Network Analysis (Project Report)

Riccardo Mioli - 983525
riccardo.mioli2@studio.unibo.it

July 10, 2022

1 Abstract

The object of this report consists in a whole network study of the CryptoPunks' network and the study of the bipartite network formed by the NTFs and their owners.
By analyzing the main characteristics of the CryptoPunks' network it was found that some strongly cohesive groups are present, even if the global network isn't heavily clustered. Another distinctive network detail is that owners are heterophile in respect to their degree and the number of rare tokens exchanged.
Finally, the bipartite network was analyzed to verify if the degree of an NFT could be used as a proxy to express its rarity; what was found is that the two variables have a very low correlation coefficent, therefore the degree can't be used to infer the rarity of a token.

2 Context

In 2017 the CryptoPunks were launched on the Ethereum blockchain and from that moment the concept of NFT was born. NFTs are digital collectables whose ownership and trading history are recorded on the blockchain, a decentralized distributed digital ledger, using a smart contract. Since their creation, NFTs have become popular products and the prices of certain CryptoPunks have reached several million dollars.

3 Problem and Motivation

To this day, there is a very little number of studies about the networks that are established from NFT trades. This study will help to consolidate and validate what we know about those networks.
Furthermore, the existing studies do not take into consideration how the rarity of the tokens impacts on connections between users. For example, a question like "Does owners of rare tokens only have interactions between them?" still has no answer. By responding to this question we can find out if there is any form of homophily between the users related to the rarity of the tokens.
Another undiscussed topic is the one regarding the formation of cliques and clusterization of the network; so far, studies haven't taken this aspect into consideration, which can have a non-marginal impact on the NFTs' market manipulation.

Type	Frequency
Human	98.79% (9870)
Zombie	0.88% (88)
Ape	0.24% (24)
Alien	0.09% (9)

Table 1: CryptoPunks types distribution.

4 Datasets

The datasets used in this study are two: the historical log of the CryptoPunks exchanges and the set of the tokens' characteristics. In both cases, the data is from secondary sources and is publicly available.
The first dataset used was directly gathered by querying the APIs offered by Etherscan [1], which is the best way to acquire data from the blockchain if you don't have a full Ethereum node. The data obtained is composed of the events emitted by the smart contract managing the CryptoPunks and consists of 230 JSON files with 1,000 events per file.
The second dataset [2] used is formed by 10,000 rows and it lists the attributes of each CryptoPunk. The two datasets were downloaded by using the `requests` Python 3 library.
A first cleanup of the first dataset was performed by parsing all the events and maintaining only those referring to token exchanges; the filtering process used the `topic` field, which is a hexadecimal string that specifies what the event concerns. After this phase, only 47,000 events were kept and saved as an edge list with the following parameters: `timestamp, event_type, source, target, punk_id`. Subsequently, the `punk_id` was used as a key to associate the `punk_type` attribute from the second dataset to the exchange data; for this process, `pandas` was used due to the powerful merge methods that are offered by this library. Finally, the refined data was saved in a file named `all_exchanges.csv`. This file is used to construct a graph containing all the exchanges between the users and a bipartite graph containing the NFTs linked to the users that exchanged them.
Additionally, two attributes were added to the nodes by scraping the initial CSV:

- `token_rarity`: The rarity of the token expressed as an ordinal number. The value is assigned based on how many tokens of a specific type exists (i.e. the lower is the frequency of a type, the higher its rarity is). Since there are 4 types of CryptoPunks (Human, Zombie, Ape, Alien), values from 0 to 3 were assigned. In table 1 it is reported how the types are distributed.

- `rare_freq`: The sum of the transactions performed by each owner involving rare tokens (i.e. non Human CryptoPunks).

5 Validity and Reliability

The data retrieved from the blockchain represent the exact history of the CryptoPunks' exchanges, on the other hand, keeping an immutable history of transactions is the main objective of the blockchain. However, logs emitted by a smart contract could be subject to errors that can undermine the validity of a study, which is our case, and more specifically it happens when an exchange is made by accepting a buying offer from a user. Due to a bug in the code, when an

offer is accepted, the offer variable is reset and the bidder address is set to the `NULL address`, which then corresponds to `0x0000000000000000000000000000000000000000`.
After these operations, two events are thrown: `PunkBought`, which contains the `punk id`, the `bid value`, the `seller` and the `NULL address` as the `buyer`, and `PunkTransfer`, which contains only the previous owner address and the new owner address. To fix the receiving address of the `PunkBought` event the field `to` of the `PunkTransfer` event was used, this is possible since the two events are emitted in the same transaction, therefore they are "linked" together.
Regarding the validity of the study, another aspect that needs to be discussed is the one concerning the exclusion of certain addresses from the analysis. Ethereum addresses can be classified as externally owned, which means that they are controlled by humans, and contract addresses, which are controlled by code. Non-human addresses (i.e. the `NULL` address and more generally the contract addresses) could be excluded from the analysis, but this would impact in two ways:

1. non-human addresses could be recipient of NFTs in the same way as human addresses, in fact, there is no difference between these two types of accounts except in how they are controlled;

2. some non-human addresses have specific functions (e.g. the NULL address mints the tokens, assigns them to the users and is also used by the users for destroying tokens), therefore excluding them from the analysis only because they are not directly controlled by humans would give us an incomplete view of network phenomena.

Regarding the reliability of the study, only verifiable and repeatable measures were used. Furthermore, the source code for the data acquisition, refinement and analysis is available publicly [3]. However, the network evolves over time so measures' results could vary in the future. Nevertheless, this problem is circumvented by publishing also the data that was analyzed, in this way the same analysis can be repeated and verified.

6 Measures

The measures in this section are split in two categories: measures performed on the owners' network and measures performed on the bipartite network.

6.1 Owners' network

As previously mentioned, the owners' network is the network that originates from NFT trades carried out by Ethereum addresses. This network consists of more than 7,000 nodes and 47,000 edges.
The measures in this section were chosen to conduct an exploratory analysis of the network described above.

6.1.1 Degree distribution

The degree of a node is the number of edges the node has, consequently the degree distribution is a function that describes how degree values are distributed in the network. One of the most interesting aspects of a network is if its degree distribution is scale-free. In scale-free networks, the degree distribution follows a power-law, which is defined as $P_d = Cd^{\alpha}$ (where alpha typically is between 2 and 3).

To verify if the network is scale-free, the Python library `powerlaw` will be used to fit data according to the previous definition and to compare how well a power-law fits the data in relation to other distributions. This comparison process uses the Complementary Cumulative Distribution Function (CCDF) to calculate how the probability for a node to have a degree d evolves across the various distributions, then the log-likelihood ratio is used to compare the distributions and find what approximates better the data.

6.1.2 Average path length

The average path length is defined as the average of all the shortest paths between nodes, it represents the mean number of hops needed to reach every node in the network.
This measure helps us to identify whether users in the network are close to each other or have long paths separating them, in fact, the lower the average shortest path is, the lower is the separation between users.

6.1.3 K-cores

By studying the K-cores one can find out if groups of buyers and sellers that interact mainly between them are present in the network.
A K-core is a subgraph in which every participant has at least K edges with other participants; a higher value of K indicates that a group of nodes has a central position in the network and that members of the core are highly cohesive.

6.1.4 Clustering coefficient

The clustering coefficient is the number of paths of length two that are closed, divided by the total number of paths of length two in the network.
The clustering coefficient value spans between 0 and 1, values closer to 0 mean that neighbours of a node tend to be disconnected, on the contrary values near 1 indicate a very clustered network.

6.1.5 Assortativity

Assortativity is a measure that shows if nodes tend to connect following some sort of preference. This measure is calculated as a Pearson Coefficient in which the covariance of the network is computed on specific node attributes, in this study the degree of the node and the number of rare exchanges made by the node will be used.
Assortativity values range between -1 and 1. Values closer to -1 indicate that we have a negative correlation between the two variables in exam, values closer to 0 indicate that there is no correlation, finally, positive values show a positive correlation.

6.2 Bipartite network NFTs - Owners

In the bipartite network, addresses are linked to their traded NFTs. Contrary to the section 6.1, the following measures weren't chosen to perform an explorative analysis, but to assess whether the node degree is a predictor of rarity or not.

6.2.1 Degree distribution

As for the node degree analysis of the owners' network, a degree analysis of the NFTs nodes in the bipartite network NFTs - owners will be executed. In this way the most traded tokens and how their degrees are distributed in the network could be identified.

6.2.2 Pearson coefficient

The second aspect studied is the one regarding degree as a rarity predictor. In order to assess if degree can be used as a rarity proxy, the Pearson Coefficient between the degree of the node and the rarity of the node will be calculated.
As mentioned previously in the assortativity section the Pearson Coefficient is calculated as the covariance of two variables divided by the product of the standard deviations.

7 Results

7.1 Owners' network

7.1.1 Degree distribution

In figure 1 can be seen the plotting for the degree distribution of nodes in the owners' network.

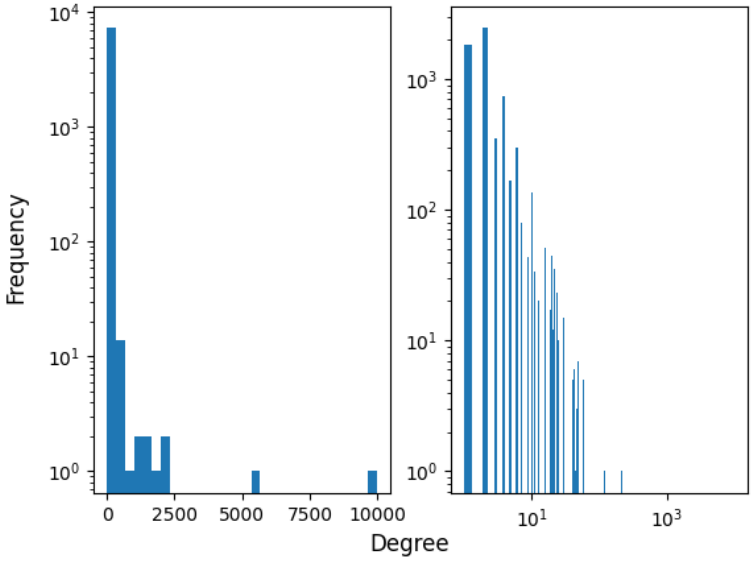

Figure 1: Degree distribution of the owners' network in semilog and log-log scale.

As shown by the first histogram, the majority of nodes in the network have a very small number of edges, while a minority of them have a high degree. It's interesting to note that 80% of the exchanges are performed by the 12% of nodes, this means that the minority of the nodes with high degrees performs the majority of the exchanges.
The histogram was also plotted in log-log scale and its linear decreasing trend seems to be compatible with a power-law distribution, however by using the Python package `powerlaw` different probability distributions were fitted to the data and in figure 2 the CCDF for each distribution is visible.

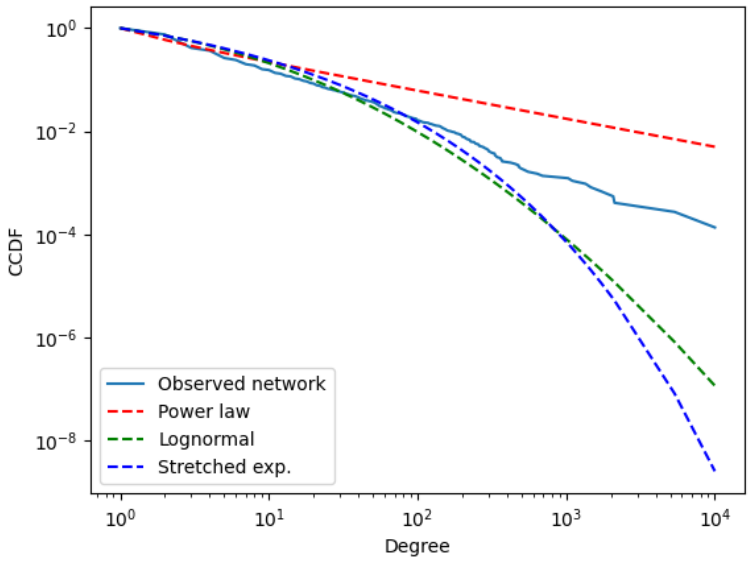

Figure 2: Comparison between CCDFs for different distributions.

Distributions	R
Powerlaw vs. Lognormal	-14.72
Powerlaw vs. Stretched exp.	-11.96
Lognormal vs. Stretched exp.	13.63

Table 2: Log-likelihood ratio for the different distributions. Positive numbers indicate that the distribution on the left approximates data better than the one on the right.

Even if the data observed in the network seems to follow a linearly decreasing trend, as the reference power-law plotted in red, two aspects have to be considered: the estimated alpha for the power-law fitted to the network is 1.54, so its value is out of the usual range that distinguishes power laws, secondly, the log-likelihood ratio reported in table 2 indicates that the best fitting distribution is the stretched exponential.

7.1.2 Average path length

The average path length for the network is equal to 4.1 and considering the number of nodes in the network it can be said that the value obtained is quite low; however, the result is consistent with the characteristics of the observed network, considering that there is no barrier which separates the users (i.e. any Ethereum user is free to trade with every other user) and that there are nodes with a very high degree that shorten the path between other owners. Despite this, some long paths between the users still exist, in fact the diameter of the network is equal to 16.

7.1.3 K-cores

K-cores were evaluated for different values of K, in figure 3 the number of nodes for a core and the corresponding value of K are plotted.
What emerges from the data is that in the network some highly cohesive groups of traders are present and for the maximum degree (K = 14) the subgraph is composed of 121 nodes.

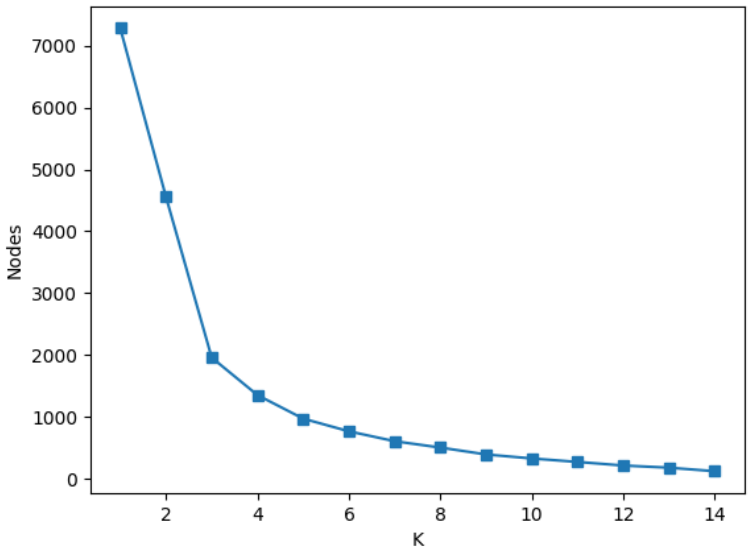

Figure 3: Number of nodes in the K-cores for different values of K.

7.1.4 Clustering coefficient

The clustering coefficient for the network is equal to 0.11, as we said in section 6.1.4 values close to 0 indicate that the neighbours of a node tend not to be connected.
The tendency to form closed triads is typical of networks of friends, colleagues, etc., in the CryptoPunks' network the owners of the tokens do not know each other and they are not tied to geographic zones that help them to meet, consequently closed triads are formed in a purely random way with token exchanges.
The low clustering coefficient value supports also the previous average shortest path result: in highly clustered networks nodes are linked with their neighbours and because of that paths to their neighbours are short, on the contrary, reaching far nodes requires more steps[4]. This situation happens because nodes are linked only to other nodes in their cluster, so many steps are needed to reach nodes of other clusters if edges between different clusters are not present. Nevertheless, this isn't the case of this network, in fact, the CryptoPunks' network is not heavily clustered and because of that any node can reach the other nodes in a few steps.

7.1.5 Assortativity

The assortativity for the network was calculated corresponding to two values: the degree of a node and the number of rare exchanges made by any node. For the first parameter the observed assortativity is -0.07 and for the second is -0.065.
As mentioned in section 6.1.5, assortativity values are between -1 and 1, these values indicate that users tend to prefer doing exchanges with users having a lower degree and with non-rare token owners. This happens because the minority of nodes with high degrees does the majority of the exchanges on the network, but those exchanges are performed with nodes that in most cases have a low degree. Finally, the owners in the network are heterophile to their degree and the rarity of the tokens traded.

Distributions	R
Powerlaw vs. Lognormal	-63.82
Powerlaw vs. Stretched exp.	-65.22
Lognormal vs. Stretched exp.	8.76

Table 3: Log likelihood ratio for the different distributions. Positive numbers indicate that the distribution on the left approximates data better than the one on the right.

7.2 Bipartite network NFTs- Owners

7.2.1 Degree distribution

As can be seen in figure 4 the degree distribution for the NFTs in the bipartite graph is very different when compared to the degree distribution of the owners' network; watching the histogram on the left is evident that the degree distribution for the NFTs has a more linear decreasing trend compared to figure 1. This linearity implies that 52% of the NFTs are involved in the 80% of the exchanges.

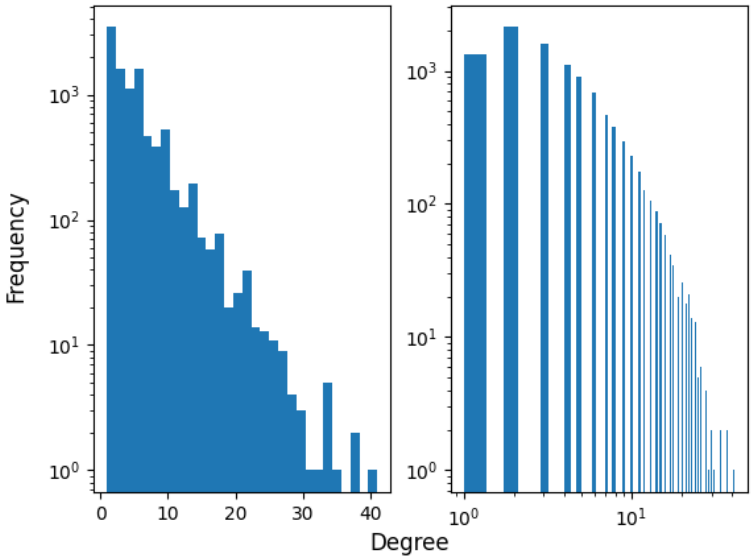

Figure 4: Degree distribution in semilog and log-log scale.

Regarding the degree distribution, the `powerlaw` package estimated an alpha of 1.51 and in figure 5 the CCDFs for the different distributions are shown. From a first visual analysis it is difficult to say whether distribution between lognormal and stretched exponential approximates the data in a better way, nevertheless from log-likelihood ratio analysis (see table 3) it emerges that stretched exponential provides a better approximation.

7.2.2 Pearson coefficient

The main goal of the analysis of the bipartite network is to assess if the degree of an NFT could be used as a proxy for expressing the rarity of a token. From what we can presume, due to their scarcity, rare tokens should be treated as very valuable assets by their owners. For this reason, we expect rare NFTs to have lower degree values when compared to common tokens.

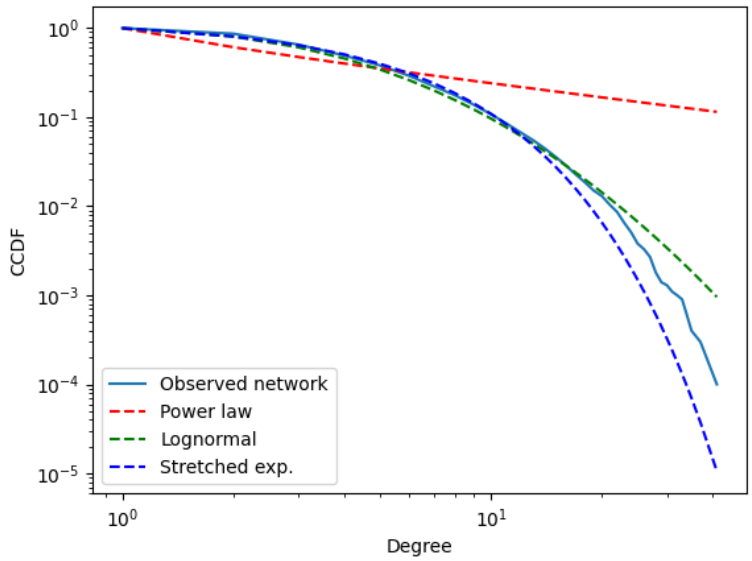

Figure 5: Comparison between CCDFs for different distributions.

The plot 6 shows the rarity level on the x axis and the degree of the token on the y axis. A visual interpretation is that rare tokens ($x \geq 1$) tend to be traded less when compared to common tokens. However, the Pearson Coefficient for the two variables is equal to -0.01, which means that the correlation between the two variables is negligible. The result could be explained by the fact that rare tokens are less traded than common Punks, but there is also a consistent number of common tokens that has the same low degree, as can be seen in the figure. Ultimately the degree of a token cannot be used as a proxy to express its rarity.

8 Critique

The measures performed on data have contributed to clarify how the CryptoPunks' network is structured.
In the present study, it has been proved that some central and very cohesive groups exist in the CryptoPunks' network, even if the entire network is not very clustered. Furthermore, it has been found that big CryptoPunks collectors tend to trade tokens with smaller traders. Finally, the degree of an NFT was found not to be a good indicator of its rarity.
Although the objectives of the study have been fully achieved, additional measures could be calculated to understand the network more extensively; for example, sigma can be used to verify if the network exhibits the small world effect, the eigenvector centrality could be calculated for more central nodes to support the results obtained from assortativity by degree (i.e. we can expect that more central nodes are connected to nodes that have a low degree, therefore they are not very popular).
Furthermore, even if removing externally owned addresses (i.e. addresses not controlled by humans) implicates an incomplete view of network phenomena, it would be useful to do that to evaluate how the network changes and what the impact of these addresses on the community is.

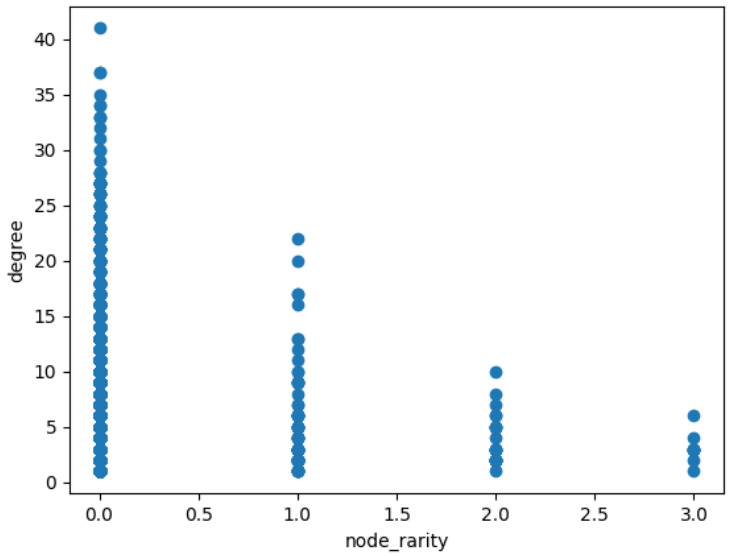

Figure 6: Scatterplot of the NFTs' values for degree and rarity values.

References

[1] Etherscan: APIs, `https://docs.etherscan.io/`.

[2] Gerald Bauer : punks.attributes, `https://github.com/cryptopunksnotdead/punks.attributes/tree/master/original`.

[3] Riccardo Mioli : SNA Exam, `https://github.com/NorwegianGoat/sna_exam`.

[4] S.P. Borgatti, M.G. Everett, J.C. Johnson: Analyzing Social Networks. SAGE, 2018.