Abstract cryptocurrency sectorisation through clustering and webscraping: Application to Systematic Trading

Babak Mahdavi-Damghani[1,*] , Robert Fraser* , James Howell** , and Jon Sveinbjorn Halldorsson***

*University of Oxford, **Imperial College London & ***University of Warwick

***Abstract*— Though it is presented as a hypothesis, we discuss the historical events that have led to the rise of Cryptocurrencies as a legitimate new asset class. We also discuss issues around cryptocurrency fundamentals as a means to explain the lack of sectors which exists for the other asset classes such as equities or commodities. To address this issue we propose a new methodology based on a hybrid approach between k-Means and Hierarchical Clustering (HC) with alternative data gathered from web-scraping. We then reintroduce a couple of mathematical models, namely Risk Parity (RP) and Momentum. We finally test our geopolitical hypothesis through a long only strategy using RP, and test our abstract sectorisation through a Long/Short (L/S) strategy.**

Keywords: Cryptocurrency, Geopolitics, Trading, Long/Short, Market Neutral, k-Means, Hierarchical Clustering, Momentum, Web-scraping, Alternative Data.

I. Introduction

A. Preamble: Finance as an Ecosystem

A small caveat: we do not approve or disapprove of the way historical events are portrayed in general in this document. We however, believe that enough market participants view historical events exactly as depicted in this document. Their number is such that their actions in the ecosystem of market participants influence the prices of the underlying financial securities and therefore change the way portfolios should be constructed. In some sense, real history does not matter as much as its perception for trading decisions. If you are interested in knowing more about ecosystem modelling for financial applications, please refer to [11], [34], [35].

B. Brief Relevant Geopolitical History

The destruction of life induced by the world wars of the previous century culminated with the Bretton Woods agreement. This addressed the concerns raised by the political leaders of the time, more specifically when it came to agreeing on a world currency backed by Gold. Issues surrounding Gold as a reserve currency (or its lack thereof) has been a recurring theme in world history going back thousands of years. It seems that all empires decide, at some point, to move away from Gold[1] as a reserve currency. It also seems that all empires fail shortly after that (eg: the fall of the Roman empire [15]). Though, the Bretton Woods agreement was arguably imperfect, it did however bring stability to the world economy. This lasted until 1971, when the USA, pressured financially by the Vietnam war, unilaterally terminated convertibility of the USD to gold, rendering it a fiat currency. In the next couple of years the UK and France also decided to move towards a system of fiat currencies and soon the whole world followed. As seen in Figure 1, a period of financial instability followed. This issue of fiat currency is a central theme of realpolitik. For instance, because commodities are priced in USD, many nations (Venezuela, Libya etc.) rich in natural resources felt naturally compelled (the USD purchasing power dropped by around 90% since 1971) to fight this alternative banking model with their own competing model (eg: Gaddafi's gold-backed African currency [21]). These attempts at escaping this USD domination, have been systematically contained with a combination of conflicts. This makes the creation of cryptocurrencies, an anonymous currency model, a logical response. In fact the creation of an e-cash model was predicted by Nobel Prize recipients in Economics, such as Milton Friedman [17]. Many see the USD as an unfair system but proposing to go back to the gold standard could be seen as a war declaration on the USD [42]. Supporting cryptocurrencies is therefore a way to reject the USD while staying anonymous at the same time (and therefore escaping military interventions). In fact, though not perfect, cryptocurrencies address the fundamental issues that most countries have with the post Bretton Woods fiat currency model. Many countries facing embargoes have started investigating cryptocurrencies as a way to escape the USD. For instance, Venezuela has

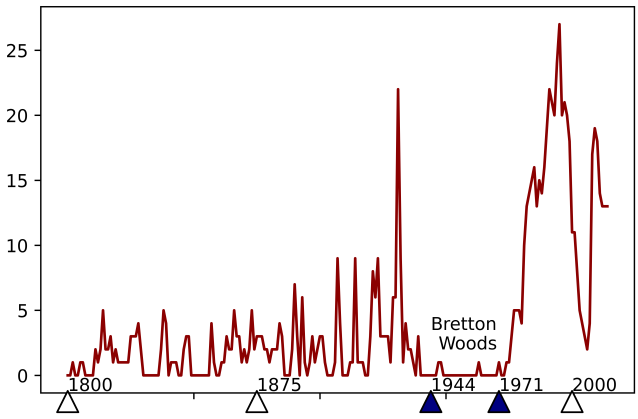

Fig. 1: Countries having a banking crisis count each year.

* babak.mahdavidamghani@oxford-man.ox.ac.uk

[1]or temper with the purity of its gold coins (at the dawn of the Roman Empire.)

recently launched the "Petro" [50], its own cryptocurrency backed by oil. Other countries are also currently developing their own cryptocurrencies [20], for example Russia, China, Japan etc. The biggest and most prestigious banks have been at the forefront of misrepresenting the importance of cryptocurrencies as a legitimate alternative to the USD [44]. This lack of integrity has been noticed by Cryptocurrency supports. A meme has emerged as a result [13], [23] (see figure 2).

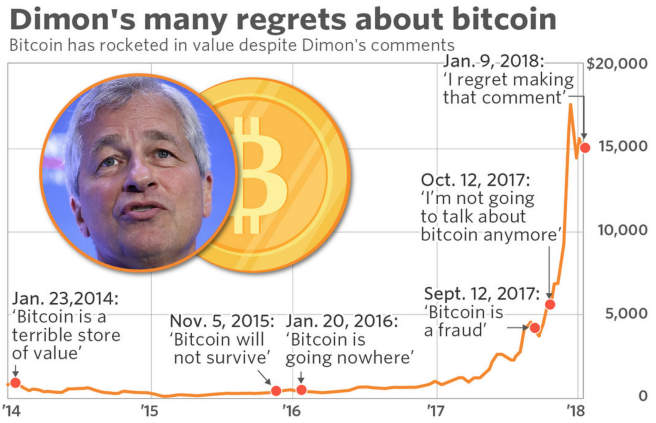

Fig. 2: Jamie Dimon vs Bitcoin internet meme example [13]

In any case despite propaganda belittling the emergence of cryptocurrencies as a legitimate long term alternative to the USD [28] the cryptocurrency market cap is steadily increasing [49]. Though it remains small compared to USD, it is fast catching up to the USD in circulation as well as to the Gold market [49]. It may be unlikely to be immediately adopted by large corporations (eg: Tesla [29]), it is however very likely that it will be the case in the coming decade. We can already see price volatility decreasing. Without delving more into geopolitical theories, cryptocurrencies are another inflation resistant tool that cannot be created out of thin air, which ensures sustainable value in the long term.

C. Cryptocurrencies' Fundamentals

Note that the geopolitics of cryptocurrencies is in itself its most important fundamentals. However, we felt compelled to include a more traditional section of their fundamentals. More specifically, each cryptocurrency has its own benefits and each has its own limitations. We could almost say that they have their own economic niche, the same way equities or the commodities markets have their own sectors. For instance, some cryptocurrencies are better in their speed of transactions. Some other cryptocurrencies can accommodate more transactions at the same time. Few are paradoxically bank friendly. Others, yet, are actually a generalised "decentralisers". The ecosystem of cryptocurrencies can be quite complex and their numbers is constantly increasing through forking (see Figure 11).

Remark A fork, in the context of blockchains, is defined when a blockchain diverges into two potential paths forward or if there is a change in protocol or other similar situations. For instance by the end of 2020, the biggest cryptocurrency exchanges were trading more than 50 different cryptocurrencies. These include:

Bitcoin (BTC): the first decentralized digital currency and the biggest by market cap. Invented by an unknown person or group of people under the name Satoshi Nakamoto in 2009, it is currently the least volatile of the main cryptocurrencies.

Ethereum (ETH): the second Cryptocurrency by market cap, has potentially a bigger upside when compared to BTC as it also offers the possibility to build decentralized applications (though questions were also raised about its security, scalability as well as the programming language used).

Ethereum Classic (ETC): was created as a result of a internal dispute fueled by ETH vulnerability.

Litecoin (LTC): built originally by a Google engineer on the premise that BTC was too slow, LTC main advantage is its relative increased speed but makes it harder to mine.

Ripple (XRP): system was designed to eliminate BTC's reliance on centralized exchanges. It also uses less electricity than BTC and performs transactions faster than BTC.

Dash : called XCoin and Darkcoin in the past, was initially designed to be the most user friendly cryptocurrency.

Tron (TRX): like ETH aims at building build decentralized applications. As of mid 2020 and though it is less known than ETH, TRX prouds itself by being able to accommodate transaction almost at a scale 10 times faster than ETH and it is also able to achieve this with a much more well known language than the latter (Python for TRX as opposed to Solidity for ETH).

Neo : is described as the ETH of China. NEO has however suffered from a slower adoption than ETH. It uses what few consider a revolutionary consensus method called dBFT which ensures a fast and ethical transaction process especially when the network increases in size.

Omisego (OMG): is best known for the next ETH scaling Solution. It is important to note that OMG is not a new project but an add-on to ETH. Its existence is related to ETH fees issues. Indeed users can pay a higher fees in order to have their transaction prioritized which goes against the spirit of equality as derived from decentralisation. It used to be called Omisego but re-branded into OMG upon the hiring of a new CEO.

Monero (XMR): is sometimes referred as privacy coin. The concept of privacy is con-substantial to cryptocurrencies however, there exists degrees of how this is engineered. For instance, BTC and most other cryptoccurencies have a traceable transaction history. This makes them easily traceable. This is not necessarily a bad thing if the trace goes back to a hack or illegal activities. However, these traces could be abused. XMR addresses these issues through the concept of stealth addresses. The IRS, DEA and FBI have all put

bounty on any other external agency that would be able to crack XMR down.

Zcash (ZEC): is, like XMR, another privacy coin. Originally called ZeroCash, it was renamed to Zcash for marketing purposes

AirSwap (AST): is a decentralised exchange aiming at freeing users from the need of a centralised exchange that would charge transaction fees or fees associated to other services.

Monacoin (MONA): was created out of a hard fork with litecoin. Designed to be a payment token instead of a speculative one, it is also referred as the digital cash of Japan.

Verge (XVG) relies on the technology of The Onion Router (TOR) and the Invisible Internet Project (I2P) to protect users' identities (instead of relying on cryptographic techniques). TOR ensures hiding the user's identity better.

Horizen (ZEN) offers privacy shielded Z-Addresses and public T-Addresses that work similarly to Bitcoin. However, sending funds from a Z-Address to a T-Address will show the amount received. Horizen also boasts a vast node network, which helps to improve anonymity.

D. Problem Formulation

How can we take advantage of this newly formed asset class if we are an asset manager or a hedge fund manager? The latter two have different trading constraints. More specifically we know that asset managers are constrained by strategies that are long only. On the other hand hedge fund managers are strongly encouraged to have a hedging element in their strategies, more specifically by being market neutral.

1) Trading for asset managers: most of the geopolitics of cryptocurrencies is arguable labelled as fake news or conspiratorial in nature. How can we filter out the truth from the fake? Or do we even need to accomplish such objective? Can we construct a strategy that would be long only, abide by the constraints of asset management, yet keep drawdowns under control?

2) Trading for hedge fund managers: classic fundamentals, sector driven, long/short (L/S) strategies initially tailored for the equities or commodities market do not translate naturally for the cryptocurrencies market. For instance BP and Shell may compete in the energy sector of the equities market because their business model and categorization is clear. Comparing each company and coming up with a reasonable trading strategy based on this categorisation becomes easier. However, for cryptocurrencies the market is not that easy. The market is very immature and there is no developed categorisation as a result. Each cryptocurrency tries to address the limitations of the others while at the same time trying to bring a unique twist for what they can offer. The trend for finding clear separations for cryptocurrencies has recently even touched the domains of art and pseudo-science (see Figure 3). However this way of viewing cryptocurrencies is very simplistic because many cryptocurrencies' features are not mutually exclusive. For example XRP is both scalable and also categorised as Digital Cash (see figure 4). Sectorisation[2] is "fuzzier".

E. Agenda

In section II we discuss data related issues. In section III we summarize the mathematical models used in this backtesting of our two trading strategies. In this section we also discuss classic clustering methodologies that we use later in the paper. The results of our trading backtesting are summarized in section IV. The latter is subdivided in two subsections. First, in subsection IV-A, a long only strategy keeping in mind that drawdowns above 20% are usually not well received. Second, in subsection IV-B we examine a more sophisticated strategy by ensuring that our overall strategy remain market neutral.

II. Data

A. Traditional Data

All of the price data sourced for the strategy was taken from CoingGecko [31] using the free data API [30] they provide. We chose the coins based on their market cap, more specifically taking the one-hundred largest coins based on market cap as of 21^{st} of December 2020 (see Figure 5). Using the API provided, we downloaded as much hourly data as we could starting from the 15^{th} of December 2018 at 00:00, running until the 17^{th} of December 2020 at 00:00. The 15^{th} December 2018 00:00 is chosen as the start date since this appears to be the earliest date from which one can download hourly data from CoinGecko.

Remark Not all of the coins offer hourly data as early as this, but we have managed to get at least one and a half thousand data points for each of the coins in hour sample. Of the one-hundred coins we collected price data for, twelve were dropped from the final sample. Of these twelve, eleven were dropped because their short-hand names were incompatible with our clustering algorithm, and the final coin was dropped because the API returned no price data over our specified date range.

B. Alternative Data

As mentioned previously, the formalisation of clear sectors is currently an open problem in cryptocurrency ecosystem modeling. Multiple attempts have been made [1], [2] but none of these studies is making consensus. The absence of the latter consensus forces us to engineer one. We start by collecting all features of these cryptocurrencies using our tailor made web-scraping algorithm. Web-scraping is the process of algorithmically downloading data from a format that is not necessarily immediately usable for analysis (eg: text on a website) into a usable format (eg: excel spreadsheet). Figure 6 illustrates the process. Web-scraping has to be bespoke for every website we considered. This is because the structure of each of these websites is different and because

[2]Perhaps a slight abuse of language: we mean as the process of creating a sector.

Fig. 3: Artistic and pseudo scientific representation (as a periodic table) of cryptocurrencies' categorisations [18].

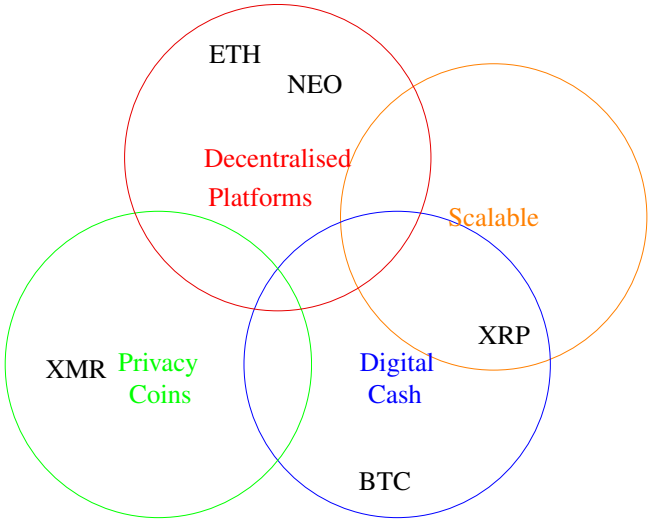

Fig. 4: Abstract illustration of cryptocurrencies sectorisation.

these structures can change through time. We built these bespoke algorithms and also made sure the scraping was done responsibly. Indeed the scraping should never disrupt the proper functioning of the website. We have performed a correlation study for the features of one of these website[3]. The list of the features can be quite intuitive (eg: "privacy coin", "smart contract platform") to more advanced (for example the of algorithm used: cryptonight, equihash, sha256 etc.). Though the feature correlation study can perhaps help reduce the dimension of the problem, it does not however give us any information on how to form cryptocurrency sectors. We need to, instead, study the correlation between the features of different cryptocurrencies. Though not ideal this can be done using 1s and 0s to signify "has/does not have this feature". We have done this study with a very small sample size (see Figure 8). We can see, for instance here, that Bitcoin and Bitcoin Cash or Etherium and Etherium classic are quite "close" in terms of features. So we could, based on that simple study, put BTC and BCH in "abstract sector 1" and ETH and ETC in "abstract sector 2". However, this way of proceeding, quickly leads to a dead end. More specifically as the the number of cryptocurrency increases[4], the more unclear this correlation study becomes.

III. Mathematical Modelling

A. Risk Parity

There exists several methodologies for adding different sources of alpha to each other. Throughout this paper we will use Risk Parity (RP), sometimes known as equally-weighted risk contributions for this purpose. Formally:

Definition We consider a portfolio of N assets: $x_1, ..., x_N$ with corresponding weights $w_1, ..., w_N$ and volatilities $\sigma_1, ..., \sigma_N$. If our weights are given by equation (1),

$$w_p = \frac{\tilde{w}_p}{\sum_{i=1}^{N} \tilde{w}_i}, \tag{1}$$

where $\tilde{w}_p = \frac{\sum_{i=1}^{N} \sigma_i}{\sigma_p}$, then we say that the risk contribution for each orthogonal source of alpha composing our portfolio abides by the rules of Risk Parity.

[3]This can be explored in the on Figure 7 from our appendix.

[4]See Figure 9 from our appendix

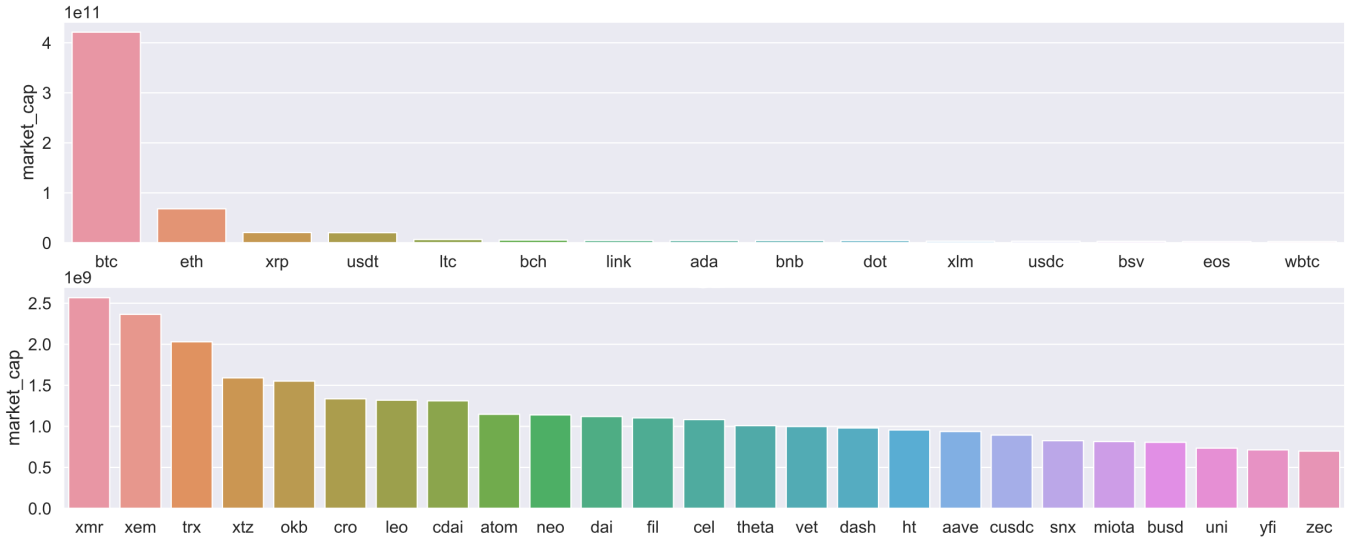

Fig. 5: Market cap of the 40 coins used in the analysis (USD).

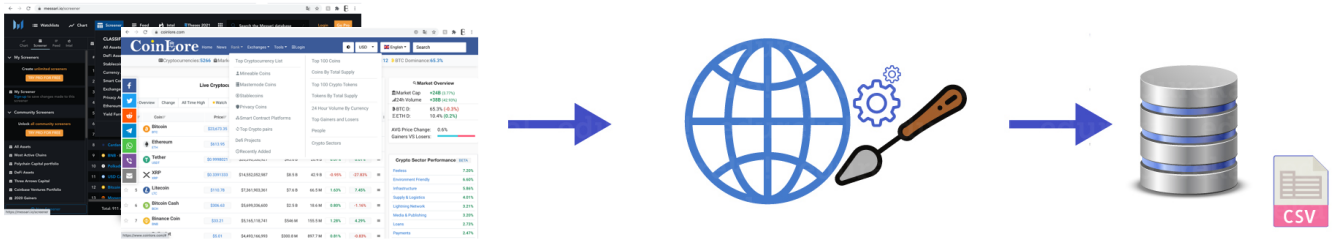

Fig. 6: Flowchart of our web-scraping algorithm.

Remark Note that if we call by Σ the covariance of $X = (x_1, ..., x_N)$, the volatility is defined by $\sigma(w) = \sqrt{w'\Sigma w}$. The weights are allocated following the minimization optimization given by equation (2)

$$\arg\min_{w} \sum_{i=1}^{N} \left[w_i - \frac{\sigma(w)^2}{(\Sigma w)_i N} \right]^2 \tag{2}$$

We explore in the next section how to implement this result in a simple fashion.

B. Momentum

1) Overall Definition: momentum is perhaps the most studied and most widely celebrated trading strategy [41], [26], [47], [7]. It disputes the Efficient Market Hypothesis [16] rooted in Bachelier [9] doctoral dissertation which has led part of 20th century quantitative finance. The idea of momentum is rooted on the idea that best performing securities over the more or less recent past tend to continue to perform well over the subsequent period. The strategy is usually longing these best performing securities and shorting the worst ones.

2) Index Construction: when discussing a momentum strategy the existence of a benchmark is assumed.

Definition We call I the index composed of N assets $x_1, ..., x_N$ with similar characteristics but corresponding volatilities $\sigma_1, ..., \sigma_N$ and weights $w_1, ..., w_N$ from definition III-A. The index value at time t will be given by

$$\pi_{t-\tau:t} = \prod_{t=\tau}^{T} \left[\left(\sum_{i=1}^{N} w_{i,t} r_{i,t} \right) + 1 \right] \tag{3}$$

We will also call $\pi_{\tau:t}$ the rolling window of length τ, and $r_{i,:t-\tau:t}$ the cumulative return of security i between $[t-\tau, t]$.

Remark Note that the term characteristics is meant to be a generalisation of the "sector" concept in equities. It is meant to be a rough method to gather securities that are similar into a group. In equities an example of sectors would be "energy" (eg: BP, Shell etc.) or "technology" (eg: Apple, Google etc ...). Note that in commodities "energy" means something else (eg: WTI, Brent etc.). Other example of commodities without name ambiguity (with the equities market) would be "precious metals" (eg: Gold, Silver).

Definition We call $\tilde{r}_{i,\tau:t}$ the normalised cumulative returns of asset i between $[t-\tau, t]$ such that

$$\tilde{r}_{i,t-\tau:t} = \frac{r_{i,t-\tau:t}}{I_{t-\tau:t}} \tag{4}$$

Lately we need to define the window of relevance and a measure for our momentum. We need to engineer the idea that there is a consistent departure from a security to its benchmark index. One way to engineer this concept is through a couple of snapshots in time.

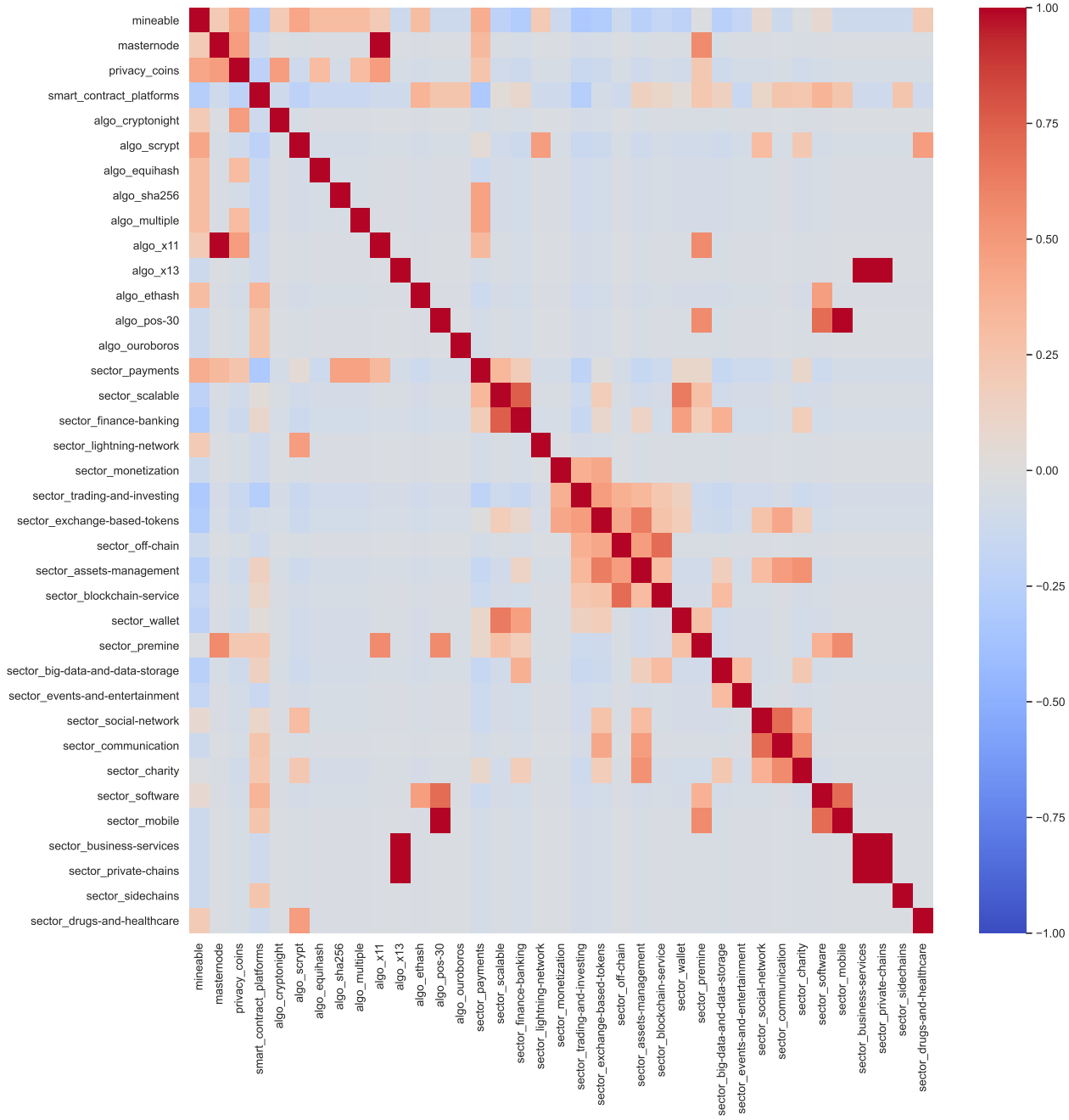

Fig. 7: Correlation matrix of 37 cryptocurrencies features.

Definition Let τ_l, τ_m and τ_s be the long, medium and short term time window, the our signal is given by equation (5):

$$S_{i,t} = 1_{\delta_{i,\tau:t}=1} - 1_{e_{i,\tau:t}=-1} - \left(1_{\delta_{i,\tau:t}=-1} - 1_{e_{i,\tau:t}=1}\right) \quad (5)$$

where

$$\delta_{i,\tau:t} = \begin{cases} -1 & \text{if } (\tilde{r}_{i,\tau_l:t} > \tilde{r}_{i,\tau_m:t}) \ \& \ (\tilde{r}_{i,\tau_m:t} > \tilde{r}_{i,\tau_s:t}), \\ 1 & \text{if } (\tilde{r}_{i,\tau_l:t} < \tilde{r}_{i,\tau_m:t}) \ \& \ (\tilde{r}_{i,\tau_m:t} < \tilde{r}_{i,\tau_s:t}), \\ 0 & \text{otherwise,} \end{cases}$$

and

$$e_{i,\tau:t} = \begin{cases} -1 & (\tilde{r}_{i,\tau_m:t} > \tilde{r}_{i,\tau_s:t}), \\ 1 & (\tilde{r}_{i,\tau_m:t} < \tilde{r}_{i,\tau_s:t}), \\ 0 & \text{otherwise.} \end{cases}$$

Remark Our signal can be decomposed into its entry strategy $\delta_{i,\tau:t}$ and its exit strategy given by $e_{i,\tau:t}$.

Remark Also note that the value of $\tau_i i \in l, m, s$ is function of the application and the trading frequency.

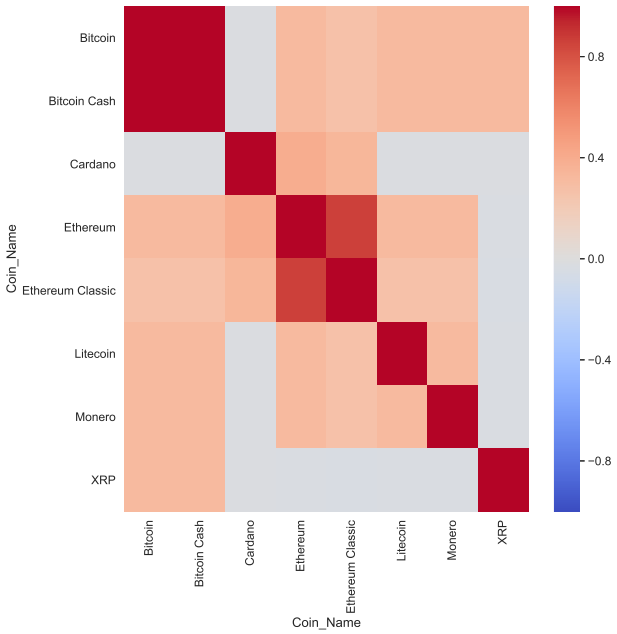

Fig. 8: Correlation matrix of 8 cryptocurrencies, using the characteristics from web-scraping.

C. Clustering algorithm

1) Literature Review: greedy algorithms are a family of algorithms that employs practical methods that are not guaranteed to be optimal. Hierarchical clustering algorithms are part of that family [22]. The steps of the algorithm are as follows. We assume that a pair of groups (or clusters), as large as one wishes but as small as singletons, are merged at each time. The methodology is recursive in spirit. The trivial partition consists of just one class. Each class is merged two by two up to its root in order to create a binary tree. The tree in this context is usually called a dendrogram [10], [32], [38], [24], [6], [8], [19], [25], [46]. One major theme of the clustering literature is time complexity, the earlier versions of the clustering algorithms being relatively slow: $O(n^2)$ [48], [45], [14]. Innovations was therefore led by these speed limits [12], [27]. A survey of these improvements can be found in [37], [37]. The family of k nearest neighbor (NN) chains (NNC) models have emerged as the ones combining the best overall speed and accuracy ratio. A NNC consists of an arbitrary point followed by its NN, followed by the NN from among the remaining points for this second point. This is recursively applied to the rest of the chain. In order to allocate similarities we calculate the points dissimilarities [4] first. This done using a Euclidean distance from equation (6).

$$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2} \tag{6}$$

Remark Note that the Euclidean distance is one out many others, for example:

- Squared Euclidean distance $\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
- Manhattan distance $\|a-b\|_1 = \sum_i |a_i - b_i|$
- Maximum distance $\|a-b\|_\infty = \max_i |a_i - b_i|$
- Mahalanobis distance $\sqrt{(a-b)^\top S^{-1}(a-b)}$ where S is the Covariance matrix

A major drawback of the k-NN rule is the high variance when dealing with sparse prototype datasets in high dimensions. Most techniques proposed for improving k-NN classification rely either on deforming the k-NN relationship by learning a distance function or modifying the input space by means of subspace selection. This has led for the construction of the boosted K-NN [43].

2) K-Means: K-NN remains a supervised clustering methodology. Perhaps the closest unsupervised version of the algorithm is the k-Means algorithm described by algorithm (1). This procedure is repeated until we are able to produce the elbow plot (see figure 10).

Algorithm 1 K-Means

Inputs:
  $C^{1,2,\dots N}_{f_1,f_2,\dots,f_F}$
Initialize:
  $\forall (i,j) \in N, i < j,$
  $d(C_i, C_j) \leftarrow 0,$
  $\mathcal{C}_{1,2,\dots K} \leftarrow C^{1,2,\dots K}_{f_1,f_2,\dots,f_F}$
//——- ▷ Calculate distance between cryptocurrencies
for i = 1 to N **do**
  for j > i to N **do**
    for f = 1 to F **do**
      $D(C^i, C^j) \leftarrow D(C^i, C^j) + \sqrt{(C^i_f - C^j_f)^2}$
    end for
  end for
end for
//———————— ▷ Compare each crypto to K centroids
for i=1 to K **do**
  for j = 1 to N **do**
    $E(\mathcal{C}_i, C^j) \leftarrow \sqrt{(\mathcal{C}_i - C^j)^2}$
  end for
end for
//———————— ▷ Assign each crypto to its set
for j = 1 to N **do**
  $\Omega_{1,2,\dots K} \leftarrow \arg\min_i \sqrt{(D(C^i, C^j) - \mathcal{C}_{1,2,\dots K})^2}$
end for
//———————— ▷ Recalculate Centroids
for i = 1 to K **do**
  for j = 1 to N **do**
    $\mathcal{C}_i \leftarrow C^i \times 1_{C^j \in \Omega_i}$
    $N_i \leftarrow N_i + 1_{C^j \in \Omega_i}$
  end for
end for
for i = 1 to K **do**
  $\mathcal{C}_i \leftarrow \frac{\mathcal{C}_i}{N_i}$
end for
//———————— ▷ Repeat until stable & elbow plot

3) Hierarchical clustering: Agglomerative Hierarchical Clustering (AHC) [39], [40] have also emerged as one of the

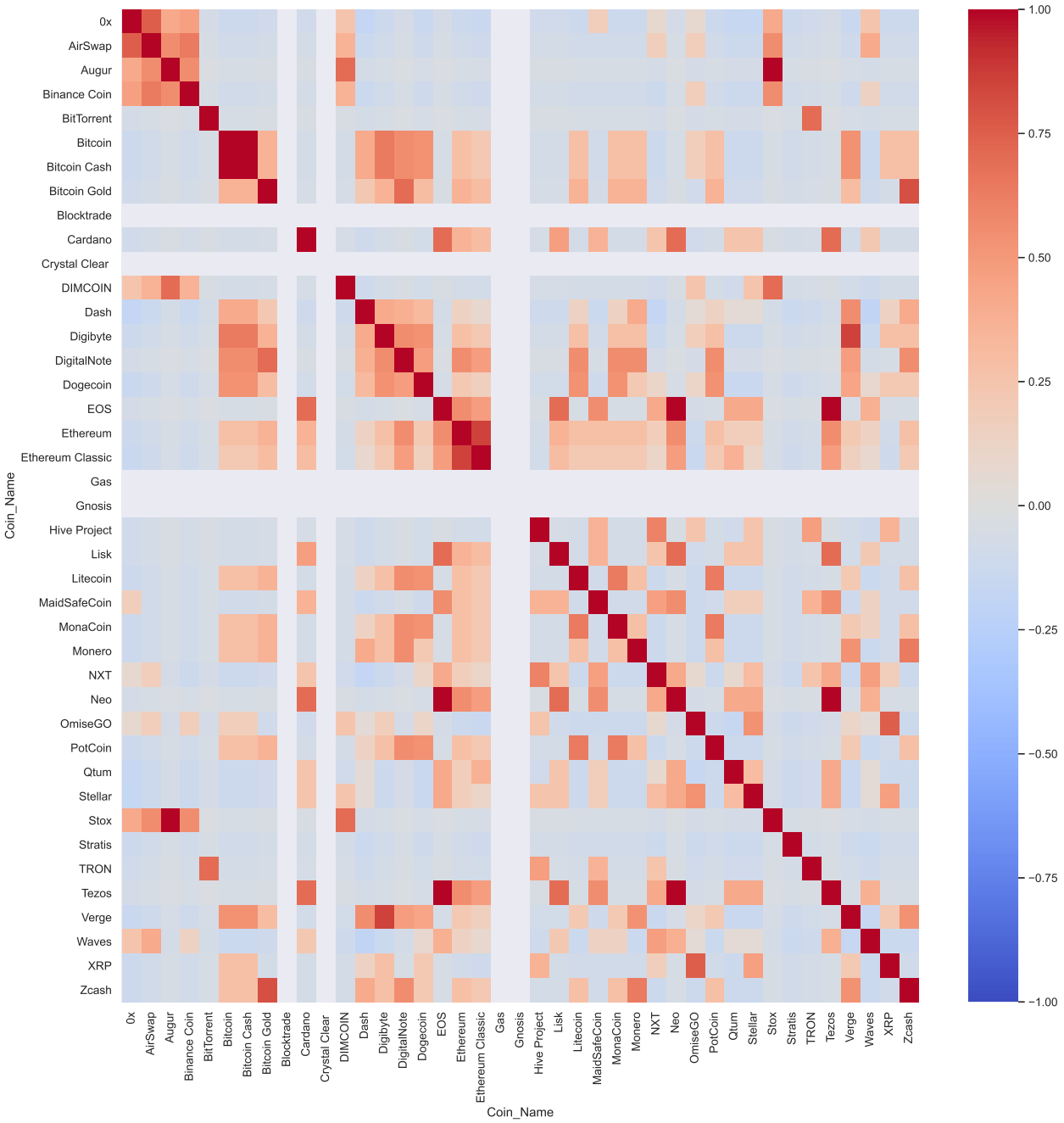

Fig. 9: Correlation matrix of 40 cryptocurrencies, using the characteristics from web-scraping.

most useful unsupervised clustering methodologies for our application. This method starts by assigning each observation to its own cluster, the distances between the clusters are then calculated and the two most similar clusters are joined. This process is repeated until there is only one cluster left. Before clustering the function used to create the distance matrix is specified. We used a Ward's linkage criterion with Euclidean distances to determine the distances between the clusters, which in the two cluster case is given by equation (7):

$$D_{12} = \sqrt{\frac{2|k||l|}{|k| + |l|}} \cdot ||\mathbf{x} - \mathbf{y}||, \tag{7}$$

where k and l are the number of observations in each cluster and $\mathbf{x}$ and $\mathbf{y}$ are the position of the centroids of the clusters. The AHC methodology uses a sorting algorithm, not central as long as they converge. We refer the motivated reader to studies dedicated to this problem [3], we have chosen the merge sort algorithm 28 because of its simplicity and it widespread availability. The core of the AHC methodology is

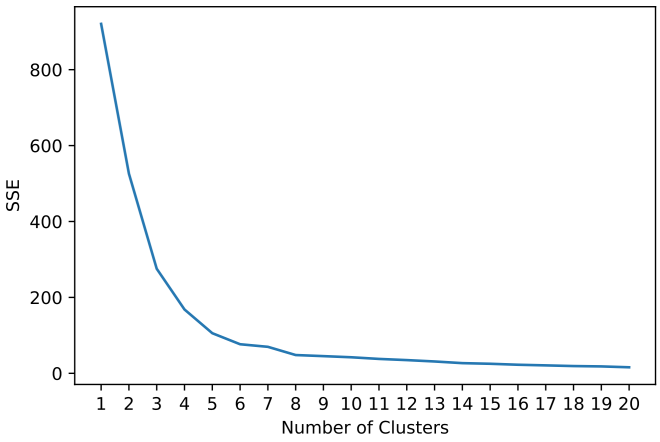

Fig. 10: Elbow plot for our k-Means algorithm.

described in pseudo code in algorithm 3. In general the K-means is more autonomous in terms of the size of the clusters thanks to the elbow methodology but the AHC feels more natural when applied to cryptocurrencies mostly because of the concept of "forking" which representation is a tree like the AHC (see Figure 11). We tested two methodologies in this paper: K-means and AHC. Their differences have been highlighted in the table I.

	K-Means	**Hierarchical**
Dataset Size	Large	Small
Method	Heuristic	Agglomerative or Divisive
Grouping	Specify number	No decision needed
Speed	Fast	Can be slow
Cluster number	Automatic	Manual

TABLE I: Comparison of clustering algorithms.

IV. Trading Application

A. Long only strategy

1) Description: in this section we implement the RP methodology to the problem described in section I-B. More specifically we describe what we have labelled as the "dollar hedge strategy" (DHS). The latter, given the context described in section I-B, is composed of the most likely alternatives to the USD, weighted based on the RP principle. In order to encapsulate abstraction we have subdivided this DHS in two sub-strategies: one cryptocurrency related and the other precious metals related. Both abide by the rule of RP. Their volatilities were estimated using rolling standard deviations.

Remark Though not perfect, the advantage of this methodology is to keep the benefit to complexity ratio high. However there are ways to improve the methodology (but increasing complexity at the same time). Incorporating market cap or forecasting volatility using traditional methods [5] are two ideas out of many.

2) Performance: Figure 13 represents the P&L of these two strategies. The overall exposure to the cryptocurrency

Algorithm 2 Merge Sort

Inputs:
A, p, q, r
//———- ▷ A (array), p (left), q (middle), r (right)
Initialize:
None
$n_1 = q - p + 1$
$n_2 = r - q$
Let $L[1 \dots n_1 + 1]$ and $R[1 \dots n_2 + 1]$ be new arrays
for $i = 1$ to n_1 **do**
  $L[i] = A[p + i - 1]$
end for
for $j = 1$ to n_2 **do**
  $R[i] = A[q + j]$
end for
$L[n_1 + 1] = \infty$
$R[n_2 + 1] = \infty$
$i = 1$
$j = 1$
for $k = p$ to r **do**
  if $L[i] < R[j]$ **then**
    $A[k] = L[i]$
    $i = i + 1$
  else if $L[i] > R[j]$ **then**
    $A[k] = R[j]$
    $j = j + 1$
  else
    $A[k] = -\infty$ ▷ We mark the duplicates with the largest negative integer
    $j = j + 1$
  end if
end for

Algorithm 3 Agglomerative Hierarchical Clustering

Inputs:
$C^{1,2,\dots N}_{f_1,f_2,\dots,f_F}$
Initialize:
$\forall (i,j) \in N, i < j,$
$d(C_i, C_j) \leftarrow 0,$
$\mathcal{C}_{1,2,\dots K} \leftarrow C^{1,2,\dots K}_{f_1,f_2,\dots,f_F}$
//——- ▷ Calculate distance between cryptocurrencies
for i = 1 to N **do**
  for j > i to N **do**
    for f = 1 to F **do**
      $D(C^i, C^j) \leftarrow D(C^i, C^j) + \sqrt{(C^i_f - C^j_f)^2}$
    end for
  end for
end for
//———— ▷ Rank Distances and merge accordingly
$S \leftarrow$ **Sort**$(D(.,.))$
//———— ▷ Merge pairwise
for $i = 1$ to N^2 **do**
  Merge()
end for
//———— ▷ Repeat until one cluster

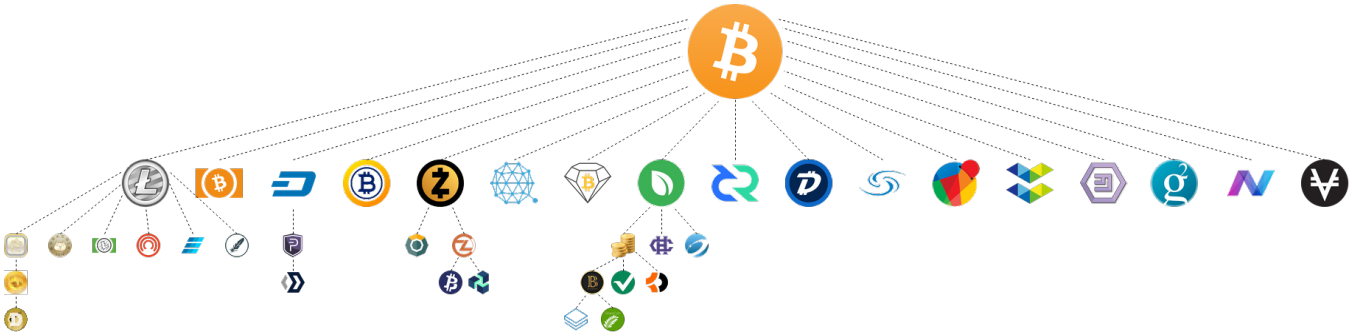

Fig. 11: Bitcoin Fork representation [36].

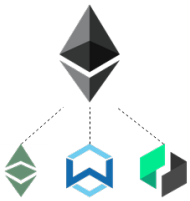

Fig. 12: Etherium Fork representation [36].

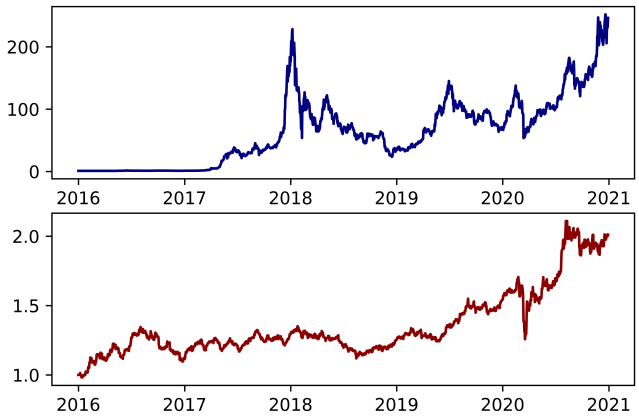

Fig. 13: Crypto (blue) & Precious metals (red) P&Ls.

market is done by weighting the basket to reflect both an overall exposure to the market but also with a small bias towards the most promising of these cryptocurrencies. The precious metal strategy is constructed through a similar methodology. The overall strategy consists of these two sub-strategies dynamically weighted in order to abide by the rules of RP proxied through a mixture of realised recent volatility. Though the independent conjunction of these two strategies yield substantial drawdowns, their combination has had substantial historical returns with more reasonable drawdowns than long only cryptocurrency related strategies (Figure 14). Though this is expected from diversification, it cannot be explained by this alone. More specifically, this is another argument supporting the rational expressed in section I-B. Table II summaries the statistics of this strategy split in a 12 months rolling windows.

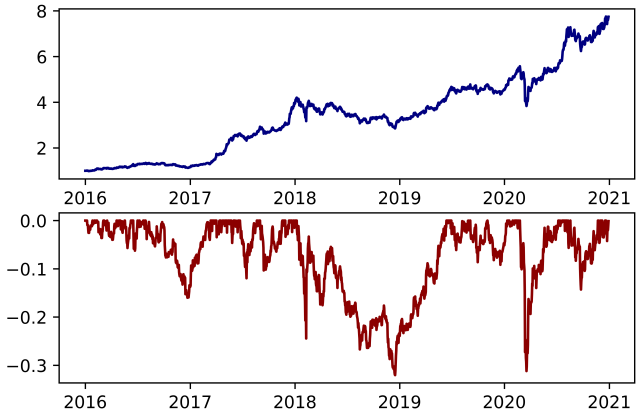

Fig. 14: DHS P&L (blue) & its maximum drawdown (red).

Start	End	Returns	Sharpe Ratio	Drawdown
2016/01/01	2016/12/31	16.2%	0.9	-16.0%
2017/01/01	2017/12/31	126.3%	4.9	-12.0%
2018/01/01	2018/12/31	-15.6%	-0.5	-32.0%
2019/01/01	2019/12/31	37.0%	2.1	-9.2%
2020/01/01	2020/12/31	58.8%	1.8	-31.2%
Yearly Averages		44.2%	1.7	**-20.1%**

TABLE II: Statistical summary of dollar hedge strategy.

B. *Market neutral strategy*

1) Description: we examine in this next section a market neutral strategy, this time composed exclusively of cryptocurrencies. The assumption is that we have already built a coherent momentum strategy (see section III) and would like to leverage on it by applying the signal to different sectors within the crypto space. Note that this is a classic implementation in the equities market[5]. The issue with cryptocurrencies is that they do not have sectors. And organising the categorisation of these potential abstract sectors cannot possibly be achieved using a correlation matrix (as seen in sector I-B). More advanced Machine Learning (ML) methodologies are necessary to achieve this goal. We have implemented the k-Means and the AHC described in section III with the clustered data described in section II. We preferred the AHC

[5]If the cumulative return of a company within a sector changes, perhaps something different is being done in terms of fundamentals within this company. We would like to leverage on that trend until the exit signal.

as its mathematical formalisation resembled the closest to the concept of forking (see Figure 11) but had no idea of what would be the appropriate number of sectors so we used the k-Means there (see elbow plot from Figure 10). Figure 15, 16 and 17 represents 3 dendrograms resulting from this hybrid method. Through this hybrid methodology we are able to

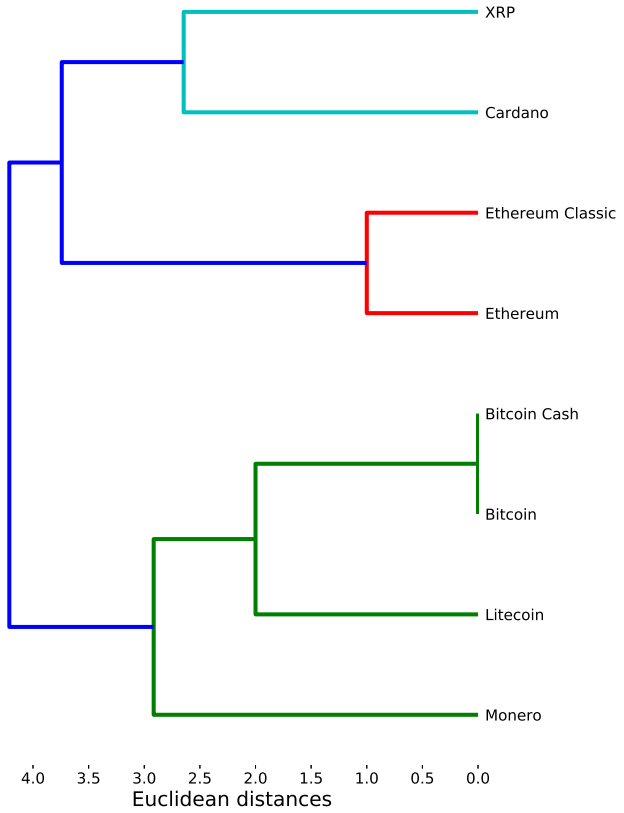

Fig. 15: Reduced Dendrogram.

engineer the proximity of cryptocurrencies with respect to each other and therefore formalise abstract sectors. More specifically, a dendrogram, induced by hierarchical models, can help us classify these various cryptcurrencies based on how far they are from each other. For instance, in Figure 15 we can see that Bitcoin and Bitcoin Cash or Ethereum and Ethereum Classic remain extremely close. When we increase the sample space (Figure 16 and 17) we see that this abstract classification remains coherent. We can also see that the decentralised platforms end up in the same green area. We can also see that Ripple is a sector different of the above two mentioned, which based on the fundamentals, makes a great deal of sense. There seems to be a great deal of overlap with this specific simplistic model [33] laid out in Figure 18.

2) Performance: Figure 19 illustrates respectively the overall P&L (in red) and the maximum drawdown (in blue) of the strategy recomposed with its 3 abstract constituents. The table below summarises the basic statistics of this strategy split yearly rolling windows.

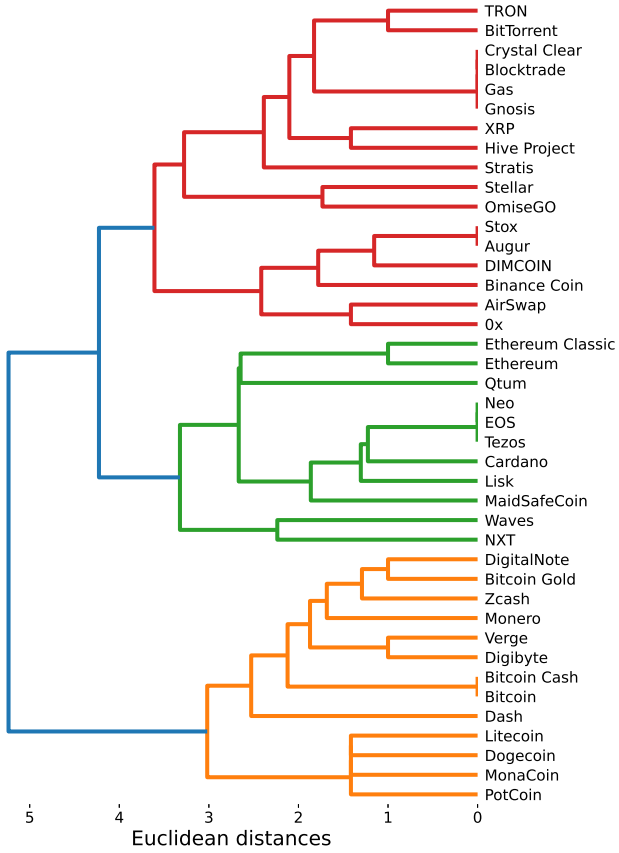

Fig. 16: Fully labelled dendrogram of 40 Cryptocurrencies using hierarchical model fed by the scraping of data contained in few cryptocurrency fundamentals website.

Start	End	Returns	Sharpe Ratio	Drawdown
2018/01/01	2018/12/31	58.0%	2.5	-10.7%
2019/01/01	2019/12/31	37.0%	1.7	-9.6%
2020/01/01	2020/12/31	39.6%	1.9	-17.4%
Yearly Averages		44.9%	2.0	-12.6%

TABLE III: Performance summary of the L/S Strategy.

V. Conclusion

A. Summary

We have shown how Bretton Woods was a key agreement in bringing world monetary stability and how withdrawing from it has launched a chain of reaction that has ultimately propelled Cryptocurrencies to a place where it is now a legitimate asset class. We have exposed one peculiarity of this asset class which is its lack of sectors or at least introduced a methodology in which one could gather this fastly growing list of cryptocurrencies into groups of similar features. A proposal to this issue was given. Namely we gather data, through web-scraping from various cryptocurrency classification sites, merge it into a matrix that we feed into few clustering methodology. Ultimately we decide to use a hybrid method between k-Means and AHC in which we use the number of clusters suggested by the k-Mean but use the classification from AHC. We then apply this

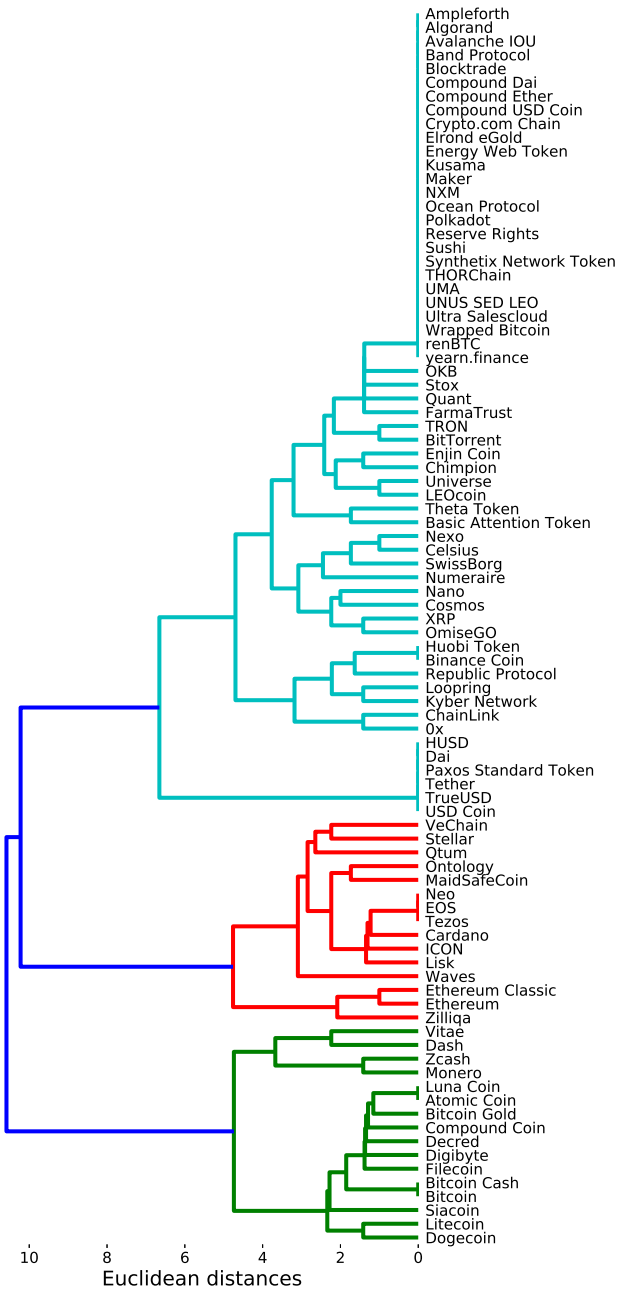

Fig. 17: Larger dendrogram.

alternative data to a couple of traditional strategies. The first strategy we performed the backtests on we labeled the dollar hedge. We have shown that diversification of two random sources of alpha cannot explain on its own the significant reduction in the overall risk. The weighted combination of precious metals (PM) and cryptocurrencies support some of the geopolitical hypothesis (or its perception) raised in the introduction. The second strategy, L/S in nature, has even better results supporting the idea that momentum[6] captures some of the inefficiencies of this new market.

[6]as of early 2021 (it is important to mention the date as inefficiencies in the market are not always consistent.)

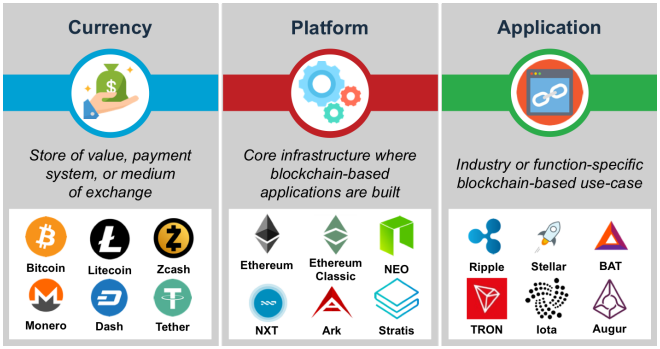

Fig. 18: Categorisation from acceleratingbiz [33].

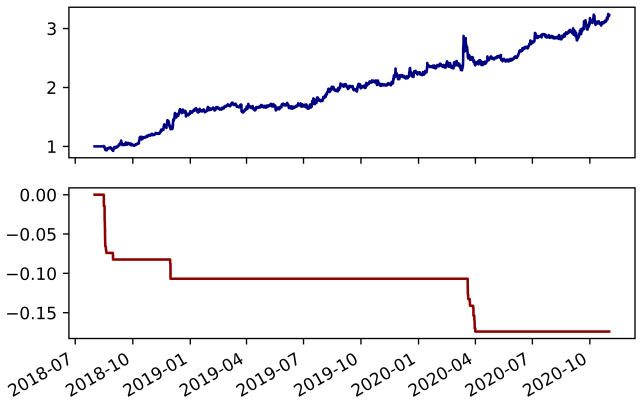

Fig. 19: L/S P&L (in blue) and maximum drawdown (red).

B. *Future work*

Below, we list ideas on how to improve our model:

- Our model input is function of the website from which we scrapped the data from. We have used Messari and Coinlore [2], [1] so far but there exists plenty of other websites
- There are also plenty of videos on youtube and other social medias in which people produce comparative studies on cryptocurrencies fundamentals.
- We have mentioned few other clustering methodologies in section III. It would have been more ideal to compare more clustering methodologies
- within these different clustering methodologies, different measures of distance exists. We have used primarily the Ward measure.
- We have also different number of clusters that are produced through the AHC. We have used the number associated to the k-Means algorithm but this is a heuristic method.
- Though the creation of alternative data was the biggest hurdle for this project, a great deal of improvement could be produced as a result of improving the quality and frequency of our traditional data.
- We used a limited amount of cryptocurrencies. Adding more we bring more confidence in our results.
- The idea of forking was not leveraged on enough. More

specifically incorporation some of the information of Figure 11 will improve our results.

- The AHC methodology is sometimes advertised on forums to do poorly with true and false statements. However we used 1s and 0s in our data. We did not experience this issue with our project. However incorporating an unbiased weighting methodology could become beneficial.

APPENDIX

Table 1.1: Coin Universe pt.1		
Coin	No. Observations	First Observation
ZRX	21383	2018-07-15 02:00:00
AAVE	4310	2020-06-25 11:00:00
ALGO	13200	2019-06-21 01:00:00
AMPL	13031	2019-06-28 02:00:00
AVAX	2177	2020-09-22 08:00:00
BAL	16595	2019-01-30 14:00:00
BNANA	11052	2019-09-18 13:00:00
BAND	21383	2018-07-15 02:00:00
BAT	11012	2019-09-20 05:00:00
BNB	18569	2018-11-09 08:00:00
BTC	21383	2018-07-15 02:00:00
BCH	21383	2018-07-15 02:00:00
BTG	16573	2019-01-31 12:00:00
BTT	10153	2019-10-26 00:00:00
STX	21383	2018-07-15 02:00:00
ADA	11897	2019-08-14 08:00:00
CDAI	19267	2018-10-11 06:00:00
CEL	21383	2018-07-15 02:00:00
LINK	7024	2020-03-04 09:00:00
CETH	4536	2020-06-16 01:00:00
COMP	11422	2019-09-03 03:00:00
CUSDC	16033	2019-02-23 00:00:00
ATOM	17291	2019-01-01 14:00:00
CRO	9584	2019-11-18 17:00:00
DAI	21383	2018-07-15 02:00:00
DASH	21383	2018-07-15 02:00:00

Table 1.2: Coin Universe pt.2		
Coin	No. Observations	First Observation
DCR	21383	2018-07-15 02:00:00
DGB	21383	2018-07-15 02:00:00
DOGE	2629	2020-09-03 12:00:00
EGLD	2422	2020-09-12 03:00:00
EWT	21383	2018-07-15 02:00:00
EOS	21383	2018-07-15 02:00:00
ETH	21383	2018-07-15 02:00:00
ETC	21383	2018-07-15 02:00:00
FIL	1617	2020-10-15 16:00:00
FTT	12273	2019-07-29 16:00:00
SNX	21383	2018-07-15 02:00:00
HT	21383	2018-07-15 02:00:00
HUSD	11032	2019-09-19 09:00:00
ICX	21383	2018-07-15 02:00:00
MIOTA	21383	2018-07-15 02:00:00
KSM	11029	2019-09-19 12:00:00
KNC	21383	2018-07-15 02:00:00
LSK	21383	2018-07-15 02:00:00
LTC	21383	2018-07-15 02:00:00
LRC	21383	2018-07-15 02:00:00
MAID	21383	2018-07-15 02:00:00
MKR	21383	2018-07-15 02:00:00
XMR	21383	2018-07-15 02:00:00
NANO	21383	2018-07-15 02:00:00
XEM	21383	2018-07-15 02:00:00
NEO	21383	2018-07-15 02:00:00
NEXO	21383	2018-07-15 02:00:00
NXM	3902	2020-07-12 11:00:00
OCEAN	14359	2019-05-03 18:00:00
OKB	21383	2018-07-15 02:00:00
OMG	21383	2018-07-15 02:00:00
ONT	21383	2018-07-15 02:00:00
PAX	19647	2018-09-25 10:00:00
DOT	2998	2020-08-19 03:00:00
QTUM	21383	2018-07-15 02:00:00
QNT	21383	2018-07-15 02:00:00
RENBTC	4715	2020-06-08 14:00:00
REN	21383	2018-07-15 02:00:00
RSR	13908	2019-05-22 13:00:00
XRP	21383	2018-07-15 02:00:00
SC	21383	2018-07-15 02:00:00
XLM	21383	2018-07-15 02:00:00
SUSHI	2769	2020-08-28 16:00:00
CHSB	21383	2018-07-15 02:00:00
LUNA	14270	2019-05-07 11:00:00
UST	1939	2020-10-02 06:00:00
USDT	21383	2018-07-15 02:00:00
XTZ	21383	2018-07-15 02:00:00
THETA	21383	2018-07-15 02:00:00
AETH	4308	2020-06-25 13:00:00
ALINK	4283	2020-06-26 14:00:00
AUSDC	1918	2020-10-03 03:00:00
BUSD	21383	2018-07-15 02:00:00
BSV	21383	2018-07-15 02:00:00

Table 1.3: Dropped Coins		
Coin	No. Observations	First Observation
ESD	6396	2020-03-30 13:00:00
HBAR	11080	2019-09-17 09:00:00
HBTC	1701	2020-10-12 04:00:00
LEO	13960	2019-05-20 09:00:00
NEAR	1651	2020-10-14 06:00:00
USDN	21383	2018-07-15 02:00:00
GRT	0	None
WAVES	21383	2018-07-15 02:00:00
WBTC	16584	2019-01-31 01:00:00
YFI	3757	2020-07-18 12:00:00
ZEC	21383	2018-07-15 02:00:00
ZIL	21383	2018-07-15 02:00:00
RUNE	12496	2019-07-20 09:00:00
TRX	21383	2018-07-15 02:00:00
TUSD	21383	2018-07-15 02:00:00
UMA	5672	2020-04-29 17:00:00
UNI	2303	2020-09-17 02:00:00
USDC	19431	2018-10-04 10:00:00
VET	21121	2018-07-26 00:00:00
VITAE	20440	2018-08-23 09:00:00

REFERENCES

[1] Coinlore. https://www.coinlore.com/crypto-sectors, November 2021.

[2] Messari. https://messari.io/screener, November 2021.

[3] Khalid Alkharabsheh, Ibrahim Alturani, Abdallah Al Turani, and Nabeel Zanoon. Review on sorting algorithms a comparative study. *International Journal of Computer Science and Security (IJCSS)*, 7, 01 2013.

[4] M.R. Anderberg. Cluster analysis for applications. *Academic Press, New York*, 1973.

[5] Beth Andrews. Rank-based estimation for garch processes. *Econometric Theory*, 28(5):1037–1064, 2012.

[6] Hubert L.J. Arabie, P. and G. De Soete. Clustering and classification. *World Scientific, Singapore*, 1996.

[7] Clifford S. Asness, Tobias J. Moskowitz, and Lasse Heje Pedersen. Value and momentum everywhere. *The Journal of Finance*, 68(3):929–985, 2013.

[8] Mirkin B. Mathematical classification and clustering. *Kluwer, Dordrecht*, 1996.

[9] Louis Bachelier. Theorie de la speculation. *Annales Scientifiques de l'École Normale Supérieure*, 3 (17):21–86, 1900.

[10] J.P. Benzécri. L'analyse des données. i. la taxinomie. *Dunod, Paris (3rd ed.)*, 1979.

[11] Jean-Philippe Bouchaud. Economics needs a scientific revolution. *Nature*, 455:1181, 2008.

[12] C. de Rham. La classification hiérarchique ascendante selon la méthode des voisins réciproques. *Les Cahiers de l'Analyse des Données*, 5:135–144, 1980.

[13] Mark DeCambre. Jpmorgan chase boss tells fox business network he regrets calling the cryptocurrency a 'fraud'. *Market Watch*, January 2018.

[14] D. Defays. An efficient algorithm for a complete link method. *The Computer Journal*, 20:364–366, 1977.

[15] Jeff Desjardin. Currency and the collapse of the roman empire. *Visual Capitalist*, February 2016.

[16] Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970.

[17] Milton Friedman. Milton friedman predicts bitcoin in 1999, 2018.

[18] Daniel Frumkin. The periodic table of cryptocurrencies. *an overview of the cryptocurrency market*, 2020.

[19] A.D. Gordon. Classification. *2nd ed., Champman and Hall*, 1999.

[20] Amelia Heathman. Move over bitcoin, these countries are creating their own digital currencies. *Money Matters*, September 2017.

[21] Brad Hoff. Hillary emails reveal true motive for libya intervention. *Foreign Policy Journal*, 2016.

[22] Ellis Horowitz and Sartaj Sahni. *Fundamentals of Data Structures*. 1976, 1982, 1983.

[23] Joon Ian Wong and Jason Karaian. Bitcoin keeps hitting record highs, and jamie dimon doesn't want to talk about it. *RAGEQUIT*, October 2017.

[24] A.K. Jain and R.C. Dubes. Algorithms for clustering data. *Prentice-Hall, Englewood Cliffs*, 1988.

[25] M.N. Jain A.K., Murty and Flynn P.J. Data clustering: a review. *ACM Computing Surveys*, 31:264–323, 1999.

[26] Narasimhan Jegadeesh and Sheridan Titman. Momentum. *Annual Review of Financial Economics*, 3:493–509, 2011.

[27] J. Juan. Programme de classification hiérarchique par l'algorithme de la recherche en chaîne des voisins réciproques. *Les Cahiers de l'Analyse des Données*, 7:219–225, 1982.

[28] Arjun Kharpal. Polish central bank paid youtube stars to make a video about a cryptocurrency crash. *CNBC*, February 2018.

[29] Steve Kovach. Tesla buys $1.5 billion in bitcoin, plans to accept it as payment. *CNBC*, February 2021.

[30] TM Lee and Bobby Ong. coingecko API. https://www.coingecko.com/en/api, 2021.

[31] TM Lee and Bobby Ong. coingecko website. https://www.coingecko.com/en, 2021.

[32] I.C. Lerman. Classification et analyse ordinale des données. *Dunod, Paris*, 1981.

[33] Monchester Macapagal. Cryptocurrencies can be broadly categorized as medium for exchange, platform enabler, or means to acquire services. *Cryptocurrency Categories*, May 2018.

[34] Babak Mahdavi-Damghani. Introducing the HFTE model: a multi-species predator prey ecosystem for high frequency quantitative financial strategies. *Wilmott Magazine*, 89:52–69, 2017.

[35] Babak Mahdavi-Damghani. Data-Driven Models Mathematical Finance: Apposition or Opposition? Technical report, 2019.

[36] Killian McGrath and Matt Godshall. The ultimate list of bitcoin and alt-cryptocurrency forks. *Unhashed*, 2018.

[37] F. Murtagh. Multidimensional clustering algorithms. *Physica-Verlag, Würzburg*, 1985.

[38] F. Murtagh and A. Heck. Multivariate data analysis. *Kluwer Academic, Dordrecht*, 1987.

[39] D. Müllner. Modern hierarchical, agglomerative clustering algorithms. 2011.

[40] D. Müllner. Fastcluster: Fast hierarchical, agglomerative clustering routinesfor randpython. *Journal of Statistical Software*, 53, 2013.

[41] Bozhidar Nedev and Boryana Bogdanova. Dynamics of the momentum effect on the nyse from the perspective of behavioral finance. page 020018, 01 2018.

[42] Alex Newman. Gadhafi's gold-money plan would have devastated dollar. *The New American*, 2011.

[43] Paolo Piro, Richard Nock, Wafa Bel haj ali, Frank Nielsen, and Michel Barlaud. *Boosting k-Nearest Neighbors Classification*, pages 341–375. 01 2013.

[44] Jonathan Ponciano. Bitcoin losses near $200 billion as jpmorgan warns it's the 'least reliable' dollar hedge. *Forbes*, January 2021.

[45] F.J. Rohlf. Single link clustering algorithms. 1982.

[46] Xu Rui and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16:645–678, 2005.

[47] Alan Scowcroft and James Sefton. Understanding momentum. *Financial Analysts Journal*, 61(2):64–82, 2005.

[48] R. Sibson. SLINK: an optimally efficient algorithm for the single link cluster method. *The Computer Journal*, 16:30–34, 1973.

[49] Konstantinos Stylianou and Nic Carter. The size of the crypto economy: Calculating market shares of cryptoassets, exchanges and mining pools. *Journal of Competition Law and Economics*, 2020.

[50] Alexandra Ulmer and Deisy Buitrago. Enter the 'petro': Venezuela to launch oil-backed cryptocurrency. *Energy & Enviroment*, December 2017.

Dr. **Babak Mahdavi-Damghani** (BMD) completed his PhD in Machine Learning for Quantitative Finance at the University of Oxford. He has a broad range of work experiences in the financial industry notably having worked for Citigroup, Socgen, Cantab Capital Partner, Credit Suisse, LCH, Oxford Algorithmic Trading Programme and other smaller hedge funds. He is also the author of numerous publications, including cover stories of Wilmott magazine.

Dr. **Robert Fraser** (RF) is a research scientist with expertise in analyzing and predicting the behavior of large data sets. He holds a doctorate in Physics from the University of Oxford, where he used a combination of computational methods and physical theory to investigate a variety of nonlinear systems. He has been integral in developing an equities analytics platform that is now used by several major financial institutions. He has also implemented Machine Learning to design several Alpha Capture based trading strategies.

Mr. **James Howell** (JH) is a Quantitative analyst with a MSc from Imperial College in risk management and financial engineering (core modules: risk & trading strategies). JH also graduated with a BSc in Economics from Bristol University. He also worked for Minter Capital in Australia where he traded Australian interest rate derivatives. JH also consulted for CrossBorder Capital by leading their strategists in implementing machine learning algorithms to make more efficient investment decisions.

Mr. **Jon Sveinbjorn Halldorsson** (JSH) brings his experience in Risk Management, having previously worked in Iceland at Islandsbanki and at ALM Securities. He has also taught Financial Mathematics at Reykjavik University. JSH has done his undergraduate and postgraduate studies in Financial Mathematics (core modules: stochastic calculus, machine learning, statistics) at the University of Warwick where he graduated with a distinction.