Journal of Database Management
Volume 32 • Issue 1 • January-March 2021

Privacy Preserving Data Mining as Proof of Useful Work:
Exploring an AI/Blockchain Design

Hjalmar K. Turesson, York University, Canada
Henry Kim, blockchain.lab, York University, Canada
Marek Laskowski, blockchain.lab, York University, Canada
Alexandra Roatis, Aion Network, Canada

ABSTRACT

Blockchains rely on a consensus among participants to achieve decentralization and security. However, reaching consensus in an online, digital world where identities are not tied to physical users is a challenging problem. Proof-of-work provides a solution by linking representation to a valuable, physical resource. While this has worked well, it uses a tremendous amount of specialized hardware and energy, with no utility beyond blockchain security. Here, the authors propose an alternative consensus scheme that directs the computational resources to the optimization of machine learning (ML) models – a task with more general utility. This is achieved by a hybrid consensus scheme relying on three parties: data providers, miners, and a committee. The data provider makes data available and provides payment in return for the best model, miners compete about the payment and access to the committee by producing ML optimized models, and the committee controls the ML competition.

KEYWORDS

Blockchain Mining, Distributed Systems, Machine Intelligence, Predictive Analytics, Proof-of-Useful-Work

INTRODUCTION

Bitcoin (Nakamoto, 2009) presented a workable solution to the problem of double spending of electronic cash without a controlling central entity such as a bank. Launched in 2009, the Bitcoin network implements a peer-to-peer network of computers that maintains a distributed ledger, tracking all the network participants' cryptocurrency balances. In an open network of pseudonymous participants, reaching consensus about what transactions to include in the ledger is challenging – a simple voting scheme won't work since an individual can get an unfair influence by pretending to be an arbitrarily large number of individuals in a "Sybil attack" (Douceur, 2002). For Bitcoin, Sybil-resistance was achieved by requiring participants to expend real-world resources for a chance to append new transactions to the ledger, a scheme known as Proof-of-Work (PoW) (Back, 2002; Nakamoto, 2009; Dwork & Naor, 1993).

DOI: 10.4018/JDM.2021010104

BACKGROUND

PoW "proves" that the important task of appending the next block to the ledger is given to someone – a miner – who is "rich" enough that they cannot be corrupted to tolerate a Sybil attack. Wealth is proxied by the miner's access to resources; for Bitcoin, that is the abundant amount of electricity and computational resources required to solve a very difficult mathematical puzzle before others do. However, for every block, it follows that the vast amounts of energy expended by the winning miner and the numerous losing miners are wasted (Vries, 2018; O'Dwyer & Malone 2014; Budish 2018). There have been some attempts at ameliorating this shortcoming by instead securing the blockchain with useful work via a Proof-of-Useful-Work (PoUW) scheme. PoW requires miners to collective expend vast computational resources to solve a mathematical problem whose solution has no other purpose. PoUW entails solving a mathematical problem whose solution is useful to a third-party external to the blockchain. Early examples were Primecoin (King, 2013), where the work required was to search for chains of prime numbers, and Permacoin (Miller et al., 2014), intended to direct mining resources to distributed storage of archival data. However, these efforts have failed to reach wide adoption possibly due to the limited utility of the work performed. More recent efforts have attempted to solve the orthogonal vectors problem useful for graph theory analysis (Ball et al., 2017) or perform computational tasks for executing Software Guard eXtensions (SGX) instructions on Intel chips (Zhang et al., 2017).

Here we take a different approach and focus on a specific, but common, task: privacy-preserving data mining. Our approach also results in work towards providing a dual-purpose scheme that is useful for domains of blockchain (consensus mechanism) and AI (data mining).

WHY PRIVACY-PRESERVING DATA MINING

The application of machine learning (ML) to important problems in medicine and finance often results in an apparent contradiction: Training the models requires access to large and varied data sets under industry or regulatory expectation that security and privacy will be preserved, even though the size and scope of the data collected makes it attractive to hackers and increases likelihood of malicious or even unintended privacy breaches. Recent news reports have highlighted data security and privacy failures (Armeding, 2018; Cameron, 2017; Subramanian & Malladi, 2020). To mitigate this seeming contradiction and limit data leaks, a popular scheme obfuscates the raw data and applies machine learning on the transformed data, enabling data-driven discovery ("mining") of insights while ensuring that the data remain private. This scheme which preserves privacy yet maintains data utility and modeling accuracy is called privacy-preserving data mining (Thuraisingham, 2005).

Given the popularity of AI (Siau & Wang, 2020; Wang & Siau, 2019), it is attractive to conceptualize a blockchain's proof-of-work mathematical problem as a data mining problem. However, proof-of-work is most compelling for blockchain use cases in which the proof of access to resources is a proxy for proof of incorruptibility amongst untrusted potential validators (Nakamoto, 2018). Bitcoin and Ethereum are blockchain networks that exemplify this "trustless," "permissionless" context. Clearly, raw, un-obfuscated data cannot be provided to third party validators (cryptocurrency miners) to do data mining on such open blockchains; miners may be trusted to do transparent, straightforward validation, but they cannot be trusted with raw data. Hence, our PoUW solves a privacy-preserving data mining problem, not a generic data mining problem using raw data.

Given a training data set of numerical features and a categorical or continuous target (output variable) associated with each example, the PoUW task can be set up as an ML competition, where miners compete to predict the targets given some inputs. The miner that best predicts the test targets wins. A standard ML competition relies on a trusted party, the organizer, who has full access to the data set and withholds a subset of the targets from the competitors. The organizer releases two sets of data to the competitors: a complete set of inputs-target pairs for model training (the training data)