Tension between GDPR and Public Blockchains: A Data-Driven Analysis of Online Discussions

Zeynep Chousein
zozkalkan@gmail.com
Ankara Yıldırım Beyazıt Üniversitesi
Ankara, Turkey

Hacı Yakup Tetik
haciyakuptetik@gmail.com
Ankara Yıldırım Beyazıt Üniversitesi
Ankara, Turkey

Rahime Belen Sağlam
r.belen-saglam-724@kent.ac.uk
University of Kent
Canterbury, UK

Abdullah Bülbül
mabdullahbulbul@gmail.com
Ankara Yıldırım Beyazıt Üniversitesi
Ankara, Turkey

Shujun Li
s.j.li@kent.ac.uk
University of Kent
Canterbury, UK

ABSTRACT

Since coming into effect in May 2018, the EU General Data Protection Regulation (GDPR) has raised serious concerns among users of public (permissionless) blockchain systems. Such concerns are triggered by a tension between some unique characteristics of public blockchain systems and some new data subject rights introduced in the GDPR, e.g., the data immutability and the "right to erasure" (a.k.a. "the right to be forgotten"). The aim of this work is to understand how service providers and developers behind public blockchain systems have communicated about such GDPR-related challenges to their users and how the users have perceived such GDPR-related issues. To this end, for 50 public blockchain systems whose corresponding cryptocurrency had a capital market size over $150 million, we analyzed relevant communications and discussions on the following three online channels: blog and forums posts, GitHub repositories, and discussions on Twitter. Our results show that service providers and developers of the selected public blockchain systems did not play an active role in GDPR-related online discussions on Twitter. They also did not communicate with their users about GDPR on their forums and blogs frequently, where we could identify only 56 posts out of 17,821 posts for the period we studied. Our study also reveals that only an extreme minority of the studied systems (4) mentioned GDPR in their GitHub repositories. Our work adds new evidence on the lack of transparency and active communications of the public blockchain sector on the challenging GDPR compliance issue of public blockchain systems.

KEYWORDS

blockchain, distributed ledger, permissionless, GDPR, data protection, privacy, law, Twitter, GitHub, online forum, blog, machine learning, NLP, natural language processing, classification, topic modeling

SINCONF'20, November 4–7, 2020, Istanbul, Turkey

ACM ISBN 978-1-4503-8751-4…$15.00
https://doi.org/10.1145/3433174.3433587

ACM Reference Format:
Zeynep Chousein, Hacı Yakup Tetik, Rahime Belen Sağlam, Abdullah Bülbül, and Shujun Li. 2020. Tension between GDPR and Public Blockchains: A Data-Driven Analysis of Online Discussions. In *Proceedings of 13th International Conference on Security of Information and Networks (SINCONF'20).* ACM, New York, NY, USA, Article 17, 8 pages. https://doi.org/10.1145/3433174.3433587

1 INTRODUCTION

Since the first public blockchain system (and a cryptocurrency) Bitcoin [26], its underlying technique, blockchain, has attracted lots of attention from researchers, and practitioners and users. It has become one of the most recent emerging technologies in many application areas such as the FinTech (financial technology) industry, and many new blockchain-based systems and cryptocurrencies have been created especially in the past five years.

Technically, a blockchain is a decentralized peer-to-peer (P2P) network where each node keeps a full copy of a distributed database called a distributed ledger for storing transaction data. Because the distributed ledger is at the core of a blockchain, the technology is also called distributed ledger technology (DLT). Data (e.g., financial transactions) on a distributed ledger are stored following a distributed consensus protocol (such as proof of work or proof of stakes) among participating nodes called miners[1] who compete or collaborate to append data to the existing ledger (which is called mining). The data is appended in such a way that the whole distributed ledger forms an increasingly long chain and historical records cannot be amended in a normal setting. This leads to the unique (and desired) attribute of blockchains: data immutability, i.e., data on a blockchain can be permanently stored there without the worry of their being manipulated.[2] Depending on who can access the data, blockchain systems can be classified into three types: *public* (*permissionless*), consortium (permissioned), and private. Public blockchain systems allow any (normally pseudo-anonymous) node to read and write to a blockchain without the need to seek a permission from anyone. In contrast, in a permissioned blockchain system, a membership service (which can be existing members collectively) controls which nodes to participate in the system and what rights

[1] In some systems miners are called forgers because the data creation process is less about block/coin mining. In this paper we will use the term "miner" in a broader sense.
[2] Editable or redactable blockchains have actually been studied [18], but have not been widely accepted or deployed, so we will not consider this in our paper.

they have (writing to the ledger and/or validating the transactions). A private blockchain system is simply a distributed ledge used by a single party (e.g., a large organisation). It has been known that a tension with the GDPR exists more for public blockchains [22], so for the rest of the paper we will focus on public blockchains only.

Public blockchains can be used to support many practical applications. Among all its applications, virtual (or crypto) cryptocurrencies is probably the first and the most popular one. Virtual cryptocurrencies are digital assets enabled by a blockchain system, and they can be exchanged between different parties without going through a centralized financial institution such as a bank [19]. Although a public blockchain system does not have to be equipped with a cryptocurrency, most existing systems do have one or more cryptocurrencies incorporated in them. Since BTC (the cryptocurrency behind Bitcoin), thousands of such cryptocurrencies have been proposed, such as Ethereum, Ripple, Litecoin and Dash, with a total market value of more than 250 billion US dollars as of July 2020 according to CoinMarketCap[3], a cryptocurrency market information website.

The European Union (EU)'s General Data Protection Regulation (GDPR) was drafted in 2016 and became effective in the whole EU in May 2018 [21]. Although being a EU legislation, it aims to protect privacy of any data subjects in the EU (not just EU citizens), and any personal data that are collected or processed in the EU even if the data subjects are not in the EU (Article 3). This broad territorial scope makes GDPR compliance a global concern due to the increase of the globalization of information-centric businesses and public services. The GDPR defines data protection principles and lawful bases for collecting and processing personal data, and also specifies a number of data protection rights of data subjects (individuals), such as the right to erasure (to be forgotten), the right to rectification, and the right to access to personal data. Moreover, the GDPR imposes a number of obligations on data controllers and data processors (i.e., organizations collecting and processing personal data). For instance, they are required to get explicit consents from users if the lawful basis for collecting and processing data is based on consent, to keep records of data processing activities, and to inform data subjects about automatic decisions based on such processing. An infringement of such obligations can be subject to a fine of up to 20 million or 4% of worldwide turnover.

With the aim of offering some common understanding and future development directions on how to tackle the challenges around GDPR compliance of blockchain systems, the EU Blockchain Observatory & Forum published a report in 2018 [22], which clearly identified a tension between the GDPR and blockchain systems especially for public blockchain systems. One of the biggest challenges was given as the right to erasure. The data immutability of public blockchain systems makes it technically difficult or impossible to delete or correct their personal data once stored on blockchain. Another concern was given regarding the identification of data controllers and data processors. Due to the (pseudo)anonymity mechanism and the distributed nature of public blockchain systems, it is hard to identify roles, and thus, to assign responsibilities. In addition, even if data controllers and data processors can be identified, it is also difficult to gather explicit consents from users due to the (pseudo)anonymity mechanism and the territorial scope rules of the GDPR.

Although it has been more than 2 years since the GDPR came into effect, there are still very limited research about how (public) blockchain developers and service providers perceive the GDPR compliance issue and how they communicate related challenges to their users for transparency purposes. The only work we are aware of is our previous research reported recently [27], in which we focused on how public blockchain developers and service providers communicated the GDPR compliance issue to their users by examining two online communication channels: 1) legal documents including privacy policies, T&C (Terms and Conditions) documents and other similar legal documents published on systems' official websites; and 2) public tweets of their official Twitter accounts. Their study revealed a systematic lack of transparent and detailed communications by public blockchain developers and service providers to their users, and discovered questionable statements about GDPR compliance in their public communications.

In this study, we extend our work in [27] by examining three other public communication channels not covered before: 1) source code repositories (Github), 2) blogs and web forums, and 3) public tweets of a much more diverse group of Twitter accounts, including not just official accounts of public blockchain systems but also any Twitter accounts that mentioned at least one of such public blockchain systems we selected to study. For the third channel, we automated the analysis of tweets by using machine learning based classification and NLP (natural language processing) based topic modeling techniques to extract GDPR-related discussions from a corpus with more than 11 million tweets. Analysis of the GitHub repositories, blogs and web forums of selected public blockchain systems was done more manually because it was much harder to automate the process, and the findings were checked by at least two independent human encoders. Our analysis revealed a finding similar to what we reported in [27]: only a minority of the systems communicated about the GDPR through the channels investigated in this study. It was also observed that blockchain companies did not play an active role in blockchain- and GDPR-related discussions on Twitter. In addition, they rarely used their official forums or blogs to communicate with their users about the GDPR. Finally, we also observed a lack of systematic details about the GDPR compliance on GitHub repositories, where we had expected to identify some explanations in end user licence agreements (EULAs) of blockchain software.

The rest of the paper is organized as follows. In the next section, the challenges between the GDPR and the blockchain technology are summarized. Section 3 explains details about how we collected and processed the data we used for the study. Finally, our main findings are reported in Section 4. The paper is concluded by the last section, with limitations of the work and future research briefed.

2 BACKGROUND AND RELATED WORK

The GDPR specifies data protection rights for individuals and also defines principles and the lawful bases for processing personal data. In this section, we introduce a subset of these elements that are relevant for key open issues in the GDPR-compliance issue of (public) blockchain systems.

[3]https://coinmarketcap.com/

The "data minimisation" principle requires data controllers to ensure that personal data processed is adequate, relevant and limited to what is necessary in relation to the processing purpose. The spirit of this principle is profoundly at odds with data storage on a public blockchain where data remains part of the whole chain permanently once it is added. Another data protection principle highlighted in the GDPR dictates that data controllers and processors should not store personal data for longer than they actually need it. It depends very strongly on the purpose or purposes determined for the processing. However, in the case of a public blockchain system, data are stored on the chain permanently for the technical purpose of making the system work, which may be seen contradicting with data subjects' rights.

The GDPR specifies eight data protection rights of individuals: the right to be informed, the right of access, the right to rectification, the right to erasure, the right to restrict processing, the right to data portability, the right to object, and rights in relation to automated decision making and profiling. The right of access enables individuals to get a copy of their personal data with additional information when they request. This right aims to allow individuals to understand how and why their personal data is being used. The right of erasure, also known as 'the right to be forgotten', is one of the most challenging rights for public blockchain systems. It mandates that data controllers and processors must delete data if there is no longer a lawful basis for processing or if the data subject withdraws the consent. This is one of the most-cited aspect of the tension between the GDPR and public blockchain systems since the immutable nature of public blockchains makes it impossible for data controllers and processors to delete any data.

In addition to principles and rights for individuals, the GDPR also specifies the lawful basis for collecting and processing personal data. The most common way of obtaining a lawful basis is to ask for the explicit consent from a data subject for the processing to occur for one or more specific purposes explicitly. Here, an explicit consent means freely given, specific, informed and unambiguous indication of the data subject's preferences about the processing of personal data relating to him or her. In addition, a data subject has the right to withdraw consent at any time. In a public blockchain system, it is essential to gather consents before the download or execution of the blockchain software since once a transaction has been made on the blockchain, the same set of data will be processed by the all nodes in the chain.

One of the major requirements enforced by the GDPR for personal data processing is that the underlying IT systems should follow the concept of "Data Protection by Design" (Article 25). It requires data protection to be considered as a default setting of every new IT systems and should be built into systems from the design stage. This may be interpreted that personal data must not be stored in plaintext on a public blockchain. When considering this principle in our work, we focused on explicit mentions of related terms including "privacy by design", "data protection by design", "data protection by default" and other spellings of those words (e.g., "privacy-by-design").

Even though those conflicts have been widely covered in the literature, there have been very limited past studies that focused on online communications of public blockchain systems' service providers and developers to their users on the GDPR compliance issue.

In one of the related studies [23], Gruzd et al. studied GDPR-related discussions to examine public opinions and organizational public relations (PR) strategies about the GDPR. For this purpose, they collected all public tweets mentioning the #GDPR hashtag during a period of 6 months. It was reported that the GDPR was being actively discussed by a variety of stakeholders, especially by cyber security and IT-related firms and consultants. However, some of the stakeholders that were expected to have a more active role were salient, which included companies that store or process personal data, government and regulatory bodies, mainstream media, and academics [23].

Another more closely related study was reported by us in 2020 [27]. We analyzed public online communications of public blockchain systems' developers and service providers. We focused on legal documents, including privacy policies, T&C (Terms and Conditions) documents and other similar legal documents published on systems' official websites, and public tweets of their official Twitter accounts. We concluded that most of the systems they investigated had not communicated about GDPR. The legal documents they provided on their websites lacked an explicit acknowledgment and warnings to users on the legal challenges introduced by the underlying blockchain technology [27].

In this follow-up study, we focus on further channels and enrich the Twitter dataset with a much large number of tweets that mentioned at least one of the public blockchain systems investigated in this study. Our goal is to provide further consolidated evidence to achieve a fuller understanding of the communication practices of public blockchain systems' developers and service providers, in order to identify ways to motivate them to be more transparent and active in keeping users aware of the GDPR compliance issue and possible solutions.

3 DATA USED

3.1 Selection of Public Blockchain Systems

In this study, we have followed the approach used in [27] while selecting public blockchains systems. Given the lack of well maintained list of public blockchain systems and the necessary indicators that could be considered while selecting such systems, we focused on cryptocurrencies with a large market capitalization size. We used CoinMarketCap to decide the market capitalization size of cryptocurrencies, from which we also identify the underlying public blockchain systems. Due to the amount of manual work for examining data, we decided to limit our study to focus on top 50 cryptocurrencies. This led us to cover public blockchain systems whose associated cryptocurrencies have a market size greater then 150 million US dollar at the time of our study.

3.2 Online Communication Channels

There are wide-ranging information formats that could be covered to observe GDPR-related discussions. In this study, we aimed to explore the sources including posts on web forums, blog articles and GitHub repositories of the systems, and tweets published by their official accounts and by other accounts that mention at least one of the studied systems. We identified the links to access their

Table 1: Basic statistics of collected documents

Channel	#(Blockchain Systems)	#(Documents)
Twitter	41	13,605,080
Forum and Blog	7/39	17,821
Github	35	970

official Twitter accounts, GitHub repositories and forum and blog posts exploring their websites manually. While working on the GitHub repositories, our initial aim was to access EULAs which were overlooked in the previous study [27] due to the complexity of collecting such information. However, we noticed that GDPR is mentioned in some other documents as well and extended the scope of our search to the whole repositories of the systems. In addition, due to our observations on the passive role of blockchain service providers and developers on relevant discussions on Twitter, we expanded our Twitter dataset to include any tweets that mention their names in order to enrich our findings. The details of our datasets can be seen in Table 1.

3.2.1 Blog and Forum Data. Web forums are online places for Internet users to interact with each, and some are question-and-answer systems used by people to get quick responses from other peer users on the Internet. Blogs are diary-like websites run by individuals and organizations with a present online. The boundary between web forums and blogs are not a clear cut as many blog systems allow comments and discussions to take place. There are several studies that analyzed web forums and blogs for different purposes, however, we did not notice any studies concerning the use of web forums for GDPR discussions in the context of public blockchain services and their users. In this study, we covered web forum and blog posts to explore how blockchain service providers communicate with their users in the context of GDPR and to observe what type of concerns are raised by the users in those discussions. As we identified top 50 cryptocurrencies, we developed customized tools to gather web forum and blog data of those systems using two software libraries – BeautifulSoup [1] and Selenium [16]. In order to eliminate irrelevant posts, we removed the ones that do not cover the word "GDPR" or the full title "General Data Protection Regulation". The total number of web forums, blogs and posts covered in the study can be seen in Table 2.

Table 2: Web forums and blog posts studied

Type	Number
Total number of web forums covered	18
Total number of blogs covered	38
Total posts collected	17,821

3.2.2 Twitter Data. In addition to the approach we used in our previous study [27], where only the tweets posted by the official accounts of the blockchain service providers and developers were investigated, in this study we also focused on blockchain users' perspectives and collected tweets that mention the names of the systems disregarding the accounts they were posted. Using the library GetOldTweets3 [5], we collected tweets published between 1 January 2018 and 1 May 2020 for this purpose.

This approach led us to retrieve more than 13 million tweets from 39 official accounts and Twitter accounts of many users. A majority of those tweets (more than 9 million) mentioned the largest blockchain system #bitcoin. During the preprocessing phase, we performed several actions to remove irrelevant data. First, in order to eliminate tweets automatically posted by bots, we set a threshold for tweets that could be published by an account and removed the tweets of the accounts who posted more than 5,000 posts during a day. Secondly, we filtered the tweets by searching keywords as done in [27] where the researchers had removed tweets that did not contain GDPR. However, in this study, in order not to loose important tweets that may cover a GDPR-related context without explicitly mentioning the name of it, we prepared a list of seed words considering the GDPR elements that challenge blockchain technologies (right to be forgotten, data minimisation, explicit consent etc.). For this purpose, we revisited the GDPR document and the ICO's guide [24] to the GDPR and identified 87 main terms manually (see Table 3).

Table 3: 87 GDPR-relevant seed words used for collecting tweets

Access	Accountability	Accuracy
Adequate	Automated Decision	Certification
Codes of conduct	Confidential	Consent
Contract	Controller	Correct
Criminal	Cyber	Cyberattack
Dataintegrity	Dataleak	Datasafety
Delete	Encryption	Erase
Erasure	Fair	Fairness
Forget	forgotten	Format
GDPR	Hold	Impact
Inform	Law	Lawful
Legal	Legitimate	Limit
Limitation	Long	Machine Readable
Minimisation	Minimum	Needed
Object	Obligation	Offence
Outside	PECR	Period
Personal	Probability	Principle
Processing	Processor	Profiling
Protection	Public Task	Purpose
Rectification	Relevant	Remove
Request	Restrict	Retain
Retention	Personal	Revoke
Right	Secure	Security
Sell	Sensitive	Sold
Special Category	Storage	Third Party
Third Parties	Transfer	Transmit
Transparency	Transparent	Update
Vital Interest	Websites	Withdraw
Years	Confidentiality	GDPR

After removing the tweets that do not contain any of the seed words, we ended up with 1,606,269 tweets. At this stage we decided

to make use of hashtags in the tweets in order to access the most relevant ones. We identified 1,282 hashtags that were used with the hashtag #GDPR. We ranked them in order by their frequencies to detect most popular ones. We identified 14 popular hashtags to be used in further analysis (see Table 4).

Table 4: Hashtags covered in the study

#GDPR	#DataProtection	#DataPrivacy
#privacy	#compliance	#trust
#regulation	#PersonalData	#Regulations
#biometrics	#law	#security
#legal	#brokageofpersonaldata	

There were 36,482 tweets that cover at least one of those hashtags given in Table 4. In order to understand the main themes in those tweets, we ran topic modelling on this set. However, the results were too broad and meaningless which led us to apply further techniques to remove irrelevant data. We randomly selected 4,000 tweets out of 36,482 tweets where we tried to generate a representative sample by covering tweets from each hashtag proportionate to its frequency in the main corpus. We manually labeled them as GDPR-related or non-related and trained a classifier using Naive Bayes algorithm and ran it on the rest of the dataset (32,482 tweets) to identify GDPR-related tweets. Our classifier detected 1,417 GDPR-related tweets. We then ran topic modelling algorithm, on this set without validating the accuracy of our classification approach due to the high volume of data. Topic modelling was used to discover patterns of word use within documents, and it has been frequently applied on Twitter data in the literature. Its aim is to identify topics, which are typically defined as a distribution of words, with documents modelled as mixtures of topics. This whole process is summarized in Figure 1.

3.3 Github Data

GitHub [12] is a code hosting repository based on the Git version control system. It is widely used by the software development community for the reuse of code. Among the 50 cryptocurrencies, we could identify and access repositories of 35 systems using the GitHub API [13]. Using the search pools method, we downloaded 970 GitHub repos belonging to 35 systems.

4 RESULTS

4.1 Web Forum and Blog Data

Blog posts are generally used by the systems to announce their news such as release information or to provide information about the events organised by them. On the other hand, web forums are used to engage with the users of the system and answer their questions in different topics. Some popular topics detected in the investigated forums can be given as discussions regarding the current value of cryptocurrencies, estimation of the future values, rights of individuals or announcement of events organised by the service providers. However, the GDPR discussions were quite limited where we could identify only 56 posts among the 17,821 posts investigated in this study.

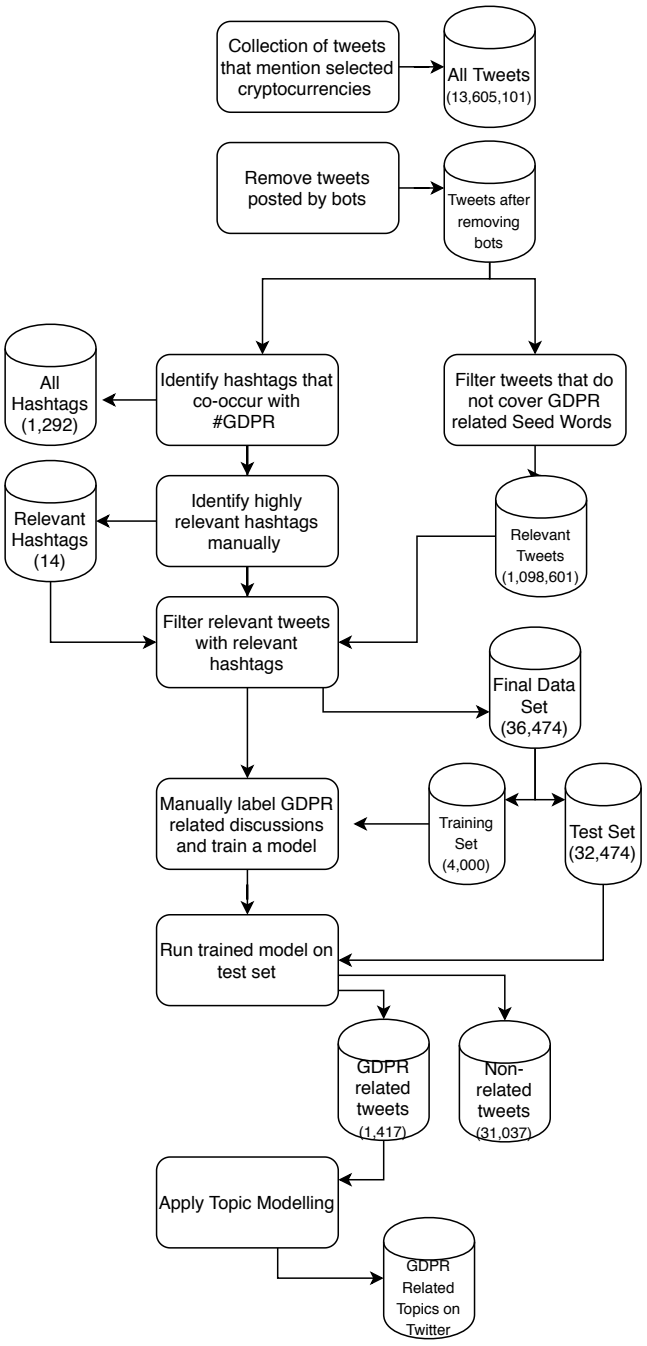

Figure 1: Processes on twitter data.

In the majority of the posts (33 posts out of 56 posts) that cover the GDPR, the discussions are limited to generic information given about the GDPR. However, it is possible to observe some further details regarding a number of GDPR elements in a smaller set of systems. In this section, we have covered the GDPR elements which are discussed in forum or blog posts.

The right to be forgotten is the element which received the highest amount of attention in the blog or forum posts. We identified nine posts, three of are limited to very brief statements. These three posts were written by the administrators of the systems where it was claimed that users of their systems have this right, without any further explanation. In three other posts, users are provided with some further generic information about this right, however it is not explicitly mentioned how users can exercise it while using their systems. Finally, the rest three posts cover clear information about the challenges in deleting information in a public blockchain and it is explicitly stated that the right to be forgotten is not supported by their systems. Surprisingly, we could not identify any question regarding this right asked by the users to the system admins.

The right to access was even less frequently covered in blog and forum posts. We identified 4 posts, two of which provide vague information posted by the system administrators. These two posts are unhelpful to understand how users can exercise this right. We detected a relevant question asked by a user as follows:

"*Where is this data stored? I thought the HoloVault would act as the local source-chain in this case, but could be wrong ... A capability token to access this 'vault' would open up a whole bunch of GDPR-complaint use-cases.*"

However, we could not identify an answer for this question given by the system admins. In another post, published by IOTA [4], it was stated that

"*MAM channels will be used to guarantee ownership, privacy and access control of journey and transport data respectively, for users and providers. ... Future versions of Trinity will also allow users to manage their own journey histories, thus guaranteeing full GDPR compliance.*"

It seems that even though this right may not be supported by the system at the time of this study, it was recognised and planned to be supported to assure the GDPR compliance.

According to the GDPR, the data controllers are responsible for providing information to data subjects about the transfer of their personal data to third parties, third countries or international organizations. We identified only 2 forum posts regarding this issue. In one of them, a user asks a question to clarify an issue in T&C document as

"*Could you please specify exactly in the Terms and Conditions section what data you will be sharing with other participants? Does it include name or email address?*"

However, this question was not answered. In the post which was posted in the blog of Binance [2], the users are informed about third party data sharing as follows:

"*The FATF Travel Rule requires Virtual Asset Service Providers (VASPs) to share Personal Identifiable Information (PII) and Know-Your-Customer (KYC) data between qualifying institutions when executing transactions for senders and receivers.*"

We identified one forum post about right to withdraw consent. In their forums, Cardano stated that

"*More specifically, stake pools will host their own metadata, which is where personal identifiers may be present, and where data management including acting on withdrawal of consent can be managed fully off-chain.*"

It is promising to observe off-chain solutions in practice which are highly recommended by researchers in the literature to assure GDPR compliance. In addition, regarding the gathering explicit consent, IOTA [14] stated that

"*As many others in this space, we envision a new way of approaching data control, enforcing opt-in policies instead of an opt-out approach. This means that users would have to provide consent allowing companies to process their data, and only to the level the user has agreed upon. Whenever the user changes his / her decision, the company is not allowed (or only allowed with reduced capabilities) to process the data.*"

Actively opting in is important to meet the standard of an unambiguous indication by clear affirmative action in Article 4 in the GDPR. It is also quite promising to observe those details in posts of blockchain systems even if it is not given as an existing feature in their systems.

According to the GDPR, storing and processing personal data should be is adequate and limited to what is necessary. This principle, namely data minimisation, is only explicitly covered by IOTA [3]. It was stated that

"*During the hackathon we discovered that GDPR requirements and our'less is more' perspective were nicely in sync with respect to the handling of privacy related data like more decentralized processing, the user in control, only collect data necessary for the service delivered, enabling anonymization to remove barriers for sharing data that can aid 'less energy' services and more efficient1 usage of the energy infrastructures.*"

This post is quite relevant to this principle with its statement "*only collect data necessary for the service delivered*", however, it is surprising not to find any explanation about the conflict between decentralised data and data minimisation.

Data protection by design and default is another GDPR element which was mentioned only in one post. This is again a post published by IOTA [15] stating that

"*IOTA provides data protection by default and by design, it relies on a trustless model. A model that allows people to operate directly with one another, trust with any of the actors in the ecosystem. Nodes in the network do not have authority over any other node, hence a decentralised model of a distributed ledger.*"

4.2 GitHub Data

The discussions about the GDPR in GitHub repositories are mainly the legal documents such as privacy policies or user interfaces that are developed to display information to the end users of the systems. We did not identify any source code that mentions the GDPR explicitly other than ones that are used to communicate with users through graphical user interfaces.

In total, we identified 27 files in the investigated repositories that belong to four systems: Ethereum [20], Holochain [7], Hydrogen-dev [8] and Zackbitcoin [9]. It is mainly the privacy policy of Ethereum that covers the GDPR mentions in its repository whereas Holochain mentions GDPR in one of its json files providing a generic information as

"*The GDPR aims primarily to give control back to citizens and residents over their personal data and to simplify the regulatory environment for International business by unifying the regulation within the EU.*"

We identified two other files belonging to the same system, where the personal information is claimed to be deleted or edited open

request. However, the conditions under which these procedures will be applied or how the users can exercise it are not given.

Hydrogen-dev and Zackbitcoin covered statements about their future plans to comply with the GDPR in their user interfaces. Zackbitcoin stated that

"*Europe – GDPR (data protection regulations) come into effect May 25th. Still unclear how exchanges should respond. ShapeShift hasn't released their policy re: GDPR but promises an update for the community soon – can probably piggyback on their strategy.*"

and Hydrogen-dev covered GDPR compliance as a risk and stated the following:

"*To combat the risk of low adoption and competition, it is imperative for the following things to happen:*
- Acceptance by consumers of the concept of decentralization of data
- Continued privacy pushes, such as GDPR, by central governments globally
- Education about the risks of centralized document storage."

Ethereum's privacy policy was the only document in the investigated repositories that covers detailed statements about the conflicts between the right to be forgotten and the immutable nature of the public blockchain systems [6]. It is stated that

"*... In most cases though it is both dangerous and in some cases illegal (according to EU GDPR rules for example) to record Identity Claims containing Personal Identifying Information (PII) on an immutable public database such as the Ethereum blockchain.*"

The policy even covers the cases where minor's information is stored unintentionally, and states that parents have right to erasure this type of information.

Even though the right of access has been reported to be entirely compatible with the blockchain technology [25], only three systems mention this right in their repositories. In one of the identified file [7], it is stated that

"*You have the right to obtain from the Foundation free information about your personal data stored at any time and a copy of this information.*"

Similar brief statements were observed for the other systems regarding this right. However, those statements are not very helpful in means of guiding users while exercising it.

Consent is one of the GDPR elements that received the least attention in GitGub repositories. We identified only one file mentioning the explicit consent and another one file mentioning right to withdraw consent both which provide very brief generic statements about those GDPR elements. Similarly data portability was observed in one of the systems' repositories where it was stated that

"*HoloVault puts you in control of how your information is used and allows you to share the same information to many different apps.*"

Storage limitation was also observed to be overlooked by the systems where we identified only one document giving very generic information about this principle as follows:

"*The Foundation will process and store the personal data of the data subject only for the period necessary to achieve the purpose of storage, or as far as this is granted by the applicable laws or regulations. If the storage purpose is not applicable, or if a storage period prescribed by the applicable laws expires, the personal data is routinely erased in accordance with the legal requirements.*"

4.3 Twitter Data Analysis

After pre-processing step summarized in Figure 1, we ran several experiments and generated different number of topics with different number of words using the LDA algorithm. We achieved better results when we generated 5 topics each of which consist of 10 words. Those results can be seen in Table 5.

Table 5: Results of topic modelling

Topic	Words
Topic 1	essential ico decentralized tokensale crypto dapps invest crowdfunding access right
Topic 2	security privacy ico bit data project new pany GDPR capitaltechnologiesresearch
Topic 3	data personal pdata secure crypto opiria yourblock io brokageofpersonaldata opirium
Topic 4	ico crypto news io tokensale smart contract technology platform right
Topic 5	ico crypto tokensale project join forget dont token secure security

In order to interpret those topics, we explored the tweets that contain the topics generated. It is possible to conclude that, since the GDPR-related tweets are limited, the discussions can easily be dominated by systems that advertise themselves by posting tweets that announce their GDPR compliance. Yourblock [17] is one of them whose tweets resulted Topic 3 to be generated. One of the GDPR-related tweets of this system can be given as

"*Yourblock will help you with your right to erasure also known as 'the right to be forgetten'.*"

Another system that dominated the GDPR discussions in the investigated dataset is Opiria which is a global decentralized marketplace for the secure and transparent buying and selling of personal data [11]. Those finding reveal that it is not possible to identify rich GDPR discussions neither in the timeline of the blockchain systems and service providers nor in the tweets that mention them. In addition, as inline with the finding in the literature [23, 27] GDPR discussions in blockchain context were not observed to be posted by the blockchain systems investigated in this study. The hashtag #ico was observed in 355 tweets which yielded it to appear in the generated topics. Here, it is noteworthy that this hashtag stands for Initial Coin Offering not Information Commissioner's Offices [10]. Therefore, the tweets with this specific word were not directly related to the GDPR.

5 CONCLUSION AND FUTURE WORK

In this study, we aimed to deepen our understanding of how blockchain sector sees the challenges introduced by the GDPR. We have conducted a data driven study based on a large database of the public communications regarding the GDPR in blockchain context. Due to the technical challenges in processing GitHub repositories and Forum posts, we have limited our study to top 50 cryptocurrencies with a market size greater than $150 million at the time of this study. It is possible to report that the lack of explicit acknowledgment and warnings to users on the legal challenges introduced by the nature of the blockchain technology still remains the same as reported in

[27] one year ago. We could not identify necessary warnings in GitHub repositories even though we had expected to see some in EULAs in those repositories. It was the same on Forum and blog posts where we could not identify satisfactory communications of blockchain systems and service providers on GDPR with their users. Our study also confirms that those systems do not have an active role in the GDPR discussion on Twitter. Given the immutable and distributed nature of the public blockchain technologies, we keep our call for more research into the interfaces between data protection law and the blockchain technology.

REFERENCES

[1] [n.d.]. Beautiful Soup Documentation — Beautiful Soup 4.9.0 documentation. https://www.crummy.com/software/BeautifulSoup/bs4/doc/. (Accessed on 07/05/2020).

[2] [n.d.]. Binance Joins Shyft Network to Launch Global FATF Travel Rule Solution | Binance Blog. https://www.binance.com/en/blog/421499824684900436/Binance-Joins-Shyft-Network-to-Launch-Global-FATF-Travel-Rule-Solution. (Accessed on 07/05/2020).

[3] [n.d.]. Blockchaingers 2018 — Boosting Co-Creation with IOTA. https://blog.iota.org/blockchaingers-2018-boosting-co-creation-with-iota-8086c215caad. (Accessed on 07/07/2020).

[4] [n.d.]. Developers community update: IOTA & +CityxChange - IOTA. https://blog.iota.org/iota-cityxchange-community-update-85f43894bcca. (Accessed on 07/07/2020).

[5] [n.d.]. GetOldTweets3 · PyPI. https://pypi.org/project/GetOldTweets3/. (Accessed on 06/02/2020).

[6] [n.d.]. GithubEtherium. https://github.com/ethereum. https://github.com/ethereum.

[7] [n.d.]. GithubHolochain. https://github.com/holochain.

[8] [n.d.]. GithubHydrogen. https://github.com/Hydrogen-dev.

[9] [n.d.]. GithubZack. https://github.com/zack-bitcoin.

[10] [n.d.]. ICO. https://ico.org.uk/.

[11] [n.d.]. Opiria. https://www.opiria.com/.

[12] [n.d.]. Privacy is not a currency - IOTA. https://github.com/.

[13] [n.d.]. Privacy is not a currency - IOTA. https://developer.github.com/v3/.

[14] [n.d.]. Privacy is not a currency - IOTA. https://blog.iota.org/privacy-is-not-a-currency-63018fc45920. (Accessed on 07/05/2020).

[15] [n.d.]. Privacy is not a currency - IOTA. https://blog.iota.org/privacy-is-not-a-currency-63018fc45920. (Accessed on 07/07/2020).

[16] [n.d.]. selenium · PyPI. https://pypi.org/project/selenium/. (Accessed on 07/05/2020).

[17] [n.d.]. Yourblock. https://www.yourblock.io/faq.

[18] Giuseppe Ateniese, Bernardo Magri, Daniele Venturi, and Ewerton R. Andrade. 2017. Redactable Blockchain – or – Rewriting History in Bitcoin and Friends. In *Proceedings of the 2nd IEEE European Symposium on Security and Privacy*. 111–126. https://doi.org/10.1109/EuroSP.2017.37

[19] Shaen Corbet, Brian Lucey, Andrew Urquhart, and Larisa Yarovaya. 2019. Cryptocurrencies as a financial asset: A systematic analysis. *International Review of Financial Analysis* 62 (2019), 182–199.

[20] ethereum.org. [n.d.]. ethereum · GitHub. https://github.com/ethereum.

[21] European Parliament. 2016. Regulation (EU) (2016) 2016/679 of the European Parliament and of the Council of 27 April on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union 59(L 119). https://eur-lex.europa.eu/eli/reg/2016/679/oj

[22] European Union Blockchain Observatory & Forum. 2018. Blockchain and the GDPR. thematic report. https://www.eublockchainforum.eu/sites/default/files/reports/20181016_report_gdpr.pdf

[23] Anatoliy Gruzd, Deena Abul-Fottouh, and Atefeh Mashatan. 2020. Who is Influencing the# GDPR Discussion on Twitter: Implications for Public Relations. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*.

[24] ICO, UK. [n.d.]. Guide to the General Data Protection Regulation (GDPR). https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/. (Accessed on 05/20/2020).

[25] Florian Martin-Bariteau. 2018. Blockchain and the European Union General Data Protection Regulation: The CNIL's Perspective. Blckchn.ca Working Paper Series.

[26] Satoshi Nakamoto. 2008. Bitcoin: A peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf.

[27] Rahime Belen Sağlam, Çağrı B. Aslan, Shujun Li, Lisa Dickson, and Ganna Pogrebna. [n.d.]. A Data-Driven Analysis of Blockchain Systems' Public Online Communications on GDPR. ([n. d.]).