Data Cooperatives for Neighborhood Watch

Abiola Salau
Department of Computer Science and Engineering
University of North Texas
Denton, TX, 76207, USA
AbiolaSalau@my.unt.edu

Ram Dantu
Department of Computer Science and Engineering
University of North Texas
Denton, TX, 76207, USA
Ram.Dantu@unt.edu

Kritagya Upadhyay
Department of Computer Science and Engineering
University of North Texas
Denton, TX, 76207, USA
KritagyaUpadhyay@my.unt.edu

2021 IEEE International Conference on Blockchain and Cryptocurrency (ICBC) | 978-1-6654-3578-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICBC51069.2021.9461056

***Abstract*—The increasing proliferation of user data is moving the world from the era of 'big data' to a new era of shared data, and some are considering data as a factor of production that is paving the way for a new business economy. In this paper, we propose a solution that uses blockchain technology as a platform for online neighborhood watch using a form of data cooperative among individuals or organizations in the sharing of data through a peer-to-peer mechanism. We prove the concept by implementing a distributed phishing data sharing system that will maintain a community ledger of reported phishing activities with a consensus-based approval of the phishing transaction and a novel reputation scoring system thereby adding reliability to the system and effectively tackling the phishing problem. The data cooperative provides a way for timely multi-party sharing of phishing data among anti-phishing organizations and users of the internet eliminating the current approach of each organization maintaining its database. Our results show that blockchain is effective in complementing the existing methods of phishing detection and serves as a platform for sharing phishing data with respect to scalability, cost, and memory consumption. Also, our results further show that transaction times on the Ropsten test net follow a Gamma distribution. Our approach can be extrapolated to other data sharing systems like medical data, spam calls, discussion forums, etc.**

***Index Terms*—Cybersecurity, blockchain, phishing, distributed ledger, data cooperative, smart contract, data sharing, peer-to-peer.**

I. Introduction

Real-time sharing of data and information is becoming critical in the fight towards a more secure online practice as shared data is the backbone of the knowledge economy. Sharing the right information at the right time in a systematic way with the right stakeholders permits the effective protection of assets, intellectual property, and business processes [1]. Despite the common consensus that individuals own their data, in most cases, these individuals do not have control over the data and companies collect, store, use, and even sell these data for personal gain. With the volume of data generated daily by users of the internet and especially the social media networks, individuals are increasingly concerned about the privacy of their data and how the social network platforms handle their data [2] [3]. The decline in trust is increasingly evident and the World Economic Forum (WEF) in the report on rethinking personal data [3] gave three recommendations. (i) Redefining transparency policies on data practices in a way that is more understandable and relevant to the individuals (ii) improving accountability by ensuring stakeholders are held accountable throughout the value chain and (iii) giving the individual the power to have a say in how their data is being used and have the capacity to use the data for their purposes.

With these three recommendations from WEF, we propose in this paper a concept of distributed data ledger for real-time neighborhood watch. The concept of a data neighborhood watch system, a form of data cooperative based on trust, on the blockchain such that we can have an online community of people come together with a common goal of sharing data and information for the safety and benefit of its members. The system will integrate the various aspect of data sharing like sharing of spammers' phone numbers [4] or emails addresses, phishing data sharing [5], societal news, discussion groups within an organization, and even re-engineering of social media networks such that the menace of misinformation and spreading of fake news can be tackled by verifying the credibility of a news post by a consensus algorithm before being disseminated to the members of the community. The system will also integrate algorithms for phishing, spam, and fact-checking to verify the credibility of news posts or social media messages. A data cooperative is a member-owned legal organization constructed to collaborate in the pooling of data for the benefit of members with trust in the use of its data [6] [7] [8].

A. Problem Statement and Motivation

A lot of data is generated daily by end-users. This data is often not in the control of the owner and sometimes not accessible to them. Today, data is seen as a central part of businesses and governments as an increasing number of businesses get driven by data, and governments rely on data for economic and infrastructural development. For instance, an effective solution to the eradication of the COVID-19 pandemic requires data in many different facets e.g for contact tracing, the design of AI models for the prediction of trends, manufacture of vaccines, etc [10].

Unfortunately, these kinds of data needed for research and development are in the hands of only a few people and as

978-1-6654-3578-9/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on April 15,2023 at 08:16:21 UTC from IEEE Xplore. Restrictions apply.

such, the data owners are at the mercy of those with their data to come up with solutions.

Empowering end-users with control over their data and having them involved in how their data is being used will improve the diminishing trust between end-users and social networks and provide more opportunities for improved availability of data to researchers. To illustrate, a patient in most cases does not have unrestricted access to its medical data [8]. If the patient visits a new medical provider, a request will have to be made to the former provider to send the patient's records and in some cases, the patient may have to repeat some procedures at a fee if there is no access to the records.

TABLE I
Some Existing Anti-Phishing Products [15]

Product	**Approach Used**	**Mode of Operation**
AntiPhish	Restricted form filling	Stand-alone
B-APT	Machine Learning	Stand-alone
eBayAccount Guard [13]	Blacklist, heuristics	Server
McAfee Site Advisor [12]	Rates the site with their own tests	Server
Microsoft smart screen filter	Blacklist, heuristics	Server
PhishTank site checker	Open database	Server
Web of Trust (WOT)	Blacklist, crowdsourcing	Third-party
Verisign EV green bar extension	Domain popularity	Server
Virtual browser extension	Blacklist, heuristics, visual similarities	Third-party
Netcraft	Blacklist, heuristics, user rating	Stand-alone
Passpet	Restricted form filling	Server
SpoofGaurd	Heuristics	Stand-alone
TrustWatch	Blacklist	Server
PhishProof	Blacklist, Whitelist, Heuristics	Server
GoldPhish [14]	Visual Similarities	Third-party

Table 1 shows some existing phishing detection products including their mode of operation and the approaches adopted by each mode.

From table 1, the modes of operation presented either require a central server to be up to date with the latest phishing data or a database has to be constantly updated which may not be timely, failing to detect some phishing instances whereas, in a data cooperatives, a community as a whole decides who is a spammer because the fingerprints and features vary city to city or country to country. More so, it is not transparent to the user how the algorithm decided to classify a website/an email address as phishing or not. Users must be able to visualize the series of transactions that affected the reputation of a website/email address and be able to trace the origin of a phishing email. Also, the integrity of such data must be preserved, and any third-party provider should not be able to change the data [4].

Various approaches have been used in the detection and prevention of phishing leading to the existence of several anti-phishing products available today as seen in table 1. Most of these approaches, however, only focus on the detection and prevention of phishing but not on the effective sharing of phishing data which can reduce the cost of updating the individual databases the anti-phishing solutions maintain since the phishing data will be readily available. Blockchain technology is a possible solution to an efficient sharing of data to enable timely broadcast of phishing data to all participants and improving the overall cost of maintenance and quick disposal of phishing activities.

B. Why Blockchain

Some of these advantages are discussed below with respect to this work.

1. Immutability: Data stored on the blockchain cannot be altered unlike when stored on a centralized server which may even be maintained by a third-party. This is important so that no participating peer can alter the reputation score of a website or email address and eliminates third-parties.

2. Transparency: In a phishing detection system based on reputation, users' trust is important. With blockchain technology, phishing transactions reported can be stored on the ledger, accessed, and reviewed by all the participants. All the participants can read not only the final state of transactions but also the history of past states. This visibility builds trust among the users and improving users' participation in the phishing sharing system.

3. Distributed: The distributed nature of the blockchain, allowing every member access to data and transactions, makes all the reported phishing transactions readily available to peers promptly. Unlike in a centralized server where a failure results in unavailability of data. Multiple phishing detection techniques can be used to update the ledger, which can be used to calculate email addresses, website ratings. The feature vectors for phishing can be SMTP paths, SMTP relays, IP addresses of phishers, time of phishing, length of the phishing emails, and more [16].

4. Audit-ability: Every phishing transaction reported can be audited by users before it becomes part of the ledger. Users can trace the history of past states and the current state of a transaction. The data is only posted after validation by a mechanism such as proof of work preserving the integrity of data present on the ledger. Thus, regardless of the phishing detection mechanism adopted by a user off-chain, he will be able to update a global phishing ledger and retrieve reputation scores of email addresses and websites already on the blockchain at any moment they require it.

C. Our Contributions

An objective of this work is to evaluate the blockchain as a transparent, secure, and cost-effective platform for sharing of data among sharing data among members of a community such as an enterprise, a city, or a category of people such

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on April 15,2023 at 08:16:21 UTC from IEEE Xplore. Restrictions apply.

as patients with cancer, etc. We took a look at the problem of phishing in online settings and implement a blockchain-based ledger for the sharing of phishing data that complements the existing phishing detection mechanisms. Despite groups like the APWG and the Openphish platform monitoring and making available phishing intelligence to the public, [11] reports that an average phishing attack costs an organization $3.86 Million and an average email user receives 16 malicious emails per month. This shows that the problem of phishing is still very devastating and the current approaches are insufficient. More so, current platforms are centralized and are subject to a single point of failure leaving users to no information or wrong information if the platform is attacked or unavailable. The decentralized and distributed architecture of the blockchain makes it a viable solution against such attacks.

This work does not aim at providing a phishing detection algorithm but it;

1. provides a complementary solution to the existing approaches of phishing detection,
2. evaluates the performance of the blockchain as a platform for peer-to-peer sharing of phishing data
3. provides a novel reputation tracking scheme for participating peers sharing the phishing data.

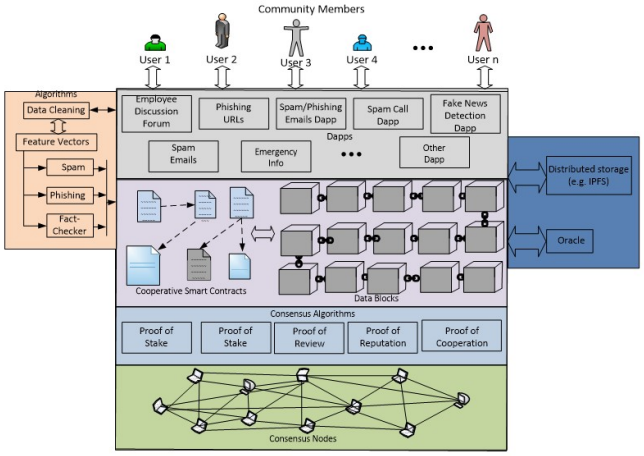

Fig. 1. Overview of the Distributed Neighborhood Watch System

II. Literature Survey

Paper [17] classified anti-phishing products into two major categories, the content-based and the non-content-based. It discussed that content-based phishing detection involves analyzing the content of a website such as checking the HTML code of the website, the grammar and spelling, JavaScript content of the web pages, etc. Zhang H, et al in their paper [17], applied a Bayesian model approach to the content-based phishing detection while [18] used TF-IDF (Term Frequency/Inverse Document Frequency) to retrieve information about webpages. The non-content-based category simply focuses on other attributes of a web page rather than the content. A behavior-based approach was discussed in paper [19].

Machine Learning (ML) and AI algorithms have been well used in the detection and classification of phishing websites and emails. Papers [20] [21] used supervised ML algorithm in the detection of phishing or malicious URLs. Some other Anti-Phishing solutions found in literature use pattern matching in the detection of phishing URLs where the DNS information of a URL is verified to identify malicious content [22] [23]. The rule-based mining approach was investigated in paper [24]. An approach using case-based reasoning was presented in paper [25].

Since the development of blockchain technology, researchers and cybersecurity professionals have been exploring the advantages of the inherent attributes of blockchain technology in the prevention, detection of phishing emails, and sharing of phishing data among peers. A blockchain-based anti-phishing solution can detect phishing activity at the DNS level because blockchain has its naming system like Namecoin, Bitforest, etc. in addition to the decentralized and distributed nature which makes every participating peer have a copy of the ledger making it easy to update [15].

In paper [5], the authors presented a phishing data sharing mechanism based on Hyperledger fabric. The mechanism used four different types of nodes in the blockchain network; the reporting nodes, accounting nodes, servicing nodes, and supervising nodes. The authors do not regard the 'citizens of the net' as reporters because they believe citizens may become rogue and spam the system with invalid reports. The mechanism sets a supervision cycle of 30 days before each reporting node gets its performance score and consequently gets penalized if need be. This is more than enough time for a rogue node to have spam the network with false reports and also there is no performance check for the supervising nodes themselves which may not be good for the system as they could get compromised.

Like phishing emails, the use of spam calls is another approach adversaries use to trick users to divulge personal information. Paper [4] presented a blockchain-based ledger for the logging of spammers' phone numbers such that taking advantage of the distributed nature of the blockchain, a spammer's phone number is swiftly broadcast on the network for all participating nodes to validate. However, it does not consider the possibility of a user falsely reporting a phone number as spam which may be a major cause for concern as a user maliciously report fake spam activities thereby spamming the system.

Paper [26] presented an email protocol based on blockchain whereby the sender of an email pays a processing cost using crypto-currency in form of a sending deposit. This deposit is returned if the email is received normally at the receiving end but if it was a malicious email then the sender forfeits the deposit. It claims the possibility of losing the sending deposit will prevent spammers from sending spam emails.

Our work differs from the works mentioned above by

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on April 15,2023 at 08:16:21 UTC from IEEE Xplore. Restrictions apply.

addressing the lapses identified in their approaches. We incorporate the citizens of the net as a potential reporter since they are also a target of phishing attack [11] in addition to other corporate enterprises and organizations that tackle phishing, while also providing a novel approach for checking and penalizing any rogue user through a reputation scoring and penalty scheme. This methodology can be applied to other resource-sharing systems like medical data sharing, consumer news cooperative, spam calls, etc.

III. System Architecture

A. Participants in the Blockchain

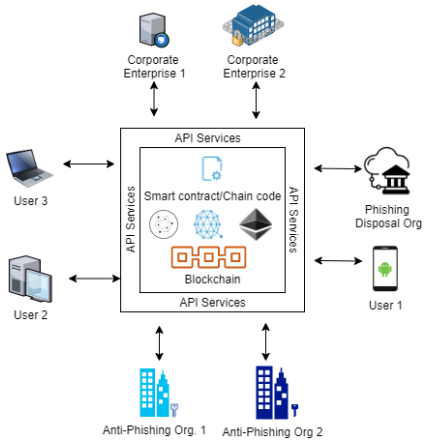

Fig. 2. Participants in the Blockchain

The blockchain comprises users of email services with an email address, corporate enterprises with stakes in internet security such as financial institutions, anti-phishing organizations such as APWG, and phishing disposal organizations such as security software manufacturers. Each node will be able to communicate with the blockchain through an HTTP web client using REST API to report a phishing email address or query the blockchain for the reputation score of an email address or website. The API services will be designed using a web3.js client in Node.js. Through this, the user can connect to the blockchain to post phishing transactions or request reputation scores. Ganache is used as the blockchain during the development stage of the smart contract, but performance results are measured on the Ropsten test net. Metamask plug-in is used to interact with the blockchain for user accounts.

B. Smart Contract

To achieve the goal of peer-to-peer phishing data sharing, we implemented three key functions in the smart contract.

One function registers email addresses of new users, the second function contains the logic for the report phishing transaction and verification algorithm while the third function computes the reputation score for the reporting email address and the reported email address. The logic flows implemented in the smart contract are detailed in Figure 3 and Figure 4.

- ***Registering Email Addresses***. A participating node registers its email address with the service creating a profile on the blockchain. An initial reputation score of 1 is assigned to the profile the registration flag is set to 1.

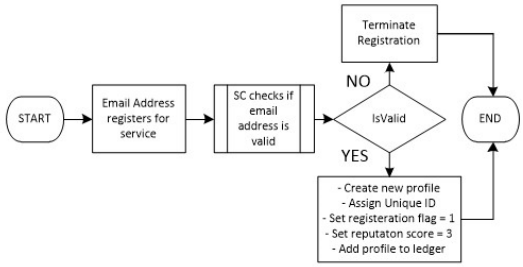

Fig. 3. Flowchart for New User Registration

The email address can be verified by sending a verification email to that address and the user confirming before the profile is created to avoid robots spamming the system. The profile includes details such as the email address, reputation score, domain of service provider, user's unique ID on the blockchain, and the number of phishing transactions reported.

- ***Phishing Reporting.***

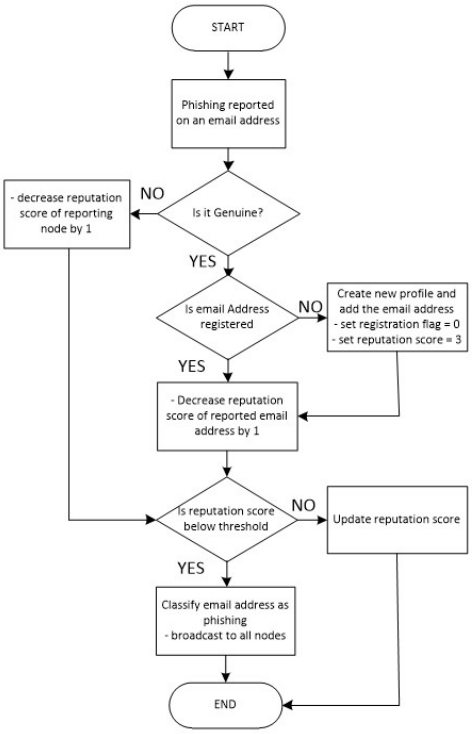

Fig. 4. Flowchart for Report Phishing and Reputation Score Penalty

The phishing activity is reported in a predefined format, "Reporting Node UserID", "Reported Email Address", " Reported Email Address Domain", "Hash of Email from

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on April 15,2023 at 08:16:21 UTC from IEEE Xplore. Restrictions apply.

IPFS". The flow of events in a report phishing transaction is shown in Figure 4.

A reporting node can be an average member of the cooperative, enterprise companies such as banks, anti-virus, and anti-phishing solution providers or organizations like APWG. These organizations and companies by default incline to tackle phishing as a way to protect their customers and their brand. They do not require any special incentives to report phishing activities. However, in practice, the reputation system will motivate the users to share phishing emails since their reputation will be stagnant if they do not, and reward shared from the joining fee of new users are shared based on members' current reputation score.

The reporting node uploads the entire email received/to be reported to IPFS and only submits the link(hash) to the stored email to the blockchain.

- ***Phishing Verification***. To verify the credibility of a phishing email reported, a smart contract logic written in solidity is used to retrieve the email from the IPFS storage using the hash stored on the blockchain, features common to phishing emails [27] [28] [29] like the number of hyperlinks in the body of the email, number of dots in the domain if the message ID domain matches the sender domain, if the URL contains'@', the message size, number of attachments, number of receivers, image maps used as hyperlinks, E-Shape analysis, etc. are then extracted from the email header and body. The information gained is aggregated using the different features identified from the email and the decision is made as either phishing or non-phishing based on a threshold value.
- ***Reputation Score and Penalty***. Reputation score obtainable is an integer initially starting at 1 for self-registered users and 0 for reported phishing users. This reputation score increments by an additive factor of 1 for every validated phishing report made by the user and decrements by a multiplicative factor of 2 for every false phishing email reported. So, when a user reports a phishing transaction, its authenticity is verified by as discussed above before the reputation scores are updated. This ensures that a reporting node does not maliciously report a non-phishing email address. If over 51% of the consensus nodes approve the transaction, the reputation score of the reporting node is increased and that of the phishing email address is decreased. The transaction is then added to the block and broadcast to all participating nodes but if the transaction is found to be malicious on the part of the reporting node, it is penalized by reducing its reputation score by dividing by 2 and that of the reported email address remains as it was before the transaction then added to the block.

IV. Experimental Setup

The phishing data sharing application is developed on the blockchain network and a smart contract is written using solidity. A graphical user interface is developed for interaction with users and for this to work, all participants will have a copy of the blockchain running on their computer and the GUI will be created using HTML/CSS/JavaScript/React. Web3.js will be used to interact with the local or remote Ethereum node.

Technology software/tools that were used for this project include: i) Remix Web IDE ii) Truffle.js iii) Web3.js and Web3.py iv) Node.js v) Ganache vi) Ropsten test net vii) Solidity language

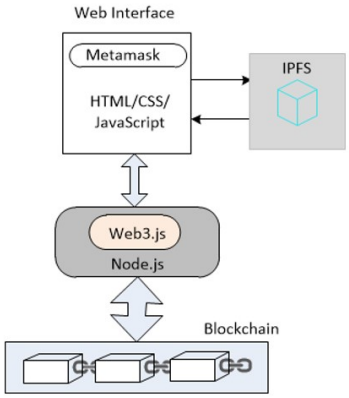

Fig. 5. Test Setup Diagram

The Dataset is a combination of a collection of more than 2500 Fraudulent emails and another collection of about 12000 Fraud email dataset already pre-classified as phishing and non-phishing from CLAIR collection of fraud email [30]. The data set include features such as sender email add, return path, reply-to, date, subject, mailing platform used, etc.

V. Results and Discussion

Different experiments are carried out on the Ropsten test net and results are collected to investigate the performance of the blockchain as a platform for peer-to-peer sharing of phishing data. In our previous work [4], results obtained showed that the performance on the Ropsten test net is identical to what is obtainable on the Ethereum main net with only increased transaction receipt times on the main net that can be attributed to the volume to transaction in the pool at the time of making the transaction. Metrics such as the cost of executing the application, running time of the execution, and how well the application scales with an increasing number of users, ledger memory consumption are measured.

The results presented in this section are preliminary and will be revised based on production experience. In subsequent works, We plan to do further analysis on the performance of different reputation schemes and also further optimize the system to improve on the results and its performance.

A. Time and Gas Cost

To examine the time taken for a phishing data sharing transaction, we set up the experiment and deployed using Ropsten Test Net as the blockchain and a dataset comprising 25 phishing emails and 25 non-phishing emails are selected

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on April 15,2023 at 08:16:21 UTC from IEEE Xplore. Restrictions apply.

from the two available datasets. Ten user profiles are created and registered, and each user reported different phishing emails.

For each of these transactions, we record the transaction receipt times as well as the gas used. We also recorded transaction receipt times for a query transaction to retrieve the reputation scores for different users.

Observations:

The transaction time varied between 3.81 seconds to 39.14 seconds with an average of 19.07 seconds for the entire data sample to report a phishing email while the time taken for the reading of reputation score of an email vary between 60 ms to 90 ms with an average of 70 ms over a data sample of 50 email addresses.

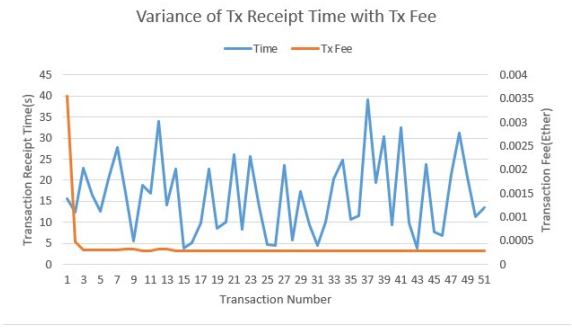

Fig. 6. The Variance of Transaction Receipt Time Vs Transaction Fee on Ropsten Test Net

The observed fluctuation in the transaction time of figure 6 may be attributed to the number of pending transactions in the transaction pool as well as the time it takes to mine the block.

The fee per transaction remained somewhat constant after the initial smart contract deployment fees indicated by transaction number 1 of Figure 6. This is attributed to the use of a hash pointer to the email content on IPFS rather than the email text itself which size is a factor in the volume of gas used. Using a pointer to the file on IPFS reduced cost because the IPFS hash is a string of constant length of 46 characters regardless of the size of the file uploaded [31].

Transaction Fees was calculated using the formula,

$$Transaction\,Fee = GasPrice \times GasUsed$$

Where Gas Price = 0.000000002 Ether (2 Gwei) [32].

Table 2 shows a summary of the gas cost for the smart contract deployment and the two write functions in the smart contract. The smart contract was deployed once and it consumed 1765074 Wei of gas while the phishing reporting function consumed gas ranging from 142453 Wei to 239139 Wei and the register email function consumed gas ranging between 41025 Wei to 74374 Wei.

TABLE II
Gas Used Per Function

Function	Average Gas Used(Wei)	Minimum Gas Used(Wei)	Maximum Gas Used(Wei)
Contract Deployment	1765074	-	-
Phishing Reporting	146667	142453	239139
Register Email	52318	41025	74374

The time taken to retrieve a reputation score from the Ropsten test net is close to real-time which is important because a user needs to be able to know the reputation of the sender of an email. Even as data stored increases and the blockchain grows, since the unique email IDs are the lookup keys on the ledger, it will be easy and fast for any user to verify the reputation of any email address on the blockchain. In addition, the delay associated with retrieving data from the blockchain is negligible as can be seen from Figure 7.

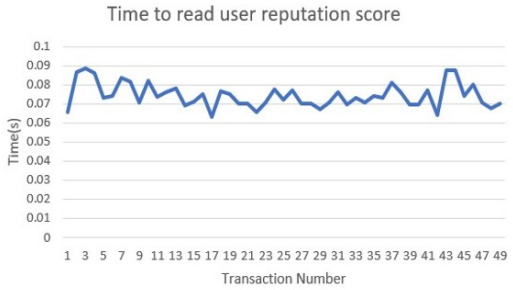

Fig. 7. Time taken to read a reputation score for a queried userID. .

The next set of experiments is aimed at investigating the statistical distribution of the transaction receipt times of each of the transactions on the blockchain network. Since the time taken for a transaction to be mined completely on the blockchain often are not dependent on the application itself, other factors like network issues, the total number of pending transactions, and the competitiveness of the gas price offered by the reporter also affect the transaction times.

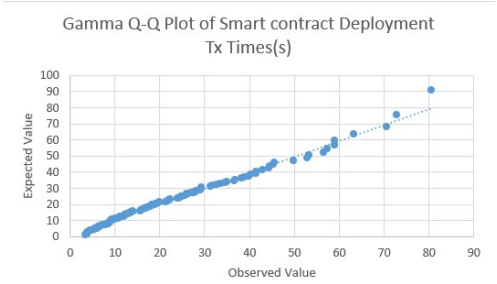

Fig. 8. Q-Q Plot for the Smart Contract Deployment Transaction Times

The purpose of the experiment is to observe the trend of the transaction receipts times on the blockchain network and be able to forecast based on the observed probability distribution,

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on April 15,2023 at 08:16:21 UTC from IEEE Xplore. Restrictions apply.

the duration a transaction may take to complete. For this experiment, we ran 114 instances of the contract deployment and other functions in the smart contract and examined the distribution of the transaction receipt times on a Q-Q plot, and then plotted its probability distribution curve.

The Q-Q plot is a technique used to informally visualize whether a set of random samples plausibly came from some theoretical distribution such as a Normal or Exponential [33]. For instance, if we suspect after running some statistical analysis that the dependent variable seems to be normally distributed, plotting the data on a normal Q-Q can help confirm the suspicion.

Observations:

The Q-Q plot of Figure 8 shows that the transaction receipt times distribution is a gamma distribution and the probability density function curve is plotted on the distribution histogram in Figure 9. The gamma distribution is a right-skewed, continuous probability distribution with two positive parameters, α and β corresponding to shape and scale respectively [34] [35]. Gamma distributions are particularly useful in real-life scenarios where the data sample naturally has a minimum of 0 like time in our case. The gamma distribution probability density function (PDF) is given as:

$$P(x) = \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}$$

where,

$$\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}\ dt\ for\ values\ of\ x > 0$$

$$\alpha \approx \frac{\mu^2}{variance}\ and\ \beta \approx \frac{variance}{\mu}$$

The result shows that given α and β values for some random transaction times on the blockchain network, we may be able to predict the behavior of the network and consequently forecast the completion time for new transactions. To validate this observation, we repeated the same experiment for the main functions of the smart contract running 114 instances of each function, and similar behavior was observed in the probability distribution of the transaction completion times.

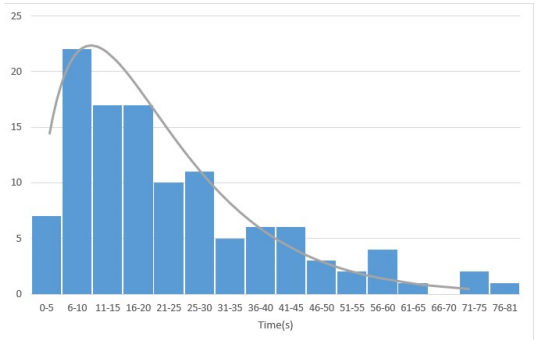

Fig. 9. Gamma Distribution Plot for the Smart Contract Deployment Transaction Times

B. Scalability

To examine the scalability of the Enterprise-Wide Phishing Data Sharing Dapp on how well it responds to an increase in the number of users to memory consumption, gas consumption, and transaction receipt times, we created 100 user accounts on the Ropsten test network and measured the gas consumption as the number additional profiles are added onto the blockchain network. We started with the creation of 10 accounts, repeated the experiment for 20 accounts till we finally repeated it for the 100 user accounts.

Observations:

239311 Wei of gas was consumed in the creation of 10 user accounts simultaneously and it was observed that as the number of users on the network increases, the gas consumption increases linearly.

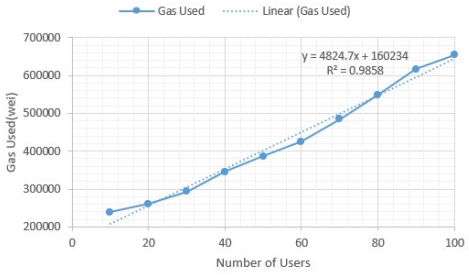

Fig. 10. Gas Consumption for Creating New User Profile

To further examine the effect of the number of simultaneous users on the transaction times as phishing emails are being reported we conducted another experiment. For this experiment, we started with 1 user and took the user through the steps of successfully registering and reporting a phishing email onto the blockchain with a gas price offer of 2 Gwei. If the process completes successfully with the gas price we offered for the transaction, we add 1 more user but if the process fails to complete, we increase the offered gas price by 2 Gwei and repeat the experiment.

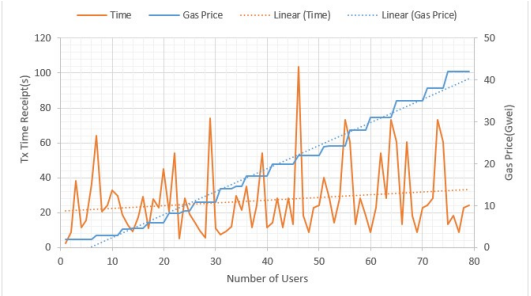

Fig. 11. Variance of Transaction Receipt times as Number of Users Increase

It was observed that the experiment required more gas after every addition of about 5 new users. It can be seen from the plot of Figure 11 that the gas price is directly proportional to the number of users.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on April 15,2023 at 08:16:21 UTC from IEEE Xplore. Restrictions apply.

The number of users may linearly affect the time taken to process each phishing transaction. However, there are additional variations likely due to other network issues, such as the total number of pending transactions and the competitiveness of the gas price offered by the reporter. These additional variations are reflected in the sudden spikes in the plot of Figure 11.

C. Memory Consumption and Money Conservation

From the APWG phishing activity trends quarterly reports from the last quarter of the year 2015 to the third quarter of the year 2020, February 2016 had the highest number of unique phishing emails reported by consumers with a total of 229,315 emails [36]. Statistics of the average daily transactions globally on the Ethereum blockchain network reported by Statistica, reports an average of 11,922 Transactions per day during the third quarter of the year 2020 transacted on the Ethereum blockchain globally [37]. If we compute the memory consumption of the transactions based on the data for the month with the highest number of unique phishing email reported in the APWG report, that is about 229,315 unique phishing e-mail reports, this gives an average of 7,644 reports daily which is much lower than the average number of daily transactions from the Statistica report.

In this work, the maximum gas used for a phishing report transaction cost 239139 Wei and default transaction payments from the Ethereum yellow paper [38], show that 21000 Wei is paid for every transaction, and 16 Wei is paid for every non-zero byte of data or code for a transaction. Hence, we can compute the memory consumed by a transaction on the blockchain by

$$(239139 - 21000)/16 = 13.6kB$$

Therefore for 7,644 reports, memory consumed will be

$$13.6kB \times 7644\, reports = 103.9MB$$

of reports per day. Although the result shows that our application can handle this amount of daily unique phishing emails report, we expect a fewer number of emails to be reported since it is an enterprise application and only its citizens can share phishing data on it. Thus, having a similar average on our solution over 5 years, we will have a ledger growth of about 6.5GB of data on the blockchain.

The user profile stored on the blockchain only contains details such as its user ID, email address(restricted to 64 characters), user account address, registration flag, reputation score, and phishing transaction counts. Since we know the amount of gas used to create and store a user profile from Figure 10, we can compute the memory size of a profile on the blockchain using the transaction fees from [36].

$$(23931 - 21000)/16 = 183B$$

Thus, 183 Bytes of memory is required per user profile on the blockchain and for 100,000 users registered on the network, the ledger will only grow by 18.3 MB (100,000 * 183 Bytes) of storage space.

VI. Conclusion and Future Work

Blockchain technology provides a cost-effective and scalable platform for a data cooperative among members that can be used in the sharing of data and information. Our results show that the blockchain added functionalities to the existing techniques of phishing detection and promotes a peer-to-peer phishing data sharing mechanism with decentralization, auditability, and transparency. Users will be able to share and update reputation scores without the involvement of a third-party service provider. We discuss conclusions on various aspects of the system below.

Ledger Access Times: The read times of the email's reputation scores were achieved in real-time, which is important for the blockchain to facilitate users to identify the email address as a phishing email address when an email is received from the address.

Concurrency of transactions: From the APWG report, in June 2020 [36], a total number of 44,497 unique phishing emails were received from consumers. On average, 1 phishing email is received per minute. According to the Ethereum Transaction growth chart [32], the highest number of the 1,406,016 transactions occurred on Thursday, September 17, 2020, achieving a transaction rate of 16.27 tx/sec. This shows that the Ethereum blockchain can serve as a preferred platform for peer-to-peer phishing data sharing.

Transaction Approval Times: The transaction approval times when a phishing email is reported is a variant that cannot be determined for certain since it directly depends on the miner activity and several other transactions running on the blockchain. Acceptable averages of approximately 19.07 seconds were achieved on the Ropsten test net. This can be addressed by increasing the gas price. However, this is not a limitation as this need not be in real-time.

Memory Consumption: The ledger size increases with an increasing amount of reported phishing activity on the blockchain. This was handled by using IPFS [31] to store the email data while the blockchain stores only the hash of the data. The blockchain still maintains the data about the reputation score of users.

Future Work. We would like to further extend our work to evaluate the Additive Increase and Multiplicative Decrease(AIMD) approach to computing the user reputation scores based on the paper [9]. Another approach discussed in [39] to compute reputation on P2P networks where reputations are computed using either a debit-credit reputation computation (DCRC) or a credit-only reputation computation (CORC) will be explored. Also, we aim to evaluate the security of the system against possible attacks like a Sybil attack where a rogue user can register on the platform with different email addresses thereby spamming the system with false reports. A possible candidate solution that has been well used in P2P systems is the trusted certifying authority (CA), where the CA helps to validate the authenticity of a user before joining the network [40].

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on April 15,2023 at 08:16:21 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] World Economic Forum,(October 2020) "Cyber Information Sharing: Building Collective Security,"[Online], http://www3.weforum.org/docs/WEF-Cyber-Information-Sharing-2020.pdf

[2] M. Madden, "Public Perceptions of Privacy and Security in the Post-Snowden Era," November 2014,

[3] World Economic Forum, "Rethinking Personal Data: A New Lens for Strengthening Trust," May 2014, http://reports.weforum.org/rethinking-personal-data. http://www.pewinternet.org/2014/11/12/public-privacy-perceptions/.

[4] A. S. Muttavarapu, R. Dantu and M. Thompson, "Distributed Ledger for Spammers' Resume," 2019 IEEE Conference on Communications and Network

[5] Dongjie Liu, Wei Wang, Yang Wang, and Yaling Tan. 2019. "PhishLedger: A Decentralized Phishing Data Sharing Mechanism". In Proceedings of the 2019 International Electronics Communication Conference (IECC '19). Association for Computing Machinery, New York, NY, USA, 84–89. DOI:https://doi.org/10.1145/3343147.3343154

[6] A. Penland, D. Shrier, T. Hardjono, and I. Wladawsky-Berger (2016) "Towards an internet of trusted data: A new framework for identity and data sharing."[Online] MIT Connection Science.

[7] Data Futures. " Research to shift power through data governance" [Online]. Available: https://foundation.mozilla.org/en/initiatives/data-futures/data-for-empowerment/what-is-a-data-cooperative/ [Accessed 16 Dec 2020]

[8] Hardjono, T. and Pentland, A. (2020). 4. Empowering Innovation through Data Cooperatives. In Building the New Economy. https://doi.org/10.21428/ba67f642.0499afe0.

[9] Z. Zaccagni and R. Dantu, (2020). "Proof of Review (PoR): A New Consensus Protocol for Deriving Trustworthiness of Reputation Through Reviews". Cryptology ePrint Archive, Report 2020/475,

[10] Smith, Charlotte D, and Jeremy Mennis. "Incorporating Geographic Information Science and Technology in Response to the COVID-19 Pandemic." Preventing chronic disease vol. 17 E58. 9 Jul. 2020, doi:10.5888/pcd17.200246

[11] CloudPhish (2020). https://cloudphish.com/wp-content/uploads/2020/09/Cloudphish_WhitePaper.pdf [Accessed 25 Feb 2021]

[12] SiteAdvisor: MCAfee Site Advisor. (2006). https://en.wikipedia.org/wiki/McAfeeSiteAdvisor. [Accessed Oct 7, 2020].

[13] eBay Toolbar and Account Guard. http://pages.ebay.in/help/account/toolbar-account-guard.html. [Accessed 7 Oct 2020]

[14] Dunlop M, Groat S, Shelly D. "Goldphish: using images for content-based phishing analysis". In: 2010 Fifth international conference on internet monitoring and protection. 2010. pp.123–128. https ://doi.org/10.1109/ICIMP .2010.24

[15] S. Chanti, T. Chithralekha, "Classification of Anti-phishing Solutions", SN COMPUT. SCI. 1, 11 (2020). https://doi.org/10.1007/s42979-019-0011-2

[16] S. Palla (2006). A Multi-Variate Analysis of SMTP Paths and Relays to Restrict Spam and Phishing Attacks in Emails.

[17] Zhang H, Liu G, Chow TW, Liu W. "Textual and visual content-based anti-phishing: a Bayesian approach". IEEE TransNeural Netw. 2011;22(10):1532–46. https ://doi.org/10.1109/TNN.2011.21619 99

[18] Zhang Y, Hong JI, Cranor LF. Cantina: "A content-basedapproach to detecting phishing web sites". In: Proceedings of the16th international conference on world wide web. WWW '07,ACM, New York, NY, USA, 2007. pp. 639–648. https ://doi.org/10.1145/12425 72.12426 59.

[19] Arun Vishwanath. (2017). "Getting phished on social media". Decis. Support Syst. 103, C, 70–81. DOI:https://doi.org/10.1016/j.dss.2017.09.004

[20] Hajgude, J. and L. Ragha. "Phish mail guard: Phishing mail detection technique by using textual and URL analysis"." 2012 World Congress on Information and Communication Technologies (2012): 297-302.

[21] Mohammed Al-Janabi, Ed de Quincey, and Peter Andras. 2017. "Using supervised machine learning algorithms to detect suspicious URLs in online social networks." In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (ASONAM '17). Association for Computing Machinery, New York, NY, USA, 1104–1111. DOI:https://doi.org/10.1145/3110025.3116201

[22] Naresh, Undeti, U. Sagar and C. Reddy.(2013) "Intelligent Phishing Website Detection and Prevention System by Using Link Guard Algorithm." IOSR Journal of Computer Engineering 14 : 28-36.

[23] Usuff Rahamathunnisa, Manikandan, N., Kumaran U.S. and Niveditha, C.. (2017). "Preventing from phishing attack by implementing url pattern matching technique in web." International Journal of Civil Engineering and Technology. 8. 1200-1208.

[24] Jeeva, S.C. and Rajsingh, E.B.2016. "Intelligent phishing URL detection using association rule mining." Human-centric

[25] Abutair, Hassan and Belghith, Abdelfettah. (2017). "Using Case-Based Reasoning for Phishing Detection." Procedia Computer Science. 109. 281-288. 10.1016/j.procs.2017.05.352.

[26] Nakayama, Koichi and Moriyama, Yutaka and Oshima, Chika. (2018). "An Algorithm that Prevents SPAM Attacks using Blockchain" International Journal of Advanced Computer Science and Applications. 9. 10.14569/IJACSA.2018.090729

[27] S. Smadi, N. Aslam, L. Zhang, R. Alasem and M. A. Hossain, "Detection of phishing emails using data mining algorithms," 2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), Kathmandu, Nepal, 2015, pp. 1-8, doi: 10.1109/SKIMA.2015.7399985.

[28] Sroufe Paul, (2009), E-Shape Analysis, (Unpublished Master's Thesis). University of North Texas, TX

[29] Palla, Srikanth (2006), A Multi-Variate Analysis of SMTP Paths and Relays to Restrict Spam and Phishing Attacks in Emails, (Unpublished Master's Thesis). University of North Texas, TX

[30] Radev, D. (2008), "CLAIR collection of fraud email," ACL Data and Code Repository, ADCR2008T001, Available: http://aclweb.org/aclwiki

[31] Juan Benet "IPFS - Content Addressed, Versioned, P2P File System https://github.com/ipfs/papers/raw/master/ipfs-cap2pfs/ipfs-p2p-file-system.pdf

[32] Ethereum (ETH) Blockchain Explorer [Online]. Available: https://etherscan.io/chart/tx

[33] B. Das and S. I. Resnick (2008) "QQ Plots, Random Sets and Data from a Heavy Tailed Distribution, Stochastic Models," 24:1, 103-132, DOI: 10.1080/15326340701828308

[34] William L. Hosch. (2017), "Gamma distribution"[Online]. Available: https://www.britannica.com/science/gamma-distribution [Accessed 14 Dec 2020]

[35] Stephanie Glen. "Gamma Distribution: Definition, PDF, Finding in Excel" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/gamma-distribution/ [Accessed 14 Dec 2020]

[36] APWG (2020), Phishing Activity Trends Reports," [Online]. [Accessed 19 Feb 2021]

[37] Statistica (2020), "Number of daily Ethereum transactions worldwide from 1st quarter 2016 to 3rd quarter 2020,"[Online]. Available: https://www.statista.com/statistics/730818/average-number-of-ethereum-transactions/ [Accessed 17 Nov 2020]

[38] Gavin Wood (2001), ETHEREUM: A Secure Decentralised Generalised Transaction Ledger Petersburg Version. Available at https://ethereum.github.io/yellowpaper/paper.pdf

[39] Minaxi Gupta, Paul Judge, and Mostafa Ammar. (2003), "A reputation system for peer-to-peer networks". In Proceedings of the 13th international workshop on Network and operating systems support for digital audio and video (NOSSDAV '03). Association for Computing Machinery, New York, NY, USA, 144–152. DOI:https://doi.org/10.1145/776322.776346

[40] B. N. Levine, C. Shields, and N. B. Margolin (2006), "A survey of solutions to the Sybil attack", University of Massachusetts Amherst, Amherst, MA.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on April 15,2023 at 08:16:21 UTC from IEEE Xplore. Restrictions apply.