**Analysis of Bitcoin Heist Ransomware Address**

ECE 552 DL1 - Big Data Technologies| Prof. Erton Boci, Ph.D.

George Mason University Fall 2021

Team 14

Sandeep Kumar Yedla, Sahul Krishna, Madhvi Singh Rajawat

**Abstract**

The recent spread of cryptocurrencies that allow for pseudo-anonymous transactions has led to an increase in different e-crime activities, including bitcoin payments in ransomware attacks that encrypt sensitive user data. Currently, the majority of hackers utilize Bitcoin to make payments, and existing ransomware detection technologies rely solely on a few heuristics and/or time-consuming data collection methods. We offer a uniquely efficient and tractable framework for automatically predicting new ransomware transactions in a ransomware family, based on recent breakthroughs in Topological Data Analysis, given only limited data of prior transactions. Furthermore, our novel technique has shown to be extremely useful in identifying the introduction of new ransomware families, that is, ransomware that has no previous transaction data. (Akcora, n.d.)

**Introduction**

The emergence of blockchain-based technology has been a feature of this decade. Blockchain is, at its heart, a distributed public ledger that records transactions between parties without the need for a recognized central authority. Two strangers can establish an unchangeable action on a cryptocurrency that also is permanently recorded mostly on the ledger and visible to the public. The Bitcoin cryptocurrency was the first Blockchain application. The triumph of Bitcoin has ushered in a new era known as Blockchain, with over 1000 Blockchain-

based coins. The use of bitcoins for ransomware payments appears to be far more common than previously thought.

According to Hernandez-Castro et al. [Hernandez-Castro et al., 2014], "the prevalence of Ransomware is significantly greater than predicted" among respondents to their poll.

The percentage of CryptoLocker victims who claim to have consented to pay the ransom to retrieve their files, 41% appears to be substantially higher than expected 3 percent predicted by Symantec, 0.4 percent predicted by Dell SecureWorks. As a result, comprehending ransomware payments and their total economic impact has emerged as a crucial social concern.

**Big Data Ecosystem Architecture and Block Diagram**

Logo, company name

Description automatically generated

Figure 1

The block diagram briefly defines the data flow for our project.

Diagram, schematic

Description automatically generated

Figure 2

In the first, we read the data into pyspark, and the cleaning is performed on the data, the null and duplicate values are removed, also the labels are changed for some columns in the dataset file. The income column is in Satoshi amount (1 bitcoin = 100 million satoshis) so we divided the complete column by 100 million to get the bitcoin currency.

Data transformation is performed in which after reading the data in pyspark the original excel file is converted to a Parquet file. This transformation can make the data reading faster. We had first read the data using pandas. It took 2.23 seconds, later we read the data after transformation to the Parquet file, and it only took 364 milliseconds.

Sub setting using spark SQL is performed on two columns in our dataset which are year and label. The year data is subset to 2011-2014 as our data is more populated for these years, and the label is subset by only ransomware family e.g., Cryptxxx, CryptoLocker, etc. Ransomware addresses are taken from three widely adopted studies: Montreal, Princeton, and Padua.

Data Visualization with matplotlib is performed and we have derived 2 visuals one bar plot showing transactions per year. Second pie chart showing labels in ransomware family.

Machine learning algorithms, we have used Decision tree and Linear regression to generate the model and derive the accuracy of our model.

**Software’s used:**

* Apache Spark: version 2.4.7
* Apache Parquet is used for data
* transformation.
* Jupyter Notebook for running queries and performing visualizations.

**Literature Review:** (iospress, n.d.)

On Bitcoin, this paper present's the notion of k-chainlets, which extends the concepts of patterns and graphlets to Blockchain topologies. Chainlet analysis gives a better understanding of the Cryptocurrency local topological features and the function of those local relatively high topologies in Bitcoin price development.

This paper  present explains the notion of k-chainlets, which extends the concepts of patterns and graphlets to Blockchain topologies. Chainlet analysis gives a better understanding of the Cryptocurrency local topological features and the function of those local relatively high topologies in Bitcoin price development. (Myre, n.d.)

This paper present's the cryptocurrency payment system which allows users to get several anonymous credentials, known as bitcoin addresses, that may be used to send and receive payments. However, previous study has shown that the system's usage of such addresses may reveal some information about their owners. Because all of the system's transactions are publicly available on the blockchain for examination, it's possible to group distinct addresses belonging to the same user and classify specific applications

**DataExplanation** (https://www.researchgate.net/publication/281773799\_Research\_and\_Challenges\_on\_Bitcoin\_Anonymity)

Our dataset has approximately 3.5 million records and 10 attributes. The data has captured the Bitcoin transaction graph from 2009 January to 2018 December. Using a time interval of 24 hours, daily transactions on the network have been extracted and formed the Bitcoin graph and filtered out the network edges that transfer less than B0.3, since ransom amounts are rarely below this threshold. Ransomware addresses are taken from three widely adopted studies: Montreal, Princeton and Padua.

(Cooper, n.d.)

**Attributes**

Address: (String) Bitcoin address.

Year: (Integer). Year.

Day: (Integer). Day of the year. 1 is the first day, 365 is the last day.

Length: (Integer) designed to quantify mixing rounds on Bitcoin, where transactions receive and distribute similar amounts of coins in multiple rounds with newly created addresses to hide the coin origin.

Weight: (Float) quantifies the merge behavior of input-output addresses. The weight feature represents information on the amount of transactions.

Count: (Integer) designed to quantify the merging pattern. count feature represents information on the number of transactions.

Looped: (Integer) intended to count how many transactions occur.

Neighbors: (Integer)

Income: (Integer) Satoshi amount (1 bitcoin = 100 million satoshis).

Label: (Category String). Name of the ransomware family (e.g., Cryptxxx, cryptolocker etc) or white (i.e., not known to be ransomware).

**Analysis and Visualizations:**

**First plot:**

 The plot is derived by counting the number of years that means for each year the sub setting is performed and the transaction count for a particular year is found. The graph is then plotted between the year on the x-axis and the number of transactions on the y-axis.

Chart, bar chart

Description automatically generated

Figure 3

From the graph, we can say the most number of transactions is made in the year 2012 and the least number of transactions is made in the year 2014. The plot shows the transaction count in ascending order.

**Second plot:**

The pie chart is plotted by sub-setting the label column by ransomware family and the pie is derived showing the family bifurcations in percentage form. The other family in the label is white bitcoins. White Bitcoin addresses are capped at 1K per day (Bitcoin has 800K addresses daily). When compared to non-ransomware addresses, ransomware addresses exhibit more profound right skewness in distributions of feature values that is why we have sub-set the ransomware family bitcoins.

Chart, pie chart

Description automatically generated

Figure 4

The pie chart shows that most addresses of the ransomware are from PaduaCryptoWall (30.5%), followed by PrincetonCerber at (24.3%), MontrealCryptoLocker at (21.8%), PrincetonLocky at (17.1%), and the least number of addresses are from MontrealCryptXXX at (6.3%).

**Third Plot**

The correlation matrix shows relations between each attribute, we can see and analyses if the following feature is positively or negatively correlated.

**Bar chart

Description automatically generated with medium confidence**

Figure 5

**Machine Learning Algorithms**

We have split the data into 70 / 30 ratio for the testing and training of data to run the algorithms.

**Linear regression:**

In the model we have derived the correlation to length as it is pour target variable in respect to all other attributes.

If the value is closer to 1 it is strongly correlated and if it is further from 1 it is not strongly correlated. That is the reason the correlation to length is 1 and so it is highly correlated and to income it is not strongly correlated.

*import six*

*for i in house\_df.columns:*

*if not( isinstance(house\_df.select(i).take(1)[0][0], six.string\_types)):*

*print( "Correlation to Length for ", i, house\_df.stat.corr('length',i))*

*Output:*

*Correlation to Length for year 0.09177169340486724*

*Correlation to Length for day 0.04342247835202402*

*Correlation to Length for length 1.0*

*Correlation to Length for weight 0.060180331529211345*

*Correlation to Length for count 0.6627173924217697*

*Correlation to Length for looped 0.32688044270582417*

*Correlation to Length for neighbors 0.08099163536800867*

*Correlation to Length for income 0.009641611840837548*

**Intercept of Linear regression model.**

***from*** *pyspark.ml.regression* ***import*** *LinearRegression  
lr* ***=*** *LinearRegression(featuresCol* ***=*** *'features', labelCol****=****'length', maxIter****=****10, regPa lr\_model* ***=*** *lr****.****fit(train\_df)  
print("Coefficients: "* ***+*** *str(lr\_model****.****coefficients))  
print("Intercept: "* ***+*** *str(lr\_model****.****intercept))*

*Output:*

*Coefficients: [0.0,0.0,0.0,0.0,0.0,0.993352837149316]*

*Intercept: 0.1747814056096046*

**Root mean square value of Linear regression model.**

trainingSummary **=** lr\_model**.**summary  
print("RMSE: %f" **%** trainingSummary**.**rootMeanSquaredError) print("r2: %f" **%** trainingSummary**.**r2)

*Output:*

RMSE: 0.283893

r2: 0.999956

**Applying model**

We have applied our model on test dataset we are getting prediction as 129, and original length is 130.

*Table

Description automatically generated*

Figure 6

**Decision Tree**

The length is our target variable (dependent variable). Same testing and training data is used for decision tree and the model is run.

**Root Mean Square Value of Decision Tree**

***from*** *pyspark.ml.regression* ***import*** *DecisionTreeRegressor  
dt* ***=*** *DecisionTreeRegressor(featuresCol* ***=****'features', labelCol* ***=*** *'length') dt\_model* ***=*** *dt****.****fit(train\_df)  
dt\_predictions* ***=*** *dt\_model****.****transform(test\_df)  
dt\_evaluator* ***=*** *RegressionEvaluator(*

labelCol**=**"length", predictionCol**=**"prediction", metricName**=**"rmse") rmse **=** dt\_evaluator**.**evaluate(dt\_predictions)  
print("Root Mean Squared Error (RMSE) on test data = %g" **%** rmse)

*Output:*

*Root Mean Squared Error (RMSE) on test data = 2.39521*

**Conclusion**

We have found from analyzing various visualization and machine learning algorithms that:

The bar plot showed the greatest number of transactions were from year 2011.

The pie chart showed the most ransom ware addresses in the transactions were from PaduaCryptoWall.

The scatter matrix showed the correlation between each attribute.

The linear regression model has RMSV of 0.28 and the Decision Tree has RMSV of 2.39. Which shows the best fit model is the linear regression model. (Marsh, n.d.)

# References

Akcora, C. G. (n.d.). Retrieved from https://www.wsj.com/articles/u-s-to-target-crypto-ransomware-payments-with-sanctions-11631885336

Cooper. (n.d.). Retrieved from bitcoin-price-boom-is-helping-sustain-ransomware-hackers: https://www.marketwatch.com/story/bitcoin-price-boom-is-helping-sustain-ransomware-hackers-fbi-says-11637089437

https://www.researchgate.net/publication/281773799\_Research\_and\_Challenges\_on\_Bitcoin\_Anonymity. (n.d.). Bitcoin Ransomware.

iospress, c. (n.d.). *content iospress*. Retrieved from https://content.iospress.com/articles/risk-and-decision-analysis/rda02

Marsh. (n.d.). *Marsh*. Retrieved from https://www.marsh.com/us/services/cyber-risk/insights/ransomware-paying-cyber-extortion-demands-in-cryptocurrency.html

Myre, G. (n.d.). Retrieved from https://www.npr.org/2021/06/10/1004874311/how-bitcoin-has-fueled-ransomware-attacks