AALBORG UNIVERSITY

Sixth International Workshop on Foundations and Applications of Blockchain

In conjunction with VLDB'24

August 30th, 2024

Data Management and AI for Blockchain Data Analysis: A Round Trip and Opportunities

FAB 2024 Keynote

Arijit Khan

Department of Computer Science

Aalborg University, Denmark

arijitk@cs.aau.dk

Outline

1) **Introduction**
   1.1 Blockchain Data Analysis
   1.2 Applications and Challenges
   1.3 Blockchain Data Extraction

2) **Account-based Blockchain Graphs Analysis**
   2.1 Local and Global Graph Property Analysis
   2.2 Temporal Graph Analysis

3) **Advanced Data Analytics for Blockchain Graphs**
   3.1 Topological Data Analysis on Blockchain Graphs
   3.2 Machine Learning on Blockchain Graphs
   3.3 Higher-order Structural Analysis on Blockchain Graphs

Data Management and AI for Blockchain Data Analysis

4) **Blockchain Datasets and Analysis Tools**

Blockchain Datasets and Tools for Data Science Ecosystem

5) **Open Problems**

Reference

WWW 2020: Xi Tong Lee, Arijit Khan, Sourav Sen Gupta, Yu Hann Ong, and Xuan Liu, "*Measurements, Analyses, and Insights on the Entire Ethereum Blockchain Network*", in Proc. of The Web Conference 2020.

WWW 2021: Lin Zhao, Sourav Sen Gupta, Arijit Khan, and Robby Luo, "*Temporal Analysis of the Entire Ethereum Blockchain Network*", in Proc. of The Web Conference 2021.

WSDM 2022: Voon Hou Su, Sourav Sen Gupta, and Arijit Khan, "*Automating ETL and Mining of Ethereum Blockchain Network*", in Proc. of the Web Search and Data Mining Conference 2022.

CIKM 2022: Arijit Khan and Cuneyt Gurcan Akcora, "*Graph-based Management and Mining of Blockchain Data*", in Proc. of the ACM International Conference on Information and Knowledge Management 2022.

Frontiers in Blockchain 2024: Jason Zhu, Arijit Khan, and Cuneyt Gurcan Akcora, "Data Depth and Core-based Trend Detection on Blockchain Networks", in Frontiers in Blockchain, section Blockchain Economics, 2024.

ArXiv 2024: Poupak Azad, Cuneyt Gurcan Akcora, Arijit Khan, "*Machine Learning for Blockchain Data Analysis: Progress and Opportunities*", CoRR abs/2404.18251, 2024.

Blockchain Data Analysis, Applications, and Challenges

Blockchain Data Analysis

- Data stored in a public blockchain can be considered **big data.**

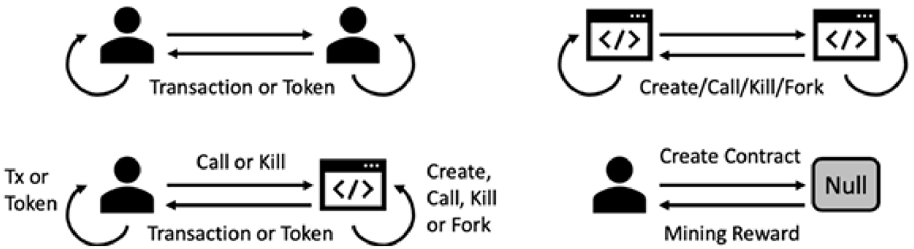
- **Volume:** Ethereum archive nodes that store a complete snapshot of the Ethereum blockchain, including all the transaction records, take up to **4TB of space.**

  https://decrypt.co/24779/ethereum-archive-nodes-now-take-up-4-terabytes-of-space

- **Velocity:** Ethereum blockchain has processed more than **1.1 million transactions per day** in July 2021.

  https://www.statista.com/statistics/730838/number-of-daily-cryptocurrency-transactions-by-type/

- **Veracity:** Ethereum contains a vast number of **heterogeneous interactions**, e.g., user-to-user, user-to-contract, contract-to-user, and contract-to-contract across multiple layers via external and internal transactions, ether, tokens, dAapps, etc.

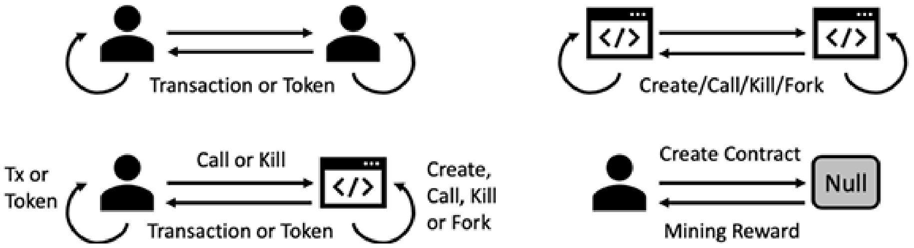


Interactions in the Ethereum Blockchain Network

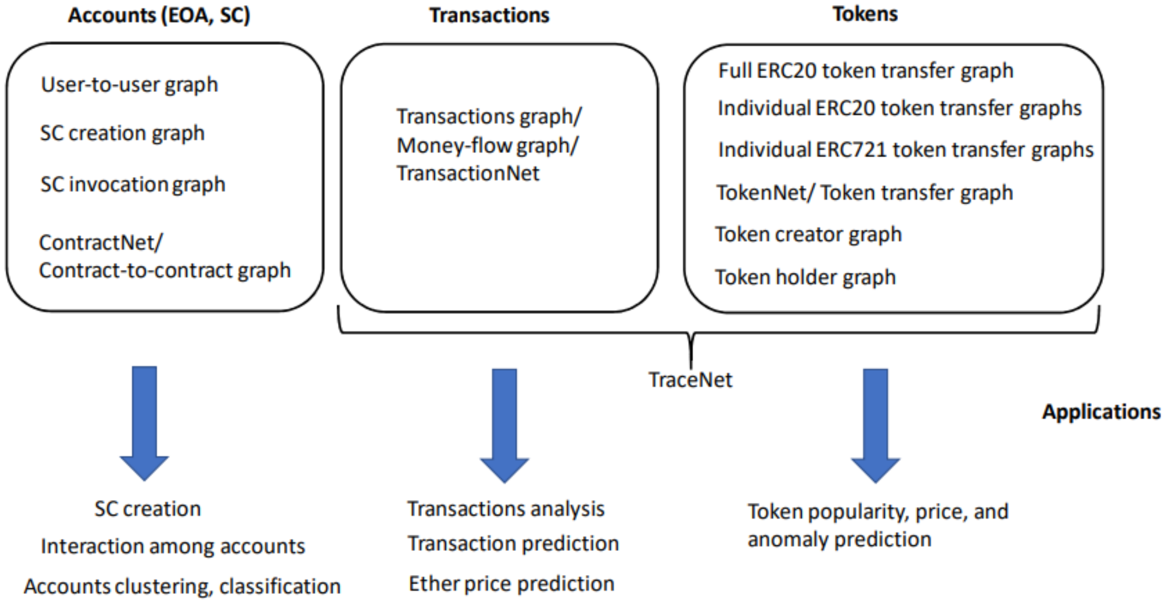
Graph-based Blockchain Data Analysis

- **Data analytic methods** can be applied to extract knowledge hidden in the blockchain.

- Several recent research works performed **graph analysis** on the publicly available blockchain data to reveal insights into its transactions and for important downstream tasks, e.g., **cryptocurrency price prediction**, **address clustering**, **phishing scams**, and **counterfeit tokens detection**.



Interactions in the Ethereum Blockchain Network



Various graphs created from interactions between accounts, transactions, token transfers; as well as their common applications

Blockchain Data Analysis: Applications

- Bulk of the works conducted graph analysis to gain insights into transaction and token transfers.
- Some of them considered downstream tasks, e.g., node classification, link prediction, anomaly detection, token price prediction.
- Most tools for blockchain data are related to e-crime or financial (e.g., price, investor) analytics.
- From ransomware payment detection to sextortion discovery, transaction graph analysis has proven useful to study blockchain address importance and to cluster them.

Oggier, F., Datta, A. and Phetsouvanh, S., 2020. **An ego network analysis of sextortionists**. *Social Network Analysis and Mining*, *10*(1), pp.1-14.

Bistarelli, S., Mercanti, I. and Santini, F., 2018, August. **A suite of tools for the forensic analysis of bitcoin transactions: Preliminary report**. In *European Conference on Parallel Processing* (pp. 329-341). Springer, Cham.

Wu, J., Lin, K., Lin, D., Zheng, Z., Huang, H., and Zheng, Z. (2022). Financial crimes in web3-empowered metaverse: taxonomy, countermeasures, and opportunities. IEEE Open J. Comput. Soc. 4, 37–49.

Intelligent Blockchain Ecosystem

- Assess health of crypto eco-systems, data mining and analytics skills to help clients avoid transaction risks.
- Network features of cryptocurrencies transactions as a proxy for market sensing.
- Companies to build better blockchain ecosystems, blockchain intelligence (https://blockchaingroup.io), blockchain-based social networks (Steemit) and blockchain search engines

J. Zhu, A. Khan, and C. G. Akcora (2024). **Data depth and core-based trend detection on blockchain transaction networks**. Front. Blockchain 7:1342956. doi: 10.3389/fbloc.2024.1342956

Blockchain Data Analytics: Challenges

- **Anonymity:** tracking addresses and analyzing transaction patterns difficult.
- **Limited visibility:** compiled binary of smart contract code visible on the blockchain.
- **Blockchain data:** Volume, velocity, Veracity (Big Data).
- **Adversarial behaviors:** long-range attacks, manipulations, malicious smart contracts, abusive users.
- **Machine Learning Challenges:** skewed distribution, lack of ground truth, new attacks, distribution drift, external influences, black-box ML models.

P. Azad, C. G. Akcora, A. Khan, "**Machine Learning for Blockchain Data Analysis: Progress and Opportunities**", CoRR abs/2404.18251, 2024.

Data Management and AI for Blockchain Data Analysis

Sixth International Workshop on Foundations and Applications of Blockchain

In conjunction with VLDB'24

August 30th, 2024

Blockchain Data ETL

Graph-based Blockchain Data Analysis
- Account-based graphs, UTXO graphs

Advanced Data Analytics for Blockchain Graphs
- Topological data analysis
- Graph machine learning
- Higher-order structural analysis

Graph ML
transaction graph

Temporal ML
transaction, price

Sequential ML
transaction, smart contract, social data

Code ML
smart contract

Text ML
social data

A. Khan and C. G. Akcora, "Graph-based Management and Mining of Blockchain Data", CIKM 2022.

P. Azad, C. G. Akcora, A. Khan, "**Machine Learning for Blockchain Data Analysis: Progress and Opportunities**", CoRR abs/2404.18251, 2024.

Blockchain Data Extraction

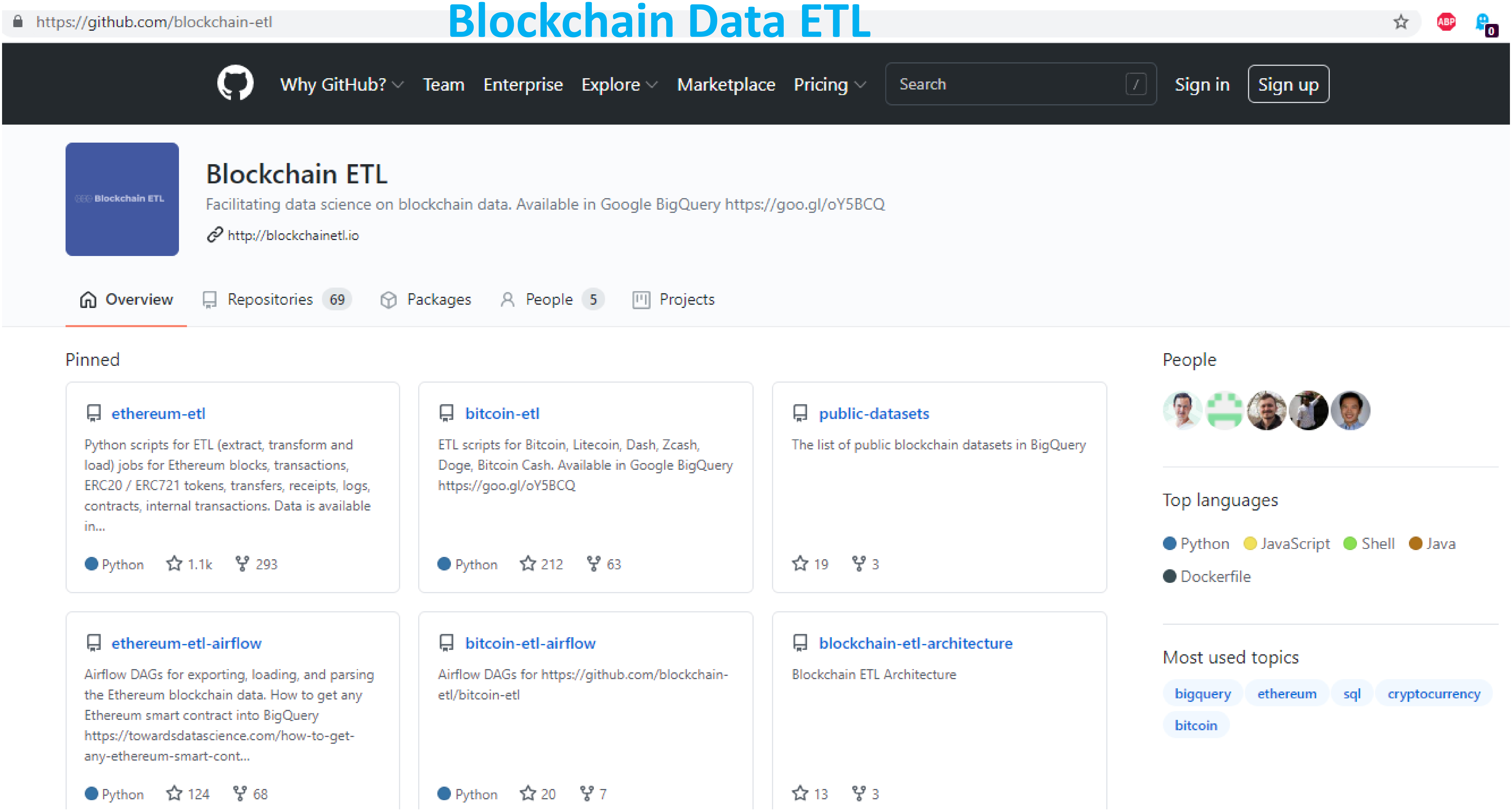
Data Extraction Methods

- **Run a full-node on the blockchain to collect all historic transactions – e.g., Bitcoin-Core, Geth, and Parity.**
  - Massive-storage and hardware requirement; more than a week to fully synchronize entire data at a newly connected node.
  - Not good for ad-hoc queries.

- **Web3 services and APIs for data extraction – e.g., Infura, SoChain, and Quicknode.**
  - high costs if users want to extract large amounts of data; paid and slow APIs.
  - Blockchain data is stored at clients in heterogeneous, complex data structures, in binary or in encrypted format, which cannot be directly used for exploration, mining, or visualization.

- **Well-processed blockchain datasets – e.g.,**
  - **Google Big Query** (https://cloud.google.com/blog/products/data-analytics/introducing-six-new-cryptocurrencies-in-bigquery-public-datasets-and-how-to-analyze-them)
  - **https://xblock.pro/#/** (Sun Yat-sen University and others)
  - ETL (extract-transform-load) can still be an issue.

Blockchain Data ETL



Source – Google BigQuery

table_id	utc_created_date	utc_modified_date	rows_millions	size_gb
blocks	2019-01-15 13:30:29.658	2021-05-06 05:29:23.607	**11.72**	**12.07**
token_transfers	2019-01-15 13:28:07.793	2021-05-06 05:31:55.894	**595.69**	**171.88**
traces	2019-01-15 13:55:23.777	2021-05-06 05:22:25.641	**2775.28**	**1626.74**
transactions	2019-01-15 13:29:49.289	2021-05-06 05:28:48.798	**985.76**	**455.64**

These four tables from Google BigQuery are the most important sets of data from the Ethereum blockchain in terms of the primary **“interaction networks” between User and Contract accounts**.

V. H. Su, S. S. Gupta, A. Khan. **Automating ETL and mining of ethereum blockchain network**, WSDM 2022.

ETL Problem to Solve

Convert this

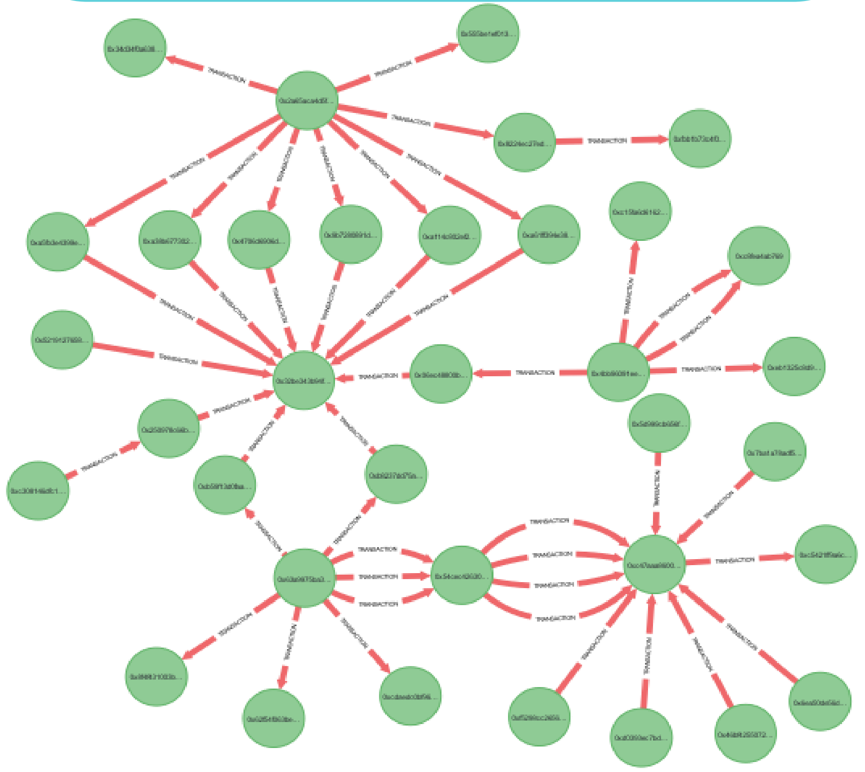
Tabular Representation

from_address	to_address	edge_data	block_number
0xd3b1fad...	0x1625a9f...	...	0
0x4bc3c20...	0xfe611a3...		1
0x40af81b...	0x5716678...		2
0x9786a24...	0xa25a8dc...		3

How to perform this step?

To this

Graph Representation



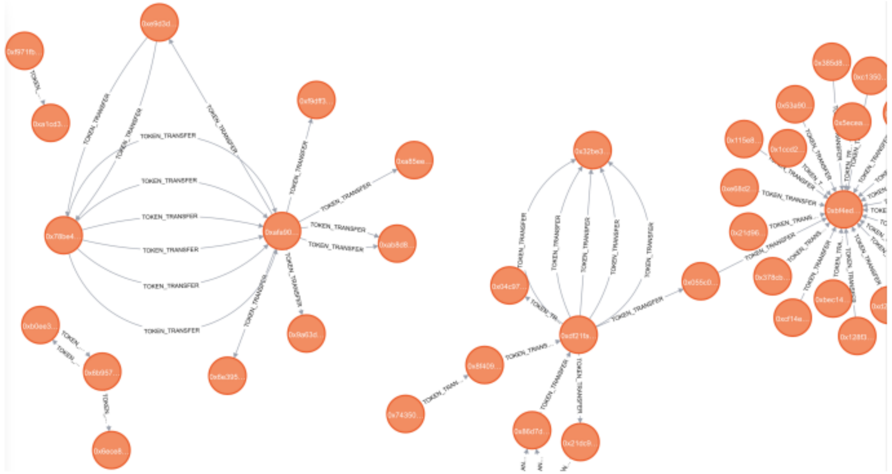
V. H. Su, S. S. Gupta, A. Khan. **Automating ETL and mining of ethereum blockchain network**, WSDM 2022.

Demonstration – Notebook Interface

```
# Connect to EtherNet Core
from ethernet.client import Client
ec = Client(core_host="192.168.1.99",
            core_grpc_port=9090, core_http_port=8080)

# Create a token_transfers graph given a block range
response = ec.create_token_transfers_graph(2000000, 2000500)

# Switch to the Neo4j graph that has just been created
dbs = ec.get_databases()
ec.switch_database(dbs[0])
```



Output visualization for the constructed graph in Neo4J

V. H. Su, S. S. Gupta, A. Khan. **Automating ETL and mining of ethereum blockchain network**, WSDM 2022.

Account Graphs: Ethereum

Graphs Constructed

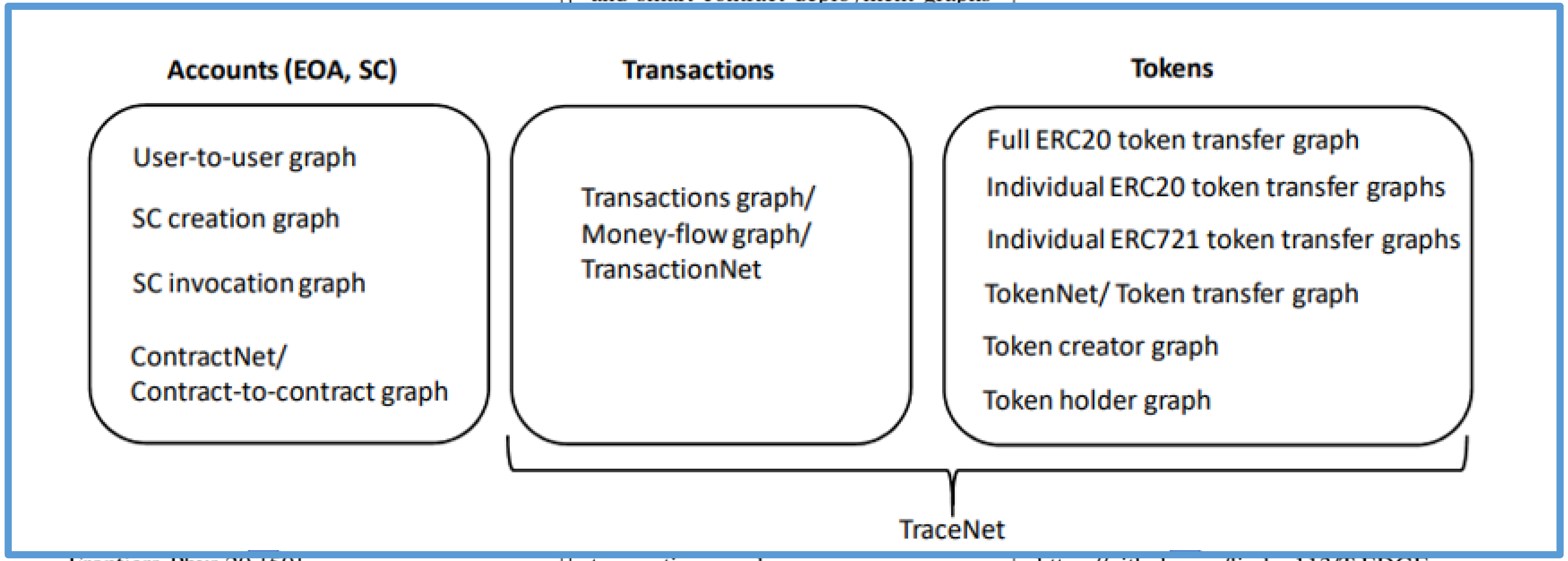
- **Survey:** A. Khan, "**Graph analysis of the Ethereum blockchain data: a survey of datasets, techniques, and future direction**", IEEE International Conference on Blockchain 2022

paper	constructed graphs	links to data and/or code
INFOCOM18 [36]	money flow graph, contract creation graph, contract invocation graph	https://github.com/brokendragon /Ethereum_Graph_Analysis
PLOS ONE18 [37]	transaction graph	https://dataverse.harvard.edu/dataset.xhtml? persistentId=doi:10.7910/DVN/XIXSPR
Complex Sys18 [38]	(full) ERC20 tokens transfer graph	not given
NTMS18 [39]	user-to-user, user-to-smart contract, and smart contract deployment graphs	not given
FC19 [40]	(individual) ERC20 token transfer graphs	not given
ICDMW19 [41]	Storj token transfer graph	not given
Appl. Netw. Sci.19 [42]	transaction graph	not given
Inf. Sci.19 [43]	transaction graph	not given
WWW20a [44]	trace graph, contract graph, transaction graph, token graph	https://github.com/sgsourav /blockchain-network-analysis
SDM20 [45]	(individual) ERC20 token transfer graphs	https://github.com/yitao416/EthereumCurve
WWW20b [23]	ERC20 token creator, holder, and transfer graphs	http://xblock.pro/#/
Sci Rep20 [46]	(individual) ERC20 token transfer graphs	not given
ACM Meas. Anal. Comput. Syst.20 [47]	ERC20 token creator, holder, and transfer graphs for counterfeit tokens	not given
Concurr. Comput. Pract. Exp.20 [48]	transaction graph	not given
IEEE Trans. Circuits Syst.20 [49]	transaction graph	https://github.com/lindan113/T-EDGE
Frontiers Phys.20 [50]	transaction graph	https://github.com/lindan113/T-EDGE
J. Complex Networks20 [51]	transaction graph	not given
Networking20 [9]	user-to-user, contract-to-contract, and user-contract graphs	not given
SBP-BRiMS20 [52]	(full) ERC20 tokens transfer graph	not given
WWW21 [8]	trace graph, contract graph, transaction graph, token graph	https://github.com/LinZhao89 /Ethereum-analysis
ECML PKDD21 [10]	(individual) token transfer graphs, stacked as a multi-layer network	https://github.com/tdagraphs
PAKDD21 [53]	transaction graph	https://github.com/fpour/SigTran
ACM Trans. Internet Techn.21 [55]	transaction graph	http://xblock.pro/#/
Blockchain21 [56]	(individual) ERC721 token transfer graphs	https://github.com/epfl-scistimm /2021-IEEE-Blockchain
IEEE Trans. Syst. Man Cybern. Syst.22 [54]	transaction graph	http://xblock.pro/#/

Graphs Constructed

- **Survey:** A. Khan, " **Graph analysis of the Ethereum blockchain data: a survey of datasets, techniques, and future direction** ", IEEE International Conference on Blockchain 2022

paper	constructed graphs	links to data and/or code
INFOCOM18 [36]	money flow graph, contract creation graph, contract invocation graph	https://github.com/brokendragon/Ethereum_Graph_Analysis
PLOS ONE18 [37]	transaction graph	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/XIXSPR
Complex Sys18 [38]	(full) ERC20 tokens transfer graph	not given
NTMS18 [39]	user-to-user, user-to-smart contract, and smart contract deployment graphs	not given
Frontiers Phys.20 [50]	transaction graph	https://github.com/lindan113/T-EDGE
J. Complex Networks20 [51]	transaction graph	not given
Networking20 [9]	user-to-user, contract-to-contract, and user-contract graphs	not given
SBP-BRiMS20 [52]	(full) ERC20 tokens transfer graph	not given
WWW21 [8]	trace graph, contract graph, transaction graph, token graph	https://github.com/LinZhao89/Ethereum-analysis
ECML PKDD21 [10]	(individual) token transfer graphs, stacked as a multi-layer network	https://github.com/tdagraphs
PAKDD21 [53]	transaction graph	https://github.com/fpour/SigTran
ACM Trans. Internet Techn.21 [55]	transaction graph	http://xblock.pro/#/
Blockchain21 [56]	(individual) ERC721 token transfer graphs	https://github.com/epfl-scistimm/2021-IEEE-Blockchain
IEEE Trans. Syst. Man Cybern. Syst.22 [54]	transaction graph	http://xblock.pro/#/



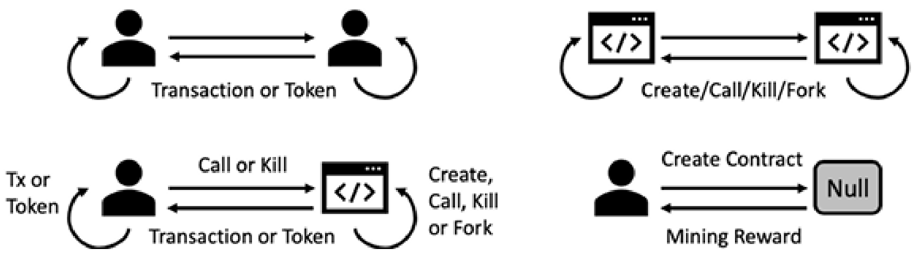
- Static graphs
- Dynamic graphs
- Temporal snapshot graphs
- Directed graphs
- Weighted graphs (?weight)
- Simple and multi-graphs
- Attributed graphs
- Multi-layer networks

Graphs between Accounts:

- Ethereum has two types of accounts:

  - **Externally owned accounts (EOAs)** are accounts controlled by private keys. If a participant own the private key of an EOA, the participant has the ability to send ether and messages from it.

  - **Smart contract code controlled accounts** have their own code, and are controlled by the code.



- **User-to-User Graph**
- **Smart Contract Creation Graph**
- **Smart Contract Invocation Graph**
- **ContractNet/ Contract-to-Contract Graph**

- T. Chen, Y. Zhu, Z. Li, J. Chen, X. Li, X. Luo, X. Lin, and X. Zhang, "**Understanding Ethereum via graph analysis**," in INFOCOM, 2018.
- A. Anoaica and H. Levard, "**Quantitative description of internal activity on the Ethereum public blockchain**," in NTMS, 2018.
- Q. Bai, C. Zhang, Y. Xu, X. Chen, and X. Wang, "**Evolution of Ethereum: a temporal graph perspective**," in IFIP Net. Conf., 2020.
- X. T. Lee, A. Khan, S. S. Gupta, Y. H. Ong, and X. Liu, "**Measurements, analyses, and insights on the entire Ethereum blockchain network**," in WWW, 2020.
- L. Zhao, S. S. Gupta, A. Khan, and R. Luo, "**Temporal analysis of the entire Ethereum blockchain network**," in WWW, 2021.

Sixth International Workshop on Foundations and Applications of Blockchain

In conjunction with VLDB'24

August 30th, 2024

Graphs Based on Transaction of Ether:

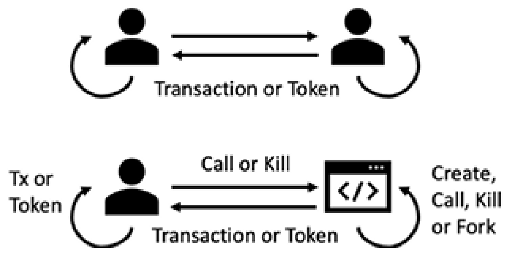
○ **Regular**, or **external transaction** denotes a transaction with the sender address being an EOA.

○ **Internal transaction** refers to a transfer that occurs when the sender address is a smart contract, e.g., a smart contract calling another smart contract or an EOA.

○ **Token transfer** is an event log for transfer of tokens only.

- ➢Token transfers can be considered as internal transactions. Internal transactions are not broadcast to the network in the form of regular transactions.

○ **Transaction Graph/ Money Flow Graph/ TransactionNet**



○ T. Chen, Y. Zhu, Z. Li, J. Chen, X. Li, X. Luo, X. Lin, and X. Zhang, "**Understanding Ethereum via graph analysis**," in INFOCOM, 2018.

○ J. Liang, L. Li, and D. Zeng, "**Evolutionary dynamics of cryptocurrency transaction networks: an empirical study**," PLoS ONE, vol. 13, no. 8, p. e0202202, 2018.

○ D. Guo, J. Dong, and K. Wang, "**Graph structure and statistical properties of Ethereum transaction relationships**," Inf. Sci., vol. 492, pp. 58–71, 2019.

○ S. Ferretti and G. D'Angelo, "**On the Ethereum blockchain structure: a complex networks theory perspective**," Concurr. Comput. Pract. Exp., vol. 32, no. 12, 2020.

○ D. Lin, J. Wu, Q. Yuan, and Z. Zheng, "**Modeling and understanding Ethereum transaction records via a complex network approach**," IEEE Trans. Circuits Syst., vol. 67-II, no. 11, pp. 2737–2741, 2020.

○ X. T. Lee, A. Khan, S. S. Gupta, Y. H. Ong, and X. Liu, "**Measurements, analyses, and insights on the entire Ethereum blockchain network**," in WWW, 2020.

○ L. Zhao, S. S. Gupta, A. Khan, and R. Luo, "**Temporal analysis of the entire Ethereum blockchain network**," in WWW, 2021.

Graphs Based on Transfer of Tokens:

- Full ERC20 token transfer graph
- Individual ERC20 token transfer graphs
- Individual ERC721 token transfer graphs
- TokenNet/ Token transfer graph
- Token creator graph
- Token holder graph

- S. Somin, G. Gordon, and Y. Altshuler, "**Network analysis of ERC20 tokens trading on Ethereum blockchain**," in Complex Systems, 2018.
- F. Victor and B. K. L¨uders, "**Measuring ethereum-based ERC20 token networks**," in Financial Cryptography and Data Security, 2019.
- Y. Chen and H. K. T. Ng, "**Deep learning Ethereum token price prediction with network motif analysis**," in ICDM Workshops, 2019.
- W. Chen, T. Zhang, Z. Chen, Z. Zheng, and Y. Lu, "**Traveling the token world: A graph analysis of Ethereum ERC20 token ecosystem**," in WWW, 2020
- Y. Li, U. Islambekov, C. G. Akcora, E. Smirnova, Y. R. Gel, and M. Kantarcioglu, "**Dissecting Ethereum blockchain analytics: what we learn from topology and geometry of the Ethereum graph**?" in SDM, 2020.
- B. Gao, H. Wang, P. Xia, S. Wu, Y. Zhou, X. Luo, and G. Tyson, "**Tracking counterfeit cryptocurrency end-to-end**," Proc. ACM Meas. Anal. Comput. Syst., vol. 4, no. 3, pp. 50:1–50:28, 2020.
- X. T. Lee, A. Khan, S. S. Gupta, Y. H. Ong, and X. Liu, "**Measurements, analyses, and insights on the entire Ethereum blockchain network**," in WWW, 2020.
- L. Zhao, S. S. Gupta, A. Khan, and R. Luo, "**Temporal analysis of the entire Ethereum blockchain network**," in WWW, 2021.
- D. Ofori-Boateng, I. Segovia-Dominguez, C. G. Akcora, M. Kantarcioglu, and Y. R. Gel, "**Topological anomaly detection in dynamic multilayer blockchain networks**," in ECML PKDD, 2021.
- S. Casale-Brunet, P. Ribeca, P. Doyle, and M. Mattavelli, "**Networks of Ethereum non-fungible tokens: a graph-based analysis of the ERC-721 ecosystem**," in Blockchain, 2021.

Graph Analysis on Blockchain Graphs

- X. T. Lee, A. Khan, S. S. Gupta, Y. H. Ong, and X. Liu, "**Measurements, analyses, and insights on the entire Ethereum blockchain network**," in WWW, 2020.
- L. Zhao, S. S. Gupta, A. Khan, and R. Luo, "**Temporal analysis of the entire Ethereum blockchain network**," in WWW, 2021.

Ethereum Network Properties

- Basic Network Properties
- Local Network Properties
- Global Network Properties
- Temporal Network Properties

Motivation

- Blockchain is a fascinating ecosystem of humans and autonomous agents.
- Not like conventional social networks, where the players are human users.
- Not like cryptocurrencies, where all interactions are transfer of value/asset.

Blockchain network is closer to the Internet or Web, where users interact with one another, as well as with programs.

We study a public permissionless blockchain network as a **complex system**, and we choose **Ethereum**, the most prominent blockchain network, for this purpose.

AALBORG UNIVERSITY

Ethereum

- Introduced an automation layer on top of a blockchain through contracts.
- Facilitates a decentralized computing environment across the blockchain.

Transaction-based state machine. Global state made up of accounts. Transfer of value/information between accounts cause transitions in the state. Recorded in the blockchain.

We target the **network of interactions** between the User and Contract accounts that make up the global state of Ethereum, and study them as **complex systems**.

Networks

1 TraceNet

v : user and smart contract addresses
a : all successful traces/transactions

2 ContractNet

v : only smart contract addresses
a : all successful traces/messages

3 TransactionNet

v : user and smart contract addresses
a : all successful transactions by users

4 TokenNet

v : user and smart contract addresses
a : all successful transaction of tokens

While **TraceNet** presents a global view of interactions, **ContractNet** focusses on the multi-agent network of contracts. While **TransactionNet** depicts all of basic ether transactions, **TokenNet** focusses on the rich and diverse token ecosystem.

Network Data

Source : Google Cloud Platform BigQuery
bigquery-public-data.Ethereum_blockchain.

Data extracted/mined : Block #0 till #7185508
Blocks recorded upto 2019-02-07 00:00:27 UTC
Seven different tables in the Ethereum dataset.

	Size of Dataset	Row Count
blocks	8 GB	7 185 509
contracts	15.7 GB	12 950 995
transactions	190 GB	388 018 489
traces	500 GB	974 766 498
logs	160 GB	289 552 838
tokens	11.4 MB	126 181
token transfers	58 GB	173 421 940

Data cleaning : Removing failed traces and handling Null addresses appropriately.

Basic Network Properties

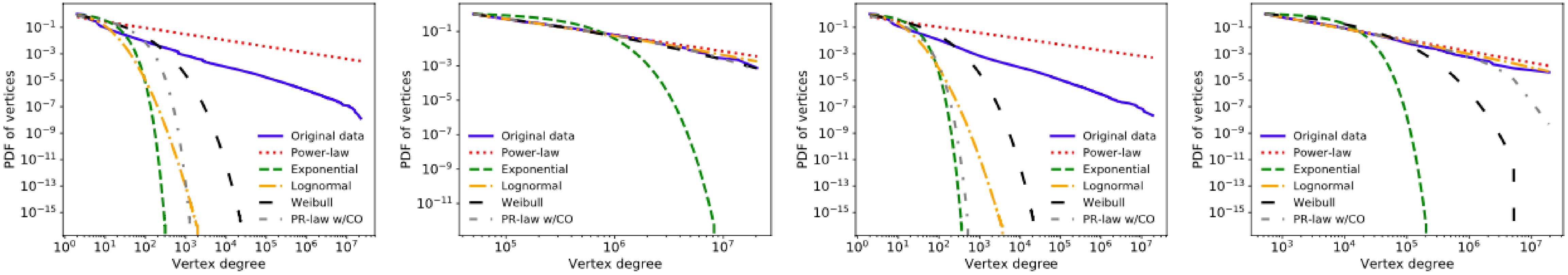
Vertices and Arcs, Self-Loops and Density

	# Vertices	MultiDigraph			Simple, undirected graph		
		# Arcs	# Self-loops (% of Arcs)	Density	# Arcs	# Self-loops (% of Arcs)	Density
TraceNet	75 807 179	768 813 599	3 036 915 (0.40%)	1.34×10^{-7}	191 901 321	178 241 (0.09%)	0.67×10^{-7}
ContractNet	11 332 750	317 967 546	**2 521 670 (0.79%)**	$\mathbf{24.8\times10^{-7}}$	19 608 452	**63 234** (0.32%)	3.05×10^{-7}
TransactionNet	45 527 529	388 018 489	515 245 (0.13%)	1.87×10^{-7}	128 368 878	115 007 (0.09%)	1.24×10^{-7}
TokenNet	30 429 099	173 421 940	326 557 (0.19%)	$\mathbf{1.87\times10^{-7}}$	93 844 445	36 950 (0.04%)	$\mathbf{2.03\times10^{-7}}$

We observe that self-loop percentage in ContractNet MultiDiGraph is significantly higher than that in the three other networks. Moreover, the number of self-loops in its MultiDiGraph is **almost 40 times** than that in its own simple, undirected graph, indicating that a lot of **smart contracts make multiple calls to itself**.

Local Network Properties

Vertex Degree Distribution



(a) TraceNet (b) ContractNet (c) TransactionNet (d) TokenNet

We compare power-law distribution model against (i) exponential, (ii) log-normal, (iii) power-law with exponential cutoff, and (iv) stretched exponential or Weibull.

We see that for our larger networks, TraceNet and TransactionNet, three of the four alternative heavy-tailed distributions are better fit than the power-law.

Local Network Properties

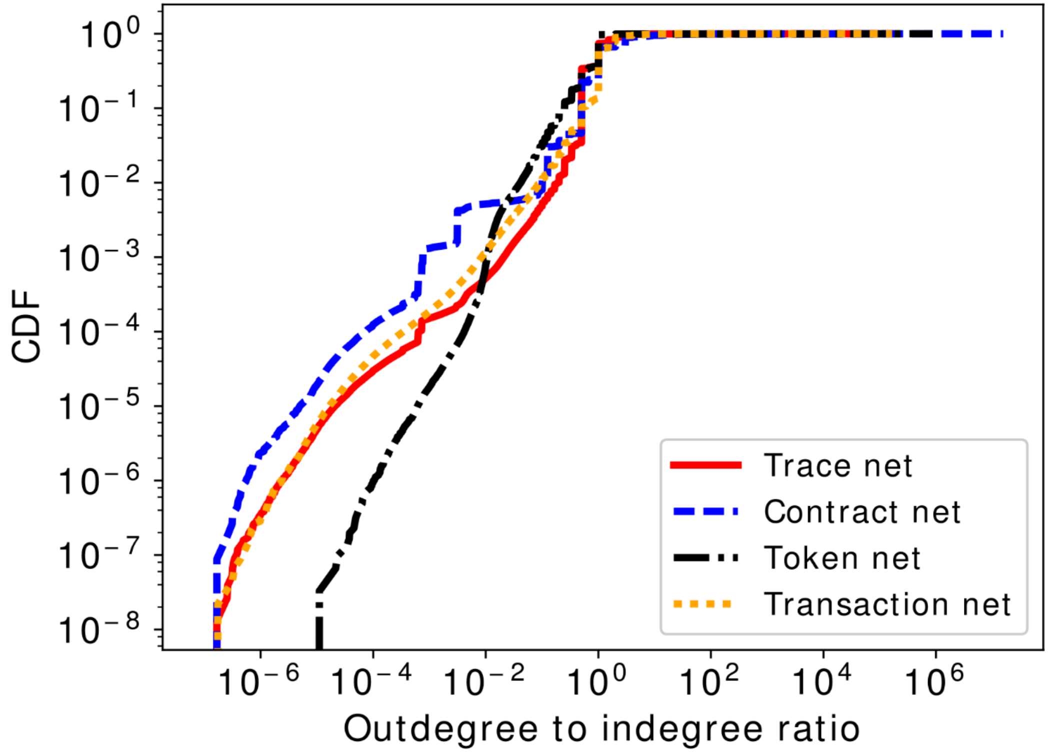
Indegree and Outdegree Correlation

Indegree and outdegree of vertices in the four network MultiDiGraphs.

≈ 50% have similar in and out.

≈ 30% have significantly higher in (ICO smart contracts appear a lot in the to_address).

≈ 20% have significantly higher out (mining pools and mixers generally appear a lot in the from_address).



This is similar to the Web, involving hubs and authorities, and it is unlike the case of standard social networks.

Global Network Properties

Reciprocity and Assortativity

Reciprocity: Measure of vertices being mutually linked in network.

Assortativity: Measure of vertices being linked to similar-degree ones.

Network (#vertices, #arcs)	Reciprocity	Assortativity
TraceNet (76M, 198M)	0.06	-0.13
ContractNet (11M, 22M)	**0.21**	**-0.64**
TransactionNet (46M, 130M)	0.03	-0.12
TokenNet (30M, 95M)	0.03	-0.13

Unlike social networks, all four of our blockchain networks are Disassortative.
Negative assortativity implies relatively more scenarios of addresses (vertices) with different degrees transacting with each other in the blockchain networks.

Global Network Properties

Strong and Weakly Connected Components

Simple, directed networks (#vertices, #arcs)	# Strongly connected components	Largest strongly connected component (#vertices, #arcs)	# Weakly connected components	Largest weakly connected component (#vertices, #arcs)
TraceNet (**76M, 198M**)	35 215 962	40M, 116M	**7 324**	**76M, 192M**
ContractNet (**11M, 22M**)	**9 013 144**	2M, 4M	**12 555**	**11M, 20M**
TransactionNet (**46M, 130M**)	15 560 831	30M, 76M	**8 181**	**46M, 128M**
TokenNet (**30M, 95M**)	16 980 001	13M, 56M	**54 271**	**30M, 94M**

Number of WCC is significantly lesser than the number of SCC in their respective networks, due to lesser bidirectional edges between majority pairs of vertices.

ContractNet has the least # of SCC in the networks, indicating relatively stronger connectivity within smart contracts. Similar to the Web, the blockchain networks have a single, large SCC, with about 98% of the remaining vertices within reach.

Global Network Properties

Core Decomposition

k-core is the maximal subgraph, where each vertex is connected to at least **k** other vertices within the subgraph.

Largest Weakly Connected Component (#vertices, #arcs)	# Cores	Innermost core (#vertices, #arcs)
TraceNet (76M, 192M)	98	(221, 12 058)
ContractNet (11M, 20M)	**264**	**(1071, 143 352)**
TransactionNet (46M, 128M)	105	(682, 55 926)
TokenNet (30M, 94M)	**218**	(475, 57 124)

ContractNet and TokenNet have larger core indices for vertices in the innermost cores, indicating higher density of their innermost cores. ContractNet's innermost core is the largest, implying more vertices participating in denser substructures.

Global Network Properties

Sixth International Workshop on Foundations and Applications of Blockchain

In conjunction with VLDB'24

August 30th, 2024

Triangles, Transitivity, Clustering Coefficients

Transitivity is quite low. This suggests that in the blockchain networks, we do not have a conducive environment for creation of triangles. **Indeed, non-social networks have lower transitivity coefficients.**

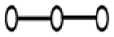
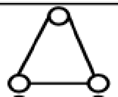
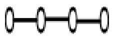
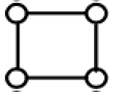
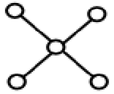
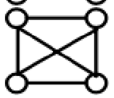
	Largest strongly connected comp. (Simple, undirected)			Largest weakly connected comp. (Simple, undirected)		
	# Triangles	T	C	# Triangles	T	C
TraceNet	4 008 794	10.0×10^{-7}	0.099	5 813 165	1.2×10^{-7}	0.077
ContractNet	405 265	$\mathbf{38.0\times10^{-7}}$	**0.212**	871 359	6.7×10^{-7}	0.078
TransactionNet	1 908 138	8.3×10^{-7}	0.064	4 550 517	$\mathbf{12.4\times10^{-7}}$	0.100
TokenNet	2 803 894	8.6×10^{-7}	0.209	5 296 640	5.5×10^{-7}	**0.175**

High-degree vertices are often “loner-star”, that is, connected to mostly low-degree vertices, resulting in lack of community structure in blockchain graphs.

Global Network Properties

Higher-Order Motifs Counting

The most frequent motifs in the blockchain graphs are primarily chain and star-shaped. Counts for more complex patterns, e.g., cliques and cycles, are less.

	#	Motif density		#	Motif density
	13 669	1×10^{-1}		2 214	2×10^{-2}
	17 081	3×10^{-3}		60 297	9×10^{-3}
	387 816	12×10^{-3}		2 578	4×10^{-4}

We check the density of a motif, the ratio of its count to its count in a complete graph having same number of vertices as the innermost core. The densities for more complex patterns are quite less, indicating lack of community structure.

Global Network Properties

Articulation points, Adhesion, Cohesion, Average path lengths, Radius, Diameter

	# Articulation points (% of all vertices)	Largest strongly conn. comp.		Largest weakly conn. comp.		Largest weakly connected component		
		Adhesion	Cohesion	Adhesion	Cohesion	Avg. path length	Radius	Diameter
TraceNet	1 214 137 (1.6%)	1	1	1	1	5.25	5 002	8 267
ContractNet	28 309 (0.2%)	1	1	1	1	5.94	14	27
TransactionNet	1 337 527 (2.9%)	1	1	1	1	5.33	5 002	8 267
TokenNet	75 513 (2.5%)	1	1	1	1	3.87	82	164

Adhesion and Cohesion for all blockchain networks are 1, indicating that removal of the only one vertex or only one arc disconnects the respective SCCs and WCCs.

Interestingly, similar to social networks, blockchain graphs are also small-world. However, in both our larger networks, TraceNet and TransactionNet, there are vertices which are far apart, making the radius and the diameter quite large.

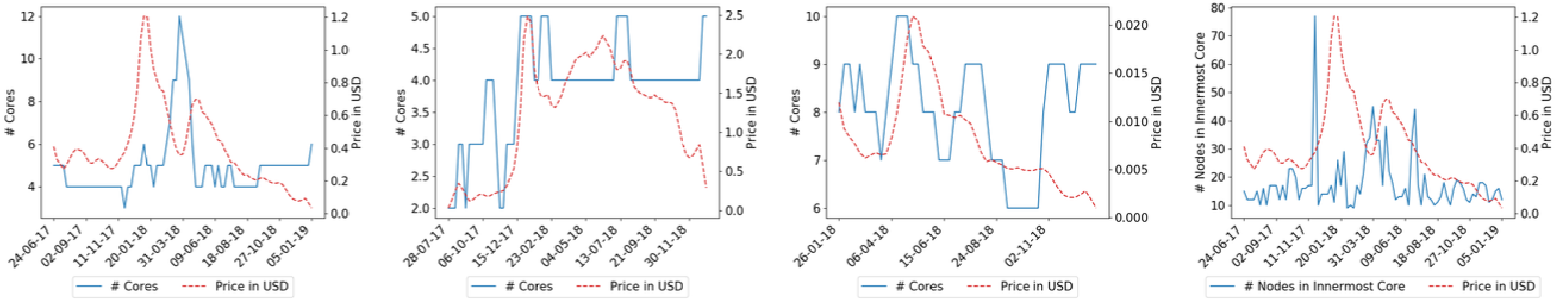
Temporal Network Properties

Sixth International Workshop on Foundations and Applications of Blockchain

In conjunction with VLDB'24

August 30th, 2024

Progress of Core Decomposition in Token Networks



(a) Bancor : Number of Cores vs. Price (b) Binance Coin : Number of Cores vs. Price (c) Zilliqa : Number of Cores vs. Price (d) Bancor : Vertices in Inner Core vs. Price

We study temporal evolution of the number of cores in token subgraphs against the corresponding evolution of price of the token in the cryptocurrency market. Observations clearly show a significant relationship between activity and price.

Summary of Observations

the Web

- In/Out-degree characteristics are very similar to the Web (hub/authority).
- The blockchain networks are disassortative, having very low transitivity.
- Complex motifs occur quite less, indicating lack of community structure.
- Removal of one vertex or arc can disconnect the entire largest SCC/WCC.

social network

- Blockchain networks are surprisingly small-world and well-connected.

both networks

- Networks contain a single, large SCC, with 98% of the vertices reachable.
- ContractNet and TokenNet yield larger core indices for vertices in the innermost cores, indicating higher density of their innermost cores.

financial

- Significant relationship between temporal relationship of inner cores of prominent token networks and the price of the tokens in the market.

https://github.com/sgsourav/blockchain-network-analysis

Motivation and Research Questions

- Investigate the **evolutionary nature of Ethereum interaction networks** from a temporal graph perspective

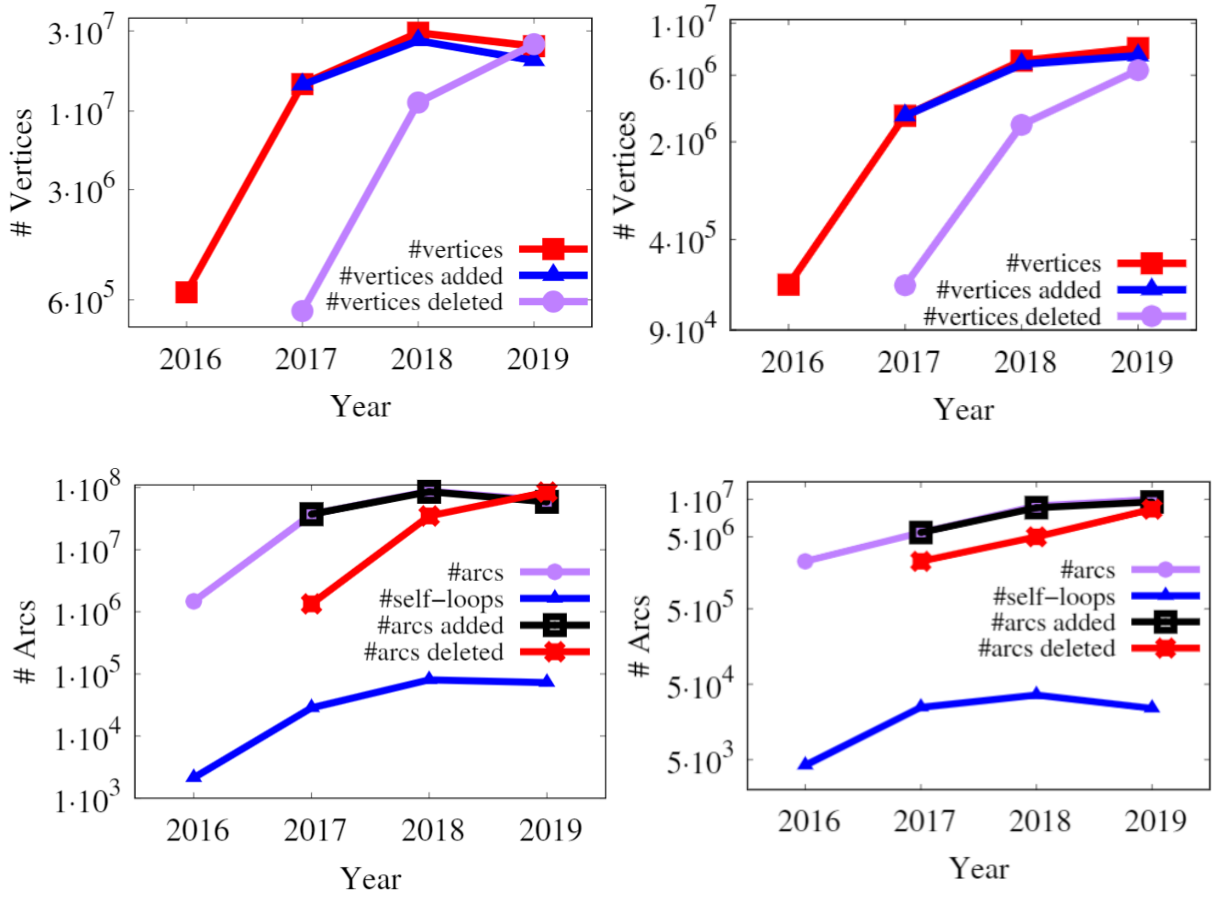
L. Zhao, S. S. Gupta, A. Khan, and R. Luo, "**Temporal analysis of the entire Ethereum blockchain network**," in WWW, 2021.

- Address 3 main questions:
  - How do Ethereum network evolve over time?
  - How network properties changes over time, what is the right "time granularity" for such temporal analysis?
  - Detect meaningful communities and forecast the survival of communities in succeeding months.

Evolution of Ethereum Network (Vertex)

- The number of new vertices and arcs added is almost of the same order of total number of vertices and arcs at that time => Ethereum interaction networks growing at a fast speed. (highly active network).

- Vertices which are disappeared keep increasing.



(a) TransactionNet (b) ContractNet

Network Growth Model

The increasing percentage (3rd column) indicates:

- As the Ethereum network matures, more accounts remain active.
- And more than half of new vertices participate in interaction with old vertices.

Table 3: TransactionNet: New vertices connecting with old vertices

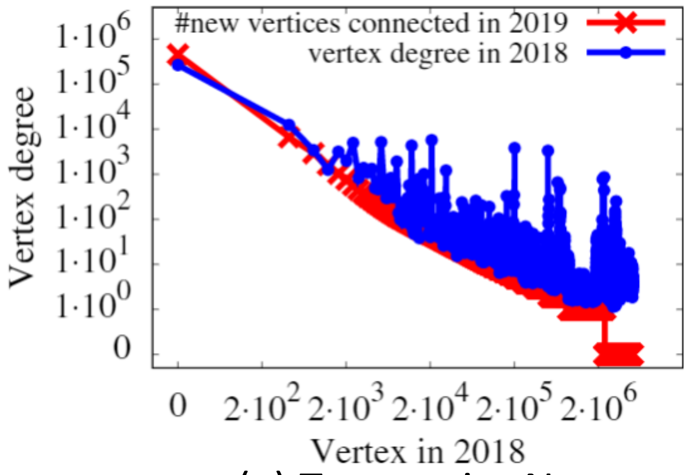
year	# old vertices	# new vertices	# new vertices with arc to old vertices (% of new vertices)	# new vertices without arc to old vertices (% of new vertices)
2017	163982	14789934	5646964 (38.18%)	9142970 (61.82%)
2018	3599770	28583252	14279239 (49.96%)	14304013 (50.04%)
2019	5060613	21240780	14807280 (69.71%)	6433500 (30.29%)

Table 4: ContractNet: New vertices connecting with old vertices

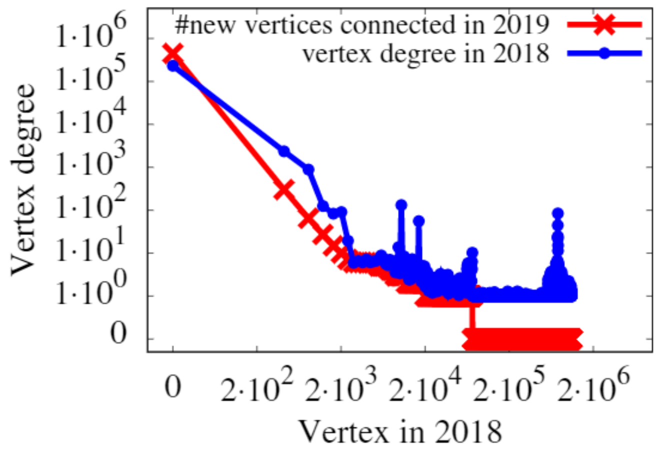
year	# old vertices	# new vertices	# new vertices with arc to old vertices (% of new vertices)	# new vertices without arc to old vertices (% of new vertices)
2017	1859	3070553	182920 (5.96%)	2887633 (94.04%)
2018	426000	7196954	2927928(40.68%)	4269026 (59.32%)
2019	1108567	8266061	6086678(73.63%)	2179383 (26.37%)

Network Growth Model

- Correlation between old vertex degree in previous year (2018) to its number of new connections in the current year (2019).
- High degree vertices are highly likely to have more new vertex connections in next year.
- The observation indicates that the Ethereum graphs follow the **preferential attachment** growth model.



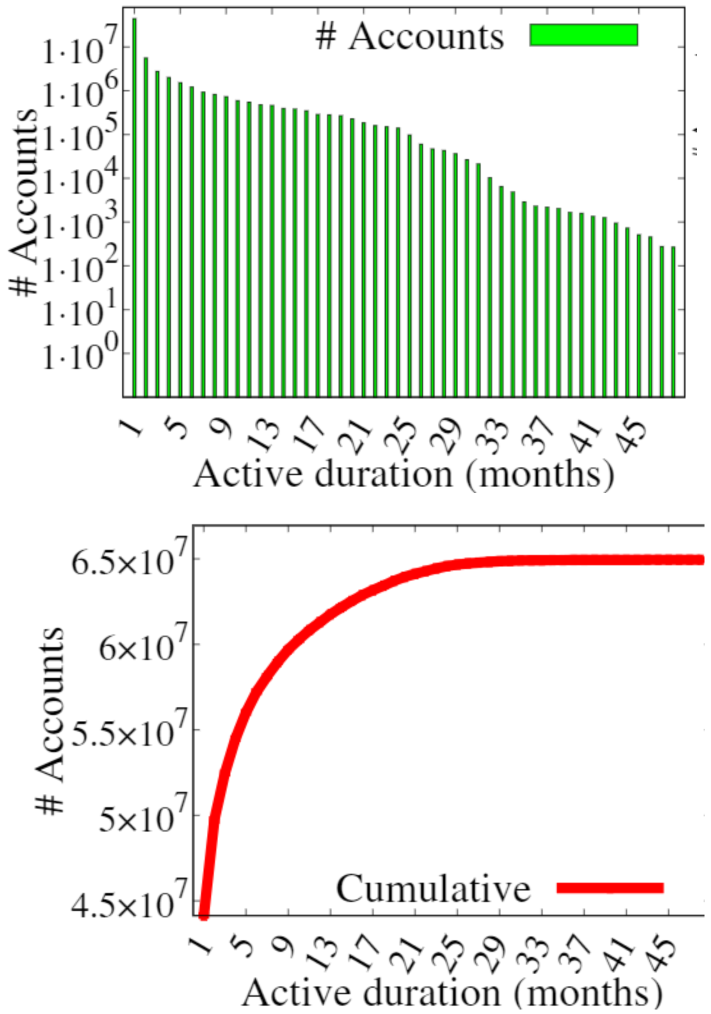
(a) TransactionNet



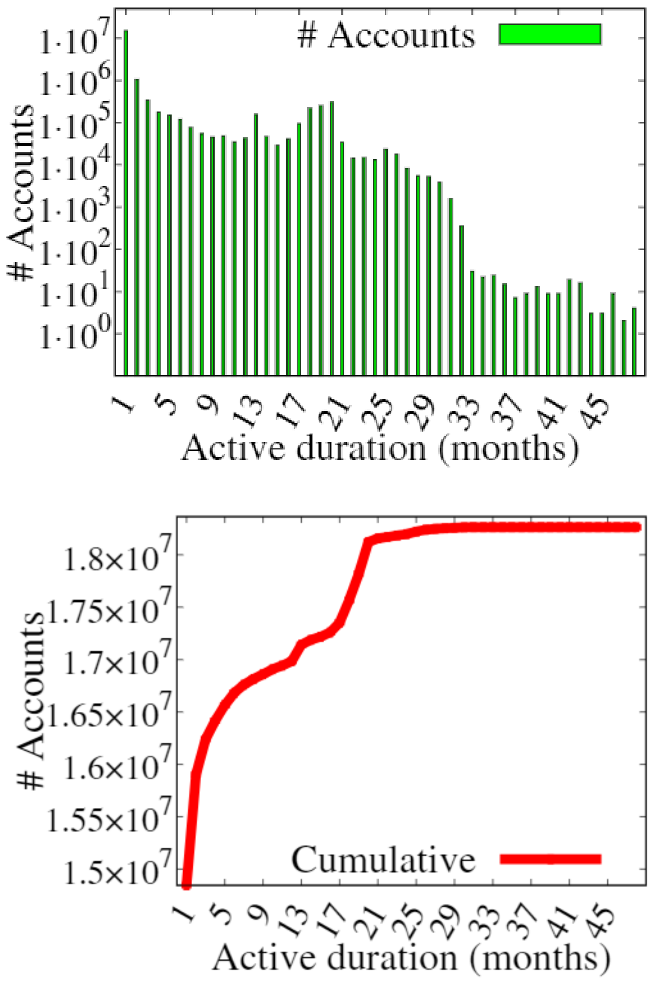
(b) ContractNet

Average Activity Period of Vertices

- **Active period** = duration (month) from its first transaction to the last transaction between Jan 2016 and Dec 2019.
- **ContractNet:** 91% has no more than 6 month active period.
- **TransactionNet:** Longer active period.
- In general, 88% of accounts have an active period of no more than 6 months, and up to 68% of accounts are only active within a month.



TransactionNet



ContractNet

Temporal Evolution of Network Properties

- Investigate network properties changes over time to understand how the network is connected and changed over time.
- Reveal any anomaly (beyond average) occurred in a specific time duration.
- A good time granularity as the shortest time duration by which we can detect an anomaly.

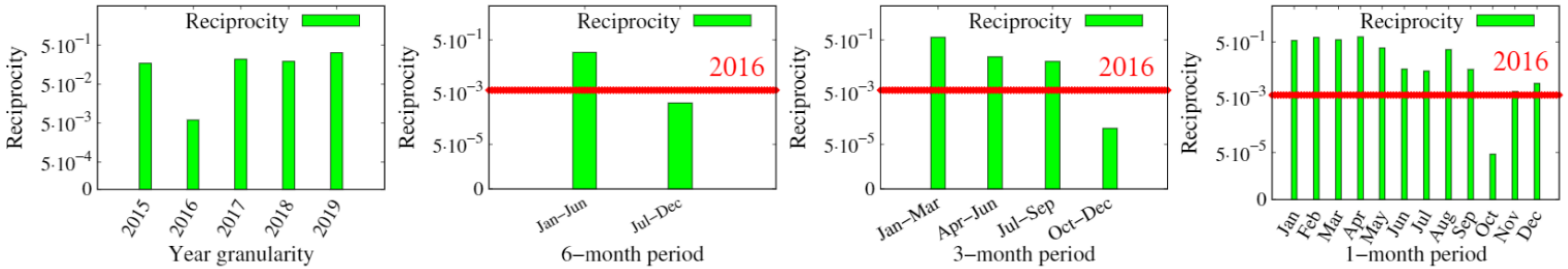


Figure 8: Time granularity analysis for reciprocity; ContractNet 2016

Temporal Evolution of Network Properties

- **Oct 2016:** Plenty of positive news on Ethereum in the media → a lot of tokens were deployed on the network, which increased the number of one-directional arcs to the token contracts.

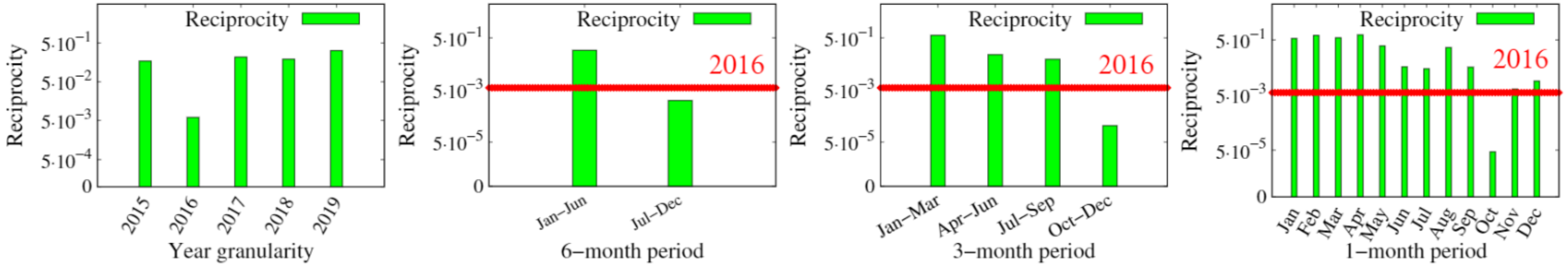
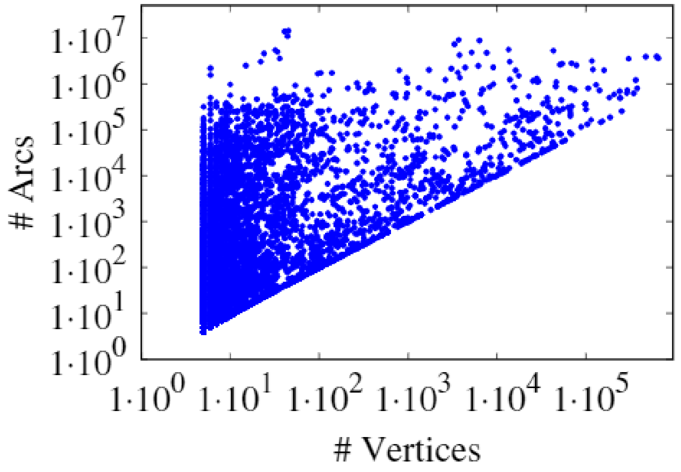


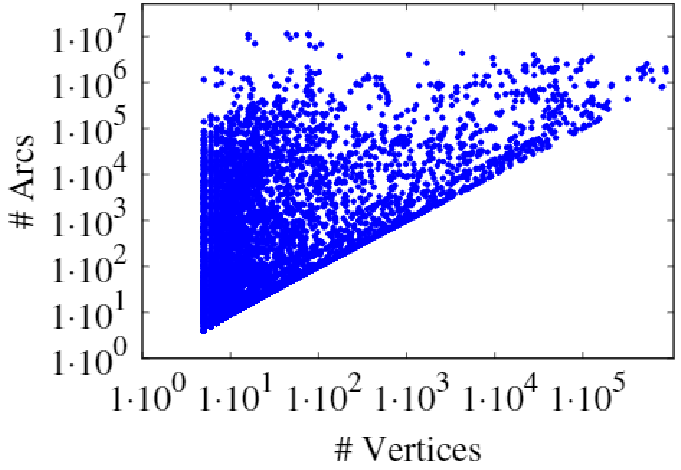
Figure 8: Time granularity analysis for reciprocity; ContractNet 2016

Detection of ContractNet Communities

- Multilevel algorithm scales well over large-scale datasets and produce good-quality communities.
- Consider multi, undirected version of graph .
- # vertices and arcs in each community obtained over ContractNet 2018 and 2019 networks.
- The size of the communities follows power-law: **a few large communities followed by a long-tail of remaining small communities.**



(a) 2018

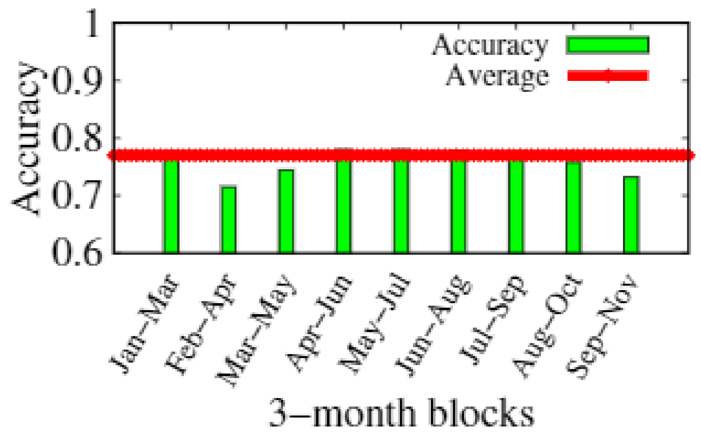


(b) 2019

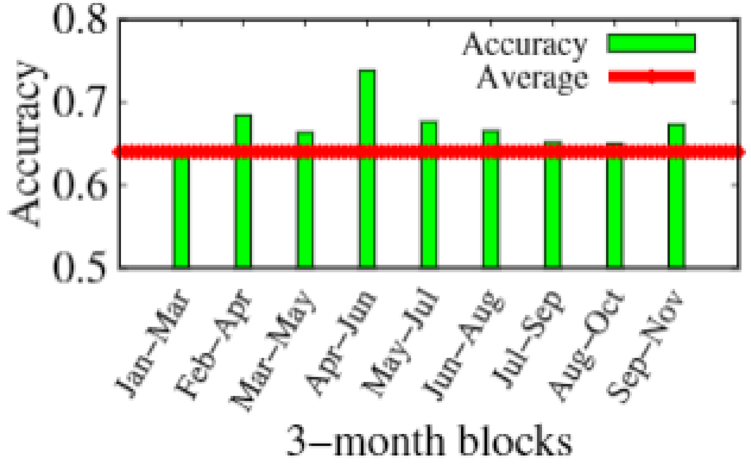
V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. 2008. **Fast unfolding of communities in large networks.** Journal of Statistical Mechanics: Theory and Experiment 2008, 10 (2008), 10008.

Community Continuation Prediction

- Data preparation: window size of 3 months and slide stride of 1 month.
- Training dataset: the network properties of communities existing in 3-month period dataset.
- Aim: predict whether the communities still exists in next 1 month.
- Model: Logistic Regression & Random Forest.



Logistic Regression prediction accuracy for ContractNet 2019



Random Forest prediction accuracy for ContractNet 2019

Summary of Observation

- Ethereum interaction network grows at a fast speed.
- Networks follow the preferential attachment growth model.
- User accounts remain active much longer than smart contracts.
- Reveal anomalies occurred in a specific time duration and correlate them with external 'real-life' aspects of network.
- Detect meaningful communities in Ethereum network using multilevel algorithm.
- Forecast the continuation of communities in succeeding months leveraging on the relevant graph properties and ML models. Achieving up to 77% correct predictions for continuation.

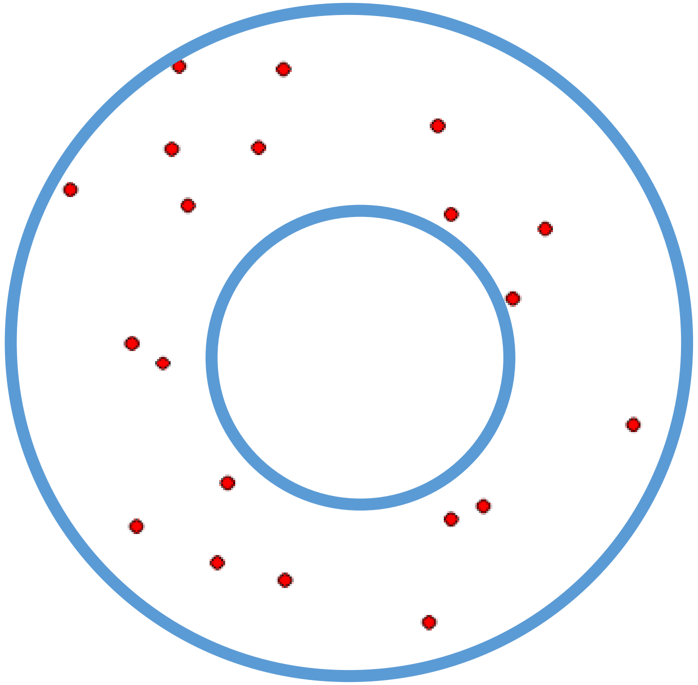
https://github.com/LinZhao89/Ethereum-analysis

Topological Data Analysis (TDA)

- Data depth (Multivariate analysis/ statistics)
- Persistent homology
- TDA mapper

F. M. Taiwo, U. Islambekov, C. G. Akcora. Explaining the Power of Topological Data Analysis in Graph Machine Learning. CoRR abs/2401.04250 (2024)

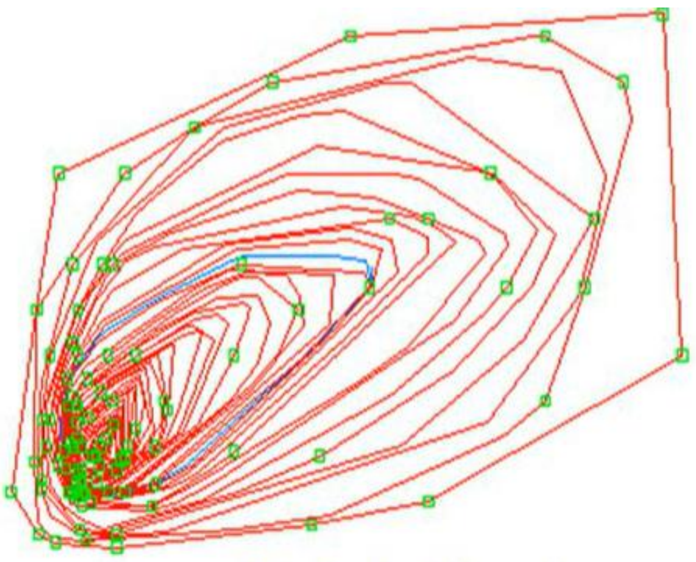
Why TDA?



What is the true shape of this data?

- capture intricate shapes and their persistence.
- robust in handling noisy and high-dimensional datasets.
- expensive computation.

Data Depth



- measures how deep a data point is relative to a data cloud.
- deals with the shape of the data.
- Nodes with high property values (e.g., large edge weights) generally have a low depth, while nodes with low property values (e.g., most blockchain nodes that trade small amounts of tokens) often have a high depth.
- Community structure around the node also plays a role.

Mahalanobis depth to the origin:

$$MhDO_F(\mathbf{x}) = \left(1 + \mathbf{x}^\top \Sigma_F^{-1} \mathbf{x}\right)^{-1}$$

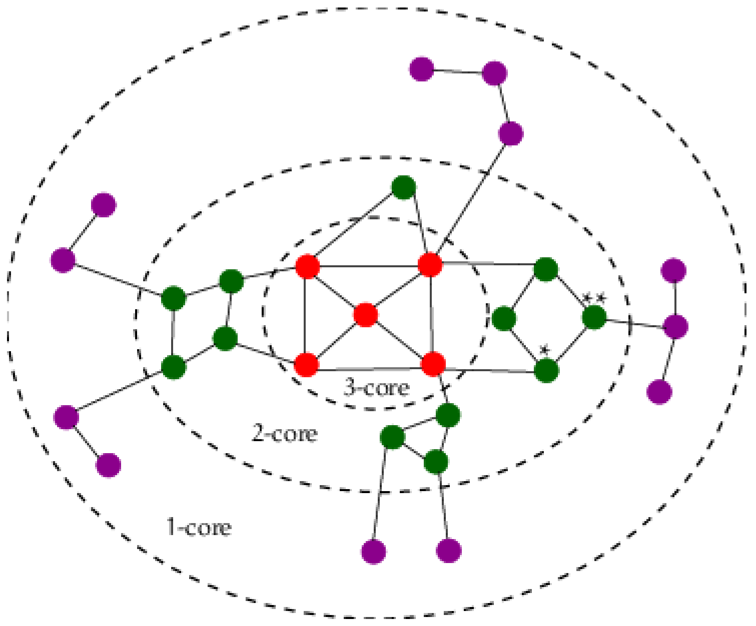
J. Zhu, A. Khan, and C. G. Akcora (2024). **Data depth and core-based trend detection on blockchain transaction networks**. Front. Blockchain

Graph Core Decomposition

Sixth International Workshop on Foundations and Applications of Blockchain

In conjunction with VLDB'24

August 30th, 2024



Classic k-core decomposition

What if a node has multiple features?

Function	Definition
$N(v)$	neighbors of v
$N_{out}(v)$	neighbors reachable with outgoing edges from v
$N_{in}(v)$	neighbors reachable with incoming edges to v
$deg(v)$	edges to/from v (Degree)
$deg_{out}(v)$	outgoing edges from v (Out-Degree)
$deg_{in}(v)$	incoming edges to v (In-Degree)
$S(v)$	sum of edge weights incident to a node (Strength)
$S_{out}(v)$	sum of outgoing edge weights (Out-Strength)
$S_{in}(v)$	sum of incoming edge weights (In-Strength)

J. Zhu, A. Khan, and C. G. Akcora (2024). **Data depth and core-based trend detection on blockchain transaction networks**. Front. Blockchain

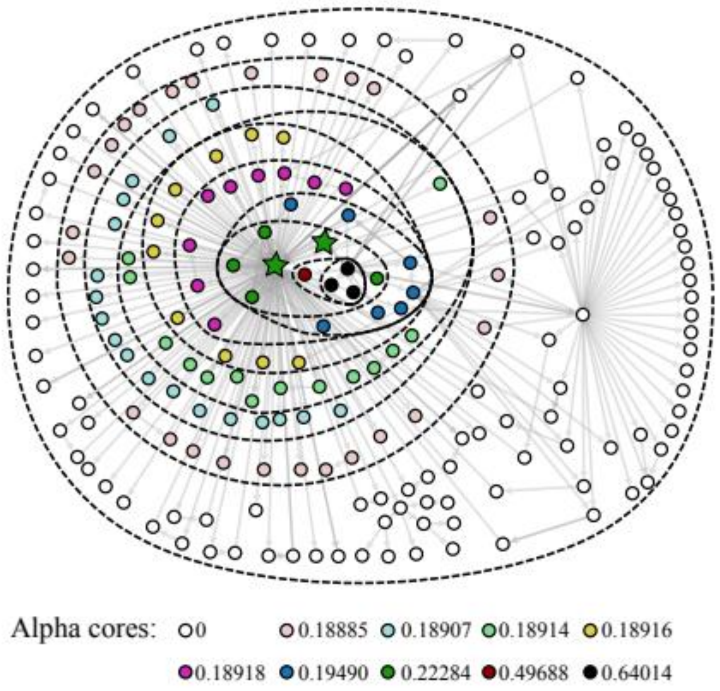
Data Depth-based Core Decomposition

- Nodes with high property values (e.g., large edge weights) generally have a low depth, while nodes with low property values (e.g., most blockchain nodes that trade small amounts of tokens) often have a high depth.

- a data depth threshold $\epsilon \in [0, 1]$ is applied to remove high-depth nodes iteratively.

- Nodes are in the $\alpha = (1 - \epsilon)$-core if their depth, relative to themselves, is no more than ϵ.

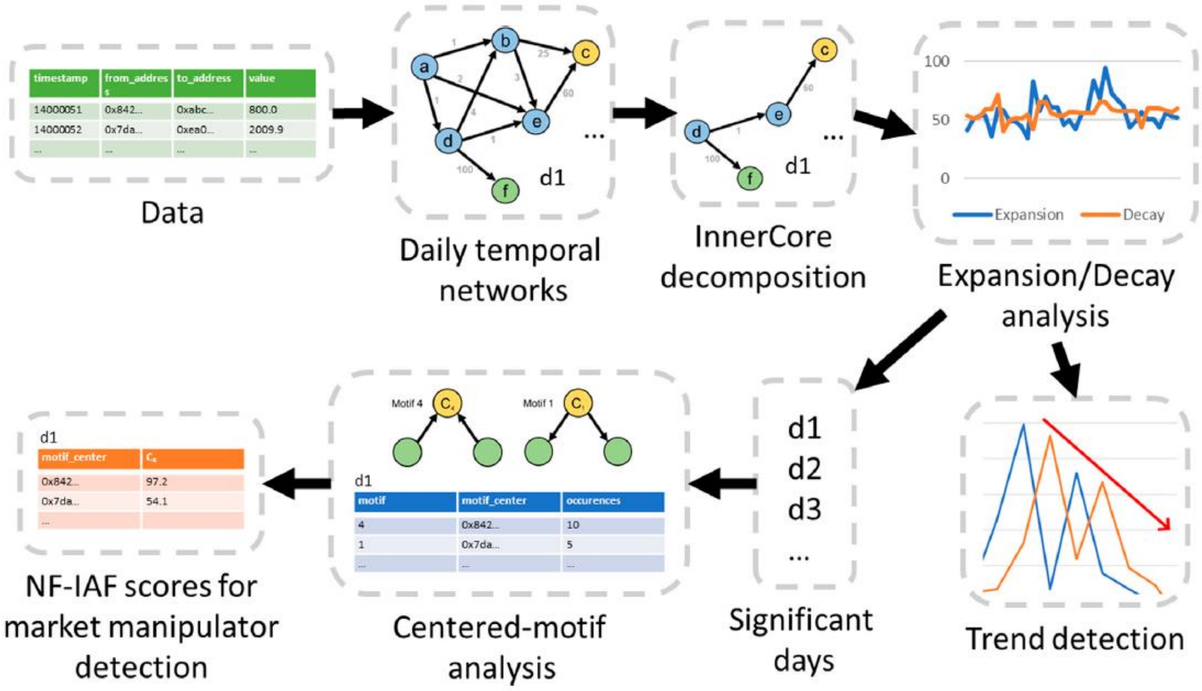
- We are interested in finding the innermost core (**innerCore**) by setting ϵ to a small value.



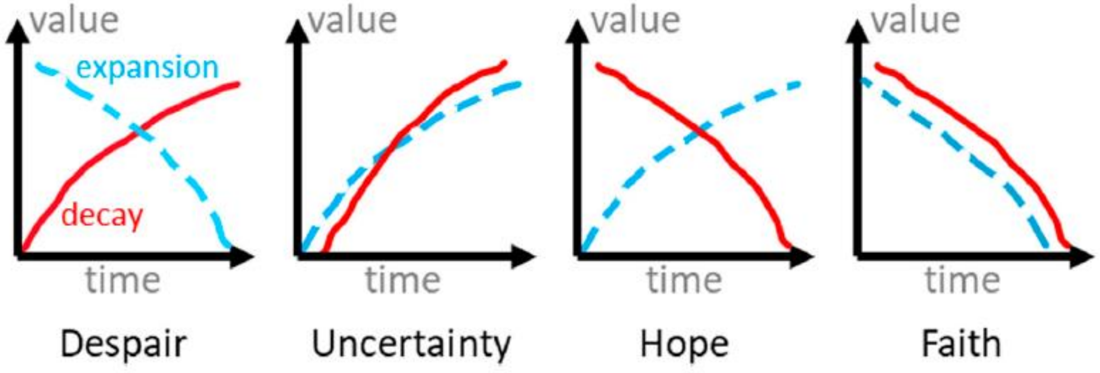
F. Victor, C. G. Akcora, Y. R. Gel, M. Kantarcioglu. AlphaCore: Data Depth based Core Decomposition. KDD 2021.

J. Zhu, A. Khan, and C. G. Akcora (2024). **Data depth and core-based trend detection on blockchain transaction networks**. Front. Blockchain

Inner Core Expansion and Decay



Flowchart of our methodology



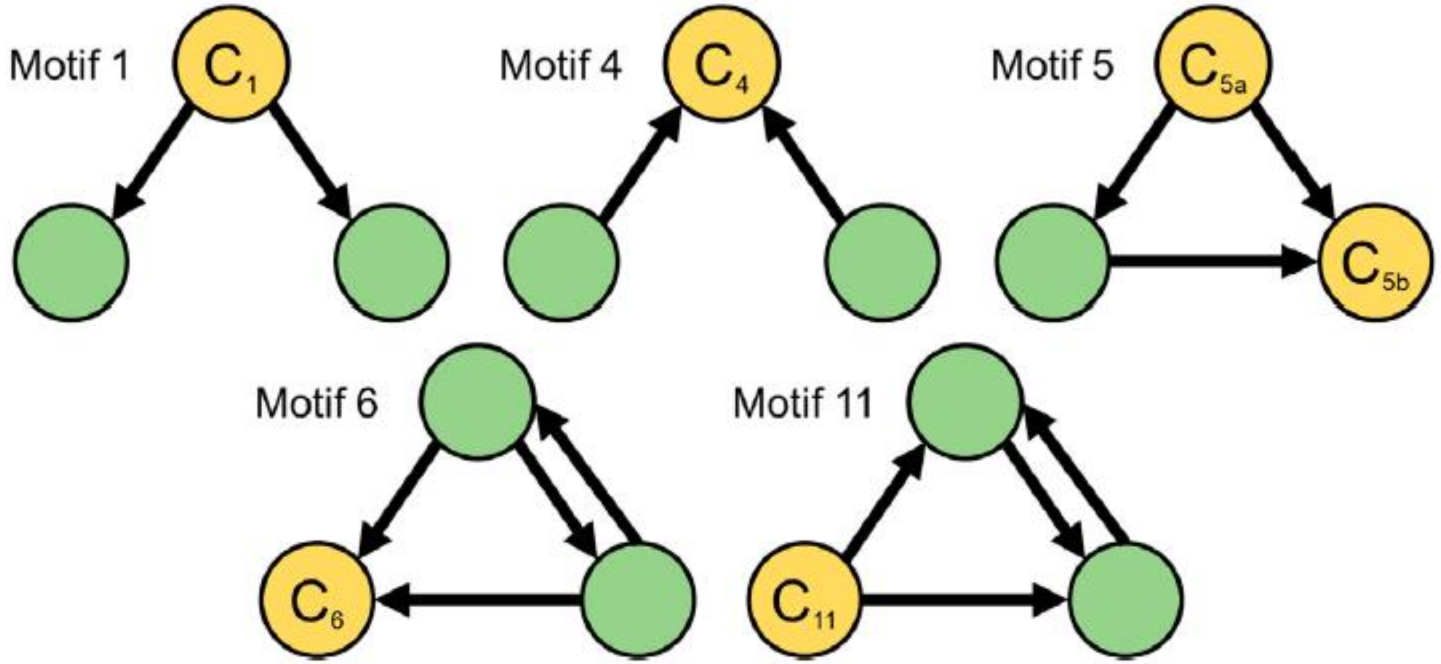
Behavioral patterns based on innerCore expansion and decay over time

$$(\text{Expansion}).\ \mathbb{E}_t = |\mathcal{V}_t^{inner} \setminus \mathcal{V}_{\bigcup(t-i)}^{inner}|$$

$$(\text{Decay}).\ \mathbb{D}_t = |\mathcal{V}_{\bigcup(t-i)}^{inner} \setminus \mathcal{V}_t^{inner}|.$$

J. Zhu, A. Khan, and C. G. Akcora (2024). **Data depth and core-based trend detection on blockchain transaction networks**. Front. Blockchain

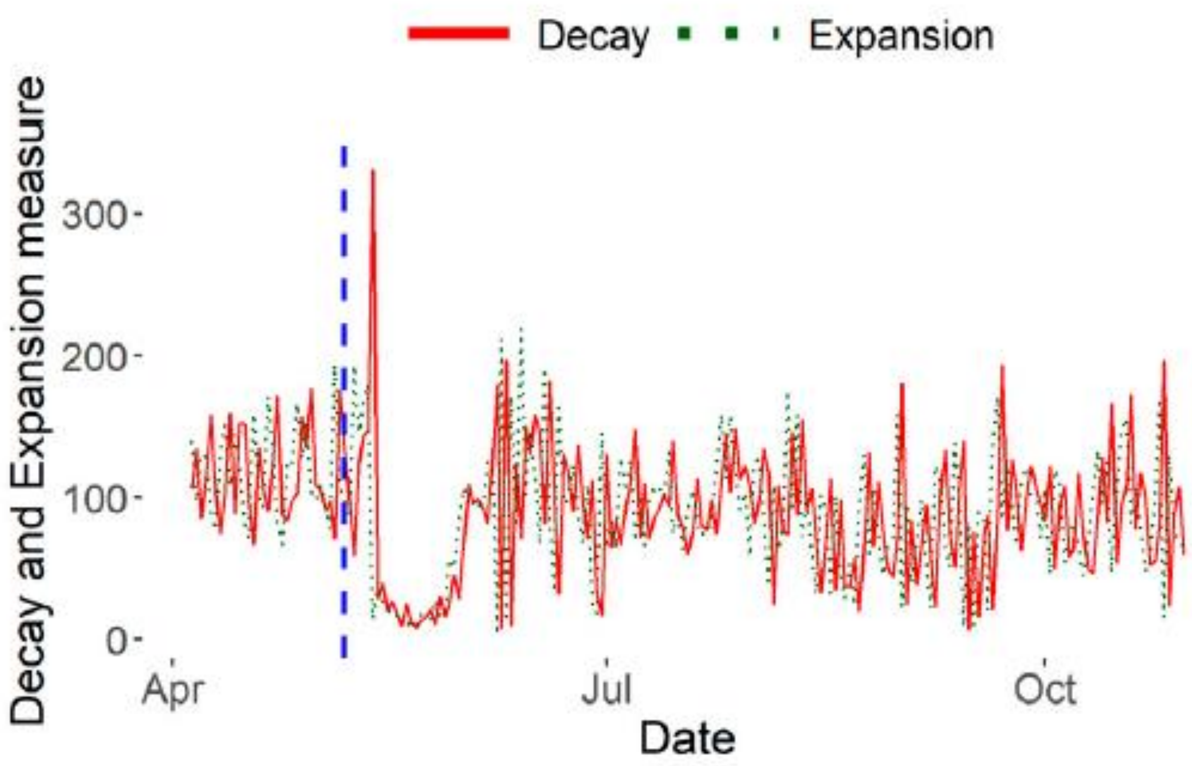
Inner Core Motif Analysis



Five 3-node motifs exhibiting buy and sell behaviors. Nodes labeled C denote the center where a center with an in-degree = 2 indicates buy behavior and an out-degree = 2 indicates sell behavior

J. Zhu, A. Khan, and C. G. Akcora (2024). **Data depth and core-based trend detection on blockchain transaction networks**. Front. Blockchain

The Collapse of LunaTerra



Stablecoin decay and expansion measures. On May 8 (shown with the vertical blue line), UST loses its $1 peg and falls to as low as 35 cents.

- For approximately 2 weeks afterward, a consistent **behavioral pattern of faith** is characterized by low expansion and low decay. During this period, few new traders entered or left the stablecoin network. There was still faith in the remaining traders that perhaps a large stablecoin such as UST could rebound and restore its peg with USD and thus, they refrained from engaging in any transactions.

- There is a **delayed reaction** from traders when a significant unannounced event occurs due to indecision, and there is a general **trend of inactivity** in the following period.

J. Zhu, A. Khan, and C. G. Akcora (2024). **Data depth and core-based trend detection on blockchain transaction networks**. Front. Blockchain

The Collapse of LunaTerra

- Before the LunaTerra collapse, nodes exhibiting both high selling and buying behaviors, could have influenced the initial phase of the crash.

- We identify the addresses that occurred most frequently as motif centers in InnerCores.

- Exchanges are well-known intermediary hubs to facilitate transfers between traders, hence not very interesting in our context.

- addresses that are not exchanges are mostly owned by traders and thus, the existence of such addresses as motif centers is interesting.

	# Unique addresses	# Exchange addresses
Motif 1	1,221	15
Motif 4	1762	15
Motif 5	1,447	17
Motif 6	1,513	4
Motif 11	939	11

Numbers of center addresses in motifs identified by our motif analysis method that are known exchanges. Motif centers identified from InnerCores have a high ratio of non-exchange addresses to exchange addresses (≈99%). This shows the effectiveness of our method to identify potentially meaningful addresses in a network different from high-traffic exchange addresses.

J. Zhu, A. Khan, and C. G. Akcora (2024). **Data depth and core-based trend detection on blockchain transaction networks**. Front. Blockchain

The Collapse of LunaTerra

LunaTerra addresses on May 7						
Address/Motif Center	C_1	C_4	C_{5a}	C_{5b}	C_6	C_{11}
Celsius	-	81	79	-	-	-
hs0327.eth	30	4	28	28	4	-
Smart LP: 0x413	-	69	-	-	95	-
Token Millionaire 1	85	81	73	-	67	89
Token Millionaire 2	35	100	99	-	99	38
masknft.eth	97	94	82	-	93	92
Heavy Dex Trader	54	17	-	-	32	-
Oapital	94	83	62	62	72	92
Hodlnaut	40	99	90	-	99	-

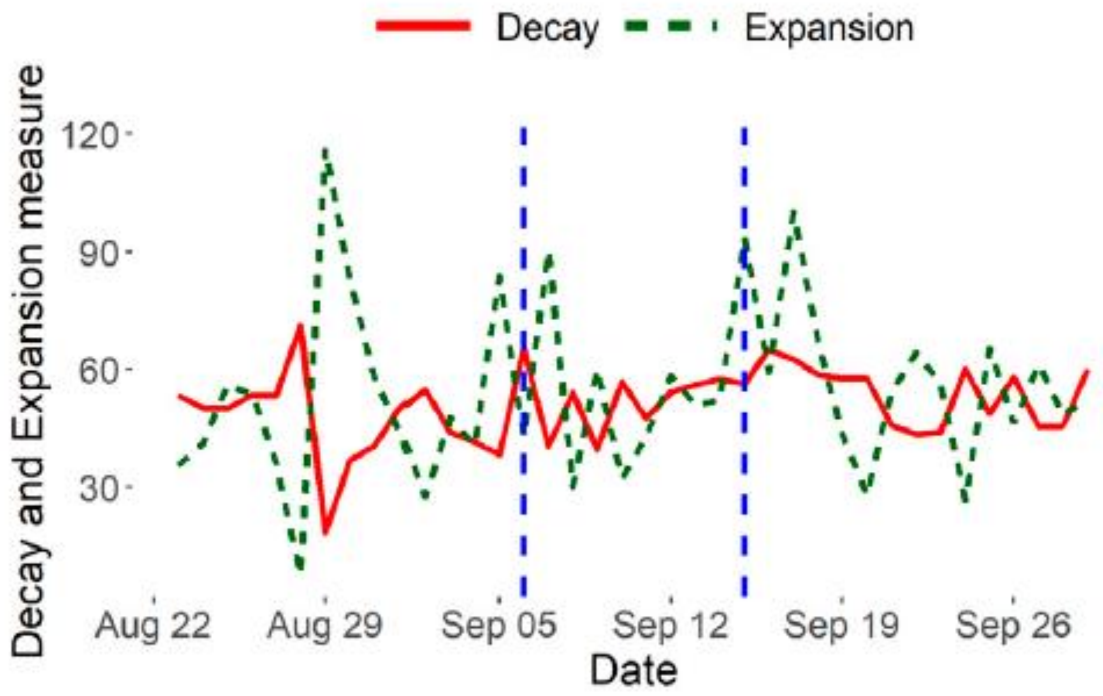
LunaTerra addresses on May 8						
Address/Motif Center	C_1	C_4	C_{5a}	C_{5b}	C_6	C_{11}
Celsius	-	81	79	-	-	-
hs0327.eth	88	67	70	96	82	-
Smart LP: 0x413	-	68	-	-	95	-
Token Millionaire 1	85	90	86	-	74	89
Token Millionaire 2	70	100	99	-	99	38
masknft.eth	91	91	82	-	93	92
Heavy Dex Trader	71	96	-	-	81	-
Oapital	92	79	58	61	72	93
Hodlnaut	40	99	91	-	99	-

LunaTerra addresses on May 9						
Address/Motif Center	C_1	C_4	C_{5a}	C_{5b}	C_6	C_{11}
Celsius	-	80	77	-	-	-
hs0327.eth	95	66	68	95	79	-
Smart LP: 0x413	-	67	-	-	95	-
Token Millionaire 1	83	89	85	-	73	88
Token Millionaire 2	67	100	99	-	99	88
masknft.eth	90	90	81	-	92	92
Heavy Dex Trader	70	93	-	-	80	-
Oapital	94	78	57	63	71	94
Hodlnaut	39	99	90	-	99	-

Nansen (https://www.nansen.ai/) is a prominent blockchain analytics platform that conducted a thorough analysis of the LunaTerra collapse in May 2022 and identified 11 important addresses that played central roles. We have captured 9 of 11 externally owned addresses (EoAs) identified by Nansen.ai that occurred as center addresses for our motif types on days immediately leading up to the LunaTerra collapse. We notice that the importance score percentile ranks of these addresses are higher compared to that of other center addresses for the same motif type on the same day, indicating that these addresses were important traders contributing to the buy or sell behavior associated with the motif on the day.

J. Zhu, A. Khan, and C. G. Akcora (2024). **Data depth and core-based trend detection on blockchain transaction networks**. Front. Blockchain

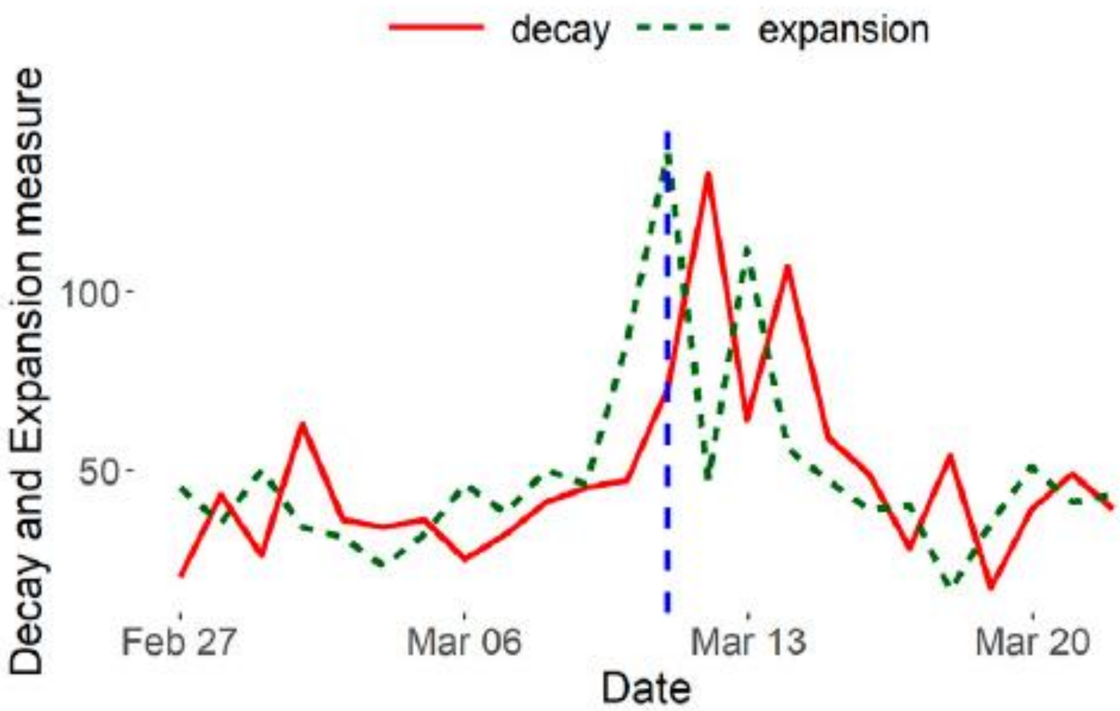
Ethereum's Switch to PoS



The move of Ethereum to Proof-of-Stake mining took place in two stages, indicated by 2 vertical blue lines (September 6 and 15, 2022). An expansion peak on 5 Sep 2022 detects the anomaly 1 day before the first stage commenced. A pattern of hope is observed.

J. Zhu, A. Khan, and C. G. Akcora (2024). **Data depth and core-based trend detection on blockchain transaction networks**. Front. Blockchain

USDC’s Temporary Peg Loss

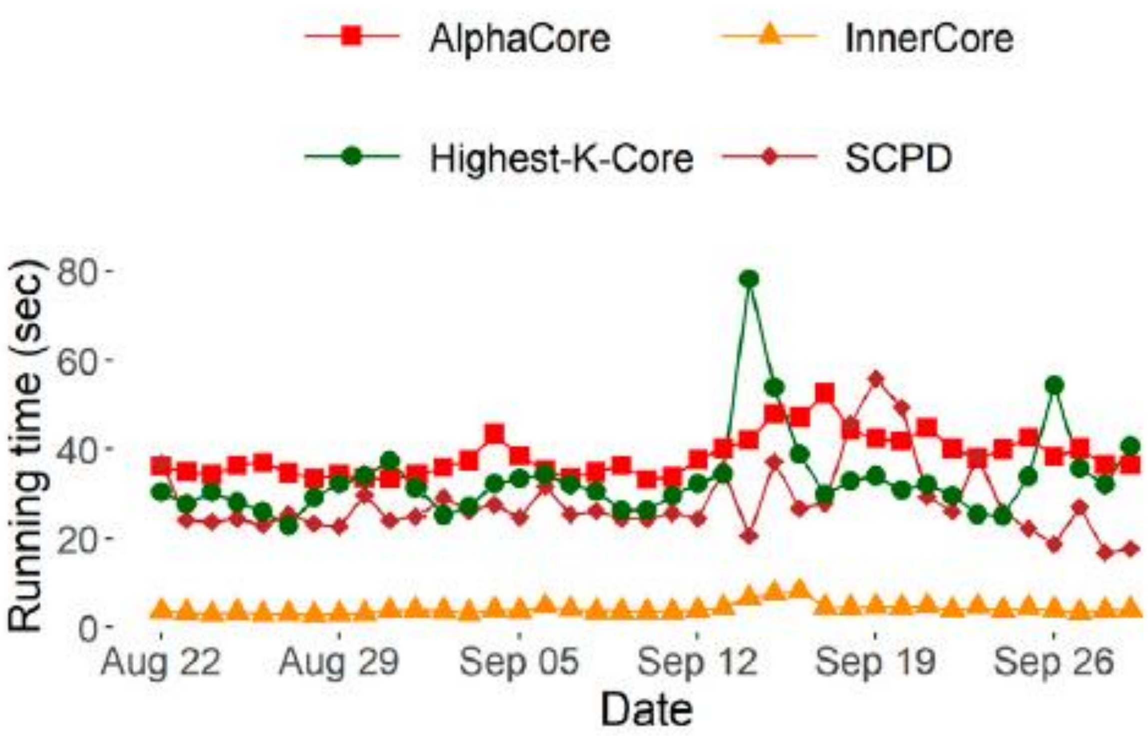


On 11 Mar 2023 (shown with the vertical blue line), USDC loses its $1 peg and falls to as low as 87 cents. A sudden surge in expansion on 11 May 2023 happened due to many traders liquidating their USDC holdings in response to the stablecoin’s all-time low value. In the subsequent 3 days following the temporary loss of USDC’s peg, a distinct series of behavioral patterns emerged, characterized by alternating signals of despair, hope, and despair again, before eventually stabilizing. During this 3-day period, Circle’s reassurances regarding the recovery of lost reserves gradually restored trust among its traders.

J. Zhu, A. Khan, and C. G. Akcora (2024). **Data depth and core-based trend detection on blockchain transaction networks**. Front. Blockchain

AALBORG UNIVERSITY

Efficiency Results



Our method InnerCore is also the fastest compared to existing methods. Innercore requires approximately 0.10 times the average computation time of AlphaCore, 0.12 times the average computation time of the highest graph k-core, and 0.14 times the average computation time of SCPD.

J. Zhu, A. Khan, and C. G. Akcora (2024). **Data depth and core-based trend detection on blockchain transaction networks**. Front. Blockchain

Summary

- InnerCore expansion and decay provide a **sentiment indicator** for the networks and explain trader mood.

- The centered-motif analysis in the InnerCore can detect **market manipulators**.

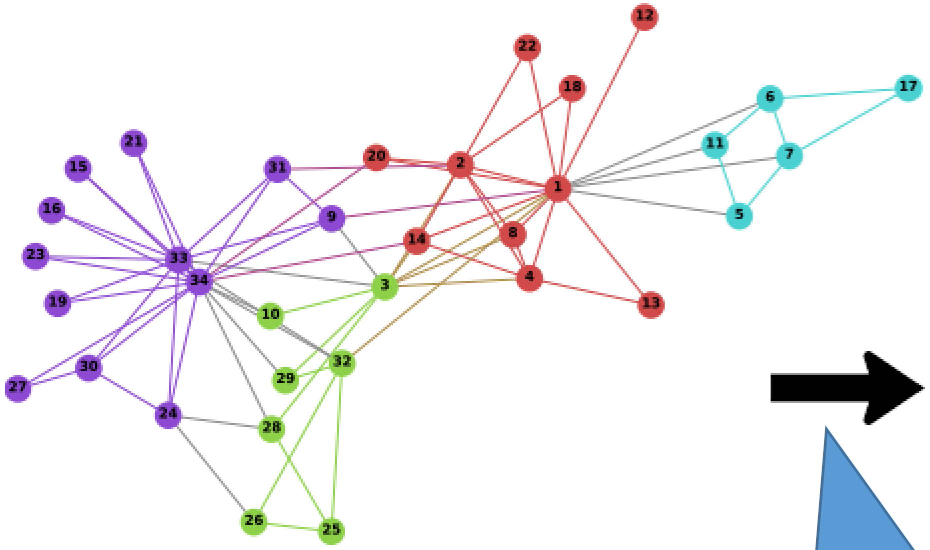
- The **scalability** and computational **efficiency** of InnerCore discovery make it well-suited for analyzing large temporal graphs

https://github.com/JZ-FSDev/InnerCore

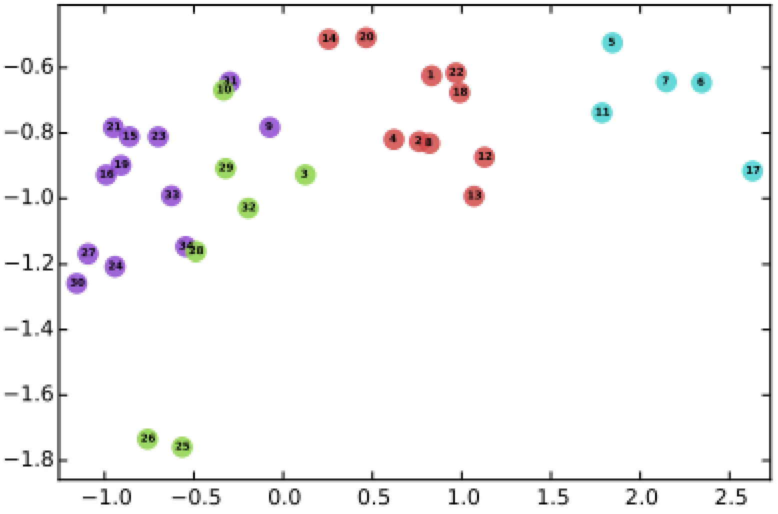
J. Zhu, A. Khan, and C. G. Akcora (2024). **Data depth and core-based trend detection on blockchain transaction networks**. Front. Blockchain

Machine Learning on Blockchain Graphs

Graphs Representation Learning



Graph



Node embedding/ vectors

Node classification
Link prediction
Graph classification
Entity resolution
Question Answering
… … …

Downstream tasks

Matrix factorization
Random walk sampling + Skip-Gram learning ✔
Graph convolutional neural networks (GCN) ✔

Machine Learning on Blockchain Graphs

Paper	Embedding Method	Downstream Task
D. Lin, J. Wu, Q. Yuan, and Z. Zheng. ***Modeling and understanding Ethereum transaction records via a complex network approach***. IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II: EXPRESS BRIEFS, VOL. 67, NO. 11, NOVEMBER 2020.	Random walk sampling + Skip-Gram learning	Transaction (link) prediction
D. Lin, J. Wu, Q. Yuan, and Z. Zheng. **T-EDGE: Temporal WEighted MultiDiGraph Embedding for Ethereum transaction network analysis.** Front. Phys., 2020, Sec. Social Physics.	Random walk sampling + Skip-Gram learning	Transaction (link) prediction
F. Poursafaei, R. Rabbany, and Z. Zilic. **SIGTRAN: Signature vectors for detecting illicit activities in Blockchain transaction networks**. PAKDD 2021.	Random walk sampling + Skip-Gram learning + Feature	Detecting illicit activities (node classification)
J. Wu , Q. Yuan, D. Lin , W. You, W. Chen, C. Chen**. Who are the phishers? Phishing scam detection on Ethereum via network embedding.** IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS 2020.	Random walk sampling + Skip-Gram learning	Phishing scams detection (node classification)
L. CHEN, J. PENG, Y. LIU, J. LI, F. XIE, and Z. ZHENG. ***Phishing scams detection in Ethereum transaction network.*** ACM Trans. Internet Technol. 2021.	Graph convolutional neural networks (GCN)	Phishing scams detection (node classification)
T. Yu , X. Chen, Z. Xu, and J. Xu. ***MP-GCN: A phishing nodes detection approach via graph convolution network for Ethereum***. Appl. Sci. 2022.	Graph convolutional neural networks (GCN)	Phishing scams detection (node classification)

Survey about Machine Learning on Blockchain Data

Survey	Graph ML	Seq. ML	Code ML	Temp. ML	Text ML
A Survey on Blockchain Anomaly Detection Using Data Mining Techniques [Li *et al.*, 2020a]	✓	✗	✓	✓	✗
Knowledge Discovery in Cryptocurrency Transactions: A Survey [Liu *et al.*, 2021a]	✓	✓	✓	✓	✗
A Survey on Blockchain Data Analysis [Hou *et al.*, 2021]	✓	✓	✓	✗	✗
Analysis of Cryptocurrency Transactions from a Network Perspective: An Overview [Wu *et al.*, 2021]	✓	✗	✓	✓	✓
Anomaly Detection in Blockchain Networks: A Comprehensive Survey [Hassan *et al.*, 2022]	✓	✓	✓	✓	✗
Graph Analysis of the Ethereum Blockchain Data: A Survey of Datasets Methods and Future Work [Khan, 2022]	✓	✗	✓	✓	✗
A survey on machine learning approaches in cryptocurrency: challenges and opportunities [Mujlid, 2023]	✗	✓	✗	✗	✗
Blockchain Data Mining with Graph Learning: A survey [Qi *et al.*, 2023]	✓	✓	✓	✓	✗
Machine Learning for Blockchain Data Analysis: Progress and Opportunities **[ours]**	✓	✓	✓	✓	✓

P. Azad, C. G. Akcora, A. Khan, "**Machine Learning for Blockchain Data Analysis: Progress and Opportunities**", CoRR abs/2404.18251, 2024.

Higher-order Structural Analysis on Blockchain Graphs

Blockchain Hypergraphs

Flow of coins creates a hyper-edge that connects more than two nodes, providing a more nuanced view of asset transfers.

Flow of coins between seemingly different addresses which are owned by the same user, creating hyper-edges.

S. Ranshous , C. A. Joslyn, S. Kreyling, K. Nowak, N. F. Samatova, C. L. West, and S. Winters. **Exchange Pattern Mining in the Bitcoin Transaction Directed Hypergraph.** International Financial Cryptography Association 2017.

S. Kim, M. Choe, J. Yoo, and K. Shin. Reciprocity in Directed Hypergraphs: Measures, Findings, and Generators. ICDM 2022.

Blockchain Datasets and Tools

Blockchain Datasets

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

BitcoinHeistRansomwareAddressDataset

Download: Data Folder, Data Set Description

Abstract: BitcoinHeist datasets contains address features on the heterogeneous Bitcoin network to identify ransomware payments.

Data Set Characteristics:	Multivariate, Time-Series	**Number of Instances:**	2916697	**Area:**	Computer
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	10	**Date Donated**	2020-06-17

The Elliptic Data Set: Working With the Community to Combat Financial Crime in Cryptocurrencies

Chartalist

First Blockchain ML-Ready Dataset Platform

Live Graph Lab

Smart Contract Sanctuary

SmartBugs: A Framework for Analysing Ethereum Smart Contracts

AALBORG UNIVERSITY

Blockchain Data Analytic Tools

Sixth International Workshop on Foundations and Applications of Blockchain

In conjunction with VLDB'24

August 30th, 2024

○ **Bartoletti et al.** developed a Scala framework for blockchain data analytics. This can integr relevant blockchain data with data from other sources, and organize them in a database, eith SQL or NoSQL.

○ **GraphSense** is an open-source platform for analyzing cryptocurrency transactions.

○ **BlockSci** loads the parsed data as an in-memory database, which the user can either query directly or through a Jupyter notebook interface.

○ **Industry:** https://santiment.net/ , https://www.nansen.ai/ , https://www.blockchain.com/ , https://www.chainalysis.com/ etc.

M. Bartoletti, S. Lande, L. Pompianu, A. Bracciali. **A general framework for blockchain analytics**. SERIAL@Middleware 2017.
B. Haslhofer, R. Stütz, M. Romiti, R. King. *GraphSense:* **A general-purpose cryptoasset analytics platform**. CoRR 2021.
H. A. Kalodner, M. Möser, K. Lee, S. Goldfeder, M. Plattner, A. Chator, A. Narayanan. **BlockSci: design and applications of a blockchain analysis platform**. USENIX Security Symposium 2020.

Blockchain Data Analytic Tools

- **Information on User Accounts:** https://etherscan.io/, https://cryptoscamdb.org/, https://tutela.xyz/ - fraud detection and classifying accounts.
- **Static code analysis, machine learning on smart contracts** are popular for code reuse checking, contract classification, and ponzi schemes detection.
- **LATTE** provides a novel visual smart contract construction system. This will benefit non-programmers to easily construct a contract by manipulating visual objects and without writing Solidity code.
- **BiVA** is a graph mining tool for the bitcoin network visualization and analysis and transaction pattern analysis.

T. Hu, X. Liu, T. Chen, X. Zhang, X. Huang, W. Niu, J. Lu, K. Zhou, Y. Liu. **Transaction-based classification and detection approach for Ethereum smart contract**. Inf. Process. Manag. 58(2): 102462 (2021).

S. Tikhomirov, E. Voskresenskaya, I. Ivanitskiy, R. Takhaviev, E. Marchenko, Y. Alexandrov. **SmartCheck: static analysis of Ethereum smart contracts**. WETSEB@ICSE 2018.

S. Ducasse, H. Rocha, S. Bragagnolo, M. Denker, C. Francomme. **SmartAnvil: open-source tool suite for smart contract analysis**. Blockchain and Web 3.0: Social, Economic, and Technological Challenges. 2019.

T. Durieux, J. F. Ferreira, R. Abreu, and P. Cruz. **Empirical review of automated analy-sis tools on 47, 587 ethereum smart contracts**. In ICSE, 2020

S. S. Kushwaha, S. Joshi, D. Singh, M. Kaur, and H.-N. Lee. **Ethereum smartcontract analysis tools: A systematic review**. IEEE Access, 10:57037–57062, 2022.

S. Tan and S. S. Bhowmick and H.-E. Chua and X. Xiao. **LATTE: visual construction of smart contracts**, SIGMOD, 2020.

F. E. Oggier, A. Datta, and S. Phetsouvanh. **An ego network analysis of sextortionists**. Soc. Netw. Anal. Min., 10(1), 2020.

Blockchain Data Analytic Tools

○ **Visualization of blockchain data:** BitConeView, BitConduite, Bitcoinrain, Ethviewer, ...

Survey: N. Tovanich, N. Heulot, J.-D. Fekete, P. Isenberg. **Visualization of Blockchain data: a systematic review**. IEEE Trans. Vis. Comput. Graph. 27(7): 3135-3152 (2021)

Z. Zhong, S. Wei, Y. Xu, Y. Zhao, F. Zhou, F. Luo, and R. Shi. **Silkviser: A visual explorer of blockchain-based cryptocurrency transaction data**. In IEEE Conference on Visual Analytics Scienceand Technology, 2020.

○ **Natural language processing and sentiment analysis** using tweets, online articles, cryptocurrency prices and charts, Google Trends about blockchain.

- ➢ M. S. Tash, O. Kolesnikova, Z. Ahani, and G. Sidorov. **Psycholinguistic and emotion analysis of cryptocurrency discourse on x platform**. Scientific Reports,14(1):8585, 2024
- ➢O. Kraaijeveld and J. D. Smedt. **The predictive power of public Twitter sentiment for forecasting cryptocurrency prices**, 2020, Journal of International Financial Markets, Institutions and Money, 65.
- ➢ A.-D. Vo and Q.-P. Nguyen and C.-Y. Ock, **Sentiment analysis of news for effective cryptocurrency price prediction**, International Journal of Knowledge Engineering, 5(2), 2019.
- ➢ Abraham and D. Higdon and J. Nelson and J. Ibarra. **Cryptocurrency price prediction using tweet volumes and sentiment analysis**, SMU Data Science Review, 2018.

Open Problems

Open Problems

- Multilayer graphs would be an expressive model of real-world activities such as external and internal transactions, token transfers, dApps and DeFi usage, cross-chain analysis.

- Multimodal data could integrate information across diverse modalities
  - blockchain transactions, smart contract code, bytecode, price, social data.

- Due to highly dynamic nature of accounts and transactions, employed ML models must deal with data and model drifts.
  - Drift detection, incremental learning, machine unlearning and continuous learning would be useful.

- Deep learning models.
  - Black-box: adding explainability and human-in-the-loop, reducing bias.
  - Real-time detection.

L. Cheng, F. Zhu, Y. Wang, R. Liang, H. Liu. Evolve Path Tracer: Early Detection of Malicious Addresses in Crypto currency. KDD 2023

- LLMs for Blockchain data analysis.
  - LLMs for understanding natural language query, interacting with transaction and contract data, and generating source code.

Y. Gai, L. Zhou, K. Qin, D. Song, A. Gervais. Blockchain Large Language Models. CoRR abs/2304.12749 (2023)