Article

BNS: A Detection System to Find Nodes in Bitcoin Network

Ruiguang Li *, Liehuang Zhu, Chao Li, Fudong Wu *, Dawei Xu

Posted Date: 23 November 2023

doi: 10.20944/preprints202311.1477.v1

Keywords: bitcoin; reachable nodes; unreachable nodes; node activity; dicision tree model

Article

BNS: A Detection System to Find Nodes in Bitcoin Network

Ruiguang Li [1,2], Liehuang Zhu [1], Chao Li [2], Fudong Wu [3,*] and Dawei Xu [1]

[1] School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081, China
[2] National Computer Network Emergency Response Technical Team/ Coordination Center, Beijing 100029, China
[3] School of Cyberspace Science and Technology, Beihang University, Beijing 100191, China
* Correspondence: lrg@cert.org.cn, wufudong@buaa.edu.cn

Abstract: Bitcoin has been launched for over a decade and made an increasing impact on the world's financial order, which attracted extensive attention of researchers. Bitcoin system runs on a dynamic P2P network, containing tens of thousands of nodes including reachable nodes and unreachable nodes. In this article, a detection system BNS (Bitcoin Network Sniffer) was prososed, which could collect as many Bitcoin nodes as possible. For reachable nodes, the authors designed an algorithm BRF (Bitcoin Reachable-nodes Finding) based on node activity evaluation, which reduced the nodes to be detected and greatly shortened the detection time. For unreachable nodes, the authors trained a dicision tree model BUF(Bitcoin Unreachable-nodes Finding) to identify unreachable nodes based on attribute features from massive node addresses. Experiments showed that BNS performed better than the website "Bitnodes" in total number and efficiency. Based on the experimental results, the authors analyzed the real network size, node "churn" and geographical distribution.

Keywords: Bitcoin; reachable nodes; unreachable nodes; node activity; dicision tree model

1. Introduction

Bitcoin was first proposed by Satoshi Nakamoto in 2008 [1] and has been working steadily for over a decade. By now, it's the most successful cryptocurrency in the world. With the outbreak of COVID-19 in 2019, much currency flooded into the Bitcoin market and raised the Bitcoin's price which attracted more people to join the Bitcoin operation.

The Bitcoin system can be divided into the transaction layer and the network layer. Most previous studies had focused on the transaction layer, but less on the network layer. The Bitcoin network has the characteristics of decentralization and anonymity. Decentralization means there is no central organization or trust center in the network. Participants gain trust through message interaction. Anonymity means Bitcoin users' accounts and addresses are encrypted to ensure the privacy and security. All the transactions are stored in the block-chain in order of time and broadcast to all participants. Nodes in the Bitcoin network recorded all block-chain data. The decentralization and anonymity of Bitcoin brings difficulties to the supervision, because the transactions are anonymous and difficult to track. Therefore, it's worth making deep studies on the Bitcoin network.

The main contributions of this article are as follows:

1) The authors designed an detecting algorithm BRF (Bitcoin Reachable-nodes Finding) based on node activity evaluation, which greatly reduced the nodes to be detected and improved detection efficiency.

2) Using node attribute features, the authors trained a dicision tree model BUF(Bitcoin Unreachable-nodes Finding) to identify unreachable nodes from massive node addresses.

3) Based on detection experimental results, the authors analyzed the real network size, node "churn" and geographical distribution.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2023 doi:10.20944/preprints202311.1477.v1

2. Bitcoin Network

Bitcoin network is a typical P2P network which has no centralized organization and the topology is dynamically changed. Each node works independently according to the agreed protocols, shaking hands, broadcasting addresses, verifying transactions, packaging blocks and competing mining.The Bitcoin network is composed of both reachable nodes and unreachable nodes [2], as shown in Figure 1.

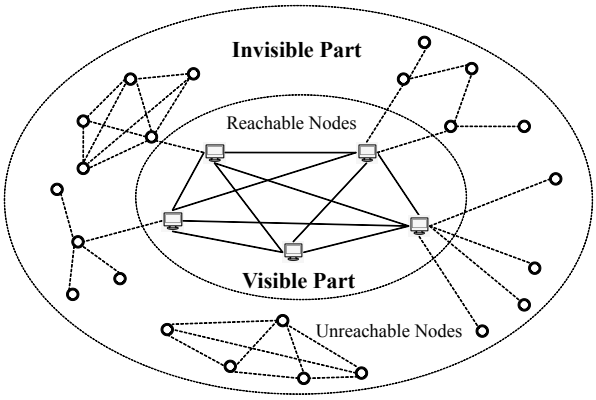

Figure 1. The Structure of Bitcoin Network.

The reachable nodes receive connection requests from external peers and provide public services [3] to the network, forming the visible part of the Bitcoin network. Most reachable nodes are full nodes [4], storing the complete transaction ledger and constituting the backbone of the Bitcoin network. In the early days, academic research on the Bitcoin network primarily focused on the detection of reachable nodes [5–7]. As research on the Bitcoin network progressed, it became apparent that the reachable nodes are only a part of the network, and there is also a significant portion of network nodes that cannot be directly connected but still actively participate in network operations. These nodes are referred to as "unreachable nodes".

The unreachable nodes do not accept incoming connection requests from external peers and do not provide public services to the network, forming the invisible part of the Bitcoin network. The unreachable nodes are usually deployed behind NAT or firewalls and cannot be discovered through active probing methods. Cause unreachable nodes play a crucial role in block storage, message forwarding, and competitive mining, it's necessary to understand the number and attributes of these nodes. By far, we knew that the number of unreachable nodes was more than that of the reachable nodes and they hold significant value for research on transaction tracing and user identification.

3. Related Work

In the previous work, Bitcoin researchers had focused on reachable nodes. Joan et al. measured the Bitcoin network [8] from Nov 2013 to Jan 2014, collected 872,000 nodes using Bitcoin-Sniffer, and analyzed node properties like geographic distribution, node stability, network transmission delay, etc. Fadhil et al. measured the Bitcoin network [7] during one week, collected 313,676 nodes and 6430 stable online nodes. Sehyun Park et al. measured the Bitcoin nodes [5] in 2018 and carried out a comparing research. They collected nearly 1 million nodes in 37 days, and compared the result with previous works. From these related works, we can find that the number of nodes was closely related to the measurement time.

Because observers cannot establish a direct connection with unreachable nodes, the previous methods to find unreachable nodes mainly relied on passive collection of network propagated messages. Biryukov et al. conducted a de-anonymization study [9] and found a large number of nodes that could not be connected in the network. Neudecker et al. identified two main categories of roles for unreachable nodes: standard clients in NAT or miners in mining pools [10]. Wang et al. measured the unreachable nodes in Bitcoin and developed a detecting tool called bcclient [11]. They deployed

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2023 doi:10.20944/preprints202311.1477.v1

102 probe nodes worldwide to collect connection requests and discovered 189,000 active IPv4 nodes within a week. Assuming each unreachable node maintains 3.5 outgoing connections, they estimated that the number of unreachable nodes within a 6-hour interval is about 155,000.

Grundmann et al. conducted studies on unreachable nodes in Bitcoin and proposed a passive announcement listening(PAL) method [12,13]. They extracted unreachable nodes by receiving broadcast addresses in the network, recording data from 2016 to 2020. They stated that there were approximately 31,000 active unreachable nodes per day at the end of 2020. In addition, they deployed a testing node to validate the correctness of PAL. Stouten conducted probing of the Bitcoin network in passive mode [14] and discovered 86,741 unreachable nodes in a span of 6 days in May 2020.

However, the limitations of existing methods are as follows [15]: 1) Low coverage rate. Due to the clustering characteristics of the Bitcoin network, the range of the probes were usually limited, making it difficult to collect total unreachable nodes. 2) Low collection efficiency. Cause passively waiting for messages, existing methods usually took several weeks or months to obtain satisfactory results. 3) Lack of validation methods. Nodes that could not be connected do not necessarily mean they are unreachable nodes. Reachable nodes may appear as "unreachable" due to network delay or reaching the maximum connections threshold. Offline nodes look like "unreachable" but they are never working in the network. There are lack of effective validation methods by far.

4. Problem Statement

4.1. Node Address Category

Bitcoin node addresses can be classified into five categories: Reachable nodes, Churn reachable nodes, Unreachable nodes, Offline nodes, and Fake nodes, as shown in Figure 2.

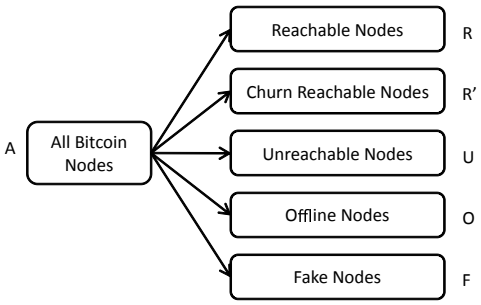

Figure 2. Bitcoin Node Address Categories.

Set R represents reachable node addresses, which corresponds to the currently online reachable nodes that the detecting system can establish connections to.

Set R' represents churn reachable node addresses, which corresponds to currently "unreachable" reachable nodes that temporarily show an "unreachable" state due to network latency or maximum connection limits.

Set U represents unreachable node addresses, which corresponds to online unreachable nodes that do not accept external connection requests. Here, we don't distinguish whether the unreachable nodes are in "churn" state, because a unreachable node can never be actively connected.

Set O represents offline node addresses, which corresponds to nodes that have gone offline either due to IP address changes or physical device shutdowns. Due to the lack of a regular cleaning mechanism for offline nodes in the Bitcoin network, these offline node addresses are stored in the addrman of network nodes for a long time with an very older timestamp.

Set F represents fake node addresses, which are not real Bitcoin nodes but injected into the network by attackers. We have discovered some abnormal node addresses in our experiments that have obvious arrangement patterns indicating that they are likely fake node addresses injected into the network by attackers.

 doi:10.20944/preprints202311.1477.v1

Classifying Bitcoin node addresses into categories will help us better detect and study online reachable nodes and unreachable nodes.

4.2. Node Attribute

During the interaction with Bitcoin nodes, the detection system obtained a large number of node attributes, shown in Table 1. On one hand, the returned ADDR messages showed node information such as: service type (Service), port number (Port), and the timestamp (Time). On the other hand, the detection system recorded many working parameters such as total records of one target IP (IP_Count), the time of sending GETADDR message (Send_Time), the time of receiving returned ADDR message (Receive_Time), the byte length of ADDR message (ADDR_Length), and the returned times of different ADDR messages(ADDR_Num).

Table 1. Node Attributes.

Node Attributes	**Meaning**
Service	Service type number
Port	Port number
Time	The timestamp of node address
IP_Count	Total records of one target IP
Send_Time	Time of sending GETADDR message
Receive_Time	Time of receiving returned ADDR message
ADDR_Length	The byte length of an ADDR Message
ADDR_Num	Returned times of different ADDR Messages

Bitcoin nodes can be classified into five categories. Different categories of nodes will reflect different statistical characteristics of attributes due to different service capabilities, different connection quality, and different software versions. Nodes in different categories have different statistical features in their attributes, making it possible to apply machine learning methods to classify them automatically.

4.3. Node Activity Parameters

The previous detection system usually connected to the node addresses one by one to verify whether it was connectable. In Figure 2, the addresses in set O account for a very large proportion. Due to the large amount of offline node addresses, traditional detection system would cost a very long time to complete one round of detection. In fact, it's useful for us to only detect online nodes, and useless to detect offline nodes.

We can simply judge a node address whether refers to a online node by evaluating it's activity. The node activity can be evaluated by some parameters. In this article, we proposed an evaluating model based on information entropy, which included parameters as: Ci (IP_Count), Si (Service), Pi (Port), Ti (Time) and Di (Receive_Time - Send_Time), where "i" stands for node i. The meaning of attributes are showed in Table 1. These parameters have close relationship with node activity.

1) Ci represents the total number of i-th node address records collected by the detection system. The more influential a node in a Bitcoin network, the wider its node address spreads in the network. Therefore, when the detection system requests inventory node addresses from remote nodes, active nodes' addresses will be more counted.

2) Si represents the Service Type value of the i-th node. Different Si values correspond to different combinations of service identifiers, including NODE_WORK, NODE_WITNESS, NODE_NETWORK_LIMITED, NODE_BLOOM, NODE_COMPACT_FILTERS, et al. Among them, NODE_WORK identifies whether this node has stored a complete copy of blockchain(this node is a full node). Full nodes are more likely to be active nodes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2023 doi:10.20944/preprints202311.1477.v1

3) Ti represents the fressness of i-th node. The fresh node often indicates to a high level of activity.

4) Pi represents the port number of i-th node. Most Bitcoin nodes open 8333 port to receive conncetions. A node with 8333 port opening indicates a high level of activity.

5) Di represents the delay of sending GETADDR message and receiveing ADDR message. The smaller the delay, the stronger the service capability or good connection quality of the node. The larger the delay, the weaker the service capability or poor connection quality.

By evaluating nodes' activety, we can select active nodes to detect, which will greatly reduce the number of nodes in the queue and enhance the detection efficiency greatly.

5. Methodology

As for node detection, reachable nodes and unreachable nodes are very different, so we proposed two different methods: BRF (Bitcoin Reachable-nodes Finding) to find reachable nodes and BUF (Bitcoin Unreachable-nodes Finding) to identify unreachable nodes.

5.1. Detecting Reachable Nodes

To address the problem of long scanning cycles and low detection efficiency in the detection of all Bitcoin reachable nodes, the authors proposed a reachable node detection algorithm BRF based on evaluating node's activity. It uses an entropy method to calculate node's parameters to get a score, and only detect the node whose score exceeding the threshold. By BRF, the detection system can reduce the number of nodes to be detected from millions to thousands and improve the detection efficiency greatly.

5.1.1. Parameter Normalization

To use the entropy method to calculate node activity one by one, it is necessary to normalize the parameters firstly. These parameters include: Ci, Si, Pi, Ti and Di. In the following formulas, uppercase letters represent the normalized evaluation value, and lowercase letters represent the variable value. Suppose node set N has n nodes, and j is any node.

1) Ci represents the total number of node i's addresses collected by the detection system. The normalization formula for Ci is:

$$C_i = \log c_i / \log \max_{1 \le j \le n} c_j \tag{1}$$

2) Si represents the service type of node i. If node i is a full node, the Si value can only be 1037, 1033, 1101, 1, 3, and 5. Therefore, the normalized formula for Si is:

$$S_i = \begin{cases} 1 \; s_i \in 1033, 1037, 1101, 1, 3, 5 \\ 0 \; s_i \notin 1033, 1037, 1101, 1, 3, 5 \end{cases} \tag{2}$$

3) Ti represents the difference between the timestamp and the current time. The normalized calculation formula for Ti is:

$$T_i = 1 - \log t_i / \log \max_{1 \le j \le n} t_j \tag{3}$$

4) Pi represents the port number of the node i. A node with 8333 port opening will be more likely a active node. The normalized formula for Pi is:

$$P_i = \begin{cases} 1 \; p_i = 8333 \\ 0 \; p_i \neq 8333 \end{cases} \tag{4}$$

5) Di represents the delay between sending a GETADDR message and receiving the ADDR message. This delay reflects the service capability of the target node. The normalized formula for Di is:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2023 doi:10.20944/preprints202311.1477.v1

$$D_i = 1 - \log d_i / \log \max_{1 \le j \le n} d_j \tag{5}$$

5.1.2. Node Activity Evaluation

Next, we calculate the comprehensive score of node i to evaluate it's activity. Here we use a method based on information entropy.

Suppose node set N has n nodes, where i stands for any node and j stands for any evaluation parameter. The variables of evaluation parameters are: ci, si, ti, pi and di. In the previous section, we had got the normalized evaluation parameters for node i: Ci, Si, Pi, Ti and Di. Then we calculate the weight of each parameter j.

$$w_j = \frac{1 - e_j}{n - \sum_{j=1}^{n} e_j} \tag{6}$$

In (6), e_j stands for the entropy of parameter j and w_j stands for the weight of parameter j. We could calculate the weights of different parameters on set N: w_c, w_s, w_t, w_p and w_d. Finally, we calculated the comprehensive score of node i:

$$Score_i = w_c * C_i + w_s * S_i + w_t * T_i + w_p * P_i + w_d * D_i \tag{7}$$

We carried out experiments for many times to get the threshold score for an active node. We found that if a node's comprehensive score exceeded 0.1, it's much likely to be an active node. So we send the nodes whose comprehensive score greated than 0.1 to detection queue.

The next detection process was not significantly different from the previous method. However BRF had filtered the to be detected and reduced the target nodes greatly, the detection efficiency is improved dramatically.

5.2. Idendifying Unreachable Nodes

In order to solve the problem of unable to actively detect unreachable nodes in the Bitcoin network and lack of effective verification methods, the authors proposed a model BUF for identifying unreachable nodes based on attribute features. It extracted attributes such as node service type, port number, and total number of records to build feature vectors. It constructed a decision tree model through training on a large number of inventory node addresses to automatically classify and identify real unreachable nodes.

5.2.1. Dataset and Feature Extraction

The selection of samples has a significant impact on the classification performance. In this article, the dataset D consisted of positive and negative samples, randomly chosen from the node address database, with a total of 20,000 records. Positive samples: The detection system recorded all received broadcast ADDR messages on a day, and extracted all node addresses from them. After removing all reachable nodes, the real online unreachable node addresses were left. Choose 10,000 records randomly from them as positive examples. Negative samples: The detection system recorded all node addresses that failed to connect on the same day. After removing the known reachable and unreachable node addresses, offline nodes and fake nodes addresses were left. Choose 10,000 records randomly from them as negative examples. After mixing the positive and negative samples in 1:1 arbitrarily, 14000 records were selected as training data D_T, and the remaining 6000 records were selected as validating data D_V.

We have introduced many node attributes (see Table 1) and explained different statistical features according to node categories. Based on these attributes, we could extract some features to train a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2023 doi:10.20944/preprints202311.1477.v1

machine learning model, and automatically classify node addresses into different categories. The selected features were shown in Table 2.

Table 2. Feature Extraction.

Notation	Features
f_S	Service
f_P	Port
f_T	Now-Time
f_C	IP_Count
f_D	Receive_Time-Send_Time
f_L	ADDR_Length
f_N	ADDR_Num

In this article, we applied Gini Index criterion to choose optimal features. It is most commonly used in machine learning. Suppose feature "a" has "V" possible values a_1, a_2, ..., a_V. If "a" is used to partition the sample set D, "V" branches will be generated. The "v"th branch contains all the samples in D with attribute "a_v", denoted as D_v. So the Gini Index obtained by using attribute "a" can be calculated as:

$$Gini(D,a) = \sum_{v=1}^{V} \frac{|D_V|}{|D|} Gini(D_V) \tag{8}$$

By calculating the Gini Index of every feature, we selected the ones with higher Gini Index as the optimal features. The Gini Index reflects the probability of data inconsistency. The larger the Gini index, the greater the uncertainty and disorder in the data.

5.2.2. Classification Model

This article proposed a model BUF (Bitcoin Unreachable-nodes Finding), which could extract typical features from sample nodes' attributes, train a machine learning model and automatically classify unreachable nodes from massive collected node addresses. The structure and data-processing of BUF is shown in Figure 3.

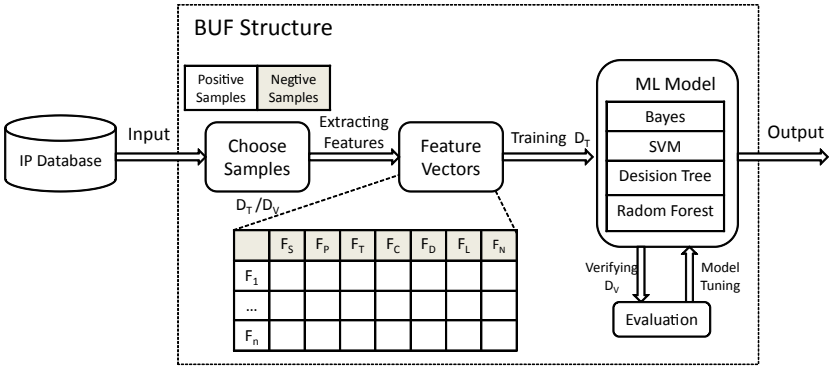

Figure 3. BUF Structure.

The most commonly used machine learning classifiers include Naive Bayes, Supporting Vector Machine(SVM), Random Forest and Decision Tree, etc. The author applied these classifiers at default parameters to evaluate the classification performance. Several experiments were carried out in PyCharm environment and the Precision, Recall and F1 of these models were compared. The Comparison of different models are shown in Table 3. Decision tree model got best classification performance at default parameters.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2023 doi:10.20944/preprints202311.1477.v1

Table 3. Comparison of Different Models (Default Parameter).

Classification model	Precision	Recall	F1
Naive Bayes	0.701	0.336	0.501
SVM	0.852	0.899	0.844
Random Forest	0.861	0.987	0.864
Decision tree	0.889	0.945	0.884

In this application scenario, the sample features were small in number, clear in meaning and had a certain correlation. Each feature measurement value had a great impact on the classification result. So the decision tree model was more applicable. Here the theoretical analysis was consistent with the preliminary experimental results.

6. Experiments

We developed a detection system BNS (Bitcoin Network Sniffer) and carried out experiments to detect reachable and unreachable nodes in Bitcoin network from April 30th to May 14th, 2023.

6.1. Bitcoin Network Sniffer

The BNS system is divided into five main parts: Main Thread, Node detecting module, IP Database module, Real-time Analysis module and Data Processing module. The system structure is shown in Figure 4.

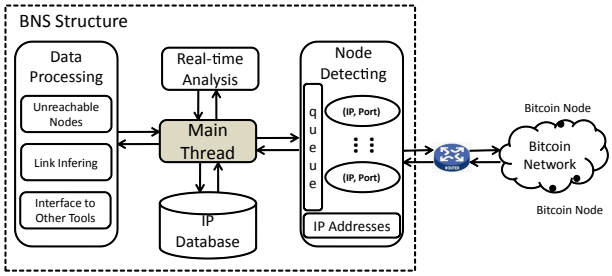

Figure 4. BNS Structure.

The Main Thread is the core of BNS, responsible for system controlling, socket driving, multi-thread applying and database managing, etc. The Node Detection module reads the node IP address and port number from the node queue, establishes multi-thread connections with the target nodes, and completes message interaction in independent pipelines. The IP Database module is responsible for storing the Bitcoin node addresses collected by the detection system. Each node address record includes basic information such as IP address, port number, service type, timestamp, as well as working parameters such as the total number of records, the time to send GetADDR message, the time to receive ADDR message, the length of ADDR message packet and the times of different ADDR messages returned. The Real-time Analysis module is mainly responsible for processing returned messages, including calculating node activity and counts node attributes, and so on. The Data Processing module receives the formatted information, performs feature extracting, link inferring and communicating to other third-party tools.

6.2. Detection Experiment

We carried out detection experiment by BNS from April 30th to May 14th, 2023, and recorded the total found reachable nodes, unreachable nodes. Also, the time cost of daily experiment was recorded.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2023 doi:10.20944/preprints202311.1477.v1

During the experimental period, BNS found an average of 18284 reachable nodes and identified 29339 unreachable nodes per day, with an average time cost of 1 hour and 23 minutes, as shown in Table 4.

Table 4. Detection Experiment.

Num	Date	Reachable Nodes	Unreachable Nodes	Time Cost
1	April 30th	17609	29049	1h21min
2	May 1st	18446	28498	1h 42min
3	May 2nd	18466	29154	1h 38min
4	May 3rd	18339	29869	1h 29min
5	May 4th	18446	29407	1h 51min
6	May 5th	18440	29736	1h 3min
7	May 6th	18680	29418	1h 31min
8	May 7th	18357	29648	1h 42min
9	May 8th	18092	29796	1h 29min
10	May 9th	18143	29197	1h 4min
11	May 10th	18416	30081	1h 27min
12	May 11th	18491	29271	1h 4min
13	May 12th	18323	29197	1h 5min
14	May 13th	17965	29048	1h 25min
15	May 14th	18052	28720	58min
Average		18284	29339	1h 23min

"Bitnodes" is currently an authoritative third-party website in the field of Bitcoin measurement. The authors compared the experimental results with Bitnodes' real-time data at the same time, as shown in Figure 5. The blue curve represents the daily change nodes of Bitnodes, while the red curve represents the daily change nodes of BNS. During the experimental period, BNS daily found more reachable and unreachable nodes than Bitnodes, showing the superiority of the algorithm.

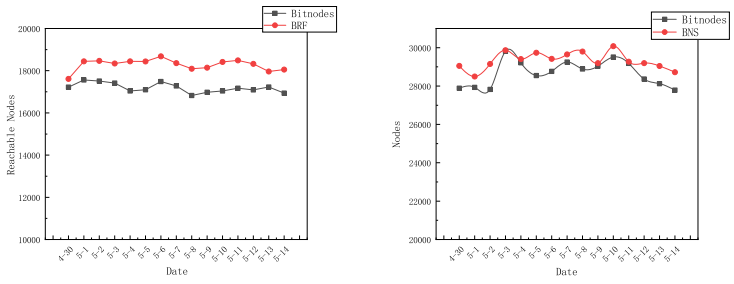

Figure 5. Comparison of Experiment Results to Bitnodes.

In terms of detection efficiency, BNS also had more advantages. Table 5 shows the daily found nodes and time cost. According to the description of Bitnodes website, it scanned for reachable nodes every 2 hours and collectted unreachable nodes information every 4 hours. Our BNS completed one whole network scanning in an average of 1h 23min, and found more nodes than Bitnodes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2023 doi:10.20944/preprints202311.1477.v1

Table 5. Comparison of Detection Efficiency.

	Reachable Nodes	Unreachable Nodes	Time Cost
Bitnodes	17191	28677	4h
BNS	18284	29339	1h 23min

7. Discussion

7.1. Bitcoin Network Size

In the previous work, Bitcoin researchers had known the number of reachable nodes. However, unreachable nodes cannot be actively detected, people do not know the exact number of unreachable nodes by far. Therefore, it is difficult to estimate the whole network size of Bitcoin system.

Previous work [11,12,14] carried out passive collection of unreachable nodes, but the total number of unreachable nodes still unclear. Bitcoin network often reflects clustering characteristics, that is, nodes present a certain degree of aggregation. Broadcast node addresses propagate rapidly within a cluster of nodes, but are slow and limited outside the cluster. Passive collection cannot obtain all unreachable nodes, thus the estimation of Bitcoin network size is not accurate.

In this paper, the authors collected the inventory addresses of reachable nodes, and used a decision tree model to automatically identify unreachable nodes from them. Reachable nodes are usually important nodes in a node cluster, storing all node addresses broadcasting in this cluster. Our method collected the inventory addresses of reachable nodes all over the world, and obtained more node addresses than previous work.

We present the number of reachable nodes, unreachable nodes, and total nodes of Bitcoin network in Table 6. The total nodes in Bitcoin network is about 45,000-50,000 currently, and the ratio of reachable nodes to unreachable nodes is about 1:1.6. Compared to Bitnodes, our method BUF showed an advantage in the total number of discovered nodes.

Table 6. Size of the Bitcoin network.

Tools	Reachable Nodes	Unreachable Nodes	Total Nodes
Bitnodes	17191	28677	45868
BNS	18284	29339	47623

7.2. Churn of Nodes

The Bitcoin network is a dynamic P2P network. Some nodes in the network exhibit intermittent "churn" state due to network latency or other reasons. The authors analyzed the "churn" phenomenon in the Bitcoin network.

We analyzed all nodes from April 30th (Day1) to May 14th (Day15). From Day1 to Day15, the total number of reachable nodes fluctuated around 18000, with a total of 9878 nodes consistently online within 15 days, as shown in the left of Figure 6; The total number of unreachable nodes fluctuated around 27000, with a total of 10942 nodes consistently online, as shown in the right of Figure 6. In the figure, the blue curve represents the daily change in the total number of nodes, while the red curve represents the daily stable number of nodes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 November 2023 doi:10.20944/preprints202311.1477.v1

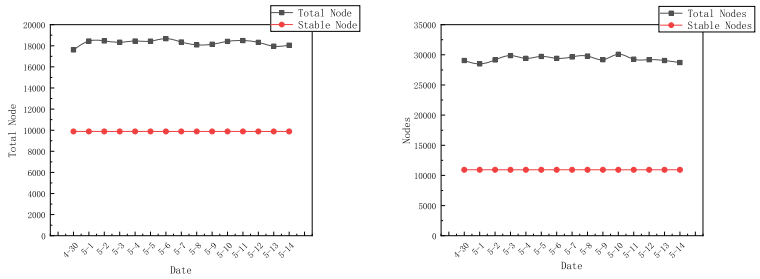

Figure 6. Stable Reachable and Unreachable Nodes.

Furthermore, the author analyzed the daily change proportion of nodes. Similarly, use the daily discovered nodes from Day1 to Day15 to calculate the daily change proportion of nodes. The daily variation ratio of reachable and unreachable nodes is shown in Figure 7 (The left for reachable nodes and the right for unreachable nodes). The curve represents the change in the number of daily nodes, and the bar chart represents the proportion of changes in daily nodes compared to the previous day, with red representing a decrease and blue representing an increase.

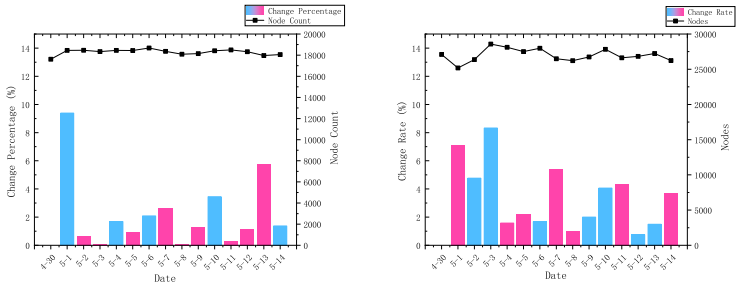

Figure 7. Stable Reachable and Unreachable Nodes.

7.3. Geographic Distribution

We searched for the longitude and latitude information of node IP addresses through the cyberspace search engine Zoomeye, and calculated their distribution proportions on various continents worldwide, as shown in Table 7. As can be seen from the table, Bitcoin nodes are most distributed in Europe, America, and Asia, accounting for over 98% of the global total nodes.

Table 7. Geographic Distribution of Bitcoin Nodes.

Regions	Reachable nodes	Unreachable node
Europe	58.07%	50.21%
America	30.91%	29.91%
Asia	9.43%	18.18%
Oceania	1.36%	1.16%
Africa	0.22%	0.54%
Total	100%	100%

An interesting phenomenon is that the reachable nodes in Asia accounted for 9% of the total reachable nodes, while the unreachable nodes in Asia accounted for 18% of the total. This may be due to the large population in Asia and the large number of Bitcoin clients.

8. Conclusions

In this article, the authors discussed how to collect Bitcoin nodes as many as possible. To address the problem of long scanning cycles and low detection efficiency in Bitcoin reachable nodes detection, the authors proposed an algorithm BRF which reduce the number of nodes to be detected from millions to thousands and improve the detection efficiency greatly. To solve the problem of unable to actively detect unreachable nodes in the Bitcoin network, the authors proposed a model BUF for identifying unreachable nodes based on attribute features. Experiments showed that two methods performed better than the website "Bitnodes" in total number and efficiency.

Based on the experimental results, the authors analyzed the real network size, node "churn" and geographical distribution. The total nodes in Bitcoin network was about 45,000-50,000 in 2023, and the ratio of reachable nodes to unreachable nodes was about 1:1.6. Everyday there were at most 9% online nodes churn. Most nodes located in Europe, America, and Asia, accounting for over 98%.

Funding: This work was supported by National Key Research and Development Program of China (Grant No.2020YFB1006100).

References

1. Nakamoto S. Bitcoin: A peer-to-peer electronic cash system. *Decentralized business review* **2008**.
2. Eisenbarth J P, Cholez T, Perrin O. A Comprehensive Study of the Bitcoin P2P Network[C]// 2021 3rd Conference on Blockchain Research and Applications for Innovative Networks and Services(BRAINS). IEEE, 2021: 105-112.
3. Antonopoulos A M, Media O. Mastering Bitcoin: Unlocking Digital Crypto-Currencies. O'Reilly Media, Inc. 2015.
4. Li R, Zhu J, Xu D, et al. Bitcoin network measurement and a new approach to infer the topology[J]. China Communications, 2022, 19(10): 169-179.
5. Park S, Im S, Seol Y, et al. Nodes in the bitcoin network: Comparative measurement study and survey[J]. IEEE, 2019, 7: 57009-57022.
6. Imtiaz M A, Starobinski D, Trachtenberg A, et al. Churn in the bitcoin network: Characterization and impact[C]//2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC). IEEE, 2019: 431-439.
7. Fadhil M, Owenson G, Adda M. A bitcoin model for evaluation of clustering to improve propagation delay in bitcoin network[C]//2016 IEEE intl conference on computational science and engineering (CSE) and IEEE intl conference on embedded and ubiquitous computing (EUC) and 15th intl symposium on distributed computing and applications for business engineering (DCABES). IEEE, 2016: 468-475.
8. Donet Donet J A, Pérez-Sola C, Herrera-Joancomartí J. The bitcoin P2P network[C]//International conference on financial cryptography and data security. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014: 87-102.
9. A. Biryukov, D. Khovratovich, and I. Pustogarov, "Deanonymisation of clients in bitcoin P2P Network," in Proc. ACM SIGSAC Conf. Comput.Commun. Secur. (CCS), 2014, pp. 15-29.
10. Neudecker T, Andelfinger P, Hartenstein H. Timing analysis for inferring the topology of the bitcoin peer-to-peer network[C]//2016 Intl IEEE Conferences on Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC /ATC /ScalCom /CBDCom /IoP /SmartWorld). IEEE, 2016: 358-367.
11. Wang L, Pustogarov I. Towards better understanding of bitcoin unreachable peers[J]. arXiv preprint arXiv:1709.06837, 2017.
12. Grundmann M, Amberg H, Hartenstein H. On the estimation of the number of unreachable peers in the Bitcoin P2P network by observation of peer announcements[J]. arXiv preprint arXiv:2102.12774, 2021.
13. Grundmann M, Amberg H, Baumstark M, et al. Short Paper: What Peer Announcements Tell Us About the Size of the Bitcoin P2P Network[C]//International Conference on Financial Cryptography and Data Security. Cham: Springer International Publishing, 2022: 694-704.

 doi:10.20944/preprints202311.1477.v1

14. Stouten T. Hide and seek: different scan methods to analyse peer-to-peer based blockchain networks[D]. University of Twente, 2020.
15. Stouten T. Hide and seek: different scan methods to analyse peer-to-peer based blockchain networks[D]. University of Twente, 2020.