APTGen: An Approach towards Generating Practical Dataset Labelled with Targeted Attack Sequences

Yusuke Takahashi
NEC Corporation

Shigeyoshi Shima
NEC Corporation

Rui Tanabe
Institute of Advanced Sciences, Yokohama National University

Katsunari Yoshioka
Graduate School of Environment and Information Sciences, Yokohama National University

13th USENIX Workshop on Cyber Security Experimentation and Test 2020

Outline

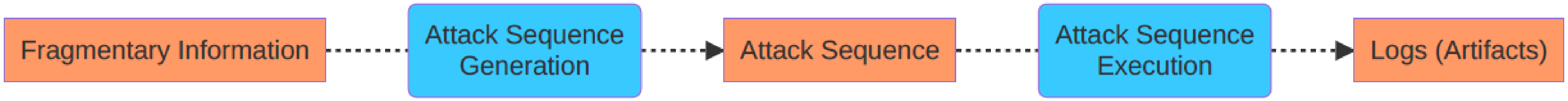
- Motivation
- Overview of APTGen
- Related Work
- Problems
  - Reproducibility
  - Diversity
- Design
- Evaluation
- Limitations
- Conclusion

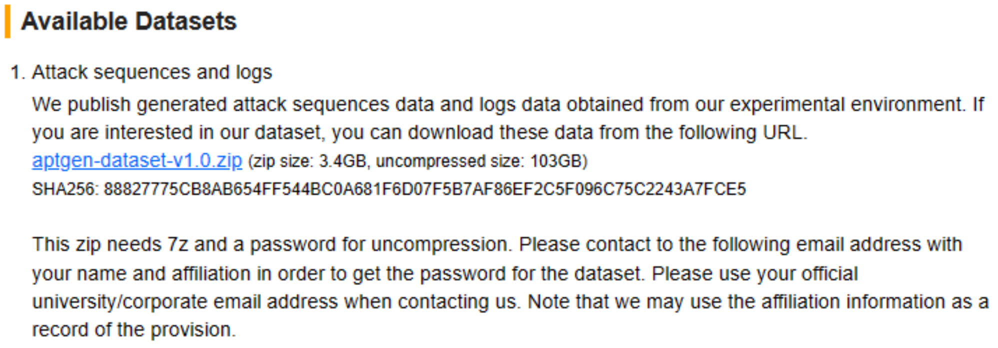
Motivation

- CSIRT
  - Not only investigates the attack methods
  - But also investigates the sequence of these attack methods and the attacker's purpose
- Developing technologies to automate investigating attack sequences
  - In order to evaluate these technologies
  - It is important to prepare a dataset
- Obtaining logs of actual attacks and use them as a dataset
  - Get the log from organizations victimized by attacks
  - Observe attacks purposely initiated in an observation environment

Overview of APTGen

- An approach for generating attack sequences and executing for building a dataset
  - Generates attack sequences from incident reports and security articles
  - Develops tools that execute attack sequences
  - Analyzes the relationship between the generated attack sequences by visualization
  - Releases a dataset with ground-truth





Available Datasets

1. Attack sequences and logs

   We publish generated attack sequences data and logs data obtained from our experimental environment. If you are interested in our dataset, you can download these data from the following URL.
   aptgen-dataset-v1.0.zip (zip size: 3.4GB, uncompressed size: 103GB)
   SHA256: 88827775CB8AB654FF544BC0A681F6D07F5B7AF86EF2C5F096C75C2243A7FCE5

   This zip needs 7z and a password for uncompression. Please contact to the following email address with your name and affiliation in order to get the password for the dataset. Please use your official university/corporate email address when contacting us. Note that we may use the affiliation information as a record of the provision.

Related Work

- Automatically generating attack trees
  - EG. Falco, A. Viswanathan, C. Caldera, and H. Shrobe. AMaster Attack Methodology for an AI-Based Automated Attack Planner for Smart Cities. IEEE Access, 6:4836048373, 2018.
- Automation techniques of red team
  - Andy Applebaum, Doug Miller, Blake Strom, Chris Korban, and Ross Wolf. Intelligent, Automated Red Team Emulation. In Proceedings of the 32Nd Annual Conference on Computer Security Applications, ACSAC ’16, pages 363–373. ACM, 2016.
  - Suneel Randhawa, Benjamin Turnbull, Joseph Yuen, and Jonathan Dean. Mission-Centric Automated Cyber Red Teaming. In Proceedings of the 13th International Conference on Availability, Reliability and Security, ARES 2018, pages 1–11. Association for Computing Machinery, August 2018.
- Planning techniques and frameworks
  - Joerg Hoffmann. Simulated Penetration Testing: From "Dijkstra" to "Turing Test++". In Twenty-Fifth International Conference on Automated Planning and Scheduling, 2015.
  - Joseph Yuen. Automated Cyber Red Teaming. Technical Report DSTO-TN-1420, DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION EDINBURGH (AUSTRALIA) CYBER AND ELECTRONIC WARFARE DIV, 2015.

Problems

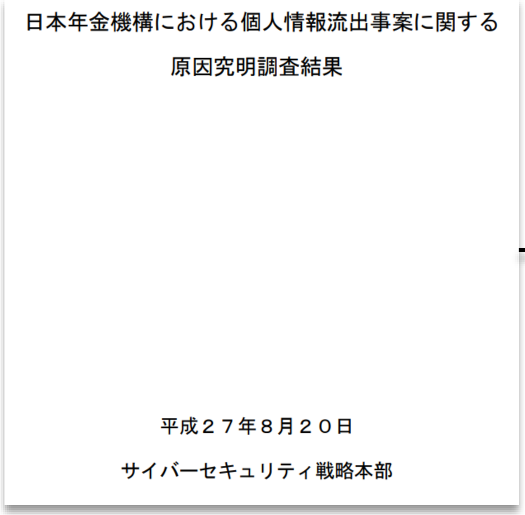
- Reproducibility
  - Require unified attack sequence information which is treated as the ground truth

- Diversity
  - Focus on generating a dataset (information as much as possible)
  - There are multiple sequences to achieve an attacker’s goal

- Reality(remained)
  - How realistic they are compared to the logs obtained from the real attacks

Design

- Reproducibility
  - Identify the reports or articles.
  - Mapping attack methods to Techniques in ATT&CK
  - Generate an attack sequence (Compensate for the lack of a Technique)
  - Executable codes in the experimental environment

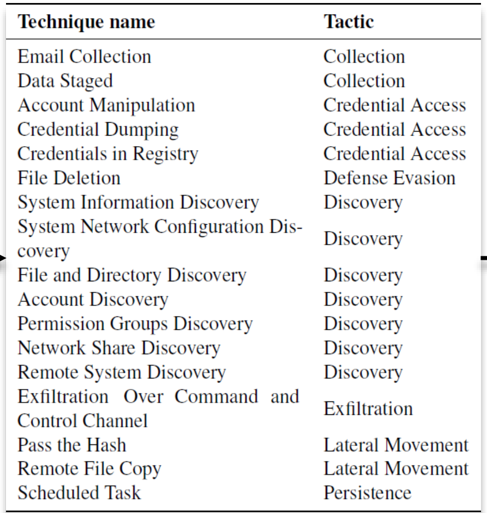


日本年金機構における個人情報流出事案に関する

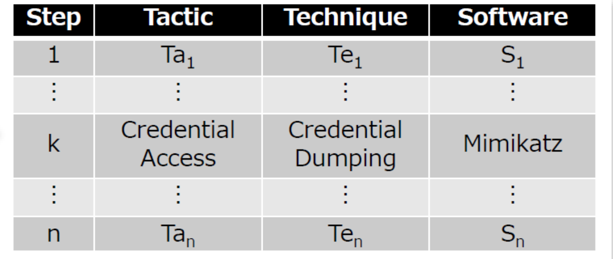
原因究明調査結果

平成２７年８月２０日

サイバーセキュリティ戦略本部



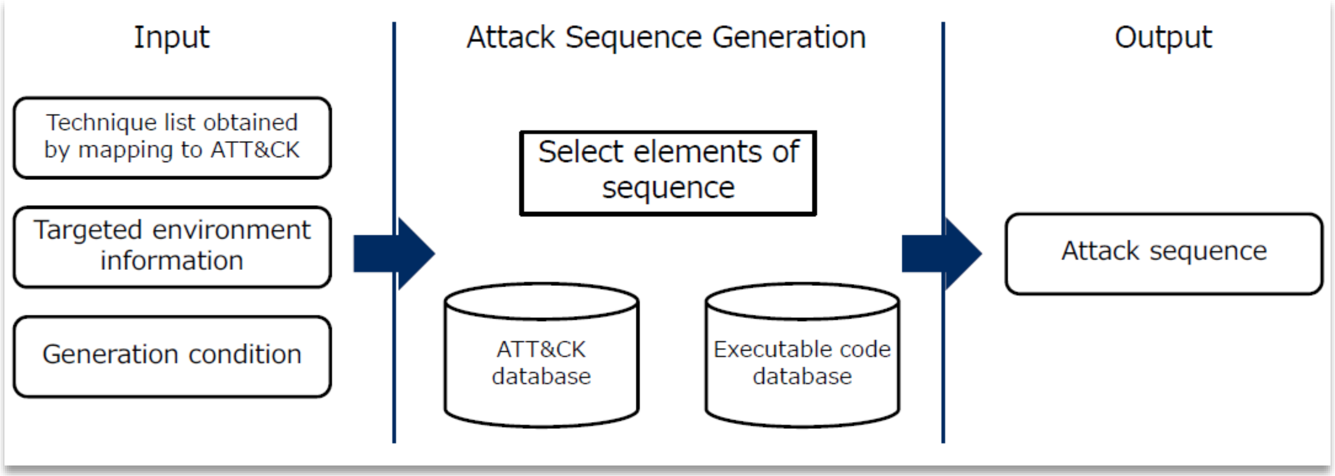
Technique name	Tactic
Email Collection	Collection
Data Staged	Collection
Account Manipulation	Credential Access
Credential Dumping	Credential Access
Credentials in Registry	Credential Access
File Deletion	Defense Evasion
System Information Discovery	Discovery
System Network Configuration Discovery	Discovery
File and Directory Discovery	Discovery
Account Discovery	Discovery
Permission Groups Discovery	Discovery
Network Share Discovery	Discovery
Remote System Discovery	Discovery
Exfiltration Over Command and Control Channel	Exfiltration
Pass the Hash	Lateral Movement
Remote File Copy	Lateral Movement
Scheduled Task	Persistence



Step	Tactic	Technique	Software
1	Ta$_1$	Te$_1$	S$_1$
⋮	⋮	⋮	⋮
k	Credential Access	Credential Dumping	Mimikatz
⋮	⋮	⋮	⋮
n	Ta$_n$	Te$_n$	S$_n$

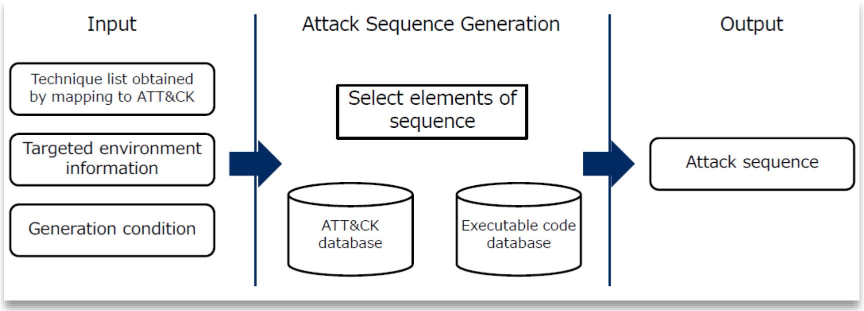
Design

- Diversity
  - It is difficult to generate many sequences manually
  - Developed attack sequence generation tool
  - Exclude Tactics： Initial Access， Execution， Command and Control， Impact Tactic


Input
Technique list obtained by mapping to ATT&CK
Targeted environment information
Generation condition
Attack Sequence Generation
Select elements of sequence
ATT&CK database
Executable code database
Output
Attack sequence

- Concerns
  - Tactic, Technique, and Software correspond to each other
  - Techniques are not chained into a logical sequence
  - Not be suitable for the experimental environment (OS)

Design


Input

Technique list obtained by mapping to ATT&CK

Targeted environment information

Generation condition

Attack Sequence Generation

Select elements of sequence

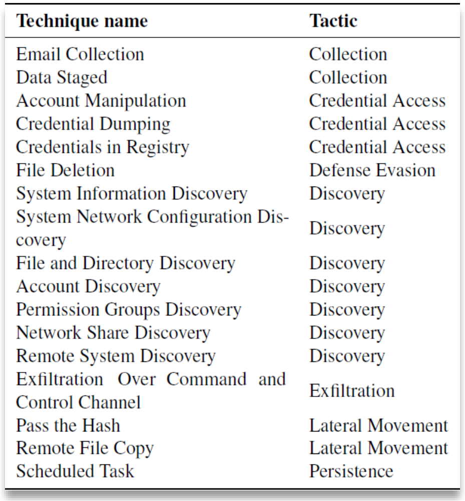
ATT&CK database

Executable code database

Output

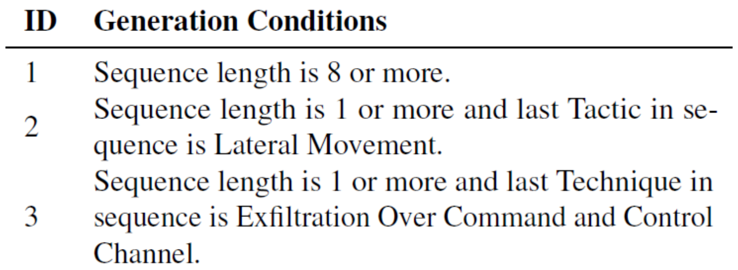
Attack sequence

Technique list

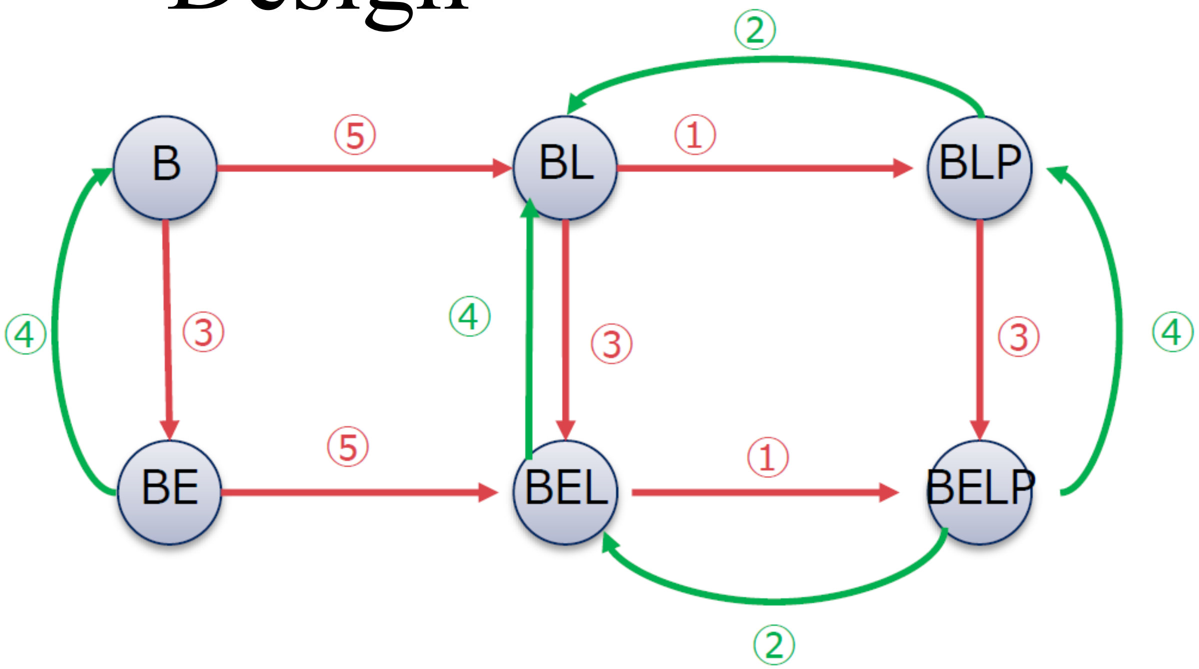


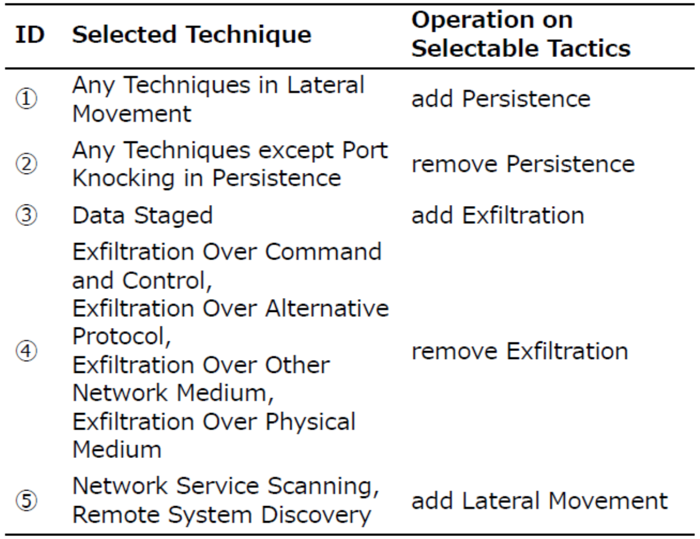
Technique name	Tactic
Email Collection	Collection
Data Staged	Collection
Account Manipulation	Credential Access
Credential Dumping	Credential Access
Credentials in Registry	Credential Access
File Deletion	Defense Evasion
System Information Discovery	Discovery
System Network Configuration Discovery	Discovery
File and Directory Discovery	Discovery
Account Discovery	Discovery
Permission Groups Discovery	Discovery
Network Share Discovery	Discovery
Remote System Discovery	Discovery
Exfiltration Over Command and Control Channel	Exfiltration
Pass the Hash	Lateral Movement
Remote File Copy	Lateral Movement
Scheduled Task	Persistence

Generation conditions



ID	Generation Conditions
1	Sequence length is 8 or more.
2	Sequence length is 1 or more and last Tactic in sequence is Lateral Movement.
3	Sequence length is 1 or more and last Technique in sequence is Exfiltration Over Command and Control Channel.

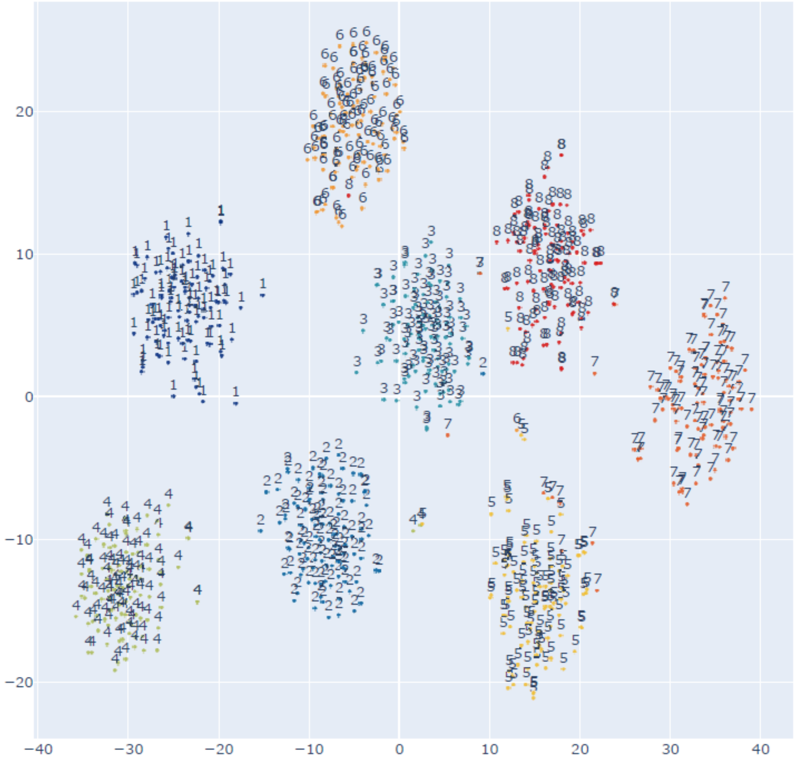




ID	Selected Technique	Operation on Selectable Tactics
①	Any Techniques in Lateral Movement	add Persistence
②	Any Techniques except Port Knocking in Persistence	remove Persistence
③	Data Staged	add Exfiltration
④	Exfiltration Over Command and Control, Exfiltration Over Alternative Protocol, Exfiltration Over Other Network Medium, Exfiltration Over Physical Medium	remove Exfiltration
⑤	Network Service Scanning, Remote System Discovery	add Lateral Movement

Evaluation

- Atomic Red Team
- Are these generated attack sequences executable?
- How realistic they are compared to the logs obtained from the real attacks
- Confirm diversity of generated attack sequences (whether achieve the attack goal)
  - ➢ TF-IDF
  - ➢ T-SNE

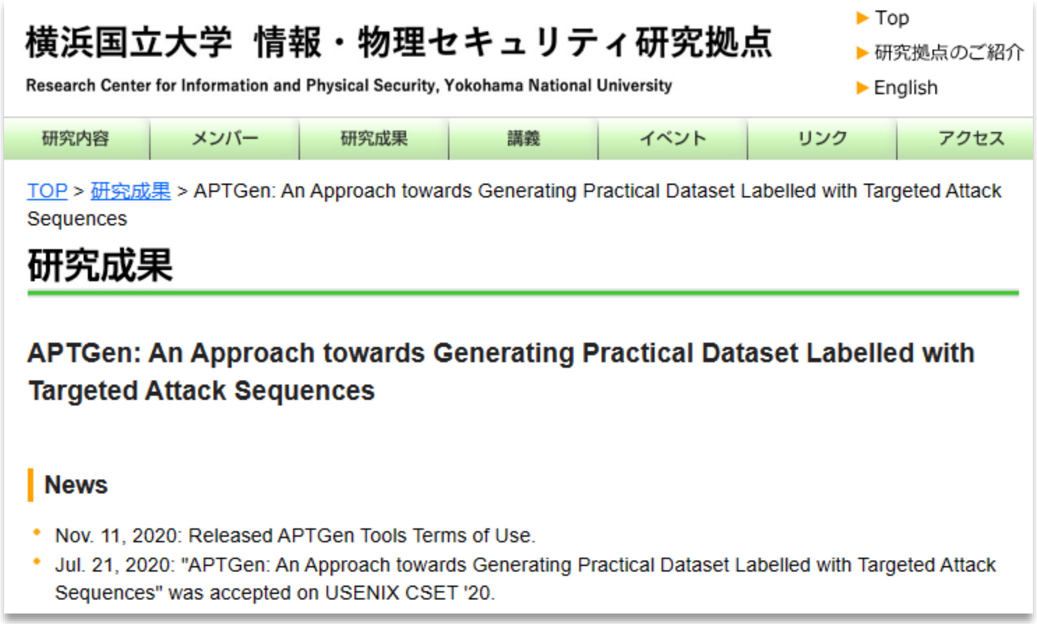


Limitations

- Do not contain logs related to benign background activity
- Do not use software attackers built such as malware and attack tools
- Do not include attackers’ characteristics such as attacking slowly over a long period
- Due to Ethics problems, do not publish the generate tool
- Binaries and executable codes used in the executable code database will need to be prepared

Conclusion

- APTGen, a method of generating attack sequences of targeted cyber attacks
- Building a dataset for research of investigating them
- Published 800 generated attack sequences and logs as a dataset



横浜国立大学 情報・物理セキュリティ研究拠点

Research Center for Information and Physical Security, Yokohama National University

TOP > 研究成果 > APTGen: An Approach towards Generating Practical Dataset Labelled with Targeted Attack Sequences

研究成果

APTGen: An Approach towards Generating Practical Dataset Labelled with Targeted Attack Sequences

News

- Nov. 11, 2020: Released APTGen Tools Terms of Use.
- Jul. 21, 2020: "APTGen: An Approach towards Generating Practical Dataset Labelled with Targeted Attack Sequences" was accepted on USENIX CSET '20.