Analyzing the Ethereum NFT Market using on-chain data

Authors

- Sophia Arakelyan
- Andrei Dolgolev
- Neeraj Kashyap
- Nana Landau
- Daria Navoloshnikova
- Tim Pechersky
- Yhtyyar Sahatov
- Sergei Sumarokov

Document history

First draft: October 13, 2021

Abstract

We present the Ethereum NFTs dataset, a representation of the activity on the Ethereum non-fungible token (NFT) market between April 1, 2021 and September 25, 2021 constructed purely from on-chain data.

We then consider the following questions:

1. Who owns NFTs on the Ethereum blockchain? Are NFTs for a small number of wealthy elite or are they for the masses?
2. How should one measure the utility of a non-fungible token?

We observe from the data that the distribution of the number of NFTs owned by Ethereum addresses is Zipfian in nature. This indicates that the NFT market is indeed an open market, free for anyone to participate in and with low barriers to entry.

We consider the distribution of owners on a per-contract basis and observe that most NFTs have few owners relative to the number of tokens in their token supply.

We observe that the distribution of owners over the tokens in a contract reflects the utility of that contract. We propose an entropy-based measure of utility for NFT contracts -- their **ownership entropy**.

Introduction

A non-fungible token, or NFT, is a blockchain asset which serves as a distinct representation on the blockchain of some object or concept. It is non-fungible in the sense that it immutably and exclusively represents that object or concept. It will never change what it represents and the represented object or concept admits no *other* representation on the blockchain.

The global market for NFTs has seen a massive boom over the past 3 months. Visual artists and other content creators are digitizing their creations as NFTs to distribute their work to patrons. Game producers are tokenizing assets in computer and mobile games as NFTs to create shared worlds with *other* content creators.

Conventional digital representations of physical works can be replicated aribtrarily and indefinitely. For example, if you own a digital copy of a book, you could in principle make arbitrarily many clones of that copy and distribute it to anyone who asked for it.

In contrast, NFTs naturally reflect the scarcity of the objects they represent. This essential scarcity makes NFTs a perfect tool to globally conserve value when transferring ideas and assets from one digital reality to another. NFTs allow people to create common representations of scarce resources across multiple realities.

Given the recent boom in the NFT market, there is a wild variance in utility across the NFTs that are being released. Similarly, there is a growing number of first-time NFT buyers. This paper analyzes these variances and builds profiles that can be used to classify NFTs and NFT purchasers.

The Ethereum NFTs Dataset

The majority of recent NFT hype has been centered around the Ethereum blockchain. We will therefore restrict our analysis to NFT operations on the Ethereum blockchain.

Contracts, tokens, and events

A non-fungible token is a digital asset representing a distinct idea or physical object.

On the Ethereum blockchain, these tokens are created using [Ethereum smart contracts](#) which represent entire collections of non-fungible tokens. The most famous examples of such contracts are [CryptoPunks](#) and [CryptoKitties](#).

Every time an NFT is transferred on the Ethereum blockchain, it emits a `Transfer` event which is stored on the blockchain. This dataset was constructed by crawling the emitted Transfer events that were emitted in the period of time represented by the dataset.

ERC721

Anyone interested in creating non-fungible tokens is free to implement their tokens in any manner whatsoever *as long as their implementation satisfies the non-fungibility condition*. The most common implementation follows [the Ethereum ERC721 protocol](#).

There is real value in following a standard such as ERC721. There is a growing ecosystem of secondary protocols that only allow NFTs which follow these standards to participate. Secondary markets for NFTs, such as [Nifty Gateway](#) and [OpenSea](#), make the fundamental operational assumption that the tokens they list follow these standards. As a consequence, ERC721-compliant tokens account for the overwhelming majority of Ethereum NFTs.

Technical details

Smart contracts that comply with the ERC721 standard store the following events into the global Ethereum state every time a token is transferred from one party to another:

```
/// @dev This emits when ownership of any NFT changes by any mechanism.
///  This event emits when NFTs are created (`from` == 0) and destroyed
///  (`to` == 0). Exception: during contract creation, any number of NFTs
///  may be created and assigned without emitting Transfer. At the time of
///  any transfer, the approved address for that NFT (if any) is reset to none.
event Transfer(address indexed _from, address indexed _to, uint256 indexed _tokenId);
```

These `Transfer` events can have one of the two following application binary interfaces (ABIs):

```
[
    {
        "anonymous": false,
        "inputs": [
            {"indexed": false, "name": "from", "type": "address"},
            {"indexed": false, "name": "to", "type": "address"},
            {"indexed": false, "name": "tokenId", "type": "uint256"},
        ],
        "name": "Transfer",
        "type": "event"
    },
    {
        "anonymous": false,
        "inputs": [
            {"indexed": true, "name": "from", "type": "address"},
            {"indexed": true, "name": "to", "type": "address"},
            {"indexed": true, "name": "tokenId", "type": "uint256"},
        ],
        "name": "Transfer",
        "type": "event"
    }
```

```
]
```

These ABIs allow us to identify when we encounter an Ethereum transaction which emitted `Transfer` event.

<mark>The core of our dataset is all the `Transfer` events</mark> that we crawled from the Ethereum blockchain between April 1, 2021 and September 25, 2021.

Our dataset is a single [SQLite](#) database. The dataset was crawled using [Moonstream](#), an open source blockchain analytics tool. The code is available on GitHub: https://github.com/bugout-dev/moonstream.

Core relations

We have partitioned the `Transfer` events we crawled into two relations -- `mints` and `transfers`. A mint represents the creation of an NFT and is signified by a `Transfer` event with a zero `from` address. A transfer represents the transfer of ownership of an NFT and is signified by a `Transfer` event with a *non-zero* `from` address.

These tables share the same schema:

```
.schema mints
sqlite> .schema mints
CREATE TABLE mints
    (
        event_id TEXT NOT NULL UNIQUE ON CONFLICT FAIL,
        transaction_hash TEXT,
        block_number INTEGER,
        nft_address TEXT REFERENCES nfts(address),
        token_id TEXT,
        from_address TEXT,
        to_address TEXT,
        transaction_value INTEGER,
        timestamp INTEGER
    );

.schema transfers
sqlite> .schema transfers
CREATE TABLE transfers
    (
        event_id TEXT NOT NULL UNIQUE ON CONFLICT FAIL,
        transaction_hash TEXT,
        block_number INTEGER,
        nft_address TEXT REFERENCES nfts(address),
        token_id TEXT,
        from_address TEXT,
        to_address TEXT,
        transaction_value INTEGER,
        timestamp INTEGER
    );
```

The columns are as follows:

- `event_id` - a unique ID associated with each event, generated when we created the dataset
- `transaction_hash` - the hash of the [Ethereum transaction](#) in which we observed the event
- `block_number` - the number of the [Ethereum transaction block](#) in which the transaction containing the event was mined
- `nft_address` - the address of the smart contract containing the NFT that the event describes
- `token_id` - the identifier that represents the NFT that the event describes within the context of the smart contract with address `nft_address`
- `from_address` - the address that owned the NFT *before* the `Transfer` event denoted by this row in the relation (it is **not** the address that initiated the transaction in which the NFT transfer took place)
- `to_address` - the address that owned the NFT *after* the `Transfer` event denoted by this row in the relation (it is **not** the address that was the recipient the transaction in which the NFT transfer took place)
- `transaction_value` - the amount of WEI that were sent with the transaction in which the `Transfer` event took place
- `timestamp` - the Unix timestamp at which the Ethereum transaction block with number `block_number` was mined into the Ethereum blockchain

There are 6,667,282 `mints` and 4,514,729 `transfers` in our dataset. Together, these events represent the activity of 9,292 NFT contracts which were actively used on the Ethereum blockchain **between April 1, 2021 and September 25,**

2021.

Derived relations

For ease of analysis, we have also included several derived relations in the Ethereum NFTs dataset:

- `nfts` - available metadata (from [Moonstream](#)) about the NFT contracts represented in the dataset
- `current_market_values` - the current (estimated) market value of each NFT, in WEI
- `current_owners` - the current owner of each NFT
- `transfer_statistics_by_address` - transfers in and out of every address that was involved in an NFT transfer
- `transfer_values_quartile_10_distribution_per_address` and `transfer_values_quantile_25_distribution_per_address` - for each 10th or 4th quantile of each NFT collection, these relations give the proportion of the value of the most valuable token in that collection that the quantile represents

Here, "current" means "as of September 25, 2021".

The analysis in this paper will make use of `nfts`, `current_market_values`, `current_owners`, and `transfer_statistics_by_address`.

Caveats

The Ethereum NFTs dataset is constructed purely from events on the Ethereum blockchain. It does not include any data from Layer 2 networks like Polygon. Nor does it include any data from centralized APIs like the OpenSea API. It does not account for events or data from any non-ERC721 smart contracts associated with these platforms on the Ethereum blockchain.

This means that two parties could exchange a positive amount of funds for a transfer off-chain and conduct the transfer on-chain and we would not be able to distinguish the transfer from a gift.

It is also possible for a single transaction to involve multiple NFT transfers. Here is an example of such a transaction involving several NFT layers on top of a Loot token:
https://etherscan.io/tx/0xa578aac6db19c464f69492747fa147985006281a57d77e46316fe09fb406deb2

If a transaction involves multiple NFT transfers and has a non-zero value, it is difficult to understand whether that value is related to the transfers and, if so, how it distributes over the transfers.

For that reason, in the first version of our dataset, the valuation numbers should only be treated as a rough estimate of the actual value for each NFT.

Accessing the dataset

The complete Ethereum NFTs dataset is available on Kaggle under a creative commons license:
https://www.kaggle.com/simiotic/ethereum-nfts

If you would like access to this dataset and it is not convenient to access it from Kaggle, you can discuss it with us on the [Moonstream community Discord](#) or email [Neeraj Kashyap](#).

If you use this dataset in your own research, please use the following citation:

```
Sophia Arakelyan et al., "Ethereum NFTs." Kaggle, 2021, doi: 10.34740/KAGGLE/DSV/2698517.
```

Who is buying NFTs?

Given the hype that exists around NFTs, it is difficult to understand how much of the hype is manufactured and how much of it reflects the situation on the market. Is there a small number of people who each carry significant NFT holdings who are driving the hype around the market and carrying the market with them? Or is the market dominated by many different people each of whom own a relatively small number of NFTs?

For each $n > 0$, let A_n denote the number of addresses that assumed ownership of exactly n NFTs between April 1, 2021 and September 25, 2021.

The histogram below plots A_n against n on a logarithmic scale:
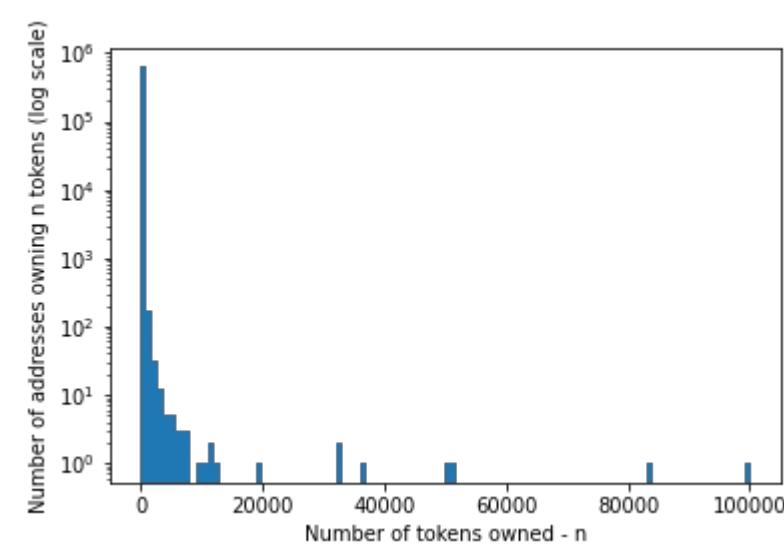
Of course, the NFT owners in the full dataset includes the addresses of smart contracts which act as exchanges and clearinghoues for NFTs and work with thousands and even tens of thousands of NFTs at a time. It also includes the addresses of bots which may not be implemented as smart contracts but which automatically submit transactions based on their triggering logic.

Let us zoom in on the histogram by considering only those addresses which assumed ownership of at most 1500 NFTs between April 1, 2021 and September 25, 2021:
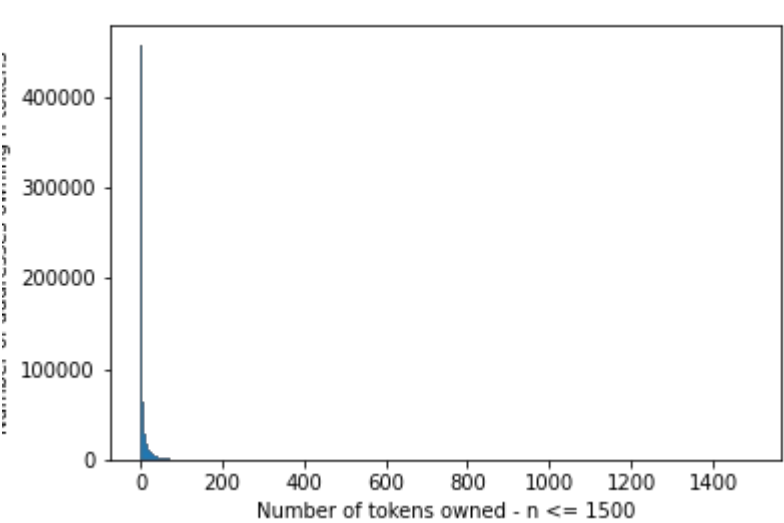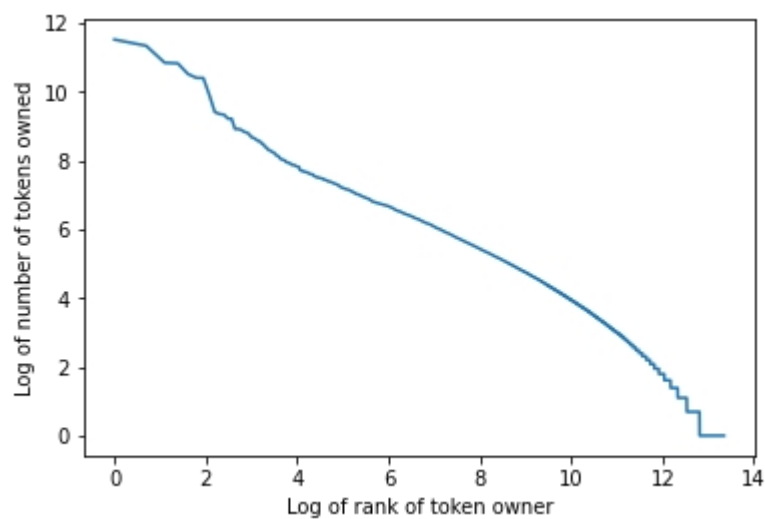

Even this graph is better viewed on a lograthmic scale:
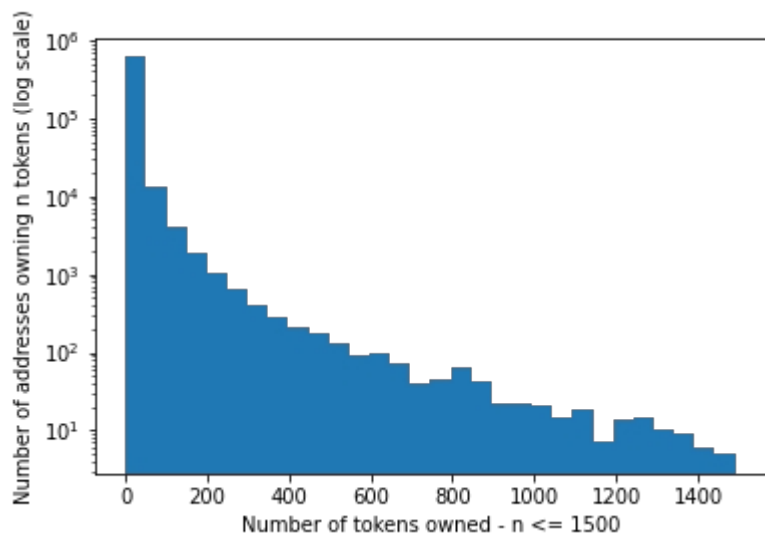
These statistics suggest one additional hypothesis - that the distribution of the number of NFTs per owner follows a [Zipf distribution](#).

This hypothesis is clearly supported by the following plot of the log rank of owners versus the number of NFTs they own:

To put this into hard numbers:

NFTs owned (n)	Number of addresses	Percentage of addresses	Number of tokens owned by addresses	Percentage of tokens owned by addresses
$n \geq 1$	625354	100%	7020950	100%
$1 \leq n \leq 1000$	625107	99.96%	6112780	87.07%
$1 \leq n \leq 100$	615658	98.45%	4036089	57.49%
$1 \leq n \leq 10$	520834	83.29%	1335177	19.02%
$1 \leq n \leq 5$	456399	72.98%	842892	12.01%
$1 \leq n \leq 2$	348948	55.8%	438090	6.24%
$n = 1$	259806	41.55%	259806	3.7%

83.29% of the addresses which assumed ownership of an NFT at the end of the window between April 1, 2021 and September 25, 2021 assumed ownership of only a handful of tokens ($n \leq 10$).

It is possible that there is a small number of people or organizations who are creating a distinct wallet for each NFT they purchase, but doing so at a scale that would our analysis would be technologically and operationally complex enough, and expensive enough, that it is virtually impossible.

What this data shows us is that the Ethereum NFT market is open in the sense the vast majority of its participants are small-time purchasers who likely make their purchases manually. There are few barriers to entry for those who wish to participate in this market.

There is also a great inequality in the Ethereum NFT market in the sense that the top 16.71% of NFT owners control 80.98% of the NFTs. This latter statistic does require a little more nuance in its interpretation, however, as many of those owners are marketplaces and clearinghouses like OpenSea, Nift Gateway, and other platforms of the same ilk. We plan to expand on this analysis in a future report.

The utility of an NFT

People buy NFTs for different reasons. Some buyers may purchase an NFT to support their favorite artists or communities. Others may prefer to purchase NFTs that bring them extrinsic utility. A good example of this kind of utility is the [Ethereum Name Service](), which allows anyone to create a human-friendly name (such as `vitalik.eth`) associated with the Ethereum addresses and more. The associations are represented as NFTs on the ENS contract, and many services (e.g. Coinbase, Metamask, etc.) support resolution of ENS names as part of transfers and other blockchain operations.

The question is, how do we distinguish between NFTs that represent intrinsic and perhaps subjective value such as artwork by someone's favorite artist and those which have clear extrinsic utility such as the Ethereum Name Service or governance tokens for decentralized protocols?

We propose that this kind of extrinsic utility should be measured at the level of an NFT collection rather than the individual tokens. For tokens that do represent extrinsic utility for a large pool of users, any other tokens deployed as part of the same contract are also likely to represent similar utility. This is in contrast to collections like CryptoKitties, in which the extrinsic utility of the kitties is unknown and limited, and the most expensive kitties dwarf the cheaper ones in value by multiple orders of magnitude.

Let us consider a few different possible statistics that could act as measures of extrinsic utility for NFT contracts. We will start by considering statistics that perform poorly as measures of utility and address their faults to arrive at good candidates.

Maximum token value

We could attempt to use the maximum value of a token in an NFT contract as a measure of its utility. However, this statistic does not capture distributional information about other tokens in the same contract. We would not be able to use this statistic to understand if all the tokens in the same contract seemed to have similar utility. This makes it a poor statistic for the purposes of measuring external utility.

Distribution of value over the tokens in a collection

This statistic has an advantage over the maximum token value in that it encodes information about all the tokens in a collection.

It suffers from three problems:

1. It is not a scalar statistic. We would need to calculate several moments of the token value distribution over the collection in order to capture all the information it contains, and this could make it awkward to work with.

2. It requires us to estimate the value of the tokens in a contract. The estimation of value from on-chain data is difficult because people are not required to exchange monetary value on the blockchain. It would be a simple matter for two parties to exchange money off-chain and then exchange their NFTs on chain.

This second problem is a practically insurmountable obstacle to the use of any statistic based on the distribution of values over the tokens in an NFT contract.

Distribution of number of transfers over the tokens in a collection

This statistic has an advantage over the previous statistics in that it doesn't require us to estimate the value of NFTs in a collection.

Like the previous candidate, this too is not a scalar statistic. It would require a great deal of care in analysis in interpretation to use this distribution of transfers to draw conclusions about the utility of an NFT contract from this distribution.

Perhaps a more serious concern is that both the value of tokens and the number of times they are transferred is dependent on the particular form of their extrinsic utility. One can imagine use cases in which tokens derive utility through being transferred or through being volatile in value, and other use cases in which tokens derive utility through being held or through being stable in value.

Because the *form* of utility could have such a drastic effect on this candidate and the previous one, neither is an ideal candidate for a measure of utility. Our measure of utility should be *independent* of the form of the utility. We cannot predict *how* people will derive utility from their NFTs in the future, but we would like to be aware of when they start deriving it!

Distribution of ownership over the tokens in a collection

And now we narrow in on an invariant. We discussed why it is sensible to associate extrinsic utility with an NFT contract - a full collection of NFTs - rather than with the individual tokens. Because extrinsic utility applies to large populations of users rather than to a small number of individuals.

This means that, if an NFT collection has extrinsic utility, then it should have many distinct owners relative to its number of tokens.

Suppose that a few parties strike out to purchase most of the tokens. Then the tokens would gain monetary value, and would become good vehicles for investment. But this represents a gain in *intrinsic* utility and a *reduction* in extrinsic utility. So we see that the dynamics whereby extrinsic utility is traded off for intrinsic utility correspond to a an increased concentration of ownership among a few addresses as compared to a dispersion of ownership across many addresses.

The level of dispersion of ownership across the tokens in an NFT contract is invariant to the particular form that the external utility of the tokens takes. If the form involves many transfers, for example, it still doesn't significantly affect the dispersion at any single point in time.

This notion of dispersion of ownership is an invariant of the NFT contract under different forms of extrinsic utility which nonetheless captures how attractive NFTs in that contract are to the general Ethereum community.

The notion of [information theoretic entropy](#) formalizes this concept of dispersion. We propose a statistic called **ownership entropy**, defined below, as a measure of the external utility of the tokens in an NFT contract.

Ownership entropy

Let π be a probability distribution on the sample space ${1, 2, \ldots, n}$ for $n \geq 1$. Denote $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$, where π_j is the probability associated with the event j.

Then the entropy of π is defined as: $$H(\pi) = \sum_{j=1}^n -\pi_j \log(\pi_j).$$

(Here, $\log$ represents the logarithm for base 2, although the constant is not so important.)

The entropy $H(\pi)$ is maximized for the distribution π which assigns equal probability to all its outcomes. In fact, from [Jensen's inequality](#),

$$H(\pi_1, \ldots, \pi_n) \leq \log(n),$$

with the maximum achieved if and only if $\pi_1 = \pi_2 = \ldots = \pi_n = \frac{1}{n}$.

$H(\pi)$ is an information theoretic measurement of how well distributed the probability mass of π is over its sample set, and is maximized when the probability mass is evenly distributed. The units of entropy are *bits* (as in binary digits).

This makes it a natural candidate to measure the dispersion of ownership over the tokens of an NFT contract.

For an NFT contract C, let T denote the set of tokens (represented by their token IDs) present in C. For each token $t \in T$, let A_t denote the address owning that token. It is possible for A_t to be *any* Ethereum address,

including the 0 address.

We can think of C as a probability distribution over its tokens whereby we select each token t with probability $\frac{1}{|T|}$. This induces a probability distribution π_C on the set $\mathcal{A}$ of all generated Ethereum addresses whereby, for any address $A \in \mathcal{A}$, the probability of π_C selecting that address is: $$\pi_{C,A} = \frac{|\{t \in T : A_t = A\}|}{|T|}.$$

We define the **ownership entropy** of C to be the entropy $H(\pi_C)$ of this probability distribution that C induces on $\mathcal{A}$.

Ownership entropy on the Ethereum blockchain

Below is a histogram of contracts in the Ethereum NFTs dataset by their ownership entropies:
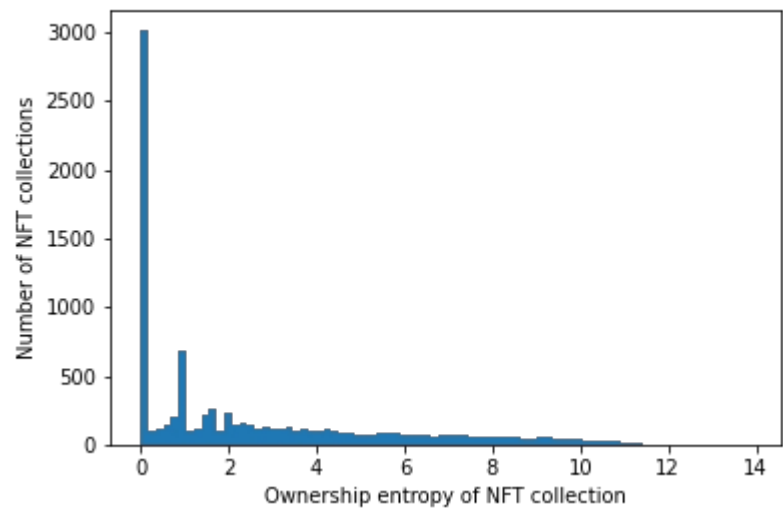
NFTs which were active between April 1, 2021 and September 25, 2021 exhibit ownership entropies ranging from 0 (all tokens owned by a single address) to 13.864019.

The contracts with the highest ownership entropy are:

NFT contract address	Ownership entropy	What is this contract?
0x57f1887a8BF19b14fC0dF6Fd9B2acc9Af147eA85	13.831032	Ethereum Name Service
0x60F80121C31A0d46B5279700f9DF786054aa5eE5	13.864019	The Rarible governance token (details)
0xC36442b4a4522E871399CD717aBDD847Ab11FE88	13.742724	The Uniswap v3 position NFT, representing non-fungible liquidity positions
0xabc207502EA88D9BCa29B95Cd2EeE5F0d7936418	13.714889	Badges for Yield Guild Games
0x5537d90A4A2DC9d9b37BAb49B490cF67D4C54E91	13.285761	OneDayPunk

Interestingly, the Rarible and Yield Guild Games tokens show high entropy because the companies behind the tokens have invested a substantial amount into airdropping their tokens to many different addresses on the Ethereum blockchain. Given the level of investment into these tokens and how recently they have been dropped, we will have to see in the next installment of this report whether or not they realized their potential utility. We will do this through an analysis of the information gain (measured via ownership entropy) of the smart relevant smart contracts.

The contracts with zero entropy represent contracts that were recently minted but either have not launched or have stagnated in activity.

More interesting are the contracts with low but non-zero entropy:

NFT contract address	Ownership entropy	What is this contract?
0x08CdCF9ba0a4b5667F5A59B78B60FbEFb145e64c	2.004886	World Cup Token
0xA4fF6019f9DBbb4bCC61Fa8Bd5C39F36ee4eB164	2.003856	instigators

NFT contract address	Ownership entropy	What is this contract?
0xB66c7Ca15Af1f357C57294BAf730ABc77FF94940	2.003756	Gems of Awareness Benefit
0x5f98B87fb68f7Bb6F3a60BD6f0917723365444C1	2.002227	SHADYCON (associated with Eminem)
0x374DBF0dF7aBc89C2bA776F003E725177Cb35750	2.001823	WyldFrogz

All these NFTs have very low rates of adoption and either the marketing efforts behind them have stalled or they have not yet started gaining any momentum.

Finally, let us look in the middle of the range:

NFT contract address	Ownership entropy	What is this contract?
0x0ae3c3A1504E41a6877De1B854C000EC64894bEa	6.021144	Circleorzo, a collection of images of procedurally generated circles.
0x1ECA43C93D8e06FB91489818B4967014D748Da53	6.017002	Cowboy Punks, which appeal to NFT owners that prefer westerns to cyberpunk
0xc57605Bef27ef91DbECc839e71E49574b98857Fc	6.011324	The Enigma Project, tokens that control access to puzzle games
0xd3f69F10532457D35188895fEaA4C20B730EDe88	6.010405	RTFKT Capsule Space Drop (independent-ish blog post)
0xba61aEF92ebF174DbB39C97Dd29D0F2bd3D83d33	6.009679	Dommies

From this analysis of the ownership entropy spectrum, we see that the ownership entropy does indeed seem to capture the external utility of NFT contracts. The cases of Ethereum Name Service and the Uniswap position NFT are particularly interesting because holders realize the value of those NFTs in very different ways - an address is much more likely to hold onto an ENS token and much more likely to trade their liquidity position on Uniswap v3. Despite the differences between these two contracts in the *form* of their utility, they both rise to the top when we consider their ownership entropies.

The large scale airdrops - the Rarible governance token and the Yield Guild Games Badges - are concerning because although a lot of money has been put into their distribution, it remains to be seen whether or not the market realizes any extrinsic utility from these tokens. This is something we will track closely for the next installment of this report.

We also see that, at the lower ranges, the ownership entropy serves to measure just plain adoption event for tokens which have no extrinsic utility (like Cowboy Punks and Dommies versus World Cup Token and instigators).

This analysis also highlights the importance of considering ownership entropy as a time series, and of tracking the differences in ownership entropy over time to reflect changing market perceptions of NFTs.

Finally, all the high entropy contracts contain very detailed explanations of utility to for their users, in the form of papers and documentation. This is something we do not see lower down in the the spectrum. This care that is put into documentation for external and fresh users is a clear sign of extrinsic utility.

Conclusions

Our analysis paints a picture of the Ethereum NFT market as an open and free market which exhibits the same kinds of wealth disparities as conventional markets.

It also shows the viability of ownership entropy as a means of identifying NFT contracts that actually contain extrinsic utility. This is not of much value for NFT investors, who are typically making bets about the *intrinsic value* of NFTs. However, regulators are more and more concerned with establishing the level of utility of tokens on the Ethereum blockchain and the ownership entropy is a tool perfectly suited to this purpose.

In future versions of this report, we plan to:

1. Conduct analysis of the openness of the Ethereum market over time. Rather than only considering data in a single window of time (in the case of this report, April 1, 2021 to September 25, 2021), we will consider the time series of the same statistics generated at frequent intervals from 2016 until the time of publication of the report.
2. Expand the analysis of onwership entropy into an analysis of ownership information gain - the change in ownership entry over time.

3. Enrich our dataset and our analyses with information about the addresses which *funded* NFT transfers.
4. Enrich our dataset and our analyses with side information about NFT valuations from centralized sources (like the OpenSea API).

Collaboration

The calculations and analysis for this report are available as a Kaggle notebook: https://www.kaggle.com/simiotic/ethereum-nft-analysis

All the software we build is free software, and all the data we produce is available under a creative commons license.

If you have burning questions about NFTs or Ethereum smart contract activity in general, and would like to collaborate with us to find answers, join us on Discord: https://discord.gg/K56VNUQGvA