Categorizing Cryptoassets: A Return-Driven Cluster Analysis

What does cluster analysis reveal about Bitcoin, Ripple and other large digital assets?

—

Binance Research - April 26th 2019

KEY TAKEAWAYS

- Using a correlation matrix along with hierarchical clustering, digital assets can be grouped into several sub-segments.
- Based on weekly returns, large cryptoassets such as Bitcoin and Ethereum exhibit the highest correlations, but Ripple displays a lower correlation than in our previous report and is an exception as the best diversifier amongst digital assets with a market cap above $3 billion.
- Bitcoin forks (Gold and Cash), Ethereum Classic and Litecoin form a single cluster whereas other potential groups around the following effects:
  - “Binance effect”: Tezos and Dogecoin, two assets not listed on Binance, each form a single child cluster.
  - Potential geographical effects such as a dichotomy between American and Asian cryptoassets.
  - “Coinbase listing effect”: some assets that were reported to be listed or investigated by Coinbase seemed to belong to similar clusters.
  - Some privacy coins (Dash and Monero) form a single cluster.

- On the other hand, performing K-Means clustering on risk-return profiles of each cryptoasset did not return any meaningful results. One potential explanation is that return & volatility profiles are not related to underlying price co-movements over the study period.

In a previous report, we used a cross-sectional method to analyze the internal correlations of the cryptoasset market to observe the cyclical patterns of various assets. We found that low
internal correlations between cryptoassets are often due to idiosyncratic factors, in addition to a coin’s consensus mechanism and a potential “Binance Effect”.

However, the overall correlation observed across the cryptoasset market has increased, which may be due to the rise of stablecoin volume, and corresponding increase in pair offerings, in all cryptoasset markets. In this report, cluster analysis over the twelve previous trading months (March 2018 to March 2019) was conducted to distinguish whether cryptoasset clusters, constructed using an unsupervised learning approach, can be created and more importantly, interpreted.

1. Methodology

“Cluster analysis is a technique used to group sets of objects that share similar characteristics. It is common in statistics, but investors will use the approach to build a diversified portfolio. Stocks that exhibit high correlations in returns fall into one basket, those slightly less correlated in another, and so on, until each stock is placed into a category.”

Source: Investopedia

1.1 Dataset

For the top 30 cryptoassets by market capitalization, USD-equivalent prices were retrieved as reported by CoinMarketCap. Stablecoins are excluded from the analysis along with any cryptoasset that is backed by other assets, whether they are digital or physical. Specifically, 30-day rolling average market capitalizations were computed and the 30 “non-backed” largest cryptoassets (as of March 31st 2019) were selected.

The data collection period covers a full year between March 31st, 2018 and March 31st, 2019.

1.2 Algorithm selection

Cluster algorithms are one of the sub-segments of unsupervised learning algorithms. The table below highlights key differences between two of the most common cluster algorithms, k-means clustering and hierarchical clustering.

---

1 Indices, stablecoins or cryptoassets claiming to be backed by collateral (e.g. gold) are excluded from this analysis.
2 http://www.cs.utah.edu/~piyush/teaching/4-10-print.pdf
### Table 1 - Comparison between K-Means and Hierarchical clustering methods

<table>
<thead>
<tr>
<th></th>
<th>K-MEANS CLUSTERING</th>
<th>HIERARCHICAL CLUSTERING</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of the dataset?</td>
<td>Large datasets</td>
<td>Small datasets</td>
</tr>
<tr>
<td>Selection of amount of clusters?</td>
<td>Predefined manually</td>
<td>Automatic</td>
</tr>
<tr>
<td>Partitionings?</td>
<td>Single</td>
<td>Multiple</td>
</tr>
<tr>
<td>Approach?</td>
<td>Heuristic</td>
<td>Bottom-up (Agglomerative) Top-down (Divisive)</td>
</tr>
</tbody>
</table>

### 1.2.1 Hierarchical clustering methodology

1. Data is pre-processed using a correlation matrix based on 1-year weekly returns, thus representing 52 observations. This correlation matrix captures variable interactions in a normalized way. As a result, it **simplifies calculations by inherently eliminating a lot of irrelevant information.**

2. A dendrogram illustrates **the composition of each cluster by drawing a U-shaped link between a non-singleton cluster and its children.** The height of the top of the U-link represents the distance between his groups of children but also the cophenetic distance between the original observations in the two groups of children.

3. Euclidean distance (i.e. $L^2$ distance) is used and the linkage method relies on the “Ward” method which **minimizes within-cluster variance.**

### 1.2.2 K-means clustering methodology

1. Weekly returns are calculated for the largest 30 assets by market capitalization.

2. From these weekly returns, the annualized volatility is computed along with the annualized average return.

   $$\text{annualized return} = (1 + \text{average weekly return})^{52} - 1$$

   $$\text{annualized volatility} = \text{standard deviation of weekly returns} \times \sqrt{52}$$

3. Data is pre-processed, using a feature scaling, such that each single value is normalized based on the variance($s$) and mean($\mu$):

   $$Z = (X - \mu) / s$$
4. Eventually, the analysis relies on a three-process methodology:
   a. **Initialization**: k-initial centroids are generated randomly.
   b. **Assignment**: k-clusters are created by matching each observation with the nearest centroid.
   c. **Update**: the centroid of the clusters becomes the new mean.

   The second and third steps above are repeated until they converge to a solution that **minimizes the sum of squared errors between points and their respective centroids**.

5. The k-optimal amount of clusters is selected based on the “elbow curve” methodology that selects an optimal value that **minimizes the distance to the centroid** for each center while minimizing the amount of clusters.

2. Results

2.1 Selection of assets

The eligible assets based on the methodology described in 1.1 results in the following digital assets being selected are the following:

**Top 10 cryptoassets selected:**

- Bitcoin (BTC),
- Ethereum (ETH),
- Ripple (XRP),
- Litecoin (LTC),
- EOS,
- Bitcoin Cash (BCHABC),
- Binance Coin (BNB),
- Stellar (XLM),
- Tron (TRX),
- Cardano (ADA).

**Top 11-20 cryptoassets selected:**

- Monero (XMR),
- IOTA,
- DASH,
- Maker (MKR),
- NEO,
- Ethereum Classic (ETC),
- Ontology (ONT),
- NEM,
- Tezos (XTZ),
- ZCash (ZEC).

**Top 21-30 cryptoassets selected:**

- Waves (WAVES),
- Basic Attention Token (BAT),
- Dogecoin (DOGE),
- Bitcoin Gold (BTG),
- Qtum (QTUM),
- OmiseGo (OMG),
- Decred (DCR),
- Lisk (LSK),
- ChainLink (LINK),
- 0x (ZRX).

**Chart 1 - 30-day 10 largest average market capitalizations (USDbn) as of March 31st 2019**

![Chart showing market capitalizations of top 10 cryptoassets as of March 31, 2019.](chart1.png)

Bitcoin, Ethereum and Ripple account for most of the industry total market capitalization.
As explained in the first section, the top 30 digital assets, based on the 30-day mean market capitalization as of March 31st 2019, were considered for this analysis.
As highlighted in our previous reports:

- Correlations are extremely high among large-cap digital assets.
- Ether and Bitcoin also exhibited an extremely high correlation (0.872) between each other.
- **POW assets exhibited higher correlations** with each other than with non-POW assets.
- **Observation of a potential “Binance effect”**: Tezos and Dogecoin, the only two assets not listed on Binance, exhibited lower correlations with other cryptoassets.

However, a few additional observations were noted:

- Dogecoin (DOGE), Tezos (XTZ), Ripple (XRP) exhibited the **lowest correlations with other digital assets** across this one-year period. Notably, Ripple is less correlated in the long-term than what our previous analysis suggested across several 3-month time-periods, using daily returns.
- **Ripple is highly correlated with Stellar (0.73)**. While Stellar was initially built on the Ripple protocol, its code was forked quickly forked and revamped. As of today, Stellar and Ripple code does not rely on the same common core. Yet these two digital assets still share several similarities as they both aim to "reshape the global remittance industry."

2.3 Hierarchical clustering results

Figure 2 - Dendrogram based on correlation matrix (based on squared euclidean distances)

Based on the above dendrogram, some clusters seem to share similar characteristics such as:

- **Geographical affinities**: the popularity of the coins in specific countries and the team’s own location could affect the clusters. For instance, Qtum (QTUM), Cardano (ADA), NEO and OmiseGo (OMG) are projects based in Asia, and most of their coin-holders are located in this region. Similarly, Ripple (XRP), Basic Attention Token (BAT) or Dogecoin (DOGE) are digital assets where most of the teams and investors are located in America.

- **Privacy coins belong to the same sub-cluster** such as DASH and Monero (XMR).

- Stellar (XLM) and NEM are payment systems and they ultimately form a single cluster.

- **“Coinbase listing effect”**: Ripple (XRP) and Basic Attention Token (BAT) are two digital assets that got listed on Coinbase over the study period. Furthermore, Zcoin (ZEC) was

---

3 For instance, Cardano’s ICO was aimed at Japanese investors who bought up to 95% of the total issue. Furthermore, Emergo is based in Tokyo, Japan.  
https://www.worldcryptoindex.com/cardano/  
https://emurgo.io/#/ja
listed on Coinbase in November 2018, and followed by Stellar (XLM) and Maker (MKR) a few months after⁴. These three digital assets reside in a common sub-group as well, supporting the notion that depending on the timing of listings on the same exchanges, coins may exhibit similar trends in the same market conditions.

- **Hard forks and “code forks”**: Litecoin (LTC), Ethereum Classic (ETC), Bitcoin Cash (BCHABC), Bitcoin Gold (BTG) all share a common history, whether it is on the chain or not, with the top 2 largest digital assets: Ethereum (ETH) and Bitcoin (BTC).
  - While Litecoin is not a fork of Bitcoin, its code was initially forked with very little changes, from the Bitcoin Github repository. On the other hand, Bitcoin Cash and Bitcoin Gold were forked from Bitcoin.
  - Conversely, Ethereum Classic and Ethereum shares the same genesis block. Whereas Ethereum Classic may be the original chain, Ethereum has become, by far, the most used and active of the two.
  - Here, Bitcoin Cash is in the same child cluster with Bitcoin Gold whereas Litecoin is grouped with Ethereum Classic.

- **Binance Coin represents its own child group and first-degree parent group** but is in a 2nd degree parent group with EOS, Tron (TRX), Lisk (LSK) and Decred (DCR).

- **Potential “Binance effect”**: Dogecoin (DOGE) and Tezos (XTZ), the only two digital assets not listed on Binance, are each their only component in their child group.

- **“Largest market capitalization”**: Bitcoin (BTC) and Ethereum (ETH), the two largest digital assets, belong to the same sub-cluster.

However there are clear limitations to any of the interpretation above:

- **The timing of the listings on Coinbase does not match**. For example, Ripple(XRP) and Basic Attention Token(BAT) were listed more than five months apart.

- **Matches of Misfits**. Waves (WAVES), Ontology (ONT) and Tezos (XTZ) fundamentally, share little in common. By default, they were grouped together, as they all exhibited lower correlations with other assets.

- IOTA is grouped with Bitcoin (BTC) and Ethereum (ETH), but its market capitalization is not close to any of them. However it is arguable that:
  - Bitcoin: the first Blockchain 1.0 for digital money
  - Ethereum: the first Blockchain 2.0 with the introduction of smart contracts
  - IOTA could be seen as the first Blockchain 3.0 for the Internet of Things

- If EOS, Tron (TRX) and Lisk (LSK) are all on a similar market same segment (“blockchains with smart contracts”), they don’t have much in common with Decred (DCR) that defines itself as an “autonomous digital currency”.

---

⁴ [https://blog.coinbase.com/stellar-lumens-xlm-now-available-on-coinbase-37bc730ec79a](https://blog.coinbase.com/stellar-lumens-xlm-now-available-on-coinbase-37bc730ec79a)
2.4 K-means analysis

**Figure 3 - Elbow curve - Selecting the optimal amount of clusters**

Based on the above figure, the optimal amount of clusters appears to be around 6, as the marginal improvement in the total within-cluster sum of squared distances becomes very small beyond 6 clusters.

As a result, six clusters are selected.

**Table 2. Risk-return profiles for first cluster (A)**

<table>
<thead>
<tr>
<th></th>
<th>BNB</th>
<th>BAT</th>
<th>MKR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annualized return (%)</td>
<td>134</td>
<td>242</td>
<td>186</td>
</tr>
<tr>
<td>Annualized volatility (%)</td>
<td>91</td>
<td>132</td>
<td>116</td>
</tr>
<tr>
<td>Cluster</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>
Table 3. Risk-return profiles for second cluster (B)

<table>
<thead>
<tr>
<th></th>
<th>BTC</th>
<th>LSK</th>
<th>DCR</th>
<th>OMG</th>
<th>QTUM</th>
<th>BTG</th>
<th>ZEC</th>
<th>NEM</th>
<th>ETC</th>
<th>NEO</th>
<th>XLM</th>
<th>ETH</th>
<th>XMR</th>
<th>LTC</th>
<th>DASH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volatility (%)</td>
<td>65</td>
<td>102</td>
<td>97</td>
<td>116</td>
<td>117</td>
<td>106</td>
<td>98</td>
<td>100</td>
<td>94</td>
<td>113</td>
<td>106</td>
<td>96</td>
<td>94</td>
<td>104</td>
<td></td>
</tr>
</tbody>
</table>

Cluster B is constituted by digital asset with high volatility but negative average returns over the period.

Table 4 - Risk-return profiles for the 4 other clusters (C-F)

<table>
<thead>
<tr>
<th></th>
<th>XRP</th>
<th>EOS</th>
<th>DOGE</th>
<th>IOTA</th>
<th>TRX</th>
<th>XTZ</th>
<th>ADA</th>
<th>ZRX</th>
<th>BCHABC</th>
<th>LINK</th>
<th>ONT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average return (%)</td>
<td>29</td>
<td>95</td>
<td>70</td>
<td>-33</td>
<td>55</td>
<td>-8</td>
<td>1</td>
<td>63</td>
<td>29</td>
<td>10</td>
<td>420</td>
</tr>
<tr>
<td>Volatility (%)</td>
<td>136</td>
<td>152</td>
<td>150</td>
<td>130</td>
<td>132</td>
<td>129</td>
<td>128</td>
<td>126</td>
<td>119</td>
<td>192</td>
<td>143</td>
</tr>
<tr>
<td>Cluster</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
</tr>
</tbody>
</table>

Based on a traditional risk-return attributes of cryptoassets, many assets share the same risk-return profiles (clusters B and C). Cluster C is constituted by assets with extremely high volatility but annualized returns either positive or slightly negative (the exception being IOTA).

Clusters C, D and E are made of one single digital asset such as:
- Cluster is constituted by Bitcoin Cash (BCHABC): the second highest-volatility asset in the group but with an extremely small positive return. This risk-profile is explained by the Bitcoin Cash fork (in November 2018) that led to the creation of Bitcoin SV (BSV/BCHSV), creating a common event in the coins’ mutual histories.
- Clusters E and F are each made of a single asset that exhibit extremely high returns:
  - ChainLink (LINK) forms cluster E.
  - Ontology (ONT) forms cluster F.

In general, the drawback of this method is that two assets might belong to the same cluster without having any correlation. Precisely, two assets may display a negative correlation while simultaneously sharing an exact same risk-return profile within the market, resulting in them being part of the same cluster.

---

5 As Ontology weekly returns are not normally distributed (with large extreme positive values), the annualized average return is strongly biased upward.
In the future, our approach must be extended to include additional characteristics such as: market capitalization, volume and turnover ratio.

Native blockchain metrics such as hashrate, active addresses, on-chain transactions or amount of active nodes could also be used as inputs in cluster analysis to analyze what are the underlying dynamics among cryptoassets.

3. Conclusion

Cluster analysis is an unsupervised learning technique that provides flexibility in classification of objects in groups without introducing human bias. For digital assets, cluster analysis has not been extensively studied in research and this report represents one of the first attempts at classifying cryptoassets from an unsupervised approach.

Hierarchical cluster analysis revealed potential groups of cryptoassets such based on characteristics such as asset function (e.g. privacy tokens), chain history (e.g. Bitcoin forks) or a potential dichotomy between Asian and US-based cryptoassets. Eventually the potential existence of a “Binance effect” strikes again along with a newer potential effect related to Coinbase-related news (e.g. Ripple, Basic Attention Token).

Whereas some of the results appear to be consistent with industry-defined fundamental approaches, the difficulty of finding trustworthy data may hold investors back from completing a thorough analysis on this topic. In comparison, traditional equity markets offer plenty of metrics (e.g P/E ratio, turnover, ROE) that are routinely used in research reports.

Additional cluster analysis of the digital asset industry could be performed from different perspectives, and further research on different crypto market cap segments (e.g. mid and small caps) with alternative inputs such as hashrate or on-chain transactions may help paint a more complete picture for the cryptoasset market as a whole.

Disclaimer
This communication has been prepared solely for informational purposes and is subject to change. This communication should not be relied upon as the basis for making any investment decision or be construed as a recommendation to engage in any transaction or be construed as a recommendation of any investment strategy. Reports prepared by Binance Research, a division of Binance, are not related to the provision of advisory services regarding investment, tax, legal, financial, accounting, consulting or any other related services and are not recommendations to buy, sell, or hold any asset. Binance Research does not provide reports on companies that issue securities or on securities generally; rather this report relates solely to digital assets that are not securities. This communication is not a personal recommendation and does not take into account whether any transaction is suitable for a particular person. The information contained in this communication is based on sources considered to be reliable, but not guaranteed, to be accurate or complete. Binance Research is not responsible for information obtained or derived from third-party sources or established services. Any opinions or statements expressed herein reflect a judgment made at the time of this data, and are subject to change without notice. Neither Binance Research nor any of its directors, officers, employees, representatives or agents accepts any liability whatsoever for any direct, indirect or consequential loss (whether civil, contract, tort or otherwise) arising from the use of this communication or its contents or reliance on the information contained herein.

Binance Research provides in-depth analysis and data-driven insights of digital assets by generating unbiased, institutional-grade research reports for investors in the crypto space.