Analysis and Visualization of Ethereum Transaction Data
Social Network Analysis

Group number: **09**
Authors:

- Marieke Bouma (S3142558)
- Robin Sommer (S2997592)
- Lazar Popov (S3340473)

1 Introduction

From their advent, cryptocurrencies have been increasingly popular. With Bitcoin as a first mover in 2009 and followed by Ethereum in 2015. Many other cryptocurrencies have appeared ever since with variable degrees of success.

When Bitcoin first emerged 13 years ago one of the biggest critiques was that it would not stay. Ethereum was considered ephemeral but yet it is still here 7 years later. Both of those cryptocurrencies have more or less established themselves and possibly created a standard tested through those years that other newer cryptocurrencies could follow.

Ethereum is the frontier of the smart chains by extending the architecture of Bitcoin to create a decentralized Turing complete virtual machine where programs with variable complexity called smart-contracts can be hosted and executed. As a frontier many of the smart contracts on Ethereum are copied to newer smart chains.

Since Ethereum is the oldest smart chain, it could be argued that the smart-contracts on the network have evolved in such a way that they belong to a subset of smart contracts that will appear on every smart-chain. On top of that, a subset of the smart-contracts on the Ethereum network can be part of a set that is necessary for every smart chain.

In this project we perform an exploratory analysis on the Ethereum blockchain transaction data from the perspective of Social Network Analysis.

2 Dataset

We use transaction data from the real Ethereum blockchain [1]. Since this data set is very very large, we opt to use data from only four days a year - the first day of each quarter, for the past five years: the first day of January, April, July, and October for years 2018, 2019, 2020, 2021, 2022.

We quickly noticed that quite a few addresses receiving transactions were NaN. In an attempt to figure out what this means, we checked the comments of some addresses frequently sending transactions to NaN, on Etherscan. We found that this is most probably a smart-contract creation call that could be from the cryptocurrency exchange Poloniex. Nevertheless, the transactions having NaN address were removed as they did not aid us in creating a graph.

In order to obtain graphs that we can actually visualise properly and work with, we filter the data significantly. We first find the intersection of all nodes across the 4 four days from years 2018, 2019, 2020, 2021 and 2022, to obtain those addresses who have sent and/or received at least one transaction each year. Then, for each node in the intersection, we use only those edges to other nodes in the intersection. We believe that this is a good summary of the "core" of the Ethereum network, as the nodes describe wallet addresses, smart-contracts, or exchanges that have persisted on the Ethereum network for the last 5 years.

Repeated transactions between the same two addresses in the same year were flattened to weighted edges. We obtain one directed graph for each of our five years, which we combine into a multiplex directed graph to ease visualisation. The code used to obtain the dataset, make graphs, and analyse them, is available in our GitHub page[1].

[1] https://github.com/Robin1711/SNA_ethereum_network

3 Network statistics - Metrics

In this section, we discuss several metrics in varying degrees of detail.

Year	2018	2019	2020	2021	2022	All (Multiplex)
Edges	32025	20259	18665	20147	12279	101386

Table 1: Overview of graphs and their edges. We consider the same 15671 nodes in each graph, and only the edges between those.

3.1 Degree distribution (in-degree and out degree)

In table 2 we can see the average degrees and the average weighted degrees for each year and for the Multiplex. Especially in the weighted degrees there is lots of variation between the different years. In section 3.7 we perform a more detailed analysis regarding the many nodes and the difference in (in-out) degree between the nodes.

Year	2018	2019	2020	2021	2022	All (Multiplex)
Degree	4.09	2.59	2.39	2.57	1.57	13.19
In/out-degree	2.04	1.29	1.19	1.29	0.78	6.60
Weighted degree	163.46	50.20	110.03	185.77	16.87	526.32
Weighted in/out-degree	81.73	25.10	55.01	92.89	8.44	263.14

Table 2: Overview of the average degrees of the nodes in each year.

3.2 Centrality indices

The centralities that we considered are the degree centrality, eigenvector centrality, betweenness centrality, and the closeness centrality. For each centrality we would need to dive deep into nodes in the network, which is infeasible with a network with 15671 nodes. We considered averages and maximum values, but without context these values would be meaningless.

The degree centrality is almost equal to the degree of the nodes. This is analysed in detail in section 3.7. The eigenvector centrality is similar to the degree centrality, but is a more recursive definition where importantness of the nodes are considered. We regarded closeness centrality as unimportant in our network as we can see from the other metrics that the network is divided in mainly two categories: big nodes with a high degree and small nodes with low degree connected to the big nodes. The closeness centrality will note that these nodes have the highest closeness scores as they are connected to the outermost nodes. Similarly we have the betweenness scores, which are very high for the big nodes. Removing one of these nodes will massively impact the network and will increase the number of connected components as lots of nodes are only connected to a big node.

3.3 Clustering coefficient

Similarly to the centrality scores, it is infeasible to dive into the many nodes of the network. For all nodes the clustering coefficients will be very low as most nodes are connected to the big nodes, but are not connected to each other, resulting in low clustering coefficients. We delve deeper into the network to find clusters, communities, and/or cliques in section 4

3.4 Network diameter

Here we show the average shortest paths in each network as well as the diameter of the networks - the longest shortest paths. The values are noted in table 3. What stands out is that all values are very low for a graph with 15671 nodes. The average shortest path lengths show that in many cases nodes can hop to other nodes in around 4 steps. The longest shortest paths are also around length

10. Both values indicate either that the network is tightly knit together, or that the network has a few big authorities, that every smaller node is connected to.

Year	2018	2019	2020	2021	2022	All (Multiplex)
Avg. shortest path	4.67	3.44	3.67	4.58	2.82	4.42
Diameter (max shortest path	11	10	10	12	5	12

Table 3: Overview of the densities of the graphs over all the years.

3.5 Density

Density is the fraction of the edges over all possible edges. From the density we can deduct how close the network is to being complete, i.e. all nodes are connected through links. We can see the density of the graph over several years in table 4. We observe that all values are very small, and year 2018 has the largest density. Interestingly, the multiplex, has a lower density than all the other graphs.

Year	2018	2019	2020	2021	2022	All (Multiplex)
Density	$1.30 * 10^{-4}$	$8.25 * 10^{-5}$	$7.60 * 10^{-5}$	$8.20 * 10^{-5}$	$5.00 * 10^{-5}$	$4.21 * 10^{-5}$

Table 4: Overview of the densities of the graphs over all the years.

3.6 Connected components

A connected component (CC) is a component in which all nodes are connected through some path. Every node in the set can reach another node in the same set through following the edges. We are using a directed graph, so we have *weakly* connected components and *strongly* connected components. The difference is which edges can be followed for creating a path from one node to the other. A *weakly* connected component means that each node is reachable via an edge, regardless the direction of the edge. A *strongly* connected means that, in the set of nodes of a strongly connected component, each node is reachable via a directed edge path along the graph.

Year	2018	2019	2020	2021	2022	All (Multiplex)
# strongly CC	14644	15532	15579	15345	15646	
largest strongly CC	944	120	73	295	5	
# weakly CC	2183	3063	3445	3460	5991	
largest weakly CC	13422	12535	12141	12124	9547	

Table 5: Overview of the connectedness of the graphs in each year.

From table 5 we observe that the number of strongly connected components are very big, indicating lots of one-way transactions. It can be deducted that many of these are single node components as the numbers are close to the total number of nodes (15671) in the graphs. Consequently, this also leads to a small number for the largest strongly connected component for each year. Over all graphs the largest connected component is 944, which is 6.4% of possible nodes. Incredibly the largest connected component for 2022 is 5 nodes, from 15671 possible nodes.

From the table we can also observe that the number of weakly connected components is significantly lower. Looking at the largest weakly connected component of all years, in year 2018, we see that the number of nodes in the component is very close the number of nodes in the graph 13422 nodes of 15671 possible nodes result in 85.6% of the graph being involved in one single weakly connected component.

Combining the numbers for strongly and weakly connected components, it points to the fact that in each year there is probably one single big component with lots of outgoing or incoming links. As the number decreases over the years we expect that the largest player in the network decreases in size.

3.7 Degrees: weighted vs. unweighted

Since we use weighted graphs, we can analyse both the weighted and unweighted degrees. For example, the unweighted out-degree of a node describes how many different addresses it has sent transactions to. In contrast, the weighted out-degree describes the total number of outgoing transactions for that node. We first analyse the various degrees in the combined multiplex graph.

Address	Label	Deg_{in}	Deg_{out}	Degree
0x...ec8	Ethermine mining pool	2	23942	23944
0x...3b5	NanoPool	6	11734	11740
0x...0be	Binance Exchange	3514	2186	5700
0x...ec7	USDT Clipboard scam	3733	0	3733
0x...b98	Bittrex Exchange	7	2738	2745
0x...cc2	WrappedEther Clipboard scam	2222	0	2222
0x...4f3	Yobit Exchange	1768	326	2094
0x...ced	Binance Exchange	4	1892	1896
0x...819	DEX	1808	0	1808

Table 6: Nodes with the highest unweighted degrees in the multiplex graph. These addresses are all in either the top 5 nodes with highest in-degree or out-degree. The address in bold is in the top 5 for both.

We notice that most nodes have either a very high in-degree and very low out-degree, or vice versa. In our context, this corresponds to miners and token contracts, respectively. The small in-degrees of miner addresses correspond to the mining rewards. One exception with a high total degree is the one boldened in Table 6. This Ethereum address is an exchange address.

	Address	Label	Deg^w_{in}	Deg^w_{out}	Deg^w
1	0x...ec7	USDT Clipboard scam	1571248	0	1571248
2	**0x...0be**	Binance Exchange	17092	597168	614260
3	0x...ced	Binance	223	337763	337986
4	0x...4ed	Bitstamp	0	266553	266553
5	0x...428	Bitstamp	266553	0	266553
6	0x...ec8	Ethermine mining pool	3	252565	252568
7	0x...b98	Bittrex Exchange	11	241234	241245
17	0x...2e2	Tron	96244	0	96244
18	0x...db0	Eos Clipboard Scam	93534	0	93534
21	0x...c07	OMG Network	77002	0	77002

Table 7: Nodes with the highest weighted in- or out-degrees in the multiplex graph (top-5). The leftmost column describes their ranks when sorting by total weighted degree.

	Address	Label	Deg^w_{in}	Deg^w_{out}	Deg^w
31	0x...819	OMG Network	43588	0	43588
45	0x...4f3	Yobit Exchange	22034	767	22801
94	0x...cc2	WrappedEther Clipboard scam	8226	0	8226

Table 8: Weighted degrees of multiplex nodes with high unweighted degrees.

The weighted in-, out-, and total degrees tell a slightly different story. Firstly, we observe that many of the addresses with high unweighted degrees, also send or receive many transactions per neighbour, but some are lower in the top 100 with the highest weighted degrees (see Table 8). The latter thus have, on average, fewer transactions per neighbour than those ranked 1-7 in Table 7.

Moreover, it seems that having a high weighted in- or out-degree does not always correspond to a high total degree. The three nodes at rank 17, 18, and 21 in Table 7 are in the top 5 of highest

	Address	Label	Deg^w_{in}	Deg^w_{out}	Deg^w
8	0x...ab3	Huobi 9	4	153120	153124
9	0x...f2b	Huobi 1	6	153093	153099
10	0x...24f	$15 M wallet	12	152999	153011

Table 9: Weighted degrees of multiplex nodes with out-degrees just outside of the top 5 highest.

Year	2018	2019	2020	2021	2022	Multi
No. of cliques	21441	17822	16204	15397	15134	29069
> 1	15483	11006	9399	8943	6608	26088
> 2	1818	540	466	571	22	6480
> 3	62	0	0	0	0	312
> 4	1	0	0	0	0	3
> 5	0	0	0	0	0	0
Largest clique	5	3	3	3	3	5

Table 10: Overview of cliques in our data.

weighted in-degrees, but are surpassed in the total weighted degree ranking by nodes which have a high weighted out-degree, while just outside of the top 5 (such as those in Table 9). This suggests that the most active addresses in our Ethereum data set are largely miners and exchanges, while the single most active address in our data set is a contract. Looking at the Etherscan comments for this address, it seems that this is a clipboard scam and the address has been blocked.

Lastly, we note that some nodes in our graph have exactly opposing in- and out-degrees; for example, the addresses with rank 4 and 5 in Table 7. These nodes also both only have unweighted degrees of 5, but very high weighted degrees. We comment on this in Section 4.

4 Communities

It seems that our data does not contain many cliques large enough to spring to the eye in the data or the visualised graphs, which makes sense considering the nature of our data. We note that directed graphs are in general less likely to contain cliques than their undirected counterparts, since subgraphs need to be complete in order to be cliques. Luckily, we can find the largest cliques for each of our graphs using a NetworkX function. The results are shown in Table **??**.

This shows that there are in fact some cliques, with most of them simply being pairs. This is something we noticed earlier in our analysis already: pairs of nodes that have exactly the same (large) value, one for its in-degree and one for its out-degree. This suggests that these addresses belong to the same company, where one is a contract frequently mined by the other. For example, Etherscan indeed shows that the two addresses mentioned in the latter part of Section 3.7 both belong to Bitstamp, a Luxembourg-based cryptocurrency exchange.

Much more scarce are cliques of larger size. In most years, only a few triads exist, and no cliques bigger than that. Looking at the addresses involved in some of the cliques, it seems that cliques, especially those bigger than 3 nodes, describe mining pools. For example, the clique of size 5 in 2018 consists of the Binance exchange and four trivial addresses that do not pop up elsewhere in our analysis.

Again due to the nature of our data, we opt out of analysing homophily and bridges specifically.

5 Longitudinal analysis

We can perform a brief longitudinal analysis by comparing the graphs and their metadata over the years. Per year, we analyse the weighted degrees for the 10 nodes with highest weighted degree overall (Table 11). This shows how the activity of each address over time. As one would expect, we see some addresses remaining steadily active, while others increase or decrease rapidly in terms

of activity. Specifically, the two nodes of the Binance Exchange both have weighted degree 0 in 2022. We have not been able to find out why this is the case.

Address	Label	2018	2019	2020	2021	2022
0x...ec7	USDT Clipboard scam	565	280	1	1	1
0x...0be	Binance Exchange	1	3	4	2	13369
0x...ced	Binance	5	6	15	3	14635
0x...4ed	Bitstamp	3	7	21	5	3
0x...428	Bitstamp	4	8	22	4	2
0x...ec8	Ethermine mining pool	8	1	3	14	4
0x...b98	Bittrex Exchange	2	4	19	16	5
0x...ab3	Huobi	33	48	13	6	29
0x...f2b	Huobi	37	49	8	10	14
0x...24f	$15 M wallet	41	45	14	7	16

Table 11: The 10 nodes with highest weighted degree overall from Section 3.7, and their "ranks" based on highest weighted degree per year. A node with rank 1 is the node with highest degree for that year.

In the visualised graph of the network over time, we observe, as hypothesized, a slight shift from the main cluster to several mining pools. Figure 1 shows this.

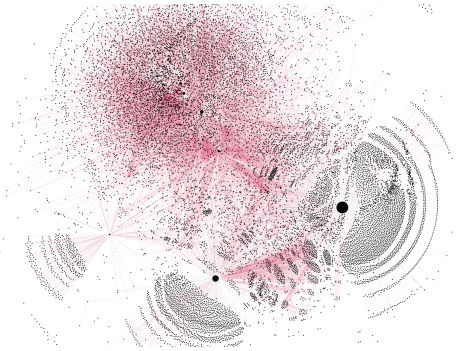
6 Link Analysis

In this section, we analyse our data with both the Page Rank and HITS algorithms for link analysis. Page Rank indicates the importance of a node based on the number of other nodes pointing at it and their importance. Hub score indicates how much a node acts as a Hub, i.e. a node that points to many different nodes with different weights to the edges. An Authority score score is computed as well, which is similarly to Page Rank but it has the Hub score included in the calculation.

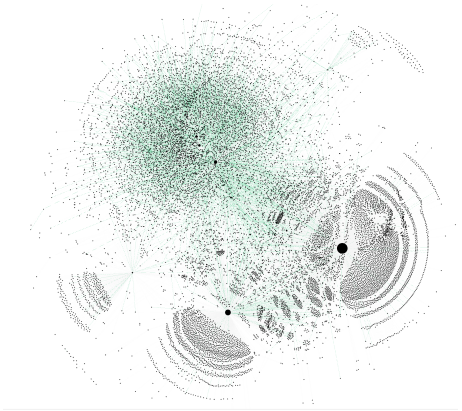
Address	Label	Page rank
0x...ec7	USDT Clipboard scam	0.020144
0x...cc2	WrappedEther Clipboard scam	0.008379
0x...0be	Binance Exchance	0.007371
0x...819	DEX	0.006675
0x...d88	Poloniex Exchange	0.005394
0x...208	IDEX DEX	0.004648
0x...6ca	Token Contract chainLink	0.003125
0x...4f3	Yobit Exchange	0.002473
0x...359	Token Contract Maker	0.002353
0x...c07	OMG Network	0.002316

Table 12: Page rank on multiplex nodes

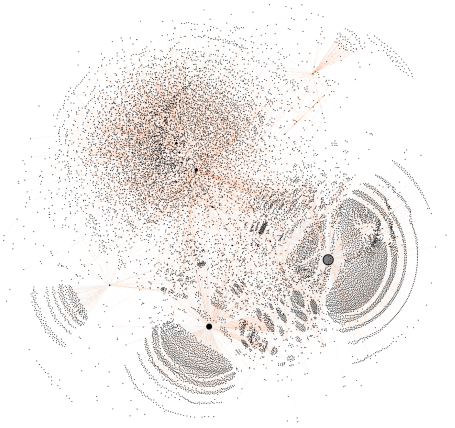
Table 12 shows the 10 addresses with highest page rank in the multiplex graph and their labels. We see that the top 2 are two contracts that are involved in Clipboard scam. It makes sense that those scams are with a high page rank as they have received a lot of funds from different people. Moreover, it checks out that exchanges are with high page rank score since people send their cryptocurrencies to exchanges in order to exchange them for another cryptocurrency. Figure 2 shows the three nodes with highest page rank visually.



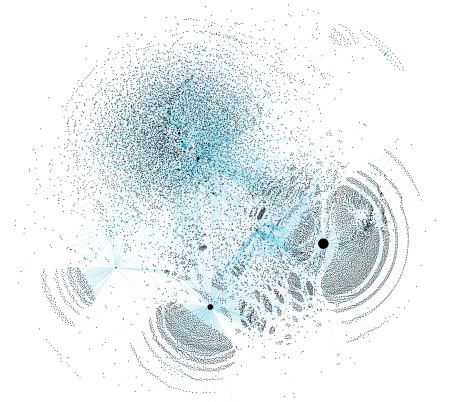
(a) 2018



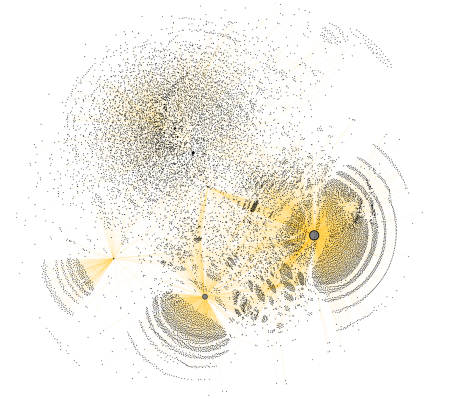
(b) 2019



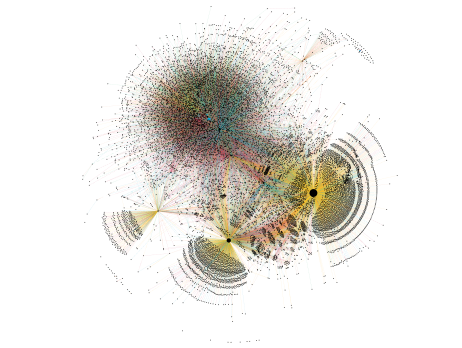
(c) 2020



(d) 2021

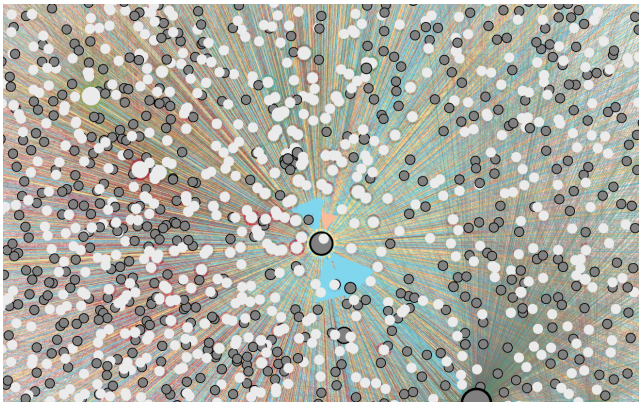


(e) 2022

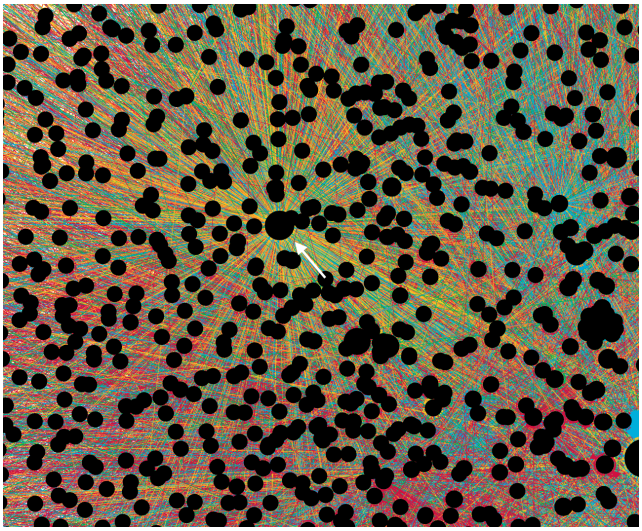


(f) Multiplex graph

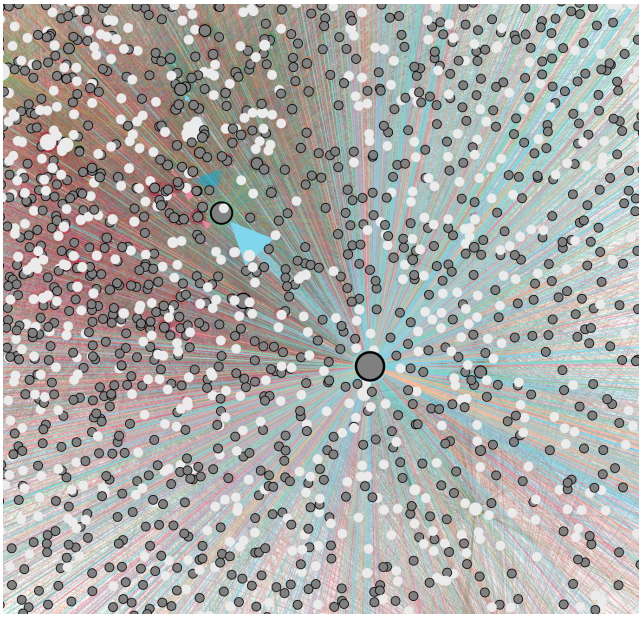
Figure 1: Full graphs per year and the multiplex graph.



(a) USDT Clipboard scam address

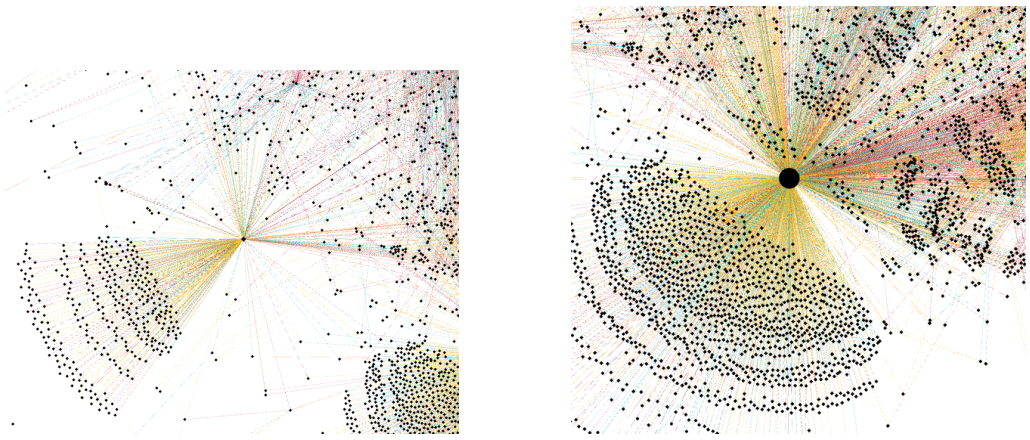


(b) Wrapped Eth Clipboard scam



(c) Binance exchange

Figure 2: High page rank nodes.



(a) F2Pool (b) Nanopool

Figure 3: High Hub score.

Address	Label	Hub
0x...9e8	Mining	0.010689
0x...d88	Poloniex Exchange	0.00819
0x...0be	Binance Exchance	0.005064
0x...ced	Binance	0.004854
0x...4f3	Yobit Exchange	0.003692
0x...6e8	HitBTC Exchange	0.00359
0x...dc0	Kraken Exchange	0.00351
0x...830	F2Pool	0.003026
0x...cfa	Bitfinex Exchange	0.002094
0x...fc7	Hotbit Exchange	0.001972

Table 13: Hub score on multiplex nodes

Table 13 shows the 10 addresses with highest Hub scores in the multiplex graph and their labels. We see that we have Mining pools and exchanges. This makes sense since people can get out their fund of exchanges after done trading. Also it makes sense that mining pools are considers Hubs since people collect their rewards from mining. Figure 3 shows these nodes visually.

In Table 14 we can see the top 10 addresses and their labels ordered by Authority scores. We see that on the top we have several user wallets followed by mining pools, exchanges, Chainlink Token contract, and finally USDT Clipboard scam which might sound ironic. The user wallets are unknown wallets that could not be labelled but they non the less were shown to have the highest authority score. They could be indeed user's wallet but more likely is that they are unlabelled exchange addresses.

Address	Label	Authority
0x...7b6	User wallet	0.016955
0x...cbe	User wallet	0.016945
0x...343	User wallet	0.016739
0x...692	User wallet	0.016722
0x...3b5	NanoPool	0.006169
0x...b98	Bittrex Exchange	0.003434
0x...359	Token Contract Maker	0.003159
0x...c07	OMG Network	0.000974
0x...6ca	Token Contract chainLink	0.000681
0x...ec7	USDT Clipboard scam	0.00068

Table 14: Authority score on multiplex nodes

7 Discussion and conclusions

We performed a network analysis on a 4 years, 20 day summary of the Ethereum Network. Our summary included an intersection of all the active nodes over those days. We hoped to get a better understanding of the core of Ethereum and other smart-chains.

The main observation we made was that the biggest players in the Ethereum network are exchanges, mining pools and scam addresses. We also found some contracts like Tron, Chainlink and the OMG Network, which are protocols that are built on top of the Ethereum blockchain.

What is more, we also found a possible trend of creating more edges over the recent years toward mining pools which shows the emergence of mining pools as a concept. There are many interesting direction that a follow-up project could take.

Implicitly our goal was to see what the Ethereum Decentralised Machine is made of. In other words, we wanted to see what are the main smart contracts and dApps running on Ethereum. We did not manage to get to that level of analysis in the scope of this project, since Ethereum is largely used as a speculative tool. Exchanges, mining pools and scams are the most prevalent addresses. A better analysis would be to filter those addresses out and what would be left would be what we were looking for initially. The contacts mentioned above like Tron, Chainlink and OMG Network are a testimony to the fact that digging a little deeper might pay off.

Another interesting project we could branch out could be observing the future behaviour of mining pools since Ethereum has stitched mining protocols. A thing we could improve would be adding txnHash data so we could more accurately understand the different nodes.

References

[1] Allen Day, Evgeny, and Google BigQuery. Ethereum blockchain dataset, 2018.

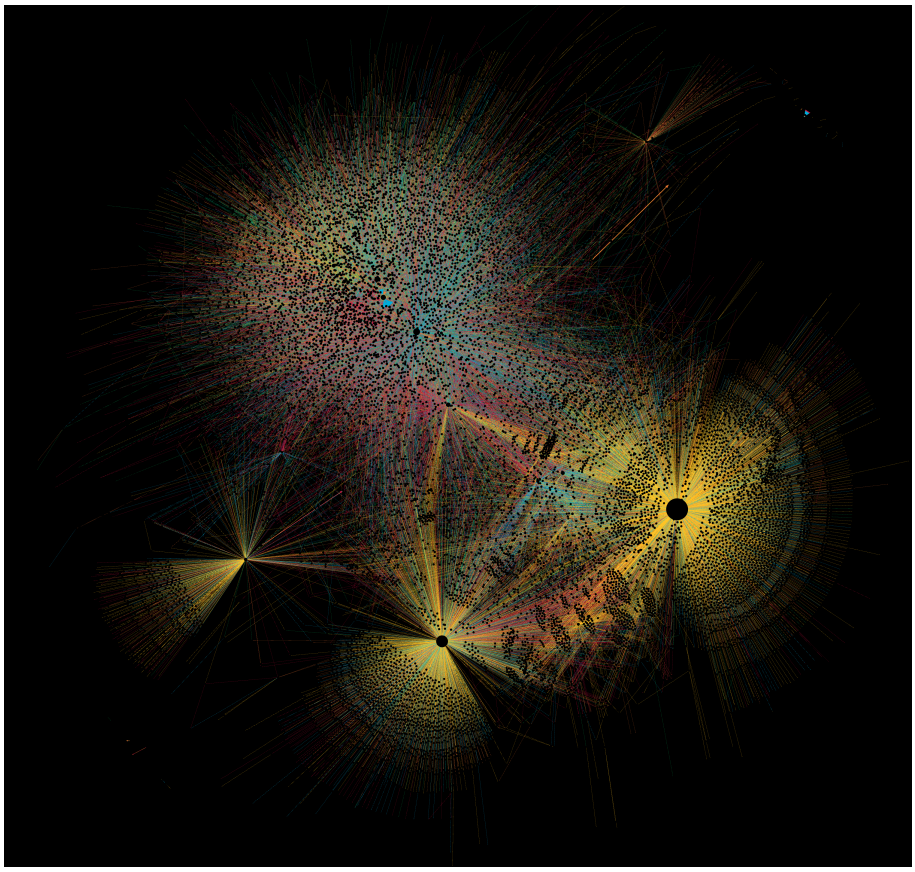


Figure 4: Beautiful network