Bitcoin Ransomware Detection using Deep Neural Networks

Abhishek Aditya BS*
Department of Computer Science
PES University, Bangalore, India
abhishek.aditya10@gmail.com

Vinay P Naidu*
Department of Computer Science
PES University, Bangalore, India
vinaypurushothamnaidu@gmail.com

Vishal R*
Department of Computer Science
PES University, Bangalore, India
vishalramesh01@gmail.com

***Abstract*—Ransomware is a sophisticated malware that has grown quickly in recent years, that prevents users from accessing their data using encryption techniques until a ransom is paid to the attacker. Traditional machine learning algorithms tend to be biased towards the more frequently occurring class categories, and fail to capture the general pattern or structure of the ransomware bitcoin addresses. They tend to over fit to the data provided thereby performing poorly on the real world data instances, which have unknown class category labels. Hence we believe that traditional machine learning algorithms do not perform well enough to be used in such a critical data security problem. Deep Neural Networks have proven to work well with time-series data as well as multi-class classification and clustering problems, and it captures various levels of granularity of the underlying structure in the data set at different layers of the model architecture.**

***Index Terms*—Bitcoin, Ransomware, Detection, Deep Neural Networks, Deep Learning, Cryptocurrency**

I. Introduction

Ransomware is a sophisticated malware that has grown quickly in recent years, that prevents users from accessing their data using encryption techniques until a ransom is paid to the attacker. This results in huge losses for businesses and individuals. An Internet of Things architecture generally consists of a wide range of Internet-connected devices or things such as Android devices, and devices that have more computational capabilities (e.g storage capacities) are likely to be targeted by ransomware authors [1]. One of the main risks include the ability to use the compromised devices to further attack the underlying architecture. Consequences include temporary or permanent loss of sensitive information, disruption of regular operations, such as to restore systems and restore an organisation's reputation.

Ransomwares can be classified into numerous types, the most violent and virulent being crypto ransomware. Crypto ransomware not only encrypts user data, but it also tries to encrypt information on both mapped and unmapped network devices, putting a whole department or company to a standstill if just one machine is infected [2]. In this type of attack, the attacker does not benefit by selling stolen information on underground markets but from the value of items that victims of attack assign to their locked data and their willingness to pay a fee to regain access. Crypto ransomware prefers Bitcoin network as during the ransomware attack, the victim's system remains fully functional thus allowing the victim to pay the ransom in Bitcoins on the system [3].

Bitcoin is a peer-to-peer online communication protocol which was introduced by group of developers [4]. This framework can be used to make electronic payments since it servers as a virtual currency. Bitcoin network is decentralised and the transactions are stored in a public ledger that is distributed to all the nodes in the network. Transactions recorded in the ledger are irreversible and is not under any influence of a single authority. A Bitcoin address consists of alphanumeric characters which are obtained from public and private keys of the user. This makes the user details pseudo-anonymous.

The traditional ransom payment methods have a number of drawbacks such as restricted geographic availability reduces the number of paying victims, and they are operated by businesses subject to local laws, which may force them to reverse transactions or monitor ransom receivers [5]. To overcome this the attackers have adapted to Bitcoin .

Like other types of malware, ransomware spreads through a multitude of channels like malicious email attachments, pay-per-install networks and existing weaknesses in network services to spread inside a LAN [5]. Once the ransomware is executed on the host it encrypts the files and documents and displays a ransom message on the screen saying files will be decrypted upon paying ransom in Bitcoins. The ransom message includes ransom addresses of Bitcoin wallets victims are supposed to pay into. Once the payment is done ransomware decrypts the files and documents which were held for ransom. The ransomware attackers then exchange the bitcoins for cash and other fiat currencies.

Cryptocurrencies have ganied popularity over past few years and this trend has attracted cyber-threat actors to exploit the existing vulnerabilities and infect their targets. The malicious actors use cryptocurrency malware to perform complex computational tasks using infected devices. It is a challenging task to detect this type of threat by a manual or heuristic method. Various Machine Learning algorithms [6]–[8] have been proposed to detect the Ransomware Bitcoin addresses,

but they do not generalize enough to classify Bitcoin addresses belonging to different Malware families and the test results need to be validated on more recent Bitcoin addresses.

This paper goes into further details on ransomware works in section II, related works in section III, and our approach to the solution in section V

II. Background

Ransomware attack begins when the targeted users receive suspicious mail-attachments or if any of their network services, pay-per-install have any vulnerabilities which can be exploited [5]. The ransomware software starts executing on the host system and encrypts the files and documents valuable to the user. Once the encryption is done, the ransomware attacker who requests payments in Bitcoin will broadcast a Bitcoin address to which the victim needs to send money to [9]. The ransomware displays a ransom message on the user's screen, demanding a ransom to be paid in Bitcoins to regain access to their files by decrypting them. This address is a ransom payment address from which clustering heuristics in the Bitcoin network can be computed.

The ransom message instructs the victim to purchase the Bitcoins from specific exchanges and online services that allow the conversion of Bitcoins to fiat currencies. These exchanges operate either globally or regionally and some of them cab be centralized while others allow direct exchange facility. Some of the Ransomware families like Locky and Cerber will generate a unique ransom address which can be used to identify the paying victims, while WannaCry and CryptoDefense ransomware families reuse the same address, but the victim needs to send hashed payment transaction so that the attacker can verify the paying victims. On the payment of ransom the victim's files and documents are either automatically decrypted by the ransomware or the attacker sends the decryption keys which can be used to decrypt their files.

To convert the Bitcoins obtained by ransomware attacks into fiat currencies (eg. USD) the attackers deposit the bitcoins into exchanges. Some of the exchanges require them to provide information about their clients, so they often deposit the Bitcoins into mixers, which are services that intermix bitcoin inflow from various sources, to conceal the bitcoin trails.

Crypto Ransomware can be classified into three types [10], [11]:

- Symmetrical Cryptosystem Ransomware : uses a symmetrical encryption algorithm like AES or DES for encrypting user's files. It uses the same public key for both encryption and decryption. So the victim can regain access to his files by applying techniques like reverse engineering or memory scanning techniques.
- Asymmetrical Cryptosystem Ransomware : uses a public key included inside the ransomware file or obtained during contact with the command and control (C&C) server to encrypt the victim's files. Since the attacker owns the private key, victim cannot decrypt unless he obtains the private key by paying the ransom.
- Hybrid Cryptosystem Ransomware : uses a dynamically generated symmetric public key which itself is encrypted by the public key embedded in the ransomware file. So it inherits the benefits of both types of crypto ransomware encryption.

There has been extensive research and study on detection of ransomware addresses which includes some approaches like using different heuristics for grouping the bitcoin addresses into maximal subsets or clusters that can be associated to some real-world actors. The multiple-input heuristics [12] takes into account that two addresses used as inputs in the same transaction must be controlled by the same real-world actor. If one input address is used in another transaction along with other input addresses, they can all be linked to the same real-world actor, clustering of Bitcoin addresses and tagging addresses with attribution data are two central features that are nowadays supported by modern cryptocurrency analytic tools (e.g. Chainalysis, Elliptic, GraphSense, Bitcluster) [9].

Some other approaches include identifying patterns of specific features within a malware code or behaviour to distinguish malware from non-malicious applications, for example, NLP could be used to analyse the function calls, installation activities etc [1]. Most ransomware families attempt to link to command and control servers before executing their destructive payloads so network-based approaches can be useful in identifying ransomware assaults.

III. Related Works

Masarah Paquet-Clouston et. al, in their paper [9] used a data driven approach to identify and gather data on bitcoin transactions related to suspicious activities based on the footprints left on the public bitcoin blockchain. They also found out that the market is highly skewed as there are only a few players who are responsible for the majority of the payments made in bitcoin. Monetary flows were traced by computing network representations and calculating the summary statistics (such as number of transactions and the estimated value that's flowing between two addresses) for each directed edge. Cluster graphs were used to partition addresses into maximal subsets or clusters that belonged to the same real-world actors. Using the above graphs, authors also estimate the lower bound direct financial impact of each ransomware family and conclude that total ransom amounts are relatively low.

Ahmad O. Almashhadani et. al. in their paper [2] designed a network based intrusion detection system with two independent classifiers, packet classifier and flow-based classifier, working in parallel on packet and flow levels to detect the packet-level and flow-level feature vectors coupled with a decision unit which detects any suspicious activity. A dedicated

testbed was built and Locky, one of the malware families, was taken for the case study. On a packet-based data set using Random Forest Machine learning algorithm they were able to achieve a F1 score of 0.979 and an accuracy of 98.72 % . On a flow-based data set, Bayes Net algorithm achieved a F1 score of 0.971 and an accuracy of 99.83%.

Danny Yuxing Huang et. al. in their paper [5] performed a two-year end-to-end study on ransomware ecosystem, including ransomware payments, victims, and operators on a wide scale. The ransomware families produce a unique ransom Bitcoin wallet address for the paying victims. Since the Bitcoin blockchain is a public sequence of timestamped transactions involving wallet addresses, to discern transactions belonging to ransom activity, they designed a methodology to trace known victim payments (victims who reported on public forums) and cluster them with previously unknown victims (victims who have not reported) to filter out the transactions not linked to ransom payments. They found that Cerber and Locky ransomware families generated most income and over the course of 22 months, we were able to track $16,322,006 USD in 19,750 suspected victim ransom payments for 5 ransomware families.

Cuneyt G. Akcora in their paper [13] used Topological Data Analysis (TDA) for detection of ransomware transaction patterns on the Bitcoin blockchain. They modeled the publicly available Bitcoin transaction ledger into a Bitcoin transaction graph, with nodes and edges where nodes can either belong to address type or transaction type, and edges connecting the address and transaction nodes. They used Co-spending heuristic [14], as one of the baseline methods for detecting ransomware addresses. Compared to conventional detection methods like DBSCAN clustering algorithm they achieved an accuracy of 69% with TDA, with overall gain of 213.8% and compared to pairwise Cosine similarity algorithm they were able to achieve an accuracy of 78% with a gain of 2.9%.

Authors of [15] used a new approach which detects the ransomware before it encrypts the user files. This pre-encryption detection algorithm (PEDA) consists of two phases. Phase 1 incorporates a learning algorithm (LA), which recognizes suspicious behaviour based on the API pattern recognition. PEDA generates a signature of the malicious program and stores it in the signature repository. In phase 2 the signature will be compared with known crypto-ransomware in the signature repository using SQL queries. If a match is found an alert is given to the user else it is allowed to execute in a sandbox environment capturing all APIs and then analyzing all the collected APIs to make a prediction. PEDA-Phase-I was compared to three machine learning algorithms, Random Forest (RF), Naive Bayes (NB) and Ensemble. It was observed that PEDA-Phase-I performs the best with AUC of 0.9930 and lowest test error of 0.0295. Their approach has few shortcomings, it may achieve a high false positive rate (FPR) and this method could only detect known crypto ransomwares.

Kirat Jadhav et. al. in his paper [6] showcases the impact of several supervised machine learning techniques on the successful detection of Bitcoin payments for Ransomware attackers. The data set in question is a multi-class data set that is very unbalanced. The Gradient Boosting and XGBoost algorithms successfully recognised more attack types with an accuracy of 99% and average F-Measure of 0.98 compared to other classifiers evaluated, such as the Naive Bayes, Multilayer Perceptron, k-Nearest Neighbor, and Random Forest Classifiers. However the algorithms need to be tested on more such data sets which have imbalanced class proportions. Further, emphasis needs to be done on classifiers which take into account the minority classes from the training set for making more accurate predictions and more work needs to be done on validating the results on more recent spurious Bitcoin transactions involving ransomware payments.

Authors of [7] propose a Pre encryption detection algorithm for detecting crypto ransomware prior to the occurrence of any encryption, there are two levels, first is to compare the file with a signature of a known crypto ransomware stored in a signature repository, second one is a machine learning approach to train a predictive model using data from the API. Cuckoo Sandbox analysis system to capture all the API requests, information from these requests was extracted and converted into data set format for machine learning training and testing. They also propose six different metrics apart from the conventional ones like precision, recall, F score, etc. The new ones are likelihood ratio, diagnostic odds ratio, Youden's index, number needed to diagnose, number needed to misdiagnose and net benefit, these metrics were proposed for better insight and the paper provides a range of values in which the performance is optimal.

Authors of [8] propose a deep Recurrent Neural Network (RNN) model for identifying cryptocurrency malware threats. The opcodes from Windows applications' are analysed using the RNN. A data set that contains 500 cryptocurrency malware samples and 200 benign ware samples is used. The malicious executable was collected from repositories like VirusTotal and Virus Share. The files were unpacked, decompiled using the object-dump tool, and then a script was applied to obtain only the opcodes; the sequences of these opcodes were used for training the model. Since the sequence length was high, tokenization was applied which maps phrases to numbers. Several models were trained and evaluated using 10 fold cross validation and the best configuration achieves about 98% detection accuracy.

Authors of [16] propose a Software Defined Networking based approach that utilises the network communication. The analysis of HTTP message sequences and their respective content sizes were used to detect threats. Communications of two ransomware families, namely CryptoWall and Locky were analysed. It was found that both of them use a custom protocol and they were similar, this characteristic feature is used for ransomware detection. The method is based on the size of the

data inserted by the victim in the outgoing messages. There are 3 phases for the proposed scheme, the learning phase to extract the characteristic features of the outgoing messages, a fine-tuning phase to adjust the parameters, and the detection phase where a list was created consisting of the size of the generated HTTP messages, the distance from this vector to the centroid of the family was found, if it was smaller than the limit distance that was found during the learning phase, the system identified this as a sign of ransomware infection. The experimental results obtained using real ransomware samples proved that even such a simple approach is feasible and offers good efficacy. They were able to achieve detection rates of 97–98% with 1–2% or 4–5% false positives. The experimental results confirm that the proposed approach is feasible and efficient.

Authors of [1] propose a machine learning based approach to detect ransomware attacks by monitoring power consumption of Android devices. The proposed method monitors the energy consumption patterns of different processes to classify ransomware from non-malicious applications. A tool called Power-Tutor was used to monitor and sample power usage of all running processes in 500 ms intervals, this can be considered as time series data. Energy consumption logs of both good-ware and ransomware (via VirusTotal API) apps were generated, and the power usage was normalised while creating the dataset. Four ML algorithms, namely K nearest neighbour, Support Vector machine, Neural Network and Random forest were used. Metrics such as precision, recall, accuracy, F score were used to evaluate the models. These conventional models weren't as promising and hence a method is proposed to overcome high distribution of features, power usage samples are divided into subsamples prior to using different classification techniques. This new technique produces much better results and achieves detection rate of 95.65% and a precision rate of 89.19%.

IV. Problem Statement

Our goal is to identify and classify ransomware addresses amongst the different types of Bitcoin addresses in the dataset. Dataset is downloaded and parsed from the entire Bitcoin transaction graph from 2009 January to 2018 December. Using a time interval of 24 hours, daily transactions on the network is recorded and used to form the Bitcoin graph. Network edges that transfer less than B0.3, are filtered out since ransom amounts are rarely below this threshold. Ransomware addresses are taken from three widely adopted studies: Montreal, Princeton and Padua.

The dataset consists of 10 attributes in which the last attribute is the target label of the bitcoin addresses. The features are described below [13] :

- address: (String) The Bitcoin address.
- year: (Integer) Year of the transaction.
- day: (Integer) Day of the year. 1 denotes first day, similarly 365 is the last day.
- weight: (Float) Sum of the fraction of coins that originate from a starter transaction.
- count: (Integer) Number of starter transactions connected to the address node of interest through a acyclic directed path.
- looped: (Interger) Number of starter transactions connected to the address node of interest by more than one directed path.
- neighbors: (Integer) Number of transactions which have the the address node of interest as the output address.
- income: (Integer) Satoshi amount (1 bitcoin = 100 million satoshis), total amount of coins output to the address node of interest.
- label: (Category String) Name of the ransomware family (e.g., Cryptxxx, cryptolocker etc) or white (i.e., not known to be ransomware).

The dataset imposes few challenges like skewed distribution of the class labels i.e the dataset is imbalanced which might affect the quality of the classifier models, its performance metrics, and its generalization to real world ransomware address detection. Also the results obtained from the classfier model cannot be validated on more such datasets because the data that contains the ransomware addresses with proper target labels is relatively smaller than the non-ransomware addresses. Since the data is heavily skewed the model gives more bias towards the majority classes and does not take into account the minority classes which results in overfitting of the data.

One of the most critical problems in today's digital era is dealing with ransomware attacks. Unless one keeps a backup and take measures such as avoiding suspicious email attachments, keeps the software up to date, taking regular backups, turning the system down at the sign of malware attack to minimize the loss suffered, protecting oneself from the ransomware is hard. Since the generated data volume is increasing exponentially and more users are using internet, they become more vulnerable to such attacks. The traditional ransom payment methods have drawbacks, the attackers are now demanding the ransom in the form of cryptcurrency mainly Bitcoin, which results in huge financial losses on the individuals and establishments. Since cryptocurrency is decentralized and there is no proper laws enforcing the transactions, such attacks takes place from different countries where the laws are weak or non-existent, making crypto ransomware attacks more frequent.

So mitigating such attacks can be proven to be beneficial to many individuals and establishments, as it prevents the loss of assets in the form of data or money. By predicting the type of ransomware address in advance in the initial stage of the attack can be helpful since it minimizes the loss endured. It can also be helpful in other areas like cybersecurity, and data security in general.

V. Approach

Since the problem statement requires classification as its core functionality, we are using machine learning approach to solve the problem. Traditional machine learning algorithms tend to be biased towards the more frequently occurring class categories, and fail to capture the general pattern or structure of the ransomware bitcoin addresses. They tend to over fit to the data provided thereby performing poorly on the real world data instances, which have unknown class category labels. On exploration of other research papers based on machine learning approaches, we found that results are not up to the mark. Hence we believe that traditional machine learning algorithms do not perform well enough to be used in such a critical data security problem.

Deep Neural Networks have proven to work well with time-series data as well as multi-class classification and clustering problems. To tackle the class proportion imbalance we are exploring different approaches like gathering additional data on ransomware bitcoin addresses so that the model does not have high bias, another approach being reducing the number of instances for the majority class label and performing model hyper-parameter optimization by tuning various hyper-parameters like learning rate, regularization, drop out rate and so on. Since we will be using deep neural networks, it captures various levels of granularity of the underlying structure in the data set at different layers of the model architecture.

We would also like to perform various clustering analysis like K-means, Heirarchial clustering, Co-Spending Heuristic [14] based clsutering and multiple input based clustering to identify to better understand the bitcoin graph structure. We would like to also transform the dataset using clsutering and train the neural network on these clusters to acheive better results. Accuracy, prcision, Recall, F1 score are chosen to be our metrics for model evaluation and model slection.

Follow our work on Github : https://github.com/iVishalr/Bitcoin-Ransomware-Detection

References

[1] A. Azmoodeh, A. Dehghantanha, M. Conti, and K.-K. R. Choo, "Detecting crypto-ransomware in iot networks based on energy consumption footprint," *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, no. 4, pp. 1141–1152, 2018.

[2] A. O. Almashhadani, M. Kaiiali, S. Sezer, and P. O'Kane, "A multi-classifier network-based crypto ransomware detection system: A case study of locky ransomware," *Ieee Access*, vol. 7, pp. 47 053–47 067, 2019.

[3] R. Richardson and M. M. North, "Ransomware: Evolution, mitigation and prevention," *International Management Review*, vol. 13, no. 1, p. 10, 2017.

[4] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," *Decentralized Business Review*, p. 21260, 2008.

[5] D. Y. Huang, M. M. Aliapoulios, V. G. Li, L. Invernizzi, E. Bursztein, K. McRoberts, J. Levin, K. Levchenko, A. C. Snoeren, and D. McCoy, "Tracking ransomware end-to-end," in *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2018, pp. 618–631.

[6] K. Jadhav, "Investigating machine learning approaches for bitcoin ransomware payment detection systems."

[7] S. Kok, A. Azween, and N. Jhanjhi, "Evaluation metric for crypto-ransomware detection using machine learning," *Journal of Information Security and Applications*, vol. 55, p. 102646, 2020.

[8] A. Yazdinejad, H. HaddadPajouh, A. Dehghantanha, R. M. Parizi, G. Srivastava, and M.-Y. Chen, "Cryptocurrency malware hunting: A deep recurrent neural network approach," *Applied Soft Computing*, vol. 96, p. 106630, 2020.

[9] M. Paquet-Clouston, B. Haslhofer, and B. Dupont, "Ransomware payments in the bitcoin ecosystem," *Journal of Cybersecurity*, vol. 5, no. 1, p. tyz003, 2019.

[10] A. Liska and T. Gallo, *Ransomware: Defending against digital extortion*. " O'Reilly Media, Inc.", 2016.

[11] M. M. Ahmadian, H. R. Shahriari, and S. M. Ghaffarian, "Connection-monitor & connection-breaker: A novel approach for prevention and detection of high survivable ransomwares," in *2015 12th International Iranian Society of Cryptology Conference on Information Security and Cryptology (ISCISC)*. IEEE, 2015, pp. 79–84.

[12] F. Reid and M. Harrigan, "An analysis of anonymity in the bitcoin system," in *Security and privacy in social networks*. Springer, 2013, pp. 197–223.

[13] C. G. Akcora, Y. Li, Y. R. Gel, and M. Kantarcioglu, "Bitcoinheist: Topological data analysis for ransomware prediction on the bitcoin blockchain," in *Proceedings of the twenty-ninth international joint conference on artificial intelligence*, 2020.

[14] S. Meiklejohn, M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voelker, and S. Savage, "A fistful of bitcoins: characterizing payments among men with no names," in *Proceedings of the 2013 conference on Internet measurement conference*, 2013, pp. 127–140.

[15] S. Kok, A. Abdullah, N. Jhanjhi, and M. Supramaniam, "Prevention of crypto-ransomware using a pre-encryption detection algorithm," *Computers*, vol. 8, no. 4, p. 79, 2019.

[16] K. Cabaj, M. Gregorczyk, and W. Mazurczyk, "Software-defined networking-based crypto ransomware detection using http traffic characteristics," *Computers & Electrical Engineering*, vol. 66, pp. 353–368, 2018.