The Information Content of ICO White Papers

David Florysiak* Alexander Schandlbauer‡

January 15, 2019

Abstract

White papers are the most important source of information provided to potential ICO investors. We decompose their general textual content into informative and standard contents following the methodology of Hanley and Hoberg (2010). We find that white paper document similarity and the informative and standard contents are related to factors such as the number of industry peers or subjectivity of the text. Analyzing whether information content matters, we find that potential fraud firms adjust information content to the product market environment and varying market conditions. User expert ratings on ICO listing sites are practically unaffected by the information content in white papers. Easy-to-extract information such as team size, having KYC, or the number of social media channels determine expert ratings. This suggests that most expert ratings do not produce any excess information and experts probably even do not read or understand the white paper. Initial listing returns are partly related and trading volume is unrelated to white paper information content. For some of our results, we document differences in the structure of information content in response to variations in ICO market conditions, i.e. ICO wave effects.

JEL classifications: G30, M13, O16, K29, K40.

Keywords: Initial coin offerings, ICOs, white paper, textual analysis, information content, potential fraud, ICO expert ratings.

*University of Southern Denmark and Danish Finance Institute, Campusvej 55, 5230 Odense, Denmark. Phone: (+45) 65503940. Email: florysiak@sam.sdu.dk

‡University of Southern Denmark and Danish Finance Institute, Campusvej 55, 5230 Odense, Denmark. Phone: (+45) 93507001. Email: alsc@sam.sdu.dk

1 Introduction

How do risky ventures get financing? Besides venture capital (CV) funding (e.g., Gompers and Lerner, 1999) or initial public offerings in OTC markets (e.g., Ljungqvist, 2007), initial coin offerings (”ICOs”) are an alternative way to raise capital using cryptocurrencies. Financing through ICOs has risen tremendously since 2016. The total amount raised accelerated from $0.9 billion in 2016 to $11.4 billion in 2018. The relevance of this new form of financing is further underlined by the fact that, according to the Forbes magazine, the ICO market was 45% of the traditional initial public offerings (IPO) market and and 31% of the venture capital market during Q2 2018.[1] Several reasons for the popularity of ICOs exist: low costs, time to finance, and regulation and costly financial intermediaries (such as venture capitalists, banks, and stock exchanges) can be avoided.

Compared to IPOs of equity stock, where pre-market information through prospectuses is typically scarce, ICOs disclose information through a ”white paper”. A white paper is voluntary disclosure of information that typically consists of a business idea description, a road-map including key milestones, the intended use of proceeds, the team, and a time schedule for the token sale. This document is the main channel to inform potential ICO investors. However, no standardized format exists and white papers differ tremendously in their length, style, and content. In this paper, we seek to understand if white papers actually convey relevant information and whether this information content matters. More precisely, we test whether information content is related to the detection of potential fraud, ICO expert ratings, and performance and trading volume.

Following the methodology of Hanley and Hoberg (2010), we use a textual analysis to

[1]Moreover, according to TechCrunch, the leading online publisher of technology industry news, ICOs delivered 3.5 times more capital to blockchain startups than VC since 2017.

decompose a white paper's information content into a standard and an informative part. White papers that share many textual elements of both recently published white papers and of industry peer white papers have a high standard content measure. This can also be interpreted as to which degree related white papers borrow or copy text from each other. The remainder of the text, that is not standard, is thus informative. This can also be interpreted to which degree white papers contain new textual elements, and thus excess information, not used by others.

The IPO financing literature documents that IPOs occur in "waves" in which "hot" and "cold" market phases prevail, i.e., the number of firms that go public dramatically changes over time (e.g., Pástor and Veronesi, 2005). Different explanations for IPO waves have been proposed (Lowry et al., 2017). For example, an increase in investor sentiment may inflate equity prices and cause investors to overpay for newly public companies during "hot" periods. Sentiment thereby allows riskier or lower quality firms to raise capital. Alternatively, a firm's demand for capital can change as a result of variations in macrcoeconomic conditions. We also observe a "wave" pattern in the ICO market. Figure 1 shows the run-up in the number of ICOs until early 2018 and a sharp decrease thereafter. As described in a recent Bloomberg article (August 13, 2018), akin to that is the Ethereum price surge in 2017 and the sharp decline thereafter, i.e. the price run-up and the subsequent dip of the cryptocurrency that is (technically) underlying almost all ICOs, which is possibly caused by crypto-related investor sentiment (Drobetz et al., 2018).[2,3]

We predict that the structure of information content in ICO white papers is different

[2]In untabulated results, we document that the Ethereum price Granger-causes the number of ICOs, i.e. the number of ICOs is influenced by lagged Ethereum price but not vice versa.

[3]Ethereum is "a blockchain with a built-in Turing-complete programming language, allowing anyone to write smart contracts and decentralized applications," according to Vitalik Buterin's initial white paper. Similar price run ups have been observed for other crypto currencies such as Bitcoin.

in hot and cold ICO markets. For example, in cold markets, firms may provide both more standard and informative content to attract potential investors. In hot markets the relative importance of standard content is higher and informative content is lower. In the latter case, investor sentiment reduces the importance of information production.

To analyze the information content in ICO white papers, we closely follow Hanley and Hoberg (2010). Our main data source is information that is provided on the ICO listing site www.ICOBench.com, which is generally regarded as one of the leading ICO listing websites. We first ask the question of whether those who draft a white paper use content from other sources that is available at the time of the ICO. Stated differently, is the content "standardized" across different white papers? To answer this question, we construct a variable that measures the similarity between different white papers. We document that white paper similarity is related to several determinants such as being in the same industry or having country restrictions that regulate which investors may participate in the ICO. Additionally, white papers of ICOs with a similar team size and product idea, as measured by the number of milestones, are more similar.

Next, we decompose the general content into standard and informative parts. While the standard content measures how much of the information in a white paper can be related to either other recent white papers or to those within the same industry, the informative part captures any additional information that is not covered by recent or industry peers. We document that different factors determine informative and standard content. For example, the availability of a bounty program or the number of characters per page positively influence the informative content, but not standard content, indicating that different ICO characteristics result in differences in information content. Additionally, the information content changes as market conditions change, e.g., more positively written white papers have less informative content in bad market conditions, i.e., in 2018, but the sentiment polarity does not matter in good market conditions.

We subsequently examine whether the documented differences in informative and standard contents are related to potential fraud, expert ratings of ICOs, performance, and trading volume. We start off by hand-collecting data on potential fraud ICOs and document that the information content in ICO white papers of potential fraud firms responds to the product market environment and varying market conditions. Potential fraud firms control the way in which they write their white papers if they try to scam in small or large industries, measured by the number of peer firms within the industry, or in good vs. bad market conditions.

Expert ratings, that are provided by users on the ICO listing site ICOBench, are practically unaffected by differences in informative and standard contents. Hence, experts disregard any informative or standard content in their rating assignment. Obvious and easy-to-extract information such as team size, having KYC, a bonus program, or the number of social media channels determine expert ratings while the white paper information content is disregarded. This suggests that most expert ratings do not produce any excess information, in fact, self proclaimed "experts" might not even read or understand the white paper.

Not all successful ICOs are subsequently listed; for a subsample of ICOs we are able to obtain historical pricing information. We then regress both the 22-day returns and the 22-day cumulative volume from the initial listing date on our informative and standard content measures. We find that informative content is negatively and standard content positively related to 22-day returns. Hence, this association could be related to potential fraud ICOs that engage in pump and schemes, which quickly result in valuations close to zero. However, more informative content does not necessarily mean that the business idea of the ICO is good. An alternative interpretation is that more informative content allows the market to more efficiently price the digital coin, also called token. The 22-day trading volume is unrelated to information content, only team size is positively correlated with trading volume in our sample. These findings suggest that, first, team members that have retained a significant amount of tokens act as market makers at the exchange and provide liquidity. Second, team members sell off their tokens to exit or scam investors, especially in the first days from the initial listing. Moreover, longer term, 66-day returns or volume are unrelated to the information content in ICO white papers.

This paper makes several contributions to the emerging literature on ICOs: to the best of our knowledge, we are the first to analyze the information content of ICO white papers and its influence in several applications. Beyond documenting that information content matters in potential fraud detection, is unrelated to expert ratings, and correlates with initial returns and trading volume, we believe our findings have implications for the currently discussed regulation of ICOs. Providing information to investors through white papers serves a similar purpose as does the information conveyed in IPO prospectuses. However, important differences in the ICO market exist. We document that on the one hand the information content matters and on the other hand it does not matter where it actually should matter. Our results are thus relevant to regulators in order to design a mandatory structure and information criteria for ICO white papers that serves its purpose best, i.e. providing relevant information to potential ICO investors.

Since ICOs are a very recent phenomena, a limited, yet quickly growing, number of papers have so far analyzed different factors that can influence the success of ICOs. Looking at the empirical side, Adhami et al. (2017) provide one of the first comprehensive description of the ICO phenomenon and of the determinants of token offerings. The authors document that the secondary market for ICO tokens is liquid on the first day of trading. Moreover, the initial underpricing is positive, 24%. Following up on these findings, Momtaz (2018) examines the dynamics of 302 ICOs between 2016 and 2018 and shows that ICOs create, on average, investor value in the short run. In an influential paper, Howell et al. (2018) examine the relationship between issuer characteristics and measures of success, with a focus on liquidity, and find that liquidity and trading volume are higher when issuers offer voluntary disclosure. Also looking at other aspects that influence the success of an ICO, Fenu et al. (2018) show that both the country of origin and a high overall rating play a positive role. Moreover, Amsden and Schweizer (2018) analyze whether the fact that a token is subsequently listed on an exchange and is traded actively influences the success of an ICO. The authors find that venture uncertainty is negatively related to ICO success while venture quality has a positive influence. In spirit of these papers, Fisch (2018) examines the dynamics of ICOs by looking at 238 ICOs in 2016 and 2017. The author highlights that several characteristics of the ICO campaign, the underlying technology, and the language of the white paper determine the amount raised.[4]

To the best of our knowledge there is only one other paper that also looks at the content of ICO white papers. In a concurrent working paper, Feng et al. (2018), use 355 white papers and create an index based on 3 readability factors (the blockchain platform originality, whether the token is related to the company's blockchain platform, and the technicality of the writing) to analyze the amounts raised in an ICO. We differ from this paper in a number of aspects, most crucially, we focus on the similarity of the different white papers and we use the readability of the white papers to explain applications that go beyond the amount raised, i.e., potential fraud or ratings.

There also exists a growing theoretical literature studying the economics of crypto-

[4]Our paper is also related to the literature on voluntary disclosure around equity offerings (see e.g., Lang and Lundholm, 2000; Healy and Palepu, 2001, for seminal papers).

currencies and of ICOs for launching a peer-to-peer platform. Sockin and Xiong (2018) highlight the benefits of crypto-currencies in facilitating decentralized trading among participants of a platform, whereas Li and Mann (2018) present a model that rationalizes the use of ICOs. The latter paper also provides several implications for policy makers and argues against a universal ban that was adopted recently by China and Korea. Catalini and Gans (2018) explore how entrepreneurs can use IPOs to fund venture start-up costs and to elicit demand information through generating buyer competition for the tokens. Similarly, Chod and Lyandres (2018) discusses the potential reasons for why an ICO can dominate traditional venture capital financing.

In section 2, we highlight the importance of the white paper in the ICO process. Section 3 contains our analyzes of the information content of white papers. Section 4 uses our measures for standard and informative content and relates them to several applications. Section 5 concludes.

2 The Role of White Papers in Initial Coin Offerings

Initial coin offerings, also called token sales or crowd-sales, are a mechanism to raise external funding through the emission of crypto-currency tokens, which conceptually are entries on a blockchain. The blockchain publicly records all transactions made in the crypto-currency and the owner of the token has a key that lets her re-assign the token ownership (Yermack, 2017). In an ICO, the token is offered to the public for the first time. Therefore, conceptually, ICOs are similar to conventional initial public offerings of equity stock.[5] However, in contrast to IPOs, tokens typically do not convey voting rights or dividends but may provide access to products or services of the issuing firm (Momtaz, 2018).

[5]See, e.g., Ljungqvist (2007) for a detailed overview of the literature on IPOs.

An ICO typically begins when the firm, that is issuing the token, publishes a document called a "white paper". A white paper is the primary public tool to describe the project.[6] It often includes information on the project's objective, a road-map including key milestones, the intended use of proceeds, the team, and a time schedule for the token sale. Moreover, as many technology start-ups conduct ICOs, their white papers often include a technical description of the underlying technology for which funding is sought. However, ICOs are not limited to technology start-ups but can be used by any firm to raise external funding. White papers are published on the projects' webpages and on ICO listing sites.

The white paper's aim is generally twofold: To provide the information to potential investors and to promote the token. Hence, white papers are similar to IPO prospectuses and include information typically provided in financial statements or business plans (Purda and Skillicorn, 2015). However, both the content and the structure of white papers exhibit clear differences, for example as there is no underwriter involved and no road show to potential investors conducted. In fact, no standardized disclosure format exists and white papers usually reveal very little about the issuing firm (e.g., they often fail to give a postal or other contact address) (Zetzsche et al., 2018). Moreover, white papers are published voluntarily, without any legal, regulatory or exchange related requirements.

Information in white papers need to be credible as they provide the most important source of information to investors (Feng et al., 2018). The SEC recently took an active role in ensuring the white papers' rightfulness. One recent example is Centra Tech, which was charged with orchestrating a fraudulent ICO that raised more than $32 million from thousands of investors. "We allege that Centra sold investors on the promise of new digital technologies by using a sophisticated marketing campaign to spin a web of lies about their

[6]Besides publishing a white paper, projects are commonly also advertised and discussed on various social media channels. See Section 3.2 for details.

supposed partnerships with legitimate businesses," said Stephanie Avakian, Co-Director of the SEC's Division of Enforcement.[7]

3 Information Content Analysis

3.1 Data and Methodology

Our data consists of all ICOs that were listed on www.ICOBench.com ("ICOBench") between August 2015 and September 2018. This web-page provides a comprehensive overview of ICOs and is generally regarded as one of the biggest ICO listing and rating sites. For each of the 4.053 ICOs, we download all available data using an application programming interface (API). This includes variables such as the name, token, initial price, money raised (in USD) if the ICO has finished, but also textual descriptions of e.g. milestones or team members. Importantly, a link to the white paper is provided, which we subsequently download. If a white paper link no longer works, we search for it using several other web-pages or via the archived internet pages.[8] Our analysis is limited to those ICOs for which we are able to collect the white paper. Moreover, we require a white paper have at least 15.000 characters.[9] ICOBench typically provides several dates for each ICO: A pre-ICO start and end date and an ICO start and end date.[10] We supplement this information with the date an expert rating was provided and with the creation and modification dates of a white paper

[7] See www.sec.gov/news/press-release/2018-53 for more details.

[8] Examples include www.cryptorating.eu/whitepapers, www.icosbull.com, www.icorating.com or www.archive.org.

[9] This requirement corresponds to dropping less than 5% observations and it is done because short white papers typically consist of pictures, which are harder to analyze. Alternatively, some ICOs provide shorter summary versions of the full white paper, which would distort the analysis. As a rule of thumb, the average single spaced page (12pt) contains about 3.000 characters.

[10] At a pre-ico start date, investors are allowed to buy tokens before the commencement of the official crowd sale. This price is typically cheaper than at the ICO. 50% of all ICOs in our sample offer such pre-sale.

pdf. To determine when the general market first learned about an ICO, and given that ICOs frequently fail to provide exact dates, we define the date as the earliest time that one of the aforementioned variables exists.[11] Our final sample for which we are able to collect white papers and all other relevant variables, and that also incorporates the below mentioned time and industry restrictions, consists of 2.425 ICOs.

We extract the text and further pdf-specific characteristics such as names of the authors, the producing software, the number of pages, etc. from the white paper. First, the texts are pre-processed such that only words remain and these are lematized. Subsequently, a corpus of all white paper texts is defined. For each document in the corpus, we transform the text into a sequence of unigrams, bigrams, trigrams, and named entities and record the term frequencies and filter stopwords. We then transform the tokenized documents into a sparse document-term matrix of shape (# white papers, # unique terms). We represent term vectors using the "tfidf" (term frequency–inverse document frequency) statistic. White paper-specific, local, term frequencies are multiplied by their corpus-wide, global, inverse document frequencies. Terms appearing in many white papers have higher document frequencies and thus lower weights. We normalize the term vectors by the L2 norm. We filter terms that are both very rare, by requiring a minimum frequency of 5, and that are very common, by disregarding terms that have a relative frequency larger than 95%.

We closely follow Hanley and Hoberg (2010), who develop a methodology to determine the information content of IPO prospectuses, to decompose white papers into either standard or informative content. We deviate in some smaller details. The standard content of a specific white paper is determined by its exposure to other white papers that occurred either just

[11] Despite this categorization, we were not able to obtain a date for 93 observations, out of which 21 have a white paper. Since ICOBench uses a unique and continuous identifier, we proxy for the missing date by choosing the nearest, lagged, ICO.

prior to the ICO, or alternatively, that occurred in the same industry. Hence, to estimate ICO i's exposure to the content of other recent ICOs, we calculate a variable called $norm_{rec,i}$. We denote the normalized term vectors as $norm_{tot,k}$. $norm_{rec,i}$ is calculated as the average of the normalized term vectors for the K ICOs that were filed in the 30 days preceding $i's$ ICO.[12] The formula for $norm_{rec,i}$ is given by:

$$norm_{rec,i} = \frac{1}{K} \sum_{k=1}^{K} norm_{tot,k}. \tag{1}$$

The second component of the standard information content, the variable $norm_{ind,i}$, is calculated as the average of the normalized term vectors of the white papers of the P ICOs that occurred within the same industry as ICO i is in. To be an eligible industry peer, the ICO has to occur within a two months window, that is, at least 30 days and at most 90 days prior. This time choice ensures that the recent and the industry information content do not overlap. $norm_{ind,i}$ is calculated as

$$norm_{ind,i} = \frac{1}{P} \sum_{p=1}^{P} norm_{tot,p}. \tag{2}$$

One challenge is to determine in which industry an ICOs is active. We observe that ICOs choose up to 17 industry classifications from a list of 29 categories that ICOBench provides. While the average is relatively low, 2.9 industry categories, we determine the appropriate peers by applying an approach that is similar to the text-based network industry classification of Hoberg and Phillips (2010). That is, we calculate a similarity score based

[12]We limit K to at most 30 ICOs to account for the fact that the average founder of an ICO is unlikely to read/screen more white papers than that. This results in the average firm being subject to 30 recent ICOs. For each ICO that has more potential recent peers, we randomly choose 30 peers. This restriction reduces noise and if it were relaxed, the average number of recent ICOs would increase to 248.

on the chosen industry classifications. Moreover, to account for the fact that some of these categories are more universal and occur more often as a chosen category (e.g., platform, 18% or cryptocurrency, 14%) than others (e.g., banking, 4% or health, 2%), we down-weight these generic categories to 20% while more informative industry classifications receive a weight of 100%.[13] We use a 75% similarity cut-off to determine and allow each firm to have a maximum of 30 peers. This results in firms having on average 19.7 peers, though the distribution is right-skewed (the median is 21 peers).[14]

For each ICO, we then run the following regression:

$$norm_{tot,i} = \alpha_{rec,i} norm_{rec,i} + \alpha_{ind,i} norm_{ind,i} + \epsilon_i \tag{3}$$

The standard content, which measures the proportion of standard words in i's white paper, is then defined as the sum of the coefficients $\alpha_{rec,i}$ and $\alpha_{ind,i}$, whereas the informative content is defined as the sum of the absolute residuals of regression (3).

$$\alpha_{standard,i} = \alpha_{rec,i} + \alpha_{ind,i} \tag{4}$$

The interpretation of $\alpha_{rec,i}$ and $\alpha_{ind,i}$ is as follows: Given the corpus of all words, a normalized term vector contains the relative frequency of the words used in the document. The coefficient on recent is 0.47. This means that if the relative usage of all words in the term vector increases by a marginal unit, this translates into an increase of 0.47 times the

[13] The generic categories include: platform, cryptocurrency, software, internet, smart contract, big data, artificial intelligence, virtual reality, and others. These account for 54% of all industry classifications.

[14] We first sort firms on their similarity measure and chose the 30 most similar firms. If the similarity measure is the same for more than thirty peers, a random sample of 30 firms with the same similarity is chosen. Notice that the maximum of 30 peers restriction does not significantly alter the number of peers. Without it, the average firm would have 24 peers, with a maximum of 65 peers.

average word usage of the average white paper. For instance, if the average white paper's word usage was 0.1, it becomes 0.1047 (0.10 + 0.47*1%).

Panel A of Table 1 presents the summary statistics for the estimates of our standard and informative content. The average document in our sample has a standard content coefficient of 0.8 which indicates that the average white paper is relatively similar to the average past white paper. The intuition behind this is the following: If the standard deviation of $norm_{tot,i}$, $norm_{rec,i}$, and $norm_{ind,i}$ are close to zero, $\alpha_{rec,i}$ and $\alpha_{ind,i}$ should be close to 0.5. However, a higher standard deviation in the the word vectors causes the standard information to be less than one. The influence from recent ICOs is slightly larger than that of past industry ICOs. The average of the informative content measure, i.e., the average of the sum of the regression residuals, is 11.4.[15] Moreover, the standard deviation of both the standard and the informative content is seizable (0.3 and 3.1), indicating that a large cross-sectional variation exists.

3.2 Descriptive Statistics

Table 1 also reports the summary statistics for a large number of variables: Panel B describes the market environment, Panels C and D the ICO and team characteristics, Panel E the white paper characteristics, Panels F to H the financing volume, terms, and the social media activity.

As the general perception from the popular press indicates, the number of ICOs has increased sharply in the last two years. 867 ICOs occurred in 2017 and 1.558 occurred in the first 10 months of 2018.[16] While 56% of the ICOs are successful in raising funds, the typical

[15]The reason for why the informative content is compatibly large is twofold. Both the the tfidf representation of the term vectors and the L2 normalization produce informative values larger one. Notice also that the former reason implies a positive correlation between informative and standard content.

[16]By contrast, there were only 51 ICOs in 2016 according to coinschedule.com, however, for most of these

proceeds from conducting an ICO are relatively small, the average (median) is \$21.6 (\$6.22) million. 43% of the tokens are distributed and around a half of all ICOs have a country or a Know Your Customer (KYC) restriction, the latter being a process where companies can verify the identity of their investors. The average ICO team consists of 13 members (10% have 23 or more) and, typically, the proposed future progress is measured via several, 8, milestones. ICOs also actively use 6.5 out of 9 different social media channels. The most common usage is Twitter and Facebook, 96% and 89%, respectively. 85% of the ICOs have a telegram account, while 73% use Youtube. The least popular social media outlets are Github and Reddit (56% and 61%).

The typical white paper is 34 pages long, though this number ranges from 17 to 56 when looking at the 10^{th} and 90^{th} percentile. It has 42.950 characters, or, stated differently, 1.279 characters per page, which corresponds to roughly 150 words per page. As a rule of thumb, a text written in Times New Roman font in 12 points is 500 words for a single spaced page and 250 words for a double spaced page. Hence, the number of characters/words per page highlight that white-papers frequently use figures, graphs, pictures, or other visual techniques to attract potential investors.

Readability measures the ability of a reader to decipher an intended message. It is, however, difficult to precisely define and several popular measures exist.[17] One of the oldest measures, is the Gunning Fog Index, which measures readability based on sentence and word length. It is defined as a combination of average sentence length and the proportion of complex words (i.e., words with three or more syllables).[18] The average Gunning Fog Index is

ICOs no detailed data is available.

[17]While some authors have pointed out that the file-size is a good proxy for the readability of 10-K filings (e.g., Loughran and McDonald, 2014), the usage of figures and pictures/graphs/figures is likely to distort this (some may or may not require a lot of file storage).

[18]Originally, it was developed to measure the years of formal education required to comprehend a narrative. More recently, it has also been used in finance applications. For instance Li (2008) finds that firms

13.3, which is significantly lower than that of 10-K, which is 18.9 (Loughran and McDonald, 2014). This indicates that the text in white papers is written in a simpler language whereas 10-K's are deems as unreadable.[19]

To measure the sentiment (or the tone) of the white papers we use a natural language processing tool to determine the subjectivity and polarity. This is typically based on a word list, or a "dictionary", which is a collection of items, each with an associated attribute. The subjectivity measure differentiates between objective and subjective words (-1.0 to 1.0) and the polarity measure differentiates between negative and positive words (0.0 to 1.0). One then simply counts the number of words that are associated with a particular sentiment and scales this by the total number of words in the document. Thus, for example, higher proportions of negative words in a document indicate a more pessimistic tone. The white papers' average subjectivity measure is 0.43 while the polarity one is 0.11.

3.3 Sources of Content

To assess the source of content of white papers and to examine in how far white papers use common language, we examine the similarity of different documents. Following Hanley and Hoberg (2010) and Lang and Stice-Lawrence (2015), we calculate the cosine similarity between document i and j, defined as the dot product of the word vectors, which corresponds to the correlation between the document-term vectors. The value ranges between 0 and 1,

---

with lower reported earnings tend to have annual reports that are harder to read as measured by a high Fog Index and by the amount of words used. However, its usage to analyze financial documents has also been questioned. Loughran and McDonald (2014) point out that the usage of complex words can be misleading as the usage of multisyllable words decreases the readability measure. However, many such words are frequently used in financial documents (e.g., the terms "corporation", "management", and "operations" are not difficult for an average investor to comprehend). One alternative to those readability measures is the file size (in megabytes), others are the Flesch–Kincaid or the Flesch Reading Ease score.

[19]Documents with a Gunning Fog Index above 18 are generally considered unreadable since more than 18 years of schooling, i.e., a master's degree, is needed to understand the text.

and where 1 indicates that two white papers are exactly the same. This measure is widely regarded as one of the most popular similarity measures applied to text documents and it allows us to explore whether those who draft a white paper use content from other sources. Hence, we seek to understand to degree to the content of white papers is "standardized".

Figure 2 first plots the document similarity measure for our sample period, which highlights two aspects: it remained relatively stable over the years and there is a rather big cross-sectional variation, as shown by the 95% confidence intervals.

Next, Table 2 examines all unique possible ICO pair observations and regresses the document similarity measure on four sets of explanatory variables that describe the market environment, different restrictions, the team and product idea, and last, the characteristics of the white paper. Those independent variables capture how similar the characteristics of ICO i and j are. Thus for 2425 ICOs, $\frac{2425^2-2425}{2}$ unique pairs exist, and hence, 2.939.100 observations appear in any regression. To ensure unbiased results, given the repeated usage of each white paper, all regressions include ICO fixed effects and the standard errors are clustered at the ICO level.[20]

The results in column 1 first show that white papers of those ICOs that are conducted in the same market environment are more similar. For instance, ICOs from the same industry are likely to share common cultural or economic characteristics which may influence the writing style. ICOs that are in the same industry might share similar product markets and are thus described similarly in their white paper. Interestingly, white papers of ICOs that are conducted within in a close time-wise proximity, as measured by either a 3 month time window or by the absolute date difference, are less similar. This indicates that, on average, white papers of recent ICOs do not serve as a guide. To interpret this finding further, we

[20]For expositional purposes, all coefficients are multiplied by 100.

split the sample into the year 2017 and 2018. While the time difference is positively and significantly related to the document similarity in 2017, the opposite is the case in the year 2018. Hence, when the popularity of ICOs experienced rapid growth, two white papers were more similar if they were written around the same time. This effect reversed as the ICO industry became more mature. Looking at either KYC or country restrictions, we see that white papers potentially written to comply to regulatory issues, are indeed more similar. KYC can proxy for the quality of the product (Blaseg, 2018) and hence ICOs with similar disclosure about the product quality also write more similar white papers. Country restrictions, however, only influenced the the document similarity in 2018, indicating that they did not play a role when the market was rapidly expanding. The team and product characteristics also influence the document similarity. If both ICOs have a similar team size or milestones number, this indicates that the way the ICO is planed is similar (i.e., more milestones can proxy for a more elaborate planing), they tend to write their white papers more similarly. Last, looking at the differences in the lengths of white papers, we see that white papers of similar length also contain similar information.

To interpret these numbers, we calculate the percentage change in the explanatory variables if the document similarity changes by one standard deviation (SD). The document similarity's standard deviation is very similar for both time periods: 0.070 for 2017 and 0.073 for 2018. Looking at the coefficient for the same country in 2017, 0.356, we see that IPOs that are from the same country have an overall document similarity that is 5.1% of one standard deviation higher than those with entirely different country. Looking at 2018, the coefficient is comparable, 0.337, and the economic effect is 4.6%. The economic magnitude of being in the same industry is much bigger and its importance has increased over time: If ICOs are within the same industry, they have a document similarity that is 30.5% (for

2017) and 36.2% (for 2018) of one standard deviation higher than those that are in different industries. ICOs that are occurring within a 90 day window have similarities that are 2.8% of one SD higher in 2017 vs. that are 1.2% lower in 2018. Hence the importance of ICOs that have been conducted just prior is reduced by 4 percentage points. Similarly, albeit less significant, having a KYC became less important when market conditions got better: 4.3% of 1 SD higher vs. 3.8%. Looking at the white paper characteristics, the bottom two lines show that white papers whose absolute length difference is bigger are less comparable, and the importance of this has stayed rather constant, 1.7% of 1 SD higher for the characters difference in 2017 vs. 1.9% of 1 SD higher in 2018. Hence, white papers of comparable length increase the document similarity.

To summarize, Table 2 indicates that white papers contain standardized information from e.g. restrictions and from ICOs' white papers that are in the same industry and country.

3.4 Standard versus Informative Content

Next, we examine the determinants of standard and informative content by regressing both variables on several explanatory variables. To account for time-invariant effects, quarter-year fixed effects are included and all the standard errors are adjusted for clustering by quarter-years.

Table 3 depicts first the overall results (columns 1 and 2) and subsequently splits the sample into two sub-periods: 2017 (columns 3 and 4) and 2018 (columns 5 and 6). Looking at the difference between the determinants of informative and standard content for all years, we see that only a few variables influence one but not the other. A noticeable example is that the percent distributed only influences the standard information content but not the informative one, indicating that firms that distribute more have more comparable white papers. Having a bounty program positively relates to the informative content but not the standard one, and, similarly, the number of characters per page also has a significant positive influence on the informative content variable. These findings indicate that ICOs try to differentiate themselves via bounty programs or via the lay-out of the white papers.

Several interesting results emerge when comparing better (i.e., year 2017) and worse (i.e., 2018) market conditions. The number of ICOs that occurred in the quarter and the number of industry peers plays a bigger role for both the standard and for the informative content in 2017 compared to 2018. In line with this, in 2017, the industry success rate, as measured by the number of firms that were successful in raising money divided by the total amount raised in that industry, is an important determinant for standard and informative content. In 2018, this does not play a role. Hence, as the market matures, the standard and informative content of white papers can be explained less by these market and industry variables. The one month ethereum return is is negatively and significantly correlated with standard and informative in 2018 while the long-term one-year return has a positive influence in 2018. Neither variable influences either standard or informative content of white papers in 2017. Looking at the variables that describe sentiment and readability, we again observe interesting differences between 2017 and 2018. A white paper that is more difficult to read as measured by the Gunning-Fog index reduces the standard and informative information content in 2018. The more positive and less subjective a white paper is written, the higher the information content, both for the standard and the informative part in 2017. In 2018, the sentiment subjectivity, but not the polarity, is significant.

Last, we split the overall regressions into seven separate regressions based on different groups of characteristics.[21] Tables 7 and 8 in the Online Appendix contain these results.

[21] These groups are: (1) market environment, (2) restrictions, (3) team and product idea, (4) white paper characteristics, (5) financing volume, (6) financing terms, and, (7) social media and disclosure.

Several notable differences emerge, e.g., the number of ICOs that occurred in the last quarter is positively related with standard content, whereas no relation is found with informative content, indicating that firms that do an ICO in more busy times have more standardized white paper content. Moreover, those that distribute more, i.e retain less tokens and increase free-float, also have higher standard content while informative content does not change.

4 Applications

4.1 Potential Fraud Detection

Fraudulent activities in stock listed firms can be observed across all firm sizes and exchange segments. Firms listed both on secondary exchanges and in OTC markets are especially subject to fraud. For example, they are prone to pump and dump schemes. According to the website www.news.bitcoin.com, 46% of the ICOs that were initiated in 2017 have already failed, i.e. the market valuation is close to zero. For a number of ICOs, the money that was collected was subsequently stolen or e.g., the team has disappeared. This has sparked the attention by both investors and regulators. For instance, the U.S. Securities Exchange Commission has issued a news statement saying that "while these digital assets and the technology behind them may present a new and efficient means for carrying out financial transactions, they also bring increased risk of fraud and manipulation because the markets for these assets are less regulated than traditional capital markets" (Aug. 21, 2018). A recent example of an ICO that was halted in January 2018 by a court order obtained by the SEC on charges of fraud is AriseBank, that raised $600 million in just 2 months.[22]

[22]Shamoil T. Shipchandler, the director of the SEC's Fort Worth Regional Office, recently stated "attempting to conceal what we allege to be fraudulent securities offerings under the veneer of technological terms like "ICO" or "cryptocurrency" will not escape the Commission's oversight or its efforts to protect

We hand-collect information on potentially fraudulent ICOs using three specialized webpages.[23] On these webpages, users provide different information, such as ”scam”, ”pump and dump scheme”, or ”ponzi and pyramid schemes” for various ICOs. By using this information, we identify 71 potential fraud cases that are also covered in our ICOBench sample. We end up with 43 potential fraud ICOs for which we also have white papers. This number is relatively small for two potential reasons. First, ICOBench screens firms before they get listed on the platform. This is a first hurdle that more obvious fraudulent firms are unlikely to take. Second, ICOBench deletes potential or detected fraud firms from their database. Even though we are able to backfill part of these deletions with the help of an earlier data retrieval, this leads to a selection of ICOs where fraud is scarce and difficult to detect. Our 43 fraud cases are potentially not yet detected fraud cases or classified as no potential fraud by ICOBench. Due to the small sample size, we match only industry peer ICOs to the fraud case sample. This leads to 444 observations in total, 41 fraud ICOs and 403 industry peer ICOs. Potential fraud firms are relatively evenly distributed across time and industry.[24]

In Table 4, we run probit regressions of a potential fraud dummy on our informative and standard content measures and further controls. In column 1 we run the full sample regression and find no association between potential fraud and informative or standard content. Among the control variables, the variable ”number of characters/page count” measures average text per page - i.e., is small if many pictures instead of text is used in the white paper. The coefficient for this variable is negative and highly significant, i.e the more pictures are used

investors.” For more details, see: www.sec.gov/news/press-release/2018-8.

[23]The websites are www.icoscams.net, www.coinopsy.com, and www.deadcoins.com.

[24]Table 9 in the Online Appendix shows the summary statistics for the potential fraud and non-fraud ICOs. A number of noticeable differences emerge. While standard content is similar across the two groups, potential fraud ICOs rely more on past industry content rather. Moreover, these ICOs are active in smaller industries, have a smaller team size and have shorter, albeit more dense, white papers. Interestingly, the amount raised by potential fraudulent ICOs is higher.

instead of text, the more likely the ICO is potential fraud. Moreover, we find that both a larger team size and having a pre-ICO investment possibility actually reduces the likelihood of potential fraud. Instead, longer ICOs are more prone to fraud.

If we run the regression conditional on having a small number of industry peers, i.e. up to 10 peers, the coefficient on informative content in column 2 is positive. An ICO with a white paper that has higher informative content is thus more likely to be potential fraud if the number of industry peers is small. Potential fraud firms may provide more informative content, i.e. use words that are not used in peer ICO white papers, to separate themselves from the small number of industry peers to actively attract more investors. At the same time, an ICO is more likely to be potential fraud if the white paper contains less standard content, i.e. its words are less related to words used in recent and peer ICO white papers. For ICOs with many industry peers, i.e. more than 10 peers, the coefficients on informative and standard in column 3 have flipped signs compared to those with a small number of peers in column 2. This means that more informative content leads to a lower likelihood of being potential fraud if there are many industry peers. One interpretation of this could be that potential fraud ICOs try to reduce informative content and increase standard content to hide among the many peers within their industry.

In the last two columns, 4 and 5, we split our sample into firms that conduct their ICO prior to 2018, i.e. in a good market condition, vs. in 2018, i.e. a bad market condition which prevailed as the Ethereum to USD exchange rate experienced a significant decrease. In 2017, potential fraud is positively related to information content while it is negatively related in 2018. That is, the higher informative content, the lower is the likelihood of potential fraud in a bad market environment (2018). One potential explanation for this finding is that in the 20180downward trend of the ICO, potential ICO investors might require more information about ICOs. This makes it harder for potential fraud firms to scam investors as investors are likely to scrutinize information about ICOs more carefully. This effect is similar as in the case in which there are many industry peers (column 3), where the potential fraud ICO strategy is to hide among their many peers by providing less informative and more standard content. Thus, having many peers or the pressure to provide more information about an ICO in a bad market environment leads potential fraud firms to provide less informative content and more standard content. The opposite is true for having just a little number of industry peers and being in a good market condition. In those cases, potential fraud firms provide more informative content and less standard content, potentially following the strategy of active promotion when potential ICO investors do have less information about peers or do not scrutinize carefully but buy almost any story.

4.2 Expert Ratings

On ICOBench, as well as on other ICO listing platforms, quality ratings are frequently assigned to new ICOs. The main goal of the quality ratings is to help potential investors make more informed investment decisions decisions. However, those ratings have been subject to various discussions. One the one hand, they are subjectively provided both by self-proclaimed experts, i.e., individuals that have experience with ICOs and the respective industry. However, the incentives of those experts are often unclear and a number of users have pointed out that individual expert ratings can actually be purchased by ICOs.[25] On the other hand, ICOBench itself provides ratings by using a rating algorithm.[26] According to ICOBench, while staff is screening the white paper for information on the product presentation, they

[25]See www.medium.com/alethena/this-is-how-easy-it-is-to-buy-ico-ratings-an-investigation-13d07e987394 for a discussion of how easy it is to buy experts ratings.

[26]A description of the rating methodology used by ICOBench can be found here: www.icobench.com/ratings#experts_methodology.

are not evaluating the content and its originality - rather the availability of the information is of key importance. This information enters the ICOBench rating.[27]

Ratings range from 1 to 5 and a higher rating is associated with a better ICO. However, it is obviously difficult to define precisely what makes an ICO "better". Reading through brief explanations that are usually provided alongside a rating, experts comment on various dimensions such as the quality of team members or the business idea. ICO ratings are probably comparable to a Morningstar fund rating. The idea of having a certification of how well ICOs score on properly defined rating categories could help to reduce information asymmetries between investors and ICO team members. Beyond structured information provided on ICOBench (or any other ICO listing site), the white paper should be the most important source of information for a rating. We should thus expect that providing not only more informative, but also standard content, in the white paper should reduce information asymmetry, leading to a higher rating or lower discount of the rating.

The average rating in our sample is 3.2. While some ICOs are rated very infrequently, others are rated by more than 60 experts. On average a rating constitutes of 5.3 experts. If an ICO has just one rating (39.9%), it is ICOBench itself that provides the rating.

Column 1 in Table 5 first focuses on the entire sample and shows that the informative content of a white paper is negatively related to the rating while standard content is positively related. When differentiating the sample between ICOs that have one rating and those that have more than one rating, we can partly separate the effect that comes from the ICOBench platform rating and that from other external experts. Column 2 focuses on the platform's own rating when there are no expert ratings. In this case, more informative content leads

[27]Interestingly, the rating does not directly influence the visibility of the ICO on the ICOBench webpage. Rather, this depends on how much money an ICO is willing to pay to ICOBench - for instance, ICOs can purchase a placement the newsletter or they can influence the visibility on the competitor's profile pages while competitors will not be shown on the profile pages of paying ICOs.

to worse ratings and more standard content leads to better ratings. One potential reason is that when ICOBench's staff is trying to extract the information based on those white papers whose words are less common in peer or recent white papers, the rating is slightly discounted. Intuitively, experts are expected to do a better job in interpreting informative content. However, the results in column 2 reveal that experts disregard any informative or standard content in their assessment. Obvious and easy-to-extract information such as team size, having KYC, a bonus program, or the number of social media channels determine expert ratings while the white paper information content is disregarded. This suggests that most expert ratings do not produce any excess information and experts may not even do not read the white paper. Analyzing expert ratings in good (column 4) vs. bad market environments (column 5) reveals that informative content is unrelated to expert ratings. Standard content leads to increase in ratings in good market conditions while it is unrelated to ratings in bad market conditions.

Evidence in the literature of a positive influence of expert ratings on the likelihood to successfully raise funds are probably driven by investors relying on ratings as supposedly credible source of information. This, however, makes this kind of expert ratings that do not produce any information in markets with very high information asymmetries particularly detrimental.

4.3 Performance and Trading Volume

For a smaller subsample of ICOs, we are able to match historical pricing information from www.coinmarketcap.com.[28] This sample selection only contains ICOs that successfully

[28]The website www.coinmarketcap.com is the major provider of historical pricing information. Our sample is a snapshot that contains information until October 2018, however, it does not cover all of the ICOs in our sample.

raised funds and subsequently list their tokens on an exchange. We first regress the 22-day return[29] right after the listing on informative and standard contents as well as on a number of control variables. Column 1 of Table 6 depicts that while more informative content is associated with lower initial 22-day returns, an increase in standard content is associated with higher returns. One interpretation of this could be related to our findings on potential fraud ICOs in Section 4.1 that engage in pump and schemes and quickly result in valuations close to zero. More informative content does not necessarily mean that the business idea of the ICO is good and an alternative interpretation is more informative content allows the market aggregation function to more efficiently price the token. Other factors, such as having a pre-ICO decreases the 22-day return whereas having a higher free float, as measured by the fraction of distributed tokens, or the number of characters in the white paper increase 22-day returns.

Column 2 in Table 6 contains the regression of the cumulated 22-day trading volume, measured by the value of tokens in USD traded within the 22 days after the initial listing. The information content of the white paper does not influence trading volume. Interestingly, the team size is the only factor that determines trading volume in our sample. This could indicate two things. First, team members that have retained a significant amount of tokens act as market makers at the exchange and provide liquidity. Second, team members sell off their tokens to exit or scam investors, especially in the first days from the initial listing. Analyzing the corresponding 66-day returns (column 3) and the 66-day trading volume (columns 4), the information content of ICO white papers do not influence returns or trading volume. The trading volume is as for the 22-day volume only significantly and persistently positively determined by the number of team members.

[29]Note that trading tokens can take place every day as exchanges are not restricted to trading hours of traditional exchanges. The 22-day return is three weeks after the initial listing.

5 Conclusion

The ICO market has grown significantly over the last two years. As such, the market has matured, the rapid growth of 2017 has slowed down, fundraising grows more competitive, more established companies are entering the market. At the same time, the market remains unregulated and ICOs still rely primarily on their white papers to convey information to attract potential investors. This raises the question of what information is actually conveyed in white papers.

Our paper seeks to answer this question by decomposing the overall information into a standard part (1), that measures how much information a given white paper retrieves from other white papers that have been published either recently or that describe a project that is in the same industry, and into an informative part (2). The latter measures the additional information that a white paper has that is not covered by the standard measure.

We then look at three applications: The likelihood of potential fraud, the expert ratings and ICO listing return. We find that the information content matters or does not matter where it actually should matter. For some of our results, we document that the information content in ICO white papers responds to the product market environment and varying market conditions, i.e. ICO wave effects.

ICOs are likely going to be regulated in the near future. Providing information to investors through white papers is similar to the information conveyed in IPO prospectuses, but still different as the market for ICOs is different due to different firms using it and different market characteristics apply. Our results should support regulators to design a mandatory structure and information criteria for ICO white papers that serves its purpose best, i.e. providing relevant information to potential ICO investors.

References

Adhami, S., G. Giudici, and S. Martinazzi (2017). Why Do Businesses Go Crypto? An Empirical Analysis of Initial Coin Offerings. *Working Paper*.

Amsden, R. and D. Schweizer (2018). Are Blockchain Crowdsales the New 'Gold Rush'? Success Determinants of Initial Coin Offerings. *Working Paper*.

Blaseg, D. (2018). Dynamics of Voluntary Disclosure in the Unregulated Market for Initial Coin Offerings. *Working Paper*.

Catalini, C. and J. S. Gans (2018). Initial Coin Offerings and the Value of Crypto Tokens. *NBER Working Paper*.

Chod, J. and E. Lyandres (2018). A Theory of ICOs: Diversification, Agency, and Information Asymmetry. *Working Paper*.

Drobetz, W., P. P. Momtaz, and H. Schröder (2018). Investor Sentiment and Initial Coin Offerings. *Working paper*.

Feng, C., N. Li, B. Lu, M. F. Wong, and M. Zhang (2018). Initial Coin Offerings, Blockchain Technology, and White Paper Disclosures. *Working Paper*.

Fenu, G., L. Marchesi, M. Marchesi, and R. Tonelli (2018, mar). The ICO Phenomenon and its Relationships with Ethereum Smart Contract Environment. *Working Paper*.

Fisch, C. (2018). Initial Coin Offerings (ICOs) to Finance New Ventures: An Exploratory Study. *Working Paper*.

Gompers, P. A. and J. Lerner (1999). What Drives Venture Capital Fundraising? *NBER Working Paper*.

Hanley, K. W. and G. Hoberg (2010). The Information Content of IPO Prospectuses. *Review of Financial Studies 23*(7), 2821–2864.

Healy, P. M. and K. G. Palepu (2001). Information Asymmetry, Corporate Disclosure, and the Capital Markets: A Review of the Empirical Disclosure Literature. *Journal of Accounting and Economics 31*(1-3), 405–440.

Hoberg, G. and G. Phillips (2010). Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis. *Review of Financial Studies 23*(10), 3773–3811.

Howell, S. T., M. Niessner, and D. Yermack (2018). Initial Coin Offerings: Financing Growth with Cryptocurrency Token Sales. *Working Paper*.

Lang, M. H. and R. J. Lundholm (2000). Voluntary Disclosure and Equity Offerings: Reducing Information Asymmetry or Hyping the Stock? *Contemporary Accounting Research 17*(4), 623–662.

Lang, M. H. and L. Stice-Lawrence (2015). Textual Analysis and International Financial Reporting: Large Sample Evidence. *Journal of Accounting and Economics 60*(2-3), 110–135.

Li, F. (2008). Annual Report Readability, Current Earnings, and Earnings Persistence. *Journal of Accounting and Economics* (45), 221–247.

Li, J. and W. Mann (2018). Initial Coin Offering and Platform Building. *Working Paper*.

Ljungqvist, A. (2007). IPO Underpricing. *Handbook of Corporate Finance: Empirical Corporate Finance 1*, 375–422.

Loughran, T. and B. McDonald (2014). Measuring Readability in Financial Disclosures. *Journal of Finance 69*(4), 1643–1671.

Lowry, M., R. Michaely, E. Volkova, et al. (2017). Initial Public Offerings: A Synthesis of the Literature and Directions for Future Research. *Foundations and Trends in Finance 11*(3-4), 154–320.

Momtaz, P. P. (2018). Initial Coin Offerings. *Working paper*.

Pástor, L. and P. Veronesi (2005). Rational IPO Waves. *Journal of Finance 60*(4), 1713–1757.

Purda, L. and D. Skillicorn (2015). Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection. *Contemporary Accounting Research 32*(3), 1193–1223.

Sockin, M. and W. Xiong (2018). A Model of Cryptocurrencies. *Working paper*.

Yermack, D. (2017). Corporate Governance and Blockchains. *Review of Finance 21*(1), 7–31.

Zetzsche, D. A., R. P. Buckley, D. W. Arner, and L. Föhr (2018). The ICO Gold Rush: It's a Scam, It's a Bubble, It's a Super Challenge for Regulators. *Working paper*.

A Data Library

We provide a data library to this paper at www.florysiak.com/data. Document similarity measures (i.e., the cosine similarity measure), the information content measures (i.e., the standard and informative content) and the industry classification measure for each ICO in our sample can be found in this data library.

B Figures and Tables

Figure 1. Number of ICOs over time

This figure shows weekly Ethereum prices in USD (red, dashed line) as well as the numbers of ICOs that are part of our sample and that are registered on ICOBench (blue, solid line).

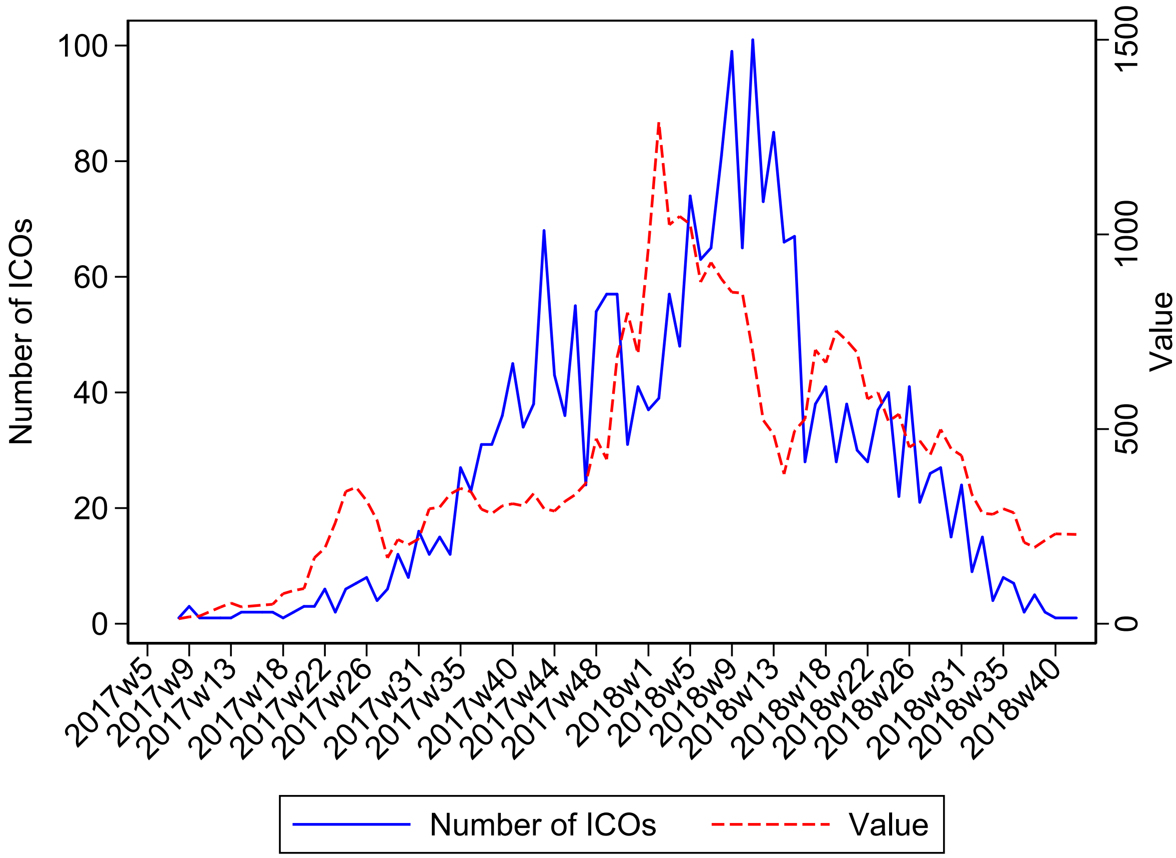

Figure 2. Document similarity

This figure shows monthly average document similarity (solid line), as well as the 95% confidence interval (dashed lines).

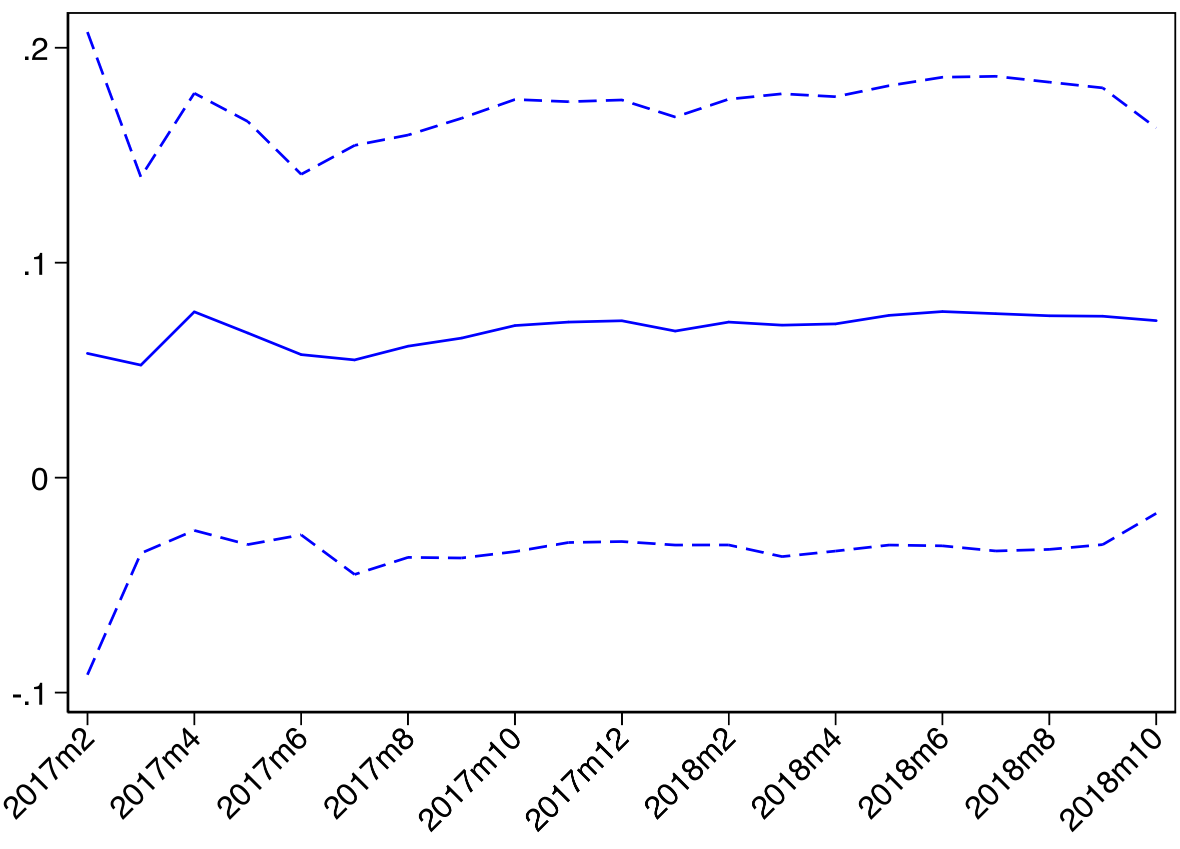

Table 1. ICO summary statistics
This table presents the summary statistics of the variables that are related to Initial Coin Offerings (ICOs). In total, 2425 ICOs were completed on or before Sept. 15, 2018 for which all variables are available. 867 ICOs occurred in 2017 and 1558 occured in 2018. For each variable, the table depicts the the number of non-missing observations, along with the cross-sectional mean, standard deviation, 10th, 50th, and 90th percentile values.

	Count	*Mean*	*St.d.*	*P10*	*Median*	*P90*
Panel A: Standard and informative						
Recent content	2425	0.47	0.28	0.13	0.46	0.83
Past industry content	2425	0.34	0.29	0.03	0.28	0.74
Standard content	2425	0.81	0.30	0.45	0.80	1.20
Informative content	2425	11.37	3.06	7.61	11.38	15.24
Panel B: ICO market environment						
Count nr. of industry peers	2425	13.33	12.38	1.00	9.00	31.00
Industry success rate	2143	0.51	0.25	0.20	0.50	1.00
1-month ethereum return	2425	0.09	0.50	-0.39	-0.06	0.74
Panel C: ICO restrictions						
Restriction (dummy)	2425	0.43	0.49	0.00	0.00	1.00
KYC (dummy)	2425	0.51	0.50	0.00	1.00	1.00
Panel D: Team and product idea						
Team size	2425	13.11	8.33	4.00	12.00	23.00
Number of milestones	2425	8.00	4.80	3.00	7.00	14.00
Panel E: White paper characteristics						
Number of pages	2425	34.16	16.99	17.00	30.00	56.00
Number of characters	2425	42951	29377	17252	37074	73195
Number of characters / page count	2425	1279	650	789	1211	1777
Sentiment subjectivity	2425	0.43	0.07	0.38	0.43	0.47
Sentiment polarity	2425	0.11	0.05	0.07	0.11	0.15
Gunning-Fog index	2425	13.33	30.18	10.19	12.49	14.55
Panel F: Financing volume						
Amount raised (in mil. $)	807	21.34	161.44	0.50	6.22	32.90
Success (amount raised $> 0, \leq$ 15.09.2018)	1445	0.56	0.50	0.00	1.00	1.00
Hardcap (dummy)	2425	0.75	0.43	0.00	1.00	1.00
Softcap (dummy)	2425	0.55	0.50	0.00	1.00	1.00
Token (in bil. $)	2425	9335	459324	0.00	0.06	1.00
Distributed (in percent)	2425	0.43	0.30	0.00	0.50	0.80
Panel F: Financing terms						
Length until start of ICO (in days)	2425	4807	8892	36	123	21658
Pre-ico (dummy)	2425	0.53	0.50	0.00	1.00	1.00
Bonus (dummy)	2425	0.48	0.50	0.00	0.00	1.00
Bounty (dummy)	2425	0.35	0.48	0.00	0.00	1.00
Panel G: Social media and disclosure						
Number of social media channels	2425	6.46	1.84	4.00	7.00	8.00

Table 2. Sources of content

This table depicts the regressions in which the dependent variable is the document similarity of two ICO white papers. One observation is one pair of ICO i and j. All all unique possible ICO pair observations (excluding pairs in which $i = j$). For the sample of 2425 ICOs, $\frac{2425^2-2425}{2}$ unique pairs exist, and hence, 2.939.100 observations appear in any regression. Document similarity is the cosine similarity between document i and j). The independent variables measure how similar the characteristics of ICO i and j are. For expositional purposes, all coefficients are multiplied by 100. ICO fixed effects are included and the t-statistics are adjusted for clustering at the ICO level.

	(1)	(2)	(3)
Coefficient estimates are multiplied by 100.	Document similarity	Document similarity	Document similarity
	all years	*2017*	*2018*
Market environment:			
Same country	0.343***	0.356***	0.337***
	(15.677)	(8.272)	(13.430)
Same industry	2.543***	2.151***	2.641***
	(26.727)	(14.733)	(23.034)
Within same 90 days	-0.053***	0.198***	-0.091***
	(-6.497)	(11.155)	(-10.080)
Abs. date difference	0.010**	-0.103***	0.040***
	(2.401)	(-12.724)	(8.885)
Restrictions:			
Both have KYC	0.244***	0.305***	0.279***
	(21.920)	(4.862)	(24.971)
Both have restrictions	0.173***	0.058	0.202***
	(14.906)	(0.710)	(17.192)
Team and product idea:			
Milestones difference	-0.025***	-0.020***	-0.025***
	(-31.146)	(-11.515)	(-28.670)
Team size difference	-0.028***	-0.031***	-0.027***
	(-46.562)	(-25.621)	(-40.042)
White paper characteristics:			
Abs. length difference (pages)	-0.037***	-0.040***	-0.036***
	(-81.323)	(-36.606)	(-73.530)
Abs. length difference (characters)	-0.136***	-0.118***	-0.140***
	(-20.866)	(-8.875)	(-19.057)
Observations	2939099	473798	2465301
Adjusted R^2	0.267	0.252	0.270

Table 3. Standard vs. informative content
This table presents the regressions for 2425 ICOs where the dependent variable is either the standard content or the informative content of the white paper based on the following first-stage regression for each ICO i: $norm_{tot,i} = \alpha_{rec,i} norm_{rec,i} + \alpha_{ind,i} norm_{ind,i} + \epsilon$. The standard content is the sum of the coefficients $\alpha_{rec,i}$ and $\alpha_{ind,i}$. The informative content is the sum of the absolute residuals. All regressions include quarter-year fixed effects and the t-statistics are adjusted for clustering at the quarter-year level.

	(1)	(2)	(3)	(4)	(5)	(6)
	Standard	Informative	Standard	Informative	Standard	Informative
	All years		*2017*	*2017*	*2018*	*2018*
Number of ICOs in quarter	0.056***	0.186***	0.099**	0.360*	0.052***	0.149**
Number of industry peers (log)	0.070***	0.524***	0.081***	0.535**	0.066***	0.515***
Industry sucess rate	0.020	0.160	0.043***	0.237**	−0.002	0.088
1-month ethereum return	−0.024	−0.307*	0.015	0.020	−0.045**	−0.479***
1-year ethereum return	0.000	0.004*	−0.001	−0.001	0.001***	0.006**
Restriction (dummy)	0.015	0.008	−0.016	−0.401**	0.022	0.098
KYC (dummy)	0.016*	0.165**	0.029***	0.244*	0.020**	0.261**
Team size (log)	−0.013*	−0.156**	−0.006	−0.051**	−0.020**	−0.248***
Number of milestones (log)	0.026***	0.114*	0.014	0.055	0.033**	0.130
Number of characters (log)	0.155***	2.710***	0.157***	2.816***	0.158***	2.678***
Number of characters / page count	−0.000	0.000**	−0.000*	0.000	−0.000	0.000
Sentiment subjectivity	−0.484***	−7.014***	−0.259*	−4.234***	−0.668***	−9.231***
Sentiment polarity	0.105	0.259	−0.055***	−1.277*	0.391	3.389
Gunning-Fog index	−0.001***	−0.013***	−0.000	−0.036	−0.001**	−0.013**
Hardcap (dummy)	0.042**	0.334***	0.029	0.368***	0.057	0.283
Softcap (dummy)	−0.015**	−0.137	−0.005	−0.012	−0.023**	−0.210
Number of token (log)	−0.000	−0.006*	−0.000	−0.002	−0.000	−0.011***
Distributed (in percent)	0.053***	0.282	0.057**	0.456	0.054**	0.236
Length ICO (in days)	−0.003	−0.021	−0.005*	−0.037*	−0.001	−0.011
Pre-ico (dummy)	0.030***	0.276***	0.047*	0.426**	0.022**	0.194*
Bonus (dummy)	0.002	−0.040	−0.001	0.005	0.006*	−0.038
Bounty (dummy)	0.015	0.169**	0.002	0.059	0.025*	0.277***
Number of social media channels	0.004	0.007	0.009	0.028	−0.000	−0.010
Observations	2425	2425	867	867	1558	1558
Adjusted R^2	0.174	0.320	0.196	0.349	0.163	0.307

Table 4. Relation of potential fraud and standard and informative content
This table depicts probit regressions that relate potential fraud to informative and standard content and control variables. All regressions include quarter-year fixed effects and the t-statistics are adjusted for clustering at the quarter-year level.

	(1)	(2)	(3)	(4)	(5)
	Potential fraud	Potential fraud	Potential fraud	Potential fraud	Potential fraud
		< 10 peers	*≥ 10 peers*	*2017*	*2018*
Informative content	−0.010	0.403***	−0.124**	0.144**	−0.657***
Standard content	−0.157	−4.486**	1.318***	−1.509**	4.779***
Number of ICOs in last quarter	−3.285***	15.605***	−2.013***	−0.770**	1.865
Number of industry peers (log)	−0.287	−0.347	1.097***	0.178	−0.132***
Industry sucess rate	0.515	−0.314	1.071	1.075***	1.414***
1-month ethereum return	−0.005	−0.174	1.037**	−0.007	−0.735*
1-year ethereum return	0.005	−0.022	−0.010	0.009*	−0.003
Restriction (dummy)	−0.418	−1.395***	−0.866***	0.000	−1.169***
KYC (dummy)	−0.298	1.157***	0.458	1.623***	−2.603***
Team size (log)	−0.219***	0.116	−0.356***	−0.252*	0.110***
Number of milestones (log)	−0.073	−0.610***	0.015	−0.032	−0.406**
Number of characters (log)	0.341	−0.155	0.082	−0.027	1.840***
Number of characters / page count	−0.001***	−0.003***	−0.001	−0.001**	−0.005***
Sentiment subjectivity	0.811	6.112***	−1.452	−1.535	−0.738
Sentiment polarity	0.794	2.952	−1.275	1.091	−1.265
Gunning-Fog index	−0.002	0.092	0.050**	0.014	0.056***
Hardcap (dummy)	0.039	1.098***	−0.489**	0.559*	−2.140***
Softcap (dummy)	−0.207	−1.349**	0.276	−0.349*	0.233
Number of token (log)	0.012	0.092***	−0.020	0.019	−0.004***
Distributed (in percent)	−0.340	−0.868	−0.559	−0.444	−0.434***
Length ICO (in days)	0.054**	0.156**	0.062***	0.007	0.305***
Pre-ico (dummy)	−0.303***	−1.015**	−0.201	−0.323	0.167
Bonus (dummy)	0.251	0.967**	0.145	0.112	0.469***
Bounty (dummy)	0.283	−3.046***	0.843	−1.320***	2.638***
Number of social media channels	−0.010	0.002	−0.000	−0.078***	0.058
Observations	444	167	253	309	102

Table 5. Relation of expert ratings and standard and informative content
This table depicts OLS regressions that relate expert ratings to informative and standard content and control variables. All regressions include quarter-year fixed effects and the t-statistics are adjusted for clustering at the quarter-year level.

	(1)	(2)	(3)	(4)	(5)
	Ratings	Ratings	Ratings	Ratings	Ratings
		1 rating	*≥ 2 ratings*	*≥ 2 ratings*	
				2017	*2018*
Informative content	−0.010**	−0.016***	0.001	−0.013	0.006
Standard content	0.106**	0.132***	0.017	0.164**	−0.030
Number of ICOs in quarter	−0.010	−0.046***	−0.060**	−0.082**	−0.137**
Number of industry peers (log)	−0.001	0.012	−0.007	−0.007	−0.016
Industry sucess rate	0.011	−0.023	0.047	0.019	0.081
1-month ethereum return	0.076***	0.087***	0.067	0.149**	0.017
1-year ethereum return	−0.001	−0.001	−0.001	−0.001	−0.001
Restriction (dummy)	0.048	0.048	0.050	0.089**	0.036
KYC (dummy)	0.129***	0.096***	0.133***	0.105*	0.127*
Team size (log)	0.316***	0.262***	0.350***	0.333***	0.370***
Number of milestones (log)	0.100***	0.119***	0.059*	0.104**	0.013
Number of characters (log)	0.139***	0.077	0.136**	0.218***	0.073
Number of characters / page count	−0.000	0.000	−0.000	−0.000	0.000
Sentiment subjectivity	−0.073	−0.146	−0.030	−0.059	−0.003
Sentiment polarity	0.199	0.485**	−0.042	0.079	−0.240
Gunning-Fog index	−0.000**	−0.001**	−0.000	0.001	−0.000
Hardcap (dummy)	0.095***	0.068***	0.084***	0.082	0.070
Softcap (dummy)	−0.036*	−0.020	−0.040**	−0.026**	−0.040
Number of token (log)	0.002**	0.002	0.002	0.005***	−0.000
Distributed (in percent)	−0.043	−0.094*	0.013	−0.071	0.078*
Length ICO (in days)	−0.032***	−0.030***	−0.030***	−0.026**	−0.031*
Pre-ico (dummy)	0.057***	0.057***	0.034	−0.013	0.051
Bonus (dummy)	0.009	0.055*	−0.034	−0.030	−0.042
Bounty (dummy)	0.115***	0.107***	0.103***	0.121**	0.092
Number of social media channels	0.170***	0.149***	0.161***	0.145***	0.177***
Observations	2425	967	1458	583	875
Adjusted R^2	0.651	0.633	0.583	0.624	0.510

Table 6. Relation of stock returns or trading volume and standard and informative content
This table depicts OLS regressions that relate stock returns or trading volume to informative and standard content and control variables. All regressions include quarter-year fixed effects and the t-statistics are adjusted for clustering at the quarter-year level.

	(1)	(2)	(3)	(4)
	22-day return	22-day volume	66-day return	66-day volume
Informative content	−0.168**	0.192	0.177	0.072
Standard content	1.214**	−1.723	−1.061	−0.630
Number of ICOs in last quarter	−0.251**	−0.313	0.125	−0.232
Number of industry peers (log)	0.078	−0.163	0.130	−0.152
Industry sucess rate	0.276	−0.320	1.402	−0.204
1-month ethereum return	0.152	−0.019	1.713	0.322
1-year ethereum return	−0.003	0.001	−0.020	−0.003
Restriction (dummy)	−0.048	−0.449	−0.168	−0.366
KYC (dummy)	−0.201*	0.020	0.230	0.222***
Team size (log)	0.063	0.521***	0.543	0.623***
Number of milestones (log)	0.085	−0.224	0.044	−0.054
Number of characters (log)	0.496***	−0.171	−0.805	−0.042
Number of characters / page count	−0.000	−0.000	−0.000	−0.000
Sentiment subjectivity	−4.479*	−0.661	3.552	−2.451
Sentiment polarity	2.158**	1.018	0.147	1.783
Gunning-Fog index	−0.014	0.021	0.001	0.026
Hardcap (dummy)	0.076	−0.219	1.118	−0.037
Softcap (dummy)	−0.072	0.324	−0.820	0.158
Number of token (log)	0.015**	0.012	0.038	0.012
Distributed (in percent)	0.457	−0.357	3.082	−0.393
Length ICO (in days)	0.000	−0.084	−0.013	−0.124**
Pre-ico (dummy)	−0.203*	−0.182	−0.981	−0.220
Bonus (dummy)	−0.004	−0.066	−0.060	−0.314
Bounty (dummy)	−0.218*	−0.646	−0.086	−0.701*
Number of social media channels	−0.085	−0.114	−0.372	−0.085
Observations	474	472	432	432
Adjusted R^2	0.042	0.084	0.000	0.097

C Online Appendix

Table 7. Appendix: Determinants of standard content

This table presents the regressions for 2425 ICOs where the dependent variable is the standard content of the white paper based on the following first-stage regression for each ICO i: $norm_{tot,i} = \alpha_{rec,i} norm_{rec,i} + \alpha_{ind,i} norm_{ind,i} + \epsilon$. The standard content is the sum of the coefficients $\alpha_{rec,i}$ and $\alpha_{ind,i}$. All regressions include quarter-year fixed effects and the t-statistics are adjusted for clustering at the quarter-year level. Columns (1) to (7) describe different groups of characteristics: (1) market environment, (2) restrictions, (3) team and product idea, (4) white paper characteristics, (5) financing volume, (6) financing terms, and, (7) social media and disclosure.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Standard	Standard	Standard	Standard	Standard	Standard	Standard
Number of ICOs in last quarter	0.044***						
Number of industry peers (log)	0.067***						
Industry sucess rate	0.011						
1-month ethereum return	−0.023						
1-year ethereum return	0.000						
Restriction (dummy)		0.034*					
KYC (dummy)		0.061***					
Team size (log)			0.038***				
Number of milestones (log)			0.036***				
Number of characters				0.163***			
Number of characters / page count				−0.000			
Sentiment subjectivity				−0.482***			
Sentiment polarity				0.112			
Gunning-Fog index				−0.001***			
Hardcap (dummy)					0.072***		
Softcap (dummy)					−0.009		
Number of token (log)					0.001		
Distributed (in percent)					0.056***		
Length ICO (in days)						−0.004	
Pre-ico (dummy)						0.047***	
Bonus (dummy)						0.019**	
Bounty (dummy)						0.037***	
Number of social media channels							0.020***
Observations	2425	2425	2425	2425	2425	2425	2425
Adjusted R^2	0.055	0.031	0.037	0.116	0.032	0.032	0.034

Table 8. Appendix: Determinants of informative content

This table presents the regressions for 2425 ICOs where the dependent variable is the informative content of the white paper based on the following first-stage regression for each ICO i: $norm_{tot,i} = \alpha_{rec,i} norm_{rec,i} + \alpha_{ind,i} norm_{ind,i} + \epsilon$. The informative content is the sum of the absolute residuals. All regressions include quarter-year fixed effects and the t-statistics are adjusted for clustering at the quarter-year level. Columns (1) to (7) describe different groups of characteristics: (1) market environment, (2) restrictions, (3) team and product idea, (4) white paper characteristics, (5) financing volume, (6) financing terms, and, (7) social media and disclosure.

	(1) Informat.	(2) Informat.	(3) Informat.	(4) Informat.	(5) Informat.	(6) Informat.	(7) Informat.
Number of ICOs in last quarter	−0.028						
Number of industry peers (log)	0.504***						
Industry sucess rate	0.024						
1-month ethereum return	−0.297						
1-year ethereum return	0.005*						
Restriction (dummy)		0.210					
KYC (dummy)		0.829***					
Team size (log)			0.629***				
Number of milestones (log)			0.173*				
Number of characters				2.695***			
Number of characters / page count				0.000**			
Sentiment subjectivity				−7.042***			
Sentiment polarity				0.252			
Gunning-Fog index				−0.013***			
Hardcap (dummy)					0.676***		
Softcap (dummy)					−0.065		
Number of token (log)					0.001		
Distributed (in percent)					0.073		
Length ICO (in days)						−0.006	
Pre-ico (dummy)						0.408**	
Bonus (dummy)						0.107	
Bounty (dummy)						0.338***	
Number of social media channels							0.212***
Observations	2425	2425	2425	2425	2425	2425	2425
Adjusted R^2	0.034	0.030	0.042	0.294	0.021	0.022	0.030

Table 9. ICO summary statistics: Potential fraud vs. non-fraud
This table presents the summary statistics of the variables that are related to Initial Coin Offerings (ICOs). We differentiate between potential fraud (41 ICOs) and non-fraud cases (403 ICOs).

	Potential fraud		*Non-fraud*	
	Mean	*St.d.*	*Mean*	*St.d.*
Recent content	0.53	0.26	0.64	0.25
Past industry content	0.36	0.28	0.25	0.23
Standard content	0.89	0.33	0.89	0.22
Informative content	12.16	3.25	12.22	3.03
Count nr. of industry peers	12.66	8.13	20.40	22.32
Industry success rate	0.64	0.22	0.70	0.21
1-month ethereum return	0.32	0.64	0.74	0.60
Restriction (dummy)	0.17	0.37	0.10	0.32
Team size	13.71	8.69	15.90	10.89
Number of milestones	7.72	5.67	5.70	3.83
Number of pages	33.12	15.85	47.80	21.81
Number of characters / page count	1435	771	959	342
Sentiment subjectivity	0.43	0.08	0.42	0.03
Sentiment polarity	0.10	0.08	0.10	0.03
Gunning-Fog index	12.50	1.78	12.06	1.70
Amount raised (in mil. $)	35.06	292.83	14.57	15.51
Hardcap (dummy)	0.67	0.47	0.80	0.42
Softcap (dummy)	0.40	0.49	0.40	0.52
Distributed (in percent)	0.39	0.33	0.44	0.33
Pre-ico (dummy)	0.47	0.50	0.40	0.52
Bonus (dummy)	0.44	0.50	0.60	0.52
Bounty (dummy)	0.14	0.35	0.30	0.48
Number of social media channels	6.81	1.66	7.00	1.41

Table 10. ICO summary statistics: One rating vs. more than one rating
This table presents the summary statistics of the variables that are related to Initial Coin Offerings (ICOs). We differentiate between ICOs that have one expert rating (967 ICOs) and those that have more expert ratings (1458 ICOs).

	1 rating		*>1 rating*	
	Mean	*St.d.*	*Mean*	*St.d.*
Recent content	0.50	0.26	0.48	0.27
Past industry content	0.36	0.27	0.30	0.25
Standard content	0.86	0.32	0.77	0.30
Informative content	11.69	3.16	10.58	3.43
Count nr. of industry peers	14.00	8.30	13.73	14.66
1-month ethereum return	0.28	0.57	0.28	0.56
Restriction (dummy)	0.19	0.39	0.20	0.40
Team size	11.63	7.92	10.34	9.45
Number of milestones	7.31	4.98	6.54	4.51
Number of pages	31.03	14.95	34.46	26.02
Number of characters / page count	1349	632	1105	301
Sentiment subjectivity	0.43	0.07	0.45	0.13
Sentiment polarity	0.10	0.07	0.11	0.09
Gunning-Fog index	12.42	1.96	11.71	3.03
Hardcap (dummy)	0.65	0.48	0.54	0.50
Softcap (dummy)	0.41	0.49	0.32	0.47
Distributed (in percent)	0.38	0.33	0.34	0.31
Pre-ico (dummy)	0.49	0.50	0.41	0.50
Bonus (dummy)	0.42	0.49	0.49	0.51
Bounty (dummy)	0.14	0.35	0.20	0.40
Number of social media channels	6.18	1.98	5.80	2.33