
Hashing

Ključne besede
Data fingerprinting, Similarity digests, Fuzzy hashing, TF-IDF, Cosine-similarity

1. UVOD
Živimo v obdobju shranjevanja ogromnih količin podatkov. Pri forenzičnih preiskavah se pogosto zgodi, da je pridobljenih datotek preveč za ročno pregledovanje. Digitalni forenziki se tako soočijo s problemom avtomatizacije preiskave datotek. Možna rešitev so algoritmi, kot so `ssdeep`, `sdhash` in `FbHash`, ki poskusijo filtrirati vnaprej znane "slabe" oziroma "dobre" datoteke. Ti algoritmi (angl. *Approximate Matching algorithms*) ugotavljajo delež ujemanja datotek s pomočjo (nekriptografskih) zgoščevalnih funkcij. Algoritma `ssdeep` in `sdhash` lahko preslepi aktivni napadalec, ki pametno napravi majhne spremembe na določenih mestih datoteke. Učinkovitega napada na algoritem `fbhash` ne poznamo.[3]

V 2. poglavju predstavimo predhodnike algoritma `FbHash`. V 3. poglavju podrobneje predstavimo algoritem `FbHash` in našo implementacijo. V 4. poglavju opišemo izvedene eksperimente in v 5. poglavju opišemo rezultate. V 6. poglavju povzamemo narejeno delo in rezultate.

2. SORODNA DELA
Prvi algoritem, namenjen iskanju približnih ujemanj, je bil objavljen leta 2002 pod imenom `dcfldd`. Ta algoritem je razvil N. Harbour kot izboljšano verzijo ukaza `dd`[4]. Izboljšana različica tega algoritma je`ssdeep`. Pomembnejša predhodnika algoritma `FbHash` sta tudi `mvHash-B` in `mrsh-v2`.

2.1 ssdeep
Algoritem `ssdeep` je implementacija kontekstno sprožene kosovno zgoščevalne funkcije (angl. *Context Triggered Piecewise Hash*, CTPH), ki jo je predstavil J. Kornblum septembra 2006 v članku [5]. Algoritem temelji na detektorju neželene elektronske pošte `spamsum`, ki lahko zazna sporočila, ki so podobna znanim neželenim sporočilom.

CTPH uporablja zgoščevanje po kosih (angl. *piecewise hashing*), kar pomeni, da se zgoščena vrednost izračuna na posameznih kosih fiksne dolžine. Za razliko od algoritma `dcfldd` CTPH uporabi poljubno zgoščevalno funkcijo.

Zgoščevalna funkcija z drsečim oknom (angl. *rolling hash*) preslika zadnjih nekaj zlogov (bajtov) v psevdonaključno vrednost. Vsakega naslednika je možno hitro izračunati iz predhodno izračunane vrednosti.

Postopek CTPH se začne z izračunom zgoščenih vrednosti z drsečim oknom. Ob določeni sprožilni zgoščeni vrednosti (angl. *trigger value*) se vzporedno s tem sproži še algoritem zgoščevanja po kosih. Ob ponovni pojavitvi sprožilne vrednosti se dotlej zbrane vrednosti druge zgoščevalne funkcije zapišejo v končni prstni odtis. Tako se ob lokalni spremembi v datoteki sprememba pozna le lokalno tudi v prstnem odtisu.

Sledi primerjava prstnih odtisov datotek, ki temelji na uteženi Levenstheinovi razdalji (angl. *edit distance*), ki je nato še skalirana in obrnjena, da predstavlja 0 povsem različna prstna odtisa.

Algoritem `ssdeep`, ki je implementacija CTPH, se izkaže pri primerjavi podobnih besedilnih datotek in dokumentov [5]. Po drugi strani pa lahko aktivni napadalec popravi "slabe" datoteke na tak način, da se izognejo črni listi [3].

2.2 sdhash
Nekineki [6]

Po drugi strani pa lahko aktivni napadalec spremeni "slabe" datoteke na tak način, da se izognejo črni listi oziroma "dobre" datoteke tako, da se obdržijo na beli listi [**?**].

2.3 mvHash-B

Nekineki [1]

2.4 mrsh-v2

Nekineki [2]

3. ALGORITEM

4. NAŠI EKSPERIMENTI (NAME IN PROGRESS)

5. REZULTATI

6. ZAKLJUČEK

7. ZAHVALA

Mogoče zahvala avtorjem za narjeno delo al kej.

8. REFERENCES

[1] F. Breitinger, K. P. Astebøl, H. Baier, and C. Busch. mvhash-b - a new approach for similarity preserving hashing. In *2013 Seventh International Conference on IT Security Incident Management and IT Forensics*, pages 33–44, March 2013.

[2] F. Breitinger and H. Baier. Similarity preserving hashing: Eligible properties and a new algorithm mrsh-v2. In *Digital Forensics and Cyber Crime,*, pages 167–182, October 2013.

[3] D. Chang, M. Ghosh, S. K. Sanadhya, M. Singh, and D. R. White. Fbhash: A new similarity hashing scheme for digital forensics. In *The Digital Forensic Research Conference*, volume 29, pages S113–S123. DFRWS, July 2019.

[4] N. Harbour. Dcfldd. defense computer forensics lab. *online*, 2002.

[5] J. Kornblum. Identifying almost identical files using context triggered piecewise hashing. *Digital Investigation*, 3:91–97, September 2006. The Proceedings of the 6th Annual Digital Forensic Research Workshop (DFRWS '06).

[6] V. Roussev. Data fingerprinting with similarity digests. *IFIP Advances in Information and Communication Technology*, 337:207–226, September 2010. Advances in Digital Forensics VI. DigitalForensics.