LETTER

Discovering Message Templates on Large Scale Bitcoin Abuse Reports Using a Two-Fold NLP-Based Clustering Method

Jinho CHOI[†a)], ***Student Member***, **Taehwa LEE**[†], **Kwanwoo KIM**[†], **Minjae SEO**[†], **Jian CUI**[†], ***and*** **Seungwon SHIN**[†b)], ***Nonmembers***

SUMMARY Bitcoin is currently a hot issue worldwide, and it is expected to become a new legal tender that replaces the current currency started with El Salvador. Due to the nature of cryptocurrency, however, difficulties in tracking led to the arising of misuses and abuses. Consequently, the pain of innocent victims by exploiting these bitcoins abuse is also increasing. We propose a way to detect new signatures by applying two-fold NLP-based clustering techniques to text data of Bitcoin abuse reports received from actual victims. By clustering the reports of text data, we were able to cluster the message templates as the same campaigns. The new approach using the abuse massage template representing clustering as a signature for identifying abusers is much efficacious.

key words: *Bitcoin abuse, text clustering, text template campaign*

1. Introduction

The rise of cryptocurrency over the past 10 years, after introducing Bitcoin [1], has made the digital currency available to everyone without much effort. This ease of access to digital asset made it popular, but it also attracts the attention of cyber criminals. Specifically, cyber criminals noticed that cryptocurrency avoid the central authority and thus possibly guarantee that trading digital assets is under a pseudonym.

By exploiting the advantages of cryptocurrency, cyber abuse has constantly threatened Internet users and practically inflicted huge financial losses [2], [4]. For this reason, delving into actual cases of the real victims and identifying the cyber abuse behavior is necessary. However, previous studies only focused on cryptocurrency abuse instances' specific cases while using a vendor-specific dataset in the perspective of attackers [3]. Additionally, previous studies were more likely tend to center around a specific case of cryptocurrency abuse ecosystem from the abuser side. It denotes that there was a restriction to investigating various types of abuse within a small set of data and was difficult to comprehend the actual victims' cases. Therefore, it is imperative that we need to collect a large-scale of public report data from actual victims to analyze and identify cyber abuse and criminals in terms of the victim side.

For this purpose, cryptocurrency addresses have been a significant signature to identify cyber abuse and criminals. Specifically, Bitcoin addresses have been solid linkage signatures because profit sharing is undeniable evidence. However, abusers also know this fact; thus, they continuously change their cryptocurrency addresses and employ various ways of launder criminal funds to withdraw victims' assets so that they could easily bypass the address-based detection. In addition, Email addresses have also been a helpful signature, but in our public report data, we noticed that cyber abusers tend to deceive the source Email addresses frequently. We cannot rule out the possibility of a false Email address. Therefore, we need to focus on inevitable and unchangeable patterns of cyber threatening, and one of our findings is a message template. In our public report data, Bitcoin abuse text templates in the abusing messages are recycled and tended to be sent to victims with few modifications to abuse as many victims as possible for the efficiency. Therefore, we believe that this finding can be a robust signature and further discovers the characteristic of cyber abuses.

To this end, we have investigated the abuse cases of Bitcoin, the most popular cryptocurrency these days, by massively collecting the real-world Bitcoin abuse reports and analyzing them. Even though some recent studies have used bitcoin abuse report data, they have simply employed the report data as confirmation or ground truth [5]–[7]. In order to improve the understanding cyber abuse and identify the criminals, we have captured 200,597 reports presenting Bitcoin abuse cases [8]. Also, we conducted a study to evaluate, measure, and analyze data on Bitcoin abuse reports from the focusing on the real victim-side as the first attempt. Subsequently, we introduced a text clustering method, leveraging NLP techniques, for analyzing those data in an automated manner. This analysis method enables us to identify various sorts of malicious campaigns, use Bitcoin for malicious purposes, and also help us detect text templates that can be used for detecting those malicious campaigns easily.

Contributions The main contributions in this paper are summarized as follows:

- Propose a data processing method for analyzing Bitcoin abuse reports
- Propose a method for dectecting a message template as a abuse signature
- Evaluate the proposed ideas with real world data and show practical result of clustering

Manuscript received October 1, 2021.
Manuscript revised December 14, 2021.
Manuscript publicized January 11, 2022.

[†] The authors are with Korea Advanced Institute of Science and Technology, Korea.
a) E-mail: cyberz@kaist.ac.kr
b) E-mail: claude@kaist.ac.kr

DOI: 10.1587/transinf.2021EDL8092

2. Methodology

We made the first trial for massage templates clustering in abusing domain with the Bitcoin abuse report. Because most text data were copied and pasted the abuse messages received by the victims, the raw text data is large-scale but contains influenceable noise for the analysis. Therefore, we had several research challenges to overcome.

First, we have to preprocess the text, and then the dataset needs to be pruned to apply the NLP technique to text data effectively. Next, for the efficiency of applying the NLP technique, it was divided into two stages. In order to reduce the computational load, (1) we create a partition that shares discriminating topics with tf-idf and k-means. (2) text similarity and DBSCAN are applied to each partition for fine-grained clustering to identify abuse message campaigns. The work flow is shown in Fig. 1

2.1 Data Collection

To collect real world Bitcoin abuse reports, we have collected user reports from *bitcoinabuse.com* [8], the most popular public database including Bitcoin addresses used for malicious purposes. We employ the *Complete Download API*, provided by *bitcoinabuse.com*, to collect all reports within the specific time period, and finally we obtained 200,597 Bitcoin abuse reports spanned from May 16, 2017, to December 31, 2020. The summary of collected information is presented in Table 1.

2.2 Preprocessing

Our text data in the description of reports is mostly alphabet character, but there were plenty of non-ASCII codes interfering the analysis. Thus, we address the most prominent problem by applying the data preprocessing technique.

As observed in Fig. 2, the non-ASCII character ratio sampled from 30% to 70% of the text data identified in English in the first language analysis before convergence and then creating a dictionary that can convert 729 analogous alphabetic non-ASCII codes into the alphabet of ASCII codes.

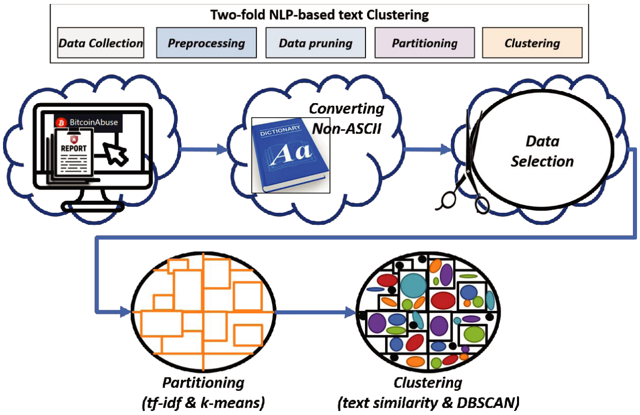


Fig. 1 The work-flow of two-fold NLP-based clustering from data collection to the clustering for identifying the campaigns

Table 1 Data collection from the *bitcoinabuse.com*.

Collection	Description	Count
# of User report	The abuse report	200,597
# of Bitcoin address	The reported Bitcoin address	57,935
# of IP address	The IP address of the reporter	140,612
# of User ID	The registered user's ID (not mandatory)	1,338
Period	May 16, 2017 ~ Dec 31, 2020 (44 months)	

2.3 Data Pruning

After converting the non-ASCII characters, only data whose language is identified in *English* is targeted, although other languages using alphabet characters might also be clustered. However, our goal of the research is the identification of a fine-grained message template campaign. For the accurate clustering results, it is convinced that excluding a small number of other languages does not affect the results as shown in Table 2.

Still, there are factors that cause noise. The too-short text is not accurately language detected, which increases noise, such as being classified as *Unknown*. Therefore, for accurate clustering, only text data of a certain length or longer is selected with the heuristic threshold as over 50 bytes of text length. As a result, our pruned data is a total of 163,185 text data, and the average text length is 495.35 bytes.

2.4 Two-Fold NLP Based Clustering Processing

Even after data pruning, it is impossible to cluster the reports into campaign level at once, mainly because of its massive data size. Thus, we introduce a two-fold approach to perform accurate clustering on massive signature reports: data

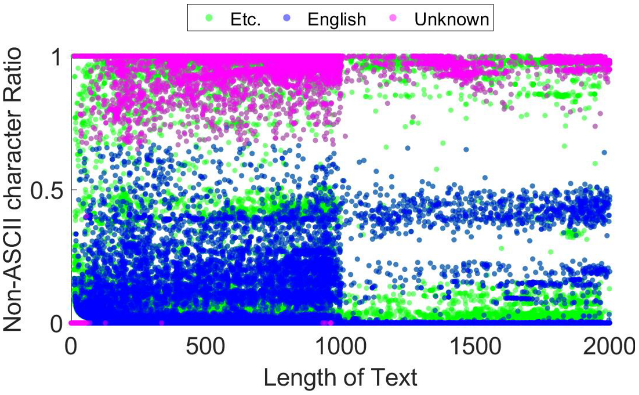


Fig. 2 Distribution by length of Text and non-ASCII Ratio. The text input limit has been expanded from 1,000 bytes to 2,000 bytes since April 2020.

Table 2 Top-10 language distribution text data from reports

Language	Count	%	Language	Count	%
English	16,1458	80.49%	German	2,400	1.2%
Unknown	8,844	4.4%	Polish	1,648	0.82%
French	3,204	1.6%	Italian	1,299	0.65%
Russian	2,885	1.44%	Portuguese	1,217	0.61%
Spanish	2,644	1.32%	Dutch	1,001	0.5%

partitioning and clustering in each partition. The data partitioning is to vaguely group similar posts into multiple partitions to reduce the computation of clustering. After the partitioning, each partition becomes small enough to perform algorithms with time complexity of a square of the number of data points, so the distance between every pair of reports can be calculated in partitions. In the second phase of the clustering, more accurate clustering, such as hierarchical or density-based methods, can be done on each partition.

2.4.1 Partitioning Data

Accurate clustering methods without vectorized data require more than a square of the number of data points. Thus, scale-downing the data size is essential to find out similar contents on the whole dataset accurately. This methods should have high false negatives, which means that the points are in the same partition for the same campaign, but noisy points in a partition do not much matter because there is a backup stage.

To do this, we adopt the k-means algorithm on the tf-idf vectors of the report's data. A significant advantage of the k-means algorithm is that it has relatively low time complexity (*nkdm*, where *n* is the number of data points, *d* is the dimension of the vector, and *m* is the number of k-means). However, the algorithm requires vectorized points of the data samples, so the tf-idf score is calculated on top-1000 frequent words of the whole preprocessed dataset on the strength of preprocessing step before. It is enough to gather templates in the same campaign into a partition and scale down the size.

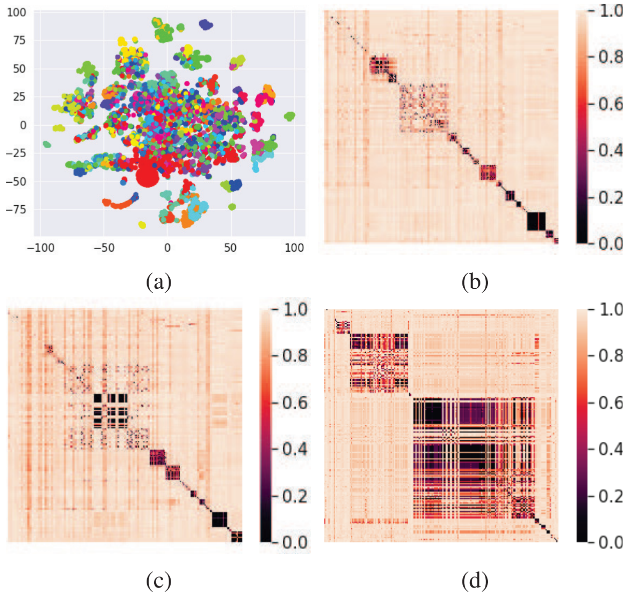
2.4.2 Clustering in Each Partition

After the partitioning, it is small enough to calculate text similarity one by one. For the text similarity, the ratio of edit distance over length of the text is chosen to consider intentional perturbation on the text template. We then perform the DBSCAN clustering algorithm to cluster similar templates, where half of the texts are precisely the same, and leave negative data points not clustered. It guarantees a negligible value of false-positive rate, which can help the manual analysis of the campaigns for security practitioners.

3. Result

As can be seen in Fig. 3 (a), we were able to obtain 100 partitions through a two-fold NLP-based clustering method. The average number of text data included in each partition is 1999.37, which is clustered into an average number of 29.54 clusters. However, finding the association between the partition and the clusters is limited due to blended clusters of various tendencies.

In fact, the meaning of clustering results as a campaign is the most important, because partitioning was a process to reduce the computational overload of fine-grained clustering with text similarity (An example shown as Fig. 3 (b)–(d)). We have obtained 2,819 clusters, and the average number of the text data in each cluster is 57.14.



(a) (b)

(c) (d)

Fig. 3 The visualization of the two-fold NLP-based clustering. (a) shows the t-SNE for 100 partitions with tf-idf and k-means. (b), (c) and (d) denote examples of the heatmap for partitions (ID 6, 12 and 21) by text similarity and DBSCAN.

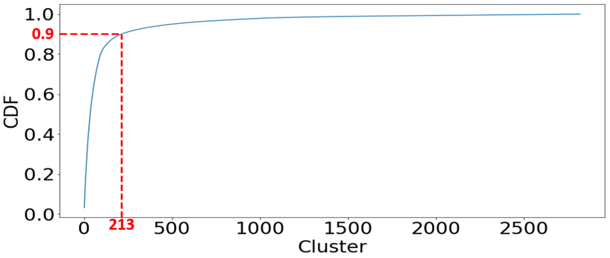


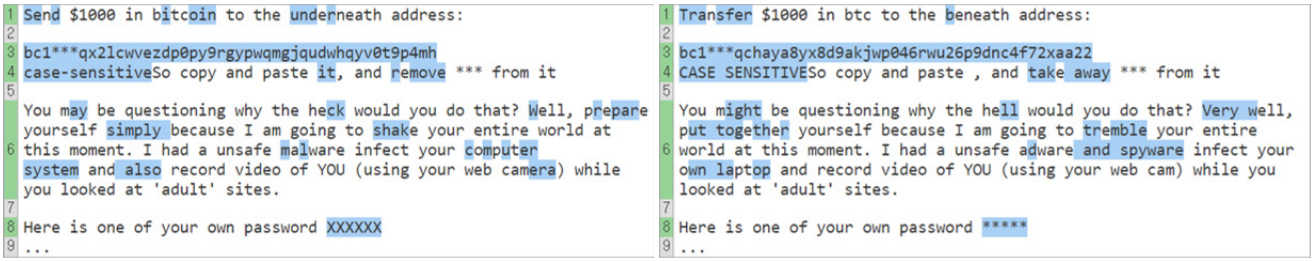
Fig. 4 CDF graph for text data by clusters

Nevertheless, the 213 clusters, which is the upper 7.5% of entire clusters, occupy 90% of total text data as shown in Fig. 4. This upper 7.5% of clusters contained more than 41 text data, with a maximum number of 3,760, an average number of 107. On the other hand, there are 1,375 single clusters that have only a single text data, which indicates they are not campaigns.

Among the 294 clusters with only two text data, there are 97.6% of campaigns share the same template. In order to estimate the accuracy of the clustering result, a complete manual inspection and voting were performed by three researchers for a week. Finally, we can identify 1,081 clusters on the 95% confidence interval range as abuse message template campaigns with 62,872 text data that is 31.34% of total victim's reports.

4. Signature Example

As a use case of our method, Fig. 5 shows abuse messages, which belong to a single cluster and share a specific text



```
Send $1000 in bitcoin to the underneath address:

bc1***qx2lcwvezdp0py9rgypwqmgjqudwhqyv0t9p4mh
case-sensitiveSo copy and paste it, and remove *** from it

You may be questioning why the heck would you do that? Well, prepare
yourself simply because I am going to shake your entire world at
this moment. I had a unsafe malware infect your computer
system and also record video of YOU (using your web camera) while
you looked at 'adult' sites.

Here is one of your own password XXXXXX
...
```

```
Transfer $1000 in btc to the beneath address:

bc1***qchaya8yx8d9akjwp046rwu26p9dnc4f72xaa22
CASE SENSITIVESo copy and paste , and take away *** from it

You might be questioning why the hell would you do that? Very well,
put together yourself because I am going to tremble your entire
world at this moment. I had a unsafe adware and spyware infect your
own laptop and record video of YOU (using your web cam) while you
looked at 'adult' sites.

Here is one of your own password *****
...
```

Fig. 5 The presentation of comparison with text template samples (a) and (b) found in cluster ID 551.

template. The first significant difference between these two messages is Bitcoin address. In this case, the Bitcoin addresses differs from each other, even though the messages show certain similarities which could be associated with the identity of campaign. This is a clear illustration of rapid altering strategy of Bitcoin addresses by abusers as mentioned before. In addition to the Bitcoin addresses, some words are modified to synonym words (e.g., Transfer/Send, take away/remove), and verb also changed from the present tense to the past tense (e.g., may/might). Although the messages were perturbed in various ways, it is possible to confirm sharing template as the signature of the same campaign.

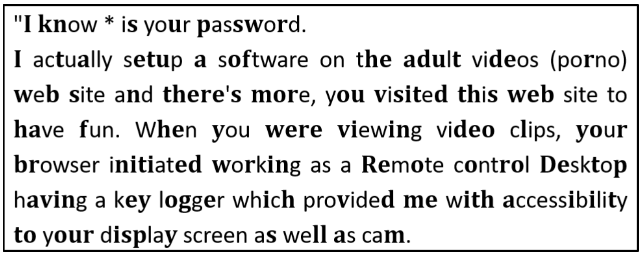
5. Discussion

Furthermore, the majority of description text is a direct copy and paste of the messages received. We trust this as precious data for discovering characteristics of threat groups. For example, it identified the abnormal usage pattern of non-ASCII characters in the original text message templates before converting several campaigns of the upper 5% cluster. Hence, it could be inferred that some attackers tended to use non-ASCII characters to avoid security system filters (as shown in Fig. 6), and this causes the text data keywords to be very noisy and error-prone.

Homoglyph letters (symbols, numbers, letters) look very similar but have completely different encoding; thus, it is used in unicode watermark embedding/ Although there has been previous academic research to detect and convert homoglyphs, we are the first to discover the homoglyphs used in cyber abuse in the wild to the best of our knowledge. To overcome this challenge, we extracted non-ASCII characters used in all homoglyphs techniques and created a dictionary that can convert 729 analogous alphabetic non-ASCII characters into the alphabet of ASCII characters. Although our efforts are partially included in this work, we are convinced that this technique can be used as an additional signature to identify threat groups. Moreover, tracking the sharing of identical message templates post-homoglyph conversion can be an future research topic.

6. Conclusion

The public report data, collected from actual victims around the world, is the latest live data providing a holistic understanding of Bitcoin abusing. However, little research has been conducted on victim reports. Even now, cyber abusers would constantly be sending messages to innocent victims with their templates, which are recycled without much effort for each attack.



"I know * is your password.
I actually setup a software on the adult videos (porno) web site and there's more, you visited this web site to have fun. When you were viewing video clips, your browser initiated working as a Remote control Desktop having a key logger which provided me with accessibility to your display screen as well as cam.

Fig. 6 Example of non-ASCII character in texts

We discovered the value of the public report data for cyber abuse investigation from the victim-side and verified the insight of the abuse message campaign as the signature using NLP-based text clustering. Therefore, it is possible to track the cyber threat groups with signatures for identifying abusers include text their message templates.

References

[1] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," Decentralized Business Review, 21260, 2008.

[2] M. Paquet-Clouston, M. Romiti, B. Haslhofer, and T. Charvat, "Spams meet cryptocurrencies: Sextortion in the bitcoin ecosystem," Proc. 1st ACM Conference on Advances in Financial Technologies, pp.76–88, Oct. 2019.

[3] S. Kethineni, Y. Cao, and C. Dodge, "Use of bitcoin in darknet markets: Examining facilitative factors on bitcoin-related crimes," American Journal of Criminal Justice vol.43, no.2, pp.141–157, 2018.

[4] S. Lee, C. Yoon, H. Kang, Y. Kim, Y. Kim, D. Han, S. Son, and S. Shin, "Cybercriminal minds: an investigative study of cryptocurrency abuses in the dark web," Network & Distributed System Security Symposium, Internet Society, 2019.

[5] F. Oggier, A. Datta, and S. Phetsouvanh, "An ego network analysis of sextortionists," Social Network Analysis and Mining, vol.10, pp.1–14, 2020.

[6] P. Xia, H. Wang, X. Luo, L. Wu, Y. Zhou, G. Bai, G. Xu, G. Huang, and X. Liu, "Don't fish in troubled waters! characterizing coronavirus-themed cryptocurrency scams," arXiv preprint arXiv:2007.13639, 2020.

[7] M. Wang, H. Ichijo, and B. Xiao, "Cryptocurrency address clustering and labeling," arXiv preprint arXiv:2003.13399 (2020).

[8] T.B. Team, "Bitcoin abuse database," https://www.bitcoinabuse.com/, 2020.