Protecting Personal Data with Blockchain Technology

19

Alexander Masluk and Mikhail Gofman

Abstract

Service providers depend on the ability to host, analyze, and exchange the personal data of users. Legal and contractual frameworks aim to protect the rights of users regarding this data. However, a confluence of factors render these rights difficult to guarantee. This paper evaluates the potential of blockchain technology as a mechanism for achieving transparency and accountability in the realm of personal data collection.

Keywords

Security · Cybersecurity · Neural networks · Machine learning · Network security · Intrusion detection · Artificial intelligence

19.1 Introduction

Unless a user takes extraordinary protective measures, typical computer use results in the creation of vast quantities of personal and private data and metadata [1–4]. Each day, people upload terabytes of data to storage platforms like Dropbox, Google Drive, and social networking sites like Facebook. In addition, environments such as Google and Windows 10 passively collect even greater quantities of metadata in the form of keystroke dynamics, usage duration logs, and the like. By applying machine learning and data analytics to this data and metadata, service providers are able to refine the interfaces and content of their products, and derive profit through targeted advertising platforms such as Google's AdWords [3, 4, 10].

A. Masluk · M. Gofman (✉)
Department of Computer Science, California State University, Fullerton, CA, USA
e-mail: alex.masluk@csu.fullerton.edu; mgofman@fullerton.edu

The benefits both user and provider enjoy in this arrangement are difficult to overstate. Users have access to a wide range of robust, free-to-use internet services, and service providers procure significant profits [5–7]. Yet, with regular data breaches in the news and an increasing public awareness of the depth of information that can be gained through data mining, this arrangement has become a cause of deep concern for users fearing that their privacy may be compromised or their data misused.

Privacy policies and other mechanisms behold service providers to certain practices regarding the data they collect [1–4, 8]. These practices limit the circumstances under which they may retain a user's personal data, share the data with third parties, and so on. However, several challenges impede the efficacy of these agreements and associated enforcement mechanisms. The users may find the privacy policies difficult to understand and draw unwarranted conclusions about the degree to which their data really is private. The service provider may inadvertently or intentionally violate policy without leaving any sort of record of having done so [9].

Our motivation is to better understand the specific data collection and use practices of service providers according to a close reading of the relevant policy agreements, and propose a mechanism by which a greater degree of provider transparency and accountability may be achieved to aid in enforcing said policy. To this end, we will examine the application of blockchain technology as an accounting mechanism for personal data collection, retention, and exchange. We will see that defining a set of data transaction types and recording these in a publicly verifiable ledger helps to achieve this accountability.

The organization of the paper is as follows. Section II establishes the preliminary concepts of data collection and blockchain technology; section III discusses the limitations of allowable use of collected data. In the fourth section we present the concept of our proposed solution for applying

S. Latifi (ed.), *Information Technology – New Generations*, Advances in Intelligent Systems and Computing 738,
https://doi.org/10.1007/978-3-319-77028-4_19

blockchain to personal data lineage, while in the fifth we touch upon related research. The sixth section concludes the paper.

19.2 Preliminaries

19.2.1 Data Collection Scope

Necessary to any proposed scheme is a realistic view of the breadth and depth of data collected. To that end, we examined the data collection policies of four prominent service providers: Apple, Facebook, Google, and Microsoft.

Each of these entities differ in terms of the business model under which they operate and the nature of services they provide. Nonetheless, we discover a uniformly maximalist approach to the type of data they each afford themselves license to collect, with noteworthy differences.

All four provide cloud-based repositories for data to be stored privately or shared with others. The providers reserve the right to analyze both varieties. So too do each of these providers log and analyze metadata in many or all of the following forms: frequency and duration statistics; details of the hardware, operating system, and file system contents of the device used to access a service; location data; data generated by a device's input peripherals such as keyboards, touch screens, microphones, and webcams; and other sources too numerous to list exhaustively [1–4].

Many web-based services connect and interact with one another. This offers providers additional data collection vectors. For example, Facebook owns the virtual-reality platform Oculus and collects data generated by use of that platform [1]. In general, using one's identity on one of these platforms to access another service allows the provider to collect and link data from each service [1, 2]. Apple and Microsoft distinguish themselves here by enumerating safeguards against linking the data collected from third party or subsidiary services to the user's identity [3, 4].

In the case of social media platforms, information provided about one user by another user may be collected. For example, if Alice uploads a photo of Bob to Facebook, Facebook can link the photo with information it has collected directly from Bob [1].

Additionally, through use of "unique application numbers", Google can link together multiple accounts that have been accessed through a single app installation. This may indicate multiple accounts held by an individual, or a connection between multiple individuals who used the same device [2]. In its own privacy policy, Microsoft specifies that it declines to collect this particular type of data [3].

Though noteworthy differences exist, four privacy policies we examined all took a broad approach to delineating the types of data their policies permit them to collect, and the uses to which the data can be put.

19.2.2 Blockchain

Blockchain is the distributed ledger technology which guarantees the value and controls the inflation rate of Bitcoin. It achieves this through mechanisms such as proof of work and mutual consensus [11].

All Bitcoin transactions are recorded in the blockchain ledger. A transaction records an amount to be transferred from one Bitcoin "wallet" to another. Since a given wallet's balance is calculated as the running summation of ledger-recorded transactions involving that wallet, the ability to guarantee the validity of each transaction guarantees the currency's viability and value [11, 12].

The actual process of recording new ledger transactions takes place in a distributed manner, with multiple nodes competing to conduct the validation process necessary to add the transaction to the ledger. To win the competition, a node must solve a computationally difficult problem ("proof of work"), the answer to which can be evaluated by other competing nodes for correctness. Nodes will attempt to append or "chain" new transactions to previous transactions that they agree are correct ("mutual consensus"). Unless the majority of blockchain nodes are controlled by hostile actors, this system guarantees the validity of ledger transactions [11, 12].

We are particularly interested in blockchain's ability to facilitate trustworthy record keeping in a distributed environment.

19.3 Data Usage Limitations

We hope to propose a mechanism for ensuring transparency and accountability regarding personal data collection. To this end, we must understand what rights and limitations exist regarding the data collection sphere. These are the limitations whose enforcement we intend to support.

Insofar as a user licenses through agreement the provider's right to various uses of their data, we turn again to the privacy policies of four prominent providers. We also examine the EU-U.S. Privacy Shield, a self-certification framework that prescribes additional obligations on the part of the provider.

19.3.1 Limitations Through Organization Policy

Although the four privacy policies we examined tended towards similarity regarding the breadth of data the providers enjoy license to collect, greater distinction can be discovered in the uses to which the data is put. Google and Facebook make significant use of the ability to draw connections between a user's activity across different services, platforms, and accounts in order to build a unified data profile [1, 2]. By contrast, Microsoft and Apple may decline to collect this information; in Apple's case, the privacy policy elaborates the safeguards Apple employs to confound the potential of doing so [3, 4].

Similarities exist as well. Each privacy policy reserves the provider's right to transfer a user's data to a third party entity for data processing purposes. We are assured in each case that their privacy agreement extends to any third party thus employed. None of the policies made it clear how the identity of these potential third parties can be discovered [1–4].

The four providers distinguish between sensitivity levels of the data they collect, subjecting more sensitive data to more stringent protective measures. For example, Facebook promises not to display advertisements based on a user's medical condition. Apple makes a broad distinction between "personal" and "non-personal" data. The policies provide little in the way of a concrete methodology for evaluating and assigning the sensitivity level of data collected [1–4].

Regarding the issue of data retention, each provider's policy stipulates that the provider will remove any individual user datum upon request. However, each also includes a similar caveat, providing for circumstances in which they may fail to fulfill the request. Facebook may decline to remove data if a "good faith belief" implies that the information could be "necessary to ... protect ourselves, you and others" [1]. Google may not remove information that resides on their backup systems [2]. Microsoft may elect to "access, transfer, disclose, and preserve personal data" when it believes doing so will "protect our customers" and "protect the rights or property of Microsoft" [3]. Apple can retain data for a "longer retention period" than specified in their privacy policy if doing so is "permitted by law" [4]. We discovered no official company documentation regarding how the decisions to retain or delete data are made.

19.3.2 Limitations through the EU-U.S. Privacy Shield

Many service providers transfer data across state and national boundaries, and in doing so fall subject to a variety of regional privacy laws too numerous to summarize here. In general, European Union member nations possess privacy laws of greater stringency than the U.S [13]. This presents EU nations with the challenge of protecting the privacy rights of its citizens while permitting them to use services that may result in their data residing in more unrestricted nations such as the United States. The EU-U.S. Privacy Shield framework addresses this dilemma [8].

The framework operates under the principle of self-certification. An organization self-certifies that it meets the Privacy Shield's requirements. Among other specifications, these require a member organization to make available a transparent and complete privacy policy, including an exhaustive list of all the uses to which the collected data is put. They must provide a mechanism for responding to user complaints if deviations from the policy should be suspected, and only transfer user data to organizations with equal or greater privacy protection [8]. Microsoft, Facebook, and Google are Privacy Shield member organizations; Apple is not [14].

The specifications of the Privacy Shield strike a balance between permissiveness and restrictiveness. This makes intuitive sense. Given the intent of bounding allowable uses of personal data, such a framework must be restrictive in some degree in order to perform any useful function. Conversely, fewer organizations will find their mode of operation compatible with overly restrictive regulations, and will decline to adopt the certification.

It is easily understood that the Privacy Shield derives its authority from the degree to which a broad number of U.S. organizations attain membership. EU member nations cannot exert any legal authority over U.S. organizations, but an EU data protection authority (DPA) may prohibit data trade with non-certified U.S. organizations. It becomes less advantageous to exert such prohibitions in a circumstance where few U.S. organizations are Privacy Shield members, since it prohibits trade with a vast majority of potential business partners. On the other hand, trade restrictions imposed by DPAs provide a market incentive for U.S. organizations to obtain certification.

A weakness of the Privacy Shield framework stems from our core problem of accountability. How might a user determine, for example, that their data was transferred to a third party that failed to fulfill the "equal or greater" privacy standard, or that the provider neglected to delete data upon request? The framework might be strengthened by insisting the adherence of member organizations to a mechanism guaranteeing transparency in this regard, supposing such a mechanism could be implemented with sufficiently non-restrictive impact.

19.4 Blockchain for Personal Data

How can blockchain be used to promote transparency and accountability in the data collection sphere? Here we will outline the general concept of how we propose to apply the technology.

19.4.1 Premise

For our purpose, we consider the Bitcoin concept of the "wallet". For Bitcoin, the blockchain ledger supports two simple wallet operations: adding or removing funds to and from a wallet. These wallets are pseudonymous and do not reveal the identity of their owner.

We propose to extend the concept of the "wallet" into a hierarchy of data repositories (DRs). At the higher level, a master data repository (MDR) refers to a service provider's primary or backup storage; unlike Bitcoin's wallets, the MDR's owner is publicly known. Sub-DRs correspond to the individual users whose data falls under the provider's possession.

In this context, our blockchain should support the following transaction types: WRITE, COPY, ERASE, and PROCESS. These transactions contain information corresponding to the relevant DR, user datum, and transaction time. Tracing these transactions reveals a sequential chain of custody for each datum, allowing data lineage to be established. The service provider exposes an interface for users or enforcement entities to examine the ledgers and gain knowledge of these custody chains..

Examples of how the flow of data translates into ledger transactions can be seen in Fig. 19.1. Provider 1 (P1) collects data from users, which result in a series of WRITE transactions in the provider's primary storage, MDR_P. From MDR_P they are also stored in backup, resulting in a COPY entry for MDR_P and a WRITE entry for the backup storage, MDR_B. P1 associates its data with the users that originated them, so we also see a series of COPY operations for MDR_P and corresponding WRITE operations for the SDRs corresponding to each user.

Suppose P1 outsources analysis of its data to a third party, P2. The transfer required for this outsourcing results in COPY transactions for P1's MDR, and WRITE transactions for P2's. Additionally, we will see PROCESS transactions corresponding to P2's MDR, indicating the nature of the operations performed on the data.

Transactions of type PROCESS also indicate instances where processing data results in the creation of derived data distinct from—but based upon—the original data. In this case, the PROCESS ledger entry will indicate both the original data and any data derived from the original data, allowing data lineage to be traced through derivations.

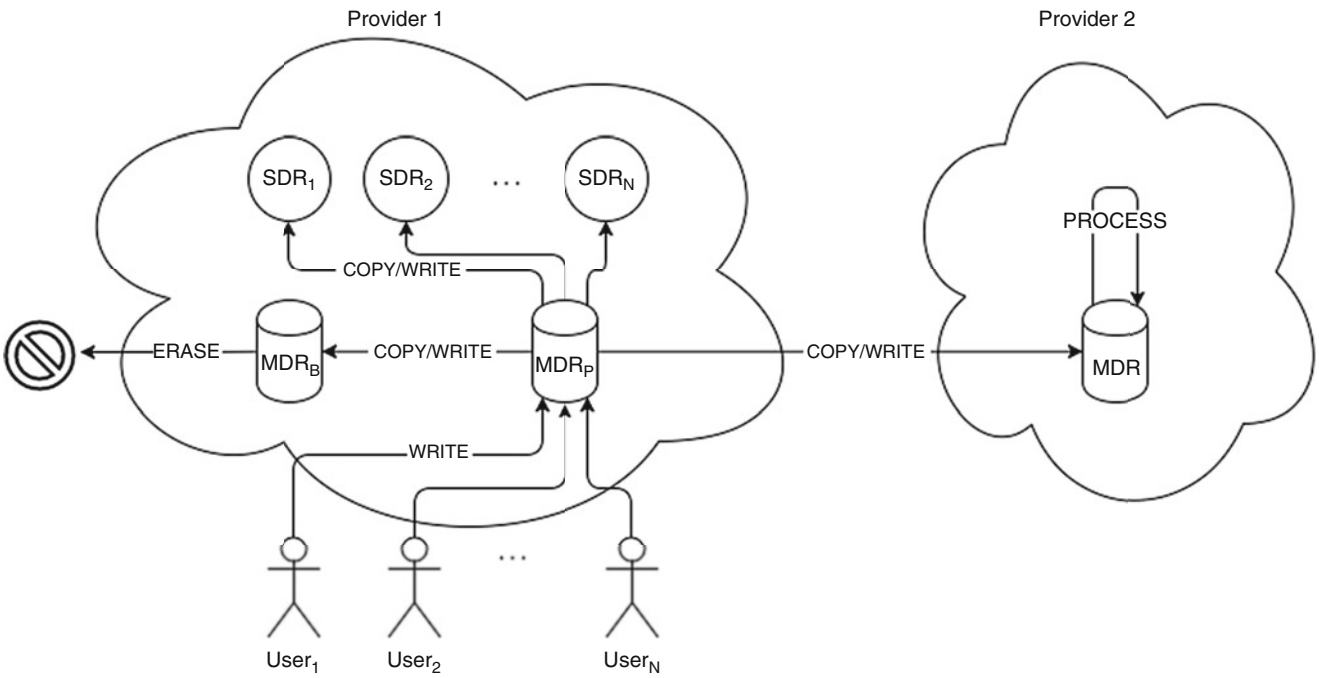

Fig. 19.1 How the flow of data translates into ledger transactions

The complete contents of a ledger transaction include the following:

1. Transaction timestamp
2. Data identifier
3. Transaction type
4. Source MDR
5. Destination MDR
6. Process performed (if applicable)
7. Derived data identifier (if applicable)

This design provides accountability for several typical data collection scenarios.

19.4.1.1 What Data Has Been Collected?

We can examine the ledger of WRITE operations that a given provider has recorded to see exactly what data they have collected about us. The hierarchical nature of the data repositories give us an access control mechanism to allow a user to track the data belonging to their SDRs alone, whereas an enforcement agency may be given broader access to MDR ledger information.

19.4.1.2 Has My Data Been Deleted?

The ERASE ledger entries allows us to verify that data requested for removal has indeed been removed. Since copying data from primary storage to backup storage results in a COPY and WRITE operation in the two respective high-level DRs, we can verify that the data no longer resides in backup by finding the corresponding ERASE entry in the backup DR ledger if necessary.

19.4.1.3 With Whom Has My Data Been Shared?

An examination of the COPY entries allows us to trace each instance in which our data was transferred to a third party DR. The third party may be audited for corresponding ERASE transactions to ensure they have not retained any data.

19.4.1.4 Have Multiple Accounts of Mine Been Correlated?

Because low-level DRs must correspond to individuals, data collected from disparate but correlated accounts must belong to the same low-level DR. We are thus able to confirm the extent to which our activity on various accounts and services have been compiled into a single data profile.

19.4.1.5 What Exactly Are you Doing with My Data?

Because Privacy Shield-compliant organizations must publish an inclusive list of applications for which our data is being used, we speculate that each such organization could publish a list of keywords associated with each. If Google analyzes our search strings for use in AdWords, the corresponding PROCESS ledger entry may include the keyword "ADWORDS" or, more generally, "MARKETING". This allows users to verify their data are being used exclusively for the purposes for which they have been licensed.

19.4.2 Challenges of Implementation

There exist a few generic implementation challenges when it comes to generalizing blockchain from cryptocurrency to other applications. Several scale-related obstacles could affect the viability of implementing our proposal: the Bitcoin network can handle a maximum of seven transactions per second; each copy of the blockchain ledger occupies 50 gb of storage space; the process of validating new transactions and adding them to the chain consumes energy on the order of $15 million USD per day. A blockchain capable of handling the number and frequency of transactions associated with big data collection and processing would be prohibitively resource-intensive if implemented using the same strategy [12, 15, 16].

A few strategies can be leveraged to mitigate the issue of scale. The major factor that causes the Bitcoin network to perform slowly and consume great amounts of energy is the computationally intensive task necessary to validate a block of transactions, the "proof of work" (PoW). Alternatively, we can make use the faster and less expensive "proof of stake" (PoS). Rather than competing to solve a computationally difficult problem, nodes are awarded the right to add transactions to the chain using a lottery-based system [12]. Kiayias et al. present a provably secure implementation of the PoS model [17, 20].

The Bitcoin network uses a single, permissionless ledger, but this may not be necessary in our case. It may be reasonable to allow each provider to implement a local, closed-participation blockchain that meets a given set of requirements. Using such a stratified approach would help alleviate scaling issue of the storage size required for the ledger itself, since service provider A would not need to store a copy of service provider B's data transactions. This may weaken the efficacy of the design since a provider, controlling all of the nodes participating in the blockchain, could generate counterfeit ledger entries. Nonetheless, an internally distributed blockchain protects against rogue individuals within an organization attempting to falsify records, and against security breaches in which a minority of participating nodes fall under hostile control.

Another issue is the variety of datatypes we see collected, which range from small strings of text to large video files. Hashing can be used to uniformly transform these data into a fixed, manageable size. Existing blockchain implementations make use of this strategy [15]. The strategy could be

extended from individual data to datasets. The total set of all data pertaining to a user hashes to a single unique value; any subset of that data transferred or operated upon would hash to a single value as well.

Though beyond the scope of this research, a more detailed design would be needed in order to formally evaluate the efficiency, and therefore viability, of our proposed system.

19.5 Related Works

In the course of our research, we encountered these proposed personal data accountability mechanisms, and applications of blockchain besides cryptocurrency.

Mehmood et al. provide a useful survey of the various strategies proposed and in use for augmenting privacy in the big data context. These consist largely of cryptographic and anonymization techniques. We store user data in an encrypted form to protect against data breaches, and we present data in an anonymized form so that it can be analyzed without revealing the identity of the data's originator. A discussion of integrity verification obliquely touches upon the issue of accountability, but only in terms of providing a mechanism to guarantee the data has not been altered [18].

These concerns, while important, may at odds with the our use cases. Generalizing and suppressing datasets inevitably leads to a loss of integrity in the original dataset; this trade-off may be appropriate when exposing the data to third parties for analysis, but users expect the data we upload to the cloud to be preserved exactly as we created it. Encryption, especially when implemented with a multitude of keys relative to the user's identity or attributes, introduces additional overhead that does not lend itself well to the sort of parallel computing strategies big data consumers use to expediently process data.

Gao and Iwane [19] present a model for social networking with robust privacy features. The model works by introducing trusted intermediaries between social networking providers that guard a user's data and only permit access to the data as specified by the user's preferences. These intermediaries or "virtual-network control centers" (VCCs) facilitate data transfer between participating social media organizations (POs). The VCCs maintain an anonymized copy of the user data to supply to other POs upon request, should the request be permissible given the access controls assigned to the user data. The participating organization must implement connectivity with these VCCs and expose the implementation to the user, who may decline to make use of the system and simply expose their data directly from the PO using the traditional approach. Though Gao and Iwane discuss their model in terms of intentionally created data, the approach likely generalizes to automatically collected metadata as well; a user could, for example, toggle "location history" as a private field in the VCC interface, which will then block social media platforms from sharing this information.

Gao and Iwane's approach addresses the accountability concern with regards to data transfer between parties: if a social media provider wishes to transfer data to a third party entity, it must do so through user-obedient VCCs. It does not, however, promote transparency regarding initial data collection or data retention. The VCC has no awareness of what a PO may be doing with a user's data until a situation arises for which the VCC must act as intermediary.

A potential issue arises from how the VCCs themselves will be implemented, whether through centralized trusted authorities, or an open-participation distributed model similar to certain blockchain implementations. Relying on a trusted third-party to act as intermediary for the huge amounts of data characteristic of social media may create a central bottleneck and point of failure, whereas a distributed model introduces security and privacy concerns.

Setting big data privacy and accountability aside [21, 22], we examine how blockchain technology has been applied to spheres other than cryptocurrency in order to achieve accountability. The organization Everledger attempts to apply distributed ledger technology to luxury goods such as diamonds and fine art. The organization Factom is developing a protocol that they hope will be used to apply distributed ledger methodology in any sphere where record-keeping accountability is needed [15]. Factom and Everledger apply to scenarios in which records are already being kept and trustworthy records must be separated from counterfeit or otherwise illegally-produced records. In the realm of personal data collection, we are concerned not only with the trustworthiness of records, but with ensuring that records are being kept in the first place, and kept in a way that provides data lineage transparency.

As well as use and retention, one of our primary goals is achieving to achieve accountability with regards to data lineage. Backes et al. propose a framework for building data lineage transparency into data exchange at the bit level. This is achieved through "robust watermarking". A data sender alters the data before sending it so that the data contains a detectable watermark uniquely associated with the data receiver. The watermark has the following properties: it preserves the original data's information in such a way that it can be processed according to the original intent, and can coexist with multiple other watermarks such that the order in which they were applied to the data can be discerned [23].

For data lineage, this is in ways a superior approach to our proposed blockchain model, since it functions at the protocol level and avoids the significant overhead of maintaining transaction ledgers; the ledger information is instead stored

in the data itself. Our scheme could perhaps be augmented if the concept of robust watermarking could be extended to describe not just data transfers, but various distinguishable data processing operations.

19.6 Conclusion

Computer users enjoy access to a range of robust and inexpensive services and applications. The providers of these rely upon the ability to collect, share, and analyze large amounts of personal data in order to financially support themselves and deliver quality products. Users have the expectation that providers will adhere to a set of rules regarding the collection, use, and retention of their data, but these rules are difficult to enforce given the lack of a mechanism guaranteeing transparency and accountability regarding the life of a given piece of data.

We turn to the distributed ledger technology of blockchain to provide such a mechanism. By suggesting a logical organization of collected data into data repositories, and defining a simple set of data transactions between them, we have shown how a publicly visible ledger of these transactions would facilitate data accountability and empower enforcement agencies.

References

1. Data Policy. Facebook, 2017, www.facebook.com/about/privacy
2. Apple Legal—Legal—Privacy Policy—Apple. Apple, Apple Legal, 2017, www.apple.com/legal/privacy/en-ww/
3. Privacy Policy—Privacy & Terms—Google. Google, 2017, www.google.com/policies/privacy
4. Privacy—Microsoft Privacy. Microsoft, 2017, privacy.microsoft.com/en-US
5. Google's ad revenue from 2001 to 2016 (in billion U.S.dollars). Statista, 2017, www.statista.com/statistics/266249/advertising-revenue-of-google
6. Tam, Donna. Facebook processes more than 500 TB of data daily. CNET, 2012, www.cnet.com/news/facebook-processes-more-than-500-tb-of-data-daily
7. J. Dean, S. Ghemawat, MapReduce: Simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)
8. Privacy Shield Framework. Privacy Shield, 2017, www.privacyshield.gov/EU-US-Framework
9. Minelli, Michael, et al., *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*, Hoboken, Wiley, 2012, pp. 151–167
10. Your guide to AdWords. Google, 2017, support. google.com/adwords/answer/6146252hl=en&ref_topic=3119071,3181080,3126923
11. Shute, Jeff, et al., F1: the fault-tolerant distributed RDBMS supporting Google's Ad business, in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 2012, pp. 777–778
12. Protocol Buffers. Google, 2017, developers. google.com/protocol-buffers/docs/overview
13. C. Garcia, Demystifying MapReduce. Procedia Computer Science **20**, 484–489 (2013)
14. Nakamoto, Satoshi. *Bitcoin: A Peer-to-Peer Electronic Cash System*, Bitcoin, 2008, bitcoin.org/bitcoin.pdf
15. Judmayer, Aljosha, et al, *Blocks and Chains: Introduction to Bitcoin, Cryptocurrencies, and Their Consensus Mechanisms*, 2017
16. Greenleaf, Graham. *Global Data Privacy Laws: 89 Countries, and Accelerating. Privacy Laws & Business International Report, no. 115*, 2012., papers.ssrn.com/sol3/papers.cfm?abstract_id=2000034
17. Participant Search. Export.gov, International Trade Administration, 2017, www.export.gov/participant_search
18. S. Underwood, Blockchain beyond Bitcoin. Commun. ACM **59**(11), 15–17 (2016)
19. J. Yli-Huumo et al., Where is current research on Blockchain technology?-a systematic review. PLoS One **11**(10), e0163477 (2016)
20. A. Kiayias et al., Ouroboros: A provably secure proof-of-stake Blockchain protocol. Advances in Cryptology **10401** (2017)
21. Mehmood, Abid, et al, Protection of big data privacy. Access, IEEE, vol. 4, 2016, pp. 1821–1834
22. Gao, C., & Iwane, N., A social network model for big data privacy preserving and accountability assurance. in *Consumer Communications and Networking Conference (CCNC), 2015 12th Annual IEEE*, pp. 19–22
23. Backes, Michael, et al, Data lineage in malicious environments. in *Dependable and Secure Computing, IEEE Transactions On, vol. 13, no. 2*, 2016, pp. 178–191