Innovative R&D by NTT

MWS 2014 意見交換会

D3M (**D**rive-by **D**ownload **D**ata by **M**arionette) 2014

NTTセキュアプラットフォーム研究所
ネットワークセキュリティプロジェクト
高田 雄太、秋山 満昭

2014年05月19日

はじめに

- 研究を進める上で評価用の実データは非常に重要
  - 実データに触れること傾向を把握できる
  - 研究を客観的に評価できる

- セキュリティ系の研究における評価用データとは
  - マルウェア検体
  - マルウェア実行ログ
  - 悪性通信データ
  - などといった「悪性コンテンツ」

- 「悪性コンテンツ」を収集するには…
  - ハニーポットを設置して集める？（安全性担保できる？）
  - ブラックマーケットで購入？

研究コミュニティとデータセット

Innovative R&D by NTT

攻撃コード、マルウェアなど

- iseclab (http://www.iseclab.org/)
  - コミュニティ内でのデータセット共有
- Honeynet Project (http://www.honeynet.org)
  - コミュニティ内でのデータセット共有

悪性 URL, Domain リスト

- Malware domain list (http://www.malwaredomainlist.com/)
  - 悪性 URL を公開
- Google safe browsing
  - URL のハッシュ値のみ公開（NDA により実際の URL 共有？）
- StopBadware (https://www.stopbadware.org/)
  - 一部のURL を公開、パートナーシップ契約（有料）により詳細情報を共有
- hphosts (http://www.hosts-file.net/)
  - 悪性ドメインを公開

- 各研究コミュニティにおいて、データセット共有による研究開発
- データセット共有までの道のりは長い
- **MWS は日本人であればコミュニティの参加が比較的容易**

NTT

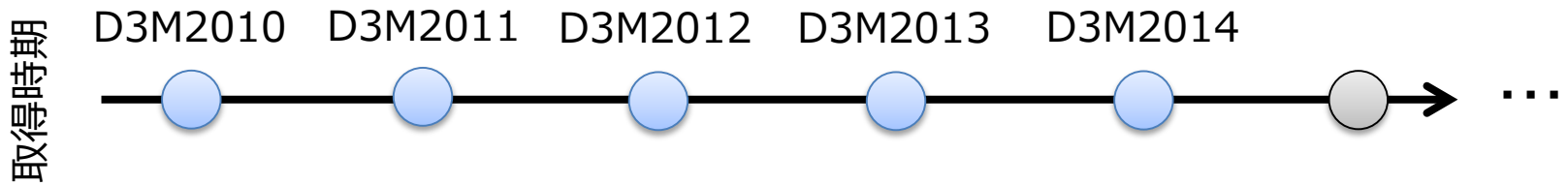
攻撃の複雑化、巧妙化

- マルウェア感染経路の変化
  - 能動的な攻撃（ネットワーク経由）から受動的な攻撃（メールや Web 経由）へ
- 脆弱性の多様化
  - ブラウザ（IE 6/7/8/9/10/11, Firefox, Opera …）
  - プラグイン（Acrobat 8/9, Flash 9/10/11, Java 6/7 …）
- エクスプロイトキット
  - 難読化処理、クローキング、多段リダイレクト、フィンガープリント
- 短命な悪性サイト
  - 攻撃者は、攻撃コードやマルウェアの解析を妨害するべく、悪性サイトを削除してしまう（公開ブラックリストに掲載されると特に早い）

- 複雑化・巧妙化するドライブバイダウンロード攻撃は、**データセットの収集自体が非常に困難**
- D3M では、ドライブバイダウンロード攻撃による一連の攻撃通信およびマルウェア通信データが記録されており、**様々な攻撃・マルウェアデータが含まれている**

D3M の概要

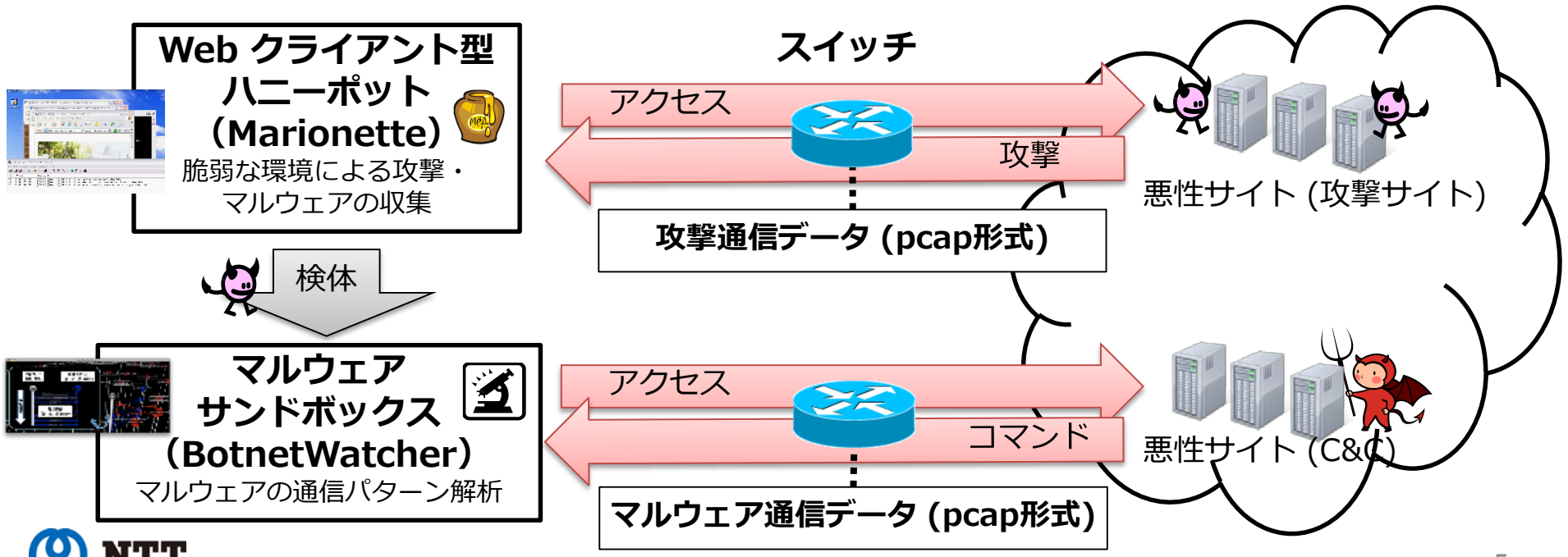
- データセットの内容
  - 攻撃通信データ
    - 悪性 URL を巡回した際に得られたドライブバイダウンロード攻撃の通信データ
  - マルウェア
    - ドライブバイダウンロード攻撃によってホスト上にダウンロードされた実行形式のファイル
  - マルウェア通信データ
    - 取得して24時間以内にマルウェアサンドボックス上で実行した際の通信データ
    - マルウェアサンドボックスはインターネットに接続可能（攻撃通信は遮断）



D3M2014 には、D3M2013, 2012, 2011, 2010 が同梱されている
5年分のデータを解析することで、攻撃の変化傾向がわかる（かも・・・）

D3M の取得環境

- ドライブバイダウンロード攻撃に関連する URL をブラウザへ投入し、自動的に発生する一連の Web 通信、ダウンロードしたマルウェアの通信を収録
- 取得手順
  1. 独自に収集した悪性 URL リストを Web クライアント型ハニーポット（Marionette）で巡回
  2. 検知した URL を直ちに再巡回し、その際の通信データを記録
  3. 2. で取得したマルウェア検体を、マルウェアサンドボックス（Botnet Watcher）で解析し、その時の通信データを記録



D3M に含まれる情報

- 提供されるデータ形式
  - pcap（ドライブバイダウンロード攻撃通信、マルウェア通信）
  - Binary（マルウェア検体）

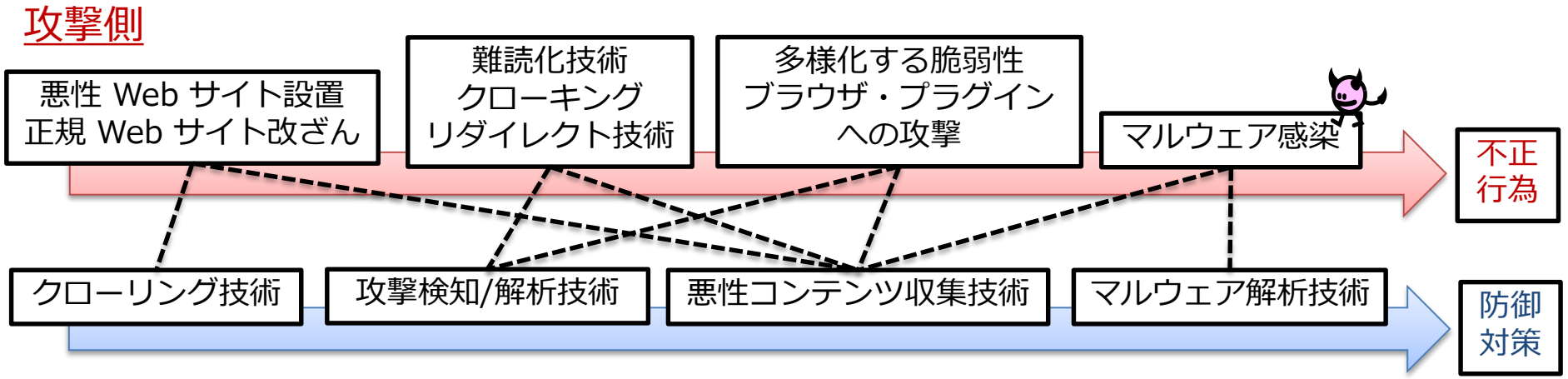
- 攻撃を行う URL, ドメイン名, IP アドレス
  - ブラックリスト巡回のため、攻撃サイトへの直接アクセス
  - **入口サイトの情報は含まれていない**
- 難読化された JavaScript
- 攻撃コード（HTML, JavaScript, PDF, Jar, …）
- マルウェア検体
- マルウェア実行時の通信

研究の観点

- どのような観点でデータセットを解析すればよいか？
- トップカンファレンスを調査し、世の中の研究動向を把握する

- 学術系カンファレンス
  - ACM CCS
  - IEEE S&P
  - USENIX Security
  - NDSS
  - RAID
  - ACSAC
  - DIMVA
  - etc.
- 産業系カンファレンス
  - Blackhat
  - CanSecWest
  - PacSec
  - CODEBLUE
  - etc.

ドライブバイダウンロード攻撃の対策フェーズ



- クローリング技術
  - 悪性サイトの特徴に基づいた効率的な Web 空間探索技術の研究
- 攻撃検知／解析技術
  - 多様化する脆弱性やさまざまなブラウザ、プラグインへの攻撃の対策研究
- 悪性コンテンツ収集技術
  - Web クライアント型ハニーポットの研究やそれを用いた大規模な実態調査
- マルウェア解析技術
  - ドライブバイダウンロードに限らず、広く研究されている
  - BHO 化や MITB などブラウザに寄生するマルウェアの解析

クローリング技術

- WebCop *[USENIX LEET 2010] (Microsoft research)*
  - マルウェア配布URLのリンク・被リンクを辿ることで、マルウェア配布に関わる悪性な入口URLを発見
- Structural Neighborhood URL Lookup *[SAINT 2011] (NTT SC研)*
  - 既知の悪性URLの構造的な近隣を中心に検査することで効率的に未知の悪性URLを発見する方法
- PoisonAmplifier *[RAID 2012]*
  - SEOを行う既知の悪性URLに対して、SEO特有の文字列を抽出してキーワード検索を行うことで、SEOを行う未知の悪性URLを発見する方法
- EVILSEED *[IEEE S&P 2012] (UCSB)*
  - ハイパーリンク、URL構造、SEO、ドメイン登録情報、DNSクエリ情報などを使って未知の悪性URLを発見する方法
  - 上記３論文の手法の合わせ技

攻撃検知／解析技術 1/2

- JavaScript解析
  - JSUnpack（*http://jsunpack.jeek.org*）
  - JSAND *[WWW 2010] (UCSB, Wepawet)*
    - JavaScript 関数呼び出しや ActiveX コンポーネントの使用を基に機械学習し、アノマリ検知
  - Prophiler *[WWW 2011] (UCSB)*
    - HTML や JavaScript、URL を静的解析
  - Revolver *[USENIX Security 2013] (UCSB)*
    - JavaScript AST のクラスタリング
  - ZOZZLE *[USENIX Security 2011] (Microsoft research)*
    - JavaScript AST を用いた静的解析
  - ROZZLE *[IEEE S&P 2012] (Microsoft research)*
    - マルチパス実行によるコードカバレッジ向上
  - Cujo *[ACSAC 2010] (TU-Berlin)*
    - 静的解析と動的解析による特徴抽出を行い、機械学習による悪性コンテンツ検知
  - ICESHIELD *[RAID 2011] (Ruhr-Univ.)*
    - JavaScript 関数呼び出しや DOM 挿入といった特徴を基に機械学習による悪性コンテンツ検知

攻撃検知／解析技術 2/2

- HeapSpray検知
  - NOZZLE *[USENIX Security 2007] (Microsoft research)*
  - Heap Inspector *[Blackhat USA 2011]*
- Flash解析
  - Analyzing and detecting malicious flash advertisements *[ACSAC 2009] (UCSB)*
  - FlashDetect *[RAID 2012] (UCSB)*
- PDF解析
  - Detection of Malicious PDF Files Based on Hierarchical Document Structure *[NDSS 2013]*
- Jar解析
  - Jarhead *[ACSAC 2012] (UCSB)*

悪性コンテンツ収集技術

- Web クライアント型ハニーポット
  - 高対話型
    - HoneyMonkey *[NDSS 2006] (Microsoft research)*
    - Argos / Shelia *(VU Amsterdam)*
    - Capture-HPC *(Honeynet project)*
    - BLADE *[ACM CCS 2010] (Georgia Tech)*
    - Marionette *[SAINT 2012] (NTT SC 研)*
  - 低対話型
    - HoneyC *(Honeynet project)*
    - Caffeine Monkey *[Blackhat 2007]*
    - PhoneyC *[USENIX LEET 2009] (Honeynet project)*
    - Thug *(Honeynet project)*
  - ハイブリッド型
    - HoneySpider *(CERT Polska)*

実態調査

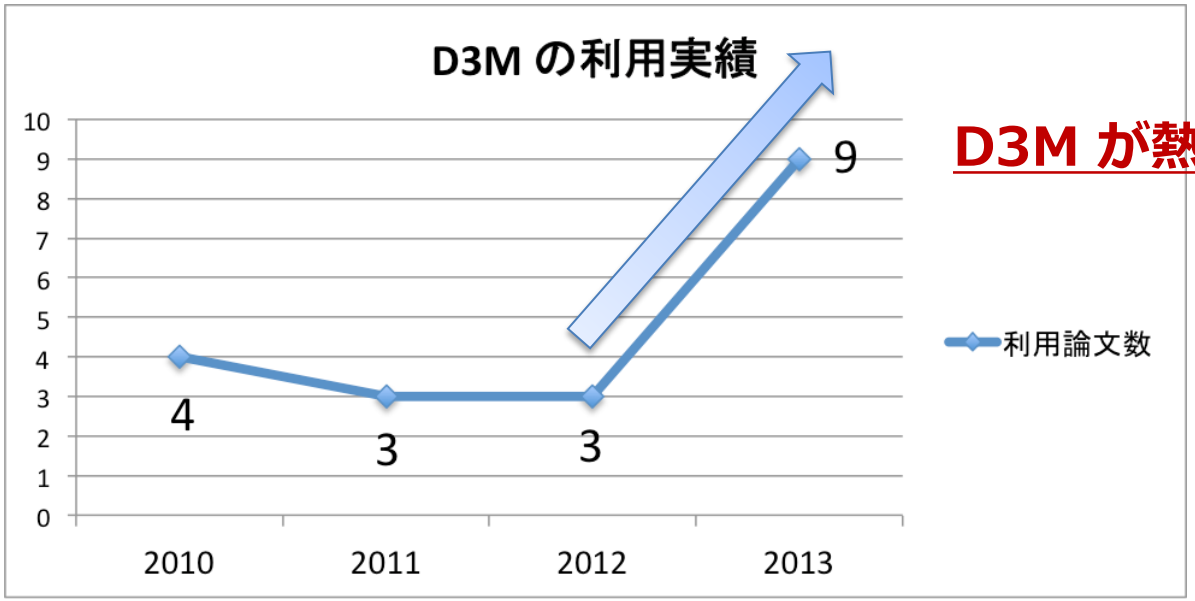
- A Crawler-based Study of Spyware on the Web *[NDSS2006] (Washington Univ.)*
- Know Your Enemy: Malicious Web Servers *(http://www.honeynet.org/papers/mws/, 2007) (Honeynet project)*
- All Your iFRAMEs Point to Us *[USENIX Security 2007] (Google)*
- Manufacturing Compromise: The Emergence of Exploit-as-a-Service *[ACM CCS2012]*

セキュリティにおける理想的な研究開発サイクル

検知手法 ＞ 実態調査 ＞ 検知手法改良 ＞ 実態調査 ＞ ・・・

MWS の研究動向 1/2

- MWS における D3M の研究利用実績は、**昨年３倍増**
  - ドライブバイダウンロード攻撃に関する研究が多く行われている



D3M が熱い！

MWS の研究動向 2/2

- JavaScript 解析
  - 抽象構文解析木による不正なJavaScriptの特徴点抽出手法の提案 *[MWS 2011] (*セキュアブレイン 神薗ら, *MWS2011優秀論文賞)*
  - 難読化されたスクリプトにおける特徴的な構文構造のサブツリーマッチングによる同定 *[MWS 2011] (*奈良先端大 *Gregory*ら*)*
  - 抽象構文木を用いた Javascript ファイルの分類に関する一検討 *[MWS 2011] (*東大 宮本ら*)*
- PDF 解析
  - 動的解析を利用した難読化JavaScriptコード解析システムの実装と評価 *[MWS 2010] (*セキュアブレイン 神薗ら, *MWS2010優秀論文賞)*
  - PDF の構造検査による悪性 PDF の検知 *[MWS2013] (NISC* 大坪ら*)*
- Exploit kit 解析
  - Drive-by-Download攻撃における通信の定性的特徴とその遷移を捉えた検知方式 [MWS 2013] (NTT データ 北野ら)
  - Exploit kit の特徴を用いた悪性 Web サイトの検知手法 *[MWS 2013] (NICT* 笠間ら*)*
- リダイレクト解析
  - 検知を目指した不正リダイレクトの分析 *[MWS 2010] (*富士通研 寺田ら*)*
  - パスシーケンスに基づく *Drive-by-Download* 攻撃の分類 *[MWS 2010] (*東海大 桑原ら*)*

まとめ

- セキュリティ系の研究用データセットを収集することは非常に難しいが、MWS では様々なデータセットが提供されている

- ドライブバイダウンロード攻撃対策研究
  - 攻撃検知／解析技術、クローリング技術、マルウェア解析技術、実態調査

- MWS において、ドライブバイダウンロード攻撃に関連する研究は増加傾向にある
  - D3M が熱い！