Personal Archive Service System using Blockchain Technology: Case Study, Promising and Challenging

by

Yixuan Zhu

yzhu@mercymavericks.com

A thesis submitted to the Faculty in the

Department of Mathematics and Computer Sciences

in partial fulfillment of the requirements for the degree of

Master of Science in Cybersecurity

Mercy College

Dec 2017

Thesis Advisor

Professor Chen Zhixiong

Department of Mathematics and Computer Sciences

School of Liberal Arts

Mercy College

Abstract

Key words – *blockchain*

I. INTRODUCTION

Blockchain is a continuously growing list of records, called blocks, which are linked and secured using cryptography. Each block typically contains a hash pointer as a link to a previous block, a timestamp and transaction data. By design, blockchains are inherently resistant to modification of the data. A blockchain is "an open, distributed ledger that can record transactions between two parties efficiently and in a verifiable and permanent way." For use as a distributed ledger, a blockchain is typically managed by a peer-to-peer network collectively adhering to a protocol for validating new blocks. Once recorded, the data in any given block cannot be altered retroactively without the alteration of all subsequent blocks, which requires collusion of the network majority.

The simplest and most popular application based on blockchain is digital currency, such as Bitcoin. Unlike traditional currencies, bitcoins are entirely virtual. There are no physical coins or even digital coins per se. The coins are implied in transactions which transfer value from sender to recipient. Users of bitcoin own keys which allow them to prove ownership of transactions in the bitcoin network, unlocking the value to spend it and transfer it to a new recipient. Those keys are often stored in a digital wallet on each user’s computer. Possession of the key that unlocks a transaction is the only prerequisite to spending bitcoins, putting the control entirely in the hands of each user.

Nowadays, because of blockchain is decentralization, trustless, collectively maintain, reliable database, more and more companies and entrepreneurs start to

use blockchain in finance record, medical records, and other records management activities, such as identity management, transaction processing, documenting provenance, or food traceability.

The paper is structured with section two being a background of blockchain in Bitcoin and its variants. Section three shows how Personal Archive Service System using Blockchain Technology works and what's different between traditional Third-party verification agencies. Section four outlines my demo for this system include private blockchain, biological information, smart contract, Dapp. Section five encompassing future and related works which should be conducted to further this experiment. My report then terminates post acknowledgements, reference sections, and two appendices.

II. BLOCKCHAIN OVERVIEW

The first blockchain was conceptualized in 2008 by an anonymous person or group known as Satoshi Nakamoto with the publication of a paper titled "Bitcoin: A Peer-to-Peer Electronic Cash System", and implemented in 2009 as a core component of bitcoin where it serves as the public ledger for all transactions. This technology makes bitcoin become a completely de-centralized electronic cash system that does not rely on a central authority for currency issuance or settlement and validation of transactions.

A. The Block

1. Introduction

The blockchain data structure is an ordered back-linked list of blocks of transactions. The blockchain can be stored as a flat file, or in a simple database. Blocks are linked “back”, each referring to the previous block in the chain. The blockchain is often visualized as a vertical stack, with blocks layered on top of each other and the first block ever serving as the foundation of the stack.

Every block constitute of block head and any kind of information which length decided by block header. The information can be message, finance record, medical record, certificate, it’s all base on what this blockchain designed for. In bitcoin, the information means transactions.

2. Block Header

The block header consists of three sets of block metadata. First, there is a reference to a previous block hash, which connects this block to the previous block in the blockchain. It should include version, previous block hash, merkle root, timestamp, difficulty target, nonce (difficulty target and nonce only for the blockchain which use proof-of-work, proof-of-stake as consensus).

3. Block Identifiers

The primary identifier of a block is its cryptographic hash, a digital fingerprint, made by hashing the block header twice through the SHA256 algorithm. The resulting 32-byte hash, is called the block hash, but is more accurately the block header hash, as only the block header is used to compute it. Block’s hash is computed by each node as the block is received from the network. The block hash

may be stored in a separate database table as part of the block's metadata, to facilitate indexing and faster retrieval of blocks from disk.

A second way to identify a block is by its position in the blockchain, called the block height. The first block ever created is at block height zero. A block can identify two ways, either by referencing the block hash, or by referencing the block height. Each subsequent block added "on top" of that first block is one position "higher" in the Blockchain, like boxes stacked one on top of the other. The block height may also be stored as metadata in an indexed database table for faster retrieval.

4. Linking Blocks in the Blockchain

Bock is linked in the form of a chain by referencing the block header hash of the parent block

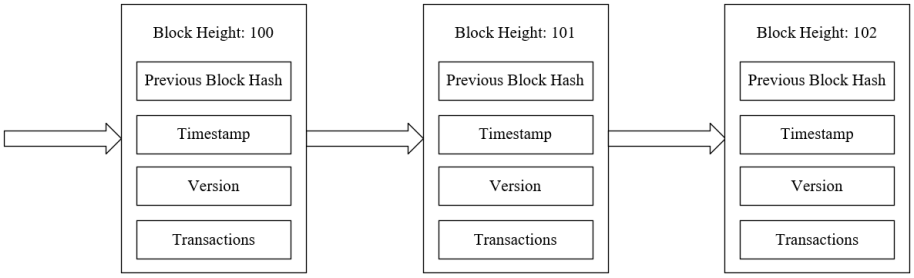

Figure 1. Linking blocks

For the simplest explain for Blockchain is blocks link each other by every block have "previous block hash". If change data in one block, for ensure it's still a chain, it needs to change every block which after this block, and combine with consensus, it makes all the data in Blockchain immutable.

B. Mining and Consensus

1. Instruction

Mining and consensus is the most important process in Blockchain. It also secure the bitcoin system against fraudulent transactions or transactions spending the same amount of bitcoin more than once, known as a double-spend.

Consensus gives the rule accept and confirm the new block in the whole Blockchain community. Follow the consensus, mining is a way that miner use to compete with each other and get the right to write the new block.

Mining is the invention that makes bitcoin special, a de-centralized security mechanism that is the basis for peer-to-peer digital cash. The reward of newly minted coins and transaction fees is an incentive scheme that aligns the actions of miners with the security of the network, while simultaneously implementing the monetary supply.

2. De-centralized Consensus

Blockchain is more like the global public ledger of all transactions and everyone in the Blockchain network accepts as the authoritative record of ownership.

Trustless is also an important feature of Blockchain. Without having to trust anyone, the traditional way depend on a trust model that has a central authority providing a clearinghouse service. Blockchain has no central authority, but every node has a complete copy of a public ledger that it can trust as the authoritative record. Blockchain is not created by a central authority, but is assembled

independently by every node in the network. Every node in the network, acting on information transmitted across insecure network connections can arrive at the same conclusion and assemble a copy of the same public ledger as everyone else.

The most creative idea in Blockchain is the decentralized mechanism for emergent consensus. Consensus is an emergent artifact of the asynchronous interaction of thousands of independent nodes, all following simple rules. Independent verification of each transaction, by every full node, based on a comprehensive list of criteria. Independent aggregation of those transactions into new blocks by mining nodes, coupled with demonstrated computation through a Proof-of-Work algorithm. Independent verification of the new blocks by every node and assembly into a chain. Independent selection, by every node, of the chain with the most cumulative computation demonstrated through Proof-of-Work.

3. Independent Verification of Transactions

While clients need some new transactions write into next block, all they need to do is sending these transactions to any node in the Blockchain network. Before broadcast these transactions to its neighbors, every bitcoin node that receives a transaction will first verify every transaction against a long checklist of criteria. This ensures that only valid transactions are propagated across the network, while invalid transactions are discarded at the first node that encounters them. Then every node builds a pool of valid new transactions, roughly in the same order.

4. Aggregating Transactions into Blocks

The node that write the last block, will aggregate these transactions which will be written in the next block from the transactions pool into a candidate block. it is not yet a valid block, as it does not contain a valid proof-of-work. The block becomes valid only if the miner succeeds in finding a solution to the Proof-of-Work algorithm.

5. Transaction Age, Fees, Reward and Priority

For every transaction wants to add in the Blockchain, there is a priority metric to each transaction and adding the highest priority transactions first. One way to decide the priority is the "age" of the UTXO (Unspent Transaction Output), old and high-value inputs to be prioritized over newer and smaller inputs.

The priority of a transaction is calculated as the sum of the value and age of the inputs divided by the total size of the transaction:

$$Priority = Sum\ (Value\ of\ input \times Input\ Age) \div Transaction\ Size$$

Other way is give more transaction fee, most of the transactions combine with transaction fee. Miner can get these fee if they success get the right of write next block. Thus most of the miners will prioritize those with the highest fee per kilobyte of transaction.

If there is any space remaining in the block, mining node may choose to fill it with no-fee transactions. But some miners may choose to ignore transactions without fees.

Finally, for every node which successfully calculates the new block, the node have the rights to generate new UTXO, separate and send to any address node want.

6. Proof-of-Work Algorithm

A hash algorithm takes an arbitrary-length data input and produces a fixed-length deterministic result, a digital fingerprint of the input. It is also virtually impossible to select an input in such a way as to produce a desired fingerprint, other than trying random inputs. With SHA-256, the output is always 256 bits long, regardless of the size of the input. Thus, POW (Proof-of-Work) uses these features.

The miner constructs a candidate block filled with transactions. Next, the miner calculates the hash of this block's header and sees if it is smaller than the current target. If the hash is not less than the target, the miner will modify the nonce and try again.

The formula to calculate the difficulty target from this representation is:

$$target = coefficient \times 2^{(8 \times (exponent - 3))}$$

Furthermore, the number of participants in mining and the computers they use will also constantly change. To keep the block generation in a stable time, the difficulty of mining must be adjusted to account for these changes. The equation can be summarized as:

$$New\ Difficulty = Old\ Difficulty \times (Actual\ Time\ of\ Last\ N\ Blocks \div N\ minutes)$$

Sometimes to avoid extreme volatility in the difficulty, the retargeting adjustment will be less than a factor of N per cycle.

7. Proof-of-Stake Algorithm

For POW algorithm, miner needs to waist too much calculation, then get a right answer. If a miner can have enough processing power, in theory more than 51% processing power in the whole Blockchain network, it can control all the transactions and do the double-spend attack. The key problem is this miner has the right to create new block continuously. Thus, some new Blockchain networks, like Ethereum, starting use POS (Proof-of-Stake) algorithm.

Unlike POW, POS algorithm not only based on processing power, but also based on transaction age. Difficulty of mining is different to every miner, the more stake miner gets, the less difficulty of mining is. Stake is calculated by amount of coin miner has and transaction age. Every time when a miner finds the new block, its coin's transaction age will clear to zero again, and its almost impossible that this miner can create the new block continuously.

C. Blockchain Community

1. Blockchain Forks

Because the Blockchain is a decentralized data structure, different copies of it are not always consistent. Blocks may arrive at different nodes at different times, causing the nodes to have different perspectives of the Blockchain. To resolve this, each node always selects and attempts to extend the chain of blocks that represents the most POW, also known as the longest chain or greatest cumulative difficulty chain. By summing the difficulty recorded in each block in a chain, a node can

calculate the total amount of Proof-of-Work that has been expended to create that chain. As long as all nodes select the longest cumulative difficulty chain, the global bitcoin network eventually converges to a consistent state. Forks occur as temporary inconsistencies between versions of the Blockchain, which are resolved by eventual re-convergence as more blocks are added to one of the forks.

Ideally the network has a unified perspective of the Blockchain, with the blue block as the tip of the main chain.

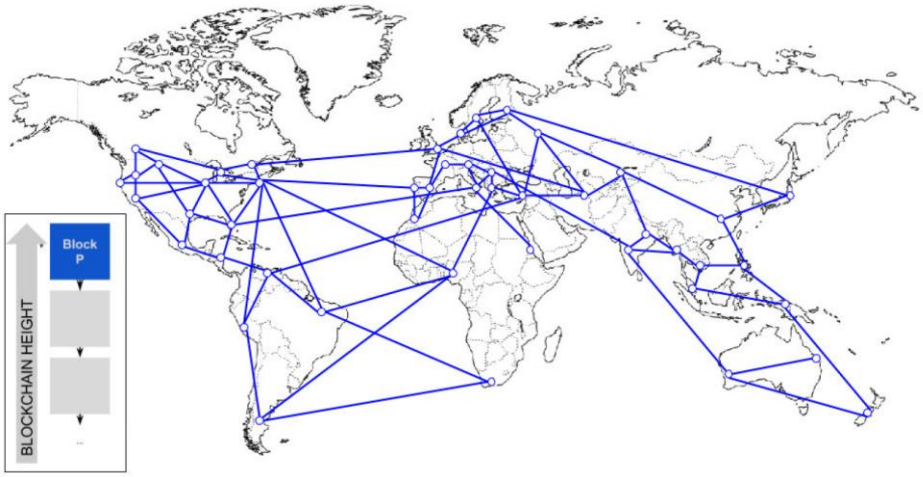

Figure 2. Before the Fork

For some reasons, like two nodes get the new block at the same time or network delay problem, sometimes several new blocks will broadcast in the Blockchain network at the same time. As the two blocks propagate, some nodes receive block “red” first and some receive block “green” first. The network splits into two different perspectives of the Blockchain, one side topped with a red block, the other with a green block.

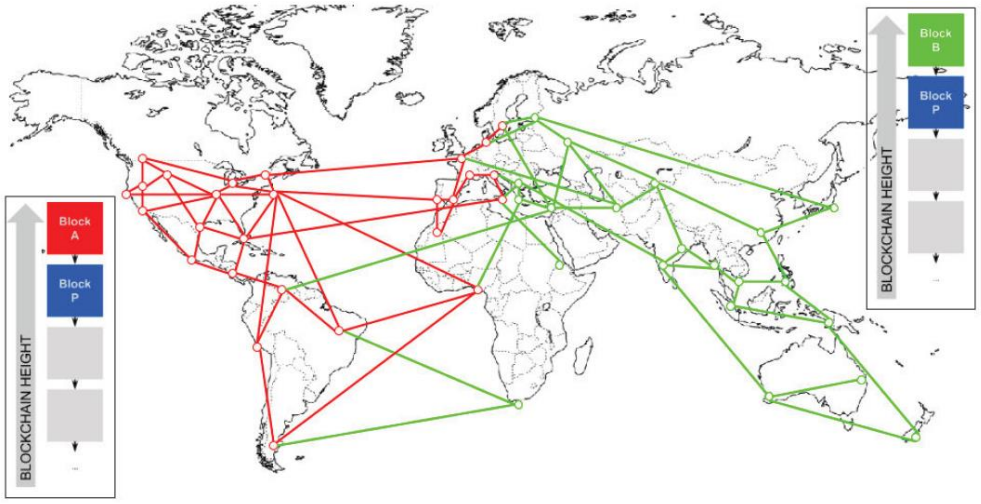

Figure 3. Two blocks propagate, splitting the network

Any miners that saw “red” first will immediately build candidate blocks that reference “red” as the parent and start trying to solve the POW for these candidate blocks. The miners that accepted “green” instead, will start building on top of “green” and extending that chain.

Forks are almost always resolved within one block. As part of the network’s hashing power is dedicated to building on top of “red” as the parent, another part of the hashing power is focused on building on top of “green”. Even if the hashing power is almost evenly split, it is likely that one set of miners will find a solution and propagate it before the other set of miners have found any solutions.

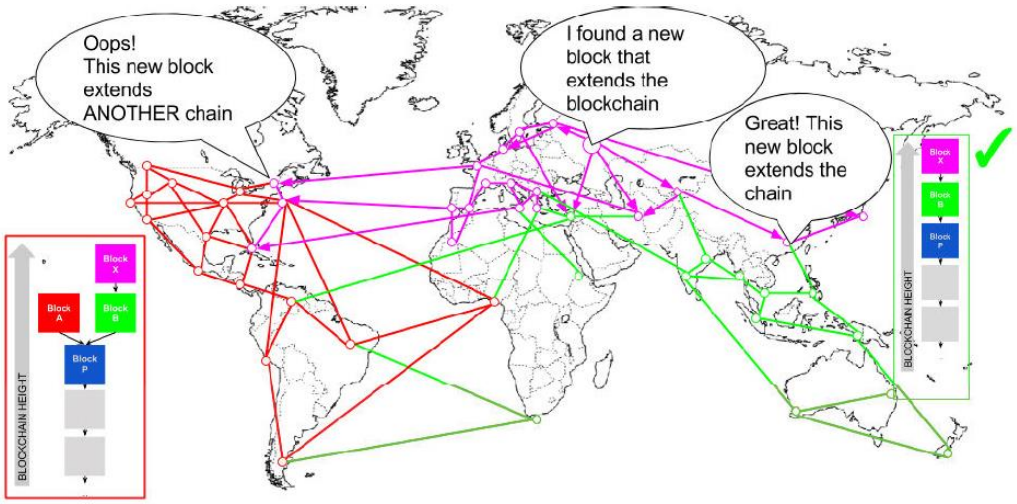

Figure 4. re-converges on a new longest chain

It is theoretically possible for a fork to extend to two blocks, if two blocks are found almost simultaneously by miners on opposite “sides” of a previous fork. However, the chance of that happening is very low. Whereas a one-block fork may occur every week, a two-block fork is exceedingly rare.

2. Mining Pools

When the Blockchain network grows up, individual miners working alone don’t stand a chance, it represents a gamble, like playing the lottery. By participating in a pool, miners get a smaller share of the overall reward, but typically get rewarded every day, reducing uncertainty.

Successful blocks pay the reward to a pool bitcoin address, rather than individual miners. The pool server will periodically make payments to the miners’ bitcoin addresses, once their share of the rewards has reached a certain threshold. Typically, the pool server charges a percentage fee of the rewards for providing the pool mining service.

Miners participating in a pool, split the work of searching for a solution to a candidate block, earning “shares” for their mining contribution. The mining pool sets a lower difficulty target for earning a share, typically more than 1,000 times easier than the bitcoin network’s difficulty. When someone in the pool successfully mines a block, the reward is earned by the pool and then shared with all miners in proportion to the number of shares they contributed to the effort.

3. Managed Pools

Most mining pools are “managed”, meaning that there is a company or individual running a pool server. The owner of the pool server is called the pool operator and they charge pool miners a percentage fee of the earnings.

Pool miners connect to the pool server using a mining protocol such as Stratum or GetBlockTemplate. Each pool miner then mines using the block template, at a lower difficulty than the bitcoin network difficulty and sends any successful results back to the pool server to earn shares.

4. P2Pool

Managed pools create the possibility of cheating by the pool operator, who might direct the pool effort to double-spend transactions or invalidate blocks. P2Pool works by de-centralizing the functions of the pool server, implementing a parallel blockchain-like system called a sharechain. Each of the blocks on the sharechain records a proportionate share reward for the pool miners who contribute work, carrying the shares forward from the previous share block. P2Pool mining is more complex than pool mining, as it requires that the pool miners run a dedicated

computer with enough disk space, memory and internet bandwidth to support a full bitcoin node and the p2pool node software.

III. RealID System

A. Introduction

Identification management, authentication and authorizations are being studied ever since digital systems were emerged, from the still effective user name and password pair to multi-factors, multi-channels and multi-devices authentication, to social authentication, and to open authentication and authorization service.

OpenID Connect is an open standard and decentralized authentication protocol. It enables relying parties (RP) to verify the identity of an end-user based on the authentication performed by OpenID provider (OP) or Authorization Server, as well as to obtain basic profile information about the End-User in an interoperable and REST-like manner. By allowing users to be authenticated through a third party service, it eliminates the need for webmasters to provide their own ad hoc login systems and a separate identity and password.

B. OAuth

OAuth is an open protocol to allow secure authorization from web. It enables a third party application to obtain specific access rights through http service. But it does not provide ID management and verification. OpenID Connect is a simple identity layer on top of the OAuth 2.0 protocol.

While OAuth enables third party authorization service, it is still centralized in terms of OpenID Provider. And users have to reveal many private information in order to be authenticated. It also poses a single point of failure. Moreover, it is against the very nature of internet that not a single central authority that controls who can do what.

C. RealID System

Recently, block chain technology, first introduced in bitcoin as an innovative payment network and a new kind of money, is being applied to many fields such as supply chain, manufacture, Internet of Things in addition to financial related industrials.

The key features from the block chain technology are its decentralized, consensus based shared ledger, smart contract and privacy. Hence, it can realize the needed features for ID management and ID service such as secure, decentralized, anonymity, transparency and immutability. The proposed RealID management and service in this paper is to exploit such features.

In addition, the RealID is also to utilize biometric specific information to differentiate human being from robots, which is particularly useful in the social media applications in which a user can be anonymous but needs to be stamped as a real human rather than a bunch of social bots. Therefore, results such as sentiment or poll from a social media application can represent the real situations and not skewed by social bots.

The following section talks about the general requirement of an ID management and services. Digitized biometric information collection and distribution are discussed thereafter. After that, the proposed RealID system is detailed and its application is followed. The last section is devoted to the discussion and future applications.

1. ID System Requirements

W3 identifies general requirements for a global identity management service as follows:

- Portability and Interoperability
- Extensibility
- Negotiated Privacy and Security
- Accountability
- Distributed Registration Authority
- Distributed Certification Authority
- Independent Governing Authority

Anonymity and pseudonymity for protection of personal privacy is under the purview of privacy and security.

2. Biometrics and Uniqueness

Biometrics in computer science is the discipline of using metrics related to human characteristics to do identification, authentication and access control. Fingerprint, palm veins, face recognition, DNA, palm print, hand geometry, iris

recognition, retina are some examples. Other metrics can be related to behaviors such as voice, typing rhythm.

As technology advances, biometrics accuracy improves greatly. The time needed to get such information is much faster. The way to get them is easier. The price is much affordable.

One main direction in biometric authentication is to detect fake biometrics quickly so that such system can be used for real time authentication. In a different application scenario in which biometrics is being used to differentiate human being from robots. The challenging is to detect efficiently non-human biometrics such as digital modification from vast amount of human biometrics. This is particularly useful in applications that require anonymity but identifiable human natures. We are investigating various methods such as principal component analysis, wavelets, and correlations to test such classification.

In the proposed RealID, biometrics is being used but not stored. The one way function makes it hard to get original biometrics.

3. ID System

We now to describe the genera sequence for a user to get identified. Figure 5 demonstrates the work flow.

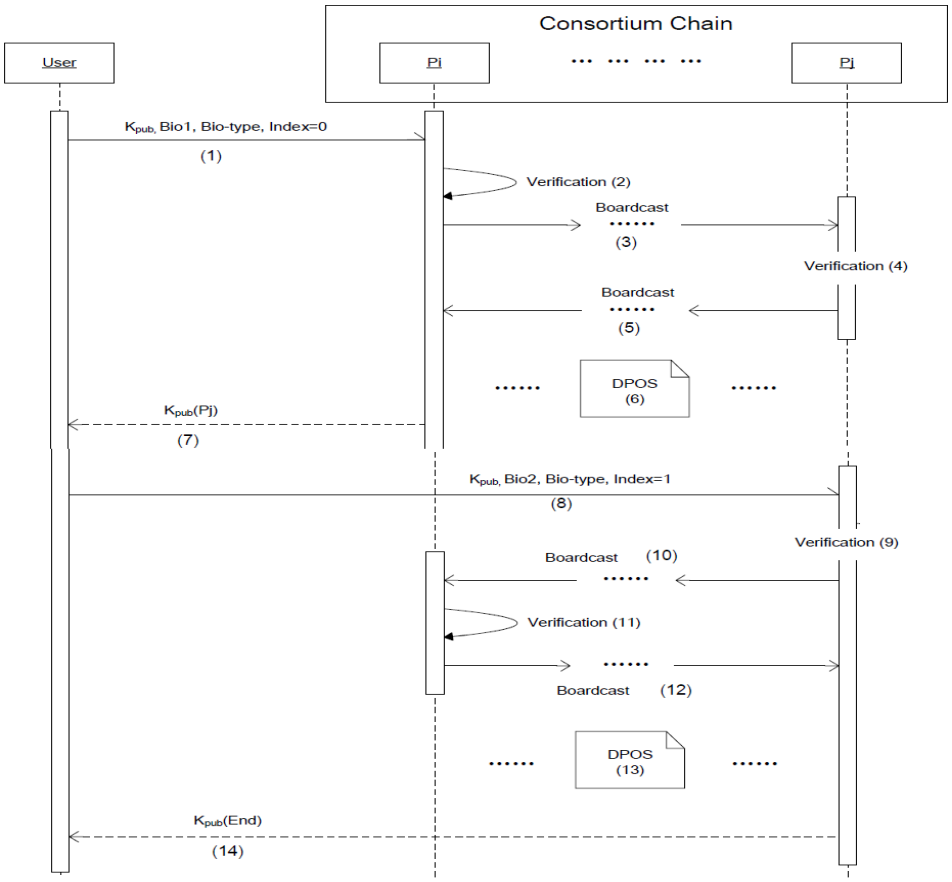

Figure 5. RealID ID System

(1) A user starts to send its pubic key, original biometric information, the biometric type and a counter =1 via HTTPS to randomly picked node in the consortium block chain, say Pi. The counter implies the verification sequence. Counter = 1 means it is the first time the user asks for a verification. If it is bigger than 1, the node P selection needs to follow a particular instruction (see step 9).

(2) Pi receives the request and starts to verify the information. It looks into the biometrics is in line with the type of biometrics implied. For example, biometrics for a facial is quite different from those for finger print. If it is in line, Pi translates the biometrics into a set of characteristics. The transformation is one way in nature so that no one is able to recover original biometrics at least in theory. If it is not in

line, then Pi returns a false and stops the verification process. Next, Pi compares the characteristics with the existing ones in the block. If the characteristics is closely matched, Pi returns false and stops. If not, Pi checks if count =1. If it is 1, Pi then do a final check if the public key is registered. If it is not, Pi moves to step 3. If it is registered already, Pi returns a false and stops. It the counter is not 1, Pi does a SHA256 to the characteristics and selects the last four digits to see if it (Pi) has the right to do this verification. If it is, Pi moves to step 3. If not, it returns false and stops.

(3) Pi broadcasts the type of biometrics, the characteristics, the user's public key, Pi's unique identifier such as its public key or pre-assigned unique number, and the counter.

(4) Receiving nodes from Pi will do the verification. They need to verify if the biometric information is in type and unique like in step 2. They also need to verify Pi has the right to do such verification. It all yes, they move to step 5. If not, they returns false and stops.

(5) These nodes in step 4 continue broadcasting to the next level of nodes and these new nodes do the same verification like 4. And the propagation keeps going.

(6) When enough nodes returns positive response, the consortium uses the Delegated Proof of Stake (DPOS) consensus model to choose a node to write the information sent at step 3 into the block.

(7) DPOS returns a positive signal to the user along with the next node to be selected by the last four digits from SHA256 the characteristics.

(8) User moves to get next verification by sending its information similar to step 1 to the assigned node, say Pj. The counter will increase by 1.

(9) Pj will do the similar tasks like in step 2. In addition, Pj will check if the counter reaches its threshold. If it is not, Pj will do the same steps from 3 -7. If it is, Pj will create a unique UTXO that represents the user passes the verification process. UTXO can be generated by selecting the first couple digits from SHA256 to the recorded biometric characteristics.

(10) Pj broadcasts the UTXO

(11) Similar to 4, receiving nodes will verify the biometric information and double check the correctness of UTXo calculation. If all of them are right, they moves to step 11. If one of them is not, they return false and stop.

(12) Like step 5, broadcasting continues.

(13) Finally, similar to step6, the UTXO is recorded.

(14) The user gets a notification Applications.

4. ID Service

Internet social media lacks a layer of quality control to any postings, partially it has little background information on who is in the circle or no control to whom allowed into the circle. Assessing trustworthiness of such postings is therefore based on postings themselves and their related information or signals. Most researches focus on by exogenous signals such as hyperlink structures. Another research is on endogenous signals such as correctness of factual information on postings. Such signals can place high quality postings, mostly by experts in a

relatively high ranking otherwise lost in a sea of postings. Recent research and practices try to related posting quality with individual account, for example, reacting to things in Facebook, scores and reputation, for examples, some chat groups use reputation to decide who is allowed to make comments in, personal online ratings based on the aggregated digital identity.

Some investigations and research are on whom post what and when. During the process of sign in, some applications employ CAPTCHA, photo-based social authentication, rate-limit violation and account connection property.

With the id management system, new ID service can be provided in high confidence with the desired features a) real human being, b) no more alias account, c) anonymous and secure, d) transparent, e) immutable, f) consensus based de-centralized and, g) revocable.

Figure 6 lists the steps that the flow of authorization service.

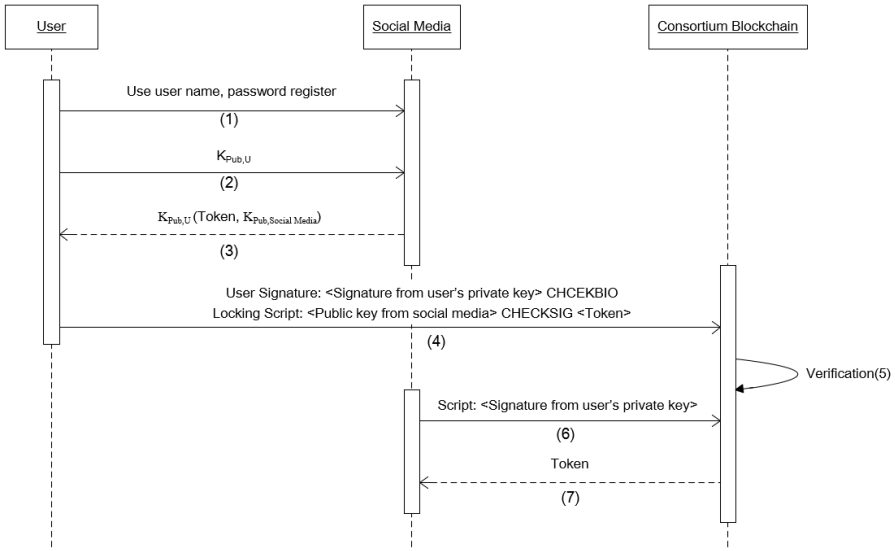

Figure 6. RealID ID Service

(1) A user (U) registers to a social media application, SM,

(2) U sends its public key to SM to let it to verify U's identity.

(3) SM generate a secret, say a token, and sends encrypted token via U's public key to U.

(4) U sends its unlocking script (or signature) along with a locking script to any node in the consortium block chain. The unlocking script is used to unlock U's UTXO. It manifests only the owner of the UTXO can have the right to use it. The locking script states only SM can check it.

(5) The Node will make verification using the scripts it received. It check the corresponding public key using the unlocking script. If it has the UTXO, it will broadcast its locking script and unlocking script.

(6) SM sends its unlocking script to any node in the consortium block chain, The node will check if matched from both scripts. It yes, the node will return the Token to SM.

(7) SM can request more verifications from other nodes in consortium block chain.

5. DISCUSSION

I have illustrated a novel approach of ID management and authorization service using block chain technology. I noticed that the block chain technology is fit for ID management and service very well. The system is immune many ID attacks.

A prototype implementation has demonstrated that RealID is feasible and serves our original research purpose, that is, how to identify social bots from social media applications.

More robustness tests needs to be done, especially such system involves cryptographic applications.

The extraction of biometric information is another area needs to be addressed. It is true as technology advances, biometrics accuracy improves greatly. The time needed to get such information is much faster. The way to get them is easier. The price is much affordable. But, it takes equipment and financial burden to get such information. We hope RealID can provide such services that not only attract users to register and use it but also to prevent ID misuse by ID theft. Luckily, many such devices are being developed in the market.

Lastly, the RealID is a demonstration of ID management and service. Its idea can be extended and exploited to any other systems that require both anonymity and accountability.

IV. Personal Archive Service System

A. Introduction

Personal archive service system (PASS) is defined as a collection of digital artifacts as well as a collection of service tools. A personal digital artifact (PDA) is a digital form of personal achievements with evidentiary documents (PAE) or of personal identifications (PID) that can be used to uniquely identify a person.

Examples of PAE are such as personal education, experiences, training, degree, diploma, academic transcripts and various certificates. Personal wealth like property value, bank balance, investments can also be examples of PAE. Typical examples of PID are biometric measures from the person like finger prints, iris and vein. Other information known only to the person, physical objects owned by the person can be examples of PID. Behavior patterns, personal reflection and characteristics are also considered as examples of PID.

The collection of PAEs is similar to personal portfolio (PP). It is associated with timeline that assembles a specific characteristic of a person. PP is distinct from resume in that it is more of inclusive and expanding than a line of statements in resume.

To be useful and trustful, these PDAs need to be verified by a third party who should have direct knowledge of these claimed artifacts. For example, a university can provide official transcripts to their graduates. A landlord is able to provide a letter of reference to their tenants. No other entities can do that.

We define various roles in this process. We use the term “subject” to be a person or a user who is building his or her collection of PDAs, “certifier” to be an institute or an entity that provides certification, “inquisitor” to be an agent or organization who provides service of investigation and obtaining relevant proof of any particular subject, and “client” to a person or organization using professional services provided by “inquisitor”. For example, for a potential candidate to be hired by a company, the candidate is a subject, the company is a client who needs to hire

a third party to verify all the information provided by the subject. The third party is the inquisitor and the company is a client of the inquisitor. The inquisitor needs to contact various certifiers such as universities who gain education, companies who used to work, or organizations who issue other certifications.

This process of verifying PDAs by inquisitors is often time consuming. It is also repeated every time a client makes a request. In some cases, it obtains information beyond the consent given by the subject. Therefore, it poses a threat of privacy. Moreover, it is essentially a model of centralized system that can potentially cause a point of failure, a point of bottleneck, a point of confusion and a point of abuse. For example, clearance for teaching when hiring an adjunct by a third party takes time. In some extreme cases we have experienced that the semester starts while the clearance is still pending. Another example is about name change by a subject. Without revealing its previous name used, the subject is hard to get verified. We see a case that a subject wants an institution to issue a new diploma due to name change. It is usually a mission impossible task although in this case the subject has all other IDs that remain the same just name difference.

But, the process of verification is becoming increasing necessary and a must. We have observed resume padding is becoming more epidemic. Resume padding is to add false or exaggerated information to a resume to enhance credentials for a job. More than 40% of resumes inflate their salary, 33% inaccurate job description, 29% altered employment dates, 27% falsified references, and 21% fraudulent degrees. Hence, we can image that the burden of requiring verification is very heavy.

Recently, Blockchain technology, first introduced in bitcoinas an innovative payment network and a new kind of money, is being applied to financial related industrials. Many other fields such as supply chain, manufacture, Internet of Things are exploiting it as well. The key features from using the Blockchain technology are its decentralized, consensus based shared ledger, smart contract, high transparency yet high privacy. The idea to have secured storage and transmission of digitally signed documents with a super audit trail in immutable document exchange networks is emerging in trade finance, shipping, and insurance, where everyone has the same demand to validate the identity of people and assets. It does not need a third party as an intermediary or authority for verification, cleaning house or any other purposes. It is being used in building a secure anonymous yet transparent immutable ID Service.

Personal Archive Service System (PASS) proposed in this paper is to use Blockchain technology to exploit its desirable features such as immutable, transparency, anonymous and public consensus. The subject controls its own PDAs and makes decision to whom to release. Figure 1 illustrates the general infrastructure and architecture view of a PASS under Blockchain. The subject, represented by icon has its own repository or wallet that aggregates all its relevant PDAs. The certifiers issue certificates to its owners as well as to a trusted network. It does such certificates once for everyone involved and should not be bothered anymore. There is no need to have a third party or an inquisitor. They also pass the certificates to a consortium oriented block chain network that the trust is developed

in a delegated proof of stake. A client makes a request to the subject and gain access to those granted PDAs.

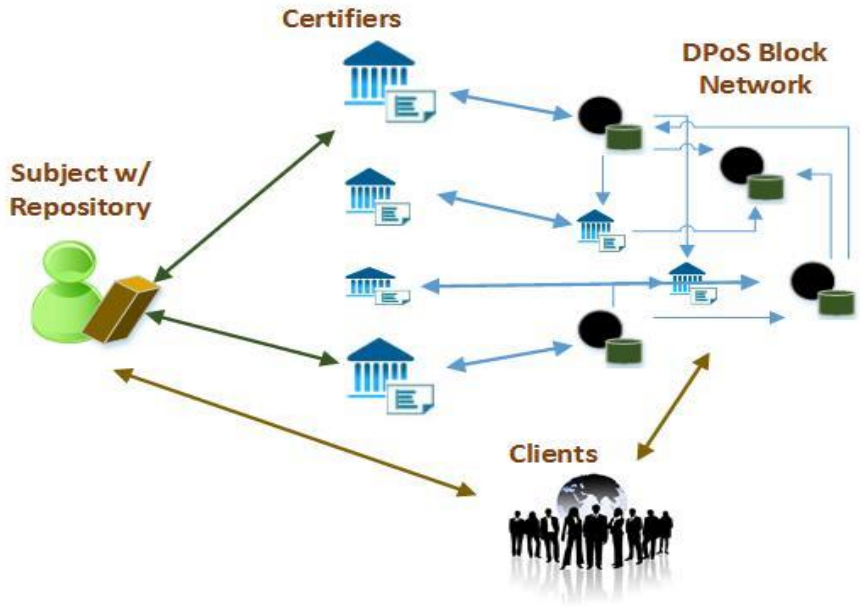

Figure 7. Architecture of PASS

Our main contribution is to propose a framework of using Blockchain technology to build a personal archive that associated certifications. It preserves the authenticity, accuracy and transparency while keeps privacy. Inquisitors are no longer needed. A subject can decide what to reveal to whom and when based on the nature of the request. When a request is more of academic achievement, any PDAs associated with such tag can be unlocking and revealed. When a request is of identification, the subject can unlock those PDAs such as biometrics. A client can gain access to the PDAs immediately and makes decision accordingly. Moreover, tools and services provided by the PASS can make any projects related to personal growth and timeline much easier.

The following section reviews the latest development of the concept “proof of X” using Blockchain technology where X can be anything like identity, property

ownership, specific transaction, college degree, medical records, and academic achievements. The next section will discuss the PASS, a personal archive service system upon which subjects can build their personal archives timely and accurately. They are certified so that they do not need to be verified every time to be used. The PASS is transpierce while maintains privacy. The last portion is devoted to the discussion on opportunity, challenging and future applications.

B. Literature Review

Main issues of building personal digital artifacts (PDAs) are the verification process and the trustworthiness. Statistically as indicated in section I, more than 30% of job seekers modified their diploma information. 70% inflate their achievement.

Various efforts are made to build a level of verification and certification around academic achievement. Open Badge idea is one of them. The standard working group for open badges as it stated in its web site communicates skills and achievements by providing visual symbols of accomplishments packed with verifiable data and evidence that can be shared across the web. Open Badges enable subjects manage their own learning achievements. The goal of the working group is to move the specifications more towards a standard that is clearer and easier to align with, maintain, and build from. Acclaim is a digital badging platform based on the Open Badge Standard and backed by Pearson. Acclaim has issued millions of badges from reputable organizations like IBM for career-advancing achievements that help individuals move forward professionally. A badge issued

through Acclaim is a digital representation of a learning outcome, experience or competency. These badges can be shared and verified online in a way that is easy and secure. They contain detailed information that provides context around what exactly was achieved, which organization recognizes the achievement, and the individual who earned the recognition. The services provided have their merits. The main shortcoming is its centralized model.

Sony Global Education announces that it will develop technology using Blockchain for open sharing of academic proficiency and progress records on 2016. It is going to leverage Blockchain's secure properties to realize encrypted transmission of data - such as an individual's academic proficiency records and measures of progress - between two specified parties. The actual working flow is yet to be seen and needs to be tested.

ID system is an aged long problem. W3 identifies seven general requirements for a global identity management service: portability and Interoperability, extensibility, negotiated privacy and security, accountability, distributed registration authority, distributed certification authority and independent governing authority. Such service must use globally unique identifiers in a common interchange format, support extensible mapping to these identifiers from other commonly used identifiers and use a common protocol for asserting and authenticating a global identity. It must support global vocabulary definition as well as distributed local vocabulary definition. It needs support anonymity and pseudonymity for protection of personal privacy. So the anonymity and

pseudonymity for protection of personal privacy is under the purview of privacy and security.

ID system should support both hierarchical and peer-to-peer registration models. The governing authority should be chartered as an international non-profit organization so it is industry-, vendor-, and government-neutral in all respects. It should set both technical and operational standards for the service, as the two are tightly intertwined. It should manage global vocabulary development for universal identity attributes and global protocol control structures. It should set the accountability terms for all agents, including registration and certification authorities. It should serve as an impartial root authority for hierarchical registration or certification models.

OpenID with OpenID Connect is to define and develop an open standard and decentralized authentication protocol. It enables relying parties (RP) to verify the identity of an end-user based on the authentication performed by OpenID provider (OP) or Authorization Server, as well as to obtain basic profile information about the End-User in an interoperable and REST-like manner. By allowing users to be authenticated through a third party service, it eliminates the need for webmasters to provide their own ad hoc login systems and a separate identity and password.

OAuth is an open protocol to allow secure authorization from web. It enables a third party application to obtain specific access rights through http service. But it does not provide ID management and verification. OpenID Connect is a simple identity layer on top of the OAuth 2.0 protocol.

While OAuth enables third party authorization service, it is still centralized in terms of OpenID Provider. And users have to reveal much private information in order to be authenticated. It also poses a single point of failure. Moreover, it is against the very nature of internet that not a single central authority that controls who can do what.

Multiple biometric specific information can be used to identify human being in a very high accuracy. This is particularly useful in the social media applications in which a user can be anonymous but needs to be stamped as a real human rather than a bunch of social bots. Therefore, results such as sentiment or poll from a social media application can represent the real situations and not skewed by social bots.

Some web applications like taskstream provide user a platform to build personal portfolio. It can track the progress in particular projects and record reflection on particular topics. These applications can serve as a way to show a trace of personal growth that can provide a provenance of particular skills.

The author proposed ideas of using Blockchain to prove identities in the digital world without centralized database in that it leads to a natural monopoly that everyone has to use. Whereas a decentralized system follows a common protocol that allows individuals to add new transactions and distribute them using peer-to-peer architecture with a super audit trail. A subject can give controlled key usage to clients such as banks, insurers, or governments who want to inspect the documents with smart contracts. A smart contract is a piece of code recorded in the

common blocks. The contract can restrict the number or timing of inquisitions and record them all for the subject.

C. Structure of PASS

1. Artifacts, Portfolio and Archive

A personal digital artifact (PDA) is a piece of signed digital document that contains the following tags or fields:

- PDA-ID: An index generated by the system
- PDA-Type: Either personal achievements with evidentiary documents (PAE) or personal identifications (PID). It follows the sub-type such as diploma, degree, skills, and biometric measure. They should be standardized codes eventually.
- PDA-Description: a string to describe the nature of the PDA
- PDA-Subject: Name of the person who owns the PDA
- PDA-Certifier: Person or organization who issues certification toward this PDA
- PDA-Date-Start: The starting date of the PDA
- PDA-Date-End: The completion date of the PDA
- PDA-Date-Validation: optional date that the PDA is valid
- PDA-Comment: Optional field
- PDA-Key-Cert: The public key associate with the certifier

- PDA-Key-Sub: The public key generated for the subject and this particular PDA. The private key is kept by the subject in its application
- PDA-Unique: Used by the certifier to verify the ownership of this PDA. Date of Birth, Student ID or Universal ID, and/or something known only to the subject. It can be multiple fields required by the Certifier
- PDA-Data: Data needed by the certifier. For PID type, it could be specific biometric information. For PAE, it could be a scan of original certificate (hard copy)

Personal portfolio (PP) is a collection of PDAs whose PDA-type is under PAE (personal achievements with evidentiary documents). Personal identifications (PID) is a collection of PDAs whose PDA-type is under PID. Therefore PA is a union of PP and PIDs. Symbolically,

$$PA = \{P_p, P_{id}\}$$

where

$$P_p = \{PDA_{PAE1}, PDA_{PAE2}, \cdots, PDA_{PEAn}\}$$

and

$$P_{id} = \{PDA_{PID1}, PDA_{PID2}, \cdots, PDA_{PIDm}\}.$$

2. Service Tools

A service tool is a piece of code that performs a specific task required by the PP. It should include the following fields.

- ST-ID: An index generated by the system
- ST-Type: The nature of the service such as inquisitor, manager, or others.

- ST-Description: A string to describe the nature of the service
- ST-Input: Service conditions and parameters
- ST-Output: Service results
- ST-Condition: Under what condition and/or privilege the service should be performed
- ST-Creator: Person or organization who issues the tool
- ST-Date: The date the service in place
- ST-Comment: Optional field
- ST-Language: Specific language that implements the tool.

A collection of service tools is denoted as T. It covers a variety of the services needed to make the PA workable. For example, requesting for a certificate, aggregating and synchronizing PP in local repository or wallet from public ledger, granting permission to a specific party to inspect PDA. Again, we can use the following notation to illustrate the set of tools.

$$T = \{ST_1, ST_2, \cdots, ST_n\}.$$

Hence, PASS can be symbolized as

$$PASS = \{PA, T\}$$

3. Subject, Inquisitor, Certifier, Client, Stake

The paper uses the following terms to describe a Blockchain based distributed peer to peer network.

- The term “subject” is a person that is being discussed about his or her PDAs

- The term “certifier” is an institute or an entity that provides certification
- The term “inquisitor” is an agent or organization to investigate and get relevant proof
- The term “client” is a person or an organization that uses professional services (by “inquisitor”).
- The term “stake node” is a special node in a consortium oriented block chain network. The trust is developed in a delegated proof of stake. It is like “Mining and Verification” in a bitcoin network.

For example, for a potential candidate to be hired by a company, the candidate is the subject, the company needs to hire a third party to verify all the information provided by the subject. The third party is the inquisitor and the company is a client of the inquisitor. The inquisitor needs to contact various certifiers such as universities who gain education, companies who used to work, or organizations who issue other certifications.

4. Initial Trust and Block of the Consortium Network

In general, the initial block or the first block in a Blockchain, coined as initial state of a p2p network, is a difficult task. It needs an acceptable assumption by all peers. PASS adopts a consortium network in which states with certain reputation are included or voted for. Therefore, we can assume such initial block includes data structure in PASS and initial stakes exclusively. A Certifier can be one of the stakes but it does not need to be. To become a trusted certifier, the certifier needs to provide Stakes with its identity.

5. Delegated Proof of Stake (DPoS)

Proof of work (PoW) is a consensus algorithm used in bitcoin network. PoW is done through mining that is the main process of the decentralized clearing house, by which transactions are validated and cleared. Mining secures the bitcoin system and enables the emergence of network-wide consensus without a central authority. The competition to solve the proof-of-work algorithm to earn reward and the right to record transactions on the Blockchain is the basis for bitcoin security model.

A more efficient consensus algorithm is Delegated Proof of Stake (DPoS). It is a variant of the Proof of Stake (PoS). Both were developed in order to reduce the cost and inefficient electricity usage associated with PoW. PoS allows every wallet which contains coins to 'stake', that is to participate in process of validating transactions and forming the distributed consensus and to earn coins in return. While in DPoS, every wallet which contains coins is able to vote for delegates, and it is these delegates who perform the function of validating transactions and maintaining the Blockchain and take the transaction fees as profit. It is more efficient than PoS.

PASS uses DPoS. Delegated stake holders make global ledger.

6. Biometrics and Uniqueness

Biometrics in computer science is the discipline of using metrics related to human characteristics to do identification, authentication and access control. Fingerprint, palm veins, face recognition, DNA, palm print, hand geometry, iris recognition, retina are among them. Other metrics can be related to behaviors such

as voice, typing rhythm and walking patterns and reaction to certain stimulant. As technology advances, biometrics accuracy improves greatly. The time needed to get such information is much faster. The way to get them is easier. The price is much affordable.

One main direction in biometric authentication is to detect fake biometrics quickly so that such system can be used for real time authentication. In a different application scenario in which biometrics is being used to differentiate human being from robots. The challenging is to detect efficiently non-human biometrics such as digital modification from vast amount of human biometrics. This is particularly useful in applications that require anonymity but identifiable human natures. We are investigating various methods such as principal component analysis, wavelets, and correlations to test such classification.

Like other PDAs, the actual biometrics are being used but not stored in the network. The one way function makes it hard to get original biometrics. In addition, it needs multiple biometric information to reduce the probability of false positive. Thus, it makes impossible to have two persons to have two the same biometrics within certain thresholds.

D. COMMUNICATION OF PASS

We now to describe the genera sequence for a subject to get a PDA. Figure 1 in section I demonstrates the work flow. The main idea is to use consortium **B**lockchain in which registered nodes with delegated proof of stake can record the

result in a global ledger. These nodes can be a collection of reputable organizations and companies. Certifiers need not be a node in the network. Their main function is to provide a certificate to the subject. For example, a graduate can ask an official transcript to be one of its PDA. The graduate is the subject. The university is the certifier. The Blockchain network can be composed by collections organizations that are reputable and have incentive to maintain the blocks. Universities, textbook publishers and training societies are ideal candidates. Clients can be any employers who are going to offer a job to the subject.

Figure 8 illustrates the time sequence for PDA-PAE. It is for illustration and has no detailed data structure exchange. Here, the subject makes a request of getting a certificate of one particular achievement to a certifier. The certifier is registered and authorized to issue such certificate. These information is encrypted by the pubic key of the certifier so that no one else can view the PDA. When the certifier gets the request, it verifies the request information in the PDA first and responds either denied or a certificate. In reality, communication between the subject and the certifier is more complicated than it is illustrated. It may ask for more evidences and specific information related to the PDA. It may ask for a proof of payment for the service before issuing the certificate. The certifier also sends the certificate to other nodes in the consortium Blockchain network. The certificate includes information like the certifier digital signature with a unique number, a public key for the PDA, etc. can be viewed by other participants. Any node

received the certificate will do the verification. It broadcast its verification to other nodes. The certificate will be recorded in the global ledger by a delegated stake.

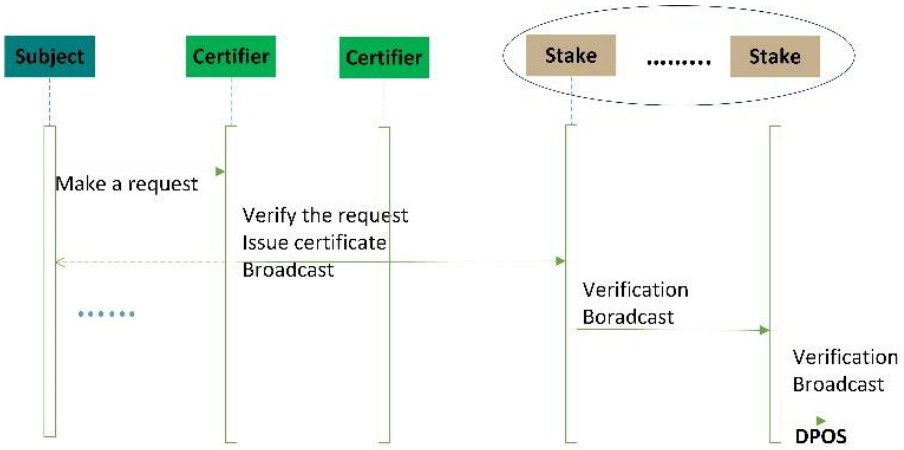

Figure 8. Time Sequence for PDA-PAE

Figure 9 demonstrates another example of the time sequence for PDA-PID. The PID is of biometric data or other id related information that the subject owns, knows and acts. Here we use the biometric measurement for illustration. Since it is shown statistically using two or more types of biometric information will reduce collision dramatically, the subject is asked to present at least two different types of biometric measures. In the end the ledger has all the verifiable subject information via consensus. So it requires a counter to count different types of biometric information.

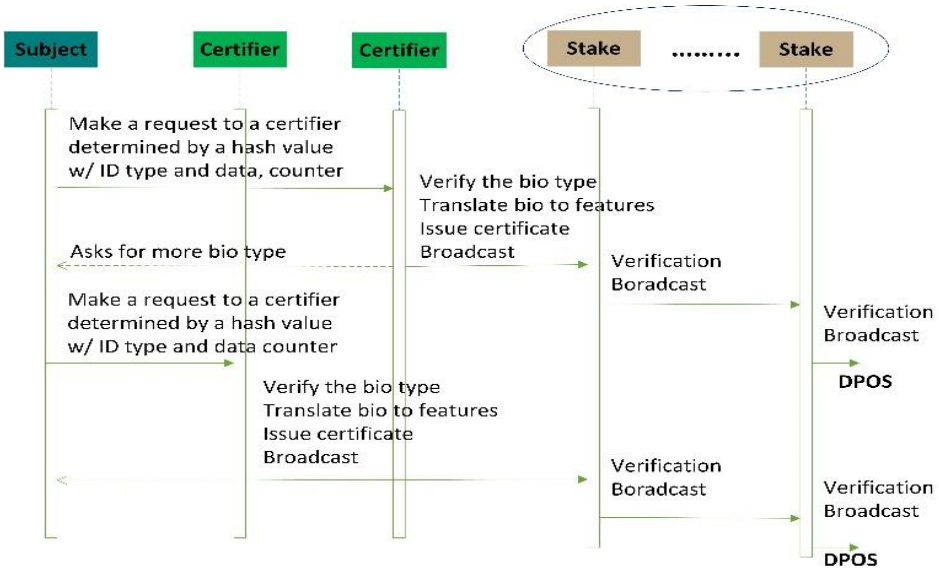

Figure 9. Time Sequence for PDA-PID

Detailed postfix operations can be referenced in. Remember, there are various ways to realize the time sequence in figure 8 and figure 9.

E. SERVICE SYSTEM

An organization or entity (client) who needs some or all of the subject's PDAs in PA can negotiate with the subject directly rather than hires an inquisitor to do the verification. With a smart contract between the client and the subject, the client is able to inspect requested signed documents such as degree, transcript, working experiences, identifications, etc.

Figure 10 is an example of using PASS in the case of internet social media. Internet social media lacks a layer of quality control to any postings, partially it has little background information on who is in the media circle or no control to whom allowed into the circle. Assessing trustworthiness of their postings is therefore based on postings themselves and their related information termed as signals. Most

researches focus on exogenous signals such as hyperlink structures. Recent research is on endogenous signals such as correctness of factual information on postings. Such signals result in placing high quality postings, mostly by experts in a relatively high ranking otherwise lost in a sea of postings. We have observed some novel practices on quality control of posting used in the internet social media are to associate with individual account. Examples are reacting to things, scores and reputation, personal online ratings based on the aggregated digital identity. Some investigations and research are on whom post what and when. During the process of sign in, some applications employ CAPTCHA, photo-based social authentication (Facebook), rate-limit violation (twitter, GitHub, Redit, etc.) and account connection property.

With the PASS, ID service can be provided in high confidence with the desired features a) real human being, b) no more alias account, c) anonymous and secure, d) transparent, e) immutable, f) consensus based de-centralized and g) revocable.

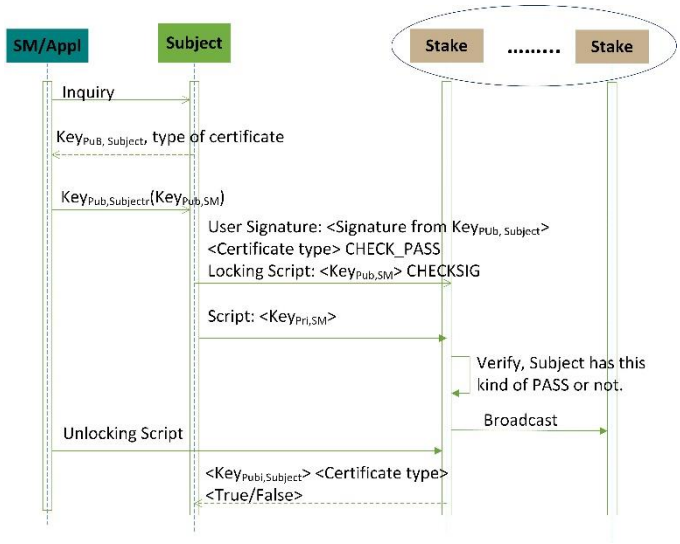

Figure 9. PASS Service

The steps are as follows:

(1) An application such as an internet social media makes a request to the subject for feature verification;

(2) The Subject sends its public key and relevant PDAs to the application;

(3) The application returns its public key signed by subject's public key;

(4) The subject passes its signed unlocking script, signature, list of PDAs with their types and methods to gain inspection of PDAs;

(5) The subject also sends a locking script that includes the public key of the application for signature check. It is unlocked only when the signature matches the application;

(6) A stake in the consortium P2P network makes check if the subject has such PDAs; It continues when they are existed;

(7) The stake broadcasts to the rest of the stakes so they all can use this for verification;

(8) The application sends the unlocking script with its signature;

(9) The stake uses the unlocking script to pass back to the application if the verification is done successfully or failed.

As we see, such sequence of PASS can be generalized to any other applications that need to verify PDAs.