Transparent Cloud Privacy: Data Provenance Expression in Blockchain

Gabriel Hogan and Markus Helfert
ADAPT Centre, Dublin City University, Ireland

Keywords: Cloud, Privacy, Data Provenance, Transparency, W3C, PROV, Blockchain, Distributed Ledger Technologies.

Abstract: The development of Cloud processing and 'Big Data' have raised many concerns over the use to which data is being put. These concerns have created new demands for methodologies, and capabilities which can provide transparency and trust in data provenance in the Cloud. Distributed ledger technologies (DLTs) have been proposed as a possible platform to address cloud big data provenance. This paper examines the W3C recommendation for data provenance PROV and if the blockchain DLT can apply the core primary PROV attributes required to satisfy data provenance. The research shows that not all data provenance expressions can be provided by blockchain. Instances of data provenance which rely on circular references are not possible as the blockchain DLT is a single linked list.

1 INTRODUCTION

Provenance is a well-established and well understood concept which seeks to establish the origin, lineage, history, transactions on, and ownership of, an artefact and has been applied in many domains, including art, antiquities, finance, and procurement, to name just a few, over many centuries.

Data Provenance applies the concept of provenance to the digital data domain. This has application in nearly all the current digital domains where data and content are being produced and transacted at an ever-increasing rate, but particularly in the Cloud based 'big data' domain. The importance of tracking provenance is widely recognized, as witnessed by significant research in various areas including: e-science (Janowicz et al, 2018), (Sigurjonsson, 2018); data warehousing (Hambolu et al, 2016); democratic decision making (Aragón et al, 2014), (Beris and Koubarakis, 2018); e-Health (Masi and Miladi, 2018); digital forensics (Ulybyshev et al, 2018), (Zawoad et al, 2018); security (Cha and Yeh, 2018); news checking (Huckle and White, 2017); and information theory (Lemieux, 2016), (Lemieux and Sporney, 2017), to name just a few. As Cloud based processing, storage and 'big data' has become ubiquitous, privacy concerns have become common to all these areas, raising the same problems that data provenance seeks to address: where did this data originate, what is its history and how can these be shown?

For ordinary people this has many specific use cases including identity theft, breach of copyright, digital anonymity, and the ability to see, and gain control over, how individuals' personal data is transacted, used or misused.

For organisations collecting and using personal data, the ability to organise, audit, and verify compliance with legislation such as Sarbanes-Oxley (Congress of the United States, 2002), Health Insurance Portability and Accountability (Congress of the United States, 1996), Gramm-Leach-Bliley Financial Services Modernization (Congress of the United States, 1999), are key requirements in business today. This creates new challenges to organisational strategies and the management of data provenance in their Cloud and 'big data' infrastructure and management.

The increased public awareness of the use and misuse of big data has raised many privacy concerns, particularly in opaque Cloud based technologies (Zou, 2016), (Pahl et al, 2018). These public concerns have resulted in the introduction of specific new legislative concepts and laws which seek to mitigate and address these issues, such as General Data Protection Regulation (GDPR) (European Commission, 2016), which seeks to not only regulate what data can be used and how it can be used, but also where it can be used. This along with the Payment

Hogan, G. and Helfert, M.
Transparent Cloud Privacy: Data Provenance Expression in Blockchain.
DOI: 10.5220/0007733404300436
In *Proceedings of the 9th International Conference on Cloud Computing and Services Science (CLOSER 2019)*, pages 430-436
ISBN: 978-989-758-365-0

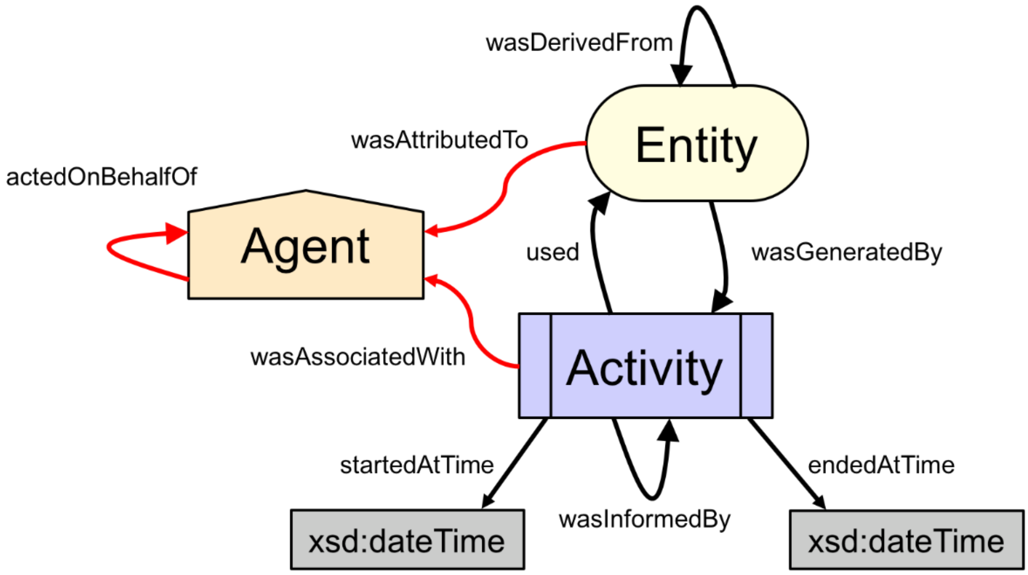

Figure 1: PROV Ontology Fundamental Classes and Relations.

Services Directive (PSD2) (European Commission, 2015), in Europe, which has also highlighted these privacy issues in other jurisdictions, presents specific challenges for privacy in the global Cloud.

The introduction of these new legal requirements have also brought challenges to regulatory bodies charged with the investigation, auditing and enforcement of these new laws. Current practices are manually intensive and linear expansion of human resources to address the ever increasing demand for their services are unsustainable. Existing provenance approaches are challenged when faced with Cloud based distributed systems, and with the volume, velocity, variety, variability, and veracity requirements of big data and (Wang et al, 2015). New strategies, new approaches and new capabilities are required to enable provenance strategy and management in the era of Cloud big data.

This paper is constructed as follows: Section 2 outlines the literature review and the methodology used to conduct the review; section 3 presents the background to the paper based on the review; section 4 presents the research question; section 5 presents the results; section 6 outlines the limitations and section 7 outlines the conclusion and future work.

2 SELECTION METHODOLOGY

The academic database sources AIS eLibrary, ACM Digital library, IEEE Xplore, Inspect Engineering Village, Science Direct, Web of Science, Wiley Online Library, along with Semantic Scholar, Google Scholar, grey and non-peer reviewed resources such as arXiv were searched using the following search criteria:
“(PROV-O OR W3C) AND (blockchain OR distributed ledger) AND (provenance OR lineage OR pedigree)”.

The queries to these databases returned over 350 publications of which 201 were non duplicate publications. Manually filtering the title and abstract for specific mentions of provenance or blockchain excluded 112 of these publications, leaving 89 publications to be full text screened. The manual full text screening, which examined the context relevance of each of these publication to the subject matter of this paper, further filtered the body of publications down to a review set of 27 papers with relevant contributions to the subject in question.

A search for citations of these 27 papers provided an additional set of publications which were reviewed in a forward search, examining publications that cited any of the review set for relevant contributions. This provided another 12 relevant publications.

In addition new search queries for each of the authors of the publications in the review set were carried out on the databases to see if any previous or subsequent publications by the authors contributed further to the subject of this paper. A review of the references and the authors previous publications provided additional 15 relevant papers after repeating the screening protocol above.

3 BACKGROUND

A number of different approaches and recommendations for provenance are represented in the literature, including: the Open Provenance Model (OPM) (Moreau et al, 2011), the Dublin Core Metadata Initiative (Dublin Core Metadata Initiative, 2014), both of which culminated in the W3C Provenance Recommendation (World Wide Web Consortium, 2013a, 2013b). The W3C Working Group on Provenance provide the following definition "Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness". In addition, they specify a standard model, PROV-DM (World Wide Web Consortium, 2013a) with documentation for supporting data provenance, and a provenance ontology PROV-O (World Wide Web Consortium, 2013b), Figure 1.

Other approaches recognise that provenance is an aspect of information theory (Lemieux, 2016), specifically in the context of record keeping and the auditing of these records. A suite of ISO/IEC standards also addresses provenance in the context of information management including: security techniques; privacy; records management; audit and certification; information security management; and information technology in business (International Standards Organisation, 2011, 2012, 2013, 2015, 2016). In short there are many interpretations of, and perspectives of provenance due to the contextual nature of provenance.

Cloud based Distributed Ledger Technologies (DLTs) emerged from a number of incremental innovations in the concept of digital and electronic money, the development of blockchain, and distributed Cloud technology, which underpinned the invention of Bitcoin (Nakamoto, 2008). DLTs potentially offer an alternative model for enabling data provenance in both the digital and physical realms. The DLT domain has evolved to include further concepts, addressing: standards for privacy and trust (Anjum et al, 2017); public, private, permissionless or permissioned DLTs (El Ioini and Pahl, 2018); hybrid DLTs (Lemieux, 2016); and federated DLTs (Wood, 2013). In addition there are several competing implementations of DLTs including Blockchain (through which Bitcoin is implemented), Ethereum (Wood, 2013), Hyperledger, ConsenSys, and R3, though this list is not exhaustive and new implementations are being released on an ongoing basis with different application areas and capabilities.

Additionally the application of smart contracts which allows for the implementation of logic through the execution of code in DLTs has enabled new applications and scenarios for both physical and digital provenance (Fotiou and Polyzos, 2018). The relative recentness of DLT and its rapid rate of development have outpaced the academic community's study of the topic, particularly with reference to the strategies and management of data provenance.

Multiple individual instances, use cases and examples of provenance capable blockchains have been proposed including SPADE (Gehani and Dawood, 2012), SECPROV (Zawoad et al, 2018), health data provenance (Masi and Miladi, 2018), jewellery provenance (Orenge, 2018) which illustrate the topical conversation on this topic and the current

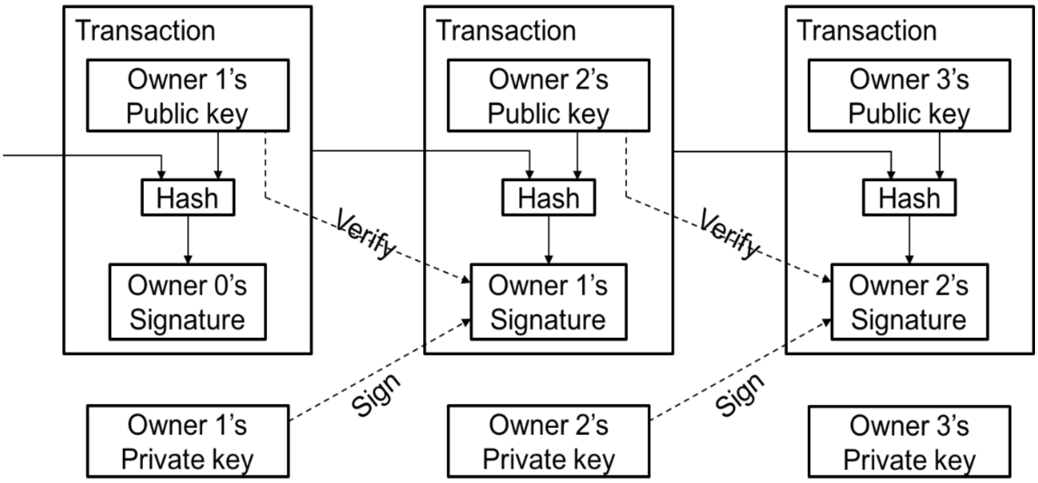

Figure 2: Fundamental concepts of the Bitcoin Blockchain.

efforts being made to express provenance through the Blockchain DLT.

4 THE RESEARCH QUESTION

The background literature shows data provenance being expressed for specific use cases in Cloud based Blockchains but do not address if data provenance can be expressed for all use cases in Blockchain.

The research approach is to examine a specialised case by using the Blockchain DLT and questions the degree to which this most adopted implementation, Blockchain aligns with the W3C PROV recommendation for data provenance (Problem 1).

4.1 Problem 1: Alignment of Blockchain with PROV

The provenance (PROV) family of documents specify the W3C recommendations for provenance information on the world wide web. The PROV-DM data model and PROV-O ontology (World Wide Web Consortium, 2013a, 2013b) outline the classes, properties and relationships between these that are used to describe a provenance instance.

Table 1: PROV-DM types and relations.

PROV Concepts	PROV-DM types or relations	Name
Entity	PROV-DM Types	Entity
Activity		Activity
Agent		Agent
Generation	PROV-DM Relations	wasGeneratedBy
Usage		used
Communication		wasInformedBy
Derivation		wasDerivedFrom
Attribution		wasAttributedTo
Association		wasAssociatedWith
Delegation		actedOnBehalfOf

The primary classes of Agent, Entity and Activity and their relationships, shown in Figure 1, form the core concept which allows provenance to be described. PROV-O provides the following definitions for each of the fundamental classes:

- "An prov:Entity is a physical, digital, conceptual, or other kind of thing with some fixed aspects; entities may be real or imaginary."
- "An prov:Activity is something that occurs over a period of time and acts upon or with entities; it may include consuming, processing, transforming, modifying, relocating, using, or generating entities."
- "An prov:Agent is something that bears some form of responsibility for an activity taking place, for the existence of an entity, or for another agent's activity."

Four basic types of technologies for DLT are identified: blockchain; tangle; hashgraph; and sidechain, with underlying technologies of a) linked-list/ list of linked-list or b) directed acyclic graph, with blockchain described as a linked-list (El Ioini and Pahl, 2018). The immutable quality of blockchain restricts the linked list type that blockchain can be, to a singly linked list (Scriber, 2018), or single backward-linked list (van den Hooff et al, 2014).

The Bitcoin implementation of Blockchain (Nakamoto, 2008) is outlined in Figure 2 showing its fundamental concepts and relationships. Its main concepts are Transactions, Owners and Hashes. The Bitcoin 'thing' is defined as a "chain of digital signatures". Coin transfers are made by means of transactions which are initiated and executed by, and between Owners.

In this specific case, for the single linked list Blockchain implementation of Bitcoin to provide data provenance there should be alignment between the Blockchain fundamental concepts and the core PROV-DM types and relations.

5 RESULTS

In naive terms, there is a simple alignment between the PROV data model (PROV-DM) fundamental classes, shown in Table 1 and the Bitcoin implementation of blockchain. At the highest level the core PROV-DM (prov:) type align logically to a Blockchain(bc:) equivalent:

- bc:Transaction is equivalent to a prov:Activity and has a temporal start and end which are equivalent to prov:startedAtTime and prov:endedAtTime respectively.
- bc:Owner is equivalent to a prov:Agent both as the initiator, and the receiver, of a transaction;
- bc:Hash (chain of hashes) is equivalent to a prov:Entity as the object of a transaction

In each of the cases above it can be stated that blockchain satisfies the basic PROV-DM types.

Looking at the PROV-DM relations between their origin and endpoint:

Table 2: Alignment of PROV and Blockchain types and concepts.

Core PROV-DM types and relations	Name	Blockchain type
PROV-DM Types	Entity	Hash (Chain)
	Activity	Transaction
	Agent	Owner
PROV-DM Relations	WasGeneratedBy	Verify, Sign, (Signature), HashFunction
	Used	Verify, Sign, (Signature), HashFunction
	WasInformedBy	-
	WasDerivedFrom	-
	WasAttributedTo	Verify, publicKey, (Signature), privateKey
	WasAssociatedWith	Verify, Sign, (Signature), publicKey, HashFunction
	ActedOnBehalfOf	-
xsd:dateTime	startedAtTime,endedAtTime	TimeStamp

- prov:Entity[wasGeneratedBy]Activity is equivalent to bc:Hash[Verify, Sign, (Signature), HashFunction]Transaction as these form the relationships between the new Hash and the previous transaction.
- prov:Activity[used]Entity is equivalent to bc:Tranaction[Verify, Sign, (Signature), HashFunction]Hash as these form the relationships between the new Transaction and the previous Hash.
- prov:Activity[wasInformedBy]Activity has no equivalent. There is no Blockchain Transaction which allows another Transaction to be created without the involvement of a Hash (Entity) and Owner (Agent).
- prov:Entity[wasDerivedFrom]Entity has no equivalent. There is no Blockchain Hash (other than the initial creation of the first Hash) which allows another Hash to be created without the involvement of a Transaction (Activity) and Owner (Agent).
- prov:Entity[wasAttributedTo]Agent is equivalent to bc:Hash[Verify, publicKey, (Signature), privateKey]Owner as these form the relationships between the Hash (Entity) and the Owner/s (Agent).
- prov:Activity[wasAssociatedWith]Agent is equivalent to bc:Tranaction[Verify, Sign, (Signature), publicKey, HashFunction]Owner as these form the relationships between the Transaction (Activity) and the Owner/s (Agent).
- prov:Agent[actedOnBehalfOf]Agent has no equivalent. As the Blockchain depends on unique identity in the cryptography of the hash, there is no Blockchain Agent which allows another 'proxy' Agent to act on the Owners behalf.

The full mapping of the Core PROV-DM types to the equivalent blockchain type is shown in Table 2. Each of the PROV-DM relations which have no blockchain equivalent are circular (self-referring) references to a single PROV-DM type for which there is no equivalent in a single linked list blockchain. Data provenance instances which reply on these circular reference relations are not implementable in a single linked list DLT such as blockchain.

6 LIMITATIONS

The findings of this paper are limited in scope to the Blockchain specific instance of Distributed Ledger Technology. In addition it is further confined to the Bitcoin implementation on Blockchain. Other altchain instances of Distributed Ledger Technologies such as Hashgraph, Tangle, and Sidechain, or alternative implementations of of Distributed Ledger Technologies such as Ethereum and Hyperledger are not considered.

In addition, only the core classes and concepts of PROV is considered. The extended classes and relationships of PROV-DM and the PROV-O ontology are more complex and are outside the scope of this paper.

7 CONCLUSIONS AND FUTURE WORK

In this paper we showed that there are equivalencies in Blockchain to the core PROV-DM types - Agent, Activity and Entity, namely Owner, Transaction and Hash respectively. This paper also showed that there are instances of the core PROV-DM relationships which have equivalent Blockchain relationships.

We conclude that it is possible to express a subset of the total instances of data provenance in the Blockchain using the Blockchain equivalents of the PROV-DM core types and equivalent relationships.

This paper also showed that there are instances of the core PROV-DM relationships which have no equivalent Blockchain relationships, namely prov:wasInformedBy, prov:wasDerivedFrom and prov:actedOnBehalfOf, as they self-refer, or loop back, to a single class without the involvement of another class, which Blockchain as a single linked list does not support.

It is reasonable to propose that any data provenance instances which rely on these PROV relationships may not be expressed in Blockchain.

Future steps include further examination of the PROV model and investigation of possible extensions that may facilitate closer alignment of blockchain to PROV. Other flavours of distributed ledgers such as directed acyclic graph based Tangle or Hashgraph may provide different results to Blockchain should also be investigated.

ACKNOWLEDGEMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) and is co-funded under the European Regional Development Fund under Grant Number 13/RC/2106.

REFERENCES

Anjum, A., Sporny, M., Sill, A., 2017. Blockchain Standards for Compliance and Trust. In: *IEEE Cloud Computing,* 4(4):84-90.

Aragón, P., Bria, F., de Filippi, P., Halpin, H., Korhonen, J., 2014. D-CENT D4.3 Technical Design of Open Social Web for Crowdsourced Democracy. [Available online: https://dcentproject.eu/wp-content/uploads/2014/10/D4.3-final.pdf]

Beris, T., Koubarakis, M., 2018. Modeling and Preserving Greek Government Decisions Using Semantic Web Technologies and Permissionless Blockchains. In *ESWC 2018, The Semantic Web - 15th International Conference*, pages 81-96. Springer.

Cha, S., Yeh, K., 2018, An ISO/IEC 15408-2 Compliant Security Auditing System with Blockchain Technology. In *CNS'18, IEEE Conference on Communications and Network Security*, pages 1-2. IEEE

Congress of the United States (1996). *Health Insurance Portability and Accountability Act, Public Law No: 104-191*. Washington D.C.

Congress of the United States (1999). *Financial Services Modernization Act, Public Law No: 106-102*. Washington D.C.

Congress of the United States (2002). *Sarbanes-Oxley Act, Public Law No: 107-204*. Washington D.C.

Dublin Core Metadata Initiative 2014, *DCMI Metadata Provenance Task Group*. [Available online: http://dublincore.org/groups/provenance/]

El Ioini, N., Pahl, C., 2018. A Review of Distributed Ledger Technologies. In *On the Move to Meaningful Internet Systems, Proceedings of OTM 2018 Conferences*, pages 277-288. Springer

European Commission (2015) *Payment Services Directive.* Luxembourg: Office for Official Publications of the European Communities.

European Commission (2016) *General Data Protection Regulation.* Luxembourg: Office for Official Publications of the European Communities.

Fotiou, N,. Polyzos, G.C., 2018. Smart Contracts for the Internet of Things: Opportunities and Challenges. In *EuCNC'2018, European Conference on Networks and Communications,* pages 256-260. IEEE.

Gehani, A., Dawood, T., 2012. SPADE: Support for Provenance Auditing in Distributed Environments. In *Middleware'12, Proceedings of the 13th International Middleware Conference*, pages 101-120. Springer

Hambolu, O., Yu, L., Oakley, J., Brooks, R.R., Mukhopadhyay, U., Skjellum, A., 2016. Provenance threat modelling. In *PST'14, 14th Annual Conference on Privacy, Security and Trust,* pages 384-387. IEEE

Huckle, S. and White, M., 2017. Fake News: A Technological Approach to Proving the Origins of Content, Using Blockchains. *Big Data,* 5(4), pages 356–371. Mary Ann Liebert

International Standards Organisation 2011, *Information technology -- Security techniques -- Privacy framework,* ISO/IEC 29100:2011, International Standards Organisation, Geneva.

International Standards Organisation 2012, *Space data and information transfer systems -- Audit and certification of trustworthy digital repositories,* ISO 16363:2012, International Standards Organisation, Geneva.

International Standards Organisation 2013, *Information technology -- Security techniques -- Information security management systems – Requirements,* ISO/IEC 27001:2013, International Standards Organisation, Geneva.

International Standards Organisation 2015, *Information technology -- Business operational view -- Part 4:*

Business transaction scenarios -- Accounting and economic ontology, ISO/IEC 15944-4:2015, International Standards Organisation, Geneva.

International Standards Organisation 2016, *Information and documentation -- Records management -- Part 1: Concepts and principles,* ISO 15489-1:2016, International Standards Organisation, Geneva.

Janowicz, K., Hitzler, P., Regalia, B., Mai, G., Delbecque, S., Frohlich, M., Martinent, P., Lazarus, T., 2018. On the prospects of blockchain and distributed ledger technologies for open science and academic publishing. In *Semantic Web, 9(5),* pages 545-555. IOS Press

Lemieux, V.L., 2016. Trusting records: is Blockchain technology the answer? In *Records Management Journal, 26(2),* pages 110-139. Emerald.

Lemieux, V.L., Sporny, M., 2017. Preserving the Archival Bond in Distributed Ledgers: A Data Model and Syntax. *WWW '17, Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1437-1443. ACM

Masi, M., Miladi, A., 2018. *Using PROV and Blockchain to Achieve Health Data Provenance*. Working Paper, University of Southampton Institutional Repository. [Available online: https://eprints.soton.ac.uk/421292/]

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., Van den Bussche, J., 2011. The Open Provenance Model core specification (v1.1). In *Future Generation Computer Systems, 27(6),* pages 743-756. Elsevier.

Nakamoto, S., 2008. *Bitcoin: A Peer-to-Peer Electronic Cash System*. [Available online: http://bitcoin.org/bitcoin.pdf]

Orenge, A.O., 2018. *Blockchain-based Provenance Solution for Handcrafted Jewellery*. Masters Thesis. University of Tartu.

Pahl, C., Ioini, N.E., Helmer, S., Lee, B., 2018. An architecture pattern for trusted orchestration in IoT edge clouds. In *FMEC'2018, Third International Conference on Fog and Mobile Edge Computing,* pages 63-70. IEEE.

Scriber, B. A., A Framework for Determining Blockchain Applicability, in *IEEE Software, 35(4):*70-77. IEEE.

Sigurjonsson, S.M.K., 2018. *Blockchain Use for Data Provenance in Scientific Workflow*. Master Thesis. KTH Royal Institute of Technology.

Ulybyshev, D., Villarreal-Vasquez, M., Bhargava, B., Mani, G., Seaberg, S., Conoval, P., Pike, R., Kobes, J., 2018. (WIP) Blockhub: Blockchain-Based Software Development System for Untrusted Environments. In *IEEE 11th International Conference on Cloud Computing,* pages 582-585. IEEE.

van den Hooff, J., Kaashoek, M.F., Zeldovich, M., 2014. VerSum: Verifiable Computations over Large Public Logs. In *CCS'14, Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security,* pages 1304-1316. ACM.

Wang, J., Crawl, D., Purawat, S., Nguyen, M.H., Altintas, I., 2015. Big data provenance: Challenges, state of the art and opportunities. In *Big Data'2015, IEEE International Conference on Big Data,* pages 2509-2516. IEEE

Wood, G., 2013. *Ethereum: A Secure Decentralised Generalised Transaction Ledger*. [Available online: https://gavwood.com/paper.pdf]

World Wide Web Consortium 2013a, *PROV-DM: The PROV Data Model*, W3C Recommendation, World Wide Web Consortium, Geneva.

World Wide Web Consortium 2013b, *PROV-O: The PROV Ontology,* W3C Recommendation, World Wide Web Consortium, Geneva.

Zawoad, S., Hasan, R., Islam, M.K., 2018. SECProv: Trustworthy and Efficient Provenance Management in the Cloud. In *INFOCOM'2018, IEEE Conference on Computer Communications:* pages 1241-1249. IEEE

Zou, J., 2016. *Accountability in cloud services*. PhD Thesis. Macquarie University, Sydney.