scientific data

Check for updates

OPEN

DATA DESCRIPTOR

# Crypto-asset trading on top of Ethereum Blockchain comprehensive dataset

Shahar Somin[1,2] ✉, Yaniv Altshuler[2] & Alex Pentland[2]

**Blockchain technology, once limited to niche technological communities, has seen widespread global adoption in recent years, with the potential to reshape financial and social systems. Launched in July 2015, the Ethereum blockchain introduced programmable Smart Contracts. This innovation enabled the creation of user-defined crypto-assets adhering to the ERC-20 standard, supporting a wide range of decentralized applications beyond simple value transfer. We present a large-scale, temporally annotated dataset of ERC-20 token transactions recorded on the Ethereum blockchain. Spanning from November 2015 to December 2024, the dataset encapsulates the trading activity of 216,336,529 users trading 1,138,136 unique tokens, offering a detailed view of crypto-market activity over time. Uniquely, it enables the analysis of a financial ecosystem from its inception, providing rare insights into its structural evolution, participant dynamics, and emergent behaviors. As the largest publicly available resource of its kind, it supports research in blockchain analytics, market dynamics and temporal network analysis. The full dataset and accompanying code are released for public use.**

## Background & Summary

Blockchain technology provides a decentralized framework for recording transactions across a distributed network, eliminating the need for a central authority. Transactions are grouped into blocks, validated through consensus protocols, and permanently appended to the ledger. Each entry is cryptographically signed, ensuring data integrity and tamper resistance. Ethereum, launched in 2015[1,2], extends the original design of the Bitcoin blockchain[3] by introducing a built-in computational layer known as the Ethereum Virtual Machine (EVM). This allows users to deploy and interact with *Smart Contracts*[4], autonomous code executed on the blockchain. Through this mechanism, Ethereum enables the creation of decentralized applications and user-defined digital assets, including fungible tokens adhering to the ERC-20 standard. Ethereum operates on an account-based model, in which each address maintains a persistent balance and transaction history. Unlike Bitcoin's UTXO-based model, this structure facilitates address-level modeling and the direct analysis of trader behavior over time. The Ethereum ecosystem has since evolved into a cornerstone of decentralized finance (DeFi)[5], enabling a wide range of activities such as trading, lending, fundraising, and resource sharing. These capabilities have positioned Ethereum as a valuable resource across diverse research domains, including network science[6,7], economic behavior analysis[8–11], privacy and security aspects[12–14], fraud and illegal activity detection[15–17], financial regulation[18], and decentralized governance[19–21].

Despite the importance of the Ethereum blockchain to a broad range of research fields, a comprehensive, large-scale dataset capturing the full scope of ERC-20 token trading remains largely unavailable to date. While several studies have focused on subsets of token activity or limited timeframes[22], these datasets often exclude critical temporal information[23], or restrict access to only a small fraction of the token ecosystem[24–26]. Part of this gap stems from the technical complexity involved in extracting ERC-20 transaction data: unlike native ETH transfers, ERC-20 token movements are implemented as function calls to Smart Contracts and are not recorded as direct transactions between wallets. Instead, they are observable only through emitted *Transfer* events, which reside within transaction logs and require dedicated parsing from the blockchain's event layer, making large-scale extraction of ERC-20 activity non-trivial. Given the rapid expansion and heterogeneity of ERC-20 tokens, ranging from high-value assets to experimental or fraudulent contracts, there is a clear need for a unified, temporal dataset that reflects the complete financial ecosystem.

[1]Bar-Ilan University, Ramat-Gan, Israel. [2]MIT, Cambridge, MA, USA. ✉e-mail: shaharso@mit.edu

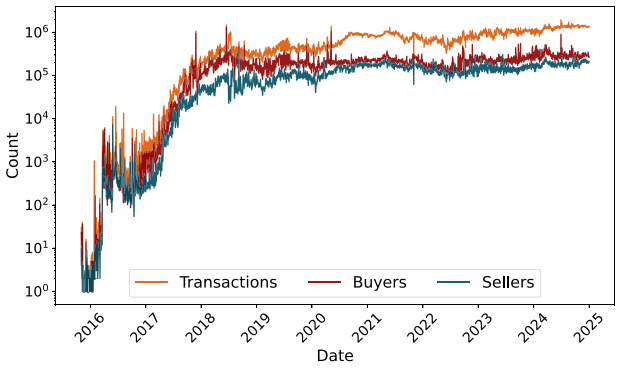


**Fig. 1** Crypto-token trading dynamics. Daily transaction count of ERC-20 crypto-tokens over the Ethereum blockchain (orange curve), and weekly buyers and sellers count depicted by red and green curves, correspondingly.

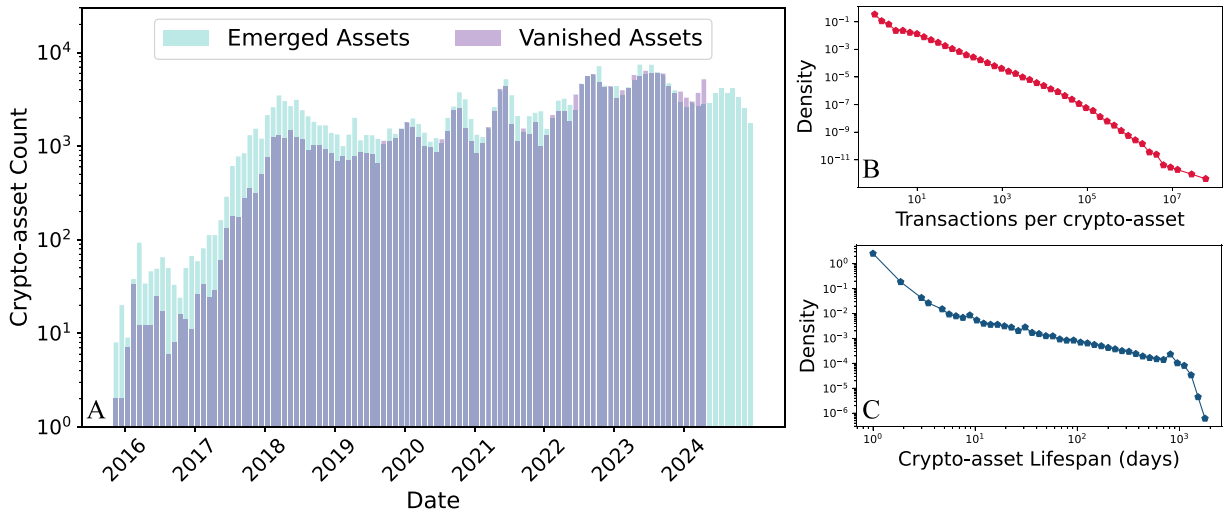


**Fig. 2** Crypto-token lifespan and characteristics. Panel A presents monthly counts of new emerged crypto-tokens (light blue bars), alongside monthly counts of vanished crypto-tokens (no longer traded, purple bars), indicating the high crypto-tokens turnover level. This analysis is limited to data up to June 2024, as we define a token as vanished if it has not been traded for at least six months. Panel B depicts the distribution of transaction amount per crypto-token, presenting a long-tailed distribution, Panel C depicts the distribution of crypto-token lifespan (period between first and last trade, in days).

To this end, we introduce an open-source dataset capturing all ERC-20-compliant token transactions conducted on the Ethereum blockchain between November 2015 and December 2024, based on a curated list of verified ERC-20 token contracts from Etherscan (https://etherscan.io/), consider the Methods section for details on the verification criteria. This dataset encompasses 216,336,529 users and 1,138,136 distinct crypto-tokens. Figure 1 presents weekly transaction amounts alongside weekly amounts of unique buyers and sellers. Figure 2A (light-blue bars) presents the monthly dynamics of crypto-token emergence, showing the number of tokens traded for the first time during each month. Furthermore, we examine the number of tokens that were traded for the last time in each month, serving as an approximation for the number of vanished tokens. These counts are represented by the purple bars in Fig. 2A. This analysis is limited to data up to June 2022, as we define a token as vanished if it has not been traded for at least six months. Figure 2C presents the distribution of token lifespan (period between first and last trade, in days) and Fig. 2B presents the distribution of the number of transactions performed with each token. The dataset offers a detailed, time-resolved view of activity across the ERC-20 ecosystem. By spanning the entire lifecycle of the market—from its inception through multiple cycles of growth and contraction, and across major global events such as the COVID-19 pandemic, shifts in crypto regulation, and the Russia-Ukraine war—this dataset enables comprehensive analysis of how decentralized financial systems evolve under both routine and extreme conditions. It provides a rare opportunity to study the evolution of a decentralized financial ecosystem at scale, enabling analysis of market-wide dynamics that are seldom observable in traditional financial systems and providing a unique lens into the development of complex digital economies.

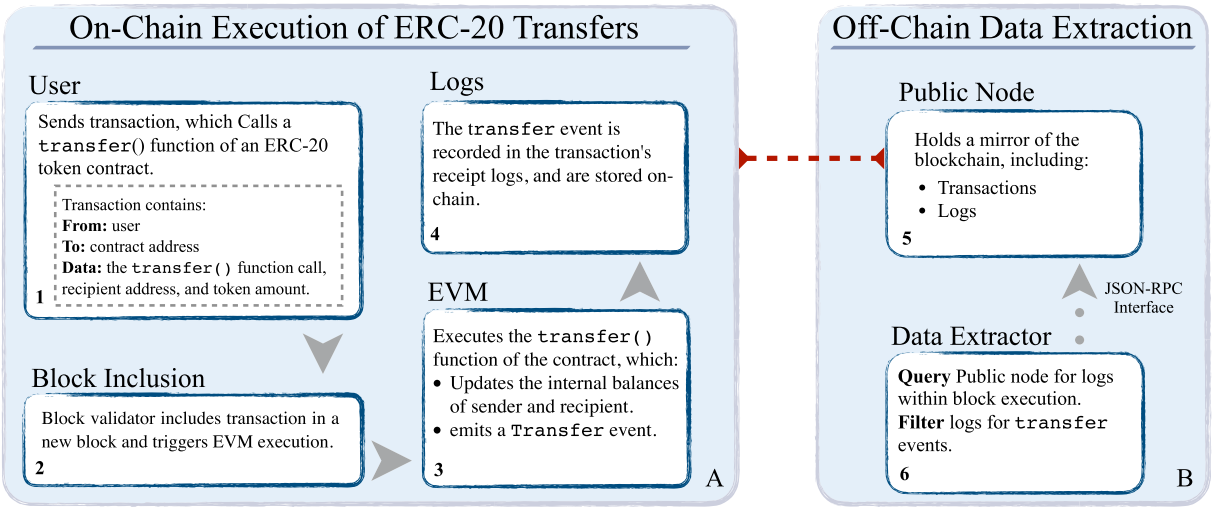


**Fig. 3** ERC-20 transfers On-chain execution and their extraction. Panel A presents the operational flow of ERC-20 token transactions over the Ethereum blockchain. A diagram that illustrates the internal process by which ERC-20 token transfers are executed and recorded on-chain. A user initiates a transfer by sending a transaction to a token smart-contract, specifying the recipient and amount (stage 1). The transaction is validated and included in a block (stage 2), after which the Ethereum Virtual Machine (EVM) executes the transfer() function within the smart contract (stage 3), the smart-contract updates token balances and emits a Transfer event, which is recorded in the transaction's receipt logs (stage 4). Panel B presents the data extraction process, occurring off-chain. A public node, which hold a mirror of the blockchain, is queried for logs by our data extraction code. The code parses these logs and filters *transfer* events associated to ERC-20 trading.

## Methods

Ethereum's infrastructure eliminates the need for centralized control by relying on cryptographic protocols and distributed consensus. Each user operates through a wallet address derived from a public-private key pair, where the private key is used to sign transactions and the public key allows others to verify them. All transactions, including those involving ERC-20 tokens, are propagated to the network and validated independently by all participating nodes. Once confirmed, transactions are grouped into blocks and appended to the blockchain through a consensus mechanism. To maintain user anonymity on the Ethereum Blockchain, transactions exclude any personally identifiable information. Each user interacts through one or more wallet addresses, which are generated by applying the Keccak-256 hash function to their public key. While a wallet address is often treated as a proxy for a user, this is an approximation, as individuals may control multiple addresses[11].

The Ethereum Blockchain supports various types of transactions, including the transfer of Ether between wallets, the deployment of new Smart Contracts, and the invocation of existing contract functions. Smart Contracts, being immutable code stored on the Blockchain, are also assigned unique addresses. Interacting with a Smart Contract involves sending a transaction to its address, which triggers its autonomous execution across all nodes in the network, following the logic encoded in the contract and the input data provided in the transaction. ERC-20 tokens are implemented as Smart Contracts that adhere to a standardized interface governing both token transfers and data accessibility. A core requirement of this standard is the implementation of a transfer function, which facilitates the movement of tokens between wallets. As a result, each ERC-20 token transfer is executed by sending a transaction to the appropriate Smart Contract. This transaction includes a call to the transfer function within its data section, specifying the recipient's address and the amount to be transferred. The transaction modifies the internal state of the Smart Contract by updating the token balances maintained in its storage and emits a *transfer* event which is recorded in the transaction's logs, saving essential details about the transaction. Panel A in Fig. 3 presents the operational flow of ERC-20 token transfers over the Ethereum blockchain.

In order to retrieve ERC-20 token transfers, we first compiled a list (denoted $L_C$) of verified ERC-20 token contract addresses. To this end, we followed a two-stage validation process. First, we queried the Ethereum public dataset hosted on Google BigQuery to retrieve all unique token addresses involved in token transfers (*bigQuery_token_adds. csv*). Specifically, we executed:

```
SELECT DISTINCT token_address
FROM 'bigquery-public-data.crypto_ethereum.token_transfers`
```

This query yields all contract addresses that have emitted events matching the canonical ERC-20 Transfer(address indexed from, address indexed to, uint256 value) signature. However, the presence of this log structure alone does not guarantee full ERC-20 compliance or verification status. To address this limitation, we cross-referenced the resulting token addresses using the Etherscan API. For each address, we requested the verified contract ABI using the getabi endpoint. Etherscan assigns verified status to contracts whose source code

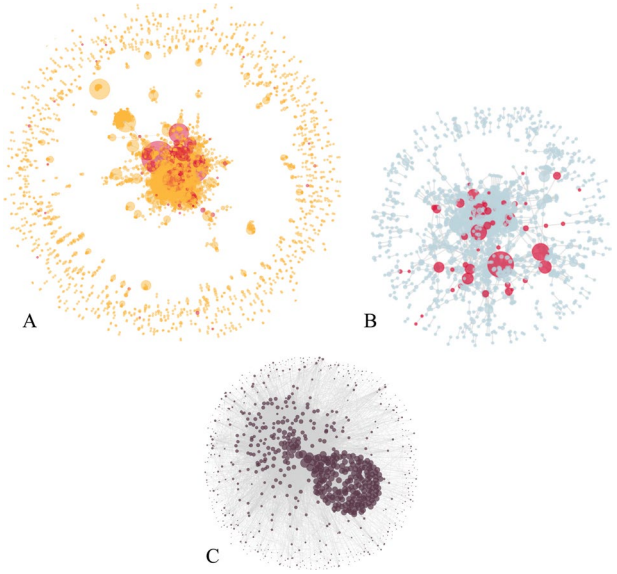


**Fig. 4** Different network types. Panel A presents a wallet-to-wallet network, with edges representing all crypto-token trading (buy or sell), during a 30 minutes period. Panel B presents a wallet-to-wallet network, trading a single crypto-token (USDC) throughout a single hour. Panel C presents token-to-token network, where two crypto-tokens are connected if were bought by the same trader, during a 30 minutes period. throughout the different panels, node size is proportional to its degree, and red nodes represent crypto-exchanges.

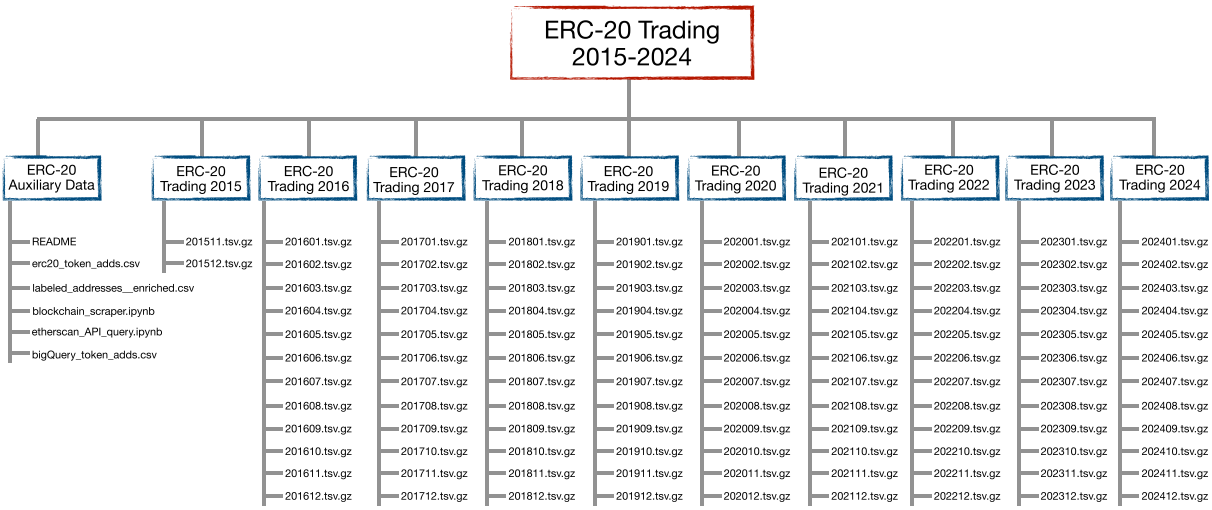


**Fig. 5** Diagram of dataset structure.

has been publicly submitted and successfully matched against the deployed bytecode. This guarantees that the contract's logic conforms to its declared interface. We parsed the ABI and confirmed the presence of key ERC-20 methods (totalSupply, balanceOf, and transfer) to ensure functional compatibility. Only addresses passing this verification check were retained in $L_C$. While $L_C$ covers the vast majority of widely used ERC-20 tokens, it may exclude contracts that were never submitted for verification or that do not strictly adhere to the standard interface. This list is included in the public repository accompanying the dataset (named *erc20_token_adds. csv*). The script used for querying the Etherscan API (*etherscan_API_query. ipynb*) and the initial token list (*bigQuery_token_adds. csv*) are included in the auxiliary data supplied with this dataset .

Next, we used the Ethereum infrastructure provider Infura (https://infura.io/) as our backend node. Infura exposes a standard Ethereum full node via the JSON-RPC API, and we accessed it using the Python web3.py library. For each contract $c \in L_C$, we performed *eth_getLogs* queries in small block intervals, filtering by the canonical *Transfer* event topic hash. While the *Transfer* event is used in both ERC-20 and ERC-721 standards and shares the same topic hash, our dataset construction methodology ensures that only ERC-20 transfers are included. Specifically, by querying *eth_getLogs* over a predefined list of verified ERC-20 token contract addresses, we ensure only ERC-20 transactions are collected. From each log, we extracted the emitting contract address,

| block-id | Transaction hash | Time | Token Address | Token Name | Token Symbol | Token Decimals | Sender | Recipient | Value |
|---|---|---|---|---|---|---|---|---|---|
| 5 | 1 | 13-10-2021 22:11:32 | 0x1i876 | Tether | USDT | 18 | 0x987j0 | 0x8754h | 100 |
| 6 | 2 | 04-04-2019 12:22:01 | 0xo98n6 | 0x token | ZRX | 18 | 0x8764h | 0x5326j | 23 |
| 7 | 3 | 12-05-2016 10:33:05 | 0x87b75 | SHIBA INU | SHIB | 18 | 0x98753 | 0x7643g | 45 |
| 8 | 4 | 22-02-2021 01:55:08 | 0xu65b3 | Wrapped BTC | WBTC | 18 | 0x8764h | 0x86bl9 | 43 |

**Table 1.** Example of dataframe and the columns in the ERC-20 dataset.

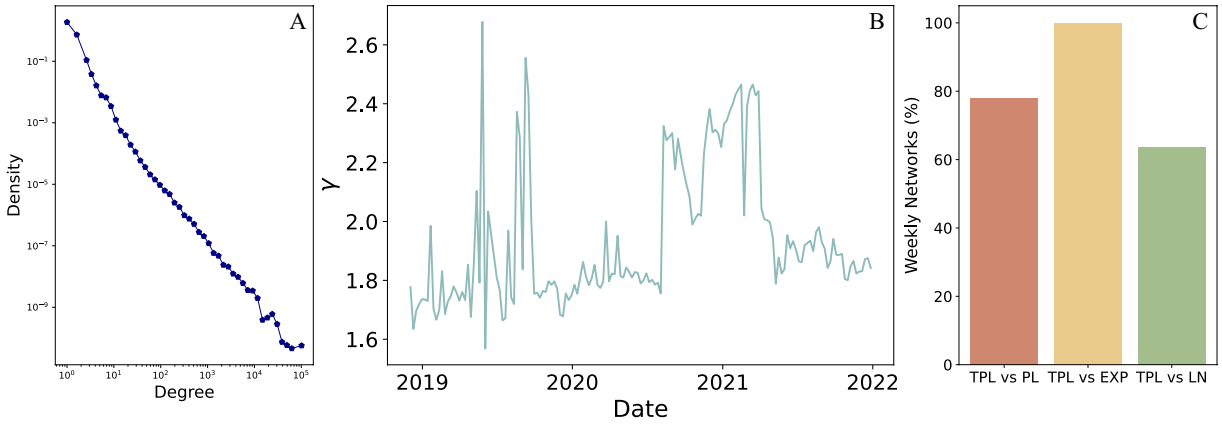


**Fig. 6** Long-tailed degree distributions. Panel A depicts the long-tailed degree distribution of a weekly wallet-to-wallet all crypto-tokens trading network. Panel B depicts the dynamics of $\gamma$, the truncated power-law parameter of the degree distributions, along time. Panel C depicts goodness-of fit tests, standing for the weekly percentage of networks whose Log-Likelihood Ratio is positive and statistically significant ($p\text{-}value < 0.1$), when compared to power-law, exponential and lognormal models. The weekly networks present high agreement with the truncated power-law model.

sender and recipient addresses, transferred value, block number, and transaction hash. To enrich the dataset with token-level metadata, we issued additional *eth_call* instructions to each contract, invoking the standard ERC-20 methods name(), symbol(), and decimals(). Although we used a paid Infura plan to allow long-range, sustained querying, we provide code in the auxiliary data repository (*blockchain_scraper. ipynb*) which is fully compatible with Infura's free tier and can also reproduce smaller portions of the dataset, subject to Infura's rate limitations. The following auxiliary files are included in the public dataset repository:

1. The list of verified contract addresses (*erc20_token_adds. csv*).
2. The list of initial contract addresses (*bigQuery_token_adds. csv*).
3. The Etherscan API querying script (*etherscan_API_query. ipynb*).
4. The data extraction script (*blockchain_scraper. ipynb*).

The data extraction script is extensible and can also be used to include unverified or newly deployed ERC-20 contracts.

Accordingly, we collected the set of ERC-20 token transactions conducted between November 2nd, 2015 (block id: 477958) to December 31st, 2024 (block id: 21525890), totaling in 1,943,438,828 trades, executed by 216,336,529 users and 1,138,136 different crypto-tokens. The data extraction methodology ensures that the dataset includes all Transfer events emitted by ERC-20 contracts, regardless of whether they were initiated by externally owned accounts or triggered as part of a contract-to-contract interaction. Any transaction that resulted in the emission of a Transfer log is captured by our extraction method, consistent with the observable *Transfer* event recorded on the Ethereum blockchain.

Additionally, since Smart Contracts on the Ethereum blockchain are immutable, any update to a token's logic necessitates the deployment of a new Smart Contract, resulting in a new Contract Address. Consequently, a crypto-token may be associated with multiple addresses over time. However, at any given moment, each token corresponds to a single active Contract Address. Thus, the total number of distinct contract addresses observed should be interpreted as an upper bound on the actual number of unique tokens.

**Network construction.** The ERC-20 dataset enables the construction of various types of networks, each offering a distinct representation for analyzing the ERC-20 financial ecosystem. The most natural representation is the wallet-to-wallet transaction network, where nodes represent wallet addresses and edges indicate token transfers between them. Figure 4A presents a wallet-to-wallet network, with edges representing trading across

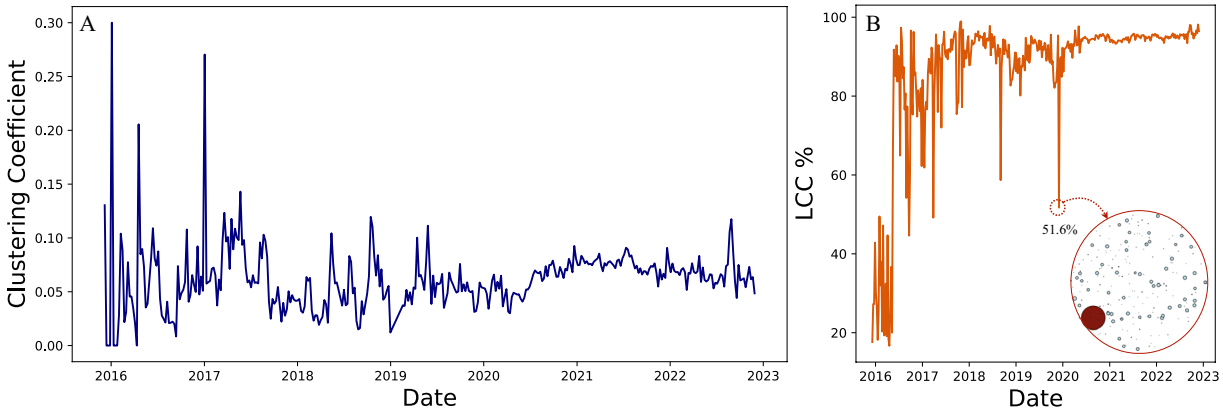


**Fig. 7** Node-level and network level connectivity dynamics. Panel A presents node-level connectivity, as indicated by the averaged node clustering coefficient along time, demonstrating its stabilization process. Panel B presents the network level connectivity, as the percentage of nodes within the largest connected component (LCC) over time in weekly wallet-to-wallet networks, presenting the formation of a giant connected component. Inset image presents the different components of an outlier weekly network, obtaining an LCC percentage as low as 51.6%, with the LCC depicted by the red node.

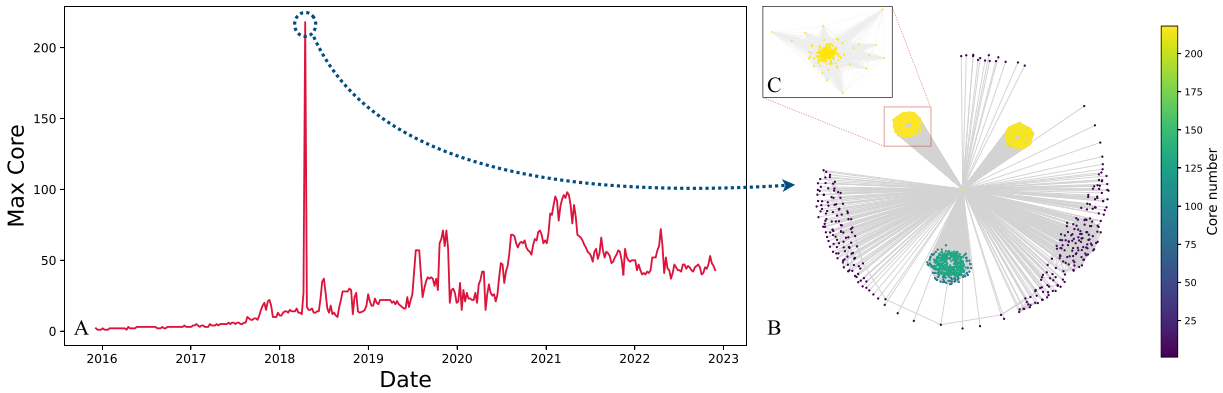


**Fig. 8** Core-number dynamics. Panel A depicts the maximal core-number dynamics, presenting an increase along time. Panel B presents the connected component responsible for the anomalously high core number associated with the weekly network starting at April 16th, 2018, with node color representing the node's core-number. Panel C presents the structure of nodes with the highest core-number.

all possible tokens. These networks can be constructed statically or at various temporal resolutions (e.g., daily, weekly, monthly), and can be either aggregated across all tokens or filtered by specific assets. Figure 4B presents a wallet-to-wallet network with edges representing trading in a specific crypto-token. Edges may be weighted by the number of transfers or the cumulative volume transferred, which could support analysis of flow intensity, centrality, and hub structures within the ecosystem. This representation illustrates potential use-cases, including identifying influential traders, detecting anomalous behavior, and examining patterns of wealth redistribution.

Beyond direct transfers, the dataset also supports the construction of higher-order or multi-modal networks. For instance, a bipartite network can be built linking wallets to the tokens they interact with, enabling analysis of token co-ownership or specialization. Alternatively, one may define a token-to-token similarity network, where edges link tokens that share overlapping traders. Figure 4C presents a sample token-to-token network. Dynamic network representations can also be derived by constructing time-evolving sequences of transaction graphs, allowing for the study of structural shifts, market fragmentation, and coordination patterns over time. These diverse network views allow researchers to represent the ERC-20 token space not merely as a collection of transactions, but as a dynamic, evolving financial ecosystem.

## Data Records

The ERC-20 trading dataset is available at the Harvard Dataverse[27–37]. It contains all ERC-20 trades during 7 years. Figure 5 describes the dataset structure and its organization. The origin is the "ERC-20 Trading 2015-2022" dataverse. Each of the blue rectangles stands for different datasets within this dataverse. The first one holding all auxiliary data, including a README file, the list of all ERC-20 crypto tokens, a list of exemplary labeled

wallet addresses, pertaining to service providers such as crypto-exchanges, and a jupyter notebook containing the scraping code we have utilized for extracting the data, in order to facilitate further data extraction. The rest of the datasets hold ERC-20 trading data for each year, divided by months.

Table 1 presents an example of a monthly transactions file. Each record refers to a single ERC-20 transaction and contains the following fields:

1. **Token Address**: indicating the address of the Smart Contract governing the crypto-token.
2. **Token Name**: String identifying the crypto-token.
3. **Token symbol**: The abbreviated shorthand of the crypto-token, 3 to 4 characters long.
4. **Token Decimals**: The number of decimal places a token supports, determining its smallest transferable unit.
5. **Value**: denoting the transferred token amount.
6. **Sender**: wallet address representing the seller.
7. **Recipient**: wallet address representing the buyer.
8. **Time**: the time at which the transaction took place.
9. **Transaction hash**: a unique transaction identifier.
10. **Block-id**: the unique integer identifier of a block in the Ethereum blockchain.

We also release a supplementary dataset (*labeled_addresses__enriched. csv*) of labeled Ethereum wallet addresses, intended to support downstream tasks such as wallet classification, behavioral analysis, and entity prediction. The dataset includes 10,627 addresses associated with 52 known entities, spanning centralized exchanges, decentralized finance (DeFi) protocols, and other blockchain-based service providers. To construct this dataset, we first identified labeled wallet addresses directly from Etherscan, which annotates certain addresses with verified entity names. For each of these labeled addresses, we then manually searched for additional information online, including the project's official website and service type, in order to enrich the dataset with contextual metadata. By providing ground-truth labels for a diverse set of actors in the Ethereum ecosystem, this resource enables supervised learning approaches and enhances interpretability in network-based or transaction-based analyses. Each record in this dataset refers to a single Ethereum wallet, containing the following fields:

1. **Address**: the address of the labeled Ethereum wallet.
2. **Name**: the name identifier of the Ethereum wallet.
3. **Type**: the service this address is associated with.
4. **URL**: a url address associated with the service provider.

## Technical Validation

To demonstrate the structural richness and temporal consistency of the dataset, we present a series of validations on the derived wallet-to-wallet transaction networks. These validations aim to demonstrate that the dataset not only captures the full breadth of ERC-20 token activity but also reflects meaningful patterns and dynamics that evolve over time. By presenting key network-theoretic properties, including core number distributions, clustering coefficients, and degree distributions, we show that the data preserves both the micro- and macro-level behaviors expected in complex networks. These dynamics further illustrates the dataset's capacity to capture structural variations in the ERC-20 token ecosystem across different market periods. Due to computational considerations, all analyses in this section are restricted to data from 2015 through 2022 (inclusive).

Specifically, prior studies have consistently shown that the degree distributions of complex networks, particularly in economic systems, tend to exhibit long-tailed behavior[38–41]. As illustrated in Fig. 6A, the degree distribution of a representative weekly wallet-to-wallet transaction graph (over all ERC-20 tokens) similarly displays a heavy-tailed structure, aligning with these established findings. As an example of dataset validation, we apply established statistical methodologies[42] to illustrate how truncated power-law provides the best fit to the observed degree distributions, similarly to a variety of other complex systems. Specifically, we perform a goodness of fit test calculating the Log-Likelihood Ratio (LLR) of the different models and the corresponding p-values. Panel C in Fig. 6 presents the comparison of the truncated power-law to the power-law, exponential and the lognormal models for each of the weekly networks. The bars signify the percentage of networks for which the LLR is positive and the test achieved $p - value < 0.1$. The results suggest that the truncated power-law model better fits the majority of networks compared to all heavy-tailed models (78%, 99% and 63% for comparing with power-law, exponential and lognormal correspondingly).

Furthermore, illustrating the evolution of connectivity in transaction networks can help demonstrate how the dataset may support studies of systemic robustness and the potential for information or value propagation. In the context of ERC-20 trading, local and global connectivity measures serve as examples of how fragmentation or cohesion can be observed over time using the dataset. Fig. 7 illustrates these dynamics in weekly wallet-to-wallet transaction networks. Panel A shows the average clustering coefficient over time, illustrating a gradual stabilization of local connectivity among wallets. This stabilization over time suggests that the microstructure of wallet interactions reaches a steady regime, consistent with patterns that may emerge in mature ecosystems. Such stabilization may point to the maturation of the ecosystem, where transactional behaviors become more structured and less volatile, even as the network continues to grow. This phenomenon also implies that localized trading relationships or community structures remain relatively stable, which could be relevant for exploring the efficiency, resilience, and vulnerability of the system to shocks in future work. Panel B depicts the emergence of network-level connectivity through the proportion of nodes present in the largest connected component (LCC), illustrating its formation along time. The inset figure highlights an outlier network presenting a

structural fragmentation, where its LCC contained only 51.6% of wallets. The rapid growth of the LCC indicates that a substantial share of wallets became mutually reachable early in the network's development. This early formation suggests that the conditions for large-scale interaction were established relatively quickly, allowing for system-wide patterns, such as collective trading behavior or shared responses to market events, to begin emerging even in the initial phases of the ERC-20 ecosystem.

Presenting another notion of connectivity, we analyze the core-number[43] dynamics of weekly wallet-to-wallet networks. Figure 8A presents the maximal core number obtained by nodes on a weekly basis, manifesting an increasing dynamics over time. The increasing maximal core number over time reflects the growing presence of densely connected substructures within the network. Rather than remaining loosely organized, certain subsets of wallets become more deeply embedded in the transactional fabric, providing an example of how the dataset may support examination of connectivity, centralization, or vulnerability to targeted disruptions. Panel B depicts the anomalous weekly network, responsible for a maximal core-number of over 200, where panel C presents the structure of nodes obtaining this high core-number.

## Code availability

No custom code was generated for this work.

Received: 22 April 2025; Accepted: 18 July 2025;
Published online: 12 August 2025

## References

1. Buterin, V. *et al*. A next-generation smart contract and decentralized application platform. *white paper* **3**(37), 2–1 (2014).
2. Wood, G. *et al*. Ethereum: A secure decentralised generalised transaction ledger. *Ethereum project yellow paper* **151**(2014), 1–32 (2014).
3. Nakamoto, S. Bitcoin: A peer-to-peer electronic cash system. (2008).
4. Voshmgir, S., *Token economy: How the Web3 reinvents the internet*, volume 2. Token Kitchen, (2020).
5. Harvey, C. R., Ashwin Ramachandran, & Joey Santoro. *DeFi and the Future of Finance*. John Wiley & Sons, (2021).
6. Kim, S. K. *et al*. Measuring ethereum network peers. In *Proceedings of the Internet Measurement Conference 2018*, 91–104, (2018).
7. Somin, S., Altshuler, Y., Gordon, G., Pentland, A. S. & Shmueli, E. Network dynamics of a financial ecosystem. *Scientific reports* **10**(1), 4587 (2020).
8. Chen, C. *et al*. When digital economy meets web3. 0: Applications and challenges. *IEEE Open Journal of the Computer Society* **3**, 233–245 (2022).
9. Somin, S., Altshuler, Y., Pentland, A. S. & Shmueli, E. Beyond preferential attachment: falling of stars and survival of superstars. *Royal Society Open Science* **9**(8), 220899 (2022).
10. Somin, S. *et al*. Remaining popular: Power-law regularities in network dynamics. *EPJ Data Science* **11**(1), 61 (2022).
11. Victor, F. Address clustering heuristics for ethereum. In *Financial Cryptography and Data Security: 24th International Conference, FC 2020, Kota Kinabalu, Malaysia, February 10–14, 2020 Revised Selected Papers 24*, 617–633 (Springer, 2020).
12. Chen, H., Pendleton, M., Njilla, L. & Xu, S. A survey on ethereum systems security: Vulnerabilities, attacks, and defenses. *ACM Computing Surveys (CSUR)* **53**(3), 1–43 (2020).
13. Kushwaha, S. S., Joshi, S., Singh, D., Kaur, M. & Lee, H.-N. Systematic review of security vulnerabilities in ethereum blockchain smart contract. *Ieee Access* **10**, 6605–6621 (2022).
14. Somin, S., Erhardt, K. & Pentland, A. S. Temporal fingerprints: Identity matching across fully encrypted domain. *arXiv preprint arXiv:2407.04350*, (2024).
15. Chen, W. *et al*. Detecting ponzi schemes on ethereum: Towards healthier blockchain technology. In *Proceedings of the 2018 world wide web conference*, pages 1409–1418, 2018.
16. Wu, J. *et al*. Who are the phishers? phishing scam detection on ethereum via network embedding. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **52**(2), 1156–1166 (2020).
17. Somin, S., Cohen, T., Kepner, J. & Pentland, A. Echoes of the hidden: Uncovering coordination beyond network structure. *arXiv preprint arXiv:2504.02757*, 2025.
18. Zetzsche, D. A., Arner, D. W. & Buckley, R. P. Decentralized finance. *Journal of Financial Regulation* **6**(2), 172–203 (2020).
19. DuPont, Q. Experiments in algorithmic governance. *Bitcoin and beyond*, page 157 (2017).
20. Hassan, S. & De Filippi, P. Decentralized autonomous organization. *Internet Policy Review* **10**(2), (2021).
21. Lee, J. Y. A decentralized token economy: How blockchain and cryptocurrency can revolutionize business. *Business Horizons* **62**(6), 773–784 (2019).
22. Cui, W. & Gao, C. Wteye: On-chain wash trade detection and quantification for erc20 cryptocurrencies. *Blockchain: Research and Applications* **4**(1), 100108 (2023).
23. Wang, Q. *et al*. Ex-graph: A Pioneering Dataset Bridging Ethereum and X. *The Twelfth International Conference on Learning Representations* (2023).
24. Huang, S. *et al*. Temporal graph benchmark for machine learning on temporal graphs. *Advances in Neural Information Processing Systems* **36**, 2056–2073 (2023).
25. Weber, M. *et al*. Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. arXiv preprint arXiv:1908.02591 (2019)
26. Victor, F. & Lüders, B. K. Measuring ethereum-based erc20 token networks. In *Financial Cryptography and Data Security: 23rd International Conference, FC 2019, Frigate Bay, St. Kitts and Nevis, February 18–22, 2019, Revised Selected Papers 23*, pages 113–129 (Springer, 2019).
27. Somin, S. *ERC-20 Trading 2015*. Harvard Dataverse. https://doi.org/10.7910/DVN/REODCK (2025).
28. Somin, S. *ERC-20 Trading 2016*. Harvard Dataverse. https://doi.org/10.7910/DVN/GE0KWI (2025).
29. Somin, S. *ERC-20 Trading 2017*. Harvard Dataverse. https://doi.org/10.7910/DVN/APVWMK (2025).
30. Somin, S. *ERC-20 Trading 2018*. Harvard Dataverse. https://doi.org/10.7910/DVN/JZNWIC (2025).
31. Somin, S. *ERC-20 Trading 2019*. Harvard Dataverse. https://doi.org/10.7910/DVN/F8Z1IN (2025).
32. Somin, S. *ERC-20 Trading 2020*. Harvard Dataverse. https://doi.org/10.7910/DVN/8YO2VZ (2025).
33. Somin, S. *ERC-20 Trading 2021*. Harvard Dataverse. https://doi.org/10.7910/DVN/C1AR9V (2025).
34. Somin, S. *ERC-20 Trading 2022*. Harvard Dataverse. https://doi.org/10.7910/DVN/5P82QC (2025).
35. Somin, S. *ERC-20 Trading 2023*. Harvard Dataverse. https://doi.org/10.7910/DVN/KVTEPR (2025).
36. Somin, S. *ERC-20 Trading 2024*. Harvard Dataverse. https://doi.org/10.7910/DVN/H6R61C (2025).
37. Somin, S. *ERC-20 Auxiliary Data*. Harvard Dataverse. https://doi.org/10.7910/DVN/MBF0GC (2025).

38. Liu, Y.-Y., Nacher, J. C., Ochiai, T., Martino, M. & Altshuler, Y. Prospect theory for online financial trading. *PloS one* **9**(10), e109458 (2014).
39. Pan, W., Altshuler, Y. & Pentland, A. Decoding social influence and the wisdom of the crowd in financial trading network. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 203–209. IEEE, 2012.
40. Kitzler, S., Victor, F., Saggese, P. & Haslhofer, B. Disentangling decentralized finance (defi) compositions. *ACM Transactions on the Web* **17**(2), 1–26 (2023).
41. Lee, X. T., Khan, A., Sen Gupta, S., Ong, Y. H. & Liu, X. Measurements, analyses, and insights on the entire ethereum blockchain network. In *Proceedings of The Web Conference 2020*, pages 155–166, 2020.
42. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM review* **51**(4), 661–703 (2009).
43. Carmi, S., Havlin, S., Kirkpatrick, S., Shavitt, Y. & Shir, E. A model of internet topology using k-shell decomposition. *Proceedings of the National Academy of Sciences* **104**(27), 11150–11154 (2007).

## Acknowledgements

Research was sponsored by the United States Air Force Research Laboratory and the Department of the Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

## Author contributions

Y.A. was pivotal in Blockchain data extraction. S.S. performed data analysis, technical validation and wrote the manuscript with input from all the authors. A.P. and Y.A. reviewed manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to S.S.