ARTICLE IN PRESS

KICS The Korean Institute of Communications and Information Sciences

Available online at www.sciencedirect.com

ScienceDirect

ICT Express xxx (xxxx) xxx

www.elsevier.com/locate/icte

Camp2Vec: Embedding cyber campaign with ATT&CK framework for attack group analysis☆

Insup Lee, Changhee Choi*

Cyber Technology Center, Agency for Defense Development, Seoul 05771, Republic of Korea

Received 19 September 2022; received in revised form 21 March 2023; accepted 26 May 2023
Available online xxxx

Abstract

As the cyberattack subject has expanded from individual to group, attack patterns have become a complicated form of cyber campaigns. Although detecting the attack groups that operated the cyber campaigns is an important issue, complex methods such as deep learning are difficult to use due to the lack of campaign data. This paper proposes Camp2Vec, a lightweight statistics-based embedding for cyber campaigns, enabling attack group detection. The proposed method models a relationship between a campaign and techniques in the ATT&CK® framework as a document and words. Experimental results with expert-labeled datasets prove that Camp2Vec identifies representative attack groups successfully.

Keywords: Cyber campaign; Cyber threat intelligence; MITRE ATT&CK; TF-IDF; TTP sequence

1. Introduction

Even though the rapid development of communication and technology has brought lots of conveniences, it also leads to security threats from emerging attack surfaces [1]. These threats have caused enormous damage to the network and host levels, which motivates recent research to address them using machine learning and deep learning [2–7]. For instance, machine learning has been employed to improve the performance of intrusion detection systems in network [2,3] and host [4] environments. Deep learning has received lots of attention since it extracts important features automatically. To prepare a sufficient amount of training data for deep learning, there also have been studies on data augmentation for packets [5], binary files [6], and sequence data [7].

The scale of the cyberattacks has expanded from individual to group and nation levels, which leads to an advanced form of a *cyber campaign* or advanced persistent threats (APT). For instance, the Lazarus group has caused severe social disruptions by operating several campaigns such as the Sony Pictures hack, the WannyCry ransomware attack, and the abuse of the SWIFT banking systems [8]. Compared to traditional cyberattacks, it is more challenging to detect attack groups in sophisticated campaigns since the groups keep changing their patterns by adapting tactics, techniques, and procedures (TTPs) to prevent detection. Nevertheless, detecting and analyzing the groups are significant problems considering the critical damages from the campaigns.

To cope with the campaigns and APT attacks, MITRE has presented the ATT&CK® (Adversarial Tactics, Techniques, and Common Knowledge) framework [9], which models the attacks with a cyber kill chain and conceptualizes using TTPs. The ATT&CK framework promotes cyber threat intelligence (CTI) studies [10–16], and we focus on the attack attribution for groups that operated the cyber campaigns. Preparing a reliable dataset is challenging due to the low frequency of cyber campaigns. Thus, we focus on simplicity and lightweightness, rather than utilizing deep learning for attack group analysis. Motivated by the previous attempts to model attacks with embedding [17,18], we approach the group detection problem with campaign embeddings. Mohaisen et al. [17] proposed CHATTER, which initiated attack representation using dense vectors with cost-effective and flexible embeddings. CHATTER considers the high-level order to extract features from behavioral artifacts, followed by the n-gram technique. Shen et al. [18] presented Attack2Vec, which conducts embedding

☆ A preliminary version of this paper appeared in the Proc. of the Korea Institute of Military Science and Technology, 2022 (Lee et al., 2022).

* Corresponding author.
E-mail addresses: dlstjq0711@add.re.kr (I. Lee), changhee84@add.re.kr (C. Choi).

Peer review under responsibility of The Korean Institute of Communications and Information Sciences (KICS).

https://doi.org/10.1016/j.icte.2023.05.008

Please cite this article as: I. Lee and C. Choi, Camp2Vec: Embedding cyber campaign with ATT&CK framework for attack group analysis, ICT Express (2023), https://doi.org/10.1016/j.icte.2023.05.008.

for each Common Vulnerabilities and Exposures (CVEs) to examine the context of complex cyberattacks. Attack2Vec focuses on vectorizing the targeted vulnerabilities, without campaign modeling with TTPs and ATT&CK framework.

In this context, we introduce an effective group analysis method even with a limited amount of training data. We present Camp2Vec, which vectorizes cyber campaigns based on a statistical approach called term frequency-inverse document frequency (TF-IDF) with ATT&CK framework. Since Camp2Vec models high-level information (tactics and techniques) rather than specific vulnerabilities, Camp2Vec can represent group-level campaigns in terms of the cyber kill chain. To validate the superiority, we prepare a dataset labeled by security experts with domain knowledge. Through visualized results and quantitative analysis, we show that the proposed method shows the potential to distinguish attack groups successfully.

2. Background

This section describes the preliminaries for dataset preparation and the attack model.

2.1. MITRE ATT&CK

MITRE ATT&CK® framework [9] is a standardized matrix for various attack procedures by analyzing the malicious behaviors from the perspective of the cyber kill chain. The framework aims to model the behaviors with tactics, techniques, and procedures (TTP) to examine attack patterns. There are 14 tactics (e.g., Initial Access, Lateral Movement, and Impact) and 218 techniques (e.g., Process Injection, PowerShell, and Masquerading) in MITRE ATT&CK v10.1. We focus on the fixed version (v10.1) in this paper since the detailed information, e.g., the number of techniques, changes under the corresponding version. Our dataset is labeled using the TTP information such as 'TA0001.T1566.002', where 'TA0001' is a tactic (Initial Access), 'T1566' is a technique (Phishing), and '002' is a sub-technique (Spearphishing Link).

2.2. Data collection

We use 855 security reports from APT & Cybercriminals Campaign Collection [19] as source data. Each report corresponds to a specific cyber campaign with information about an attack group that has operated the campaign. Although there have been several methods to label TTP for the reports automatically such as rcATT [20] and TRAM [21], they have limitations of biased distribution due to the training data insufficiency.

To address the insufficiency issue, we employ security experts to tag TTP labels for the reports using their domain knowledge. When tagging the TTP labels, the experts consider the context of APT attack phases where techniques occurred. For instance, suppose a report with the phrase 'Keylogging to capture passwords otherwise obscured from viewing.' The Keylogging maps to 'T1056.001', and the 'T1056.001' is associated with multiple tactics such as TA0006 (Credential Access) and TA0009 (Collection). Due to the context of data collection, the security experts give a label of 'TA0009.T1056.001.' Likewise, a security report includes a sequence of TTP. We assume that the tagged TTP sequence represents a cyber campaign and focus on the sequence data in our experiments.

2.3. Attack model

This paper assumes APT groups that operate campaign-level attacks to achieve their objectives such as data leakage. We design two cyber campaign scenarios for well-known attack groups, Lazarus and menuPass, based on relevant security reports [22,23].

Scenario for Lazarus (SCN-L). In SCN-L, adversaries use document malware called ThreatNeedle to compromise industries. They begin by sending spear-phishing emails with COVID-19 information to gain initial access. Once in, they use a Remote Access Trojan (RAT) to generate persistence and move laterally to Windows servers on the same network. The RAT also activates victims' ssh protocols and collects data, ultimately resulting in data leakage.

Scenario for menuPass (SCN-m). Attackers in SCN-m gain access to victims using remote access service via PsExec (Sysinternals Suite). Next, they download penetration tools to discover compromised network and host information using well-known tools such as nbtscan or PowerShell script. They then extract confidential information using Mimikatz and send it to the attack server, followed by lateral movement to the next victim.

These scenarios include three key features: Lateral movement (TA0008), working on X86 operating systems, and ease of testing and detection. We use the scenarios for evaluation in Section 5, and the proposed Camp2Vec will be detailed in the next section.

3. Proposed method

As illustrated in Fig. 1, we propose an attack group analysis consisting of three steps: (i) TTP sequence data preparation, (ii) campaign embedding by Camp2Vec, and (iii) calculating group similarities with group score. This section describes the proposed embedding method called Camp2Vec, and how to measure the similarity between attack groups after the embedding.

3.1. Preprocessing

Camp2Vec assumes an input as a cyber campaign, which has a form of raw TTP sequence. The three steps to preprocess the sequence are summarized as follows.

- **Minimum TTP length.** Camp2Vec filters the input sequence by setting the minimum length, defined as the number of TTPs in the input. Fig. 2 illustrates the length distribution for 855 campaigns (i.e., TTP sequences), which shows that the number of campaigns per TTP

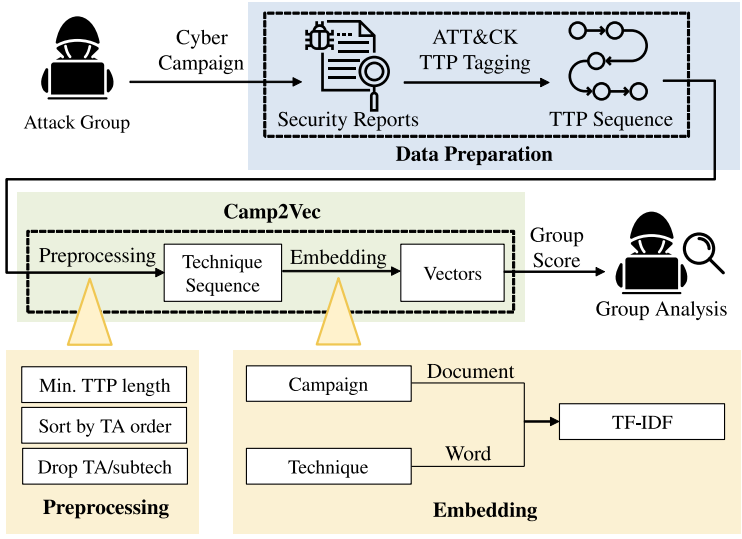

Fig. 1. Attack group analysis using Camp2Vec with ATT&CK framework. The overall process has three phases: campaign data preparation, campaign embedding with Camp2Vec, and measuring group scores for group estimation.

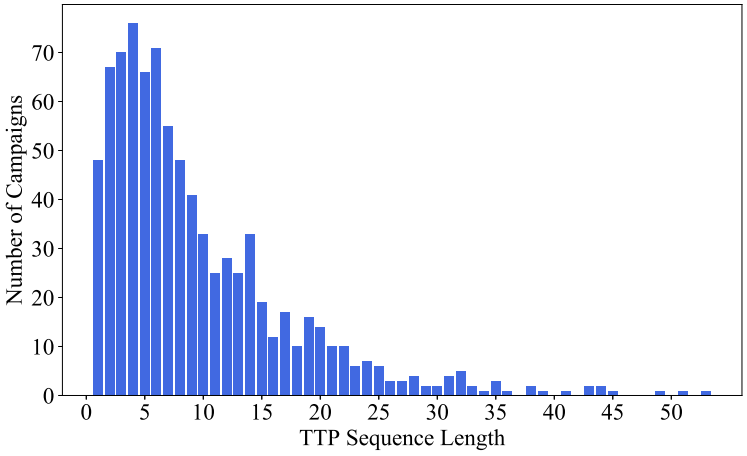

Fig. 2. TTP length distribution for campaigns.

length tends to increase until a length of 4 and then decreases. Setting the minimum length too high leads to insufficient data, while setting it too low may cause group characteristics to fade. Considering both aspects, we choose a minimum length of 5. It allows us to embed campaigns with group characteristics while avoiding data shortage.

- **Sort by tactic order.** The tactic order in the MITRE ATT&CK implies an attack flow modeled by a cyber kill chain. Since an input TTP sequence for Camp2Vec originates from contents in security reports, the listed order of TTPs does not reflect the kill chain process. To consider the process, we reorder the input sequence by the tactic order such as Initial Access (TA0001), Execution (TA0002), Persistence (TA0003), etc.
- **Drop tactics and sub-techniques.** The excessive number of TTP categories may interfere with fitting Camp2Vec due to the campaign data insufficiency. Although there are some cases where a specific technique is included by multiple tactics, we exclude tactics and sub-techniques to focus on the most important features provided by techniques.

After the preprocessing, an example of a technique sequence with a length of 6 is shown in Table 1.

3.2. Cyber campaign embedding

Camp2Vec embeds technique sequences into the corresponding dense vectors, utilizing the term frequency-inverse document frequency (TF-IDF) which is effective in calculating document similarity. The TF-IDF is a statistical method to express the importance of a specific word in sets of documents. The core idea for Camp2Vec is that we can view a relationship between a document and words as between a campaign and techniques. They are similar in that a document is a sequence of words whereas a campaign consists of techniques. We adopt the embedding without learning, TF-IDF, because the data insufficiency for campaigns is unsuitable for learning-based embedding.

Table 1
Technique sequence after preprocessing.

Type	TTP Sequence
Raw	TA0001.T1566.002 → TA0005.T1070.003 → TA0002.T1204.002 → TA0003.T1547.001 → TA0008.T1021.002 → TA0011.T1071.001
Preprecessed	T1566 → T1204 → T1547 → T1070 → T1021 → T1071

The TF-IDF is calculated by multiplying tf (the frequency a word occurs in a document) by idf (the measure of how much information the word presents; the rarity of a word across documents is proportional to the presented information). The tf and idf in Camp2Vec are defined as follows:

$$tf(t, c) = log\,[1 + freq(t, c)]\,, \tag{1}$$

$$idf(t) = log\,[(1 + n)/(1 + df(t))]\,, \tag{2}$$

where t is the technique, c is the campaign, $freq(t, c)$ is the frequency of technique in the campaign, n is the total number of campaigns and the $df(t)$ is the document frequency, i.e., the number of campaigns in the campaign set which include the technique. The TF-IDF has a higher value when a word exists in the specific document frequently and less in other documents. That is, Camp2Vec performs the embedding on the premise that the essential technique often occurs in a particular campaign and less in other campaigns.

3.3. Attack group analysis

Since an attack group operates multiple cyber campaigns, we approach the group similarity from the campaign similarities, measured with campaign vectors embedded by Camp2Vec. We adopt cosine similarity as the metric to measure campaign distance as follows:

$$similarity(x, y) = cos(\theta) = \frac{x \cdot y}{\|x\|\|y\|}, \tag{3}$$

where x and y are embedded vectors of cyber campaigns. Next, we assume the set of attack groups as follows:

$$G = \{G_1, G_2, \ldots, G_m\}\,, \tag{4}$$

where m is the number of groups in the dataset and each G_i has a different number of campaigns. We define *attack group analysis* as deriving the most likely attack group that carried out the input campaign (technique sequence). We term *group score* to examine the group, calculating similarities from all past campaigns and averaging them per attack group. The group score of G_i for the input campaign is defined as follows:

$$score_i = \frac{1}{N_i} \sum_{n=1}^{n=N_i} similarity(vec_{input}, vec_n), \tag{5}$$

where N_i is the number of campaigns in G_i, and vec is the embedded campaign vector. Among the all group scores about the input campaign, we can guess the attack group whose score is the highest has conducted the input campaign.

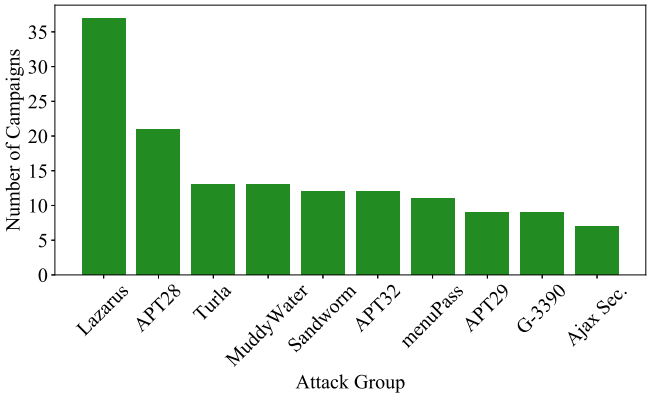

Fig. 3. Number of campaigns for each attack group.

4. Performance evaluation

We prepare a dataset consisting of 855 campaigns, labeled by security experts as described in Section 2. After preprocessing with a minimum length of 5, we use 594 campaigns from 329 attack groups as our base data. Note that there are not many attack groups that have carried out multiple campaigns; 256 attack groups have done only one campaign in our dataset. As shown in Fig. 3, we list the top 10 attack groups based on the number of campaigns. We focus on the top 10 attack groups, e.g., Lazarus, APT28, and Ajax Security, who has operated 37, 21, and 7 campaigns, respectively.

In addition to the 594 campaign data, we also prepare two scenarios for test data, designed by security experts via constructing TTP sequences as described in Section 2.3. The scenarios choose two representative attack groups, Lazarus and menuPass, since they have operated multiple campaigns and have caused tremendous impacts on society. A scenario for Lazarus (SCN-L) has a length of 24, whereas the scenario for menuPass (SCN-m) has 18. Specifically, we analyze attack groups with three phases: (i) fitting the Camp2Vec with the 594 campaigns from security reports, (ii) vectorizing the campaigns from top-10 attack groups and two scenario campaigns (SCN-L and SCN-m) using the fitted Camp2Vec model, and (iii) measuring group scores between the embedded scenarios and top 10 attack groups.

To start with an intuitive understanding, we visualize the Camp2Vec-embedded vectors on a two-dimensional plane. As shown in Fig. 4, we use t-distributed Stochastic Neighbor Embedding (t-SNE) with an iteration number of 1000 for dimension reduction from higher dimensions. The 594 dots are base data, the two crosses are the scenario data, and the colors are group information, i.e., Lazarus and menuPass are colored red and blue, respectively. Overall, the distribution of embedded vectors does not show a clear pattern since

Table 2
Group score between input scenario and attack group.

	Lazarus	APT28	Turla	MuddyWater	Sandworm	APT32	menuPass	APT29	G-3390	Ajax Sec.
SCN-L	**0.2403**	0.1744	0.1801	0.1526	0.1527	0.1692	0.1326	0.1092	0.1353	0.1842
SCN-m	0.0833	0.0406	0.0395	0.0385	0.0499	0.0166	**0.1908**	0.0398	0.1094	0.0637

Table 3
Attack group prediction using group score according to minimum TTP length in preprocessing.

Min. TTP length	SCN	Predicted groups			
		top 1	top 2	top 3	top 4
0	L	**Lazarus** (0.2063)	Ajax Sec. (0.1842)	Sandworm (0.1361)	APT28 (0.1334)
	m	**menuPass** (0.1645)	G-3390 (0.0807)	Lazarus (0.065)	Ajax Sec. (0.0543)
5	L	**Lazarus** (0.2403)	Ajax Sec. (0.1842)	Turla (0.1801)	APT28 (0.1744)
	m	**menuPass** (0.1908)	G-3390 (0.1094)	Lazarus (0.0833)	Ajax Sec. (0.0637)
10	L	**Lazarus** (0.3)	APT28 (0.2103)	Ajax Sec. (0.2072)	APT32 (0.2007)
	m	**menuPass** (0.1942)	MuddyWater (0.1731)	G-3390 (0.1068)	Sandworm (0.0855)
15	L	**Lazarus** (0.3662)	MuddyWater (0.2571)	Ajax Sec. (0.2057)	APT32 (0.2011)
	m	**menuPass** (0.3135)	G-3390 (0.1363)	Sandworm (0.1206)	Lazarus (0.1099)

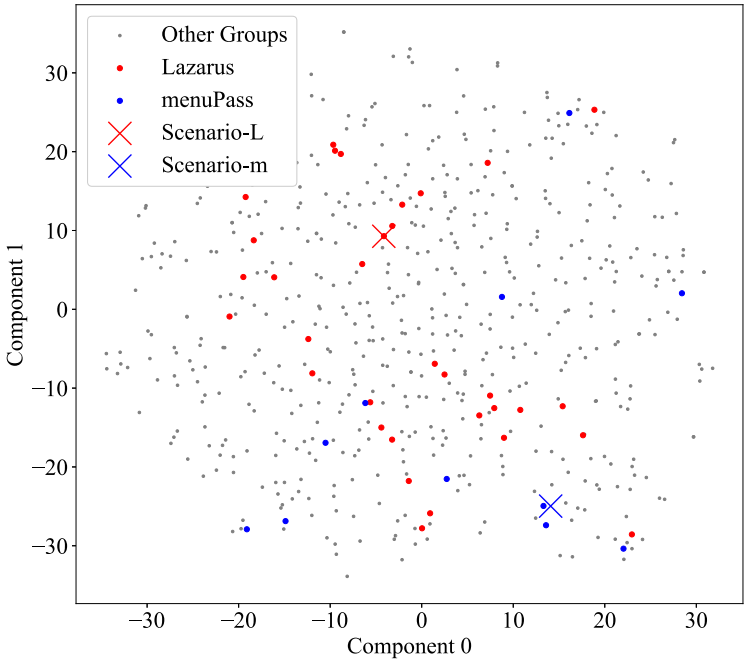

Fig. 4. Visualization using t-SNE for Camp2Vec-embedded campaigns. Lazarus (red) and menuPass (blue) tend to show a distinct distribution.

attack groups have become intelligent, leading to sophisticated campaigns. However, we can observe that SCN-L tends to be close to embedded Lazarus vectors while SCN-m to menuPass vectors, which needs an additional quantitative analysis.

Therefore, we conduct a quantitative analysis using a group score defined in Eq. (5). As shown in Table 2, we measure group scores between input scenarios (SCN-L and SCN-m) and the selected 10 attack groups. For example, a group score between SCN-L and Lazarus is an average of campaign similarities between SCN-L and 37 Lazarus campaigns. The results show that SCN-L scores from 0.1092 (APT29) to 0.2403 (Lazarus) and SCN-m scores from 0.0166 (APT32) to 0.1908 (menuPass). Considering the cosine similarity has values between 0 and 1, the overall results seem not to achieve high scores. However, we observe that the predicted group (i.e., the group which attains the highest group score) for SCN-L is Lazarus with a score of 0.2403 and for SCN-m is menuPass with a score of 0.1908, which are the correct answers.

Furthermore, we evaluate the impact of minimum TTP length in Camp2Vec's preprocessing on attack group analysis. When the lengths are 0, 5, 10, and 15, the numbers of used campaign data are 855, 594, 313, and 169, respectively. As shown in Table 3, we examine the top-4 predicted attack groups for each 8 (4×2; 4 for lengths and 2 for the scenarios) test case. Group scores increase gradually with the longer TTP

length since too short sequences that adversely affect group analysis are excluded. Note that the higher value for minimum TTP length indicates the limited amount of campaign data. The experimental results show that all the test cases can classify attack groups successfully regardless of minimum length, which means Camp2Vec can predict attack groups for input campaigns even with insufficient data.

5. Conclusion

We have presented Camp2Vec, which vectorizes cyber campaigns to address attack group detection. To design a lightweight and efficient campaign embedding, we applied the TF-IDF by considering the campaign as a document and the technique as a word. Experimental results with expert-labeled datasets have demonstrated that Camp2Vec successfully detects attack groups even with a limited amount of campaign data.

CRediT authorship contribution statement

Insup Lee: Conceptualization, Methodology, Investigation, Validation, Writing – original draft, Writing – review & editing. **Changhee Choi:** Methodology, Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Agency For Defense Development, Republic of Korea.

References

[1] M. Wazid, A.K. Das, V. Chamola, Y. Park, Uniting cyber security and machine learning: Advantages, challenges, and future research, ICT Express 8 (2022) 313–321.

[2] T. Thilagam, R. Aruna, Intrusion detection for network based cloud computing by custom RC-NN and optimization, ICT Express 7 (2021) 512–520.

[3] I. Lee, H. Roh, W. Lee, Encrypted malware traffic detection using incremental learning, in: Proc. IEEE International Conference on Computer Communications Workshops, 2020, pp. 1348–1349.

[4] S.R.T. Mat, M.F.A. Razak, M.N.M. Kahar, J.M. Arif, A. Firdaus, A bayesian probability model for android malware detection, ICT Express 8 (2022) 424–431.

[5] P. Wang, S. Li, F. Ye, Z. Wang, M. Zhang, PacketCGAN: Exploratory study of class imbalance for encrypted traffic classification using CGAN, in: Proc. IEEE International Conference on Communications, 2020, pp. 1–7.

[6] C. Choi, S. Shin, I. Lee, Opcode sequence amplifier using sequence generative adversarial networks, in: Proc. IEEE International Conference on ICT Convergence, 2019, pp. 968–970.

[7] S. Shin, I. Lee, C. Choi, Anomaly dataset augmentation using the sequence generative models, in: Proc. IEEE International Conference on Machine Learning and Applications, 2019, pp. 1143–1148.

[8] P. Kalnai, M. Poslusny, Lazarus group: a mahjong game played with different sets of tiles, in: Proc. Virus Bulletin International Conference, 2018.

[9] Mitre, MITRE ATT & CK®, 2021, Available: https://attack.mitre.org.

[10] I. Lee, C. Shin, S. Shin, S. Seo, C. Choi, Analyzing cyberattack campaign similarity via TTP sequence embedding, in: Proc. Korea Institute of Military Science and Technology, 2022, pp. 1431–1432.

[11] Y. Shin, K. Kim, J. Lee, K. Lee, Focusing on the weakest link: A similarity analysis on phishing campaigns based on the ATT & CK matrix, Secur. Commun. Netw. (2022).

[12] Y. Huang, C. Lin, Y. Guo, K. Lo, Y. Sun, M. Chen, Open source intelligence for malicious behavior discovery and interpretation, IEEE Trans. Dependable Secure Comput. 19 (2) (2022) 776–789.

[13] C. Xiong, T. Zhu, W. Dong, L. Ruan, R. Yang, Y. Cheng, Y. Chen, S. Cheng, X. Chen, CONAN: A practical real-time APT detection system with high accuracy and efficiency, IEEE Trans. Dependable Secure Comput. 19 (1) (2022) 551–565.

[14] K. Kim, Y. Shin, J. Lee, K. Lee, Automatically attributing mobile threat actors by vectorized ATT & CK matrix and paired indicator, Sensors 21 (19) (2021).

[15] Z. Jadidi, Y. Lu, A threat hunting framework for industrial control systems, IEEE Access 9 (2021) 164118-164130.

[16] S.Y. Enogh, Z. Huang, C. Moon, D. Lee, M. Ahn, D. Kim, HARMer: Cyber-attacks automation and evaluation, IEEE Access 8 (2020) 129397-129414.

[17] A. Mohaisen, A.G. West, A. Mankin, O. Alrawi, Chatter: Classifying malware families using system event ordering, in: Proc. IEEE Conference on Communications and Network Security, 2014, pp. 283–291.

[18] Y. Shen, G. Stringhini, Attack2Vec: Leveraging temporal word embeddings to understand the evolution of cyberattacks, in: Proc. USENIX Conference on Security Symposium, 2019, pp. 905–921.

[19] APT & CyberCriminal Campaign Collections, Available: https://github.com/CyberMonitor/APT_CyberCriminal_Campagin_Collections.

[20] V. Legoy, M. Caselli, C. Seifert, A. Peter, Automated retrieval of ATT & CK tactics and techniques for cyber threat reports, 2020, arXiv preprint arXiv:2004.14322.

[21] The center for threat-informed defense, 2021, TRAM is an open-source platform designed to advance research into automating the mapping of cyber threat intelligence reports to MITRE ATT & CK, Available: http://github.com/center-for-threat-informed-defense/tram/.

[22] Lazarus targets defense industry with ThreatNeedle, Available: https://ics-cert.kaspersky.com/publications/reports/2021/02/25/l-azarus-targets-defense-industry-with-threatneedle/.

[23] APT10: Tracking down the stealth activity of the A41APT campaign, Available: https://media.kasperskydaily.com/wp-content/uploads/sites/86/2021/02/25140359/greatidea_A41_v1.0.pdf.