通信の多様化に向けた 生物の環境適応性に基づく Web サイトのぜい弱性スキャン検知

○久世尚美[1]、石倉秀[1]、八木毅[2]、千葉大紀[2]、村田正幸[1]

[1] 大阪大学大学院情報科学研究科

[2] NTT セキュアプラットフォーム研究所

Web サイトへの攻撃

- Web サービスの社会インフラ化
- Web サービスの起点となる Web サイトへの攻撃が多発
  - Linux worm（2013 年 11 月～）
  - Apache Struts（2013 年 11 月～）
  - Shellshock（2013 年 11 月～）など

- **Web サイトへの攻撃を収集、解析して、未知の攻撃への対策に活用することが必須**
  - 全サービスのぜい弱性を特定、Web サイトへの攻撃を防ぐことは困難
  - 既知のぜい弱性に基づいて攻撃への対策を行うだけでは不十分

ハニーポットによる攻撃収集

- Web サーバ型ハニーポット
  - Web サイトへの攻撃を収集するために、攻撃ベクトルに応じて構築されるおとりのシステム
  - 低対話型と高対話型に大別
    - 低対話型
      - 特定の OS やアプリケーションをエミュレートして監視を行う
      - 低コストではあるが、あらゆる攻撃への対応をエミュレートすることは困難
    - **高対話型**
      - 実際の OS やアプリケーションを用いて監視を行う
      - 比較的高コストではあるものの、おとりであることが攻撃者に感知されにくく、多くの情報の収集が可能
- **収集した情報の中から目的となる攻撃情報を分類する必要性**
  - 検索エンジンデータベース作成のためのクローラなどによる正常な通信が多く含まれる
  - 特に Web サーバへの攻撃の準備動作として Web サイトのぜい弱性を確認するぜい弱性スキャンの識別は攻撃への対策のために重要
    - **様々なプログラムに対して様々な入力値を試行するクローラと特徴が類似**

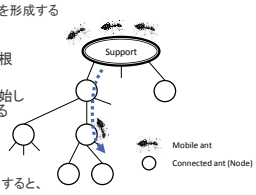
通信の多様性とスキャン識別

- 従来の攻撃識別 [1]
  - 初めにクローラによる通信の識別を行い、それ以外の通信を攻撃と判定
  - クローラの識別においては、Google などの有名クローラに似た挙動を示すものをクローラと判断
- Web サービスの多様化
  - 攻撃者からの通信だけでなく、クローラなどからの正常な通信も多様化
  - 通信の多様化に適用可能な識別手法を検討する必要がある
- **生物由来のクラスタリング手法をクローラ識別へ適用**
  - 多様なデータ群の分類が可能
    - 生物は、全体の情報を用いることなく、個々の個体が知覚可能な情報のみに基づいて動作を決定し、結果として全体の秩序や機能を創発
    - 特定の特徴を持ったデータの判別のみでなく、それぞれ異なる特徴を持ったクラスタへの分類が可能
  - 多量なデータの取り扱いに有用
    - 生物は、局所情報のみを用い単純なルールに基づいて動作を決定するため、高い拡張性を有する

[1] J. P. John, F. Yu, Y. Xie, A. Krishnamurthy, and M. Abadi, "Heat-seeking honeypots: Design and experience," in *Proceedings of the 20th International Conf. on World Wide Web*, Mar. 2011, pp. 207-216.

AntTree [5]

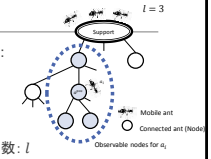
- アリが互いに連結して木構造を構築する集団的行動に着想を得たクラスタリング手法
  - アリを模したデータが互いに類似度に基づいて木構造を構築
    - データの一つ一つを ant と呼ばれるモバイルエージェントとみなす
    - Ant 同士が互いに連結することで木構造を形成する
- Ant による木構造構築
  1. 初期状態では、全ての ant はツリーの根（support）に存在
  2. Support から ant が一つずつ移動を開始しツリー上を移動しながら自身と類似する ant（ノード）を探索
  3. 類似するノードに到達すると、ant はそのノードの子となり、移動を終了
     - 移動中の ant がノードとなり、移動を終了すると、support 上の ant が一つ移動を開始



[5] H. Azzag, N. Monmarche, M. Slimane, and G. Venturini, "AntTree: a new model for clustering with artificial ants," in *Proceedings of IEEE Congress on Evolutionary Computation (CEC2003)*, vol. 4, Dec. 2003, pp. 2642-2647.

Ant の設計

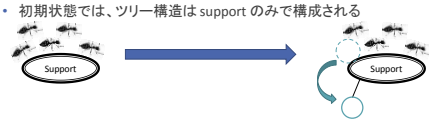
- クラスタリング対象となる ant（データ）の集合：
  $\{a_1, a_2, \ldots, a_N\}$
  - 個々の ant がデータに対応
  - Ant は自身が現在存在するノード（ant）と その隣接ノードの情報のみを用いて動作を決定
  - ノードとなった ant の連結可能な子ノードの最大数：l



- 類似するノードの探索
  - Ant a_i、a_j の類似度：$Sim(a_i, a_j)$
  - Ant a_i が他の ant との類似度、非類似度を測る閾値：
    $T_{Sim}(a_i), T_{Dissim}(a_i)$
    - $Sim(a_i, a_j) \geq T_{Sim}(a_i)$ であれば、a_i は a_j が類似した特徴を持つと判断
    - $Sim(a_i, a_j) < T_{Dissim}(a_i)$ であれば、a_i は a_j が異なる特徴を持つと判断
    - Ant a_i はツリー構造上の移動を行う過程で閾値の更新を行い、適切な値を学習
      - 初期値として $T_{Sim}(a_i) = 1$、$T_{Dissim}(a_i) = 0$

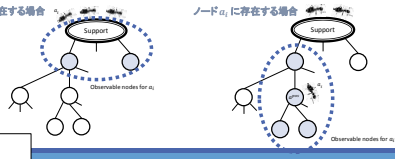
ツリー構造の構築

- 最初に移動を開始した ant は support の子となり、移動を終了
  - 初期状態では、ツリー構造は support のみで構成される



- 後続の ant は局所的な情報に基づき動作

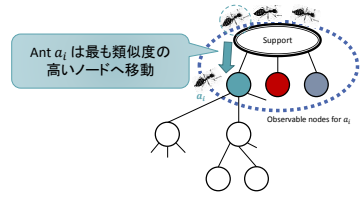
Support に存在する場合 / ノード a_i に存在する場合



Mobile ant
Connected ant (Node)

Support からの移動アルゴリズム (1/3)

1. Ant a_i が support から移動を開始したとき
   - Ant a_i は自身と support の子ノードとを比較
   a. Support の子ノードに a_i と類似したノードが存在する場合



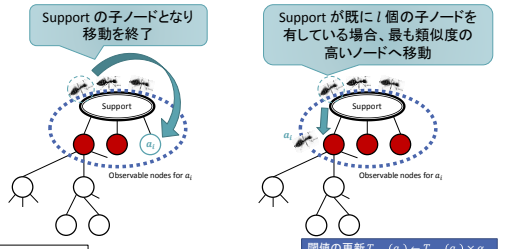
- a_i と類似したノード
- a_i と異なる特徴を持つノード
- 上記のどちらにも当てはまらないノード

Mobile ant
Connected ant (Node)

Support からの移動アルゴリズム (2/3)

1. Ant a_i が support から移動を開始したとき
   - Ant a_i は自身と support の子ノードとを比較
   b. Support の子ノードが全て a_i と異なる特徴を持つ場合

類似ノード
非類似ノード
どちらでもないノード



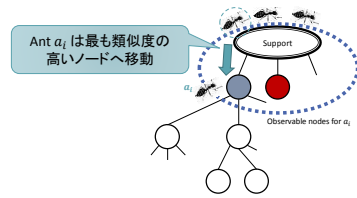
閾値の更新 $T_{Sim}(a_i) \leftarrow T_{Sim}(a_i) \times \alpha_1$

Mobile ant
Connected ant (Node)

Support からの移動アルゴリズム (3/3)

1. Ant a_i が support から移動を開始したとき
   - Ant a_i は自身と support の子ノードとを比較
   c. a., b. のどちらにも当てはまらない場合

類似ノード
非類似ノード
どちらでもないノード



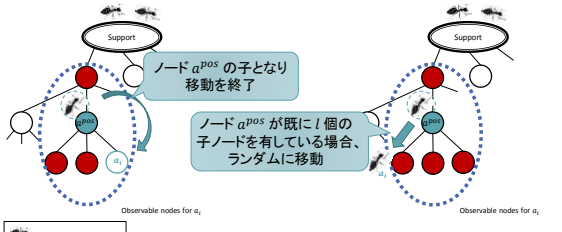
閾値の更新
$T_{Sim}(a_i) \leftarrow T_{Sim}(a_i) \times \alpha_1$
$T_{Dissim}(a_i) \leftarrow T_{Dissim} + \alpha_2$

Mobile ant
Connected ant (Node)

ノード a^{pos} からの移動アルゴリズム (1/3)

2. Ant a_i が ノード a^{pos} へ到着したとき
   a. Ant a_i がノード a^{pos} と類似している場合
   i. ノード a^{pos} の隣接ノード全てが ant a_i と異なる特徴を持つ場合

類似ノード
非類似ノード
どちらでもないノード

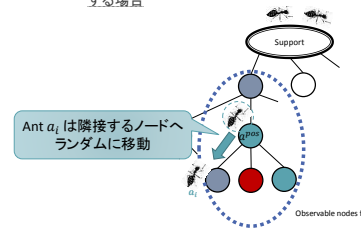


Mobile ant
Connected ant (Node)

ノード a^{pos} からの移動アルゴリズム (2/3)

2. Ant a_i が ノード a^{pos} へ到着したとき
   a. Ant a_i がノード a^{pos} と類似している場合
   ii. ノード a^{pos} の隣接ノードに ant a_i と異なる特徴を持たないノードが存在する場合

類似ノード
非類似ノード
どちらでもないノード

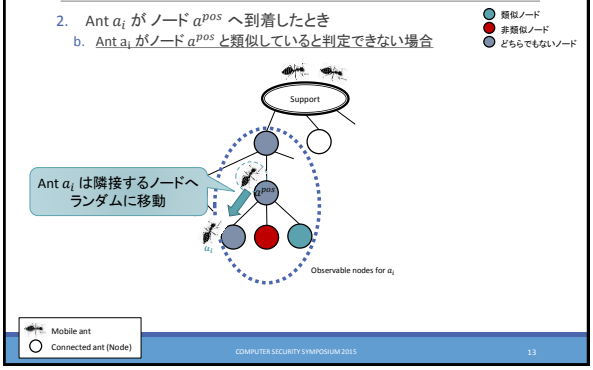


閾値の更新
$T_{Sim}(a_i) \leftarrow T_{Sim}(a_i) \times \alpha_1$
$T_{Dissim}(a_i) \leftarrow T_{Dissim} + \alpha_2$

Mobile ant
Connected ant (Node)

ノード a^{pos} からの移動アルゴリズム (3/3)

2. Ant a_i が ノード a^{pos} へ到着したとき
   b. Ant a_i がノード a^{pos} と類似していると判定できない場合



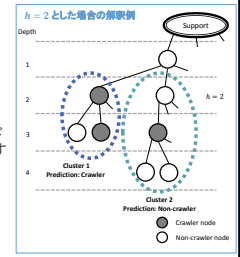
AntTree のクローラ識別への適用

- Ant a_i と a_j との類似度 $Sim(a_i, a_j)$
  - Ant a_i は M 個の特徴量からなる特徴ベクトル $\{v_{i_1}, \ldots, v_{i_M}\}$ を持つ

$$Sim(a_i, a_j) = 1 - \sqrt{\frac{1}{M}\sum_{k=1}^{M}(v_{i_k} - v_{j_k})^2}$$

Ant a_i と a_j の特徴ベクトル空間上のユークリッド距離

- クラスタの解釈
  - 構築されたツリーにおいて、深さh のノードを根とした部分木を一つのクラスタとみなす
  - 各クラスタをそのクラスタに属する最多のノード種別（クローラ or 非クローラ）へと分類する



評価実験

- 生物の仕組みに着想を得たクラスタリング手法 AntTree（教師なし学習）によるクローラ識別の精度について評価
  - 比較対象
    - 有名クローラの特徴からその他のクローラの識別を行う従来手法 [1]
      - 学習アルゴリズムとして、Random Forest（教師あり学習）を用いる
  - 使用データ
    - ハニーポット [14] 37 台により、2013 年 8 月 29 日～2014 年 1 月 14 日の期間に収集した HTTP 通信を解析した際のログ
  - 評価指標
    - 再現率: 同一ラベルを付加されたデータの内、正しく分類されたデータの割合
    - 適合率: 同一カテゴリへ分類されたデータの内、正しく分類されたデータの割合

再現率$= \frac{|L_A \cap C_A|}{|L_A|}$, 適合率$= \frac{|L_A \cap C_A|}{|C_A|}$

L_A: ラベル A を付加されたデータの集合
C_A: カテゴリ A へと分類されたデータの集合

[1] J. P. John, F. Yu, Y. Xie, A. Krishnamurthy, and M. Abadi, "Heat-seeking honeypots: Design and experience," in *Proc. of the 20th International Conf. on World Wide Web*, Mar. 2011, pp. 207-216.
[14] T. Yagi, N. Tanimoto, and T. Hariu, "Intelligent high-interaction web honeypots based on url conversion scheme," *IEICE transactions on communications*, vol. 94, no. 5, pp. 1339-1347, May 2011.

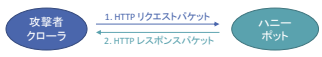
データセット

- ハニーポットで収集された通信ログ
  - 各通信ログに対して、以下のラベル付けを行う
    - ***Google***: Google 使用するクローラによる通信ログ
      - Google による公開情報（UserAgent、送信元 IP アドレス）により判別
    - ***Crawler***: Google 以外が運用するクローラによる通信ログ
      - 研究者、技術者の手動解析により判別
    - ***Non-crawler***: クローラ以外による通信ログ
      - ぜい弱性スキャンをはじめとした悪意のある通信ログが含まれる

- 提案手法（AntTree）、従来手法の評価用データセット
  - Crawler ログ、Non-crawler ログからそれぞれ 1,502,254 個サンプリング
- 従来手法の学習用（教師）データセット
  - Google ログ、Non-crawler ログからそれぞれ 1,502,254 個サンプリング

特徴ベクトル

- HTTP 通信ログに着目し、通信の識別を行う



- 本実験で使用した特徴量
  - リクエストパケット
    - ハニーポットが外部から受信した HTTP リクエストパケットに含まれる情報
      - リクエスト部: リクエスト URL、通信メソッド（GET、POST）
      - ヘッダ部: UserAgent、referer、送信側・受信側ポート番号、通信プロトコル（HTTP、HTTPS）
      - ボディ部: ボディ部の長さ
  - リクエストに対する応答
    - ハニーポットがリクエストパケットを受信した際の動作
      - ハニーポットの応答種別 : StatusCode（200、404、etc.）
      - レスポンスパケットの情報: コンテキスト種別（HTML、CSS、etc.）、コンテキストの文字コード（UTF-8、ISO-8859-1、etc.）

評価結果

<パラメータ設定>
木構造の深の数 t: 5、クラスタ探索時の深さ h: 3
類似度、葉類似度閾値更新時のパラメータ (α_1, α_2): $(0.95, 0.2)$

- 従来手法、AntTree を用いた際の Crawler の識別精度
  - AntTree を用いることで、従来手法と比較して高い精度での識別が可能
    - 再現率、適合率ともに AntTree を用いた場合の方が高い
    - AntTree により教師ラベルを用いずとも高い識別精度が達成される
  - クローラの多様化により、クローラ同士が必ずしも類似した特徴を持つとは限らない

従来手法

		予測		再現率
		Crawler	Non-Crawler	
ラベル	Crawler	1,241,437	260,817	82.64%
	Non-Crawler	105,952	1,396,302	92.95%
	適合率	92.14%	84.26%	

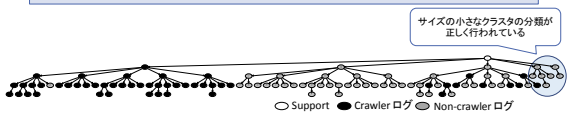
AntTree

		予測		再現率
		Crawler	Non-Crawler	
ラベル	Crawler	1,259,976	242,278	83.87%
	Non-Crawler	76,417	1,425,837	94.91%
	適合率	94.28%	85.48%	

AntTree の特徴

- 比較的サイズの小さなクラスタの分類精度が高い
  - AntTree においては、各データがツリー上を移動しながら局所的な情報を用いて探索を行う

- データ全体から見たそれぞれの通信の特徴の重要度にかかわらず分類が可能



AntTree による識別の例
Crawler ログ、Non-crawler ログそれぞれランダムに 50 個ずつサンプリングしたものを実験用データセットとしている

まとめと今後の課題

- まとめ
  - アリの仕組みに着想を得たクラスタリング手法をクローラに適用
  - 実際にネットワークから収集されたデータを用いて評価実験
    - 従来手法と比較して高い精度でクローラの識別が可能
    - データ全体として、比較的サイズの小さなクラスタを高い精度で分類可能

- 今後の課題
  - AntTree に関して通信の傾向の変化を考慮した解析、実験
  - 統計的な情報を用いた識別に関しての検討
    - 同一の IP アドレスや AS からの通信を一つの集合として、統計的な特徴を抽出

補足資料: クローラとの通信ログ

- Google クローラとの通信ログ
  - Google の公開情報（送信元 IP アドレス、UserAgent）から容易に判別可能
  - 本研究では有名クローラとして Google クローラを採用

- Google 以外のクローラとの通信ログ
  - Microsoft、Baidu など
  - 研究者、技術者により手動で判別されたものを正解ラベルとして使用