Open Algorithms as Smart Contracts: Enabling Future Data Markets using Blockchain Technology

Thomas Hardjono, Keeley Erhardt and Alex Pentland (MIT Media Lab)

Introduction

Data is crucial for the proper functioning of the current and future digital society. It is fundamental to the day-to-day running of societies, governments and businesses. Furthermore, data increases in value when combined cross domains and verticals, and previously unconsidered insights can be obtained when data is combined (Pentland, et. al, 2013). However, the current reality is that cross-organizational data sharing is prohibitive both from the regulatory perspective and from the business risk perspective.

One of the key findings of the 2011 World Economic Forum (WEF) report on personal data is that the current ecosystems that access and use personal data is fragmented and inefficient. For many participants, the risks and liabilities exceed the economic returns. Furthermore, personal privacy concerns are inadequately addressed. Current technologies and laws fall short of providing the legal and technical infrastructure needed to support a well-functioning digital economy.

The MIT Open Algorithms (OPAL) project offers an alternative paradigm for information sharing based on a number of key principles (Pentland, et. al 2016). The first principle is that of *moving the algorithm to the data*, which means algorithms (i.e. queries) are sent to the location of the relevant data (i.e. "data-repository") and be executed there. This is in contrast to the current approach of collecting or collating raw data into one massive repository and performing analytics there, an approach which works only if the collator has legal access to all needed data (typically from one vertical or industry).

The second principle is that *raw data must never leave its (distributed) repository* and must always be under the control of its owner or the owner of the data repository. This principle is consistent with the privacy requirements of the GDPR (European Commission, 2016). Advanced encryption techniques such as homomorphic encryption and multi-party secure computation (Zyskind et. al, 2015) can then be applied to data in their repositories.

Thirdly, *only vetted algorithms should be executed by the data repository*. The term "vetted" means that algorithms must be studied and vetted by domain-experts to be "safe" from bias, discrimination, privacy violations and other unintended consequences.

These combined principles lend itself to the establishment of industry-based *data sharing consortiums*, where a group of data providers (data owners) across verticals or sectors *can share results without sharing raw data*. Members of a data sharing consortium must collectively write/author and vet algorithms which are acceptance by all members and which they agree to execute within their own backend data repositories (behind their firewalls).

The concept for Open Algorithms (OPAL) emerged from several MIT research projects over the past decade which focused the future data-driven society (Pentland et. al, 2015). It was increasingly becoming evident that an individual's privacy could be affected through correlation using only a small data set (de Montjoye, et. al 2013) and a new model for *user-centric personal data stores* to be developed and championed in industry.

In this paper we extend the OPAL concepts by expressing vetted algorithms as smart contracts to be executed by nodes within the peer-to-peer (P2P) net-work of nodes within a blockchain system or an implementation of a distributed ledger technology (DLT):

- *Vetted algorithms expressed as smart-contracts*: Once an algorithm has been vetted to run against a given data-set, it is expressed as a combined code and legal-prose, digitally signed and distributed on the nodes of the blockchain system. Depending on the type of blockchain system (permissioned, permissionless, semi-anonymous) the algorithm itself may be publicly readable. The combined code and legal-prose is also referred to in legal circles as "dual integration".

- *Identity-based invocation of smart-contract*: An entity (organization or individual) must present sufficient identity information as part of the smart-contract invocation. The degree of identity revelation will depend on the type and use-case of the OPAL smart contract model.
- *Escrowed payments*: The querier (caller) must invoke the smart-contract algorithm accompanied with payment, which will be escrowed until the completion of the execution of the algorithm upon the intended data-set.
- *Identity-based response encryption*: The node on the P2P belonging to the data repository (referred to as the data node) has the option to encrypt the response for the querier (caller), using the public-key supplied during the identity-based invocation of the smart-contract algorithm.
- *Smart-contract legal prose:* The "dual integration" of executable code and legal prose into a compact smart-contract allows the data repositories (data owners) to state a Terms of Use for the resulting response for privacy and regulatory compliance purposes. This includes prohibiting the querier from collating too many responses from multiple data repositories -- with the goal of re-identifying subjects (e.g. individuals) whose personal data maybe part of the data-set.
- *Personal Data Nodes:* The architecture should be independent of the size of the data-set within the data repositories (data nodes). This allows personal data stores (e.g. PDS) to be participated by individuals (owners) as legitimate OPAL-based endpoints on the blockchain.
- *Data sharing coins:* New forms of digital currencies ("coins") can be created on the blockchain based on the value of the data-sets within the P2P network, and can therefore incentivize data owners and node operators to collaborate on the blockcian.

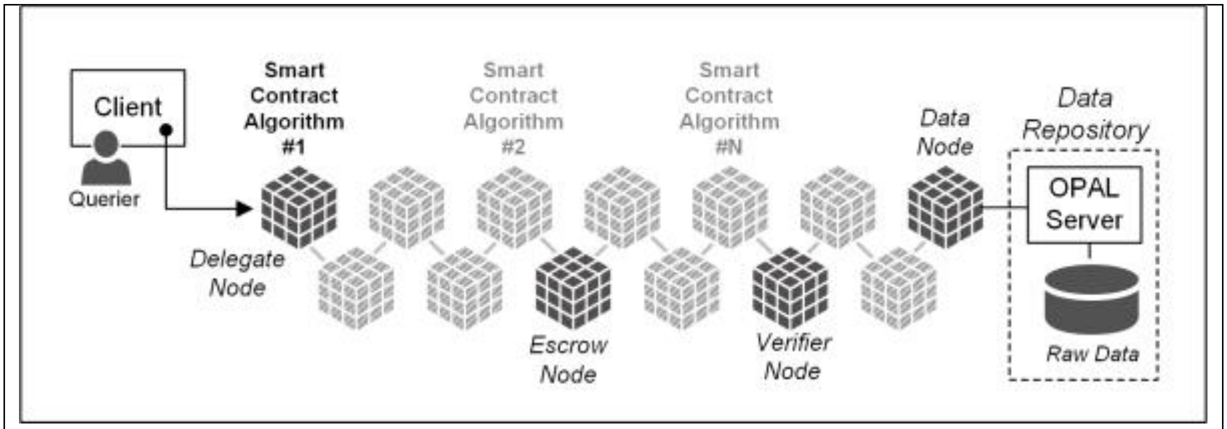

Figure 1. OPAL Smart Contracts - Overview

There are a number of motivating reasons for expressing OPAL algorithms as smart contracts:

- *New model for data markets*: The use of smart contracts to express vetted algorithms in combination with the P2P network of nodes in a blockchain system allows queriers to "meet" data providers through the network. Nodes of the P2P network can make available a variety of OPAL smart contracts, where each contract reflects a specific algorithm intended to be executed by a specific data provider. Queriers are able to "scan" through these nodes (e.g. using bots) to search for OPAL vetted algorithms that suit their current need. The fact that vetting has been performed by the data provider and/or a data-consortium or federation provides a strong degree of assurance to the querier regarding the origins of the algorithm.
- *New revenue source for data repositories*: Data repositories who process incoming queries (vetted algorithms) obtain guaranteed remuneration through the payment attached to the smart

contract. In many cases, a data provider may not care who the identity of the querier may be, so long as payment has been made (e.g. via an escrow entity).

- *Subscription model*: For data-sets that accumulate over time and grow in value, a data provider may offer a subscription service in which an algorithm is executed at regular intervals (e.g. daily).
- *Algorithm compositions*: By populating nodes of the P2P network with OPAL smart contracts, queriers have a broader choice of vetted algorithms from which to choose. They are then able to compose a set of queries – possibly across different data domains/types – and obtain interesting insights into these domains.
- *Smart-contract legal prose*: The Terms of Use of the response in the legal prose provides additional legal protection against misuse of information by queriers.
- *Decentralized algorithms for personal data stores*: Individuals can get together as a unified group or community and define (and vet) the algorithms which they collectively feel to be suitable for their personal data stores (i.e. preserve their privacy). This will be relevant in the case of consumer IoT data (e.g. household data generated by smart appliances and home sensors).

The MIT Open Algorithms (OPAL) Architecture

The MIT OPAL architecture seeks to implement the OPAL principles, effectively moving the computation of algorithms to the location of the data. Se Figure 2.

The *Data Provider* is the entity that holds the raw data in its data repository. The *Querier* is the entity seeking to obtain information from the data provider's data set. The data provider publishes one or more *vetted algorithms* that are specific to its data repository. A group of *data providers* may collaborate to establish a Data Federation in the form of a consortium that agrees on the set of algorithms which they collectively vet and share.

Before requesting the execution of a given algorithm at its intended data provider, the querier must first select the algorithm from the published list and then "parameterize" it with the relevant input values. The querier then digitally signs the request and sends it to the data provider (optionally with some form of payment).

Upon receiving the signed request, the data provider must authenticate the request by validating the Querier's signature. The data provider must then verify that the indicated algorithm (or algorithm-identifier) is one that it recognizes from the published list of vetted algorithms.

After executing the algorithm (i.e. against its data set) the data provider has the opportunity to further filter or refine the response to ensure the information in the response conforms to the policy of the data provider (e.g. privacy policies, corporate rules, regulatory, etc.). Here there is opportunity for a data provider to apply advanced techniques of analysis, such as using AI, machine learning and other approaches to ensure the safety of the response. The point of this step is to provide a data provider with control over what it returns as a response.

Several "Living Labs" are currently being established for Proof of Concepts (PoC) using the OPAL paradigm and MIT software, for specific data-domains, uses-cases and countries (OPALproject, 2017).

OPAL Smart Contracts

Smart contracts offers an attractive means to represent algorithms in the context of information sharing across organizations. The following are a summary of the relevant fields of an OPAL smart contract, independent of the specific blockchain implementation and smart contracts expression language:

- *Algorithm-ID*: This is the Algorithm-ID as known in the OPAL ecosystem.
- *OPAL Web Contract location*: This is the location of the full OPAL Web Contract document. This value may be a hash when used with hash-based resolver systems (e.g. BNS, IPFS, etc.).
- *Querier address*: This is the blockchain address or public-key of the querier entity.

- *Data provider address*: This is the blockchain address or public-key of the data provider who is being asked to execute the algorithm (Algorithm-ID).
- *Payment*: This is proof that payment has been made to an escrow node.

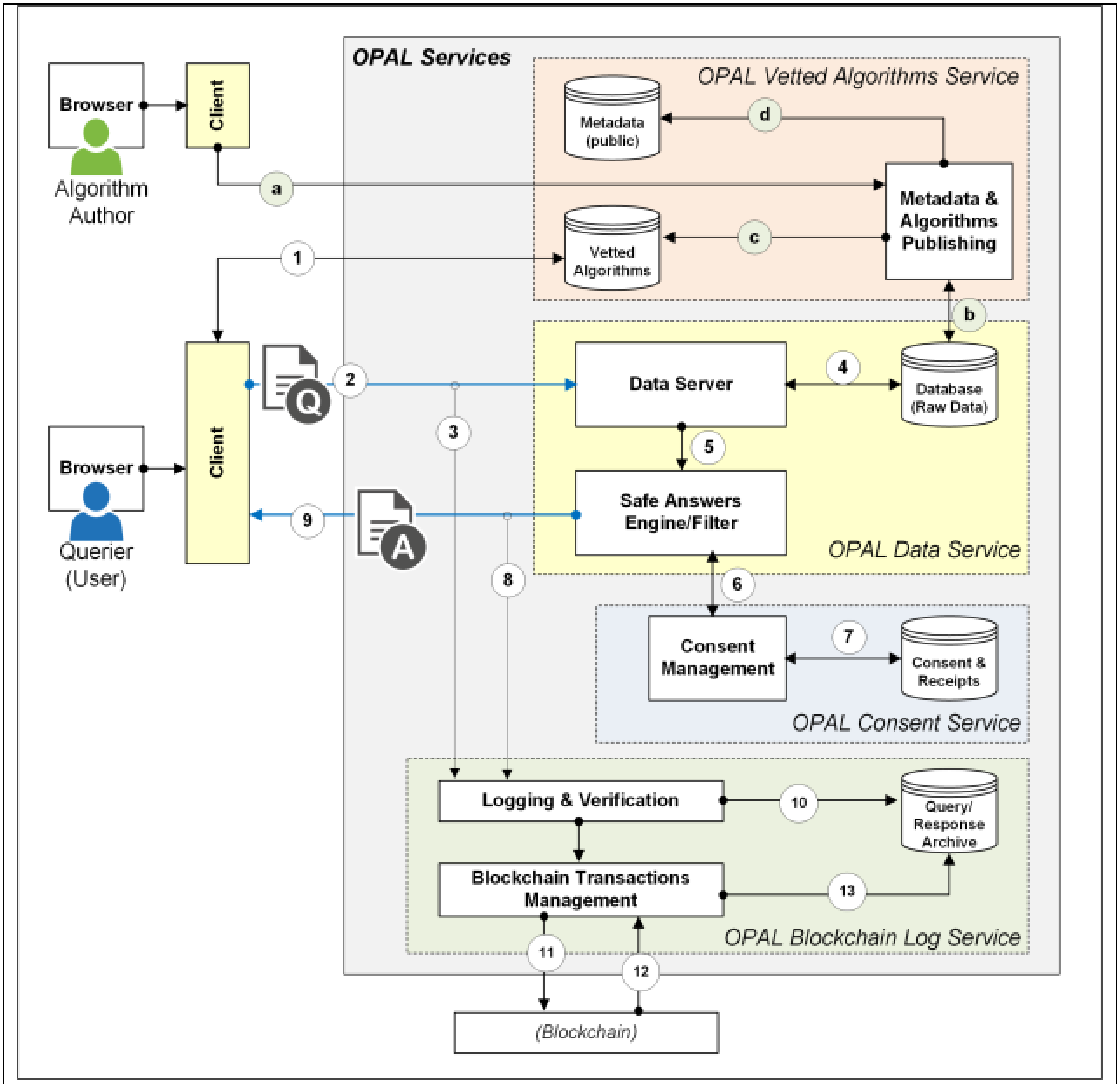

Figure 2. Overview of the MIT OPAL Architecture and Functions

References

Buterin, V. 2014.A "Next-Generation Cryptocurrency and Decentralized Application Platform". *Bitcoin Magazine*. (https://bitcoinmagazine.com/articles/ethereum-next-generationcryptocurrency-decentralized-application-platform-1390528211/)

de Montjoye, Y. A., Quoidbach, J., Robic, F., & Pentland, A. 2013. "Predicting personality using novel mobile phone-based metrics". In *Social computing, behavioral-cultural modeling and prediction* (LCNS Vol. 7812) (p. 48-55). Springer.

de Montjoye, Y. A., Shmueli, E., Wang, S., & Pentland, A. 2014. "Open-PDS: Protecting the Privacy of Metadata through SafeAnswers". *PLoS ONE* 9(7), 13-18. (https://doi.org/10.1371/journal.pone.0098790)

European Commission. 2016. "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)". *Official Journal of the European Union*, L119 , 1-88.

Hardjono, T., & Maler, E. 2017. *Blockchain and Smart Contracts Report*. Kantara Initiative.(https://kantarainitiative.org/confluence/display/BSC/Home)

Lizar, M., & Turner, D. 2017. *Consent Receipt Specification Version 1.0*, Kantara Published Specification. (https://kantarainitiative.org/confluence/display/infosharing/Home)

Nakamoto, S. 2011. *Bitcoin: A Peer-to-Peer Electronic Cash System*. Retrieved from https://bitcoin.org/bitcoin.pdf

Norton Rose Fulbright. 2016. *Can smart contracts be legally binding contracts* (Report). Norton Rose Fulbright. (http://www.nortonrosefulbright.com/knowledge/publications/144559/can- smart-contracts-be-legally-binding-contracts)

OPAL Project, 2017. The OPAL Project. http://www.opalproject.org

Pentland, A. 2014. "Saving Big Data from Itself". *Scientific American*, pp65-68.

Pentland, A. 2015. *Social Physics: How Social Networks Can Make Us Smarter*. Penguin Books.

Pentland, A., Reid, T., & Heibeck, T. 2013. *Big Data and Health - Revolutionizing Medicine and Public Health: Report of the Big Data and Health Working Group 2013* . World Innovation Summit for Health, Qatar Foundation. (http://www.wish-qatar.org/app/media/382)

Pentland, A., Shrier, D., Hardjono, T., & Wladawsky-Berger, I. 2016. "Towards an Internet of Trusted Data: Input to the Whitehouse Commission on Enhancing National Cybersecurity". In T. Hardjono, A. Pentland, & D. Shrier (Eds.), *Trust::Data - A New Framework for Identity and Data Sharing* (p. 21-49). Visionary Future.

Szabo, N. 199. *Smart Contracts: Building Blocks for Digital Markets*.

World Economic Forum. 2011. *Personal Data: The Emergence of a New Asset Class*. (http://www.weforum.org/reports/ personal-data-emergence-new- asset-class)

Zyskind, G., Nathan, O., & Pentland, A. 2015. "Decentralizing privacy: Using blockchain to protect personal data". In *Proceedings of the 2015 IEEE Security and Privacy workshops*. IEEE.