The 24th International Conference on Information and Communications Security (ICICS 2022)

◎ Session 6: Attack and Vulnerability Analysis I

DeepC2: AI-powered Covert Command and Control on OSNs

Zhi Wang, Chaoge Liu, Xiang Cui, Jie Yin, Jiaxi Liu, Di Wu and Qixu Liu

Canterbury, UK
September 2022

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

中国科学院大学
University of Chinese Academy of Sciences

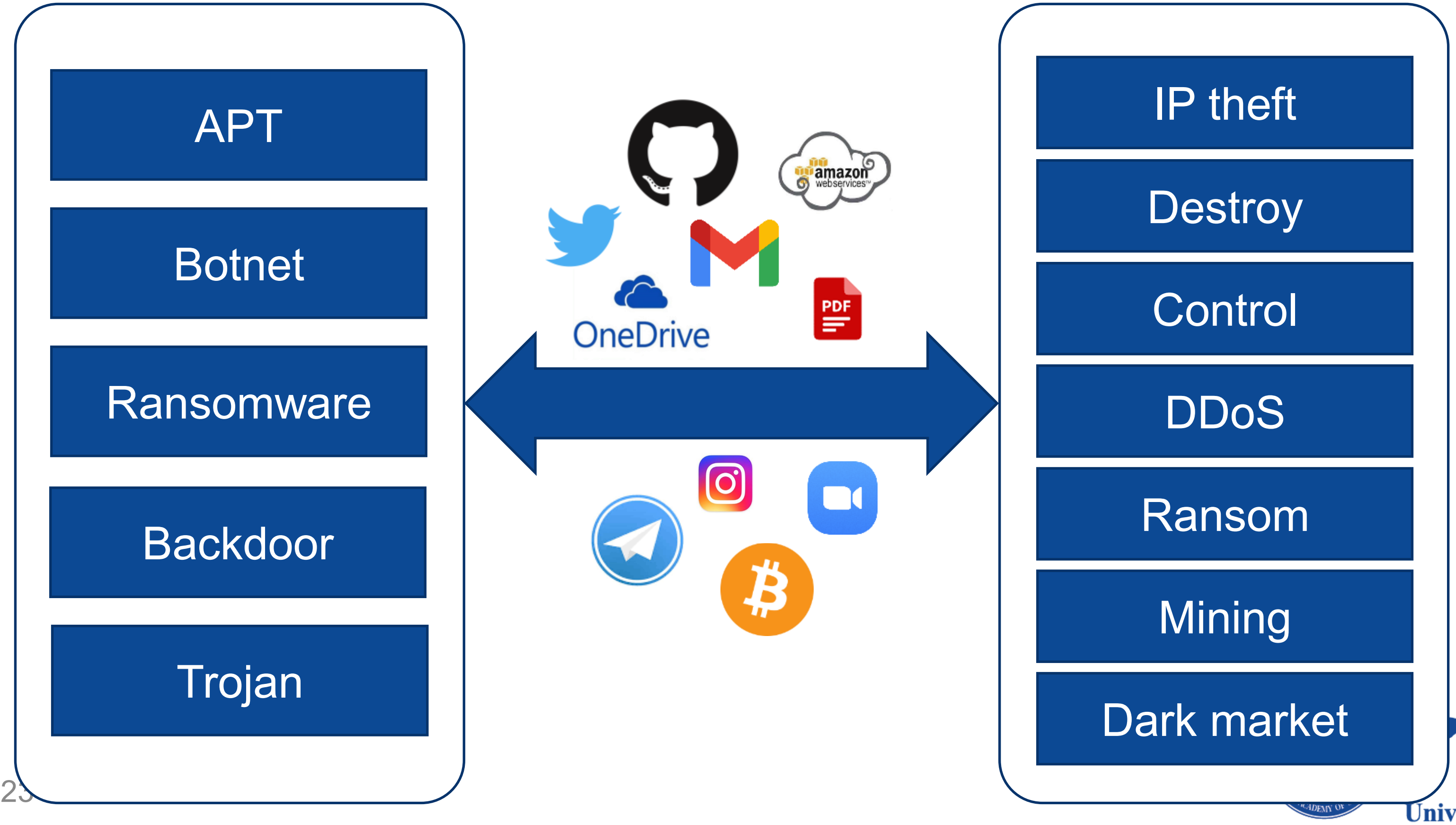
广州大学
GUANGZHOU UNIVERSITY

Contents

- Background and Motivation
- Technical Design
- Experiments and Evaluation
- Mitigation

Background

- Command and control (C&C) plays an essential role in an attack.
- During an advanced attack, the attacker needs to communicate with the malware to send the commands or payloads, and the malware also needs to send feedback to attackers.



APT

Botnet

Ransomware

Backdoor

Trojan

IP theft

Destroy

Control

DDoS

Ransom

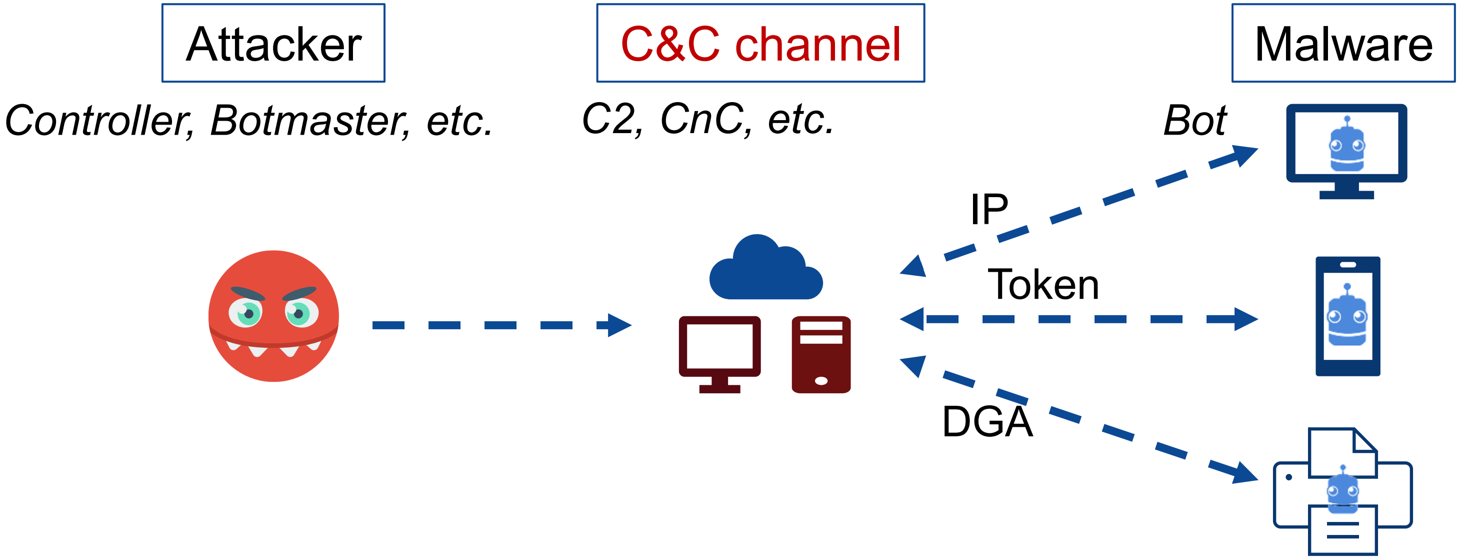
Mining

Dark market

Background

Basic structure of a C&C communication

- There are three main components in a C&C communication: the attacker, C&C channel, and malware.
- The attacker publishes the commands to the channel, and the malware fetches them.
- The process for the malware to find the commands is **addressing**.



中国科学院大学
University of Chinese Academy of Sciences

Background

Development of C&C communication

Eggdrop Bot

IRC

1993

Benign application, still in maintenance

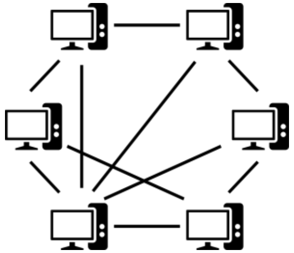
GTBot

IRC

1998

First well-known botnet

Slapper



P2P

2002

First P2P botnet

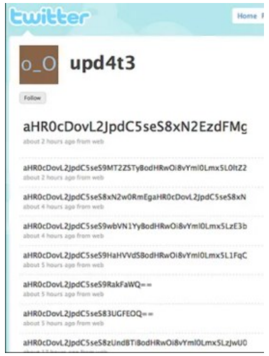
Bobax

HTTP

2004

First HTTP botnet

upd4t3



o_O upd4t3

aHR0cDovL2JpdC5seS8xN2EzdFMg
about 2 hours ago from web

aHR0cDovL2JpdC5seS9MT2Z5TyBodHRwOi8vYml0Lmx5L0ltZ2
about 2 hours ago from web

aHR0cDovL2JpdC5seS8xN2w0RmEgaHR0cDovL2JpdC5seS8xN
about 4 hours ago from web

aHR0cDovL2JpdC5seS9wbVN1YyBodHRwOi8vYml0Lmx5LzE3b
about 4 hours ago from web

aHR0cDovL2JpdC5seS9HaHVVdSBodHRwOi8vYml0Lmx5L1FqC
about 5 hours ago from web

aHR0cDovL2JpdC5seS9RakFaWQ==
about 5 hours ago from web

aHR0cDovL2JpdC5seS83UGFEOQ==
about 5 hours ago from web

aHR0cDovL2JpdC5seS8zUndBTiBodHRwOi8vYml0Lmx5LzJwU0

Web 2.0

2009

First Web-based C&C

Glupteba

Blockchain

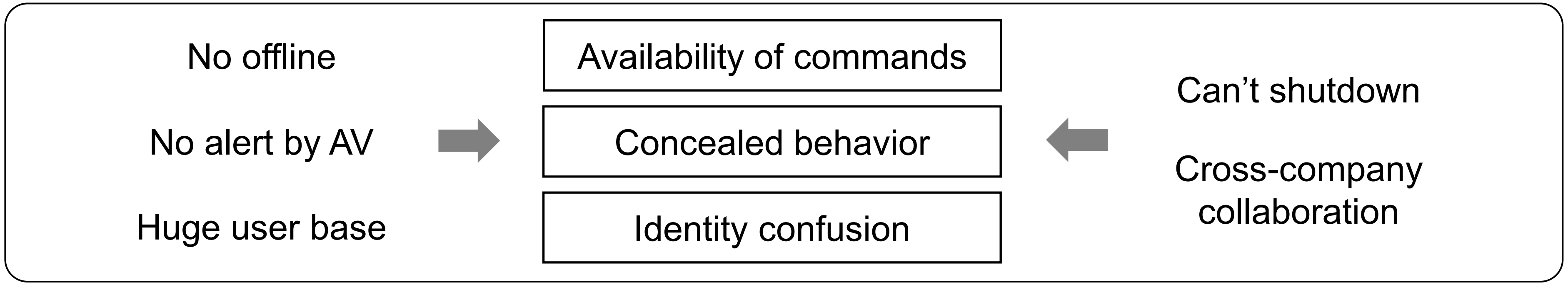
2019

First blockchain C&C

- Single point failure
- Sybil pollution attack

Background

Advantages of using online social networks (OSNs)



No offline

No alert by AV

Huge user base

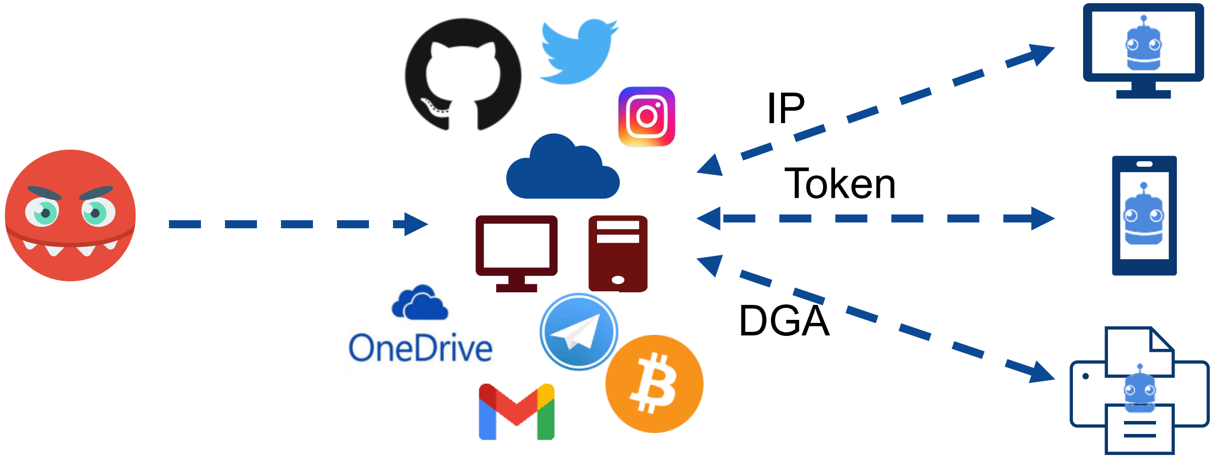
Availability of commands

Concealed behavior

Identity confusion

Can’t shutdown

Cross-company collaboration



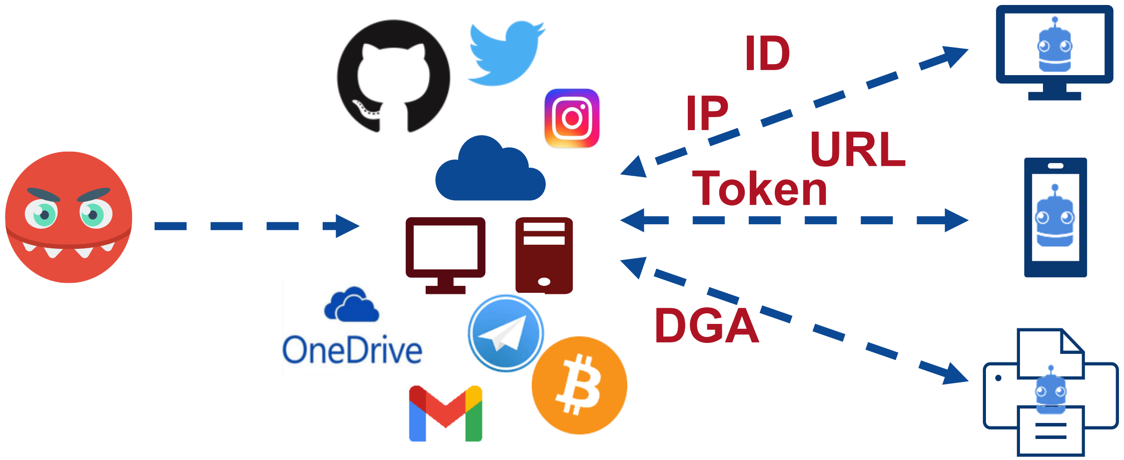
Year	Name	Platform
2009	upd4t3	Twitter, Tumblr
2014	Garybot	Twitter
2015	Hammertoss	Twitter, GitHub
2015	MiniDuke	Twitter
2017	ROKRAT	Twitter, Yandex
2017	PlugX	Pastebin
2018	Comnie	GitHub, Blogspot
2018	HeroRat	Telegram
2019	DarkHydrus	Google Drive
2019	Glupteba	Bitcoin
2019	Pony	Bitcoin
2019	IPStorm	IPFS
2020	Turla	Gmail

Background

However, there are also two problems.

1

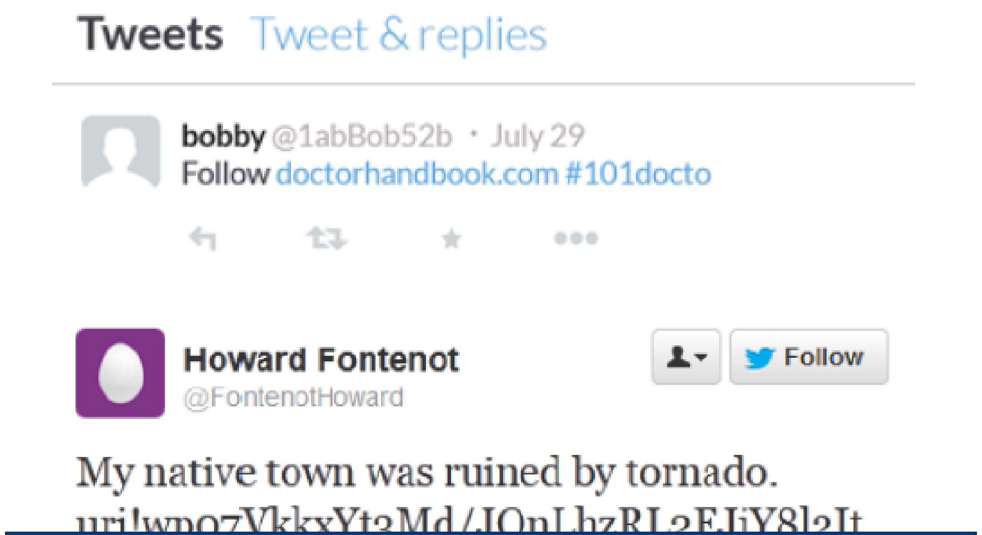
- The malware has two addressing methods.
  - Static methods like IP, ID, URL, Token, etc.
  - Dynamic generation algorithms (DGAs).
- They are **reversible**.
  - Defenders can block the accounts before commands are published.



How to eliminate reversible hardcoding?

2

- The commands are published on OSNs.
  - Plain text, encoded form, encrypted form, etc.
- They are **abnormal contents**.
  - The abusive behavior may trigger restrictions on the accounts and contents.



bobby @1abBob52b · July 29
Follow doctorhandbook.com #101docto

Howard Fontenot
@FontenotHoward

My native town was ruined by tornado.
uri!wpo7VkkxYt3Md/JOnLhzRL2EJiY8l2It

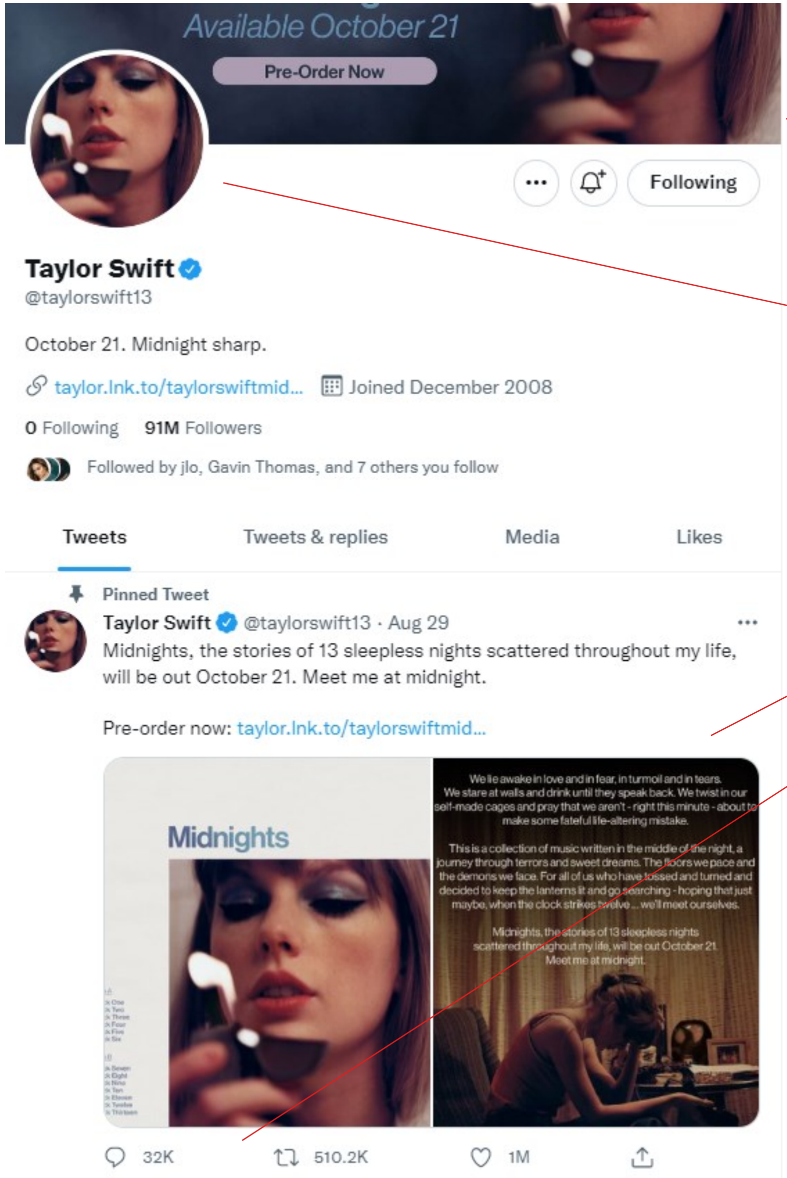
How to eliminate abnormal content?

University of Chinese Academy of Sciences

Technical Design

	Problem	Idea	Method
1	Reversible hardcoding	Irreversible methods	Neural network
2	Abnormal content	Readable content	Data augmentation & hash collision

Technical Design



Available October 21

Pre-Order Now

··· Following

Taylor Swift

@taylorswift13

October 21. Midnight sharp.

taylor.lnk.to/taylorswiftmid... Joined December 2008

0 Following 91M Followers

Followed by jlo, Gavin Thomas, and 7 others you follow

Tweets | Tweets & replies | Media | Likes

Pinned Tweet

Taylor Swift @taylorswift13 · Aug 29

Midnights, the stories of 13 sleepless nights scattered throughout my life, will be out October 21. Meet me at midnight.

Pre-order now: taylor.lnk.to/taylorswiftmid...

Midnights

We lie awake in love and in fear, in turmoil and in tears. We stare at walls and drink until they speak back. We twist in our self-made cages and pray that we aren't - right this minute - about to make some fateful life-altering mistake.

This is a collection of music written in the middle of the night, a journey through terrors and sweet dreams. The floors we pace and the demons we face. For all of us who have tossed and turned and decided to keep the lanterns lit and go searching - hoping that just maybe, when the clock strikes twelve ... we'll meet ourselves.

Midnights, the stories of 13 sleepless nights scattered throughout my life, will be out October 21. Meet me at midnight.

32K 510.2K 1M

Header photo

Avatar

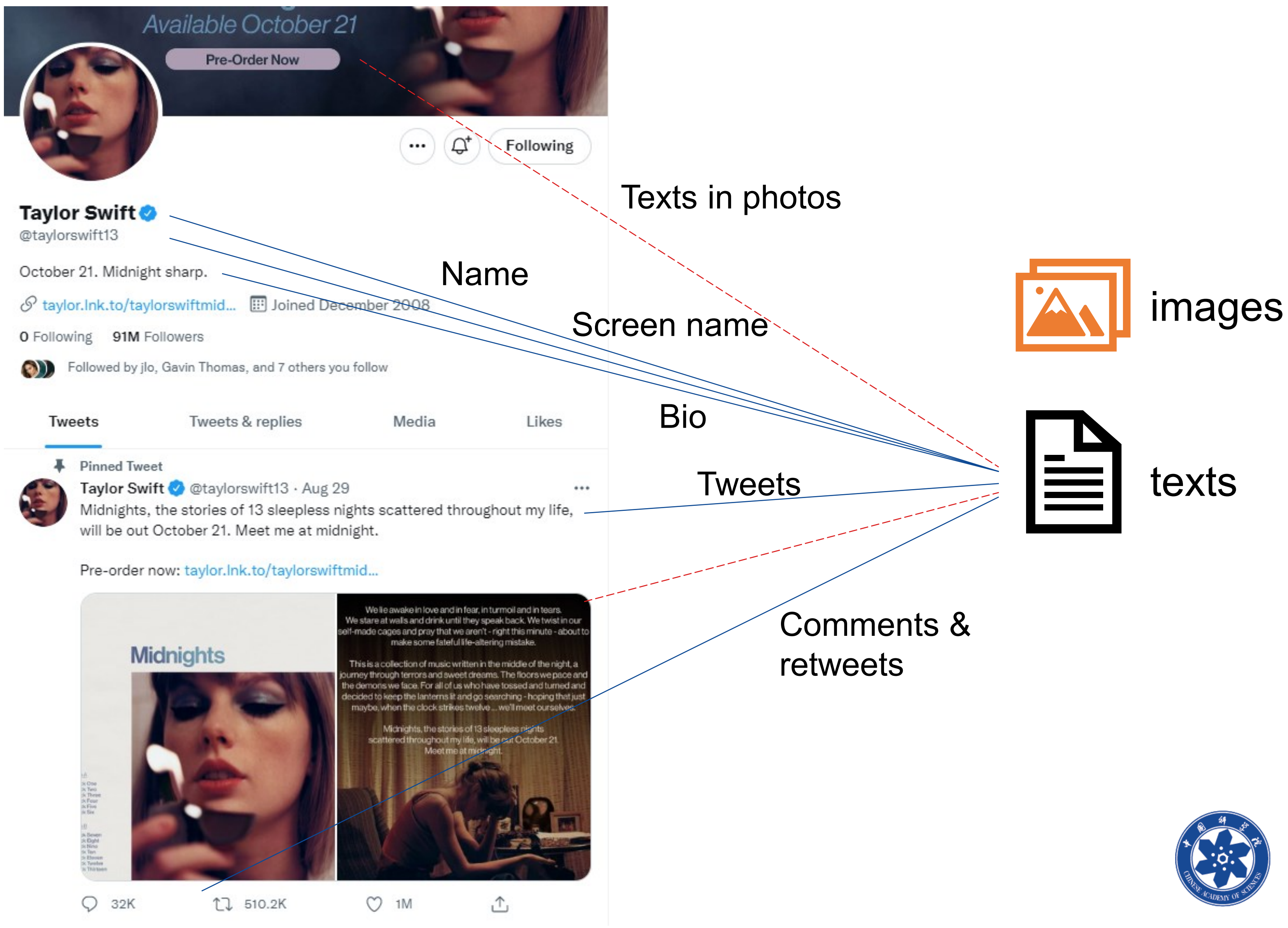
Shared photos

Comments & retweets with photos

images

texts

Technical Design



Available October 21

Pre-Order Now

Following

Taylor Swift
@taylorswift13

October 21. Midnight sharp.

taylor.lnk.to/taylorswiftmid... Joined December 2008

0 Following 91M Followers

Followed by jlo, Gavin Thomas, and 7 others you follow

Tweets | Tweets & replies | Media | Likes

Pinned Tweet

Taylor Swift @taylorswift13 · Aug 29

Midnights, the stories of 13 sleepless nights scattered throughout my life, will be out October 21. Meet me at midnight.

Pre-order now: taylor.lnk.to/taylorswiftmid...

Midnights

We lie awake in love and in fear, in turmoil and in tears. We stare at walls and drink until they speak back. We twist in our self-made cages and pray that we aren't - right this minute - about to make some fateful life-altering mistake.

This is a collection of music written in the middle of the night, a journey through terrors and sweet dreams. The floors we pace and the demons we face. For all of us who have tossed and turned and decided to keep the lanterns lit and go searching - hoping that just maybe, when the clock strikes twelve ... we'll meet ourselves.

Midnights, the stories of 13 sleepless nights scattered throughout my life, will be out October 21. Meet me at midnight.

32K 510.2K 1M

Texts in photos

Name

Screen name

Bio

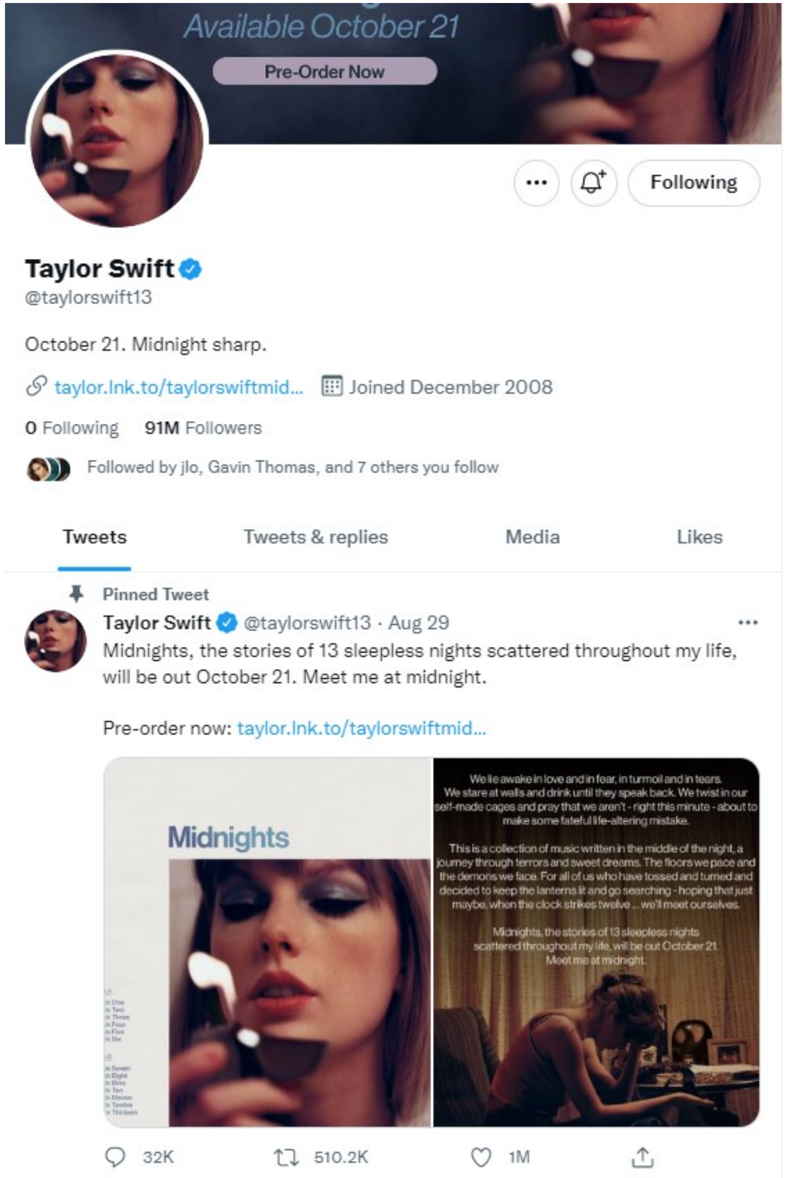
Tweets

Comments & retweets

images

texts

Technical Design



Available October 21

Pre-Order Now

··· Following

Taylor Swift
@taylorswift13

October 21. Midnight sharp.

taylor.lnk.to/taylorswiftmid... Joined December 2008

0 Following **91M** Followers

Followed by jlo, Gavin Thomas, and 7 others you follow

Tweets | Tweets & replies | Media | Likes

Pinned Tweet

Taylor Swift @taylorswift13 · Aug 29 ···

Midnights, the stories of 13 sleepless nights scattered throughout my life, will be out October 21. Meet me at midnight.

Pre-order now: taylor.lnk.to/taylorswiftmid...

Midnights

We lie awake in love and in fear, in turmoil and in tears. We stare at walls and drink until they speak back. We twist in our self-made cages and pray that we aren't - right this minute - about to make some fateful life-altering mistake.

This is a collection of music written in the middle of the night, a journey through terrors and sweet dreams. The floors we pace and the demons we face. For all of us who have tossed and turned and decided to keep the lanterns lit and go searching - hoping that just maybe, when the clock strikes twelve ... we'll meet ourselves.

Midnights, the stories of 13 sleepless nights scattered throughout my life, will be out October 21. Meet me at midnight.

32K 510.2K 1M

1. Using a neural network to recognize the images

   images

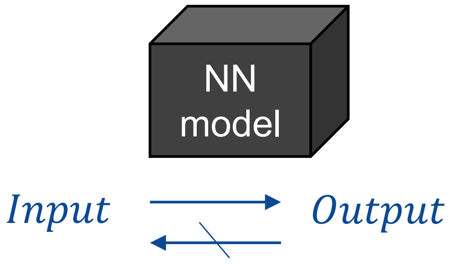
   texts

2. Hiding commands into readable contents

Neural Network Model

Why neural networks?

Reversible hardcoding?

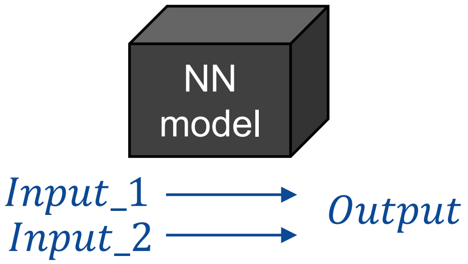


NN model

$$\mathit{Input} \;\underset{\not\longleftarrow}{\longrightarrow}\; \mathit{Output}$$

The calculation of neural networks is hard to reverse. Combined with intentionally introduced losses, it is hard to get attacker’s identifiers in advance.

Compressed images?



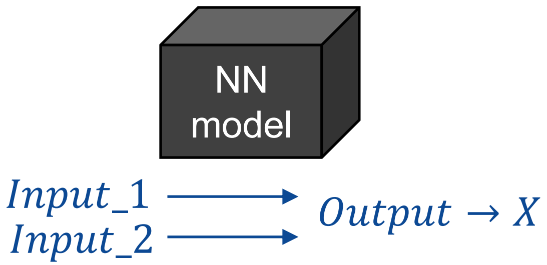
NN model

$$\begin{array}{l} \mathit{Input\_1} \longrightarrow \\ \mathit{Input\_2} \longrightarrow \end{array} \mathit{Output}$$

$$\mathit{Input\_1} \approx \mathit{Input\_2}$$

Neural network is fault-tolerance that similar inputs will generate similar outputs.

Unknown avatars?



NN model

$$\begin{array}{l} \mathit{Input\_1} \longrightarrow \\ \mathit{Input\_2} \longrightarrow \end{array} \mathit{Output} \rightarrow X$$

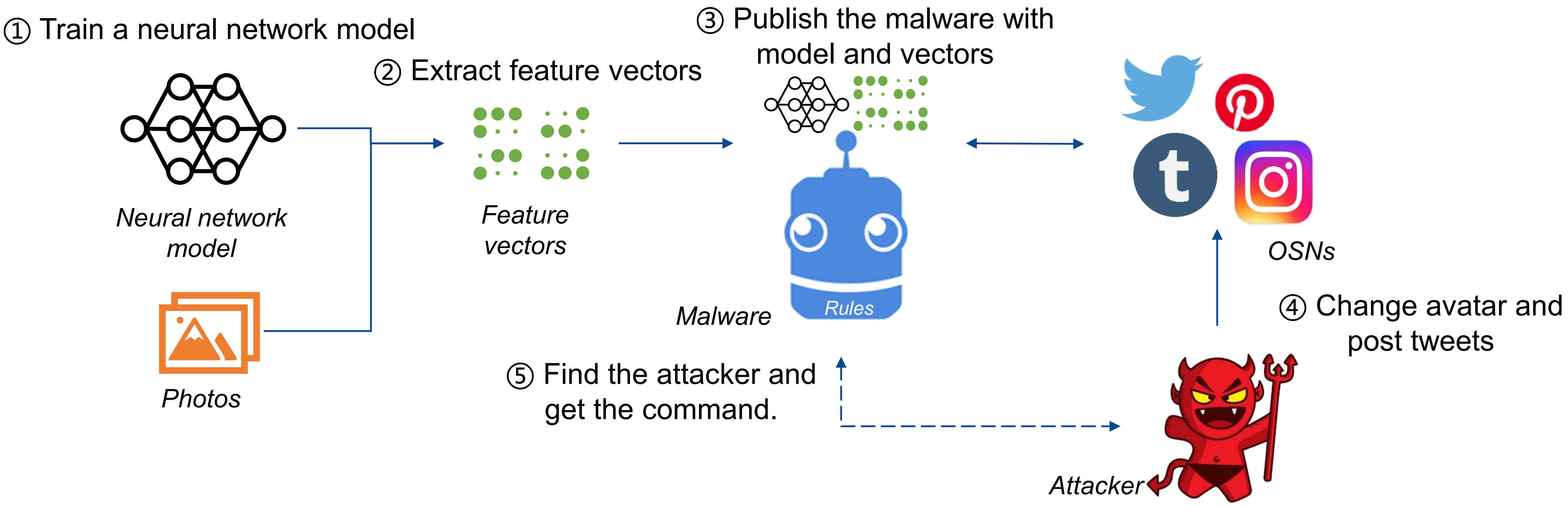
$$\mathit{Input\_1}, \mathit{Input\_2} \in X$$

Neural network has a good generalization ability. It can recognize the attacks accurately and not mistakenly identify someone else as the attacker.

Neural Network Model

How to use it?

Twitter, tweets, and avatars

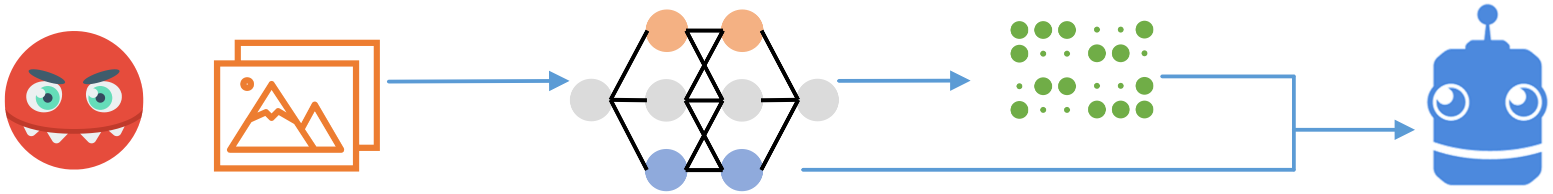


① Train a neural network model
② Extract feature vectors
③ Publish the malware with model and vectors
Neural network model
Feature vectors
Photos
Malware
Rules
OSNs
④ Change avatar and post tweets
⑤ Find the attacker and get the command.
Attacker

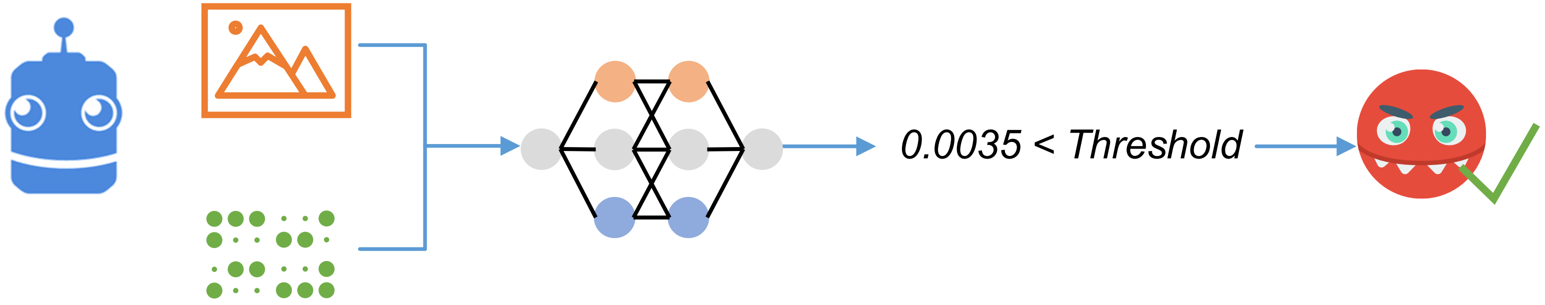
Neural Network Model

How to use it?

Attacker

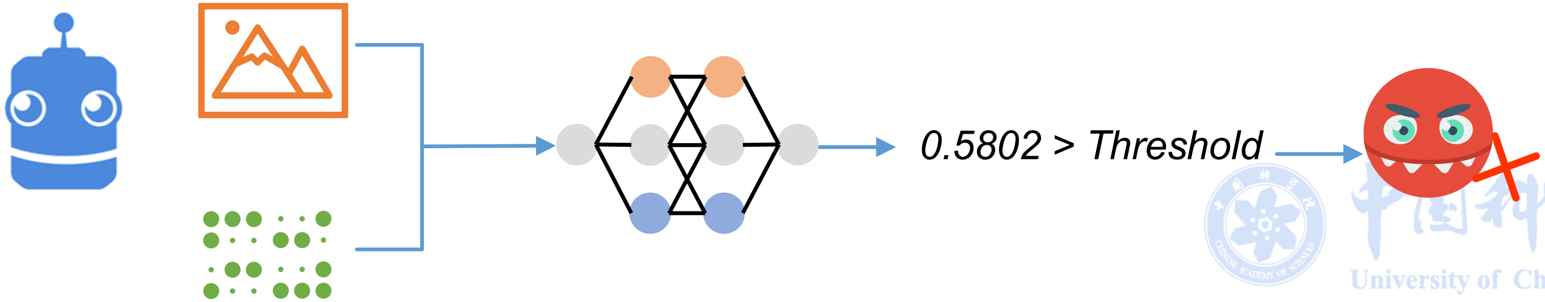


(a) The attackers use the model to extract feature vectors from pictures



Malware

(b) The malware uses the model to identify the attacker.
If the distance of inputs is below a threshold, the attacker is found.

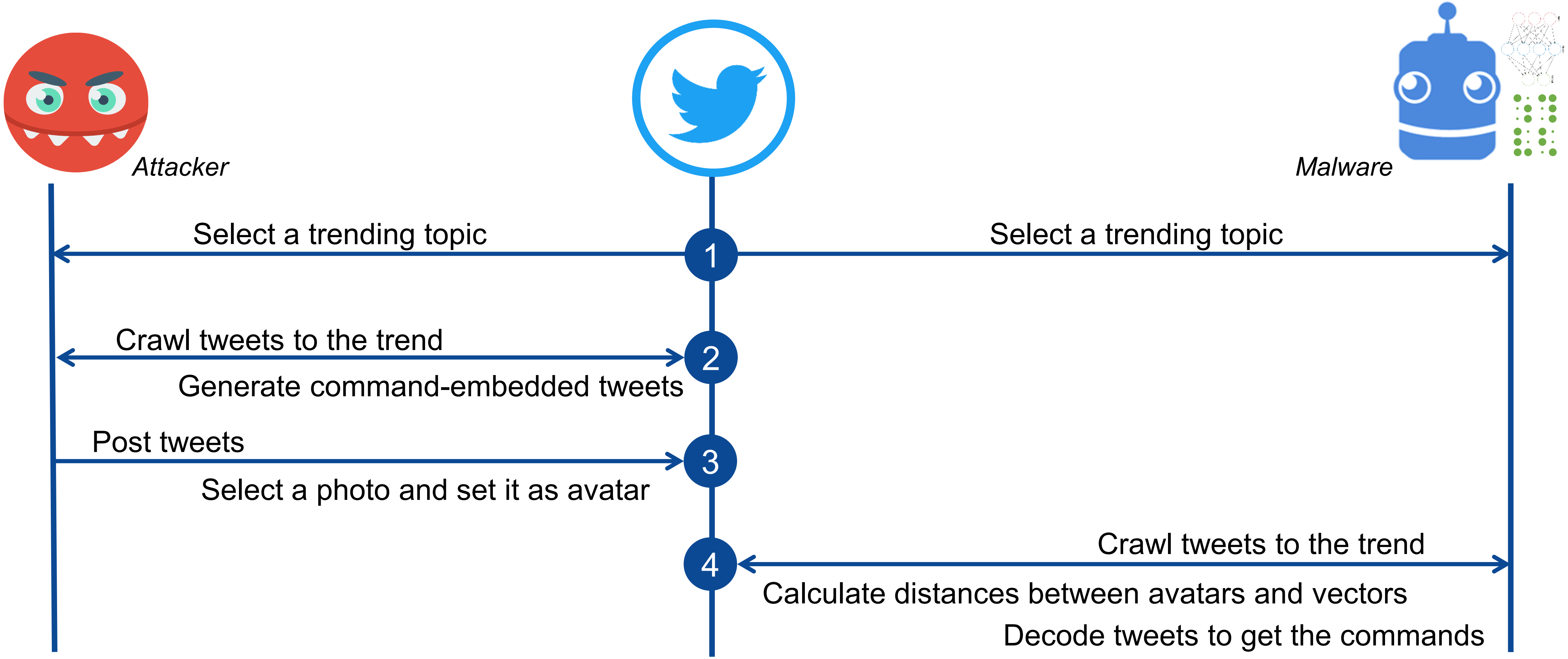


中国科学院大学
University of Chinese Academy of Sciences

Technical Design

How do they meet?

Twitter Trends



Attacker

Malware

1. Select a trending topic / Select a trending topic
2. Crawl tweets to the trend
   Generate command-embedded tweets
3. Post tweets
   Select a photo and set it as avatar
4. Crawl tweets to the trend
   Calculate distances between avatars and vectors
   Decode tweets to get the commands

Twitter Trends

Why Twitter Trends?

Meeting point

It is not easy for malware to find an attacker among Twitter users. Twitter Trends provides a meeting point for them.

Identity confusion

Twitter Trends contains numerous discussions on top topics. The attacker can hide among them and achieve identity confusion.

Hard to predict

Twitter Trends changes with the tweet volume and is updated every five minutes, which is not easy to predict.

Hash Collision

How to convert commands to tweets?

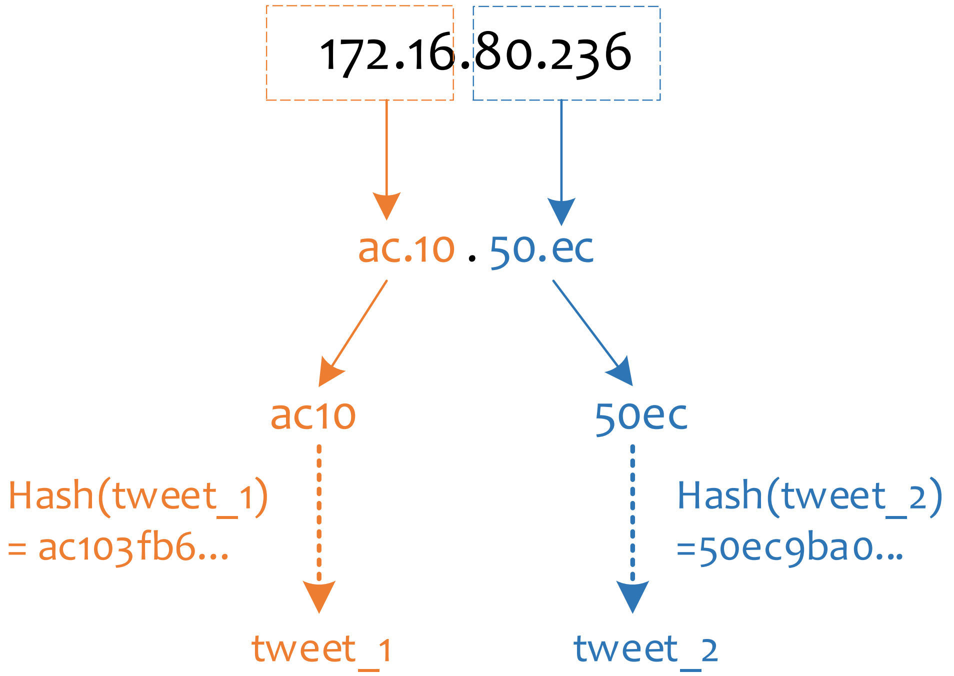
We take publishing an IP address as an example. Attackers can also publish other commands in this way.

Step 1: Split the command into two-byte chunks.

Step 2: Change them to hex form.

Step 3: Calculate the hash of the tweets and compare the first two bytes with a command chunk.

Step 4: Collect all the collided tweets and post them on Twitter.

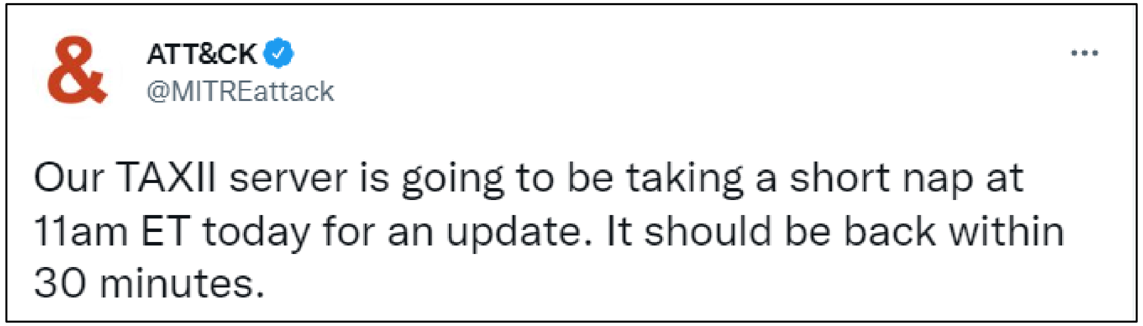


University of Chinese Academy of Sciences

Data Augmentation

How to generate tweets for hash collision?

- Data augmentation is a technique to solve the insufficiency of training data.
- Easy data augmentation (EDA) uses four ways to get new sentences:
  - Synonym Replacement (SR)
  - Random Insertion (RI)
  - Random Swap (RS)
  - Random Deletion (RD)



ATT&CK

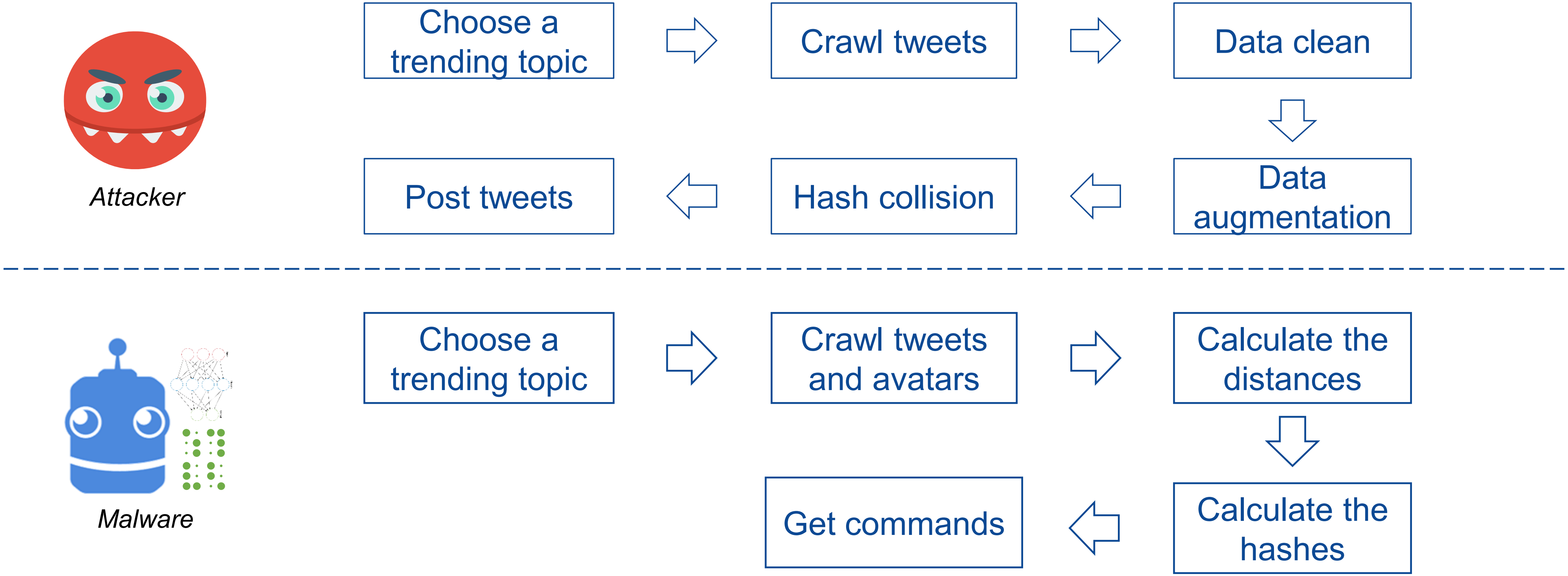
@MITREattack

Our TAXII server is going to be taking a short nap at 11am ET today for an update. It should be back within 30 minutes.

Op.	Sentence
None	Our TAXII server is going to be taking a short nap at 11am ET today for an update.
SR	Our TAXII server is ***endure*** to be taking a short nap at 11am ET today for an update.
	Our TAXII server is going to be ***conduct*** a short nap at 11am ET today for an update.
RI	Our TAXII server is going to be taking a short nap at 11am ***cat sleep*** ET today for an update.
	Our TAXII server is going to be taking a short ***circuit*** nap at 11am ET today for an update.
RS	Our ***short*** server is going to be taking a ***TAXII*** nap at 11am ET today for an update.
	Our TAXII server is going to be ***today*** a short nap at 11am ET ***taking*** for an update.
RD	Our server is to be taking a short nap at 11am ET today for an update.
	Our TAXII server is going to taking short 11am ET today for an update.

Workflow

Workflow when issuing commands



Attacker

Choose a trending topic → Crawl tweets → Data clean → Data augmentation → Hash collision → Post tweets

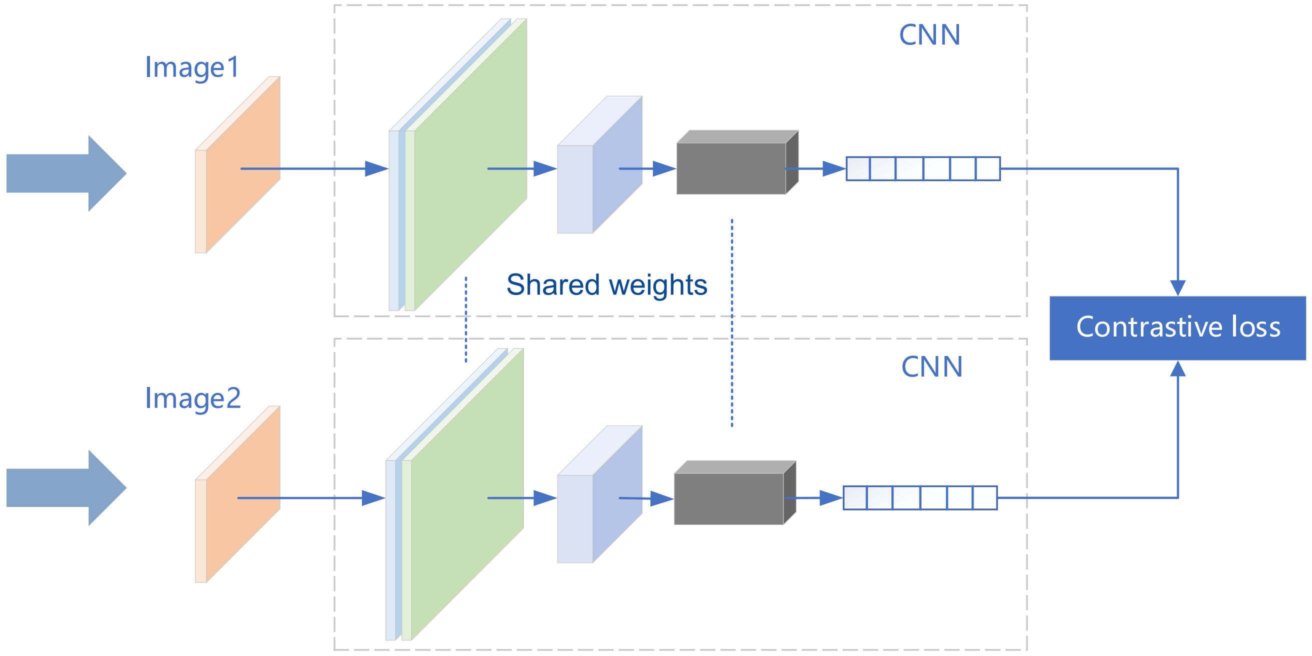
Malware

Choose a trending topic → Crawl tweets and avatars → Calculate the distances → Calculate the hashes → Get commands

Implementation

Siamese neural network

The Siamese neural network is effective in measuring the similarity between two inputs.



Contrastive loss $L = (1 - Y)\frac{1}{2}(D_w)^2 + Y\frac{1}{2}(\max(0, m - D_w))^2$

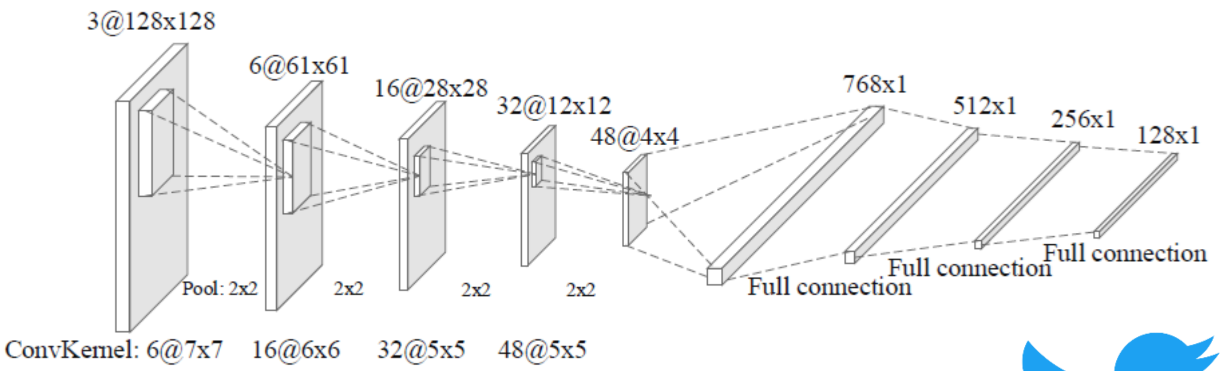
Photo

```
[0.06141704320907593, 0.11299607157707214,
0.13662077486515045, -0.13357725739479065,
...
0.17597267031669617, -0.0214485302567482,
0.04336101561784744, 0.07453791797161102]

[0.030405446887016296, 0.05502897500991821,
0.14236226677894592, -0.12090344727039337,
...
0.10791455209255219, 0.018605416640639305,
0.017460424453020096, 0.05878069996833801]

[0.06956829130649567, 0.09473420679569244,
0.15777051448822021, -0.1374780535697937,
...
0.14949743449687958, -0.0038978923112154007,
0.03145717829465866, 0.052630871534347534]
```

Feature vector

Implementation

Convolutional neural network



Layer	Input	Output	Kernel
conv1	128×128×3	122×122×6	7×7×6, 1
Tanh			
pool1	122×122×6	61×61×6	2×2×1, 2
conv2	61×61×6	56×56×16	6×6×16, 1
Tanh			
pool2	56×56×16	28×28×16	2×2×1, 2
conv3	28×28×16	24×24×32	5×5×32, 1
Tanh			
pool3	24×24×32	12×12×32	2×2×1, 2
conv4	12×12×32	8×8×48	5×5×48, 1
Tanh			
pool4	8×8×48	4×4×48	2×2×1, 2
fc1	1×768×1	1×512×1	
ReLU			
fc2	1×512×1	1×256×1	
ReLU			
output	1×256×1	1×128×1	
CNN size	**2.36MB**	**SNN size**	**2.42MB**

Twitter avatars

Label 0
200x200 & 400x400
from the same user

Same

1 : 2

Label 1
400x400 from
different users

Not the same

Experiment

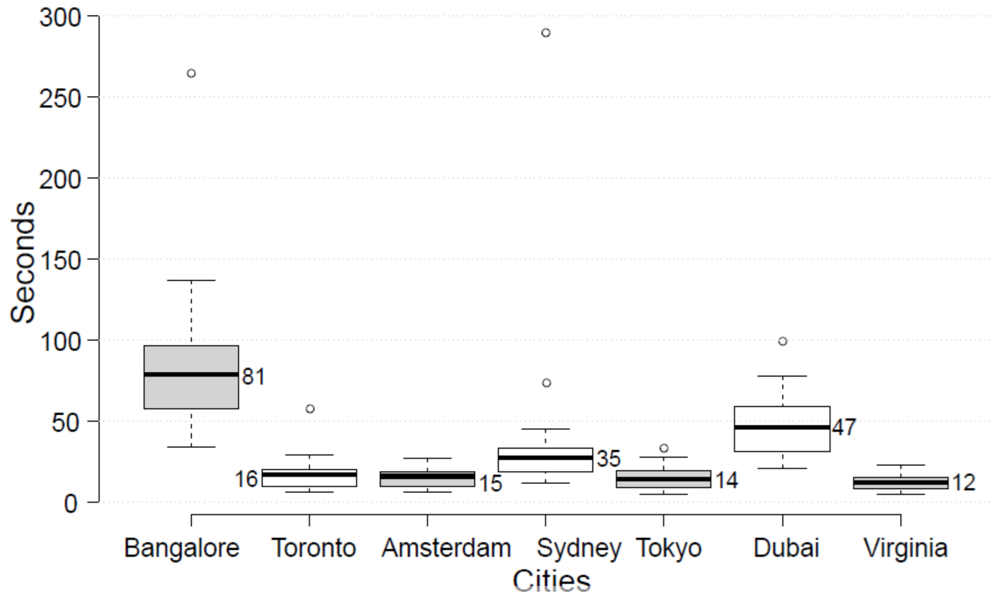
Settings

- 8 VPS (Ubuntu 18.04 x64, 1 GB ROM & 1 vCPU) to simulate the bots and attacker.
- One Twitter account to publish 47 commands.
- Last trending topic above 10K discussions from Johannesburg, South Africa.

Results

Bots' distribution and time cost for addressing

Location	**Time cost/s**		
	Avg.	Min.	Max.
Bangalore	81.51	34	267
Tokyo	14.59	5	36
Toronto	16.56	6	60
Virginia	12.13	5	23
Amsterdam	15.19	6	27
Sydney	35.26	12	292
Dubai	46.92	21	102



Time cost for addressing

Malware

Attacker

- Average time for the attacker to generate tweets and calculate hash is **13.8s**.
- All commands were obtained by the malware accurately.

Evaluation

Tweets generation

- Topic completeness after data augmentation
- Efficiency of tweets generation

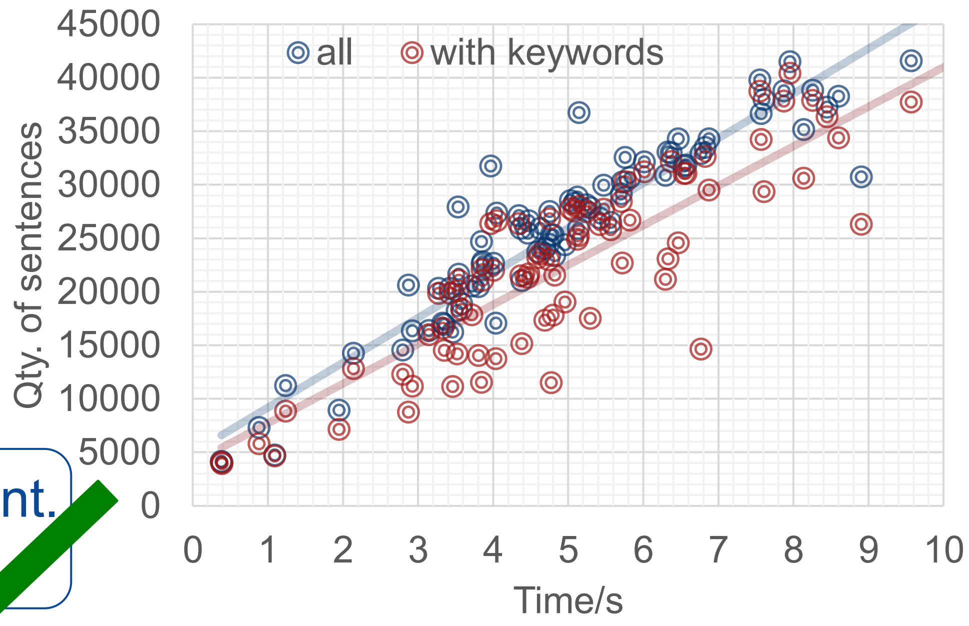
- Collect 79 trending topics from four big cities.
- Crawl 1,000 tweets per topic.
- Clean the tweets and generate 50 more sentences per tweet.

Completeness of topics in new sentences

Word(s)	Quantity	Completeness
1	55	89.54%
>1	24	77.55%

Efficiency of tweets generation

Time/s	1	2	3	5	10	15	20
Qty.	10262	14232	18202	26142	45993	65843	85694
Qty.	10K	20K	30K	50K	100K	150K	200K
Time/S	0.93	3.45	5.97	11.01	23.60	36.20	48.79



Efficiency of tweets generation

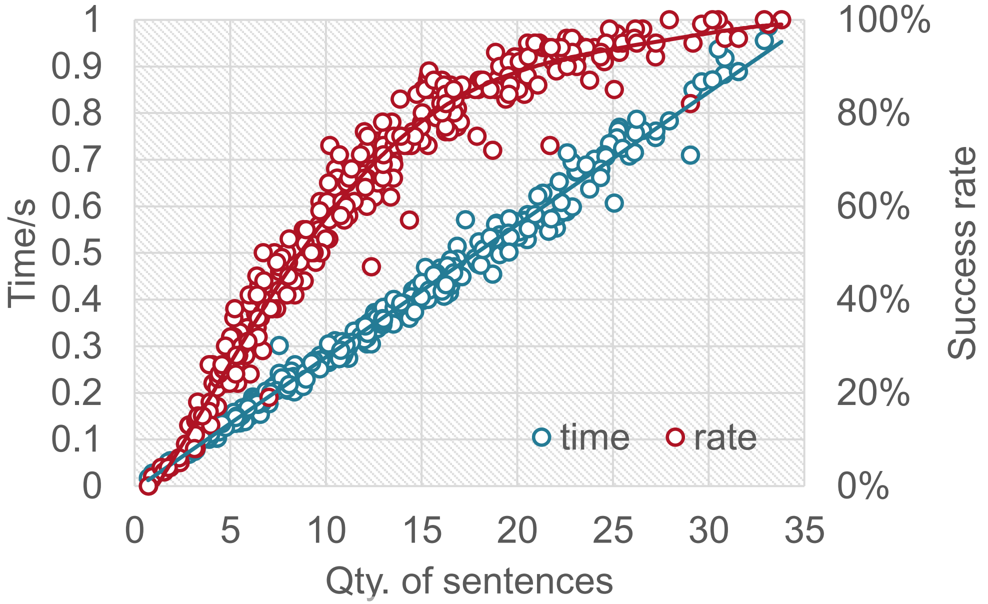
- Sentences with the complete trending word are sufficient.
- The attacker needs **3~10s** to generate the sentence

Evaluation

Hash collision

- Time cost
- Success rate

- Transformation to get enough sentences
  - Add punctuations
  - Convert cases
- > 400K sentences & 100 commands (IP)
- SHA-256, hashlib, Python, single thread



Efficiency of hash collision

- Time cost: < 1s
- Success rate:
  - 140K sentences, 75%
  - 210K, 90%
  - 330K, ~ 100%
  - 219,335 in Twitter experiment, 90.28%

✓

Evaluation

Avatar recognition

- Time cost

- 40 feature vectors & 1,000 avatars
- 11.92s for extracting vectors and calculating distances

✓

Crawling tweets

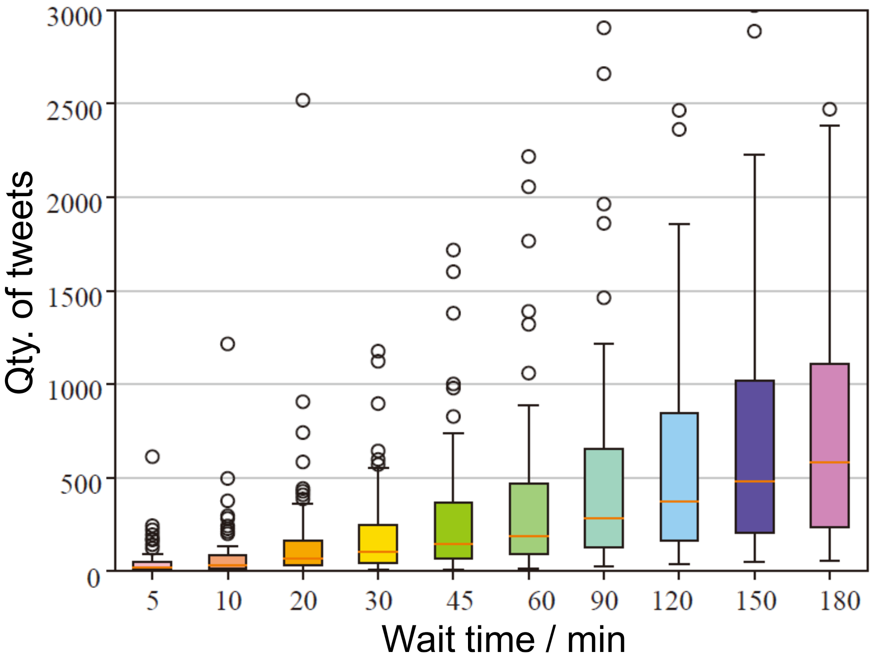
- Number of tweets with different wait times

In the Twitter experiment, after choosing a trending topic, the malware waited 5 minutes and then crawled 1,000 tweets.

- After the attacker tweets, the malware waits at different times and then crawls the tweets to find the attacker.
- 5, 10, 20, 30, 45, 60, 90, 120, 150, and 180 minutes

Wait time / min	Probability	
	1,000 tweets	3,000 tweets
5	100%	100%
60	88%	98%
180	68%	89%

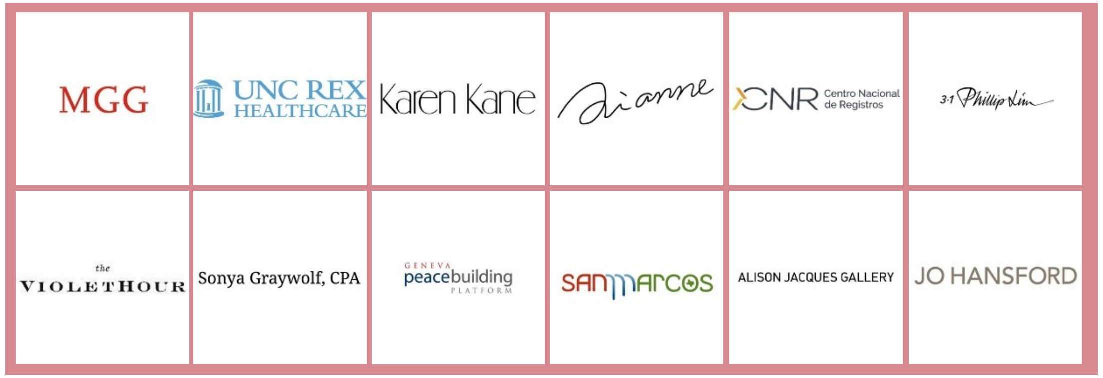
✓



Crawl number with different wait times

Security

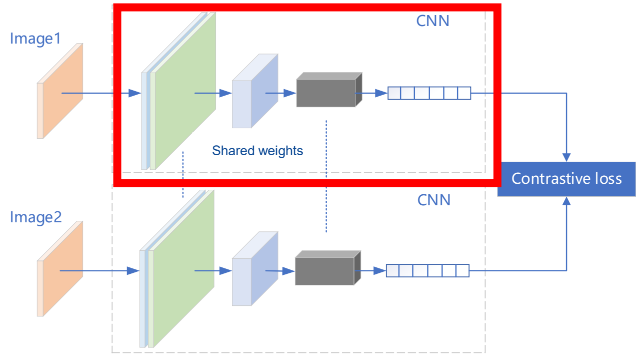
- Reuse an avatar
  - Each avatar and feature vector is used only once.
  - Only affects the malware that missed some commands.
  - Cannot affect the C&C channel.
- Collide an avatar
  - Each value comes from a continuous interval (-0.350, 0.264), which is hard to collide.
  - 600M calculations between 115,887 avatars.
    - < 0.02, 2050 pairs, 0.00031%
    - < 0.01, 81 pairs, 0.000012%
    - Mainly with logo

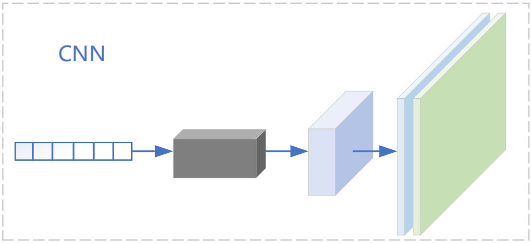


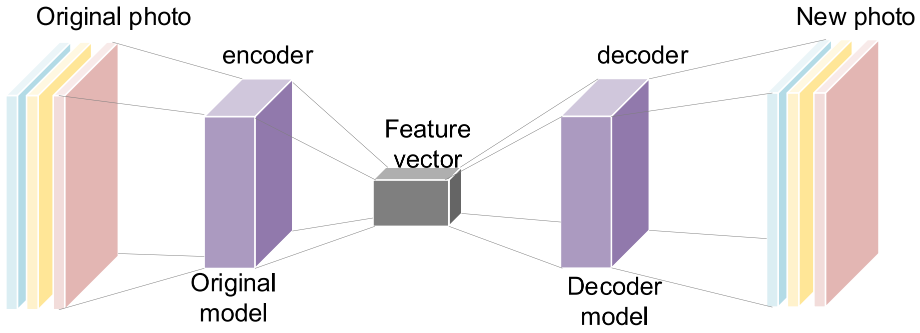
- Train a GAN
  - The avatars are too divergent to be capable of GAN.
  - Insufficient training set.

Security

- Train a decoder
  - We aimed to build a decoder that can generate an image from a vector, with a small distance between the new image and the vector.







  - Minimum distance is 0.0504, greater than the threshold.
- Attack the model
  - Only affects malware in the lab, not malware in the wild.
- Use adversarial samples
  - White-bot non-targeted adversarial attack.
  - 128 outputs are not 128 classes, and changes to the values will result in higher distances.

Enhancement

- Model
  - Feature vectors can be longer than 128.
  - More losses can be introduced in image processing.
- Addressing
  - Choosing more topics.
  - Using other fields, e.g., comment, retweet, bio.
  - Using more platforms.
- Maintenance
  - High-value accounts.
  - Behave like a human.

Countermeasures

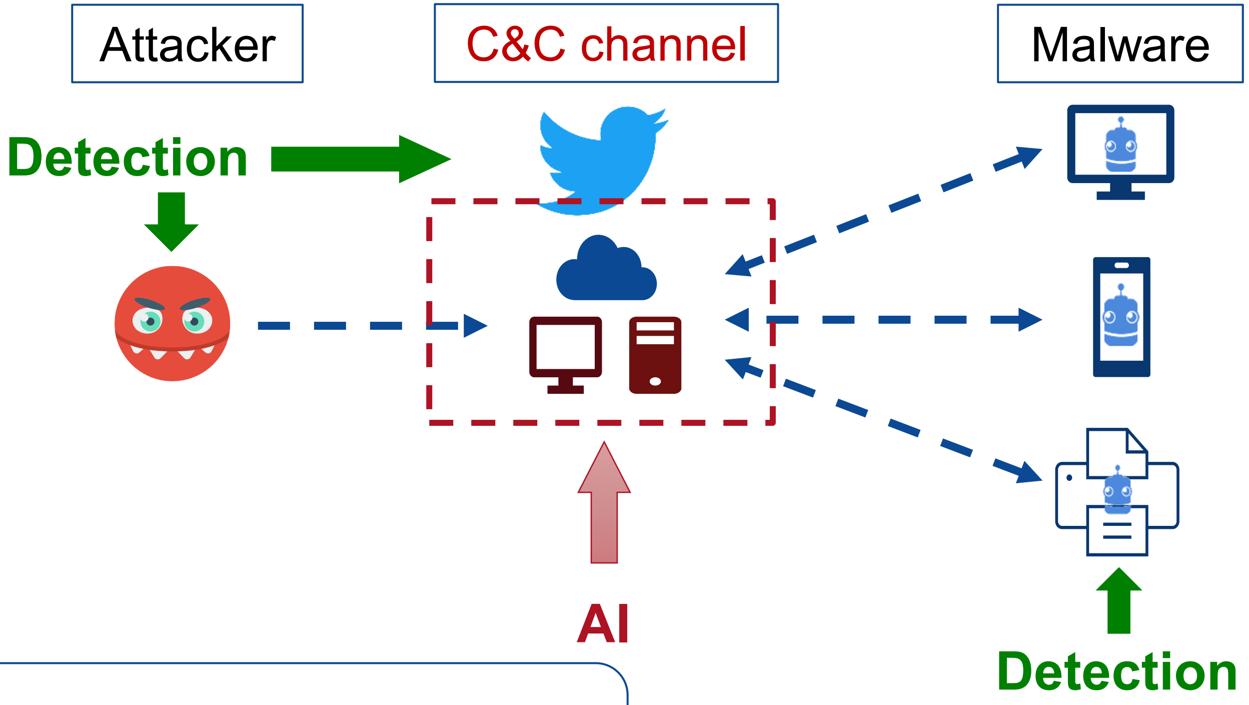
Risk control

- Verification of risky operation
- Limit content exposure for low-credit users
- Crack down on illegal account transaction

Anomaly behavior

+

Behavior similarity



Spatiotemporal similarity

+

Periodic behavior

Network, API call, CPU, etc.

Collaboration

- Security community shares the malware sample, model, and vectors to OSNs
- OSNs check the new avatars.

Conclusion

Method

AI-powered C&C channel

- Irreversible addressing by neural network
- Readable content by hash collision and data augmentation

Evaluation

Feasibility

- Siamese neural network
- Data augmentation
- Hash collision
- Avatar recognition
- Tweets crawling
- Security analysis

Mitigation

Possible countermeasures

- Malware side
- OSNs side

DEEPC2: AI-POWERED COVERT COMMAND AND CONTROL ON OSNS

Q&A

https://github.com/oicid/DeepC2