Abstract

Background: This research uses exploratory graph analysis to analyze the transaction data of the Ethereum network. This is achieved through network visualization and mathematical and statistical modelling of the network data.

Methods: The dataset used in this study was extracted from the Ethereum in the BigQuery public dataset, specifically selected transactions in July 2019. The transactions were firstly modelled as network graphs and then visualized using the Kamada-Kawai and force-directed graphs layouts. Further modelling was explored with classical random graph and network block, with emphasis on network cohesion, hierarchical clustering and community membership.

Results: Looking at the network visualization and hierarchical clustering of the data, the network shows 170 clusters, the largest having 135 members. Through random graph modelling the optimum number of clusters is shown to be 95. Referring to the generated dendrograms, notable large transactions center around the DRINK token, the Maximine Exchange, the Upbit2 Exchange and the IDEX Exchange, identified through public disclosure of their Ethereum addresses. The network graphs tend to go towards the DRINK smart contract and the Maximine Exchange, indicating deposit actions, while it is the opposite for the IDEX Exchange. Further analysis also shows a different number of communities than the expected number. Falling short of the expected 170 clusters, the model is not able to capture additional mechanism that may be present at the density and social interaction distribution level of the network. On the other hand, network block modelling shows only four major clusters out of the 170 expected clusters, an indication that the model is not able to capture the network sufficiently.

Conclusions: The study was able to capture and model the interconnectedness of the system with its notion of elements, in this case, the transactions on the network.
Introduction
A blockchain is a distributed and replicated data structure in which a linked chain of digital information known as a block, is stored as a public database or ledger. Each of the blocks in the chain may contain zero or more records of transactions or exchanges. If another exchange happens on the network, a copy of the transaction is added to the record of every member on the network.1

Blockchain has several characteristics that make it so compelling. Firstly, blockchain is immutable. It cannot be corrupted, changed, or altered.2 A blockchain is also decentralized if it is a public one. There is no governing authority overseeing the network. In addition, information on the blockchain is cryptographically hashed, thus preserving privacy and integrity of the data that is stored.2 Ethereum is an open-source blockchain, known for its decentralized smart contract platform. It is programmable and allows for the fabrication and distribution of decentralized applications (DApps).3

This research investigates the transaction data of the Ethereum network, through a graph analysis approach. This is achieved through network visualization and mathematical and statistical modelling of the network data. The work presented here provides a targeted analysis of the Ethereum network, capturing the interconnectedness of the system with its notion of elements, in this case, the transactions on the network.

Related works
Small-world models were proposed by Watts et al.4 These creators were interested in the way that numerous networks in the genuine work show significant levels of clustering, yet with the little distance between most vertices. They recommended rather starting with a network graph with a cross-section structure, and afterward arbitrarily 'reworking' a little the level of the edges. Assuming we have many N vertices that are instigated on an intermittent style, each of the vertex joins its neighbors to their respective side.5

Barabási et al. proposed a preferential attachment model for modeling networks.5 The more associated a vertex is, the more probable it is to get new connections. A vertex with a further extent has a more grounded capacity to get joins added to the network.5

Exponential random graph models (ERGMs) have relations to generalized linear models. However, the appropriate specification and fitting of exponential random graph models can be more unobtrusive than that of a standard generalized linear model. Furthermore, a significant part of the standard inferential foundation accessible for generalized linear models, based upon asymptotic approximations to fitting chi-square distributions, is not in the consideration for exponential random graph models.7

Stochastic block models (SBM) are a class of models in the statistical data analysis of the network or graph data, and they can be utilized to find or comprehend the structure and the latent of a network graph for clustering. This model will in general deliver a graph containing networks, subsets described by being associated with edge densities.7

Methodology
Dataset
The dataset of transaction records of Ethereum was taken from the BigQuery public dataset (Ethereum in BigQuery).8 The attributes from_address and to_address will be used as nodes (vertices) and the relationship between them will be the edges of the nodes in this study. The attributes and their descriptions are shown in Table 1. In this exploratory study, only arbitrarily selected 1000 contiguous rows of transactions in July 2019 are considered.15

Software for analysis
The analysis is performed in R version 4.0.2, utilising the igraph R package for network analysis. The source code used can be found at Zenodo.16

Network graph
A graph \( G = (V, E) \) is a set of \( E \) of links (edges), a set \( V \) of nodes (vertices), and where components of \( E \) are unordered pairs \( \{u, v\} \) of distinct vertices \( u, v \in V \). The number of vertices \( N_v = |V| \) and the number of edges \( N_e = |E| \) are also called the order and size of the graph \( G \), respectively.7
Visualizing the network

To visualize a large network having over 1000 vertex and edge attributes, few methods exist. Kamada, et al. proposed a technique to draw general graphs.\(^9\) This technique tends to be broadly utilized in the network structures that are managed by the system. The mathematical displacement between the node in the drawing can be identified by graph-theoretic displacement. The spring algorithm proposed by Kamada, et al. has good properties such as symmetric drawings, edge crossing with a relatively small number of edges, and isomorphic graphs with almost congruent drawings.\(^9\)

The DrL method provides two-dimensional visualizations of exceptionally huge theoretical chart structures.\(^10\) The graph is drawn employing a force-directed algorithm based on recreated strengthening. This clustering is utilized to create a coarsened graph (less vertex) which is at that point redrawn. This handle is rehashed until an adequately small graph is produced.\(^10\)

Charactering network cohesion

One way to deal with characterizing the network cohesion of a specific network graph is through the identification of the subgraph of interest. The standard case is that of finished subgraphs and henceforth are subgroups of vertices that are completely strong, as in all vertices inside the subgroup are associated by edges. Large groups of subgraphs are considered rare in practice because they need the required network graph itself to be dense.\(^7\)

\[
\text{den}(H) = \frac{|E_H|}{\binom{|V_H|}{2}}
\]

is the general formula of the density of a subgraph \(H = (V_H, E_H)\), that the recurrence of acknowledged edges is comparative with possible edges, in an undirected graph \(G\) which has no self-loops and various edges. In the situation that if \(G\) is a directed graph, \(|V_H|(|V_H| - 1)\) will be replaced with the denominator of the equation above. The value of zero and one in \(\text{den}(H)\) is to justify how near the subgraph \(H\) is to be becoming a clique in the network.\(^7\)

In graph clustering, the relative frequency also can be one of the methods to defining a cluster. The standard use of the term clustering coefficient normally means the quantity \(cl_T(G) = \frac{3t_3(G)}{2t_2(G)}\), where \(t_3(G)\) is the number of triangles in the graph \(G\), and \(t_2(G)\) is the number of connected triples. The value of \(cl_T(G)\) is the transitivity of the graph and is also referred to as the fraction of transitive triples A graph \(G\) is fully connected if each node is reachable from all other nodes, and that a graph’s connected component is a maximally connected subgraph.\(^7\)

Hierarchical clustering

Various methods have been proposed for clustering, varying in essentially by the way they assess the nature of the proposed clustering and quality of the enhancement brought about by the algorithms.\(^11\) These strategies adopt a greedy strategy to looking through the states of all potential partitions \(C\), through altering progressive applicant partitions iteratively.

One of the most popular measures is modularity.\(^10\) Let \(C = \{C_1, \ldots, C_k\}\) as a given candidate partition and define \(f_{ij} = f_{ij}(C)\) to be the fraction of edges in the original network that connect vertices in \(C_i\) with vertices in \(C_j\). The modularity of

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>hash</td>
<td>Hash of the transaction</td>
</tr>
<tr>
<td>nonce</td>
<td>The number of transactions made by the sender prior to this one</td>
</tr>
<tr>
<td>transaction_index</td>
<td>Integer of the transactions index position in the block</td>
</tr>
<tr>
<td>from_address</td>
<td>Address of the sender</td>
</tr>
<tr>
<td>to_address</td>
<td>Address of the receiver. Null when it’s a contract creation transaction</td>
</tr>
<tr>
<td>value</td>
<td>Value transferred in Wei</td>
</tr>
<tr>
<td>gas</td>
<td>Gas provided by the sender</td>
</tr>
<tr>
<td>gas_price</td>
<td>Gas price provided by the sender in Wei</td>
</tr>
<tr>
<td>input</td>
<td>The data sent along with the transaction</td>
</tr>
<tr>
<td>receipt_cumulative_gas_used</td>
<td>The total amount of gas used when this transaction was executed in the block</td>
</tr>
</tbody>
</table>
$C$ is the value mod $C = \sum_{k=1}^{2} \left( f_{kk}(C) - f_{kk}^* \right)^2$, where $f_{kk}^*$ is the expected value of $f_{kk}$ under some model of random edge assignment.\(^{10}\)

**Classical random graph modelling**

Graph modelling refers to a collection $\{P_\theta(G), G \in g : \theta \in \omega\}$, where $g$ is a collection of possible graphs, $P_\theta$ is a probability distribution on $g$, and $\theta$ is a vector of parameters, ranging over possible values in $\omega$. A collection $g$ and a uniform probability $P(\cdot)$ over $g$ is usually used to refer to a random graph model.\(^{7}\) Erdos et al. established a series of seminal papers on the classical theory of random graph models.\(^{8}\) Erdos et al.’s model specifies a collection of $g_{Nv, Ne}$ of all graph $G = (V, E)$ with $|V| = N_v$ and $|E| = N_e$, and assigns probability $P(G) = \left( \begin{array}{c} N_e \\ N_v \end{array} \right)^{-1}$ to each $G \in g_{Nv, Ne}$, where $N = \left( \begin{array}{c} N_v \\ 2 \end{array} \right)$ is the total number of distinct vertex pairs.\(^{8}\)

**Network block models**

In a random graph, $G$, let $G = (V, E)$. The graph can be categorized into one of $Q$ classes. Each element of the block model for a particular graph can be specified as each element of the adjacency matrix $Y$, labelled as $q$ and $r$ of vertices $i$ and $j$, respectively, with a probability of Bernoulli random variable. Snijders, et al. introduced a parametric limit of the form.\(^{13}\)

**Results and discussions**

The network graph (Figure 1) shows a total number of 170 clusters among 1000 rows of data. The top three clusters inside the network graph consist of the size of 135 vertices, followed by 55 vertices, and 39 vertices. There are 103 clusters having a pair of vertices, that is the smallest cluster size in this network.

The Kamada-Kawai layout uses a combination of spring model and the technique local minimization of global energy.\(^{15}\) The graph is shown in Figure 2.

Clauset proposed a fast and greedy approach to optimization through an agglomerative type of hierarchical clustering algorithm.\(^{11}\) Figure 3 shows the outputs from the clustering, while Figure 4 shows the dendogram from the clustering. The clusters are the same as the ones shown from the network graph (Figure 1). There are also many two-member communities in this network, which may indicate disposable addresses after single use.

Looking into the clusters, Figures 5 and 6 show the membership of the largest cluster, with 135 addresses. We notice that all the addresses point toward address 0x0089659fd609933d16a5cd6c2be1a5dc1be24ad. Looking at the Ethereum blockchain Explorer, Etherscan,\(^{14}\) the address belongs to the DrinkChain (DRINK) token.

---

**Figure 1. Network graph.**
Figures 7 and 8 show the membership of the second-largest cluster which contains a total number of 55 addresses. Members of the clusters point toward the address 0x8e766f57f7d16ca50b4a0b90b88f6468a09b0439. This address belongs to the token, Maximine Coin (MXM).

Figures 9 and 10 show a cluster with 39 addresses, they all point to the address 0xBA826fEc90CEFdf6706858E5F-baFcb27A290Fbe0. This address belongs to Upbit2, a digital asset exchange in South Korea. This cluster is most probably showing withdrawals as the network graphs are pointing out from the center.
Figures 11 and 12 show a cluster with 20 addresses, they all point to the address 0x2a0c0DBEcC7E4D658f48E01e3fA353F44050c208.

This address belongs to IDEX, a decentralized exchange.

From hierarchical clustering, it is shown there are 170 communities.

To examine this further, network modelling was considered. In this case, two assumptions were made. The first one, the same order of vertex and size, 800 and 1000 respectively, was maintained. The second is the same as the first, but with a fixed degree of sequence.

In random graph modelling, Monte Carlo methods were used to generate approximations quickly over 1000 trials. For each of the trials, community detection was employed. Figure 13 shows the number of communities. It is shown both assumptions detected 24 and 95 communities, respectively. This is nowhere close to the actual 170 number of communities. There could be additional mechanism at work which complicate the density and the social interaction distribution in modeling the network.
Figure 6. Dendogram of the DrinkChain cluster.

Figure 7. Maximine Coin cluster.
Figure 8. Dendogram of the Maximine Coin cluster.

Figure 9. Upbit2 Coin cluster.
Figure 10. Dendogram of the Upbit2 cluster.

Figure 11. IDEX cluster.
**Figure 12.** Dendogram of the IDEX cluster.

**Figure 13.** Relative frequency vs number of communities.
Figure 14. Integrated conditional likelihood (ICL) vs classes.

Figure 15. Reorganized adjacency matrix.

Figure 16. Summary of class connection probabilities.
Figure 14 shows the integrated conditional likelihood (ICL), with its corresponding classes $Q$. It is shown that the fitted model has 4 classes, which suggests a suitable latitude of 4 for this model.

To investigate further, we estimated the posterior probability of each of the class members. We noticed that the evidence for the class membership assignment shows a very strong relationship between vertices with a maximum posterior probability of 99%. From the reorganized adjacency matrix (Figure 15), 0.57894737 and 0.38596491 will be the larger proportions and the rest the smaller class proportions.

Figure 16 shows the visual summary of the class connection probabilities and the concordance of the 4 classes. Falling short of the expected 170 clusters, the modelling effort was not able to capture additional mechanism that may be present at the density and social interaction distribution level of the network.

Conclusions
The paper presented a targeted analysis of the Ethereum network, capturing the interconnectedness of the system with its notion of elements, in this case, the transactions on the network. The study takes into account only selected transactions in July 2019, which may not be generally representative of the Ethereum network. This will be addressed in future work.

Data availability
Underlying data

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Extended data

Archived analysis code as at time of publication: https://doi.org/10.5281/zenodo.5336085.16

License: GNU General Public License v3.0.

References
10. Martin SB: Dr.L: Distributed Recursive (Graph) Layout. Computer software. 2007.
The benefits of publishing with F1000Research:

• Your article is published within days, with no editorial bias
• You can publish traditional articles, null/negative results, case reports, data notes and more
• The peer review process is transparent and collaborative
• Your article is indexed in PubMed after passing peer review
• Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com