Monthly Research

アンチウイルスの検知率の一例と未検知検体の類似度

株式会社ＦＦＲＩ
http://www.ffri.jp

Ver 2.00.01

アジェンダ

- 背景と目的
- 調査対象のマルウェア検体
- マルウェア対策研究用のデータセット FFRI Dataset
- アンチウイルスの検知率の一例
- 検知率に関する考察
- Fuzzy hashing による検体の類似度調査
- まとめ

背景と目的

- 2014年5月とあるアンチウイルスベンダーの幹部による「アンチウイルスソフトは死んだ」という発言が話題になった

- 本リサーチでは、アンチウイルスの検知率の現状について調査し、一例として示す

- 検知率を向上させる手法を検討するために、検知できなかった検体の類似度について Fuzzy hashing を用いて調査した

調査対象のマルウェア検体

- 性質
  - 一般的なアンチウイルスで検出すべきと考えられる既知のマルウェア
  - アンチウイルスで検知困難とされる標的型や未知のマルウェアなどではない
- 収集期間
  - 2014年1月から4月
- 件数
  - 3,000 件
- 検知率の調査時期
  - 2014年4月24日および2014年7月7日

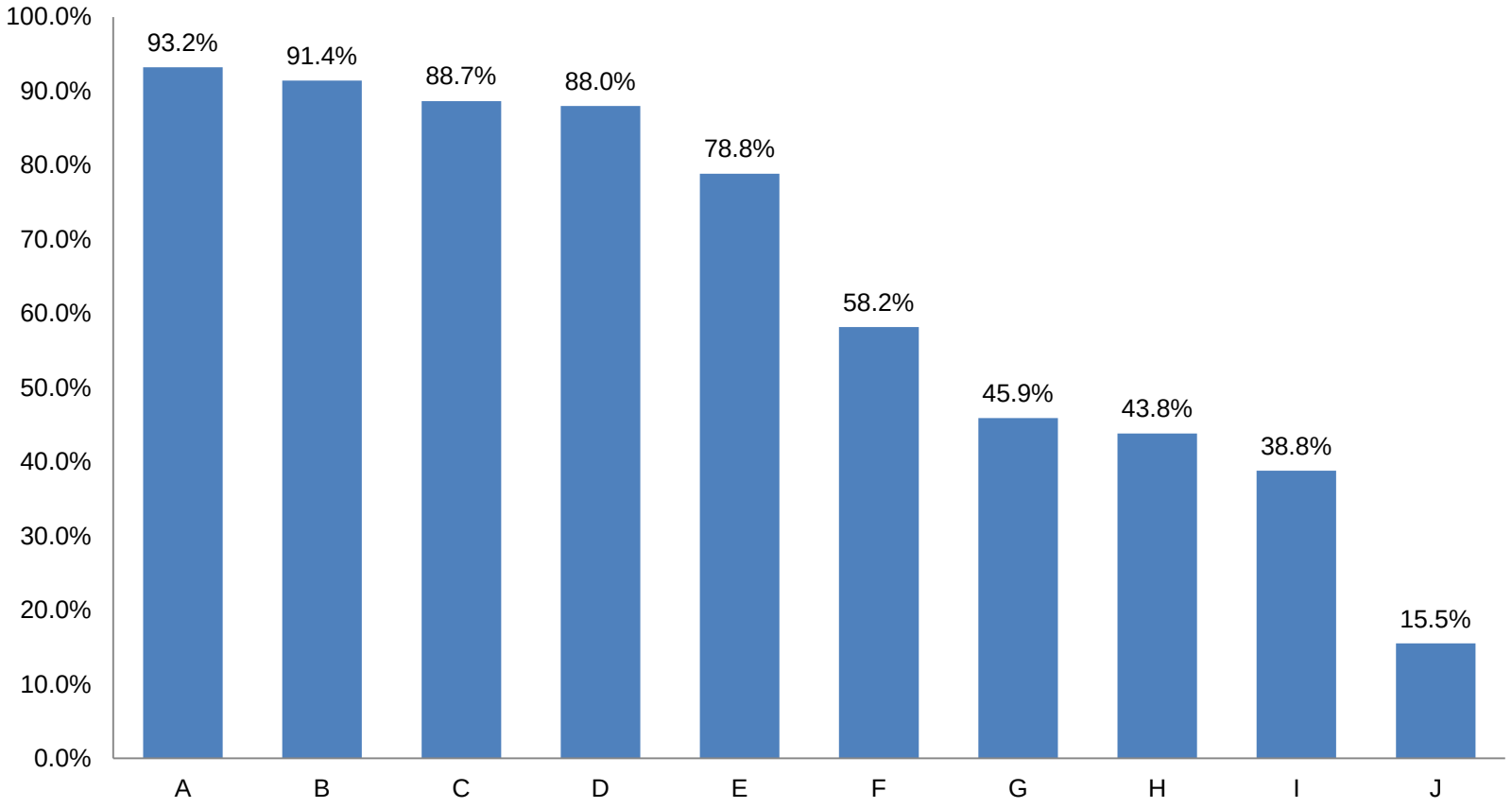
マルウェア対策研究用のデータセット FFRI Dataset

- FFRI はマルウェア対策の研究のためのデータセット FFRI Dataset を MWS というワークショップに提供している
  - MWS(マルウェア対策研究人材育成ワークショップ)
    - http://www.iwsec.org/mws/2014/

- FFRI Dataset 2014 の概要
  - 前述のマルウェア検体 3,000 件の動的解析ログ(マルウェアの動作情報)
  - 動的解析は Cuckoo Sandbox と FFR yarai analyzer Professional で実施
  - 具体的なデータ項目は下記参照
    - http://www.iwsec.org/mws/2014/files/FFRI_Dataset_2014.pdf

アンチウイルスの検知率の一例

- FFRI Dataset 2014 のマルウェア 3,000 件の検知率を調査



検知率に関する考察と留意点

- 考察
  - 予想以上に検知率に差があった
  - 2ヶ月後に検知率を調査したがほとんど変化がなかった
  - レピュテーションベース検知ロジックよりもパターンマッチによるジェネリック検知を中心にしているベンダのほうが検知率が高い傾向にあった
  - 無償製品の検知率は低かった
- 留意点
  - 本結果はあくまで研究用サンプルに対する検知率であり、アンチウイルスの包括的な性能の評価結果ではない
  - データセットを作成した2014年4月下旬の静的スキャンのみの検知結果
    - 静的スキャン以外のロジックで検知・防御する製品もある
    - より現実的な性能評価には、サンプル数を増やしたり、動的な検知率の調査も必要

Fuzzy hashing による検体の類似度調査

- Fuzzy hashing は類似ファイルの特定に利用できる技術である
  - これを用いることでバイナリ的に似ているマルウェアを特定できる
  - 概要については、2014年3月の Monthly Research 参照

- Fuzzy hashing を用いて以下を調査した
  - 未検知だった検体同士がどの程度類似しているか？
  - 未検知検体群に、検知できた検体と類似しているものがどの程度あるか？

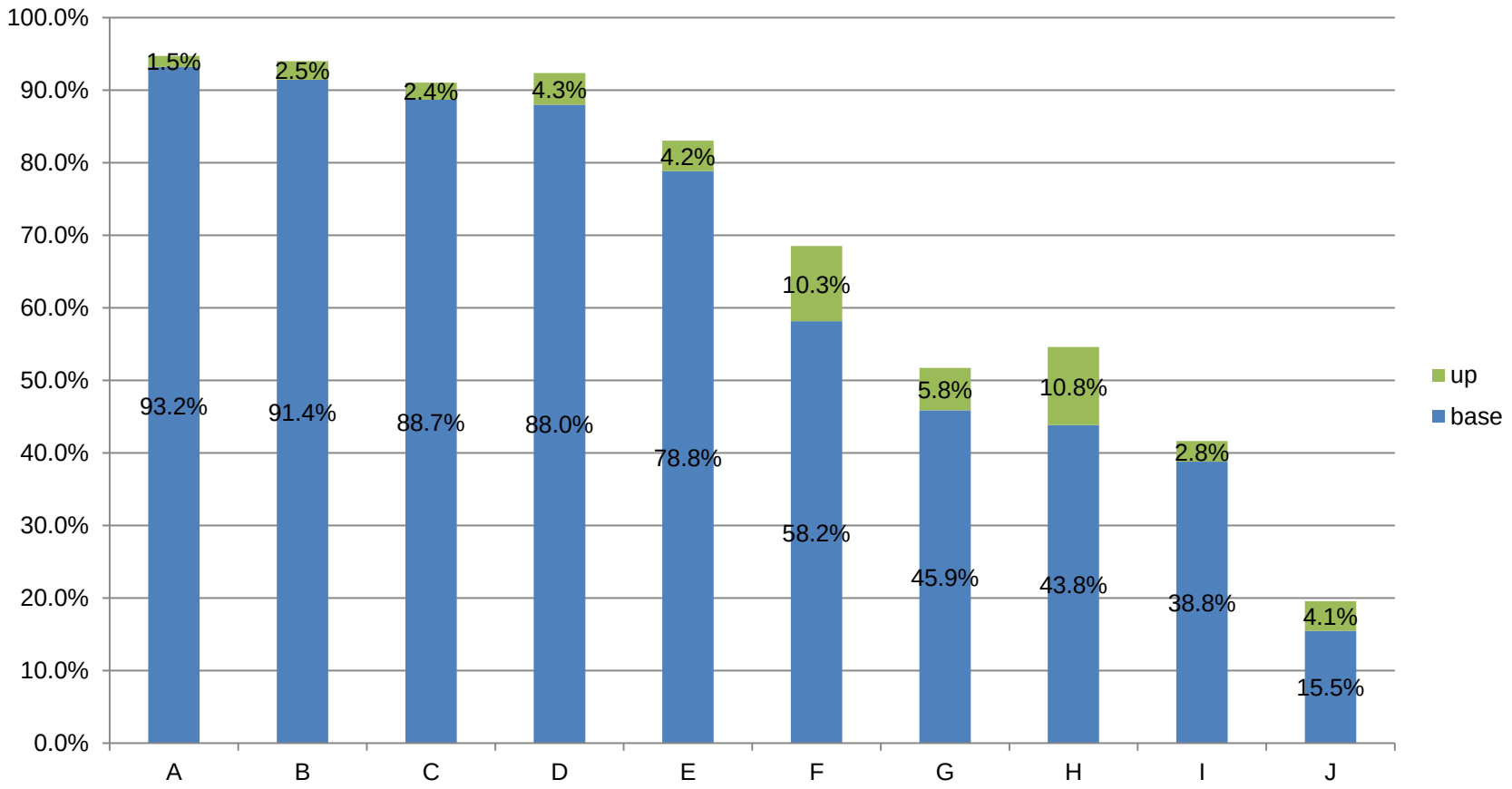
- 今回は sdhash という Fuzzy hashing ツールを用いた
  - http://sdhash.org
  - sdhash で類似度スコア21以上のペアを類似しているとした

未検知検体の類似度調査結果

ベンダ	未検知検体数	平均類似件数	検知した検体との類似件数
A	197	2.89	44
B	253	8.11	75
C	340	2.29	71
D	360	3.72	130
E	628	4.33	124
F	1250	6.51	309
G	1620	16.01	175
H	1670	11.30	320
I	1835	15.95	85
J	2580	14.24	106

- 平均類似件数：未検知検体を1個サンプリングした場合にそれに類似する未検知検体が平均何個あるか

検知できた検体と類似度の高い未検知検体の割合



未検知検体の類似度調査結果（考察）

- 未検知検体の多くを検知するには、ほとんどのベンダーは100～200パターンの追加が必要と思われる

- 検知した検体と類似する未検知検体を検出した場合、検知率にして1.5～10.8%の向上に値する。しかしながら、大半のベンダーでは5%未満の向上にしかならない

まとめ

- 研究用データセットに用いた検体に対するアンチウイルスの静的なスキャンの検知率は、ベンダー間で大きな差がある
- 未検知検体の多くを検知するためには、ほとんどのベンダーは100～200のパターンあるいはロジックの追加が必要と考えられる
- 検知できた検体とバイナリ的に類似している未検知検体は少ない
- アンチウイルスによる静的な検知は限界を迎えており、振る舞い検知が必要と考えられる

Contact Information

E-Mail ： research-feedback@ffri.jp
Twitter： @FFRI_Research