Formulation of Association Rule Mining (ARM) for an Effective Cyber Attack Attribution in Cyber Threat Intelligence (CTI)

Md Sahrom Abu[1] Aswami Ariffin[4]
Malaysian Computer Emergency Response Team
Cybersecurity Malaysia
Cyberjaya, Selangor DE, Malaysia

Siti Rahayu Selamat[2] Robiah Yusof [3]
Faculty of Information Technology and Communication
Universiti Teknikal Malaysia Melaka
Durian Tunggal, Melaka, Malaysia

***Abstract*—In recent year, an adversary has improved their Tactic, Technique and Procedure (TTPs) in launching cyberattack that make it less predictable, more persistent, resourceful and better funded. So many organisation has opted to use Cyber Threat Intelligence (CTI) in their security posture in attributing cyberattack effectively. However, to fully leverage the massive amount of data in CTI for threat attribution, an organisation needs to spend their focus more on discovering the hidden knowledge behind the voluminous data to produce an effective cyberattack attribution. Hence this paper emphasized on the research of association analysis in CTI process for cyber attack attribution. The aim of this paper is to formulate association ruleset to perform the attribution process in the CTI. The Apriori algorithm is used to formulate association ruleset in association analysis process and is known as the CTI Association Ruleset (CTI-AR). Interestingness measure indicator specially *support* (*s*), *confidence* (*c*) and *lift* (*l*) are used to measure the practicality, validity and filtering the CTI-AR. The results showed that CTI-AR effectively identify the attributes, relationship between attributes and attribution level group of cyberattack in CTI. This research has a high potential of being expanded into cyber threat hunting process in providing a more proactive cybersecurity environment.**

***Keywords*—Cyber threat intelligence (CTI); association rule mining; apriori algorithm; attribution; interestingness measures**

I. Introduction

As the Tactic, Technique and Procedure (TTPs) used by an adversary become unpredictable, determined, imaginative, funded, far more coordinated and financially motivated, acquiring useful information from threat information sharing is essential for cyberattack attribution. Cyber Threat Intelligence (CTI), as one of threat information sharing frameworks, has received a lot of media attention in mitigating and reducing cyberattack infection. However, one of the common issues in CTI is the quality of voluminous data from shared information and there is scarce literature in discussing the meaning of quality, basic methods and tools for assessment [1]. A huge volume of data in the CTI consists of raw data without a meaningful relationship between the data. This voluminous data can lead to the ineffectiveness of identifying cyberattack attribution levels due to a lack of useful data from various data sources. Cyberattack attribution process can provide a meaningful relationship between data by identifying the attribution level and hidden knowledge behind the data to assist organizations in decision making [2]. However, the current cyberattack attribution technique is ineffective in handling the voluminous data in CTI because it relies heavily on the manual process performed by the security analyst and is strictly related to the analyst's knowledge, creating human bias and error-prone [3].

This paper highlight the data mining process in solving the voluminous data issue that can help security analyst to find the relationship between datasets and perform the cyberattack attribution process in CTI. The proposed study was to formulate an association ruleset for cyberattack attribution process in CTI. This ruleset would enable the discovery of hidden knowledge behind the raw data in identifying the attribution level.

The remaining of the paper is organized as follows: Section II presents the research background and related work based on association rules mining in CTI. Section III describes the proposed methodology that includes data collection using CTI feeds, dataset for CTI feeds, association rules mining in CTI framework and formulation of association ruleset using the Apriori algorithm. While Section IV represents the outcome for association ruleset formulation in CTI and evaluate the ruleset generated using interestingness measures. Finally, Section V provides a brief conclusion for this paper.

II. Research Background and Related Works

A. Cyber Threat Intelligence (CTI) for Threat Attribution

There has been a lot of studies in the area of data mining to discover its insights in terms of large groups of items or objects in transactional databases, relational databases, or other information repositories using Association Rule Mining (ARM) technique. Association Rule Mining (ARM) is an important research branch of data mining which has attracted many data mining researchers due to its capability to discover useful and interesting patterns from extensive, noisy, fuzzy and stochastic data. The concept of ARM was introduced by Agrawal and Srikant [4]. In the data mining field, ARM can be utilized as a part of cyberattack attribution process in CTI to discover the hidden knowledge behind raw data. A critical issue for cyberattack attribution in CTI is how to successfully and effectively extract the hidden knowledge from the

voluminous data and feasibly create the association ruleset for cyberattack attribution to assist security analysts in decision making.

Since the introduction of the first concept of ARM by Agrawal et al. [5], a wide variety of efficient ARM algorithms for generating association rules have been proposed over time. Some of the well known and most important algorithms are Apriori, Apriori-TID, SETM, Apriori Hybrid, AIS and Fp-growth [6].

Currently, the most widely used algorithms in ARM is Apriori Algorithm. Agrawal and Srikant developed this algorithm to study customers' purchasing behavior in supermarkets where goods are often purchased together by customers [4]. Besides, the Apriori Algorithm has also been used successfully in many areas of daily life, including energy, recruitment, communication protocol, monitoring and network traffic behavior [7]. Hence, the implementation of the Apriori Algorithm in determining malicious network traffic behavior can help security analysts to study attacker behavior in conducting cyberattack.

Apriori algorithm has been implemented in various fields. Khalili and Sami [8] proposed an industrial intrusion detection approach to mitigate threats to cyber physical systems that utilise sequential patterns extracted by the Apriori algorithm to aid experts in identifying critical states. The study showed Apriori could be employed in the extraction of sequential patterns for industrial process monitoring. A study conducted by Hsiao et al. investigated the use of the Apriori algorithm to track adversaries transitioning through sequences of hosts to launch an attack [9]. Data are retrieved from network packets to determine the host sequence. The Apriori algorithm is proven to be suitable for this study. Meanwhile Liu et al. have utilized Apriori and MS-Apriori algorithm to investigate the relationship of data for network footprint (NFP) which consists of DPI data from ISPs and Crawler data from Web for App usage analysis [7]. The result provides insights for mobile application developers to recommend other applications for their users based on their interest and usage pattern. Adebayo and Abdul Aziz presented a novel knowledge-based database discovery model that utilizes an improvised apriori algorithm with Particle Swarm Optimization (PSO) to classify and detect malicious android application [10]. The usage of several rule detectors can maximize the true positive rate of detecting malicious code, whereas the false positive rate of wrongful detection is minimized. The use of the Apriori algorithm outside the cybersecurity domain has also been explored. It is used for smart health services in a study conducted by Jung, Kim and Chung [11]. The Apriori algorithm was used for a series of patient images acquired through the surveillance technology to generate bio-sequential patient patterns. The bio-sequential patterns are then used to create a basis for a bio-sequential pattern and any deviation from this could result in a possible emergency. The study demonstrated that the Apriori algorithm is used to develop bio-sequential patterns and could be used to extract patterns from the adversary SSH command sequence. Other than that, the Apriori algorithm is also being employed in a study to discover the contributory crash-risk factors of hazardous material (HAZMAT) vehicle-involved crashes on expressways [12]. The findings from this study indicated that ARM is a feasible technique of data mining that can be used to draw correlations between HAZMAT vehicle-involved accidents and significant crash-risk factors, and has the potential to provide more easy-to-understand findings and applicable lessons for improving the expressways safety.

In this paper, we collect CTI data from current cyberattacks which contained network resources and attackers' behaviour and do association rules analysis using Apriori to generate rules. These rules would enable the discovery of hidden knowledge behind the raw data in identifying the attribution level.

III. Methodology

In this section, the experimental design to generate the association ruleset in CTI for cyberattack attribution is presented. The input of this experimental design was CTI feeds from OSINT. Data preprocessing technique were used to clean the CTI feeds and produce meaningful data that were used to generate the association ruleset. By conducting this experiment, the association ruleset could be produced to identify the hidden knowledge behind attributes in CTI feeds and identify the attribution level for cyberattack attribution in CTI. The design of experiments is shown in Fig. 1. Fig. 1 illustrates the entire process of association rule mining in CTI framework that consists i) Preprocessing network traffic data, ii) Generating logical rules using Apriori algorithm and iii) Apply the generated rule to facilitate cyber attack attribution. The Apriori Algorithm can discover groups of items occurring frequently together in lots of transactions and such groups of items are called frequent itemsets. The association rule generated from this process is measured using *support*, *confidence*, and *lift*. Given a set of transaction, the problem of mining association rules is to generate all association rules that have support and confidence greater than the user-specified minimum support (called *minsup*) and minimum confidence (called *minconf*) respectively.

To conduct Apriori algorithm on our dataset, we used R to process the filtered data and visualize the result. R is a language and environment for statistical computing, data mining and graphics.

A. Data Collection for CTI Feeds

Data collection for this paper is limited to CTI feeds from OSINT that related to network intrusion activities. For this paper, OSINT CTI feeds from Shadowserver, Lebahnet and MITRE as shown in Fig. 2 has been chosen because it can provide various types of useful information and Indicators of Compromise (IoC) for cyberattack attribution [13]–[15]. The focus of this research was to gather CTI data comprising network resources and attacker behaviour from existing cyberattack.

Fig. 2 shows data collection process for CTI feeds. An API from each CTI feeds was used to collect the data, respectively. Thus, a scraper was used to collect popular network resources such as the domain of search engines or government website, IP address of common DNS server and MD5 hash value of notorious malware from CTI feeds. The examples of attributes collected from each CTI feeds are listed in Table I.

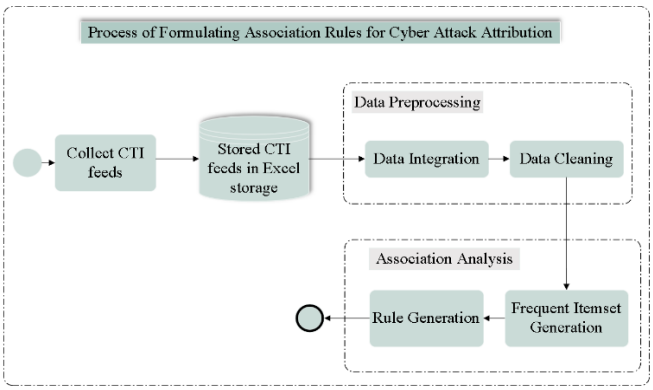

Fig. 1. Experimental Design.

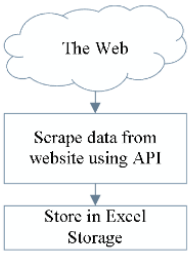

Fig. 2. Data Collection Process.

Shadowserver security feeds provided information about all the infected machines, drones, and zombies that were captured from the monitoring of IRC Command and Controls, capturing IP connections to HTTP botnets, or the IPs of spam relays. Lebahnet security feeds provided valuable supporting information such as network trends and malicious activities that were captured using a collection of distributed honeypots. Both security feeds could provide basic indicators of compromise such as IP address, domain name, URLs, hash value, malware infection type and geolocation. In contrast to Shadowserver and Lebahnet, MITRE knowledgebase was about high-level IOC that related to the behaviour of cybercriminals. MITRE datasets contained various tactics, techniques, software or tools and attackers groups that involved different stages of a cyberattack when infiltrating the network and exfiltrating data. The combination of basic IOC from Shadowserver and Lebahnet and attackers behavior from MITRE knowledgebased was essential in identifying the attribution level for cyberattack attribution in CTI.

B. Dataset for CTI Feeds

The domain of this research was limited to the cyber threat intelligence that related to network intrusion activities and the datasets limited to CTI feeds from OSINT. An API from each CTI feeds was used to collect the data, respectively. Thus, a scraper was used to collect popular network resources such as the domain of search engines or government website, IP address of common DNS server and MD5 hash value of notorious malware from CTI feeds. The CTI feeds covered the top 3 highest infections from 2018 until 2019 in order to be considered relevant cyberattack effort [16]. The summary of each dataset (DS) is depicted in Table II.

TABLE I. THE DETAIL FIELDS IN CTI FEED

(√ = attribute available, x = attribute does not available)
SS=Shadowserver, L=Lebahnet, M=MITRE

Attribute	Data Source			Description
	SS	L	M	
timestamp	√	√	x	the date and time attack captured by the sensor
hashes	x	√	x	malicious file hashes reported in the threat report associated with a particular network resource.
domains	√	x	x	malicious domains reported in the threat report associated with a particular IP resource.
subdomains	√		x	sub-domains reported in a domain resource.
av scans	x	√	x	anti-virus detections reported in the threat report for a network resource or file hash.
source IPs address	√	√	x	malicious IPs used in the attack among the IPs reported in the threat report associated with a particular network resource.
source port number	√	√	x	source of the attacker port number
destination IP address	√	√	x	source of the compromised host
destination port number	√	√	x	source of the compromised port number
URLs	√	x	x	malicious URLs used in the attack among the URLs reported in the threat report associated with a particular network resource.
GeoIP	√	x	x	country of IP or URL location
infection type	√	√	x	malware name as defined by anti-virus detection
technique	x	x	√	technique related to specific threat actors or threat groups
tactic	x	x	√	tactics related to specific actors or threat groups
software or tools	x	x	√	software or tools tactics related to specific actors or threat groups
group	x	x	√	threat actors or groups of threat actors associated with cyberattack

TABLE II. SUMMARY OF THE DATASET FOR EVALUATION

Dataset	Start Date	End Date	Data Source	Total record
DS1	01/05/2018	31/05/2018	Shadowserver	334848
DS2	01/01/2018	31/01/2018	Lebahnet	498
DS3	01/01/2018	31/01/2018	MITRE	15216
DS4	01/03/2019	30/03/2019	Shadowserver	462885
DS5	01/07/2019	31/07/2019	Lebahnet	46
DS6	01/08/2019	31/08/2019	MITRE	4356
DS7	01/06/2018	30/06/2018	Shadowserver	332874
DS8	01/11/2018	30/11/2018	Lebahnet	406
DS9	01/04/2018	31/04/2018	MITRE	21283
DS10	01/07/2019	31/07/2019	Shadowserver	933665
DS11	01/08/2019	31/08/2019	Lebahnet	46
DS12	01/09/2019	30/09/2019	MITRE	5584

Table II shows four datasets from Shadowserver, four datasets from Lebahnet and four datasets from MITRE were collected in this research. The total datasets is twelve and naming as DS1, DS2, DS3, DS4, DS5, DS6, DS7, DS8, DS9, DS10, DS11, and DS12. DS1 to DS3 used for training purposes and explain in Section III (C). While DS4 to DS12 used for evaluation and validation purposes but only result for DS4 explain in Section IV. The rest of DSs were using the same process, hence, adopting the similar explanation as DS4.

C. Association Rule Mining Algorithm in CTI Framework

After the CTI feeds have been preprocessed for producing clean and useful data, the results will be used for association analysis to formulate an association ruleset. This association ruleset is to facilitate a cyber-attack attribution process in the CTI framework to produce an effective threat attribution. The association ruleset can assist security analysts in identifying the origin of the cyberattack and cyberattack attribution level.

To have a general view on the result generated by using R, we set the minimum support value as 0.001 and the minimum confidence value as 0.5. The overall association ruleset analysis classification in CTI was shown in Table III.

The attribution level was divided into three levels namely Level 1, Level 2 and Level 3 [17]. The attributes in Level 1 consisted of IP address, malware type, hash value and port number, Level 2 was Geolocation and Level 3 needed further analysis of the attributes from Level 1 and 2 to identify the person or attack campaign used by an attacker to launch the cyberattack. However, if the dataset acquired contained the TTP about attackers' behaviour such as datasets from MITRE, then, the attribution for Level 3 was achievable without further analysis from the association ruleset in Level 1 and 2.

Based on the analysis in Table III, three attribution levels can be used to identify the identity and location of an attacker and it can be correlated to CTI type to ease a decision making in an organization.

Table IV depicts the relationship of attribution level and its attribute with CTI types that are useful for verifying the effectiveness of the proposed cyberattack attribution in CTI. Level 1 and Level 2 are parts of tactical intelligence, and the outputs can help an organization to deal quickly and accurately through threatening indicators and prioritize vulnerabilities patches. Level 3 is part of operational intelligence, and its output can improve the detection rate and prevent future incidents as attacks can be seen in a clear context. The conclusion of output from level 1,2 and 3 are part of strategic intelligence which can drive organizations' decision making in terms of security countermeasures and improved areas through comprehending the current attack trends and financial impact to organizations.

TABLE III. OVERALL ASSOCIATION RULESET CLASSIFICATION

(√ = Attribute found, x = Attribute does not found)													
Attribution Level	List of Attributte	Attribute Type									Number of Ruleset in DS		
		IP	hashvalue	URL	Infection type	GeoIP	Technique	Tactic	Software/Tools	Threat actor/Group	DS1	DS2	DS3
Level 1	'10.0.0.2', '37a98c6150d2317eb6e0df1516a5b3a4', '445', '8a4e9f688c6d0effd0fa17461352ed3e', 'Gen:Variant.Zusy.238725', '1922', '208.100.26.241', '80', 'lethic'	√	√	√	√						7	37	0
Level 2	'AM', 'MY', 'US'					√					40	0	0
Level 3	'AppCert', 'Browser', 'COM', 'Component', 'DLLs', 'Distributed', 'Doppelgänging', 'Driver', 'Execution', 'Extra', 'File', 'Hooking', 'Image', 'Injection', 'LSASS', 'Memory', 'Model', 'Mshta', 'Object', 'Options', 'Process', 'Window', 'and', 'apt33', 'cobalt', 'command-and-control', 'credential-access', 'defense-evasion', 'empire', 'execution', 'group', 'lateral-movement', 'mimikatz', 'persistence', 'privilege-escalation', 'strike'						√	√	√	x	0	0	4

TABLE IV. THE ATTRIBUTION LEVEL AND ATTRIBUTE RELATIONSHIP WITH CTI TYPE

Attribution Level	Attribute	CTI type	
Level 1: Cyberweapon	hash value, IP, domain name, URLs	Tactical	Strategic
Level 2: Geolocation	GeoIP		
Level 3: Person or Organization	TTP that consist of technique, tactic, software/tools, campaign name and threat actor name	Operational	

Based on overall association ruleset analysis classification in Table III and attribution level and attribute relationship with CTI type in Table IV, Attribution Level Group for each ruleset (ALGR) is proposed as shown in Table V.

TABLE V. ATTRIBUTION LEVEL GROUP RULESET

Attribution Level Group Ruleset (ALGR)	Description
ALGR1	This group is to represent any ruleset under attribution level 1
ALGR2	This group is to represent any ruleset under attribution level 2
ALGR3	This group is to represent any ruleset under attribution level 3

By using the association ruleset classification in Table III and the proposed ALGR from Table V, the general association ruleset can be defined as an equation (1).

$$\{LHS.A_n\} \implies \{RHS.A_n\} = ALGRn \quad (1)$$

Where, *n*= represent attribution level, Level 1, Level 2 or Level 3; *LHS.A*= Attribute from attribution level *n* from the left-hand side, *RHS.A*= Attribute from attribution level *n* from the right-hand side, and ALGR = the attribution level group ruleset. While the ruleset representation from the general equation in (1) can be;

$$\{IP, malware\ type, hash\ value\} \implies \{geolocation\} = ALGRn$$

$$\left\{\begin{matrix}195.38.137.100,\\ 7867de13bf22a7f3e3559044053e33e7,\\ gamarue\end{matrix}\right\} \implies \{RUS\} = ALGR2$$

In this paper, ALGR, as illustrated in Table V and Equation (1), are used to perform cyberattack attribution in CTI.

D. Formulation of Association Ruleset in CTI

In order to prevent cybersecurity threat from causing a significant impact on business and daily life, an actionable threat intelligence with clean data can help an organization in making a fast decision for cyberattack attribution. Cyberattack attribution is defined as a process to identify the location and identity of attackers involved in cyberattack. It is a demanding task that requires a comprehensive intelligence or context to achieve the attribution levels that are divided into three levels namely (1) Attribution to the specific hosts involved in the attack, (2) Attribution to the primary controlling host, (3) Attribution to the actual human actor and attribution to an organization with the specific intent to attack. These attribution levels can only be achieved when an effective threat intelligence framework is in place. To achieve an effective threat intelligence framework, an organization needs to think of how to build a framework deemed appropriate, specifically, in gaining the hidden information behind the raw data in CTI to assist security analysts in performing cyberattack attribution. Hence, this research focused on formulating an association ruleset in CTI framework to perform cyberattack attribution in CTI. Fig. 3 illustrates Apriori algorithm technique that was used to formulate the association ruleset from CTI OSINT feeds that were collected through CTI framework.

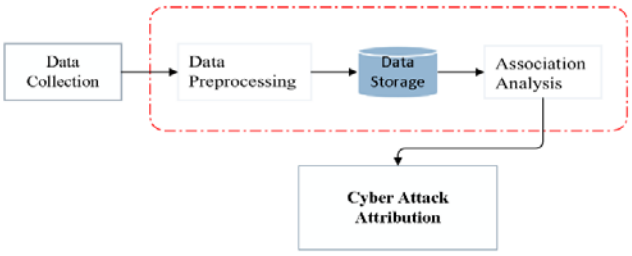

Fig. 3. The Proposed Solution for Cyberattack Attribution in CTI.

Fig. 3 shows the proposed solution to formulate an association ruleset for cyberattack attribution in CTI, which consists of data preprocessing and attribution analysis. The formulation of association ruleset in CTI name as CTI Association Ruleset (CTI-AR) is shown in Table VI.

In Table VI, the purpose and process in CTI framework show that meaningful data that are derived from the preprocessing process are used by the attribution analysis process to identify the attribute and attribution level.

TABLE VI. FORMULATION OF ASSOCIATION RULESET FOR CYBERATTACK ATTRIBUTION PROCESS IN CTI

Criteria	Purpose	Process
Data preprocessing	produce meaningful data, to provide context to raw data	Preprocess raw data
Cyberattack attribution analysis	to identify the threat attribution level, to identify attributes in attribution level	Identify attribute and attribution levels

IV. RESULT

The objective of this section is to present the result of CTI-AR implementation and its effectiveness in performing cyberattack attribution in CTI.. This CTI-AR would enable the discovery of hidden knowledge behind the raw data in identifying the attribution level and help security analyst in making a decision for cyber attack attribution. An objective interestingness measure was used to filter and rank the massive amount of association ruleset or CTI-AR generated by Apriori algorithm. This research applied three objective evaluation indicators that were frequently used in Apriori algorithm which were *support* (*s*), *confidence* (*c*) and *lift* (*l*) to measure and determine the interest of ruleset [18]. *Support* reflected the practicality or usefulness of association rules, *confidence* reflected the validity or reliability of association rules and *lift* was to complement previous two evaluation indicators by filtering and removing wrong and meaningless ruleset.

A. Association Rules Analysis for Dataset

The dataset used to mine the frequent itemset was obtained from the 'Shadowserver security feed' named "ss_2019_3.csv". The dataset, dated from 01/05/2018 to 31/05/2018, consisted of malicious network transaction data in Malaysia. It comprised 462885 rows and 35 columns of data, as shown in Fig. 4.

After performing data cleaning by removing incomplete data and filling the missing values, only eight columns of attributes were selected for discovering frequent itemsets as described in Table VII.

Fig. 5 shows a snippet preprocess data for DS4. Apriori algorithm used an iterative level-wise search technique to discover (k + 1)-itemsets from k-itemsets. First, the dataset was scanned to identify all the frequent 1-itemsets by counting each of them and capturing those that satisfy the minimum support threshold. The identification of each frequent itemset required the scanning of the entire dataset until no more frequent k-itemsets was possible to be identified. As for DS4, the minimum support threshold used was 20% or 0.2. Therefore, only the attributes that fulfilled a minimum support count of 0.2 were included in the ruleset generation process.

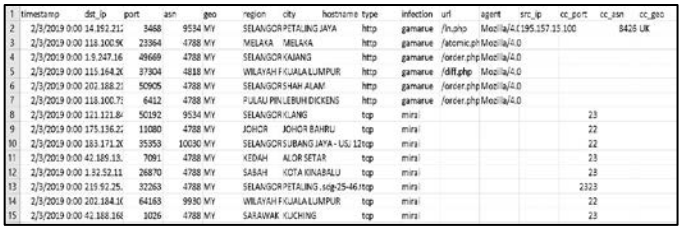

Fig. 4. Raw Dataset 4 (DS4).

TABLE VII. DESCRIPTION OF THE ATTRIBUTE FOR DS4

Attribute	Description
timestamp	Timestamp is "DAY MON DD HH:MM:SS YYYY", where DAY is the day of the week, MON is the name of the month, DD is the day of the month, HH:MM:SS is the time of day using a 24-hour clock, and YYYY is the year. The time zone is +0800
dst_ip	Destination IP for infected device
port	Source port of the victim IP connection
geo	The country where the botnet resides
infection	Malware group classification name
src_ip	The IP used by an attacker to manage (C&C) device
src_port	Server-side port for C&C IP
cc_geo	Country of the C&C server

	timestamp	dst_ip	port	geo	infection	src_ip	cc_port	cc_geo
2	3/2/2019 0:00	14.192.212.162	3468	MY	gamarue	195.157.15.100	22	UK
3	3/2/2019 0:00	118.100.90.229	23364	MY	gamarue	195.38.137.100	22	DE
4	3/2/2019 0:00	1.9.247.166	49669	MY	gamarue	195.38.137.100	22	DE
5	3/2/2019 0:00	115.164.204.231	37304	MY	gamarue	195.38.137.100	22	DE
6	3/2/2019 0:00	202.188.210.106	50905	MY	gamarue	195.38.137.100	22	DE
7	3/2/2019 0:00	118.100.73.140	6412	MY	gamarue	195.38.137.100	22	DE
8	3/2/2019 0:00	121.121.84.222	50192	MY	mirai	195.38.137.100	23	DE
9	3/2/2019 0:00	175.136.226.195	11080	MY	mirai	195.38.137.100	22	DE
10	3/2/2019 0:00	183.171.208.23	35353	MY	mirai	195.38.137.100	22	DE

Fig. 5. Preprocessed Data for Dataset 4 (DS4).

By using the frequent itemset identification process in the Fig. 1, the results of frequent itemsets for DS4 with minimum support count 0.2 were ['195.38.137.100', '22', 'AM', 'DE', 'MY', 'US', 'gamarue']. Then, these frequent itemsets were applied to ruleset generation process as in Fig. 1 to create association ruleset with the predefined minimum confidence (*minconf*) value equal to 50% or 0.5. The value of *minsup*=0.2 and the *minconf*=0.5 were adjusted manually to discover some specific and interesting rules from a large number of random rules [19]. As a result, eighty-one association rules met this threshold configuration. In order to get a realistic overview of the results, the association rules were represented in a scatter plot, as shown in Fig. 6.

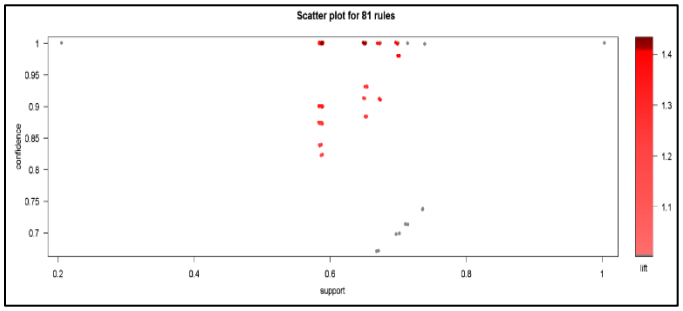

Fig. 6. The illustration of 81 Rules in Scatter plot with Minsup = 0.2 an Mincof = 0.5 for DS4.

Based on Fig. 6, support value represent x-axis and confidence value represent y-axis. For example, the first plot of association rules is located at the coordinate 0.2 for support and 1.0 for confidence. This indicate that the selected plot already meets the threshold for minimum support 0.2 and the threshold for minimum confidence 0.5. To further analyze the relationship between attributes for this association rules, the top five ruleset were selected and presented in Tables VIII, IX, and X based on three IMs; support,confidence and lift. Table VIII shows top five association rules based on support with threshold configured as *minsup* = 0.2 and *minconf* = 0.5.

TABLE VIII. TOP 5 RULES BASED ON SUPPORT MEASURE WITH MINSUP = 0.2 AND MINCONF=0.5 FOR DS4

(Attribution level: 1=Level 1, 2=Level 2, 3=Level 3, x=Did not meet the requirement to include in attribution level)

No	Ruleset	Support	Attribution Level
R1	{22} ⟹ {MY}	0.74	x
R2	{MY} ⟹ {22}	0.74	x
R3	{DE} ⟹ {MY}	0.71	x
R4	{MY} ⟹ {DE}	0.71	x
R5	{195.38.137.100} ⟹ {DE}	0.7	2

Support could measure the usefulness of association ruleset based on the frequency of itemsets occurring together in the data transaction [20], [21]. The top five rules in Table VIII summed up the combination of rules among port number 22, geolocation MY, DE and IP 195.38.137.100 which indicated that there was a strong association among these four items that frequently occured together. However, ruleset number R1 to number R4 did not meet the requirement to be included in

attribution level as the implication of antecedents and consequents did not provide meaningful information for decision making. In contrast, the R5 association rule indicated that IP 195.38.137.100 frequently appeared together in the dataset with geolocation DE and provided insight to the security analysts to deduce that the cyberattack possibly originated from this IP and country.

While support measures the usefulness of itemset that is occurring together in data transaction, confidence measure can indicate the strength of association ruleset generated whether it is reliable and valid for decision making [20].

TABLE IX. TOP 5 RULES BASED ON CONFIDENCE MEASURE WITH MINSUP = 0.2 AND MINCONF=0.5 FOR DS4

(Attribution level: 1=Level 1, 2=Level 2, 3=Level 3, x=Did not meet the requirement to include in attribution level)			
No	**Ruleset**	**Confidence**	**Attribution Level**
R1	{22} ⟹ {MY}	1	x
R2	{DE} ⟹ {MY}	1	x
R3	{195.38.137.100} ⟹ {DE}	1	2
R4	{195.38.137.100} ⟹ {MY}	1	2
R5	{195.38.137.100,DE} ⟹ {MY}	1	2

Table IX presents the top 5 most reliable rules with a threshold for *minsup* = 0.2 and *minconf* = 0.5. The top five rules based on confidence measurement showed that high confidence rules were usually related to port number 22, geolocation MY, DE and IP 195.38.137.100. This ruleset indicated that this IP was used by an attacker to launch a cyberattack and most probably originated from country DE. However, strong association rules are not always effective, some are not what users are interested in, and some are even misleading [21]. For this top five rules only ruleset number R3, R4 and R5 were reliable and were included in attribution level 2.

Support and confidence provided the information about useful rules based on occurrence and reliability of ruleset that occured in the dataset. Hence, lift measure was needed to complement these two IMs by helping to measure the importance of ruleset that suit the purpose of the research. Table X depicts the top five association rules for lift measure. Three categories were used to interpret the relationship of X / Y in lift measurement. If the lift is equal to 1, then, X and Y are independent. If the lift is higher than 1, then, X and Y are positively correlated. If the lift is lower than 1, then, X and Y are negatively correlated.

Based on Table X, the itemsets of 22, DE, MY, 195.38.137.100 and gamarue respectively had a positive correlation. Thus, this IP was malicious, being infected by gamarue and most probably originated from MY or DE. All ruleset met the requirement to be included in attribution Level 2.

B. Result of Evaluation and Validation for CTI-AR

This evaluation was to determine the capability of the proposed association ruleset for cyberattack attribution process in CTI. However, the number of association ruleset generated by using the proposed association rule mining could be massive and even tricky for domain specialists to study and summarize the meanings behind the ruleset. Moreover, it was also impractical to sift through a broad set of rules containing noise and irrelevant rules. Hence, the interestingness measure could be used for filtering or ranking association ruleset. This paper only focused on objective interestingness measure using *support*, *confidence* and *lift* to measure the meaningful and reliable association ruleset that were used to guide security analysts in making decisions. The thresholds for minimum support (*minsup*) and minimum confidence (*minconf*) were predefined manually by using trial and error method [7], [19], [22]. The summary of the association rules generated for all the datasets is depicted in Table XI using Apriori Algorithm.

Based on the association ruleset summary, the process of identifying the attributes in attribution level and classifying the ruleset into the respective attribution level group (ALGR) were conducted. Still, not all the generated ruleset met the requirement to be included in the respective ALGR because the ruleset must have at least one attribute from Level 1, Level 2 or Level 3 in both antecedents and consequents.

To further analyzed the findings of evaluation and validation for each association ruleset in Table XII, this paper summarize the ALGR and IM range for DS1 to DS12 in Table XI.

TABLE X. TOP 5 RULES BASED ON LIFT MEASURE WITH MINSUP = 0.2 AND MINCONF=0.5 FOR DS4

(Attribution level: 1=Level 1, 2=Level 2, 3=Level 3 x=Did not meet the requirement to include in attribution level)			
No	**Ruleset**	**Lift**	**Attribution Level**
R1	{22,DE} ⟹ {195.38.137.100}	1.43	2
R2	{22,DE,MY} ⟹ {195.38.137.100}	1.43	2
R3	{DE, gamarue} ⟹ {195.38.137.100}	1.43	2
R4	{22,DE,gamarue} ⟹ {195.38.137.100}	1.43	2
R5	{DE,gamarue,MY} ⟹ {195.38.137.100}	1.43	2

TABLE XI. SUMMARY OF ASSOCIATION RULESET

Dataset	Number of ruleset	Level 1	Level 2	Level 3	N/A
DS1	75	5	40	0	30
DS2	37	0	37	0	0
DS3	4	0	0	4	0
DS4	81	7	40	0	34
DS5	50	45	0	0	5
DS6	12	0	0	12	0
DS7	76	7	40	0	29
DS8	91	89	0	0	2
DS9	14	0	0	14	0
DS10	64	5	31	0	28
DS11	86	84	0	0	2
DS12	17	0	0	17	0

TABLE XII. SUMMARY OF ALGR AND IM RANGE FOR DS1-DS12

(√ = ALGR exist, x = ALGR does not exist) (support=*s*, confidence=*c*, lift=*l*, Attribution Level Group=ALGR)								
Dataset	**ALGR**			***minsup threshold***	***minconf threshold***	**Range for *IM***		
	1	2	3			***s***	***c***	***l***
DS1	√	√	x	0.2	0.5	≥ 0.28	≥ 0.52	≥ 1
DS2	√	x	x			≥ 0.27	≥ 0.52	≥ 1
DS3	x	x	√	0.05		≥ 0.06	≥ 0.53	≥ 2.06
DS4	√	√	x	0.2		≥ 0.21	≥ 0.67	≥ 1
DS5	√	x	x			≥ 0.4	≥ 0.5	≥ 1
DS6	x	x	√	0.05		≥ 0.07	≥ 0.5	≥ 1.84
DS7	√	√	x	0.2		≥ 0.24	≥ 0.52	≥ 0.83
DS8	√	x	x			≥ 0.21	≥ 0.75	≥ 0.96
DS9	x	x	√	0.05		≥ 0.04	≥ 0.5	≥ 1.11
DS10	√	√	x	0.2		≥ 0.37	≥ 0.5	≥ 1
DS11	√	x	x			≥ 0.23	≥ 0.52	≥ 1
DS12	x	x	√	0.05		≥ 0.07	≥ 0.5	≥ 2.86

Table XII shows the range of IM capture from the strongest association ruleset that was generated using the general Equation (1), the threshold used to generate the ruleset and ALGR found in DS4 up to DS12. The value of range for *support*, *confidence* and *lift* in Table XII was used to validate and verify the strong association ruleset to be included in ALGR. Support could measure the usefulness of association ruleset based on the frequency of itemset occured together in the data transaction. Confidence indicated the strength of association ruleset generated whether it was reliable and valid for decision making. At the same time, lift measure was needed to complement these two IMs by helping to measure the importance of ruleset that suit the purpose of the research, whereby to perform cyberattack attribution process in CTI. Once the list of strong association ruleset was identified and met the threshold for *minsup* and *minconf*, this list of association ruleset was included in the respective ALGR based on the presence of the attributes in each association ruleset. The steps to classify the association ruleset into ALGR are explained in the following subsection.

Table XII showed that the ruleset found in this research was effective in performing the cyberattack attribution because it could identify all ALGRs where each ALGR is mapped to different CTI type as discussed in Table IV and Table V. This CTI type was used by an organization for a specific purpose to prevent from cyberattack. For example, ALGR1 and ALGR2 were mapped to tactical intelligence subtype, hence, the outputs from these ALGRs could help an organization to deal with threat indicators and prioritize vulnerabilities patches quickly and accurately. Then, ALGR3 was mapped to operational intelligence and the output from ALGR3 could improve the detection rate and prevent future incidents as attacks could be seen in a clear context. The outputs from ALGR1, ALGR2 and ALGR3 were mapped to strategic intelligence to drive the organization decision making regarding security countermeasure and areas of improvement from the insights of current attack trends and financial impact to the organization.

The results of the evaluation and validation from the experimental approach are presented in Table XIII. Table XIII illustrates the top 5 association rulesets results from each Interestingness Measure (IM) based on support, confidence and lift measure that filtered and ranked to their respective ALGR. The ALGR grouping could provide hidden information behind the rulesets about attribution level that could help security analysts to perform cyberattack attribution process in CTI.

The association ruleset in Table XIII showed how attributes of LHS implied the attributes of RHS. For example, a ruleset {195.38.137.100,gamarue} ⟹ {22} indicated that an IP address 195.38.137.100 was infected by gamarue and had been used by an attacker to launch an attack using port 22. This ruleset provided the relationship between attribute and guidance to security analysts on the function of the attribute in the cyberattack. This knowledge can help security analysts to plan a mitigation action.

Table XIII also showed how association ruleset were divided into specific ALGR through IM. The grouping of association ruleset into ALGR was based on an attribute that was available in the particular ruleset. Table IV describes the details of attribute in each attribution level. The attributes description for attribution level in Table IV was used as a reference for distinguishing the presence of the attribute from a specific attribution level in each association ruleset. The attribute identification in ruleset could help security analysts to verify what type of attribution achieved from each ruleset. For example, a set of association ruleset in row number four from Table XIII was measured through confidence to prove the reliability of association ruleset provided the information about attribution on IP address, malware type, hash value and port number. The list of attribute found using confidence could be used by a security analyst for further investigation as it is valid and reliable.

Besides, Table XIII also summarized the list of association ruleset into respective ALGR. The classification of ruleset into ALGR was done based on discussion in Table IV and Table V. For example, ruleset classification to ALGR1 was based on the existence of the attribute from Level 1 in the ruleset. This attribute comprised IP address, hash value, malware type, domain name or URLs in the LHS or RHS of the ruleset. As for ALGR2, it required the occurrence of an attribute from attribution Level 1 and Level 2. Geolocation was an attribute of attribution Level 2.

In contrast, the classification of ALGR3 must have attribute from attribution Level 1, 2 and 3 occurred in the ruleset. However, there was also an exception in determining ALGR3, where TTPs alone was sufficient in determining the ruleset as part of ALGR3. It is because TTPs could provide the context to the association ruleset throughout the technique, tactic and procedure used by an attacker to launch the cyberattack.

The results from Table XIII indicated that the formulation of association ruleset from the proposed CTI-AR could help security analysts in making a decision about cyberattack

attribution and the details of the validation result are characterized in Table XIV.

TABLE XIII. RESULTS OF INTERESTINGNESS MEASURE (IM) FOR DS4-DS12

No	Interestingness Measure *(IM)*	ALGR	Association Ruleset	Attribution achieved
1	**Support**	1	{195.38.137.100} ⟹ {22} {22} ⟹ {195.38.137.100} {*gamarue*} ⟹ {195.38.137.100} {195.38.137.100} ⟹ {*gamarue*} {195.38.137.100, *gamarue*} ⟹ {22}	IP address, malware type and port number were found
2	**Support**	2	{195.38.137.100} ⟹ {*DE*} {*DE*} ⟹ {195.38.137.100} {195.38.137.100} ⟹ {*MY*} {*MY*} ⟹ {195.38.137.100} {195.38.137.100, *DE*} ⟹ {*MY*}	IP address and geolocation were found
3	**Support**	3	{*Cloud Service Dashboard*} ⟹ {*discovery*} {*discovery*} ⟹ {*Cloud Service Dashboard*} {*Cloud Service Discovery*} ⟹ {*discovery*} {*discovery*} ⟹ {*Cloud Service Discovery*}	Technique and tactic were fond
4	**Confidence**	1	{210.48.151.111} ⟹ {445} {7867*de*13*bf*22*a*7*f*3*e*35590440 ⟹ {*Backdoor. Agent. rke*} {*Backdoor. Agent. rke*} ⟹ {7867*de*13*bf*22*a*7*f*3*e*35590 {7867*de*13*bf*22*a*7*f*3*e*35590440 ⟹ {210.48.151.111} {7867*de*13*bf*22*a*7*f*3*e*35590440 ⟹ {445}	IP address, malware type, hash value and port number were found
5	**Confidence**	2	{195.38.137.100} ⟹ {*DE*} {195.38.137.100} ⟹ {*MY*} {195.38.137.100, *DE*} ⟹ {*MY*} {195.38.137.100, *MY*} ⟹ {*DE*} {195.38.137.100,22} ⟹ {*DE*}	IP address and geolocation were found
6	**Confidence**	3	{*Cloud Service Dashboard*} ⟹ {*discovery*} {*Cloud Service Discovery*} ⟹ {*discovery*} {*defense − evasion*} ⟹ {*Application Access Token*} {*Elevated Execution with Promp* ⟹ {*privilege − escalation*}	Technique and tactic were found
7	**Lift**	1	{*Troj. Spy. Xxp*! *c*} ⟹ {786*ab*616239814616642*ba* {786*ab*616239814616642*ba*443 ⟹ {*Troj. Spy. Xxp*! *c*} {210.48.151.111, *Troj. Spy. Xxp*! *c* ⟹ {786*ab*616239814616642*ba* {210.48.151.111,786*ab*61 6239814616642*ba*4438*df*78*a*9} ⟹ {*Troj. Spy. Xxp*! *c*} {445, *Troj. Spy. Xxp*! *c*} ⟹ {786*ab*616239814616642*ba*-	IP address, malware type, hash value and port number were found
8	**Lift**	2	{22, *DE*} ⟹ {195.38.137.100} {22, *DE*, *MY*} ⟹ {195.38.137.100} {*DE*, *gamarue*} ⟹ {195.38.137.100} {22, *DE*, *gamarue*} ⟹ {195.38.137.100} {*DE*, *gamarue*, *MY*} ⟹ {195.38.137.100}	IP address, malware type, port number and geolocation were found
9	**Lift**	3	{*Elevated Execution with Promp* ⟹ {*privilege − escalation*} {*privilege − escalation*} ⟹ {*Elevated Execution with Pr* {*Data from Cloud Storage Obje* ⟹ {*collection*} {*collection*} ⟹ {*Data from Cloud Storage C* {*defense − evasion*} ⟹ {*Application Access Token*}	Technique and tactic were found

TABLE XIV. CHARACTERIZATION OF THE EXPERIMENTAL VALIDATION RESULT

Criteria	Characteristic
Cyberattack attribution analysis	Capable of identifying the relationship of attributes Capable of identifying the attributes in attribution level Capable of identifying the threat attribution level

Therefore, using the characteristics shown in Table XIV, the CTI-AR was validated, as summarized in Table XV.

Table XV indicates the proposed CTI-AR which comprised all characteristics. The proposed CTI-AR was capable of generating the association ruleset from the frequent itemset process, identifying the relationship of attributes among the association ruleset, identifying the threat attribution level for each association ruleset and the attributes in attribution level. Based on the association ruleset and attribution level, the proposed CTI-AR was capable in performing cyberattack attribution process in CTI. These findings were then compared to the findings from the association rule mining (ARM) in existing CTI framework to validate the proposed CTI-AR as discussed in Table XVI.

Table XVI shows the comparison between the association rule mining in existing CTI framework and the proposed CTI-AR in CTI. Based on the characteristics, the ARM in the existing CTI framework is able to identify the attribution level but unable to classify and identify the complete list of attributes that belong to the attribution level. In contrast, the proposed CTI-AR in CTI is more capable in performing the attribution of cyberattacks not only by finding the relationship between the attribute but also providing additional information on the attribution level and attributes at the attribution level.

TABLE XV. SUMMARY OF RESULT VALIDATION OF THE PROPOSED CTI-AR IN CTI

(√ = characteristic exist, x = characteristic does not exist)	
Characteristic	**Proposed CTI-AR in CTI**
Capable of identifying the relationship of attributes	√
Capable of identifying the attributes in attribution level	√
Capable of identifying the threat attribution level	√

TABLE XVI. COMPARATIVE ANALYSIS WITH EXISTING ARM IN CTI

(√ = characteristic exist, x = characteristic does not exist)		
Characteristics	**ARM in existing CTI framework**	**Proposed CTI-AR in CTI**
Capable of identifying the relationship of attributes	√	√
Capable of identifying the attributes in attribution level	x	√
Capable of identifying the threat attribution level	√	√

V. CONCLUSIONS

This paper introduce an approach to overcome voluminous data issue in CTI for cyber attack attribution. The approach consist of data preprocessing, frequent itemset identification and ruleset generation that was used to formulate an association ruleset name as Cyber Threat Intelligence Association Ruleset (CTI-AR). This CTI-AR is used to assist security analyst in discovering the hidden knowledge behind the voluminous data to produce an effective cyberattack attribution in CTI. The results obtained in the experiment demonstrates the CTI-AR is able to discover the hidden knowledge behind the voluminous data in CTI that can help security analyst in performing cyber attack attribution effectively.These abilities are demonstrated through the result obtained using three Interestingness Mesures indicators: *support* (*s*), *confidence* (*c*) and *lift* (*l*). *Support* (*s*) reflected the practicality or usefulness of association rules, *Confidence* (*c*) reflected the validity or reliability of association rules and *Lift* (*l*) was to complement previous two evaluation indicators by filtering and removing wrong and meaningless ruleset. Based on the result from Interestingness Measures indicators, CTI-AR can effectively help security analyst identify the attributes, relationship between attributes and attribution level group of cyberattack in CTI. This research has a high potential of being expanded into cyber threat hunting process in providing a more proactive cybersecurity environment. For future work, more association rule algorithm and other statistical measures can be implemented to improve association ruleset effectiveness and accuracy in performing cyber attack attribution.

ACKNOWLEDGMENT

This study was kindly supported by The Ministry of Communications and Multimedia (KKMM), Cybersecurity Malaysia and Universiti Teknikal Malaysia Melaka (UTeM).

REFERENCES

[1] C. Sauerwein et al., "Threat Intelligence Sharing Platforms : An Exploratory Study of Software Vendors and Research Perspectives," pp. 837–851, 2017.

[2] S. Qamar, Z. Anwar, M. A. Rahman, E. Al-Shaer, and B.-T. Chu, "Data-driven analytics for cyber-threat intelligence and information sharing," Comput. Secur., vol. 67, pp. 35–58, 2017.

[3] E. C. L. L. W. E. Karafili, "An Argumentation-Based Approach to Assist in the Investigation and Attribution of Cyber-Attacks," arXiv Comput. Sci., 2019.

[4] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules (expanded version). Research Report IBM RJ 9839," Proc. 20th Intl. Conf. VLDB, pp. 487--499, 1994.

[5] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," no. December, pp. 207–216, 1993.

[6] P. Prithiviraj and R. Porkodi, "A Comparative Analysis of Association Rule Mining Algorithms in Data Mining: A Study," Open J. Comput. Sci. Eng. Surv., vol. 3, no. 1, pp. 98–119, 2015.

[7] Y. Liu, K. Yu, X. Wu, Y. Shi, and Y. Tan, "Association rules mining analysis of app usage based on mobile traffic flow data," in 2018 IEEE 3rd International Conference on Big Data Analysis, ICBDA 2018, 2018, pp. 55–60.

[8] A. Khalili and A. Sami, "SysDetect: A systematic approach to critical state determination for Industrial Intrusion Detection Systems using Apriori algorithm," J. Process Control, vol. 32, no. April 2018, pp. 154–160, 2015.

[9] H.-W. Hsiao, H.-M. Sun, and W.-C. Fan, "Detecting stepping-stone intrusion using association rule mining," Secur. Commun. Networks, vol. 6, no. March, pp. 1225–1235, Mar. 2013.

[10] O. S. Adebayo and N. Abdul Aziz, "Improved Malware Detection Model with Apriori Association Rule and Particle Swarm Optimization," Secur. Commun. Networks, vol. 2019, pp. 1–13, Aug. 2019.

[11] J. C. Kim and K. Chung, "Sequential-index pattern mining for lifecare telecommunication platform," Cluster Comput., vol. 22, no. 4, pp. 1039–1048, 2019.

[12] J. Hong, R. Tamakloe, and D. Park, "Application of association rules mining algorithm for hazardous materials transportation crashes on expressway," Accid. Anal. Prev., vol. 142, no. 3, pp. 105–497, 2020.

[13] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, and R. Beyah, "Acing the IOC Game : Toward Automatic Discovery and Analysis of Open-Source Cyber Threat Intelligence," in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 755–766.

[14] Z. Zhu and T. Dumitras, "FeatureSmith: Automatically Engineering Features for Malware Detection by Mining the Security Literature," in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16, 2016, pp. 767–778.

[15] C. Sabottke, O. Suciu, T. Dumitraș, C. Sabottke, and T. Dumitras, "Vulnerability Disclosure in the Age of Social Media: Exploiting Twitter for Predicting Real-World Exploits," Proc. 24th USENIX Secur. Symp., 2015.

[16] J. R. Scanlon and M. S. Gerber, "Automatic detection of cyber-recruitment by violent extremists," Secur. Inform., vol. 3, no. 1, pp. 1–10, 2014.

[17] Jawwad A. Shamsi, S. Zeadally, F. Sheikh, and A. Flowers, "Attribution in cyberspace: techniques and legal implications," Secur. Commun. NETWORKS, 2016.

[18] D. S. S. Mrs. M.Kavitha, "Association Rule Mining using Apriori Algorithm for Extracting Product Sales Patterns in Groceries," Int. J. Eng. Res. Technol., vol. 8, no. 3, pp. 5–8, 2020.

[19] S. Mahmood, M. Shahbaz, and A. Guergachi, "Negative and positive association rules mining from text using frequent and infrequent itemsets," Sci. World J., vol. 2014, 2014.

[20] X. Niu and X. Ji, "Evaluation methods for association rules in spatial knowlegde base," ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci., vol. II–4, no. May, pp. 53–58, 2014.

[21] C. Ju, F. Bao, C. Xu, and X. Fu, "A Novel Method of Interestingness Measures for Association Rules Mining Based on Profit," Discret. Dyn. Nat. Soc., vol. 2015, no. 2, pp. 1–10, 2015.

[22] L. Yan, Y. Ke, and W. Xiaofei, "Association Analysis Based on Mobile Traffic," in 2014 4th IEEE International Conference on Network Infrastructure and Digital Content, 2014.