International Journal of Pure and Applied Mathematics
Volume 118 No. 20 2018, 391-396
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Anomaly Detection In Blockchain Using Clustering Protocol

RACHANA KUMARI
M.TECH (ISCF)
SRM UNIVERSITY
Kattanakulathur, India
ruhikumar129@gmail.com

MONICA CATHERINE
Assistant Professor, IT Department
SRM UNIVERSITY
Kattanakulathur, India
monicacatherine@ktr.srmuniv.ac.in

Abstract—**With the growth of technology every day, new technologies are coming and evolving day by day. Blockchain is one such technology. Blockchain, one of the revolutionary technology, has proved its potential of being used in numerous fields. The fields include digital money or crypto-currency, IoT, Product tracing, Smart Contracts etc. Blockchain is a distributed database which allows us to share or process data between multiple parties over a network of non-trusted users securely. One of the biggest advantage of blockchain is that it is fully decentralized that means that there is not any particular authority who has hold of it or is governing it. However, blockchain has its own set of disadvantages also. Some of these problems are its complexity, network size, large energy consumption etc. But one of key problem is that there is no way to find the anamolous nodes i.e nodes which may be malicious. In blockchain, while most of the nodes behave normally there are chances that few nodes may try to do some illegal activity in the network or may have some illegal interest. For this reason, we need to monitor the behaviour pattern of each node. But if we try to this manually foe each node, it will take a tremendous amount of time and effort and is nearly impossible to do.**
So, in this paper we introduce a novel solution for the above-mentioned problem. This problem can be solved by clustering the nodes of blockchain network. For this we will propose an algorithm which will help us in clustering the blockchain and then further for analyzing the malicious activity of nodes, if any performed. As we know, clustering means dividing the nodes or data points into a number of group depending or based on some similar traits that they may have. Therefore, the aim of our project is to segregate groups having similar traits from the blockchain network and then to cluster them so that to identify the malicious node or any illegal activity.

Keywords—Blockchain, Cluster, Clustering Algorithm, K-means clustering, Security

I. INTRODUCTION

With the growing world, everyday new and advanced technologies are coming and one such technology is blockchain. Blockchain is a decentralized and distributed database that allows to maintain the list or inventory of all the transactions that has happened or is going to happen. The term given to this ever-growing list or record is known as Block. Each and every blockchain contains the following things:

1. A timestamp
2. Relevancy to the previous block and the transaction that has happened since the last block.
3. A hash value of the previous block

In a blockchain network, since nothing is centralized, each and every node has associated degree of equal rights. Blockchain maintains a ledger of each transaction that has ever occurred and also have multiple backup of that ledger so as in case of some failure, the backup can be retrieved easily and there is no data loss.

There are two types of blockchain network:

1. Public Blockchain: In this anyone in the world can join this network and can become a node of the network.
2. Private Blockchain: In this people who have access and can join the network only can become a node.

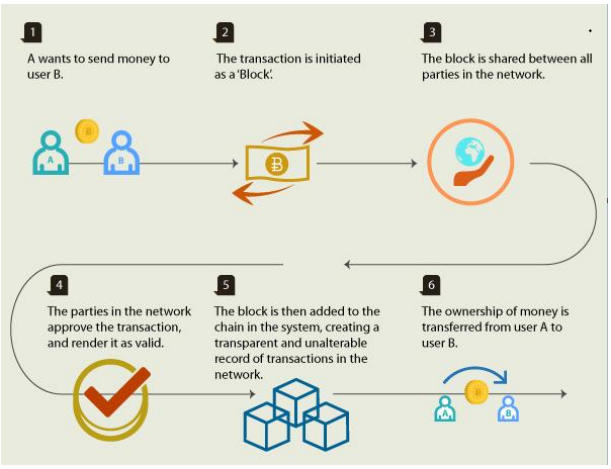


Fig: Working of Blockchain

According to the performance of nodes in blockchain transactions, nodes can be divided into honest nodes, and malicious nodes.

1. Honest nodes. When providing service to other nodes, nodes with a higher behavioural stability and reputation value, which can ensure the correctness of data forwarded to other nodes, are known as honest nodes.
2. Malicious nodes. Nodes juggle and throw away data, and even worms or Trojans are disguised as data to be forwarded to others. These behaviours of nodes can severely damage the security and stability of blockchain communication. Nodes that disseminate false data and demolish resources are known as malicious nodes.

Blockchain promises for a revolutionary technology, that has potential to look out applications in varied fields. Bitcoin module changes the currency in the future. In all the fields they have the block chain technology. Some node could manipulate the transactions of the previous one and can have abnormal pattern that can affect the one block. Its very hard to manipulate the whole network because the transaction will be happening in different blocks

It takes a lot of time to read the behaviour of the patterns of nodes. So, we propose a cluster management to analyse the behaviour patterns of all the nodes that transactions r happening. after the formation of clusters, we can select the template of patterns by using the template we can analyse the strange behaviour template by comparing the cluster template.
By making the cluster templates we can able to manage the nodes and transactions and networks and also can organize the nodes in correct form.
We can conduct experiments by to measure the effectiveness of the algorithm and their existing modules. It will show how accuracy the module is effective with the previous one

II. RELATED WORK

Storing the behavior patterns of all the nodes in a clustering network, there are two steps. First step is to capture each and every node behavior. The aim of the project is similarity measure.

A. Sequence Similarity measure

Similarity measure goal is to identify the similarity of two sequence patterns. Popular measures have been produced to find the similarity. Some of them are:

a) Euclidean distance (EDR) is one of them. Euclidean distance requires two sequences to have same length. It uses L2 norm to compute the distance between two nodes having same length.

b) Dynamic time wrapping (DTW) is similar to EDR but in DTW there is no need the sequence lenghth to be same or equal it can compute the distance of two sequence having different lengths also. Also, it can handle time shifting. To overcome this, DTW duplicates the value of previously obtained elements and then it calculates the optimal value between the sequences.

c) Longest Common SubSequence (LCSS) technique uses a threshold value to manage the noise in a sequence, which later tries to find the longest common sub-sequence between the two sequence.

A. Clustering approaches

There are different clustering methods that has been proposed over a time. But majorly they are classified into four. They are:

1. Partition Based Clustering Method
2. Density Based Clustering Method
3. Grid Based Clustering Method
4. Hierarchical Clustering Method

In partition-based clustering method, we need to divide the r labelled data into n portion and then every partition needs to corresponds to a cluster having a cluster centroid. It uses Distance based similarity measure to group the nodes into cluster. Also, it is more effective for small to medium size data set.

Hierarchical clustering makes a tree of cluster or in general term cluster hierarchy. Every cluster contains a child node, a sibling node having a common parent. This results in exploring data on different levels of granularity.
There are two types of hierarchical methods:

a) Agglomerative: This one is also known as bottom up. This method starts with one point and subsequently starts to merge two or more nodes and continues till all nodes are in a single cluster or until the criteria (requested no. of k clusters) is reached.

b) Divisive: This method is exactly opposite of how agglomerative works. This approach works as top down. In this one cluster is recursively split into the most appropriate cluster of smaller cluster.

But one of the major disadvantage of hierarchical clustering is that once executed, any split or merge is unchangeable.

Density based method [4] generally starts by forming a cluster of one node and then continuously keeps on absorbing all the neighbouring nodes till the density (the threshold value) is not reached. DBSCAN, GDBSCAN, OPTICS and DBCLASD is some of the commonly used algorithm of Density based Clustering Method.

Grid Based Clustering method [1,12] works by forming a grid structure of a finite number of cells. Then the next step in this is to calculate the cell density of each cell. After then the sorting of cells according to the cell density is done and this

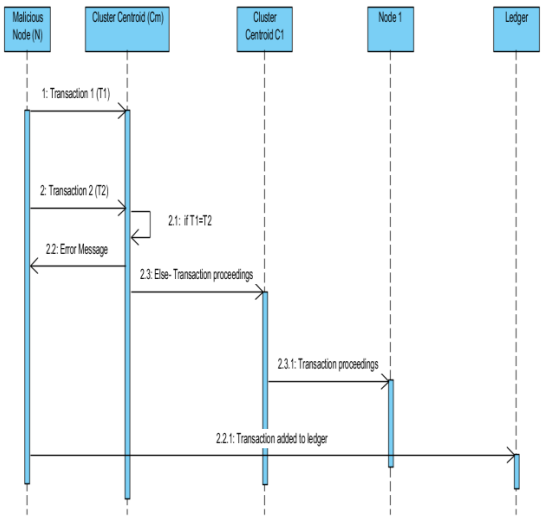
helps in identifying the cluster centres and forming a cluster properly. A n example of grid-based approach can be STING. In STING, the spatial area is divided into rectangular cells represented by hierarchical structure. These cells correspond to different levels of resolution and the information about each cell is pre-computed and is stored separately.

III. PROPOSED SOLUTION

We know that even though blockchain technology can prevent fraud and anonymous behavior, it cannot detect fraud by its own. Attacker can find one way or the other to steal currency and can commit various fraudulent crime. Also with the increase in number of attacks or vulnerabilities, the developers are trying to come up with new and innovative technique and technologies to strengthen the blockchain and also to detect the attack prior to its occurrence. Techniques such as machine learning and data mining algorithms can be very much helpful in finding and detecting malicious behavior, behaviors. So, we also will use one among these technologies to provide security against fraud and malicious behavior.

Here, in this paper we propose a system in which the blockchain will be clustered in groups based on its behavioral pattern. To analyses the behavior, the parameter that we are choose as our pattern is the time taken for one transaction and the amount involved from one node to another node. The reason for selecting this parameter is that usually the transaction amount is the most important and predominant feature of a node. For doing this the algorithm that will be used is K-means algorithm. Also, we will not use the same algorithm but a little modified version of this algorithm.

A. Sequence Diagram



The diagram shows how the clusters objects interact with each other. Through this diagram we can understand how the clustering will help in securing the blockchain. Let us assume that there is a malicious node (M) and its cluster centroid is Cm. If M initiates two transaction say T1 and T2 at the same time (attack called double spending attack) to two different nodes, then the Cm will get to know about it as the transaction will be added to the block. When Cm gets to know that malicious activity is being performed by the node M, it will discard both transactions and show an error message.

If the transaction is to happen between two genuine users and the node is not doing any malicious activity, then the transactions will be sent from the Cm to the Cluster centroids of the other node i.e. the node which has to accept the transaction. And that centroid will proceed the transaction to the node.

After all this the transaction whether valid or invalid will be added to the distributed ledger, which keeps record of all the transaction happened ever.

B. Methodology

In this section we propose the problem defined. In order to similarity measure of the cluster sequence. We first implementing the k-means algorithm and improve the features and enhance k-means algorithm.

1. Similarity Measure Selection

To find the similarity between two sequences, there are various selection techniques such as Euclidean distance (EDR), Dynamic time wrapping (DTW), and Longest Common Sub-Sequence (LCSS) (as discussed in section II-A)

But for our project we will be using DTW as this technique uses the sequence of different lengths. This is because of the reason that in a Blockchain all the blocks are of different lengths and needs not to be of equal length. Hence, we will be using DTW.
The reason why we cannot use EDR is obvious as it takes the sequence of same length and that is not possible in Case of Blockchain. And LCSS can't be uses because there is no noise between any of the sequence in a Blockchain. But LCSS techniques is majorly based on noise handling od sequence.
Therefor the possibility of using EDR and LCSS is zero.

K-Means Clustering Algorithm
K-means clustering is unsupervised learning algorithm that helps in solving the clustering problem. This algorithm divides the elements of a data set into K different groups based on similarity to one another. The main idea behind this algorithm is to define k-centroids, one for each cluster.
Below we have explained how the algorithm works:

Current Algorithm Summary:

1. First step is to initialize K points randomly which will be called as means.
2. Each object should be assigned to a group that have the closest centroid.
3. After all the objects has been assigned to a group, we need to recalculate the centroids position.
4. Step 2 and Step 3 should be repeated till the centroids stops to move. This results in separation of objects into groups in which the minimized metric can be calculated.

C. Proposed Improvisation:

- As seen in the above algorithm, the cluster centres or centroids are assigned randomly, but we don't need that for our algorithm, what we need is to sequentially sorted list of k centres.
- Another problem with the above-mentioned algorithm is that it uses Euclidian Distance between the tuples which are static, but we need dynamic distance measure, so we will be using DTW distance between the sequences.
- The above-mentioned algorithm uses the regular or the mean value of the node of each cluster to find the cluster head or as we say centroid. But that is not what we want from our algorithm, so we will select a sequence that will have the least amount of distance from its[n/k]th nearest neighbour from all the other sequences in the cluster as the centroid.

IV. RESULT

OUTPUT (BLOCKCHAIN CREATION)

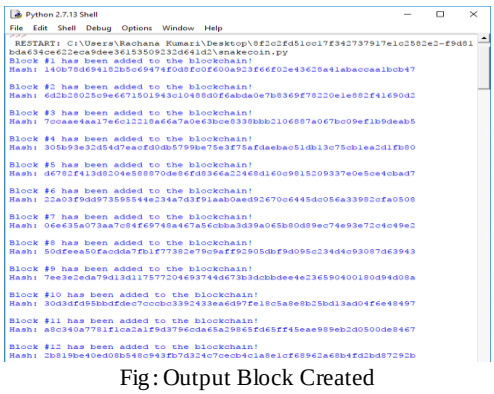


Fig: Output Block Created

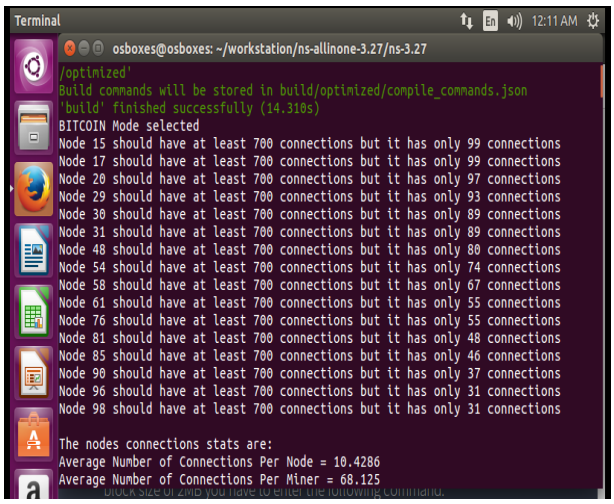


Fig: Node and Connection Creation

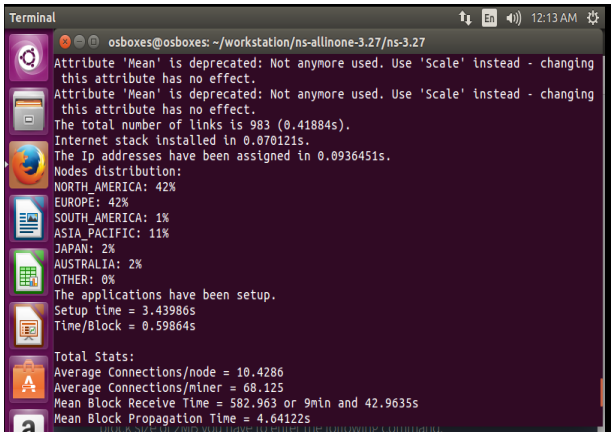


Fig: Node Deployment

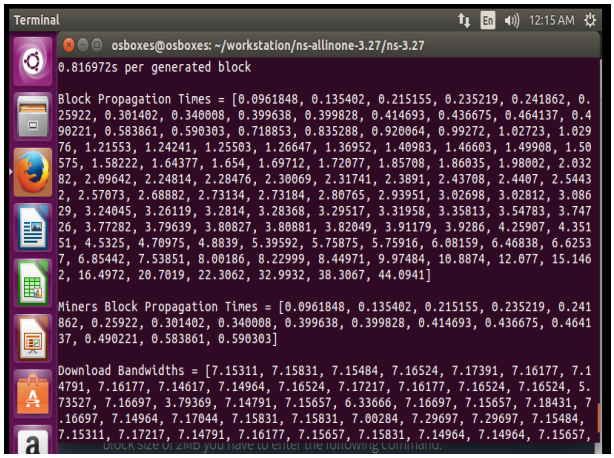


Fig: Block and Miner Propogation

CONCLUSION

This paper basically discusses about the problem of malicious activities that occur in blockchain network and then tries to solve this problem by using clustering protocol. For this reason, we need to monitor the behavior pattern of each node. But if we try to this manually foe each node, it will take a tremendous amount of time and effort and is nearly impossible to do. For doing the clustering, K-means clustering protocol has been used. But some improvisation was need in that algorithm. Hence an improvised version of k-means algorithm is implemented. This will make the blockchain more secure from any kind of illegal or anomalous behaviour.

REFERENCES

1. Christidis K, Devetsikiotis M (2016) Blockchains and smart contracts for the internet of things. IEEE Access 4:2292–2303
2. Croman K, Decker C, Eyal I, Gencer AE, Juels A, Kosba AE, Miller A, Saxena P, Shi E, Sirer EG, Song D, Wattenhofer R (2016) On scaling decentralized blockchains - (a position paper). In: Financial cryptography and data security - FC 2016 international workshops, BITCOIN, VOTING, and WAHC, pp 106–125
3. Dorri A, Kanhere SS, Jurdak R (2016) Blockchain in internet of things: challenges and solutions. arXiv:1608.05187
4. Ester M, Kriegel H, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the second international conference on knowledge discovery and data mining (KDD-96), pp 226–231
5. Garay JA (2015) Blockchain-based consensus (keynote). In: 19th international conference on principles of distributed systems, OPODIS 2015, pp 5:1–5:1
6. Guadamuz A, Marsden C (2015) Blockchains and bitcoin: regulatory responses to cryptocurrencies. First Monday 20(12)
7. Guha S, Rastogi R, Shim K (2001) Cure: an efficient clustering algorithm for large databases. Inf Syst 26(1):35–58
8. Hirano S, Tsumoto S (2003) Comparison of similarity measures and clustering methods for timeseries medical data mining. In: Data mining and knowledge discovery: theory, tools, and technology, pp 219–225
9. Karame G (2016) On the security and scalability of bitcoin's blockchain. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, Vienna, Austria, October 24–28, 2016, pp 1861–1862
10. Karypis G, Han E, Kumar V (1999) Chameleon: hierarchical clustering using dynamic modeling. IEEE Computer 32(8):68–75
11. Liao TW (2005) Clustering of time series data: a survey. Pattern Recogn 38(11):1857–1874
12. Morse MD, Patel JM (2007) An efficient and accurate method for evaluating time series similarity. In: Proceedings of the ACM SIGMOD international conference on management of data, Beijing, China, June 12–14, 2007, pp 569–580
13. Underwood S (2016) Blockchain beyond bitcoin. Commun ACM 59(11):15–17
14. Xu R, Wunsch II DC (2005) Survey of clustering algorithms. IEEE Trans Neural Networks 16(3):645–678
15. Zhang T, Ramakrishnan R, Livny M (1996) BIRCH: An efficient data clustering method for very large databases. In: Proceedings of the 1996 ACM SIGMOD international conference on management of data, pp 103–114
16. Zyskind G, Nathan O, Pentland A (2015) Decentralizing privacy: using blockchain to protect personal data. In: 2015 IEEE symposium on security and privacy workshops, SPW 2015, pp v180–184