PAPER

A Large-Scale Bitcoin Abuse Measurement and Clustering Analysis Utilizing Public Reports

Jinho CHOI[†*a)], *Student Member*, Jaehan KIM[†*b)], Minkyoo SONG[†], Hanna KIM[†], Nahyeon PARK[†], Minjae SEO[†], Youngjin JIN[†], *and* Seungwon SHIN[†c)], *Nonmembers*

SUMMARY Cryptocurrency abuse has become a critical problem. Due to the anonymous nature of cryptocurrency, criminals commonly adopt cryptocurrency for trading drugs and deceiving people without revealing their identities. Despite its significance and severity, only few works have studied how cryptocurrency has been abused in the real world, and they only provide some limited measurement results. Thus, to provide a more in-depth understanding on the cryptocurrency abuse cases, we present a large-scale analysis on various Bitcoin abuse types using 200,507 real-world reports collected by victims from 214 countries. We scrutinize observable abuse trends, which are closely related to real-world incidents, to understand the causality of the abuses. Furthermore, we investigate the semantics of various cryptocurrency abuse types to show that several abuse types overlap in meaning and to provide valuable insight into the public dataset. In addition, we delve into abuse channels to identify which widely-known platforms can be maliciously deployed by abusers following the COVID-19 pandemic outbreak. Consequently, we demonstrate the polarization property of Bitcoin addresses practically utilized on transactions, and confirm the possible usage of public report data for providing clues to track cyber threats. We expect that this research on Bitcoin abuse can empirically reach victims more effectively than cybercrime, which is subject to professional investigation.

key words: *Bitcoin abuse, clustering, cryptocurrency, cyber crime, cyber threat intelligence, public data*

1. Introduction

Cryptocurrency is one of the recent hot topics attracting researchers and practitioners with its usage of blockchain technology to realize a transparent, reliable, and secure digital asset. In addition, the emergence of some notable cryptocurrencies (e.g., Bitcoin [1] and Ethereum [2]) has accelerated the adoption of cryptocurrencies in the real world, allowing easily access to those digital assets; anyone can buy/sell/trade cryptocurrencies based on online trading service providers [3].

Indeed, there is no doubt that cryptocurrencies give a chance of realizing new decentralized financial systems. However, the anonymous aspect of cryptocurrency that allows users to trade digital assets without revealing their identities also attracts the attention of criminals as a side effect. For example, criminals opt to deal with cryptocurrencies when selling drugs or illegal products [4]. As such, there are some recent trends in adopting cryptocurrencies for illicit purposes.

Motivation In order to analyze illicit cryptocurrency abuse, previous studies have tried to measure and understand the abuse instances [5]–[7]. However, these studies mostly focus on specific cases with a small and narrow dataset, which limits the analysis from covering the complete cryptocurrency abuse behavior. Therefore, it is imperative that a large-scale measurement is conducted for in-depth analysis with a wide range of domains with the public data which meets the research demands but is rarely addressed.

Goal In this paper, we aim to assess the value of public data with the statistical measurement and analysis of reported abuse methods. Ultimately, we investigate the transaction of abused bitcoin addresses to demonstrate the possibility of providing clues to track the cybercrime threat groups.

Research Strategy Our workflow is shown as Fig. 1. First, we collect real-world bitcoin abuse reports submitted by victims across the world and extract this large-scale public report data into several attributes for an empirical analysis from diverse perspectives, including long-term timeline and geo-distribution. Also, to certify the bias of data by specific factors, the presence of heavy reporters and their impact are assessed. Next, we explore the reported abuse methods from text data by analyzing the abuse type and comparing extracted common keywords for assessment. Then we could collect meaningful information, including abuse-related channels, based on the extracted keywords. Finally, we investigate the transaction history of collected bitcoin addresses via the blockchain analysis of Bitcoin. We estimate the illicit bitcoin flow as financial damage of cybercrime by bitcoin abusing and certify the possibility of threat groups

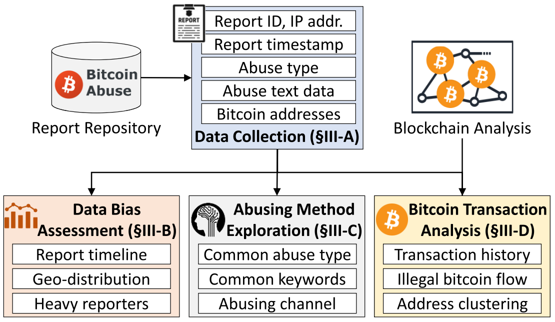


Fig. 1 Overview of work flow

Manuscript received September 2, 2021.
Manuscript revised January 14, 2022.
Manuscript publicized April 7, 2022.

[†]The authors are with KAIST, Korea.

[*]Jinho Choi and Jaehan Kim contributed equally to this work.

a) E-mail: cyberz@kaist.ac.kr
b) E-mail: rlawogks145@kaist.ac.kr
c) E-mail: claude@kaist.ac.kr

DOI: 10.1587/transinf.2021EDP7182

clustering with abused bitcoin addresses.

Contribution

More specifically, we have collected 200,507 abuse reports with various types closely related to real-world events submitted from 2017 to 2020 from 214 countries and analyze 57,935 unique cryptocurrency addresses.

Our main contributions are summarized as follows:

- The first large-scale measurement with public abuse reports collected from real victims all around the world
- A rational investigation on the relationship between real-world events and abuse reports
- Confirmation that the data input from reporters without guidelines can contaminate the data
- Discovery the increasing of bitcoin abuse via social media channels
- A clustering analysis through a trace on the cryptocurrency transaction history of real abusers
- Proof that public data can be utilized to collect clues for tracking cybercrime

2. Background and Related Work

2.1 Cryptocurrency

Cryptocurrency is defined as a decentralized and partially confidential alternative currency [8]. While it is classified as a subset of digital currency [9], it has become more prevalent than digital currency. Bitcoin, the first distributed digital cryptocurrency introduced by Satoshi Nakamoto, relies on the blockchain with cryptographic algorithms and peer-to-peer networks managed via an entirely distributed ledger without a central authority [1]. According to the algorithm, new cryptocurrency blocks are created and awarded to computer users (miners) who solve a predetermined mathematical problem involving cryptographic hashes.

The cryptocurrency network is maintained by the peer-to-peer network of nodes that verify transactions based on their trusts without mediums, and it possesses a great advantage in that it enables users to make anonymous transactions.

The identity of a cryptocurrency user is anonymized by adopting the hashed value of the public key generated from a digital signature algorithm as an address for conducting transactions. The public/private keys owned by a user are recorded in a wallet; the hashed version of the public key is called a wallet address and the private key is used to sign input of a transaction as proof of ownership. Cryptocurrency users can make payments by simply checking their ownership (i.e., private key) through their wallet software. Therefore, the payments can be transferred through the cryptocurrency network without revealing the identities of the participants involved in each transaction. Unlike traditional banking systems, the absence of a central authority means that the financial activities remain anonymous. Thus, governments or central authorities cannot manipulate or control the supply of cryptocurrency.

2.2 Cryptocurrency Abuse in Cybercrime

Cryptocurrencies have been adopted for criminal payments due to their anonymity. Bitcoin is the most popular cryptocurrency in the cybercriminal world. Even though there are various kinds of cryptocurrencies [10], large amounts of Bitcoins are being circulated in the cybercriminal world because it is a successful predecessor of the other cryptocurrencies. Ethereum, the second-largest cryptocurrency, is a decentralized, Turing-complete computing platform [2]. The flexibility of Ethereum that is enabled through the use of smart contracts attracts many users, developers, and investors, and its price reached $138 billion in 2018.

Although Ethereum is commonly referred to as a competitor to Bitcoin, around $76 billion of illegal activity per year is financed through payments in Bitcoin which take 46% of entire illicit transactions [11].

Bitcoin is a trend in darknet markets [5] and tumbler services [12] for criminal users. Likewise, it is largely used for abuse which is the most frequent method of cybercrimes on the Internet. For example, countless Bitcoin-demanding ransomwares have been deployed every second [13]. Furthermore, recent popular methods of cryptocurrency related crimes involve blackmails and scams involving sextortion [14], [15].

2.3 Bitcoin Transaction

Cryptocurrency users can establish transactions with the other parties using only the Bitcoin addresses. For example, a sender makes a new transaction with a receiver. The sender uses one or more Bitcoin addresses as inputs. The sender also includes the amount to remit and designates the receiver address as an output of the transaction. The sender signs it with his or her private key and then broadcasts it to the Bitcoin network, which prevents the transaction from being altered. In order to get rewards, miners collect broadcasting transactions, embed them into well-defined data structures called blocks by solving hashing puzzles involving blocks. A resolved block is attached to the blockchain, which is a hash chain of blocks that maintains all blocks, then all embedded Bitcoin transactions are created and verified in the Bitcoin network in turn. It could guarantee the reliability of the transactions without a central authority in the Bitcoin network because the transactions are publicly maintained through the blockchain.

Blockchain analysis tools (e.g., BlockSci [16]) can be utilized to parse and analyze the transaction data distributed in the Bitcoin network. We can check transactions between a specific sender and receiver through their addresses even though transactions with Bitcoin are presumed to be established anonymously, interconnecting addresses is possible due to the history of the blockchain. Therefore, we are able to identify the connectivity of crimes by clustering addresses that participate in the same transaction or are identified as being owned if the sender has multiple addresses.

2.4 Related Work

Sesha Kethineni et al. [5] conducted a study on how Bitcoin has coincided with cybercrime interests and how it is illegally used. Researches about Bitcoin adopted in cybercrimes have been developed in two directions. The first is identifying anomalous transactions in the Bitcoin network [6], [7], but these studies investigate the overall transaction trend using machine learning techniques rather than studying specific transactions. The second is measuring transactions of Bitcoin abusing users who actively participate in darknet markets or forums [11]. These studies also focus on tracking mixed Bitcoin and tainting [17]–[19], as it is essential to identify threat actors that share understanding through Bitcoin transactions. These methods are applied to investigate the characteristics of illicit addresses and explore related cybercrimes (e.g., ransomware [13], [20], sextortion spam [14], Ponzi [21], dark web trade [22], gambling [23]).

However, they focus on specific cases, and few works analyze the overall aspects of cryptocurrency abuse with a narrow dataset (e.g., data collected for a distinct purpose or a specific vendor), which limits the analysis of encompassing a variety of abuse instances comprehensively. Hence, a comprehensive measurement of cryptocurrency abuse is imperative to thoroughly understand the entire cybercrime behavior and ecosystem. The public data, which correspond to the research demands but is rarely addressed by analysis, inspired us.

3. Methodology

3.1 Data Collection

We adopt *bitcoinabuse.com* [24], a popular public Bitcoin abuse database that consists of a large number of abuse reports with various kinds of informative attributes. A report is categorized into six types: *ransomware, darknet market, bitcoin tumbler, blackmail scam, sextortion,* and *other* where the *other* type can be filled out freely on side blank. Attributes of a single sample of the report database are listed in Table 1; *address*, *abuse_type_other*, *abuser*, *description* are written directly by the reporter and *abuse_type_id* is selected without guidelines. We obtain 200,597 reports from May 16, 2017, to December 31, 2020, using complete download API. In addition, we contact the service to acquire 1,338 user IDs which are unique numbers assigned to the users who sign in. We parse the reports to obtain unique abused Bitcoin addresses and reporters' IP addresses and summarize the collection in Table 2.

Table 1 Attributes of an abuse report.

Attribute	Description
id	Identifier of the report
address	Abused Bitcoin address
abuse_type_id	Specific abuse type
abuse_type_other	Describing abuse type
abuser	Describing the abuser
description	Explanation in details about the abuse
user_id	ID for registered reporter
from_ip	IP address of the reporter
from_country	Country of the reporter
from_country_code	Country code of the reporter
created_at	The time when the report was created

Also Digital Currency exchanges and some related sites (e.g., Bitcoin Who's Who, CryptoScamDB, ScamAlert and Bitcoin.org) are collecting reports on abused accounts themselves. However, these sites collect reports on their own and provide relevant information passively. In other words, related reports can be obtained by querying a suspicious bitcoin address. Although some APIs are provided, this also returns only the query address result but does not offer the entire collected data. BitcoinAbuse, however, is one of the largest services based on a general and public purpose of the reports. BItcoinAbuse service has received attention in research as a trustful source of bitcoin abuse information [15].

3.2 Data Assessment

Our first process is to comprehensively investigate the statistics of our dataset with respect to the reports and reporters. We analyze fluctuations and biases in the volume of the reports over time and correlations of the results with real-world events exacerbating cryptocurrency crimes. We can only obtain the metadata of IP addresses recorded when the reports are created; thus, we cannot confirm the private IP hidden behind the global IP. Therefore, we consider one global IP address as a reporter identity who can be multiple individuals or a group who shares the same global IP address. Then, we study the geographical information from which the reports are submitted. GeoIPLookup [25] is utilized to find the specific locations of the IP addresses recorded in the reports. It is concerned that the locations might be intentionally manipulated by the users. We try to figure out IP addresses of VPN by utilizing a collection of common VPN lists [26]. However, only 2% of IP addresses in our dataset match the lists; hence, we conclude that the influence of this kind of IP addresses on our analysis is negligible. We analyze changes in the distribution of country over the years such as countries with significant degrees of report count.

A single reporter can publish an abnormally large number of reports. We refer to these users as "heavy reporters" and define them in terms of user ID and the IP addresses that they use. If a user ID exists in a report data, we con-

Table 2 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)	

sider the unique user ID as a single reporter; otherwise, the IP address itself is considered as a reporter. Although we cannot consider a user who use multiple IP addresses, we assume that it is not an obstacle to discover heavy reporters according to few IP addresses of VPN in our dataset. In the case of IPv6 addresses, the leftmost 64 bits represent a network and the remained 64 bits are for a device based on the MAC address. In this regard, we consider the leftmost 64-bit prefix as a single reporter [27]. To identify heavy reporters, we first construct two seed lists of user IDs and IP addresses consisting reporters who submit more than 300 reports. Then, we extend the seed IP list by adding the IP addresses of reporters whose user IDs are in the seed user ID list. In the same manner, we extend the seed user ID list as well. Consequently, we can obtain the complete user ID list and IP list, additionally including reports submitted by heavy reporters without user ID (i.e., without logging in).

3.3 Abusing Method Exploration

It is worth noting that reporters can select six abuse types when reporting their abused cases. For better understanding of several kinds of the topics derived from each abuse type, we conduct a text analysis through extracting meaningful keywords. Specifically, we use Term frequency-inverse document frequency (TF-IDF), which is a numerical statistic that is intended to reflect how important a word is to a document in the reports. Based on the TF-IDF, we obtain several candidate abuse types from informative keywords. We, then, select appropriate words which can be representative and independent of the other types (e.g., "malware" in *ransomware* type, "porn" in *sextortion* type, and "mix" in *bitcoin tumbler* type). In addition, we investigate the occurrences of type names and the keywords of each type in the reports to ensure whether a given type shares its topics with other types. Lastly, except for the existing six types, we regard any undefined and newly emerged types as *other* and extract the most frequently appearing unigrams and bigrams (i.e., a sequence of two adjacent words) from user inputs, which are specified in the *abuse_type_other* attribute. These results are described in Sect. 4.2.1 with more details.

The *abuser* attribute in Table 1 is used to disclose a threat actor who performs a malicious action using a Bitcoin address. Nearly half of the reports (47.68%) contain an email address which indicates the abuser. This finding is in line with the fact that an email is one of the most common methods to deliver a blackmail message or communicate with targeted victims. However, the email indicator cannot be directly related to a cryptocurrency scam because the threat actor often provides the Bitcoin address or utilizes QR code to make victims transfer the cryptocurrency [28]; thus, we try to identify evidence using several different channels (e.g., social media) using text analysis. Specifically, we focus on external communication channels (e.g., YouTube links, Instagram profiles, etc.) rather than identifying the *abuser* attribute. To this end, we randomly sample 5% of the reports and search for clues of social media platforms in a heuristic manner. Consequently, the clues of the social media are converted into regular representations so that we can extract abuse channels from the text data.

3.4 Transaction Analysis

We extract 57,935 unique Bitcoin addresses from the *bitcoinabuse.com* database. To estimate the illicit earnings of Bitcoin abusers, we filter out Bitcoin addresses that have no transactions. We then utilize BlockSci [16] to Bitcoin addresses within Dec 31, 2020, and trace the transactions and daily balances of each Bitcoin address. The peak balance is defined as the highest amount among the daily balances of each Bitcoin address. Next, we analyze the transaction count and time diff (i.e., time difference from the first transaction day to the last transaction day) of the Bitcoin addresses to study the distribution of Bitcoin addresses through transaction count and time diff (CDF) graphs, which we assume to be successful in the attack.

We evaluate the scale of Bitcoin amounts that flowed into cybercriminal proceeds with the amount of withdrawal, which is the difference between the peak balance and the current balance. We used the withdrawal ratio, which is a withdrawal amount compared to the peak balance. We also estimate the scale of forgotten wallets with no transactions for more than three months (after Sept 30, 2020).

Furthermore, we consider clustering of method with transaction history to find out the relationship between reported abuse addresses. We observe that some Bitcoins have been transferred from a reported abuse address to another reported address. This can imply that there may have been some manipulation by a malicious actor. Consequently, we cluster the linked reported addresses and study three representative clusters.

4. Results

4.1 Report and Reporter Statistics

In this section, we present an overall evaluation of the abuse reports in our dataset. We study the report counts and the geographical distribution of reports over time, reflecting the impact of real-world events in each country. Furthermore, we examine several heavy reporters who actively submit abused Bitcoin addresses.

4.1.1 Counts over Time

We present the number of each type of reports from June 2017 to December 2020 as shown in Fig. 2. In contrast to the small number of reports in 2017, Bitcoin abuses have been actively reported since 2018 and spread worldwide in 2020. Exceptionally, all types of abuse reports are proliferated in April 2020; in total, the number of reports is increased by 13 times than that of the previous month and occupies about a quarter of the entire dataset (24.1%).

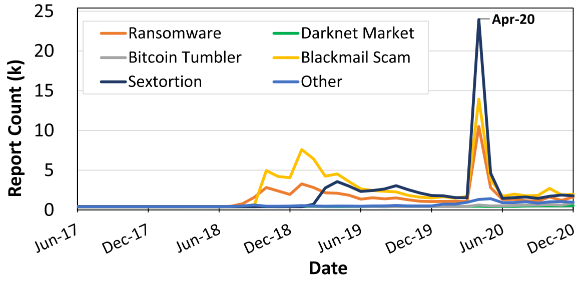


Fig. 2 Monthly report counts of each type.

A surge in Bitcoin abuses during this period is along with the claims in previous researches, which outline an increasing number of malicious activities extorting cryptocurrency (e.g., a Bitcoin scam) after the advent of the COVID-19 pandemic, since March 2020 [15], [29].

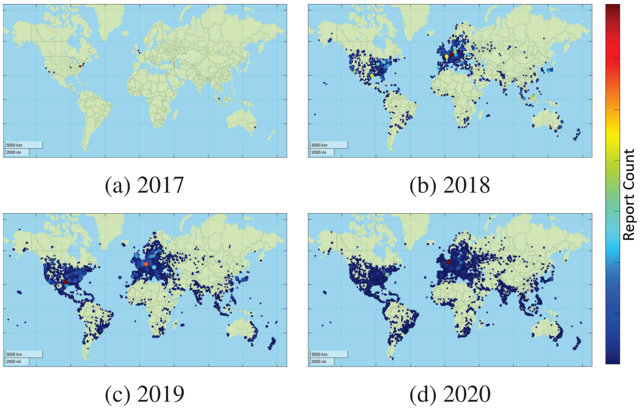
The emergence of COVID-19, which the world has faced unprecedentedly, led to rapid social aspects change and collapse balance, including the cyber security domain. Then it temporarily would let attackers judge as a good opportunity of attack. Due to the early social anxiety caused by the outbreak of the COVID-19 pandemic and rapid changes in the working environment and IT security environment such as shutdown and telecommuting of each institution, attackers also temporarily attempted a large number of attacks under the crack. One of the examples is the emergence of new abuses exploiting fears and confusions of the public, such as impersonating the World Health Organization (WHO) to ask for a fake donation [30]. Specifically in our data, almost half of the reports in April 2020 (48.7%) are about sextortion abuses. As an increasing usage of video conferencing applications (e.g., ZOOM), the abusers would take advantage of webcams to expose user privacy [31].

The influence of the COVID-19 on Bitcoin abuses has declined since May 2020 to some degree, although substantial numbers of abuses have been consistently reported. We speculate that it is due to The FBI and other security companies being immediately aware of the rapid increase of attack and the efforts of various security companies and organizations who struggle to mitigate the problematically growing threats of cryptocurrency abuses [29], [32].

4.1.2 Country Distribution

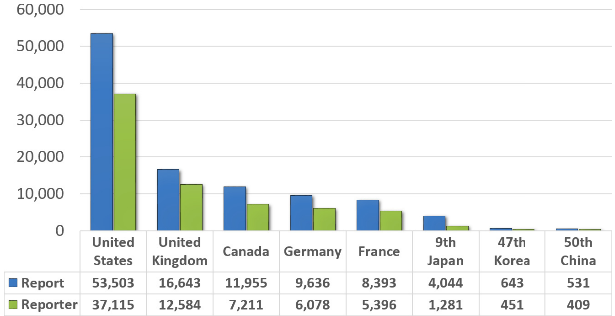
The geographical distribution of the reports in each year from 2018 to 2020 is shown in Fig. 3. Although a predominant number of reports are published from the United States (26.7%), Canada (6.0%), and the European nations, e.g., the United Kingdom (8.3%), Germany (4.8%), France (4.2%) the abuses spread worldwide (up to 214 countries) as time progressed, especially in North America and Europe.

Figure 4 represents the number of reports and reporters from the top five countries and three Asian countries. In the top 10, Japan (2%) is the only country not in North America and Europe. The majority of the reports from Japan are submitted by heavy reporters; the details are described in Sect. 4.1.3. Regarding the ratio of reporters to population, Japan is 10 to 17 times less than the top 3 countries. Korea, which has high-developed ICT infrastructures and is interested in cryptocurrency, have lower rank than expected. It is doubtful that non-english-speaking countries, especially Asian countries, are less accessible than the countries in North America and Europe. Specifically, China is ranked 50th; it is suspected that the control of cryptocurrency significantly affect to the scale of abuses [38].



(a) 2017
(b) 2018
(c) 2019
(d) 2020

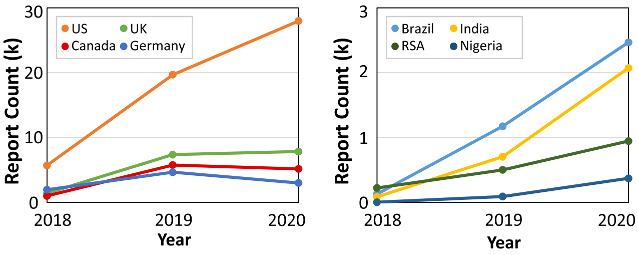
Fig. 3 Geographical distribution of reports over the years. The colors represent relative number of the reports in a given year.



	United States	United Kingdom	Canada	Germany	France	9th Japan	47th Korea	50th China
🟦 Report	53,503	16,643	11,955	9,636	8,393	4,044	643	531
🟩 Reporter	37,115	12,584	7,211	6,078	5,396	1,281	451	409

Fig. 4 Report and reporter of top 5 and Asia 3 country

Following the global proliferation of Bitcoin abuse, the impact on South America, Asia, and Africa notably increased over two years, even more drastically than North America and Europe (see Fig. 5). For example, we observe that the number of reports from Brazil, a country in South America, grows approximately 20 times. This growth is probably due to the fact that many Brazilians suffering from the high inflation of national currency started to use Bitcoin as cash in April 2019 [33]. Moreover, in Asia, abuse reports sharply increased in 2020 compared to 2019. In particular, under the advent of the COVID-19 pandemic era in March 2020, numerous data breach threats forced victims to transfer Bitcoin to protect their privacy in India [34]. The Republic of South Africa (RSA) and Nigeria show a remarkable increment of reports in 2020 as tech-savvy young people adapt to Bitcoin, undermining the value of the de facto currency [35], as with Brazil.



(a) (b)

Fig. 5 Report counts of given countries over the years. (a) shows the results of the top 4 countries ranked by the report counts and (b) shows the results of ones where the largest amount of reports are published in each place.

Table 3 Heavy reporters with user IDs and IP addresses (and the number of IP addresses in the network if more than one).

User ID	IP address (# of addresses)	# of reports	Country
N/A	176.199.5*.1*0	528	Germany
A	99 of IP, 63 of A class	379	31 Countries
B	2a00:1178:1:6*:*::*	707	Netherlands
	78.140.1*1.2*	2	
	78.106.*.6*	1	Russia
C	2600:3c00:*:*::	648	US
	24.49.3*.0	31	
D	146.115.2*2.1*3	326	US
	68.96.1*6.7*	5	
E	2405:6581:5**0:3**0:: (62)	483	Japan
	104.238.6*.0 / 24 (5)	20	
	89.187.1*1.1*7	10	
	240f:39:e**1:*::	4	
N/A same IP as E	2405:6581:5**0:3**0:: (187)	1,519	
	240f:39:e**1:*:: (5)	10	
	104.238.6*.0 / 24 (2)	2	

Table 4 Report counts and Bitcoin addresses by abuse type

Abuse type	# of reports (% of total)	Unique addr.*	Report/addr.
Blackmail scam	80,088 (**39.9%**)	26,841	2.98
Sextortion	62,644 (**31.2%**)	24,157	2.59
Ransomware	45,203 (**22.5%**)	17,310	2.61
Other	8,858 (**4.42%**)	6,162	1.44
Bitcoin tumbler	2,674 (**1.33%**)	1,826	1.46
Darknet market	1,130 (**0.56%**)	858	1.32
Total	200,597	57,935	2.07

* An address can be duplicated over the types.

4.1.3 Heavy Reporters

We identify five heavy reporters by user IDs, all of which are in the seed ID list, and one unregistered heavy reporter with an IP address of 176.199.57.150 (see Table 3). We were not able to find heavy reporters that use a single IP address with two or more user IDs.

While the majority of the reports are published from the United States as shown in Sect. 4.1.2, the heavy reporters are from various countries. A significant portion of them are located in Japan (2,048) or Netherlands (709). User "A" uses 99 different IP addresses located in various countries (e.g., Thailand, Indonesia, US, etc.), each of which are distant from each other. In this regard, this user seems to utilize multiple IP addresses on purpose (e.g., VPN).

Interestingly, user "E" submits 483 abuse reports from "2405:6581:5**0:3**0::" with 62 different devices (i.e., the same network prefix with 62 different rightmost 64 bits for different devices). Meanwhile, four IP addresses are owned by this user, and many reports (1,531) are from three of them without signs of logging in. We speculate that this heavy reporter is not an individual but a group or an organization (e.g., a security company) that submits a substantial number of reports using multiple devices.

The aggregated volume of abuse reports written by the heavy reporters (4,675) is 2.33% of the total reports. This implies that there is no heavy reporter that intentionally biases the distribution of our dataset.

4.2 Exploring Abuse Methods

This section provides analyses of various methods employed by abusers to extort Bitcoin. We study the distribution of each abuse type and its characteristics. Moreover, by using text analysis, we discover wide coverage of blackmail scam across various types. We also investigate reports in the *other* type by exploring the *abuse_type_other* attribute. Lastly, we discover various channels used for the abuse campaigns and analyze their aspects over time.

4.2.1 Abuse Types

The number of reports and Bitcoin addresses that belong to each abuse type is shown in Table 4. The most common type is *blackmail scam*, which is reported 80,088 (39.9%) times. It is then followed by *sextortion* type, reported 62,644 (31.2%) times. In these two major types, most reporters copied a body of the message from the coercive email and attached it to the description field. 45,203 (22.5%) reports have *ransomware* type and also include large amounts of texts from emails. While some are ransom notes, it is hard to figure out the source of each note (from email, compromised device, or others). While these top 3 types account for 95% of the total report, *Darknet market* and *bitcoin tumbler* types combined make up less than 2% of the total. Reports of these types are about disclosure of darknet market itself, scam in the market, or fake tumbler service, respectively. However, most of them are related to the *blackmail scam* type because they involve coercion and falsehood. In this regard, we consider that *blackmail scam* type can cover a wide range of abuse due to the broad meaning of "blackmail".

Even though the overall trend of the Bitcoin address count coincides with the volume of the reports in each type, the ratio of the report count to the Bitcoin address count

Table 5 Occurrence of each type-related-keyword in the description by type.

	Abuse Type					
Keyword	Ransom-ware	Darknet market	Bitcoin tumbler	Blackmail scam	Sex-tortion	Other
ransom	3,571	10	80	1,667	1,008	93
malware*	5,118	20	122	9,175	8,890	162
darknet	788	92	10	1,415	301	31
market	79	83	59	158	66	110
tumbler	7	0	26	8	10	5
mix*	30	1	35	67	25	12
blackmail	2,076	24	50	8,176	1,806	254
scam	3,042	232	439	9,337	3,949	4,539
sextortion	553	5	7	936	3,540	69
porn*	5,500	23	87	11,356	10,685	218

* Additional keywords to name of the type.

Table 6 Unigrams and bigrams in abuse types written by the reporters and their occurrences.

Unigram / Count		Bigram / Count	
scam	1,462	giveaway scam	1,255
giveaway	1,084	trading scam	958
trading	922	trust trading	921
trust	624	investment scam	315
bitcoin	451	youtube scam	135

Table 7 Abuse channels discovered in the text data and the number of reports where each channel is mentioned.

Channel	Ransom-ware	Darknet market	Bitcoin tumbler	Blackmail scam	Sex-tortion	Other	Total
YouTube	152	4	64	157	30	1,916	**2,323**
Telegram	77	38	98	122	7	319	**661**
Twitter	30	3	11	61	124	327	**556**
Instagram	38	11	36	48	21	181	**335**

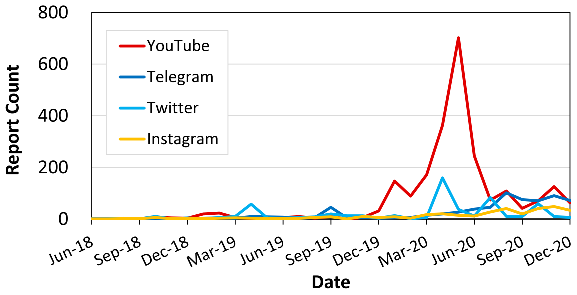


Fig. 6 The number of reports via each abuse channel over time.

is slightly different over types. The ratios of the top three major types reach 2.59 and up, being much higher than the average ratio (2.07), whereas others are less than 1.5. These results show that sexual exploitation or ransomware compromising threats are usually spread to an enormous number of victims. On the other hand, coercion using darknet market engagements and tumbling service scams tend to target a small number of individuals with a single Bitcoin address.

However, we were curious as to whether the abuse type accurately represents the abuse aspect of the report. We found some cases in which several reporters had different abuse types for the same abusing campaign with an identical text template during our review of the dataset. For that reason, the results of simple keyword matching verification on our dataset with conjoined abuse types within descriptions are shown in Table 5. Typically, each keyword appeared most frequently in blackmail scam rather than in its original type. For example, “darknet” appears about 15 times more in blackmail scam (1,415) than in darknet market (92), which is closely related to this word. In other words, a significant portion of descriptions in each abuse type shares topical meaning with blackmail scam.

The reports in the *other* type occupy less than 5% of the total. However, we can still obtain useful information from this undisclosed type as these reports are not covered by the predefined categories. We list frequent unigram and bigram keywords extracted from the attribute of the report in *other* type in Table 6. All of the unigrams appeared in the bigrams except “bitcoin”, whereas most of the top-ranked bigram keywords are related to giveaway scam (e.g., trading scam, trust trading, and investment scam), falsely promising extra returns for sending cryptocurrency. This is one of the newly emerging threats using cryptocurrency in recent days [15]. Besides, we observe topics about hack database, DDoS, cryptocurrency mining, and several words indicating abuse channels (e.g., social media) in the text.

4.2.2 Channels for Abuse Campaigns

Inspired by the evidences of abuse methods mentioned in Sect. 4.2.1, we investigate the reports containing clues of the social media platforms within the *abuser*, the *description*, and the *abuse_type_other* attributes. Table 7 shows discovered abuse channels and occurrences of the channels in the abuse reports. YouTube is the most frequently reported channel besides email, and the remaining are popular social network services (e.g., Telegram, Twitter, and Instagram).

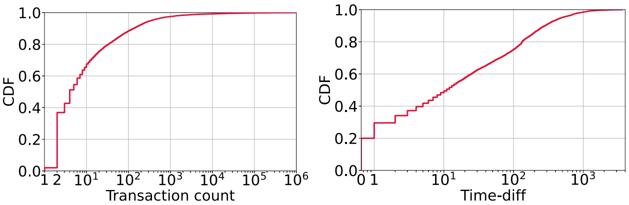
The abuse with these channels primarily belongs to the *other* type. In fact, most of them are related to giveaway scam which does not exist within the predefined abuse types, as aforementioned in Sect. 4.2.1. According to the descriptions of the reports, the abuse scenario is described as follows: (i) The abusers hack accounts of influential users (e.g., celebrities, CEOs, and politicians) on the platforms. (ii) Then, they promote a fake investment or giveaway event promising specific rewards for sending Bitcoin. (iii) They do not pay out the rewards and give no answer. Apart from that, a number of reports are significant; 18.5% of abuses via Telegram are reported as blackmail scams; 22.3% of those via Twitter are about sextortion.

The abuses through newly discovered channels were rarely reported until early 2019 but dramatically increased during the first half of 2020 (see Fig. 6). We speculate that this phenomenon is due to a surge in usage of social media with the emergence of the COVID-19 pandemic since

Table 8 Characteristics of abused Bitcoin addresses which have transaction history.

		# of addresses (% of total)	**Avg. Rpt. Cnt.**	**Avg. Rpt. Time diff**	**Avg. TX. Cnt.**	**Avg. TX.Time diff**	**Avg. Bal.** (BTC)	**Avg. Peak Bal.** (BTC)	**Avg. Withdrawal Ratio**
Total	Address	11,810	7.87	14.93	1,468.4	102.2	83.0013	167.9933	0.9520
Groups by TX Count	only 1 TX	233 (1.9%)	7.49	11.76	1	0	0.9436	0.9436	0
	only 2 TXs	4,122 (34.9%)	4.77	10.83	2	20.0	0.0204	3.0845	0.9908
	over 100 TXs	1,372 (11.6%)	1.81	20.44	12,556.2	417.4	411.8219	999.9329	0.9255
Groups by TX Time diff	only 1 day	2,355 (19.9%)	3.20	4.92	2.7	0	0.1123	2.0866	0.8929
	only 2 days	1,134 (9.6%)	4.59	6.92	4.7	1	0.0040	0.9068	0.9856
	over 100 days	2,924 (24.8%)	3.62	17.08	5,810.4	367.6	308.8685	613.9223	0.9342

∗ **Rpt.**: report, **Cnt.**: count, **Time diff**: time difference from the first transaction day to the last transaction day, **Bal.**: balance
∗ **TX Count**: transaction count, (only *n* time/times) denotes that transactions have generated *n* times
∗ **TX Time diff**: transaction time diff, (only *n* day/days) denotes that transactions have observed within *n* (time diff + 1) days



(a)

(b)

Fig. 7 CDF of (a) the transaction count and (b) transaction time diff. Both present the polarization of temporal and non-temporal abused Bitcoin addresses.

December 2019 [15]. In particular, the quantity of abuse reports via YouTube grew more than 140 times during six months (from November 2019 to May 2020). Even though the impact of abuses via social media was mitigated to some extent since the second half of 2020, these kinds of abuses still have been persistently reported.

4.3 Transaction of Abused Bitcoin Address

In this section, we analyze transactions of reported Bitcoin addresses. We delve into the characteristics by dividing them into the temporal and non-temporal groups with transaction count and time diff for each group. Then, we observe transaction flows on cybercriminal proceeds and forgotten wallets with peak balance and withdrawal ratio. Finally, we identify clusters of the addresses with transactions considering a 1-hop connection and study the clusters in detail.

4.3.1 Overview of Abused Bitcoin Transactions

Among the 57,935 abused Bitcoin addresses, we only deal with addresses which have Bitcoin transaction history. We estimate that only 11,810 addresses containing transaction records are successful in the attack. The average number of transactions generated from the addresses was 1,468, with the highest number of transactions at 3,484,897. However, the median value is 4, and the modal value is 2 in Fig. 7 (a). The average time diff of transactions for active Bitcoin addresses is 102.2, which is considerably longer than the time diff of the report (14.93). However, this finding suggests that it would be significantly polarized considering the median value, 10, in Fig. 7 (b). The address has an average balance of 83.0013 BTC, an average peak balance of 167.9933 BTC, and a withdrawal ratio of 0.9520, which is not as high as expected. Therefore, it is necessary to analyze the characteristics of each group regarding the polarized situation.

In terms of the transaction count (i.e., TX Count in Table 8), addresses with only 2 transactions account for 35% of the total, and addresses with less than 10 transactions account for 65% of the total. As for the transaction time diff (i.e., TX Time diff in Table 8), addresses with only 1 day transaction (time diff = 0) account for 20% of the total. On the other hand, addresses with a transaction time diff of more than 100 days account for 25% of the total. As a result, it can be tentatively determined that both sides are polarized into temporal and non-temporal. To analyze the polarization characteristics of the addresses identified through the CDF distribution, we divide Bitcoin addresses into groups with only 1, 2 and over 100 times in transaction count and only 1, 2 and over 100 days in transaction time diff shown as Table 8.

We identify 233 Bitcoin addresses with only 1 transaction count which assumed no withdrawal. The average balance of these addresses is 0.9436 BTC, and they are classified as forgotten wallets. The Bitcoin addresses with 2 transactions are addresses in which the paid amount is withdrawn and is no longer in use after the abuse. There are 4,122 such addresses, and 4,080 (99.0%) of them have zero balance. 25 addresses among the remaining 42 addresses whose balance is not zero after twice transactions have no transaction records for more than three months as forgotten wallets.

The group which has more than a hundred transaction count represents the characteristic of non-temporal Bitcoin addresses. It shows a large number of transaction counts and time diff. Considering the average of peak balance and withdrawal ratio, Bitcoin addresses in this group are constantly abused for malicious deployment. However, considering that the average report count of this group is remarkably small (1.81), it is likely that the group contains middle nodes used as mixing services that rarely expose Bitcoin addresses. So professional bitcoin tracking techniques are expected for more detailed analysis. In addition, as there are possible limitations in the analysis from the perspective of transaction count, we also compare and analyze the data from the perspective of the transaction time diff.

Although groups divided by the transaction count show an inversely proportional tendency in the average report count perspective, it shows a different pattern in terms of the transaction time diff. In the 1 day group, 2,355 Bitcoin addresses are discovered (including 233 of only 1 transactions described earlier), and no more transactions are observed after that day. Also, 2,073 (88.0%) addresses have zero balance, and 213 addresses among the 282 addresses do not have transaction records for more than three months. It shows a similar pattern in the only 2 days group with only 1.5 times growing trend with report and transaction perspective.

We also identified high withdrawal ratios that were not seen in the only 1, 2 times of transaction counts groups. The groups, whose transaction time diff is only 1 and 2 days, account for 30% of the addresses and represent the typical characteristics of temporal Bitcoin addresses. From the transaction time diff perspective, instead of transaction count, the over 100 days group means definite non-temporal Bitcoin addresses. There are 2,924 addresses, accounting for 25% of the entire addresses which have transactions. This group represents a lower average report count in our analysis. This finding can be a piece of decisive evidence that non-temporal addresses can be used for a long-term time diff if they are rarely exposed during the abuse.

4.3.2 Cybercriminal Proceeds of Abused Addresses

As previously mentioned, only 11,810 abused Bitcoin addresses succeeded in the attack, however, the process of realized profit to cybercriminal proceeds was needed analysis. We estimated the bitcoin amounts that flow into cybercriminal proceeds through active address peak balance and withdrawal ratio, which are assumed success in the attack. We note that the total peak balance is 2,162,083.0760 BTC and the average peak balance is 183.0722 BTC. We observe 11,382 Bitcoin addresses that have a withdrawal history. 10,342 (87.5%) of them have 1 withdrawal ratio, which means that the balance is zero since the abuser withdrew the balance completely. The total amount of the withdrawal is 1,181,837.8921 BTC, and we estimated that this amount has flowed into cybercrime proceeds, estimated to be 6 billion dollars when a bitcoin is converted to about $ 5k.

The total remaining current balance is 980,245.1838 BTC. Of these, 428 addresses remained the peak balance as the current balance that is a total of 390,282.0017 BTC in the wallet. A maximum amount of up to 94,505.8314 BTC was also found. We cannot determine the statuses of these wallets since the balance can be withdrawn in the near future. The judgment on these wallets would be deferred to a later since these could be withdrawn at any time.

Table 9 Cluster of linked Bitcoin addresses which in a transaction block

	Cluster A	Cluster B	Cluster C
# of addresses	7,164	4	2
Total Peak Bal.	2,146,162.0749	1,434.9560	498.2174
Total Withdrawal	1,166,167.8309	1,434.8142	498.2174
Total Current Bal.	979,994.2440	0.1418	0
Average TX Cnt.	2,385.2	261.5	35,151.0
Total. Rpt. Cnt.	70,750	14	2
Report Type # (%)	Blackmail Scam 31,625 (44.7%) Bitcoin Tumbler 1,685 (2.4%)	Other 12 (85%)	Other 2 (100%)

On the other hand, 496 Bitcoin addresses have no transaction history for more than three months, and the total balance is 423.6880 BTC, so that we can suppose them as forgotten wallets.

4.3.3 Clustering of Abused Bitcoin Addresses

Out of 11,810 Bitcoin addresses that have transaction history, 3,782 (32.0%) of them have no directly linked transactions between each other. The remaining Bitcoin addresses (8,028) are divided into 311 clusters. However, it leaves room for discovering new connections or combining them into larger clusters through additional transaction tracking.

We present the characteristics of the top three clusters ranked by the total withdrawal amount as shown in Table 9. We first discover a large extent cluster, "A", containing 7,164 Bitcoin addresses. 63% of bitcoin tumbler reports in our dataset belong to it. In addition, the total withdrawal for this cluster is 1,166,167.8309 BTC, and the remaining current balance is 979,994.2440 BTC. In this regard, we surely speculate that it is the center of unknown tumbler mixers.

The second cluster, "B", contains four Bitcoin addresses. However, the potential size is enormous if we include addresses that are not in our dataset as well. As for the four addresses, the peak balance (1,435.0 BTC) and the withdrawal amount (1,434.8 BTC) are relatively large, whereas the average transaction count of these addresses is only 261.5. While the average report count is 3.5, a large amount of Bitcoins are leaked as proceeds of cybercrime. Reports in this cluster mainly belong to the *other* type (85%). Besides, considering the description of the reports, the abuses are related to "Electrum 4.0 Bitcoin wallet update trick" [36].

The third cluster, "C" contains two Bitcoin addresses that have transactions with each other. Interestingly, the average transaction count of this cluster is considerably large (35,151). We observe one report for each address submitted from the same reporter. According to the description of the reports, the abuse is described as "Cloud Token Ponzi" [37]. The total peak balance is not quite huge (498.2174 BTC), but some portion of the balance has been withdrawn constantly. Specifically, the addresses received the last transaction in May 2020, then all balance was withdrawn in January 2020.

5. Disscusion

Continuous management. We provided a quantitative analysis to understand the cryptocurrency abuse cases using the public reports (see Sect. 4.2.1). While these public reports provided valuable data that could cover the entire phenomenon all over the world in the perspective of the victims, updates such as observing trends and reflecting the trends to appropriate types were necessary through continuous management and constant interaction with the public reporters.

Ambiguous type definition. While we already organized abuse types properly in Table 4, several types were ambiguous to be determined clearly (see Table 5). As such, we discovered that several types were overlapped with other types, so that the type definition was not absolute. For example, scam-related words were most frequently observed in our dataset. Thus, it could be assumed that the scam category might contain other categories as well because it had an inclusive meaning. For example, consider the case where a threat actor took victims' pictures or videos as a hostage and asked cryptocurrency. In this situation, on one hand, a victim reporter might choose the type as *blackmail scam*; on the other hand, a victim reporter might choose the type as *ransomware* by default. For this reason, it would be required of a new study for a type selection appropriateness to prevent ambiguous type choices.

Future work. For better verifying the quality of the data, we identified heavy reporters from a quantitative point of view. Besides, overlapping reports with same contents from several users were filtered by *bitcoinabuse.com* itself. Therefore, we could successfully verify the duplicated reports with same contents. However, we were still lack of metrics to determine whether the content had false information or not. Therefore, we suggest studying text data, which contains report contents, using natural language processing (NLP) in our future work. Through this, we can perform a bitcoin abuse campaign identification using text templates as a follow-up study and further improve the performance of a type selection appropriateness or analyze the contents of the reports in an automatic manner [38].

6. Conclusion

We present a large-scale measurement study of cryptocurrency abuse using 200,507 real world abuse reports. By utilizing our data, we observe that the impact of the abuses grows consistently and spreads worldwide. We investigate the factual issues to figure out the cause of the surge in cryptocurrency abuses. Besides, we demonstrate that the semantics of several types are overlapped mutually. We identify famous social media services, which are frequently used for abusive activities after the emergence of the COVID-19 pandemic. We manifest the polarization characteristics of the temporal and non-temporal Bitcoin addresses in the perspective of the transaction history. We find that a large amount of Bitcoins are leaked into the cybercriminal area and investigate influential attack groups for illicit purposes. We hope that these insights can inspire researchers and practitioners to improve the security of the cryptocurrency environment.

Acknowledgments

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-00118, Development of decentralized consensus composition technology for large-scale nodes).

References

[1] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," Decentralized Business Review, 21260, 2008.
[2] G. Wood, "Ethereum: A secure decentralised generalised transaction ledger," Ethereum project yellow paper, vol.151, no.2014, pp.1–32, 2014.
[3] P.M. Krafft, N.D. Penna, and A.S. Pentland, "An experimental study of cryptocurrency market dynamics," Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp.1–13, 2018.
[4] N. Christin, "Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace," Proceedings of the 22nd international conference on World Wide Web, pp.213–224, 2013.
[5] 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.
[6] T. Pham and S. Lee, "Anomaly detection in bitcoin network using unsupervised learning methods," arXiv preprint arXiv:1611.03941, 2016.
[7] P. Monamo, V. Marivate, and B. Twala, "Unsupervised learning for robust Bitcoin fraud detection," 2016 Information Security for South Africa (ISSA), IEEE, 2016.
[8] R. Grinberg, "Bitcoin: An innovative alternative digital currency," Hastings Science & Technology Law Journal, vol.4, p.160, 2011.
[9] D.L.K. Chuen, ed., Handbook of Digital Currency: Bitcoin, innovation, financial instruments, and big data, Academic Press, 2015.
[10] S. Ahamad, M. Nair, and B. Varghese, "A survey on crypto currencies," 4th International Conference on Advances in Computer Science, AETACS, Citeseer, 2013.
[11] S. Foley, J.R. Karlsen, and T.J. Putniņš, "Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies?," The Review of Financial Studies, vol.32, no.5, pp.1798–1853, 2019.
[12] T. de Balthasar and J. Hernandez-Castro, "An analysis of bitcoin laundry services," Nordic Conference on Secure IT Systems. Springer, Cham, 2017.
[13] M. Paquet-Clouston, B. Haslhofer, and B. Dupont, "Ransomware payments in the bitcoin ecosystem," Journal of Cybersecurity, vol.5, no.1, tyz003, 2019.
[14] M. Paquet-Clouston, M. Romiti, B. Haslhofer, and T. Charvat, "Spams meet cryptocurrencies: Sextortion in the bitcoin ecosystem," Proceedings of the 1st ACM conference on advances in financial technologies, pp.76–88, 2019.
[15] P. Xia, et al., "Don't fish in troubled waters! characterizing coronavirus-themed cryptocurrency scams," arXiv preprint arXiv:2007.13639, 2020.
[16] H. Kalodner, et al., "Blocksci: Design and applications of a blockchain analysis platform," 29th USENIX Security Symposium

(USENIX Security 20), 2020.

[17] T. Tironsakkul, M. Maarek, A. Eross, and M. Just, "Tracking mixed bitcoins," Data Privacy Management, Cryptocurrencies and Blockchain Technology. Springer, Cham, pp.447–457, 2020.

[18] Y. Zhang, J. Wang, and F. Zhao, "Transaction Community Identification in Bitcoin," 2020 13th International Symposium on Computational Intelligence and Design (ISCID), IEEE, 2020.

[19] Y. Zhang, J. Wang, and J. Luo, "Heuristic-Based Address Clustering in Bitcoin," IEEE Access, vol.8, pp.210582–210591, 2020.

[20] D.Y. Huang, M.M. Aliapoulios, V.G. Li, L. Invernizzi, E. Bursztein, K. McRoberts, J. Levin, K. Levchenko, A.C. Snoeren, and D. McCoy, "Tracking ransomware end-to-end," 2018 IEEE Symposium on Security and Privacy (SP), IEEE, 2018.

[21] M. Bartoletti, B. Pes, and S. Serusi, "Data mining for detecting bitcoin ponzi schemes," 2018 Crypto Valley Conference on Blockchain Technology (CVCBT), IEEE, 2018.

[22] 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.

[23] S. Choi, K.-S. Choi, Y. Sungu-Eryilmaz, and H.-K. Park, "Illegal Gambling and Its Operation via the Darknet and Bitcoin: An Application of Routine Activity Theory," International Journal of Cybersecurity Intelligence & Cybercrime, vol.3, no.1, pp.3–23, 2020.

[24] T.B. Team, Bitcoin abuse database, accessed 2020-10-30, available: https://www.bitcoinabuse.com/, 2020.

[25] Octolus and Zigi, GeoIPLookup, accessed 2021-05-04, available: https://geoiplookup.io/, 2020.

[26] IPQualityScore, accessed 2021-01-12, available: https://www.ipqualityscore.com/free-ip-lookup-proxy-vpn-test

[27] J. Davies, Understanding IPv6: Understanding IPv6 _p3, Pearson Education, 2012.

[28] R. Phillips and H. Wilder, "Tracing cryptocurrency scams: Clustering replicated advance-fee and phishing websites," 2020 IEEE International Conference on Blockchain and Cryptocurrency (ICBC), IEEE, 2020.

[29] H.S. Lallie, L.A. Shepherd, J.R.C. Nurse, A. Erola, G. Epiphaniou, C. Maple, and X. Bellekens, "Cyber security in the age of covid-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic," Computers & Security, vol.105, 102248, 2021.

[30] H. Partz, Scammers impersonate world health organization to steal btc covid-19 donations, accessed 2021-05-11, available: https://cointelegraph.com/news/scammers-impersonate-world-health-organization-to-steal-btc-covid-19-donations, 2020.

[31] Y. Sapkale, Sextortion scams surging during COVID-19 pandemic and work from home environment, accessed 2021-05-11, available: https://www.moneylife.in/article/sextortion-scams-surging-during-covid-19-pandemic-and-work-from-home-environment/62993.html, 2021.

[32] FBI.gov., FBI expects a rise in scams involving cryptocurrency related to the COVID-19 pandemic, accessed 2021-12-11, available: https://www.fbi.gov/news/pressrel/press-releases/fbi-expects-a-rise-in-scams-involving-cryptocurrency-related-to-the-covid-19-pandemic, 2020.

[33] J. Gil-Pulgar, Brazil: Highest inflation in 4 year propels bitcoin volume to record highs, accessed 2021-05-13, available: https://bitcoinist.com/brazil-highest-inflation-in-4-years-amid-record-bitcoin-trading-volume/, 2019.

[34] S. Shinde and N. Alawadhi, India becomes favourite destination for cyber criminals amid covid-19, accessed 2021-05-13, available: https://www.business-standard.com/article/technology/india-becomes-favourite-destination-for-cyber-criminals-amid-covid-19-121040501218_1.html, 2021.

[35] A. Akwagyiram and T. Wilson, How bitcoin met the real world in Africa, accessed 2021-05-13, available: https://www.reuters.com/article/us-crypto-currencies-africa-insight-idUSKBN25Z0Q8, 2020.

[36] C. Cimpanu, Bitcoin wallet update trick has netted criminals more than $22 million, accessed 2021-05-15, available: https://www.zdnet.com/article/bitcoin-wallet-trick-has-netted-criminals-more-than-22-million/, 2020.

[37] Googlesite, Cloud token wallet: Decentralized wallet-is cloud token Ponzi scheme or scam?, accessed 2021-05-15, available: https://sites.google.com/view/cloud-tokenwallet/blogs/is-cloud-token-ponzi-scheme-or-scam, 2019.

[38] J. Choi, T. Lee, K. Kim, M. Seo, J. Cui, and S. Shin, "Discovering message templates on large scale Bitcoin abuse reports using a two-fold NLP-based clustering method," IEICE Trans. Inf. & Syst., vol.E105-D, no.4, pp.824–827, April 2022. DOI:10.1587/transinf.2021EDL8092

Jinho Choi co-first author, is a Ph.D. student in the Graduate School of Information Security at KAIST. He received his M.S. degree from the Computer Engineering Department at Texas A&M university and B.S. degree from Computer Science Department at Korea Military Academy. His research interests mainly center on the measurement of cyber abusing with Open Source and Cyber Threat Intelligence.

Jaehan Kim co-first author, is a master student in the School of Electrical Engineering at KAIST. He received his B.S. degree in School of Electrical Engineering at KAIST. His research interests mainly focus on Data mining, Natural language processing, and Cyber threat intelligence.

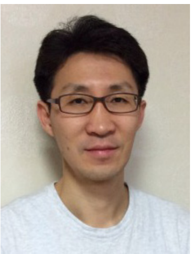
Minkyoo Song is a master student the School of Electrical Engineering at KAIST. He received his B.S. degree in Industrial & System Engineering Department and School of Electrical Engineering at KAIST. His research interests mainly focus on Big data and Cyber threat intelligence.

Hanna Kim is a master student in the School of Electrical Engineering at KAIST. She received his B.S. degree in School of Electrical Engineering and Mechanical Engineering at KAIST. Her research interests mainly focus on Data mining, Differential privacy, Deep learning and Cyber threat intelligence.

Nahyeon Park is a master student in the School of Electrical Engineering at KAIST. She received his B.S. degree in School of Electrical Engineering at KAIST. Her research interests mainly focus on Natural language processing, Image security, and Cyber threat intelligence.

Minjae Seo is a master student in the Graduate School of Information Security at KAIST. He received his B.S. degree in Computer Engineering from Mississippi State University, in 2019. His current research interests include Software-defined networking security, Network fingerprinting, and Deep learning-based network system.

Youngjin Jin is a master student in the School of Electrical Engineering at KAIST. He received his B.S. degree in School of Electrical Engineering at KAIST. His research interests mainly focus on Data mining, Natural language processing, and Cyber threat intelligence.



Seungwon Shin is an associate professor in the School of Electrical Engineering at KAIST. He received his Ph.D. degree in Computer Engineering from the Electrical and Computer Engineering Department, Texas A&M University, and his M.S. degree and B.S. degree from KAIST, both in Electrical and Computer Engineering. His research interests span the areas of Software-defined networking security, DarkWeb analysis, and Cyber threat intelligence.