The Application of Block Chain Cryptography in Health Information Technology and Health Research Data Management

By

Dr. Robert Steven Owor, Professor of Computer Science and Software Engineering
Dr. Zephyrinus Okonkwo, Professor of Mathematics
Dr. Anilkumar Devarapu, Associate Professor of Mathematics
Department of Math and Computer Science, Albany State University, Albany GA, USA 31705

Dr. Khalil Dajani, Professor of Computer Engineering
Associate Dean of Engineering and Technology Division, Ashford University, San Diego, CA

1. Introduction
The goal of this White Paper is to investigate the relationship between Block Chain technology and its application to Health IT and/or Health Related research. The paper examines elliptical curve cryptography and the principles of Block Chain technology and how it can be applied to improve health information technology and health research data management including Electronic Health Records (EHR), Electronic Patient Records, Patient Centered Outcomes Research, Electronic Prescribing (eRx) Incentive Program, Physician Quality Reporting System (PQRS), HIPPA Privacy, Security, Breach notification and Health Information interchange.

We begin by examining the current state and limitations of Health Data Management; we then discuss Elliptical Curve Cryptography, a modified Generalized Block Chain Technology Model and the potential application of this model to addressing key problem areas in Health IT and related Health Research Data. Finally we conclude by describing possible avenues for implementing this model.

2. Survey of the Current State and Limitations of Health Data Management Systems
Health Catalyst, a publication which monitors trends and developments in the Health Sector has listed 7 of the most important issues facing Health Data Management Systems today.

2.1 Incomplete Transition to Electronic Health Systems
According to Centers for Medicaid and Medicare Services (CMS.gov), "more than 257,000 eligible professional providers who are not meaningful users of certified EHR technology would have their Medicare Fee Schedule cut by one percent in 2015. Eligible professionals may also see reductions in reimbursements for noncompliance with Medicare’s Electronic Prescribing (eRx) Incentive Program and the Physician Quality Reporting System (PQRS)." The principle reason for this is the inability of Providers to comply with the US Government's HIPPA and HL7 requirements for Electronic Health Records (EHR) and eRx systems due to the complexity and cost of the effort. Clearly, major assistance is needed to help Health Providers comply with the HIPPA and HL7 requirements.

2.2 Personal Digital Health Products generating new data
The rate of personal health technological adoption is outpacing the rate at which Health Care Providers can incorporate this data into the care management of their patients. Wearable devices which measure heart rates, physical activity, weight, blood pressure, sleep patterns and many other parameters are exploding in usage. Several digital health products are also being adopted by Health Providers in Hospitals and other Practices. These records are being stored in any media from cell phones, to memory cards, to cloud storages with possible HIPPA and/or violations.

2.3 Patient Centered Care Data
Patient centered care is another area where Health IT is lagging behind market demands. Many patients are demanding Healthcare choices requiring them to have parameters and measures which are not available to them. Patients shopping for health services want some quality metrics and medical service comparison parameters which can enable them to make more informed health choice decisions for their families. This is a rapidly developing area of Health Information IT which has lagged behind other sectors.

2.4 Increased Demand for Health Data Analytics
Clinical and Administrative Workers, Patients, Insurance Companies and Governments are demanding more and more data in the form of summarized meaningful health statistics in order to assist in decision-making about important health choices, plans and policies. Health Data Analytics are currently unreliable, unavailable, in incompatible formats or restricted because of the difficult and heavy time requirements to extract anonymized and aggregate data which honors and meet HIPPA patient privacy and security requirements. There is a need for an automated conversion anonymized computational engine for health data Analytics.

2.5 Delayed Transition from ICD-9 to ICD-10
According to Health Catalyst, "In April 2014, Congress gave the entire healthcare industry in the U.S. a reprieve — a one-year ICD-10 delay before providers will be required to document care and submit payment invoices using ICD-10 codes. The new deadline is October 1, 2015." Most Health Providers are still using ICD-9 in 2016, others have adopted a strategy of using both ICD-9 and ICD-10 as they seek to convert the new 68,000 Code - ICD-10 standard from the 13,000 Code - ICD-9 System. There is a need for an automated conversion transition from ICD-9 to ICD-10.

2.6 Cybersecurity of Health Data and Systems
The year 2015 marked a heightened increase in cybersecurity attacks in the health sector. The industry saw attacks ranging from Denial of Service to Crypto locker to identity theft to false payment filings based on stolen medical records [1]. The Cybersecurity of Health Records is extremely important and there is need for a systematic approach to putting in place cybersecurity measures to counter attacks coming from different vulnerabilities.
2.7 Increased Multidisciplinary Collaboration and Research Initiatives
Increasingly Governments, Patients, Health Insurance Companies, Health Care Providers and Research Communities are demanding more multidisciplinary approaches to healthcare including nutrition, physical exercise, mental health, addiction counseling, early screenings and environmental monitoring, industrial safety and health standards, food and drug safety. This increased demand for accountability, openness and the right to know requires that more and more data be made available in the right format to different groups without violating privacy, security and safety rules.

These new requirements exceed the requirements set forth by HIPPA and HL7 among other standards and will require a new approach to Health Information Systems [2]. We now proceed to propose a data model and architecture which addresses the concerns raised above.

3.0 Introduction to Elliptic Curve Cryptography and Digital Signatures
From earlier work we have done [3], we will use the following definitions to lay the groundwork for the DFT method.

“An Elliptical curve may be defined as an equation of the form \( ay^2 + bxy = cx^3 + dx^2 + ex + f \), where a, b, c, d, e, f, x and y are for cryptographic purposes restricted to each belong to a finite field i.e. a, b, c, d, e, f, x and y are each chosen from a distinct set of integral values [2, 10].

The Elliptical curve provides desirable properties of simple and straightforward encryption computation. The inverse operation is intractable and very difficult to compute [4]. We can define a rule for adding two points \( S_1 \) and \( S_2 \) on the curve to find a third point \( S_3 \). These points are all on the curve thus forming an Abelian group [4]. The trivial case of infinity also needs to be included. The order of the curve is defined as the number of distinct points which satisfy this condition including the infinity point as follows:

\[
\prod_{t=1}^{m} S_t = \sum_{t=1}^{m} S_t
\]

\( S_3 = S_1 + S_2, \quad S_4 = S_3 + S_2, \quad S_5 = S_1 x S_2, \quad S_6 = S_1 x S_2 x S_3 \)

If we set \( b=0 \) in the equation \( ay^2+bxy=cx^3+dx^2+ex+f \) i.e. \( ay^2=dx^2+ex+f \), with conditions:

(i) \( 4a^3+27b^2 \neq 0 \)
(ii) \( b \neq 0 \)

The Discrete Logarithm Problem: At the foundation of every cryptosystem is a hard mathematical problem that is computationally almost infeasible to solve [3]. The discrete logarithm problem is the basis for the security of many cryptosystems including the Elliptic Curve Cryptosystem (ECC). Specifically, ECC relies on the difficulty of Elliptic Curve Discrete Logarithm Problem (ECDLP).

There are two geometrically defined operations over certain elliptic curve groups. These two operations are point addition and point doubling. By selecting a point in an elliptic curve group,
one can double it to obtain the point 2S. After that, one can add the point S to the point 2S to obtain the point 3S. The determination of a point mS in this manner is referred to as Scalar Multiplication of a point. The ECDLP is based upon the intractability of scalar multiplication products. In the multiplicative group $\mathbb{F}_p$, the discrete logarithm problem is:

Given elements $r$ and $q$ of the group, and a prime $p$, find a number $k$ such that $r = qk \mod p$. If the elliptic curve group is described using multiplicative notation, then the elliptic curve discrete logarithm problem is:

Given points $S_1$ and $Q$ in the group, find a number that $S_1 k = Q$; $k$ is called the discrete logarithm of $Q$ to the base $P$. When the elliptic curve group is described using additive notation, the elliptic curve discrete logarithm problem is:

Given points $P$ and $Q$ in the group, find a number $k$ such that $S_1 k = Q$. It is widely believed that the elliptic curve discrete logarithm problem is hard to computationally solve when the point $P$ has large prime order. The known methods for solving the ECDLP are [2]:

- The Pohlig-Hellman algorithm (which reduces the problem to subgroups of prime order).
- Pollard's methods (especially the parallel Pollard method of van Oorschot and Wiener).
- The Menezes-Okamoto-Vanstone (MOV) attack using the Weil pairing.
- The Frey-Rueck attack using the Tate pairing.
- The attacks on anomalous elliptic curves (i.e., elliptic curves over $\mathbb{F}_p$ which have $p$ points) due to Semaev, Satoh-Araki and Smart and Weil descent (for some special finite fields) [3].

ECDSA is short for Elliptic Curve Digital Signature Algorithm uses an Elliptic Curve Cryptograph to cryptographically sign (mark with a unique key generated by a complex mathematical operation) a piece of data, file or other object so that third parties can verify the authenticity of the signature without the Signer revealing how the signature was created.

At least two secret keys are used in the signing (private key) and verification process (public key). ECDSA has separate procedures for signing and verification. Each procedure is an algorithm composed of elliptic curve arithmetic operations [4]. The management of public and private keys is done by a trusted authority which can either be centralized or distributed depending on the protocol chosen for a particular application. The mathematics and computer implementation of elliptic curve cryptographic systems is well understood and known to be very secure.

4.0 Introduction to Block Chain Cryptography

The Modified Generalized Block Chain Model

There are two main kinds of participants: Consumers and Producers.

A Producer produces a unit good or a unit service costed at a certain number of digital currency units. A Consumer consumes certain goods or services valued at a certain number of digital currency units [5]. A Virtual System Manager (Distributed Ledger Manager) holds and
manages Digital Currency Units. Each Participant is both a producer and consumer. Every time a producer produces a unit good or a unit service and publishes it on the system, the System Manager creates digital currency equal to the value of the unit good or unit service.

Each participant in this model has a bunch of public/private key pairs to sign messages. With these messages they can sign "TRANSACTIONS", where a "TRANSACTION" is a fulfillment of a need between two parties (e.g. Purchase of a good, purchase of a service, provision of data requested, authentication of a file, etc).

**Scenario 1: A consumer wants to receive a good or service from a producer of a good or service.**

How do we know that a Consumer participant wants to, and can carry out a transaction? The Consumer must reference a previous transaction where they received a good, a service or got their need fulfilled, who fulfilled (Producer) the need, the time and how much they paid (could be a digital currency or some point value system). Initially, the system purchases or is assigned a certain amount of digital currency or points from the Virtual System Manager who issues payments to Producers. The Producer creates and publishes the good, or service that they offer, the price of such a good or service (in some digital currency or point system), the terms and conditions of provision of the good or service (smart contract) [6]. The good or service is cryptographically protected and cannot be accessed by a Consumer unless the Producer agrees to conduct a transaction with the Consumer. The details of the Transaction include at least the following: Previous Transaction Encrypted Identifier, Public Key of Producer, private key signature of the Consumer with the public key of the previous transaction which initially is a key issued by the Virtual System Manager.

**Scenario 2: How do we know that the digital currency (or Points from the Value System) from the previous transaction have not already been spent and is being re-used?**

Every so often say (5 Minutes), the Virtual System Manager iterates over the list of all transactions and combines them into a block. These blocks are distributed across the network to each Producer and Consumer who store them in a chronological chain [7]. Each Producer and Each Consumer compares this block with the previous block to ensure that no transaction is repeated. If any transaction is found to be repeated, all repeated transactions are suspended until arbitration determines which of the repeated transaction is correct. If the repeated transactions cannot be determined to be correct, all repeated transactions are rejected as null and void.

**Scenario 3: How do we ensure that the blocks are not altered illegally?**

Each block references the previous block using a cryptographic hash function which then forms a chain of blocks or Block Chain. If a hash value does not match, it means that the block chain has been altered illegally.

**Scenario 4: How can we guarantee that blocks are not replaced altogether?**

The creation of a block must be made computationally expensive so that the time it takes to
create a block and distribute it to all consumers and producers is more than the time it takes to check the validity of blocks (say 5 minutes). This can be achieved using cryptographic hash collision algorithms.

**Scenario 5: How are Cryptographic Hash Collision Values determined?**
It is mathematically possible to find a new Hash Collision Number every so often, perhaps once a week or once every two weeks such that the hash value of each block is always less than the Hash Collision Number ensuring that only one block can be created every say 5 minutes. This takes care of changing network size.

**Scenario 6: What happens when production does not match consumption?**
The laws of demand and supply come into play. Goods and services cost more. The converse is also true.

**Scenario 7: How is depreciation and obsoleseness determined?**
Different accounting methods can be used.

5. **Application of Modified Block Chain Technology to the Cybersecurity of Health Data Management Systems**

The basis of this proposed block chain health data security model is the design and analysis of a model for the systematic protection and defense of health data networks. The model focuses on the phases of: Reconnaissance, Identification of Vulnerabilities and Threats, prediction, prevention, detection and defense of attacks [8]. Risk levels, algorithms, procedures, defense mechanisms and the analysis of all the data collected and computed is also modeled. Among the health data management issues raised, Block Chain Cryptography seems to be very well suited to protection, transmission, storage, serialization, anonymization, distribution and auditing of health data. If a block chain monitoring and alert system is added to a Health Data Management System, the security of health data can be tremendously increased.

At the heart of this model is a **Block Chain Based Monitor and Alert System** which manages the entire security system. The comprehensive security model is designed to be flexible allowing for the different modules to be configured to operate different levels of alertness. Highly secure systems may turn all modules on at red alert level while others may turn only selected modules at lower levels of alertness. Security Indices for each unit will be computed. These indices include identity, time, location category, group, cluster, risk, severity, success, failure, frequency, scale, center of gravity, distribution, and propagation, measures of central tendency, measures of variation, trend, projections, correlation, behavior and pattern [9].

**The Block Chain Based Monitor and Alert System**

The Monitor and Alert System can be designed using rigorous and secure programming techniques to handle configuration, module management, security policies, security mechanism and internal and external interface connection. The system could sit inside routers, firewalls,
anti-virus programs, operating systems, server software, network management software, desktops, laptops and mobile devices [10]. In the initial design, we propose the monitor and alert system as a generic prototype system with options for specific tailoring in future.

**Current Reconnaissance and Vulnerability Assessment Methodologies**

A network penetration test, colloquially called (PenTest), is a method of investigating the vulnerabilities of a computer system or network by simulating several attacks from malicious external and internal attackers. Potential vulnerabilities are investigated, detected and exploited to gain entry into the system. These vulnerabilities may be due to poor or improper system configuration, both known and unknown hardware or software flaws at various levels from network, hardware, operation system to application levels, and operational/administrative weaknesses in procedural and/or technical specifications and implementations. The analysis is carried out from a potential attacker’s point of view [11]. Security flaws discovered during the process are presented to the organization in a report. Effective penetration testing combines this report with an accurate assessment of the potential impact to the organization, of different levels of attacks. A comprehensive range of technical and procedural countermeasures to reduce risks is also recommended by the report.

**Limitations of the Current Vulnerability Assessment Methodologies**

The limitation of these methodologies is the high level of dependence on manual steps and the frequent response after the fact rather than before the attack. In this Block Chain design we propose an automated statistical methodology to identify and analyze a set of computer systems’ security loopholes by identifying higher-risk vulnerabilities that result from a combination of lower-risk vulnerabilities exploited in a possible set of sequences using an automated reconnaissance process [12].

**Reconnaissance Vector Unit (RVU)**

Reconnaissance is the act of identifying a potential target and compiling a list of all of their strengths and weaknesses. In terms of computer security, one would look at the system infrastructure and attempt to determine all security protocols in an effort to produce or construct some type of system to bypass one or more administrative, procedural and/or access control systems in place. Reconnaissance is an intelligence gathering phase both internally and externally. The Block Chain Methodology can be used to monitor, track and record reconnaissance activities.

The Reconnaissance phase is modeled by block chain stochastic vectors. Additionally, we propose to completely automate the process of carrying out reconnaissance for known vulnerabilities and attempt to develop new vulnerability discovery algorithms based on the secure programming paradigm [13]. A Reconnaissance Risk Vector is associated with each reconnaissance action to measure the probability of success and the ranking and priority of the sequence of reconnaissance actions. Among the key reconnaissance vector algorithms to be investigated and developed are Automated Reconnaissance, Automated Hidden Vulnerabilities Discovery, Exploitation Possibility, Human centered Reconnaissance, Reconnaissance Information Sharing and Reconnaissance Indices. The Reconnaissance Vector Unit could be
programmed to run continuously or periodically depending on the Vulnerability, Threat and Attack level settings in the Monitor and Alert System.

**Vulnerabilities Vector Unit (VVU)**
The vulnerabilities gathered during the reconnaissance phase are used to populate the vulnerabilities vectors divided into internal and external categories. These are further sub-divided by origin, type, modus operandi, risk, severity and how it affects systems. Additionally, the vulnerabilities are further classified as follows: Vulnerabilities known to have been exploited and Vulnerabilities not know to have been exploited.

Among the key research areas to be specifically investigated and developed are Vulnerabilities Dynamics, Propagation, Testing, Risk, Severity, Alert and Information Sharing Algorithms. Every vulnerability is a risk, and poses either an active or passive threat to a system. The next unit examines how vulnerabilities populate the threat matrix [14].

**Threat Matrix Unit (TMU)**
A threat is an exploitable vulnerability which an attacker can use to gain access to a computer system and/or cause harm to it. This harm could include the theft of valuable information, the introduction of viruses and other malware, the disruption of systems or programs, and the corruption or wiping out of vital data, information and or programs among others. During the Reconnaissance phase, one reconnaissance action can lead to the discovery of more than one vulnerability. For each Reconnaissance Row Vector \( R \) of size \( N \), whose unit action is \( r[i] \), we define a column vector \( V \) of size \( M \), whose unit vulnerability is \( v[j] \). The Threat Matrix is populated by an \( M \times N \) matrix of vulnerabilities. These are subdivided into Active Threats and Passive Threats. In this research, an **Active Threat** is defined as vulnerability with a known attack history or vulnerability for which there are current attacks on computer systems. A **Passive Threat** is defined as a vulnerability whose attack history is unknown, and for which there are no known current threats on computer systems. A number of stochastic and statistical Threat Matrix Algorithms will be investigated and developed for both Active and Passive Threats. Specific attention will be paid to Enumeration, Risk, Severity, Ranking, Frequency, Clustering, Location, Propagation, proximity, Center of Gravity, Scale, Timing, Modus Operandi, Levels and Indices among others. Threat Matrix Algorithms can be run on a continuous or periodic basis, based on the settings in the Monitor and Alert System. Having identified threats, the next course of action is to prevent them from actualizing.

**Prevention Unit (PVU)**
The Prevention Unit maintains a matrix of preventive actions using the Threat Unit to identify and close vulnerability loopholes in the system. There are a number of algorithms to be investigated and developed including Solved Active/Passive Threat Algorithms, Unsolved Active/Passive Threat Algorithms, Threat Propagation Prevention Algorithms, Threat Latency Analysis Algorithms and Prevention Indices.

**Detection Unit (DTU)**
The Detection Unit gathers all cases of failed prevention from the Prevention Unit and runs intrusion detection algorithms, detected remnant active and passive threats and Detection Indices.
**Prediction Unit (PDU)**
All remnant active and passive threats are considered predicted attacks at this stage. The Prediction Unit runs a series of prediction analytics to determine attack likelihood measures for time, location, sequence, severity, frequency, location, proximity, trend, level and Prediction Indices. This information is passed on to the Decision and Communications Unit, and the Defense, and Attack Units.

**Attack Unit (AKU)**
The Attack Unit maintains three kinds of lists namely: (1) all predicted attacks reported by the Prediction Unit, (2) all attacks the system has actually received (reported by the Defense Unit) and (3) all other known attacks gathered from Attack Information Sharing Sources from other security systems. The Attack Unit runs a number of algorithms including the following: Attack Scenario Analysis, Combinations, Sequences, Durations, Frequencies, Severity, Scale and Level and indices. This information is passed to the Decision and Communications Unit and also to the Defense Unit [15].

**Defense Unit (DFU)**
The Defense Unit maintains a matrix of Defense Actions which include inoculation programs, encryption programs, update and software patch routines, quarantine programs, escalation of security and access privileges for certain resources, shutting down of certain systems, blocking of communications channels and a host of other possible defense mechanisms. If the Defense unit is interfaced with physical systems, defense actions could include, activating physical alarms, starting fire extinguishers, locking or opening doors or safes and starting backup. We will investigate and develop programs for the integration of defense actions into this system.

**Computational Algorithms Unit (CAU)**
The Computational Algorithms Unit provides computational services to all the modules of the stochastic cybersecurity matrix model. These computational algorithms can be regularly updated and can run in real time or periodically based on need. Figure 6 below show the Computational Algorithms Unit and some of the algorithms we will investigate and develop. Algorithms already in use will be combined in new ways to support the stochastic cyber security vectors, matrices and indices.

**Learning and Feedback Unit (LFU)**
Data and information from the other modules of the security system will be fed into the Learning and Feedback unit in order to improve the reliability, efficiency and learnability of the system. An investigation of various learning algorithms will be carried out. These include among others, the algorithms shown in Figure 1 below:
Inferential Engine and Data Mining Unit (IDU)

The Inferential Engine and Data Mining sub-units assist the Decision and Communications Unit in making informed decisions. Decision making in a stochastic scenario with only partial information requires validation of conclusions using a variety of inferential techniques. Automated decision making tools will be used in combination in order to increase the likelihood of making correct automated conclusions.
Data Mining techniques and prediction analytics will also be used to gather and organize security related data and information from a variety of information sources. This data provides a rich basis for improving all units of the security system when fed to the LFU and then back into the system.

**Decision and Communications Unit (DCU)**

The Decision and Communications Unit is the brain of this security model. All units in the security system send actionable data to the DCU Unit.

The Decision and Communications Unit (DCU) makes decisions using the IDU and LFU units and communicates its decisions to all internal and external units connected to it. It also receives and shares information with other Secure Cyberspace Monitoring and Alert Systems (SCMAS).
REFERENCES


http://www.robertowor.com/csci4911


https://www.trustwave.com/


http://books.google.com/books?id=hJOF2w5iOMC&pg=PA188&dq=types+of+reconnaissance+attacks&hl=en&sa=X&ei=IVt4UYyLBBTE0AGNYG4Ag&ved=0CC8Q6AEwAA#v=onepage&q=types%20of%20reconnaissance%20attacks&f=false.


