Categorizing and Modeling Cryptocurrency Addresses

Daniel Kim

Yale University

Senior Project

Advised by: Maximilian Schäfer, PhD

May 12, 2023

Abstract

With the increasing protection of user data, cryptocurrency can be a solution to avoid bleeding personally identifiable information to third-party companies that use this information to advertise and profile individuals. If cryptocurrency becomes a significant portion of financial transactions, then finding metrics to categorize these wallets or addresses becomes of the utmost importance in catering to the wants of individuals as well as maintaining profits. Looking at transactional behaviors can be a metric by which we can differentiate individuals without compromising the integrity of personal data.

The transactional data for cryptocurrency is publicly available. We will use the exponential moving average model (EMA) and autoregressive integrated moving average model (ARIMA) to forecast values. Which of these models works better and with what types of addresses? This project aims to look at 6 characteristics of address transactions: shares of non-positive transactions, mean value, median value, max value, periodicity, and max transaction quantity. Looking at these characteristics we use a k-means clustering algorithm in order to ascertain whether or not these metrics have any bearing on the quality of the model used. The purpose of this project is to explore potential models and categorize address behaviors and the qualities that make certain addresses more desirable for data-mining parties.

Introduction

To negate the advent of predatory data collection, user protection laws have been passed in the interest of protecting consumer privacy. However, large technology companies such as Google and Facebook source a majority of their revenue from data collection and seek to find newer methods to create efficient advertising campaigns.[1][2] With Google owning about 70% of all

[1] https://www.cnbc.com/2021/05/18/how-does-google-make-money-advertising-business-breakdown-.html
[2] https://www.techadvisor.com/article/745709/this-is-how-much-money-facebook-earns-from-your-data-each-year.html

credit card transaction information in America, sold by companies like VISA, the implications of using user data across their platforms and products can feel predatory.[3][4] Companies want to analyze consumer behavior in order to cater to and curate the best products and services for them. However, problems arise when data such as the types of purchases in transaction histories are used to serve personalized advertisements across various platforms. The advertising process feels predatory due to the nature of asymmetrical information sharing and the feeling of being tracked and profiled across sites.

Profiling and advertising based on purchasing habits and preferences are not inherently bad as consumers want to be sold commodities that they want. The main concern is that throughout the entire process, personally identifiable information is the metric to determine an individual consumer. Using availability bias, the advertisements given to consumers on the internet need to be as accurate and enticing for consumers as possible. Their solution is to use data points that are sourced from personally identifiable information such as one's age, race, ethnicity, religion, financial status, etc. It may be unsettling to be advertised extremely specific products for personal circumstances such as in one case where a woman was profiled as being pregnant through the advertisements she received before she knew she was pregnant herself.[5][6]

Removing personally identifiable information from the advertiser's consumer profiles would be a detriment to the accuracy of the advertisements. Advertisers would be unable to profile consumers into groups and only be able to cater to the preferences of the majority. Cryptocurrency wallets introduce pseudo-anonymous identifiable information that can allow consumers to feel safe by creating a barrier between their personal and public presence while still

---

[3] https://www.washingtonpost.com/news/the-switch/wp/2017/05/23/google-now-knows-when-you-are-at-a-cash-register-and-how-much-you-are-spending/

[4] https://www.forbes.com/sites/petercohan/2018/07/22/mastercard-amex-and-envestnet-profit-from-400m-business-of-selling-transaction-data/?sh=7d50ff937722

[5] https://pdfs.semanticscholar.org/2fba/08e4dfc4a341a83e9c5cb2ece24522dc2bb6.pdf

[6] https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=491e7e216668

maintaining accuracy. The wallet address would act as a name without revealing any personal information and only the receiver would know what was purchased. Therefore, advertisers will still have an edge and be able to use this metric in their profiles. Additionally, advertisers who do not have transaction data on their consumers now have access to the public-facing data.

Although integrating cryptocurrency into our world may prove a challenge, data security and privacy are headed toward constant scrutinization. Personal information is valued and the means to maintain it as such is being challenged at every benchmark. Under the assumption that our society adopts cryptocurrency as an alternative means for financial transactions, modeling the transactional behavior of these users becomes the substitution for personal information.

This project first aims to model the behavior through autoregressive, moving average, and volatility models in order to understand the accuracy of future transactions. We want to know if certain models, and their relative fits, are significant enough to forecast future spending quantities so that addresses can be profiled. And without using personal data, we create meaningful metrics based on the transactional spending quantities and frequencies. Lastly, clustering and categorization will place similar types of addresses, based on these metrics, together so that the profiling can be generalized to the point without becoming over-personalized.

Data Gathering

Etherscan API

The blockchain for any given cryptocurrency is publicly available, but traversing and collecting meaningful data relies heavily on the APIs. The limiting factor for choosing which cryptocurrency to use was the relative ease and availability of the third-party APIs. Ethereum

was chosen as the crypto due to the APIs provided by Etherscan, a block explorer and analytics platform for Ethereum.

Blockchain Exploration

In order to model the behavior of an address, we want to know when the address is sending a transaction. Etherscan provides an API that returns a list of addresses in a given transaction block. After filtering out duplicates, we can use another API that allows us to receive the transaction history. In order to find significant and meaningful addresses, we filter the transactions by the quantity of sending addresses. We look for addresses that have at least 100 sending addresses and store the transaction. The maximum number of transactions for a given address was 10,000 due to the restrictions on the API and to conserve storage space without losing significant data. The total number of addresses which fulfilled this requirement was 13,000.

We took 1,300 of these addresses at random and collected the following values based on their total transactions: share of zeroes, max value, mean value, median value, periodicity, and max entries. The share of zeroes refers to the portion of the total transactions that sent zero Ethereum. The max, mean, and median values correspond to their respective statistical measurements of Ethereum. The periodicity is given as the average number of zero transactions between two non-zero transactions. Lastly, the max entry refers to the total number of transactions that were recorded, zero and non-zero. All of the values are added by one and then logged in order to avoid negative predictions on the values. The log transformation was applied because the proportions become more stationary and looking at the results of proportions yields more meaningful results than the absolute differences.

Time Series Modeling

The models that were chosen to forecast the sending value were the exponential moving average (EMA) and the autoregressive integrated moving average model (ARIMA). Both models use a form of a moving average model which uses previous errors as predictors for future forecasts. The EMA model places heavier weights on more recent data whereas the ARIMA model implements the order of regression and also differences the data in order to make it stationary. The ARIMA model follows the format of *(p,d,q)* where *p* is the order of the autoregression, *d* is the degree of the differencing, and *q* is the order of the moving average. The main difference between the two models is that the exponential model does not need the data to be stationary while the ARIMA model will differentiate the data in order to create a stationary dataset. The corresponding packages used in R were the *forecast, tseries,* and, *pracma* packages.

The models chosen to forecast the volatility of the data were the generalized autoregressive conditional heteroskedasticity (GARCH) model and the GJR-GARCH model. The GARCH model can capture volatility clustering such that shocks to the value in previous periods affect future variance. The GJR-GARCH model has an additional leveraging effect where negative shocks have a stronger impact on the variance than positive shocks. The leveraging effect will help us determine if the volatility is impacted greater by a drop in value or an increase in value. The corresponding package used was the *rugarch* package which holds the GARCH and GJR-GARCH models.

Descriptives

Table 1: Statistics of Address Characteristics

	Min	Max	Mean	Median	Std
Share Zeroes	0	0.9997	0.5312143561	0.5154884565	0.18594976
Max Value	0.0006724762174	125999.9373	280.8979206	3.714851223	4384.13938
Mean Value	0.0000005006917028	649.8879142	2.570768326	0.09712002212	29.46869298
Median Value	0	368.7943474	0.3500080184	0	10.29705216
Periodicity	0	3332.333333	7.339082341	1.496674786	100.875267
Max Entries	103	10000	1568.525385	632	2361.678688

Table 2: Entries per Cluster

Cluster	Entries
Cluster 1	860
Cluster 2	319
Cluster 3	2
Cluster 4	112
Cluster 5	1
Cluster 6	1
Cluster 7	5

Table 3: Mean Squared Errors for EMA

EMA	Min MSE	Max MSE	Mean MSE	Median MSE	Std MSE
Cluster 1	0.00000001546701856	3.351430348	0.07898139089	0.01705777394	0.2128139346
Cluster 2	0.00000003083665	3.175862082	0.1513914349	0.03922201866	0.3166016682
Cluster 3	3.338230247	4.70393788	4.021084064	4.021084064	0.9657011285
Cluster 4	0.00000001122846926	1.732266418	0.1077411704	0.02298218946	0.2051360599
Cluster 5	5.054516158	5.054516158	5.054516158	5.054516158	0
Cluster 6	0.001075951196	0.001075951196	0.001075951196	0.001075951196	0
Cluster 7	1.606009484	5.03001529	3.175809932	2.401306922	1.484137768

Table 4: Mean Squared Errors for ARIMA

ARIMA	Min MSE	Max MSE	Mean MSE	Median MSE	Std MSE
Cluster 1	0.000000009911252547	6536.04923	25.32052921	0.05640352035	283.3500667
Cluster 2	0.00000003501401965	35577.21895	268.7238753	0.3236982979	2289.971513
Cluster 3	5917060.601	18171589.88	12044325.24	12044325.24	8665260.752
Cluster 4	5.00E-09	1.08E+05	1.08E+03	3.70E-01	10254.35635
Cluster 5	525180.5531	525180.5531	525180.5531	525180.5531	0
Cluster 6	0.02540801554	0.02540801554	0.02540801554	0.02540801554	0
Cluster 7	134859.9246	805162.3015	442722.8101	480610.9268	255731.0536

Table 5: Mean Squared Errors for GARCH

GARCH	Min MSE	Max MSE	Mean MSE	Median MSE	Std MSE
Cluster 1	0.00000002707943852	3.743274639	0.0902660939	0.01992820098	0.2287955178
Cluster 2	0.00000004022746004	3.421788299	0.1882756672	0.05564045206	0.374572186
Cluster 3	4.034321081	5.560070662	4.797195871	4.797195871	1.078867875
Cluster 4	0.000000006154046705	2.07585741	0.1308162102	0.02769396334	0.2549914046
Cluster 5	4.604261128	4.604261128	4.604261128	4.604261128	0
Cluster 6	0.00139174803	0.00139174803	0.00139174803	0.00139174803	0
Cluster 7	1.871373342	6.061765492	3.791599812	2.956803635	1.827176479

Table 6: Mean Squared Errors for GJR-GARCH

GJR-GARCH	Min MSE	Max MSE	Mean MSE	Median MSE	Std MSE
Cluster 1	0.0000153872788	3.743274629	0.09391247801	0.02140680587	0.2438165691
Cluster 2	0.0000000362363496	3.401520722	0.1849967153	0.0582770103	0.3714521117
Cluster 3	4.033035133	5.560076415	4.796555774	4.796555774	1.079781245
Cluster 4	0.000000006154046873	2.074419851	0.1457828143	0.04232333744	0.2638205939
Cluster 5	4.567445479	4.567445479	4.567445479	4.567445479	0
Cluster 6	0	0	0	0	0
Cluster 7	1.8699148	6.061646492	3.745354456	2.706459739	1.857266128

Table 7: ARIMA p,d,q mode per Cluster

Cluster	p,d,q
Cluster 1	0,0,0
Cluster 2	0,0,0
Cluster 3	0,0,0 or 5,1,0
Cluster 4	5,1,0
Cluster 5	2,0,1
Cluster 6	0,0,0
Cluster 7	2,1,1

Analysis

Data Gathering Process

During the searching and filtering process for addresses, it is important to note that we looked for one-to-many addresses or addresses that had the same origin address, regardless of the recipient. However, there are other types of addresses that we have considered. The many-to-one address is filtered by the quantity of receiving addresses and applying the models is redundant in nature. The main aversion is that it is not very meaningful to predict how much an address will receive as we have no control over how much the address will send. The only quality that it highlights is the potential for a higher-valued transaction, but not necessarily an indicator. The other type of address is the one-to-one address. The fixed destination address is a more meaningful address in the sense that, when forecasting how much an address will send, we do not have to worry about what type of address they are sending it to. However, the largest barrier is the scarcity of these addresses. Either there are not enough entries (less than 100) or the addresses act as middlemen for smart contracts and therefore are not indicative of normal behavior. These types of addresses should be categorized on a case-by-case basis and profiling them with the larger groups may overshadow their utility.

Why are there transactions with zero Ethereum? These transactions could be pre-processing and post-processing actions conducted by smart contracts, transactions for non-fungible tokens (NFTs), swaps for other types of tokens, or genuine transactions. Therefore, trying to arbitrarily filter out whether or not a transaction should be genuine, without losing real transactions, could hurt the integrity of the results. Certain types of these transactions can have a higher order of lagged periods, helping better fit some of the autoregressive models. The zero-value transactions can be used as a metric for categorizing addresses by looking at the

periodicity and proportional share. The data that we are trying to model could also be binary and return results of this nature. This means that the forecasts would be reflective of either a positive or non-positive transaction value instead of actual quantities. Even if the results were interpreted in this way, knowing whether or not the result would be positive or non-positive can be useful in forecasting, especially when these transactions are autoregressive.

The characteristics of the addresses were chosen based on their utility of value. The share of zeroes, the median value, and the mean value can help to distinguish whether or not an address sends non-negative transactions. The addresses with extremely high shares of zeroes oftentimes are smart contract addresses where the only non-negative transactions are at either the beginning or end of their active state. The addresses with extremely low shares of zeroes are often storage-type wallets, acting as banks from large crypto platforms such as Binance.

The median value acts similarly to the share of zeroes with the stark difference that when the value is positive, we can get a better representation of the values of the positive transactions. The mean value can be heavily skewed towards zero if the share of zeroes is high, however, in this situation, the mean would give us a better representation of the values of positive transactions than the median would. The max value shows whether or not an address can be capable of sending high quantities. The periodicity shows the time between a positive transaction and a non-positive transaction. The periodicity of an address can determine whether or not the address is active. High max values can be profiled into addresses that send high quantities infrequently or frequently, both of which are desirable for profiling. The max number of entries can also be useful in differentiating how active an address is.

Each period in the data is comprised of its own address. Although the timestamps of each transaction are given in Unix epoch, or the number of seconds since January 1, 1970, some

transactions are close to each other whereas others are spaced further apart. The models use time series data that require a certain level of equally spaced data, and there are very few models that take advantage of an unequally spaced data set. Transforming the dataset would also be skewed as for some datasets, a heavy amount of interpolation would be required to fill in the empty spaces. Forecasting on interpolated data hurts the integrity of the forecast and is redundant in nature. Therefore, by equally spacing the data by individual entry, we forego our ability to interpret when the forecast will be, but we will know what our forecast will be when the observation happens.

Modeling Process

The EMA model does not require the time series to be stationary. By setting the number of previous observations to 20, we get a substantial amount of previous lags such that the weights are favored toward recent data. Numbers higher than this may overtake datasets with low quantities. The mean squared error (MSE) is returned for all models to determine the fit amongst the dataset.

The ARIMA model differences data into a stationary data set regardless of if the data was stationary or not. After fitting the ARIMA model onto the dataset, we look at the respective autocorrelation and partial autocorrelation functions in order to determine if the dataset is stationary. We then make sure that the dataset has no significant value for an augmented dickey-fuller test, a unit root test for stationarity on large sample sizes. We also conduct a Ljung-Box test in order to check if there are any autocorrelations in the residuals. We reject the null hypothesis that there are autocorrelation degrees of zero; this results in ARIMA models where the order of the autoregression is greater than zero. We can use this test to further filter out whether or not an ARIMA model or a simple MA model would be a better fit for case-by-case

addresses, assuming stationarity. The difference between whether or not a dataset is autocorrelated is the difference between the order of the autoregressive aspect of the ARIMA model. Those that fail to reject the null hypothesis for the Ljung-Box test will have models where the *p* value of the ARIMA model is 0 and we can filter by the order of the ARIMA model used.

The ARIMA model itself compares with over 92 variations of the *p, d,* and *q* values, up to 5 each, to find the lowest Akaike Information Criterion score (AIC). We chose AIC since it is the most optimal in minimizing the mean squared error of the models and because the sample size is much larger than that required for the second-order AIC (AICc).

The GARCH model requires the time series to be stationary because the specifications use an ARMA(1,1) process, the most common order. We can check this by running a unit-root test as well as checking the autocorrelation and partial autocorrelation of the data prior to fitting. If the datasets are not stationary, we can still differentiate the data to make it stationary, however, every address that was selected, rejected the presence of a unit root. Due to the package that was used, there is no guarantee that a model will converge, meaning that some of the addresses, such as those with transactions that are dominantly zero value transactions (> 99%), result in convergence errors. This is not to say that all dominantly zero value transactions result in errors. These transactions were not included in the mean squared error statistics for the GARCH models.

The clustering algorithm that was used was the k-means algorithm which partitions observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. Our dataset has 6 observations: the shares of zero, max value, mean value, median value, periodicity, and max entries. We standardize the dataset in order to reflect normally distributed

data. The alternative is that the clustering will heavily favor values that have high magnitudes which are the max values and max entry variables. We choose the number of clusters to be 7; Figure 1 shows the optimal number of clusters where the inertia plateaus around 5 to 8 clusters.

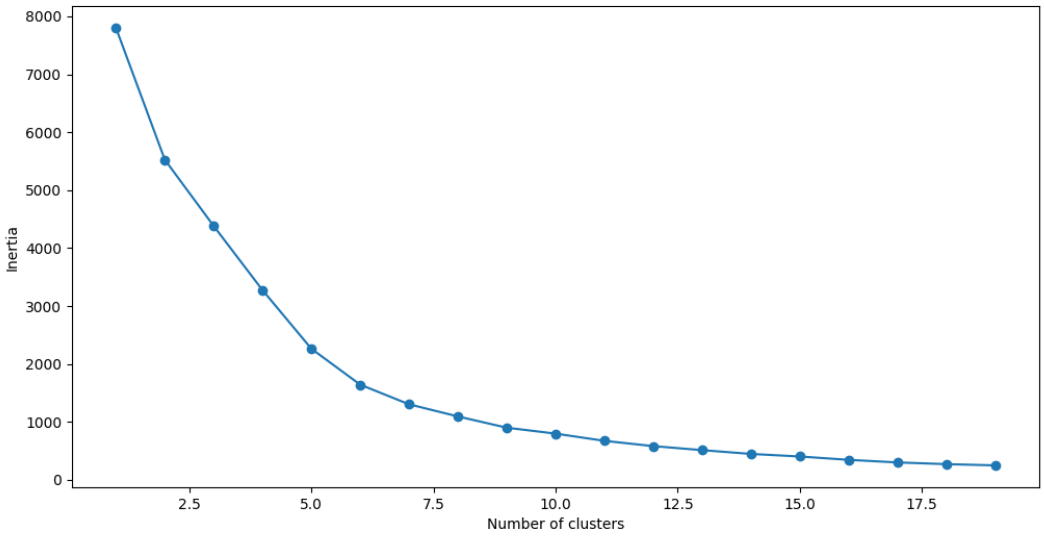

Figure 1

Summary of Results

The first table shows the wide range of characteristics that an address can encompass. It is important to note that the extreme cases for the shares of zeroes and the periodicity will lead to individual clusters. The very high standard deviation for periodicity indicates that a majority of the zero transactions are evenly spaced at low quantities. The minimum of 0 for the shares of zeroes indicates that there are addresses that are trading consistently without the use of smart contracts or other zero-value transactions. On the other hand, the address with a maximum of 0.9997 has characteristics of high frequencies of transactions but with no transaction values, which reflect smart contract addresses.

The results of the clustering algorithm shown in Table 2 do not have an even distribution for the number of entries in each cluster. Taking a look into clusters 3 to 7 we find that there are similarities in extreme values. Cluster 5 is an address with no zero value transactions that also has extremely high transaction values. Specifically, the address is not tied to a company platform which is rare for transaction and frequency values of this magnitude. Cluster 6 is the exact opposite and is an address with shares of zero-value transactions over 99 percent. Cluster 7 addresses all share extremely high maximum values although their frequency and periodicity vary. Cluster 3 addresses contain high-value transactions, even higher than cluster 5 with extremely low shares of zeroes, although their respective periodicity varies. These clusters with very low entries contain significantly extreme characteristics with specific ARIMA models that are not as commonly found in other clusters.

Looking at the clusters that have a larger number of entries, we can see that the *ARIMA(0,0,0)* or the random white noise model is the best-fit model for the majority of these clusters. Therefore, these addresses do not have any meaningful autoregressive component nor do they have any order of moving averages. However, looking at the MSEs from Tables 3 and 4, the average MSE is significantly lower in Cluster 1 than in 2. Although the fit of the *ARIMA(0,0,0)* is the majority of the addresses in these clusters, Cluster 1 has a significantly better fit for random white noise as opposed to the other clusters. The exception is cluster 6 which is a single address with only 0 transaction values and therefore an *ARIMA(0,0,0)* would fit extremely well. There are addresses in Clusters 1 and 2 that beat the MSE for Cluster 6 as indicated by the minimum MSE value. The white noise model showcases that a significant portion of the addresses cannot be modeled with autoregression and moving averages, but nevertheless, random noise is still a behavior of a type of address that can be easily modeled.

Looking at clusters through individual parameters yields substantially worse categorizations. There is not a single ARIMA model that is dominated by a single characteristic. Therefore, if we look at ARIMA models we would expect to see that there are no real clusters or patterns in the models that were selected. However, some ARIMA models are only seen in specific clusters. A little over 80% of *ARIMA(3,0,0)* addresses reside in Cluster 2, 80% of *ARIMA(5,0,0)* addresses reside in Cluster 1, and 84% of *ARIMA(0,1,4)* addresses reside in Cluster 1. There are many more ARIMA model orders that dominate certain clusters, although to what extent the value of these majority model patterns provides is uncertain. However, the emergence of patterns in tandem with low mean squared errors in itself is significant to look further into what other characteristics could model the behavior of the addresses.

The EMA model on average has a better fit for forecasting future transaction values than the ARIMA model. A majority of the addresses exhibit random noise behavior where the autoregressive component of the ARIMA model is not applied. There are some addresses where the ARIMA model is a much better fit than the EMA model and may serve to show that some of the addresses may have a better fit with the ARIMA model regardless of their white noise fit. We cannot directly compare the minimum MSEs of the two models as the address used between both models could be different. The only observation that can be made is that there are addresses where the ARIMA model fits much better than the EMA model despite the EMA model fitting better for a majority of the addresses.

The two GARCH models have insignificant differences in their fits with the regular GARCH model performing marginally better in certain clusters. The purpose of fitting the GARCH models was to ascertain whether or not volatility could be a useful characteristic for addresses. There are other clusters that have significantly better fits for the GARCH models that

similarly follow the behavior of the MSEs of the ARIMA models. The low MSEs for some GARCH models show that the addresses can be modeled and forecasted with volatility. The GARCH models showcase that these addresses can be fit with meaningful accuracy in order to use volatility as a potential characteristic for categorizing addresses.

Conclusion

The arms race between the increasing protection of user data and the ever-persistent data mining conducted by large tech companies proves that the arbitrary line of what data is protected is non-trivial. The potential adaptation of cryptocurrency as a means to protect individuals shifts the problem from individual profiling to a broader categorization of transactional behaviors. Looking at the behaviors of these addresses reveals that there are a plethora of models that can be applied to specific types of behaviors. Accurately predicting when and how much an address is going to spend may be the best indicator for advertisement firms in the absence of personal information. Regardless of significance, it still shows that broader characteristics of behaviors can be meaningfully categorized. Addresses with large quantities and high frequencies could be an indicator of higher socioeconomic statuses as opposed to lower quantities. And generalizations can be made on the types of products or services that higher socioeconomic status individuals would like. The frequency can be used as an indicator of an inclination to spend or make purchases. There can be a generalization that can be made on the level of impulsivity of an individual and perhaps their inclination to purchase large quantities of small valued items.

Volatility, although by itself only explains how great the magnitude of a transaction would be relative to the previous transactions, can be combined with frequency and average values to profile addresses. And because there exist addresses of this nature that can also be

fitted well with the GARCH models, predicting the volatility of an address can showcase an individual's impulsivity or even irrationality. Knowing that some addresses will be more inclined to have high variance in their transactions means that even addresses with low values can have the potential to have high-valued transactions. These types of addresses could be as desirable as addresses with high-valued transactions for advertising firms.

Regardless of the relative value of an address to a third-party company, looking at transactional data to categorize behaviors is not a new concept. The only difference is that the metrics we looked at are not personally identifiable and instead only reflect upon the general qualities of an address. Profiling with broader categories of transactional data can be the compromise between predatory data mining and personal information privacy.