CSS2012

マルウェアの部分コードによる類似度判定と機能推定

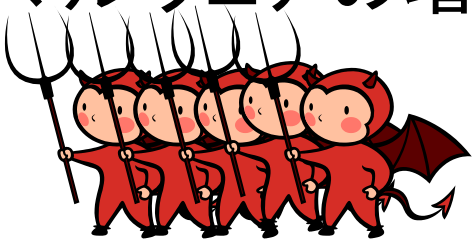
大久保諒† 森井昌克† 伊沢亮一††
井上大介†† 中尾康二††

†神戸大学大学院工学研究科
††独立行政法人情報通信研究機構

研究の背景

背景

マルウェアの増加 ⇒ 解析者の負担増大



目的

高速なマルウェアの類似度導出
マルウェアの機能推定

提案手法の概要

提案手法の概要

マルウェアの類似度導出

類似度を用いることでマルウェア間の相関関係を数値化

➡ バイトコードの分布を用いることで高速な導出

マルウェアの機能推定

類似度に基づき機能毎にポイントを与えることによって機能を推定

既存の類似度導出法 ➡

既存の類似度判定法

マルウェア間の類似度導出手法

n-gram

LCS

いずれも文字列から共通する部分を抜き出す手法

利点

マルウェア毎にどこが異なっているかが明確に示せる

欠点

亜種とされるマルウェアを探しだすのに時間がかかる

・岩村誠ら, “機械語命令列の類似性に基づく自動マルウェア分類システム,” 情処論, 2010

・東結香ら, “コードに基づいたマルウェアの機能推定に関する研究,” SCIS, 2011

提案する類似度導出法

提案する類似度導出法

0x0fの後のバイトコードの分布をもとに類似度を導出する

利点

- マルウェアの急増が問題となっている近年において
  高速に解析済みマルウェアから類似したものを
  抽出できる

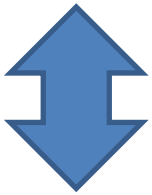
- 逆アセンブルが不可能なものに対しても一定の
  評価が可能

既存研究との比較

既存研究との比較

既存研究

1対1の比較において精度の高い類似度の導出を目的とする



提案手法

1対多の比較で迅速に類似した検体を見つけ出す

類似度導出手順

類似度導出までの手順

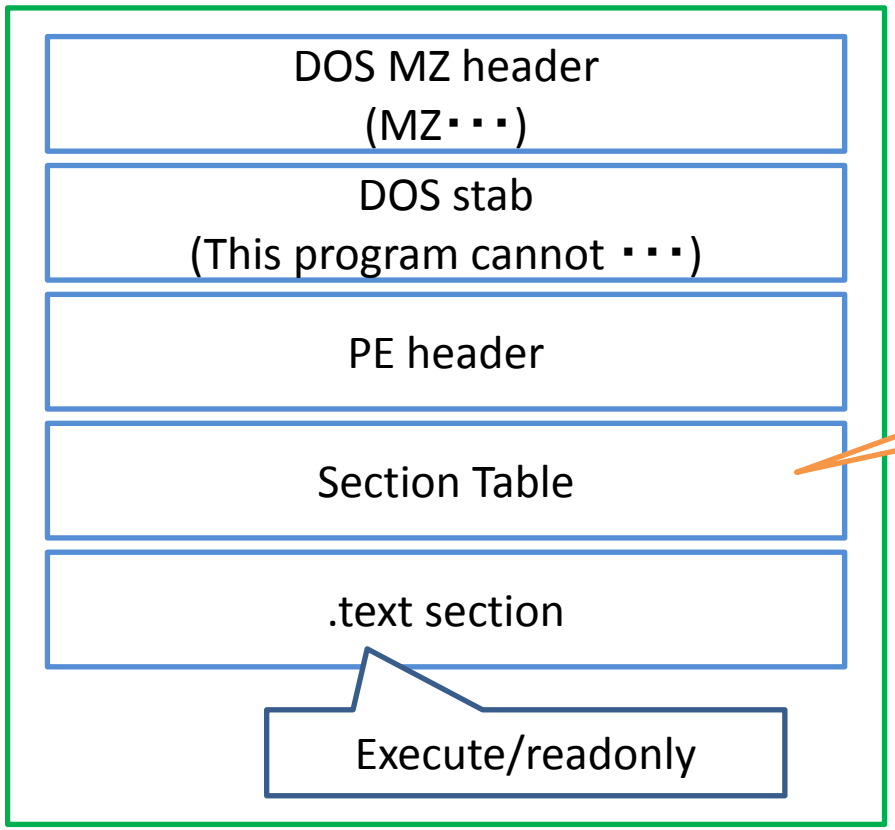
1. Section Tableから実行可能セクションを特定

2. 該当セクションから0x0fの後のバイトを抽出し，
   バイトコードの分布を取得する

3. 取得したバイトコードの分布から正規化相互相関
   を用いて比較するマルウェア間の類似度を導出

PEファイル構造

PEファイル構造



セクションに関する情報を格納している

section tableから実行コード部分を抽出

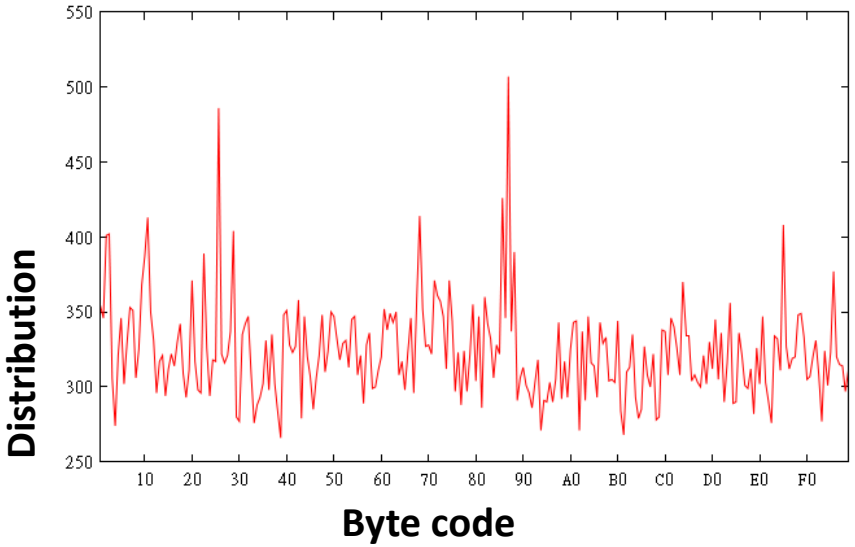
バイトコードの出現頻度を取得（0x00と0xffは除く）

類似度の偏差を大きくする

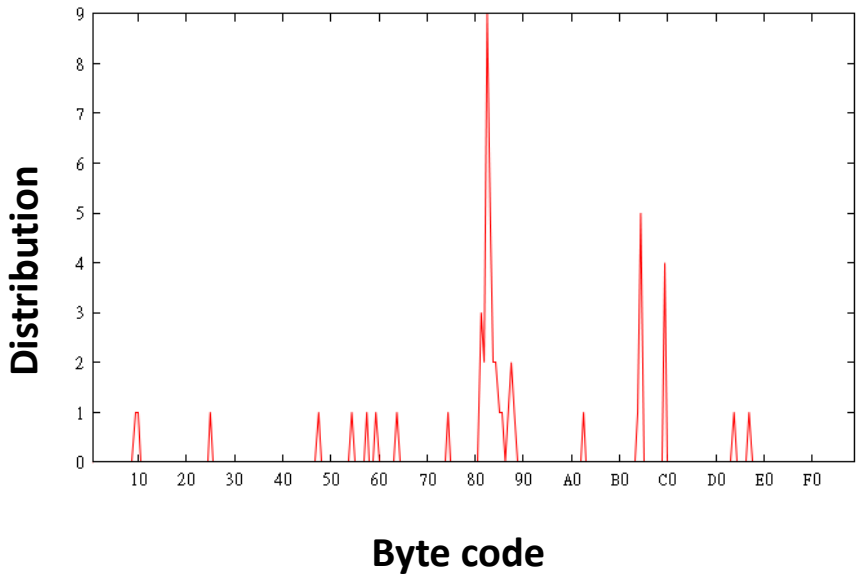
2byteオペコードに着目

2byteオペコード ➡ 0x0fから始まる

すべてのバイトコード分布



0x0fの後のバイトコード分布



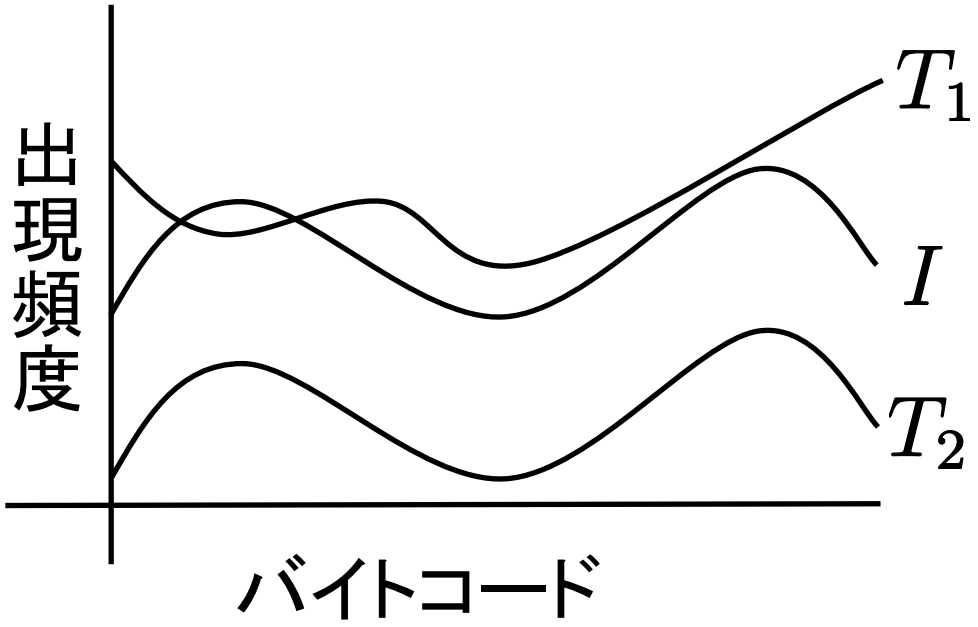
➡ より大きな偏差を持つ類似度が期待できる

正規化相互相関 ➡

正規化相互相関

分布の形状から類似性を数値化

$$R_{ZNCC} = \frac{\sum_{i=0}^{N-1}((I(i)-\overline{I})(T(i)-\overline{T}))}{\sqrt{\sum_{i=0}^{N-1}(I(i)-\overline{I})^2 \times \sum_{i=0}^{N-1}(T(i)-\overline{T})^2}}$$



	類似度
I と T_1	低
I と T_2	高

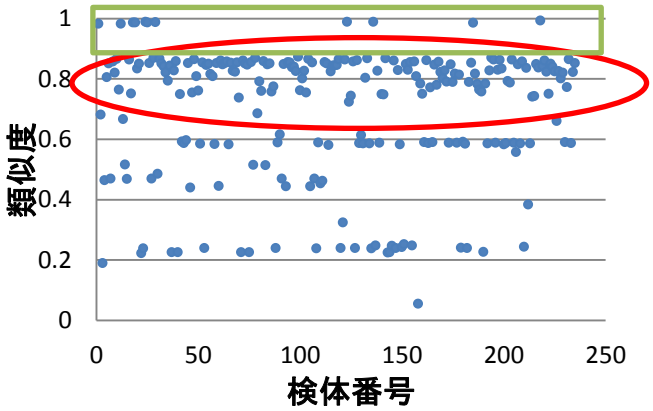
類似度導出結果

類似度導出結果

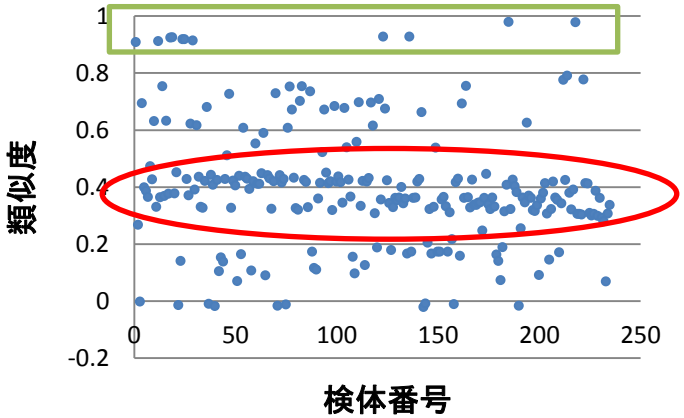
6月20日にNICTで採取されたマルウェアの内パックされていないもの230検体を対象に類似度を導出

検体の内から1検体をとりだしその他の検体との類似度を導出

科名が一致した検体



すべてのバイトコード



0x0fの後のバイトコード

0x0fの後のバイトコードの分布から導出した類似度の方がより亜種とそうでないものを明確に分類している

既存研究と提案手法の機能推定

既存研究と提案手法の機能推定

提案手法

解析済み検体から亜種とされるマルウェアを抽出したうえで，どの機能に変更が加えられている可能性があるかを提示するのが目的

既存研究

機能毎に類似度を導出することによって，検体間で共通した機能を抽出する

・東結香ら, “コードに基づいたマルウェアの機能推定に関する研究,” SCIS, 2011

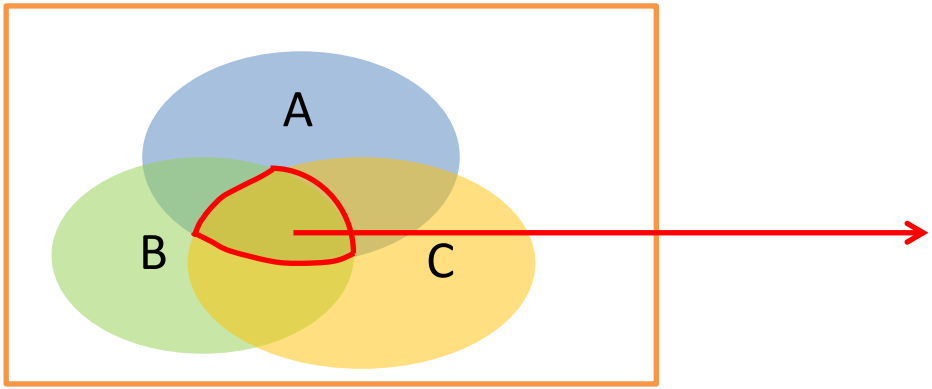
機能推定の概念

機能推定法の概念

解析対象としたマルウェアと類似性の高いマルウェアに
共通の機能がある

共通した機能は解析対象としたマルウェアも保有する
可能性が高い

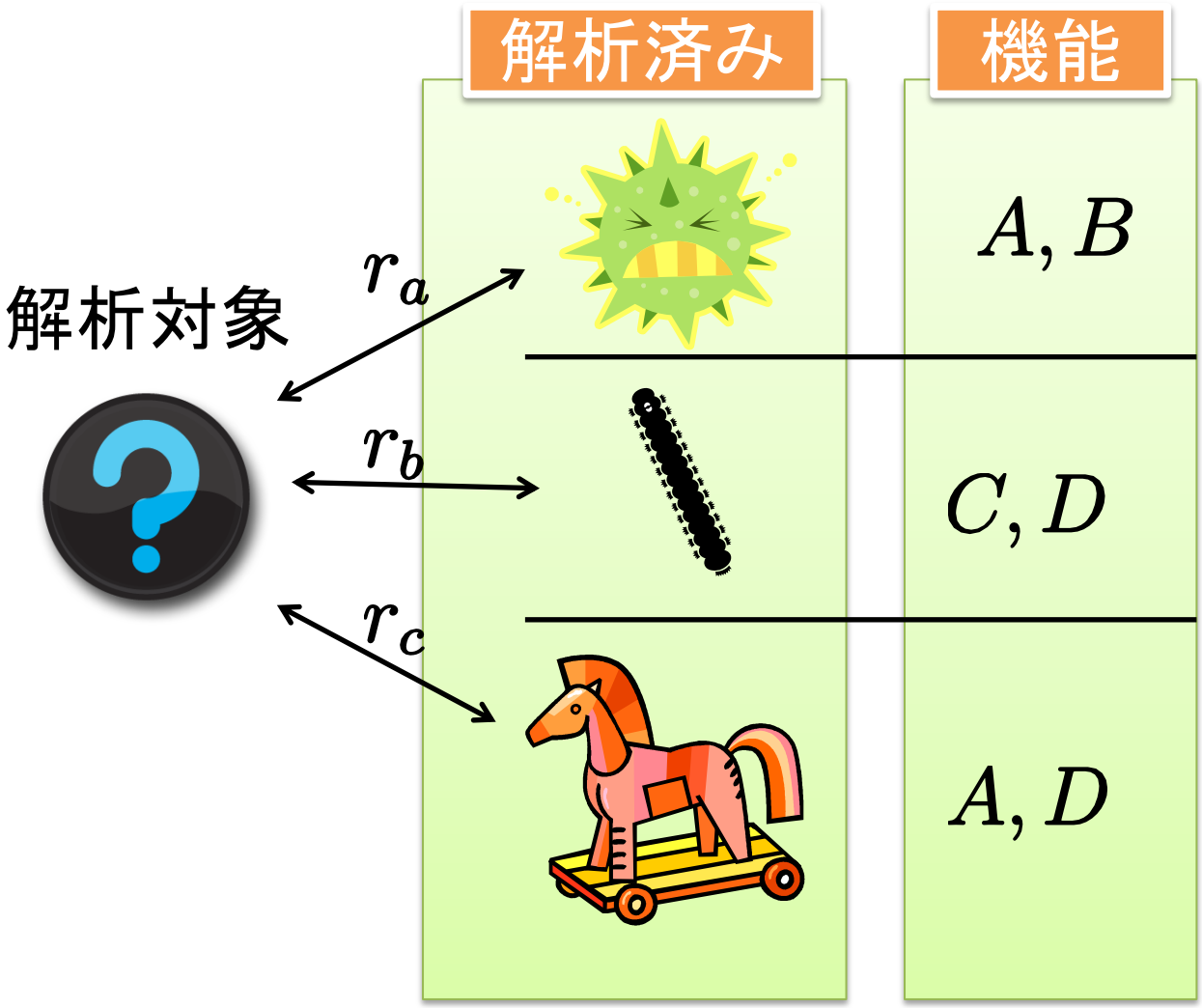
例）解析対象とするマルウェアと検体A,B,Cの類似度が高い



各検体が持つ機能群

解析対象のマルウェアが
持つ可能性の高い機能

マルウェアの機能推定



$$P(A) = F(r_a) + F(r_c)$$
$$P(B) = F(r_a)$$
$$P(C) = F(r_b)$$
$$P(D) = F(r_b) + F(r_c)$$

ポイントの高い機能を解析対象がもつ可能性が高い

機能ポイントの導出法

ポイントの算出法

類似度をもとに機能毎にポイントを加算していく

加算するポイント

$$P(k) = \frac{\sum_m |r(m)f(m,k)| r(m)f(m,k)}{\sqrt{\sum_m f(m,k)}}$$

$P(k)$: 機能 k に与えられるポイント

$r(m)$: 検体 m との類似度

$$f(m,k) = \begin{cases} 1 & (\text{検体 } m \text{ が機能 } k \text{ を保有する場合}) \\ 0 & (\text{それ以外}) \end{cases}$$

機能推定までの流れ

機能推定までの流れ

1. 解析対象とするマルウェアと解析済みマルウェアの類似度を取得する

2. 与えられた類似度から機能毎にポイントを導出

3. ポイントの高い検体から順に解析対象とする検体が持つ可能性が高い機能

機能推定の対象とした検体

機能推定実験

データベース内にある230検体を用いて機能を推定
NICTの動的解析結果を利用

機能
createService
search
copyFile
alterFile
execute
delete itself
addReg
createFile
createMutex
create,deleteReg
alterTxtFile
newHash
readFile
open,attrFile

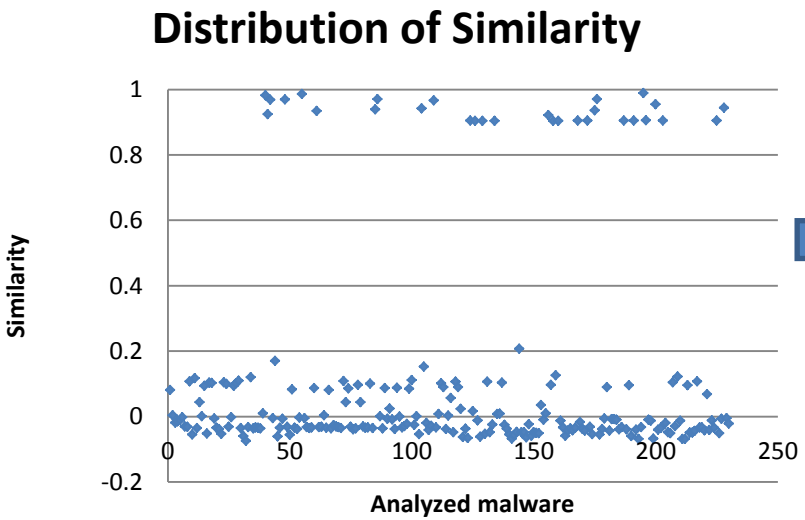
機能
openProcess
search
connect(DNS)
backdoor
alterProcess
movefile
createFile
sendMail
openwindow
connect(web)
createDir
auto start
alterFile

データベース内にあるマルウェアの機能一覧

機能推定結果

検体Aの機能推定

検体A：0ea635*



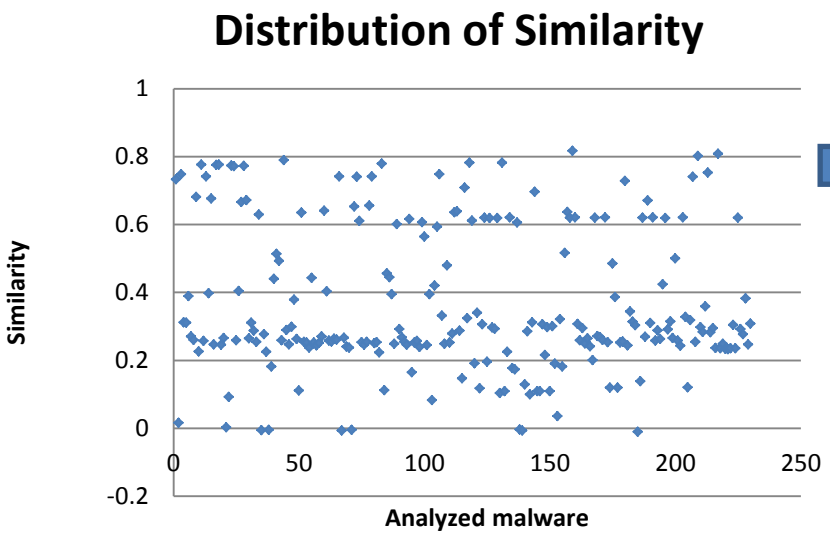
解析済みマルウェアとの類似度導出結果

createService	3.662	open,attrFile	1.388
search	2.893	openProcess	0.874
copyFile	2.341	search	0.790
alterFile	2.300	backdoor	0.771
execute	2.097	connect	0.771
deleteFile	1.911	alterProcess	0.763
alterTxtFile	1.749	createFile	0.021
addReg	1.734	movefile	0.018
newHash	1.727	sendMail	0.009
createFile	1.726	openwindow	0.003
createMutex	1.718	createDir	0.003
create,delReg	1.686	connect(web)	0.001
readFile	1.420	auto start	0.000
		alterFile	-0.003

機能推定結果

検体Bの機能推定

検体B：06acb0c*



create,delReg	2.694824968	open,attrFile	1.975260084
createMutex	2.485896529	alterTxtFile	1.974307418
createFile	2.476608583	readFile	1.972594359
addReg	2.450036323	openProcess	1.845620392
execute	2.413946806	search	1.788134163
connect	2.351873406	openwindow	0.950161253
backdoor	2.351873406	createService	0.809335896
alterProcess	2.345044456	movefile	0.725387543
deleteFile	2.209506233	createFile	0.663090661
alterFile	2.170214001	sendMail	0.642905834
copyFile	2.119917723	createDir	0.501532304
newHash	2.073169659	auto start	0.431885553
search	1.981075361	alterFile	0.243784116
		connect(web)	0.036322805

解析済みマルウェアとの類似度導出結果

中村らの研究との連携

中村らの研究との連携

中村らの手法

アンパック手法の提案

提案手法

アンパック過程において，一部が欠損してたとしてもバイトコードの分布を用いることで類似度導出，機能推定が可能

まとめ

まとめ

提案手法

- バイトコードの出現頻度を用いてマルウェア間の類似度を導出
- 類似度を用いた機能推定法

考察

- 類似度を用いることによって未知の検体の機能を推定することは可能
- 現状ではヒューリスティックに保有機能を判別する必要がある

御清聴ありがとうございました