ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed
Computing

Measuring Cryptocurrency Networks

Bachelor's Thesis

Hyun-Min Chang
changh@ethz.ch

Distributed Computing Group
Computer Engineering and Networks Laboratory
ETH Zürich

Supervisors:
Dr. Lucianna Kiffer, Lioba Heimbach
Prof. Dr. Roger Wattenhofer

January 27, 2023

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisors, Dr. Luciana Kiffer, and Lioba Heimbach for their invaluable guidance and support throughout the development of this thesis.

I am deeply grateful for their patience, encouragement, and enthusiasm, which have been a constant source of motivation for me. Their expertise and insights have been instrumental in helping me to shape my research and to bring this project to fruition.

I would also like to thank Prof. Dr. Roger Wattenhofer and the Computer Engineering and Networks Laboratory for allowing me to conduct this study.

Finally, I would like to acknowledge the support and encouragement of my family and friends, who have always believed in me and helped me to stay focused on my goals.

Abstract

In this Bachelor's thesis, I aimed to gain a broad overview of the node activity for multiple cryptocurrencies by gathering data from various node explorers and consolidating the data for analysis. The objective of this study is to understand the distribution of nodes across different cryptocurrencies and identify any patterns or trends that may exist. The results show that when looking at the overall volume of cryptocurrencies, Bitcoin is leading by a large margin. This comes as no surprise as Bitcoin represents the most well-known cryptocurrency today and consistently has the largest market capitalization. I observed that the majority of nodes for almost all of the observed cryptocurrencies are located in either the USA or Germany and also noticed a general downward trend in node activity for various cryptocurrencies during the data collection period. It is interesting to note, however, that the distribution of the total number of nodes and that of the number of active nodes show clear discrepancies in the data on Bitcoin. Surprisingly, I also observe that two explorers that report on Bitcoin node data, and therefore should theoretically report similar information, have large discrepancies on the nodes they report. Reasons for this could lie in the methodology used by the node explorer to gather the data, or that the data is received differently depending on the location where the data is gathered.

In this paper, I will present our findings, discuss the limitations of the study, and conclude with recommendations for future research.

Contents

Chapter 1

Introduction

1.1 Background and Context on the Peer-to-Peer Network Architecture

Cryptocurrency networks rely on peer-to-peer (P2P) network architecture to enable decentralized and distributed systems. P2P networks allow for equal peers, who can send and receive transactions and blocks without the need for a central authority. P2P networks are decentralized, which means that there is no central point of control or failure. This makes them resistant to censorship because there is no single point that can be targeted to shut down the network. Additionally, if one node in the network goes down, the network can continue to function because the data and functionality are distributed across multiple nodes. This makes P2P networks more robust in the face of failures or attacks, as the network can continue to operate even if some nodes are unavailable.

Blockchain, a key technology in cryptocurrencies, utilizes a P2P network for a secure and transparent recording of transactions. This distributed database, known as a blockchain, contains blocks of multiple transactions which cannot be altered or deleted once added, making it an immutable and auditable ledger. This feature can potentially revolutionize various industries by allowing for secure and efficient P2P transactions. The seamless flow of transactions and blocks within the P2P network is crucial for the proper functioning of a cryptocurrency.[1]

Cryptocurrency networks consist of a decentralized network of nodes that participate in verifying and confirming transactions. These nodes can be run by individuals or organizations, and they contribute to the network by providing computational power and storage space. The process of validating transactions and adding them to the blockchain is called mining, and it requires the use of specialized software and hardware. To ensure the security and integrity of the network, cryptocurrencies rely on consensus algorithms, which are protocols that enable nodes to reach an agreement on the state of the blockchain. There are various types of consensus algorithms, such as proof-of-work, proof-of-stake, and delegated proof-of-stake, each with its trade-offs and benefits.[2][3]

Given the importance of P2P network architecture in cryptocurrency, it is crucial to understand the current state and trends of these systems and the factors that drive changes in node activity. This Bachelor's thesis aims to contribute to this understanding by collecting and analyzing data on the node activities of multiple cryptocurrencies from publicly available node explorers.

1.2 Research Objectives

The main objective of this Bachelor's thesis is to gain a comprehensive understanding of cryptocurrency networks through the measurement and analysis of node activity data.

To achieve this, the following objectives were set:

- Collect data on the node activities of multiple cryptocurrencies from publicly available node explorers and API sources, using self-written Python scripts.
- Clean and plot the data in a sensible manner.
- Gain an understanding of how to gather, process and analyse data using Python.

1.3 Related Work

There has been extensive research on the peer-to-peer network and block propagation mechanisms of various cryptocurrencies. For example, Decker and Wattenhofer [4] study the spread of information in the Bitcoin network, while Kiffer et al. [5] examine the inner workings of the Ethereum network.

Various research has also focused on the security of these systems. For example, Heilman, Ethan, et al. [6] present and analyze an attack on the peer-to-peer network of Bitcoin in their work, while Gervais, Arthur, et al. [7] study the security of 'Proof of Work' blockchains in their research. These studies provide insight into the vulnerabilities and potential weaknesses of these decentralized systems and are important for understanding the overall security of cryptocurrencies.

While these studies provide valuable insights into the functioning of individual cryptocurrency networks, this research aims to take a broader perspective and gain an overview of node activity across multiple cryptocurrencies.

Chapter 2

Data Collection

2.1 Overview of Data Sources and Methods

For this study, data on the node activities of multiple cryptocurrencies were collected from publicly available node explorers and API sources. The node explorers used in this research were Bitnodes [8] for Bitcoin, Ethernodes [9] for Ethereum, and Etcnodes [10] for Etherium Classic. Data from Bitnodes was acquired with their API, while data from Etcnodes, and Ethernodes was scraped with web scrapers. Additionally, data from Blockchair [11] was obtained through the use of their API, which provides access to the node activities for multiple cryptocurrencies, including Bitcoin, Bitcoin Cash, Dogecoin, Dash, Zcash, Litecoin, and Groestlcoin. All the code for data collection, as well as data analysis and data plotting, were written in Python.

As both Bitnodes and Blockchair collect data on the Bitcoin network, one would expect that their data would be consistent. Any discrepancies observed may indicate the influence of additional factors.

Out of the four data sources used (Bitnodes, Ethernodes, Etcnodes, and Blockchair), only Bitnodes included information on nodes utilizing the TOR network. This is a valuable piece of information as it can provide insight into potential disparities in activity between the TOR and non-TOR networks.

The process of obtaining data from the Bitnodes and Blockchair API was straightforward, as it allowed for direct access to the current status of the respective networks in the form of JSON files. These files were subsequently converted and saved as CSV files.

Obtaining data from Ethernodes and Etcnodes required the use of web scraping techniques, as these sources did not offer an API. Initially, an attempt was made to extract the data by parsing the HTML of the websites using the Python request library and searching for '</table>' entries. However, this approach was unsuccessful as the tables were generated using JavaScript and were not present in the HTML data when accessed using a Python script.

To overcome this challenge, a tutorial provided by Zoltan Bettenbuk 'Build a Javascript Table Web Scraper With Python in 5 Steps'[12] was utilized to learn how to directly access the JSON from which the tables were generated by using the browser's Developer Tools (accessed by pressing F12 in Google Chrome) to find the 'Request-URL' and scraping it directly. This approach successfully extracted the necessary data from Ethernodes and Etcnodes.

2.2 Data Collecting

To collect a comprehensive dataset for this study, data was collected from multiple sources of cryptocurrency networks. Four custom-written web scrapers were developed, one for each of the sources: Ethernodes, Etcnodes, Bitnodes, and Blockchair. The Blockchair scraper was designed to go through all the cryptocurrencies available in the Blockchair API.

A shell script was developed to run all the web scrapers to streamline and automate the data collection process. This script was scheduled to run automatically using crontab on a Linux machine at the Swiss Federal Institute of Technology in Zurich. The data collection process was configured to run every hour, starting from December 4th, 2022 and ending on January 10th, 2023. The script was run every hour during testing to ensure accurate and up-to-date data collection, as it was observed that the data changed hourly.

The collected data was stored in directories named after the source and in CSV format. The files were named according to the website that was scraped, with the date and hour appended to the end of the file name in the format <sourcename yyyy-mm-dd hhmmss.csv>. In addition to the data fields provided by each source, a field for "Creation Date" was added to the dataset. This field refers to the date the CSV file was created and was added as an additional measure to ensure that the data collection process and the data integrity can be accurately tracked.

IP
qp6ro3mnogsi7manj3gdt5xhic43n45fumvg527z5uvtoy3vyp7re6yd.onion:8333
\[2a01:4f8:222:16d6::2\]:8333

Protocol Version	User agent	Connected since	Services	Height
70016	/Satoshi:24.0.1/	2023-01-09	1037	771476
70015	/Satoshi:0.18.1/	2022-12-01	1037	771476

Hostname	City	Country Code	Latitude	Longitude
			0	0
2a01:4f8:222:16d6::2		DE	51.3	9.49

AZN	Organization name	Creation Date
TOR	Tor network	2023-01-11 20:00:00
AS24940	Hetzner Online GmbH	2023-01-11 20:00:00

Table 2.1: Example of Bitnodes CSV file format: header and the first couple of rows

IP	User agent	Country code	Height
1.116.110.123:8333	/Satoshi:24.0.0/	CN	765714
1.234.82.86:20000	/Satoshi:0.15.1/	KR	0

Flags	Creation Date
1032	2022-12-03 13:00:00
12	2022-12-03 13:00:00

Table 2.2: Example of Blockchair CSV file format: header and the first couple of rows

ID	IP	Port	Client	Client version	OS
00174acc4bbc...	3.15.229.64	30303	besu	22.10.1	linux
001e4616c1c1...	63.35.183.159	30303	geth	1.10.25	linux

Last Update	Country	In Sync	ISP	Creation Date
2022-11-30...	United States	0	Amazon.com	2022-12-03...
2022-11-29...	Ireland	1	Amazon.com	2022-12-03...

Table 2.3: Example of Ethernodes CSV file format: header and the first couple of rows

ID	Client	Fork ID	Local IP
48600f9c...	CoreGeth/v1.12.9...	2144451365	159.203.56.33:30369
51506735...		2144451365	159.203.56.33:47958

IP	Host Name	ETH Version	Snap Version
177.75.4.25:27254		66	1
3.80.12.255:30753	ec2....amazonaws.com	66	

Last Seen	Postal Code	City	Region	Country
03.12.2022	72300-000	Brasília	Federal District	Brazil
03.12.2022	20147	Ashburn	Virginia	United States

Coordinates	Time Zone	Organisation	Creation Date
-15.8872,-48.1508	America/Sao Paulo	AS28178 ...	2022-12-03 ...
39.0437,-77.4875	America/New York	AS14618 ...	2022-12-03 ...

Table 2.4: Example of ETC-Nodes CSV file format: header and the first couple of rows

The data collection process generally ran smoothly, however, there were two instances where data was not collected. The first instance of data loss occurred on December 18th, 2022 from 8:00 to 16:00, where an outage on the Blockchair API resulted in no data being collected. The second instance occurred from January 4th, 2023 18:00 to January 5th, 2023 9:00, where no data was recorded for Bitnodes, Blockchair-Bitcoin, Ethernodes, or Etcnodes. Despite efforts to determine the cause of this data loss, the reason is still unknown.

It is worth mentioning that the instances of data loss, while unfortunate, represented a small fraction of the overall dataset, and did not have a significant impact on the study's conclusions. Specifically, the total amount of data lost was 116 CSV files out of 9360, which represents roughly 1.23% of the dataset.

2.3 Data Cleaning

2.3.1 Identification and Removal of Missing or Duplicate Values

As mentioned in Section 2.2, there were two instances of data loss during the data collection process.

It was decided that these instances of data loss would not significantly impact the study's overall findings as they represent a small proportion of the overall dataset and were limited to a few hours rather than days. Therefore, this data was not considered during the analysis. Additionally, it was also found that there were missing values for certain entries, particularly for information that was not considered crucial for the analysis of this study, such as the client version, city, coordinates, or timezone as can be seen in tables 2.1 and 2.4.

It is worth noting that in the process of consolidating the data for Etcnodes, the port information had to be removed. This was due to the observation that the port number associated with a given IP address seemed to change on a daily basis for the majority of entries. This behaviour was in contrast to the other data sources where the port information was consistent.

Despite being only available on Bitnodes, data on nodes utilizing the TOR network was retained as it has the potential to provide valuable insights into the activity on the TOR network and its potential differences from activity on the normal network.

2.3.2 Handling of Inconsistencies

In the process of data cleaning, some inconsistencies were identified in the country data. A Python script was used to compare the country data of the dataset with that of a geo-location service, such as ipinfo.io [13]. It was found that for a small number of entries (e.g. 100 out of 15'000 for bitnodes), the country information was not consistent. To address this issue, a Python script was used to update all inconsistencies with the country data obtained from ipinfo.io. It is important to note that entries for which an error occurred during updating were removed as there were only a maximum of 10 such occurrences per CSV file.

Chapter 3

Data Processing

3.1 Data Pre-Processing

To facilitate analysis, ‘merge_csv_per_day.py’ (A.2.3) was used to consolidate the data for each cryptocurrency on a daily basis using the IP address as a key to eliminating duplicate entries. This process resulted in separate CSV files for each day, containing the consolidated data for each cryptocurrency. Additionally, to provide an overall comprehensive view of the dataset, ‘merge_csv_all.py’ (A.2.3) was used to repeat the process on all the collected data.

3.2 Data Visualization

The data was visualized through the implementation of several Python scripts. These scripts were utilized to analyze the data and generate visualizations that provided insight into the state and trends of the cryptocurrency networks.

Firstly, ‘count_nodes_per_day.py’ (A.2.5) was written to count the number of active nodes per day for each cryptocurrency using the prepared consolidated data. The results were then plotted as a line graph, with individual plots generated for each cryptocurrency and a single plot that included all cryptocurrencies for comparison of scale.

Additionally, for Bitnodes two additional plots were created. One compares the number of nodes using TOR in Bitnodes with the number of nodes not using TOR, and another compares the number of nodes not using TOR in Bitnodes with the number of nodes in Blockchair-Bitcoin, as both provide data on nodes not using TOR for Bitcoin.

To supplement the line graphs, bar charts were created to show how many nodes were active for how many days. This provided a visual representation of the distribution of active nodes over time and helped to identify patterns in node activity, such as how many nodes were only active for a short or extended period of time.

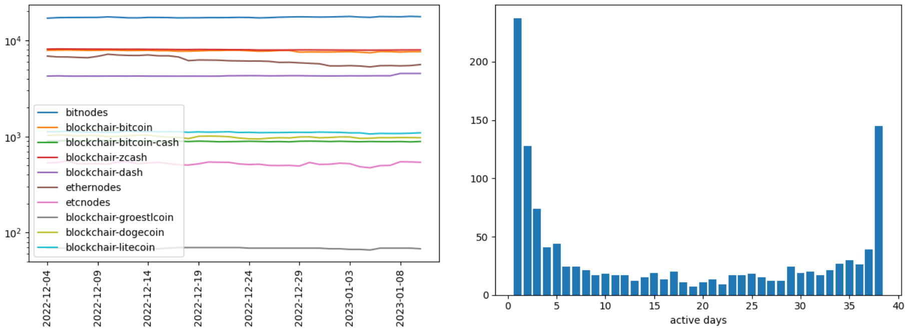

Figure 3.1: **Left**: Active nodes per day for all cryptocurrencies. **Right**: Duration of node activity for Bitnodes.

Then, to further analyze the active nodes per country, ‘count_countries.py’ (A.2.6) was written to go through all the consolidated data for all cryptocurrencies and count the number of entries per country or country code, depending on what information was available. ‘convert_country_code_to_country.py’ (A.2.6) was then used to convert all entries of country codes to the corresponding full name of the country using the ’countries.csv’ provided by developers.google.com [14] dataset. The resulting CSV files were then saved and prepared for further analysis, after which they were visualized as bar charts. To improve readability, the graphs only show the top 10 countries for each cryptocurrency. Finally, two bar charts were generated to compare Bitnodes and Blockchair-Bitcoin, as both show data on Bitcoin and should have matching data.

By analyzing the data on a per-day basis, it was possible to identify any unexpected patterns or trends in terms of the distribution of active nodes across different countries.

To enhance the visual representation of the country data, interactive heatmaps were created using the Python package ‘folium’. However, due to a lack of experience with the package, the heatmaps only display one dot per country and do not cover the entire geometry of each country.

To get the coordinates of each country, again ‘countries.csv’ was used. However, during this step, several problems emerged. Firstly, some countries had different names in the data from ethernodes.org compared to the names used in other node explorers, such as ‘Czechia’ instead of ‘Czech Republic’ or ‘Republic of Lithuania’ instead of ‘Lithuania’. To address this issue, a script was written to go through all the data from ethernodes.org and standardize the country names to match those used in the other node explorers.

However, this approach was later discarded as all the country information for all cryptocurrencies was updated with the script (A.2.4) using ipinfo.io as the common source for the data.

Unfortunately, some nodes were located in countries that were not included in ‘countries.csv’ used to obtain the coordinates. To resolve this problem, the coordinates for Curaçao[15] and Andhra Pradesh[16] had to be added manually.

More scripts (A.2.8) were written to count the number of nodes that were active for a certain number of days for each cryptocurrency. The results were plotted as a bar chart, providing insight into the level of activity of the nodes and allowing for the identification of any patterns or trends in the data, such as nodes that were active for only a short time.

Finally, ‘check_for_double_ip.py’ (A.2.9) was used to analyze the frequency of IP addresses appearing across multiple cryptocurrencies. The script processed the consolidated data for each cryptocurrency and counted the number of occurrences for each IP address. The resulting data were then plotted as a bar chart, where the x-axis represented the number of occurrences and the y-axis represented the number of unique IP addresses with that number of occurrences.

Chapter 4

Results

4.1 Total Node Count

From Figure 4.1, it can be seen, that the most widely used cryptocurrency is Bitcoin, then followed by Zcash, Ethereum, and Dash. These are in the thousands of nodes while Litecoin, Dogecoin, Bitcoin-Cash and Ethercoin are around the 1'000 range. The least used currency is Groestlcoin with a number under 100.

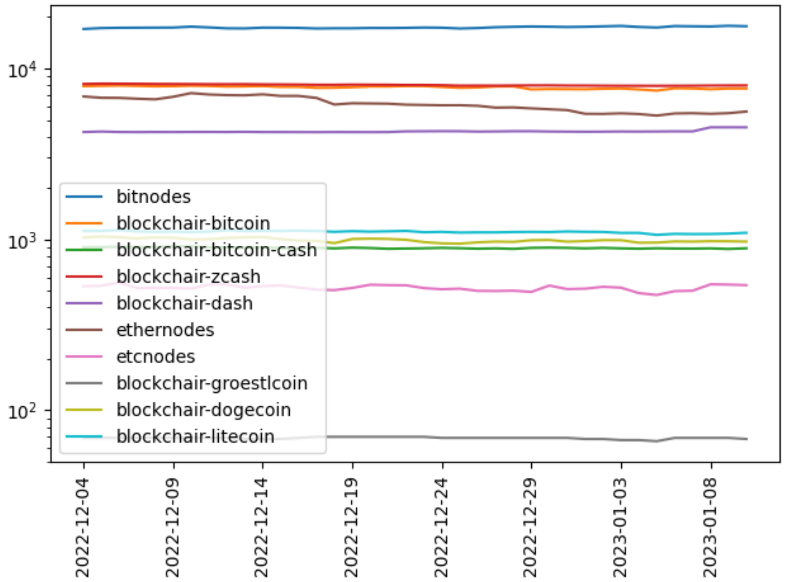

Figure 4.1: Activity of all cryptocurrencies on a Log scale in the y-axis

The plots in Figure 4.2 show that the overall number of active nodes for Bitcoin is experiencing a steady upward trend and that the increase stems from an increase in nodes utilizing TOR, as the number of active nodes not utilizing TOR is relatively stable, or potentially even decreasing.

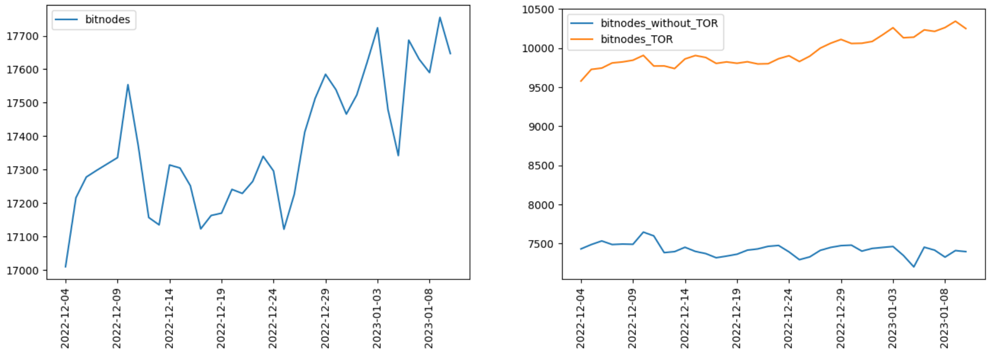

Figure 4.2: **Left**: Node activity recorded by Bitnodes. **Right**: Node activity recorded by Bitnodes split into nodes that use and do not use TOR

Comparing Bitnodes to Blockchair-Bitcoin shows that they do not record the same data, as the lines do not overlap.

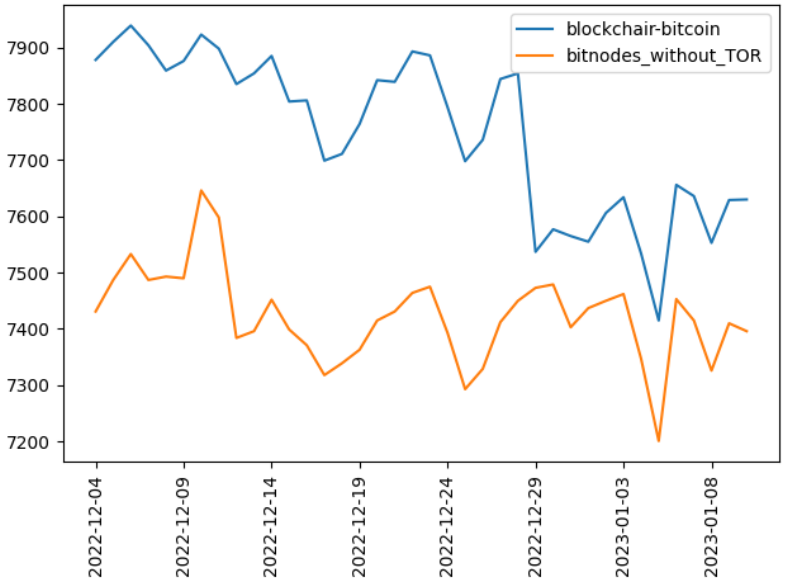

Figure 4.3: Comparison of node activity between Bitnodes and Blockchair-Bitcoin

When comparing all recorded crypotcurrencies, the node activity seems to be decreasing, as 6 out of the 9 recorded cryptocurrencies show a clear downward trend in activity.

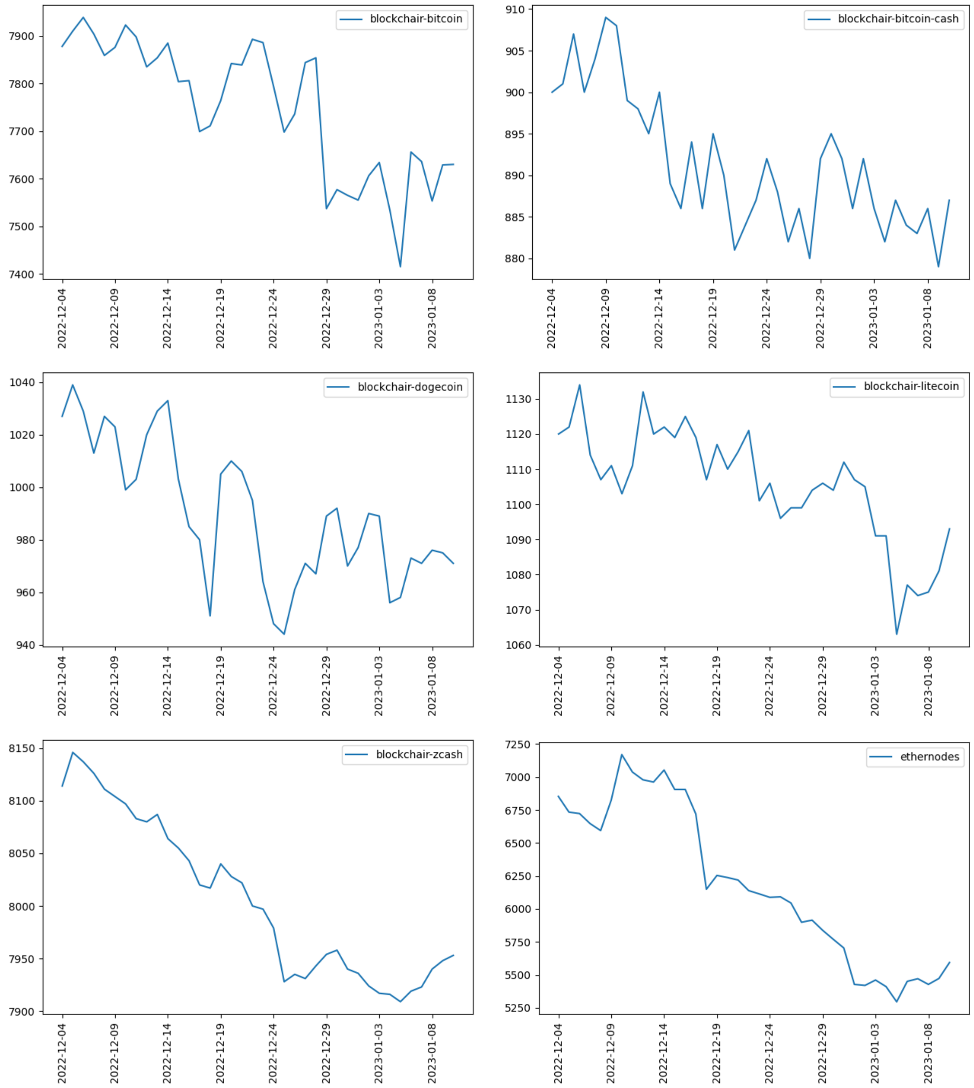

Figure 4.4: Activity of various cryptocurrencies on a downward trend

In Figure 4.5 we see that Blockchair-Dash and Etcnodes, both have seen a sudden increase. However, more data would be needed to conclude if these are only temporary or indicative of a sustained increase in activity.

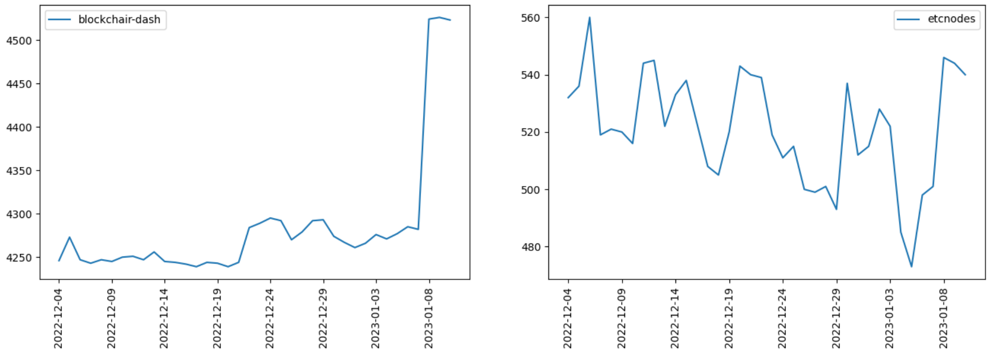

Figure 4.5: **Left**: Node activity of Blockchair-Dash. **Right**: Node activity of Etcnodes.

4.2 Total Country Data

In the analysis of node distribution by country, only the top 10 countries with the highest number of nodes are represented in the visualizations, with all remaining countries grouped together and labeled as "Other". It is crucial to acknowledge that the total number of nodes in each country in the "Other" category is significantly lower than that of the individually named countries.

Overall the USA (in yellow) has the highest amount of total nodes over all cryptocurrencies, closely followed by Germany (in light blue). These two countries hold the top 2 places in almost all cryptocurrencies as seen in Figure 4.2 and Figure 4.2.

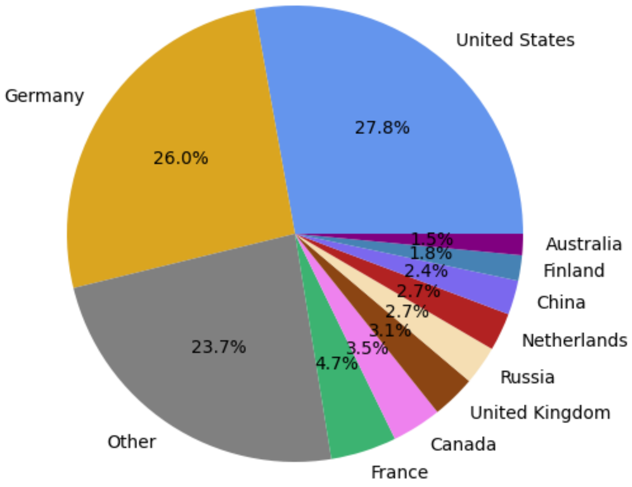

Figure 4.6: Total nodes per country over all cryptocurrencies combined.

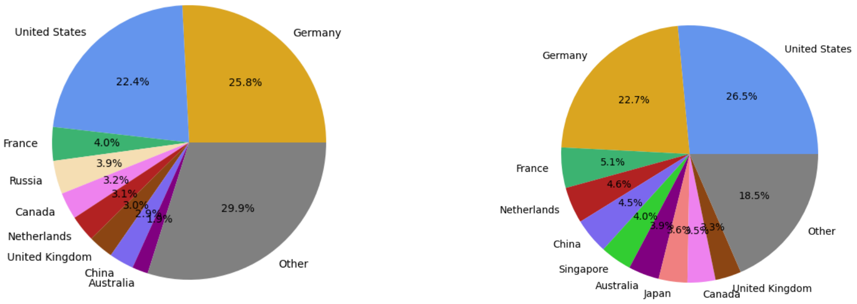

Figure 4.7: **Left**: Blockchair-Bitcoin. **Right**: Blockchair-Bitcoin-Cash.

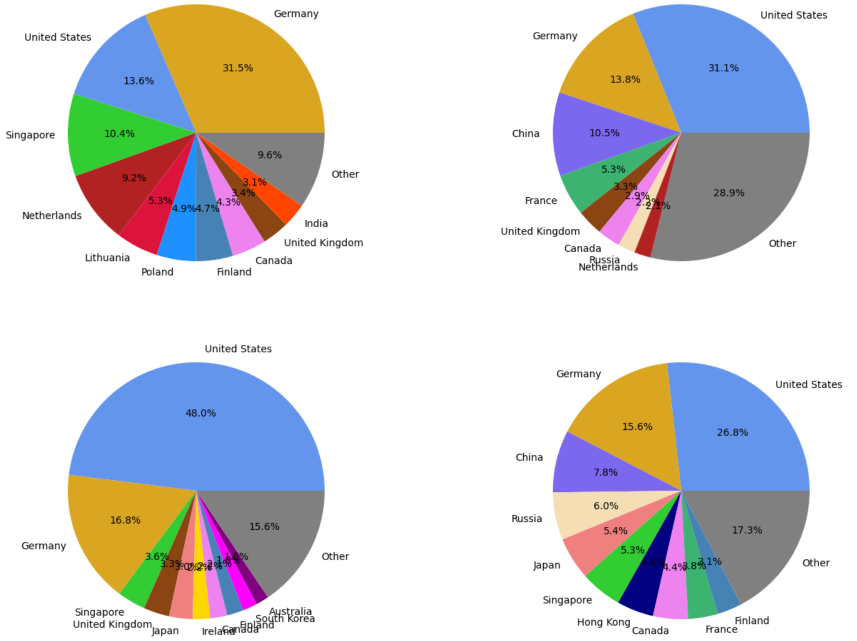

Figure 4.8: **Top-Left**: Blockchair-Dash. **Top-Right**: Blockchair-Dogecoin. **Bottom-Left**: Ethernodes. **Bottom-Right**: Etcnodes.

We can also make this observation by looking at the heatmap in Figure 4.9. The two bright red spots over the USA and Germany indicate the highest amount of activity.

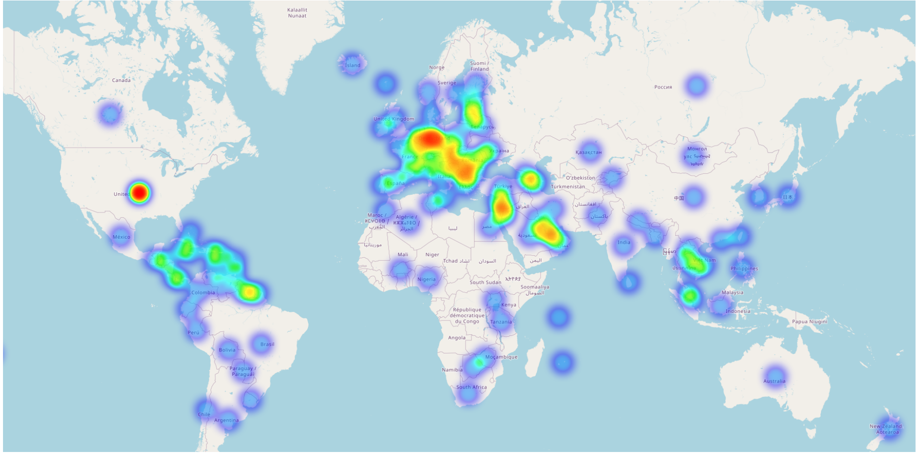

Figure 4.9: Heatmap of all cryptocurrencies combined.

4.3 Daily Country Data

These plots show the amount of daily active nodes in the top 10 countries for each cryptocurrency.

When comparing the daily country data with the total country data from the previous subsection, we can see that the USA and Germany have swapped places for Bitnodes, and Blockchair-Bitcoin.

To differentiate between the relative magnitudes of the various data, the total country data was represented in a bar chart format.

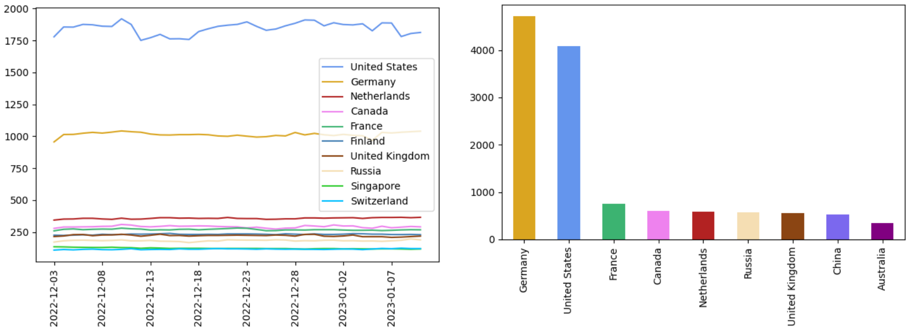

Figure 4.10: Comparison between daily (active) and total data for Bitnodes.

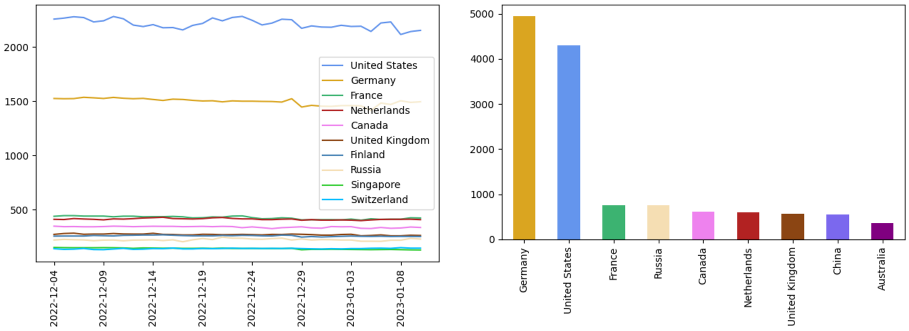

Figure 4.11: Comparison between daily (active) and total data for Blockchair-Bitcoin

4.4 Duration of Node Activity

The following plots show how many nodes were active for how many days.

A notable observation is that for Blockchair-Bitcoin-Cash and Blockchair-Dash, a significant proportion of the nodes exhibit a high degree of activity throughout the duration of the study as seen in Figure 4.12 from the peaks on the right.

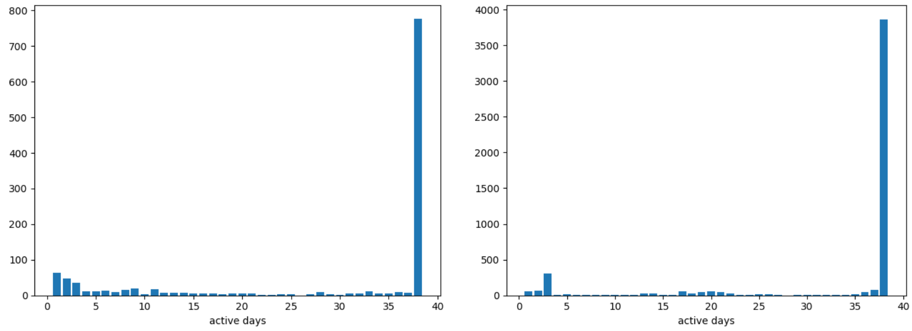

Figure 4.12: **Left**: Activity of Blockchair-Bitcoin-Cash. **Right**: Activity of Blockchair-Dash.

Contrarily, Blockchair-Bitcoin, and Etcnodes exhibit a large amount of activity, spanning only a few days, which can be seen in Figure 4.13 from the big peaks on the left.

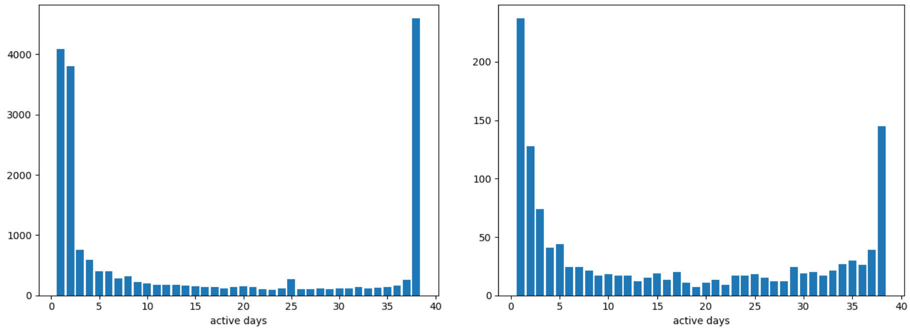

Figure 4.13: **Left**: Activity of Blockchair-Bitcoin. **Right**: Activity of Etcnodes.

4.5 IPs Appearing in Multiple Cryptocurrency Networks

In Figure 4.14 it can be observed that about 17’000 IPs appear in multiple cryptocurrencies but the majority of those are IPs that appear in both Bitnodes and Blockchair-Bitcoin, which is expected as both gather the same data.

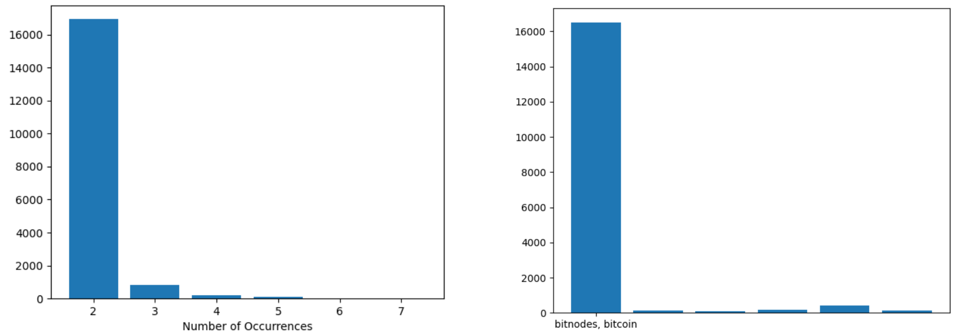

Figure 4.14: **Left**: This plot shows how many IPs appear in how many different node explorers.
Right: This plot shows which combinations appear how often. All labels except ’bitnodes,bitcoin’ were removed for clarity. ’bitcoin’ refers to Blockchair-Bitcoin, which was simplified to ’bitcoin’ for readability of the plot

Further visualizations can be found in the appendix in Section B.6.

Chapter 5

Discussion and Conclusion

5.1 Interpretation of Results

From Section 4.1 we can clearly see that looking at the overall volume of cryptocurrencies, Bitcoin is leading by a margin. This comes to no surprise as Bitcoin represents the most well-known cryptocurrency today.

Section 4.2 shows that the majority of nodes are located in either the USA (27.8%) or Germany (26.0%). It is interesting to note, however, that the distribution of the total number of nodes in Section 4.2 and that of the number of active nodes in Section 4.3 show clear discrepancies. While the country with the most Blockchair-Bitcoin nodes is Germany with 26.0%, the number of active Blockchair-Bitcoin nodes is visibly higher in the USA (Figure 4.11).

One possible explanation can be given by comparing the activity of nodes. In Section 4.4 we can see that more than half of the nodes recorded for Bitcoin have been active for only a couple of days. If many of these inactive nodes are those located in Germany, it would explain why Germany has a lower number of active nodes than the USA, despite having a higher number of total nodes.

It is also important to note, the lines in Figure 4.3 do not perfectly match. As both Bitnodes and Blockchair-Bitcoin focus on the same cryptocurrency, they should theoretically gather the same data. One reason for the discrepancy could lie in the methodology used by the node explorer to gather the data. Another explanation could be that the data is received differently depending on the location where the data is gathered.

This would coincide with the following discovery from the paper "Under the hood of the ethereum gossip protocol." [5]

> “We also find that a node’s location has a significant impact on when it hears about blocks, and that the precise behaviour of this has changed over time (e.g., nodes in the US have become less likely to hear about new blocks first).”

5.2 Limitations of the Study and Future Work

The code used for data collection, analysis and plotting within this Bachelor's thesis, is all self-written. As I do not specialize in web scraping, nor data collection through APIs, the code is prone to have inefficiencies and insecurities, and thus should not be used for operational purposes.

Working with large files (approx. 5GB of CSV files), also made it more difficult to properly process the data and maintain data integrity. This translated into a need for more complex data transformations to visualize and plot the data. While I believe that the data collection and transformation process was properly implemented, the code and the overall data collection and preparation process should be iteratively reworked and optimized.

The data for the active nodes per country was taken from ipinfo.io. This research paper assumes that the underlying information of the provider is accurate. However, this might not be the case. Particularly for IP addresses that are associated with virtual private networks (VPNs) or proxies, ipinfo.io's service might not always provide accurate results. As such, the results should not be taken for granted and should be subject to constant scrutiny. To ensure the accuracy of the data, more research into ipinfo.io's methodology would be required.

The same is true for the overall data collection process. For this research paper, only publicly available node explorers were used. This approach, due to its simplicity, facilitated the data collection. However, it also added unpredictable variables, as the use of node explorers prevented direct control over the data collection methodology. One could circumvent this problem by obtaining the data directly from the network. This would also ideally eliminate or resolve the identified discrepancies between the data from Bitnodes and Blockchair-Bitcoin.

By gathering data directly from the networks, it would also be possible to gather data on TOR usage for other cryptocurrencies, not just for Bitnodes, which would allow for a more conclusive and complete analysis.

This said, I believe that this paper has achieved its goal of providing an overview and understanding of the current landscape of cryptocurrencies, while also identifying potential discrepancies, which build the basis for further research.

Bibliography

[1] "Peer-to-peer blockchain networks: The rise of p2p crypto exchanges," https://learn.bybit.com/bybit-p2p-guide/peer-to-peer-blockchain-network/, 2022.

[2] . Blockchains, "Know everything about blockchain proof of work (pow)," https://101blockchains.com/blockchain-proof-of-work/, 2019.

[3] J. Howell, "Proof of stake vs delegated proof of stake," https://101blockchains.com/proof-of-stake-vs-delegated-proof-of-stake/, 2022.

[4] C. Decker and R. Wattenhofer, "Information propagation in the bitcoin network," in *IEEE P2P 2013 Proceedings*. IEEE, 2013, pp. 1–10.

[5] L. Kiffer, A. Salman, D. Levin, A. Mislove, and C. Nita-Rotaru, "Under the hood of the ethereum gossip protocol," in *International Conference on Financial Cryptography and Data Security*. Springer, 2021, pp. 437–456.

[6] E. Heilman, A. Kendler, A. Zohar, and S. Goldberg, "Eclipse attacks on bitcoin's peer-to-peer network," in *24th USENIX Security Symposium (USENIX Security 15)*, 2015, pp. 129–144.

[7] A. Gervais, G. O. Karame, K. Wüst, V. Glykantzis, H. Ritzdorf, and S. Capkun, "On the security and performance of proof of work blockchains," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 3–16.

[8] bitnodes, "bitnodes.io," https://bitnodes.io/api/.

[9] bitfly gmbh, "ethernodes.org," https://ethernodes.org/nodes/.

[10] E. Cooperative, "etcnodes.org," https://etcnodes.org/.

[11] blockchair, "bitnodes.io," https://blockchair.com/.

[12] B. Zoltan, "Build a javascript table web scraper with python in 5 steps," https://www.scraperapi.com/blog/scrape-javascript-tables-python/, 2022.

[13] I. LLC, "ipinfo.io," ipinfo.io.

[14] G. Developers, "countries.csv," https://developers.google.com/public-data/docs/canonical/countries_csv.

[15] LatLong.net, “Curaçao geographic information,” https://www.latlong.net/place/curacao-8711.html.

[16] LatLong.net, “Vijayawada, andhra pradesh, india geographic information,” https://www.latlong.net/place/vijayawada-andhra-pradesh-india-2126.html.

Appendix A

Source files

A.1 Gitlab Repository

All of the code and the data are available on the following Gitlab Repository:

https://gitlab.ethz.ch/disco-students/hs22/changh-measuring-cryptocurrency-networks

A.2 Scripts

Here are short descriptions of the scripts mentioned in the research. The scripts are listed in order of execution. Please refer to the GitLab repository to see the complete code.

A.2.1 Scrapers

- **bitnodes_scraper.py**: Gets the latest data from the Bitnodes-API by getting the JSON file from 'https://bitnodes.io/api/v1/snapshots/latest/', and compiling the data to a CSV file named 'bitnodes yyyy-mm-dd hhmmss.csv', which gets saved into a folder 'bitnodes' which the script creates.
- **blockchair_API_scraper.py**: Gets the latest data from the Blockchair-API for Bitcoin, Bitcoin-Cash, Dash, Dogecoin, Groestlcoin, Litecoin and Zcash, by looping through a list with the names of the cryptocurrencies and requesting the data from 'https://api.blockchair.com/{crypto}/nodes' where 'crypto' is the current cryptocurrency. The data gets compiled into CSV files named 'blockchair-{crypto} yyyy-mm-dd hhmmss.csv', which get saved into the folders 'blockchair/{crypto}' which the script creates.
- **ethernodes_scraper.py**: Gets the latest data from ethernodes.org by looping and requesting information about 100 nodes at a time (I couldn't

manage to request all the data at once and the maximum that worked for me was 100) from

```
'https://ethernodes.org/{incredibly long url}/
    start={index}&length={index+100}&search[value]=&
    search[regex]=false&\_={unix timestamp}'
```

' The data gets compiled into a CSV file named 'ethernodes yyyy-mm-dd hhmmss.csv', which gets saved into a folder 'ethernodes' that the script creates.

- **etcnodes_scraper.py**: Gets the latest data by requesting the page from 'https://peers.etccore.in/v5/nodes.json', which then gets compiled into a CSV file named 'etcnodes yyyy-mm-dd hhmmss.csv', which gets saved into a folder 'etcnodes' that the script creates.

A.2.2 Bitnodes IPv6 cleaner

- **clean_ipv6_bitnodes.py** removes all '[' ']' from the IPv6 entries of 'csv_combined_all/bitnodes.csv'.

A.2.3 Mergers

- **merge_csv_all.py**: Gets all the data gathered by the scrapers, combines all CSV files of one cryptocurrency into one pandas dataframe, removes the duplicate IP entries and saves the dataframes into a folder 'csv_combined_all/', as CSV files named '{crypto}.csv'.

- **merge_csv_per_date.py**: Gets all the data gathered by the scrapers, combines all CSV files of one date for one cryptocurrency into one pandas dataframe, removes the duplicate IP entries and saves the dataframes into folders 'csv_combined_per_day/crypto/', as CSV files named '{crypto_yyyy-mm-dd}.csv'.

A.2.4 Country Data Updaters

- **update_countries_{crypto}.py** get their respective CSV file from the folder '/scv_ combined_all/' and update the country data, by using the IP addresses to compare the data to the information provided by 'https://ipinfo.io/{IP}?token={API-Key}" and saving the result as ' {crypto}_updated.csv'. There is one script for every cryptocurrency as this can take a very long time, and by having multiple scripts, they can be executed simultaneously.

A.2.5 Total Node Count

- **count_nodes_per_day.py**: Goes through '/csv_combined_per_day/' and creates a CSV file for every cryptocurrency that contains data on how many nodes were active per day and saves into 'daily_node_count_analysis/data/'.
- **combine_data.py**: Combines the data in 'daily_node_count_analysis/data/' into one CSV file 'combined_data.csv'
- **create_plot.py**: Uses 'combined_data.csv' to create the plots as seen in Section 4.1.

A.2.6 Total Country Data

- **get_countrienumber.py**: Gets the 'countries.csv' from google.developers.com[14] adds two entries and saves the updated CSV file.
- **count_countries.py**: Goes through '/csv_combined_all/' and for all cryptocurrencies counts how many times each country code appears and saves the result to 'country_data_analysis/country_count/{crypto}_country_count.csv'
- **convert_country_code_to_country.py**: Converts country codes in 'country_data_analysis/country_count/{crypto} _country_count.csv' to country names using 'countries.csv' as a reference.
- **compare_bitnodes_blockchair-bitcoin.py**: Creates a CSV file that contains all the nodes that are in both Bitnodes and Blockchair-Bitcoin. Also updates country codes to country names.
- **create_bar_chart.py**: Creates the bar charts presented in Section 4.2
- **create_bar_chart_bitcoin_comparison.py**: Creates the bar chart comparing the country data of Bitnodes with the data of Blockchair-Bitcoin
- **create_pie_chart.py**: Creates the pie charts
- **create_heatmap.py**: Creates the heatmaps presented in Section B.7

A.2.7 Daily Country Data

- **get_countries_csv.py**: Gets the 'countries.csv' from google.developers.com[14] adds two entries and saves the updated CSV file.

- **count_countries_daily.py**: Counts the amount of nodes active per country per day per cryptocurrency and saves the daily country count in 'country_data_analysis_daily/country_count_per_day/{crypto}/' as CSV files.
- **convert_country_code_to_country.py**: Converts the country code entries of all CSV files in 'country_data_analysis_daily/country_count_per_day/{crypto}/' to country name.
- **combine_per_crypto.py**: Combines the daily CSV files into one matrix per cryptocurrency with the rows being the date and columns being the country and save them to 'country_data_analysis_daily/country_count_matrices/'.
- **plot_country_daily.py**: Plot the data in 'country_data_analysis_daily/country_count_matrices/' as presented in Section 4.3.

A.2.8 Duration of Node Activity

- **count_ip_occurences.py**: Goes through '/csv_combined_per_day/' and counts how many IPs have entries for how many days and saves the result in 'active_days/json/json_{crypto}.json'
- **count_ip_occurences_without_tor.py**: Goes through '/csv_combined_per_day/bitnodes/' and counts how many IPs have entries for how many days and saves the result in 'active_days/json/json_bitnodes_no_tor.json'
- **plot_ip_occurences.py**: Plots the data in 'active_days/json/' as presented in Section 4.4

A.2.9 IPs Appearing in Multiple Cryptocurrency Networks

- **check_for_double_ip.py**: Goes through '/csv_combined_all/' and counts which IPs appear how many times in which cryptocurrencies and saves this in '/check_for_double_ip/duplicated_ips.csv'.
- **plot_dubplicated_ips.py**: Plots '/check_for_double_ip/duplicated_ips.csv' as presented in Section 4.5

A.2.10 Check for Discrepancies

The scripts in the folder '/check_for_discrepancies/' were used at the beginning stages of the research to check how large the number of discrepancies in the country data was compared to the data given by ipinfo.io. Initially, those discrepancies were just deleted, but as in the final data processing and analysis the country data were updated with the data from ipinfo.io these scripts became redundant.

Appendix B

All Plots

B.1 Total Node Count

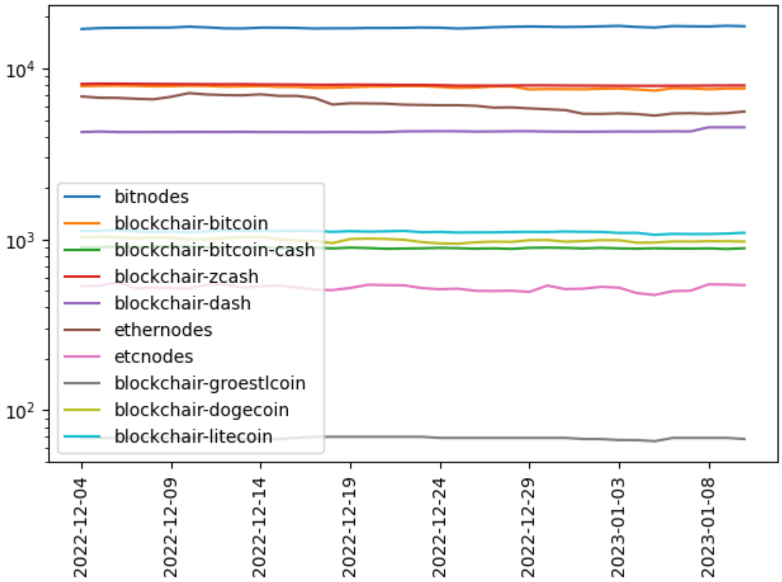

Figure B.1: Activity of all cryptocurrencies

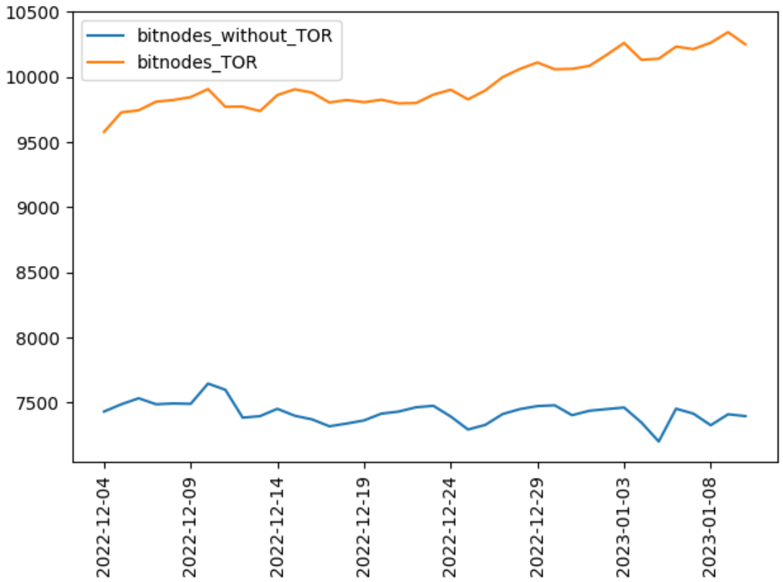

Figure B.2: Activity of all cryptocurrencies

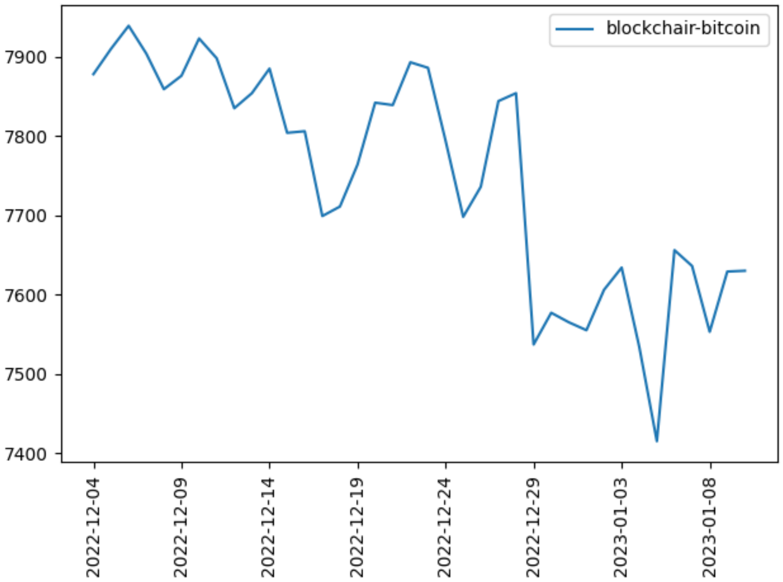

Figure B.3: Activity of Blockchair-Bitcoin

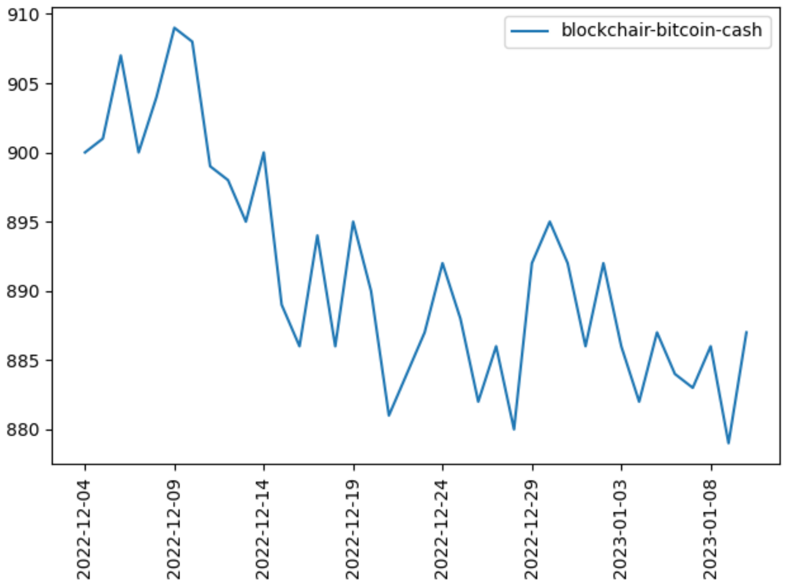

Figure B.4: Activity of Blockchair-Bitcoin-Cash

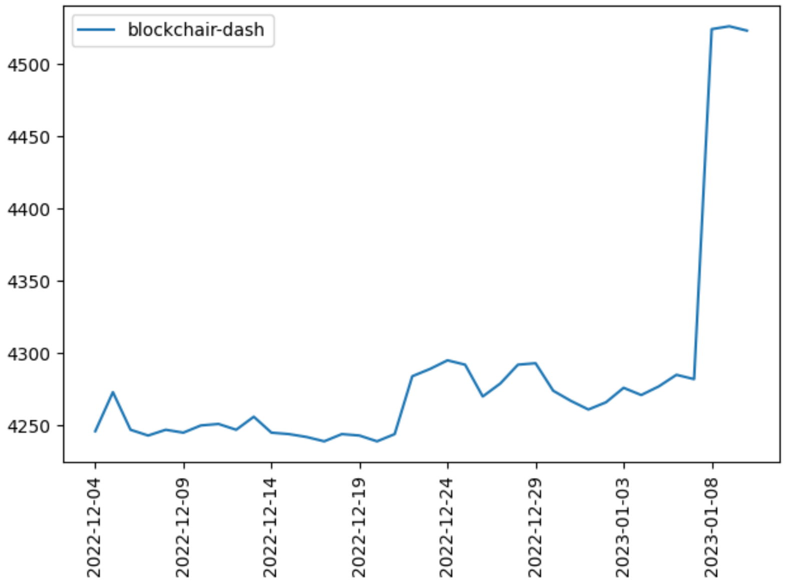

Figure B.5: Activity of Blockchair-Dash

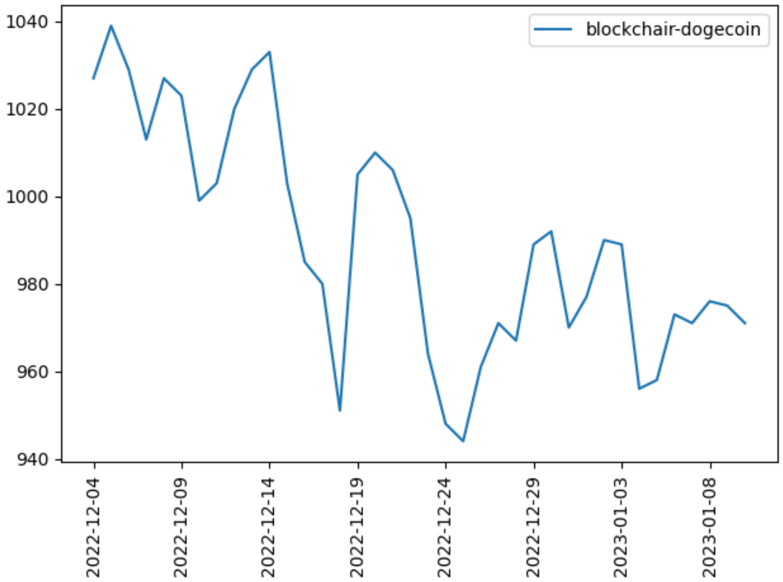

Figure B.6: Activity of Blockhair-Dogecoin

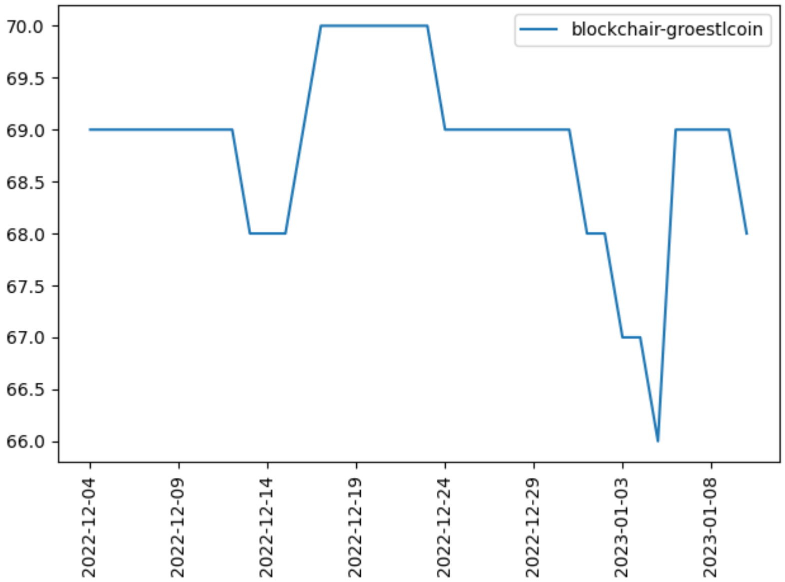

Figure B.7: Activity of Blockchair-Groestlcoin

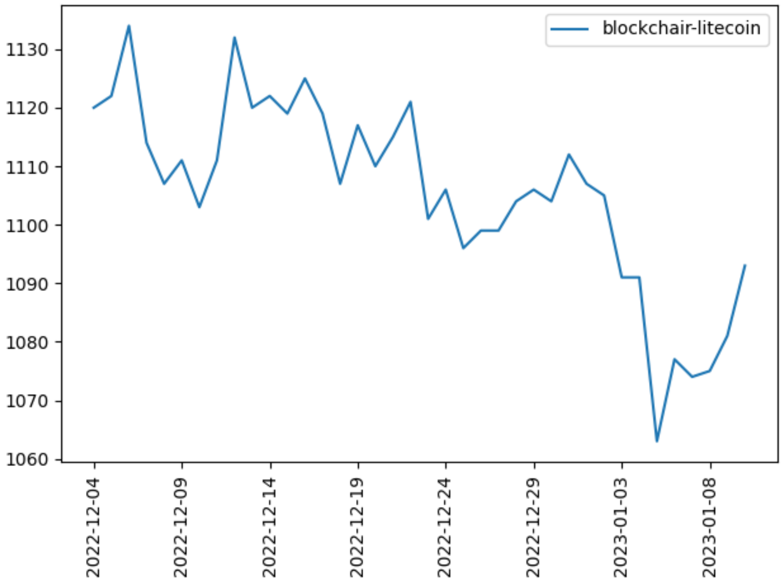

Figure B.8: Activity of Blockchair-Litecoin

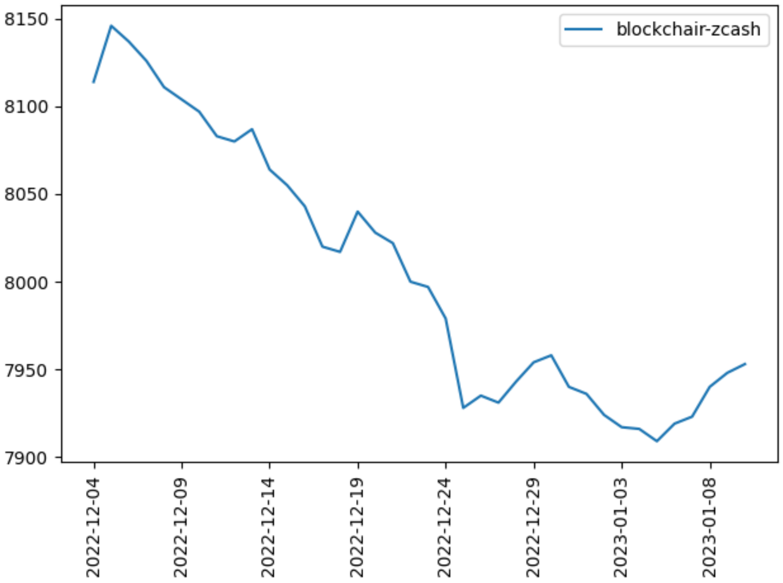

Figure B.9: Activity of Blokchair-Zcash

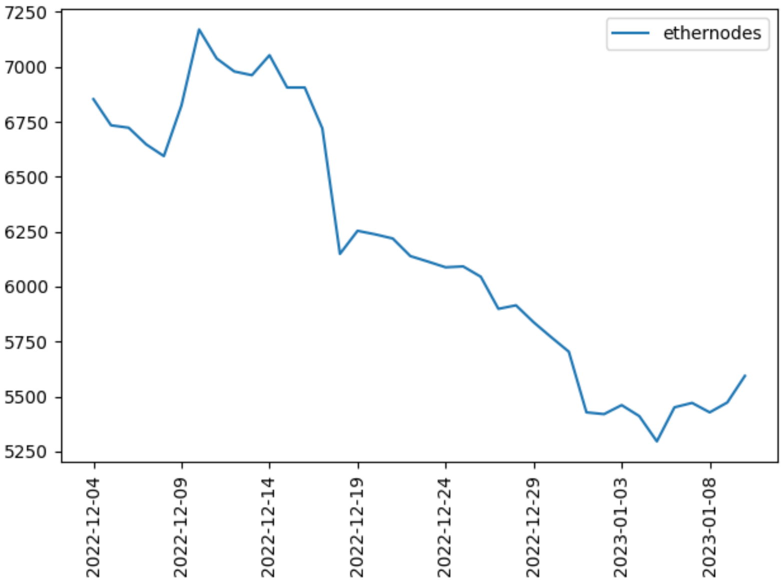

Figure B.10: Activity of Ethernodes

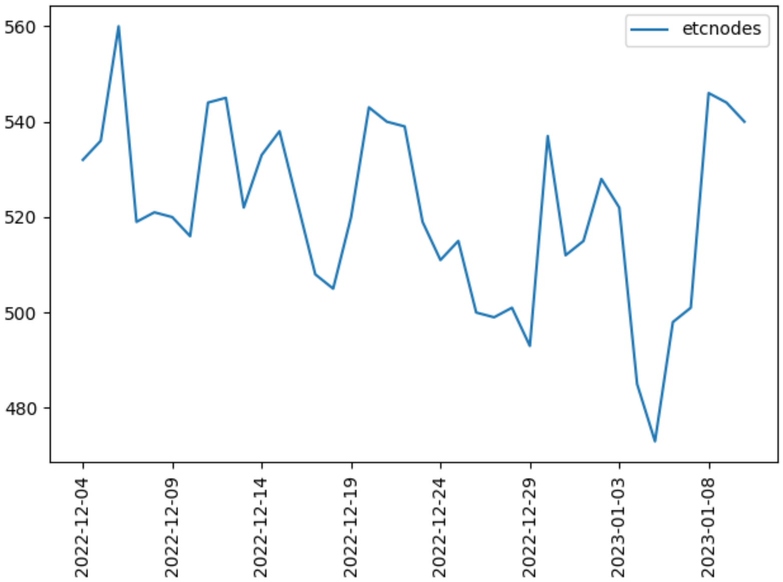

Figure B.11: Activity of Etcnodes

B.2 Total Country Data Barcharts

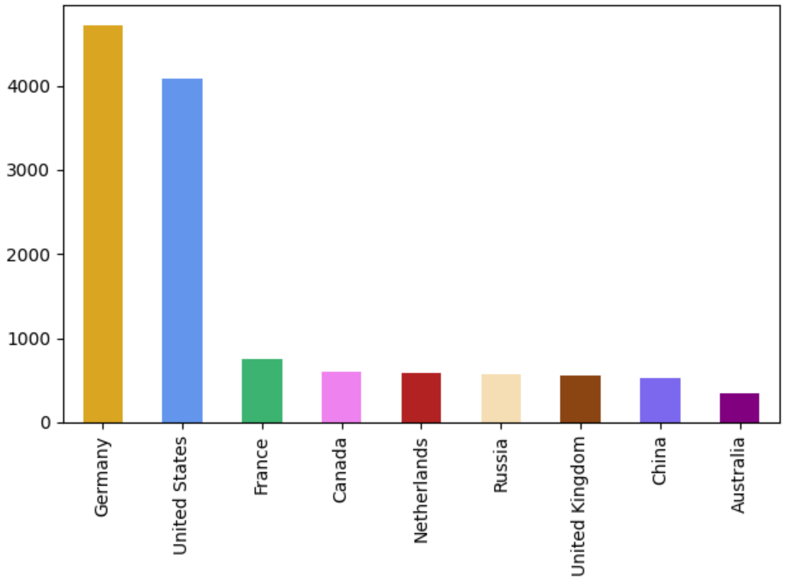

Figure B.12: Bitnodes country data as a barchart

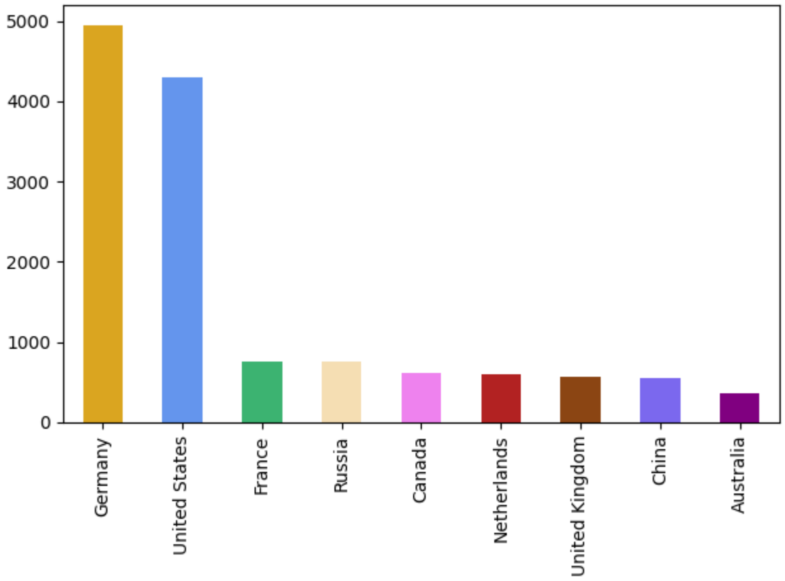

Figure B.13: Blockchair-Bitcoin country data as a barchart

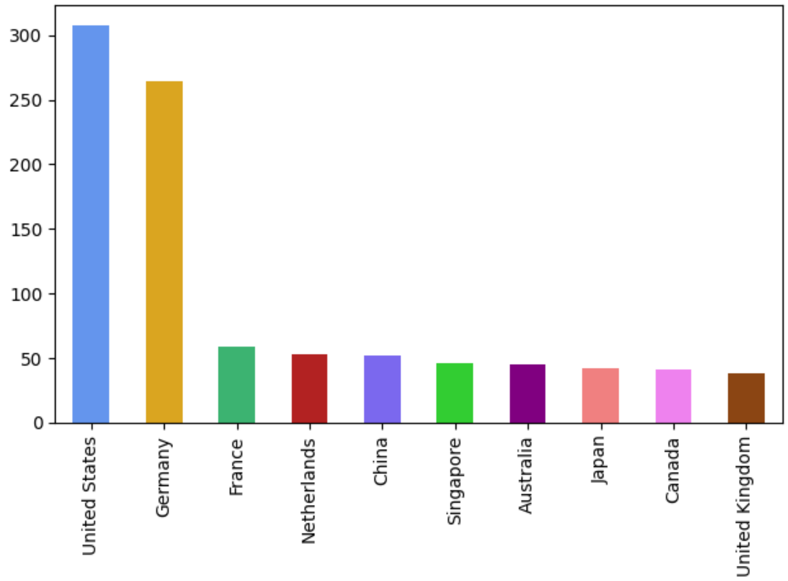

Figure B.14: Blockchair-Bitcoin-Cash country data as a barchart

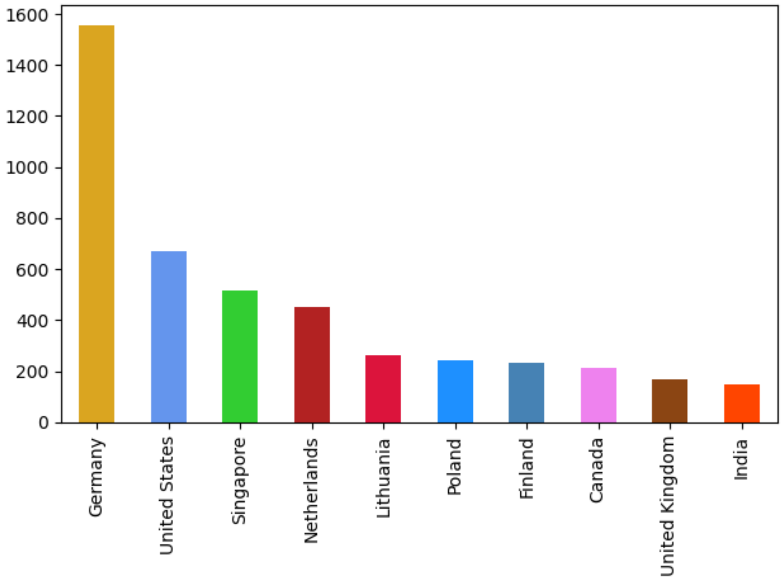

Figure B.15: Blockchair-Dash country data as a barchart

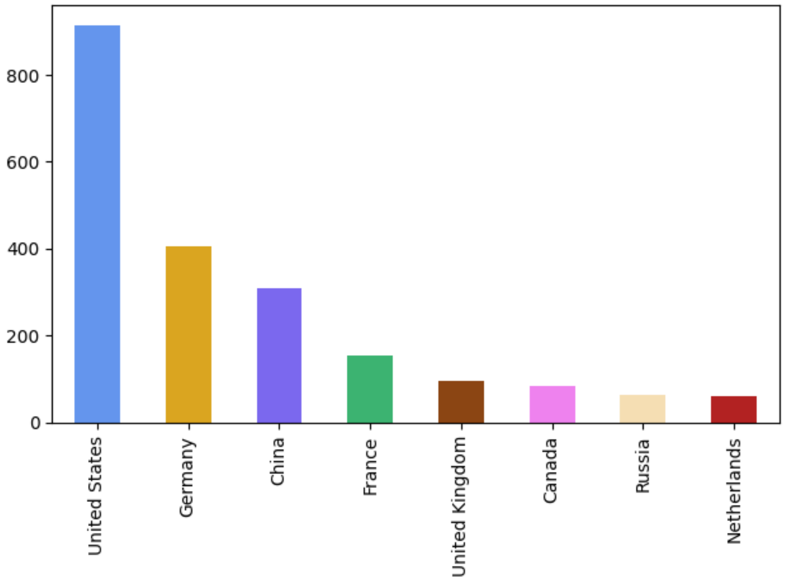

Figure B.16: Blockchair-Dogecoin country data as a barchart

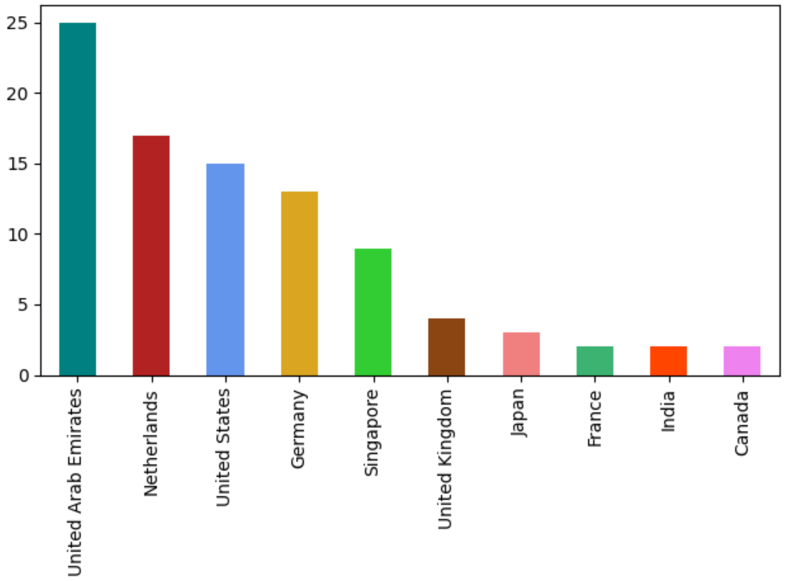

Figure B.17: Blockchair-Groestlcoin country data as a barchart

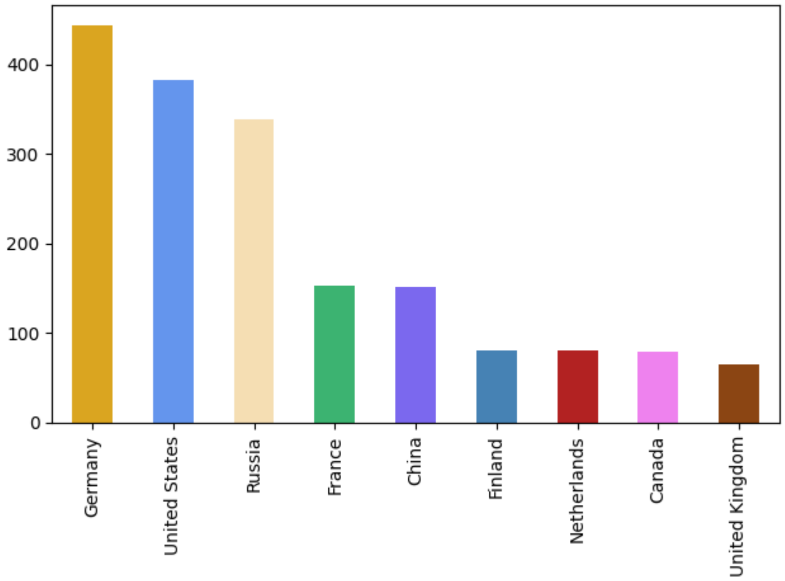

Figure B.18: Blockchair-Litecoin country data as a barchart

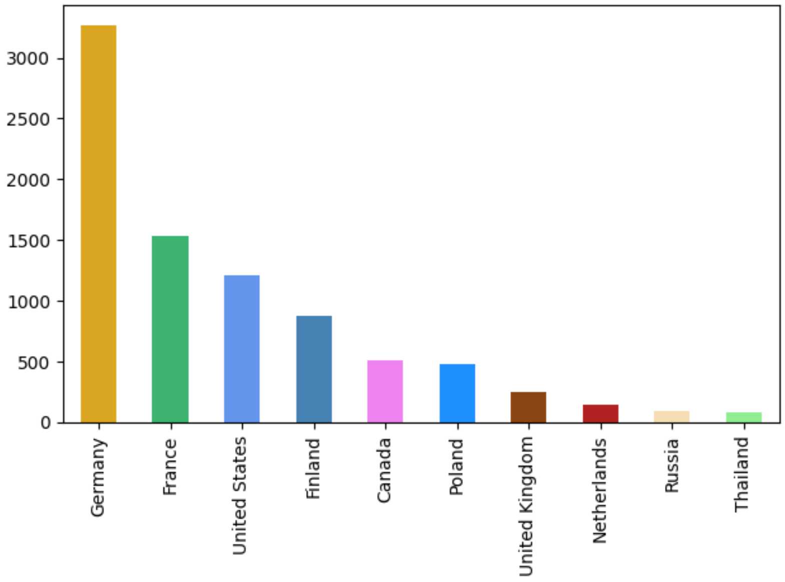

Figure B.19: Blockchair-Zcash country data as a barchart

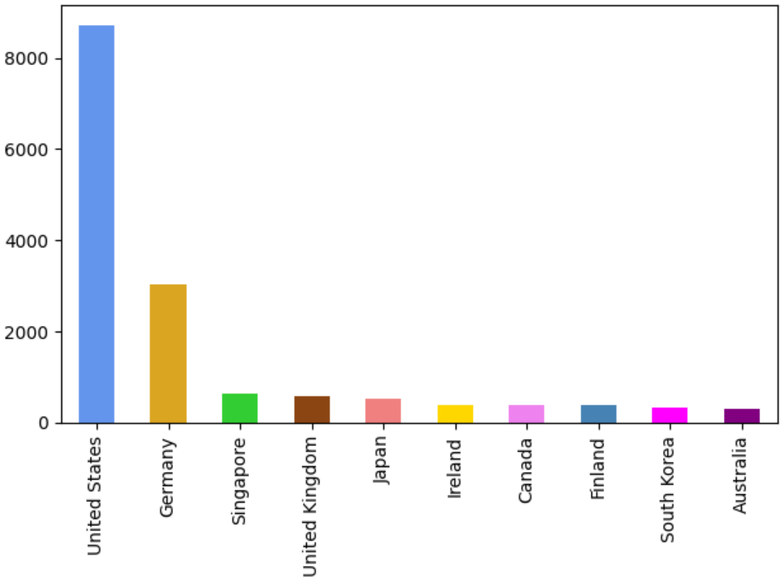

Figure B.20: Ethernodes country data as a barchart

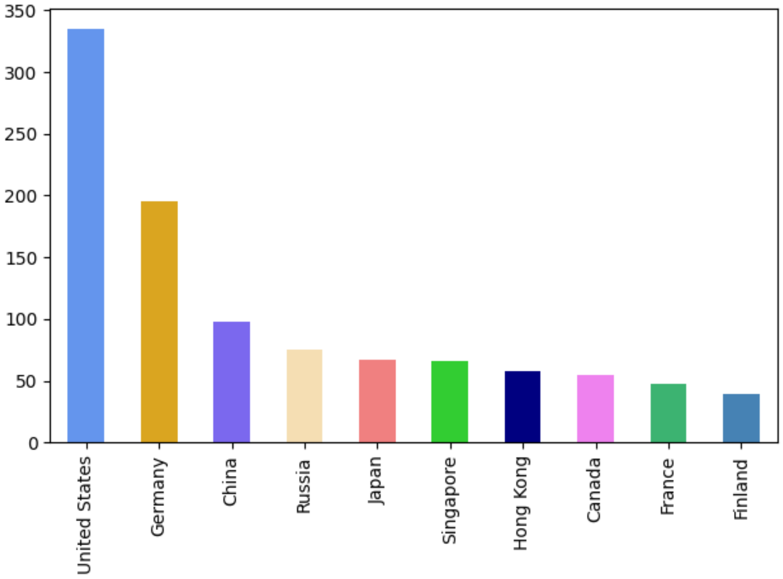

Figure B.21: Etcnodes country data as a barchart

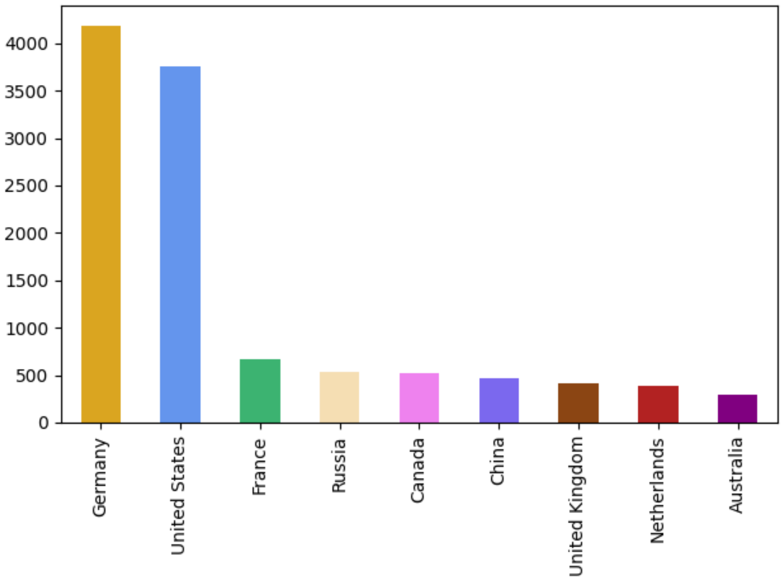

Figure B.22: Etcnodes country data as a barchart

B.3 Total Country Data Piecharts

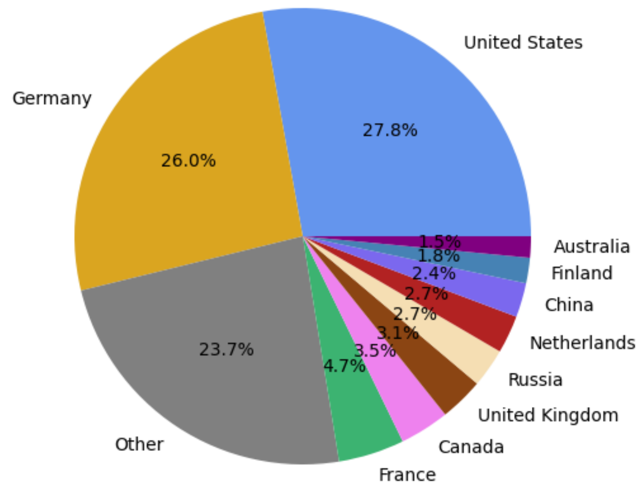

Figure B.23: Country data of all cryptocurrencies combined as a piechart

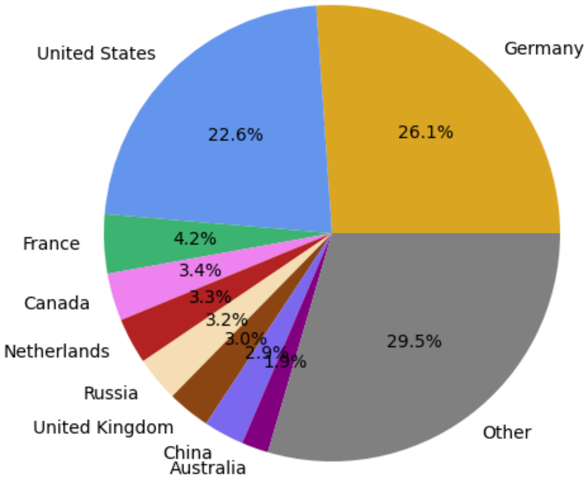

Figure B.24: Country data of Bitnodes as a piechart

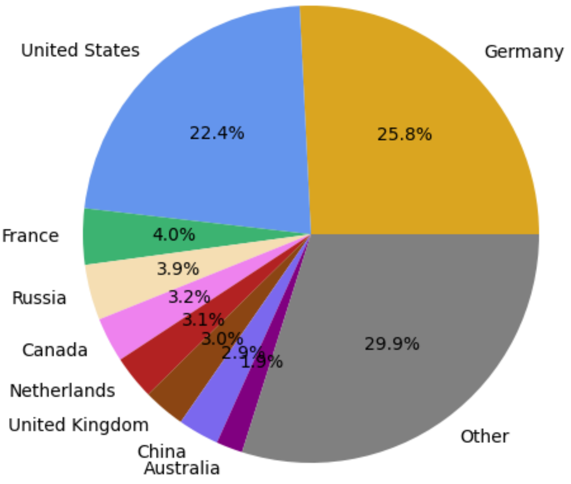

Figure B.25: Country data of Blockchair-Bitcoin as a piechart

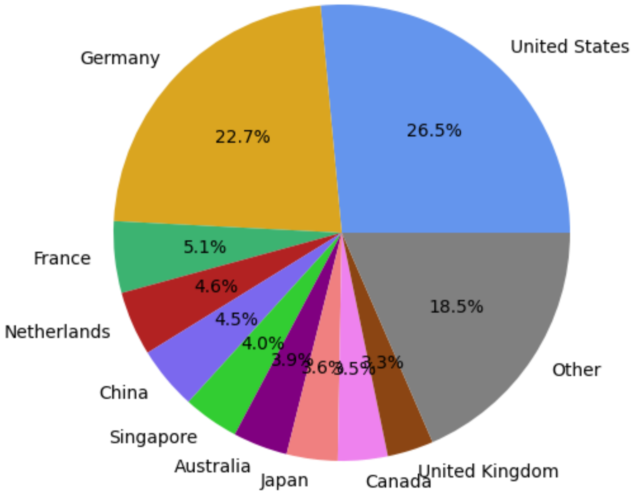

Figure B.26: Country data of Blockchair-Bitcoin-Cash as a piechart

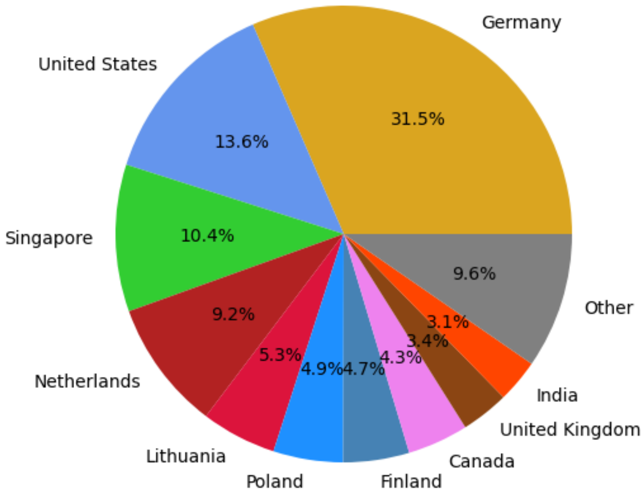

Figure B.27: Country data of Blockchair-Dash as a piechart

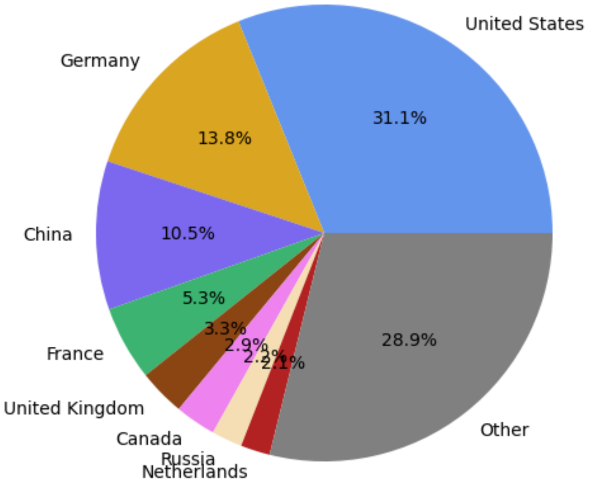

Figure B.28: Country data of Blockchair-Dogecoin as a piechart

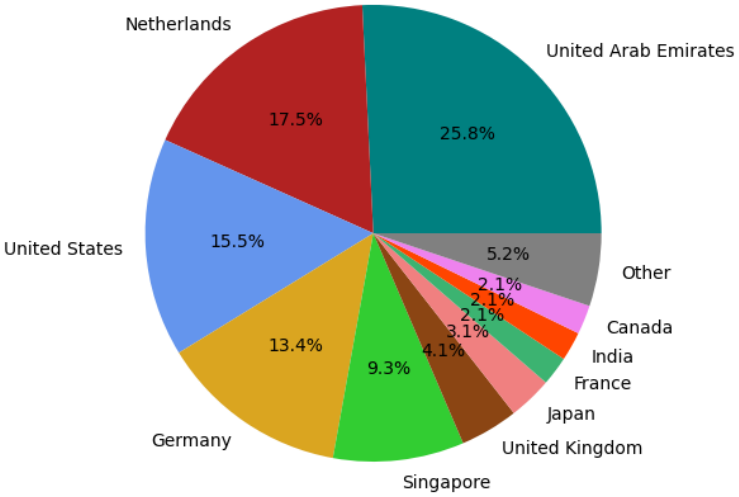

Figure B.29: Country data of Blockchair-Groestlcoin as a piechart

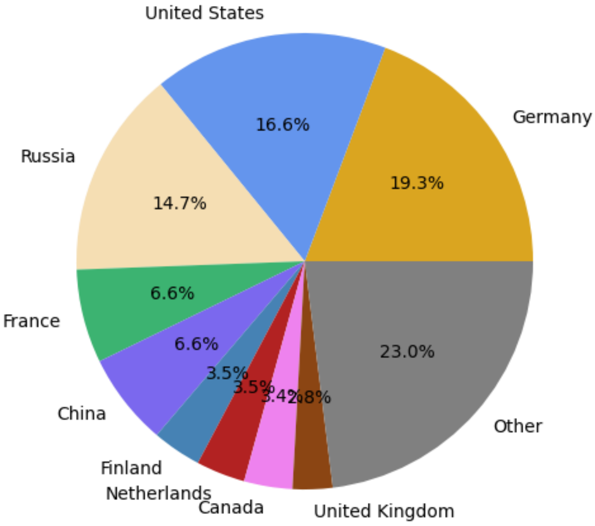

Figure B.30: Country data of Blockchair-Litecoin as a piechart

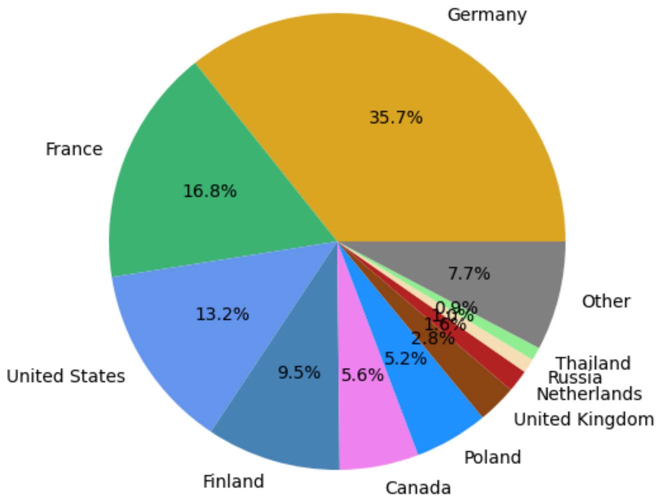

Figure B.31: Country data of Blockchair-Zcash as a piechart

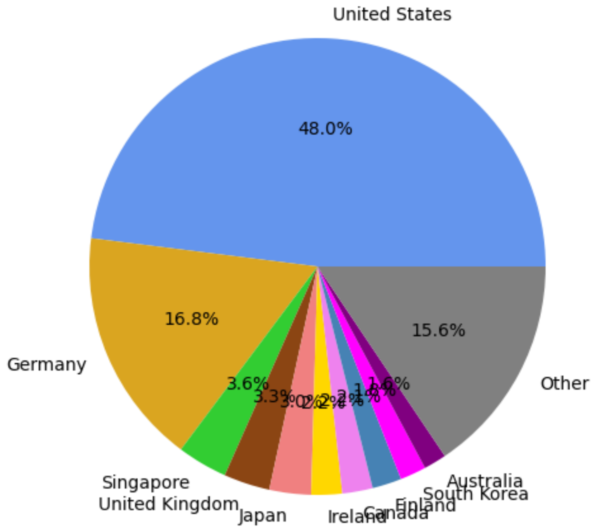

Figure B.32: Country data of Ethernodes as a piechart

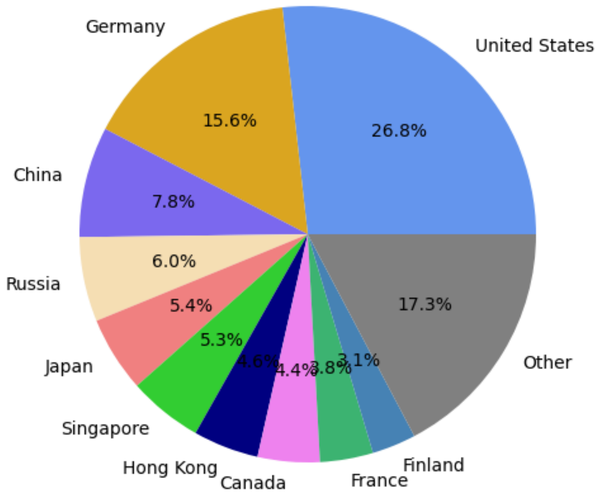

Figure B.33: Country data of Etcnodes as a piechart

B.4 Daily Country Data

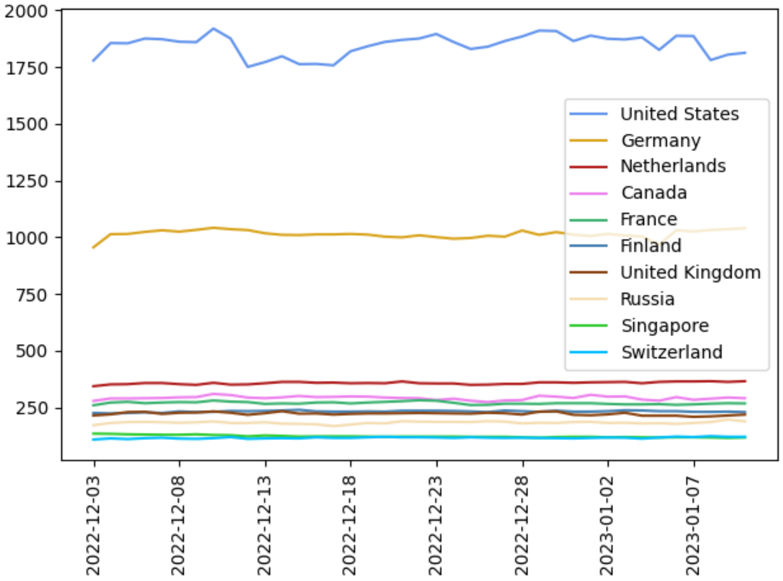

Figure B.34: Daily Bitnodes country data

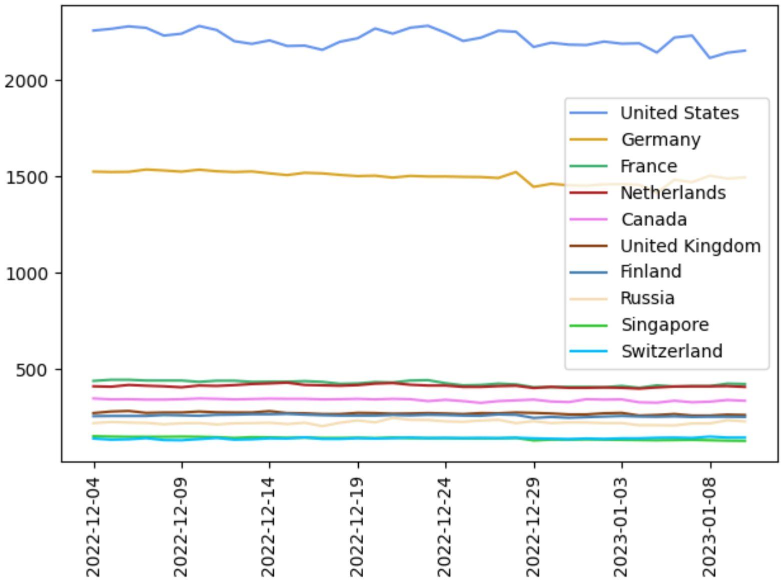

Figure B.35: Daily Blockchair-Bitcoin country data

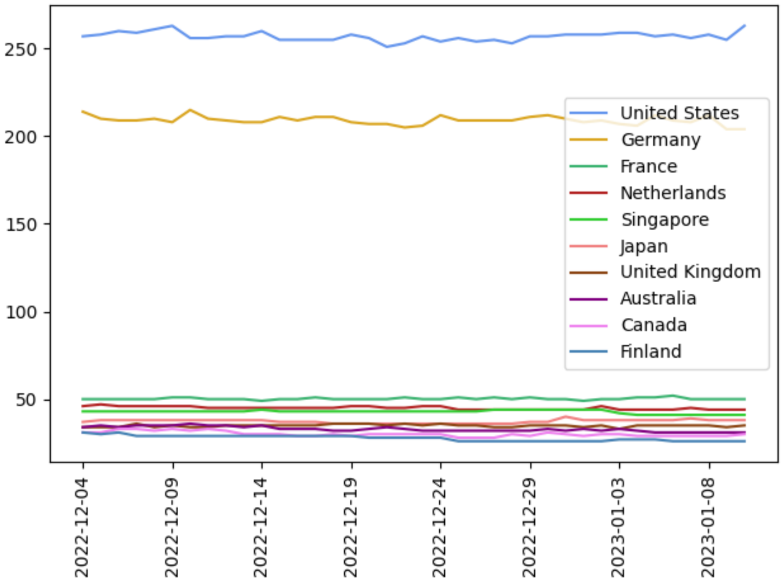

Figure B.36: Daily Blockchair-Bitcoin-Cash country data

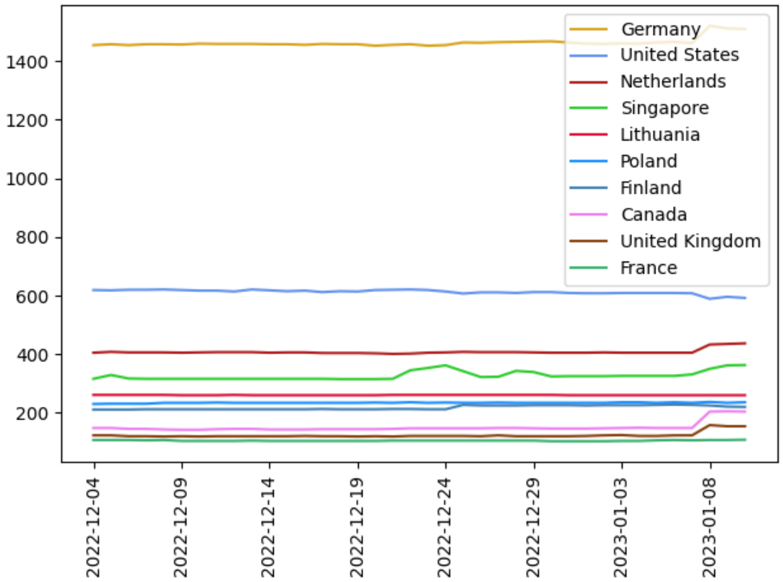

Figure B.37: Daily Blockchair-Dash country datat

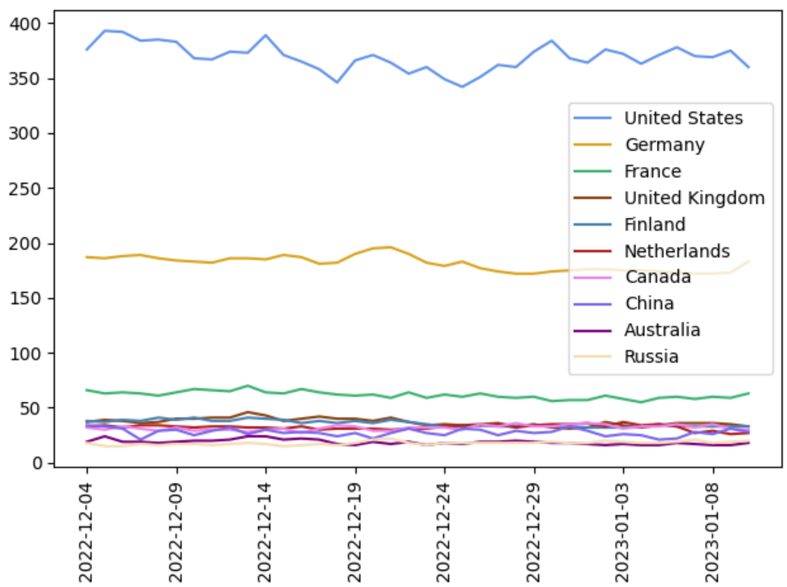

Figure B.38: Daily Blockchair-Dogecoin country data

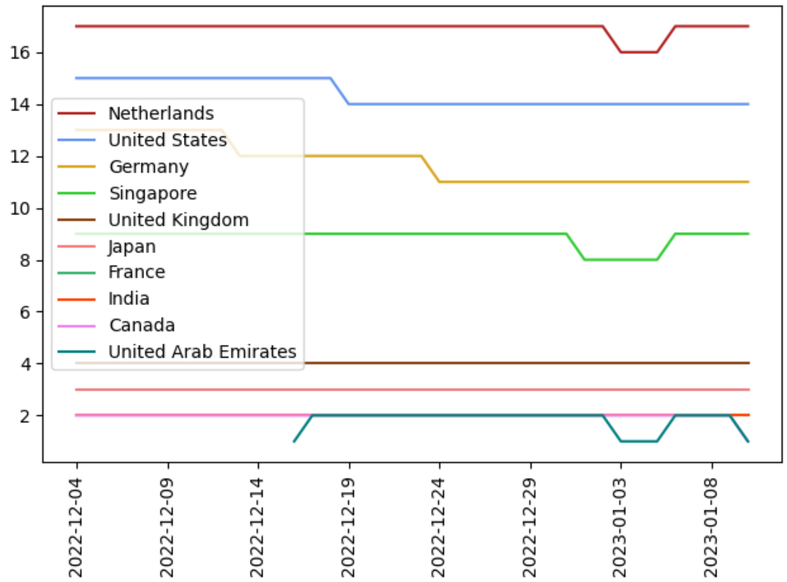

Figure B.39: Daily Blockchair-Groestlcoin country data

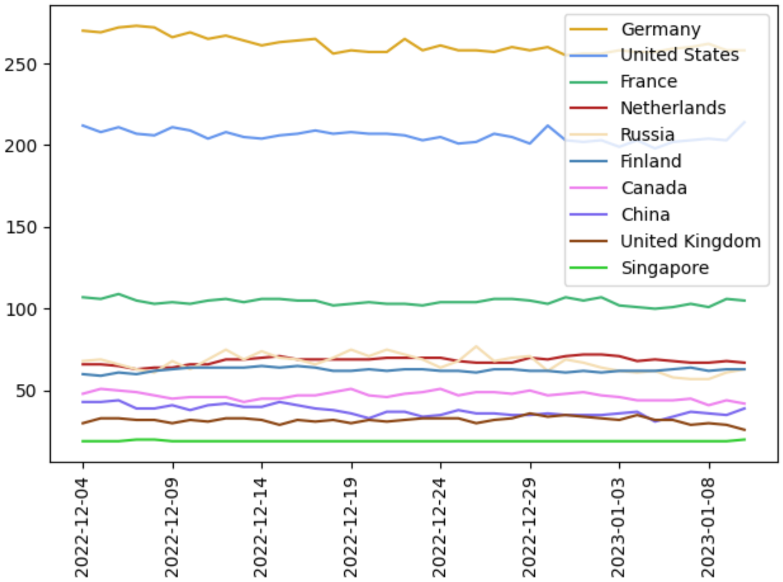

Figure B.40: Daily Blockchair-Litecoin country data

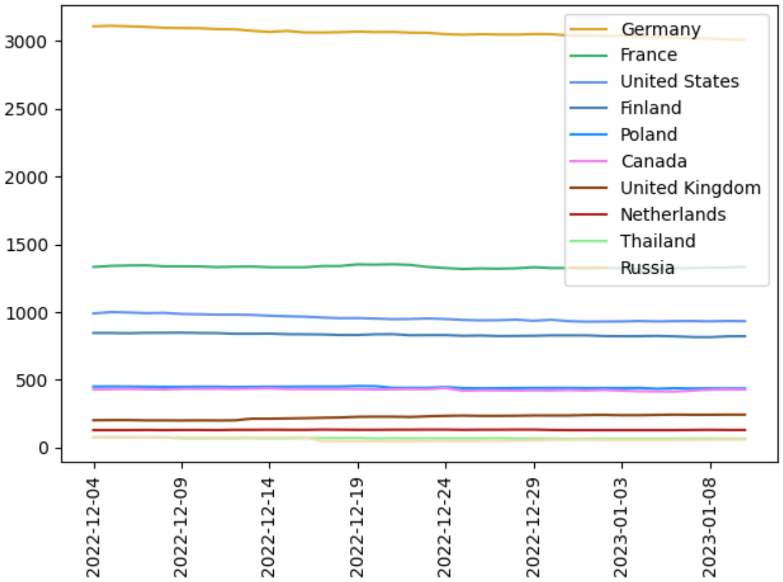

Figure B.41: Daily Blockchair-Zcash country data

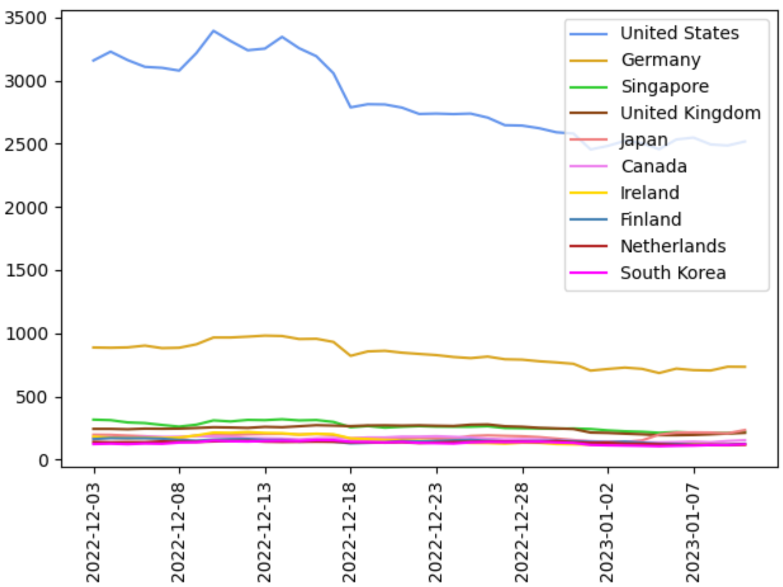

Figure B.42: Daily Ethernodes country data

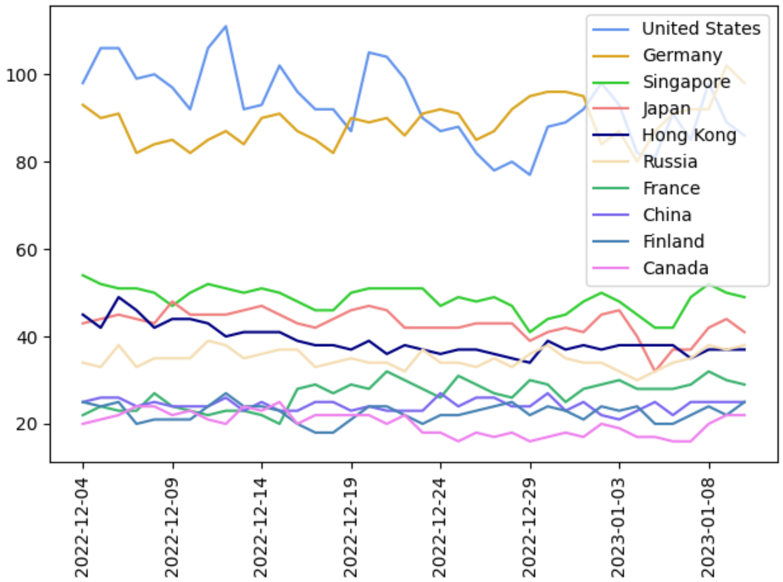

Figure B.43: Daily Etcnodes country data

B.5 Duration of Node Activity

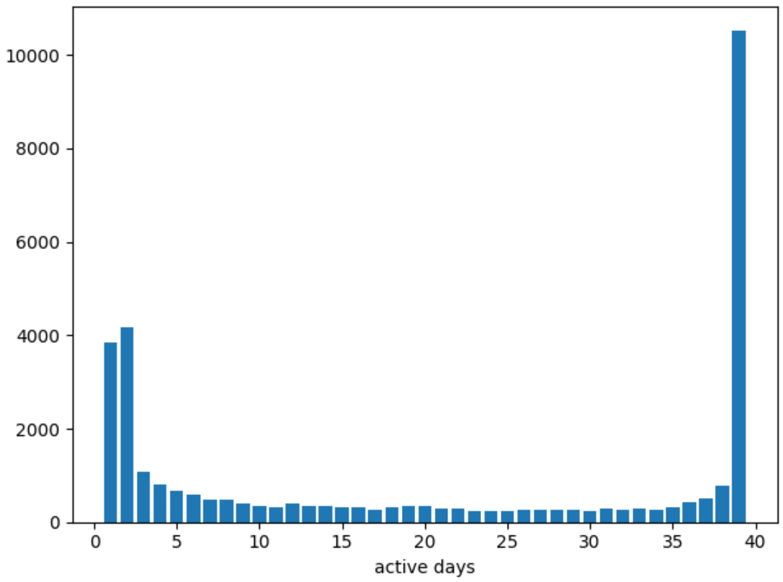

Figure B.44: Activity of Bitnodes

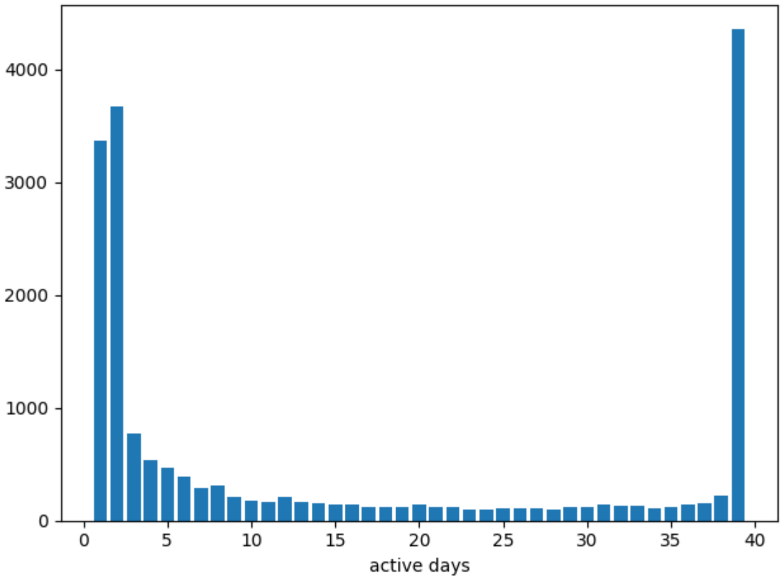

Figure B.45: Activity of Bitnodes without TOR nodes

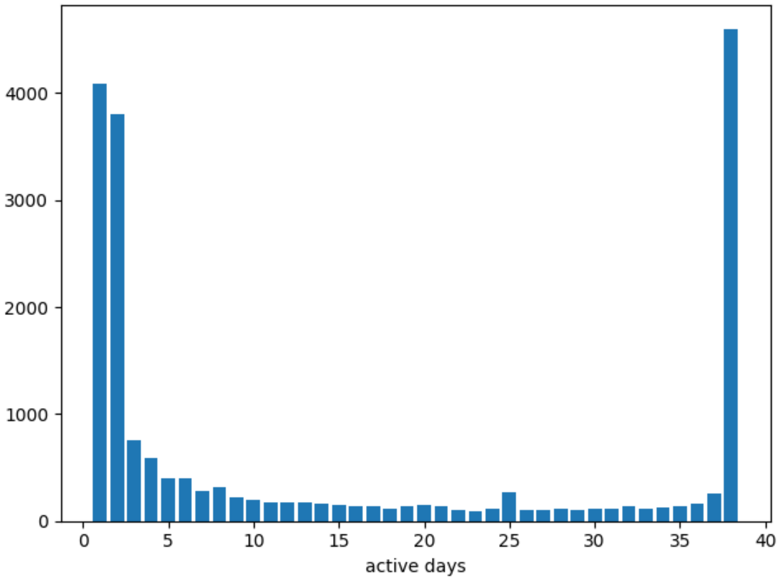

Figure B.46: Activity of Blockchair-Bitcoin

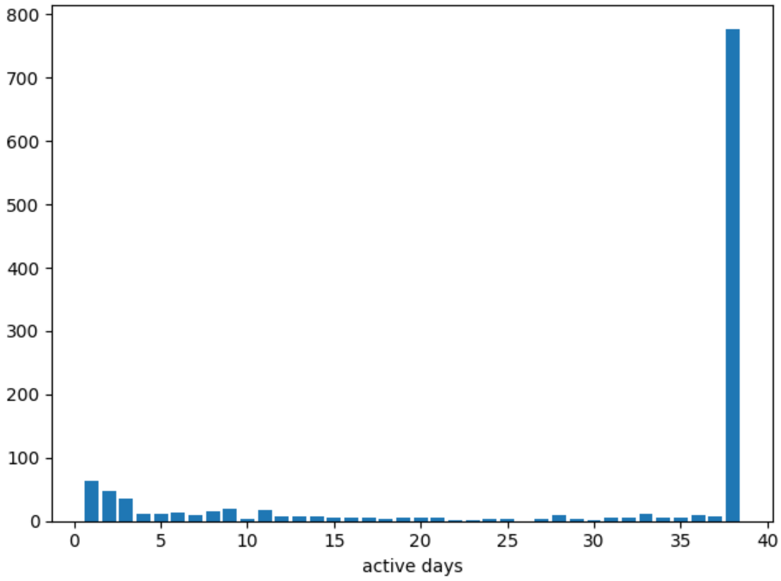

Figure B.47: Activity of Blockchair-Bitcoin-Cash

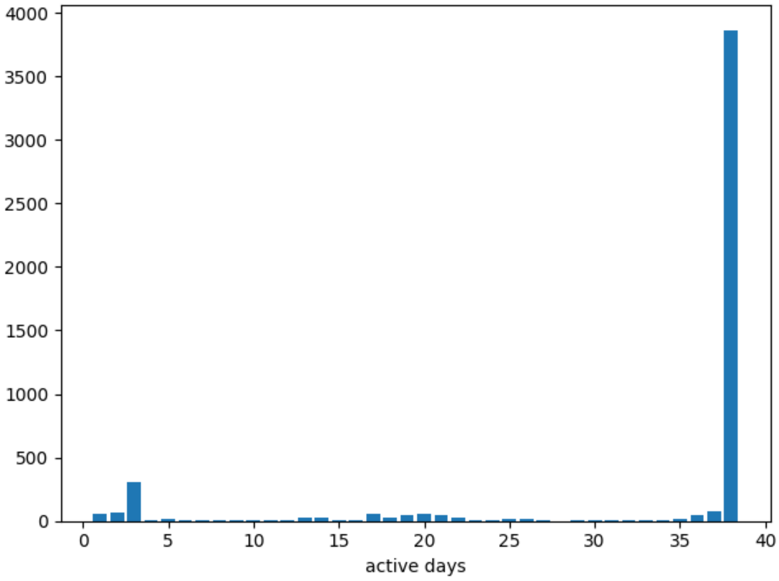

Figure B.48: Activity of Blockchair-Dash

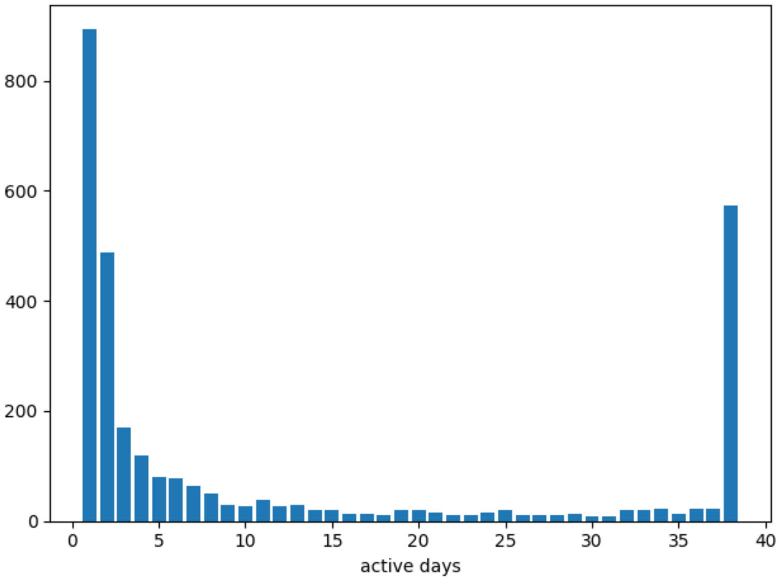

Figure B.49: Activity of Blockchair-Dogecoin

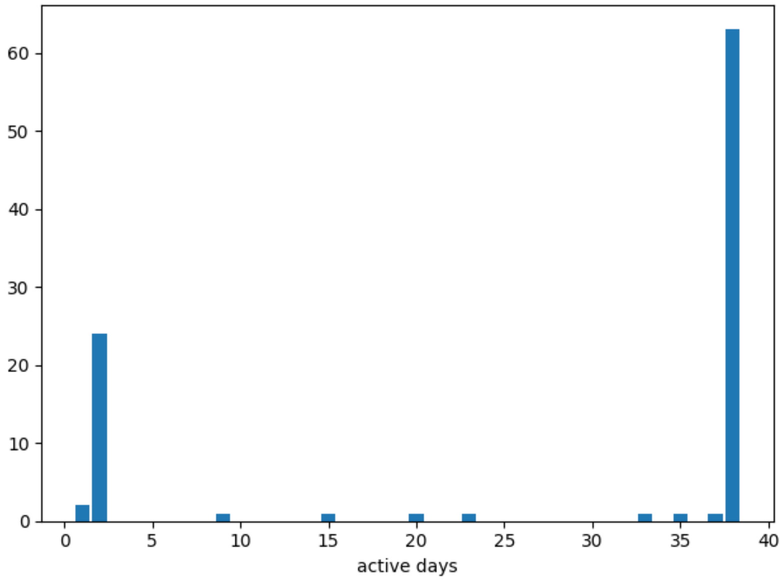

Figure B.50: Activity of Blockchair-Groestlcoin

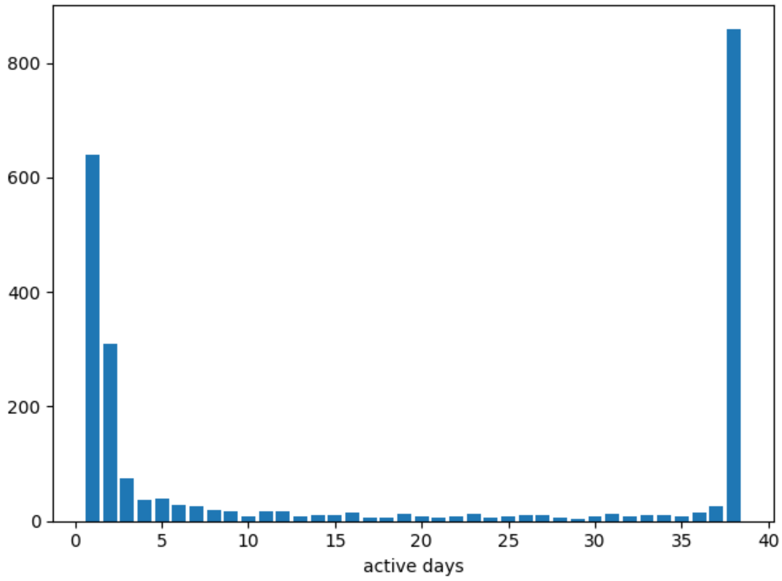

Figure B.51: Activity of Blockchair-Litecoin

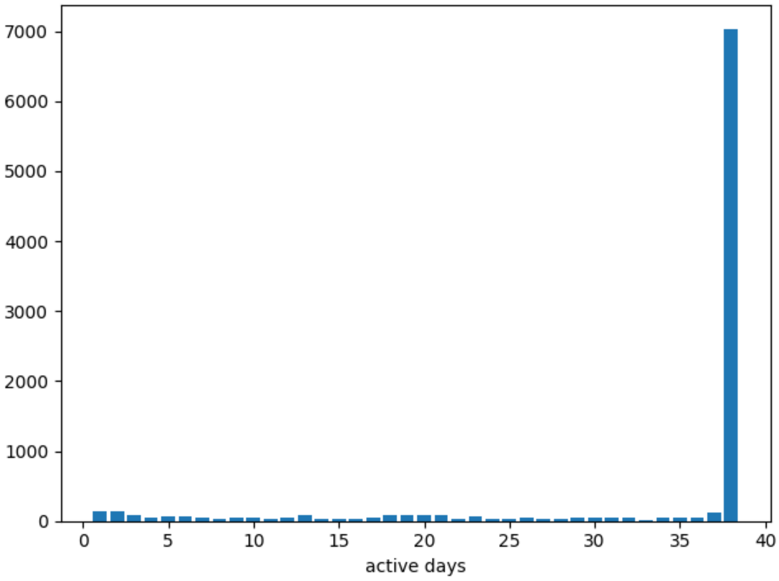

Figure B.52: Activity of Blockchair-Zcash

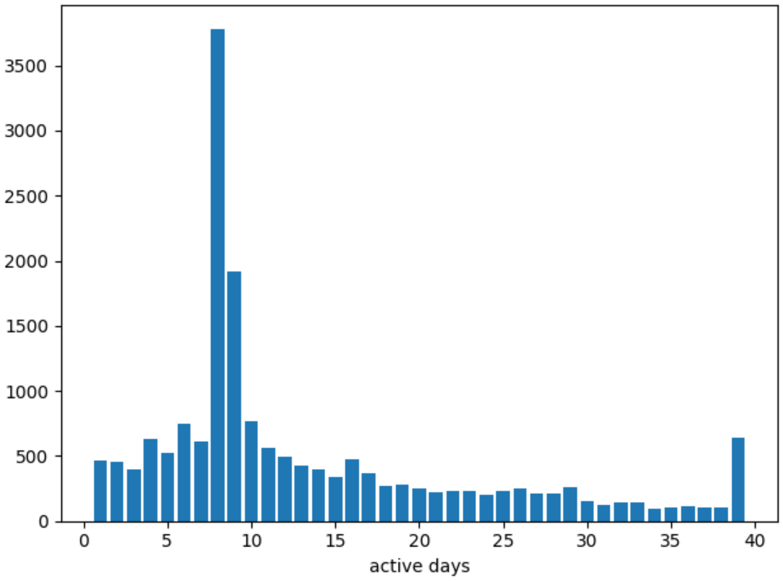

Figure B.53: Activity of Ethernodes

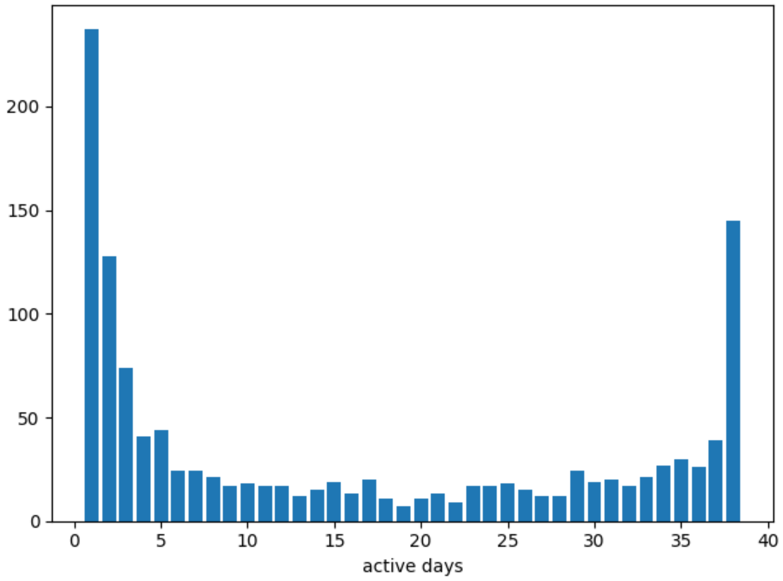

Figure B.54: Activity of Etcnodes

B.6 IPs Appearing in Multiple Cryptocurrency Networks

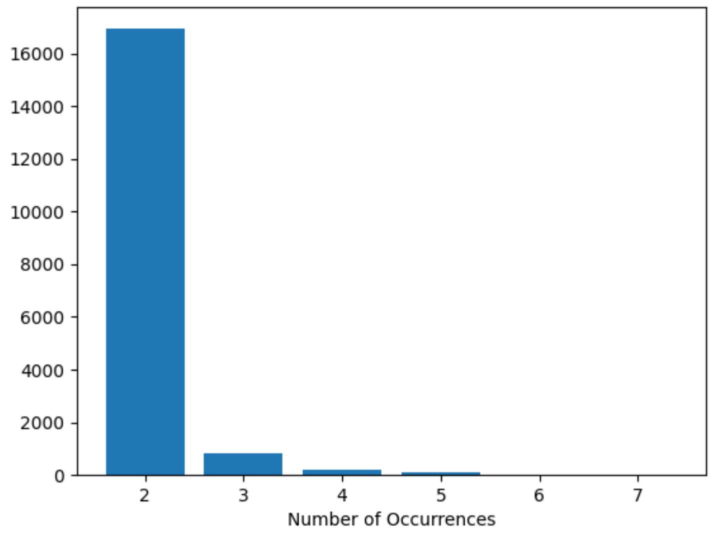

Figure B.55: Count of number of IP occurrences in different node explorers

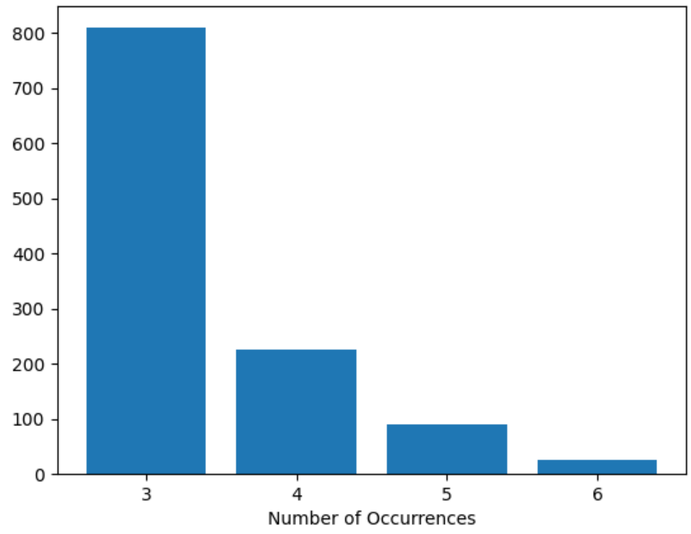

Figure B.56: Count of number of IP occurrences in different node explorers starting from 3 occurences

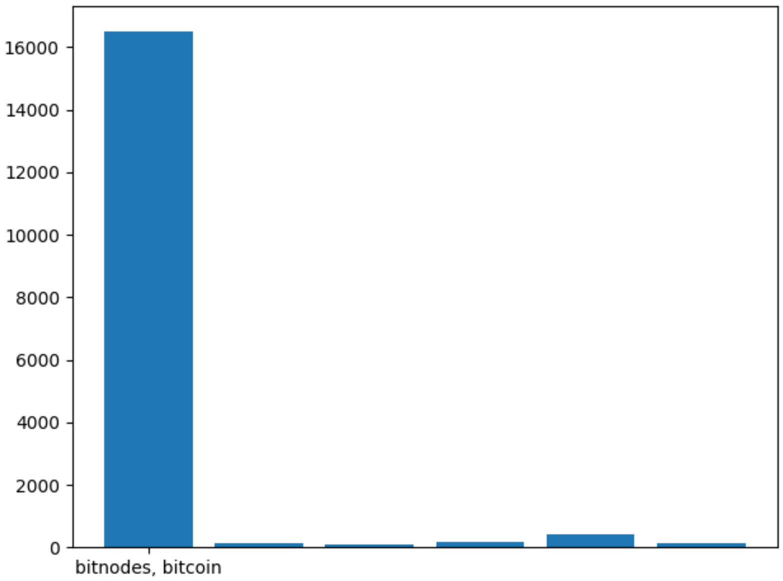

Figure B.57: Count of IP occurrences for specific combinations

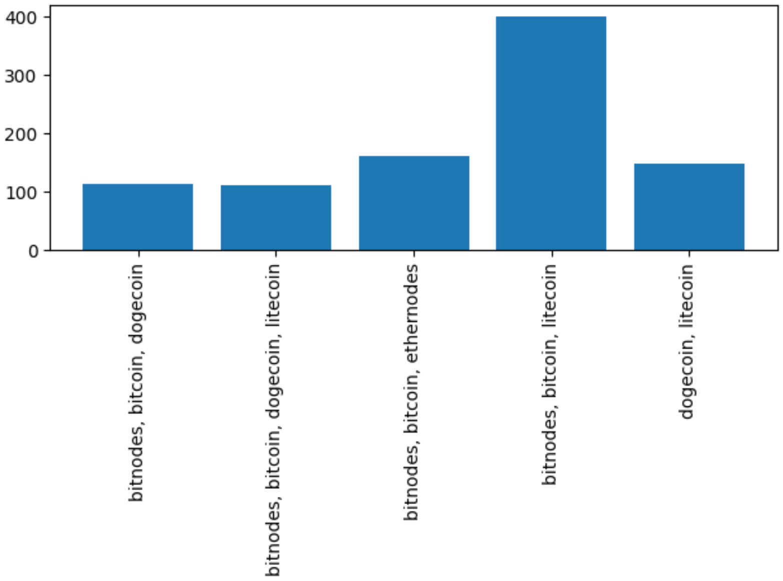

Figure B.58: Count of IP occurrences for specific combinations, excluding 'bitnodes,bitcoin'

B.7 Heatmaps

These are the Heatmaps created with the help of folium. As these are interactive HTML files please go to the Gitlab repository to view them in detail.

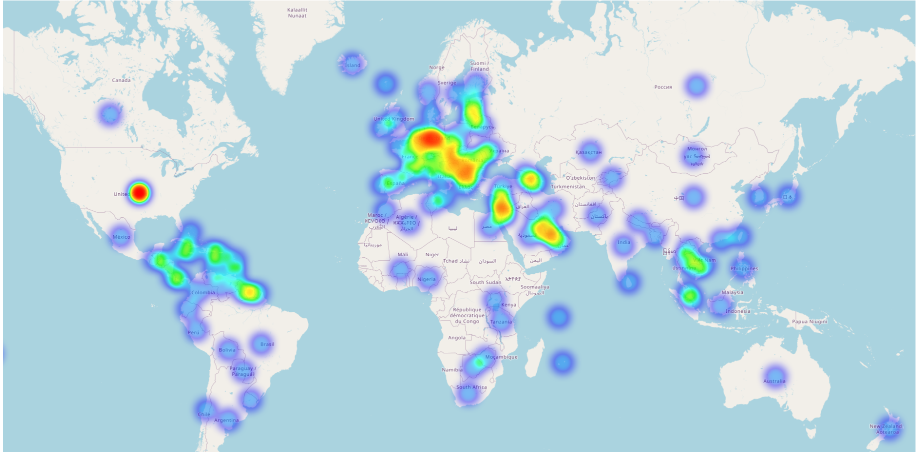

Figure B.59: Heatmap of all data combined

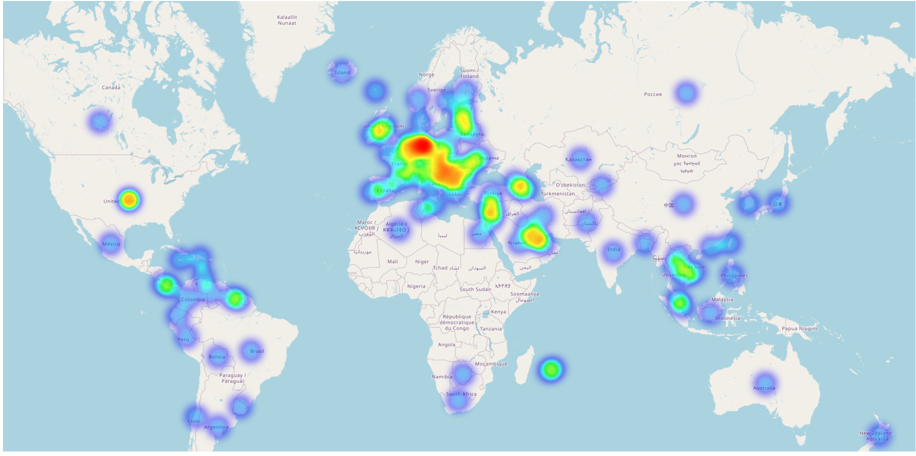

Figure B.60: Heatmap of Bitnodes

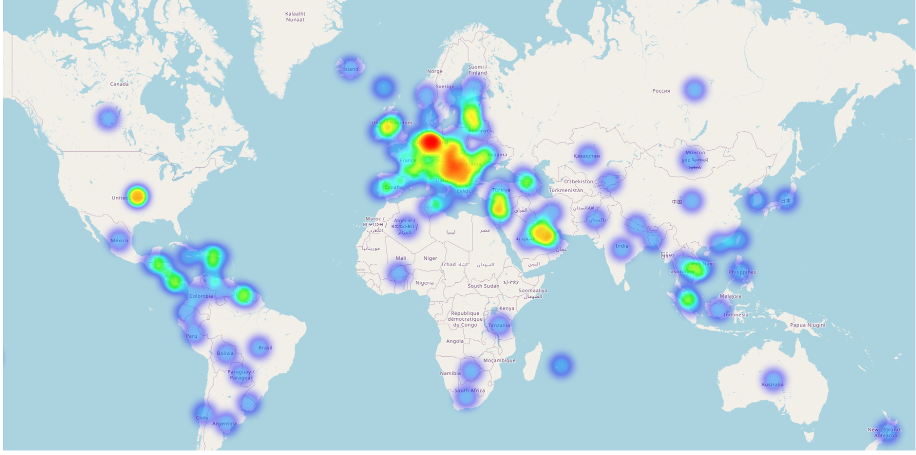

Figure B.61: Heatmap of Blockchair-Bitcoin

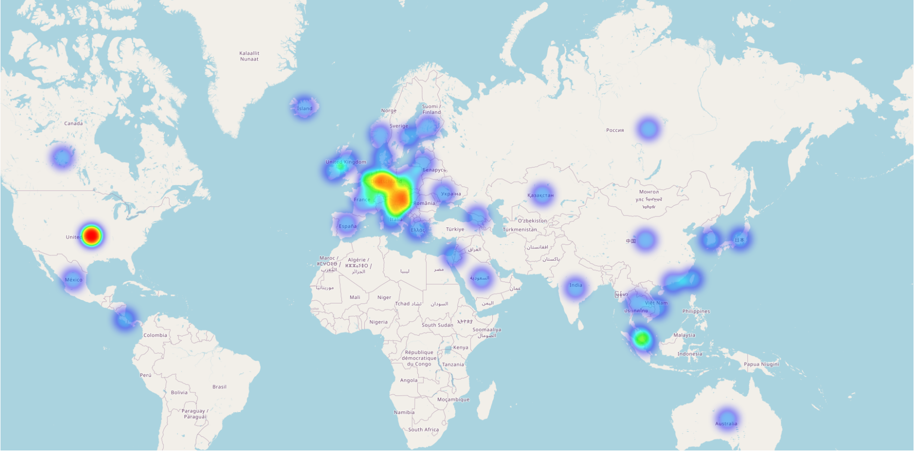

Figure B.62: Heatmap of Blockchair-Bitcoin-Cash

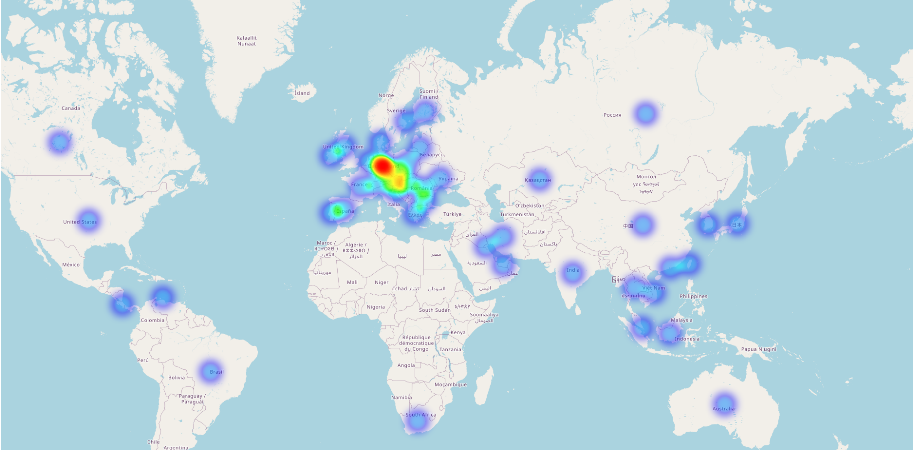

Figure B.63: Heatmap of Blockchair-Dash

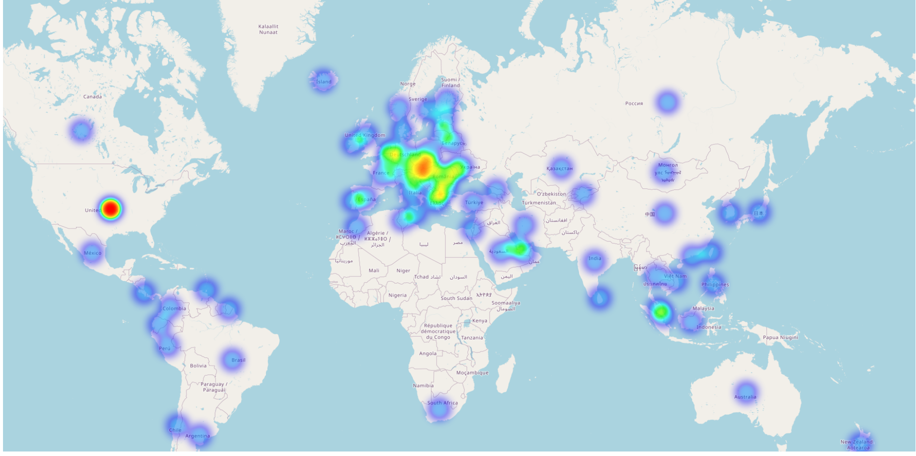

Figure B.64: Heatmap of Blockchair-Dogecoin

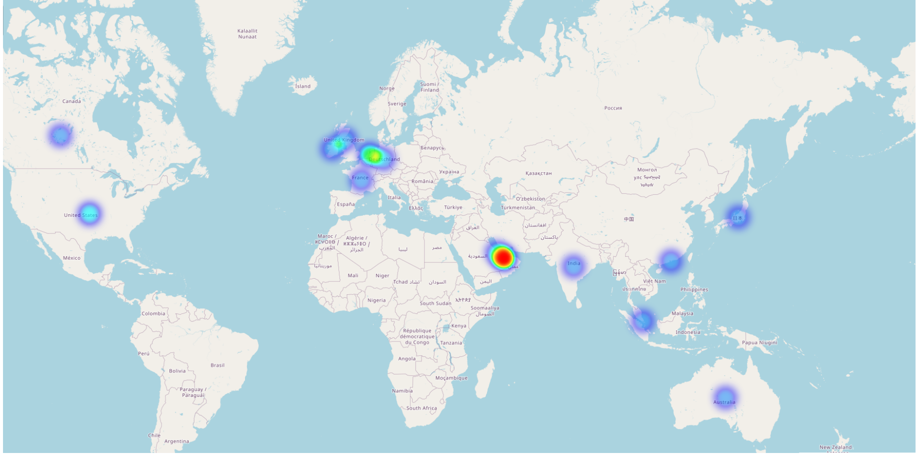

Figure B.65: Heatmap of Blockchair-Groestlcoin

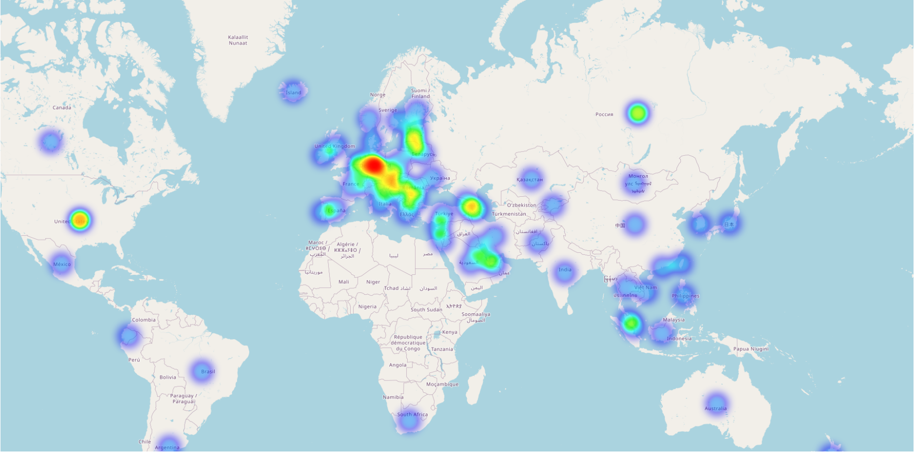

Figure B.66: Heatmap of Blockchair-Litecoin

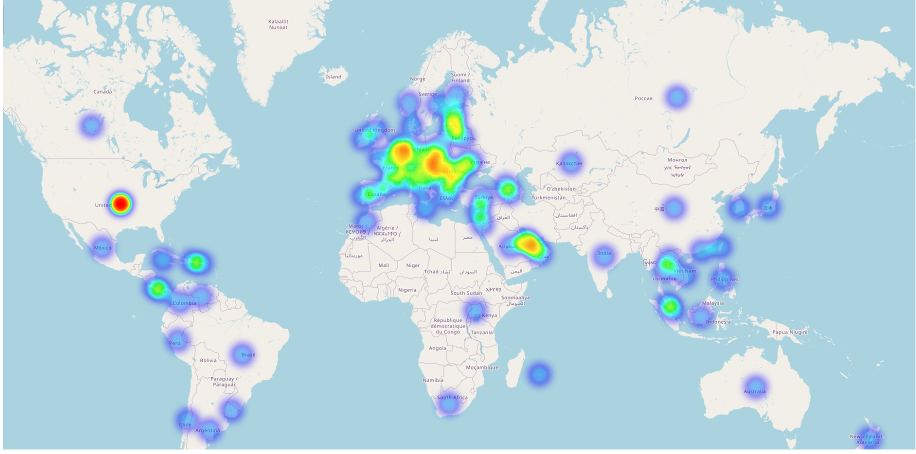

Figure B.67: Heatmap of Ethernodes

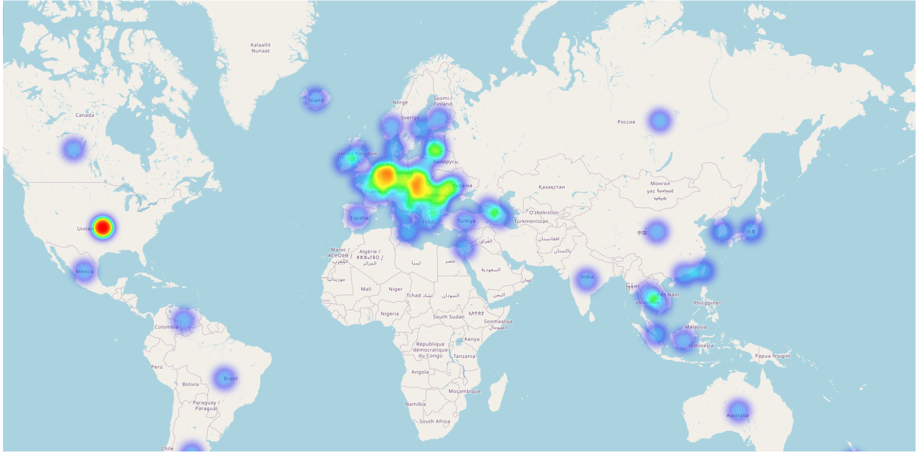

Figure B.68: Heatmap of Etcnodes