Hide and seek - different scan methods to analyse peer-to-peer based blockchain networks

T. Stouten
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
t.stouten@student.utwente.nl

ABSTRACT

Blockchain-based peer-to-peer networks have demonstrated that such mechanisms are able to provide a secure and trustworthy way to perform transactions without the need for an old-fashioned 3rd party. However, not many studies have focussed on tools for analyzing such systems on reliability. In this research, we have focussed on two such tools, a passive and active node scanner. Such scanners can be used to discover the discoverability and reachability of nodes in a blockchain-based network, with which the entire network can be mapped.
We have placed these scanners in both the Bitcoin and Litecoin network, after which we have analysed and compared the different logs produced by these scanners. Both these scanners have shown their worth. The active scanner takes 20 minutes and is able to give an overview of the network while being unable to establish many connections with the discovered nodes. The passive scanner, which was placed in the network for 6 days, discovered more nodes within the Bitcoin network and was able to establish a connection with roughly 72% of these nodes. The passive scanner was unable to discover more nodes than the active scanner in the Litecoin network. However, it was still able to connect with roughly 76% of the discovered nodes. Both of these scanners produce capable and usable datasets. therefore it is task of researchers to make a choice based on the available time for research and on the need for reachability of these nodes.

Keywords

blockchain analysis, scanner nodes, active scanner nodes, passive scanner nodes, peer-to-peer network, Bitcoin, Litecoin

1. INTRODUCTION

The blockchain is a mechanism which has originally been designed to establish secure transactions between two parties. Securing a transaction is done by sending a block with this transaction to peers, who will use this block to calculate the next block which is sent to their peers, making it near impossible to change the values once a transaction has been done. The blockchain, in essence, is a trusted third party, like the banks we use with day to day transactions and payments. Since then it received attention from multiple businesses who are interested in the usage of blockchain as an alternative to existing solutions.

Multiple businesses have shown interest in the use of a blockchain to distribute ownership with this blockchain as 3rd party. Relying on the blockchain results in a decrease of need for old-fashioned 3rd parties. Such an old-fashioned 3rd party could have their own agenda, such that transmission of ownership to the receiver could fail, while the sender has transferred the ownership to this 3rd party. With the use of a blockchain, the majority of users must agree with the transaction, which results in a significantly lower chance of transaction loss or malfunction.

Due to this increase of interest in the blockchain, research has to be done on the stability and reliability of peer-to-peer networks for such a system. A peer-to-peer network is reliable when all nodes behave as expected and are reachable at all times. This, however, is rarely the case in a larger network. Most peers disconnect when their task in the network is complete, which makes a network less reliable as a transaction might fail when counts on a disconnected peer. Many others are hidden behind a firewall or a NAT, which makes them more difficult to reach.[2] Such nodes could still show up as neighbours while refusing connections, which in essence has the same effect as nodes which disconnect from the network.

To be able to make claims about the stability and reachability of such a peer-to-peer network we must analyze these networks with scanners. Such scanners are disguised as nodes of such networks and will log the status of each discovered node. Multiple different types of scanners have been developed over the years, each with their own benefits and drawbacks. In this research, we will analyze the logs produced by two such scanners.

2. PROBLEM STATEMENT

Due to the fact that blockchain is a decentralized network, the total amount of nodes, users in a network which can be used as steps between two endpoints, in this network is initially unknown. To gather statistics and data about all nodes in a network, different searching techniques have been developed. Two of such searching techniques are the active scanner node and the passive scanner node.
The active scanner node tries to discover as many nodes as possible within 20 minutes by requesting all discovered neighbours from its direct neighbours. While this produces a quick overview of a peer-to-peer network, it will contain many expired nodes, nodes which are no longer connected to the network.
A different approach to scan a network is with the implementation of a so-called passive scanner node. This

scanner behaves like a normal node with the emphasis of accepting as many incoming connections as possible.
It is expected that the data collected by these two different implementations will be rather different. The active scanner is likely to discover more unreachable nodes compared to reachable nodes, as the received neighbour-lists could be several hours old. The passive scanner has an emphasis on accepting all incoming connections without the pressure of collecting as many neighbours as possible. Due to this, it is expected that this scanner will discover a much higher percentage of reachable nodes and should be able to establish connections with unreachable nodes discovered by the active scanner.
At this stage, it is unclear which of the two scanning methods should be used for analysing a peer-to-peer blockchain network and if the type of blockchain network has an influence on this choice.

3. RESEARCH QUESTIONS

What is the best method to analyse a blockchain-based peer-to-peer network with respect to the physical location of the nodes and to specific infrastructure?
In this paper, we will research two different methods of analysing a network. One method uses a passive scanning algorithm, the other uses an active one. The datasets collected by these two different methods will be analysed for differences and comparisons based on the collected IP-addresses, port numbers and reachability of these nodes.
It is expected that an active scan is a good enough tool to get a quick overview of all nodes which use the standard port in a network, however, the passive scanner should be able to discover more nodes due to its longer scan duration. We will compare the number of nodes discovered by the two scanners, the used port numbers and the reachability of these nodes. Depending on these statistics, we will try to give suggestions on which type of scanner one should use in what circumstances.

What is the difference between analyzing a network with an active scanner node and a passive scanner node?
A network can be analyzed with a multitude of methods. In this paper, we will be using the datasets created by a passive node scanner and the datasets created by an active node scanner. As stated prior, we expect rather obvious differences between the two methods. The active scanner tries to discover as many nodes as possible by requesting every node for their neighbours.
The passive scanner mainly focuses on getting discovered by other nodes. While the scanner still has the functionality of a normal node, which means that it does try to connect and discover other nodes, it is nowhere near as aggressive as the active scanner. This method is likely to discover many more nodes than the active scanner due to the longer runtime and should be able to have more reachable "hard to discover nodes" due to its passive nature.
The expected difference between these methods is the amount and quality of the nodes it has discovered. The active scanner tries to discover as many nodes as possible in 20 minutes, which are possibly unreachable. The passive scanner relies mostly on nodes connecting to the scanner, which should result in a higher reachable node list, especially in the harder to find category.

"Do all stale nodes end up being purged by a peer to peer network?"
Nodes which are discovered by the scanners could either be reachable or unreachable. It is unknown if those unreachable nodes have been active in the last few days or if these nodes have been expired for a long time.
A peer-to-peer network is created on the basis of reachable neighbours. Setting up a connection between nodes could be done directly or via other nodes. If an enormous amount of nodes inside a network are expired but still advertised, then setting up a connection via in between nodes takes much longer and this defeats the benefits of a peer-to-peer network.
We will try to discover if stale nodes get purged with the use of datasets created by the active scanner in January 2020 and May 2020.
It is expected that a small number of nodes will show up in both datasets, as users will often mine in a pool for a longer period of time. Therefore, a small number of stale nodes could be present in datasets from both months when the users only connect to mining pools during specific times.

How many nodes inside a peer-to-peer blockchain network are hidden or otherwise hard to reach?
A peer-to-peer network consists of a large number of nodes. Some of these nodes are reachable, while others are hidden or protected in one way or another. For example, a node which originates from behind a NAT. This node will be able to connect to nodes in the peer-to-peer network while refusing to accept incoming connections. Such nodes are identifiable by having a port number higher than 1024 while not using the standard port for either Bitcoin or Litecoin.
Since most of these nodes are only in the network for a short while or refuse any incomming connections, it is unlikely to find a large number of these nodes with use an active scan. The passive scanner, however, should be able to discover and connect to more of these nodes, because of its long runtime and focus on incoming connections.

4. RELATED WORK

Blockchain-based p2p networks have been analysed numerous times to make estimations of the stability, security and overall health of such networks. In one of these analyses, it has been discovered that a large number of nodes (48%) [3] in such a p2p network fail to contribute anything due to having incorrect underlying protocols.

The data used by Kim et. al.[3] has been conducted by NodeFinder, which is a passive scanner which accepts all incoming connections and collects the Data Acces Object (DAO) of all peers, after which the connection will immediately be terminated. NodeFinder reconnects periodically to discovered nodes to track longitudinal properties.

New technologies for scanning entire networks are continuously developed. One of such scanning applications is ZMap[1]. Flooding a network with requests for data is unacceptable behaviour, which is why this modular application has been designed to scan addresses according to a random permutation. The practices outlined by ZMap to prevent unacceptable behaviour are useful to take into consideration when comparing different types of scanners.

A.Miller et. al. [4] have created an implementation of their AddressProbe technique, which is able to identify influential nodes in a network. This implementation could be used to compare the physical location of a node with the influential nodes and conclude whether or not these hard to reach nodes can be of great importance for the entirety of a p2p network.

S.Sariou et. al. [5] presented a measurement study on peers of two large file-sharing systems, Gnutella and Napster. These measurements include the availability of each node in the network. In reality, only 20% of the peers inside a network had an IP-level uptime of 93% or higher. Similar results can be expected within a p2p network for the blockchain.

5. METHODOLOGY

In this paper, large datasets containing information about multiple blockchain networks will be analyzed. This data has been collected with the use of both active and passive scanners.

5.1 Active scanner node

The active scanner is based on a normal peer-to-peer node. This scanner tries to connect to all direct neighbours. If the connection is successful, the active scanner node requests their list of neighbours. Once these potential neighbours have been logged, the active scanner tries to establish a connection with these neighbours. If our scanner succeeds, it requests a new list of neighbours and the cycle starts repeating itself. In twenty minutes a large part of the network will be scanned.
The active scanner has created multiple datasets for both the Bitcoin and the Litecoin network. These datasets contain a list of discovered nodes and whether or not the active scanner was able to connect with these discovered nodes. We will analyse this data based on IP-addresses and port number. We will get rid of all duplicate entries and check if the active scanner has been able to connect to these discovered nodes. Each node will be categorised by connectivity and by the used port. Categorisation by port number is important because each node which uses a port number higher than 1024 and does not use the standard port, 8333 for Litecoin and 9333 for Bitcoin, is possibly situated behind a NAT or is difficult to reach or discover.

5.2 Stale nodes

We will check how long a stale node, a node which is unable to establish a connection, will remain in a network. We have two scans per network per day. As such, we are able to generate a graph with the number of stale nodes which use the normal port and are also in the first scan of the dataset. While we will be unable to say for sure if these stale nodes have been online while our scanner was not, it should give us an indication of how many of these stale nodes stay advertised in the network and how many will get purged.

5.3 Network usage

For each node that has been discovered with the usage of the active scanner, we will determine the network usage. The usage type will be determined by looking up each IP-address in the ip2location database. This data could be used to identify the userbase of a peer-to-peer network. Are most nodes hosted in a datacentre or are these nodes hosted by a home-network? These results will be shown in a piechart.

5.4 Passive scanner node

The passive scanner is, as the active scanner, based on a normal peer-to-peer node. The important difference is that this passive node tries to accept and maintains each incoming connection. Each discovered node will be saved with the accompanying timestamp. If this scanner node establishes a connection, the type of this connection will be logged. The type is either incoming, a connection initiated by the discovered node, or outgoing, a connection established by the scanner. The scanner tries to reconnect to these nodes in a specified interval to see how long a node stays available in the network. This scanner will stay in the network for several days, after which the scanner node gets terminated after which the log gets processed.
Due to the different structure of the datasets created by the passive scanner node compared to the active scanner node, slightly different analysis methods will be used. The passive scanner has a single dataset which contains all nodes which have been discovered in six days, compared to 20 minutes. These discovered nodes will be categorised in the same manner as the active data has been categorised. The discovered data will be shown in a bar chart.
The connection type is important for our research. The passive scanner node logs for each established connection if this connection is requested by the scanner, outbound, or by the discovered node, inbound. With this data, we should be able to discover if the passive nature of this scanner results in more discovered and connected nodes which are using a high non-standard port.

5.5 Comparison

The findings for passive and active data will be compared against each other. Because the active scanner has a runtime of only 20 minutes, we have combined the active scans which have been collected in the same timespan as the passive scanner before the comparison could start. These two datasets will be compared within the categories which have been established earlier. We will also compare a single active scan against the passive scanner, to discover if a single 20-minute active scan can hold up to a 6-day passive scan. These findings will be shown in bar charts.

5.6 conclusion

Once all the analysis has been done, the data will be used to construct answers to our research questions and, if possible, we will give suggestions on which type of scanner should be use in what circumstances.

6. RESULTS

6.1 overview active scanner

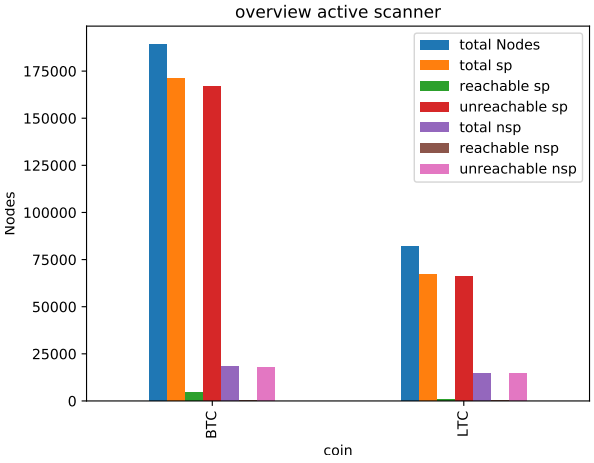

Figure 1.

Figure 1 has been constructed by taking the average of

ten separate Bitcoin datasets and the average of twenty separate datasets which were created in January. It shows how many of the discovered nodes are using the standard port (sp) and how many nodes are using the non-standard port (nsp).
This data gives us an overview of the discovered nodes in both cryptocurrency networks. In both cases, we discover many more nodes which use the standard port compared to nodes which use the non-standard port. It is very interesting to see that our active scanner is only able to establish a connection with 2.5% of all discovered nodes for Bitcoin and with 1.6% of all discovered nodes for Litecoin.

6.2 overlap active scanner

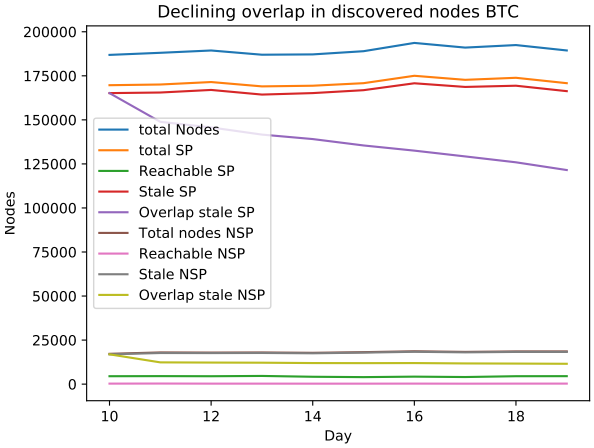

Figure 2.

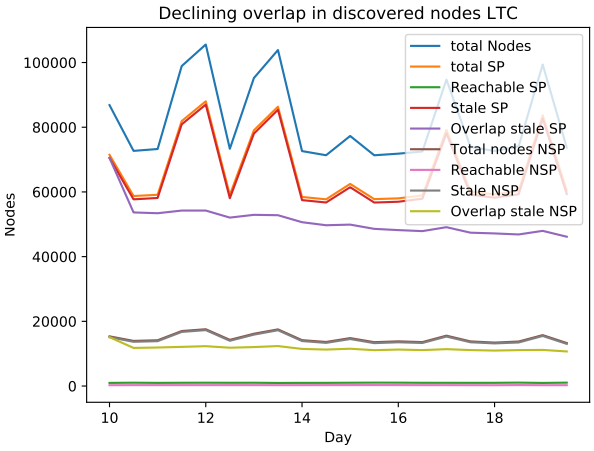

Figure 3.

We have discovered that most nodes will get purged from the network when they have been inactive for a certain amount of time. More than 54% of the reachable node will stay in the network for at least four months, while only 32.7% or less of all unreachable Bitcoin nodes and 18.7% of all unreachable Litecoin nodes will remain in the network for more than four months.
With the use of figure 2 we can conclude that the number of nodes which have been discovered in the first scan and each individual scan, slowly decreases in the Bitcoin network. For Litecoin we have a similar trend for the overlapping nodes, figures 3. However, we have an unstable amount of discovered nodes. This could be an indication that Litecoin users enter the network every other day or that many users only show up once in a the expire time has passed.

	Bitcoin	Litecoin
Total nodes January	285488	242465
total nodes May	366355	135621
Reachable January	12177	2121
Reachable May	12580	2297
Unreachable January	273311	240344
Unreachable May	353775	133324
Total Overlap	93335	46317
Overlap reachable	6594	1303
Overlap unreachable	86741	45014

Table 1. Overlap in nodes January and May

This same trend is visible in table 1. The bitcoin data has been created by combining six scans conducted between January 10 and January 15 at 12:00 and combining six scans conducted between May 3 and May 8 at 12:00. The Litecoin data has been created by combining 12 scans, each conducted at 8:00 and 17:00, between January 10 and January 10 and between May 3 and May 8.
Table 1 shows that only 32.7% of all Bitcoin nodes which have been discovered in January have also been discovered in May. 92.9% of this overlap was unreachable and 54.2% of all reachable nodes in January were also reachable in May.
In contrast with the Bitcoin network, the Litecoin network shows that we were able to discover way fewer nodes in may compared to January. Only 18.7% of all nodes discovered in January were also discovered in May. 61.4% of all reachable nodes in January were also reachable in May.

6.3 Network Usage

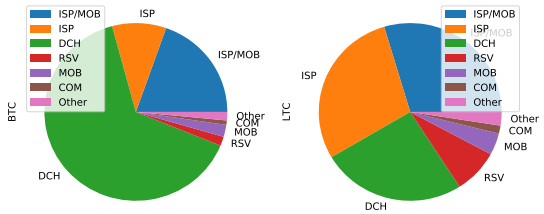

Figure 4.

Bitcoin and Litecoin have a different userbase, based on the type of networks which have been used, which can be seen in figure 4. Bitcoin has a large number of nodes which are web-hosted or which are hosted by a data centre, 64.6%. Litecoin shows the opposite. The Litecoin network has more nodes which are hosted on ISP or ISP/MOB networks, 58.5%, compared to nodes which are hosted by data centres or which are web-hosted, 28.9%.
This tells us that the Litecoin userbase is using more machine in a home environment, while the Bitcoin userbase prefers using web-hosted or data centre hosted nodes. This difference is important to take into account when we compare the working of the scanners on the different networks, as nodes which are hosted in a home environment are more likely situated behind a NAT.

6.4 overview passive scanner

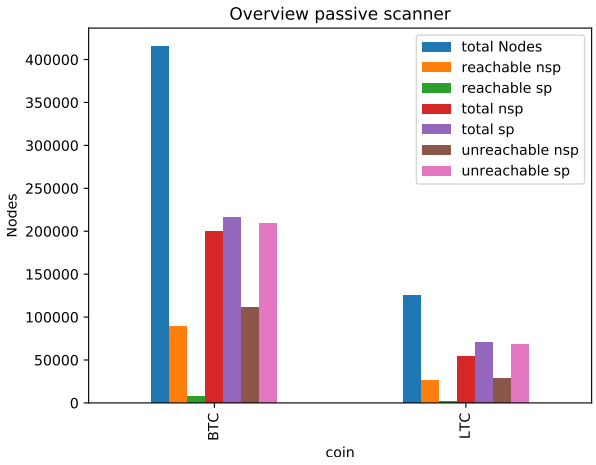

Figure 5.

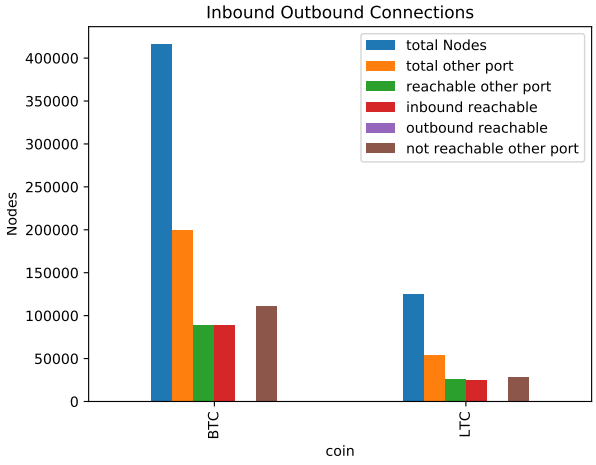

Figure 6.

The passive scanner, which has been run from the 3rd of May 2019 until the 8th of May 2019, is able to discover a larger percentage of non-standard port nodes compared to the active scanner. We have also discovered that the passive scanner has been able to reach more non-standard port nodes compared to the active scanner. The differences between using the passive scanner on the Bitcoin and the Litecoin are rather small. (figure 5) A slight larger percentage of nodes in the Bitcoin network is a non-standard port node and the passive scanner has been able to discover a larger amount of nodes within the Bitcoin network, compared to the Litecoin network. This shows the difference in size of these two networks.
The passive scanner does a much better job at establishing incoming connections compared to outgoing connections, as can be seen in Figure 6. Most of the connected non-standard port nodes, 44.7% for Bitcoin and 48% for Litecoin, are established by an incoming connection, 99.2% for Bitcoin and 96.3% for Litecoin. This means that the Passive scanner relies most on incoming connections, which is the complete opposite of the active scanner. Due to relying on incoming connections, a very small percentage of normal port nodes were in fact reachable by our scanner, as these nodes rarely initiated the connection.

6.5 comparison

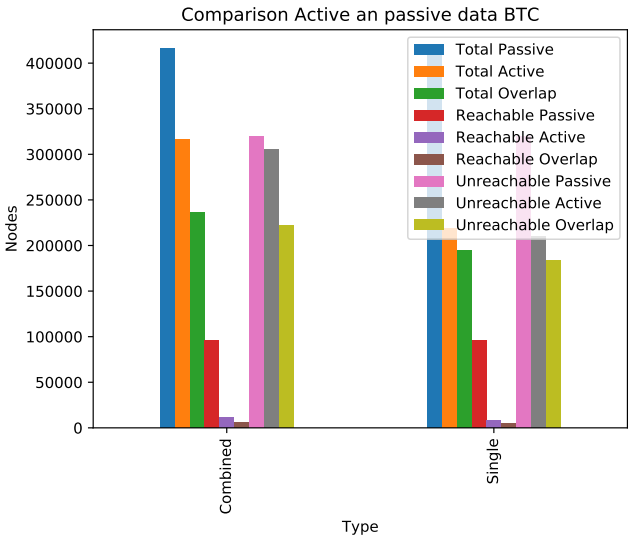

Figure 7.

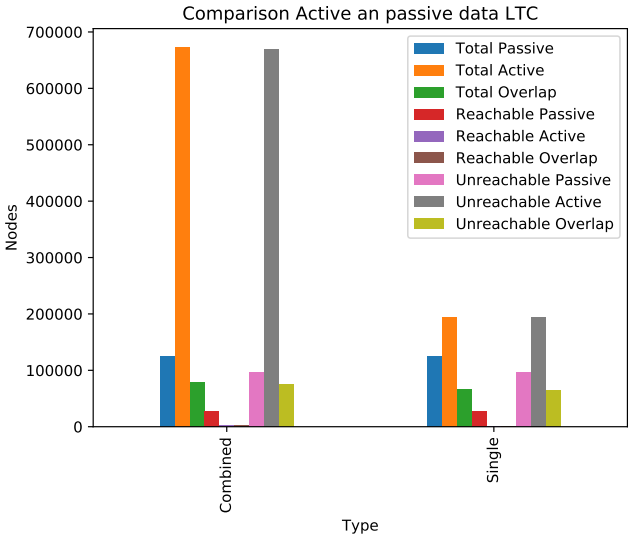

Figure 8.

In this Comparison, data created by the passive scanner is compared with data created by the active scanner. To make this a fair comparison, we compare the passive scans against a single active scan and a combination of six scans for Bitcoin and twelve scans for Litecoin.
The passive scanner is the better choice for Bitcoin when we base it on data alone (Figure 7). 56.9% of the nodes discovered by the passive scanner have been discovered by the active scanner. Of all discovered nodes, the passive scanner was able to reach 23.2% of the nodes, while the active scanner was able to reach 3.5% of these nodes. Comparing the passive dataset with a single active dataset brings a similar result. However, the amount of discovered nodes in the active dataset is less than 50% what the passive scanner has discovered.

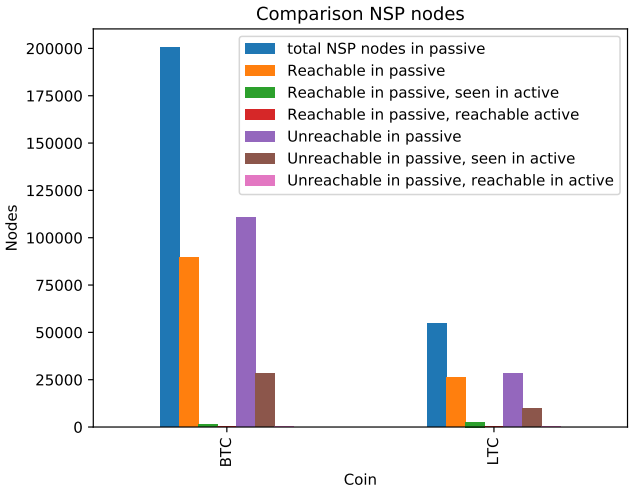

Figure 9.

The choice between the active and passive scanner is different for the Litecoin network. (Figure 8) The number of discovered nodes by the active scanner is more than 5 times larger than the number of nodes discovered by the passive scanner. The passive scanner was able to reach 22.4% of all discovered nodes, while the active scanner was able to reach 0.3% of all discovered nodes. When we compare the passive dataset to a single active dataset, we find similar results, however, the number of nodes discovered by this single active scan is only 1.5 times higher than the amount discovered by the passive scanner.)
When we purely look at the non-standard port nodes, then the active scanner is only able to reach 2% (BTC) and 11.2% (LTC) of all nodes reachable by the passive scanner. The active scanner is able to discover 25.7% (BTC) and 35.2% (LTC) of all unreachable non-standard port nodes which are discovered by the passive scanner. It appears that the active scanner has been able to discover a higher percentage of non-standard port nodes in the Litecoin network compared to the Bitcoin network. This is likely due to the fact that the Litecoin userbase is using ISP of ISP/MOB networks more often than the Bitcoin network is, thus hidden or hard to reach nodes are more common in the Litecoin network.

7. DISCUSSION

What is the difference between analyzing a network with an active scanner node and a passive scanner node?
Two different types of scanners have been used to scan two different cryptocurrency networks, the Bitcoin and the Litecoin network. These two scanners have vastly different ways to scan a network.
The active scanner is able to scan a network within 20 minutes. While it discovers quite a number of nodes, only a small number of nodes are reachable, 2.5% for the Bitcoin network and 1.6% of the Litecoin network.
The passive scanner scans a network over the span of several days, in this research we set the limit on 6 days. The passive scanner is able to discover more nodes in the Bitcoin network and fewer nodes in the Litecoin network, compared to the active scanner. Due to the fact that the passive scanner has an emphasis on incoming connections, the amount of reachable nodes is much higher compared to the active scanner. 71.7% (BTC) and 75.9% (LTC) of all discovered nodes were reachable.
Thus, the main differences between these two scanners are the time it takes to complete, the number of nodes which are discovered and the number of nodes the scanner can reach.

"Do all stale nodes end up being purged by a peer to peer network?"
We have compared all data collected in January of 2020 with all data collected in May 2020. The results are fascinating. The amount of nodes which have been discovered within the Bitcoin network has grown by a factor 1.28x, while the amount of discovered nodes discovered in the Litecoin network has shrunk with a factor 0.56x. The most interesting statistic here is that, for Bitcoin, 86741 unreachable nodes have been discovered in both the dataset of January an May. This means that only 30.4% of all discovered unreachable nodes from January have stayed within the Bitcoin network for over 4 months.
This number is even lower for Litecoin. 18.6% of all unreachable nodes discovered in January have stayed in the network until at least May. This statistic could, however, be misleading due to the fact that the number of discovered nodes in May is much lower than in January.
We can say for certain that at least 70% of the discovered stale nodes will be removed from the network within 4 months. Whether or not the remaining nodes have been connected to this network within these four months is unknown.

How many nodes inside a peer-to-peer blockchain network are hidden or otherwise hard to reach?
This is tricky to answer. Based on our active data, we must conclude that only 9.56% (BTC) and 18.05% (LTC) of the discovered nodes do not use the standard port. However, when we base our answer on the date created by our passive scanner, the amount of nodes which do not use normal port is 48.16% (BTC) and 40.80% (LTC).
Our passive scanner stays in the network for multiple days while out active scanner stays in the network for 20 minutes. Therefore, we can only conclude that when logging all discovered nodes for the duration of 6 days, the amount of such nodes is between 40% and 50%. To verify if this percentage is accurate for all blockchain networks, our passive scanner should be used in multiple other networks.

What is the best method to analyse a blockchain-based peer-to-peer network with respect to the physical location of the nodes and to specific infrastructure?
Based on the data we have used for this research, it is difficult to state which of the two scanners is the scanner of choice for all cases.
We have discovered that a single run of the active scanner is able to discover a higher number of nodes in the Litecoin network compared to the number of nodes discovered by the passive scanner with a single run. However, the opposite is true for the Bitcoin network. The passive scanner is able to discover more nodes than 6 separate active scans.
While the active scanner is able to discover more nodes than the passive scanner in the Litecoin network, the passive scanner is better at discovering nodes which are reachable. This is the case for both Bitcoin and Litecoin network. The passive scanner is able to reach between 22% and 24% of all nodes it discovered, while the active scanner is not able to connect to more than 4% of all discovered

nodes.
The run-time of the scanners should not be forgotten in this comparison. If a quick overview of all discoverable nodes in a peer-to-peer blockchain network is all that is required, an active scanner might sufficient. However, if multiple days are available for scanning the network and it is important that the discovered nodes are reachable and if the goal of the research is to discover nodes which are most likely behind a NAT, then the use of the passive scanner is recommended.

8. CONCLUSION

In this research, we have taken a look at two different peer-to-peer scanners and have established a recommendation which scanner is the best option under what circumstances.
The passive scanner is the best choice when the researcher has several days of time to run the scan and requires that most discovered nodes are reachable.
The active scanner is the best choice is the user only requires a quick overview of all nodes once present in the network, without being able to establish a connection with most of these discovered nodes.

9. FURTHER WORKS

In this study we have discovered major differences between the active scanner and the passive scanners for the Bitcoin and Litecoin blockchain network. Further research has to be conducted to ensure that the given recommendation does not change over time and to establish recommendation for other blockchain networks.
More data should be collected by running both scanners again in the Bitcoin and Litecoin network, as well as in different blockchain networks. These result should be compared to our findings to ensure a good recommendation.

10. REFERENCES

[1] Z. Durumeric, E. Wustrow, and J. A. Halderman. Zmap: Fast internet-wide scanning and its security applications. In *22nd USENIX Security Symposium (USENIX Security 13)*, pages 605–620, Washington, D.C., Aug. 2013. USENIX Association.

[2] D. Kegel, P. Srisuresh, and B. Ford. State of peer-to-peer (p2p) communication across network address translators (nats), 3 2008.

[3] S. K. Kim, J. Mason, Z. Ma, A. Miller, S. Murali, and M. Bailey. Measuring ethereum network peers. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC*, pages 91–104. Association for Computing Machinery, 10 2018.

[4] A. Miller, J. Litton, A. Pachulski, N. Gupta, D. Levin, N. Spring, and B. Bhattacharjee. Discovering bitcoin ' s public topology and influential nodes. 2015.

[5] S. Saroiu, K. P. Gummadi, and S. Gribble. A measurement study of peer-to-peer file sharing systems. *Proc SPIE*, 03 2002.

.