RANKING CRYPTOCURRENCY EXCHANGES BY TRUSTWORTHINESS

by

CARTER PERKINS

A DISSERTATION

Presented to the Department of Computer and Information Sciences
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Bachelor of Science

June 2021

DISSERTATION ABSTRACT

Carter Perkins

Bachelor of Science

Department of Computer and Information Sciences

June 2021

Title: Ranking Cryptocurrency Exchanges by Trustworthiness

As many new traders seek to earn their share in the rapidly emerging cryptocurrency domain, greater reliance is placed on digital currency exchanges to facilitate this significant demand. With malicious users establishing fake exchanges to commit fraudulent crimes, there is a great need to classify the trustworthiness of exchanges. Both research studies and practical applications have aimed to characterize features of credible exchanges, but may not be sufficient to reflect the perception of their trustworthiness. In this thesis, we introduce a metric for evaluating exchanges based on direct user sentiment. We explore the effectiveness of our metric by utilizing machine learning tactics to compare existing ranking lists to observe how well our ranking performs.

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Jun Li for encouraging me to pursue research in the first place and supporting me throughout the construction of this thesis. Second, I would like to thank the University of Oregon Computer and Information Sciences department for giving me the challenges and opportunities to grow as a student. Finally, I want to thank my friends and family for giving me the most fulfilling and exciting four years of my life.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

LIST OF ALGORITHMS

Algorithm Page

CHAPTER I

INTRODUCTION

Cryptocurrency exchanges — broker services between buyers and sellers of digital currency and fiat money — are a critical facilitator of worldwide trading. For example, well known exchanges such as Coinbase have millions of verified users across the globe that have traded hundreds of billions of dollars in cryptographic money. Unlike stock markets that are regulated by central banking systems, exchanges can be established by anyone as digital currencies are fundamentally decentralized and highly insusceptible to government manipulation. Cryptocurrency trades do not need to pass through governmental frameworks to be validated. Only the two parties of a trade — the buyer and the seller — are needed to authorize a transaction. In addition to pairing buyers and sellers, intermediary exchange services simplify trading by maintaining the software needed to operate in the crypto domain. This is a significant appeal to traders as the vast majority are not interested in setting up the required software on their own. With around 10% of Americans claiming to own cryptocurrencies, the growing demand for their services is apparent *Financial Terms Dictionary* (n.d.).

While exchanges market themselves to consumers as the best platform to maximize consumer profits, these advertising practices can easily be manipulative and deceitful for consumers. As exchanges can be created by anyone, new traders turn to news, blogs, and websites to source their rationale for choosing an exchange. The vast majority of these exchange ranking list methodologies are driven by user traffic, total volume traded, and variety of cryptocurrency offerings. However, devious exchanges exploit these methodologies by manipulating their own data to climb higher in the rankings so they can influence new traders to join their platform. In 2019, Bitwise Asset Management (Bitwise) reported to the United States Securities and Exchange Commission (SEC) that 95% of exchange volume for Bitcoin was fake despite being widely cited in prominent newspapers like the Wall Street Journal and the New York Times Fusaro and Hougan (2019). Even more consequential, malicious exchanges can give the illusion that users are trading in a secure environment where in reality they are not. For instance, BitKRX, a defunct exchange, deviously marketed their likeness as an official branch of Korean Exchange which is the most distinguished financial exchange in the South Korea Futures Exchange and South Korea Stock Exchange (KOSDAQ). BitKRX used this similarity to scam traders out of their cryptocurrencies as traders thought they were trading with a well-known corporation Young (2017). Identifying trustworthy exchanges is an imperative step to stopping fraud and cultivating public trust in the cryptocurrency domain.

Ranking exchanges by trustworthiness is a rather unexplored research area with most contributions consisting of theoretical scoring systems and commercial ranking lists (e.g. CoinMarketCap, CoinGecko, etc.) that focus on economic factors. These ranking methodologies are limited as they rely on web traffic information which may not be satisfactory to coincide with exchange reputation as evidenced by fake platforms like BitKRX. Furthermore, these methodologies reason their “trust” scores without utilizing any user opinions which we believe is a significant shortcoming. Trustworthiness is dynamic and fluctuates over time as users are influenced by elements that alter their experience (e.g. security breaches, reliability of the platform, their financial returns, etc.), so it is essential that this be incorporated into any methodology that observes trust.

In our research, we address this issue by incorporating credible user sentiment data directly into a quantitative "trust" score. We collect the user sentiment data from the Revain feedback platform for two reasons: the reviews are relevant as the target audience is specific to the crypto domain, and the reviews are credible as authors are weighted by an authenticity rating. We favor Revain as other review platforms, such as Trustpilot, have controversial histories of containing fake/spam reviews and deleting reviews at the request of businesses Kelion (2021). Revain stores data using blockchain technology which enforces transparency as reviews cannot be deleted or modified by anyone.

From the Revain data, we introduce the main body of work for the thesis: the *trustworthiness metric*. The metric is a quantitative formula that takes user sentiment and a set of existing exchange ranking lists as inputs and returns a numerical score. The score can be sorted in descending order to get the ranking list represented by the metric. On this note, we contribute three items.

First, an objective definition of a trustworthy exchange from which we derive the formula for the trustworthiness metric.

Second, we address the limitations of existing methodologies by including credible user sentiment data as part of a fact-based metric.

Third, we evaluate the metric using machine learning (ML) tactics to determine which existing exchange ranking lists most closely align with the definition of trustworthiness. Based on these experiments we outline the following research questions:

- **RQ1**: How does the trustworthiness metric compare to the existing exchange ranking lists?

- **RQ2:** How well does the metric predict future rankings in the existing ranking lists?

We organize our thesis into the chapters below:

- In the Related Work chapter, we go over research and studies in this area.
- In the Revain Exchange Ranking Methodologies chapter, we discuss the commercial ranking schemes.
- In the Methodology chapter, we provide our definition of trustworthiness and the design choices behind our quantitative metric.
- In the Experimentation chapter, we explore our results and findings from our experiments.
- In the Discussion chapter, we discuss the limitations of our approach and future work.
- In the Conclusion chapter, we summarize our entire thesis.

CHAPTER II

RELATED WORK

In the *Bitwise* report to the *SEC* on exchange volume, they showcased that exchanges with real trading volume exhibit consistent and similar patterns in bid-ask orders with respect to changes in business hours. Trade size histograms show that the vast majority of all trades occur in small quantities of cryptocurrencies with spikes occurring at whole trades sizes (i.e. more common to see trades at 1 Bitcoin (BTC) than 1.2 BTC). On the other hand, *Bitwise* claims suspicious exchanges inconsistently follow these bid-ask order patterns and do not follow the trading volume patterns that trustworthy exchanges share. While this report is extensive in the economic legitimacy, it does not go into detail on consumer perception to well-known and suspicious exchanges Fusaro and Hougan (2019).

In the *Schueffel and Groeneweg* thesis, they address the issue of consumers not knowing which cryptocurrency exchange to select by outlining a multicriteria scoring system for evaluating an exchange. *Schueffel and Groeneweg* categorize the different factors into the following four categories: user experience, fees & costs, trustworthiness, and support. For the trustworthiness category, they consider qualitative factors such as legal and operative jurisdiction, centralization, and reliable "good-faith" efforts. Despite this thesis having empirical test results, the trustworthiness factors are quite subjective Schueffel and Groeneweg (2019).

The *Chainalysis* 2020 Crypto Crime Report contains numerous suggestions for exchanges to implement to reduce scams and improve consumer experience. Among these suggestions include recommending exchanges be responsive and transparent when crypto wallets are hacked (e.g. using social media to broadcast such incidents to the public), KYC practices, increased suspicion of trades that utilize mixers,, and flagging transactions before completion. *Chainalysis* does a good job in identifying patterns in cryptocrimes and coming up with potential solutions for them, but it remains unclear if these contributions are reflective of user sentiment *Crypto Crime Summarized: Scams and Darknet Markets Dominated 2020 by Revenue, But Ransomware Is the Bigger Story* (n.d.).

The *SEC* conducted a study that identified common red flags of fraud from real court cases of Ponzi scheme exchanges. Some commonalities among fraudulent exchanges was promising high returns with no risk on investments, exchanges not formally registered with any regulatory body, difficulty cashing out on cryptocurrencies, and overly consistent returns. The *SEC* alert focuses primarily on the behavior of the schemes themselves, which may not be generalizable as crypto related crimes are often much more subtle than traditional Ponzi schemes SEC Office of Investor Education and Advocacy (n.d.).

CHAPTER III

EXISTING EXCHANGE RANKING METHODOLOGIES

In this chapter, we discuss existing exchange ranking methodologies. These methods are critical for understanding the most popular, commercial solutions for ranking exchanges. Additionally, they are used as the baseline for evaluating the generated exchange ranking list from the trustworthiness metric.

3.1 CoinMarketCap Methodology

CoinMarketCap's ranking methodology is a score based on exchange liquidity – the ease of which cryptocurrencies can be converted to other cryptocurrencies or cash at stable and transparent prices – to help users easily find and understand the best cryptocurrency exchanges. The liquidity score focuses on slippage of orders. For example, high slippage would mean that a buy and sell order was processed at a dramatically different price than expected – an indication of high volatility of the exchange. This slippage is calculated based on the size of the buy or sell order, and the percentage difference between the final price of the order and the mid price of all other buyers and sellers. To avoid placing bias on certain traders, varying order sizes are binned at intervals between $100 and $200,000. Finally, the slippage at each bin is summed together and the final score is normalized from 0 to 1,000 where 1,000 indicates low slippage for orders of the maximum bin and a 0 means high slippage for orders less than the smallest bin C. CMC (n.d.); G. CMC (n.d.); Jay (n.d.).

3.2 CoinGecko Methodology

The CoinGecko ranking system is divided into several components representing a proportion of the final score: liquidity (50%), scale of operations (30%), and Application Programming Interface (API), an intermediary software that allows multiple software applications to communicate, technical coverage (20%). Additionally, estimated cryptocurrency reserves and exchange regulatory compliance are calculated but are a work in progress and not used in direct calculation of the score.

The liquidity computation is based on the following: the normalized-reported volume ratio (NRR), average bid-ask spread, active trading pair ratio (ATR), and trading pair trust score. NRR is defined as the normalized volume of an exchange – as determined by its web traffic, average daily user trading volume (ADUTV), and median ADUTV of the Bitwise ten real volume exchanges (the ten exchanges found not to have falsely reported trading volume and the best indicator of the overall cryptocurrency economy) – over its self-reported volume where a higher value indicates higher likelihood the self-reported volume is true. Average bid-ask spread – the amount an ask price (lowest price a seller will accept) exceeds the bid price (highest price a buyer will accept) of a cryptocurrency on an exchange – is calculated across every trading pair that has been successfully executed in the last hour. Lower average-bid-ask spread indicates that the exchange is relatively liquid. As defined, ATR is a simple proportion of the number of actively traded pairs in the last hour over the number of actively traded pairs in the last 24 hours which indicates the trading activity of an exchange – liquid exchanges have high ATR scores. Finally, the trading pair trust score is a measure of trust for each trading pair in an exchange. Again, we expect liquid exchanges to have a high percentage of trustworthy trading pairs.

Finally, the scale of operations and API technical coverage are computed. Scale of operations is calculated using an exchange's normalized volume percentile (highest to lowest) and its normalized depth percentile which is a measure of supply and demand. Next, the API is graded on the availability of trading information including: tickers data, historical trades, order book, open/high/low/close (OHLC) information, web socket API, API trading, and public documentation CoinGecko (n.d.); Jin (2019); Ong (2019).

3.3 Nomics Methdology

The function of the Nomics exchange transparency rating is to reflect an exchange's compliance to provide an auditable trade history to the public. Exchanges are scored using a letter-grade system where lower grades indicate no / little trade history and the highest grades are reserved for exchanges with great reliability, available information, and data integrity standards *Crypto Market Caps - Prices, All-Time Highs, Charts* (n.d.).

3.4 CryptoWatch Methodology

CryptoWatch's methodology for ranking exchanges is solely based on exchange liquidity. In their model, they define liquidity simply as the sum of all bid-ask orders within 100 points of the best price across each market-pair tracketd on CryptoWatch. Exchanges with higher liquidity will consistently have trades with low slippage whereas lower ranked exchanges will more frequently fail to give traders the best price *Bitcoin (BTC) Live Price Charts, Trading, and Alerts* (n.d.).

3.5 CryptoCompare Methodology

The CryptoCompare exchange ranking methodology is a multimetric scoring system used to evaluate exchanges. Each metric is distributed in a manner such that a single metric cannot dominantly influence the final score compared to the others. The final rank is two-fold. First, a due diligence check is done to extract qualitative information. Secondly, analysis of trades and order book data is used to determine the market quality by measuring factors like cost to trade, liquidity, stability, behavior to sentiment, and “natural” trading behavior. Together, these are aggregated and re-scaled into a 0-100 (100 being the best and 0 being the worst) based score which is the final score of an exchange.

The due diligence check is divided into six categories: geography, legal/regulatory metrics, investment size, company quality, data provision quality, and trade surveillance. In the geography section, CryptoCompare collects the country rating and the cryptocurrency regulatory stringency. For legal and regulatory data, they collect the legal company name, determine if the company or subsidiary exchange is registered as a Money Services Business (MSB), if the company or subsidiary exchange is licensed to operate, if they have Know Your Customer (KYC) or Anti-Money Laundering (AML) practices, if they are a part of a regulatory body, and if they have insurance or proof of reserves against losses. In the investment category, CryptoCompare checks if they are funded by large Venture Capital (VC) firms, large non-crypto companies, or smaller VC firms. The company section records if the company is public or private, the identity of the chief officers (i.e. CEO, CTO, CFO, COO, etc.), educational makeup of the members, years of experience of the members, and how many years the exchange has been around since its inception. In the data provision category, CryptoCompare evaluates the responsiveness and usability of the exchange API by looking at API average response times, querying historical trades, websocket connection, order book API endpoint, and API rate limits. Finally, trade surveillance simply holds whether or not the exchange has a market surveillance system in place *CryptoCompare* (n.d.).

3.6 Messari Methodology

The Messari exchange ranking methodology is ordered by the “real” (i.e. not self-reported) volume of an exchange. This volume is derived from a list of manually selected exchanges that Messari believe have legitimate volume due to their API coverage. For exchanges that were not on this list, Messari approximates their real volume by reviewing liquidity estimates, ratings, and exchange rankings from CoinGecko, CoinMarketCap, CryptoCompare, CryptoWatch, Nomics, and FTX's global volume monitor. Additionally, Messari looks at blockchain transaction data using tools like Chainalysis *Crypto Research, Data, and Tools* (n.d.).

3.7 Summary of Methodologies

While implementations vary, the six existing exchange ranking list methodologies above share two common factors – addressing exchange liquidity and exhibiting similar ranking patterns.

First, previous implementation of these scoring systems did not account for exchanges reporting fake information to manipulate their position on these exchange ranking lists. After the *Bitwise* report, the scoring systems were updated to incorporate transparency of trading history and liquidity metrics to evaluate economic health. Furthermore, a greater emphasis was placed one these scoring system to interface directly with exchanges to verify economic attributes GoodCrypto (n.d.); Kaiko (n.d.-a, n.d.-b).

Second, the existing exchange ranking lists are alike in that the top 10-20% of exchanges are in similar positions but quickly diverge away from each other the lower ranked an exchange is. These are due to the differences in the methodologies but it becomes abundantly clear that these are the top used exchanges.

Additionally, several methodologies define user trustworthiness by measuring web traffic using services like Alexa and SimilarWeb and assume that exchanges with consistently high web traffic scores have users that trust the exchange. While this may be a good indicator of the magnitude of users who trust the exchange or identify top used exchanges, it discriminates against exchanges with smaller user bases.

CHAPTER IV

METHODOLOGY

In this chapter, we outline our main contribution toward the *trustworthiness metric* by first formulating an objective definition of trustworthiness then constructing a mathematical formula that can be applied to measure any exchange's trustworthiness.

Before to establishing a metric to observe trustworthiness, it is essential to clearly define what trustworthy means in the context of cryptocurrency exchanges. This is particularly troublesome as trustworthiness is unique to each individual so constructing an objective measurement is not trivial — some users may trust exchanges that have a professional user interface and experience while others may exclusively use exchanges that have transparent security standards. Trustworthiness is only a reflection of user opinions though factors such as number of users, economic status, or security standards may reflect this as well. We utilize this assumption as an additional factor of a trustworthy exchange. Thus, we define a trustworthy exchange by two core factors: *user sentiment* and the *relative rank* of the exchange on existing exchange ranking lists.

For the user sentiment factor, we use reviews as they each have quantifiable features that we leverage for an objective measurement of trust. Second, the relative rank factor represents how popular an exchange is on existing exchange ranking lists.

4.1 Revain Exchange Rating

The rating of an exchange on the Revain platform is a weighted-average based on three concepts: (1) the recency of the review (i.e. outdated reviews have much less of an impact than recent ones so the ratings support current user sentiment), (2) the user's credibility on the platform, and (3) the popularity of the review which is derived from verified users liking or disliking the review Revain (2020). Equation (1) below shows the weighted average formula where w_{ath} is the vector of the author experience weights, w_{age} is the vector of the review age weights, and w_{pop} is the vector of the review popularity weights. Additionally, x_i represents the 5-star score, n is the total number of reviews, and i indices (subscripts) indicate the ith-review for a factor.

$$\frac{\sum_{i=1}^{n} w_{ath_i} \cdot w_{age_i} \cdot w_{pop_i} \cdot x_i}{\sum_{i=1}^{n} w_{ath_i} \cdot w_{age_i} \cdot w_{pop_i}} \tag{4.1}$$

Author experience is calculated based on four factors: the number of reviews an author has written, the author's karma which is a score representing how popular and well-received an author's reviews are from others, the consistency of how often an author publishes reviews with more consistent writing being favored, and the author's profile having content (e.g. picture, real name, location, biography, etc.) Revain (n.d.). Aggregating these four factors together yields the author level l with $l \in [1, 10]$. The author experience weight is then as follows:

$$w_{ath} = \begin{cases} 4 & \text{if } l = 10 \\ 3 & \text{if } l \geq 8 \\ 2 & \text{if } l \geq 6 \\ 1 & \text{else} \end{cases} \tag{4.2}$$

Review age weights are calculated such that recent reviews have a higher value than older ones. The weights are determined as follows:

$$w_{age} = \begin{cases} 0.2 & \text{older than a year} \\ 0.5 & \text{older than 3 months} \\ 0.8 & \text{older than a month} \\ 1 & \text{else} \end{cases} \tag{4.3}$$

Review popularity weight is calculated from the number of likes and dislikes a review has received (denote as q). For each like/dislike, the weight of the review is increased by 1% which yields the following:

$$w_{pop} = 1 + 0.01q \tag{4.4}$$

4.2 Mathematical Formulation

We introduce our trustworthiness metric — a mathematical formula that represents our trustworthiness definition. An exchange will need to maximize both components to score well and be labeled trustworthy. Additionally, we weight the components by w_0 and w_1 for further fine-tuning. When computing the trustworthiness metric for an exchange, we hold out one of the ranking lists as the *validation list* y and use the remaining lists as the *training set* X. We then compute the trustworthiness score for every exchange using the training set and sort our exchanges descending by trustworthiness score – this yields our list of exchanges ranked by trustworthiness. As the training set and test list are disjoint, we can compare our generated rank list to the test list.

4.2.1 User Sentiment Term. The first component, user sentiment, is solely based on data from the Revain exchange platform. We measure user sentiment by the number of reviews, the Revain exchange rating, and the proportion of high-rated reviews. We give higher scores to exchanges that have more reviews as we are more confident in the information that is reported from Revain. We recognize that exchanges with few reviews on Revain are penalized disproportionately but we believe this is a fair assumption to make. Unknown exchanges have minimal data to work with and differ greatly on existing exchange ranking lists compared to the top exchanges. Further, with only 10 reviews the confidence score is 0.9 which only has a 10% penalty. The confidence score α is interpreted as a percentage where $\alpha \in [0.5, 1.0)$:

$$\alpha = 1 - \frac{1}{N+1} \tag{4.5}$$

Next, the Revain exchange rating is derived from the Revain company rating (4.1) and denoted as z with $z \in [1.0, 5.0]$. We fix this number to the date which we sampled our exchange Revain data. Finally, we consider the proportion of "good" reviews, 4-star and 5-star ratings, with *Laplace smoothing*:

$$\gamma = \frac{R_4 + R_5 + 1}{\sum_{i=1}^{5} R_i} \tag{4.6}$$

where R_i is the number of i-star ratings with $i \in \{1, 2, 3, 4, 5\}$. We use a smoothing factor to account for the case where the number of 4-star and 5-star is zero to prevent a loss of information as the entire user sentiment term would be zero.

4.2.2 Relative Position Term. The second component uses the *training set* lists to measure how highly an exchange is ranked from those lists. As the user sentiment component is completely independent of existing ranking list methodologies, we chose to include a term in our trustworthiness metric that uses all the ranking list methodologies in our metric; our metric is grounded by existing data so comparisons can be made. Since every exchange does not appear on our exchange ranking lists, we incorporate a factor called *reported accuracy* β which is the percent of exchange ranking lists that contain the given exchange. The rationale behind this is well-known and used exchanges will not be penalized as we are more confident in the reported rankings. If an exchange does not appear on one of the ranking lists, it takes the value of the average of the other lists in the training set. As each exchange ranking list has its own methodology, the relative rank differs vastly and introduces noise into this term. The value of the reported accuracy is a proportion over the size of the training data as given below:

$$\beta = \frac{1}{|X|} \cdot \sum_{x \in X} \pi_x(E) \tag{4.7}$$

where $|X|$ is the cardinality of X (i.e. the size of the training set) and $\pi_x(E)$ is a binary function determining if the exchange E is in the exchange rank list x:

$$\pi_x(E) = \begin{cases} 0 & \text{if } E \notin x \\ 1 & \text{if } E \in x \end{cases} \tag{4.8}$$

Then, we score the position of exchange being evaluated across the training rank lists using *Laplace smoothing*:

$$\sum_{x \in X} \frac{\max x - x_E + b(E) + 1}{\max x} \tag{4.9}$$

where x_E is the rank of the exchange E in the rank list x, $\max x$ is the largest numerical rank value (i.e. the lowest ranked exchange), and $b(E)$ is a binary function determining if the exchange E is in the *Bitwise* 10 exchanges with real volume (*Binance*, *BitFinex*, *Kraken*, *BitStamp*, *Coinbase*, *BitFlyer*, *Gemini*, *itBit*, *Bitrex*, and *Poloniex*). We use a smoothing factor to account for the worst ranked exchanges and prevent a loss of information as the entire term would be zero.

4.2.3 Final Formulation.

Putting together the user sentiment and relative position term we get our final formula for the trustworthiness metric:

$$T_X(E) = w_0 \cdot \alpha \cdot z \cdot \gamma + w_1 \cdot \beta \cdot \sum_{x \in X} \frac{\max x - x_E + b(E) + 1}{\max x} \tag{4.10}$$

with bounds $T_X(E) \in [0, w_0 \cdot 5 + w_1 \cdot |X|]$. While the function's bounds vary depending on the weights and size of the training set, we are only concerned with the relative ordering of these numbers to generate our ranked list. We can derive our generated list $\hat{y}$ from training set X to compare to the test list y for every exchange $E \in D$:

$$\hat{y} = \{T_X(E) : E \in D\} \tag{4.11}$$

where $\hat{y}$ is a monotonic increasing sequence (i.e. $\hat{y}_i \geq \hat{y}_{i+1}$).

4.3 Similarity Score

To compare two exchange ranking lists we devise an evaluation called *similarity score*. The motivation for this was need based as metrics like cosine similarity or accuracy are not explainable in this context. The similarity score, on the other hand, tells us how closely aligned the top exchanges are to each other while being order agnostic. The motivation for this is due to the subtle ranking fluctuations between all the existing exchange ranking lists (i.e. we disregard the specific ordering). This ensures that our scores are not tremendously low and are easily interpretable. For each generated list, we define the similarity score for the top n exchanges by the following:

$$s(y, \hat{y}, n) = \frac{|\phi(y, n) \cap \phi(\hat{y}, n)|}{n} \tag{4.12}$$

with $\phi(y, n) = \{k_i : k \in y, i \leq n\}$ being the filtering function that retrieves the top n exchanges and $s(y, \hat{y}, n) \in [0, 1]$. Similarity is measured as the cardinality of the intersection of the top n exchanges of both ranking lists over n. In other words, the similarity score is simply the number of shared exchanges in the top n ranks for y and $\hat{y}$. For example, if y and $\hat{y}$ share 3 exchanges in the top 10 exchanges then $s(y, \hat{y}, 10) = \frac{3}{10}$.

CHAPTER V

EXPERIMENTATION

In this chapter, we outline the experiments we designed for evaluating the trustworthiness metric and answering **RQ1** (how well the trustworthiness metric compares to the existing exchange ranking lists) and **RQ2** (how well the trustworthiness metric predicts future rankings in the existing exchange ranking lists). First, we introduce our dataset, which includes the exchanges we sampled, the features for each exchange, and the features for each review. Second, we explore how well each generated exchange ranking list compares to its corresponding existing list at the top n exchanges. Third, we select one of the ranking lists generated by the trustworthiness metric and compare it to the existing exchange ranking list at two different dates to examine how well the generated list predicts future exchange positions.

5.1 Dataset

For our experiments, we selected the first 100 exchanges across the six existing exchange ranking lists for a total of 131 unique exchanges. We also collected each exchange's rank for the six lists and the exchange's user sentiment data from the Revain feedback platform. For each review, we collected the following seven features: (i) the level of the author, (ii) the number of reviews the author wrote, (iii) the karma of the author, (iv) the star rating (1 to 5 integer scale, inclusive), (v) the date of the review, (vi) the number of people who upvoted the review, and (vii) the number of people that downvoted the review. These features are used in the calculation of Equation 4.1 and Equation 4.10. The Revain data was collected December 2020 and the existing exchange ranking list data was collected in December 2020 and March 2021. Exchanges with missing ranking list

i	Training Lists (X_i)					Test List (y_i)
0	CW	CC	CG	CMC	MS	NM
1	CW	CC	CG	CMC	NM	MS
2	CW	CC	CG	NM	MS	CMC
3	CW	CC	CMC	NM	MS	CG
4	CW	CG	CMC	NM	MS	CC
5	CC	CG	CMC	NM	MS	CW

Table 1. Table of partitions for training and test splits. Five of the training lists are held out and used for the generated list $\hat{y}_i$ which is then compared to y_i. Glossary of acronyms: CoinMarketCap (CMC), CoinGecko (CG), CryptoWatch (CW), CryptoCompare (CC), Nomics (NM) and Messari (MS).

information are filled in and reported by the rank accuracy score in Equation 4.7. Additionally, we calculated the proportion of k-star reviews ($k \in \{1, 2, 3, 4, 5\}$).

5.2 Evaluating Existing Exchange Ranking Lists

To answer **RQ1**, we evaluate the trustworthiness metric against each existing exchange ranking list by the following experiment:

1. We utilize the *cross validation* tactic from ML to generate and evaluate the trustworthiness metric. For the set of the six existing exchange ranking lists sampled in December 2020, we partition it into two disjoint sets of size 5 and 1 respectively. First, the training set X_i is the relative rank factor for Equation 4.10. Second, the validation set y_i is held out to be compared to the generated list from Equation 4.11. In both sets, the subscript i indicates which list is being held out (see Table 1 for the complete breakdown).

2. For each X_i and every exchange E, we compute the exchange's trustworthiness score $T_{X_i}(E)$ defined in Equation 4.10 and sort these numbers in descending order as explained in Equation 4.11. The final generated list from the trustworthiness metric is defined as $\hat{y}_i$. In other words, if we are comparing to existing rank list y_i then our generated list follows from

Rank	Exchange	Trustworthiness Score
1	Binance	7.00
2	Kraken	5.97
3	Bitfinex	5.95
4	Coinbase	5.91
5	Huobi Global	5.89
6	Bitstamp	5.88
7	OKEx	5.84
8	Bittrex	5.82
9	Poloniex	5.73
10	Liquid.com	5.43

Table 2. Top 10 ranked exchanges from the generated list that is compared to CoinMarketCap.

Equation 4.11: $\hat{y}_i = \{T_{X_i}(E) : E \in D\}$ with $X_i = \{y_j : j \neq i\}$, $w_0 = 5$, and $w_1 = 2$. We generate a total of six rank lists — $\hat{y}_0, \hat{y}_1, \hat{y}_2, \hat{y}_3, \hat{y}_4, \hat{y}_5$ — that are compared to the corresponding validation list y_i. The top 10 exchanges for $\hat{y}_3$, for example, can be seen in Table 2.

3. Finally, we compute the similarity between y_i and $\hat{y}_i$ using the similarity score (see Equation 4.12) at 10 step intervals from the top 10 to all 131 exchanges. This allows us to view how similar the two lists are as more exchanges are included.

5.2.1 Results.

Our findings in Figure 1 show that our generated ranking list is, on average, about 50% similar to the existing lists for the top 20 exchange points. Looking at the most popular exchange ranking list *CoinMarketCap*, we see that our generated list was most closely aligned with the top 10% of exchanges. The visualization shows the metric converges as the number of exchanges included approaches 131. This is expected as the intersection of the two lists will be close to 131 which means the similarity will be close to one from Equation 4.12 (i.e. $\lim_{n \to 131} s(y_i, \hat{y}_i, n) = 1$).

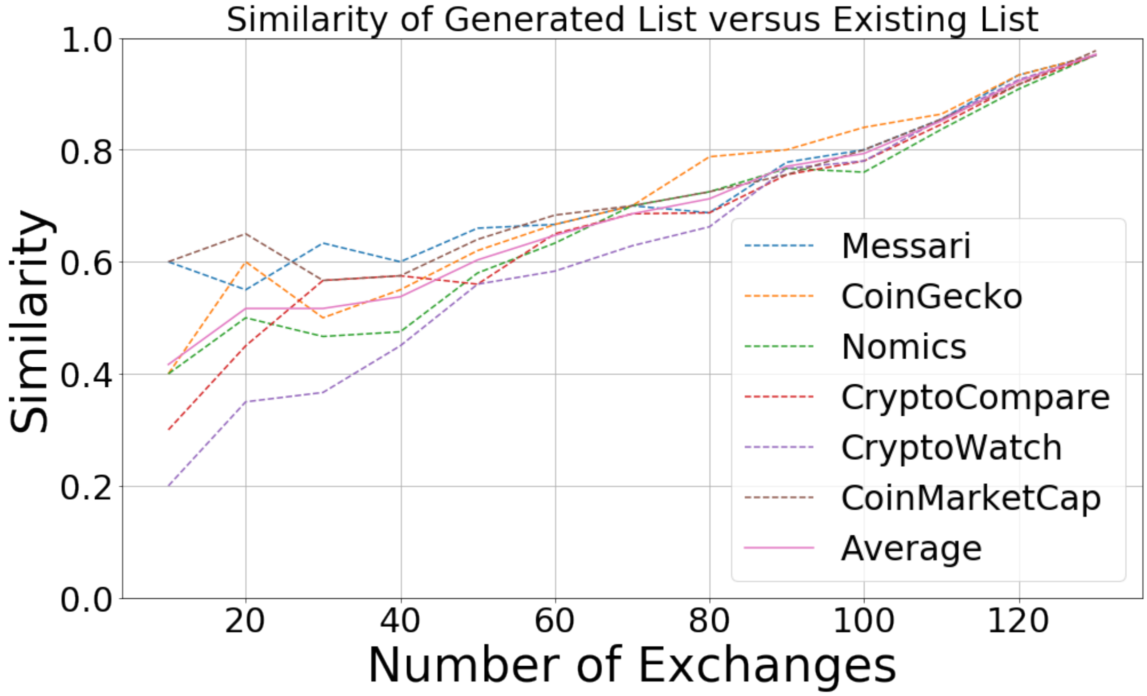

Figure 1. Line plots showing the similarity score for the six generated rank lists compared to the existing ones. On the x-axis is the number of exchanges with step 10 and the y-axis is the similarity score.

As an additional check, we recompute the experiment excluding the relative rank term to see the similarity between the two lists by changing the weights of Equation 4.10 to $w_0 = 1$ and $w_1 = 0$. Our findings in Figure 2 show that the similarity scores drop considerably for the top 20 exchanges compared to Figure 1 with the average similarity approximately 20% lower. The lower performance is further highlighted in the comparison with *CryptoCompare* as the generated list is 0% similar to the test list for the top 10 exchanges. From these results, we can conclude two points. First, the trustworthiness metric performs consistently worse without the relative rank factor. User sentiment alone is too noisy and does not align well with our expectation that the top exchanges on the existing exchange ranking lists are the most trusted. Second, the existing exchange ranking lists

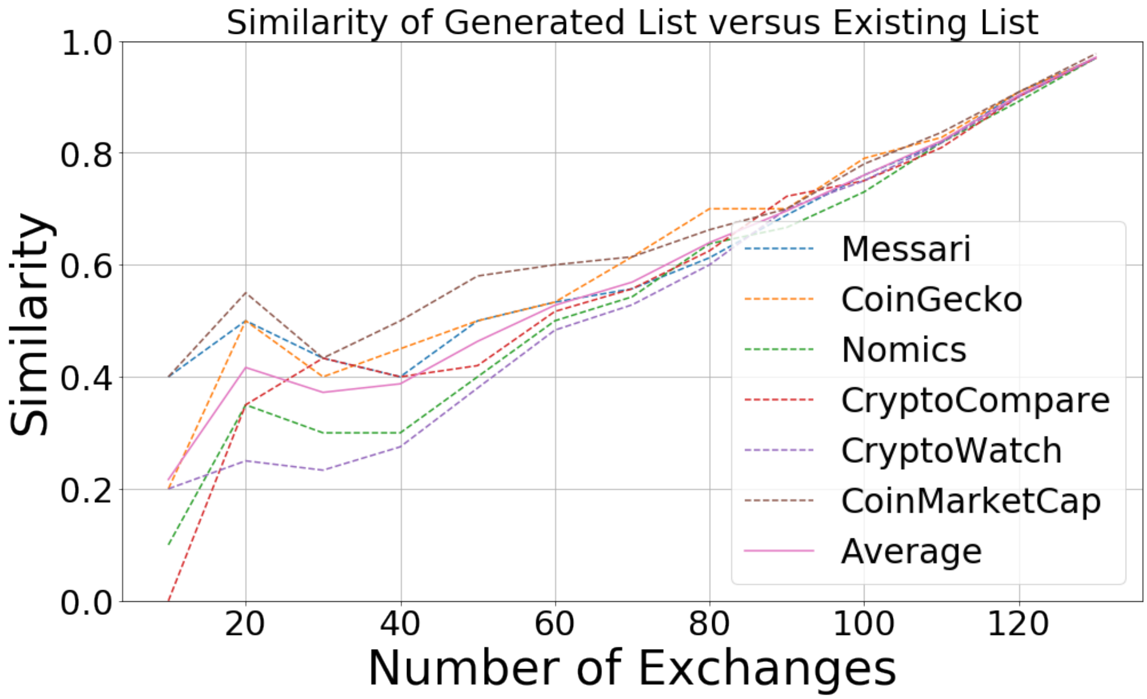

Figure 2. Line plots showing the similarity score for the six generated rank lists compared to the existing ones with just the user sentiment factor. On the x-axis is the number of exchanges with step 10 and the y-axis is the similarity score.

depend more on the shared economic factors in their methodologies than user sentiment. This makes sense as user sentiment is not found in any of the existing exchange ranking list methodologies so we expect the generated list to perform worse if it only uses user sentiment data.

5.3 Predicting Future Changes in Exchange Ranking Lists

To answer **RQ2**, we design an experiment that computes the similarity scores of the top n exchanges between our generated lists trained on the December 2020 data and the existing exchange ranking lists collected in December 2020 and March 2021. The motivation behind this setup is to evaluate the similarity of an existing exchange ranking list to our generated list for predicting future ranks of exchanges. We design the experiment as follows:

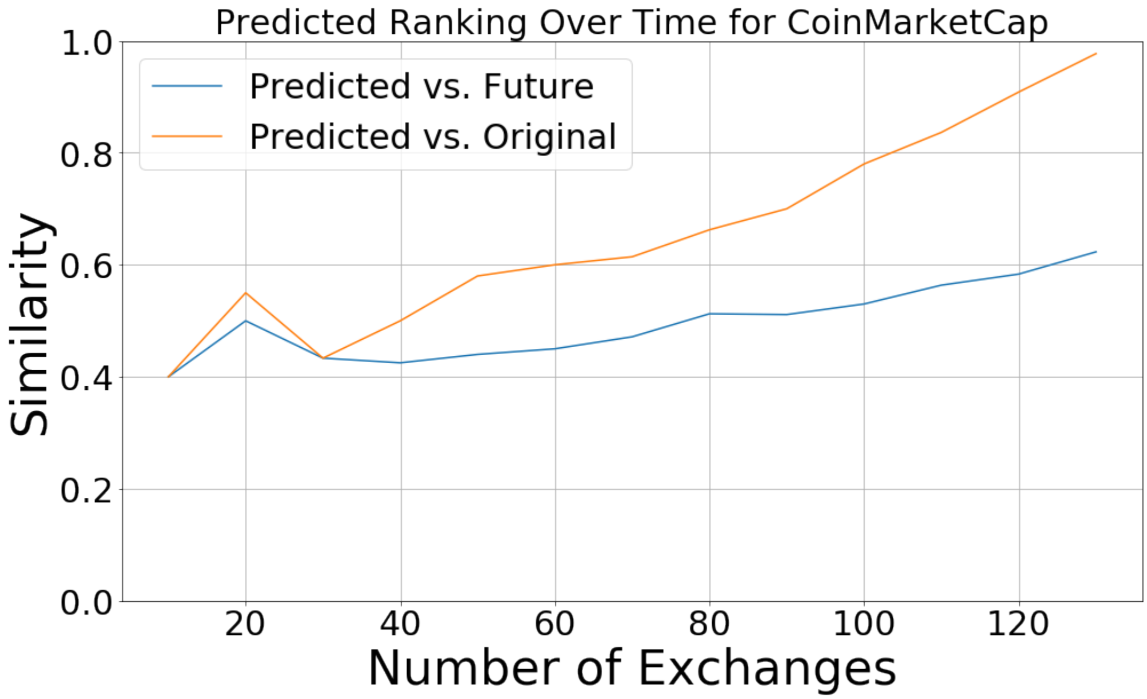

Figure 3. The line plots comparing the similarities for the top n exchanges for the generated list versus the *CoinMarketCap* list on December 2020 and March 2021.

1. Generate the exchange ranking lists as described in the previous experiment on the December 2020 data.

2. Designate the validation set to be *two* lists: $y_i^{Original}$ and y_i^{Future} which are the December 2020 and March 2021 lists respectively. We now compute the similarity between the generated list and the December 2020 list $(\hat{y}_i, y_i^{Original})$ and the similarity between the generated list and the March 2021 list $(\hat{y}_i, y_i^{Future})$.

3. Compare the similarity score between each pair..

5.3.1 Results. In Figure 3, we can see that the similarity scores of our generated list compared to both *CoinMarketCap* lists are no more than 5% different for the top 30 exchanges. This fits with the pattern of the six existing exchange ranking lists that we discussed earlier. Additionally, we find that for the top five exchanges on the generated list (Binance, Kraken, Bitfinex, Coinbase, and Huobi Global) all but Bitfinex remain in the top 5 compared to the March 2021 list and all are in the top 10. In particular, Binance was the top exchange in both lists which demonstrates predictable behavior in the generated ranking.

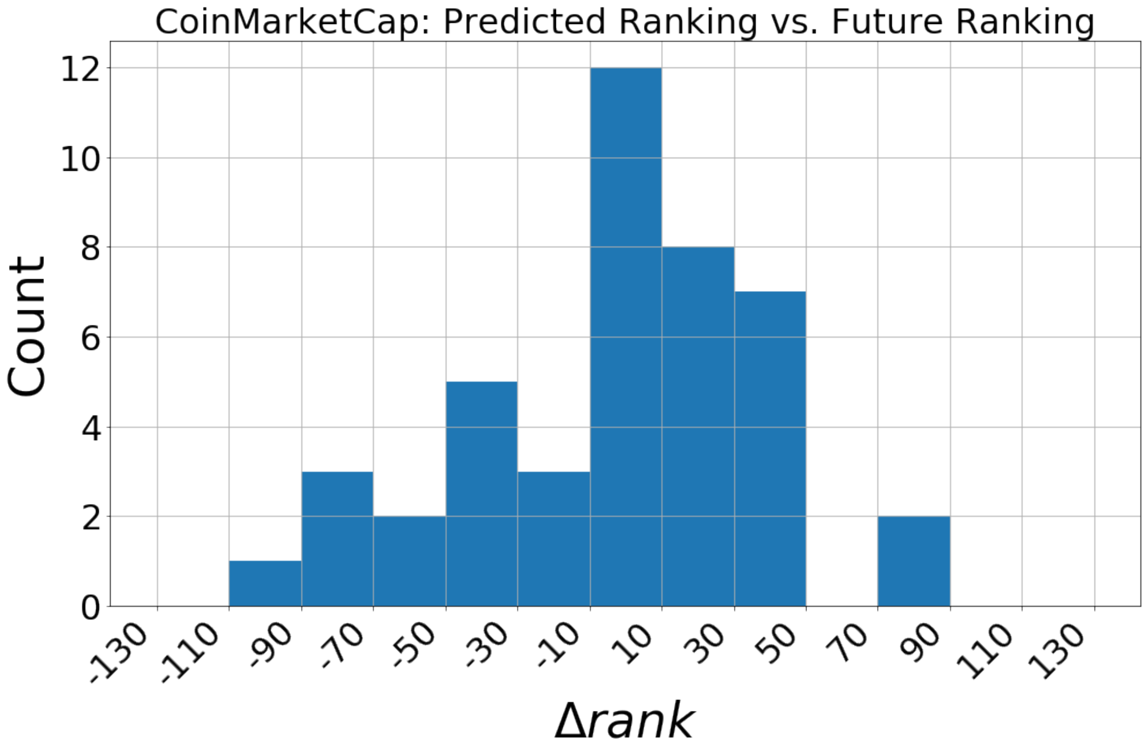

Figure 4. The histogram showing the how far off the generated list is compared to the future one. As the bins move away from zero in the positive and negative direction our trustworthiness metric marked exchanges as less trustworthy or more trustworthy, respectively.

For further inspection, we compute the change in rank for each exchange E between the generated list and the existing exchange ranking list collected on March 2021:

$$\Delta rank_E = \hat{y}_E - y_E^* \tag{5.1}$$

with $\hat{y}_E$ being the numerical rank of an exchange from the generated list on December 2020 and y^*_E as the numerical rank of an exchange from the existing exchange ranking list on March 2021. As $\Delta rank$ is computed for each exchange, we calculate and record this rank difference across all exchanges to make a distribution. The distribution (see Figure 4) provides insight of how the generated list performed compared to the existing exchange ranking list recorded on March 2021. We can determine from the distribution of $\Delta rank$ how the generated exchange ranking list behaves from three properties.

First, if the distribution right-skewed (i.e. the majority of $\Delta rank$ are negative) then an exchange will generally be ranked higher than the existing exchange ranking list. For instance, the generated list that is compared to *CoinMarketCap* had the cryptocurrency exchange *Kraken* ranked 2 but the *CoinMarketCap* list ranked it 4. The $\Delta rank$ for *Kraken* is $2 - 4 = -2$ so the generated list ranked it higher, comparatively, based on its trustworthiness score.

Second, if the distribution is left-skewed (i.e. the majority of $\Delta rank$ are positive) then an exchange will generally be ranked lower. An example of this is for the exchange *CoinBase* where the generated list had the exchange ranked 4 but the *CoinMarketCap* list ranked it 2. The rank difference is $4 - 2 = 2$ so the generated list ranked it lower, comparatively based on its trustworthiness score.

Third, if the distribution is symmetrically centered (i.e. the majority of $\Delta rank$ are approximately 0) then an exchange will generally be ranked the same as the existing exchange ranking list. For example, the exchange *Binance* was ranked 1 in both the generated list and the *CoinMarketCap* list recorded on March 2021. As $\Delta rank = 0$, the trustworthiness score behaves very closely to the existing exchange ranking list methodology.

In other words, if the distribution has a low standard deviation and has a mean around 0, then the trustworthiness metric predicts future exchange rankings very well; the more exchanges further from the mean indicates greater disparity between the trustworthiness metric and the existing exchange ranking lists. When comparing to *CoinMarketCap*, see Figure Figure 4, the distribution shows that the generated list generally finds exchanges less trustworthy than they appear in the future rankings.

CHAPTER VI

DISCUSSION

In this chapter, we discuss our findings from our experimental data and how well our trustworthiness metric performs for identifying trustworthy exchanges. We discuss the limitations of our metric and future areas of research for classifying trustworthy exchanges.

6.1 Summary of Experimental Results

The experiments show that the the trustworthiness metric is a reasonable estimator for evaluating the most trustworthy cryptocurrency exchanges. It performs well for the top 10% of exchanges in both a fixed time setting such as in the first experiment where the similarity was calculated on a single time period (December 2020), and predicting future exchange lists such as in the second experiment where the similarity was calculated in two time periods (December 2020 and March 2021). We found that the user sentiment term, although essential for addressing the shortcomings of the current, commercial exchange ranking lists, was ineffective on its own. Furthermore, we did not calculate the reverse of this (i.e. the trustworthiness metric using only the relative position factor) for the same line of reasoning. Additionally, we found that the generated list generally finds existing exchange ranking lists less trustworthy than they appear in the future. This implies that the existing lists do not consider user opinions heavily in their methodologies which we can confirm from Chapter III..

6.2 Limitations

The primary limitation when tackling the problem of identifying trustworthy exchanges was the lack of information. Cryptocurrencies are still a relatively new field and finding data of exchange characteristics is not trivial. A major focus of the crypto community is on the cryptocurrencies themselves (e.g. predicting future prices, blockchain security, mining techniques, etc.) that are much more transparent to collect data from. On the other hand, the amount of exchange user sentiment data is skewed heavily towards the most prominent companies so it becomes difficult to extend this to exchanges with small user bases. Similarly, this bias is found in existing exchange ranking lists that use web traffic scores.

The second limitation with our contribution is having a robust evaluation step. While we are able to have an interpretable metric and analysis of our generated ranking list, it remains unclear how well this actually aligns. This problem is still very new and remains unexplored so it is difficult to draw conclusions on our performance.

6.3 Future Work

Much work still needs to be done to identify trustworthy cryptocurrency exchanges, and we encourage others to continue to explore the use of user sentiment data. Our contributions are numerical formulas based on the sentiment data but exploring the use of natural language processing (NLP) for classifying individual reviews might yield further progress for this problem. A stronger understanding of the details of user opinions would provide valuable information for a deeper, robust ranking methodology. Separation of specific user concerns (e.g. security of the exchange, ease of access, customer service, etc.) would make a ranking model more explainable as these individual concerns can act as separate factors of the model's mathematical formulation. Furthermore, analyzing the distribution of these components may provide further patterns and insights on trustworthy and untrustworthy exchanges.

CHAPTER VII

CONCLUSION

In this thesis, we address the problem of identifying trustworthy exchange through our contribution of the trustworthiness metric. Unlike current, commerical exchange ranking lists, our metric incorporates user sentiment to base trustworthiness on rich user opinions. Our design is explainable and generalizable to any cryptocurrency exchange as we define trustworthiness by two principal components: the user sentiment of the exchange and where the exchange ranks, on average, in the existing exchange ranking lists. In our experiments, we find that our metric generally aligns well with the top exchanges in both a fixed time period setting and evaluating future changes in exchange ranking lists. In both cases, we found that when computing the similarity of the generated list to the existing exchange ranking lists, the differences are overtly due to the trustworthiness metric finding exchanges to be less trustworthy than current ranking lists. We have established the trustworthiness metric as a useful tool for further understanding user behavior which is essential to identify fraudulent exchanges and cultivate user trust in the cryptocurrency domain.

REFERENCES CITED

Bitcoin (btc) live price charts, trading, and alerts. (n.d.). Retrieved from https://cryptowat.ch/

CMC, C. (n.d.). *CoinMarketCap Revamps Market Pairs Ranking to Empower Users Against Volume Inflation - CoinMarketCap Blog.* Retrieved 2020-08-19, from https://blog.coinmarketcap.com/2020/05/29/coinmarketcap-revamps-market-pairs-ranking-to-empower-users-against-volume-inflation/

CMC, G. (n.d.). *An In-Depth Look at CoinMarketCap's Newly Improved Liquidity Score for Finding the Best Crypto Exchanges - CoinMarketCap Blog.* Retrieved 2020-08-19, from https://blog.coinmarketcap.com/2020/05/08/an-in-depth-look-at-coinmarketcaps-newly-improved-liquidity-score-for-finding-best-crypto-exchanges/

CoinGecko. (n.d.). *Methodology | CoinGecko.* Retrieved 2020-08-19, from https://www.coingecko.com/en/methodology

Cryptocompare. (n.d.). Retrieved from https://www.cryptocompare.com/

Crypto crime summarized: Scams and darknet markets dominated 2020 by revenue, but ransomware is the bigger story. (n.d.). Retrieved from https://blog.chainalysis.com/reports/2021-crypto-crime-report-intro-ransomware-scams-darknet-markets#:~:text=In2020,thecriminalshare,tripledbetween2019and2020.

Crypto market caps - prices, all-time highs, charts. (n.d.). Retrieved from https://nomics.com/

Crypto research, data, and tools. (n.d.). Retrieved from https://messari.io/

Financial terms dictionary. (n.d.). Investopedia. Retrieved from https://www.investopedia.com/financial-term-dictionary-4769738

Fusaro, T., & Hougan, M. (2019). *Meeting with Bitwise Asset Management, Inc., NYSE Arca, Inc., and Vedder.* Retrieved from https://www.sec.gov/comments/sr-nysearca-2019-01/srnysearca201901-5164833-183434.pdf

GoodCrypto. (n.d.). *What is liquidity and how to find a liquid exchange? - GoodCrypto.* Retrieved 2020-09-14, from https://goodcrypto.app/blog/what-is-liquidity-and-how-to-find-a-liquid-exchange/

Jay. (n.d.). *Listings Criteria – CoinMarketCap.* Retrieved 2020-08-19, from https://support.coinmarketcap.com/hc/en-us/articles/360043659351-Listings-Criteria

Jin, S. (2019). *Trust Score 2.0: CoinGecko Updates Trust Score to Improve Exchange Transparency - CoinGecko Blog.* Retrieved 2020-08-19, from https://blog.coingecko.com/trust-score-2/

Kaiko. (n.d.-a). *Bid-Ask Spread as an Indicator of Crypto-Market Liquidity | by Clara Medalie | Kaiko Data.* Retrieved 2020-09-14, from https://blog.kaiko.com/bid-ask-spread-as-an-indicator-of-crypto-market-liquidity-b15bdc0a621c

Kaiko. (n.d.-b). *Measuring Liquidity: Spread and Market Depth (And Why Trade Volume is Unreliable) | by Kaiko | Kaiko Data.* Retrieved 2020-09-14, from https://blog.kaiko.com/measuring-liquidity-spread-and-market-depth-56290f2caa0a

Kelion, L. (2021, Feb). *Trustpilot removed 2.2 million bogus reviews in 2020.* BBC. Retrieved from https://www.bbc.com/news/technology-56100082

Ong, B. (2019). *CoinGecko Introduces "Trust Score" to Combat Fake Exchange Volume Data - CoinGecko Blog.* Retrieved 2020-08-19, from https://blog.coingecko.com/trust-score/

Revain. (n.d.). *Review authors rating ranking system on revain platform.* Retrieved from https://revain.org/authors

Revain. (2020, May). *Revain updates companies rating and ranking systems.* Retrieved from https://revain.org/blog/companies-rating-and-ranks

Schueffel, P., & Groeneweg, N. (2019). Evaluating Crypto Exchanges in the Absence of Governmental Frameworks - A Multiple Criteria Scoring Model. *SSRN Electronic Journal*, 1–28. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract{_}id=3432798 doi: 10.2139/ssrn.3432798

SEC Office of Investor Education and Advocacy. (n.d.). Ponzi schemes Using virtual Currencies Ponzi Schemes Generally. *Investor Assistance*, *800*(800), 732–330. Retrieved from https://www.sec.gov/files/ia{_}virtualcurrencies.pdf

Young, J. (2017, Dec). *South korean government concerned with scams in bitcoin market, fake exchanges.* Cointelegraph. Retrieved from https://cointelegraph.com/news/south-korean-government-concerned-with-scams-in-bitcoin-market-fake-exchanges