Future Generation Computer Systems 96 (2019) 227–242

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

A machine learning-based FinTech cyber threat attribution framework using high-level indicators of compromise

Umara Noor [a,c], Zahid Anwar [a,b], Tehmina Amjad [c], Kim-Kwang Raymond Choo [d,*]

[a] Department of Computing, School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan
[b] Mathematics and Computer Science, Fontbonne University, 6800 Wydown Blvd, St. Louis, MO 63105, USA
[c] Department of Computer Science and Software Engineering, Faculty of Basic and Applied Science (DCS&SE, FBAS), International Islamic University Islamabad (IIUI), Pakistan
[d] Department of Information Systems and Cyber Security, University of Texas at San Antonio, San Antonio, TX 78249-0631, USA

HIGHLIGHTS

- A machine learning-based FinTech cyber threat attribution framework.
- Cyber threat attribution using high-level indicators of compromise.
- Cyber threat data collection.
- Automated cyber threat attribution framework using high-level compromise indicators.

ARTICLE INFO

Article history:
Received 25 October 2018
Received in revised form 10 January 2019
Accepted 6 February 2019
Available online 14 February 2019

Keywords:
Cyber threat attribution
FinTech threat attribution
Tactics techniques and procedures
Machine learning
Deep learning neural network
Cyber threat intelligence

ABSTRACT

Cyber threat attribution identifies the source of a malicious cyber activity, which in turn informs cyber security mitigation responses and strategies. Such responses and strategies are crucial for deterring future attacks, particularly in the financial and critical infrastructure sectors. However, existing approaches generally rely on manual analysis of attack indicators obtained through approaches such as trace-back, firewalls, intrusion detection and honeypot deployments. These attack indicators, also known as low-level Indicators of Compromise (IOCs), are rarely re-used and can be easily modified and disguised resulting in a deceptive and biased cyber threat attribution. Cyber attackers, particularly financially-motivated actors, can use common high-level attack patterns that evolve less frequently as compared to the low-level IOCs. To attribute cyber threats effectively, it is necessary to identify them based on the high-level adversary's attack patterns (e.g. tactics, techniques and procedures - TTPs, software tools and malware) employed in different phases of the cyber kill chain. Identification of high-level attack patterns is time-consuming, requiring forensic investigation of the victim network(s) and other resources. In the rare case that attack patterns are reported in cyber threat intelligence (CTI) reports, the format is textual and unstructured typically taking the form of lengthy incident reports prepared for human consumption (e.g. prepared for C-level and senior management executives), which cannot be directly interpreted by machines. Thus, in this paper we propose a framework to automate cyber threat attribution. Specifically, we profile cyber threat actors (CTAs) based on their attack patterns extracted from CTI reports, using the distributional semantics technique of Natural Language Processing. Using these profiles, we train and test five machine learning classifiers on 327 CTI reports collected from publicly available incident reports that cover events from May 2012 to February 2018. It is observed that the CTA profiles obtained attribute cyber threats with a high precision (i.e. 83% as compared to other publicly available CTA profiles, where the precision is 33%). The Deep Learning Neural Network (DLNN) based classifier also attributes cyber threats with a higher accuracy (i.e. 94% as compared to other classifiers).

1. Introduction

Cyber threat attribution facilitates the identification of an attacker or his/her intermediary. This can be used in subsequent (forensic) investigation by organizations or prosecution by law

* Corresponding author.
E-mail addresses: 13phdunoor@seecs.nust.edu.pk (U. Noor), zahid.anwar@seecs.nust.edu.pk (Z. Anwar), tehminaamjad@iiu.edu.pk (T. Amjad), raymond.choo@fulbrightmail.org (K.-K.R. Choo).

https://doi.org/10.1016/j.future.2019.02.013
0167-739X/

enforcement and other relevant stakeholders. For example, the U.S. Congress enacted the Cybersecurity Information Sharing Act (CISA) into law in 2015 [1], which mandates organizations (including financial institutions) involved in cyber data breach incidents to share cyber threat intelligence (CTI) with other relevant stakeholders, particularly their customers [2]. The interpretation and practices of cyber attack attribution have evolved with time. In the earlier days, the focus of attribution was to locate the attacker(s) or intermediary(ies) launching distributed denial of service (DDoS) attacks and to help stop malicious traffic via IP traceback. Thus, cyber attack attribution refers to source trace-back techniques that work backwards to geographically locate the origin of IP packets via router traversal record [3,4]. Source trace-back techniques cannot truly attribute cyber attacks, partly due to the intrinsic limitation of IP address spoofing and anonymization. A resourceful attacker can, for example, compound the challenges of attribution by using reflection hosts, small Time To Live (TTL) values, employing botnets as stepping stones, and launching attacks over wider time frames.

In more recent times, there have been reports of advanced persistent threats (APTs), particularly state-sponsored APT groups, which comprise multi-stage campaigns targeting organizations, governments, and military for financial gain, espionage, and intellectual property theft. APTs are typically associated with sophisticated methods of breaching into a network, for example using zero-day exploits. In the first half of 2018, more than 4.5 billion data records were reportedly compromised in 945 incidents [5]. One of these high profile incident is the massive data breach involving Equifax, a Financial Technology (FinTech) firm where 148 million consumers had their personal information and credit card credentials accessed [6]. According to cyber criminologists [7], the FinTech industry is highly vulnerable to data breach that use APTs designed by financially motivated Cyber Threat Actors (CTAs). FinTech refers to technological innovations designed or deployed to automate and improve services in the banking and financial sector. Generally, CTAs use common attack patterns to compromise their target. Thus, timely identification of data breach and CTAs can provide the FinTech industry with reliable evidence during prosecution trials. Hence, there is an increased focus on the sharing of CTI as a measure of proactive defense against data breach incidents. The contextual details of a threat incident and CTAs can be obtained from CTI feeds shared with an organization from their trusted security community and security product vendors [8].

Several CTI sharing standards have been proposed [9], such as the Structured Threat Information Expression (STIX) [10]. STIX records the details of an attack incident as multi-level Indicators of Compromise (IOC) represented by observable, indicator, TTPs and exploit target constructs. The attributes of the incident, adversary, and associated campaign are also represented. The mitigation strategies based on expert knowledge are provided in the form of course of action constructs. Along with structured CTI, there are other unstructured and textual CTI reports published by vendors. These unstructured CTI reports are publicly available in the form of white papers, security experts' blogs and in news bulletins. Despite the availability of large number of CTI documents and commonality in CTAs' attack patterns, it is ironic that the security analysts are not entirely capable of detecting and attributing cyber threats in a timely fashion. For example, the Equifax data breach was undetected for 79 days [11]. According to a study carried out by FireEye's M-Trends, the median time for an organization to figure out that they have been attacked is 146 days [12]. This could be due to a number of reasons. For example, CTI standards are not adequately enforced or followed, and attack patterns are either missing or reported in textual form that machines cannot interpret. A common practice is to employ low-level threat IOCs such as IPs, ports, domains and hashes to detect data breach incidents. The network administrator feeds these low-level IOCs into firewall rule sets to block malicious traffic originating from threat sources. Security information and event management (SIEM) systems are employed to analyze and correlate these low-level IOCs in order to identify the geographic location of the attacker and obtain security alerts in real time. Unfortunately, these indicators have a very short lifespan with respect to threat defense, as they are susceptible to change having fewer chances to be reused again. A CTA constantly changes IP addresses by using new servers and domain names to facilitate or perform their attacks. Also, the IP addresses can be easily spoofed or anonymized by the attacker, which leads to inaccurate and biased attribution. To more accurately and effectively attribute cyber threats, it is necessary to identify CTAs based on those attack patterns that reflect their behavior and actions taken in different phases of the cyber kill chain model to launch a cyber attack.

Attack patterns such as tactics, techniques and procedures (TTPs), software tools and malware are also termed as high-level IOCs. Such high-level IOCs are comprehensively reported as human understandable textual descriptions in the unstructured CTI reports, which cannot be directly interpreted by machines. Manual extraction of high-level IOCs is a tedious, time consuming and error prone process, which is operationally challenging due to the large volume, variety, velocity and veracity of CTI. In order to structure high-level IOCs and employ them to profile CTAs, there is a need to define a common vocabulary and taxonomize these concepts.

The significance of high-level IOCs in cyber threat attribution is demonstrated using the democratic national committee (DNC) email hack [13]. The intention is to show how the threat actor may be determined by correlating the high-level IOCs found with previous threat incidents of a similar nature. The DNC threat incident became elusive as the CTA reportedly used a persona as a decoy to distract digital forensic investigations [14]. In the past, similar strategies of using a persona as a decoy to distract digital forensic investigations have also been used in other related data breach incidents, such as those targeting the German parliament [15], the French Television network [16] and the World Anti-Doping Agency (WADA) [17]. By analyzing these threat incidents, it can be seen that there are certain TTPs commonly found across these campaigns include the use of X-Agent spyware, spear phishing of email accounts especially Google-based and use of bit.ly URL shortening techniques [18]. Other data breach incidents having these common TTPs include the Ukranian Artillery Android app hack [19] and NATO/US White House attack [20].

A STIX representation of these cyber security incidents built from well-known sources is shown in Fig. 1, in order to highlight the common patterns found in the high-level IOCs. At the top is some anonymous threat actor that is to be determined. The threat actor is further associated with the data breach incidents under consideration. The attack incidents are connected to the TTPs, the exploit target, the indicators and the observables employed in each incident. The TTPs here represent the high-level IOCs while the indicators and observables represent the low-level IOCs. It can be seen that there are certain high-level IOCs common across these incidents. Low-level IOCs may differ between attack instances. The threat actor in the data breach incidents discussed so far was identified to be Fancy Bear [21], based on findings from CrowdStrike [19], ThreatConnect [22] and FireEye's Mandiant [23]. Fancy Bear is allegedly responsible for several attack incidents targeting government, military, media and other major organizations around the world. Based on the common TTP patterns between the DNC email hack and the other data breach incidents, it was determined that the threat actor behind the DNC email hack is Fancy bear.

The above examples highlight the potential of using high-level IOCs in cyber threat attribution. Therefore, in this paper, we focus on designing an automated mechanism to extract high-level IOCs

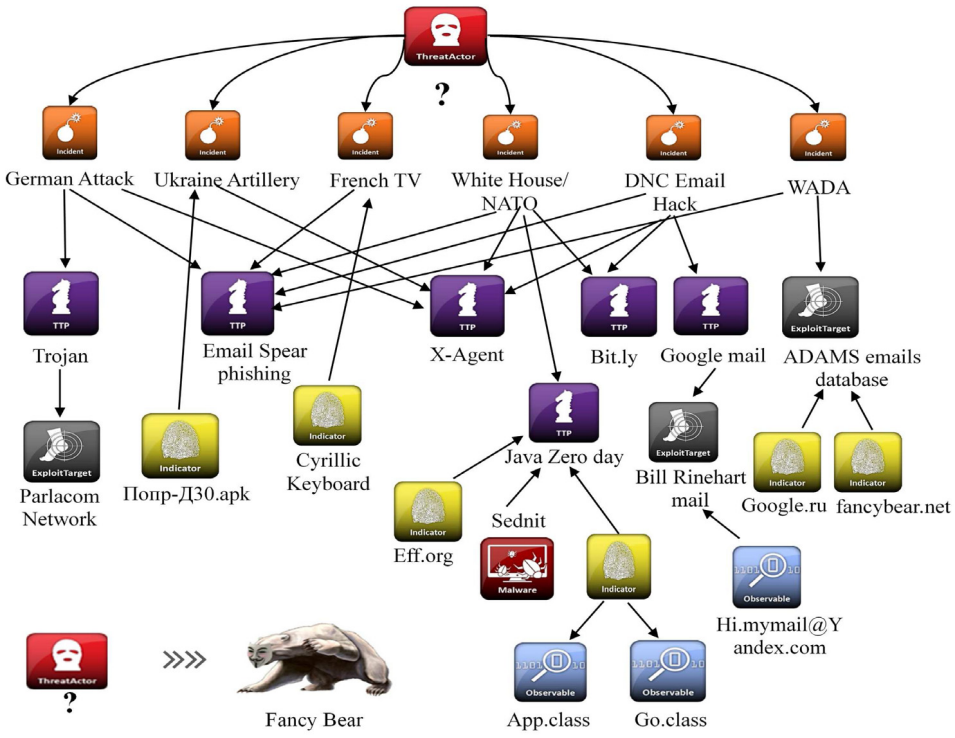


Fig. 1. DNC email hack threat actor attribution based on common patterns in high-level IOC.

from unstructured CTI documents using a common vocabulary to profile CTAs and employ these profiles to attribute perpetrators of cyber threat incidents — see Section 3. Specifically, to extract high-level attack patterns from unstructured CTI reports, we map the attack pattern query to its conceptual meaning, instead of searching for a keyword match. The labels for high-level IOC data set is prepared from the Adversarial Tactics, Techniques, and Common Knowledge (ATT&CK) [24] taxonomy provided by MITRE [24]. To semantically search for an attack pattern in unstructured CTI corpora, Latent Semantic Analysis (LSA) is used to index CTI reports. This allows analysts to use a wide variety of words to describe the high-level attack patterns, where the conceptual understanding may be concealed (since we are not able to address conceptual ambiguity between high-level attack patterns and descriptions used). The dimensionality of the sparse (with more zero entries) term-document matrix (representing the frequency of terms in the documents) is reduced using Singular Value Decomposition (SVD), which allows us to identify hidden word patterns in the documents. The relevance between the TTP query and the indexed documents is calculated by taking the cosine of the two vectors. The resulting dataset is used to attribute cyber threats to their CTAs by training five Machine Learning (ML) classifiers, i.e. Naïve Bayes, K Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF) and Deep Learning Neural Network (DLNN), using 327 CTI reports collected from publicly available CTI documents covering events from May 2012 to February 2018. As reported in Section 4, to quantify the effectiveness of the proposed framework, we use precision, recall, f-measure and FPR as the evaluation parameters. We also benchmark our performance (i.e., evaluation findings of the CTA profiles produced by our framework) with the publicly available CTA profile dataset provided by ATT&CK Mitre. The results demonstrate that the CTA profiles obtained using our framework attribute cyber threats with a higher precision, namely: 83% as compared to CTA profiles provided by ATT&CK Mitre with 33% precision. Of the five ML classifiers used, the DLNN based classifier attributes cyber threats with a higher accuracy, i.e. 94% as compared to other ML algorithms. The precision, recall, f-measure and False Positive Rate (FPR) of DLNN classifier are 90%, 89%, 0.89 and 3%, respectively. Finally, the last section concludes the proposed research.

In the next section, we will now introduce the extant literature.

2. Literature review

Cyber threat attribution based on high-level adversarial attack patterns found in CTI reports is a topic of ongoing interest, as evidenced by the number of cyber attack attribution frameworks proposed in the literature, such as those presented in [25–31]. A taxonomy to classify attribution techniques, for example, is presented by Nicholson et al. [32]. Based on their review, it was found that only two attribution techniques are currently in practice, namely: IP trace-back and honeypot data analysis. Source trace-back techniques are widely used for distributed denial of service (DDoS) attacks [3], and in many cases the attribution process is triggered manually. The factors that complicate cyber attack attribution include the ease to get access to tools that facilitate identity spoofing, the availability of reflector hosts and compromised machines (also known as zombie machines), and so on. For example, attack attribution based on attack traces obtained from cloud honeypot deployments can be clustered to profile attackers [33]. The clusters are formed according to the IP source countries. Such an approach does not identify any threat group or malware family. In addition, it is widely acknowledged in the literature that IP traceback on its own is ineffective to attribute a cyber attack [4]. We need other pieces of information, including contextual information (e.g. current political climate/landscape that may provide a hint on the possible attackers). For example, the Hunker et al. [34] suggested that one should take into consideration the set of possible actors, attributed objects, metrics to determine the confidence in attribution results, an acceptable policy for attribution, and so on. Attributions in the context of different attacks have also been discussed in the literature, including by Clark et al. [35].

Low-level IOCs acquired from malware forensic investigations have also been used in approaches such as those of Lock et al. [36]. Such low-level IOCs, such as hash values of malware binaries,

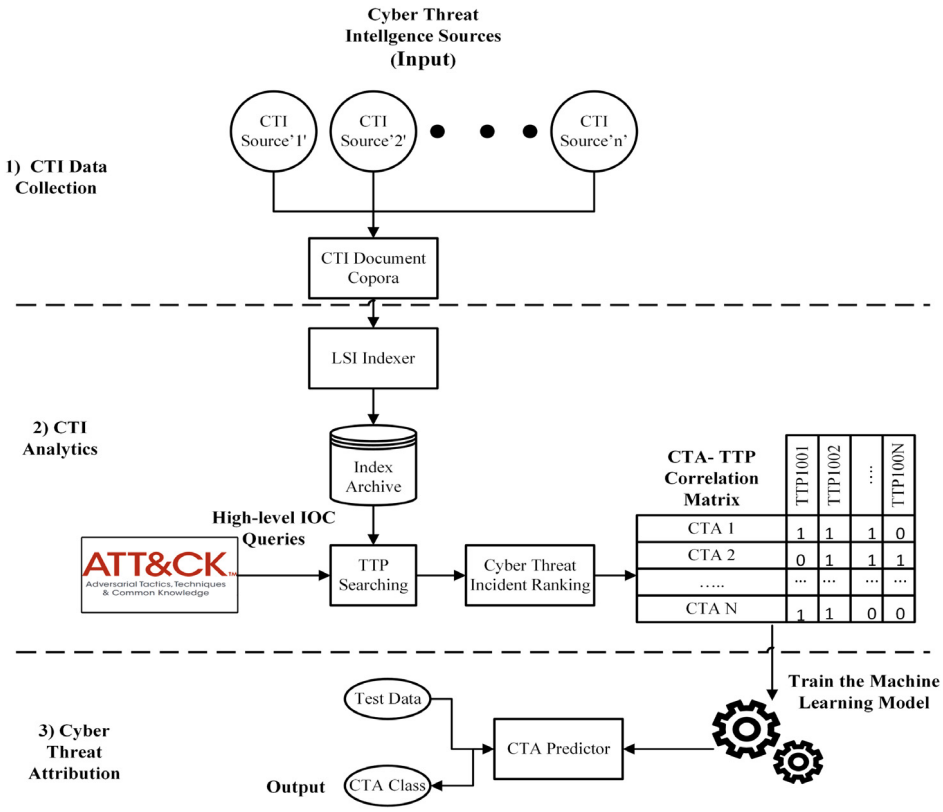


Fig. 2. Proposed cyber threat attribution framework.

are susceptible to changes (e.g. polymorphic and metamorphic malware). There have also been attempts to analyze CTA using cyber kill chain [37,38]. In more recent times, there have been attempts to introduce machine learning techniques in cyber attack attribution efforts, for example by analyzing malware used in cyber attacks [39–42].

The integration of high-level IOCs, for example by extraction them from unstructured CTI documents, with machine learning (e.g. using high-level IOCs to train machine learning classifiers) appear to be an understudied area. Hence, this is the focus of this paper - see the proposed framework in the next section.

3. Proposed framework

Our proposed framework comprises the following phases, namely: data collection, data analysis, and cyber threat attribution (see also Fig. 2). In the data collection phase, we identify reliable cyber threat sources relating to the particular CTAs and recent data breach incidents. As discussed earlier, unstructured CTI documents generally include high-level IOCs of CTAs that can be extracted. Hence, in the CTI analytics phase, a semantic search system is developed to facilitate searching for high-level IOCs in the CTI corpus. In the attribution phase, the CTA profiles developed in the analysis phase are used to train machine learning models to attribute cyber threats. The trained models predict the threat group class of unseen test data. We will now explain each phase in Sections 3.1 to 3.3.

3.1. Phase 1: Data collection

In phase 1, we need to collect data of related cyber data breach incidents from reliable threat sources. For illustrative purpose, we identified 36 cyber threat actors as summarized in Table 1. Then, publicly available documents describing the attack incidents and CTAs are collected. In many cases, one CTA can be referred to by different security companies or researchers differently; hence compounding the challenge of data acquisition. For example, CTA *Comment Crew* are also referred to as *Comment Panda* by CrowdStrike [43], *TG-8223* by Dell Secure Works [44], *APT 1* by Mandiant Fireeye [45], and *BrownFox* by isight [46].

To mitigate this challenge, we used cyber threat group alias references provided by [47] to search for and collect information about our CTAs of interest. We used a customized search engine [48] to collect the cyber threat reports. This search engine, first released in December 2015, is specifically customized to search APTs, CTAs, their operations, and malware used by these CTAs. We also used a publicly-available repository containing archived cyber threat reports dated since 2008 [49]. Using the search engine and the repository, we collected 327 unstructured CTI documents from 26 different sources that describe 36 threat actors. A list of these cyber threat sources is given in Table 2. The cyber threat reporting sources comprise security vendors, cyber security expert forums, IOC miners, security news websites, security research websites, security blogs and the U.S. Computer Emergency Readiness Team portal.

We identified 36 CTAs and a statistical summary of the number of documents for each CTA is given in Fig. 3. The highest number of CTI documents associated with a CTA is 23 (i.e. Fancy Bear) and the minimum is 3 (i.e. Gamaredon, Group5, and Fin10). The average is 9 CTI documents for each CTA, and the reference period is from May 2012 to February 2018. An excerpt of the timeline of the "Equation" threat actor is shown in Fig. 4, where 17 documents were collected. Specifically, for "Equation" threat actor the first incident was reported on February 16, 2015 by Kaspersky. Five documents reporting the incident were also found in the same month. Two documents were subsequently reported in March 2015, followed by another report in August 2016 by Kaspersky. Also, in August 2016, six documents were reported by different sources. In April

Table 1
Summary of CTAs, their country of origin and motive (s).

Cyber threat actor	Country	Motive
APT1	China	Espionage
admin338	China	Espionage, stealing trade secrets
APT 12	China	Espionage
APT 16	China	Espionage and spear phishing
APT 18	China	Espionage
APT 28	Russia	Espionage, data theft and reputation damage
APT 29	Russia	Espionage
APT 3	China	Stealing Intelligence Information
APT 30	China	Data theft for political gain
APT 32	Vietnam	Mass digital surveillance
APT 34	Iran	Espionage
Equation	USA	Espionage, data theft, system control
Fin 5	Russia	Financial gain, stealing personally identifiable information (PII) and payment card data
Fin 6	unknown	Stealing payment card data
Fin 7	Russia	Financial gain
Gamaredon	Russia	Espionage
GCMAN	Russia	Transferring money to e-currency services
Group 5	Iran	Penetrating systems and networks
Ke3chang	China	Espionage
Lazarus	North Korea	Espionage, financial loss and reputation damage
Lotus Blossom	China	Espionage
Magic Hound	Iran	Espionage
Menupass	China	Espionage and data theft
Moafee	China	Stealing trade secrets
Molerats	Middle East/Gaza	Espionage
Axiom	China	Espionage
Bronze Butler	China	Espionage
Carbanak	unknown	Financial gain
Cleaver	Iran	Data theft and service access
Copy Kittens	Iran	Espionage
Darkhotel	North Korea	Compromising personal gadgets of high-profile individuals
Deep Panda	China	Data theft
Dragonfly	Russia	Espionage
DragonOK	China	Espionage and spear phishing
DustStorm	China	Espionage and severe damage
Fin 10	unknown	Financial gain and victim extortion by stealing PII, file records and correspondences

Table 2
List of cyber threat sources used in this paper.

Cyber threat source	Description
Security vendor	Arbor [50], Zscalar [51], Netresec [52], Checkpoint [53], Novetta [54], Forcepoint [55], PWC [56], Proof Point [57], Clearskysec [58], CrowdStrike [43], Cylance [59], Trend micro [60], Symantec [61], ThreatConnect [62], Kaspersky [63], Paloaltonetworks [64] Fireeye [45]
Expert forum	Brighttalk [65]
IOC miner	Threatminer [66]
Security news website	welivesecurity [67], InfoSecurity [68], Threat Post [69]
Security research website	Citizen Lab [70]
Security blog	Krebsonsecurity [71], Security affairs [72]
USCERT portal [73]	United States Computer Emergency Readiness Team

2017, two documents were reported by different threat sources. Finally, in our document corpora, the last incident of the Equation threat actor was reported in November 2017 – see also Fig. 5.

A brief summary outlining the country of origin and motive(s) of these CTAs is given in Table 1, which can be categorized into *cyber espionage*, *data theft* and *service compromise*.

3.2. Phase 2: Data analysis

From the acquired CTI reports in phase 1, one should be able to extract details such as attack observables, indicators, Tactics, Techniques and Procedures (TTP), incidents, threat actors, campaigns, exploit targets and course of actions. Due to their unstructured form, these documents cannot be easily or directly interpreted by machines. Thus, in order to maximize the utility of the (valuable) information in these CTI reports, we will use our developed semantic search system to facilitate the extracting of high-level IOCs from these documents. This will allow the extracted high-level IOCs to be used by machines.

In the context of this paper, the CTI document corpora compiled in the previous phase is now our dataset used in this phase. As shown in Fig. 6, we can observe in the attack incident involving the Fancy Bear CTA published by CrowdStrike [74], the high-level IOCs in the text are encircled and labeled as TTPs, malware and software tools. This allows us to profile CTAs based on the IOCs. For this purpose, we propose an automated mechanism to extract high-level IOCs from unstructured CTI documents, structure them under predefined taxonomic labels and employ them to attribute CTAs.

In order to structure high-level IOCs, we have to either define a common vocabulary of metadata to taxonomize these concepts or employ an existing standard taxonomy, if it is available. One such taxonomy for the high-level adversary's IOCs is the ATT&CK from MITRE [24]. We selected this taxonomy for several reasons. It is built from a large number of cyber attack incident reports. In these reports, the security experts analyze the goal, motive, and capability of the attackers and attempt to establish their relationship with potential 'sponsors' (e.g. nation state(s)). It describes the adversary's attack patterns employed in different stages of the kill chain model. It is also constantly being updated by the research community. At the time of research, ATT&CK taxonomy comprises 188 TTPs, 146 malware and software tools used by CTAs in their attacks. The next step is to search these attack patterns in the CTI documents. Due to varied textual descriptions and choices for describing a concept, the CTI documents may not contain the exact keyword as defined in the standard taxonomy. Thus, instead of

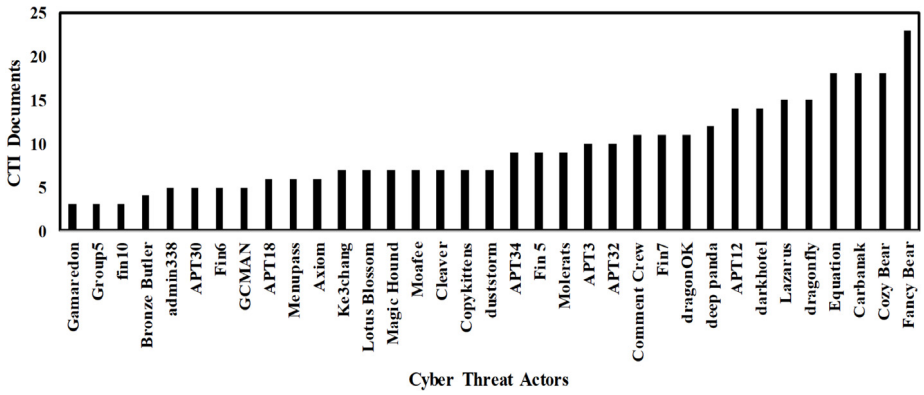


Fig. 3. Number of CTI documents associated with each CTA.

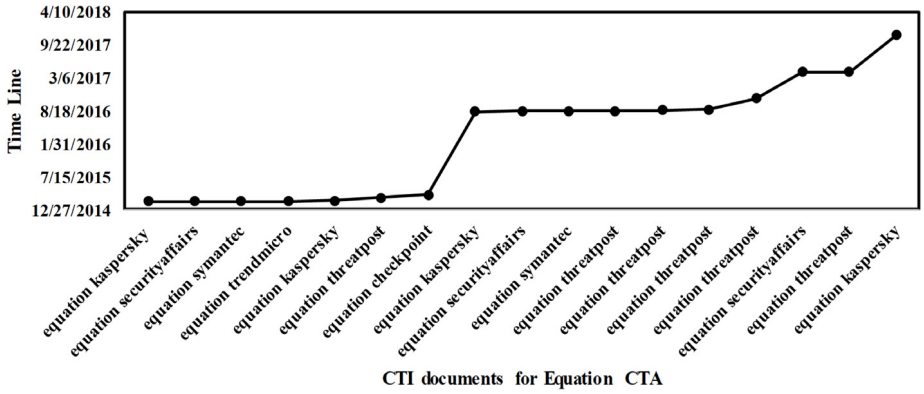


Fig. 4. Timeline of equation CTA.

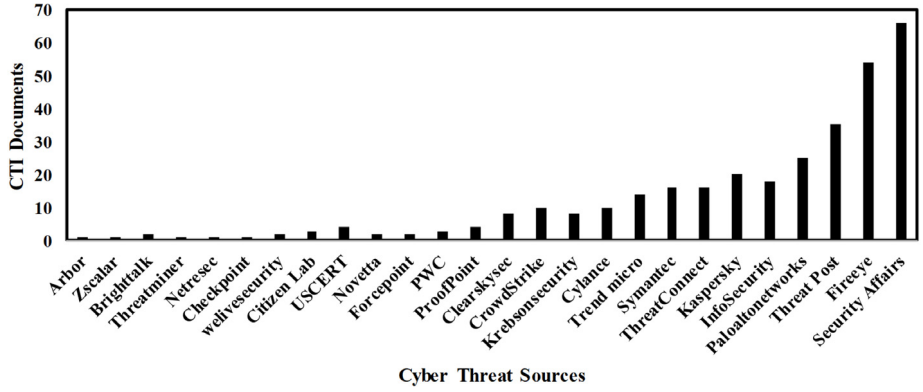


Fig. 5. Number of CTI Documents associated with each cyber threat source.

using simple keyword-based search, we develop a semantic search system based on the statistical distributional semantic relevance technique, i.e. LSA to retrieve semantically relevant documents. LSA is used to index CTI documents. The high-level IOC labels as defined in ATT&CK are searched for semantically relevant concepts or topics in the statistically derived conceptual indices. To retrieve the relevant documents, a statistical similarity measure (i.e. cosine similarity) is used. We will explain the use of semantic indexing and retrieval using an example.

In Fig. 7, the semantic mapping of TTPs present in *unstructured CTI* documents with TTP class labels in ATT&CK documents is shown. Here, we used ATT&CK taxonomy to obtain high-level adversary's attack patterns. There are five unstructured CTI documents depicting the incidents associated with Patchwork [75], Strider [76], PittyTiger [77], Oilrig [78] and Turla [79] CTAs. The labels for high-level attack patterns are obtained form the ATT&CK taxonomy are shown on the right-hand side of Fig. 7. There are four TTPs from the ATT&CK repository. Each TTP is a generalized concept with a distinguished ID attribute. The purpose of semantic mapping is to semantically connect the concepts, for example, *credential dumping* is an important TTP used by the CTAs. It is the process of obtaining account login and password information from the operating system or software. Suppose we want to find all those unstructured CTI documents where credential dumping is used, we can observe that the semantically retrieved document for the query *credential dumping* does not contain the exact words but are rather connected conceptually such as *login data database* in document 1, *use heartbleed to get valid accounts* in document 3 and *use Mimikatz* in document 4, where Mimikatz [80] is a credential dumping tool. It can be challenging for a human expert

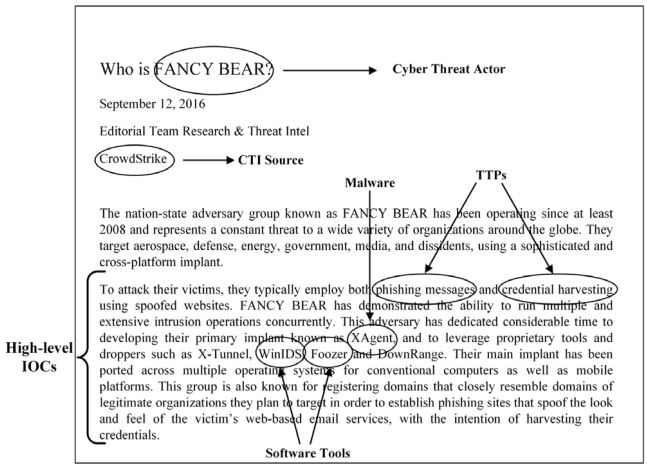


Fig. 6. An excerpt of a sample CTI document reported by CrowdStrike relating to Fancy Bear.

to identify these semantic connections between terms in the text belonging to the same class label (i.e. *credential dumping*). The conventional keyword-based lexical matching techniques retrieve information by matching query terms with the terms in the text corpora. However, they can be false matches due to inaccurate concept matching caused by synonyms and polysemous words. Thus, instead of using simple keyword search, we will use statistical distributional semantic relevance techniques to semantically connect TTPs in unstructured CTI documents to the taxonomic high-level IOC feature labels defined in ATT&CK taxonomy.

3.3. Phase 3: Cyber threat attribution

In the third phase, we use the CTA-TTP correlation dataset produced in the previous phase to train the machine learning models. These models are then used to predict the class of CTA for an unseen cyber threat incident. In this paper, we consider five widely used machine learning algorithms in this paper, namely: Naïve Bayes, KNN, Decision Tree, Random Forest and DLNN. However, this does not prevent the use of other machine learning algorithms in this phase.

- Naïve Bayes is a probabilistic classifier based on the principles of conditional probability in Bayes theorem [81]. Its implementation is very easy, and the training phase requires few data and can handle both discrete and continuous data. The model is trained in linear time, unlike iterative approaches that make approximation expensive. It is highly scalable towards the number of predictors and features. All features in the feature set are independent of each other. Due to its fast prediction quality, Naïve Bayes is generally more suitable for real-time predictions as compared to other machine learning algorithms.
- KNN is a very simple supervised machine learning algorithm, which classifies an instance based on the similarity with its neighboring instances. The input data set consists of k-closest training instances in the feature space [82]. An instance is classified based on the majority of votes of its neighbors. The instance is classified to that class which has the highest number of instances amongst the K nearest neighbors. The distance between instances is calculated using a distance function. For continuous variables, distance measures are Euclidean, Manhattan, and Minkowski. For categorical variables,

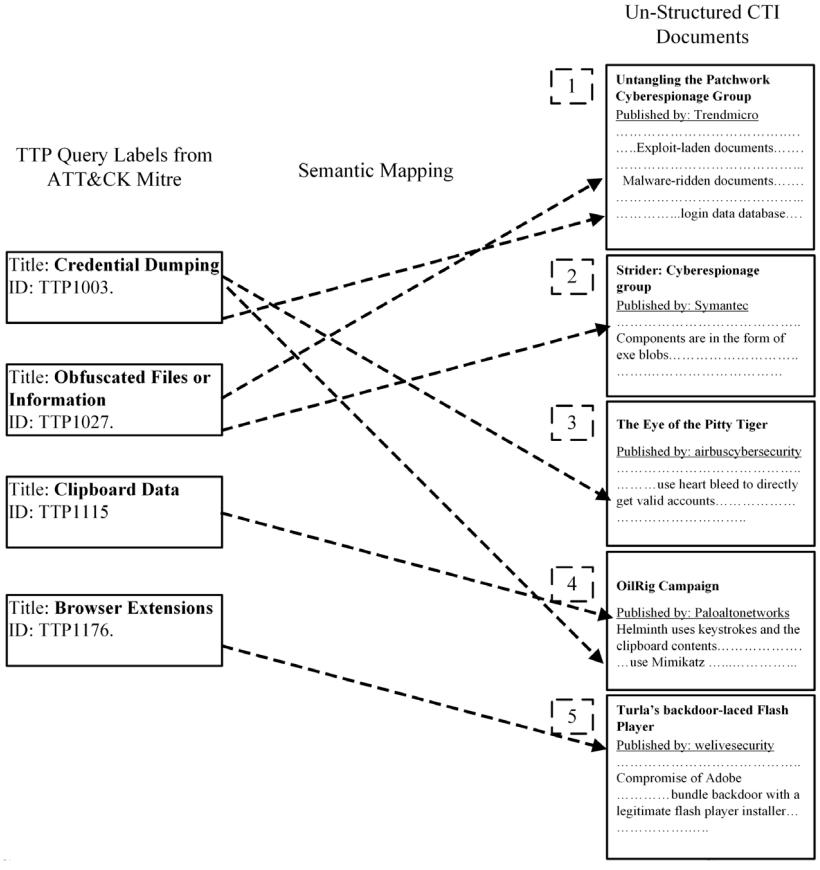


Fig. 7. Semantically mapping high-level adversary attack patterns in un-Structured CTI documents using LSA.

hamming distance is used. If K = 1, then the instance is simply assigned to the class with the smallest distance. The value of K must be odd to minimize the possibility of a tie. In order to choose the optimal value for K, the dataset must be first analyzed. For a more precise classification, a large value of K reduces the overall noise but it is not always guaranteed. Another way to determine the optimal value of K is to use cross-validation.

- The decision tree is a tree-shaped diagram used to determine a course of action [83]. Each branch of the tree represents a possible decision, occurrence or reaction. Decision trees with very deep growth tend to learn highly irregular patterns. With overfitting, one can obtain very low bias but very high variance. The problem of overfitting is mitigated using the Random Forest (RF) algorithm. RF is an ensemble learning method for classification, which constructs multiple decision trees during the training phase [84]. The decision of the majority of the trees is chosen by the RF as the final decision. For classification, RF takes the mode of the classes to produce the output. To remove overfitting, RF provides a way of averaging multiple deep trees trained with different parts of the same training set. The objective is to reduce the variance. As a result, there is a little increase in the bias and loss of interpretability, but generally it boosts the performance in the final model. RF selects a random subset of features to build different trees. One of the major advantages of RF is its high accuracy. It runs efficiently on large databases and produces highly accurate predictions. It also maintains accuracy when a large proportion of data is missing. RF has the capability of multi-object detection in complicated overlapping environments, for example when detecting different kinds of vehicles in a traffic environment. Such a feature can be useful in identifying traits of cyber criminals.
- Deep learning uses complex algorithms and deep neural networks to train a model [85]. DLNN, for example, has networks capable of learning from data that is unstructured or unlabeled, and it works similar to the functioning of a human brain. The base of a DLNN is a neural network, which is a mesh of interconnected neurons that receive some inputs, and process those inputs in layers to produce the output. Deep learning allows one to work with both structured and unstructured data, and handle complex operations. It is also scalable.

4. Evaluation and findings

In this section, we present the evaluation of our proposed framework. Specifically, the semantic search system is evaluated for its effectiveness in extracting high-level IOCs from unstructured CTI documents. The evaluation parameters considered are precision, recall, and f-measure. We compared the performance of our proposed LSA based search system with that of Apache Solr [86]. The effectiveness of the attribution results obtained by training the five machine learning models with our high-level IOC profiled CTA dataset was also evaluated. The evaluation parameters considered are precision, recall f-measure and False Positive Rate (FPR). We compared the attribution results of our dataset with those of the benchmark dataset of ATT&CK MITRE. We also observed the impact of feature selection on the attribution results. The features were selected using Information Gain (IG), which uses the entropy formula to select the best features.

4.1. Effectiveness of semantic search system

Apache Solr is a free and open-source Information Retrieval (IR) library, supported and released with the Apache license. It is based on a fuzzy search with edit distance, and commonly used in the implementation of Internet search engines and local or single-site searching. Apache Solr is a sub-project of Lucene [87]. The Lucene library is connected with the Apache Solr search server, which provides an open-source enterprise search platform. The features provided by Apache Solr are indexing, querying, mapping and ranking the outcome.

The evaluation results of Apache Solr and our LSA based semantic search system are shown in Table 3, based on the 11 tactic categories of ATT&CK Mitre. These tactic categories define the post-compromise stages of the cyber kill chain model. The high-level IOCs labels form queries for the search system. To evaluate the effectiveness of the semantic search and compare it with Apache Solr, we randomly selected 50 queries from the 188 high-level IOC queries. The precision, recall, and f-measure for both search systems were then calculated. The results demonstrated that the average precision of Apache Lucene is 90% and that of LSA based semantic search is 96%. The average recall of the Apache Solr is 22% and LSA based search system is 97%. The reason for Apache Solr's low recall is that it is strictly a keyword-based search system. In other words, only documents retrieved are those that contain the query keywords. On the other hand, our LSA based search system retrieves documents based on the semantic relevance of the query concept; thus, allowing for the retrieval of all documents that are semantically relevant to the query. The average f-measure of Apache Solr is 0.35 and LSA-based search system is 0.96. Based on the results, we conclude that the LSA based search system outperforms Apache Solr. The findings of all queries are then combined to form our dataset.

The statistics of TTPs contained in the ATT&CK taxonomy and the one identified by our proposed semantic search system for each CTA are presented in Fig. 8. The maximum number of TTPs reported in ATT&CK taxonomy is 50 for APT 16 while the minimum number of TTPs is 1 for GCMAN. On the other hand, the maximum number of TTPs identified by our semantic search system is 127 for dragonfly while the minimum number of TTPs is 5 for Moafee. The average number of TTPs for each CTA in ATT&CK taxonomy is 14, while in our dataset it is 63. The statistics clearly depict that our automated semantic search system identifies more TTPs with higher precision, in comparison to the manually compiled ATT&CK taxonomy.

4.2. Effectiveness of cyber threat attribution system

We used cross-validation technique to evaluate the five machine learning models. This technique divides the original dataset into two parts. One is a training set used to train the machine learning model and the other is the test set to evaluate the trained model. In k-fold cross-validation, the dataset is randomly and equally divided into k subsets. Of these k subsets, one subset is considered as the test data and the remaining k-1 subsets become the training data. This process is performed k times (also known as folds). Each subset has a chance to become the test data once. The evaluation results are averaged to get an estimated value. A key advantage of this method is that every single instance has an equal opportunity to be used for training and evaluation. In our evaluation, the value of k is 10.

Each CTA is represented by its features, i.e. high-level IOCs encompassing the attack techniques and software tools used by them. The machine learning models are then applied to this dataset. However, there is a need to select the most useful features that contribute to a higher precision. Therefore, there is a need to rank

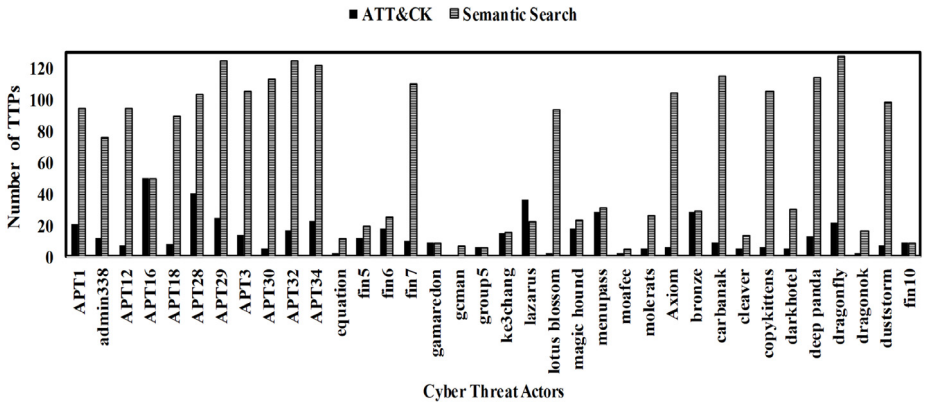


Fig. 8. TTP statistics in ATT&CK taxonomy and identified by our proposed semantic search system.

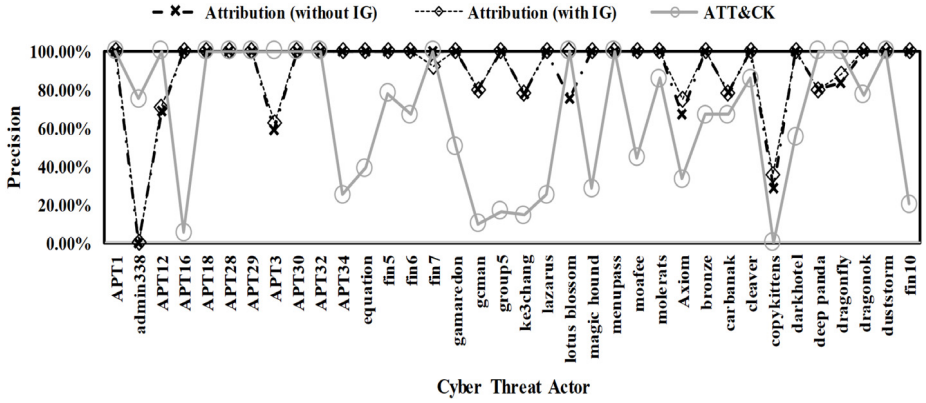


Fig. 9. Attribution precision using Naïve Bayes.

Table 3
Comparative analysis of the effectiveness of LSA based semantic search with Apache Lucene.

Tactic (Kill chain phases)	Precision		Recall		F-measure	
	Apache Lucene	Semantic search	Apache Lucene	Semantic search	Apache Lucene	Semantic search
Initial access	95%	94%	27%	99%	0.42	0.96
Execution	90%	93%	27%	98%	0.42	0.95
Persistence	80%	98%	18%	96%	0.29	0.97
Privilege escalation	75%	90%	9%	85%	0.16	0.87
Defense evasion	100%	100%	34%	100%	0.51	1
Credential access	90%	95%	23%	97%	0.37	0.96
Discovery	100%	100%	20%	96%	0.33	0.98
Lateral movement	90%	98%	21%	97%	0.34	0.97
Collection	90%	98%	23%	100%	0.37	0.99
Exfiltration	100%	89%	21%	100%	0.35	0.94
Command and control	80%	100%	20%	100%	0.32	1

feature set in order to determine which features are most useful for discriminating between the CTA classes. For this purpose, we use Information Gain (IG) – see Eq. (1) – to measure the importance of a given feature in a feature set and rank them based on the high common information.

$$IG(F, i) = E(F) - E(F \mid i) \tag{1}$$

In Eq. (1), F represents the feature set in the form of $(x_i, y) = (x_1, x_2, \ldots, x_n, y)$, where $x_i \in val(i)$. It is the value of the ith feature x, where y is the corresponding CTA class label. The Entropy (E) is defined using Eq. (2).

$$IG(F, i) = E(F) - \sum_{v \in val(i)} \frac{\mid x \in F \mid x_i = v \mid}{\mid F \mid} \cdot E(\mid x \in F \mid x_i = v \mid) \tag{2}$$

Table 4
List of TTPs discarded due to zero IG.

TTP ID	TTP title	Tactic (Kill chain phases)
1062	Hypervisor	Persistance
1118	InstallUtil	Defense evasion and execution
1121	Regsvcs/Regasm	Defense evasion and execution
1148	Hist control	Defense evasion and execution
1155	AppleScript	Execution and lateral movement
1166	Setuid & Setgid	Privilege escalation
1169	Sudo	Privilege escalation

IG ranks the TTPs that are most useful in discriminating among the classes to be learned. We discarded features with an IG of 0.

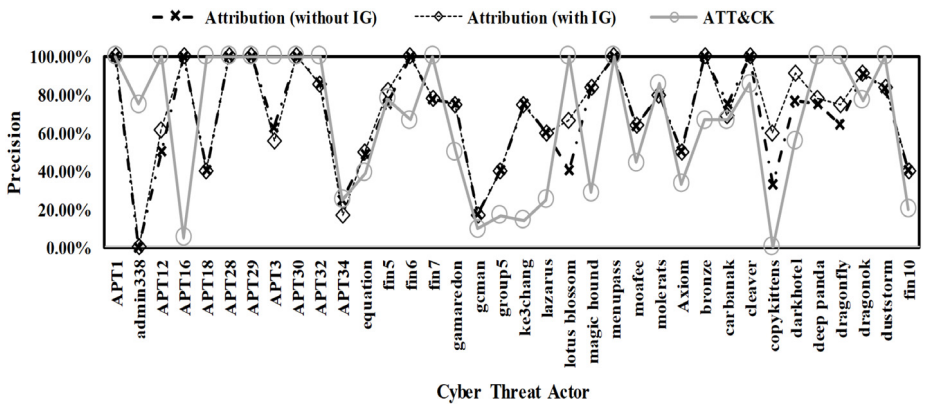


Fig. 10. Attribution precision using kNN.

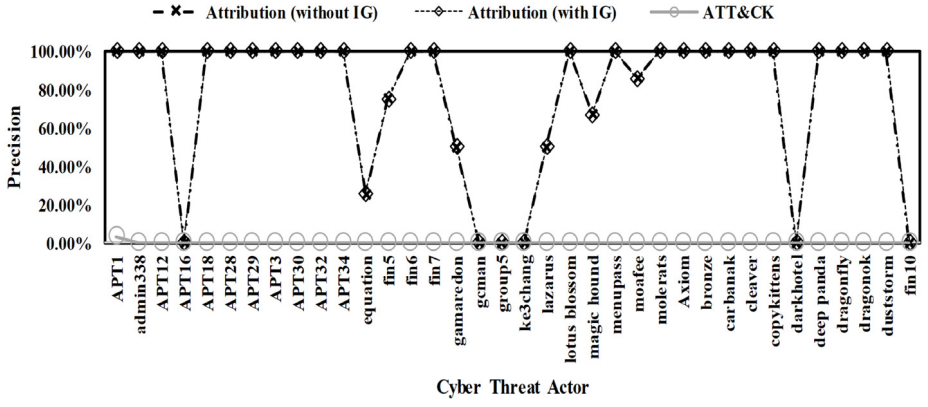


Fig. 11. Attribution precision using decision tree.

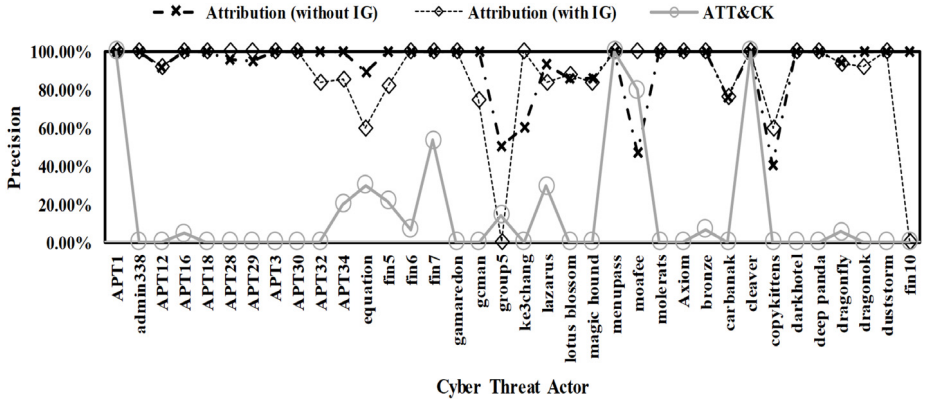


Fig. 12. Attribution precision using random forest.

We compared the results of attribution obtained with and without feature selection. The list of TTPs discarded are given in Table 4. These TTPs are not mentioned in the CTA document corpora. Also in the ATT&CK dataset, these TTPs are not used by any CTA. Based on this, we can conclude that the TTPs presented in Table 4 are currently not used by any CTA defined by ATT&CK MITRE. Similarly, the software tools and malware with zero IG are *3Para Rat, 4H Rat, Agent.btz, AutoIt backdoor, Badnews, BBSRAT, Blackcoffee, Bootrash, BS2005, CallMe, Cherry Picker, China Chopper, ComRat, Crimson, Downdelph, Duqu, Dyre, Epic, FakeM, Flame, H1N1, HDoor, HTRAN, Hacking Team UEFI Rootkit, Hizor, Janicab, Kasidet, Lslsass, Lurid, Miner-C, Mobile Order, Moonwind, Net Traveler, Nidiran, OwaAuth, P2P Zeus, Ping, Prikormka, Psylo, RaRstone, RockBoot, RTM, Regin, Remsec, Rover, Shamoon, Skeleton key, SSIMM, Sykipot, Sys10, T9000, Tinytyphoon, Taidoor, Trojan. Mebromi, UACME, Unknown logger, Uroburos, WiMM, Winnti, Wiper, Zeroaccess, httpclient, pngdowner*. We observed that these malware are either not mentioned in the document corpora or they are used by CTAs not considered in this paper. For instance, 3PARA RAT is a remote access Trojan used by CTA *Putter Panda*, which is not considered in our example.

The evaluation results are summarized in Table 5, which are the combined estimation taken by the average of individual values of accuracy, precision, recall, f-measure and FPR for each CTA. We observed that DLNN has the highest accuracy without feature selection (i.e. 94%), and RF has the highest precision without feature

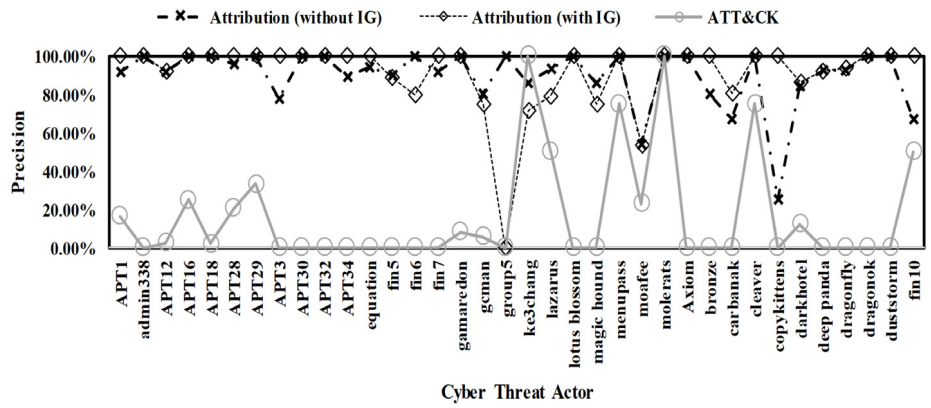


Fig. 13. Attribution precision using DLNN.

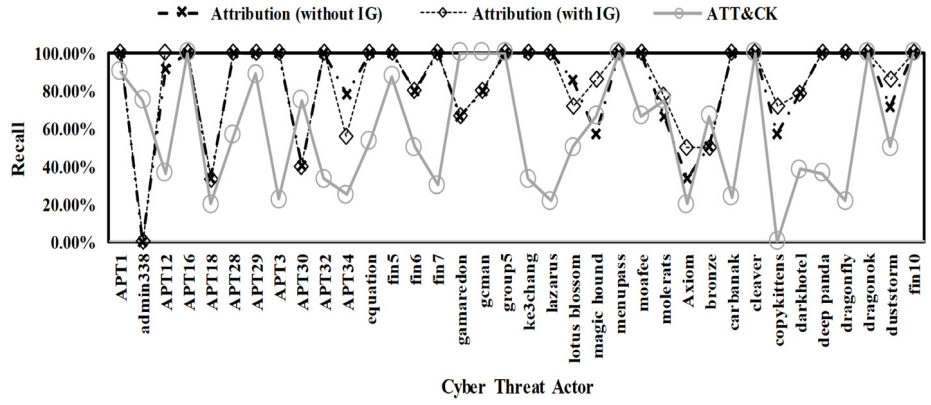


Fig. 14. Attribution recall using Naïve Bayes.

Table 5
Attribution results.

		Accuracy	Precision	Recall	F-measure	False positive rate
Naïve Bayes	Attribution (without IG)	88%	89%	82%	0.83	3%
	Attribution (with IG)	86%	90%	84%	0.84	3%
	ATT&CK	52%	66%	59%	0.53	9%
KNN	Attribution (without IG)	68%	69%	70%	0.67	7%
	Attribution (with IG)	68%	71%	71%	0.68	7%
	ATT&CK	52%	66%	59%	0.53	9%
Decision tree	Attribution (without IG)	82%	76%	74%	0.73	4%
	Attribution (with IG)	72%	76%	74%	0.73	2%
	ATT&CK	3%	0.10%	2.78%	0.002	3%
Random forest	Attribution (without IG)	88%	92%	89%	0.89	3%
	Attribution (with IG)	83%	88%	82%	0.84	2%
	ATT&CK	17%	16%	22%	0.14	15%
DLNN	Attribution (without IG)	94%	90%	89%	0.89	3%
	Attribution (with IG)	86%	91%	88%	0.88	2%
	ATT&CK	11%	17%	18%	0.13	15%

selection (i.e. 92%). The results showed that Naïve Bayes, RF, and DLNN perform well (i.e. achieving high accuracy, precision, recall, and f-measure). The FPR is also low for these models. The feature selection does not appear to improve the accuracy of attribution for all these models, but we observed improvements in precision, recall and f-measure for Naïve Bayes. However, for RF and DLNN, the precision, recall and f-measure decrease. We also trained the five machine learning models using the ATT&CK dataset. ATT&CK provides a single instance high-level IOC dataset for each CTA; thus, we are not able to use cross-validation technique to evaluate the attribution of cyber threats. Hence, we used a separate test data compiled from the recent incidents associated with the CTAs to evaluate the working of the machine learning models for ATT&CK dataset. The evaluation results show low accuracy, precision, recall, f-measure and FPR for the ATT&CK dataset.

Now, we examined the evaluation results of these five machine learning models for individual instances of CTAs. Figs. 9–13 show the precision results for each CTA for the three datasets. Naïve Bayes is unable to identify two CTAs, namely: admin338 and Copykittens. kNN is unable to identify 11 CTAs, namely: admin338, APT12, APT18, APT34, Equation, GCMAN, Group5, Lotus Blossom, Axiom, Copykittens and Fin10. Decision Tree is unable to identify nine CTAs, namely: APT16, Equation, GCMAN, Group5, Ke3chang, Lazarus, Darkhotel and Fin10. RF is unable to identify three CTAs, namely: Group5, Moafee and Copykittens. DLNN is unable to identify two CTAs, i.e., Moafee and Copykittens.

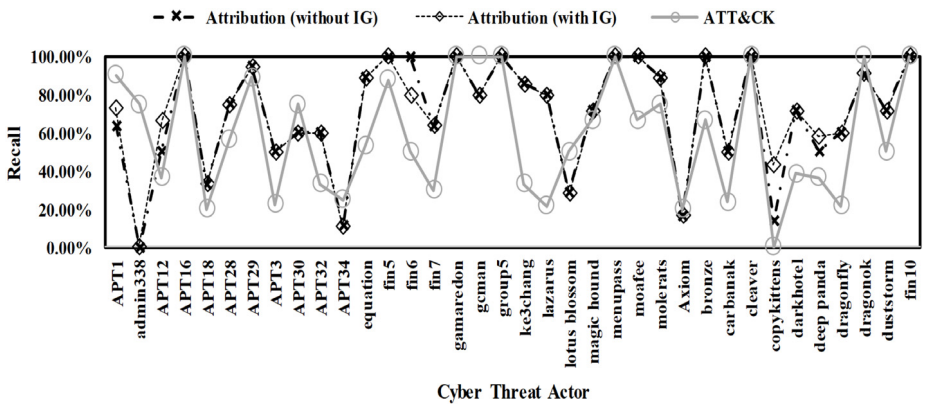


Fig. 15. Attribution recall using kNN.

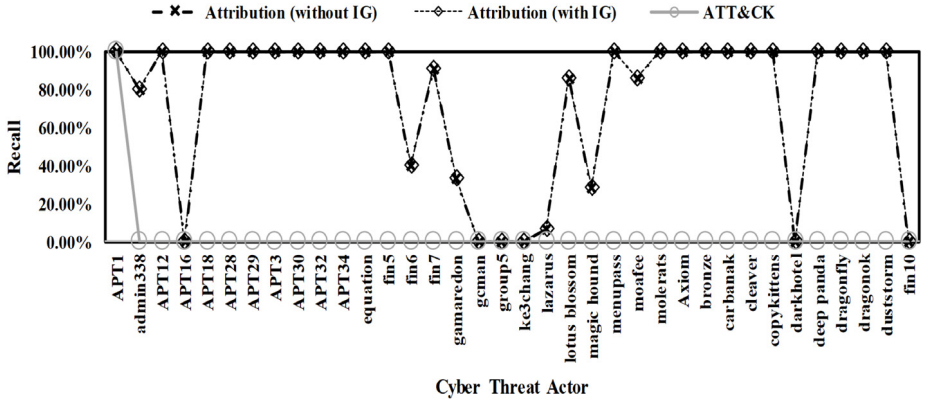


Fig. 16. Attribution recall using decision tree.

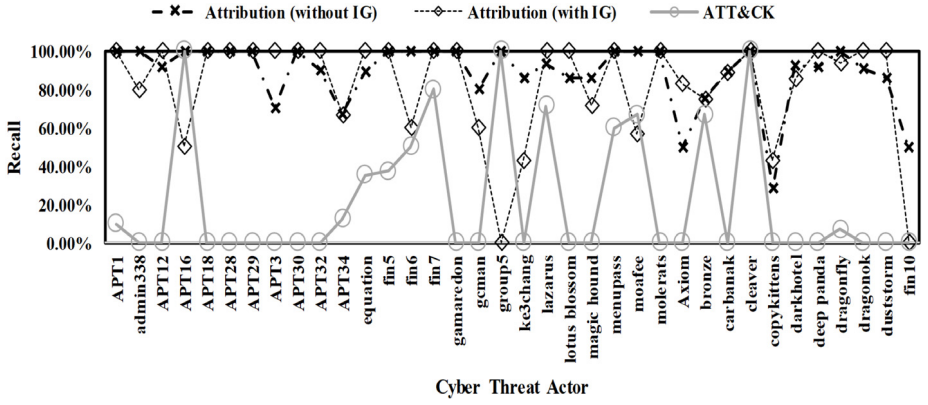


Fig. 17. Attribution recall using random forest.

We also observed that in Group5 and Moafee, a very low number of high-level IOCs are reported – see Fig. 8. This could be due to misclassification. In the case of Copykittens, we observed that it is not classified by four models. When its high-level IOCs are observed, these IOCs were found to highly overlap with IOCs of other CTAs, such as APT 32 and APT 34. Usually, the customized software tools of CTAs do not exhibit overlaps, which are distinguishing features of classification. However, in the case of Copykittens, even the customized software tools showed evidence of overlaps with those of other CTAs. One of APT 32's customized software tools "Cobalt Strike" was, for example, found to be used by Copykittens in their attack campaigns. Based on this observation, it is concluded that there is some kind of connection or partnership between the two CTAs; thus, resulting in the misclassification of Copykittens.

We analyzed the high-level IOCs of six FinTech related CTAs, namely: Fin5 [88], Fin6 [89], Fin7 [90], Fin10 [91], GCMAN [92], and Carbanak [93]. Of these six FinTech related CTAs, Fin5, Fin6, and Fin7 reportedly stole credit card data from Point Of Sale (POS) terminals using different TTPs and ram scrapping malware. Fin10 allegedly stole important corporate data and demanded ransom from the victims. GCMAN and Carbanak allegedly targeted banks to facilitate illegitimate money transfers. The evaluation results demonstrated that our cyber threat attribution framework classifies these CTAs with high precision, recall, and F-measure.

Figs. 14, 15, 16, 17, and 18 show the recall results for each CTA for the three datasets. Naïve Bayes has a low recall for five CTAs, namely: admin338, APT18, APT30, Axiom and Bronze Butler. KNN has a low recall for 13 CTAs (i.e. admin338, APT12, APT18, APT3,

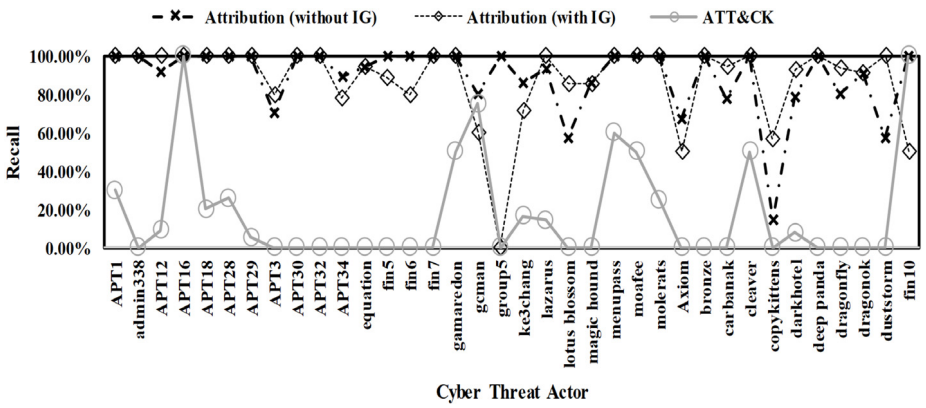


Fig. 18. Attribution recall using DLNN.

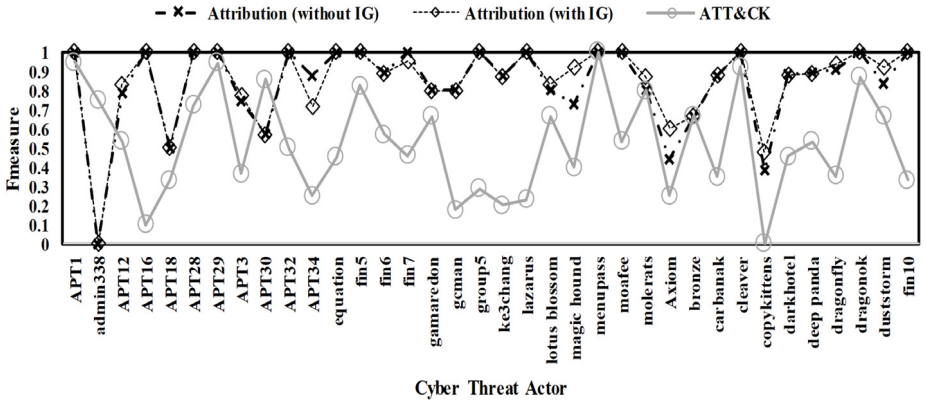


Fig. 19. Attribution F measure using Naïve Bayes.

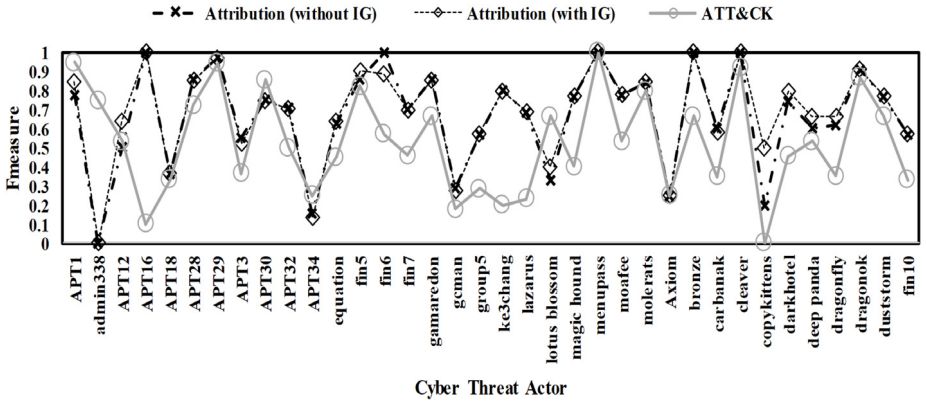


Fig. 20. Attribution F measure using kNN.

APT30, APT32, APT34, Lotus blossom, Axiom, Carbanak, Copykittens, Deep Panda and Dragonfly), and Decision Tree has a low recall for ten CTAs (i.e. APT16, Fin6, Gamaredon, GCMAN, Group5, Ke3chang, Lazarus, Magic Hound, Darkhotel and Fin10). RF has a low recall for three CTAs, namely: Axiom, Copykittens and Fin10, and DLNN has a low recall for only Copykittens. Based on these observations, we conclude that DLNN gives the best results in comparison to the four other machine learning models.

Figs. 19, 20, 21, 22, and 23 show the f-measure results for each CTA for the three datasets. Naïve Bayes has low f-measure for five CTAs, namely: admin338, APT18, APT30, Axiom and Copykittens. Both KNN and Decision Tree have low f-measure for 11 CTAs each (i.e. admin338, APT12, APT18, APT3, APT34, GCMAN, Group5, Lotus Blossom, Axiom, Carbanak and Copykittens; and APT16, Equation, Fin6, Gamaredon, GCMAN, Group5, Ke3chang, Lazarus, Magic Hound, Darkhotel and Fin10). Both RF and DLNN have low f-measure for only Copykittens.

5. Conclusion

As noted by cyber criminologists such as [7], the banking and financial sector is often the 'target of choice' for financially motivated CTAs. Thus, there is a need to ensure that FinTech is adequately secured against increasingly sophisticated APTs, including state-sponsored or state-affiliated actors. This paper highlighted the benefits of using high-level IOCs in attributing cyber threats to their perpetrators. The objective is to identify the threat actor in a timely fashion, if not in real-time so that cyber attacks can

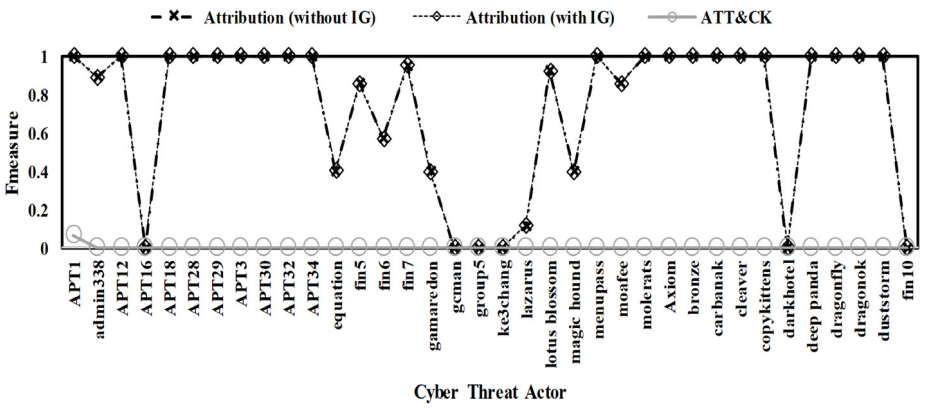


Fig. 21. Attribution F measure using decision tree.

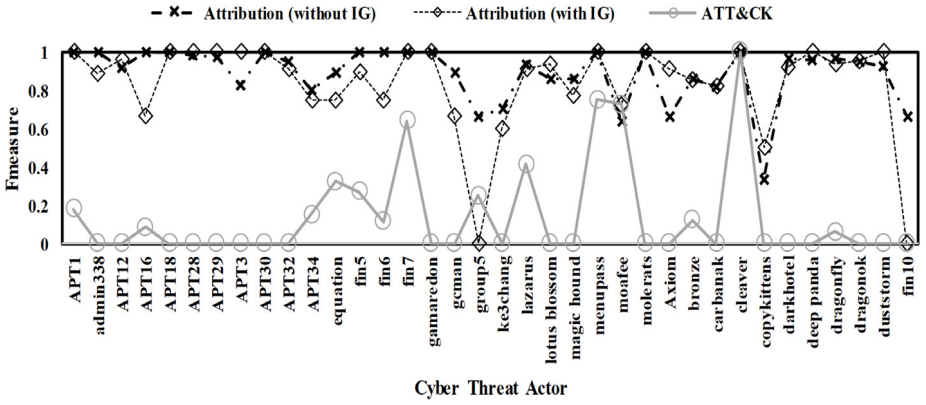


Fig. 22. Attribution F measure using random forest.

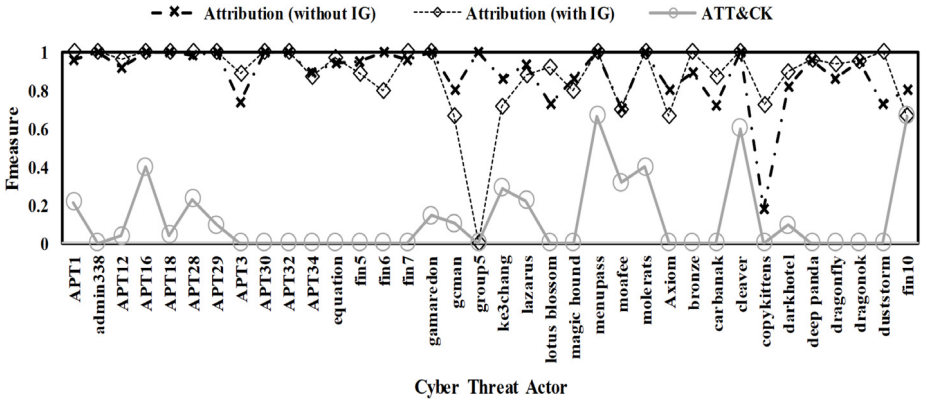


Fig. 23. Attribution F measure using DLNN.

be disrupted. The high-level IOCs relating to 36 well-known threat groups are extracted from unstructured CTI reports. The latter were collected from 26 different cyber threat reporting sources using semantic search technique. The feature labels for high-level IOCs were taken from ATT&CK MITRE. These high-level IOC features profiled the CTAs, which were then used to train the five machine learning models used in this paper (i.e Naïve Bayes, KNN, Decision Tree, Random Forest, and DLNN). The trained models were then used to attribute the threat incidents of these CTAs. Findings from the five machine learning models evaluated in this paper suggested that the DLNN model is more effective than the remaining four models. We also compared the effectiveness of our high-level IOC profiled CTA dataset with the dataset provided by ATT&CK MITRE (a single instance manually compiled for each CTA). The findings demonstrated that the machine learning models trained with our dataset attribute cyber threats with high precision, recall, f-measure, and a low FPR, as compared to the ATT&CK dataset.

However, our proposed attribution framework is not without limitation. For example, the framework is dependent on the threat data. According to a recent cyber security market survey, there are 500 independent cyber security companies around the globe [94] providing CTI feeds. Thus, for an effective attribution, there is a need to objectively and adaptively rank the reputation of these CTI sources on the functional and non-functional requirements of CTI consumers. Therefore, one future research agenda is to build a context-specific cyber threat source reputation model. For example, such a model can be used to adaptively define the stakeholders involved, the taxonomy and metadata relating to reputation

ranking. An extensive Key Performance Indicator (KPI) criteria can also be adaptively defined and appropriate weights will be assigned based on security experts' opinions and other information (e.g. based on input using deep learning techniques), in order to ensure the availability of high-quality data for cyber attribution frameworks.

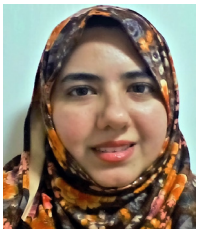
Acknowledgments

K.-K.R. Choo is supported by the Cloud Technology Endowed Professorship.

References

[1] Federal Privacy Council (FPC), Cybersecurity Information Sharing Act of 2015 (CISA), https://www.fpc.gov/19081/ (2015).
[2] U. S. Congress, S. 754 cybersecurity information sharing act of 2015 (2015).
[3] D.A. Wheeler, G.N. Larsen, Techniques for cyber attack attribution, Tech. rep., Institute For Defense Analyses Alexandria VA (2003).
[4] J. Hunker, B. Hutchinson, J. Margulies, Role and challenges for sufficient cyber-attack attribution, Inst. Inf. Infrastruct. Prot. (2008) 5–10.
[5] First half 2018 breach level index report, https://www.gemalto.com/press/pages/data-breaches-compromised-4-5-billion-records-in-first-half-of-2018.aspx, accessed: 2018-12-28.
[6] Equifax data breach affected 2.4 million more consumers, https://www.consumerreports.org/credit-bureaus/equifax-data-breach-was-bigger-than-previously-reported, accessed: 2018-3-17.
[7] K.-K.R. Choo, Cyber threat landscape faced by financial and insurance industry, Trends Issues Crime Criminal Justice 408 (2011) 1–6.
[8] D. Shackleford, S. Northcutt, Who's using cyberthreat intelligence and how? Tech. rep., SANS Institute InfoSec Reading Room (2015).
[9] G. Farnham, K. Leune, Tools and standards for cyber threat intelligence projects, Tech. rep., SANS Institute InfoSec Reading Room (2013).
[10] MITRE, Structured Threat Information eXpression (STIX) A structured language for cyber threat intelligence., http://stixproject.github.io/ (2016).
[11] J. Wattles, How the Equifax data breach happened: What we know now, http://money.cnn.com/2017/09/16/technology/equifax-breach-security-hole/index.html, accessed 2018-3-17 (2017).
[12] M-Trends 2016, https://www2.fireeye.com/rs/848-DID-242/images/Mtrends2016.pdf, accessed 2018-6-23 (2018).
[13] Dmitri Alperovitch, Bears in the Midst: Intrusion into the Democratic National Committee, https://www.crowdstrike.com/blog/bears-midst-intrusion-democratic-national-committee/(2016).
[14] BBC Trending, Conversations with a hacker: What Guccifer 2.0 told me, http://www.bbc.com/news/blogs-trending-38610402 (2017).
[15] Alliance News, Russian Hackers Suspected In Cyberattack On German Parliament, http://www.lse.co.uk/AllNews.asp?code=kwdwehme&headline=Russian_Hackers_Suspected_In_Cyberattack_On_German_Parliament (2015).
[16] Gordon Corera, How France's TV5 was almost destroyed by 'Russian hackers', http://www.bbc.com/news/technology-37590375 (2016).
[17] Josh Meyer, Russian Hackers Post Medical Files of Simone Biles, Serena Williams, https://www.nbcnews.com/storyline/2016-rio-summer-olympics/russian-hackers-post-medical-files-biles-serena-williams-n647571(2016).
[18] Bitly, THE LINK KNOWS ALL. SO CAN YOU, https://bitly.com/ (2017).
[19] Adam Meyers, Danger Close: Fancy Bear Tracking of Ukrainian Field Artillery Units, https://www.crowdstrike.com/blog/danger-close-fancy-bear-tracking-ukrainian-field-artillery-units/ (2016).
[20] Cory Doctorow, Spear phishers with suspected ties to Russian government spoof fake EFF domain, attack White House, https://boingboing.net/2015/08/28/spear-phishers-with-suspected.html (2015).
[21] Crowdstrike, Who is Fancy Bear, https://www.crowdstrike.com/blog/who-is-fancy-bear/ (2016).
[22] ThreatConnect, Russian Cyber Operations on Steroids, https://www.threatconnect.com/blog/fancy-bear-anti-doping-agency-phishing/(2016).
[23] FireEye, APT28: A Window into Russia's Cyber Espionage Operations? https://www.fireeye.com/blog/threat-research/2014/10/apt28-a-window-into-russias-cyber-espionage-operations.html (2014).
[24] Mitre, Adversarial Tactics, Techniques and Common Knowledge, https://attack.mitre.org/wiki/Main_Page (2018).
[25] E.F. Mejia, Act and actor attribution in cyberspace: a proposed analytic framework, Tech. rep., Air University Maxwell AFB AL Strategic Studies Quaterly (2014).
[26] N.C. Rowe, The attribution of cyber warfare, in: Cyber Warfare: A Multidisciplinary Analysis, Vol. 61, 2015.
[27] P. Shakarian, G.I. Simari, G. Moores, S. Parsons, Cyber attribution: an argumentation-based approach, in: Cyber Warfare, Springer, 2015, pp. 151–171.
[28] F. Stranne, U. Bilstrup, L. Ewertsson, Behind the mask–attribution of antagonists in cyberspace and its implications on international conflicts and security issues, in: International Studies Association (ISA)'s 56th Annual, Convention–Global IR and Regional Worlds. A New Agenda for International Studies, New Orleans, Louisiana, United States, 2015.
[29] B. Edwards, A. Furnas, S. Forrest, R. Axelrod, Strategic aspects of cyberattack, attribution, and blame, Proc. Natl. Acad. Sci. (2017) 201700442.
[30] L. Qiang, Y. Zeming, L. Baoxu, J. Zhengwei, Y. Jian, Framework of cyber attack attribution based on threat intelligence, in: International Conference on Interoperability in IoT, Springer, 2016, pp. 92–103.
[31] E. Nunes, N. Kulkarni, P. Shakarian, A. Ruef, J. Little, Cyber-deception and attribution in capture-the-flag exercises, in: Cyber Deception, Springer, 2016, pp. 151–167.
[32] A. Nicholson, T. Watson, P. Norris, A. Duffy, R. Isbell, A taxonomy of technical attribution techniques for cyber attacks, in: European Conference on Information Warfare and Security, Academic Conferences International Limited, 2012, p. 188.
[33] O. Thonnard, W. Mees, M. Dacier, On a multicriteria clustering approach for attack attribution, ACM SIGKDD Explor. Newsl. 12 (1) (2010) 11–20.
[34] J. Hunker, C. Gates, M. Bishop, Attribution requirements for next generation internets, in: Technologies for Homeland Security (HST), 2011 IEEE International Conference on, IEEE, 2011, pp. 345–350.
[35] D.D. Clark, S. Landau, Untangling attribution, Harv. Nat. Secur. J. 2 (2011) 323.
[36] H.-Y. Lock, A. Kliarsky, Using Ioc (Indicators of Compromise) in Malware Forensics, SANS Institute InfoSec Reading Room, 2013.
[37] K. Geers, D. Kindlund, N. Moran, R. Rachwald, World war c: Understanding nation-state motives behind today's advanced cyber attacks, FireEye, Milpitas, CA, USA, Tech. Rep., (2014).
[38] H. Mwiki, T. Dargahi, A. Dehghantanha, K.-K.R. Choo, Analysis and triage of advanced hacking groups targeting western countries critical national infrastructure: Apt28, red october, and regin.
[39] K. Rieck, P. Trinius, C. Willems, T. Holz, Automatic analysis of malware behavior using machine learning, J. Comput. Secur. 19 (4) (2011) 639–668.
[40] E. Hodo, X. Bellekens, A. Hamilton, P.-L. Dubouilh, E. Iorkyase, C. Tachtatzis, R. Atkinson, Threat analysis of iot networks using artificial neural network intrusion detection system, in: Networks, Computers and Communications (ISNCC), International Symposium on, IEEE, 2016, pp. 1–6.
[41] A. Saied, R.E. Overill, T. Radzik, Detection of known and unknown ddos attacks using artificial neural networks, Neurocomputing 172 (2016) 385–393.
[42] M.-J. Kang, J.-W. Kang, Intrusion detection system using deep neural network for in-vehicle network security, PLoS One 11 (6) (2016).
[43] CrowdStrike SaaS Endpoint Protection Threat Intelligence, https://www.crowdstrike.com/, accessed 2018-6-22 (2018).
[44] Secureworks, https://www.secureworks.com/, accessed 2018-6-22 (2018).
[45] FireEye Cyber Security & Malware Protection, https://www.fireeye.com/, accessed 2018-6-22 (2018).
[46] iSIGHT Intelligence Subscriptions, https://www.fireeye.com/solutions/isight-cyber-threat-intelligence-subscriptions.html, accessed 2018-6-22 (2018).
[47] APT Group and Operations - Cyber Peace.org, https://cyber-peace.org/wp-content/uploads/2018/APT-Groups-and-Operations.xlsx, accessed 2018-6-21 (2018).
[48] APT Groups, Operations and Malware Search Engine, https://cse.google.com/cse/publicurl?cx=003248445720253387346:turlh5vi4xc, accessed 2018-6-21 (2018).
[49] APTNotes, https://github.com/kbandla/APTnotes, accessed 2018-6-21 (2018).
[50] ARBOR SOLUTIONS DDoS Protection and Network Visibility backed by industry leading threat intelligence, https://www.netscout.com/arbor, accessed 2018-6-22 (2018).
[51] ZScaler, https://www.zscaler.com/, accessed 2018-6-22 (2018).
[52] NETRESEC Network Forensics and Network Security Monitoring, http://www.netresec.com/, accessed 2018-6-22 (2018).
[53] Check Point Software Technologies, https://www.checkpoint.com/, accessed 2018-6-22 (2018).
[54] An Advanced Analytics Company, http://www.novetta.com/, accessed 2018-6-22 (2018).
[55] Forcepoint Human Centric Cybersecurity, https://www.forcepoint.com/, accessed 2018-6-22 (2018).
[56] PWC Cybersecurity, https://www.pwc.co.uk/issues/cyber-security-data-privacy.html, accessed 2018-6-22 (2018).
[57] Proofpoint Software Company, https://www.proofpoint.com/us, accessed 2018-6-22 (2018).
[58] ClearSky Cyber Security, https://www.clearskysec.com/, accessed 2018-6-22 (2018).
[59] Cylance Artificial Intelligence Based Advanced Threat Prevention, www.cylance.com/, accessed 2018-6-22 (2018).
[60] Trend Micro Enterprise Cybersecurity Solutions, https://www.trendmicro.com/en_my/business.html, accessed 2018-6-22 (2018).
[61] Symantec Global Leader In Next Generation Cyber Security, https://www.symantec.com/, accessed 2018-6-22 (2018).
[62] ThreatConnect Security Operations and Analytics Platform, https://www.threatconnect.com/, accessed 2018-6-22 (2018).

[63] Kaspersky Lab Kaspersky Antivirus Protection & Internet Security, https://www.kaspersky.com/, accessed 2018-6-22 (2018).
[64] Palo Alto Networks, https://www.paloaltonetworks.com/, accessed 2018-6-22 (2018).
[65] BrightTALK, https://www.brighttalk.com/search?q=cybersecurity, accessed 2018-6-22 (2018).
[66] ThreatMiner Data Mining for Threat Intelligence, https://www.threatminer.org/, accessed 2018-6-22 (2018).
[67] WeLiveSecurity, https://www.welivesecurity.com/, accessed 2018-6-22 (2018).
[68] infosecurity GROUP, https://www.infosecurity-magazine.com/, accessed 2018-6-22 (2018).
[69] threat post, https://threatpost.com/, accessed 2018-6-22 (2018).
[70] THECITIZENLAB, https://citizenlab.ca/tag/cybersecurity/, accessed 2018-6-22 (2018).
[71] Brian Krebs, THECITIZENLAB, https://krebsonsecurity.com/, accessed 2018-6-22 (2018).
[72] Pierluigi Paganini, security affairs, https://securityaffairs.co/wordpress/, accessed 2018-6-22 (2018).
[73] US-CERT: United States Computer Emergency Readiness Team, https://www.us-cert.gov/, accessed 2018-6-21 (2018).
[74] Who is Fancy Bear? https://www.crowdstrike.com/blog/who-is-fancy-bear/, accessed 2018-7-4 (2016).
[75] Trend Micro Cyber Safety Solutions Team, Untangling the Patchwork Cyberespionage Group, https://blog.trendmicro.com/trendlabs-security-intelligence/untangling-the-patchwork-cyberespionage-group/ (2017).
[76] David Bizeul, Ivan Fontarensky, Ronan Mouchoux, Fabien Perigaud, Cedric Pernet, The Eye of the Tiger, https://www.symantec.com/connect/blogs/strider-cyberespionage-group-turns-eye-sauron-targets(2018).
[77] Symantec Security Response, Strider: Cyberespionage group turns eye of Sauron on targets, http://blog.cassidiancybersecurity.com/post/2014/07/The-Eye-of-the-Tiger2 (2014).
[78] Bryan Lee, Robert Falcone, OilRig Archives - Palo Alto Networks Blog, https://researchcenter.paloaltonetworks.com/tag/oilrig/ (2018).
[79] Turla Mosquito: A shift towards more generic tools, https://www.welivesecurity.com/2018/05/22/turla-mosquito-shift-towards-generic-tools/(2018).
[80] J. Mulder, M. Stingley, Mimikatz overview, defenses and detection, Tech. rep., SANS Institute InfoSec Reading Room (2014).
[81] H. Jeffreys, Scientific Inference, third ed., Cambridge University Press, England, 1974.
[82] E. Fix, J. Hodges, An important contribution to nonparametric discriminant analysis and density estimation, Internat. Statist. Rev. 3 (57) (1951) 233–238.
[83] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.
[84] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[85] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Netw. 61 (2015) 85–117.
[86] Apache Solr, http://lucene.apache.org/solr/, accessed 2018-6-21 (2018).
[87] Apache Lucene, http://lucene.apache.org/, accessed 2018-6-21 (2018).
[88] Prolific cybercrime gang favors legit login credentials, https://www.darkreading.com/analytics/prolific-cybercrime-gang-favors-legit-login-credentials/d/d-id/1322645, accessed: 2018-12-28.
[89] Follow the money: Dissecting the operations of the cyber crime group fin6, Tech. rep., Fireeye Threat Intelligence (http://www.fireeye.com) (2016).
[90] Fin7 spear phishing campaign targets personnel involved in sec filings, https://www.fireeye.com/blog/threat-research/2017/03/fin7_spear_phishing.html, accessed: 2018-12-28.
[91] Fin10 anatomy of a cyber extortion operation, Tech. rep., Fireeye Threat Intelligence (http://www.fireeye.com) (2017).
[92] Apt-style bank robberies increase with metel, gcman and carbanak 2.0 attacks, Tech. rep., Kaspersky Lab's Global Research and Analysis Team (http://www.kaspersky.com) (2016).
[93] Carbanak apt: The great bank robbery, Tech. rep., Kaspersky Lab's Global Research and Analysis Team (http://www.kaspersky.com) (2015).
[94] Cybersecurity 500, https://cybersecurityventures.com/cybersecurity-500/, accessed 2019-1-3 (2018).



Umara Noor received Gold Medal for her M.S. degree in Information Technology from Institute of Management Sciences, Pakistan in 2011. She is currently pursuing her Ph.D. degree in Information Technology from National University of Sciences and Technology (NUST), Pakistan in the domain of machine learning approaches to automated analysis of cyber security threats under the research supervision of Dr. Zahid Anwar. She is also working as a faculty member in the Department of Computer Science and Software Engineering, International Islamic University (IIU), Pakistan. Her research interests include experimenting with machine learning techniques to enhance security systems and services.

Zahid Anwar received his Ph.D. and M.S. degrees in Computer Sciences in 2008 and 2005 respectively from the University of Illinois at Urbana-Champaign. Zahid has worked as a software engineer and researcher at IBM, Intel, Motorola, National Center for Supercomputing Applications (NCSA), xFlow Research and CERN on various projects related to information security, operating systems design and data analytics. Zahid holds postdoctorate experience from Concordia University. He has worked as a faculty member at the National University of Sciences and Technology and the University of North Carolina at Charlotte. He is currently an Assistant Professor at Fontbonne University.

Tehmina Amjad is an assistant professor at the *Computer Science and Software Engineering department of the International Islamic University, Islamabad, Pakistan*. She received her Ph.D. Computer Science degree from the same institute in 2015. She attended *Indiana University, Bloomington, Indiana, USA*, under split Ph.D. Program to conduct research during Ph.D. Her research interests include information retrieval, data mining, social network analysis, probabilistic topic models, machine learning and data grids.

Kim-Kwang Raymond Choo received the Ph.D. in Information Security in 2006 from Queensland University of Technology, Australia. He currently holds the Cloud Technology Endowed Professorship at The University of Texas at San Antonio (UTSA). In 2016, he was named the Cybersecurity Educator of the Year-APAC (Cybersecurity Excellence Awards are produced in cooperation with the Information Security Community on LinkedIn), and in 2015 he and his team won the Digital Forensics Research Challenge organized by Germany's University of Erlangen–Nuremberg. He is the recipient of the 2018 UTSA College of Business Col. Jean Piccione and Lt. Col. Philip Piccione Endowed Research Award for Tenured Faculty, IEEE TrustCom 2018 Best Paper Award, ESORICS 2015 Best Research Paper Award, 2014 Highly Commended Award by the Australia New Zealand Policing Advisory Agency, Fulbright Scholarship in 2009, 2008 Australia Day Achievement Medallion, and British Computer Society's Wilkes Award in 2008. He is also a Fellow of the Australian Computer Society, an IEEE Senior Member, and Co-Chair of IEEE Multimedia Communications Technical Committee (MMTC)'s Digital Rights Management for Multimedia Interest Group.