Visualizing Cryptocurrency Transaction Data to Illustrate Illicit Activity Detection Methods

By Peter V. Stewart

Abstract-

Cryptocurrency is electronic currency that uses an encryption technology known as blockchain to create virtual tokens that are stored in a public virtual ledger but can only be decrypted using a secret key. This enables the holder of the decryption key to unlock the token to exchange it for cash or goods or other crypto tokens. Over the past decade or so the value of cryptocurrency has fluctuated wildly at times, but the popularity and value reached a peak around the fall of 2021. Despite the rise in popularity and more mainstream use of cryptocurrency as a financial tool, there is still debate about their usefulness fpr society. Although all cryptocurrency transactions are tracked in a public leger, the identities of the holders of tokens are not public and this anonymity has led to widespread use of cryptocurrency in illicit activities such as ransomware, money laundering, and myriad scams. Computer scientists have been active in researching tools for detecting illicit activities on these cryptocurrency exchanges. The goal of this project is to examine some of the data used in this research to try to create visualizations that can help illuminate the distinctions between illicit and normal data used in these projects, and to help illustrate what the advanced analytics and machine learning programs are seeing in the data to be able to make identifications around the activities of the accounts.

Introduction-

The question motivating this project is simple: is there a way to use one of the primary tools of ransomware and crypto scams in order to combat them? Ransomware is an offshoot of the earlier forms of malware like computer viruses and worms. The difference with ransomware is that the attackers use encryption technology to lock victims' data and make it inaccessible unless the victim agrees to pay a ransom fee. In 2021 ransomware attacks doubled in frequency and in a large survey, approximately 37% of global organizations said they were victims of these attacks in some form (Kerner 2021). An average of more than 4000 ransomware attacks have occurred per day since January 2016 which is a 300 percent increase since 2015 (US Justice Dept). Although estimates vary due to the difficulty in gathering data, some experts estimate that ransomware caused around $20 billion in economic losses in 2020 alone (Grauer, 2021). In that year more than 1 in 3 hospitals surveyed reported being victims of ransomware (Weiner, 2020) and the consequences of such attacks can range from loss of access to patient medical records to forcing patients to travel further for emergency services and there is almost always cost from mitigation and technology repair or replacement, whether the ransom is paid or not. Education sector entities such as local primary school districts have been a top target and taxpayers have paid a high cost. In the education sector, 2021 saw 950,129 students impacted by ransomware attacks, with a $3.65 billion lost productivity cost alone (Bischoff, 2022). And famously, one of these attacks in early 2021 shut down the Colonial gas pipeline causing oil prices to rise and causing disruption in gas supplies in parts of the US.

One commonality that is universal to ransomware attacks is the use of cryptocurrency to receive the illicit funds from the victims. And examining the nature of illicit transactions over cryptocurrency, it becomes apparent that ransomware is only the tip of the iceberg. The bar graph below from cryptocurrency research company Chainalysis demonstrates this:

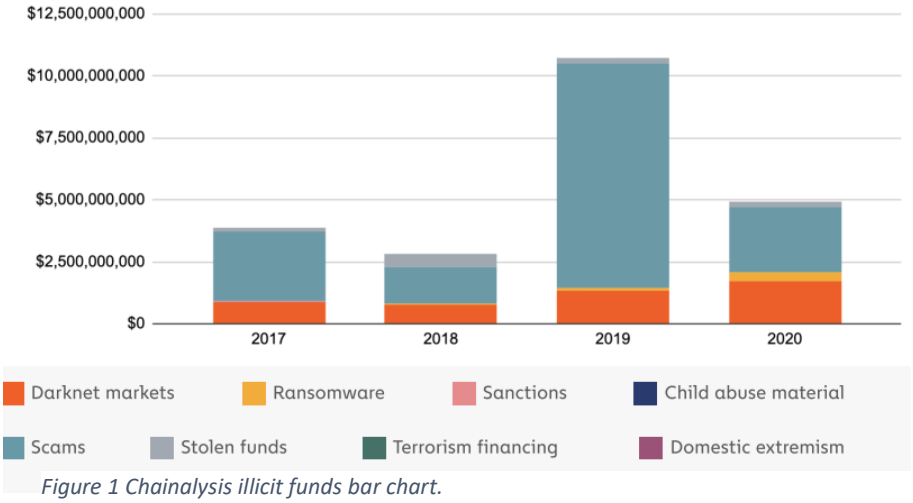


Figure 1 Chainalysis illicit funds bar chart.

In their 2021 report on crypto crime (Grauer, 2021) they detail the types of illicit transactions along several categories and how the funds move through the system. As is somewhat imprecisely demonstrated in this stacked bar chart, the total value of cryptocurrency generated from ransomware is dwarfed by the amount attributed to scams. They mentioned in their report that 2019 was a banner year for such scams. One 2019 scam alone, the PlusToken ponzi scheme, took in over $2 billion from millions of victims (Grauer, 2021). So given all these ill-gotten funds tied to cryptocurrency, one is drawn to question the inherent value of cryptocurrency in society. While this question is somewhat beyond the scope of this paper, it's of some value to offer some background on the currency before we turn to the main objective of this paper, combating the problems laid out thus far.

To understand the story of crypto currency we can start with the largest and most popular form, bitcoin. Bitcoin was started in 2009 by a mysterious anonymous developer called Satoshi Nakamoto. Although there were many attempts at creating virtual currency before bitcoin, the fundamental problem with the idea of digital currency was how to ensure that each token or coin would have a unique and non-duplicatable value. Early iterations of digital currency required a trusted third party to track and maintain the value of coins (Wallace 2011). Bitcoin's design was able to replace the trusted third party by creating a peer-to-peer network of computers that would maintain the ledger of bitcoins by a process called "mining" which is providing the computing power to track transactions and ensure a level of transparency and permanency in the bitcoin network(James, 2021). The intentions of the bitcoin's anonymous creator were made clear with the first successful mining on January 03, 2008 when Nakamoto embedded the text into the "genesis block" of bitcoin, a headline from The Times on that day: "The Times Jan/03/2009 Chancellor on brink of second bailout for banks." (James, 2021). The idea behind decentralized cryptocurrency was that it would be free from government and big bank interference. While this sentiment was certainly popular in 2009 after the great recession, the counterargument to this rationale would be that the entire modern banking system with government control over base interest rates, was designed to stabilize the value of currency. With 13 years of hindsight, one can question whether bitcoin solved any problems with modern government backed

currency or created new ones. Just looking at a simple line graph of the historic value of bitcoin makes this apparent:

Figure 2 Price history of Bitcoin

Clearly if one had bought into bitcoin around the beginning of this chart in 2016, the value of a bitcoin was around $710, by November of 2021 that $710 bitcoin would be worth around $65,000 which would be a great "investment" except that by July 2022 that bitcoin would have gone from $65,000 to about $20,000 in less than a year. This kind of volatility is the nominal reason why governments developed the modern banking system to stabilize the value of currency. Regardless of critiques of Bitcoin, there's little any person or government can do to stop it, since it is basically decentralized and only relies on the willingness of miners to continue to process the transactions, motivated by the chance of gaining bitcoin rewards. While Bitcoin may have been the first cryptocurrency to be widely adopted, many more have been created since.

One of the most popular cryptocurrencies other than Bitcoin is Ethereum. Ethereum was started by Vitalik Buterin, a Canadian developer who as a child was frustrated by centralized authority in a popular video game called World of Warcraft, so again the idea was to develop currency that does not rely on a central authority (Coinloan, 2020). Ethereum uses a similar technology to bitcoin with the addition of a feature called smart contracts which is touted as a way to add programs to the cryptocurrency code base in order to execute transactions based on predetermined conditions (101 Blockchains, 2020). A look at a similar line graph for the historic price of Ethereum shows a very similar trend to Bitcoin, in this case the graph starts in late 2018 and has similar peaks around the fall of 2021 and crashing around the late spring of 2022.

Figure 3 Price History of Ethereum

Now that we've covered the background of the Bitcoin and Ethereum, let's turn to the question of analysis and regulation of the problem of illicit financial activity on these platforms. One of the main reasons for cryptocurrency's role in illicit transactions is the anonymity provided and the lack of a central authority with the ability to interfere in transactions. At the same time, these platforms also have an inherent transparency behind them, as one of the primary features is that transactions are publicly tracked through the ledger system. In the following pages, we'll be examining two research papers with publicly available data sets who have helped to discover processes that can dissect these public ledgers to identify subtle patterns in the transactions to enable identification of unknown illicit cryptocurrency transactions. The goal of this paper is to try to demystify these processes and gain a better understanding of the aspects of cryptocurrency that are presented by using data visualizations. While the techniques used by these methods all revolve around systematic analysis of the data to detect and make predictions about cryptocurrency transactions, this paper will not delve into the actual prediction process, but rather will look at the data involved and help to illuminate what the proposed systems are actually looking at so that we can better understand why these processes work, rather than actually performing the processes on the data which the authors of these papers have already accomplished to great effect.

Bitcoin Ransomware Data Analysis, an Examination of BitcoinHeist Data-

In this section, the focus will be on a 2019 paper published on Cornell's arXivLabs site authored by a group from The University of Texas at Dallas. This paper examined millions of bitcoin addresses and created metadata around the types of transactions conducted and included various addresses that were unveiled as addresses used by ransomware families by previous studies (Gurcan Akcora, 2019). The columns of the dataset are based on analysis of graph data of these bitcoin addresses. The paper proposes to use graph data analysis to make predictions about the ownership of unknown bitcoin addresses with the goal of detecting ransomware families using bitcoin to extract ransom payments. They ultimately argue that using topological data analysis (TDA) is the most effective way to detect ransomware and in fact even claim to be able to attribute addresses based on the ransomware family they belong to. This section will start by giving a little background on TDA and then explain the features

included in the BitcoinHeist Dataset, then will go through the visualizations created and how they help to understand the features of this data.

About the dataset:

TDA is an approach to data analysis that is based on the process of organizing data using a graph that simplifies the relationships between the various features (Talebi 2022). This approach has the benefit of removing some of the noise associated with large datasets by focusing on the underlying shape of the data (Talebi 2022). The dataset we'll be looking at is composed of address entities, or essentially users of bitcoin, with features that are mostly based on analysis of the graph data for these users. Using previous research into the ransomware addresses, the dataset categorizes the addresses based on whether they are of unknown ownership or whether they are known to be associated with a ransomware family and then compiles metadata about these addresses captured in set time intervals in order to track and quantify features about the transaction over this set timeframe. The conclusion of the authors is that using a multistep TDA process was the most effective at answering most of the questions they sought to answer.

The list below is copied from the Kaggle page for the dataset and offers a brief explanation of the dataframe columns:

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.
weight: Float.
count: Integer.
looped: Integer.
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).

From: (https://www.kaggle.com/datasets/sapere0/bitcoinheist-ransomware-dataset)

Some of the features are self-explanatory but let's go through some of the key quantitative columns and try to understand their meaning. First of all, to explain how the data is organized, there are "starter addresses" and these are the beginning of the chain where the flow of coins start in a given time window, which in terms of ransomware would be the victim sending bitcoin to the address requested by the ransomware. ***Length*** is defined as the number of transactions in the chain that runs from the starter transaction to the address entity of that row. ***Weight*** is defined as the sum of fraction of coins that originate from a starter transaction and end up at the address of that row. In other words, how much of the starter transaction amounts ends up flowing down a chain to the address in question. ***Count*** is the number of starter transactions that is associated with the address in question for that row. ***Looped*** is a measure of the number of starter transactions that are connected to the address in questions through different paths. In other words, this measures the degree to which a starter transaction coins are divided up, routed through different paths, then converged back together at the address in question. ***Income*** is the total number of coins output to the address in question which is

measured in Satoshis (1 million Satoshis = 1 bitcoin). ***Neighbors*** is defined as the number of transactions which have the address in question of that row as an output. In other words, if we imagined the data in graph form, each address in the chain would form the node dots and each transaction would be lines or edges and neighbors would be how many edges connect to the address. The below image is an example network from the article which shows how these data look in graph form:

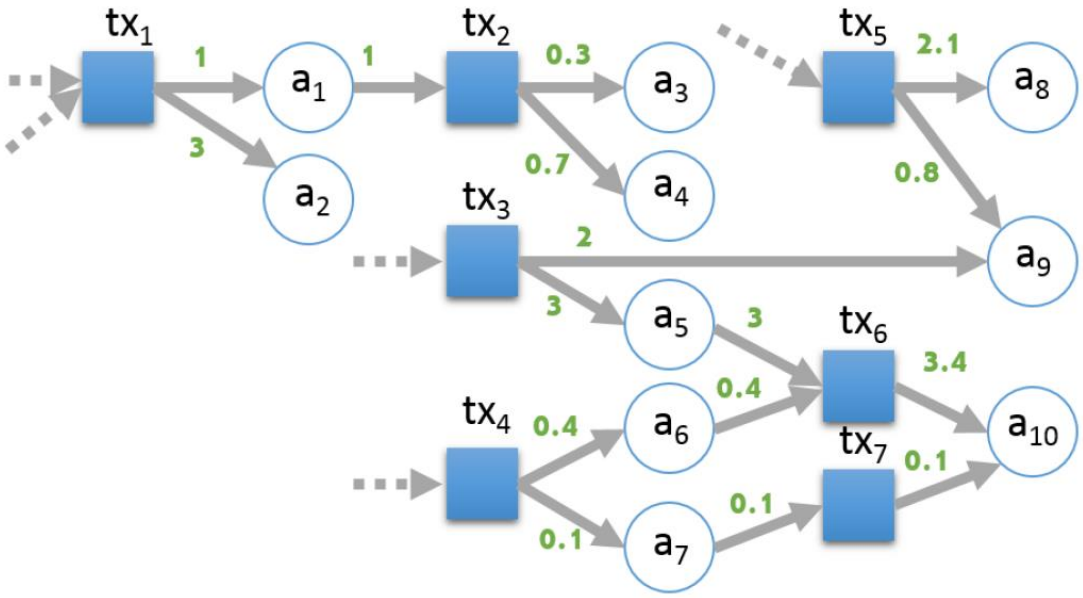


Figure 4 Example Graph of Bitcoin Transactions

In this example the blue squares are the transactions with tx1, tx3, tx4, and tx5 being starter addresses because those transactions are new to this time window, and they presumably either were created by a victim or started a previous time window. The blue circles are various addresses so we can see the connection between tx4 and a10 as a looped situation because the transaction split and then merged back at a10 later. Likewise, a10 would have 2 neighbors since there are 2 transactions pointed at that address. This is a simplified example, but we can see the level of complexity we will be dealing with as we try to analyze millions of such addresses. Now that we have some sense of the dataset, we can start to examine and visualize this data.

Data Analysis

After loading the dataset, here is what the first 5 rows look like unaltered:

	address	year	day	length	weight	count	looped	neighbors	income	label
0	111K8kZAEnJg245r2cM6y9zgJGHZtJPy6	2017	11	18	0.008333	1	0	2	100050000.0	princetonCerber
1	1123pJv8jzeFQaCV4w644pzQJzVWay2zcA	2016	132	44	0.000244	1	0	1	100000000.0	princetonLocky
2	112536im7hy6wtKbpH1qYDWtTyMRAcA2p7	2016	246	0	1.000000	1	0	2	200000000.0	princetonCerber
3	1126eDRw2wqSkWosjTCre8cjjQW8sSeWH7	2016	322	72	0.003906	1	0	2	71200000.0	princetonCerber
4	1129TSjKtx65E35GiUo4AYVeyo48twbrGX	2016	238	144	0.072848	456	0	1	200000000.0	princetonLocky

Figure 5 First 5 Rows from BitcoinHeist Dataset

What we see here is the actual address of the entity which is basically the bitcoin user being analyzed. Then we have the year and day of analysis. The features we discussed above, followed by the label which is either the study identifying the address followed by the ransomware family name, or for unidentified addresses this will have a value of 'white'. In the original study the focus was on identifying not just ransomware addresses but attempting to categorize by ransomware families, but in my case since I'm more interested in trying to visualize the distinctions they are looking for, I felt that it made more sense to reorganize the data based on all ransomware, versus unknown addresses. To that end the data was divided up into 3 new dataframes with all the ransomware address labels changed to 'black'. One dataframe was just the 'black' addresses, one was just 'white' addresses, and one was a concatenation of both, so basically a copy of the original set with the ransomware addresses changed to 'black'. One important note is that in the dataset we have a total of 2,916,697 rows and of those 2,875,284 are listed as 'white' and only 41,413 are listed as known ransomware addresses. So, the vast majority of the dataset is 'white' which means we simply don't know the nature of those addresses, certainly some are ransomware or other illicit addresses but we don't know anything about what they are. The first visual generated to try to get an overall sense of the data with the following scatterplot matrix. Due to the size limitations with this type of plot it seemed appropriate to narrow down to 4 features as follows:

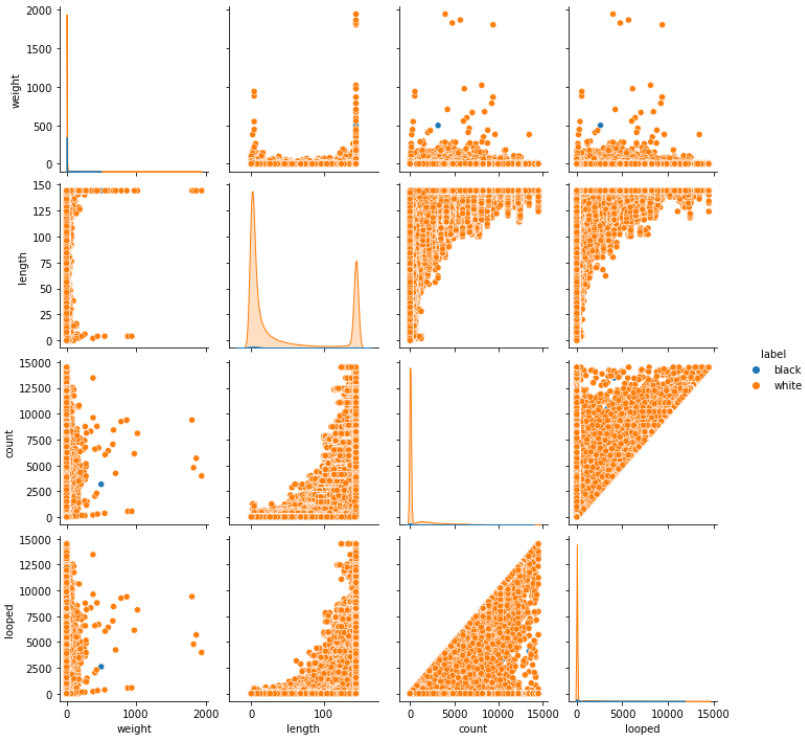


Figure 6 Scatterplot Matrix of 4 features from BitcoinHeist Dataset

Here we can see clearly that a scatter plot isn't going to tell us much about the comparison between black and white as there's just too many white datapoints to see the black data. And knowing that there are almost 3 million datapoints offers a sense of just how skewed some of the categories are since they appear as a line of points with a few scattered past the baseline. Looking at the density plots on the diagonal give a sense of very left skewed data with the vast majority of values at the very low end. For our next attempt let's try using a KDE plot in order to eliminate the occlusion problems and to be able to visualize both groups together. Also, with this one we can add in the 2 additional quantitative columns since space is not as much of an issue.

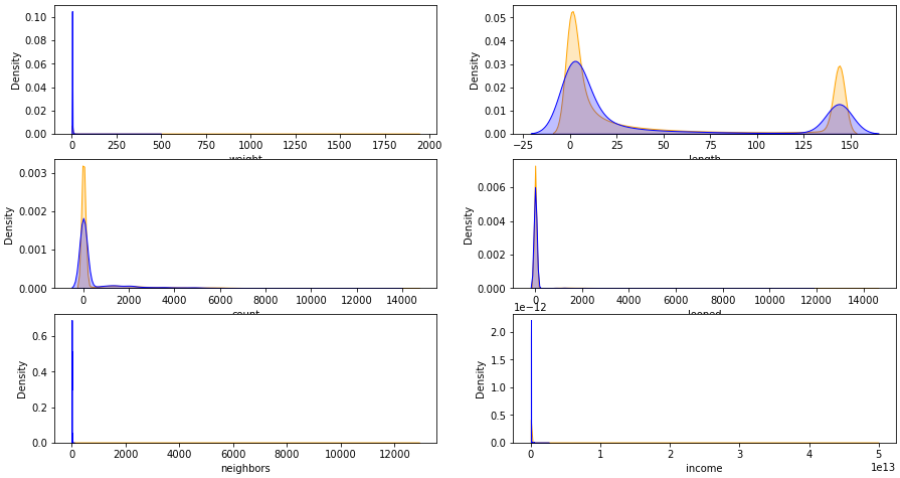


Figure 7 KDE Plot of 6 features form BitcooinHeist Dataset, Orange = "white" addresses, Blue = "black" addresses

The main takeaway from this KDE plot is that most of the categories are extremely left skewed. The length category in the upper right shows a dual peak in density which is interesting, indicating that there is a peak in length for the addresses at the very low and very high end. Also, in several dimensions we can see a difference in the density of the peak with white showing higher density values. Next let's look at a boxplot along the same 6 dimensions. This will take a lot less time to create even with such a large dataset.

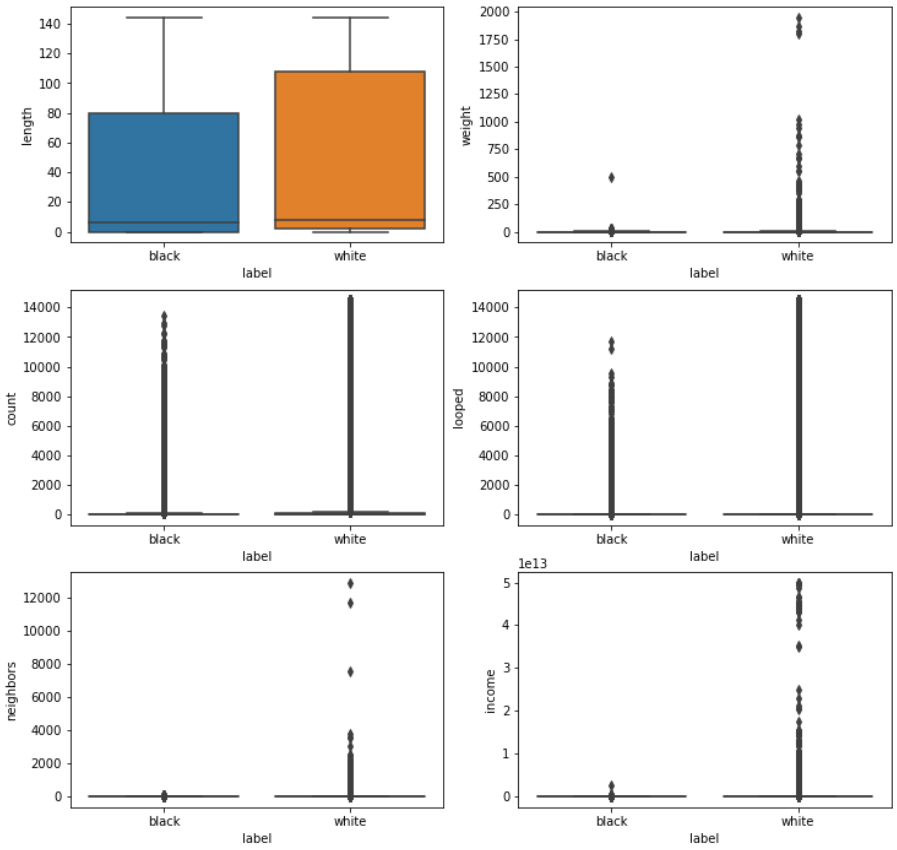


Figure 8 Boxplots of 6 features of BitcoinHeist Dataset

Again, the takeaway is that most of the categories are right skewed with most of the higher datapoints showing as outliers and all 3 quartiles at the very bottom of the plot. Again, we see a wider distribution among white than black addresses but still very left skewed. This dataset looks like it could benefit from a log transformation, but first let's see if the differences in the distribution hold up if we take an equal size sample form each group. With such a large dataset we can get a pretty good snapshot of the data from a sample of 1000 from each group. This will also help us determine whether the differences we were seeing might be due to the vast difference in the size of each group. Let's do another KDE plot with samples of 1000 rows from white and black groups:

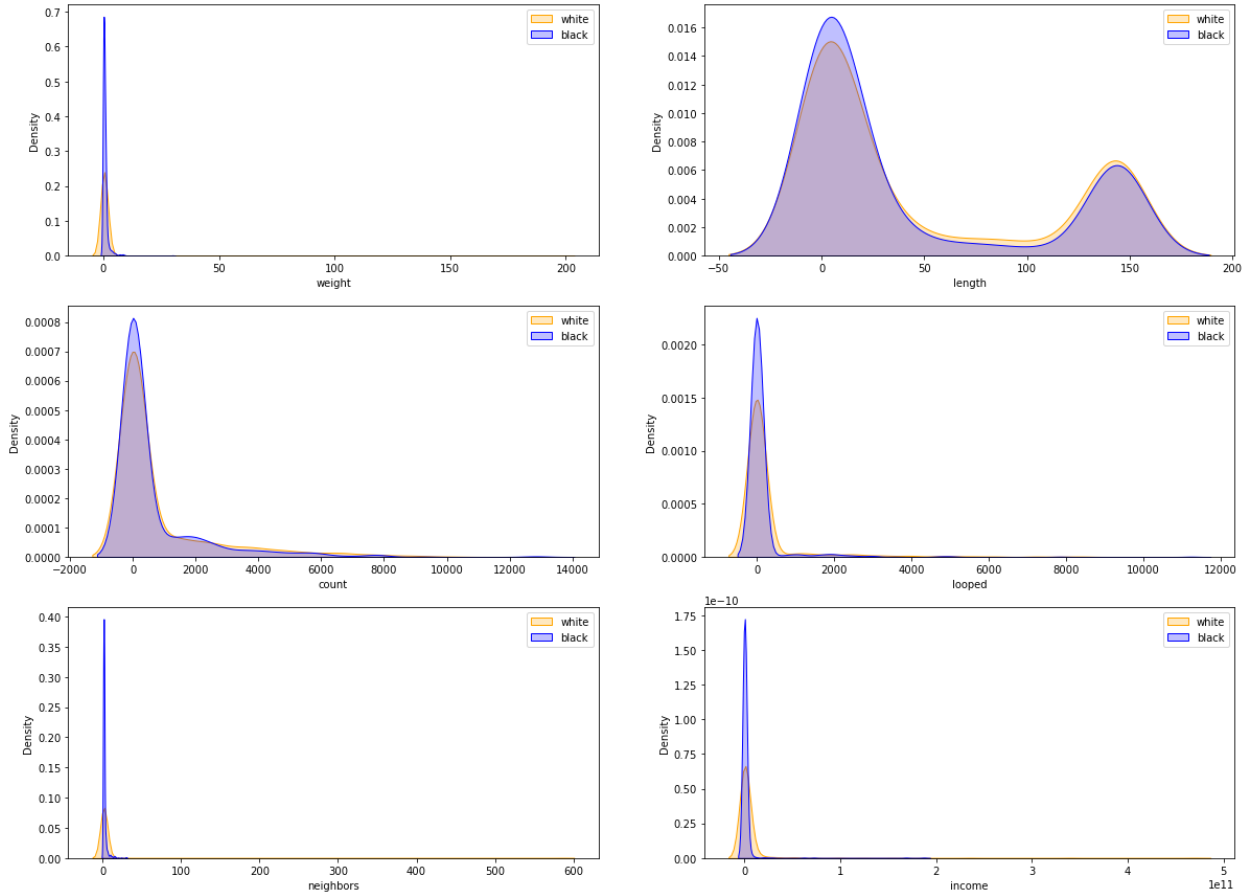


Figure 9 KDE plot of sample of 1000 for each group

After equalizing the size of the 2 groups we can now see that the density differences don't hold up. After repeating this sampling several times, we see a consistent distribution like the above where the black addresses show higher density on the low end in each measure so this may be significant in drawing distinctions between the groups. Next let's look at 4 of the features in a histogram series, we'll use a sample size of 2000 from each group:

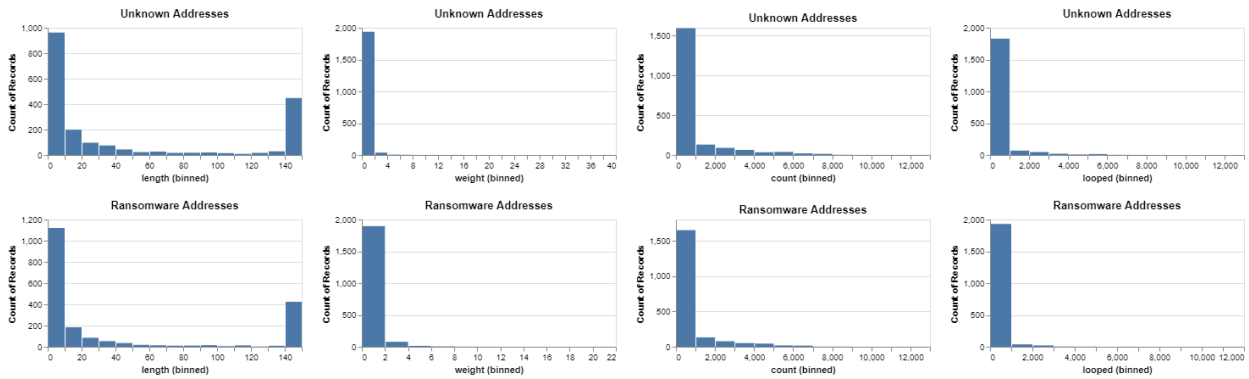


Figure 10 Histograms of groups along 4 features, sample size of 2000 each.

Here we see very similar (and left skewed) distributions but paying attention to the scales we can see that we still have a slightly higher count on the low end with respect to length among the black (Ransomware) addresses.

Now we need to try to understand the distribution better by doing a log transformation. This should get the distribution to look a little more like a normal distribution so we can better understand the overall shape. Again, let's stick with a sample of the data since the large number of rows makes it hard to work with some of the visualizations. For what we are trying to accomplish a reasonably sized sample should allow us to get a close idea of the data.

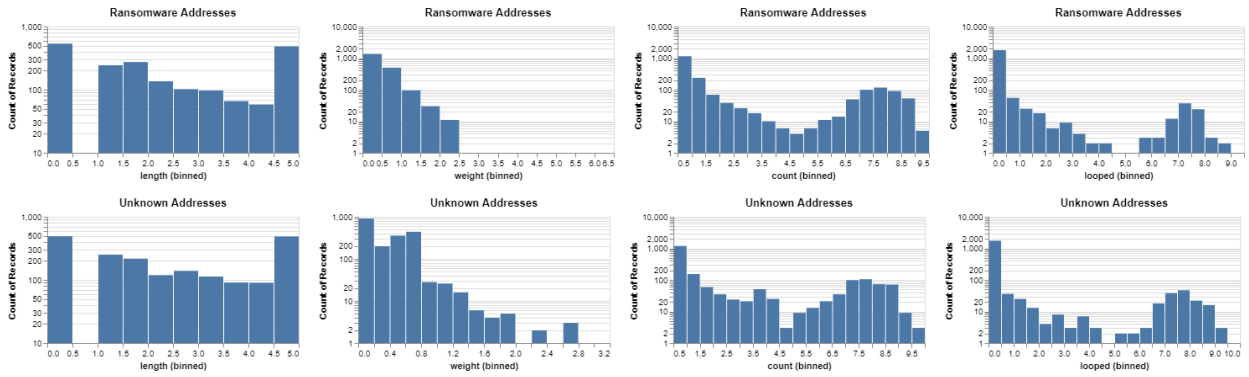


Figure 11 Histograms of log transformed BitcoinHeist Dataset

Using the log transformation does show us a lot more about the data. Overall, one of the most striking features is the similarity of both distributions. The most interesting difference is with the weight where ransomware/black addresses are spread more evenly between 0 and 2.5 where the unknown/white addresses seem more skewed to the left. This is a subtle difference but let's explore this a bit more with a close-up on that dimension:

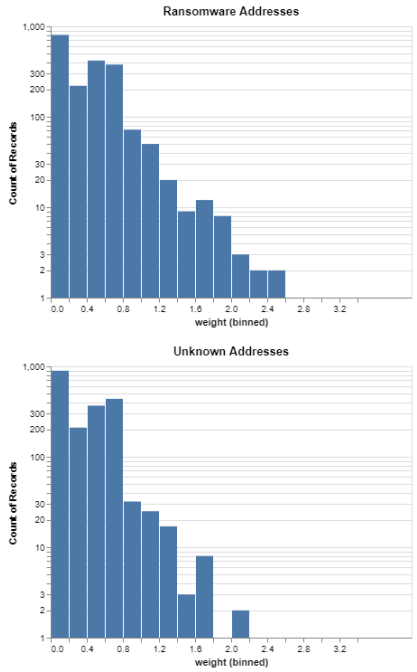


Figure 12 Weight Histogram for Bitcoin Heist Data

After running this visualization many times, we see a consistent pattern similar to above where the density tends to drop off more quickly among the white addresses than with the black. Let's try visualizing this with a jointplot showing the 2-dimensional KDE of our black and white weights:

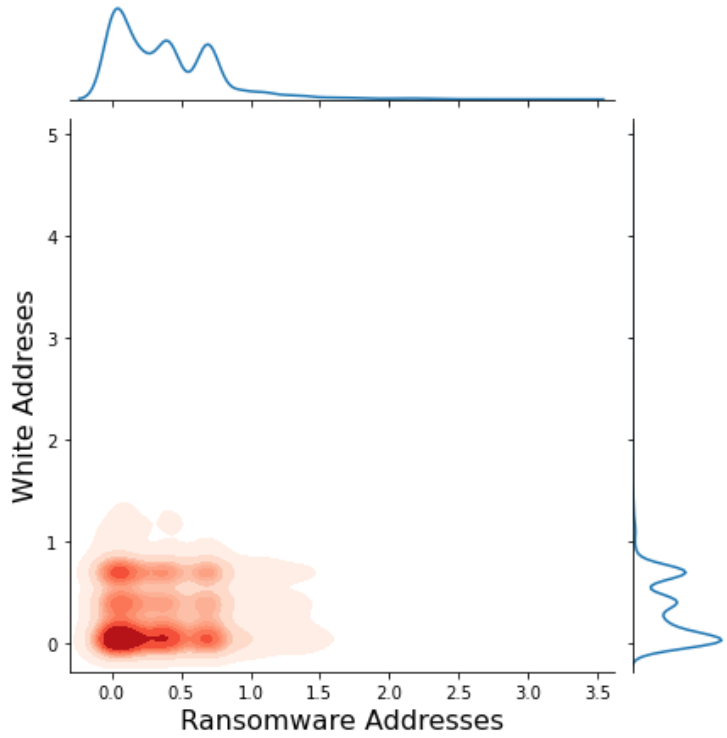


Figure 13 KDE Jointplot of BitcoinHeist Data

This one really expresses the difference in the distribution, we can see that the image shifts clearly to the right showing the more right skewed distribution of the ransomware addresses. This seems to suggest that these addresses are compiling multiple starter transaction coins at a higher level than the white addresses more often. That makes sense considering that the ransomware addresses would want to move the coins around and split them up and then re-combine them in the accounts they control. What we are picking up on here is how there can be subtle differences in our data that we can see when we examine closely but it can be difficult to identify and interpret without extensive analysis. We can manipulate the data and find these kinds of distinctions when looking at one or maybe 2 dimensions at a time, but this is where our analytical software can help. When we have a multidimensional data set like this one, we can use dimensionality reduction methods to summarize the differences in a 2-dimmensional plot. The method we are going to use here is the t-distributed Stochastic Neighbor Embedding or t-SNE. This will look at all the dimensions or our data and distill the differences down to 2 dimensions so that we can plot and show whether there are patterns that distinguish the ransomware addresses form the white addresses.

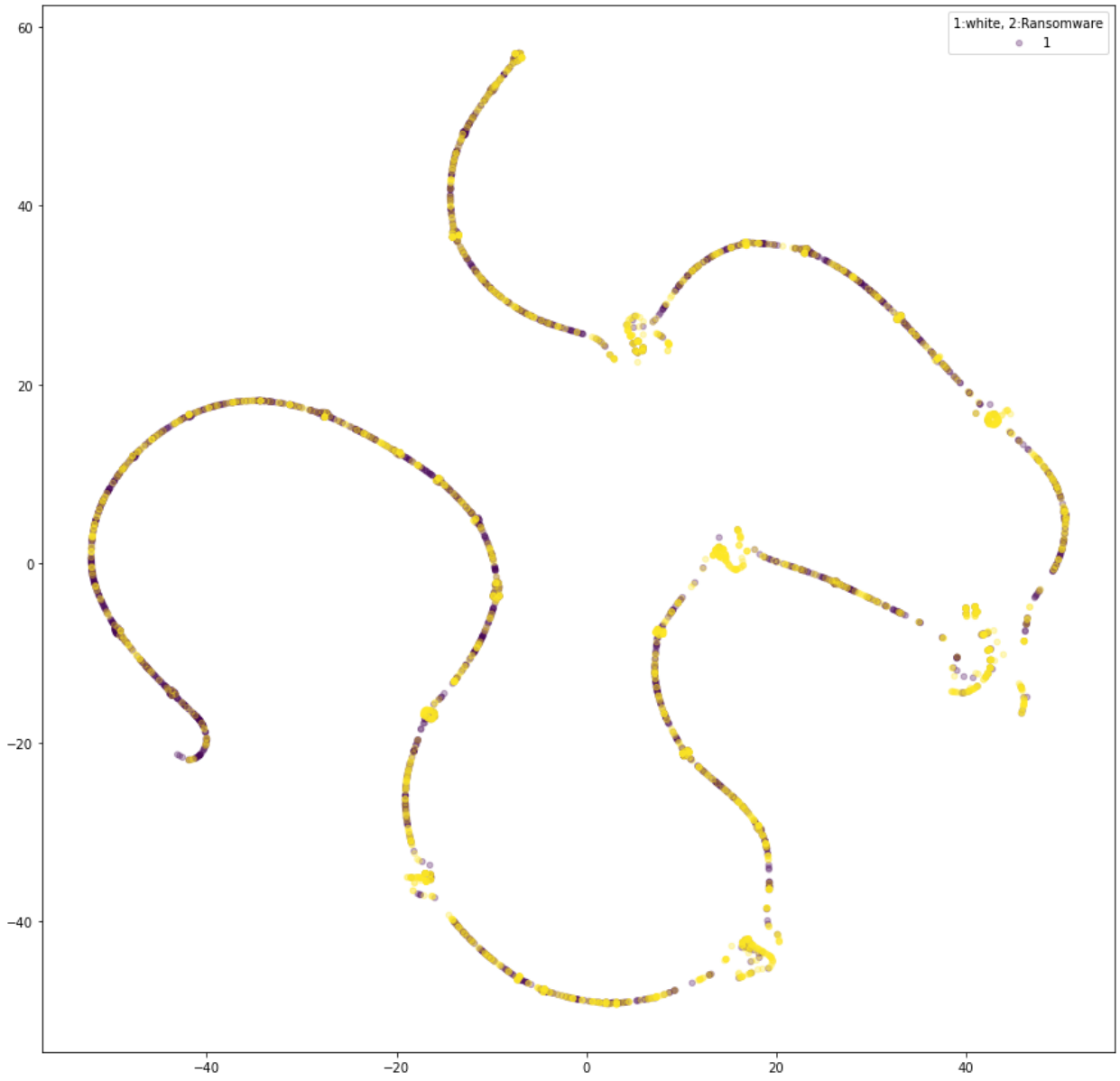


Figure 14 T-SNE Plot for Bitcoin Heist Data

This visualization really brings home the power of programmatic data analysis for recognizing patterns in the data. What we see here is a continuous line of data points where along the line we see clusters form at different places containing mostly yellow points, this means that the program is recognizing the differences in the ransomware address along the 6 dimensions that we included in the analysis. In those clusters we are seeing predominantly yellow points with a few purple mixed in. This is consistent with the idea that the analysis is finding differences in the ransomware addresses so this tool is a good indicator that our data is worth analyzing further to understand these patterns. These patterns can then be leveraged on unknown data to identify new ransomware addresses.

Detection of Elicit Accounts over Ethereum Blockchain-

The next dataset we will explore is similar in some ways to the previous one in that the authors aim to examine the use of data analytics programs to identify nefarious financial activity in cryptocurrency blockchain. This article was featured in Expert Systems with Applications in 2020 and was created by a group of 3 computer scientists from The Netherlands and Malta. The conclusions of their study were that using machine learning on this data would be effective at detecting illicit activity on the Ethereum network at the account level. They specifically advocate for the Extreme Gradient Boosting (XGBoost) machine learning algorithm due to its' ability to deal with missing values in the training data, ability to scale to large datasets, superior performance, and the fact that it ranks the features in order of importance to the analysis. The last feature is valuable for exercises like the one we are undertaking where it makes the process more accessible and understandable in terms of what factors weigh heavier into the decisions it makes. Many of the visualizations included below will focus on the top 5 most valuable features of the data according to the XGBoost analysis done by the study.

About the dataset:

The dataset that accompanies this paper has 4,681 rows each containing data about individual addresses on the Ethereum blockchain. While the overall size is smaller than the previous dataset, the number of features is a bit larger, the set has 50 columns of which 45 contain quantitative data. The data was gathered from available public ledger data on the Ethereum network. The labeling of the accounts was determined by a crowd-sourced scam reporting page called Etherscamdb along with a local Geth client which is a hardware implementation of the Ethereum program run by the study authors. For most of the visualizations below we'll focus on the aforementioned top 5 features. With the last analysis we will show a T-SNE scatterplot with all the quantitative features included and see how a larger number of dimensions effects the ability of the program to make clusters and see the differences in the data. In the paper they included a horizontal bar chart with the top 10 features according to weighting of the algorithm:

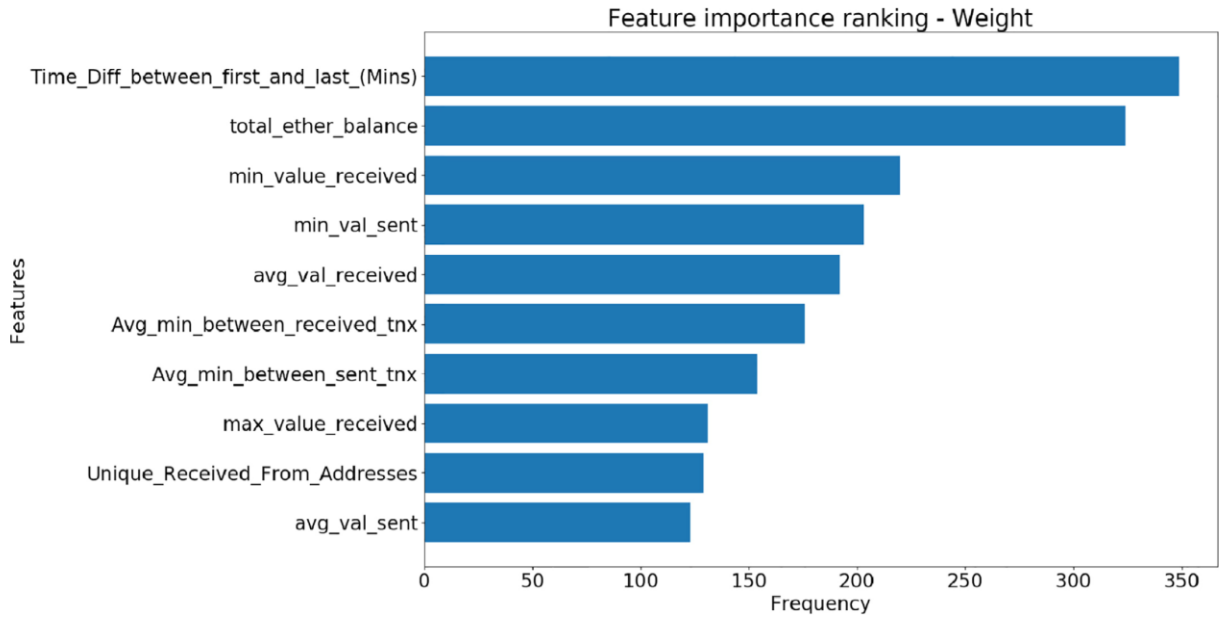


Figure 15 Feature importance ranking for Ethereum illicit account data

Time diff between first and last (mins) interestingly this was the feature with the most predictive value, this apparently shows the time differential between the first and last transactions for the account that were recorded in the data. ***Total Ether balance*** the second most important feature, ***Min value received*** involves the lowest transaction amount received by the address, ***Min value sent***, ***Average Value received***. For the purposes of again comparing the illicit to the normal account we'll use the top 5 to see what differences we can spot if any and then return to our T-SNE analysis with the full data.

Data Analysis

After loading the data from the original CSV file, the data frame was copied with the top 5 features along with the addresses and the flag which was the label for the type of account (0= normal, 1= illicit). This is what the first 5 rows look like:

	Address	FLAG	Time_Diff_between_first_and_last_(Mins)	total_ether_balance	min_value_received	min_val_sent	avg_val_received
0	0x0020731604c882cf7bf8c444be97d17b19ea4316	1	4815.43	0.001037	1.000000	1.000875	1.348445
1	0x002bf459dc58584d58886169ea0e80f3ca95ffaf	1	9622.53	0.001092	0.586269	0.585408	0.766892
2	0x002f0c8119c16d310342d869ca8bf6ace34d9c39	1	321.42	0.000924	0.001020	0.500390	0.439607
3	0x0059b14e35dab1b4eee1e2926c7a5660da66f747	1	73091.00	-0.577721	0.000784	0.000000	0.383225
4	0x005b9f4516f8e640bbe48136901738b323c53b00	1	0.00	0.000000	0.000000	0.000000	0.000000

Next the data was split into a data frame of illicit and a data frame of normal accounts and here is a KDE plot based on the data:

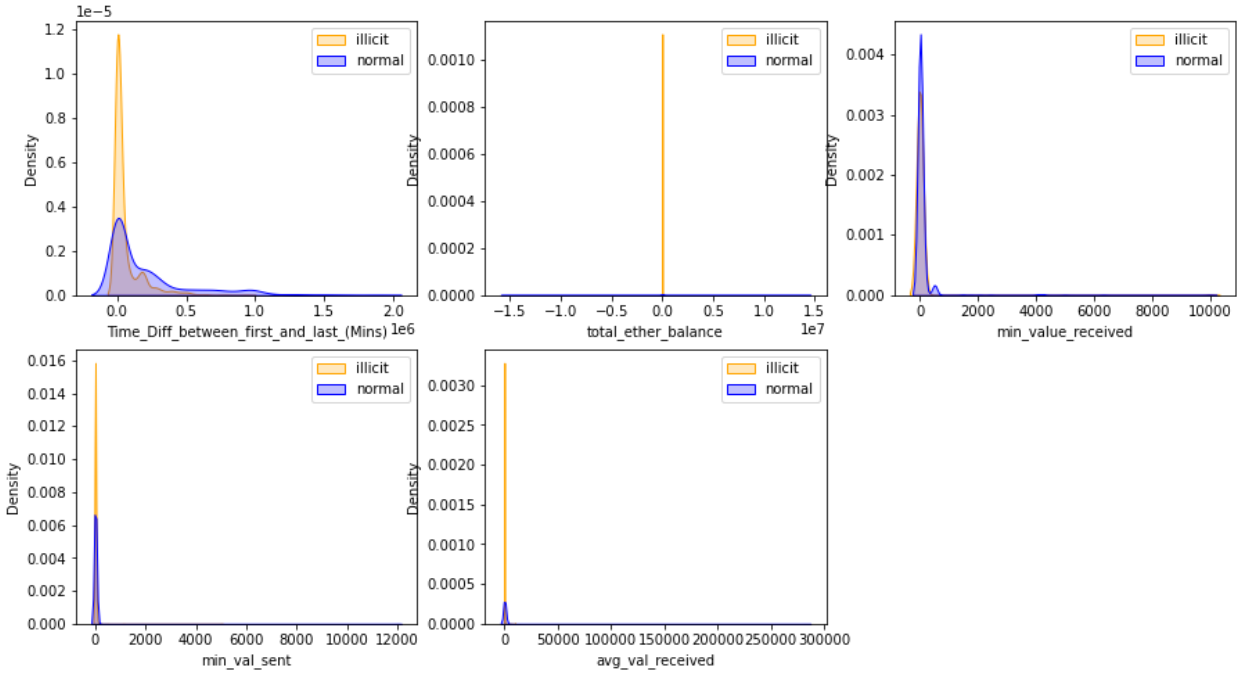


Figure 16 KDE Plot for Ethereum data (Orange = Illicit Account, Blue = Normal Account)

Again, we are seeing mostly left skewed data here. One feature that shows a clear difference in the distribution is the Time Diff feature, here we see that illicit accounts show a higher density in the low value end where the normal accounts are more spread out. Also, it seems like there is more variance with the normal accounts in the total ether balance feature than the illicit accounts but it's hard to tell for sure if that is what it shows. Let's take a look at a boxplot of each feature within each group:

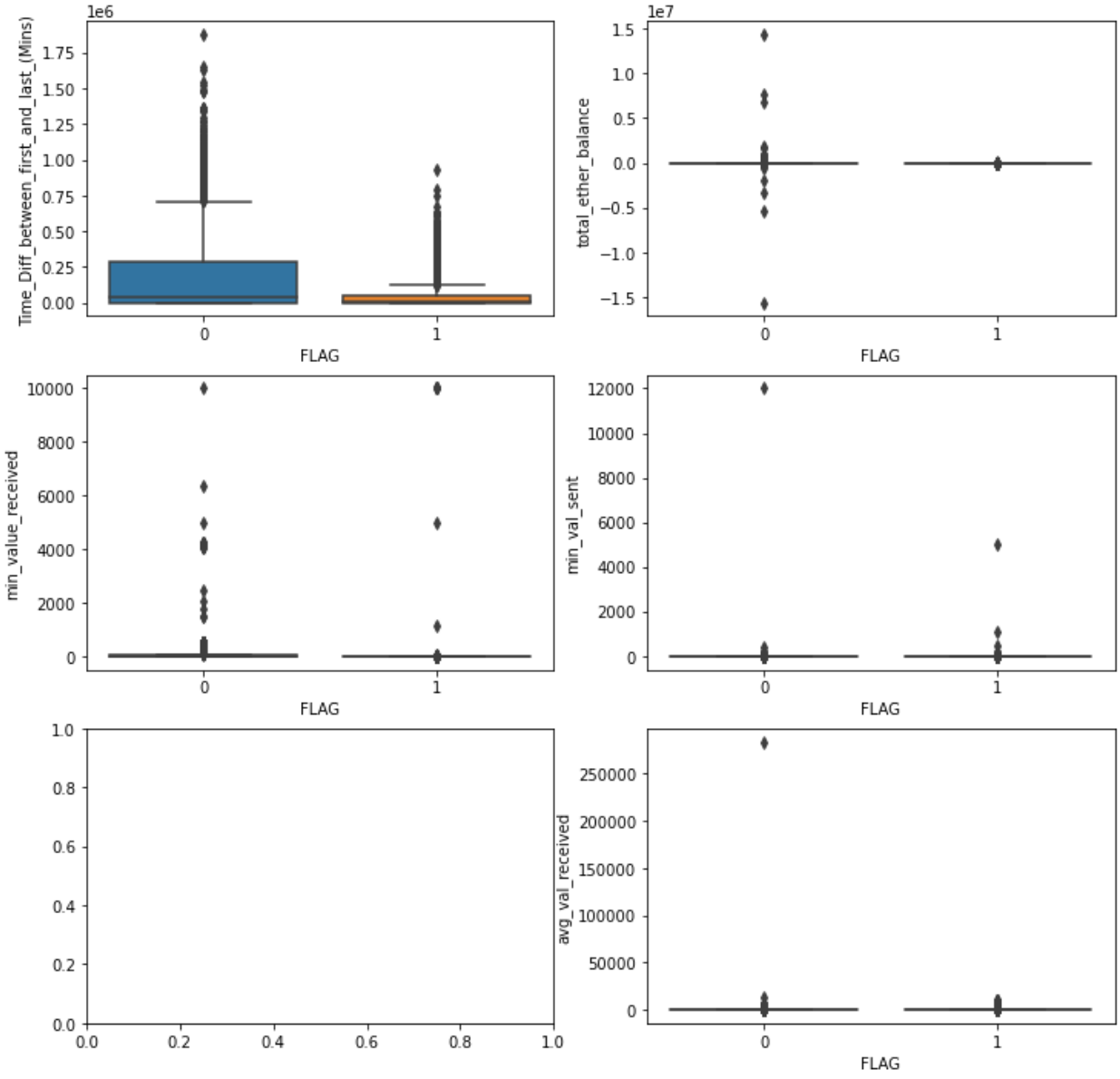
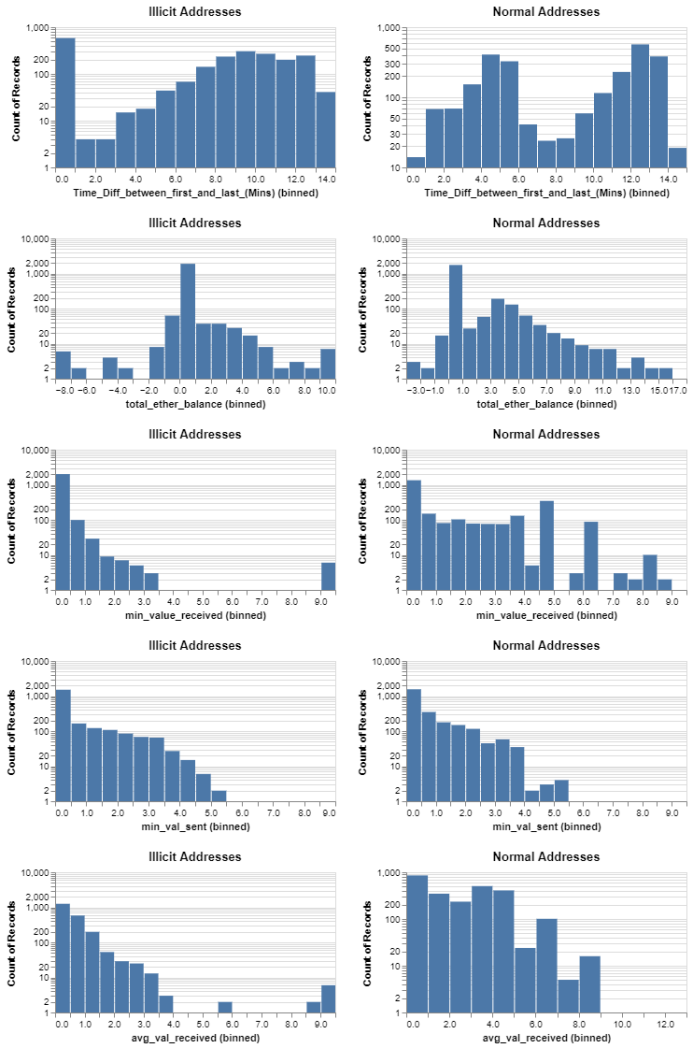


Figure 17 Boxplot of Ethereum Top 5 Features (1= normal, 0= illicit)

Again, most of these plots don't show us much due to the nature of the data but with the time diff feature we can see that the 0 flag for normal accounts is more spread, while the 1 flag for illicit account has a high density at the low end and less spread in the quartiles. It looks like this data would also benefit from a log transformation. Here is a histogram matrix with vertical columns showing each feature listed under our two groups after log transformation:



This is where log transformation really helps as it spreads the data out enough that we can now see clear distinctions in the distributions of the features. In the time diff feature we can now see a clearly much higher distribution in the 0.0 to 0.5 bin with probably indicates that a lot of the illicit accounts are showing 0 minutes between the first and last transactions. This makes sense when you think about what they are trying to do by passing the ill-gotten fund around quickly to multiple accounts in order to try to hide the source. Along the total ether balance dimension, we see that both types of accounts have a high density around 0 but with the normal accounts it's more evenly distributed where the illicit accounts seem to be clustered more at the low end. This is similar to what we see under minimum value received and average value received where normal accounts are spread more and illicit accounts are concentrated more in the low end. Let's focus in on the time diff dimension:

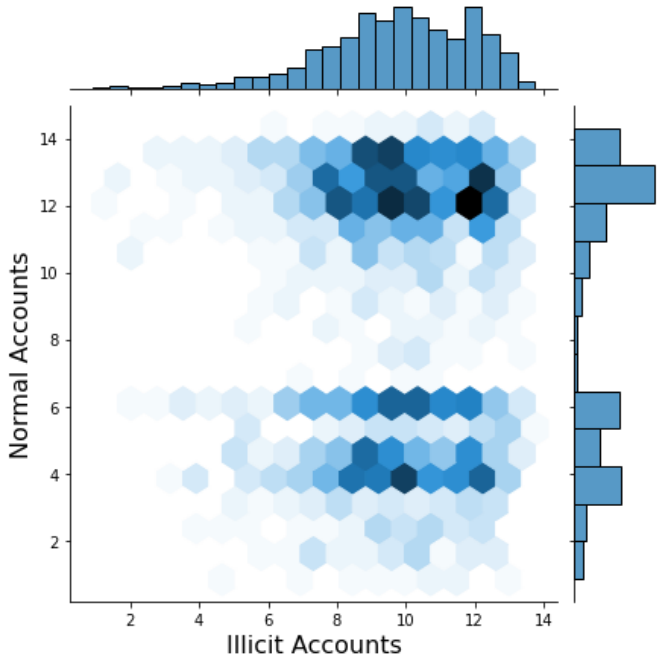
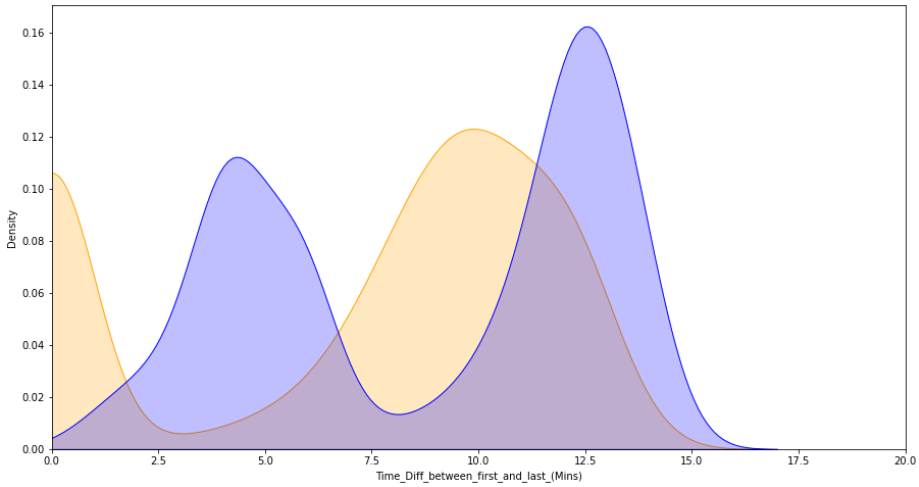


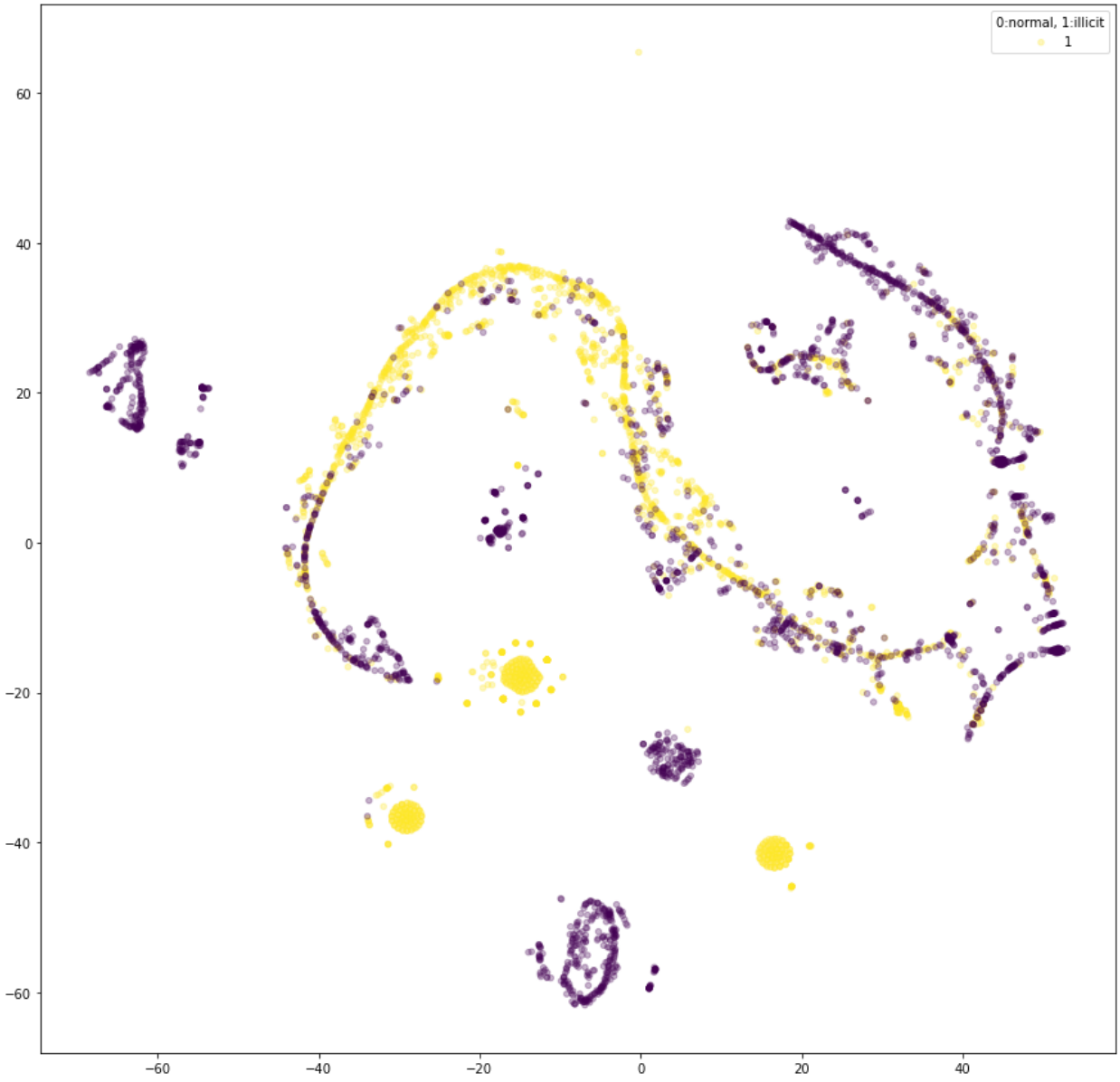
Figure 18 Hex style Jointplot of log transformed time diffs.

Here we've plotted the log transformed time difference dimensions of the normal against the illicit accounts. We had to remove zeros in order to avoid the zero division error so we should keep that in mind. Overall, we can clearly see the bi-modal distribution of the normal account weights but a more right skewed distribution for the illicit accounts. We can tell that these groups are fundamentally different on the time difference dimension. Leaving out the zeros is probably a significant issue with this plot so let's try something else.



With this KDE plot we can see the peak among illicit accounts at zero which tells us a lot about what they are trying to do. It seems like the short time difference between transactions may indicate some sort of automated transaction that is happening, or it may just be that the illicit account are being used and then dispatched quickly as the owner wants to move the funds around and try to mask the origins of the funds. This is consistent with money laundering where the illicit account owner is trying to hide

where the funds are ending up. Now we'll run another t-SNE analysis and plot the data on a scatterplot to see if our program can find some clusters or patterns in this data. This time we'll feed all 45 quantitative dimensions into the analysis.



With this pattern we have another interesting image. This time we have 3 very clear clusters with our yellow illicit address and then 3 messier but very clear clusters of normal address. We also see quite a bit of separation on the linear pattern in the middle with some sections mostly yellow and some mostly purple. What this shows us is that we certainly have distinct patterns within the data. This type of analysis can be helpful in the initial analysis of the data to determine if there are patterns here that may lend themselves to the use of machine learning or other analytic systems to make decisions and classifications in the data. This is the type of pattern you would want to see to indicate that these further analyses would be fruitful, and it helps us to recognize the utility of the proposed mechanism for in this case detecting illicit activity on the Ethereum network.

Discussion-

What this project makes clear is that blockchain transaction data is difficult to analyze in traditional ways using scatterplots or histograms. The data tends to have a uniform look regardless of the addresses, which is a consequence of the fact that ultimately, we are trying to quantify a large number of simple transactions. It's through the meta-analysis of how the transactions are connected, where the funds end up, how the accounts behave, and similar small quantiles of information that we are able to start to put together a picture of what the accounts are trying to do. And the sum of all the differences between the illicit accounts of ransomware gangs, cryptocurrency scammers, and money launderers amounts to a small amount of distinction but under advanced analytic techniques, we find patterns. I think that what is shown in this work is that if you work hard at trying to visualize those small distinctions and pay close attention to the results you can see these patterns. However human beings are not great at picking up on small distinctions spread over multidimensional datasets. We are great at picking out patterns in three-dimensional visual data but anything beyond that can be much more difficult to identify. This is where dimensional reduction techniques can help to show us in images that we can understand, the fact that these patterns exist. The next step is to find tools that can explore those patterns and take action based on them.

Future Work-

Within the two main papers analyzed here there was a direct conflict in analysis method recommended by each group. In the Bitcoin Heist paper, the authors directly said that the TDA method used was more effective than XGBoost among other machine learning methods for detecting ransomware families. One reason for this distinction may be that the original data they were working with was in graph form so it makes sense that TDA would be better at graph data analytics. It stands to reason that different tools will work better for different applications and different objectives. The studies we examined were different types of data with different features so the fact that researchers came to differing conclusions should not be surprising. While the goal of this report was to simply try to find good methods for visualizing and better understanding what the data tell us about the illicit funds on cryptocurrency markets, the next step will be to understand the tools used in taking action around these distinctions and part of that process is understanding which tools to use for which purposes. Since my area of focus is on using data analytics for cyber security objectives, I hope to further delve into this topic as these tools may be quite useful not just for cryptocurrencies but also for detecting network intrusions, software deficiencies and processing event log data. Another possible application for TDA would be any Neo4j graph database, with it's growing popularity it is valuable to recognize that these tools should work well together in highly networked data storage and analysis situations similar to those presented here.

Conclusions-

While it is hard to know what the true value of cryptocurrency is for society, it seems safe to say at this point that it's not going away any time soon. Despite the volatility and questions around the massive energy requirements of the decentralized currency system, the growth in popularity since the inception of bitcoin in the aftershock of the great recession of 2008 is undeniable. The idea of such a self-sustaining monetary system would have seemed like pure science fiction 20 years ago but now it is a fact of life and one that enables some destructive and costly activities. The financial and human toll caused by ransomware is real and as with any crime its incumbent upon society to find ways to slow it

down and bring those responsible to justice. While cryptocurrency provides anonymity to criminals and scammers, it also provides some recourse for tracking and detecting their activities. As cryptocurrencies continue to evolve it's possible that developers of crypto technology will enact measures to limit volatility and make the currency more and more attractive to everyday consumers. The topic of government regulation has been more and more prevalent as cryptocurrencies have risen in value and it's likely that regulation will be aimed at the currencies with the goal of reining in illicit activities as well as ensuring that the currencies are not vehicles for tax evasion. As these regulations come into effect it will be important to continue to improve the toolset of analytical programs that data analysts can use to continue to detect illicit transaction as the law and the perpetrators evolve and change.

References:

1. Kerner, S. (2021). *Ransomware trends, statistics and facts in 2022.* Tech Target. https://www.techtarget.com/searchsecurity/feature/Ransomware-trends-statistics-and-facts
2. United States Department of Justice. *How to Protect Your Networks from Ransomware.* https://www.justice.gov/criminal-ccips/file/872771/download#:~:text=On%20average,%20more%20than%204,000,risk%20posed%20to%20your%20organization.
3. Grauer, K and Updegrave, H. (February 16, 2021). *The 2021 Crypto Crime Report.* Chainalysis. https://go.chainalysis.com/rs/503-FAP-074/images/Chainalysis-Crypto-Crime-2021.pdf
4. Orcuttarchive, M. (January 30, 2020). *Millions of people fell for crypto-Ponzi schemes in 2019.* MIT Technology Review. https://www.technologyreview.com/2020/01/30/275964/cryptocurrency-ponzi-scams-chainalysis/
5. Fletcher, E. (June 3, 2022). *Reports show scammers cashing in on crypto craze.* United States Federal Trade Commission. https://www.ftc.gov/news-events/data-visualizations/data-spotlight/2022/06/reports-show-scammers-cashing-crypto-craze#crypto4
6. Weiner, S. (July 20, 2021). *The growing threat of ransomware attacks on hospitals.* American Association of Medical Colleges. https://www.aamc.org/news-insights/growing-threat-ransomware-attacks-hospitals
7. Bischoff, P. (June 23, 2022). *Ransomware attacks on US schools and colleges cost $3.56bn in 2021.* Comparitech. https://www.comparitech.com/blog/information-security/school-ransomware-attacks/
8. Wallace, B. (November 23, 2011). *The Rise and Fall of Bitcoin.* Wired. https://web.archive.org/web/20131031043919/http://www.wired.com/magazine/2011/11/mf_bitcoin
9. James. (November 15, 2021). *The History of Bitcoin: A Complete Timeline of the Start of Web3.* History Cooperative. https://historycooperative.org/the-history-of-bitcoin/
10. Coinloan. (2020). *The Complete History of Ethereum.* https://coinloan.io/article/the-complete-history-of-ethereum-eth/
11. 101 Blockchains. (June 8, 2020). *Ethereum Smart Contracts Ultimate Guide.* https://101blockchains.com/ethereum-smart-contracts/.

12. Gurcan Akcora, C. Li, Y. Gel, Y. Kantarcioglu, M. (June 19, 2019). *BitcoinHeist: Topological Data Analysis for Ransomware Detection on the Bitcoin Blockchain.* Cornell University arXivLabs. https://arxiv.org/abs/1906.07852
13. Talebi, S. (May 21, 2022). *Topological Data Analysis (TDA) A less mathematical introduction.* Toward Data Science. https://towardsdatascience.com/topological-data-analysis-tda-b7f9b770c951
14. Farrugia, S. Ellul, J. Azzopardi, G. (February 17, 2020). *Detection of illicit accounts over the Ethereum blockchain.* Expert Systems with Applications, Vol 150 113318. https://www.sciencedirect.com/science/article/abs/pii/S0957417420301433
15. Detection of Illicit Transactions Over Ethereum Blockchain dataset: https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/GKAQYN
16. BitcoinHeist Ransomware Dataset: https://www.kaggle.com/datasets/sapere0/bitcoinheist-ransomware-dataset