Detecting Text Reuse in Cryptocurrency Whitepapers

Andrew Morin
Tandy School of Computer Science
The University of Tulsa
Tulsa, OK, USA
anm1198@utulsa.edu

Marie Vasek
Department of Computer Science
University College London
London, United Kingdom
m.vasek@ucl.ac.uk

Tyler Moore
Tandy School of Computer Science
The University of Tulsa
Tulsa, OK, USA
tyler-moore@utulsa.edu

***Abstract*—Thousands of new cryptocurrencies have been introduced in recent years. Most are introduced with a so-called "whitepaper" containing a mix of technical documentation, legal boilerplate and marketing material. Notably, many proposed currencies reuse text from previous established cryptocurrencies. We analyze the whitepapers from 1 260 actively traded cryptocurrencies and 2 039 ICOs. We develop two measures of similarity. Moderately similar papers reuse text in a portion of the paper, often the legal disclaimers. By contrast, some highly similar whitepapers appear to copy most of the text. 4% of coin and 19% of ICO whitepapers are highly similar to those of traded coins. The fraction rises to 64% for coins and 67% for ICOs when we consider moderate text reuse.**

***Index Terms*—Cryptocurrency, Bitcoin, ICO, Text Reuse, Fraud**

I. Introduction

In 2009, Satoshi Nakamoto published a whitepaper on Bitcoin introducing his new technological idea to the world [1]. The next two cryptocurrencies, released in 2011, did not have whitepapers. Litecoin and Namecoin both made technical innovations, tweaks and improvements, achieved largely using the same code base as Bitcoin.

Now in 2021, instead of three, there are over 7 000 cryptocurrencies [2]. Does this reflect a massive explosion of innovation, or is there less there than meets the eye? The incentives to join this competitive market might be affecting the quality of its outputs. To this end, we seek to understand the landscape of cryptocurrencies by looking at the whitepapers that promote new cryptocurrencies. We focus on identifying the amount of new textual material that goes into each one. We collect as many cryptocurrency whitepapers as possible and analyze the amount of overlap between all pairs of whitepapers.

Our own paper proceeds as follows. Section II describes the datasets and how we collect and measure text reuse for each whitepaper. We empirically select text reuse thresholds and then develop two measures of similarity, one suited to identifying "legalese" and the other more extensive copying. Section III describes how we apply the detector to both datasets. We report on the prevalence of impersonation in coins and ICOs. We find that less popular coins are more likely to reuse text. We also observe that ICO rating systems do not punish ICOs that reuse text. Section IV reviews related literature, and we conclude in Section V by discussing how we have the current state of cryptocurrencies has led us to this state and whether the current approach has produced satisfactory results.

978-0-7381-1420-0/21/$31.00 ©2021 IEEE

II. Method for Identifying Text Reuse

We first describe our data sources on whitepapers, then detail multiple methods for quantifying text reuse. We develop a method for identifying two tiers of text reuse, then evaluate it against a random sample of similar whitepapers.

A. Data Retrieval

We collect whitepaper URL data from two separate sources. CoinMarketCap (coinmarketcap.com) is a large aggregation website for cryptocurrencies with up-to-date data, including technical documentation such as whitepapers. ICObench (icobench.com) is a large aggregation website for ICOs, which as of the collection date (August 2019), was actively being kept up to date. For both websites, we scraped and downloaded all whitepapers listed. For any cryptocurrency without a whitepaper, we checked the Internet archive, and the project's GitHub repository. From CoinMarketCap, we obtained 1 265 whitepapers from 2 225 cryptocurrencies. From ICObench, we obtained 2 039 whitepapers from 5 576 ICOs.

Next we translated each PDF into raw text using the pdfminer library [3]. For whitepapers with unusual formatting, we first attempted to translate them via Google Cache's raw text, stripping out passwords, running them through OCR readers, and/or looking up online character translation tables. Any whitepaper unable to be translated into text using this process was removed. The summary of our collected data can be seen in Table I.

In addition to the text on each whitepaper, we collected first-seen date, rank, and trading volume data from CoinMarketCap, and release dates and ratings for each ICO from ICObench.

B. Quantifying Text Reuse

We use two methods to measure text reuse between documents. The first method is a straightforward pairwise comparison between two documents. The other method uses winnowing [4], followed by TF-IDF cosine similarity to measure similarity. Stop words (filler words, such as 'the', 'do', 'now',

TABLE I
SUMMARY STATISTICS ON DOWNLOADED WHITEPAPERS.

	CoinMarketCap	ICObench
# listed	2 225	5 576
# with whitepaper links	1 578	5 361
# downloaded	1 265	2 597
# with usable text	1 260	2 039

etc.) are filtered out to avoid measuring meaningless similarities, as well as removal of punctuation and case. This leaves us with an ordered list of words from the original text.

Pairwise Similarity: Every list of words is converted into a sequence of n-grams using sliding windows of size n (e.g.,"the quick brown fox" decomposes to the 3-grams "the quick brown" and "quick brown fox"). We then compare every n-gram text document, A, to every other n-gram text document, B, by iterating through every n-gram in A and calculating the percentage of A's n-grams found within B. We evaluated three different sizes of n-grams: 2, 3, and 4. Any n-gram size over 4 showed negligible improvement over the preceding sizes.

Winnowing and TF-IDF: The second method employs winnowing, TF-IDF and cosine similarity. Winnowing removes whitespace and creates n-grams based on characters (i.e. the first 5-gram for "The quick brown fox" is "thequ").

The first step in winnowing is to turn the pre-processed list of words into a list of hashed n-grams called fingerprints. The TF-IDF process uses these fingerprints and adds a weight to each hash based on the number of occurrences in a document and within the corpus. Finally, cosine similarity is calculated between documents. In Section 3 we demonstrate that these automated methods are conservative and that the true extent of text reuse may be more substantial.

C. Selecting Text Reuse Thresholds

We start by examining the distribution of similarity measures. Table II shows the total number of coin pairs split into 10% wide bins. We observe a large number of comparisons in the highest bracket of similarity (91-100). Two scenarios explain why. First, many coins change their name between ICO and active trading. Second, many cryptocurrency forks will share a large portion of the same code base and whitepaper with the original cryptocurrency.

We also observe a few distinct thresholds for each method. For the N2 method, we see the vast majority of comparisons reside in the 0-20% range, just over a thousand reside in the 20-40% bins, and only a few hundred exist in the rest of the bins combined. There appears to be a moderate similarity threshold around 20%, and a second, high similarity threshold around 40%. We see the same general behavior in each of the four methods, though the location of the thresholds varies.

We then inspect a random sample of 25 whitepapers exceeding the N2 high threshold (41% $\leq$ N2), and another random 25 whitepapers between the low and high thresholds (21% $\leq$ N2 $<$ 41%). 20 whitepapers from N2 bins between 30–100% overlap were found to be copying large sections of text from another whitepaper. For example, Audiocoin and Hicoin have the exact same introduction, as well as identical descriptions of "Coin Age". In another example we find that ZoZoCoin copied virtually all of Propy's whitepaper. 25 of the remaining 30 samples showed significant overlap in at least one section.

TABLE II
THE TOTAL NUMBER OF COMPARISONS FOUND IN EACH 10 PERCENT WIDE BIN BY QUANTIFICATION METHOD.

Overlap Bin (%)	Similarity Measure N2	N3	N4	TF/IDF
0-10	3 796 594	3 804 143	3 804 853	893 166
11-20	8 774	2 115	1 547	1 955 839
21-30	1 054	267	148	832 115
31-40	138	55	60	116 657
41-50	43	35	27	8 057
51-60	21	25	16	537
61-70	24	17	10	120
71-80	15	11	12	23
81-90	19	17	16	74
91-100	323	320	316	417

Finally, we constructed an aggregate method considering all four similarity measures. Our first measure identifies **moderate similarity**. These thresholds were selected to identify localized text reuse in whitepapers. Whitepapers were deemed moderately similar if any one of the similarity measures fell within the following ranges: TF-IDF 32–59%, Pairwise 2-Gram 25–39%, Pairwise 3-Gram 20–34%, Pairwise 4-Gram 11–30%. **High similarity** was thus defined as exceeding the top end of any of those ranges.

D. Legal and Technical Text Reuse

We noticed that when only a portion of the text was reused, it was often boilerplate legal language. For example, "This is not an offer or solicitation for investment advisory services, brokerage services, or other products or services." shows up in 364 whitepapers. When more extensive text is copied, it often goes beyond legalese and includes technical details. We therefore manually labeled legal and technical text reuse in each of our 50 sample whitepapers.

We then applied the moderate and high similarity aggregate measures to this sample data. Of the 25 moderate similarity whitepapers, 16 included legal text reuse, 4 include technical text reuse, and 5 had none. By contrast, all 25 high similarity whitepapers included reuse. 21 reused technical text, while 14 copied legal text. Often, both technical and legal text was reused. For moderate similarity whitepapers, we see 96% accuracy of our framework, with a single false positive. For high similarity, we have 100% framework accuracy. We conclude that the moderate similarity measure is a good proxy for legal text reuse, while high similarity is a good indication of technical, and possibly also legal, text reuse.

To further distinguish how legal and technical language is reused, we now look at *where* the text is reused in the whitepapers. Figure 1 plots three heatmaps. The leftmost heatmap includes all 20 manually identified comparisons where legal text reuse is identified. Each comparison is a column, with

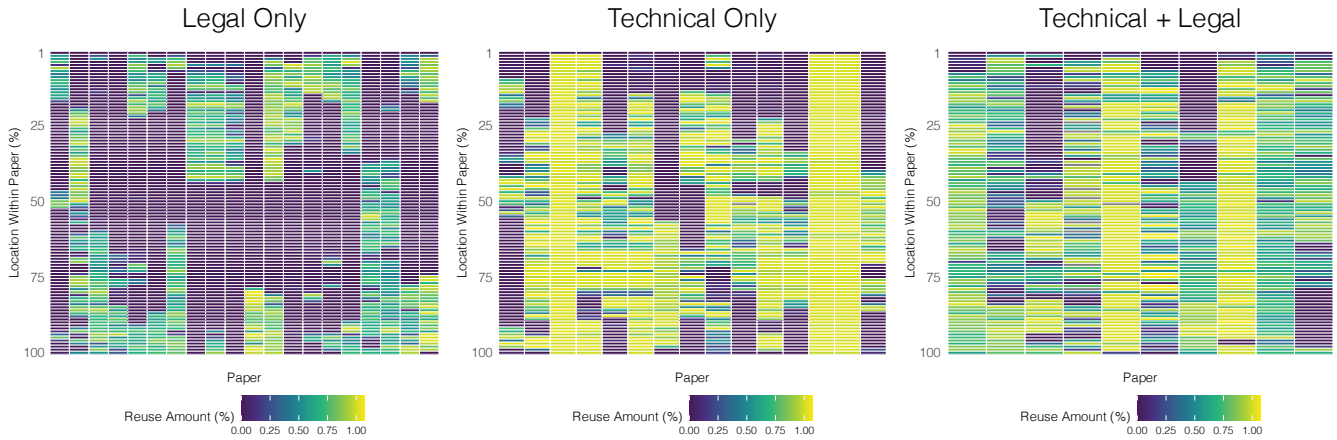

Fig. 1. Heat maps indicating location of reused text for papers reusing legal language (left), technical language (center), and both (right).

TABLE III
UNIQUE IMPERSONATIONS ACROSS SIMILARITY TYPE.

Similarity		Coins Impersonated	Impersonating	Total Pairs
High	Coin	50	53	83
	ICO	359	392	554
Moderate	Coin	568	803	15 900
	ICO	943	1 356	43 711

the papers binned in chronological order and colored by the amount of text similarity found for that bin. Lighter colors indicate higher similarity, while darker colors indicate lower text similarity. For papers that reuse legal text, this tends to happen only at the beginning and/or end of the paper. This is consistent with the pattern where whitepaper authors copy boilerplate language at the beginning or end, then differentiate their coin or token in the "heart" of the paper. By contrast, papers with technical reuse tend to copy language throughout the paper, as seen in both the center and right heat maps.

III. TEXT REUSE PREVALENCE IN COINS AND ICOS

We now take a closer look the prevalence of text reuse. We focus only on impersonations of actively traded coins, since they are the most successful cryptocurrency assets having survived a launch and remain actively traded on multiple exchanges. ICOs, by contrast, include assets that may not have yet launched, if they ever do. Moreover, we noted that whenever two ICOs share significant text overlap, in nearly all cases both ICOs reused text from the same established coin.

A. Incidence of Text Reuse

Table III reports the occurrence of text reuse across whitepapers. The first column shows the number of unique coins impersonated in each category. The second column reports distinct impersonating assets. We deem the coin traded first to be impersonated and the later one impersonating. Finally, the third column indicates the total number of pairs observed. The rows report the tallies for each dataset and similarity.

TABLE IV
COINS THAT COPIED (ROWS) BY THE COINS THAT THEY COPY (COLUMNS). WE BREAK EACH DOWN BY THE CURRENCY RANK (1 BEING BITCOIN) FROM COINMARKETCAP. ICOS ARE UNRANKED.

Rank of Text Reuser		Rank of Active Coin 1-50	51-500	501+
High Sim.	1-50	0	0	0
	50-500	1	6	9
	501+	5	20	42
	ICOs	9	130	437
Moderate Sim.	1-50	34	145	174
	50-500	351	2 085	2 932
	501+	489	3 713	5 977
	ICOs	1 335	16 531	32 245

We find that 53 (4.2%) coin whitepapers have high similarity to other actively traded coins, with 83 distinct impersonating-impersonated pairs. Meanwhile, 392 ICO whitepapers (19.2%) were highly similar to one or more of 359 impersonated coins. Based on the analysis of samples described in Section II.D, we have high confidence that these have widespread text reuse covering technical aspects of the coin.

Even more coins and ICOs exhibit moderate similarity. 64% of coins reused portions of text from 568 distinct coins across 15 900 distinct pairs. With ICO whitepapers, moderate text reuse was similarly pervasive. 1 356 ICO whitepapers, 67% of the total, had significant text reuse across 943 coins.

B. Does Coin Popularity Affect Text Reuse?

Table IV further breaks down the data on impersonating-impersonated pairs by the ranking of the impersonated coin. Among high similarities, we observe that only 15 whitepapers reuse text from the top 50 coins. Bitcoin, at rank 1, is reused 4 times. Bitsend and Bitcloud both include a copy of Bitcoin's whitepaper. GravityCoin presented Satoshi's whitepaper as their own, and Megacoin reused direct sections of the Bitcoin whitepaper. Most instances of high similarity among traded

coins originate from lower ranked coins. We observe 67 instances of coins ranked below 500 reusing text.

For ICOs, the story is similar, but at a greater prevalence. While we have less than twice as many ICO whitepapers as active coin whitepapers, we observe three times as much text reuse. More striking, is that ICO whitepapers reuse text irrespective of coin rank. There is not a statistically significantly different distribution in the rank of coins that each group copied (running a χ^2 test results in a p-value of 0.20).

C. Timing of Text Reuse

We next study the timing of text reuse in whitepapers using the original publication date and the coin's initial trading date as a proxy if publication date is missing. The median time lag is around ten months (295 days) between when a coin trades and then their impersonator starts trading. For cases of high similarity text reuse, the median lag is 394.5 days, while moderate similarity median lag is 294 days.

For the 1 382 ICO whitepapers we have release dates on, 1 354 of them released between 2017 and 2019, during the cryptocurrency boom of 2017, and the subsequent crash in early 2018. In our dataset we see 297 ICOs released in 2017, 207 of which showed moderate similarity to active coin whitepapers, and 97 exhibited high similarity text reuse. In 2018 we see the overall number of ICO's increase to 766, yet the moderate and high similarity text reuse counts are 568 and 153 respectively. Finally, in 2019, we see 291 ICOs released, with 223 high similarity and 77 moderate similarity text reuse. It appears that while text reuse has tracked closely the overall prevalence of ICOs, it has not obviously worsened.

D. Do ICO Rankings Punish Whitepaper Text Reuse?

Finally, we check whether quality rankings employed by ICO tracking websites punish ICOs that reuse text from other coins. ICObench reports two scores on a 0-5 point scale: an aggregate "expert" ranking and an automated measure using a bot, Benchy. One might expect that experts could detect and punish text reuse better. Instead, we find that the median automated rating for an ICO is 3, while the median automated rating for ICOs involved in high similarity text reuse is 3.2. Likewise, the expert rating of original ICOs is 3.2, compared to 3.3 for those with significant reuse. We therefore conclude that, for ICOs at least, there is no substantial penalty for reusing text, and that both experts and automated bots struggle to identify and penalize reuse.

IV. Related Work

A group of law researchers, Zetzche et al., analyzed cryptocurrency whitepapers [5]. They also noted the lack of consistency in these whitepapers. While our focus was looking at the technical documentation overlap between whitepapers, their focus was looking at the issuing entities for cryptocurrency businesses as well as laws applicable to each ICO, as stated in the whitepapers.

Similarly to us, others have started to analyze reuse in the altcoin economy. Reibel et al. analyse the code overlap in cryptocurrencies [6]. They found a significant amount of overlap in entire code files shared across multiple coins. Fröwis et al. analyzed the bytecode-level similarity in tokens and found, that one token system contract was deployed over eight thousand unique times [7]. This work complements our own in that we both find evidence of reuse across projects, though we focus on whitepapers rather than code. Others have researched the ICO economy more generally, including which characteristics of ICOs best predict future success [8] [9].

Our work is inspired by previous work in textual plagiarism detection using simple text-based analysis [4], [10]. It also fits in with the general literature about the use, and general usefulness, of technical documentation [11].

V. Conclusion

Our work finds that 4% of actively traded cryptocurrencies and 19% of ICOs exhibit high rates of text reuse. We find that cryptocurrencies were likely to reuse text from coin whitepapers that were released within a year of their release date. Curiously, ICOs were just as likely to copy unpopular coins as popular ones and high levels of text reuse have no effect on their website ratings.

Text reuse in cryptocurrency whitepapers is not a new concept. The first whitepaper with significant text reuse in the study is from a coin launched 2013. However, with the vast rise in new currency projects after 2017, there has been a reflexive uptick in whitepapers reusing text from older projects. More fundamentally, it calls into question the true extent of innovation that is taking place with the explosion of new coins and tokens.

We conclude by discussing two areas of improvement for the cryptocurrency community to consider.

Technical Documentation: The community seems to slowly migrating towards living, website-based technical documentation. Bitcoin released its developer guides [12]. Ethereum has a living GitHub repository on their development tutorial [13]. These living documents provide more useful (and more easily changeable) documents, and may prove more useful in the long run than the founding whitepapers.

Legal Documentation: As cryptocurrencies move towards registering with the governments that they originate in, governments will be expected to lead in dictating how cryptocurrencies document themselves and set the rules for investing in them. Again, the whitepaper legalese that we have demonstrated to be so often lifted from other projects may have outlived its usefulness.

While setting up fraudulent ICOs by copying text and concepts from other projects remains a big problem to be tackled, perhaps the bigger threat to innovation is an inability to make a compelling case via the whitepaper as currently conceived.

Acknowledgements

We gratefully acknowledge support from the US National Science Foundation Awards No. 1714291 and 1849729 and the Blavatnik Interdisciplinary Cyber Research Center (at Tel Aviv University).

REFERENCES

[1] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008.

[2] CoinMarketCap, "Cryptocurrency market capitalizations," https://www.coinmarketcap.com/. Last accessed 12 March 2021.

[3] "Pdfminer: Python pdf parser and analyzer," https://github.com/pdfminer/pdfminer.six.

[4] E. Stamatatos, "Plagiarism detection using stopword n-grams," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 12, pp. 2512–2527, 2011.

[5] D. A. Zetzsche, R. P. Buckley, D. W. Arner, and L. Föhr, "The ico gold rush: It's a scam, it's a bubble, it's a super challenge for regulators," *University of Luxembourg Law Working Paper*, no. 11, pp. 17–83, 2017.

[6] P. Reibel, H. Yousaf, and S. Meiklejohn, "Short paper: An exploration of code diversity in the cryptocurrency landscape," in *Financial Cryptography and Data Security*. Springer, 2019, pp. 73–83.

[7] M. Fröwis, A. Fuchs, and R. Böhme, "Detecting token systems on Ethereum," in *Financial Cryptography and Data Security*. Springer, 2019, pp. 93–112.

[8] S. T. Howell, M. Niessner, and D. Yermack, "Initial Coin Offerings: Financing Growth with Cryptocurrency Token Sales," *The Review of Financial Studies*, vol. 33, no. 9, pp. 3925–3974, 11 2019.

[9] S. Adhami, G. Giudici, and S. Martinazzi, "Why do businesses go crypto? An empirical analysis of initial coin offerings," *Journal of Economics and Business*, vol. 100, pp. 64–75, 2018.

[10] U. Sapkota, S. Bethard, M. Montes, and T. Solorio, "Not all character n-grams are created equal: A study in authorship attribution," in *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2015, pp. 93–102.

[11] J. Mead, "Measuring the value added by technical documentation: A review of research and practice," *Technical communication*, vol. 45, no. 3, p. 353, 1998.

[12] Bitcoin, "Developer guide," https://bitcoin.org/en/developer-guide.

[13] Ethereum, "Development tutorial," https://github.com/ethereum/wiki/wiki/Ethereum-Development-Tutorial.