data

MDPI

Data Descriptor

Dataset of Linkability Networks of Ethereum Accounts Involved in NFT Trading of Top 15 NFT Collections

Aleksandar Tošić [1,2], Niki Hrovatin [1,2] and Jernej Vičič [1,3,*]

[1] Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Glagoljaška 8, 6000 Koper, Slovenia; aleksandar.tosic@upr.si (A.T.)
[2] InnoRenew CoE, Livade 6a, 6310 Izola, Slovenia
[3] Research Centre of the Slovenian Academy of Sciences and Arts, The Fran Ramovš Institute, Novi Trg 2, 1000 Ljubljana, Slovenia
* Correspondence: jernej.vicic@upr.si

Abstract: In this paper, we present subgraphs of Ethereum wallets involved in NFT trades of the top 15 ERC721 NFT collections. To obtain the subgraphs, we have extracted the Ethereum transaction graph from a live Ethereum node and filtered out exchanges, mining pools, and smart contracts. For each of the selected collections, we identified the set of accounts involved in NFT trading, which we used to perform a breadth-first search in the Ethereum transaction graph to obtain a subgraph. These subgraphs can offer insight into the linkability of accounts participating in NFT trading on the Ethereum blockchain.

Dataset: Zenodo: Dataset of Linkability Networks of Ethereum Accounts Involved in NFT trading of Top 15 NFT Collections. Data identification number: DOI: 10.5281/zenodo.8017995. Direct URL to data: https://doi.org/10.5281/zenodo.8017995.

Keywords: NFT; Ethereum network; linkability; wash trade

Citation: Tošić, A.; Hrovatin, N.; Vičič, J. Dataset of Linkability Networks of Ethereum Accounts Involved in NFT Trading of Top 15 NFT Collections. *Data* **2023**, *8*, 116. https://doi.org/10.3390/data8070116

Academic Editors: Bijan Raahemi and Waeal J. Obidallah

Received: 31 March 2023
Revised: 9 June 2023
Accepted: 23 June 2023
Published: 28 June 2023

1. Summary

The explosion of interest in Non-Fungible Tokens (NFTs) emerged predominantly in 2021 and 2022, becoming a novel form of asset ownership and trading within the Ethereum ecosystem. Their uniqueness and digital provenance capability have significant implications, serving as the backbone for an array of innovations ranging from digital art to virtual real estate, and thereby becoming a crucial element of the Ethereum blockchain economy. The emergence of NFTs has sparked a wave of innovation, leading to the development of novel business models that leverage this transformative technology as an enabler, revolutionizing the way we create, distribute, and monetize digital assets [1]. The NFT scene and technology is quite new, so the typology of NFT scams is still evolving. Some authors, such as Kshetri [2] and Das et al. [3], concentrate mostly on direct attacks to either buyers or content creators using social engineering (phishing, tricking, false identities ...) or technological scams (exploiting security flaws, malware, and hacking). On the other hand, some authors delve into non-direct scamming in the form of wash trading or pump-and-dump schemes [4]. A more recent study analyzes whether or not wash trading is worth it by analyzing the trading data [5]. However, only NFT trades are considered, with no linkage to the underlying Ethereum transaction graph. The objective of the paper is to publish data that would aid the efforts of fellow scholars in identifying possible wash trading, fraud detection, and general network analysis of large networks. The emerging NFT market, encompassing digital assets, trading platforms, and underlying blockchain networks, is significantly susceptible to fraud. The threat actors perpetrating

fraudulent activities span a broad spectrum, encompassing individual creators, colluding buyers and sellers, and potentially larger, organized syndicates. These actors employ a variety of deceptive techniques, including fraudulent transactions, price manipulation, and the use of multiple blockchain accounts to conceal their activities. These methods exploit the unique vulnerabilities inherent to this decentralized digital ecosystem. The pseudonymous nature of blockchain transactions, the present regulatory vacuum, and the overall opacity of market activities together create an environment ripe for fraudulent exploitation. Understanding and addressing these vulnerabilities is essential to ensure the transparency, trustworthiness, and overall longevity of the NFT marketplace.

The analysis of large-scale network data indeed poses significant computational challenges. In order to address this complexity and the vast amount of data, we have created a specialized representation of Ethereum transactions that we refer to as 'linkability networks'. These networks are derived through a methodical extraction and processing of data from the Ethereum transaction graph and NFT trading graphs of popular NFT collections. This abstracted representation is designed to highlight key relationships and transaction patterns within the context of these collections, thereby making the data more manageable and conducive to further academic analysis.

This dataset enables comprehensive network analysis of NFT markets. Researchers can apply tools such as centrality measures to identify key players, and community detection algorithms to uncover clusters of interconnected accounts. Anomaly detection techniques can be used to flag potential fraudulent activities, while PageRank and other similar algorithms can help rank Ethereum accounts based on their importance within the network. Overall, these data offer a rich resource for exploring the structure, behavior, and potential vulnerabilities of NFT markets.

2. Data Description

The main graph data are organized in standardized graph description format—DOT. DOT is a graph description language. It is based on standard ASCII text format. Various programs can process DOT files, such as dot, neato, twopi, circo, fdp, sfdp, gvpr, gc, acyclic, ccomps, sccmap, and tred. The data are organized in 32 files:

- collection_metadata.txt stores basic information about each collection that was analyzed. The graph data of the collection is stored in a file named as the nickname of the collection (slug column in collection_metadata.txt). The most important columns/properties are: rank, slug—nickname, creation date, address, volume data . . . Collection rank, slug, and some statistics are shown in Table 1.
- 15 files reporting NFT ownership transfer; the names correspond to the nicknames of the collections in collection_metadata.txt. The most important columns/properties are: from/to—transfer of NFT ownership between two addresses; token_id—the identifier of the specific NFT; currency—ETH, WETH, NULL (the NULL placeholder is used if the NFT ownership transfer is not coupled to a value transfer or the value transfer occurred using exotic ERC20 tokens); the amount column gives the value transfer, which is always 0 for the NULL placeholder in the currency column.
- 15 files with graph data; the names correspond to the nicknames of the collections in collection_metadata.txt. Each file gives the address linkage graph of one of the top 15 collections according to monetary volume on the Opensea [1] marketplace, computed as presented in methodology, exclusively on the Ethereum blockchain. Figure 1 shows a visualization of an address linkage graph. The address linkage graph shows ties among externally owned accounts involved in NFT trading throughout the whole lifetime of the Ethereum network. An address linkage graph was built for each NFT collection. The graph is directed with weighted edges and is given in the DOT [2] format. The vertices of the graph are Ethereum accounts that owned NFTs of a selected collection. Each edge of the graph is the indicator of a transfer of the ETH token between the two vertices concerning the edge direction. The edge weight gives the path length of the value transfer. For instance, weight = 0 indicates a direct transfer

of the ETH token between the two vertices, and weight = 1 indicates that the source vertex moved ETH tokens to an intermediate address, which in turn moved the value to the target vertex of the edge. Due to the DOT format imposing that graph vertices' names must start with a letter, Ethereum addresses are stored without the leading 0.

- gathering_data.txt: some basic information about the construction process for each address linkage graph is stored here.

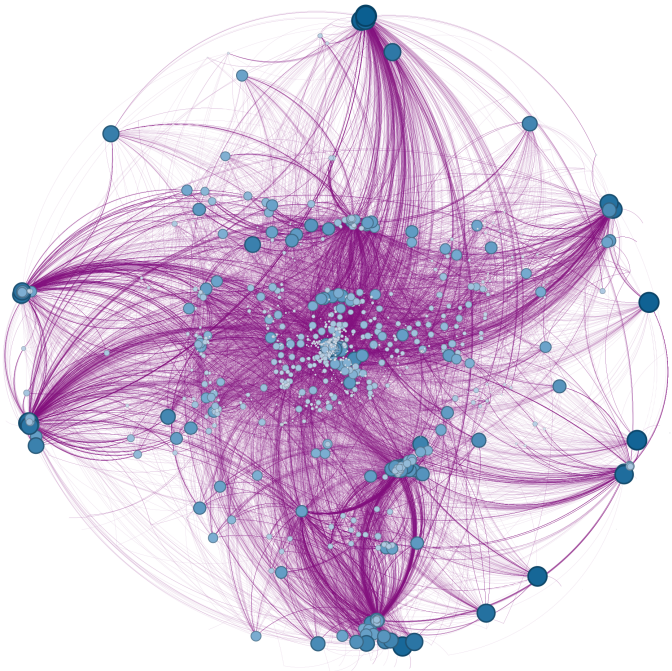

Figure 1. Visualization of the linkage graph of the VeerFriends NFT collection. The visualization consists of 2500 vertices that were selected at random from the linkage graph; nodes with degree 0 were removed. Vertices in the visualization are Ethereum accounts, and edges indicate a transfer of the ETH token between two vertices. The size color of the vertices changes with the in and out degrees (darker is larger).

Table 1. The selected NFT collections are displayed as ranked by the Opensea platform. The table shows statistics from the file collection_metadata.txt and the size of the address linkage graph per NFT collection.

Rank	Slug	Statistics				Linkability Networks	
		Total Sales	Total Supply	Market Cap (USD M)	Floor Price (USD k)	Vertexes	Edges
1	boredapeyachtclub	27,538	9998	856	82.04	19,738	23,909,149
2	mutant-ape-yacht-club	36,156	19,430	354	17.43	36,000	83,452,456
3	azuki	28,146	10,000	179	15.31	18,081	26,794,525
4	clonex	20,360	19,442	170	7.73	18,857	18,983,123
5	proof-moonbirds	10,024	10,000	111	10.43	16,575	16,989,401
6	doodles-official	24,459	10,000	89	8.32	18,613	29,169,501
7	meebits	23,841	19,999	75	3.26	17,137	13,364,581
8	cool-cats-nft	31,240	9955	30	3.39	16,768	27,941,486
9	bored-ape-kennel-club	23,697	9602	73	7.62	17,986	21,790,227
10	world-of-women-nft	25,320	10,000	26	2.29	14,571	12,976,528
11	cryptoadz-by-gremplin	22,346	7013	12	1.5	13,509	19,798,808
12	beanzofficial	23,077	19,950	27	1.19	10,213	5,655,886
13	pudgypenguins	39,350	8888	52	5.58	15,734	26,234,392
14	veefriends	5163	10,255	77	6.32	9,240	2,321,991
15	hapeprime	12,723	8192	4	0.5	14,768	8,737,874

Collection metadata obtained from the Opensea platform on 17 December 2022.
USD values were obtained by converting the historical value of the ETH token on 17 December 2022.

2.1. Linkability Network

The term 'Linkability Network' has been purposefully chosen to reflect the objective and methodology of our study. In the context of blockchains, accounts act as pseudonyms, with a single user capable of holding multiple accounts. Direct p2p transactions of native blockchain currency usually imply a relationship between the two accounts involved, therefore indicating that the two accounts—and by extension, the users managing these accounts—are related in some way. According to Pfitzmann and Hansen's comprehensive terminology work on data minimization and privacy [6], 'linkability' refers to the capacity to sufficiently distinguish whether two or more items of interest are related within the system. In our scenario, the items of interest are the blockchain accounts. Consequently, the 'Linkability Network' effectively illustrates the potential connections or 'links' between accounts, as inferred from transaction data.

In this section, we first define the Ethereum Transaction Graph as a foundation, followed by a definition of the Linkability Network derived from it.

The **Ethereum Transaction Graph** is a directed graph, denoted as $G = (V, E)$, where:

- V is the set of all unique Ethereum accounts, each account being a vertex in the graph.
- E is the set of all directed edges, where each edge $(v, u) \in E'$ represents a transaction of ETH currency from account v to account u.

In this graph, the direction of an edge signifies the direction of the transaction, from the sender to the receiver. The Ethereum Transaction Graph represents all ETH transactions that have ever occurred, providing a comprehensive view of the Ethereum network's transaction history.

Subsequently, the **Linkability Network** is a directed, weighted graph, denoted as $G' = (V', E', w)$, derived from the Ethereum Transaction Graph, where:

- $V' \subseteq V$ is the set of vertices, each representing a unique account that has owned an NFT from a specific collection at any point in time.
- E' is the set of directed edges, with each edge $(v, u) \in E'$ representing the shortest path of ETH currency transfer from u to v in the Ethereum Transaction Graph.
- $w : E' \to 1, 2, \ldots, d$ is a weight function that assigns to each edge (v, u) a weight equal to the length of the shortest path from u to v in the Ethereum Transaction Graph.

This concept of a Linkability Network allows for a structured and analyzable representation of the complex interactions within the Ethereum blockchain. Leveraging such a graph-based model can provide insightful revelations about the transaction patterns, enabling the detection and understanding of potential illicit activities, such as wash trading, in the NFT marketplace.

2.2. Value of the Data

These data are useful for discovering suspicious activity by analyzing provided data, scoring potentially malicious wallets, and detecting communities of collaborating accounts. The data can be further used for econometrics. These are the identified potential users of the data:

- Fellow scholars in the field of econometrics;
- Regulators and policy-makers in the field of NFTs;
- The whole decentralized finance community;
- Entrepreneurs dealing with NFT investments;
- Artists and content creators participating in digital content securing with NFTs;
- The entire NFT community.

The data are presented in a standardized, human-readable form. The linkability subgraphs can be further analyzed using tailored algorithms. The data are also useful for analyzing market growth, trend analysis, and projections. The methodology is further presented in Section 3. Visualizing the linkage subgraph in Figure 1, we can observe a high degree of cluster formation, which is a potential point of interest for further analysis.

See Figure 2 for a visualization of the linkability network's ability to highlight potentially suspicious activity within NFT ownership transfers in the Bored Ape Yacht Club collection.

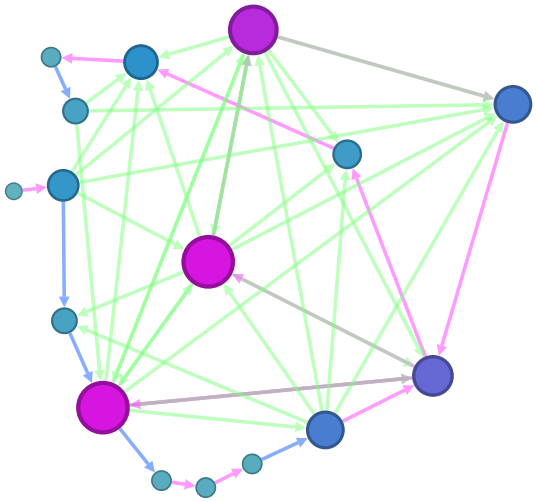

Figure 2. The figure shows the union of the linkability network and NFT ownership transfers for token ID 9202 from the Bored Ape Yacht Club collection. Each node in this graph represents an Ethereum account. Red edges are NFT ownership transfers, with blue edges representing NFT ownership transfers that occurred on the Opensea platform when selling the NFT. Green edges are from the linkability network, showing that certain addresses are interconnected, potentially indicating suspicious activity (the nodes of the same colour).

3. Methods

Data were acquired by observing the top placed NFT collections on OpenSea, which is the biggest NFT marketplace. Each NFT from the selected top collections according to monetary volume was analyzed in the Ethereum network graph, and all wallets involved in the transactions were collected. An address linkage graph was constructed for each collection.

3.1. NFT Transaction Graph

The reasoning behind the selection of the collections, the number of collections, and the selection of the blockchain is as follows:

- The Ethereum network was selected as the most popular NFT enabled blockchain;
- OpenSea was selected as the biggest NFT marketplace;
- Top collections by monetary value were selected as most of the funds are being transferred in these collections;
- 15 top collections with less than 20,000 NFTs were selected as the biggest still manageable number of NFT collections;
- Only the PFP (Profile Picture tokens) NFT category was considered, due to the high popularity of this category within the blockchain community;
- We limit our research to ERC721 tokens;
- Only collections created before June 2022 were considered (6 months before the data acquisition).

We used the opensea-scraper [3] npm package to obtain the top NFT collections' metadata from the Opensea platform. The data were filtered according to the abovementioned criteria and are given in the collection_metadata.txt file. The collection_metadata.txt was captured on 17 December 2022. The collection_metadata.txt file was then used to collect transaction data from individual NFT collections. Specifically, the address row that gives the Ethereum address of the NFT smart contract was used in a Python script for scraping data from the Etherscan [4] platform. The Etherscan platform provides APIs to access the

Ethereum blockchain data conveniently. We used the logs module of the Etherscan APIs to obtain transfers of NFT ownership by specifying the collection address, the method signature of the NFT ownership transfer, and querying the endpoint from the NFT contract creation date, 1000 Ethereum blocks at a time until 1 May 2022. The obtained transfers were integrated with the value transfer of an eventual completed auction coupled to the NFT transfer. The value transfer data was obtained using the proxy module of the Etherscan APIs, specifying the action eth_getTransactionReceipt and the transaction hash of the transaction sealing the NFT transfer of ownership. The described query returns the transaction receipt of the transaction specified by the transaction hash, and each transaction was then processed in search of the orders matched method signature of the Opensea smart contract closing the auction and carrying the information of the winning bid. We captured the value transfer of winning bids in the native token and the wrapped Ethereum token.

3.2. Ethereum Transaction Graph

The Ethereum transaction graph was obtained using the following steps:

1. We synced a go-ethereum [5] full node to the Ethereum mainnet, and we ran the ethereum-etl [6] tool for extracting all Ethereum transactions. The extraction process took nine days, producing 980 GB of output data on a dedicated server [7].
2. The large output file was then filtered using grep to keep only transactions having input 0x (Ethereum walled prefix) and value > 0. Therefore, we filtered out all smart contract transactions and value transfers with 0 amount. The filtering took 22 h, resulting in a file of 188 GB.
3. The resulting file was further filtered to remove all transactions involving addresses of the top 30 exchanges and well-known mining pools. The address list employed for the filtering process was incorporated into the data repository, accessible under the name blacklist.json.

Data size and computation times are presented in Table 2.

Table 2. Output size and elapsed time of operations performed to obtain the address linkage graphs.

Operation	Data Output Size	Time
NFT collection scraping	260 MB	13 days
Ethereum ETL	988 GB	9 days
smart contract filtering	188 GB	22 h
address linkage graphs generation	8.8 GB	27 h

3.3. Address Linkage Graph Construction

Address linkage graphs were constructed with a java program using the JGraphT [8] library. First, the Ethereum transaction graph was read in memory as a directed graph with respect to the ETH token transfer direction, and addresses owning NFTs of a selected collection were loaded. Then, each NFT owner address was used as a root vertex for a Breadth First Search (BFS) operating on the Ethereum transaction graph. The specific implementation of the BFS algorithm, including the setup to reach the desired maximal depth and to memorize visited vertices that are addresses owning NFTs of the selected collection, is outlined in Algorithm 1. The set of memorized vertices was then used to link edges in the address linkage graph from the BFS root vertex to each of the memorized vertices, and the visit depth was included as the edge weight. The address linkage graph was exported in the DOT format as a directed weighted graph. The procedure was repeated for each selected NFT collection.

Algorithm 1 Parallel BFS for Linkability Network Generation.

Input: Ethereum transaction graph G, Set of NFT collection owner addresses A, Max depth D
Output: Linkability Network LN
foreach *address a in A* **in parallel do**
  Initialize empty set *Visited* and add a to *Visited*;
  Initialize queue Q and enqueue a and enqueue *null*;
  Initialize empty graph G_a;
  $depth = 0$;
  while Q *is not empty* **do**
    v = dequeue Q;
    if $v == null$ **then**
      $depth++$;
      Q enqueue *null*;
      if $peek(Q) == null$ **then**
        break;
      end
      continue;
    end
    if $depth < D$ **then**
      foreach *neighbor u of v in G* **do**
        if *u not in Visited* **then**
          Enqueue u in Q;
          Add v to *Visited*;
          if *address v in A* **then**
            Add edge $e(a, v, depth)$ to G_a;
          end
        end
      end
    end
  end
  Synchronize: Merge G_a to LN;
end

3.4. Conclusions, Limitations and Future Work

This paper presents a dataset of connections among Ethereum blockchain accounts that were involved in Non-Fungible Token (NFT) trading. NFTs are unique identifiers recorded in blockchain that are used to certify the ownership and authenticity of tangible and digital assets. The datasets are collected by searching and storing all transactions and accounts connected with smart contracts that define targeted NFTs. The data can be used to analyze the behavior of artists, buyers and holders of the NFTs, to discover possible anomalies, and to monitor and possibly gain knowledge for future predictions on similar networks. The depth of the search was limited to a length of four because of time and space constraints. Future work will involve expanding the dataset to more than the first 15 most known collections, and also updating the dataset with new time-windows (adding new transactions and accounts).

Author Contributions: A.T.: Conceptualization, Software, Investigation, Supervision, Writing—Review and Editing. N.H.: Conceptualization, Software, Investigation, Writing—Review and Editing. J.V.: Conceptualization, Writing—Review and Editing, Data Curation. All authors have read and agreed to the published version of the manuscript.

Funding: Funding for this research was provided by European Commission through the Horizon 2020 'InnoRenew CoE' (Grant Agreement no. 739574) and SRC-EDIH—Smart, Resilient, and Sustainable Communities—European Digital Innovation Hub (Grant Agreement no. 101083351).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: The data-gathering process did not involve the use of human subjects. The only possibly problematic ethical aspect comes from the possible privacy concerns of the gathered corpus and consequently the extracted word lists. All the gathered texts are from public school materials that were priory checked and recognized as without any ethical concerns.

Data Availability Statement: Repository name: Zenodo. Data identification number: DOI: 10.5281/zenodo.8017995. Direct URL to data: https://doi.org/10.5281/zenodo.8017995. Unpack the zip files, auxiliary files are added in non-compressed form.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Notes

1 Opensea NFT marketplace: https://opensea.io/ (accessed on 16 May 2023)
2 DOT tutorial and specification: https://www.graphviz.org/documentation/ (accessed on 16 May 2023)
3 OpenSea Scraper: https://github.com/dcts/opensea-scraper (accessed on 16 May 2023)
4 Etherscan platform: https://etherscan.io/ (accessed on 16 May 2023)
5 Go-Ethereum: https://geth.ethereum.org/ (accessed on 16 May 2023)
6 Ethereum-etl: https://ethereum-etl.readthedocs.io/en/latest/ (accessed on 16 May 2023)
7 100 cores and 1TB of memory. Most of the memory was used.
8 JGraphT: https://jgrapht.org/ (accessed on 16 May 2023)

References

1. Taherdoost, H.; Madanchian, M. Blockchain-Based New Business Models: A Systematic Review. *Electronics* **2023**, *12*, 1479. [CrossRef]
2. Kshetri, N. Scams, frauds, and crimes in the nonfungible token market. *Computer* **2022**, *55*, 60–64. [CrossRef]
3. Das, D.; Bose, P.; Ruaro, N.; Kruegel, C.; Vigna, G. Understanding security issues in the NFT ecosystem. *arXiv* **2021**, arXiv:2111.08893.
4. von Wachter, V.; Jensen, J.R.; Regner, F.; Ross, O. NFT wash trading: Quantifying suspicious behaviour in NFT markets. *arXiv* **2022**, arXiv:2202.03866.
5. Bonifazi, G.; Cauteruccio, F.; Corradini, E.; Marchetti, M.; Montella, D.; Scarponi, S.; Ursino, D.; Virgili, L. Performing Wash Trading on NFTs: Is the Game Worth the Candle? *Big Data Cogn. Comput.* **2023**, *7*, 38. [CrossRef]
6. Pfitzmann, A.; Hansen, M. A Terminology for Talking about Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management, 2010. Available online: https://datatracker.ietf.org/doc/id/draft-hansen-privacy-terminology-00.html (accessed on 16 May 2023).