mathematics

MDPI

Article

BNS: A Detection System to Find Nodes in the 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; liehuangz@bit.edu.cn (L.Z.); xudawei@bit.edu.cn (D.X.)
[2] National Computer Network Emergency Response Technical Team/Coordination Center, Beijing 100029, China; lc@cert.org.cn
[3] School of Cyberspace Science and Technology, Beihang University, Beijing 100191, China
* Correspondence: lrg@cert.org.cn (R.L.); wufudong@buaa.edu.cn (F.W.)

Abstract: Bitcoin was launched over a decade ago and has made an increasing impact on the world's financial order, which has attracted the attention of researchers all over the world. The 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), which could collect as many Bitcoin nodes as possible is proposed. For reachable nodes, the authors designed an algorithm, BRF (Bitcoin Reachable-Nodes Finding), based on node activity evaluation which reduces the nodes to be detected and greatly shortens the detection time. For unreachable nodes, the authors trained a decision tree model, BUF (Bitcoin Unreachable-Nodes Finding), to identify unreachable nodes based on attribute features from a large number of node addresses. Experiments showed that BNS discovered an average of 1093 more reachable nodes (6.4%) and 662 more unreachable nodes (2.3%) than the well-known website "Bitnodes" per day. It showed better performance in total nodes 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; decision tree model

MSC: 37M10

Citation: Li, R.; Zhu, L.; Li, C.; Wu, F.; Xu, D. BNS: A Detection System to Find Nodes in the Bitcoin Network. *Mathematics* **2023**, *11*, 4885. https://doi.org/10.3390/math11244885

Academic Editor: Daniel-Ioan Curiac

Received: 21 November 2023
Revised: 30 November 2023
Accepted: 4 December 2023
Published: 6 December 2023

1. Introduction

Bitcoin was first proposed by Satoshi Nakamoto in 2008 and has been working steadily for over a decade. To date, it is the most successful cryptocurrency in the world, with the highest value and greatest influence. With the outbreak of COVID-19 in 2019, much currency flooded into the Bitcoin market and raised Bitcoin's price, which attracted more research interest in Bitcoin. The price of Bitcoin has become a topic of concern in academia and industry. Suhwan et al. [1] adopted several deep learning methods and Zi et al. [2] used Twitter comments to predict Bitcoin price. Researchers around the world have studied Bitcoin from different perspectives.

The Bitcoin system can be divided into the transaction layer and the network layer. Most previous studies 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 privacy and security. All the transactions are stored in the blockchain in order of time and broadcast to all participants. Nodes in the Bitcoin Network record all blockchain data. The decentralization and anonymity of Bitcoin complicates supervision because the transactions are anonymous and difficult to track. Therefore, thorough studies of the Bitcoin Network are worthwhile.

The Bitcoin Network consists of tens of thousands of nodes all over the world. These nodes autonomously discover neighbors and complete connection establishment, forming a

dynamic but robust P2P network. Due to the propagation delay in the wide area network, it is difficult to collect total nodes completely. A large amount of research has been conducted on detecting reachable nodes [3–6], and "Bitnodes" is the most authoritative third-party website for this purpose. In addition, detecting unreachable nodes is a major challenge due to the numerous security facilities such as firewalls existing in the network.

In this paper, the authors first introduce the structure of the Bitcoin Network and some related work then explain the basic knowledge of addresses category, node attributes, and activity evaluation parameters. Next, the authors propose two algorithms in detail: BRF, to find reachable nodes, and BUF, to identify unreachable nodes. To evaluate the performance of BRF and BUF, a detection system, BNS, was developed, and experiments were carried out from 30 April to 14 May 2023. Based on the experimental results, the characteristics of the Bitcoin Network were thoroughly studied. The main contributions of this article are as follows:

(1) The authors designed an algorithm, BRF, 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 decision tree model, BUF, to identify unreachable nodes from a large number of node addresses.
(3) The authors developed a detection system, BNS; carried out experiments; and analyzed the real network size, node "churn", and geographical distribution.

2. Bitcoin Network

The 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, as shown in Figure 1. Solid lines represent bidirectional connections, while dashed lines represent unidirectional connections from unreachable nodes to reachable nodes.

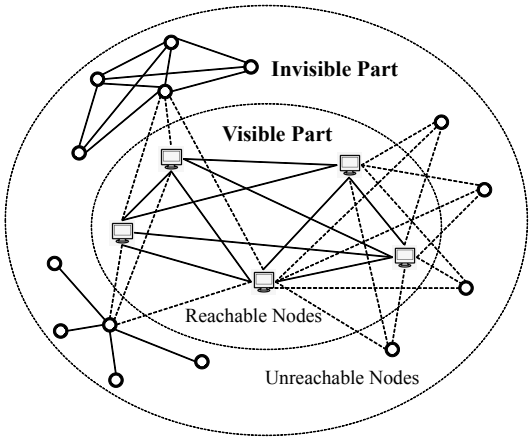

Figure 1. The structure of the Bitcoin Network.

The reachable nodes receive connection requests from external peers and provide public services to the network, forming the visible part of the Bitcoin Network. Most reachable nodes are full nodes, 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 [5,7,8]. 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. Because unreachable nodes play a crucial role in block storage, message forwarding, and competitive mining, it is necessary to understand the number and attributes of these nodes. We know that the number of unreachable nodes is more than that of the reachable nodes by far, and they hold significant value for research on transaction tracing and user identification.

3. Related Work

In the previous work, Bitcoin researchers focused on reachable nodes. Joan et al. [6] measured the Bitcoin Network from November 2013 to January 2014, collected 872,000 nodes using Bitcoin-Sniffer, and analyzed node attributes such as geographic distribution, node stability, and network transmission delay. Fadhil et al. [8] measured the Bitcoin Network over one week and collected 313,676 nodes and 6430 stable online nodes. Sehyun Park et al. [5] measured the Bitcoin nodes in 2018 and carried out comparative research. They collected nearly 1 million nodes in 37 days and compared the result with previous works. Eisenbarth et al. [4] conducted a comprehensive study on the Bitcoin Network using a crawler program similar to Bitnodes to detect the Bitcoin Network in 2020 and analyzed node activity, node churn, software versions, and security. Ruiguang et al. [3] measured the Bitcoin network, analyzed reachable node attributes, and proposed a method to infer the topology.

Researchers in previous studies typically obtained seeds of the Bitcoin network, set up connections to these seeds, then sent GETADDR messages to them. Some seeds would accept connection requests and return ADDR messages. From ADDR messages, researchers could obtain many node addresses and try to connect them one-by-one. By continuously repeating this process, researchers could gradually accumulate node addresses and discover reachable nodes in them. Due to the large number of Bitcoin nodes, scanning all addresses would take a very long time, so the detection efficiency was very low.

As for unreachable nodes, because observers cannot establish a direct connection with them, the previous methods mainly relied on passive collection of network-propagated messages. Biryukov et al. [9] conducted a de-anonymization study and found a large number of nodes that could not be connected in the network. Neudecker et al. [10] identified two main categories of roles for unreachable nodes: standard clients in NAT or miners in mining pools. Wang et al. [11] measured the unreachable nodes in Bitcoin and developed a detection tool called bcclient. They deployed 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 h interval is about 155,000. Grundmann et al. [12,13] conducted studies on unreachable nodes in Bitcoin and proposed a passive announcement listening (PAL) method. 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. Stouten [14] conducted probing of the Bitcoin Network in passive mode and discovered 86,741 unreachable nodes in a span of 6 days in May 2020.

However, the limitations of existing methods for finding unreachable nodes are as follows: (1) Low coverage rate. Due to the clustering characteristics of the Bitcoin Network, the range of the probes is usually limited, making it difficult to collect total unreachable nodes. (2) Low collection efficiency. Due to the passive wait for messages, existing methods usually take several weeks or months to obtain satisfactory results. (3) Lack of validation methods. Nodes that cannot be connected are not necessarily unreachable nodes. Reachable nodes may appear as “unreachable” due to network delay or the maximum connections threshold being reached. Offline nodes appear “unreachable”, but they are never active in the network. There is a considerable lack of effective validation methods.

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.

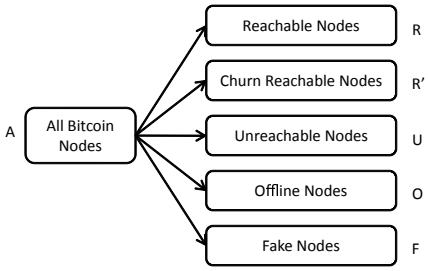

Figure 2. Bitcoin node address categories.

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 do not distinguish whether the unreachable nodes are in "churn" state because an 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 older timestamp.

Set F represents fake node addresses, which are not real Bitcoin nodes but are 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.

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

4.2. Node Attributes

During the interaction with Bitcoin nodes, the detection system obtained a large number of node attributes, shown in Table 1. On the one hand, the returned ADDR messages showed node information such as the service type (Service), port number (Port), and 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 the GETADDR message (Send_Time), the time of receiving the the returned ADDR message (Receive_Time), the byte length of the ADDR message (ADDR_Length), and the returned times of different ADDR messages (ADDR_Num).

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.

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 the GETADDR message
Receive_Time	Time of receiving the returned ADDR message
ADDR_Length	The byte length of the ADDR message
ADDR_Num	Returned times of different ADDR messages

4.3. Node Activity Parameters

We can simply judge whether a node address belongs to an online node by evaluating its activity. The node activity can be evaluated by some parameters. In this article, we propose an evaluating model based on information entropy, which includes parameters such as Ci (IP_Count), Si (Service), Pi (Port), Ti (Time) and Di (Receive_Time-Send_Time), where "i" stands for node i. The definitions of attributes are shown in Table 1. These parameters have a 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 node addresses will be counted more frequently.
(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, etc. Among them, NODE_WORK identifies whether this node has stored a complete copy of the blockchain (this node is a full node). Full nodes are more likely to be active nodes.
(3) Ti represents the freshness of the i-th node. The fresh node often indicates a high level of activity.
(4) Pi represents the port number of the i-th node. Most Bitcoin nodes open the 8333 port to receive connections. A node with an 8333 port opening indicates a high level of activity.
(5) Di represents the delay of sending the GETADDR message and receiving the 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 node activity, 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 propose 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 solve the problem of a long scanning cycle and low detection efficiency in the detection of reachable Bitcoin nodes, the authors propose a reachable node detection algorithm, BRF, based on evaluating node activity, which could reduce the number of nodes to be detected from millions to thousands and improve the detection efficiency greatly.

The previous method to find reachable nodes tried to connect to node addresses one-by-one. In Figure 2, the offline addresses (set O) usually account for a very large proportion. The traditional detection system would take a very long time to traverse all addresses. However, it is unnecessary to try every node. If we can choose online nodes (or

suspected online nodes) in advance, the detection range will be greatly reduced, and the efficiency will be improved. Here, we applied an evaluating method based on information entropy. We selected some basic parameters and calculated the information entropy of each parameter. Then, we obtained a comprehensive score of every node address. Finally, we set a threshold and only detected the node whose scores exceeded the threshold.

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 values, and lowercase letters represent the variable values. 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, or 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 an 8333 port opening will be more likely an 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:

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

5.1.2. Node Activity Evaluation

Next, we calculated the comprehensive score of node i to evaluate its 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. We obtained the normalized evaluation parameters for node i: Ci, Si, Pi, Ti, and Di, as described in the previous section. Then, we calculated 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:

$$Score_i = w_c \times C_i + w_s \times S_i + w_t \times T_i + w_p \times P_i + w_d \times D_i \quad (7)$$

We carried out numerous experiments to obtain the threshold score for an active node. We found that if a node's comprehensive score exceeds 0.1, it is highly likely to be an active node. So, we sent the nodes whose comprehensive score was greater than 0.1 to the detection queue. The process of calculating $Score_i$ is shown in Algorithm 1:

Algorithm 1: Node activity evaluation

Input: Suppose N is a node set, and i is any node in N. The evaluation parameters of node i are: c_i, s_i, t_i, p_i, and d_i. The node activity threshold is σ.
Output: Detecting node queue, Q

```
for i ∈ N do
    Normalize evaluation parameters according to (1), (2), (3), (4), and (5) and
     obtain C_i, S_i, P_i, T_i, and D_i;
    Calculate the weight of each parameter according to (6), and obtain the
     weights ω_c, ω_s, ω_p, ω_t, and ω_d;
    Calculate the comprehensive score of node i according to (7) and obtain Score_i;
    if Score_i > σ then
        input i to Q;
    end
end
```

Output: Q

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

5.2. Identifying Unreachable Nodes

In order to solve the problem of being unable to actively detect unreachable nodes in the Bitcoin Network and the lack of effective verification methods, the authors propose a model, BUF, for identifying unreachable nodes based on attribute features. It extracts attributes such as node service type, port number, and total number of records to build feature vectors. It constructs 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 study, 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. Then, 10,000 records were randomly selected 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. Then, 10,000 records were randomly selected from them as negative examples. After mixing the positive and negative samples 1:1 arbitrarily, 14,000 records were selected as training data D_T, and the remaining 6000 records were selected as validation 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 machine learning model and automatically classify node addresses into different categories. The selected features are 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 study, we applied the 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 those 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 proposes a model, BUF (Bitcoin Unreachable-Nodes Finding), which could extract typical features from sample node attributes, train a machine learning model, and automatically classify unreachable nodes from a large number of collected node addresses. The structure and data processing of BUF are shown in Figure 3.

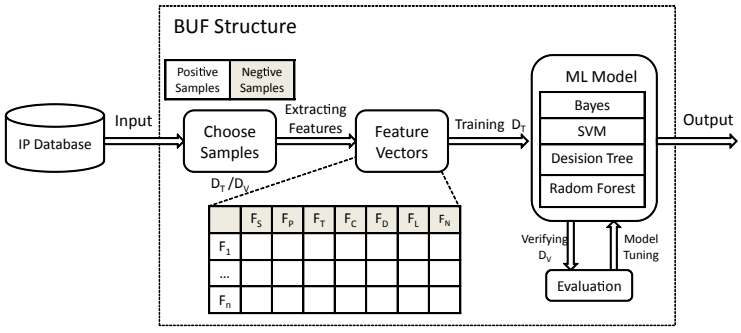

Figure 3. BUF Structure.

The most commonly used machine learning classifiers include the naive Bayes, support vector machine (SVM), random forest, and decision tree models, etc. The authors applied these classifiers at default parameters to evaluate the classification performance. Several experiments were carried out in the PyCharm environment, and the precision, recall, and F1 of these models were compared. The comparison of different models is shown in Table 3. The decision tree model attained the best classification performance at default parameters.

Table 3. Comparison of different models (default parameters).

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.

5.2.3. Decision Tree Construction and Optimization

Next, we began to construct the decision tree for Bitcoin node address classification. Starting from the root, the program calculated the optimal feature (see Table 2) to split the samples according to the GiNi criterion (8). On the next layer, the program performed the same operation, continuously splitting the tree until reaching the leaf nodes. This process was completed automatically using PyCharm.

To achieve an implementable classification performance, the decision tree must be pruned and optimized. We adjusted the depth of decision tree to observe the changes in precision. When the depth of the tree increased from 1 to 20 layers, the precision changed gradually. As shown in Figure 4, the optimal precision is achieved when the tree depth is 5. Next, we pruned the decision tree when tree depth was 5 layers. We adjusted the minimum sample number of leaf nodes to limit the growth of tree. When the minimum sample number of leaf nodes ranged from 5 to 100, the precision changed gradually. When the minimum sample number of leaf nodes reached 30, the model had the best precision. Subsequently, overfitting occurred, and the precision began to decline.

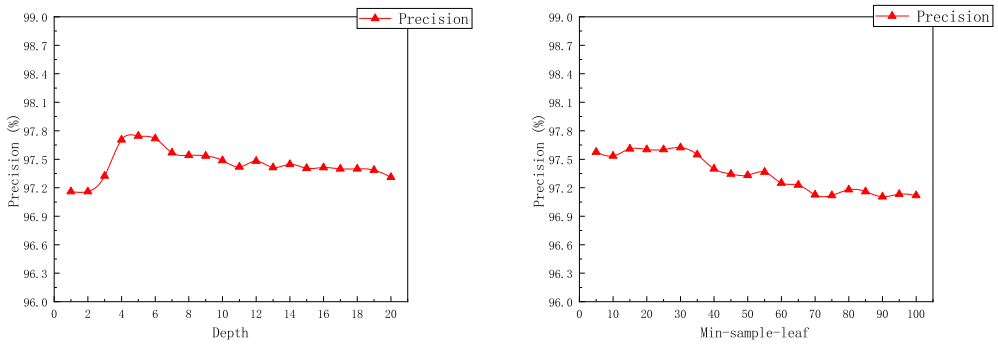

Figure 4. Decision tree optimization.

Finally, we chose a decision tree with a depth of 5 layers and a minimum sample number of 30 leaf nodes. The model had a classification precision of 97.61%, a recall rate of 96.43%, and an F1 score of 0.9757.

6. Experiments

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

6.1. Bitcoin Network Sniffer

The BNS system is divided into five main parts: the main thread, node detecting module, IP database module, real-time analysis module, and data processing module. The system structure is shown in Figure 5.

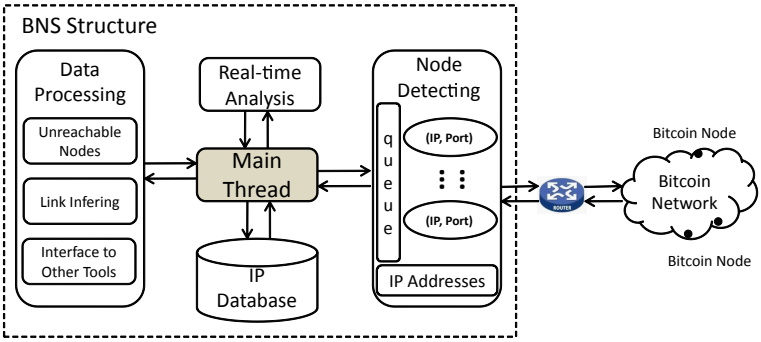

Figure 5. BNS structure.

The main thread is the core of BNS, responsible for system control, socket driving, multi-thread application, and database management, 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 the GetADDR message, the time to receive the ADDR message, the length of the ADDR message packet, and the returned times of different ADDR messages. The real-time analysis module is mainly responsible for processing returned messages, including calculating node activity, node attributes, and so on. The data processing module receives the formatted information and performs feature extraction, link inference, and communication with other third-party tools.

6.2. Detection Experiment

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

During the experimental period, BNS found an average of 18,284 reachable nodes and identified 29,339 unreachable nodes per day, with an average time cost of 1 h and 23 min, as shown in Table 4.

Table 4. Detection experiment.

Num	Date	Reachable Nodes	Unreachable Nodes	Time Cost
1	April 30th	17,609	29,049	1 h 21 min
2	May 1st	18,446	28,498	1 h 42 min
3	May 2nd	18,466	29,154	1 h 38 min
4	May 3rd	18,339	29,869	1 h 29 min
5	May 4th	18,446	29,407	1 h 51 min
6	May 5th	18,440	29,736	1 h 3 min
7	May 6th	18,680	29,418	1 h 31 min
8	May 7th	18,357	29,648	1 h 42 min
9	May 8th	18,092	29,796	1 h 29 min
10	May 9th	18,143	29,197	1 h 4 min
11	May 10th	18,416	30,081	1 h 27 min
12	May 11th	18,491	29,271	1 h 4 min
13	May 12th	18,323	29,197	1 h 5 min
14	May 13th	17,965	29,048	1 h 25 min
15	May 14th	18,052	28,720	58 min
Average		18,284	29,339	1 h 23 min

"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, as shown in Figure 6. 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.

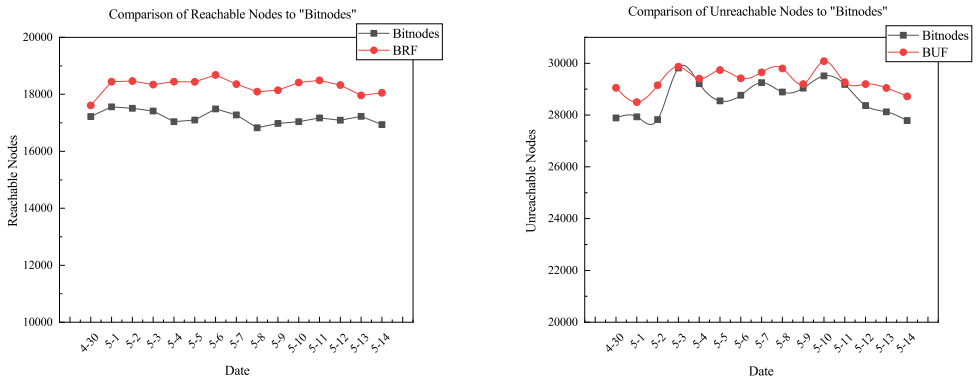

Figure 6. Comparison of experiment results with Bitnodes.

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 the Bitnodes website, it scans for reachable nodes every 2 h and collects unreachable node information every 4 h. Our BNS completed one whole network scanning in an average of 1 h 23 min, and found more nodes than Bitnodes.

Table 5. Comparison of detection efficiency.

	Reachable Nodes	Unreachable Nodes	Time Cost
Bitnodes	17,191	28,677	4 h
BNS	18,284	29,339	1 h 23 min

7. Discussion

7.1. Bitcoin Network Size

In the previous work, Bitcoin researchers determined the number of reachable nodes. However, unreachable nodes cannot be actively detected, and the exact number of unreachable nodes is not known. Therefore, it is difficult to estimate the whole network size of the Bitcoin system.

Previous work [11,12,14] carried out passive collection of unreachable nodes, but the total number of unreachable nodes still unclear. The 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 the Bitcoin Network size is not accurate.

In this study, 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 the Bitcoin Network in Table 6. The total nodes in the 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	17,191	28,677	45,868
BNS	18,284	29,339	47,623

7.2. Churn of Nodes

The Bitcoin Network is a dynamic P2P network. The vast majority (97%) of Bitcoin nodes exhibit intermittent network connectivity (churn) [7] due to network latency or other reasons. The authors analyzed the "churn" phenomenon in the Bitcoin Network.

We analyzed all nodes from 30 April (Day 1) to 14 May (Day 15). From Day 1 to Day 15, the total number of reachable nodes fluctuated around 18,000, with a total of 9878 nodes consistently online within 15 days, as shown in the left of Figure 7. The total number of unreachable nodes fluctuated around 27,000, with a total of 10,942 nodes consistently online, as shown in the right of Figure 7. 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.

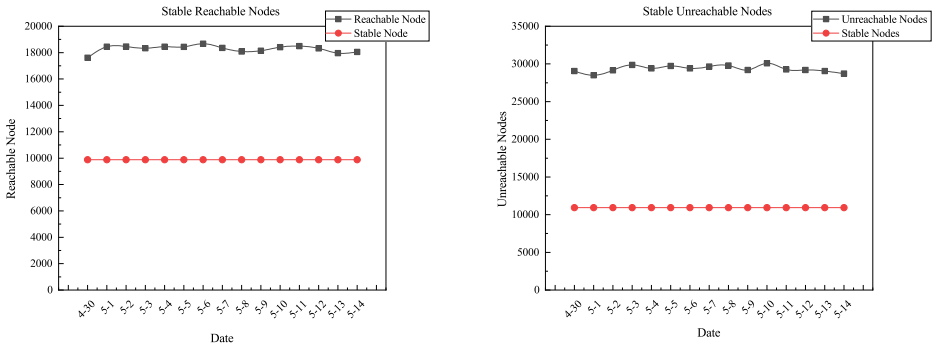

Figure 7. Stable reachable and unreachable nodes.

Furthermore, the authors analyzed the daily change proportion of nodes. Similarly, the daily discovered nodes from Day 1 to Day 15 were used to calculate the daily change proportion of nodes. The daily variation ratio of reachable and unreachable nodes is shown in Figure 8 (the left shows reachable nodes and the right 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.

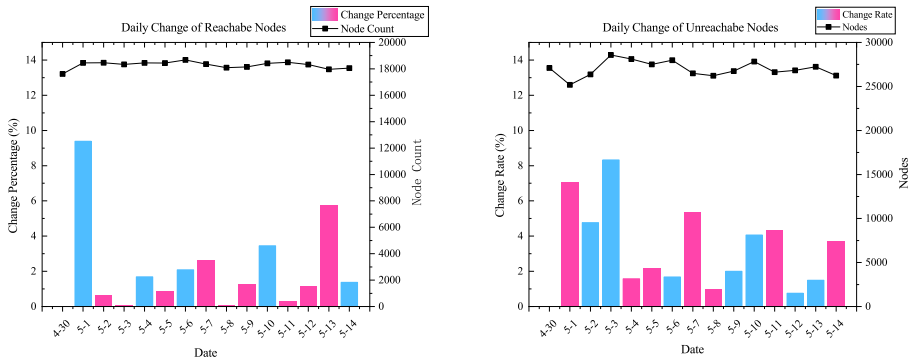

Figure 8. Stable reachable and unreachable nodes.

7.3. Geographic Distribution

We searched for the longitude and latitude information of node IP addresses through the search engine Zoomeye and calculated their distribution proportions on various con-

tinents 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 and Future Works

In this article, the authors discussed how to collect as many Bitcoin nodes as possible. To address the problem of long scanning cycles and low detection efficiency in the detection of reachable Bitcoin nodes, the authors proposed an algorithm, BRF, which can reduce the number of nodes to be detected from millions to thousands and improve the detection efficiency greatly. To solve the problem of being unable to actively detect unreachable nodes in the Bitcoin Network, the authors proposed a model, BUF, for identifying unreachable nodes based on attribute features. Applying BRF and BUF, a detection system, BNS, was developed and used to measure the Bitcoin Network in 2023. Experiments showed that BNS discovered an average of 1093 more reachable nodes (6.4%) and an average of 662 more unreachable nodes (2.3%) than "Bitnodes" per day. The time cost was reduced from 4 h to 1 h 23 min.

The experimental results further demonstrated the characteristics of the Bitcoin Network. Using days as the time window, the number of online Bitcoin nodes for one day was about 45,000–50,000 in the year of 2023, and the ratio of reachable nodes to unreachable nodes was about 1:1.6. Every day, 1% to 9% of online nodes showed a state of churn. Over 98% of online Bitcoin nodes are located in Europe, America, or Asia.

In the future, more experiments will be carried out using BNS. The authors plan to conduct a long-term observation of the Bitcoin Network. Because the actual size of the Bitcoin Network cannot be theoretically analyzed, the results of BNS will be compared with other third-party monitoring platforms. Bitcoin remains the most widely used cryptocurrency in the world, with the highest value and the greatest influence. There is a significant demand for monitoring and studying Bitcoin in academia and in industry. Detecting and mastering all Bitcoin nodes will be helpful for address anonymization analysis and user transaction traceability.

Author Contributions: Methodology, L.Z.; Software, F.W.; Validation, C.L. and D.X.; Writing—original draft, R.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (Grant No. 2020YFB1006100) and National Natural Science Foundation of China (Grant No. 62106060).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ji, S.; Kim, J.; Im, H. A comparative study of bitcoin price prediction using deep learning. *Mathematics* **2019**, *7*, 898. [CrossRef]
2. Ye, Z.; Wu, Y.; Chen, H.; Pan, Y.; Jiang, Q. A stacking ensemble deep learning model for bitcoin price prediction using Twitter comments on bitcoin. *Mathematics* **2022**, *10*, 1307. [CrossRef]
3. Li, R.; Zhu, J.; Xu, D.; Wu, F.; Gao, J.; Zhu, L. Bitcoin network measurement and a new approach to infer the topology. *China Commun.* **2022**, *19*, 169–179. [CrossRef]
4. Eisenbarth, J.P.; Cholez, T.; Perrin, O. A Comprehensive Study of the Bitcoin P2P Network. In Proceedings of the 2021 3rd Conference on Blockchain Research and Applications for Innovative Networks and Services (BRAINS), Paris, France, 27–30 September 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 105–112.
5. Park, S.; Im, S.; Seol, Y.; Paek, J. Nodes in the bitcoin network: Comparative measurement study and survey. *IEEE Access* **2019**, *7*, 57009–57022. [CrossRef]
6. Donet, J.A.; Pérez-Sola, C.; Herrera-Joancomartí, J. The bitcoin P2P network. In Proceedings of the International Conference on Financial Cryptography and Data Security, Christ Church, Barbados, 3–7 March 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 87–102.
7. Imtiaz, M.A.; Starobinski, D.; Trachtenberg, A.; Younis, N. Churn in the bitcoin network: Characterization and impact. In Proceedings of the 2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC), Seoul, Republic of Korea, 14–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 431–439.
8. Fadhil, M.; Owenson, G.; Adda, M. A bitcoin model for evaluation of clustering to improve propagation delay in bitcoin network. In Proceedings of the 2016 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC) and 15th International Symposium on Distributed Computing and Applications for Business Engineering (DCABES), Paris, France, 24–26 August 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 468–475.
9. Biryukov, A.; Khovratovich, D.; Pustogarov, I. Deanonymisation of clients in bitcoin P2P Network. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS), Scottsdale, AZ, USA, 3–7 November 2014; pp. 15–29.
10. Neudecker, T.; Andelfinger, P.; Hartenstein, H. Timing analysis for inferring the topology of the bitcoin peer-to-peer network. In Proceedings of the 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), Toulouse, France, 18–21 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 358–367.
11. Wang, L.; Pustogarov, I. Towards better understanding of bitcoin unreachable peers. *arXiv* **2017**, arXiv:1709.06837.
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. *arXiv* **2021**, arXiv:2102.12774.
13. Grundmann, M.; Amberg, H.; Baumstark, M.; Hartenstein, H. Short Paper: What Peer Announcements Tell Us About the Size of the Bitcoin P2P Network. In Proceedings of the International Conference on Financial Cryptography and Data Security, Radisson Grenada Beach Resort, Grenada, 2–6 May 2022; Springer International Publishing: Cham, Switzerland, 2022; pp. 694–704.
14. Stouten, T. Hide and Seek: Different Scan Methods to Analyse Peer-to-Peer Based Blockchain Networks. Bachelor's Thesis, University of Twente, Enschede, The Netherlands, 2020.