‘Data Poisoning’ – Achilles Heel of Cyber Threat Intelligence Systems

Thabo Mahlangu, Sinethemba January, Thulani C. Mashiane, Moses T. Dlamini, Sipho J. Ngobeni, Nkqubela L. Ruxwana
Council of Scientific and Industrial Research, Defence Peace Safety and Security, CCIW, Pretoria, South Africa.
{TMahlangu3, SJanuary, SNgobeni, CMashiane, TDlamini1,LRuxwana}@csir.co.za;

“Know yourself, the terrain of the battlefield and the enemy, you need not fear the result of a hundred battles”
Sun Tzu.

Abstract: In the cyberspace, system defenders might have an idea of their own cybersecurity defense systems, but they surely have a partial view of the cyberspace battlefield and almost zero knowledge of the attackers. Evidently, the arm’s race between defenders and attackers favors the attackers. The rise of fake news and ‘data poisoning’ attacks aimed at machine learning inspired cyber threat intelligence systems is the result of a new strategy adopted by attackers that adds complexity to an already complex and ever changing cyber threat landscape. The modus operandi and TTPs of attackers continue to change with increasing repercussions. Attackers are now exploiting a vulnerability in the data training process of AI and ML inspired cyber threat intelligence systems by injecting ‘poisoned data’ in training datasets to allow their malicious code to evade detection. The ‘poisoned’ corpus is specifically tailored and targeted to AI and ML cyber threat intelligence defense systems, especially those based on supervised and semi-supervised learning algorithms to make them misclassify malicious code as legitimate data.

This paper deals with the ‘data poisoning’ problem on different fronts. It starts by ensuring the completeness and standardization of the input datasets which, is vital to make accurate data-driven threat intelligence decisions. The input data itself is validated by using a mix of related indicators to determine its reliability. Based on the validation of input data sources, the authors make an assumption that the corpus is trustworthy and then add a security feature that prevents ‘data poisoning’ attacks. Based on these features, our model can be argued to provide a plausible solution to the ‘data poisoning’ problem of AI and ML inspired cyber threat intelligence systems. Our solution is based on working with trusted sources of input raw data. The dynamics of our solution changes completely if the input raw data comes with ‘poisoned data’ that mimic trusted data. This is one area that our future research will focus on.

Keywords: cyber threat intelligence, data poisoning, artificial intelligence, machine learning

1. Introduction

The arm’s race between cyberspace attackers and defender continues. Attackers’ tools, tactics and procedures (TTPs) evolves so quickly that cyber defence, legislation and law enforcement lag behind. On the one hand, new technology developments like cloud computing, social media, big data, machine learning, internet of things and others are continually disrupting existing business models in a global scale. Hence, there is a mad rush to adopt new business models that open up new risks. It is no coincidence that today’s cyber threat landscape reflects that the attackers are gaining more grounds than the defenders. For example, attackers are very quick to adopt the latest technologies like artificial intelligence (AI) and machine learning (ML) to detect and exploit defence systems’ vulnerabilities and, evade detection (Fortinet, 2018). Giles (2018) asserts that attackers are using AI and ML to analyse features of cyber threat intelligence systems on how they flag malware. They then remove or conceal through encryption the code snippet from their malware that could raise the red flags so that the classification algorithms cannot catch it.

The gap between attackers and defenders seems to be widening even more. This is no coincidence as today’s cyber attackers are well funded and organized; they have vast resources at their disposal; operates in a well-structured, coordinated and highly incentivized underground economy (Anstee, 2017; Dara, Zarga and Muralidhara, 2018). Today’s cyber attackers are patient and do their nefarious deeds with sophistication and targeting vulnerabilities in people, process and technology right across the globe with no respect for national boundaries. Attackers are now deploying advanced malware that leverages on cutting edge technology to not only circumvent advanced security defences but also to widen their scope and scale of their attacks (Fortinet, 2018). A glimpse to the near future could see attackers using autonomous and self-learning malware with catastrophic implications.

The anonymity or plausible deniability of cyber threats adds to the already complex threat landscape (Lundbohm, 2017). Hence, some experts argue that ballistic missiles have return addresses; yet cyber threats often emanate behind a veil of Internet anonymity that hides details of the attackers (Geer et al. 2014). This anonymity is one of the biggest challenges of deterring any defence mechanism against or retaliating to cyber threats. Hence, the difficulty to determine who exactly is behind today's cyber threats. This considerably challenges the field of cyberspace and has raised the intractable issue of cybersecurity attribution. It must be noted that not knowing the enemy is one way to lose a battle (Lundbohm, 2017). Geers (2011) projects Tzu's theory onto the cyberspace battlefield and asserts that defenders who know their defence systems, the terrain of the cyberspace battlefield and cybercriminals and their modus operandi have no reason to fear the result of a hundred cyberspace battles. However, it is also argued therein that defenders that know their defence systems and the cyberspace battlefield's terrain but not their enemy, for every victory gained, they also suffer a defeat. This means that all efforts to understand the battlefield and own defence systems cannot guarantee victory without an effort to understand the tactics of the adversary. It is also argued that defenders who know neither their enemy nor their defence systems nor the terrain of the battlefield, will always succumb in every battle that they engage in (Geer, 2011). Today's threat landscape reflects that cybersecurity defenders might know something about their defence systems, but they have a partial view of the cyberspace battlefield and have almost zero knowledge of the attackers and their modus operandi.

Furthermore, a reactive approach to defending systems also adds to the already complex threat landscape. Most organisations only act after a breach has already occurred. In the ever changing cyber threat landscape of rapid zero-days, advanced persistent threats (APTs), botnets, ransomware and state-sponsored espionage activities; a secure company today would be vulnerable by tomorrow. For example, a secured US election before the 2016 November became untrustworthy in January 2017. Hence, and as demonstrated in the recent US elections; a reactive approach to defence is totally insufficient to address today's ever changing threats of fake news and 'data poisoning' (Steinhardt, Koh and Liang, 2017).

This paper deals with the challenging issue of 'data poisoning' in cyber threat intelligence systems. Given the increasing use of predictive cybersecurity analytics and cyber threat intelligence platforms which gives system defenders a capability to somehow anticipate signatures of new malware this is inevitable. Attackers have since realised that new malware gets detected at first appearance by AI and ML inspired malware classifiers and detectors in current cyber threat intelligence systems. Hence, instead of concentrating their efforts on developing new malware, they are now investing their resources into finding ways to breach AI and ML inspired cyber threat intelligence defences.

Therefore, the main research question of this paper is, how can we solve the 'data poisoning' issue to fully leverage on a cyber threat intelligence to defend systems and respond better to the sophisticated and ever changing cyber threats? This paper argues that there is a huge need for proactive defence efforts that make use of cyber threat intelligence systems. This is to help organizations build a better situational awareness, recommend resilient cyber security controls, and learn from breaches in order to adapt and re-shape existing controls to improve cyber threat detection and system resilience. Most existing cyber threat intelligence (CTI) systems are leveraging on the current data-driven economy to collect and collate massive cyber threat data from different source feeds. Attackers on the other hand have since realized an opportunity to 'poison' cyber threat data sources to try and circumvent detection systems. 'Data poisoning' attacks leverages on the many data sources with massive corpus which make it almost impossible to validate and curate the data (Paudice et al., 2018). Hence, it is important that our solution make plausible means to validate, curate and secure input data to prevent 'data poisoning'.

The rest of the paper is structured as follows: section two delves into the related work. This is to elucidate on existing literature and show exactly how our work goes beyond the current-state-of-the-art to provide a solid contribution. Section three presents and discusses the proposed cyber threat intelligence and exchange platform (CTIEP) data pipelining model. This is presented with a specific focus on the collection, cleansing, validation, security and presentation of the collected cyber threat data. Section four discusses some of the preliminary results of our model. Section five concludes the paper and provides some information on future work.

2. Related work

Cyber threat intelligence systems are critical for organizations to achieve a strong security posture (Ponemon Institute, 2017; James, 2018). A cyber threat intelligence system is an area of cybersecurity that focuses on the collection, processing and analysis of information about potential attacks that threatens the safety of organizational information assets. A robust cyber threat intelligence system can provide highly technical metrics, countermeasures and corrective actions (Ernst &Young Global Limited, 2016). This can help organizations to get ahead of cyber criminals.

Hence, there is a rush to gather real-time cyber threat data (Tounsi and Rais, 2017). The goal is to use AI and ML algorithms in cyber threat intelligence platforms to transform threat data into actionable intelligence to help thwart cyber-attacks. This has also caused attackers to focus on targeting data-driven cyber threat intelligence systems. This section is divided into three: section 2.1 discusses related work with regards to AI and ML inspired cyber threat intelligence and section 2.2 reviews literature of data poisoning threats that target AI and ML inspired cyber threat intelligence. Section 2.3 identifies some research gaps.

2.1 AI and ML inspired Cyber Threat Intelligence

Giles (218) argues that a number of companies are now using AI and ML in their cyber threat intelligence systems to help automate and improve threat detection. However, most of the current solutions equipped with AI and ML algorithms somehow create a false sense of security (Giles, 2018). For example, some of the existing solutions are based on supervised learning algorithms (van der Walt, Eloff and Grobler, 2018) that classify code as either clean (trustworthy) or malicious (deceptive) based upon known signatures or patterns. Therefore, all that it takes for an attacker to foil such systems is to access the training data and tag malware as clean code. Such a scenario is possible, mainly because existing systems are concerned about the analysis of the data and have turned a blind eye to its protection. Giles (2018) also argues that at times attackers do not even have to corrupt the training data sets, but alter the underlying algorithms that process the data. Giles (2018) argues that AI and ML might be a new in cybersecurity, yet on the contrary it could be a dangerous weapon for attackers.

Knight (2017) argues that some existing threat intelligence systems are built based on ‘black box’ type deep learning algorithms. This is to try and solve the problems of supervised learning algorithms. However, with deep learning algorithms and the multiple layers in a neural network, it means that there is no way to know how system’ algorithms do their classification of threats in the different levels of abstraction. This is a big cause for concern because it is not completely clear how the algorithms make their decisions. Hence, Knight (2017) argues that ‘black box’ deep learning algorithms must be understandable to their creators and accountable to their users.

Treit, Stewart and Parikh (2018) investigated attackers attempting to defeat AI and ML algorithms in the next-generation anti-virus Windows Defender. Treit et al. (2018) leverage on the power of AI and ML to enable Windows Defender Advanced Threat Protection by using a nested and layered ML approach to stop the next generation type of malware from avoiding detection. The layered approach ensures that even if malware can circumvent one layer, it can still be detected by the other layers. Within each layer, there is a number of individual ML models trained to recognize new and emerging threats. Treit et al. (2018) argues that their solution provides a robust multi-faceted view of new generation of threats that no single algorithm can achieve. The solution therein also uses stacked ensemble models that take predictions from the base classifiers of ML models and combine them to create even stronger malware predictions.

Dara et al. (2018) argues that current cyber threat intelligence services violate and compromise the privacy of users. This work claims that existing privacy preserving approaches use anonymization of data. However, such approaches are prone to suffer from inferential attacks. An inferential attack analyses data to illegitimately gain confidential information about a subject in an unauthorized manner. Dara et al. (2018) then proposes an architecture for privacy preserving threat intelligence. This is based on private information retrieval using keywords and similar document retrieval that searches and retrieves data without revealing any private information about the user or document. At a glance, this looks like a credible solution. However, the proposed solution therein (Dara et al., 2018) defeats accountability and further complicates an already complex issue of cyber attribution.

Chiba et al. (2018) proposed a solution called DomainChroma - a solution based on chromatography that is used to separate a mixture into its individual components. Chiba et al. (2018) considers a pool of malicious domain names which may be involved in drive-by downloads, malware download, phishing, key-logger command and control types of cyberattacks; and uses their DomainChroma to classify them according to their characteristics in order to build actionable threat intelligence. However, this work does not make any mention of how they validate the data or the source thereof. Moreover, there is no mention of how the data is protected. Validation of sources and their data is key to make insightful decisions. Data protection is required to ensure that malicious entities cannot tamper with the training and test data.

2.2. ‘Data Poisoning’ in data driven Cyber Threat Intelligence

Van der Walt et al. (2018) make use on ML algorithms to detect deceptive identities on social media platforms. Although, this work is not based on cyber threat intelligence systems, it somehow makes a plausible effort to address one of the most intractable problems in cyber security i.e. accurate cyber attribution. However, it also does not tackle the serious issue of ‘data poisoning’.

Biggio et al. (2013) argues that clustering algorithms that are now being used in AI and ML inspired cyber threat intelligence for solving cybersecurity threats were not designed to deal with deliberate attacks that attempt to subvert the clustering process. Contrary to the above covered literature, this work (Biggio et al., 2013) demonstrates that it is indeed possible for an attacker to significantly poison the clustering process by adding a small percentage of attack samples to the input data. Furthermore, Baggio et al. (2013) also argues that an attacker may use obfuscated attack samples in existing clusters to prevent clustering algorithms from detecting their malicious code. This has raised serious concerns around the integrity of AI and ML datasets and clustering algorithms thereof. Defending against such poisoning attacks is challenging (Jagielski et al., 2018). Biggio et al. (2013) is focused on raising the concerns not necessarily on the solution. However, this work asserts that cyber threat intelligence systems must adopt cybersecurity countermeasure that embrace the secure by design principle to effectively thwart 'data poisoning' attacks (Biggio et al., 2013).

Therefore, there is a growing pool of research that attempts to address the issue of 'data poisoning' attacks in cyber threat intelligence systems (Jagielski et al., 2018; Khurana, Mittal and Joshi, 2018; Paudice et al., 2018; Roli, Biggio and Fumera, 2016). This is an indication that research in this area is intensifying. Jagielski et al. (2018) recognize the difficulty of identifying and separating 'poisoned data' samples from the legitimate corpus. Instead they choose to model their solution based on a small sample of training data, choosing only the points that are 'close' to legitimate points. Moreover, this work also made some interesting discoveries on the potential impact of 'data poisoning' in healthcare applications. For example, this study shows the effects that a small sample of 'poisoned data' points can have on the dosage of patients. The results therein show that patient dosage can increase to an estimate of 350% (Jagielski et al., 2018). This basically mean that if patient health data is 'poisoned', patients would be made to take more than required drugs. This can potentially cause drug overdose and result in deaths.

Khurana et al. (2018) proposes a reputation based system. The proposed system scores the input data based on a number of factors that determines the credibility of the source. This is also based on supervised learning algorithms which basically mean that is suffers the same fate as other supervised learning algorithms, i.e. garbage-in, garbage-out. An attack to the clustering algorithms would reverse the results to make sources that are classified 'not credible' to be credible. However, the reputation scoring is one element that the current paper would also want to leverage on.

Rubinstein et al. (2009) is one of the earliest works that use anomaly detectors to detect 'poisoned data' from legitimate corpus. This work shows at least three ways that an attacker could use to evade detection by carefully adding moderate 'poisoned data' at different intervals. Though, this approach is argued not to be worth it therein, but it performs well to throw off false-positives and negative-positive balance and hamper the efficacy of the detector (Rubinstein et al., 2009). The ANTIDOTE solution therein (Rubinstein et al. 2009) was designed to prevent 'data poisoning' attacks from shifting false-positives and false-negatives. Hence, the solution is argued to reject contaminated data. Much like all the other supervised learning systems, this too hinges on the correctness of the training data. Moreover, the inherent assumption in Rubinstein et al.'s work is that the legitimate corpus does not have contaminated data.

Paudice et al., (2018) focuses on detecting 'poisoned data' samples using an anomaly detection system. The assumption made by Paudice et al., (2018) is that attackers' 'poisoned data' points are quite different from the legitimate corpus. The claim therein is that this makes an attacker's 'poisoned data' points easy to detect as outliers using distance-based anomaly detection systems (Paudice et al., 2018). Similar to the work of Jagielski et al. (2018), Paudice et al. also use a small fraction of trusted data points to train their model instead of the entire corpus. It is important to note that Paudice et al.' and Jagielski et al.'s methods work best if the chosen small training data sample is trustworthy. Should the contrary be true, both propositions fail dismally and can even result in 'poisoned' outlier detectors. This would basically means that they will both work in reverse i.e. flagging malicious data as legitimate and legitimate data as malicious.

2.3 Summary of related work and gaps thereof

The covered related work reflects on the importance of AI and ML inspired cyber threat intelligence systems. However, the move to such promising systems has been hampered by the increasing threat of data poisoning which targets the training data of AL and ML clustering and classification algorithms with serious repercussions. For example, the covered literature shows that if unchecked data poisoning attacks could actually make models to classify malicious data as legitimate and legitimate as malicious. This is more so for supervised and semi-supervised classifiers. The 'black-box' approach of unsupervised classifiers raises trust issues because it is not so clear how the algorithms make their decisions in classifying data points. Hence, the rise of research that attempts to solve the serious 'data poisoning' issues of AI and ML based cyber threat intelligence systems. The covered literature has looked at a number of solutions such as training with a small subset of the entire corpus, using reputation-based classifiers etc. The former has huge implications if the selected data comprises mostly of 'poisoned data' which stands to poison even the data classifiers. The covered literature has also alluded to the difficulty of identifying 'poisoned data' from legitimate data. Most of the work seems to be concerned about the input data as it comes from the disparate sources. However, none of this work has touched on solutions that ensure the security of the corpus after it has been validated, cleansed and collated. This is one area that the current paper tackles to ensure that 'poisoned data' cannot be injected into the cleansed corpus. Furthermore, the covered literature does not reflect on ways of ensuring that it bases its detection on complete corpus. It seems that most of the covered literature uses just a sample of the data which raises questions around its representativeness to make generalized decisions.

The next section extends some of the related work section to propose a solution that would cover the issues raised.

3. The proposed model

The authors propose a CTIEP data pipelining model. This model is an attempt to prevent attackers from exploiting the vulnerability of 'data poisoning' in AI and ML inspired cyber threat intelligence system. This is an attempt to provide security by design approach that was first proposed in Biggio et al. (2013) and Roli, Biggio and Fumera (2016). The input raw data is received from multiple Internet sources that are indeed vulnerable to the 'data poisoning' attacks. In this paper the authors have decided to focus on addressing the cleansing, validation and normalization of the input raw data. This goes through a number of processes as depicted in figure 1.

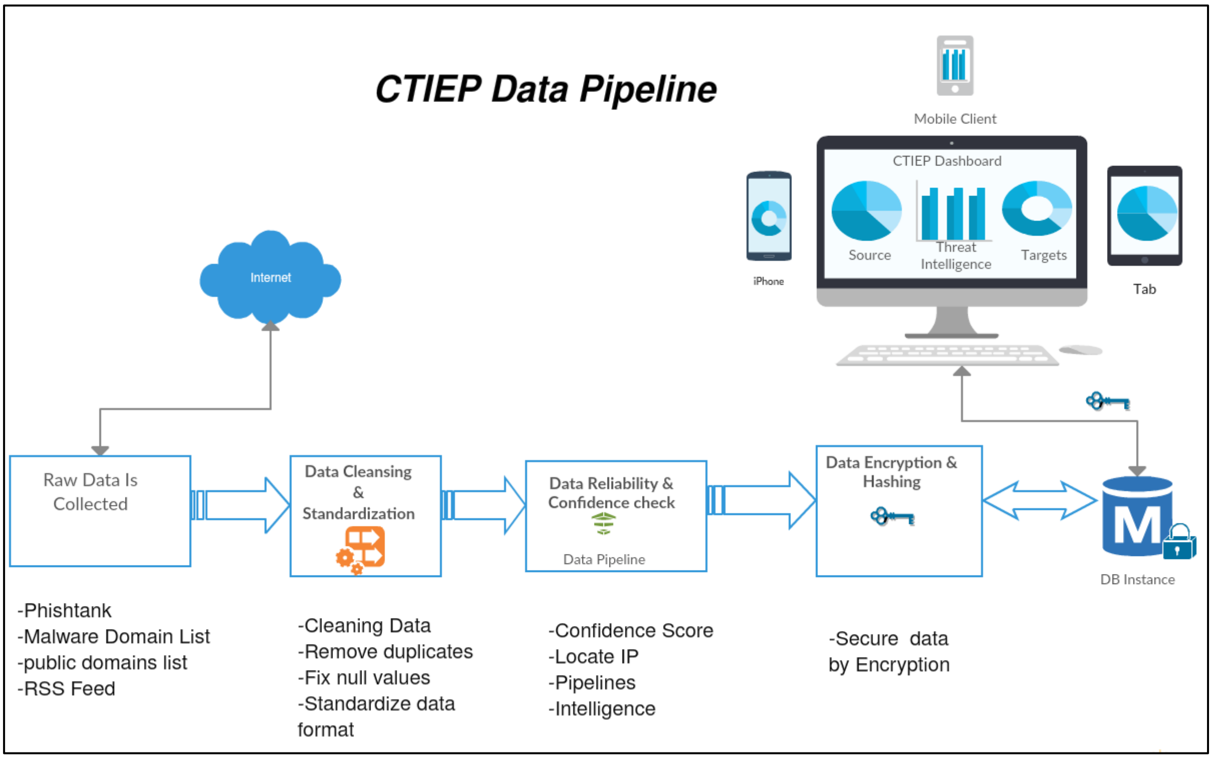


Figure 1: Cyber Threat Intelligence and Exchange Data Pipelining

3.1 Data Collection

Cyber threat data is collected from multiple vulnerable Internet sources in data objects. The authors used data objects to ensure that we collect all there is for completeness. Data objects also make it easy for our system to store the data in object oriented database. We have used sources like PhishTank, malware domain list, and other public domain lists. Though these sources may be trustworthy, they are definitely susceptible to the 'data poisoning' attacks. Currently for most of the data sources that are out there have no way to ascertain the integrity of the data. Yet, in the data-driven economy data is the life-blood of organizations that crunch it to produce allegedly critical business decisions. Hence, it is important to do a thorough exercise in cleansing the data to ensure that it does not contain contaminated data. However, as previous studies have shown, the exercise of identifying legitimate data from a contaminated data set is difficult. This paper also makes the assumption that the source data is not clean and requires careful cleansing before it can be accepted into our cyber threat intelligence system.

3.2 Data Cleansing and Standardization

Data cleansing of the corpus includes removing duplicate entries and null values. This exercise does not compromise the completeness and/or integrity of the data as obtained from the original source. However, the removal of duplicate entries ensures that there is only one record for each and every entry. This also helps in reducing the size of the corpus, more especially for storage purposes. The data also goes through a standardization process to ensure that all fields that are similar are categorised so. For example, synonyms are taken care of at this standardization process. After the standardization process, the corpus is then checked for reliability and validated.

3.3 Data Reliability and Validation

At this point the data is checked for reliability. The model builds on the bases of the work of Khurana et al. (2018) for this part. However, this model scans the Internet to see other solutions that consider the sources we have used and rate them in a scale of zero (referring to the one that is not used at all) to ten (referring to the one that is mostly used) based on the frequency of use. The sources of data are also checked if they are not 'fly-by-night' databases. This is important to ensure the reliability of the sources of threat data. The IP addresses and MAC addresses of the machines are checked for their reliability. However, this validation does not include checking for spoofing. But the validation includes checking the period such IP and MAC have been up and running, running trace-route to check if they are still in the same geographic area. This considers a public cloud storage. The validation is also done bearing in mind the DHCP protocol that changes IP addresses every time a machine is rebooted. Within this process, the model also starts the data pipelining process. The data pipelining process creates the necessary checks and balances for tracking if processes are running until completion. Should a process start and not finish, our solution flags it as an anomaly and sends the administrator an alert to let them know which process has not completed and at what point was it stopped. This is to ensure that analyses are done on complete data. This also makes it easy to trouble-shoot our solution in case of failure. Moreover, at this process the data is also enriched with other threat intelligence data from other systems. This is just to ensure that we make use other threat intelligence from other open sources systems. After enrichment, the corpus in data objects is then pushed to the security modules.

3.4 Data Security

Within this process, the cleansed, validated and standardized data objects are indexed and hashed for easy retrieval and integrity checks using SHA-1 hashing algorithm. This is the point where our model ensures that 'poisoned data' cannot be added to our corpus. At this point, the model does not allow anyone to insert any more data on the indexed and hashed objects. Should an attacker attempt to inject 'poisoned data' the hash value of the changed data object would reflect that the data has been changed. This action raises a flag and sends the administrator and alert. Once the data objects have been hashed and indexed, they are then encrypted with AES-124 for secure storage. The encryption is per data object as compared to the entire corpus. The security on the crypto comes with a performance cost in that each of the data object has to be individually decrypted before the data can be processed. The encryption process is not necessarily a problem because by the time the data is put into the database, there is not real-time requirement for it to be processed. So, the encryption is not necessarily time-bound as is the case with the decryption process. The decryption performance cost is balanced by the hash-based quick retrieval of data objects.

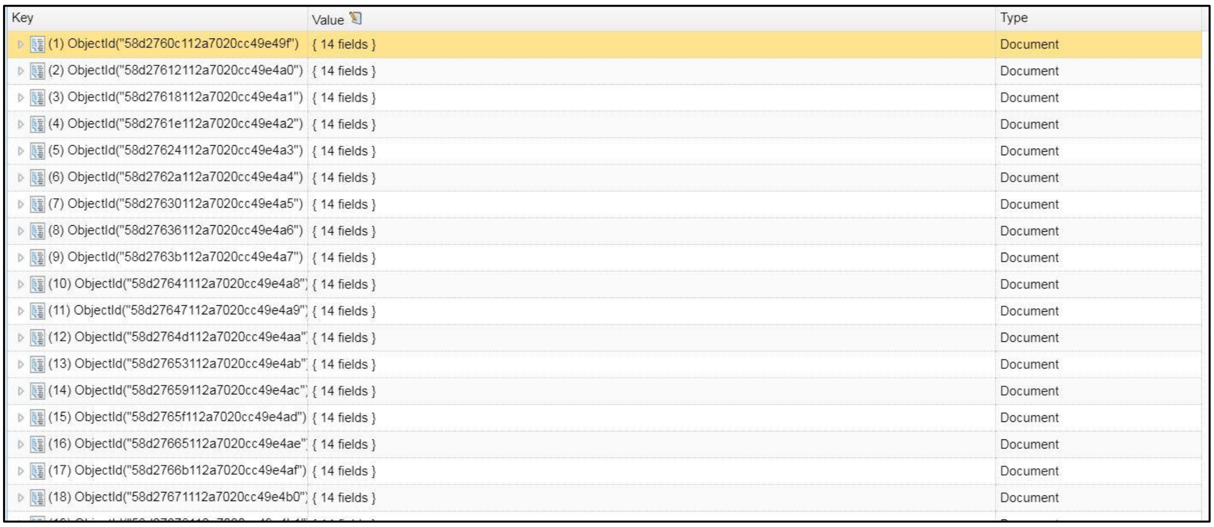
3.5 Data Storage

The hashed, indexed and encrypted data objects are stored in an encrypted database. This just adds another layer of security to prevent unauthorized access to the database. So an attacker would have to go through the database encryption before they can get to the encrypted data objects. So it takes to layers to get to the plain-text data. The system uses a need-to-know principle to restrict access to database. Hence, database access is restricted to the module of the system that does the processing and analysing of the data and the administrators only. This also has the pipelining feature to monitor incomplete processes and alert the administrators in case of incomplete processes.

The next section discusses the results of the first phase on the model.

4. Experimental Results

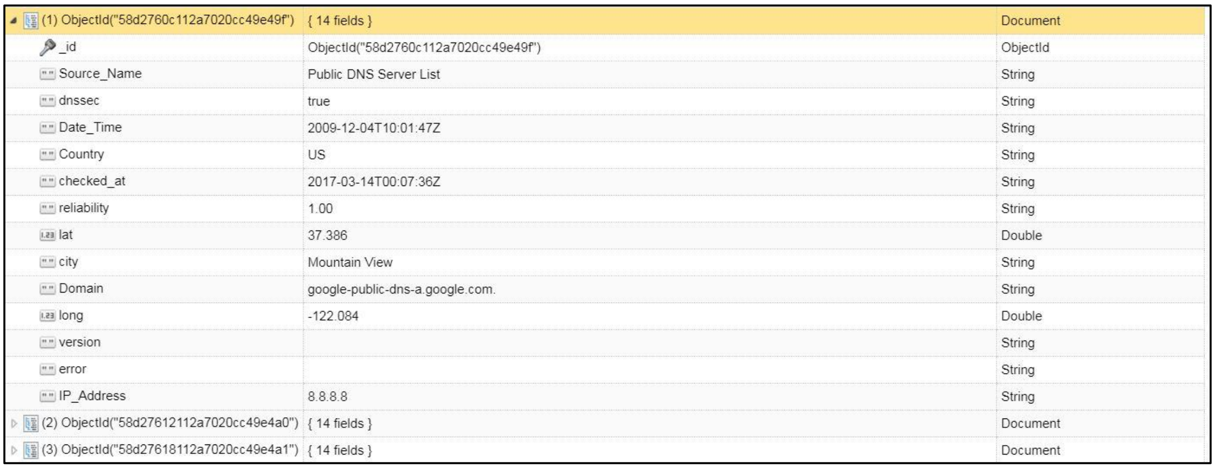
The combined raw data corpus from all sources comprised of 2,562,011 records. This eventually became 1,175,552 after the data cleansing process. About 46% of raw data is clean, which means that approximately 54% of the data was just noise. This is a great cause for concern for most of the cyber threat intelligence systems that ingest raw data without first cleansing and scrubbing it. The resultant 1,175,552 entries were stored in 344,064 data objects. This translates to about 3.4 records per data object. Figure 2 illustrates the indexed hash values of each data object. The hash value is contained in the 24-character string inside the brackets (“”). For example, the first entry (1) ObjectId (“58d2760c112a7020cc49e49f”) has 24 characters. For each of these data objects there are 14 entries from the source. This is depicted by curly {} which in this case shows {14 fields}. This indicates that from this particular source our solution was able to capture 14 unique fields for the corpus.



Key	Value	Type
(1) ObjectId("58d2760c112a7020cc49e49f")	{ 14 fields }	Document
(2) ObjectId("58d27612112a7020cc49e4a0")	{ 14 fields }	Document
(3) ObjectId("58d27618112a7020cc49e4a1")	{ 14 fields }	Document
(4) ObjectId("58d2761e112a7020cc49e4a2")	{ 14 fields }	Document
(5) ObjectId("58d27624112a7020cc49e4a3")	{ 14 fields }	Document
(6) ObjectId("58d2762a112a7020cc49e4a4")	{ 14 fields }	Document
(7) ObjectId("58d27630112a7020cc49e4a5")	{ 14 fields }	Document
(8) ObjectId("58d27636112a7020cc49e4a6")	{ 14 fields }	Document
(9) ObjectId("58d2763b112a7020cc49e4a7")	{ 14 fields }	Document
(10) ObjectId("58d27641112a7020cc49e4a8"	{ 14 fields }	Document
(11) ObjectId("58d27647112a7020cc49e4a9"	{ 14 fields }	Document
(12) ObjectId("58d2764d112a7020cc49e4aa"	{ 14 fields }	Document
(13) ObjectId("58d27653112a7020cc49e4ab"	{ 14 fields }	Document
(14) ObjectId("58d27659112a7020cc49e4ac"	{ 14 fields }	Document
(15) ObjectId("58d2765f112a7020cc49e4ad")	{ 14 fields }	Document
(16) ObjectId("58d27665112a7020cc49e4ae"	{ 14 fields }	Document
(17) ObjectId("58d2766b112a7020cc49e4af")	{ 14 fields }	Document
(18) ObjectId("58d27671112a7020cc49e4b0"	{ 14 fields }	Document

Figure 2: Indexed and hashed data objects in our database

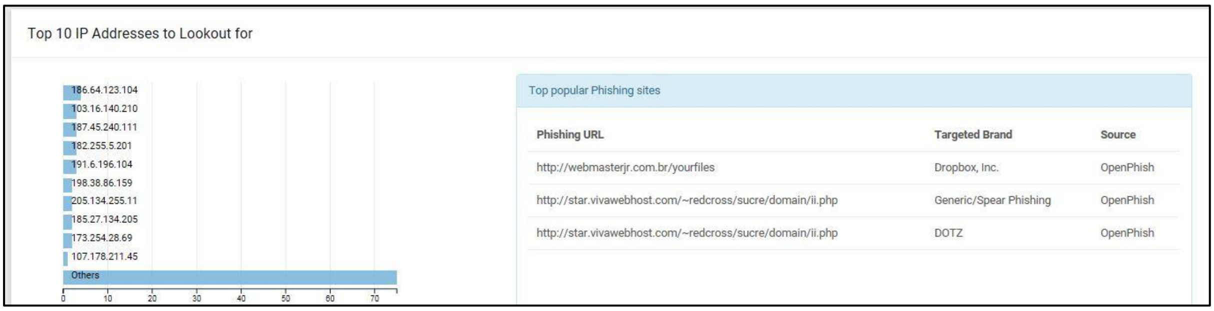
The next screenshot show the type of data that was actually captured in the data objects from figure 2 and the fields thereof. Each of the sources came with a different number of fields that could be captured. It ranges from a source with only seven fields to those with 18 fields. The variety of data from the different sources motivated our choice of an object-oriented database system. This enabled us to store any data we received and this approach also allows portability in terms of adding more sources in future.



(1) ObjectId("58d2760c112a7020cc49e49f")	{ 14 fields }	Document
_id	ObjectId("58d2760c112a7020cc49e49f")	ObjectId
Source_Name	Public DNS Server List	String
dnssec	true	String
Date_Time	2009-12-04T10:01:47Z	String
Country	US	String
checked_at	2017-03-14T00:07:36Z	String
reliability	1.00	String
lat	37.386	Double
city	Mountain View	String
Domain	google-public-dns-a.google.com.	String
long	-122.084	Double
version		String
error		String
IP_Address	8.8.8.8	String
(2) ObjectId("58d27612112a7020cc49e4a0")	{ 14 fields }	Document
(3) ObjectId("58d27618112a7020cc49e4a1")	{ 14 fields }	Document

Figure 3: The captured data inside each of the indexed and hashed data objects

As can be noted in figure 3, the data includes IP address, reliability score, a unique object ID, date of verification and others. The contents of each of the data objects vary depending on the data source. Figure 3 depicts the contents of each of the data object from the first object in figure 2. This way of capturing our data makes it very easy to identify 'data poisoning' attacks. However, this does not necessarily mean an attacker would not be able to poison our data. Instead, it means that, for attackers to do so, they would have to study each and every one of our data objects and then craft their 'poisoned data' to mimic ours. This is not an impossible thing to do for a determine attacker. But he or she would have to study the pre-processing algorithms. Figure 4 illustrates the preliminary results from the initial analysis of the data.

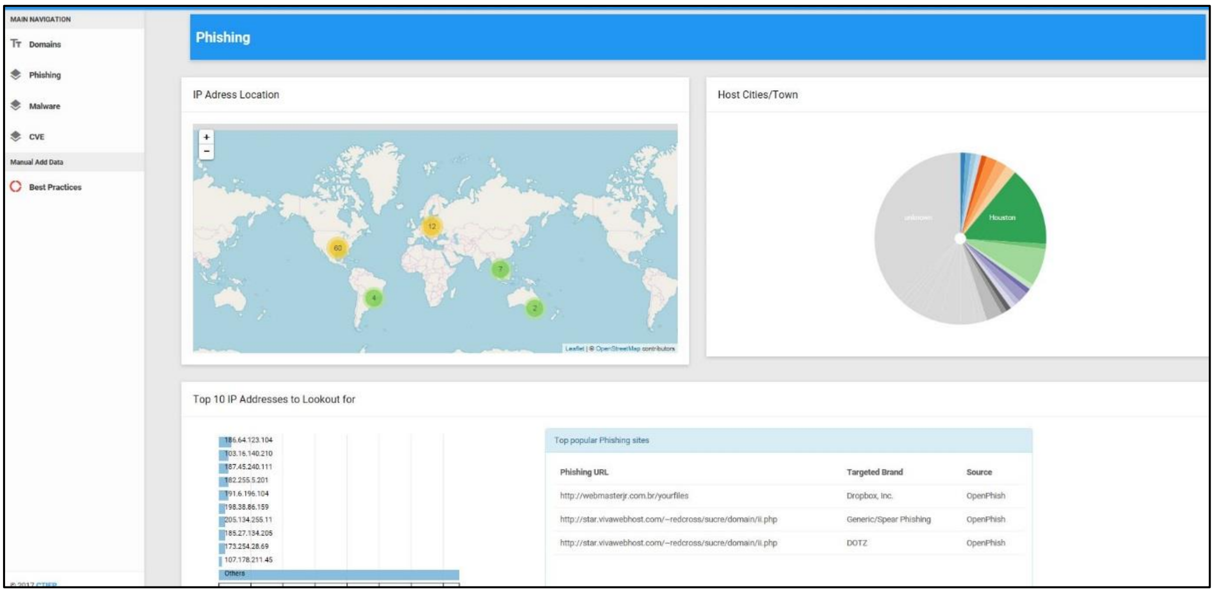


Top popular Phishing sites

Phishing URL	Targeted Brand	Source
http://webmasterjr.com.br/yourfiles	Dropbox, Inc.	OpenPhish
http://star.vivawebhost.com/~redcross/sucre/domain/ii.php	Generic/Spear Phishing	OpenPhish
http://star.vivawebhost.com/~redcross/sucre/domain/ii.php	DOTZ	OpenPhish

Figure 4: The Top 10 IP Addresses to Lookout for

This figure shows the top ten IP addresses that system and network administrators must look out for in their traffic analysis. It also shows some of the most popular phishing sites and the sources. A somehow, comprehensive screenshot is depicted in figure 5.



Top popular Phishing sites

Phishing URL	Targeted Brand	Source
http://webmasterjr.com.br/yourfiles	Dropbox, Inc.	OpenPhish
http://star.vivawebhost.com/~redcross/sucre/domain/ii.php	Generic/Spear Phishing	OpenPhish
http://star.vivawebhost.com/~redcross/sucre/domain/ii.php	DOTZ	OpenPhish

Figure 5: Overview of the Threat Intelligence Dashboard

The two quadrants on top show the geographical spread of phishing malware sources. This is a drill-down map the goes deep until the actual geo-coordinates and IP. It must be mentioned at this point that this work was done as a proof-of-concept. The next phase is to delve deep into a deep analysis to gain more insights from the data. More of our envisaged future work is covered in the next section.

5. Conclusion and Future Work

Attackers are now exploiting a vulnerability in the data training process of AI and ML inspired cyber threat intelligence systems to allow their malicious code to evade detection. On this premise, there is a growing number of research efforts that attempt to detect 'data poisoning' attacks. This paper among others has noted the difficulty of 'data poisoning' detection. Some researchers have since decided to use only a small fraction of their corpus to train their models with the hope that the small datasets would have minimal 'poisoned data'. Evidently, the results of our data cleansing process also show that more than 50% of the sourced threat data is just noise. However, using a small fraction of the corpus for data training purposes has also been shown to exhibit a great risk which may have dire consequences. For example, if it so happens that the sampled small data set has to a larger extent 'poisoned data', this approach fails with serious implications to the underlying data classification algorithms.

This paper through its proposed model ensures the completeness and standardization of the data. These are both critical to make accurate decisions. The data itself is validated by using a mix of related indicators (i.e. benchmark with other cyber threat intelligence systems and the geo-location) to determine the reliability of the raw input data. The security feature of the model provides an added value that ensures that 'poisoned data' cannot be maliciously added on our system. Based on these features, our model can be argued to provide a plausible solution to the 'data poisoning' problem of AI and ML inspired cyber threat intelligence systems. However, our solution is based on working with trusted sources of input raw data. The dynamics of our solution change completely if the input raw data comes with 'poisoned data'. This is one area that future research must focus on.

This paper has focused on the data collection, validation, cleansing, pipelining and security without really looking at the cyber threat intelligence exchange part. Future work goes beyond collection and security to include how our solution processes and analyzes the data and transform it to actionable cyber threat intelligence. It goes further to discuss how we envisage to build sustainable partnerships to exchange and share the cyber threat intelligence with other stakeholders. The fully-fledged system is also envisaged to make technical recommendations on the rules to add to firewalls and other network intrusion detection systems. Furthermore, it will analyze cyber threat patterns from existing data sources to perform predictive analytics to forecast future threats. The idea is to help anticipate future 'data poisoning' threats that try to evade detection systems.

References

Anstee, D. (September 2017). The great threat intelligence debate, *Computer Fraud & Security,* vol.2017, no.9, September 2017, pp.: 14 – 16. DOI: https://doi.org/10.1016/S1361-3723(17)30099-4

Biggio, B., Pillai, I., Bulo, S.R., Ariu, D., Pelillo, M. and Roli, F. (2013). Is Data Clustering in Adversarial Settings Secure?, *In Proceedings of the 2013 ACM workshop on Artificial intelligence and security (AISec '13)*. ACM, New York, NY, USA, pp.: 87 – 97. DOI: https://doi.org/10.1145/2517312.2517321

Chiba, D., Akiyama, M., Yagi, T., Hato, K., Mori, T. and Goto, S. (2018). DomainChroma: Building actionable threat intelligence from malicious domains, *Computer & Security,* vol. 77, no. 2018, pp.: 138 – 161. DOI: https://doi.org/10.1016/j.cose.2018.03.013

Dara, S., Zargar, S.T., Muralidhara, V.N. (2018). Towards privacy preserving threat intelligence, *Journal of Information Security and Applications,* vol. 30, no.2018, pp.: 28-39. DOI: https://doi.org/10.1016/j.jisa.2017.11.006

Ernst & Young Global Limited (2016). How do you find the criminal before they commit the cybercrime? – A close look at cyber threat intelligence*, Ernst &Young Global Limited*, EYG no. AU3750, available online: https://www.ey.com/...cybercrime/.../EY-how-do-you-find-the-criminal-before-they-commtthe-cybercrime.pdf, accessed [07-01-2019].

Geers, K., Kindlund D., Moran, N., and Rachwald, R. (2014). World War C: Understanding Nation-State Motives Behind Today's Advanced Cyber Attacks*, FireEye Report, FireEye, Inc*. available in**:** *https://www.fireeye.com/content/dam/fireeye-www/global/.../fireeye-wwc-report.pdf*, accessed [05-09-2018]

Geers, K. (2011). Strategic Cyber Security, NATO Cooperative Cyber Defence Centre of Excellence, June 2011, Tallinn, Estonia. ISBN 978-9949-9040-5-1

Giles, M. (2018). AI for Cybersecurity is a hot new thing – and a dangerous gamble, *MIT Technology Review,* available online: https://www.technologyreview.com/s/611860/ai-for-cybersecurity-is-a-hot-new-thing-and-a-dangerous-gamble/, accessed [05-09-2018].

Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C. and Li, B. (2018). Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning, *In the 39th IEEEE Symposium on Security and Privacy,* San Francisco, CA, USA, 21st – 23rd May 2018.

James, L. (2018). Making cyber-security a strategic business priority, *Network Security,* vol. 2018, no.5, May 2018, pp.: 6 – 8. DOI: https://doi.org/10.1016/S1353-4858(18)30042-4

Khurana, N., Mittal, S. and Joshi, A. (2018). Preventing Poisoning Attacks on AI Based Threat Intelligence Systems, arXiv:1807.07418v1 [cs.SI], 19 July 2018, available online*:* https://arxiv.org/pdf/1807.07418

Knight, W. (2017). The Dark Secret at the Heart of AI, *Communications of the ACM, ACM TechNews, MIT Technology Review*, available online: https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/, accessed [07-01-2019].

Lundbohm, E. (2017). Understanding nation-state attacks, *Network Security,* Vol. 2017, no.10, pp.: 5 – 8. DOI: https://doi.org/10.1016/S1353-4858(17)30101-0

Paudice, A., Munoz-Gonzalez, L., Gyorgy, A. and Lupu, E.C. (2018). Detection of Adversarial training Examples in Poisoning Attacks through Anomaly Detection, *Cornell University Library,* arXiv:1802.03041 [stat.ML], available online: https://arxiv.org/abs/1802.03041, accessed [18-10-2018].

Ponemon Institute (2017). The Value of Threat Intelligence: The Second Annual Study of North American & United Kingdom Companies, A Research Report Sponsored by Anomali, *Ponemon Institute LLC,* September 2017, pp.: 1 – 27.

Roli, F., Biggio, B. and Fumera, G. (2016). Adversarial Pattern Recognition: Learning in the Presence of Attackers, *Pattern Recognition and Applications Lab,* Montreal, June 15, 2016.

Rubinstein, B.I.P., Nelson, B., Huang, L., Joseph, A.D., Lau, S., Rao, S., Taft, N. and Tygar, J.D. (2009). ANTIDOTE: Understanding and Defending against Poisoning of Anomaly Detectors, *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement. ACM*, 2009.

Steinhardt, J., Koh, P.W. and Liang, P. (2017). Certified Defenses for Data Poisoning Attacks. *The Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, Canada, USA, 4th – 9th December 2017.

Treit, R., Stewart, H. and Parikh, J. (2018). Protecting the protector: Hardening machine learning defenses against adversarial attacks, *Microsoft Secure – Wndows Defender Research,* August 2018, available in: https://cloudblogs.microsoft.com/microsoftsecure/2018/08/09/protecting-the-protector-hardening-machine-learning-defenses-against-adversarial-attacks/, accessed: [05-09-2018].

Tounsi, W. and Rais, H. (2017). A survey on technical threat intelligence in the age of sophisticated cyber-attacks, *Computers & Security Journal,* 72(2018), pp.: 212 – 233. DOI: https://doi.org/10.1016/j.cose.2017.09.001

Van der Walt, E., Eloff, J.H.P. and Grobler, J. (2018). Cyber-security: Identity Deception Detection on Social Media Platforms, *Computers & Security (2018).* DOI: 10.1016/j.cose.2018.05.015