Monthly Research

動的情報に基づいたマルウェアのクラスタリング

株式会社ＦＦＲＩ
http://www.ffri.jp

Ver2.00.01

Agenda

1. 背景と目的
2. クラスタリング概要
3. 実験概要
4. 実験結果
5. 考察
6. まとめ
7. 今後の課題

1.背景と目的

- 近年、亜種を含めマルウェアが急増
- 全ての検体を手動で解析することは困難
  - **要求1) 優先して手動で解析すべき検体（新種等）の選定が必要**
  - **要求2) 手動解析の効率化が必要（類似マルウェア・亜種の特定）**
- 新たな検知エンジンの研究開発等においても保有する全検体を利用した研究は非効率的、または非現実的
  - **要求3) マルウェアのグルーピングと代表データの選定が必要**
- 上記より一手法として「挙動の類似性に基づいたクラスタリグ」を実施
  - 今回は試験的に一手法を試し、その結果を考察する

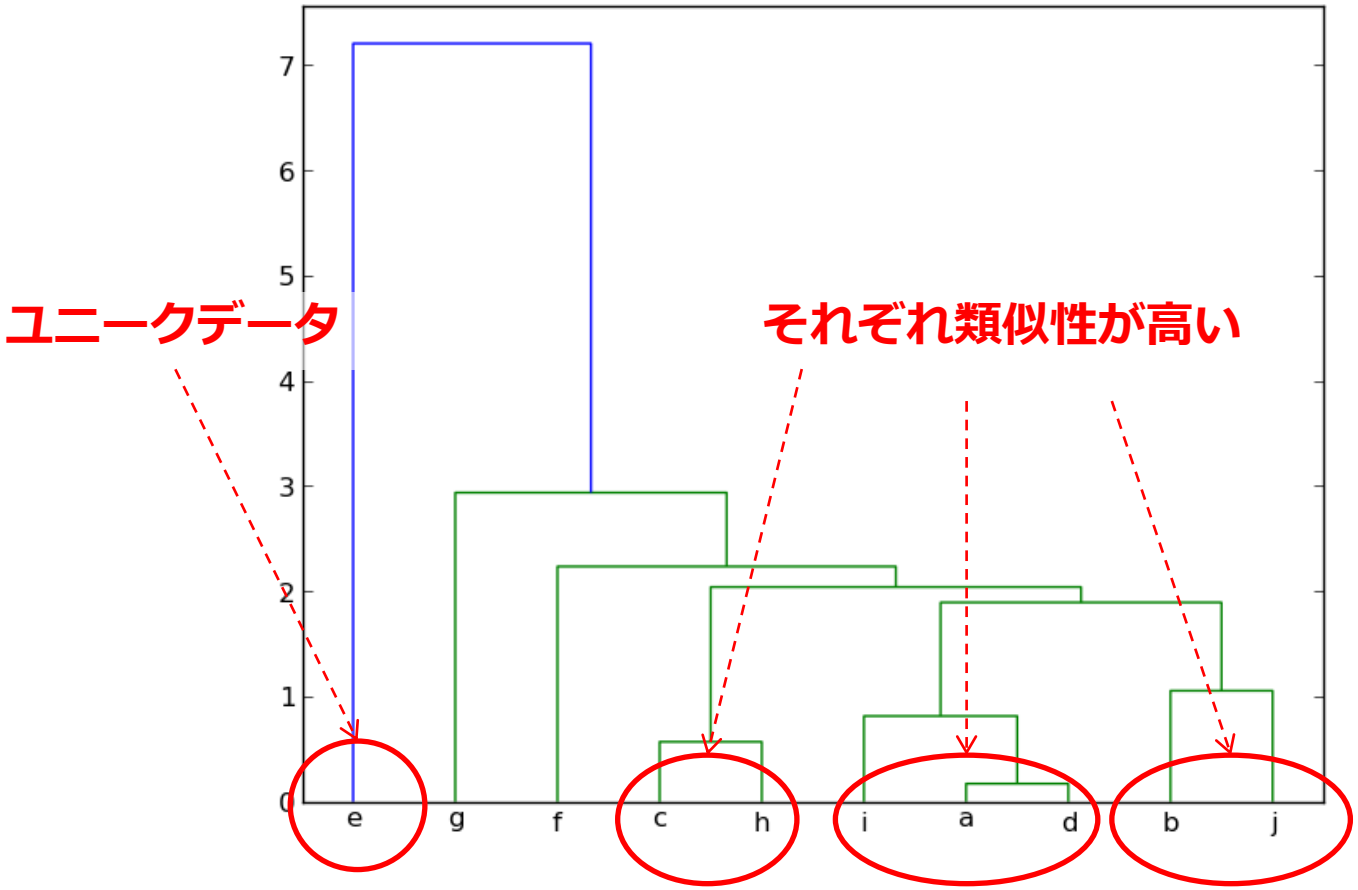
2.クラスタリング概要

- データを「特徴」に基づいて似たもの同士（クラスタ）に分割する
  - 「特徴」は人間が選ぶ必要がある

- 大きく次の2種類が存在
  - 階層的クラスタリング
    - 各データを一つのクラスタと見做して、クラスタ間の類似度等に基づいてツリーとして統合
    - 結果はデンドログラム（樹形図）として表現される
    - 代表的な手法は、最短距離法、最長距離法、群平均法、ウォード法等
  - 非階層的クラスタリング
    - データ同士の類似度に基づいて任意のグループに分割する
    - あるデータが一つのグループにのみ属するハードクラスタリングと複数のグループに属することを許すソフトクラスタリングが存在
    - 代表的な手法は、k-means、混合分布モデル、NMF等

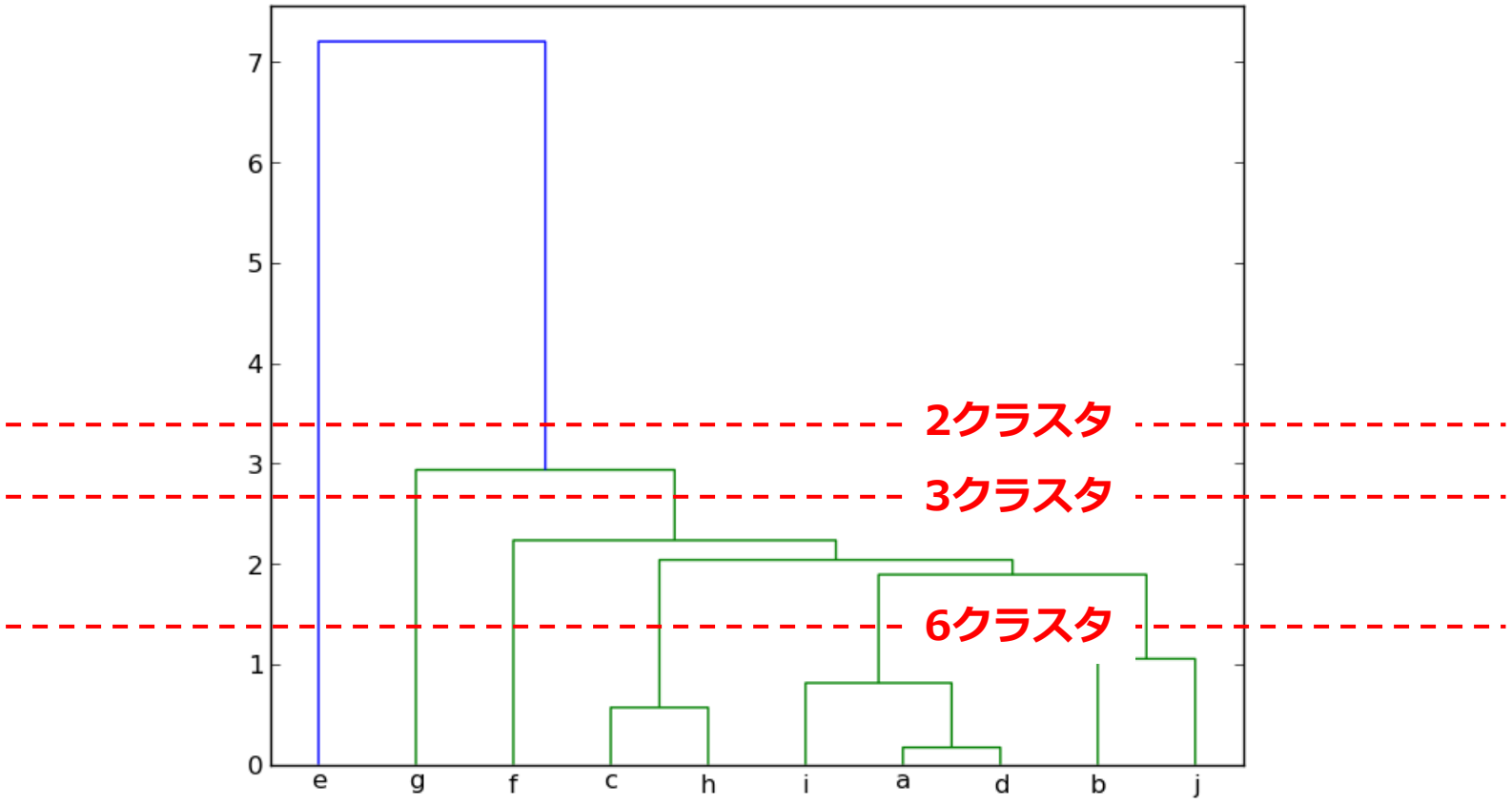
2.クラスタリング概要 / 階層的クラスタリング

■デンドログラムの例（横軸のアルファベットが個別データに対応）



2.クラスタリング概要 / 階層的クラスタリング

■ツリーの深さに応じて任意のクラスタに分割（解釈）することができる



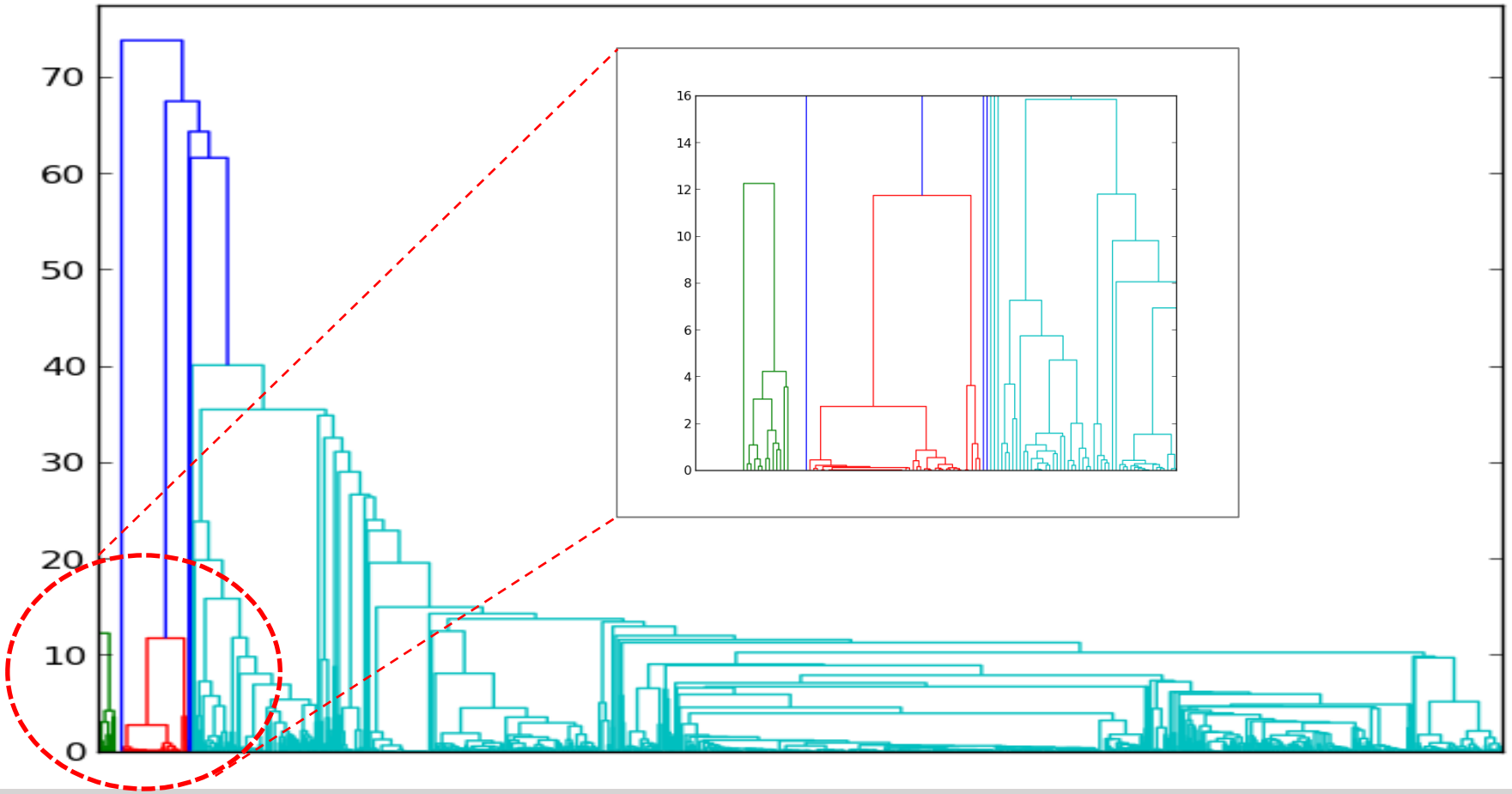
3.実験概要

- 階層的クラスタリングの代表的な手法であるウォード法を試行
  - 非階層的クラスタリングの多くはクラスタ数をユーザーが指定
    - 最適値の選択も一課題となるため階層的クラスタリングを採用
  - 特徴は下記を採用（マルウェアの機能との相関性を考慮）
    - 実行時のAPIコール(API名のみ)の3-gram出現頻度(tf-idfにて重み付け)
- 実験用データ
  - FFRI独自収集のマルウェアから1,000件を無作為抽出
    - VirusTotalを利用し、ハッシュ値に基づいて検出状況を確認
    - 各ベンダー平均して8～9割方検知
- ソフトウェア
  - Cuckoo sandbox 0.6を利用し、上記マルウェアの動的解析を実施
  - クラスタリングには、Scipy（及びMatplotlib)を利用

tf-idf: http://ja.wikipedia.org/wiki/Tf-idf
Scipy: http://www.scipy.org/

4.実験結果

- 大きく3系統に分岐（下記、緑/赤/水色の系)



5.考察

a. 機能・挙動毎の（期待した）クラスタになっているか

b. 優先して解析すべき検体の選定に有効か（要求1)

c. 手動解析の効率化に寄与するか（要求2)

d. データのサンプリングに有効か（要求3)

5.考察 / a.機能毎の（期待した）クラスタになっているか

- 全体の評価は困難なため最下位(最深)で類似したマルウェアのペアを3点抽出
  - MD5ハッシュ値及び検出名（参考）

- 下記のペアについてAPIコールログ以外の情報に基づいて機能比較を実施
  - ペア1
    - aac95e967b1ce621bd2b1a5854d0294d (HEUR:Trojan.Win32.Generic)
    - 69fcc9c0dca876307d97a64683936bad (Unknown)
  - ペア2
    - 5dfca9602289f20f13902c4ed3710fb2 (HEUR:Trojan.Win32.Generic)
    - 90c4af98638d7d9418f2e29f55ec6c9f (HEUR:Trojan.Win32.Generic)
  - ペア3
    - 9f267ae8fb419f2071795803216a3455 (Trojan.Win32.Jorik.Buterat.nwr)
    - dadcb4ab9827f66ba5bd350d78b902cc (Backdoor.Win32.Buterat.zqy)

ペア1

- 結論
  - 両検体は、亜種または設定が異なる同一生成ツールで作られたものと考えられる
- 一致点
  - 23148バイトの同一ハッシュ値のデータファイル
    C:¥Windows¥Registration配下に作成
  - アクセスするレジストリキー、作成するミューテックスが完全に一致
  - アクセスするファイル名称のエンコード（難読化）手法が同一
  - 実行ファイル中に含まれるASCII文字列の大部分が一致
  - レジストリ2箇所に自身のログオン時自動起動設定を登録
- 相違点
  - 作成する実行ファイルが異なる（ハッシュ値）
  - VirusTotalによる検出状況が大幅に異なる(Unknownはほとんど未検出)

ペア１ / 実行時にアクセスするレジストリキー（完全一致）

```
 "keys": [
            "HKEY_LOCAL_MACHINE¥¥Software¥¥Microsoft¥¥COM3",
            "HKEY_LOCAL_MACHINE¥¥Software¥¥Classes",
            "HKEY_LOCAL_MACHINE¥¥Software¥¥Classes¥¥CLSID",
            "CLSID¥¥{304CE942-6E39-40D8-943A-B913C40C9CD4}",
            "CLSID¥¥{304CE942-6E39-40D8-943A-B913C40C9CD4}¥¥TreatAs",
            "¥¥CLSID¥¥{304CE942-6E39-40D8-943A-B913C40C9CD4}",
            "¥¥CLSID¥¥{304CE942-6E39-40D8-943A-B913C40C9CD4}¥¥InprocServer32",
            "¥¥CLSID¥¥{304CE942-6E39-40D8-943A-B913C40C9CD4}¥¥InprocServerX86",
            "¥¥CLSID¥¥{304CE942-6E39-40D8-943A-B913C40C9CD4}¥¥LocalServer32",
            "¥¥CLSID¥¥{304CE942-6E39-40D8-943A-B913C40C9CD4}¥¥InprocHandler32",
            "¥¥CLSID¥¥{304CE942-6E39-40D8-943A-B913C40C9CD4}¥¥InprocHandlerX86",
            "¥¥CLSID¥¥{304CE942-6E39-40D8-943A-B913C40C9CD4}¥¥LocalServer",
            "HKEY_CLASSES_ROOT¥¥CLSID¥¥{304CE942-6E39-40D8-943A-B913C40C9CD4}",
            "HKEY_CLASSES_ROOT¥¥CLSID¥¥{304CE942-6E39-40D8-943A-
B913C40C9CD4}¥¥TreatAs",
            "HKEY_LOCAL_MACHINE¥¥Software¥¥Microsoft¥¥Rpc¥¥SecurityService",

"HKEY_LOCAL_MACHINE¥¥SYSTEM¥¥CurrentControlSet¥¥Services¥¥SharedAccess¥¥Parameters¥¥
FirewallPolicy¥¥StandardProfile",
            "HKEY_LOCAL_MACHINE¥¥software¥¥microsoft¥¥windows nt¥¥currentversion¥¥winlogon",
            "HKEY_CURRENT_USER¥¥software¥¥microsoft¥¥windows¥¥currentversion¥¥run"
        ],
```

ペア2

- 結論
  - 両検体は、亜種または設定が異なる同一生成ツールで作られたものと考えられる
- 一致点
  - 生成するファイルの件数及び各ファイルのファイルサイズ(計6件)
  - 上記のうち4件のハッシュ値が一致
  - 作成・アクセスするファイル、レジストリキー、ミューテックスが完全一致
  - wdmaud.drv, aux, mixer等のAudio関連のレジストリ値を確認
  - OSのエラー報告設定を変更
- 相違点
  - 生成するファイルの6種のうち2つが異なるファイル

ペア2 / 実行時にアクセスするレジストリキー（完全一致）

```
 "keys": [
            "HKEY_LOCAL_MACHINE¥¥Software¥¥Microsoft¥¥Windows NT¥¥CurrentVersion¥¥IMM",
            "HKEY_CURRENT_USER¥¥SOFTWARE¥¥Microsoft¥¥CTF",
            "HKEY_LOCAL_MACHINE¥¥Software¥¥Microsoft¥¥CTF¥¥SystemShared",
...
            "Drivers¥¥wave",
            "Drivers¥¥wave¥¥wdmaud.drv",
              "Drivers¥¥midi",
            "Drivers¥¥midi¥¥wdmaud.drv",
            "Drivers¥¥aux",
            "Drivers¥¥aux¥¥wdmaud.drv",
            "Drivers¥¥mixer",
            "Drivers¥¥mixer¥¥wdmaud.drv",
...
"HKEY_LOCAL_MACHINE¥¥Software¥¥Policies¥¥Microsoft¥¥PCHealth¥¥ErrorReporting",
            "HKEY_LOCAL_MACHINE¥¥Software¥¥Microsoft¥¥PCHealth¥¥ErrorReporting",

"HKEY_LOCAL_MACHINE¥¥Software¥¥Microsoft¥¥PCHealth¥¥ErrorReporting¥¥ExclusionList",

"HKEY_LOCAL_MACHINE¥¥Software¥¥Microsoft¥¥PCHealth¥¥ErrorReporting¥¥InclusionList",
            "HKEY_LOCAL_MACHINE¥¥System¥¥Setup"
        ],
```

ペア3

- 結論
  - 両検体は、亜種または設定が異なる同一生成ツールで作られたものと考えられる
  - C&CのFQDNが一部共通しているため同じ攻撃者が異なるタイミングで作成したものであることが想像される

- 一致点
  - 生成するファイルの件数及び各ファイルのファイルサイズ(計6件)
  - 上記のうち4件のハッシュ値が一致
  - 作成・アクセスするファイル、レジストリキー、ミューテックスが完全一致
  - IE及びエクスプローラの同じ設定項目を変更
  - C&Cサーバ のFQDNが一致（それぞれ3つ中1つ）
  - 実行時のプロセスツリー（外部コマンドの実行等）が同一

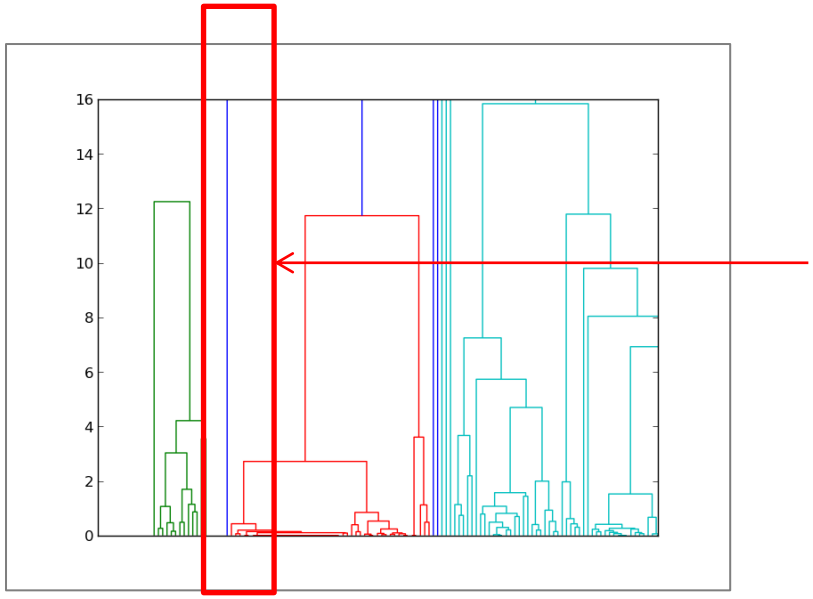
- 相違点
  - C&Cサーバの残り2つが異なる

ペア3 / 実行時のプロセスツリー（完全一致）


```
{
        "pid": 404,
        "name": "9F267AE8FB419F2071795803216A3455.bin",
        "children": [
            {
                "pid": 388,
                "name": "9F267AE8FB419F2071795803216A3455.bin",
                "children": [
                    {
                        "pid": 1832,
                        "name": "taskhost.exe",
                        "children": [
                            {
                                "pid": 1828,
                                "name": "taskhost.exe",
                                "children": []
                            }
                        ]
                    }
                ]
            }
        ]
    }
```

5.考察 / b.優先して解析すべき検体の選定に有効か（要求1)

- 上記の結果より意味のあるクラスタリングが行われている前提であれば
  下記の条件を満たすものが新種のマルウェアであると考えることができる
  - 既存のアンチウイルス製品で検出されない
  - クラスタリングの結果、ユニークまたは疎な系を形成する



こうした検体が既存の対策製品で未検出
であった場合、新種である可能性が高い

5.考察 / c.手動解析の効率化に寄与するか（要求2)

- 前述の結果より、既存の対策製品で未検出の検体でもクラスタリングにより類似した検体が存在した場合、挙動・機能も類似している可能性が高く解析のヒントになり得る

- また各社のヒューリスティックエンジンで検知された検体（HEUR～等）についてもクラスタリングすることでより正確な分類が可能なると考えられる

  - デンドログラムより同じ「HEUR～」という検体でもそれぞれ別の系に属することが複数確認された

5.考察 / d.データのサンプリングに有効か（要求3)

- クラスタリング結果より必要なデータ数に基づいて任意の深さでN個のクラスタと見做し、各クラスタからサンプリングする方法が考えられる

- 特にクラスタ間で出現傾向や全体に占める割合が異なる場合に有効
  - 層化抽出法
    - http://ja.wikipedia.org/wiki/%E5%B1%A4%E5%88%A5%E6%8A%BD%E5%87%BA

6.まとめ

- マルウェアの増加に伴い、マルウェアを資産として活用する上でその管理方法を工夫する必要がある
- 上記より今回は一試作としてAPIコールの3-gramを特徴としてウォード法によるクラスタリングを行った
- 限定的ながら機能及び挙動の類似度に基づいたクラスタになっていることを確認した
- これを利用することで下記の要求を満たすことが可能
  - 優先して解析すべき検体の選定
  - 手動解析の効率化
  - 意味のあるマルウェアのサンプリング

7.今後の課題

- 他の特徴、クラスタリング手法との比較
- マルウェアの機能・挙動の定義と比較方法の検討
  - MAEC(Malware Attribute Enumeration and Characterization)等の利用
  - http://maec.mitre.org/
- クラスタリング結果のよりユーザーフレンドリーなUI/インターフェイスの実現
- 運用を見据えた自動化及びシステム化検討

Contact Information

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