The Ethereum network: a graph analysis

Quinten Bruynseraede R0674455
Louis Van Looy R0861408

May 24, 2021

Abstract

This work investigates the Ethereum network and more particularly focuses on the transactions of Ether. Various graph analysis tools are used to obtain insights about this network. First of all, a descriptive graph analysis of the network is performed, followed by an analysis of the major players in network (called *whales*), to address their role and influence within the network. Finally, we investigate the community structure of the network through the Louvain, Leiden and Eigenvector method, complemented with a visualization of the network structure. Six different snapshots of the network are used, starting in January 2017 and ending in March 2021. This allows to investigate the importance of whales and the evolution of the network. We found that the network has a fairly stable character, with the whales playing a very prominent role in the network.

1 Introduction

The Ethereum network is a blockchain-based network for distributed processing. It contains a thriving ecosystem of distributed applications (dapps), ranging from financial tools to games and gambling. Ethereum's native token is called Ether (ETH), and it has become a acknowledged way of storing and transferring money: a so-called *store-of-value*. In this report, we analyze how ETH is distributed among users, and how it flows between them. In particular, we analyze the role of a group of very wealthy users (called *whales*), and we investigate how important these whales are for the network.

In Chapter 2, we introduce some basic concepts related to the Ethereum network. Chapter 3 outlines past work about network analysis of cryptocurrency networks, in particular the Ethereum network. In Chapter 4, we describe how we collected the necessary data. Chapter 6 analyzes the network using a set of network analysis tools, and try to show the importance of whales. Chapter 7 focuses on inequality within the Ethereum network. In Chapter 8, we try to analyze clusters in the network, and try to link these clusters to their utility in the network.

2 Ethereum

In the following section, some fundamental concepts and definitions needed to understand the Ethereum network are explained.

2.1 Ethereum

Ethereum is an open-source, blockchain-based and decentralized software platform. It makes use of its own cryptocurrency, called Ether. The main novelty of Ethereum is its ability to execute smart contracts and create decentralized, distributed applications without any downtime, fraud, control or involvement from a central party. A smart contract can be seen as a piece of code that describes the rules for a transaction between two parties, similar to a written contract. Smart contracts on the Ethereum network are Turing-complete (i.e. they can describe arbitrary computations), and are mostly programmed in the Solidity programming language.

2.2 Addresses and transactions

Users store their Ether in an account, uniquely represented by a 20-byte address, such as `0xed9a430d9a11616eb1cb07ebc28c9e20a03bd486`. Transactions contain (among other fields), a sender, a recipient and a transaction value expressed in *wei*. 1 Ether is equal to $10e^{18}$ wei. All transactions are public by default, but there is no direct link between the identify of a user and his address: the vast majority of transactions happen anonymously. However, several services use addresses that are publicly known. For example, Binance, one of the largest cryptocurrency marketplaces, has at least 20 known addresses [1].

2.3 Whales

An important element within the Ethereum network are whales. Ma gives the following definition:"The term *whale* is used to describe an individual or organization that holds a large amount of a particular cryptocurrency." [6]. It is clear that this value can fluctuate, based on the underlying value of the cryptocurrency. In the Ethereum network, a user is considered a whale if they own more than 10 000 Ethereum. These holders control approximately 70% of the total supply of Ether [2] [3]. The name 'whale'

[1] https://etherscan.io/accounts/label/binance
[2] https://coinmarketcap.com/nl/headlines/news/ethereum-eth-whales-hold-68-percent-total-supply/
[3] https://blockworks.co/ethereum-whales-gobble-70-of-eth/

refers to the fact that these users have a lot of market power within the network and are able to push the market up or down. A whale is not necessarily an individual, the term could also refer an institution or organization that holds a significant amount of cryptocurrencies within the network [6].

2.4 The network

The Ethereum network has seen tremendous growth since its inception in 2015, and contains over 350 million transactions at the time of writing. Since June 2020, the number of daily transactions exceeds 1 million [4]. Over 100 million addresses have been registered, and over 600 thousand addresses take part in transactions every day [5]. This means a full analysis is generally not feasible.

3 Related Work

It has to be said: the literature around graph analysis of the Ethereum network is quite young but strongly emerging. Guo et al. (2019) [5] were one of the first to address the properties of this network, by investigating it in the initial phase and in a more developed phase. They found out that several transaction features, such as the transaction volume, transaction relation, and component structure of the graph, exhibit the same heavy-tail property which can be approximated statistically by the power law function. Our work gives a more complete overview of the *evolution* of the network, because we analyze it at six stages instead of just two.

Chen et al. (2020) [3] performed a graph analysis on the Ethereum network. They focus on the characteristics of its users, smart contracts, and the relationships among them to get a deeper understanding of the network. Money transfers, smart contract creations and smart contract invocations are central within their work. In this work, metrics like degree distribution, degree centrality, Pearson coefficient, clustering coefficient and PageRank are used, which makes it a good benchmark for this work.

Victor & Lüders (2019) [10] focus on the trade of tokens on the Ethereum network, more specifically called ERC-20 token networks. These tokens can be seen as digital assets, built on top of ETH. These tokens strengthen the Ethereum ecosystem by driving demand for ETH, needed to power the smart contracts. They provide an overview of 64 000 ERC20 token networks and focus graphically on the top 1 000. In this work, they observe that the token networks are frequently dominated by a single hub and spoke pattern, supplemented with very small clustering coefficients. The token network looks disassortative and a lot of the network activity is directed towards exchanges whereby many owners never transfer their tokens at all. As ERC-20 tokens can be considered a smart contract, we will not spend much time investigating this aspect, but they are an indispensable part of the Ethereum ecosystem. Interestingly, we observe many identical properties when looking at the Ethereum network.

4 Data collection

Given the size of the Ethereum network, we have to resort to sampling techniques. One type of sampling uses *random walks*, that start at a given point and follow edges while collecting the necessary data. Random walks are particularly useful for networks where it is not possible to delimit a part of the network in advance [1]. Ethereum is such a network: addresses are not ordered or structured in any way. We collect transactions using a Breadth-first search (BFS). To analyze the evolution of the network, we create six separate snapshots: each snapshot spans 1 month and is limited to 1 million transactions. We describe our data collection method in pseudo-code:

```
T = ∅
Initialize Queue with 100 known addresses
while T.size ≤ 1e6 do
    1. Pop an address A from Queue
    2. Fetch all transactions by A during the
    specified time period
    3. Add all transactions to T
    4. Add all addresses that occur in these
    transactions to Queue
end
```

Although transactions contain many more features, only the columns `from`, `to`, and `value` are used. Each snapshot can then be considered a graph: nodes are addresses and edges are transactions between addresses. The value of the transaction is the weight of the edge. If multiple transactions occur between the same pair of addresses, the sum of their values is taken.

The following table shows the size of each snapshot:

Snapshot	**#txs**	**#nodes**	**#edges**
Jan '17	475 598[a]	81 242	186 942
Nov '17	988 086	548 355	657 697
Sep '18	1 000 258	528 506	635 623
Jul '19	1 006 153	550 393	673 967
May 20	1 003 812	474 030	580 773
Mar '21	1 001 864	566 733	701 482

[a] Queue was empty before 1M transactions could be collected.

To account for seasonal effects, the time between two snapshots is always 10 months. We also don't consider the state of the network before January 2017, when it was still relatively new and the number of transactions was low.

[4] https://etherscan.io/chart/tx

[5] https://bitinfocharts.com/comparison/ethereum-activeaddresses.html

5 Analysis of the Ethereum network

In this section, we look at some global characteristics of the Ethereum network. Wherever possible, we show how these characteristics evolved over time.

5.1 Degree and transaction volume

First of all, it appears that the average degree of the first snapshot in Jan'17 is high, 4.602, whereas in the following snapshots this number is significantly lower and lies within the range of 2.398 and 2.476. We suspect the smaller size and consequentially larger influence of the starting addresses cause this larger average degree in 2017.

Figure 1 shows the degree distribution for each snapshot. The distribution appears to follow a power law at first (estimated $k^{-2.1}$), but as the degree increases, the distribution is more or less uniform. In particular nodes with a degree above 5000 would be very rare in a perfect scale-free network, but we observe them often. It is noteworthy that Guo et al. [5] find a power law distribution with an exponent of $k = 3.23$, whereas we find $k = 2.1$. We suspect this difference can be explained by our data collection method, which favours nodes with a high degree (i.e. transactions with many other nodes) and consequently the phenomenon of preferential attachment.

Next, we look at the transaction volume in each snapshot. For our network, the transaction volume of a node is simply the sum of all outgoing transactions. Similar to what Guo et al. discovered [5], the transaction volume appears to follow a power law.

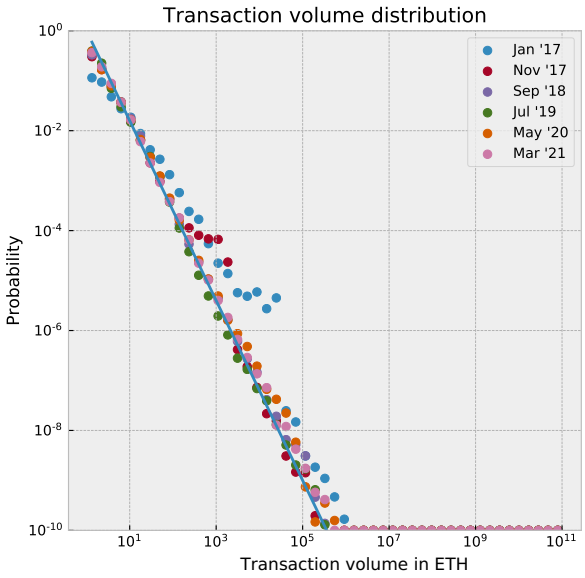

Figure 2: Distribution of transaction volumes

This logarithmic plot shows how the transaction volumes are distributed for each snapshot. First of all, it appears that transaction volumes have not changed much over time. A deviation is observed in the 2017 snapshots, in which transaction volumes were a bit higher than average. We suspect this deviation may be caused by the so-called *2017 bull run*, a period of spiking prices, which attracted many new investors. Interestingly enough, the bull run that is currently ongoing is not reflected in the transaction volumes of the March 2021 snapshot.

5.2 Small-world characteristics

Many networks can be considered *small-world networks*: they have a) a high clustering coefficient, and b) a low average path length [11] (often $L \propto log(N)$). Compared to random graphs generated with the Erdös-Rényi model, small-world graphs have many cliques or near-cliques, and many hubs: nodes that connect cliques and facilitate shorter path lengths.

The Ethereum network on the other hand, has a very low clustering coefficient: in all snapshots, we measured this coefficient as less than 1e-4. In their original paper, Watts and Duncan demonstrate small-world networks with clustering coefficients up to 0.79, so the Ethereum network is certainly not heavily clustered. Apart from January 2017, where the clustering coefficient was a bit higher, it has remained more or less constant. This results are also in line with the findings of Chen et al. (2020) [3].

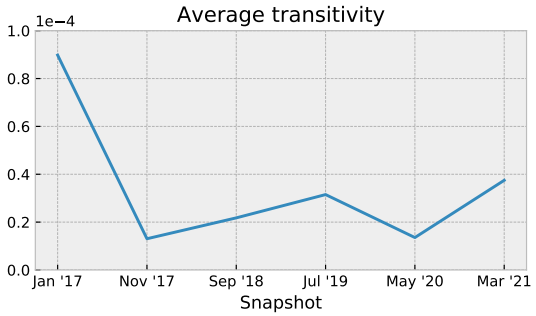

Figure 3: Evolution of transitivity over time.

The average geodesic path length, however, is very small, and appears to be in the order of magnitude that is required for a small-scale network: $L \propto log(N)$.

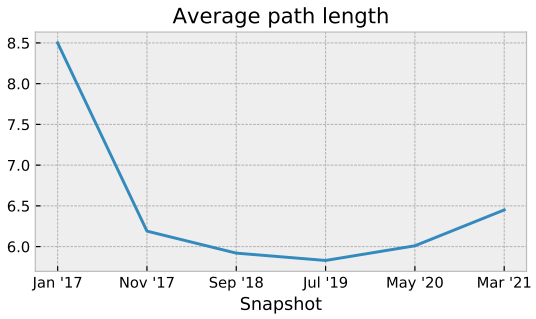

Figure 4: Evolution of transitivity over time.

Remarkably, for both transitivity and average path length, the January 2017 snapshot appears to be an outlier: we suspect that the lower number of nodes leads to a slightly higher average.

Based on these metrics, we can conclude that the Ethereum is not a typical small-world network: the network is not nearly clustered enough. However, the average path length is quite small. We suspect that the network is structured as a number

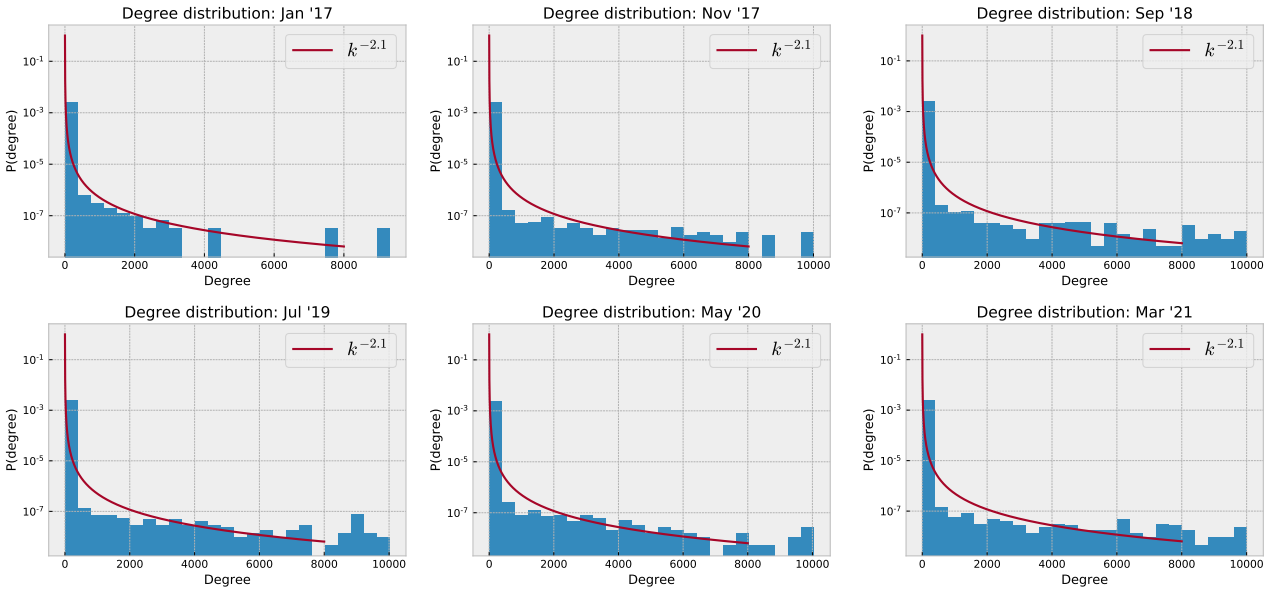

Figure 1: Degree distribution for all snapshots

of highly-active hubs (most likely exchanges), that trade with many individual users directly. These users rarely trade among themselves. This claim is also supported by the degree distribution: a very large part of the network has only one connection. Furthermore, many addresses likely belong to individual users, who have no intention of trading regularly, but just hold Ether as a long-term investment.

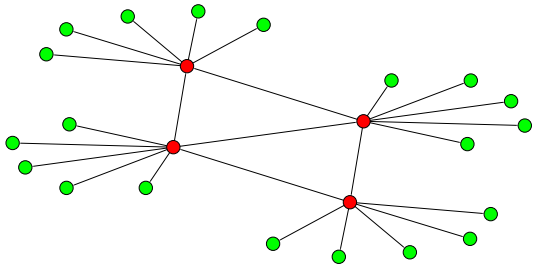

Figure 5: Structure of Ethereum network (red nodes are 'hubs', green nodes are 'users').

5.3 Whales

We now shift our focus to whales. Here we define those as addresses that posses at least 10 000 ETH. Unfortunately, it is not possible to identify whales in the past: finding out the balance of an address at a certain point in times requires backtracking the entire Ethereum blockchain, which is over 4 terabytes large. Therefore, our analysis of whales is limited to the current situation.

Currently, 1292 addresses are considered whales. Our most recent snapshot (March 2021) includes transactions from more than 500 000 addresses, of which **60** are whales. Despite making up less than 0.01% percent of the network, whales account for a trading volume of more than 25% (2.33 million out of 9.26 million ETH). Obviously, this imbalance can partially be attributed to the fact that whales have the capacity to trade large amounts, but it strengthens our suspicion that whales are very important traders, and may coincide with hubs.

We induce a subgraph by selecting all whales and the edges between them. We can now compare the full network with this subgraph.

	Full graph	**Whales**
Shortest path length	6.46	1.84
Transitivity	3.7e-5	0.06
Diameter	22	4

These differences indicate that whales are somewhat more connected between themselves than the entire network. However, these numbers have to be compared carefully, as there are only 60 whales, which naturally leads to shorter paths and a smaller diameter of the network. Transitivity, however, is a relative metric and shows an increased connectedness between whales. Still, from an absolute point of view, a transitivity of 0.06 is not high: we can definitely not conclude that many whales are directly connected.

Figure 6 shows a visualization of the 60 whales we found in March. Fortunately, some addresses are publicly known and can be linked to their organization or goal. As expected, most whales orginate from cryptocurrency exchanges and a large portion of whales remains anonymous. We also show which whales interact with each other, and the node size is proportional to the addresses' balance in March 2021.

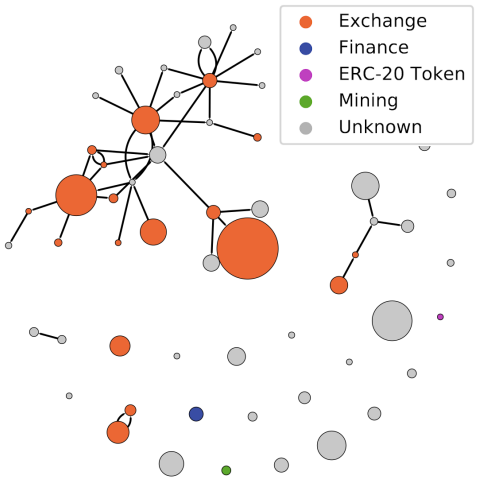

Figure 6: Visualization of whales by address type.

We conclude:

- Almost all whales are exchanges.
- All exchanges trade with other whales.
- Non-exchange whales do not trade as much with other whales.

As we expected, exchanges form an important and highly connected backbone to the network. They hold and trade large amounts of Ether. Other applications that make use of the Ethereum network (such as mining, financial products and tokens) do not play such a large role yet. This also confirms our suspicion that Ether is nowadays mostly traded as an investment; we do not yet observe a large amount of activity between users and applications on the Ethereum network.

6 Community detection

Previously, we investigated cluster-like structures among whales. We now apply proper community detection techniques to each snapshot, using the modularity measure to evaluate the clustering. Given the size of our networks, an efficient technique is desired. Additionally, we have no prior expectation about the number of clusters in our network. The following clustering approaches have been analyzed:

1. **Louvain method:** a divisive method that claims linear time complexity on sparse graphs. Its hierarchical nature may also help resolve the resolution limit problem [2].
2. **Leiden method:** an improvement to the Louvain algorithm that prevents communities that are badly connected or even completely disconnected [9].
3. **Eigenvector method:** this method expresses modularity in terms of eigenvectors of the so-called *modularity matrix* [8]. The recursive implementation of this method will keep splitting communities until no longer possible, or a fixed number of communities is reached. To prevent it from endlessly splitting, we set the maximum number to 100, a bit higher than the average number of clusters by the Leiden and Louvain methods.

Figure 7 & 8 show the results of community detection.

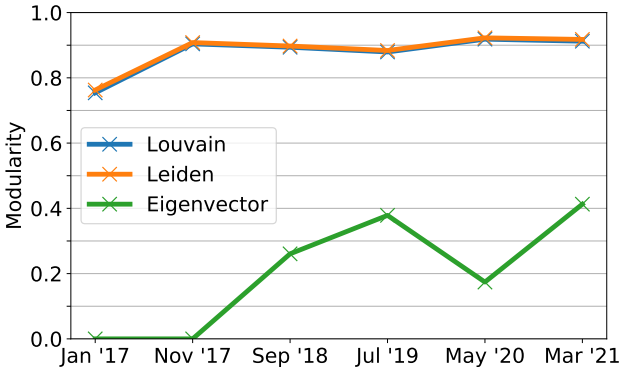

Figure 7: Modularity for three clustering methods.

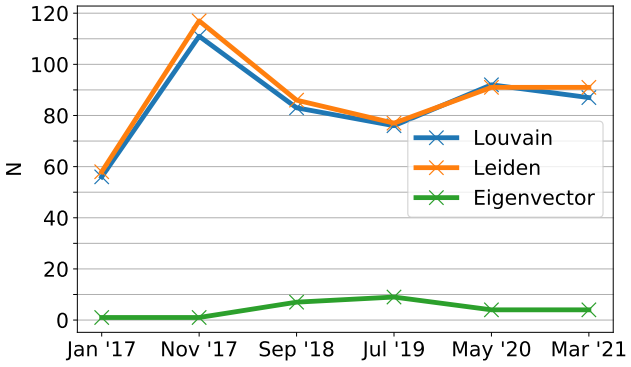

Figure 8: Number of clusters for three clustering methods.

Clearly, the Louvain and Leiden methods find very similar clusterings. This leads us to believe that the principal problem that Leiden tries to overcome (badly connected communities), does not occur in our networks. The modularity of their clusterings is very high (The number of communities is quite stable over time between 70 and 90. Again, we observe a deviation in the first snapshot, caused by the smaller network size.

The eigenvector method yields a very different result: a low number of clusters (often even 0) and a modularity under 0.4.

7 Visualization

In this section, we visualize the snapshots to see whether the characteristics mentioned in Section 5.2 and shown in Figure 5 can also be obtained visually. Gephi was used to visualize the network, mainly for its flexibility and representational power. However, Gephi's limitations became apparent when visualizing the network: the size of the snapshots was too large to be workable. That is why only a part of the network is used in the visualizations.

Figure 9 shows the network in January 2017 with 39 281 nodes and 50 154 edges. Figure 10 is a representation of July 2021 with 49 125 nodes and 50 034 edges. The used layout algorithm is OpenOrd [7] because of its ability to distinguish clusters in a convenient way and ability to work with large networks. In the second step, the Yifan Hu layout was used on the OpenOrd graph representation. This is a very fast algorithm with good empirical results on large graphs. It combines a force-directed model with a graph coarsening technique to reduce the complexity. The combination of these two gave actually a very clear representation of the network. [6]

The colors of the graph are the main clusters which are obtained from the Leiden algorithm. To keep the figure simple, only the largest seven clusters are shown. These seven clusters contain about 70% of the nodes.

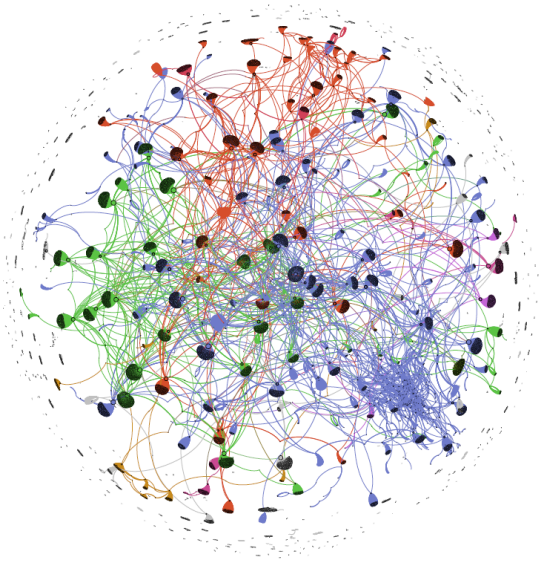

Figure 9: Visualization of the ETH network in January 2017.

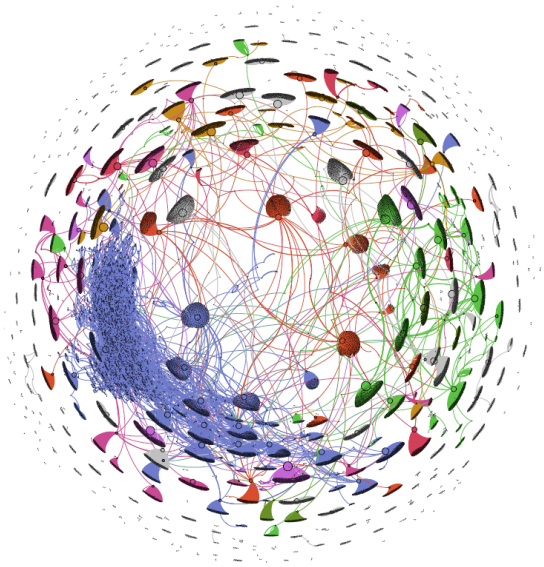

Figure 10: Visualization of the ETH network in March 2021.

[6] https://gephi.org/users/tutorial-layouts/

What is mostly notable about this visualizations is that you see some major nodes (with a high degree) are surrounded by a cloud of nodes with a smaller degree. These visualizations are similar to the ones shown in the works of Guo et al. [5] and Chen et al. [3].

8 Validity, reliability & scalability

In this section the validity, reliability and scalability of our analysis will be discussed. Validity refers to the fact to which extent our analysis accurately represents the real world. Reliability tries to investigate to which extent the results and conclusions remain consistent when different methods are used. The third criteria evaluates to which extent the proposed metrics can cope with the fact that real world networks can be huge in size and is concerned with space and/or time complexity of the methods used.

- **Validity:** when constructing the network, we had to make some compromises to get a dataset of manageable size. First of all, we limited ourselves to a graph of 1 million nodes: the full network is probably structured slightly differently than the subnetwork we analyzed. To generate a trustworthy sample, a BFS algorithm was used, but the addresses used to initialize search queue may affect the validity of the network. The size of the snapshots should compensate for this to some extent.

- **Reliability:** for each method, we compared the results of several algorithms and metrics: we looked at degree distribution and transaction volumes, investigated the small-world characteristics using transitivity and average path length, and compared whales using geodesic path length, transitivity and diameter. Furthermore, we also used visualizations to confirm some properties about whales. To detect clusters, we compared three algorithms. In each case, we saw that different metrics and methods confirm our intuitions. Finally, wherever possible, we analyzed the network at 6 different points in time, adding another dimension to our insights.

- **Scalability:** due to the size of our networks, scalability has already been an important issue during our analysis. First of all, our data collection will not scale well, because the Etherscan API severely limits the number of requests per second. It is however possible to sync a personal node to the network, and query locally. Even though we used a fairly large network, most metrics and methods used still have excellent time complexities. We summarize these in the table below. Assume V is the number of nodes, E is the number of edges, and d is the average node degree.

Degree distribution	$O(V)$
Transitivity	$O(V \times d^2)$
Shortest path length	$O(V^2 \times E)$
Diameter	$O(V \times E)$
Louvain	"near linear"
Leiden	"near linear"
Leading eigenvector	$O(E + V^2)$
OpenOrd Layout	$O(N) \times log(N)$
Yifan Hu	$O(N) \times log(N)$

All complexities can be found in the igraph reference manual [4] and the Openord paper [7].

9 Conclusion

In this work, the money flow of Ethereum was examined using tools from graph analysis. One of the first findings is that transaction volumes have not changed much over time, except for 2017, where the volumes were a bit higher. A second point of interest in the analysis was the small-world characteristic: we obtained a very low clustering coefficient within the network, but on the other hand the average geodesic path length did show signs of a small-world network. Furthermore, we looked at whales and concluded that they form a more connected backbone to the Ethereum network with many high-volume trades. Our results are in line with the results of the literature to some extent, but of course we were also limited in the examination of the entire network.

References

[1] Luca Becchetti, Carlos Castillo, Debora Donato, and Adriano Fazzone. A comparison of sampling techniques for web graph characterization, 2006.

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, Oct 2008.

[3] T. Chen, Y. Zhu, Z. Li, J. Chen, X. Li, X. Luo, X. Lin, and X. Zhange. Understanding ethereum via graph analysis. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pages 1484–1492, 2018.

[4] Gabor Csardi. igraph reference manual.

[5] Dongchao Guo, Jiaqing Dong, and Kai Wang. Graph structure and statistical properties of ethereum transaction relationships. *Information Sciences*, 492:58–71, 2019.

[6] John Ma. Cryptocurrency glossary: whale.

[7] Shawn Martin, W Michael Brown, Richard Klavans, and Kevin W Boyack. Openord: an open-source toolbox for large graph layout. In *Visualization and Data Analysis 2011*, volume 7868, page 786806. International Society for Optics and Photonics, 2011.

[8] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3), Sep 2006.

[9] Vincent Traag, L. Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 03 2019.

[10] Friedhelm Victor and Bianca Lüders. *Measuring Ethereum-Based ERC20 Token Networks*, pages 113–129. 09 2019.

[11] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *Nature (London)*, 393(6684):440–442, 1998.