Analyzing the Bitcoin Transition Network

Khushal Ramani
21110104
IIT Gandhinagar

Hirva Patel
21110154
IIT Gandhinagar

Harshit Kothari
21110070
IIT Gandhinagar

ABSTRACT

The Bitcoin transaction network operates within a decentralized system, where transactions are recorded on the blockchain ledger. This ecosystem is rich with intricate financial interactions. Our study takes a deep dive into understanding this network, focusing on key elements like its structure, identifying communities, and using spectral clustering, and identifying the most endorsed nodes.

ACM Reference Format:
Khushal Ramani, Hirva Patel, and Harshit Kothari. 2018. Analyzing the Bitcoin Transition Network. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Bitcoin, first introduced in 2009, is a decentralised payment system and electronic cryptocurrency. It has steadily grown to a market cap of around $4 trillion USD and sees over 150k transactions a day. Bitcoin operates on a peer-to-peer basis, allowing users to transact directly without the need for intermediaries. Transactions are signed using cryptographic techniques and recorded on a public, distributed ledger called a Blockchain. Working of the Blockchain can be understood using five steps:

- Transaction: The two companies agree to trade a unit of value and begin the transaction.
- Block: The transaction is grouped with other pending transactions, forming a "block," which is then transmitted to the network of participating computers in the blockchain system.
- Verification: The participating computers assess the transactions and use mathematical calculations to authenticate their validity. If 51% of participating computers authenticate the transactions,they are deemed verified.
- Hash: Each verified block of transactions receives a time-stamped cryptographic hash and references the previous block's hash, forming an immutable chain of records.
- Execution: The value is transferred from party A's account to party B's.

Blockchain is a decentralised technology, meaning it is not controlled by any one organization. Blockchain is designed to make it extremely difficult to hack the system or forge the data stored on it, thereby making it secure and immutable. Each computer in a blockchain network has a copy of the ledger to prevent single points of failure.

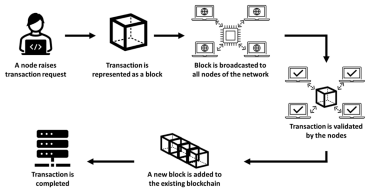

Figure 1: Bitcoin Working

2 OBJECTIVES

For our dataset we have done an in-depth analysis of the transaction graph with the specific focus on statistical properties of the network like the number of nodes, edges in-nodes, out-nodes, diameter, largest strongly connected component analysis, etc. In addition to these statistics, we have also directed our attention towards various analytical perspectives to interpret and comprehend the graph, such as:

- Community Detection: A community is a subset of nodes that are densely connected to each other and loosely connected to the nodes in the other communities in the same graph. We have detected communities using different approaches like Louvain Algorithm and Girvan-Neumann Algorithm. Different communities within the Bitcoin graph can signify various aspects of the ecosystem. For example, they may represent different types of users or entities, such as miners, investors, or exchanges.
- Spectral Clustering: Spectral clustering is a clustering technique that uses the eigenvectors of a similarity matrix to partition data points into clusters. Spectral clustering in the Bitcoin graph can provide valuable insights into the network's resilience and robustness against attacks or disruptions.
- Endorsed Nodes: We can explore how trust propagates through the network by examining which nodes endorse other highly endorsed nodes forming a recursive network of trust relationships.

For detailed work, please visit github repository: https://github.com/hirva-p/Bitcoin_Network_Analysis.

3 BASIC ANALYSIS

The Bitcoin Alpha dataset exhibits notable statistical properties, including power-law distributions in node degrees, as well as clustering coefficients indicative of network modularity. These characteristics underscore the complex nature of the Bitcoin transaction network. Here are some statistical properties of the graph:

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

Total Nodes	3783
Total Edges	24186
Total number of in-nodes	3754
Total number of out-nodes	3286
Total number of triangles	22153
Number of self-loops	0
Maximum Diameter	10
Minimum Diameter	1

Figure 2: Properties of the Network

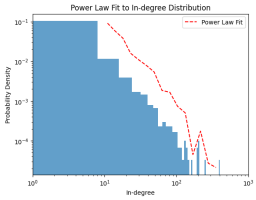

Figure 3: In-degree Power Law fit

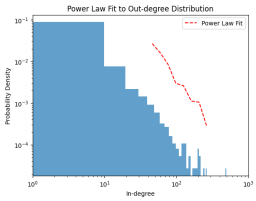

Figure 4: Out-degree Power Law fit

Figure 5: Power Law fit over histogram

The power law distribution, often observed in graphs, denotes a statistical phenomenon where the probability distribution of a variable follows a power law relationship. Specifically, in the context of graph theory, the power law states that a small number of nodes will posses a disproportionately large number of connections, while the majority of nodes have relatively few connections. Mathematically, the power law for in-degree/out-degree distribution can be expressed as follows: the probability that a node has in-degree/out-degree i is proportional to $\frac{1}{i^x}$, where x is a parameter greater than 1. This implies that nodes with higher in-degrees are less common, and the frequency of nodes with lower in-degrees decreases at a slower rate. In our dataset, we've found that the probability of a node having a certain number of connections follows a specific formula, with x values of approximately 2.28 for in-degrees and 2.78 for out-degrees, resembling how some people have lots of friends on social media while others have only a few.

Similarly, in the Bitcoin Alpha network, the out-degree distributions follow a power law pattern with an exponent of 2.72. However, it's noteworthy that at higher values of out-degree, the distribution deviates noticeably from the power law trend. This deviation suggests that nodes with a high out-degree may follow a different distribution, possibly a combination of Poisson or a mix of Poisson and power law distributions. Further investigation is required to gain a deeper understanding of this phenomenon.

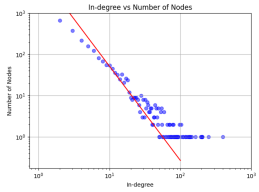

Figure 6: Indegree and Power Law

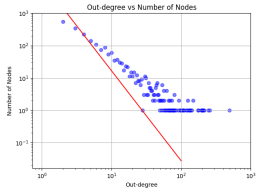

Figure 7: Out degree and Power Law

Figure 8: A figure with two subfigures

In a study similar to one by Albert, Jeong, and Barabasi, who found that most web pages are connected by just a few clicks, our research in the Bitcoin network shows a similar trend. They suggested that as the network grows, this number of connections also increases, and hence, average distance between any two randomly chose nodes showed a downward trend.

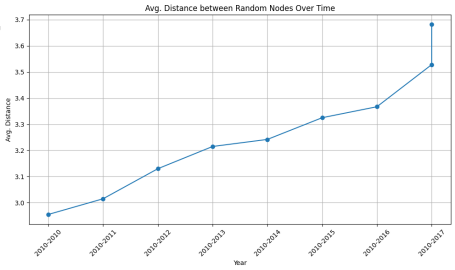

Figure 9: Avg. Distance between nodes v/s Time

Over time, the fraction of connected nodes in the Bitcoin Alpha network has increased steadily. As Bitcoin gained popularity and usage expanded, more nodes joined the network, contributing to its overall connectivity. This growth in the number of connected nodes reflects the network's increasing robustness and resilience. Additionally, technological advancements and improvements in network infrastructure may have facilitated greater node connectivity, enabling smoother and more efficient communication between participants.

Indeed, it is noteworthy that unlike typical web networks, a large portion, approximately 75%, of the Bitcoin Alpha network was already connected in its initial years. This phenomenon may be attributed to several factors. Firstly, the relatively small size of the Bitcoin Alpha dataset compared to the vastness of the entire Bitcoin network could contribute to a higher initial connectivity percentage. Additionally, the robust and decentralized nature of the Bitcoin network itself likely plays a significant role. Unlike centralized web

networks, the Bitcoin network is inherently designed to be highly interconnected, with transactions propagating and validating across the entire network.

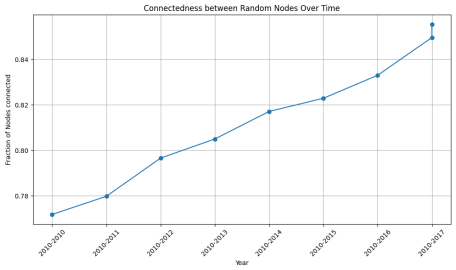

Figure 10: Fraction of nodes connected vs Time

In contrast to the web, the Bitcoin Alpha network exhibits a distinct characteristic: a significantly larger strongly connected component encompassing 99% of its nodes. This means that the vast majority of nodes in the Bitcoin network are tightly interconnected, forming a cohesive and robust core. The reason behind this phenomenon likely lies in the nature of the Bitcoin network itself. Bitcoin transactions are inherently interconnected, as each transaction typically involves inputs from previous transactions. Additionally, the network's decentralized and distributed nature fosters a high degree of connectivity, as transactions propagate through the network, validating and confirming each other.

Here are some properties of the largest strongly connected component (SCC) in the Bitcoin Alpha network.

Total Nodes	3235
Total Edges	23299
Total number of in-nodes	3235
Total number of out-nodes	3235
Total number of triangles	21575

Figure 11: Properties of the SCC

We can divide the whole network into four major parts:

- Dendrites: These are the disconnected components of the graph, meaning they are subsets of nodes that are not connected to the main component of the graph. In other words, dendrites are isolated clusters of nodes.
- SCC (Strongly Connected Component): This is the largest strongly connected component of the graph, meaning that every node in this component is reachable from every other node within the component, either directly or indirectly.
- Out: This consists of all nodes that can be reached from a given node in the SCC using forward BFS (breadth-first search). In other words, these are the nodes that can be reached by following edges in the forward direction from a starting node within the SCC.
- In: Similarly, this consists of all nodes that can reach a given node in the SCC using backward BFS. In other words, these are the nodes that can reach the starting node by following edges in the reverse direction within the SCC.

Table 1: Summary of Network Parts

Nodes in SCC	In Nodes	Out Nodes	Dendrites
3235	22	522	4

Next, we repeatedly selected 10 random nodes from a graph and performed forward and backward BFS from each node to determine the number of reachable out-nodes and in-nodes, respectively. We then recorded the occurrences of 0 in both the in-nodes and out-nodes counts across multiple iterations. The aim was to understand the network connectivity and know about the likelihood of encountering nodes with no incoming or outgoing edges. The results revealed that, on average, there were a 2 out of 10 instances where nodes had no outgoing edge, near 0 out of 10 instance where nodes had no incoming edges.

To find the most important node (or nodes) in a graph, we have used the In - degree centrality measure to quantify the importance or influence of nodes based on their network properties. The In - degree centrality is defined as the number of incoming nodes to a node divided by the total number of nodes in the network.

$$C_d(v) = \frac{\text{degree of node } v}{n-1}$$

```
Top 10 Most Important Node Sources:
Rank 1: Node Source: 1, In-Degree Centrality: 0.10523532522474881
Rank 2: Node Source: 3, In-Degree Centrality: 0.06636700158646 22
Rank 3: Node Source: 2, In-Degree Centrality: 0.0542041248016922 26
Rank 4: Node Source: 11, In-Degree Centrality: 0.0536753040719196 16
Rank 5: Node Source: 4, In-Degree Centrality: 0.0531464833421470 1
Rank 6: Node Source: 177, In-Degree Centrality: 0.0523532522474881
Rank 7: Node Source: 7, In-Degree Centrality: 0.0515600211528291 9
Rank 8: Node Source: 10, In-Degree Centrality: 0.0433632998413537 8
Rank 9: Node Source: 5, In-Degree Centrality: 0.0386039132734003 2
Rank 10: Node Source: 6, In-Degree Centrality: 0.0367530407191961 9
```

Figure 12: Important Nodes and Centrality

Importance of finding the most important node:

- Critical Points Identification: Identifying the most important node helps pinpoint critical points in the network. These nodes may serve as bridges between different network components, or they may have a high degree of influence over the network's behavior.
- Resource Allocation: Understanding the most important node allows for more efficient resource allocation. For example, in a social network, targeting influential users for marketing campaigns can lead to broader reach and impact.

4 COMMUNITY DETECTION

Detecting communities in the Bitcoin graph is important for several reasons. Firstly, it allows for a deeper understanding of the network structure, revealing clusters of nodes that share common characteristics or behaviour. This insight can aid in identifying potential patterns of transactions or interactions within the Bitcoin ecosystem.

Furthermore, the detection of communities can help in uncovering potential anomalies or fraudulent activities. Sudden changes or shifts in community structures may indicate suspicious behaviour, such as coordinated attacks or market manipulation. Communities may reflect geographical regions, regulatory jurisdictions, or specific use cases for Bitcoin.

We have used two algorithms for community detection which are Louvain Algorithm and Girvan-Neuman Algorithm. Following is the brief explanation of each algorithm and the results we have obtained.

4.1 Girvan Newman algorithm

The Girvan-Newman algorithm, developed by Michelle Girvan and Mark Newman in 2002, is a widely-used method for detecting communities within complex networks.

Here's a breakdown of how the algorithm works and how it can be applied to community detection in a Bitcoin transaction network.

- Edge Betweenness Calculation: It starts by calculating the betweenness centrality of all edges in the network. Betweenness centrality measures how often a particular edge lies on the shortest path between pairs of nodes in the network. Edges with high betweenness centrality are considered to be important bridges between different communities.

$$B_{ij} = \sum_{s \neq i \neq j} \frac{\sigma_{st}(i,j)}{\sigma_{st}}$$

Where:
- B_{ij} is the betweenness centrality.
- σ_{st} is the total number of shortest paths from node s to node t.
- $\sigma_{st}(i,j)$ is the number of those paths that pass through edge (i,j).

- Edge Removal: After calculating the betweenness centrality for all edges, the algorithm removes the edge with the highest betweenness centrality from the network. This process is repeated iteratively until the network breaks into disconnected components or until a predefined number of communities is reached.
- Community Identification: Once the network has been divided into disconnected components or communities, the algorithm can assign each node to a specific community based on the connectivity pattern of the remaining edges.

Upon applying the Newman Algorithm to our dataset, we obtained the following insights:-

The detection of six communities by the Girvan-Newman algorithm in a network suggests a structured organization characterized by distinct clusters of nodes. This result is highly influenced by the network's disconnectedness, where nodes form groups based on shared edges. The algorithm's identification of edges with high betweenness centrality, acting as bridges between communities, strengthens the separation of these clusters. Overall, the detection of six communities underscores the network's intricate organization, heavily shaped by its disconnected structure, and showcases the algorithm's effectiveness in revealing these underlying community structures.

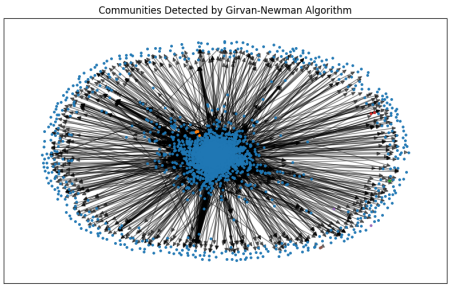

Figure 13: Communities with Girvan-Newman

No. of communities detected	6
No. of comm. having <= 11 nodes	5
No. of nodes in largest comm.	3764
No. of edges in largest comm.	24160

Figure 14: Properties of the Girvan-Newman Communities

4.2 Louvain Algorithm

The Louvain Algorithm, devised by Blondel, Guillaume, Lambiotte, and Lefebvre, efficiently detects communities in complex networks by assessing connection density. It operates in two phases: enhancing modularity by merging or separating communities, and constructing a new network where communities become nodes. Through iterative iterations, the algorithm reveals the network's community structure effectively. Upon applying the Louvain Algo-

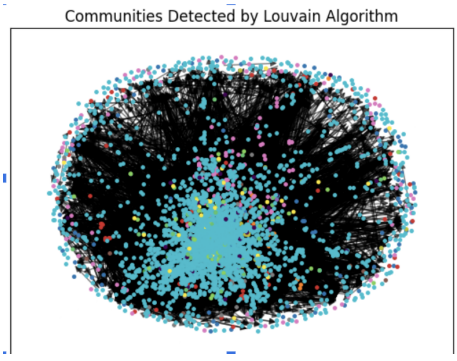

Figure 15: Communities with Louvain

No. of communities detected	484
No. of comm. having <= 15 nodes	30 approx.
No. of nodes in largest comm.	256
No. of edges in largest comm.	1312

Figure 16: Properties of the Louvain Communities

rithm to our dataset, we obtained the following insights:

- Number of Communities Detected: The algorithm identified a total of 484 distinct communities within the network. This indicates a rich and diverse structure with numerous cohesive groups of nodes.
- Distribution of Community Sizes: Approximately 30 communities contain 15 or fewer nodes. These smaller communities may represent tightly-knit groups with specific characteristics or functions within the network.
- Largest Community: The largest community comprises 256 nodes, indicating a significant portion of the network. This community likely plays a central role in connecting various parts of the network together.
- Edges in the Largest Community: With 1312 edges, the largest community exhibits a dense network of connections among its constituent nodes. This dense connectivity suggests a high level of interaction and cohesion within the community.

Based solely on the number of communities detected, the Louvain Algorithm appears to have performed better than the Girvan-Newman Algorithm. The Louvain Algorithm identified 484 distinct communities within the network, indicating a finer-grained partitioning of the network into cohesive groups. In contrast, the Girvan-Newman Algorithm detected only 6 communities.

However, it's important to consider that the effectiveness of an algorithm cannot solely be determined by the number of communities it detects. Other factors such as the quality of the detected communities, the computational efficiency, and the scalability of the algorithm also play crucial roles in assessing its performance. Therefore, a comprehensive evaluation would require a comparison of these factors between the two algorithms.

5 SPECTRAL CLUSTERING

Spectral clustering is a powerful technique for partitioning data points into cohesive clusters based on the similarity of their features. Unlike traditional clustering algorithms that rely on distance metrics, spectral clustering operates in the spectral domain, leveraging the eigenvalues and eigenvectors of a similarity matrix derived from the data. This approach offers several advantages, including the ability to handle non-linearly separable clusters and complex geometries within the data. In this article, we explore the principles behind spectral clustering and its applications across various domains.

Theory/Procedure: To find eigenvalues for clustering analysis, we first constructed the adjacency matrix or Laplacian matrix of the network, depending on the algorithm used. Next, we calculated the eigenvalues of this matrix using numerical methods such as eigen-decomposition. These eigenvalues represent the spectral properties of the network and provide insights into its underlying structure. To visualize these eigenvalues, we plotted them on a graph, with the x-axis representing the index of the eigenvalue and the y-axis representing its value. By examining the plot, we could observe patterns and trends in the distribution of eigenvalues, which can inform the selection of the optimal number of clusters, denoted as k, for the clustering algorithm. One commonly used criterion for determining k is the eigengap, which refers to the difference between consecutive eigenvalues. The most stable clustering is typically achieved by selecting the value of k that maximizes this eigengap. This criterion helps to identify a natural "break" or "gap" in the eigenvalue plot, indicating a significant change in the network's spectral properties and suggesting the presence of distinct clusters or communities. The formula is as given below:

$$\Delta_k = \left| \lambda_k - \lambda_{k-1} \right|$$

Figure 17: Ideal K

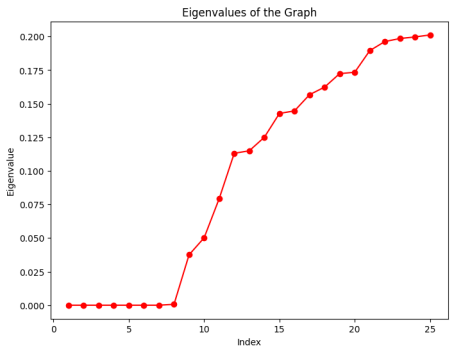

Figure 18: Eigenvalues with Unweighted Edges

As we analyze the spectral properties of the network, particularly focusing on the eigenvalues plot, we observe that the difference between consecutive eigenvalues is maximized when $k = 10$ in weighted graph. This indicates that partitioning the network into eight clusters results in the most distinct and stable clustering. By selecting $k = 10$, we achieve an optimal balance between intra-cluster cohesion and inter-cluster separation, effectively capturing the underlying structure of the network. Therefore, $k = 10$ emerges as the ideal choice for partitioning the network into meaningful clusters.

We found the following parameters:

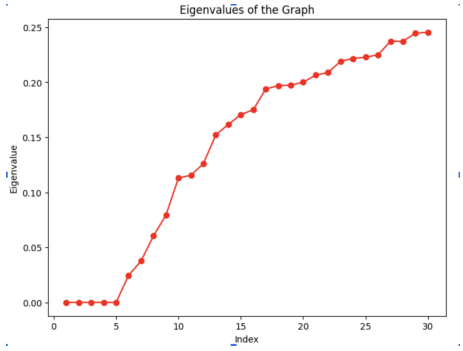

Figure 19: Eigenvalues with Weighted Edges

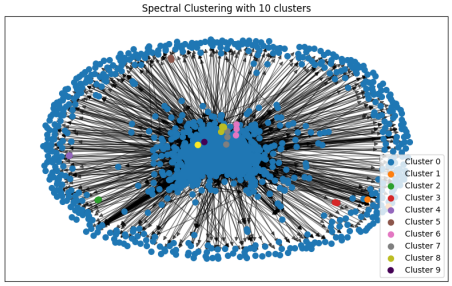

Figure 20: Spectral Clustering

Table 2: Number of Nodes in Each Cluster

Cluster	Number of Nodes
Cluster 1	3751
Cluster 2	2
Cluster 3	2
Cluster 4	3
Cluster 5	2
Cluster 6	2
Cluster 7	3
Cluster 8	3
Cluster 9	11
Cluster 10	4

6 COMMUNITY EVOLUTION

Previously, we employed spectral clustering to identify clusters within our network, representing communities of users. However, it's essential to recognize that these clusters, and thus the communities they form, evolve dynamically over time. This temporal evolution is influenced by changes in the network's structure, driven primarily by variations in the number of transactions occurring within the system. Spectral clustering relies on the calculation of the adjacency matrix and subsequent eigenvalues to partition the graph into clusters. Yet, as the eigenvalues are sensitive to changes in the network's topology, their values fluctuate with shifting transaction patterns. Consequently, the clusters identified by spectral clustering may vary over time, reflecting the evolving nature of

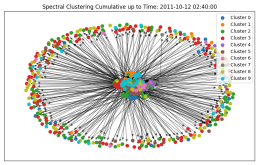

Figure 21: Clustering in year 2010

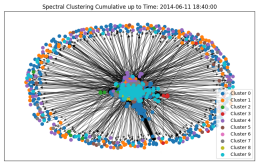

Figure 22: Clustering in Year 2013

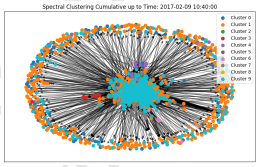

Figure 23: Clustering in Year 2016

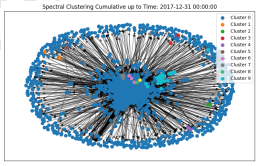

Figure 24: Clustering in Year 2017

Figure 25: A figure with two subfigures

community structures within the network. Here are few snapshots:

In our analysis, we can observe that the number of nodes in the largest cluster, as determined by spectral clustering, consistently increased over time. This observation suggests a gradual consolidation of network connectivity, wherein more nodes tend to form cohesive clusters within the Bitcoin transaction network. Such an increase in cluster size indicates a strengthening of community structures or the emergence of larger, more interconnected user groups over time. Consequently, it may signify a growing level of collaboration, transactional activity, or shared interests among network participants. This trend underscores the evolving nature of the Bitcoin ecosystem, characterized by the dynamic formation and expansion of community structures.

7 ENDORSED NODES AND ANOMALY DETECTION

7.1 Endorsed Nodes

Since our dataset is a who-trusts-whom network, we can explore how trust propagates through the network by examining which nodes endorse other highly endorsed nodes forming a recursive

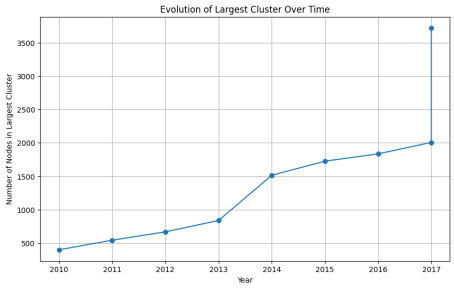

Figure 26: Evolution of largest community with Time

network of trust relationships. In our context, the most endorsed nodes are the ones that have received the highest trust rating. These nodes are considered highly reputable or trustworthy based on the collective endorsements they have received from other members of the network. Importance of endorsed nodes analysis:-

- Risk Mitigation: By knowing which nodes are endorsed, users can mitigate risks associated with interacting with unknown or untrusted entities. They can prioritize transactions or interactions with endorsed nodes, reducing the likelihood of encountering fraudulent or malicious behavior.
- Network Analysis and Insights: Analyzing endorsed nodes provides insights into the structure and dynamics of the network. It helps identify key players, influencers, or hubs that play pivotal roles in facilitating trust and connectivity within the network.

Following is the analysis of endorsed nodes performed on our dataset.

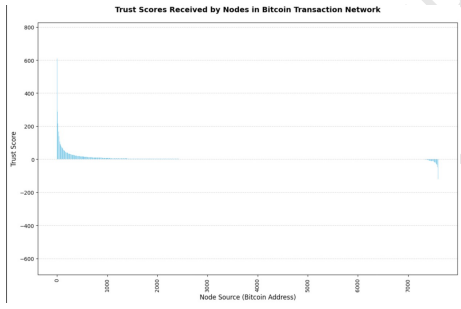

Figure 27: Trust scores vs Nodes

7.2 Anomaly Detection

In a Bitcoin Transaction Netwrok, millions of transactions happen each day, some of the transactions might cross a certain threshold and can be dangerous to the netwrok or the community. These transactions could be identified as fraudulent/scam. Anomaly Detection is a technique to find out unusual or suspicious transaction patterns.

In Fig 3, we have plotted the Transaction volume over the existence period of the network formed by our dataset. The green line represents the threshold to the number of ideal transactions that should happen over a network. We say that a point in the

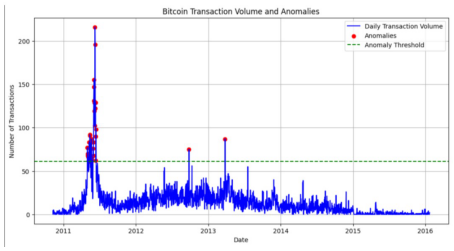

Figure 28: No. of Transactions with Time

transaction volume is an anomaly when the number of transactions above the green line changes suddenly and obtain a local maxima or minima. The red points in the graph represents the anomalous points/outliers and these might correspond to the Fraudulent/scam transactions.

8 FUTURE WORK

- Dynamic Community Detection Methods: We can explore dynamic community detection algorithms capable of identifying evolving community structures in the Bitcoin transaction network over time. These methods should adapt to changes in transaction patterns and network topology to provide more accurate and timely insights into community dynamics.
- Incorporating Transaction Attributes: We can investigate the integration of additional transaction attributes, such as transaction volume, frequency, and user behavior, into the analysis. By considering these factors alongside network structure, we can gain a more nuanced understanding of community interactions and identify emerging trends or anomalies.
- Temporal Network Analysis: Conduct a comprehensive temporal analysis of the Bitcoin transaction network to uncover temporal patterns, cyclic behaviors, and recurring motifs. This analysis can provide valuable insights into the temporal dynamics of community formation, evolution, and dissolution over different time scales.

ACKNOWLEDGMENTS

We would like to express our heartfelt gratitude to Progyan Das, our teaching assistant, and Prof. Anirban Dasgupta for their guidance and support throughout the course of our project.

9 REFERENCES

1 A. Singh, "Spectral Clustering," 2010. https://www.cs.cmu.edu/~aarti-/Class/10701/slides/Lecture21_2.pdf
2 A. Broder et al., "Graph structure in the Web," Computer Networks, vol. 33, no. 1–6, pp. 309–320, Jun. 2 https://doi.org/10.1016/
3 A. Benedikt Bünz, "Community Detection and Analysis in the Bitcoin Network CS 224W Final Report," 2015 https://snap.stanford.edu/class/cs224w-2015/projects_2015/Community_Detection_and_Analysis_in_the_Bitcoin_Network.pdf

4 D. Kondor, M. Pósfai, I. Csabai, and G. Vattay, "Do the Rich Get Richer? An Empirical Analysis of the Bitcoin Transaction Network," PLoS ONE, vol. 9, no. 2, p. e86197, Feb. 2014, doi: https://doi.org/10.1371/journal.pone.0086197

5 D. Du, "Social Network Analysis: Centrality Measures." https://ddu.ext.unb.ca/6634/Lecture_notes/Lecture_4_centrality_measure.pdf