Technology and Cryptocurrency Valuation*

Yukun Liu Jinfei Sheng Wanyi Wang

November 2022

Abstract

While various theories stress the importance of technology for cryptocurrency valuation, empirical evidence is limited. In this paper, we study whether technology aspects of cryptocurrencies matter for their valuations, using machine learning methods to construct a technology index from initial coin offering whitepapers. We then track down the performance of cryptocurrencies from their initial offering to long-term valuations. We find that the cryptocurrencies with high technology indexes are more likely to succeed and less likely to be delisted subsequently. Moreover, the technology index strongly and positively predicts the long-run performances of cryptocurrencies. Overall, the results suggest that technological sophistication is an important determinant of cryptocurrency valuations.

Keywords: Cryptocurrency, Technology, Machine Learning, Textual Analysis, FinTech, Blockchain.

*Yukun Liu is with Simon School of Business, University of Rochester. Jinfei Sheng (Corresponding Author, Email: jinfei.sheng@uci.edu) and Wanyi Wang are with Merage School of Business, University of California, Irvine. For helpful comments, we thank Ben Charoenwong (discussant), David Hirshleifer, Chong Huang, Arthur Iinuma, Alan Kwan, Jongsub Lee (discussant), Ye Li, Chuchu Liang, Fangzhou Lu (discussant), Evgeny Lyandres (discussant), Feng Mai, Asaf Manela, Kenny Phua (discussant), Daniel Rabetti (discussant), Amin Shams, Yushui Shi, Donghwa Shin (discussant), Siew Hong Teoh, Aleh Tsyvinski, David Yang, Lu Zheng, Chenqi Zhu, and conference and seminar participants at University of California, Irvine, 2019 Conference on Financial Economics and Accounting at New York University, 2020 American Finance Association Conference Poster Session, 2020 CAFR Research Workshop on FinTech, 2nd Future of Financial Information Conference, 4th Shanghai-Edinburgh Fintech Conference, 2021 Miami Research Conference on Machine Learning and Business, UWA Blockchain and Cryptocurrency Conference, 2021 Global AI Finance Research Conference, 2022 Hong Kong Conference for Fintech, AI, and Big Data in Business, and 2022 Asian Finance Association Annual Conference. All errors are our own.

1 Introduction

The rise of blockchain technology is one of the most critical innovations in recent decades. An early application of blockchain technology that has received much attention is cryptocurrency, which has experienced exponential growth since the debut of Bitcoin in 2009. To date, there are over 10,000 cryptocurrencies with a market capitalization of over 0.8 trillion dollars.[1] The rapid growth of cryptocurrency market sparks extensive debates among practitioners, policy makers, as well as academia. On the one hand, there are concerns about whether there is any fundamental value of cryptocurrencies. Speculations, price manipulations, and frauds in this space are prevalent. For example, Satis—a security token advisory firm—claims that over 80 percent of initial coin offerings (ICOs) in 2017 were scams.[2] Recent papers also find evidence of price and volume manipulations in the cryptocurrency market (Griffin and Shams, 2020; Cong et al., 2020). On the other hand, many cryptocurrency investors believe that blockchain technology is an important innovation and has intrinsic value, and cryptocurrencies represent a stake in the future of this technology.

For companies, we typically use dividends, earnings, or book value to measure their fundamental value. However, cryptocurrencies do not distribute dividends, and there is no traditional accounting information readily available. Cryptocurrencies are also different from fiat currencies in the sense that their value is not backed by any government. Therefore, it is difficult to evaluate the value of cryptocurrencies in traditional frameworks. Recent theories in cryptocurrency address these differences by emphasizing on the technology sophistication in determining the viability and valuation of coins (see e.g., Fanti et al. (2019); Irresberger et al., 2020; Iyengar et al., 2020). However, little is known whether technology aspects of cryptocurrencies matter for their valuations from the empirical side. In this paper, we develop technology indexes to measure the technology sophistication of individual cryptocurrencies and study whether investors value the technology aspects of cryptocurrencies.

[1] This is based on information from coinmarketcap.com.

[2] For the full report, please see: https://research.bloomberg.com/pub/res/d28giW28tf6G7T_Wr77aU0gDgFQ.

Measuring the technology sophistication of cryptocurrencies is challenging because of the limited information disclosed. The only widely adopted disclosures for cryptocurrencies are their whitepapers during initial coin offerings (ICOs). To attract funding, developers need to carefully describe all aspects of the initial coin offerings especially the blockchain technology employed. This feature of whitepaper gives us an unique opportunity to evaluate the technological sophistication of coins at the individual level. Following the ICOs, we can also observe the outcome and valuation of the cryptocurrencies. Therefore, the ICO market is an ideal laboratory to study investors' valuation of cryptocurrency technologies.

To measure the technological components employed in the cryptocurrencies, we use textual analysis to analyze the content of whitepapers. In particular, we use a machine learning method, word embedding, to construct a Tech Index. Our Tech Index measure is validated by other existing measures. We study the determinants of the tech index using various cryptocurrency characteristics. We find that cryptocurrencies that just use the Ethereum blockchain, have lower GitHub activities, have ambiguous whitepapers, and have less reliable teams tend to have lower tech indexes. However, the R-squared is only 0.131 when we use all the cryptocurrency characteristics, suggesting that the majority of the variation in the tech index is not captured by these characteristics.

To understand the role of technological sophistication in cryptocurrency pricing, we start by studying the relationship between the Tech Index and ICO successes. We first examine whether the technology index is related to ICO fundraising. If the entrepreneurs cannot raise any funding, the ICO is not likely to succeed, so the ability to raise funding is one of the most important steps in a successful ICO. If ICO performances are fully driven by speculations, investors would not care about the technology associated with the ICOs. Under this hypothesis, the technology index would not predict ICO successes. However, we find that ICOs with high technology indexes are more likely to raise capital and more likely to be traded in the secondary market subsequently. The economic magnitude of the effect is significant. For instance, a one standard deviation increase in the Tech Index is associated with a 10.7 percent increase in the listed probability, which is a 41

percent increase of the average. The results suggest that investors take the underlying technology of the ICOs into consideration.

Next, we investigate whether the underlying technology of cryptocurrencies is associated with subsequent performances. The process to fully incorporate technology-related information may take months due to the complexity of blockchain technology. To test this conjecture, we examine the relationship between the technology index and the long-run performance of ICOs. We measure long-run performance using cumulative post-ICO returns, abnormal returns, and liquidity measures. We find that the cryptocurrencies with higher technology indexes tend to have better performance in the long run compared to other cryptocurrencies. A one standard deviation increase in the Tech Index is associated with a 19.5 percent increase in cumulative returns at the 180-day horizon.

We also investigate whether our index helps understand cryptocurrencies failure measured by delisting. We find that the cryptocurrencies with higher technology indexes are less likely to be delisted subsequently. The economic magnitude of the effect is also large. For instance, a one standard deviation increase in the technology index leads to a 2.5 percent decrease in delisting probability, which is 25 percent of the average.

So far, we have shown that the technology index strongly and positively predicts ICO successes and subsequent performances. We argue that the results are consistent with the notion that investors care about the technological sophistication of the cryptocurrencies, but it takes time for the market to incorporate the information, leading to predictable returns. We present additional evidence in support of the delayed reaction mechanism and attempt to rule out potential alternative explanations. An implication of the delayed reaction mechanism is that investors should be able to quickly incorporate the fundamental information if the whitepapers are written clearly. Consistent with the implication, we show that among whitepapers with better readability, the long-horizon predictive power of the technology index is weaker. We also find that there is no return reversal phenomenon, suggesting that the return predictability results are unlikely to be driven by investor overreactions.

Overall, these results suggest that the underlying technology is an important determinant of cryptocurrency prices, and support the argument that investors do take the technological components in the ICO whitepapers into their consideration. However, it takes time for investors to differentiate the fundamentally sound cryptocurrencies from the others fully. The delayed reaction from investors may be caused by investor inattention and the complex nature of the technologies, both of which necessitate more time to process related information. Our results are robust to using other machine learning methods (e.g., LDA and supervised machine learning).

This paper contributes to the fast-growing literature on the economics of cryptocurrencies and digital assets in general. Yermack (2017) is the first paper to explore the financial implications of blockchain. Liu and Tsyvinski (2018) provide one of the first comprehensive analyses of the risk-return tradeoff of cryptocurrencies. Liu et al. (2019) examine the cross-section of cryptocurrency and establish a cryptocurrency three-factor model. Recently, several theoretical papers examine the rationale and mechanisms of ICOs and cryptocurrencies (Cong and He, 2019; Cong et al., 2019; Catalini and Gans, 2018; Sockin and Xiong, 2018). Our paper is closely related to Cong et al. (2019) and Sockin and Xiong (2018), which argue that the value of cryptocurrency is fundamentally anchored by the underlying utility value. In other words, their models predict that coins have fundamental values and the fundamental values are crucial for performance. However, there is little evidence showing the importance of the fundamental values of coins because it is hard to measure that empirically. A set of empirical papers study factors that contribute to ICO success, including Howell et al. (2020), Deng et al. (2018), Lee et al. (2019), and Davydiuk et al. (2022). Lyandres et al. (2020) studies the determinants of ICO sucesses and performances, and overturn some existing findings in the literature. In general, they find social media and team play a significant role in ICO success and performance. Although some prior papers touch about whitepapers (e.g., Dittmar and Wu, 2019 and Florysiak and Schandlbauer, 2019), our paper is the first paper that tries to measure the technological sophistication of cryptocurrencies using various machine learning methods and account for the relationship between whitepapers with ICO short and long-

run performance.[3] Our tech indexes appear to play a significant role in explaining ICO success, short-, and long-horizon performance, all of which are not well understood in the literature.

This paper provides support to the theoretical literature that links the technological advances of blockchain to the fundamentals and valuations of cryptocurrencies. Budish (2018), Abadi and Brunnermeier (2018), and Hinzen et al. (2019) discuss the limitations of proof-of-work technologies and the pricing implications of them. Fanti et al. (2019) show that the pricing implications of proof-of-stake. Consistent with the theoretical implications of the literature, our paper shows that the technological components affect the valuations of cryptocurrencies.

Our study also adds to the literature on machine learning and textual analysis in finance.[4] The application of machine learning in finance is a new and growing literature. Existing studies use machine learning methods to construct text-based uncertainty (Manela and Moreira, 2017), predict stock returns (Gu et al., 2020), measure corporate culture (Li et al., 2019), and analyze online reviews (e.g., Sheng, 2019). Buehlmaier and Whited (2018) measure firms' financial constraints using textual analysis of firms' annual reports. Kelly et al. (2018) use textual analysis to construct indicators of patent quality. Recently, Bybee et al. (2020) use machine learning to measure the state of the economy via textual analysis of business news. To our best knowledge, this paper is the first paper to use machine learning methods to conduct textual analyses of cryptocurrencies. Our findings from the crypto market adds new insights to the existing literature.

The rest of the paper is organized as follows. Section 2 describes the data we use. Section 3 introduces the construction and the validation of the technology index. Section 4 describes our main empirical results and Section 5 provides explanations. Section 6 documents the robustness and additional results. We conclude and discuss policy implications in Section 7.

[3]Some studies also look at the text of social media about cryptocurrencies. For example, Shams (2019) use text from Reddit to measure the connectivity among cryptocurrencies.

[4]See Tetlock (2014) and Gentzkow et al. (2019) for reviews on textual analysis. Textual analysis includes both machine learning methods and other methods, such as word count. For recent studies using the word count method, please see Liu and Matthies (2018) and Fisher et al. (2020).

2 Background and Data

2.1 Data

Our dataset consists of three different components: cryptocurrency data from coinmarketcap.com, ICO characteristics from trackico.com, and textual measures constructed from ICO whitepapers. We focus on coins issued through ICO between January 2017 and December 2018. The final sample consists of 2,916 coins, which raised more than $17 billion in total. For each coin, we collect the following information: ICO start and end date, ICO price, total capital raised, trading status, pre-ICO, bonus, platform, accepted currency, the founder team, country, industry, links of whitepapers, official website, GitHub and Twitter. Next, we merge ICO characteristics with cryptocurrency trading data from CoinMarketCap. CoinMarketCap is one of the most comprehensive price-tracking website for cryptocurrencies. By the end of 2018, CoinMarketCap provides data for over 3,600 cryptocurrencies, among which 2,070 are active and 1,583 are delisted. We collect daily opening price and 24h dollar trading volume on all coins from August 2013 to December 2018. We then use token names, ticker symbols, and website slugs to merge these variables with our ICO data. Since many coins on coinmarketcap.com were not issued through ICO, and many ICOs do not list their coins on any exchange, we get a merged sample of 765 cryptocurrencies.

We define two measures of ICO success. The first one is "CMC Trading", a dummy that equals one if a coin is listed on cryptocurrency exchanges after the ICO. The second one is a dummy that equals one if an ICO successfully raised any capital (Benedetti and Kostovetsky, 2018). Other ICO characteristics serve as control variables. "ICO length" is the number of days between the start and end of an ICO. "ICO price" is the cost per token in US dollars. "Total Raised" is the amount of money raised in millions of US dollars. "Pre ICO", "Bonus", "Ethereum Based" and "Accept BTC" are indicator variables about whether the ICO has a pre-ICO, offers bonus to investors, is built on Ethereum platform and accepts Bitcoin as a payment currency, respectively. "Team size" is calculated as the number of team members. We define "Has GitHub" and "Has Twitter" to be

indicator variables of whether the fundraiser has a GitHub or a Twitter homepage. We further control for Bitcoin price on the ICO start date or the coin's listing day as a proxy for the market sentiment. Finally, we control for quarterly, categorical and geographical (continent-level) fixed effects.

We define "First Open/ICO Price" to measure the premium on the listing day and "Delist" to characterize whether the coin is delisted from cryptocurrency exchanges. We also calculate the cumulative rate of return, Bitcoin-adjusted rate of return and 24h trading volume after the coin has been listed for 7 days, 30 days, 90 days, 180 days, 240 days and 300 days. These measures capture the short- and long-term performance and liquidity of cryptocurrencies.

The last set of variables comes from textual analysis of ICO whitepapers. Companies voluntarily disclose whitepapers to communicate with investors in the fund-raising stage, and one of the primary ways that investors evaluate coins is through whitepapers. We successfully downloaded 1,629 valid whitepapers and Table OA.1 lists why we could not obtain the whitepaper of the remaining tokens. Next, we convert PDF files into TXT format, so that it can be used as the raw input for textual analysis.Using this whitepaper corpus, we first construct our main measure of technology, which we explain in detail in Section 3. Moreover, we consider three well-known textual measures as control variables: Readability, Tone, and Uncertainty. "Readability" is characterized by the Fog Index, a widely adopted measure in finance and accounting literature. Developed by Robert Gunning in 1952, Fog Index is a linear combination of the percentage of complex words and the average number of words per sentence.[5] "Tone" is the difference between positive and negative words divided by the total number of words, and "Uncertainty" is the percentage of uncertainty words among all words used in a whitepaper. All lexical categories are defined in Loughran and McDonald (2011).

[5]The complete formula of Fog Index is: Fog Index=0.4[(words/sentences)+100((complex words)/words)]. "Complex words" are words consisting of three or more syllables.

2.2 Summary Statistics

We report the summary statistics of the sample characteristics in Table 1. Panel A of Table 1 presents summary statistics on variables related to ICO characteristics. On average, it takes 51 days to complete an ICO with a team of 11 people. 18% of the ICOs are self-reported as trading and 38% have non-zero values of capital raised. Moreover, 60% have a GitHub homepage for their project and over 90% have set up their Twitter accounts.

Panel B of Table 1 presents summary statistics on the merged sample. Consistent with the literature, we identify that 26% of ICOs have listed tokens on an exchange at some point in time. Among these listed cryptocurrencies, only 10% are delisted while the remaining 90% are still active. On average, investing in a cryptocurrency during an ICO can earn a premium of 120% on the first trading day, indicating a large amount of first-day price reaction. Moreover, the return of cryptocurrency investment increases as time goes by, from 19% during a 7-day holding period to 151% during a 300-day holding period. The 24h trading volume fluctuates with different time spans, varying from $1.5 million to $2.78 million. ICO characteristics with respect to the merged subsample are also reported in this panel.

3 Measuring the Technology Aspect of Cryptos

In this section, we discuss how we measure the technologyaspect for the cryptocurrencies based on their whitepapers. We first present how we construct the technology index using machine learning methods, and then we validate the measure.

3.1 Measure Construction

We use machine learning techniques to capture the technological components of crypto whitepapers. Specificially, we use an unsupervised machine learning method, word embedding, to measure the technological aspect of crypto whitepapers. One important advantage of unsupervised machine

learning methods is that they require little human input. In other words, they do not require researchers to have good prior knowledge about what type of words they are looking for in the texts. The results are robust to using other methods, such as Latent Dirichlet Allocation (LDA) and a supervised machine learning method (see Section 6.1). Word embedding is one of the most popular word representation methods in natural language processing (NLP) in recent years. Developed by Mikolov et al. (2013), its goal is to map words to numerical vectors, such that the semantic similarity between words is captured by the geometric distance in the vector space. How to construct such vectors? The intuition comes from the famous quotation of Firth (1957)—"You shall know a word by the company it keeps." In other words, the meaning of a word can be inferred from the context, so words appearing in similar contexts should have similar meanings.[6] Word embedding has two main advantages over traditional "bag-of-words" methods. First, it greatly reduces the number of dimensions. Word embedding vectors usually have only a few hundred dimensions, while bag-of-words models are typically sparse vectors of thousands of dimensions. Hence, it is a more efficient representation of the raw text. Second, word embedding maps synonyms to adjacent vectors, so we can use clustering methods on the vector space to divide words into different topics. We use K-means as the clustering algorithm. It is one of the simplest and most popular unsupervised machine learning methods. Given a fixed number of clusters (k), K-means seeks a partition of the dataset, such that the within-cluster sum of squared distances between each observation and its closest centroid is minimized. In the Internet Appendix, we provide details on the theoretical background of word embedding and k-means clustering and how to choose the optmial number of topics.

We find that the optimal number of topics detected by algorithm is 20. Hence, we use K-means to cluster word embedding vectors into 20 topics. Topics are mutually exclusive, so each word can only be grouped into one topic. To understand these twenty topics, we look at top words associated with each topic. This is a common approach adopted in most finance and economics literature (e.g., Hansen et al., 2018; Sheng, 2019). Table OA.2 lists the top 15 most frequent terms of each topic.

[6] Li et al. (2019) provide a good example in the Appendix to illustrate the intuition.

We assign a label to each topic based on these key terms.

To further understand the relationship between topics, we apply two machine learning techniques. The first one is hierarchical agglomerative clustering (Murtagh and Legendre, 2014), which can be used to construct a taxonomy of our topic model. Following Bybee et al. (2020), we agglomerate topics recursively according to the semantic similarity between topics, as captured by the distance between cluster centroids. Figure 1 displays the result and shows that three topics ("blockchain", "system", and "algorithm") belong to the same cluster. Another technique we use is multidimensional scaling (MDS, Torgerson, 1958), which is a non-linear dimensionality reduction algorithm such that the two-dimensional representation best preserves the distance between topics in the original space. As shown in Figure 1, "blockchain", "system" and "algorithm" are also adjacent to each other in the inter-cluster distance map. Therefore, we consider these three topics as technology-related topics. For each whitepaper, we calculate the percentage of words that belong to the "blockchain", "information" or "algorithm" topic, normalize it to zero mean and unit standard deviation, and define it as our Tech Index.

Given that this measure is new, we first validate it with other existing measures. One common measure of technology in cryptocurrency is information from GitHub. GitHub is an open-source online platform that provides repository hosting service for developers. Using data from GitHub, we obtain the number of (1) users subscribing updates of the repository (*watch*), (2) "likes" received by the repository (*star*), (3) copies made by other developers (*fork*), (4) code revisions (*commit*), (5) pointers to specific versions (*branch*), and (6) developers who have contributed to the source code (*contributor*). These measures are often used by researchers to proxy for product quality and post-ICO technology development (Deng et al., 2018; Dittmar and Wu, 2019). We compare GitHub indicators with our Tech Index. Table 2 shows that our Tech Index is positively correlated with these GitHub variables, suggesting that our Tech Index provides a good capture of the technology aspect of cryptocurrency. Compared to GitHub measures, our Tech Index is rich and captures important dimensions of the technology used in the crypto that are different from those GitHub variables.

3.2 Measure Determinants

To better understand the Tech Index, we study its determinants. We utilize cryptocurrency characteristics from several dimensions, including whether they use Ethereum blockchain, GitHub data, whitepaper information, and other characteristics. Table 3 documents the results that relate the Tech Index to these cryptocurrency characteristics. We use the Tech Index as the dependent variable. Each of columns (1)–(4) reports the determinant models based on a dimension of coin characteristics. Column (1) shows that cryptocurrencies that use Ethereum blockchain tend to have lower tech index, confirming the prior that cryptocurrencies that build their own blockchain on average have higher tech indexes. Column (2) shows that cryptocurrencies with more code revisions in GitHub have higher tech indexes. In Column (3), we find that cryptocurrencies with ambiguous whitepapers tend to have lower tech indexes. This is consistent with economic intuition because cryptos with more complicated technologies need to explain the use of technologies, which can be complex and hard to understand (i.e., low readability Fog score). In Column (4), we find that cryptocurrencies with more reliable and supportive teams have higher tech indexes. For example, team size positively predict tech indexes. Column (5) combines all the cryptocurrency characteristics and delivers consistent messages. However, the R-squared of Model (5) is 0.131, suggesting that the majority of the variation in the Tech Index is not captured by the cryptocurrency characteristics.

4 Main Results

In this section, we examine whether the Tech Index of cryptocurrency is associated with their short-term and long-term valuations. We examine several dimensions: ICO success at the fund-raising stage, short-run, and long-run returns.

4.1 ICO Success

First, we study the set of characteristics in ICO whitepapers that are most related to ICO success. We use two ways to measure ICO success. The first measure of ICO success is based on whether the cryptocurrency is listed on the coinmarketcap.com (CMC trading) and the second measure is based on whether the ICO successfully raised capital. If the entrepreneur cannot raise any funding, the ICO is not likely to succeed. Therefore, the ability to raise funding is one of the most important steps in a successful ICO. If investors care about the technological components of cryptos, we should expect that it is easier for cryptos with higher Tech Index to raise funding.

Table 4 documents the results that relate ICO whitepapers' characteristics to ICO successes. Table 4 column (1) - (2) present results based on CMC trading and column (3) - (4) present results based on whether the cryptocurrency successfully raised capital. We report coefficient estimates for the Tech Index as well as the control variables. Time, categorical, and geographic fixed effects are included in the specifications when indicated.

The first two columns show that the CMC trading indicator positively loads on the Tech Index, suggesting that when the Tech Index is high, the cryptocurrencies are more likely to be listed on coinmarketcap.com. The relationships are highly significant at the 1 percent level, with a coefficient estimate of 0.107. The economic magnitude is large. A one standard deviation increase in the Tech Index leads to an increase of the listed probability by 10.7 percent, which is a 41 percent increase of the sample average of the listing probability. In the multivariate specification with controls and fixed effects, the coefficient estimate on the composite technology index is 0.048. That is, a one standard deviation increase in the composite index is associated with an increase of the listed probability by 4.8 percent under the multivariate specification.

Column (3) - (4) measures ICO success based on whether the ICO raised capital (Success indicator). The coefficient estimates are largely consistent with the first two columns—the coefficient on the Tech Index is highly statistically and positively significant at the 1 percent level. The coefficient estimate is 0.077 in the univariate regression, which suggests that a one standard deviation

increase in the composite index is associated with a 7.7 percent increase of the probability that the ICO raised capital—a 20 percent increase of the sample average. In the multivariate specification with controls and fixed effects, the coefficient estimate on the Tech Index is 0.043. That is, a one standard deviation increases in the composite index is associated with a 4.3 percent increase in the probability that the ICO raised capital.

Further evidence that the technological component serves as an important factor for ICO success is the R^2. For example, Table 4 column (1) shows that a single variable of the Tech Index already explains 6 percent of the variation of CMC trading. Overall, the results show that when an ICO whitepaper contains more discussion on technology-related topics as captured by our index, the ICO is more likely to be successful. This suggests that technology is one of the most important factors that contribute to the success of an ICO at the fund-raising stage.

Industry Subsample

In addition, we test whether the Tech Index is a stronger predictor of ICO successes in industries that technological components are deemed more important. Table OA.3 provides a list of ICO industries. In certain industries (e.g., platform; trading), investors may scrutinize the technological components of the ICOs, while in other industries (e.g., gaming; charity), this is not the case. We categorize "platform", "cryptocurrency", and "trading" as the technology-related industry, and construct an indicator variable ("industry") to denote the technology-related industries. We test whether the technology index strongly predicts ICO successes for coins in the technology-related industries.

We present the subsample results based on industries in Table 5. Consistent with the baseline results, the Tech Index positively predicts ICO successes. The cross-terms between the technology index and the "industry" indicator are all positive and largely significant. The economic magnitude is large. The coefficient estimate doubles for the coins in the technology-related industries relative to the rest of the coins (Column (2)). These results also support the view that investors value the technological components of the cryptocurrencies, especially for the coins in the technology-

related industries.

4.2 Short- and Long-Horizon Performance

In this section, we investigate whether cryptos with higher a Tech Index tend to perform better in the short-run and long-run, measured by returns after ICO. In the equity market, initial public offerings tend to underperform in the long run (see Ritter, 1991; Loughran and Ritter, 1995). In sharp contrast, initial coin offerings perform well in the medium- to long-horizon (see Benedetti and Kostovetsky, 2018).

For our test, we look at whether the Tech Index predicts the subsequent performance of ICOs. We track the subsequent returns of ICOs over different horizons—from 7-days ahead to 300-days ahead. We regress the cumulative ICO returns on current Tech Index, controls, and fixed effects. The results are documented in Table 6 Panel A. There are three notable findings. First, we find that the Techn Index positively predicts the subsequent performances of the ICOs. Second, the point estimate steadily increases but are insignificant at short horizons. Finally, the point estimate starts to become significant in longer horizons. At the 180-day horizon, the point estimate increases to 0.195 and is statistically significant. The economic magnitude is large. It indicates a 19.5 percent increase in cumulative returns at this horizon for one standard deviation increase in the Tech Index, which is 19.5 percent increase of the sample mean. At the 240-day and 300-day horizons, the point estimate increases to 0.331 and 0.377, respectively.

A common factor that is important for the crypto market is Bitcoin returns. The ICOs took place at different times, and may be affected by the Bitcoin returns in many ways. Thus, we conduct a similar exercise with abnormal returns that are adjusted to Bitcoin returns. Table 6 Panel B reports the results of this test. It shows similar results in terms of statistical significance and economic magnitude as in Panel A. The point estimates become statistically significant at the 180-day horizon. The economic magnitude is large. It indicates a 21.6 percent increase in cumulative returns at this horizon for one standard deviation increase in the Tech Index. At the 240-day and 300-day horizons, the point estimate increases to 0.319 and 0.407, respectively.

Overall, the short- and long-horizon results are consistent with the idea that it takes time for the market to fully incorporate information about technology used in cryptos. Although cryptos with higher technology scores have a high probability of raising funds, investors undervalue these high-tech coins on average in the short-run. Investors eventually appreciate these cryptos, reflected on better long-term returns. We formally test this conjecture in Section 5.

4.3 Liquidity and Delisting

In this section, we use two additional measures to evaluate ICO performances. The first one is the liquidity measure and the second one is the delisting probability measure.

We measure coins' liquidity as the log transformation of the 24-hour trading volume. On average, we find that liquidities are higher for older coins, consistent with Howell et al. (2020). We examine the relationships between characteristics of whitepapers and coins' liquidity measures. We report the results in Table 7. In our model specifications, we include quarterly, categorical, and geographic fixed effects. We find that the technology index is positively associated with coin liquidity. These results are always statistically significant across different horizons since inception.

We then investigate the relationships between coins' delisting probability and the characteristics of the whitepapers. We define Delist as an indicator variable, which is equal to 1 if a token is delisted from CMC. The results are reported in Table 8. The results show that coins with high technology scores are less likely to be delisted subsequently. The economic magnitude of the effect is large. For instance, in the standalone specification, a one standard deviation increase in the Tech Index leads to a 2.5 percent decrease in delisting probability, which is 25 percent of the average.

The results in this section highlight that coins with high technology scores are intrinsically superior. The results provide supports to our argument that the investors in the coin market take technical aspects of the ICOs into consideration. However, as we have shown above, it takes a considerable amount of time for the market to reach the proper pricing of the ICOs eventually.

5 Explanations

In the previous section, we show that the Tech Index strongly and positively predicts ICO successes and subsequent performances. We argue that the results are consistent with the notion that investors care about the technological aspect of the cryptocurrencies, but it takes time for the market to incorporate the information leading to predictable returns. In this section, we test this conjecture and present evidence in support of the delayed reaction mechanism and attempt to rule out potential alternative explanations.

5.1 Information Processing

In the main result section, we argue that the findings are consistent with the investor delayed reaction to technological aspect of cryptos since it is complex. It would be relatively easier for investors to process information in the whitepapers if the whitepapers are easy to understand. Therefore, among the whitepapers with high readability, we should expect weaker results on long-horizon performances.

We measure the whitepaper readability using the Fog index. We construct an indicator variable ("Easy") that equals to 1 if the whitepaper has a below-median Fog index and 0 otherwise. We present the results in Table 9. Panel A of Table 9 presents results based on the rate of returns. Consistent with the baseline results, we find that the Tech Index positively and significantly predicts the long-horizon returns. The cross-terms between the Tech Index and the indicator variable ("Easy") are negative and significant at the long-horizons, suggesting that the long-horizon return predictability of the technology indexes concentrate among the cryptocurrencies with low readability. For example, the coefficient estimate on the Tech Index at the 300-day horizon is 0.913, while the cross term between the Tech Index and the indicator variable at the same horizon is -.698. This means the long-horizon return predictive power of the Tech Index mainly concentrate on the cryptocurrencies with low readability.

Panel B of Table 9 shows results based on the Bitcoin-adjusted rate of returns. Similar to the

results in Panel A, we find that the cross terms between the technology indexes and the indicator variables are negative and significant at the long-horizons across all the specifications. Overall, we confirm the implication of the investor delayed reaction mechanism: the long-horizon return predictability results are weaker for cryptocurrencies with high readability.

5.2 Return Reversal

In the previous section, we find that the Tech Index positively and significantly predicts cumulative ICO returns over the long horizons. We argue that the findings are consistent with investors' delayed reaction to the technical aspects of the cryptocurrencies. An alternative interpretation of the findings is that investors may overreact to technological aspect of the cryptocurrencies, leading to results of ICO return predictability. Barberis et al. (1998) theoretically demonstrate that investor overreaction to fundamentals can lead to overvaluation of asset values. Pastor and Veronesi (2003) show that investor learning about uncertain fundamentals can lead to a bubble-like phenomenon. A common implication of the models based on investor overreaction or learning of asset fundamentals is that the asset values would eventually reverse to the fundamental values. Technology is one important aspect of fundamental of cryptocurrencies. Therefore, we would expect return reversal if investors over-react to technological fundamentals of cryptocurrencies.

To test this alternative mechanism, in this section, we test whether there is a long-horizon return reversal phenomenon for cryptos with high technology indexes. To detect any return reversal effect, we use the technology indexes to predict crypto returns from 180 days onward. The results are documented in Table 10. Panel A and Panel B of the table document results for the rate of returns and Bitcoin-adjusted rate of returns, respectively. Overall, we do not find evidence of more subsequent return reversal for cryptos with a higher Tech Index.

5.3 ICO First-Day Price

Extensive research has shown that there is a substantial amount of first-day performance in initial public offerings in the equity market.[7] Recently, Benedetti and Kostovetsky (2018) document a similar first-day price reaction in the initial coin offering market. In this section, we study whether the technology indexes help predict not only the long-horizon phenomenon but also the first-day price reaction of ICOs.

Our measure of first-day price reaction is defined as the natural logarithm of the ratio between the first opening price and the ICO offer price. By definition, the sample only includes coins with trading records. Table 11 reports the results for ICO first-day price. Quarterly, categorical and geographic fixed effects are included in the specifications when indicated. We find that the coefficient estimates of the technology index is positive and significant at the 1 percent level. In other words, the technology indexes positively and significantly predict the first-day price reaction. The coefficient estimate is 0.368. The economic magnitude of the coefficient estimate is large. It indicates a 36.8 percent increase in first day price for one standard deviation increase in the Tech Index. The coefficient estimate remains stable in the multivariate specification with controls and fixed effects.

Overall, the Tech Index strongly and positively predict both short-horizon and long-horizon ICO performances. These two sets of results suggest that, although coin market investors take the technical aspects of coins into consideration, they fail to incorporate the information fully.

6 Additional Results and Robustness

In this section, we conduct several robustness tests. While the word embedding method is well-accepted, it is useful to construct the Tech Index with alternative machine learning methods. We use another unsupervised machine learning method, Latent Dirichlet Allocation (LDA), and a supervised machine learning method to construct the Tech Index. Also, we compare our Tech

[7] For a survey paper, see Beatty and Ritter (1986).

Index with other measures. Finally, we show a few additional tests with alternative measures of some key variables.

6.1 Tech Index from Other Machine Learning Methods

LDA is a popular method in the finance and economic literature. It has been used to analyze the structure of economic news (Bybee et al., 2020) and to detect latent topics among employee reviews (Sheng, 2019). The basic idea is that each document can be represented as a probability distribution over various topics, where each topic is a probability distribution over the vocabulary of a corpus. Similar to other textual analysis methods, LDA methods involve a step to remove stop words and then represent the text as data. In the Internet Appendix, we introduce the LDA model and describe the preprocessing procedures and the choice of topics in more detail. We find that 20 topics is optimal based on the selection process shown in .

To understand the LDA output with 20 topics, we interpret these topics by looking at top words associated with each topic. This is a common approach adopted in most finance and economics literature (e.g., Hansen et al., 2018; Sheng, 2019). Table OA.4 displays the top 15 most relevant terms of each LDA topic. We assign a label to each topic based on these key terms. Similar to word embedding, we also use hierarchical agglomerative clustering and multidimensional scaling (MDS) to understand the correlation between LDA topics. Panel A of Figure OA.1 shows the tree structure of LDA topics and Panel B shows the MDS results. These results suggest that we should group "information", "blockchain" and "system" together and define the normalized proportional attention allocated to the three topics as our LDA-based tech index.

We also use supervised machine learning methods to construct a technology index. Supervised machine learning methods learn from a training set in which both the input and the output are known. To construct the training sample, we read through 200 randomly selected whitepapers and give a score from 1 to 4 based on their technical sophistication. The process closely imitates the way investors evaluate the whitepapers. All the whitepapers emphasize on using blockchain and related technologies. Thus, these projects either employ more advanced blockchain technology

or apply existing blockchain technology to different areas. The readers assign a high score (e.g., 3 or 4) to a whitepaper when they think the ICO project involves more advanced and convincing technology. For example, Filecoin uses a novel class of Proof-of-Storage schemes called Proof-of-Replication, and receives an average score of 4. Then, we conduct preprocessing to the training set. We form all two-word phrases in the corpus, remove unigrams and bigrams that appear in less than ten documents, and convert the corpus to a document-term matrix. The final training set consists of 200 documents and 20,586 unique terms.

We consider the following supervised machine learning approaches as potential candidates: panelized linear methods (LASSO, ridge, and elastic net), dimension reduction methods (PCR and PLS), decision tree boosting methods (random forest, gradient boosting), and neural networks. In the Online Appendix, we provide a brief introduction of each supervised method. In order to tune the hyperparameters of each model and find the best model for constructing our supervised technology index, we need to quantify the performance of the model. We evaluate the model performance based on out-of-sample R^2 and we use 5-fold cross-validation to tune the parameters.[8] Table OA.5 Panel A shows the best out-of-sample R-square (R^2_{OOS}) for each supervised method and their corresponding hyperparameters. For our sample, partial least square (PLS) performs the best and has a R^2_{OOS} of 45.88%. Hence, we use the predicted technology score from PLS as our supervised technology index.

In addition, we create a composite index to aggregate the information from all tech indexes. This is done by taking the simple average of the supervised, embedding-based and LDA-based technology index. Both supervised and unsupervised machine learning methods have pros and cons. It is possible that each method capture only some aspect of the true technological components of cryptocurrencies. Then, the composite index would provide a better proxy because it aggregates the information from the three individual indexes. The composite index can potentially reduce the noise of each index, resulting in a useful proxy.

[8] Specifically, we divide the training set into five subsets, each of which contains 40 observations. Following that, each subset will be used as the validation set to evaluate the model based on R^2, while the remaining four subsets are used as the training set. The average out-of-sample R^2 is the simple average of R^2 on the five subsamples

Table OA.5 Panel B reports the correlation table of all technology indexes. The results show that the technology indexes are correlated with each other, suggesting that they all capture some technology aspects of cryptocurrencies. However, they are not perfectly correlated, indicating that each of them provides some different information. Further, Table OA.6 validates the alternative technology indexes using other machine learning methods with GitHub information. All technology indexes are positively correlated with GitHub indicators, confirming that they captures the technology aspect of cryptocurrencies.

We use these three alternative tech indexes in our tests and the results are robust. The results are shown in Table 12. Consistent with our embedding-based Tech Index, the technology index constructed from other machine learning methods positively predicts a coin's listing probability and fund-raising ability, suggesting that technology is an important determinant of cryptocurrency valuations.

6.2 Comparison of Tech Index and Other Measures

In this subsection, we compare our tech indexes with other measures that may contain information on the technological sophistication of cryptocurrencies, including a GitHub measure and a simple word count measure.

One candidate that potentially captures some information of the technological sophistication of cryptocurrencies is the GitHub measures. However, the GitHub measures are ex post measures that capture information about the successes of the ICOs. Moreover, these measures may contain information such as the hype around cryptocurrencies.

In addition, there are multiple methods to conduct textual analysis. For example, the word-count method where we can just count the number of words that belong to a dictionary is well-accepted in the finance and economic literature (e.g., Manela and Moreira, 2017; Liu and Matthies, 2018; Fisher et al., 2020). The word-count method is particularly useful when researchers have good prior knowledge about what they are looking for and the list of words is straightforward. However, cryptocurrency and blockchain are new phenomena and researchers have limited knowl-

edge about what should be a good list of words to describe the technology involved. In this case, unsupervised machine learning methods, such as LDA, are more proper and can overcome this issue. One important advantage of machine learning methods, especially the unsupervised machine learning methods such as word embedding and LDA, is that they do not require researchers to have good prior knowledge about what type of words they are looking for in the texts.

With that being said, we construct technology measures from GitHub and from a simple word count method to compare with our tech indexes. The Github measure we use is *commits*, the number of code revisions of a project on GitHub. The simple word count measure captures the percentage of technology words in a whitepaper, where the technology words are defined by a blockchain dictionary.[9] The complete word list can be found in Table OA.7. In Table 13, we present results using the tech indexes to predict CMC trading, controlling these two types of measures. Columns (1)–(3) report results using the composite tech index, the GitHub *commits* measure, and the simple word count measure, and the point estimates on the variables are all positive and significant at the 5 or 1 percent level. Columns (4) and (5) report results using the composite tech index controlling for the GitHub *commits* measure and the simple word count measure, respectively. When both the composite tech index and the GitHub measure are included, the coefficient estimates on both measures remain positive and highly statistically significant. However, the Tech Index completely subsumes the explanatory power of the simple word count measure. When all three variables are included, the coefficient estimate on the composite tech index remains positive and statistically significant at the 5 percent level, and the point estimate on the simple word count measure is insignificant.

6.3 Additional Tests

We also use an alternative measure of success, Trading, which indicates whether the token is traded on a cryptocurrency exchange. We examine whether the technology indexes predict ICO

[9] See https://consensys.net/knowledge-base/a-blockchain-glossary-for-beginners/; https://blockgeeks.com/guides/blockchain-glossary-from-a-z/; https://www.blockchaintechnologies.com/glossary/.

success under this measure and run a similar regression as in Table 4. Table 14 Panel A column (1) - (2) report the result. The coefficients on the technology indexes are positive and significant and support the same conclusion as in Table 4.

Third, we use linear regression in Table 4 where the dependent variable is a binary variable. Alternatively, we can use a Logit or Probit model. Table 14 Panel A column (3) - (4) report the results from a Logit regression and finds similar results as in Table 4. In the untabulated results, we show that the results under the Probit model are qualitatively similar.

Finally, it is well-documented that we have to impute delisted returns for equity to avoid delisting bias in the data (Shumway, 1997). The equity return data from CRSP automatically contain imputed returns for delisted stocks. For the same reason, we may need to impute returns for delisted ICOs. We set a large negative value -99% as their returns after listed for all delisted ICOs. We then redo the tests on whether the technology indexes affect short-run and long-run returns with and without adjusting Bitcoin returns as in Table 6. Table 14 Panels B and C report the results. Similar to the results in Tables 6, ICOs with higher technology indexes tend to outperform in the long-run. The economic magnitudes are also close.

7 Conclusion

There are two views about cryptocurrency and blockchain technology. The first view is that the cryptocurrency market represents bubbles and fraud. The second one believes that the value of the cryptocurrency market comes from the innovative technologies and that a stake in cryptocurrencies is an investment in the future of the technology. This study contributes to this debate by providing novel measures of technological sophistication of cryptocurrencies via textual analysis of ICO whitepapers. We construct a text-based technology index from a comprehensive sample of ICOs' whitepapers. We find that the ICOs with higher Tech-Index are more likely to succeed and less likely to be delisted subsequently. Although the Tech-index does not statistically significantly affect the short-run returns of ICOs, it has a positive impact on their long-run performance. In

short, our findings suggest that technological sophistication is an important driving force for the performances and valuations of ICOs.

Our findings have important policy implications. Although SEC has launched several initiatives on regulating ICOs, there are no clear disclosure requirements. Our results show that disclosures such as whitepapers are potentially important for the long-term development of the cryptocurrency market. Thus, it might be useful to set up a requirement or guideline for formats and necessary components in the whitepaper, which is a natural analogy for disclosure requirements for public firms (e.g., 10K) and financial firms (e.g., 497K for mutual funds).

References

Abadi J, Brunnermeier M. 2018. Blockchain economics. *Working Paper,National Bureau of Economic Research* .

Barberis N, Shleifer A, Vishny R. 1998. A model of investor sentiment. *Journal of Financial Economics* **49**: 307–343.

Beatty RP, Ritter JR. 1986. Investment banking, reputation, and the underpricing of initial public offerings. *Journal of Financial Economics* **15**: 213–232.

Benedetti H, Kostovetsky L. 2018. Digital tulips? returns to investors in initial coin offerings. *Working Paper, Boston College* .

Blei DM, Ng AY, Jordan MI. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* **3**: 993–1022.

Budish E. 2018. The economic limits of bitcoin and the blockchain. *Working paper, University of Chicago and NBER* .

Buehlmaier MM, Whited TM. 2018. Are financial constraints priced? evidence from textual analysis. *The Review of Financial Studies* **31**: 2693–2728.

Bybee L, Kelly BT, Manela A, Xiu D. 2020. The structure of economic news. *Working paper, Yale University* .

Catalini C, Gans JS. 2018. Initial coin offerings and the value of crypto tokens. *Working paper, University of Toronto and NBER* .

Cong LW, He Z. 2019. Blockchain disruption and smart contracts. *Review of Financial Studies* **32**: 1754–1797.

Cong LW, Li X, Tang K, Yang Y. 2020. Crypto wash trading. *Available at SSRN 3530220* .

Cong LW, Li Y, Wang N. 2019. Tokenomics: Dynamic adoption and valuation. *Working paper, University of Chicago* .

Davydiuk T, Gupta D, Rosen S. 2022. De-crypto-ing signals in initial coin offerings: Evidence of rational token retention. *Management Science, Forthcoming* .

Deng X, Lee YT, Zhong Z. 2018. Decrypting coin winners: Disclosure quality, governance mechanism and team networks. *Working paper, Shanghai University of Finance and Economics* .

Dittmar RF, Wu DA. 2019. Initial coin offerings hyped and dehyped: An empirical examination. *Working paper, University of Michigan* .

Fanti G, Kogan L, Viswanath P. 2019. Economics of proof-of-stake payment systems. *Working paper, MIT* .

Firth JR. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis* .

Fisher AJ, Martineau C, Sheng J. 2020. Macroeconomic attention and announcement risk premia. *Working paper, University of British Columbia* .

Florysiak D, Schandlbauer A. 2019. The information content of ico white papers. *Working paper, Available at SSRN 3265007* .

Gentzkow M, Kelly B, Taddy M. 2019. Text as data. *Journal of Economic Literature* **57**: 535–574.

Griffin JM, Shams A. 2020. Is bitcoin really un-tethered? *Journal of Finance, forthcoming* .

Griffiths TL, Steyvers M. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences* **101**: 5228–5235.

Gu S, Kelly B, Xiu D. 2020. Empirical asset pricing via machine learning. *Review of Financial Studies, forthcoming* .

Hansen S, McMahon M, Prat A. 2018. Transparency and deliberation within the fomc: a computational linguistics approach. *Quarterly Journal of Economics* **133**: 801–870.

Hinzen FJ, John K, Saleh F. 2019. Proof-of-work's limited adoption problem. *Working Paper, New York University* .

Howell ST, Niessner M, Yermack D. 2020. Initial coin offerings: Financing growth with cryptocurrency token sales. *Review of Financial Studies, forthcoming* .

Irresberger F, John K, Saleh F. 2020. The public blockchain ecosystem: An empirical analysis. *NYU Stern School of Business* .

Iyengar G, Saleh F, Sethuraman J, Wang W. 2020. Economics of permissioned blockchain adoption. *Available at SSRN* .

Kelly B, Papanikolaou D, Seru A, Taddy M. 2018. Measuring technological innovation over the long run. Technical report, National Bureau of Economic Research.

Lee J, Li T, Shin D. 2019. The wisdom of crowds in fintech: Evidence from initial coin offerings. *Working paper, University of Florida* .

Li K, Mai F, Shen R, Yan X. 2019. Measuring corporate culture using machine learning. *Available at SSRN 3256608* .

Liu Y, Matthies B. 2018. Long run risk: Is it there? *Working paper, Yale University* .

Liu Y, Tsyvinski A. 2018. Risks and returns of cryptocurrency. *Working paper, Yale University and NBER* .

Liu Y, Tsyvinski A, Wu X. 2019. Common risk factors in cryptocurrency. *Working paper, Yale University and NBER* .

Loughran T, McDonald B. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance* **66**: 35–65.

Loughran T, Ritter JR. 1995. The new issues puzzle. *Journal of Finance* **50**: 23–51.

Lyandres E, Palazzo B, Rabetti D. 2020. Ico success and post-ico performance. *Working Paper* .

Manela A, Moreira A. 2017. News implied volatility and disaster concerns. *Journal of Financial Economics.* **123**: 137–162.

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

Murtagh F, Legendre P. 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of Classification* **31**: 274–295.

Pastor L, Veronesi P. 2003. Stock valuation and learning about profitability. *Journal of Finance* **58**: 1749–1789.

Ritter JR. 1991. The long-run performance of initial public offerings. *Journal of Finance* **46**: 3–27.

Röder M, Both A, Hinneburg A. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*. 399–408.

Russell SJ, Norvig P. 2010. *Artificial Intelligence-A Modern Approach (3rd internat. edn.)*. Pearson Education.

Satopaa V, Albrecht J, Irwin D, Raghavan B. 2011. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*. IEEE, 166–171.

Shams A. 2019. What drives the covariation of cryptocurrency returns? *Working paper, Ohio State University.* .

Sheng J. 2019. Asset pricing in the information age: Employee expectations and stock returns. *Working paper, University of California Irvine* .

Shumway T. 1997. The delisting bias in crsp data. *Journal of Finance* **52**: 327–340.

Sievert C, Shirley K. 2014. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*. 63–70.

Sockin M, Xiong W. 2018. A model of cryptocurrencies. *Working paper, Princeton University* .

Taddy M. 2012. On estimation and selection for topic models. In *Artificial Intelligence and Statistics*. 1184–1193.

Tetlock PC. 2014. Information transmission in finance. *Annual Review Financial Economics* **6**: 365–384.

Torgerson WS. 1958. *Theory and Methods of Scaling.* Wiley.

Yermack D. 2017. Corporate governance and blockchains. *Review of Finance* **21**: 7–31.

Appendix: Variable Definition

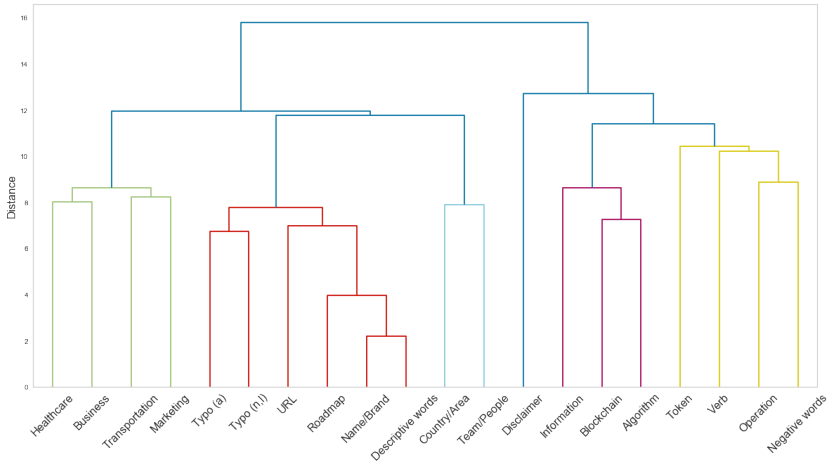
Variable	Definition
ICO Success Measures:	
CMC Trading	A dummy variable that equals to one if a cryptocurrency is shown as listed on coinmarketcap.com (CMC).
Trading	A self-reported dummy by ICO fundraisers about whether the cryptocurrency is traded on an exchange.
Success	A dummy variable indicating whether the ICO raises any capital.
Trading Variables:	
First Open/ICO Price	The ratio between the first day's opening price and the ICO price.
Delist	An indicator about whether a token is delisted from CMC.
Rate of Return	The rate of return that investors earn if they buy cryptoccurrency at the opening price on the first listing day and sell them after a certain holding period.
Trading Volume	The 24-hour trading volume in millions of USD after they have been listed on CMC for a certain period of time.
Whitepaper Measures:	
Tech_sup	The normalized predicted technology score from partial least squares (PLS), a supervised machine learning approach.
Tech_embed	The normalized percentage of words in the "blockchain", "information" or "algorithm" topics of the word embedding and clustering approach.
Tech_lda	The normalized proportional attention allocated to the "information", "blockchain" and "system" topics of the LDA topic modelling approach.
Tech_comp	The simple average of the Tech_sup, Tech_embed and Tech_lda.
Fog Index	A readability measure defined as $0.4[(words/sentences)+100((complexwords)/words)]$, where "complex words" are words with three or more syllables.
Tone	The difference between number of positive and negative words defined in Loughran and McDonald (2011) divided by the total number of words in a whitepaper.
Uncertainty	The number of uncertainty words defined in Loughran and McDonald (2011) divided by the total number of words in a whitepaper.
ICO characteristics:	
Has GitHub	A dummy variable that equals to one if the ICO project has a GitHub homepage.
Has Twitter	A dummy variable that equals to one if the ICO project has a Twitter account.
ICO Length	The number of days from the start to the end of an ICO campaign.
Team Size	The number of ICO team members.
Pre ICO	A dummy variable indicating whether if a pre-ICO exists.
Bonus	A dummy variable indicating whether the fundraiser offers bonus to investors.
Ethereum Based	A dummy variable indicating whether the ICO project is built on Ethereum.
Accept BTC	A dummy variable indicating whether the ICO accepts Bitcoin as a currency of payment.
BTC Price (ICO)	The price of Bitcoin in thousands of US dollars on the day an ICO initiates.
BTC Price (List)	The price of Bitcoin in thousands of US dollars on the day an ICO is shown as listed on CMC.

Figures & Tables

Figure 1: **Embedding Visualization**

This figure plots the relationship between embedding-based topics. Panel (a) displays the taxonomy generated by hierarchical agglomerative clustering. Panel (b) shows the similarity between topics in a two-dimensional space. The size of the circle represents the percentage of terms belonging to the topic. "Information", "blockchain" and "algorithm" are used to construct the embedding-based tech index.

(a) Taxonomy



(b) Multidimensional scaling (MDS)

Table 1: **Summary Statistics**

This table presents summary statistics on variables related to ICO characteristics, outcomes and whitepaper measures. Panel A shows descriptive statistics for 2,916 ICOs completed before December 31st, 2018. Panel B summarizes a subsample of 765 ICOs listed on coinmarketcap.com. For each variable, we show the number of non-missing observations, the mean, the standard deviation and the 10th, 50th and 90th percentile values. Please refer to the "variable definition" in the Appendix for the definition of each variable.

Panel A: Full Sample

	Obs.	Mean	SD	p10	p50	p90
ICO Success Measures						
CMC Trading	2916	0.26	0.44	0	0	1
Trading	2916	0.18	0.39	0	0	1
Success	2916	0.38	0.49	0	0	1
Whitepaper Measures						
Tech Index	1629	0	1.00	-1.02	-0.23	1.35
Fog Index	1629	16.7	12.6	13.2	15.7	18.5
Tone	1629	0.28	0.73	-0.58	0.29	1.10
Uncertainty	1629	0.75	0.39	0.35	0.67	1.25
ICO Characteristics						
Has GitHub	2916	0.60	0.49	0	1	1
Has Twitter	2916	0.91	0.29	1	1	1
ICO Length	2683	50.7	45.8	14	32	100
Team Size	2916	11.0	7.05	3	10	20
Pre ICO	2916	0.51	0.50	0	1	1
Bonus	2916	0.20	0.40	0	0	1
Ethereum Based	2916	0.83	0.37	0	1	1
Accept BTC	2916	0.40	0.49	0	0	1
BTC Price (ICO)	2669	7.80	3.07	4.23	7.28	11.3
BTC Price (List)	710	7.59	3.70	2.73	7.03	13.5
ICO Price	1684	1.57	17.8	0.01	0.10	1

Panel B: Listed Sample

	Obs.	Mean	SD	p10	p50	p90
Trading Variables						
First Open/ICO Price	413	2.20	4.66	0.16	0.97	3.79
Delist	765	0.10	0.30	0	0	1
Rate of Return						
7 Days	741	0.19	0.83	-0.45	-0.04	1.03
30 Days	730	0.30	1.84	-0.71	-0.28	1.60
90 Days	686	0.65	3.11	-0.87	-0.43	3.21
180 Days	566	1.00	5.14	-0.95	-0.64	4.00
210 Days	530	0.84	4.27	-0.96	-0.68	3.57
240 Days	486	0.69	3.89	-0.96	-0.70	3.20
270 Days	438	1.46	8.40	-0.96	-0.72	3.69
300 Days	397	1.51	8.91	-0.97	-0.74	3.85
330 Days	356	1.34	8.21	-0.98	-0.72	3.31
360 Days	289	1.60	7.74	-0.96	-0.67	4.20
Trading Volume ($ MIL)						
Listing Days	751	2.40	11.9	0.0023	0.12	3.90
7 Days	739	1.63	5.58	0.0015	0.083	3.60
30 Days	725	1.50	5.62	0.0011	0.066	2.53
90 Days	680	1.60	5.58	0.00045	0.11	3.22
180 Days	564	2.78	13.5	0.00039	0.069	3.99
210 Days	526	2.24	8.30	0.00020	0.065	3.90
240 Days	482	2.60	12.4	0.00023	0.048	3.31
270 Days	436	1.73	5.91	0.00016	0.061	3.09
300 Days	393	2.60	11.5	0.00025	0.058	3.50
330 Days	352	2.22	8.48	0.00021	0.067	3.19
360 Days	285	2.45	9.65	0.000048	0.084	3.39
Whitepaper Measures						
Tech Index	422	0.41	1.17	-0.79	0.14	2.28
Fog Index	422	17.2	18.9	13.3	15.5	18.3
Tone	422	0.20	0.72	-0.70	0.23	1.03
Uncertainty	422	0.79	0.40	0.35	0.71	1.30
ICO Characteristics						
Has GitHub	765	0.70	0.46	0	1	1
Has Twitter	765	0.96	0.19	1	1	1
ICO Length	656	34.9	42.0	2	30	63
Team Size	765	12.1	8.02	3	11	22
Pre ICO	765	0.26	0.44	0	0	1
Bonus	765	0.075	0.26	0	0	0
Ethereum Based	765	0.80	0.40	0	1	1
Accept BTC	765	0.30	0.46	0	0	1
BTC Price (ICO)	642	7.45	3.94	2.54	7.10	13.8
BTC Price (List)	710	7.59	3.70	2.73	7.03	13.5
ICO Price	420	2.37	19.9	0.01	0.12	1.22

Table 2: **Technology Index Validation**

This table validates the tech index with measures from GitHub. In each column, *watch* measures the number of users subscribing repository updates; *star* indicates the number of "likes" received by the repository; *fork* proxies for the copies made by other developers; *commit* represents the number of times the code has been revised; *branch* is the amount of pointers to specific versions of the repository; and *contributor* reflects how many developers have contributed to the source code. All GitHub measures are in logarithmic forms. The reported t-statistics are based on robust standard errors. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
	ln(watch)	ln(star)	ln(fork)	ln(commits)	ln(branch)	ln(contributor)
Tech Index	0.651***	0.792***	0.665***	0.995***	0.563***	0.580***
	(0.061)	(0.073)	(0.068)	(0.078)	(0.050)	(0.051)
Constant	1.950***	1.795***	1.389***	4.018***	1.852***	1.908***
	(0.054)	(0.060)	(0.051)	(0.080)	(0.044)	(0.046)
Observations	861	861	861	861	861	861
R^2	0.148	0.170	0.158	0.160	0.161	0.158

Table 3: **Technology Index Determinant**

This table presents the determinants of our tech index. The dependent variable is the Tech Index. Column (1) links the tech index to whether an ICO uses Ethereum blockchain; column (2) presents the relation between thetech index and GitHub commits (the number of code revisions); column (3) considers other text-based measures of ICO whitepapers; column (4) presents estimates with ICO characteristics; column (5) includes all variables. The reported t-statistics are based on robust standard errors. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels respectively.

	(1)	(2)	(3)	(4)	(5)
Ethereum Based	-0.294***				-0.172**
	(0.077)				(0.073)
ln_commits		0.122***			0.083***
		(0.013)			(0.013)
Has GitHub		-0.119**			-0.014
		(0.060)			(0.061)
Fog Index			-0.004**		-0.004
			(0.002)		(0.003)
Tone			-0.282***		-0.227***
			(0.040)		(0.039)
Uncertainty			-0.364***		-0.383***
			(0.065)		(0.065)
ICO Length				-0.002***	-0.002***
				(0.001)	(0.001)
Team Size				0.008**	0.004
				(0.004)	(0.004)
Has Twitter				0.150	0.075
				(0.124)	(0.126)
BTC Price (ICO)				-0.027***	-0.016*
				(0.010)	(0.009)
Pre ICO				-0.115**	-0.071
				(0.053)	(0.051)
Bonus				-0.103*	-0.094*
				(0.055)	(0.053)
Accept BTC				-0.134***	-0.095*
				(0.051)	(0.048)
Constant	0.247***	-0.193***	0.419***	0.202	0.569***
	(0.072)	(0.037)	(0.072)	(0.154)	(0.175)
R^2	0.012	0.097	0.042	0.040	0.131
Observations	1629	1629	1629	1483	1483

Table 4: **ICO Success**

This table examines the relationship between the Tech Index and ICO success. The dependent variable is *CMC Trading* in column (1) - (2) and *Success* in column (3) - (4). *CMC Trading* is a dummy that equals one if a cryptocurrency is shown as listed on CoinMarketCap. *Success* indicates whether the ICO raises any capital. For each dependent variable, the first column presents the univariate result, and the second column displays estimates with control variables and fixed effects. The reported t-statistics are based on robust standard errors. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels respectively.

	(1)	(2)	(3)	(4)
	CMC Trading		Capital Raised > 0	
Tech Index	0.107***	0.048***	0.077***	0.043***
	(0.011)	(0.012)	(0.012)	(0.014)
ICO Length		-0.001***		-0.001***
		(0.000)		(0.000)
Team Size		0.009***		0.009***
		(0.001)		(0.002)
Has GitHub		0.044**		0.082***
		(0.021)		(0.025)
Has Twitter		0.172***		0.092*
		(0.036)		(0.053)
BTC Price (ICO)		0.010		-0.005
		(0.006)		(0.007)
Pre ICO		-0.031		-0.008
		(0.023)		(0.028)
Bonus		0.010		0.104***
		(0.022)		(0.031)
Accept BTC		-0.011		0.070***
		(0.021)		(0.024)
Ethereum Based		-0.009		-0.007
		(0.030)		(0.034)
Fog Index		0.000		-0.000
		(0.001)		(0.001)
Tone		0.003		0.013
		(0.014)		(0.018)
Uncertainty		0.031		0.081**
		(0.028)		(0.033)
Constant	0.259***	0.521***	0.377***	0.557***
	(0.011)	(0.099)	(0.012)	(0.124)
Fixed Effects	No	Yes	No	Yes
R^2	0.060	0.324	0.025	0.256
Observations	1629	1382	1629	1382

Table 5: **ICO Success—Industry Subsample**

This table examines the relationship between the Tech Index and ICO success for technology-related industries. The dependent variable is *CMC Trading. Industry* is a dummy that equals to 1 if the ICO belongs to "platform", "cryptocurrency", and "trading" industries. For each tech index, the first column presents the univariate result, and the second column displays estimates with control variables and fixed effects. The reported t-statistics are based on robust standard errors. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels respectively.

	(1)	(2)
Tech Index	0.085***	0.028*
	(0.017)	(0.016)
Tech_embed*Industry	0.040*	0.035*
	(0.022)	(0.021)
Industry	-0.009	0.011
	(0.021)	(0.020)
ICO Length		-0.001***
		(0.000)
Team Size		0.009***
		(0.001)
Has GitHub		0.044**
		(0.021)
Has Twitter		0.169***
		(0.036)
BTC Price (ICO)		0.010
		(0.006)
Pre ICO		-0.030
		(0.023)
Bonus		0.009
		(0.022)
Accept BTC		-0.013
		(0.020)
Ethereum Based		-0.002
		(0.030)
Fog Index		-0.000
		(0.002)
Tone		0.004
		(0.014)
Uncertainty		0.043
		(0.028)
Constant	0.264***	0.468***
	(0.016)	(0.097)
Fixed effects	No	Yes
R2	0.062	0.313
Observations	1629	1382

Table 6: **Rate of Return**

This table presents the effect of the Tech Index on cryptocurrency return. In Panel A, the dependent variable is rate of return (the log transformation of gross return over a given period). In Panel B, the dependent variable is Bitcoin-adjusted return. It is calculated as the log transformation of gross return, log(1+ROR), minus the log transformation of Bitcoin gross return over the same period. Column (1)-(6) display results for six horizons: 7 days, 30 days, 90 days, 180 days, 240 days and 300 days. We include control variables related to ICO characteristics and whitepapers in all columns. Quarterly, categorical and geographical fixed effects are considered under all circumstances. The reported t-statistics are based on robust standard errors. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels respectively.

Panel A: Rate of Return

	(1)	(2)	(3)	(4)	(5)	(6)
	7 Days	30 Days	90 Days	180 Days	240 Days	300 Days
Tech Index	0.032	0.080	0.123	0.195*	0.331***	0.377**
	(0.034)	(0.059)	(0.083)	(0.105)	(0.117)	(0.150)
ICO Length	0.000	0.000	-0.001	-0.003	-0.003	0.004
	(0.001)	(0.001)	(0.001)	(0.004)	(0.004)	(0.007)
Team Size	-0.001	0.003	0.004	-0.003	-0.001	-0.003
	(0.004)	(0.007)	(0.010)	(0.013)	(0.015)	(0.018)
Has GitHub	0.026	0.000	0.136	0.201	0.189	0.332
	(0.075)	(0.134)	(0.172)	(0.235)	(0.249)	(0.374)
Has Twitter	0.063	0.019	0.031	-0.020	0.066	0.175
	(0.148)	(0.569)	(0.716)	(0.847)	(0.784)	(0.812)
BTC Price (ICO)	0.001	-0.027	-0.056**	-0.089***	-0.067**	-0.074
	(0.013)	(0.019)	(0.026)	(0.030)	(0.034)	(0.047)
Pre ICO	0.004	-0.127	0.038	0.248	0.080	0.776
	(0.079)	(0.135)	(0.178)	(0.292)	(0.366)	(0.591)
Constant	0.059	-0.492	-1.530*	-0.701	-0.722	-1.305
	(0.658)	(1.292)	(0.896)	(1.215)	(1.164)	(1.319)
Other Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.078	0.167	0.246	0.364	0.430	0.403
Observations	316	310	286	218	184	140

Panel B: Bitcoin-adjusted Rate of Return

	(1)	(2)	(3)	(4)	(5)	(6)
	7 Days	30 Days	90 Days	180 Days	240 Days	300 Days
Tech Index	0.031	0.097*	0.122	0.216**	0.319***	0.407***
	(0.033)	(0.056)	(0.080)	(0.094)	(0.109)	(0.144)
ICO Length	0.000	0.000	-0.000	-0.002	-0.002	0.005
	(0.001)	(0.001)	(0.001)	(0.003)	(0.004)	(0.005)
Team Size	-0.001	0.002	0.007	0.003	0.006	0.001
	(0.004)	(0.006)	(0.009)	(0.011)	(0.014)	(0.016)
Has GitHub	0.046	-0.057	0.072	0.143	0.098	0.138
	(0.067)	(0.125)	(0.160)	(0.196)	(0.234)	(0.340)
Has Twitter	0.193	-0.001	0.168	0.548	0.400	0.483
	(0.178)	(0.671)	(0.928)	(0.912)	(0.844)	(0.839)
BTC Price (ICO)	0.010	-0.005	-0.030	-0.069**	-0.045	-0.037
	(0.012)	(0.018)	(0.024)	(0.027)	(0.031)	(0.048)
Pre ICO	-0.002	-0.113	0.060	0.194	0.100	0.601
	(0.075)	(0.116)	(0.176)	(0.250)	(0.324)	(0.515)
Constant	-0.114	-0.679	-2.119**	-3.265***	-3.829***	-3.579***
	(0.624)	(1.207)	(1.038)	(1.237)	(1.127)	(1.257)
Other Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.100	0.164	0.179	0.338	0.326	0.324
Observations	311	305	281	213	180	137

Table 7: **Trading Volume**

This table presents the relation between the tech index and cryptocurrency liquidity. The dependent variable is the log transformation of 24-hour trading volume in USD. Column (1) displays results on the listing day. Column (2) to (7) display results for six time points: 7 days, 30 days, 90 days, 180 days, 240 days and 300 days. We include control variables in all columns. Quarterly, categorical and geographical fixed effects are considered under all circumstances. The reported t-statistics are based on robust standard errors. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Listing	7 Days	30 Days	90 Days	180 Days	240 Days	300 Days
Tech Index	0.514***	0.490***	0.489***	0.729***	0.625**	0.798***	0.780**
	(0.156)	(0.153)	(0.170)	(0.172)	(0.262)	(0.261)	(0.342)
ICO Length	-0.004	-0.005	-0.003	-0.005	-0.005	-0.000	-0.002
	(0.004)	(0.006)	(0.005)	(0.007)	(0.009)	(0.008)	(0.016)
Team Size	0.019	0.023	0.033*	0.030	0.085***	0.056**	0.094**
	(0.019)	(0.018)	(0.020)	(0.019)	(0.026)	(0.026)	(0.037)
Has GitHub	0.154	0.417	0.360	0.626	0.736	0.954*	1.068
	(0.359)	(0.382)	(0.459)	(0.504)	(0.543)	(0.563)	(0.750)
Has Twitter	-0.545	-0.112	-0.753	-1.140	0.598	-0.500	-1.203
	(1.251)	(1.207)	(1.783)	(1.799)	(2.187)	(1.615)	(1.473)
BTC Price (ICO)	0.090	0.113*	-0.021	0.034	-0.020	-0.010	0.079
	(0.070)	(0.065)	(0.073)	(0.073)	(0.090)	(0.092)	(0.118)
Pre ICO	-0.164	-0.246	-0.198	-0.652	-0.989	-1.602**	-0.063
	(0.429)	(0.496)	(0.528)	(0.595)	(0.753)	(0.790)	(1.178)
Constant	11.220***	13.273***	10.602***	8.893***	4.161	6.149**	8.349***
	(1.715)	(1.785)	(2.410)	(2.197)	(3.229)	(2.547)	(2.894)
Other Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.213	0.234	0.184	0.234	0.352	0.457	0.456
Observations	323	316	308	283	217	183	139

Table 8: **Delisting Probability**

This table presents OLS estimates of the relationship between the Tech Index and ICO delisting probability. The dependent variable is *Delist*, a dummy variable that equals to 1 if a token was shown as "inactive" on coinmarketcap.com by the end of 2018. the first column presents the univariate result, and the second column displays estimates with control variables and fixed effects. The reported t-statistics are based on robust standard errors. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels respectively.

	(1)	(2)
Tech Index	-0.030***	-0.025*
	(0.010)	(0.014)
ICO Length		-0.000
		(0.000)
Team Size		0.001
		(0.002)
Has GitHub		-0.050
		(0.039)
Has Twitter		-0.187
		(0.203)
BTC Price (ICO)		-0.012**
		(0.005)
Pre ICO		0.008
		(0.045)
Bonus		0.021
		(0.065)
Accept BTC		-0.002
		(0.034)
Ethereum Based		-0.027
		(0.052)
Fog Index		-0.001**
		(0.001)
Tone		-0.018
		(0.021)
Uncertainty		0.006
		(0.051)
Constant	0.083***	0.684**
	(0.015)	(0.332)
Fixed Effects	No	Yes
R^2	0.018	0.152
Observations	422	329

Table 9: **Long Horizon Performance—Subsample on Readability**

This table examines the long horizon performance of cryptocurrencies for different readability subsamples. The dependent variable is rate of return in panel A and Bitcoin-adjusted return in panel B. *Easy* is a dummy that equals to 1 if the whitepaper has a below-median Fog index. We include control variables related to ICO characteristics and whitepapers in all columns. Quarterly, categorical and geographical fixed effects are considered under all circumstances. The reported t-statistics are based on robust standard errors. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels respectively.

Panel A: Rate of Return

	(1)	(2)	(3)	(4)	(5)	(6)
	7 Days	30 Days	90 Days	180 Days	240 Days	300 Days
Tech Index	0.024	0.107	0.174	0.402**	0.673***	0.913***
	(0.052)	(0.092)	(0.131)	(0.186)	(0.223)	(0.202)
Tech Index*Easy	0.008	-0.082	-0.133	-0.300	-0.479**	-0.698***
	(0.061)	(0.105)	(0.155)	(0.206)	(0.231)	(0.221)
Easy	0.029	0.015	0.000	-0.018	-0.132	-0.110
	(0.071)	(0.120)	(0.168)	(0.226)	(0.249)	(0.328)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
FEs	Yes	Yes	Yes	Yes	Yes	Yes
R2	0.078	0.151	0.230	0.371	0.456	0.448
Observations	316	310	286	218	184	140

Panel B: Adjusted Rate of Returns

	(1)	(2)	(3)	(4)	(5)	(6)
	7 Days	30 Days	90 Days	180 Days	240 Days	300 Days
Tech Index	0.027	0.101	0.159	0.342**	0.605***	0.870***
	(0.050)	(0.088)	(0.121)	(0.158)	(0.190)	(0.188)
Tech Index*Easy	0.000	-0.038	-0.101	-0.173	-0.399*	-0.600***
	(0.059)	(0.102)	(0.145)	(0.174)	(0.202)	(0.200)
Easy	0.002	-0.029	-0.006	-0.016	-0.072	-0.035
	(0.068)	(0.113)	(0.154)	(0.200)	(0.234)	(0.304)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
FEs	Yes	Yes	Yes	Yes	Yes	Yes
R2	0.098	0.146	0.164	0.338	0.349	0.366
Observations	311	305	281	213	180	137

Table 10: **Is There Return Reversal?**

This table presents the effects of the Tech Index on long-term cryptocurrency returns. The dependent variable is $log(1+ROR_{180->j})$, the gross return from 180 listing days onward. Panel A displays the result on rate of returns and panel B shows Bitcoin-adjusted rate of returns. Column (1)-(6) display results for six horizons from the listing day: 210 days, 240 days, 270 days, 300 days, 330 days and 360 days. We include control variables related to ICO characteristics and whitepapers in all columns. Quarterly, categorical and geographical fixed effects are considered under all circumstances. The reported t-statistics are based on robust standard errors. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels respectively.

Panel A: Rate of Returns

	(1)	(2)	(3)	(4)	(5)	(6)
	210 Days	240 Days	270 Days	300 Days	330 Days	360 Days
Tech Index	0.048	0.086**	0.080*	0.089*	0.070	0.008
	(0.035)	(0.038)	(0.041)	(0.048)	(0.059)	(0.077)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
FEs	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.296	0.319	0.403	0.563	0.482	0.493
Observations	207	184	156	140	132	103

Panel B: Bitcoin-Adjusted Rate of Returns

	(1)	(2)	(3)	(4)	(5)	(6)
	210 Days	240 Days	270 Days	300 Days	330 Days	360 Days
Tech Index	0.045	0.037	0.076*	0.069	0.042	0.028
	(0.030)	(0.035)	(0.042)	(0.046)	(0.059)	(0.071)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
FEs	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.291	0.346	0.429	0.538	0.449	0.565
Observations	202	180	153	137	129	101

Table 11: **ICO First-Day Price**

This table presents OLS estimates of the relationship between the Tech Index and ICO first-day price. The dependent variable is Ln(First Opening Price/ICO Price), the log transformation of the ratio between the first day's opening price and ICO price. The first column presents univariate result, and the second column displays estimates with control variables and fixed effects. The reported t-statistics are based on robust standard errors. ***, **, and * indicate statistical significance at 1%, 5%, and 10% respectively.

	(1)	(2)
Tech Index	0.414***	0.368***
	(0.076)	(0.098)
ICO Length		-0.001
		(0.002)
Team Size		0.007
		(0.012)
Has GitHub		-0.128
		(0.253)
Has Twitter		-0.162
		(0.392)
BTC Price (ICO)		0.010
		(0.048)
Pre ICO		-0.205
		(0.348)
Bonus		0.128
		(0.407)
Accept BTC		-0.049
		(0.229)
Ethereum Based		-0.206
		(0.367)
Fog Index		-0.005
		(0.007)
Tone		0.105
		(0.147)
Uncertainty		0.121
		(0.328)
Constant	-0.404***	-3.392***
	(0.111)	(0.916)
Fixed Effects	No	Yes
R^2	0.111	0.328
Observations	238	199

Table 12: **Tech Index from Alternative Machine Learning Methods**

This table examines the relationship between ICO success and the supervised, LDA-based and the composite tech index. The dependent variable is *CMC Trading* in Panel A and *Success* in Panel B. *CMC Trading* is a dummy that equals one if a cryptocurrency is shown as listed on CoinMarketCap. *Success* indicates whether the ICO raises any capital. For each tech index, the first column presents the univariate result, and the second column displays estimates with control variables and fixed effects. The reported t-statistics are based on robust standard errors. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels respectively.

Panel A: CMC Trading

	(1)	(2)	(3)	(4)	(7)	(8)
	Supervised		LDA		Composite	
Tech Index	0.070***	0.039***	0.086***	0.047***	0.124***	0.066***
	(0.012)	(0.012)	(0.012)	(0.013)	(0.013)	(0.015)
Constant	0.259***	0.610***	0.259***	0.503***	0.259***	0.511***
	(0.011)	(0.094)	(0.011)	(0.099)	(0.011)	(0.097)
Controls	No	Yes	No	Yes	No	Yes
Fixed Effects	No	Yes	No	Yes	No	Yes
R^2	0.026	0.322	0.038	0.323	0.057	0.327
Observations	1629	1382	1629	1382	1629	1382

Panel B: Capital Raised > 0

	(1)	(2)	(3)	(4)	(7)	(8)
	Supervised		LDA		Composite	
Tech Index	0.061***	0.041***	0.056***	0.037***	0.091***	0.060***
	(0.012)	(0.013)	(0.012)	(0.014)	(0.014)	(0.016)
Constant	0.377***	0.633***	0.377***	0.554***	0.377***	0.545***
	(0.012)	(0.121)	(0.012)	(0.126)	(0.012)	(0.123)
Controls	No	Yes	No	Yes	No	Yes
Fixed Effects	No	Yes	No	Yes	No	Yes
R^2	0.016	0.256	0.013	0.254	0.025	0.258
Observations	1629	1382	1629	1382	1629	1382

Table 13: **Comparison with Other Measures**

This table compares the Tech Index with other alternative technology measures: GitHub commits and the simple word count. *Ln(commits)* is the logarithm of the number of code revisions on GitHub. *Simple_word_count* measures the percentage of words in a whitepaper that belongs to a technology word list. The complete word list can be found in Table OA.7. The dependent variable is *CMC Trading*. We include control variables related to ICO characteristics and whitepapers in all columns. Quarterly, categorical and geographical fixed effects are considered under all circumstances. The reported t-statistics are based on robust standard errors. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Tech Index	0.048***			0.045***	0.052***	0.051**
	(0.012)			(0.015)	(0.014)	(0.020)
Ln(commits)		0.021***		0.016**		0.016**
		(0.006)		(0.006)		(0.006)
Simple_word_count			0.024**		-0.006	-0.009
			(0.011)		(0.013)	(0.019)
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.324	0.365	0.318	0.372	0.325	0.372
Observations	1382	748	1382	748	1382	748

Table 14: **Robustness Tests**

This table displays several robustness tests. Panel A column (1) - (2) rerun the main analysis using Trading as the dependent variable. Panel A column (3) - (4) report the Logit regression version of Table 4. Besides, to mitigate the concern of survivorship bias, we impute -99% to returns of delisted cryptocurrencies and redo Table 6. Results are presented in Panel B and panel C. The reported t-statistics are based on robust standard errors. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels respectively.

Panel A: Alternative dependent variable and logit regression

	Dep. Var = Trading		Logit regression	
	(1)	(2)	(3)	(4)
Tech_embed	0.101***	0.047***	0.527***	0.345***
	(0.010)	(0.010)	(0.055)	(0.085)
Constant	0.167***	0.678***	-1.108***	-6.693***
	(0.009)	(0.068)	(0.059)	(1.010)
Controls	No	Yes	No	Yes
Fixed Effects	No	Yes	No	Yes
R^2	0.074	0.389	0.050	0.325
Observations	1629	1382	1629	1351

Panel B: Rate of return (-99% return for delisted coins)

	(1)	(2)	(3)	(4)	(5)	(6)
	7 Days	30 Days	90 Days	180 Days	240 Days	300 Days
Tech_embed	0.112**	0.156**	0.189**	0.241**	0.379***	0.396***
	(0.052)	(0.070)	(0.088)	(0.101)	(0.113)	(0.141)
Constant	-2.045	-3.035**	-2.658***	-0.747	-0.391	-0.495
	(1.246)	(1.472)	(1.015)	(1.162)	(1.252)	(1.373)
Controls & FEs	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.164	0.215	0.271	0.389	0.473	0.462
Observations	323	319	293	228	198	157

Panel C: Bitcoin-adjusted rate of return (-99% return for delisted coins)

	(1)	(2)	(3)	(4)	(5)	(6)
	7 Days	30 Days	90 Days	180 Days	240 Days	300 Days
Tech_embed	0.052	0.116*	0.137*	0.236**	0.354***	0.428***
	(0.039)	(0.060)	(0.081)	(0.096)	(0.111)	(0.146)
Constant	-0.417	-2.038	-1.991*	-3.252**	-4.174***	-3.415**
	(0.676)	(1.586)	(1.021)	(1.285)	(1.371)	(1.409)
Controls & FEs	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.097	0.166	0.175	0.333	0.341	0.342
Observations	312	308	282	217	186	144

Internet Appendix

1 Supervised Machine Learning Method

1.1 Basics

Supervised learning is "the machine learning task of learning a function that maps an input to an output based on example input-output pairs." (Russell and Norvig, 2010). Mathematically, this can be expressed as estimating a function f(·) given input variables (X) and output variables (Y), such that the mapping function Y=f(X) is satisfied as much as possible. Various machine learning models impose different constraints on the function, resulting in different optimization results.

We use the simplest regression model, ordinary least squares (OLS), as our benchmark. The objective function of OLS is:

$$\min_{\beta} ||y - x\beta||^2$$

OLS works well when there are only a few predictors, but its performance deteriorates significantly as the dimension of predictors increases. Unfortunately, high dimensionality and sparseness are both common features of text data. Hence, we apply more advanced machine learning methods to avoid the "curse of dimensionality".

The first set of methods we use are panelized linear approaches. The idea is to add a penalty term in the objective function to reduce a model's fit on noise and hence enhances prediction accuracy. We consider LASSO, ridge regression and elastic net for this approach. Another common approach to deal with high-dimensional data is dimension reduction. While panelized linear methods select a subset of predictors that have strong predictive power, dimension reduction methods combine predictors into several main components while retaining as much information as possible. We apply principal component regression (PCR) and partial least squares (PLS) in this vein. All the methods above are linear regression models, but we are also interested in using non-linear ap-

proaches to get better prediction accuracy. We consider decision trees (random forest and gradient boosting) and neural network algorithms. Next, I briefly introduce each of these machine learning methods that we consider as candidates to construct our supervised tech index.

1.1.1 LASSO

LASSO (least absolute shrinkage and selection operator) is a common approach employed to deal with high-dimensional sparse data. The objective function for LASSO is:

$$\min_{\beta} \left\{ \frac{1}{N} ||y - X\beta||^2 + \lambda ||\beta|| \right\}.$$

The first term is the same as OLS, while the second term is a penalty on non-zero coefficients, with λ representing the regularization strength. The effect of LASSO is to select only a subset of predictors by pushing other predictor coefficients to 0.

1.1.2 Ridge

Ridge regression (also known as Tikhonov regularization) is another useful method to mitigate the problem of dimensionality by adding a L2-norm regularization term as penalty. The objective function is:

$$\min_{\beta} \left\{ \frac{1}{N} ||y - X\beta||^2 + \lambda ||\beta||^2 \right\}.$$

The difference of Ridge regression from LASSO is that it shrinks the coefficients of unimportant predictors but do not set them to 0. Hence, ridge regression is a regularization approach, but not a variable selection approach.

1.1.3 Elastic Net

Elastic net is a combination of LASSO and ridge regression. It optimizes the following objective function:

$$\min_{\beta} \left\{ \frac{1}{N} ||y - X\beta||^2 + \lambda\alpha||\beta||| + \lambda\left(1-\alpha\right)||\beta||^2 \right\}.$$

α controls the weight between L1 and L2 norm penalty. If $\alpha = 1$, it is the same as LASSO; if $\alpha = 0$, it becomes ridge regression. By averaging between LASSO and ridge regression, elastic net is expected to combine the advantages of both methods.

1.1.4 Principal component regression (PCR)

Principal component regression combines standard linear regression with principal component analysis (PCA). Specifically, PCR regresses the dependent variable (Y) on principal components of independent variables (X), as opposed to regressing Y directly on X in OLS. Since the principal components are extracted based on their ability to explain the variation in X, the forecasting goal (Y) does not come into play until the final regression step.

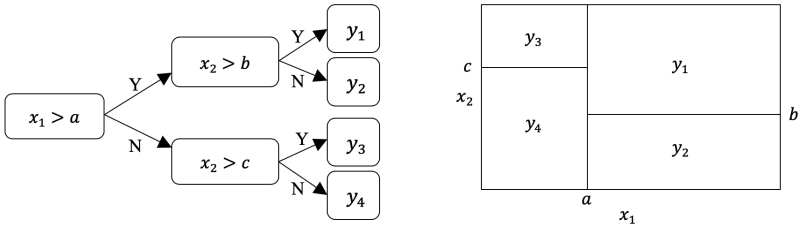
1.1.5 Partial least square (PLS)

Partial least squares (PLS) regression shares some similarities with PCR, but it constructs the principal components of X with the goal to best explain the covariance between X and Y. It first projects both the independent variables (X) and dependent variables (Y) to a new space, in which the projection of the X-space that explains the most variation of the Y-space. It then runs a linear regression model in the new space. PLS is especially helpful when predictors are more than available observations or when predictors are highly collinear.

1.1.6 Random Forest

Random forests come from decision trees. A decision tree is a set of logic conditions on input variables (X) that lead to predictions on the target output (Y). The following figure illustrates a regression tree example. The first condition used to determine y is whether x_1 is greater than a. Conditional on the answer to this question, another logic condition will be raised. This process iterates until the value of y is determined. Different from linear regressions, the regression tree is a

non-linear and non-parametric method. A random forest is an ensemble of multiple decision trees. It outputs the average prediction of each individual tree. Although a single tree may be a weak prediction model, through combination the random forest can have a strong performance.



1.1.7 Gradient Bossting

Gradient boosting is another approach to ensemble regression trees. At each step, a new tree is fitted on the negative gradient of a given loss function. Hence, new trees aim at correcting the error of preceding trees. To avoid overfitting on residuals, following trees will be discounted at each step. This process is repeated until a total number of N trees is reached.

1.1.8 Neural Network (NN)

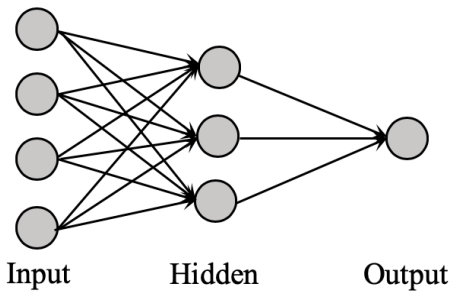
Artificial neural network is a broad set of machine learning algorithms inspired by the biological neural structure of human brains. It is a layer-by-layer structure, where each layer is composed of “neurons”, and the layers are connected by "edges". The following figure shows an example, the feedforward neural network. The input layer is the input variables (X), and the output layer is the outcome (Y). Each node of the hidden layer represents the following operation:

$$H_i = f\left(w_0 + XW\right),$$

where W, the linear weight matrix on the inputs, represents the "edges" connecting the input layer and the hidden layer. There are multiple choices for $f(\cdot)$, one of which is the sigmoid function:

$$f(x) = \frac{1}{1+e^{-x}}$$

The output of the hidden layer (H_i) can then be used as the input of the output layer or another concatenated hidden layer. This process continues until the output layer is arrived.



1.2 Hyperparameter search

We tune one hyperparameter for each of the supervised machine learning methods. For LASSO and ridge regression, we change the regularization strength (λ); for elastic net, we alter the linear weight (α); for PCR and PLS, we vary the number of principal components; for random forest and gradient boosting, we adjust the number of trees; for neural network, we tune the number of nodes of the hidden layer. Figure OA.2 presents the hyperparameter search results. Table OA.5 shows the best out-of-sample R-square ($R_{OOS}{}^2$) for each supervised method and their corresponding parameters. It may be surprising that the most popular and advanced neural network approach works the worst among all methods and even underperforms the most basic OLS. This is due to the mismatch between the high-dimensional predictors and the relatively small sample size. NN is a highly parameterized model, and we do not have enough observation to get all parameters well-tuned. This mismatch can also explain why dimension reduction methods (especially PLS)

works particularly well on our dataset. By limiting the predicting variables to only a few principal components, the number of parameters is manageable for our training set.

2 Word Embedding & K-Means Clustering

2.1 Model

2.1.1 Word Embedding model

In practice, word embedding vectors are estimated with a two-layer neural network, the Skip-gram model. Given a sequence of words w_1,w_2,... w_T, the inference problem is to maximize the average log probability of the context of w_t:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p\left(w_{t+j}|w_t\right)$$

where c denotes the size of the context. $p\left(w_{t+j}|w_t\right)$ is calculated as:

$$p\left(w_O|w_I\right)=\frac{exp\left(v'^{T}_{w_O}v_{w_I}\right)}{\sum_{w=1}^{W}exp\left(v'^{T}_{w_O}v_{w_I}\right)}$$

where $v_(w_I)$ and $v_(w_O)$ represent the input and output representation of word w_t, and W denotes vocabulary size. The embedding of w_t is the projection vector between the input and output layer.

2.1.2 K-means clustering algorithm

Given a fixed number of clusters (k), the objective function of k-means is to find a partition of the dataset, such that the within-cluster sum of squared distances between each observation and its closest centroid are minimized. Equivalently, this can be expressed as:

$$\arg\min_{S}\sum_{i=1}^{K}\sum_{x\in S_i}||x-\mu_i||^2$$

where μ_i is the average of data points in S_i.

A k-means algorithm works as follows:

1. Specify the number of clusters k. Randomly select k data points as cluster centroids (μ).

2. For each data point, assign it to the nearest centroid:

$$label\,(i) = \arg\min_j ||x_i - \mu_j||^2$$

3. Update each centroid as the average of data points in that cluster:

$$\mu_j = \frac{1}{||S_j||} \sum_{x \in S_j} x$$

4. Repeat 2) and 3), until the assignments of data points no longer change.

2.2 Preprocessing

Before estimation, we preprocess the raw text step by step to get a cleaner input. We first split all documents into words and convert them to lowercases. We then apply lemmatization to convert all words to its root form. Because word embedding uses contextual information, we do not remove individual words before estimating the vector representation, so as not to affect the sentence structure. After obtaining the word embedding vector, we drop stop-words and low-frequency words that appear less than 20 times in the vocabulary. Finally, we transform preprocessed text into numerical counts that we use as the input of word embedding estimation. The corpus is represented as a $D \times V$ document-term matrix M, where $M_(d, v)$ indicates the count of the v-th word in the d-th document. This is the "bag-of-words" representation. The underlying assumption is that the order of words does not matter. Although this is an oversimplification of reality, it retains a large amount of information while keeping the algorithm simple. The final corpus consists of 2,262 documents and 20,145 unique terms.

2.3 Choice of topics

One important step of k-means is to find the optimal number of topics. We take a data-driven approach to select the best model. To be specific, we apply the "Elbow method" to the distortion score (the sum of squared distances between each point and its assigned centroid), which is a heuristic method to find the appropriate number of clusters on a dataset. "Elbow" refers to the point where adding another cluster does not give much improvement to the model.[10] To determine the "Elbow", Satopaa et al. (2011) propose an algorithm detecting the point of maximum curvature as the elbow, where the curvature can be calculated as:

$$K_f(x) = \frac{f''(x)}{\left(1 + f'(x)^2\right)^{\frac{3}{2}}}$$

Figure OA.3 presents the results on the optimal number of topics. We find that the optimal number of topics detected by the algorithm is 20.

3 Latent Dirichlet Allocation (LDA)

3.1 LDA Model

Latent Dirichlet Allocation (LDA), developed by Blei et al. (2003), is a generative probabilistic modeling approach. The basic idea is that each document can be represented as a probability distribution over various topics, where each topic is a probability distribution over the vocabulary of a corpus. Suppose there are K latent topics, D documents and V unique terms in the corpus. LDA assumes the following data generating process for each document d:

1. Draw β_k from a multinomial distribution, where β_k (a $1 \times V$ vector) denotes the word distribution of topic k for each $k = 1, 2, \ldots, K$.

[10] Hansen et al. (2018) use the method to select the number of topics for the FOMC transcripts.

2. Draw θ_d from a Dirichlet distribution, where θ_d (a $1 \times K$ vector) denotes the topical distribution of document d.

3. For each word w in document d:

   (a) Choose a topic k from θ_d;

   (b) Choose a word w from β_k.

Intuitively, one can think of generating a document with N words as repeating the action of "generating a word" by N times, where each word is generated in two steps: first, roll a K-sided dice to select a topic; conditional on the topic being selected, roll another V-sided dice to choose a word. Note that the probability of obtaining each side is not equal. It corresponds to θ_d and β_k respectively.

Given a corpus and a latent topic number K, the inference problem of LDA is to compute the posterior distribution of hidden variables $\Theta = (\theta_1, \theta_2, \ldots, \theta_D)$ and $\mathrm{B} = (\beta_1, \beta_2, \ldots, \beta_K)$, such that the generated distribution resembles the observed distribution of words of each document. Since the distribution is usually mathematically intractable, it is solved with Gibbs sampling algorithm (Griffiths and Steyvers, 2004) in practice.

3.2 Preprocessing

Similar to word embedding, we preprocess the raw text to get a cleaner input of the LDA model. First, we split all documents into words and convert to lowercases. We then remove common stop-words like “the”, “a” and “I”, as they appear frequently in text but convey little information. Second, we convert all words to its root form, so that words like “communicates”, “communicating” all become “communicate”. Third, we identify common two-word collocations which appears more than 20 times in the corpus. For example, “machine learning” conveys a specific meaning different from “machine” and “learning”. Fourth, we drop infrequent unigrams and bigrams that appear in less than 10 documents. Finally, we convert the preprocessed text to a

document-term matrix, as what we do for word embedding analysis. The final corpus consists of 2,262 documents and 26,410 unique terms.

3.3 Choice of topics

An important yet challenging task of LDA is to find the optimal number of topics (K). As discussed in Hansen et al. (2018), there is a trade-off between model interpretability and statistical goodness-of-fit. If K is too small, the model does not fit the data well, and the topics generated are often too general and mix multiple themes. However, if K is too large, the topics are too fine-grained, which impairs the interpretability of the model. To balance the two effects, we adopt a statistical measure—topic coherence—to select K (Röder et al., 2015). A topic is said to be coherent if its top words frequently co-occur with each other. In particular, we use normalized pointwise mutual information (NPMI) that has been proved to have the largest correlation to human topic coherence ratings to calculate co-occurrence:

$$NPMI\left(w_i, w_j\right) = \frac{\log\left(\frac{P\left(w_i, w_j\right)+\epsilon}{P(w_i)P\left(w_j\right)}\right)}{-\log\left(P\left(w_i, w_j\right)+\epsilon\right)},$$

where $P(w_i)$, $P(w_j)$ and $P(w_i, w_j)$ denote the probability that w_i appears, w_i appears and w_i and w_j jointly appear in the corpus. ϵ is added to avoid taking logarithm on zero.

We consider candidates of topic numbers (K) ranging from 10 to 80 in increments of ten. Figure OA.4 shows the topic coherence of each LDA model with different specifications of K. It indicates that $K = 20$ maximizes the coherence measure and produces the best results. To understand the LDA output with 20 topics, we need to interpret the estimated topics. Since each topic is a probability distribution over all unique terms in the vocabulary, a natural way to name each topic is to read the terms with the highest probabilities and manually assign a label. However, the most frequent terms often appear in multiple topics, making it difficult to distinguish between topics. An alternative way is to look for terms that exclusively appear in a given topic. This is defined as the ratio of a term's probability within a topic to its probability across all topics (Taddy, 2012).

Bybee et al. (2020) adopts this approach to analyze the structure of economics news from the Wall Street Journal. However, this measure may put too much weight on very rare terms, which can also be hard to interpret. Following Sievert and Shirley (2014), we use the relevance measure, which is defined as the weighted average of the two measures above:

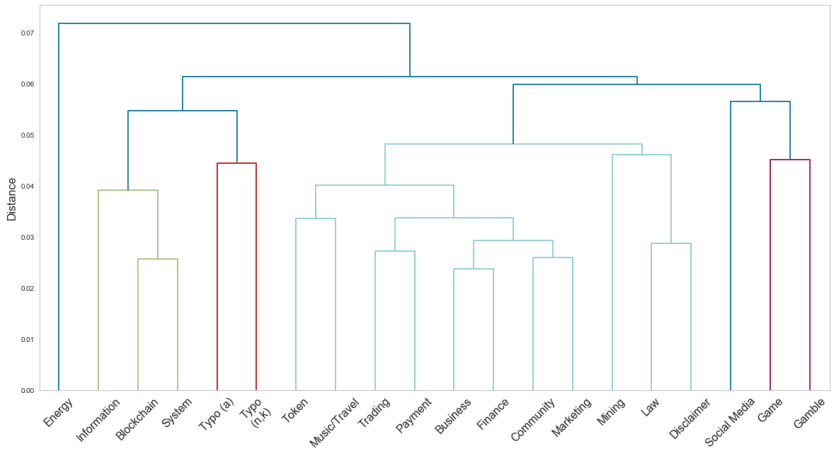
$$Relevance\left(term\ w|topic\ t\right) = \lambda \times p\left(w|t\right) + \left(1-\lambda\right) \times \frac{p\left(w|t\right)}{p\left(w\right)}$$

We find LDA topics with $\lambda = 0.6$ yields the best topic interpretability.

Figure OA.1: **LDA Visualization**

This figure plots the relationship between LDA-based topics. Panel (a) displays the taxonomy generated by hierarchical agglomerative clustering. Panel (b) shows the similarity between topics in a two-dimensional space. The size of the circle represents the relative topic prevalence in the corpus."Information", "blockchain" and "system" are used to construct the LDA-based tech index.

(a) Taxonomy



(b) Multidimensional scaling (MDS)

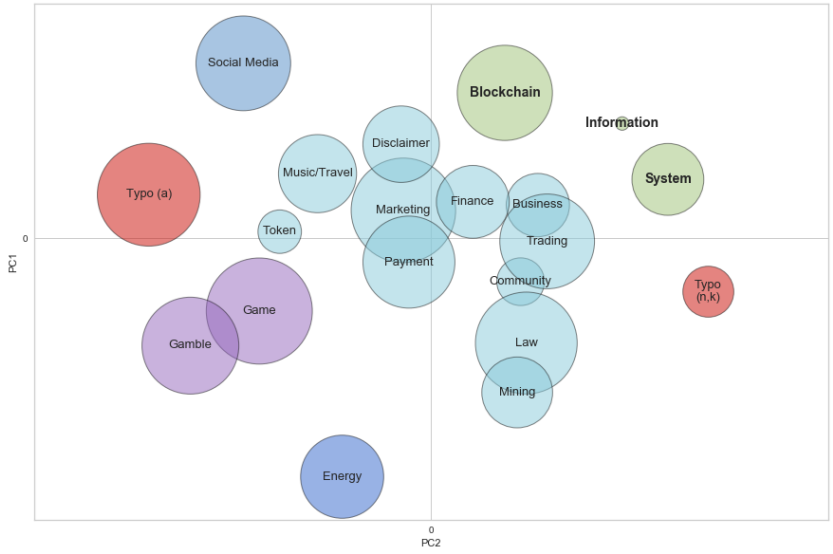


Figure OA.2: **Supervised Learning Hyperparameter Search**

This figure plots the hyperparameter search results of the supervised method. For each subplot, the solid blue line indicates how R^2_{OOS} varies with different parameter choices, and the dashed red line indicates the parameter that gives the best performance.

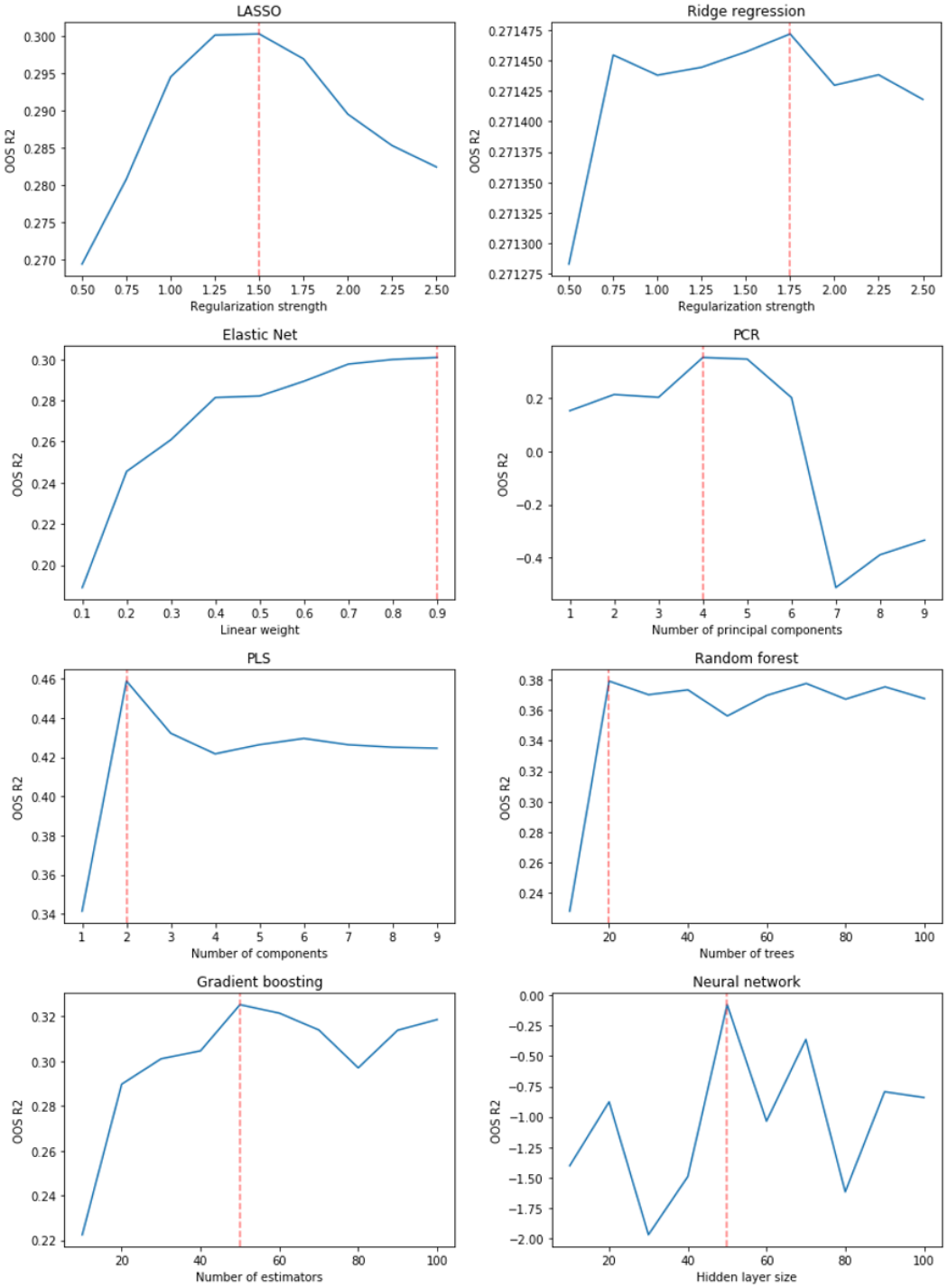


Figure OA.3: **Elbow Method**

This figure shows the elbow method used to select the most appropriate number of clusters. The blue solid line plots the elbow curve of the distortion score, and the red dashed line indicates the "elbow" detected by the algorithm.

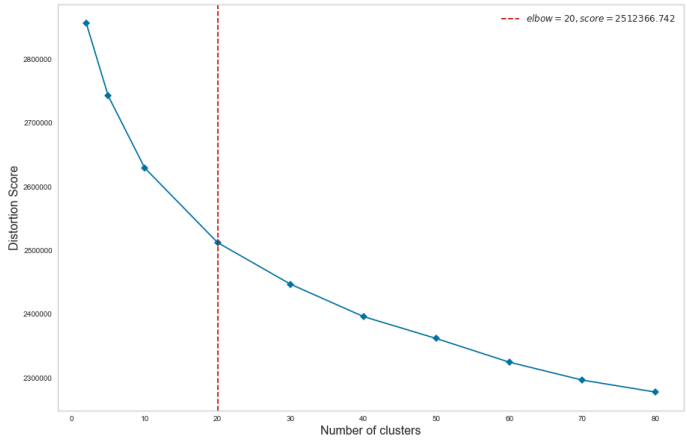


Figure OA.4: **Topic Coherence**

This figure plots the topic coherence measure with different specifications of LDA topic numbers.

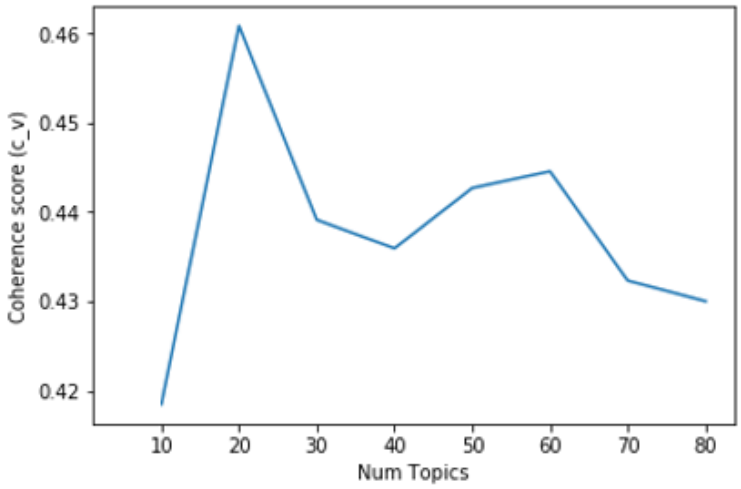


Table OA.1: **Summary Statistics on Whitepaper Status**

This table lists all possible whitepaper status and their frequencies.

	Frequency	**Percent (%)**
Downloaded.	1629	55.90
URL response: client error.	535	18.36
URL response: server error.	104	3.57
Unable to get URL response.	403	13.83
Invalid PDF files.	155	5.32
Whitepaper not found.	54	1.85
Whitepaper is accessible but not downloadable.	27	0.93
Permission is required to access.	7	0.24
Total	2914	100.00

Table OA.2: **Embedding Key Terms**

This table displays the top 15 most frequent terms of each word embedding clustering and their topic labels. The number in parentheses indicates the percentage of terms belonging to the topic.

Topic Label	Most frequent terms
Information (1.8%)	user, datum, contract, transaction, information, process, access, wallet, order, node, would, account, public, key, store
Blockchain (2.9%)	platform, blockchain, system, network, base, smart, development, application, ethereum, chain, protocol, design, developer, open, software
Algorithm (2.7%)	one, model, block, follow, level, different, bitcoin, example, two, point, state, type, function, proof, algorithm
Healthcare (1.5%)	research, tool, health, professional, quality, analysis, report, knowledge, machine, patient, test, ai, medical, al, human
Business (3.5%)	market, business, technology, project, new, also, world, high, ecosystem, community, industry, crypto, solution, digital, cryptocurrency
Transportation (2.7%)	mining, energy, small, area, production, delivery, location, car, range, hardware, retail, gold, physical, home, producer
Marketing (2.5%)	product, content, customer, com, marketing, game, app, event, online, member, social, like, medium, program, marketplace
Token (2.1%)	token, exchange, sale, time, value, fund, payment, asset, coin, purchase, price, number, currency, cost, investment
Verb (1.7%)	use, provide, make, create, include, work, offer, allow, need, develop, increase, support, share, take, require
Operation (1.7%)	service, company, security, management, financial, legal, operation, partner, trust, bank, third, individual, activity, various, foundation
Negative words (2.5%)	result, change, case, without, problem, could, control, even, possible, low, however, due, loss, reduce, therefore
Country/Area (2.2%)	year, country, group, international, university, US, united, startup, partnership, estate, China, center, states, Singapore, Asia
Team/People (2.0%)	experience, advisor, founder, co, expert, manager, director, CEO, tech, entrepreneur, executive, degree, strategic, head, science
Disclaimer (1.3%)	may, risk, party, paper, right, future, whitepaper, term, white, part, form, law, document, person, limit
Typo (a) (2.9%)	ot, ond, dnd, con, ore, os, thot, doto, sen, ho, blockchoin, ds, hove, morket, ct
Typo (n,l) (3.0%)	th, ahd, tor, ih, wiii, hot, oh, ah, biockchain, tokehs, blockchaih, ts, aii, oi, tokeh
URL (3.5%)	https, mm, ii, en, http, de, iii, st, et, _, nd, er, pdf, ng, es
Roadmap (12.4 %)	team, www, launch, plan, page, main, io, utility, roadmap, ltd, usage, introduction, copyright, overview, disclaimer
Name/Brand (18.6%)	man, ago, litecoin, ltc, forex, paypal, anol, wp, sam, proj, nakamoto, hat, wire, cite, eight
Descriptive words (28.4%)	direct, org, flow, yes, late, successfully, old, previously, additionally, pro, soon, whose, ofa, maker, ten

Table OA.3: **Distribution by ICO Industry**

This table lists all ICO industries and their frequencies.

Industry	Frequency	Percent
Business	212	7.27
Charity	15	0.51
Connectivity	40	1.37
Cryptocurrency	384	13.17
Ecology	30	1.03
Finance	219	7.51
Games & Entertainment	174	5.97
Health & Medicine	83	2.85
Internet	56	1.92
Other	223	7.65
Platform	977	33.50
Production	30	1.03
Real Estate	65	2.23
Social Media	45	1.54
Software	94	3.22
Sports	17	0.58
Study	17	0.58
Trading	158	5.42
Transport	40	1.37
Travel	37	1.27
Total	2916	100.00

Table OA.4: **LDA Key Terms**

This table displays the top 15 most relevant terms of each LDA topic. The number in parentheses indicates the relative prevalence of the topics in the corpus.

Topic Label	Most relevant terms ($\lambda = 0.6$)
Information (6.8%)	datum, health, data, patient, medical, healthcare, provider, identity, care, information, doctor, user, service, use, access
Blockchain (8.8%)	node, network, block, transaction, blockchain, proof, consensus, protocol, hash, use, chain, system, message, validator, contract
System (3.1%)	node, storage, cloud, file, quantum, datum, chain, compute, computing, de, system, application, blockchain, storage node, machine
Token (4.5%)	contract, order, exchange, trade, chain, network, asset, protocol, transaction, liquidity, user, token, fee, smart, dispute
Music/Travel (2.4%)	music, artist, travel, diamond, contract, driver, smart, smart contract, song, forest, industry, music industry, ride, cargo
Trading (6.7%)	trading, trader, market, platform, exchange, user, trade, ai, intelligence, crypto, strategy, system, service, use, development
Payment (9.8%)	payment, user, cryptocurrency, wallet, service, coin, merchant, card, exchange, transaction, use, currency, crypto, system
Business (9.3%)	business, product, token, sale, global, customer, consumer, year, blockchain, company, platform, technology, market, industry, experience
Finance (8.5%)	loan, asset, token, bank, credit, estate, platform, borrower, fund, financial, investor, lending, real estate, market, investment
Community (6.7%)	project, token, vote, community, platform, voting, fund, team, ico, user, member, reputation, bounty, crowdfunde, market
Marketing (7.2%)	token, platform, sale, user, marketing, token sale, team, online, use, social, tournament, player, service, advertising, development
Mining (3.2%)	mining, issuer, gold, der, currency, investment, die, mine, investor, EUR, crypto, holder, price, fund, coin
Law (6.1%)	may, token, company, purchaser, risk, law, include, car, purchase, regulation, party, platform, sale, jurisdiction, person
Disclaimer (3.8%)	whitepaper, distributor, statement, token, representation, forward, information, thereof, dissemination, look, constitute, risk uncertainty, person, warranty, uncertainty
Gamble (1.8%)	bet, ticket, gambling, player, casino, betting, sport, lottery, event, game, online gambling, jackpot, poker, online, gamble
Game (3.6%)	game, ad, advertiser, publisher, advertising, gamer, developer, gaming, AR, player, VR, game developer, virtual, user, games
Social Media (3%)	content, video, creator, ond, influencer, user, content creator, fan, medium, doto, viewer, social, tv, blockchoin, photo
Energy (1.6%)	energy, electricity, production, grid, water, solar, power, carbon, plant, renewable, green, waste, renewable energy, oil, fuel
typo (a) (1.4%)	dnd, ot, tor, ore, ond, sid, hove, tth, ds, cube, thot, con, hos, thdt, dny
typo (n,k) (1.7%)	ahd, tol, ih, insurance, wiii, oh, tokehs, ah, ens, blocl, ahy, to_lens, hot, blocl_chain, tol_en

Table OA.5: **Technology Indexes**

This table presents results related to tech indexes. Panel A compares various supervised machine learning methods with their out-of-sample (OOS) R^2 and corresponding hyperparameters. Panel B shows the correlation between the four tech indexes.

Panel A: Comparison of Different Supervised ML Methods

	OLS	LASSO	Ridge	ElNet	PCR	PLS	RF	GB	NN
Hyperparameter	—	$\lambda = 1.5$	$\lambda = 1.75$	$\alpha = 0.9$	$PC = 4$	$PC = 2$	$tree = 20$	$tree = 50$	$node = 50$
OOS R^2 (%)	27.14	30.02	27.15	30.08	35.40	**45.88**	37.91	32.53	-7.71

Panel B: Correlation Matrix of Technology Indexes

	Tech_sup	Tech_embed	Tech_lda	Tech_comp
Tech_sup	1.0000			
Tech_embed	0.5152	1.0000		
Tech_lda	0.4861	0.6838	1.0000	
Tech_comp	0.7929	0.8713	0.8597	1.0000

Table OA.6: **Technology Indexes Validation**

This table validates tech indexes with measures from GitHub. Panel A, B, and C display the supervised, the LDA-based, and the composite tech index respectively. In each column, *watch* measures the number of users subscribing repository updates; *star* indicates the number of "likes" received by the repository; *fork* proxies for the copies made by other developers; *commit* represents the number of times the code has been revised; *branch* is the amount of pointers to specific versions of the repository; and *contributor* reflects how many developers have contributed to the source code. All GitHub measures are in logarithmic forms. The reported t-statistics are based on robust standard errors. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels respectively.

Panel A: Supervised Index						
	(1)	(2)	(3)	(4)	(5)	(6)
	ln(watch)	ln(star)	ln(fork)	ln(commits)	ln(branch)	ln(contributor)
Tech_sup	0.561***	0.635***	0.530***	0.754***	0.429***	0.453***
	(0.064)	(0.081)	(0.073)	(0.089)	(0.052)	(0.057)
Constant	1.985***	1.842***	1.428***	4.081***	1.887***	1.943***
	(0.056)	(0.063)	(0.055)	(0.084)	(0.047)	(0.049)
Observations	861	861	861	861	861	861
R^2	0.107	0.106	0.098	0.090	0.091	0.094
Panel B: LDA-based Index						
	(1)	(2)	(3)	(4)	(5)	(6)
	ln(watch)	ln(star)	ln(fork)	ln(commits)	ln(branch)	ln(contributor)
Tech_lda	0.504***	0.600***	0.496***	0.726***	0.440***	0.454***
	(0.066)	(0.079)	(0.072)	(0.088)	(0.052)	(0.056)
Constant	1.983***	1.836***	1.424***	4.072***	1.879***	1.936***
	(0.056)	(0.062)	(0.054)	(0.083)	(0.046)	(0.048)
Observations	861	861	861	861	861	861
R^2	0.095	0.105	0.095	0.092	0.106	0.104
Panel D: Composite Index						
	(1)	(2)	(3)	(4)	(5)	(6)
	ln(watch)	ln(star)	ln(fork)	ln(commits)	ln(branch)	ln(contributor)
Tech_comp	0.796***	0.940***	0.785***	1.148***	0.665***	0.691***
	(0.072)	(0.089)	(0.083)	(0.094)	(0.059)	(0.062)
Constant	1.949***	1.796***	1.390***	4.023***	1.853***	1.908***
	(0.054)	(0.060)	(0.052)	(0.080)	(0.044)	(0.046)
Observations	861	861	861	861	861	861
R^2	0.161	0.174	0.161	0.155	0.164	0.163

Table OA.7: **Blockchain technology word list**

This table presents the complete word list that we use to count blockchain technology words as an alternative measure of technology sophistication.

accenture	DAPP	gigabyte	protocal
address	DDOS	halve	record
airdrop	DDOS attack	hard fork	relayer
altcoin	decentralize	harware wallet	reproduction
AML	decryption	hash	robustness
API	deposit	hashcash	Satoshi Nakamoto
ASIC	difficulty	hashrate	scalability
authentication	digital asset	hot wallet	scrypt
Bitcoin	digital identity	IBM	self execute
BTC	digital signature	immutable	serialization
block	distributed ledger	IPFS	server
block height	double spend	KYC	SHA-256
blockchain	EEA	ledger	shard
bounty	EIP	liquid democracy	smart contract
bug bounty	encryption	liquidity	soft fork
chain	ERC	mainnet	solidity
cipher	ETH	merkle tree	stable coin
client	Ether	multi signature	stablecoin
coin	Ethereum	NFT	testnet
cold storage	EVM	node	timestamp
cold wallet	exchange	oracle	transaction fee
collective	fiat	private key	validator
confirmation	fiat currency	public key	wallet
consensus	fork	proof	wallet address
cryptocurrency	gartner	proof of authority	workflow
cryptography	gas	proof of stake (PoS)	
DAO	genesis block	proof of work (PoW)	