Graph Analysis of the Ethereum Blockchain Data: A Survey of Datasets, Methods, and Future Work

Arijit Khan
Aalborg University
Aalborg, Denmark
arijitk@cs.aau.dk

***Abstract*—Ethereum, currently the most actively-used and the second-largest blockchain platform, consists of a heterogeneous ecosystem, cohabited by human users, smart contracts (autonomous agents), ether (native cryptocurrency), tokens (digital assets), dApps (decentralized applications), and DeFi (decentralized finance). These key actors in the Ethereum interact with each other via transactions and contract calls. Given the highly connected structure, graph-based modeling is an optimal tool to analyze the data stored in Ethereum blockchain. Recently, several research works performed graph analysis on the publicly available Ethereum 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. In this work, we conduct an in-depth survey of the existing literature. We categorize them based on publication years, venues, core ranking, and authors' affiliations, data usage and graphs construction, graph mining and machine learning techniques employed, and the new insights derived by them. We conclude by discussing our recommendations on the future work. Our article will be useful to data scientists, researchers, financial analysts, and blockchain enthusiasts.**

***Index Terms*—blockchain, ethereum, network analysis**

I. Introduction

A blockchain [1] is a distributed, digital ledger of records, stored in a sequential order. Each record, called a block, is time-stamped and is linked to the previous one; these blocks are shared openly among its users to create an immutable sequence of transactions. A blockchain can only be updated by consensus among its users (either an open or a controlled set of users), who participate in a peer-to-peer network. Thus, a blockchain contains a secure, tamper-proof, and verifiable record of every transaction ever validated in the system, and is crucial in the trust economy of the future.

The Ethereum blockchain[1], launched in July 2015, is currently the most actively-used and the second-largest blockchain platform, with market value[2] grew to over 250 billion U.S. dollars in April 2021. Ethereum, hosting ether (ETH), the second largest cryptocurrency by market capitalization, is a public blockchain that keeps records of all Ethereum related transactions. Ethereum supports decentralized applications (dApps) with its *smart contracts*, which are autonomous agents that can execute complex code across a decentralized network. Ethereum blockchain also permits creation and transaction of *tokens*, which are digital assets, through codes defined in the respective smart contracts. Therefore, Ethereum introduces a heterogeneous, financial ecosystem of humans (users) and autonomous agents (smart contracts), who transact using ether and various fungible (e.g., ERC20) and non-fungible (e.g., ERC721) tokens (digital assets), and these are recorded permanently in the blockchain ledger.

Public blockchain data are widely investigated in several downstream applications including cryptocurrency price prediction, address clustering, criminal usage detection, anti-money-laundering, business transactions analysis, and thus providing new means for financial data mining [2], [3], [4], [5]. They are critical in emerging fields such as blockchain intelligence[3], blockchain social networks [6], and blockchain search engines [7]. Data stored in a public blockchain can be considered as big data (e.g., Ethereum archive nodes that store a complete snapshot of the Ethereum blockchain, including all the transaction records, take up to 4TB of space[4]), thus data analytic methods can be applied to extract knowledge hidden in the blockchain. Ethereum blockchain has processed more than 1.1 million transactions per day in July 2021[5] and contains a vast amount 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.) that can be modeled as complex, dynamic, multi-layer networks [8], [9], [10]. In this work, we conduct an in-depth survey of the existing literature (i.e., 25 research papers published at peer-reviewed conferences, journals, and workshops in the past five years) that performed graph analysis of the Ethereum blockchain data. We compare them based on publication years, venues, core ranking, and authors' affiliations, data usage and graphs construction, graph mining and machine learning methods used, and the new insights revealed by them.

Related work. Blockchain data analytics, also known as the distributed ledger analytics (DLA), is an emerging field of research. It deals with insights from transactions and other data stored on public blockchains. For surveys and tutorials, we refer to [11], [12], [2], [5], [3], [13]. They provide more generic discussions on various blockchains, e.g., Bitcoin, Ethereum,

This work is supported by the Novo Nordisk Foundation grant NNF22OC0072415.

[1] ethereum.org/en/whitepaper/ github.com/ethereum/yellowpaper

[2] statista.com/statistics/807195/ethereum-market-capitalization-quarterly/

[3] blockchaingroup.io

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

[5] statista.com/statistics/730838/number-of-daily-cryptocurrency-transactions-by-type/

Monero, Zcash, Ripple, Iota, etc. Unlike ours, they are not specific to Ethereum, for instance, none of them considered all twenty five research papers that we survey. Different from them, we provide a more in-depth study of the Ethereum ecosystem (which is heterogeneous, cohabited by externally owned accounts, smart contracts, ether, different categories of tokens, external and internal transactions, dApps and DeFi), data extraction tools, as well as several prominent research papers published at peer-reviewed conferences, journals, and workshops in the last five years that conducted graph analysis of the Ethereum blockchain data.

Many works performed graph analysis and machine learning with other blockchain data, such as Bitcoin [14], [15], [16], [17], Litecoin [17], Monero [18], EOSIO [19], and Steem [20]. We do not survey them since we focus on Ethereum.

Our contributions and roadmap. Data on Ethereum blockchain can be modeled as graphs of different formats: static, dynamic, and historical snapshot graphs, directed and weighted graphs, simple and multi-graphs, attributed and multi-layer networks. They are critical in predicting frauds, detecting phishing scams and counterfeit cryptocurrencies. Therefore, graph-based modeling and mining of the Ethereum data is an emerging area of research. While many graph analysis works on the publicly available Ethereum blockchain data have emerged, to the best of our knowledge, *ours is the first in-depth survey of the literature that conducted graph analysis with the Ethereum blockchain data*.

We first introduce Ethereum's heterogeneous ecosystem and data extraction tools (§II). We next discuss twenty five research papers published at peer-reviewed journals, conferences, and workshops based on their (a) publication venues, years, categories, publishers, and authors' affiliations (§III-A); (b) data extraction methods, dataset durations, and the graphs constructed (§III-B); (c) graph properties, topological data analysis, machine learning methods investigated, and the target applications therein (§III-C). We conclude and discuss future directions in §IV. Our classification and description of data, models, applications, and future directions on Ethereum blockchain are *timely and critical* – this will benefit data scientists, machine learning practitioners, and financial analysts.

II. Taxonomy and the Ethereum

Ethereum is a public blockchain permitting anyone to join and use decentralized applications (dApps), created by developers, that run on the Ethereum Virtual Machine (EVM). EVM can execute codes of arbitrary algorithmic complexity, making Ethereum *Turing complete* [21]. Ethereum is a transaction-based state machine, where the state is made up of accounts, i.e., externally owned accounts (EOAs) and smart contract code controlled accounts. Transfers of values (e.g., ether and tokens transactions) and information (e.g., contract create, call, or kill) between accounts cause transitions in the global state of Ethereum, which are recorded in the blockchain.

We discuss the main components of the Ethereum in § II-A, that are important for graphs creation, followed by existing tools for Ethereum data extraction in § II-B.

A. Components of the Ethereum Blockchain

The key actors in the Ethereum ecosystem are as follows.
Ether. The native cryptocurrency of Ethereum is called the ether, or ETH, that is transferred between user accounts. Ether is also paid to run transactions, called transaction fees or *gas*, for covering the costs of computing power.
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.

Ethereum uses the *account-based transaction model* [13] that represents ether as balances within accounts, similar to bank accounts. Every account has an address, balance, storage, and code-space (in EOAs, both code-space and storage are empty) for interacting and transacting with other accounts. Every account has a publicly viewable nonce that is incremented at every transaction. If there are two transactions referring to the same account with the same nonce, then only the first one will be validated and the second one is marked as *double spending*. Also, a transaction is validated if the sending account has enough balance to pay for it.

Ethereum accounts can be of different categories, such as miners, exchanges, smart contracts for tokens, gambling games, etc. Past work [22], [23] employed various heuristic and machine learning methods to cluster Ethereum accounts and found entities that likely control multiple accounts.
Smart contracts. A smart contract is a program, usually written in *Solidity* or *Vyper* – Javascript and Python-like languages, respectively, and compiled into JVM bytecode, that runs on the Ethereum blockchain. A smart contract is deployed to a specific address on the blockchain, and constitutes a collection of code (for multiple functions) and data (its state). Smart contracts can define rules and automatically enforce them via the code. User accounts interact with a smart contract by transactions that execute a function defined on the contract. Smart contracts can also call (or, kill) each other, even itself, if processing a transaction requires some functionality within the other or in the same contract. Smart contracts can react to transactions, but cannot initiate them [24].

A smart contract can be an application, e.g., defining a token. A contract can also be a building block of a multi-tier application, such as decentralized finance. Based on an exploratory study [25], about 42% highly-active smart contracts are related to transferring, selling, and distributing tokens.
Transactions. A transaction, initiated by an EOA, transfers Ethereum-based assets (e.g., ETH, tokens) from one address to another. They can be broadly characterized into three classes (not exclusive). **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

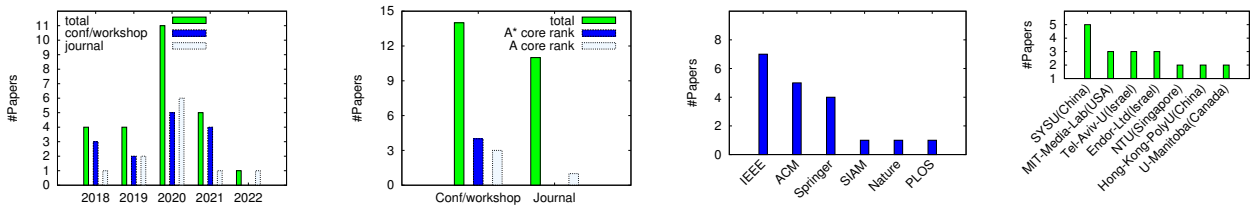

Fig. 1: Publication venues, years, categories, and affiliations of our surveyed 25 papers. (a) Number of papers vs. publication years. (b) Number of papers in conference/workshop and journal categories, as well as at core ranking A* and A venues. (c) Number of papers based on prominent publishers. (d) Number of papers based on authors' institute affiliations; if a paper has co-authors with different institute affiliations, the paper is counted under each such institution.

to the network in the form of regular transactions. A token transfer is, in fact, an update of balances in the variables of the token smart contract. Therefore, without running the EVM and executing regular transactions, we cannot observe the internal transactions, neither create the token transfer graph [13]. Using Etherscan[6], one can view the internal transactions and token transfers associated with an address.

Tokens. Tokens are digital assets or access rights provided by their issuers, managed by smart contracts and the blockchain platform. A token's smart contract specifies meta-attributes about the token, including its symbol, total supply, decimals, etc. Two most popular token standards on Ethereum are: (1) **ERC20**, a standard interface for fungible (interchangeable) tokens, such as voting tokens, staking tokens, or virtual currencies, and (2) **ERC721**, a standard interface for non-fungible tokens (NFTs), e.g., a deed for a song or an artwork.

ERC20 tokens are widely used in *initial coin offering* (ICO), a crowdfunding process to raise funds in the cryptocurrency market. The ERC721, on the other hand, introduces a standard for NFTs, such tokens are unique and can have different values than other tokens from the same smart contract. NFTs are used to represent ownership of collectible items, songs, artworks, access keys, lottery tickets, etc.

dApps and DeFi. A decentralized application (dapp) is built on a decentralized peer-to-peer network that combines smart contract(s) as backend and a frontend user interface, generally implemented via HTML5, CSS, and web3.js. dApp authors often submit their dapps to certain websites, e.g., State of the dApps and DappRadar[7] for advertisements. In Ethereum, about 70% dapps have only one smart contract, and 90% dapps have less than 3 smart contracts, while there are also some dapps having more than 100 smart contracts [26]. Exchanges, wallet, and games are the most popular dApp categories.

DeFi, or decentralized finance [27] are dApps for financial products and services, e.g., loans, savings, insurance, exchanges, liquidity, lenders, and trading, powered by decentralised blockchain technologies such as Ethereum. *DeFi protocols* are smart contracts that constitute a collection of rules similar to physical financial institutions.

B. Tools for Ethereum Data Extraction

To get all historic Ethereum transactions, one can join the Ethereum network of nodes through a client. Geth, OpenEthereum, and Parity[8] are popular software clients for running a full node on Ethereum. The Geth client stores all blockchain data on disk in LevelDB database using key-value pairs. Alternatively, users can also interact with Ethereum nodes via the web3 library using managed services, such as Infura and Quicknode[9]. In addition, some well-curated Ethereum blockchain datasets have also been released, e.g., Google BigQuery [28] and XBlock-ETH [29]. Ethereum blockchain data on Google BigQuery are updated daily and are accessible through an SQL interface. The ETL (extract-transform-load) of Ethereum data converts them into convenient formats, such as CSVs, relational databases, and graphs within a specified block range [30]. Ethereum Query Language (EQL) [31] supports SQL-like queries to retrieve information from the Ethereum blockchain data.

III. Survey of Graph Analysis with Ethereum Data

We survey twenty five research papers, published in the past five years (2018-2022), that conducted graph analysis with the Ethereum blockchain data. We do not include [32], [33], [34], [35] since they were not published at peer-reviewed venues as of the time of this writing.

A. Publication Venues and Affiliations

Figure 1 presents distributions of papers and co-authors based on publication venues, years, categories, publishers, and authors' affiliations. Among 25 papers surveyed, 11 were published in 2020, which is currently the maximum in a year (Figure 1(a)). More papers were published at conferences and workshops, than in journals. Eight papers were published at core A* and A venues (Figure 1(b)). IEEE, ACM, and Springer published majority of these papers (Figure 1(c)). Based on authors' affiliations, more papers and co-authors are from China and USA (Figure 1(d)). Prominent research groups working in this domain are from Sun Yat-sen University or SYSU (China), MIT Media Lab (USA), Tel Aviv University (Israel), Endor Ltd. (Israel), Nanyang Technological University or NTU (Singapore), the Hong Kong Polytechnic University (China), and the University of Manitoba (Canada). The SYSU group also open-sourced several well-curated Ethereum blockchain datasets [29].

[6] info.etherscan.com/understanding-an-ethereum-transaction/

[7] stateofthedapps.com/ dappradar.com/

[8] geth.ethereum.org/ openethereum.github.io/ parity.io/technologies/ethereum/

[9] infura.io/ quicknode.com/

TABLE I: Datasets and graphs in our surveyed papers.

paper	data extraction	data duration	constructed graphs	links to data and/or code
INFOCOM18 [36]	client (Geth)	2015-2017	money flow graph, contract creation graph, contract invocation graph	https://github.com/brokendragon /Ethereum_Graph_Analysis
PLOS ONE18 [37]	Etherscan APIs	2015-2017	transaction graph	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/XIXSPR
Complex Sys18 [38]	client	2016-2018	(full) ERC20 tokens transfer graph	not given
NTMS18 [39]	client (Geth)	2015-2017	user-to-user, user-to-smart contract, and smart contract deployment graphs	not given
FC19 [40]	client (Parity)	2015-2018	(individual) ERC20 token transfer graphs	not given
ICDMW19 [41]	not given	2018-2019	Storj token transfer graph	not given
Appl. Netw. Sci.19 [42]	client	2015-2019	transaction graph	not given
Inf. Sci.19 [43]	client (Geth)	2015-2019	transaction graph	not given
WWW20a [44]	Google BigQuery	2015-2019	trace graph, contract graph, transaction graph, token graph	https://github.com/sgsourav /blockchain-network-analysis
SDM20 [45]	client (Geth)	2015-2018	(individual) ERC20 token transfer graphs	https://github.com/yitao416/EthereumCurves
WWW20b [23]	client (Parity)	2015-2019	ERC20 token creator, holder, and transfer graphs	http://xblock.pro/#/
Sci Rep20 [46]	client	2016-2019	(individual) ERC20 token transfer graphs	not given
ACM Meas. Anal. Comput. Syst.20 [47]	client (Geth)	2017-2020	ERC20 token creator, holder, and transfer graphs for counterfeit tokens	not given
Concurr. Comput. Pract. Exp.20 [48]	Infura web3 service	2015-2020	transaction graph	not given
IEEE Trans. Circuits Syst.20 [49]	Etherscan APIs	2015-2020	transaction graph	https://github.com/lindan113/T-EDGE
Frontiers Phys.20 [50]	Etherscan APIs	2015-2020	transaction graph	https://github.com/lindan113/T-EDGE
J. Complex Networks20 [51]	Etherscan APIs	2015-2018	transaction graph	not given
Networking20 [9]	client (Geth)	2015-2019	user-to-user, contract-to-contract, and user-contract graphs	not given
SBP-BRiMS20 [52]	client	2016-2018	(full) ERC20 tokens transfer graph	not given
WWW21 [8]	Google BigQuery	2015-2019	trace graph, contract graph, transaction graph, token graph	https://github.com/LinZhao89 /Ethereum-analysis
ECML PKDD21 [10]	not given	2015-2020	(individual) token transfer graphs, stacked as a multi-layer network	https://github.com/tdagraphs
PAKDD21 [53]	from [54]	2015-2019	transaction graph	https://github.com/fpour/SigTran
ACM Trans. Internet Techn.21 [55]	Etherscan APIs	2015-2020	transaction graph	http://xblock.pro/#/
Blockchain21 [56]	Google BigQuery	2015-2021	(individual) ERC721 token transfer graphs	https://github.com/epfl-scistimm /2021-IEEE-Blockchain
IEEE Trans. Syst. Man Cybern. Syst.22 [54]	client	2015-2019	transaction graph	http://xblock.pro/#/

B. Datasets and Graphs

Our surveyed papers vary based on data extraction methods, dataset durations, and the graphs constructed (Table I).

Data extraction. 13 out of 25 papers employed software clients such as Geth and Parity that run a full node on Ethereum to collect all historic transactions. A few of the surveyed papers used web3 services and Etherscan APIs for data extraction [37], [48], [49], [51], [55]. Besides, [44], [8], [56] used Google BigQuery and [53] xblock.pro/#/ to access well-processed Ethereum blockchain datasets. 13 out of 25 papers provided links to their source code and/or datasets.

Constructed graphs. The graphs can be classified based on transactions and token transfers; however, there are sufficient varieties across different works. We introduce them below, highlight similarities, differences, and summarize at the end.

• Chen et al. [36] studied the **money flow graph** (MFG), **smart contract creation graph** (CCG), and the **smart contract invocation graph** (CIG) based on transactions. MFG is a *weighted, directed graph* denoting transfer of ether between accounts (both EOAs and smart contract accounts). A weight denotes the total amount of ether transferred along that edge via one or more transactions. CCG, which deals with smart contracts creation, is a *forest* having multiple trees. The root of every tree is an EOA, other nodes of the tree are smart contract accounts that are directly or indirectly created by that EOA. In contrast, CIG is a *weighted, directed graph*, an edge indicates an invocation of a smart contract, either by an EOA or by another smart contract; the edge weight counts the number of invocations, via one or more transactions.

• Liang et al. [37] constructed the **transaction graph**, which is similar to the money flow graph (MFG) in [36]; however, [37] studied it on a *monthly basis*. Nodes are added due to creation of new accounts and are removed when they are no longer involved in any transaction. New edges are inserted for transactions between two previously unconnected accounts. *Edge weight* is assigned based on the number of transactions, and not considering the total amount of ether transferred.

• Somin et al. [38] created the **full ERC20 tokens transfer graph**, with all ERC20 tokens transferred among EOAs. The edges are *directed*, but *unweighted*, that is, the number of transfers or the number of tokens transferred are not counted. In [52], the authors built *weekly* versions of the above graph for temporal analysis. Later, they studied about **1 500 individual ERC20 token transfer graphs** on *weekly* basis [46].

• Anoaica and Levard [39] analyzed **user-to-user**, **user-to-smart contract**, and **smart contract deployment** graphs, based on external transactions. The contract deployment graph is similar to the contract creation graph in [36]; however, [39] investigated these graphs on *monthly*, *weekly*, *daily*, and *hourly* basis. Each edge is *directed*, and counts both the total number of transactions and the amount of ether transferred.

• Victor and Lüders [40] studied **1 000 individual ERC20**

TABLE II: Graph properties, machine learning (ML) methods, and target applications in our surveyed papers.

paper	graph properties and ML methods	target applications
INFOCOM18 [36]	degree distribution, connected component, clustering coeff., assortativity Pearson coeff., PageRank	inferring node identity, attack forensics, anomaly detection
PLOS ONE18 [37]	monthly change of degree distribution, assortativity, clustering coeff., connected component	network growth and dynamic characteristics
Complex Sys18 [38]	degree distribution	social behavior of ERC20 token transfer
NTMS18 [39]	degree distribution, betweenness and Eigenvector centrality; their variations over time	internal activities on Ethereum blockchain, their temporal variations
FC19 [40]	degree distribution, connected component, clustering coeff., assortativity, density, shortest path to exchanges	understanding of token networks
ICDMW19 [41]	motifs count, LSTM	token price prediction
Appl. Netw. Sci.19 [42]	clustering coeff., assortativity, density, max. clique, repetition ratio, rel. growth rate of monthly graphs	structural properties, correlation between transaction graph properties with historical events and price
Inf. Sci.19 [43]	degree distribution, connected component, shortest path, diameter transaction volume distribution, bow-tie structure	insights into transaction relations
WWW20a [44]	degree distribution, centrality, density, reciprocity, assortativity connected component, core decomposition, clustering coeff., motif count, articulation points, adhesion, cohesion, diameter, shortest path	study of user-to-user, user-to-contract, contract-to-user, and contract-to-contract networks, individual token sub-networks
SDM20 [45]	persistent homology, functional data depth	price changes of crypto-tokens
WWW20b [23]	degree distribution, number of transactions PageRank, clustering coeff.	study token ecosystem, are different tokens created/ controlled by the same entity, abnormal (fake) transactions in decentralised exchanges
Sci Rep20 [46]	degree distributions and their temporal variations	dynamics of Ethereum tokens ecosystem
ACM Meas. Anal. Comput. Syst.20 [47]	number of transactions	study of counterfeit tokens on Ethereum
Concurr. Comput. Pract. Exp.20 [48]	degree distribution, connected component, clustering coeff., shortest path, diameter	study transactions in Ethereum ecosystem
IEEE Trans. Circuits Syst.20 [49]	degree distribution, temporal random walk-based graph representation learning	temporal link prediction
Frontiers Phys.20 [50]	temporal random walk-based graph representation learning	node classification (phishing vs. genuine)
J. Complex Networks20 [51]	degree distribution, number of transactions, clustering coeff., connected components, communities; their temporal variations	temporal analysis of Ethereum transaction network
Networking20 [9]	degree distribution, motifs counting, number of transactions, burstiness; their temporal variations	evolution of Ethereum
SBP-BRiMS20 [52]	degree distributions and their temporal variations	ERC20 network dynamics and predictive ability
WWW21 [8]	network growth model, density, degree distribution, number of transactions, reciprocity, assortativity, connected components, core decomposition, clustering coeff. community detection; their temporal variations	community continuation prediction, correlate anomalies with external real-life incidents, find appropriate time granularity
ECML PKDD21 [10]	clique persistent homology	topological anomaly detection in Ethereum (dynamic multi-layer networks)
PAKDD21 [53]	random-walk-based graph representation learning	Ethereum node classification (illicit vs. genuine)
ACM Trans. Internet Techn.21 [55]	GCN-based graph representation learning	Ethereum node classification (detect phishing scams)
Blockchain21 [56]	degree distribution, shortest path, diameter, assortativity PageRank, number of transactions; their temporal variations	analysis of ERC721 transactions
IEEE Trans. Syst. Man Cybern. Syst.22 [54]	random walk-based graph representation learning	Ethereum node classification (detect phishing scams)

tokens transfer graphs. Edges are *directed*; *self-edges* (an edge to the same account), *multi-edges* (multiple transfers of a token between a source and a target account), and *simple-edges* (at least one transfer of a token between a source and a target account) are counted.

• Chen and Ng [41] used *daily* **Storj token transfer networks**, which are *unweighted, directed graphs*, without considering the number or the amount of assets transferred.

• Motamed and Bahrak [42] built *monthly* **(external) transaction graphs** that are *unweighted, undirected graphs*.

• Guo et al. [43] studied the **(external) transaction graph**, they considered: (a) *directed, weighted graph*, (b) *directed, unweighted graph*, and (c) *undirected, unweighted graph*, where the edge weight is the amount of ether transferred.

• Lee et al. [44] derived four networks: (a) **TraceNet**, consisting of all successful traces with non-null from/to addresses as edges; (b) **ContractNet**, a subgraph of TraceNet, where only those edges with both from_address and to_address belonging to smart contracts, are retained; (c) **TransactionNet**, whose edges are formed by external transactions (similar to the money flow graph [36] and transaction graph [42], [43]); and (d) **TokenNet**, based on explicit transfer of tokens. The authors considered *multi-digraph* and *simple, undirected* versions of each graph. In the former, multiple edges (repeated interactions or transfers) between a pair of vertices are retained. In the later, at most one undirected edge between every pair of nodes is considered. The total amount of assets being transferred are not investigated. In [8], this research group studied temporal variations of the four networks, considering *yearly*, *6-monthly*, *3-monthly*, and *monthly* graphs.

• Li et al. [45] studied **31 individual ERC20 tokens transfer graphs** and their *daily* snapshots. The edge weight denotes dissimilarity between two nodes, that is, the larger is the transferred amount between two nodes, the smaller is the inter-node dissimilarity, and less is the edge weight.

• Chen et al. [23] built **ERC20 token creator (TCG), holder (THG), and transfer (TTG) graphs**. In TCG, an *unweighted, directed* edge denotes token creation relationship. THG is a *weighted, directed* graph, where an edge denotes holding of a token, and the edge weight represents the shares of the token being held. TTG is also a *weighted, directed* graph, indicating transfer of tokens. An edge weight denotes the total number of transfer records between the two nodes, and ignores the type and number of tokens.

• Gao et al. [47] constructed **ERC20 token creator (TCG), holder (THG), and transfer (TTG) graphs with 2 117**

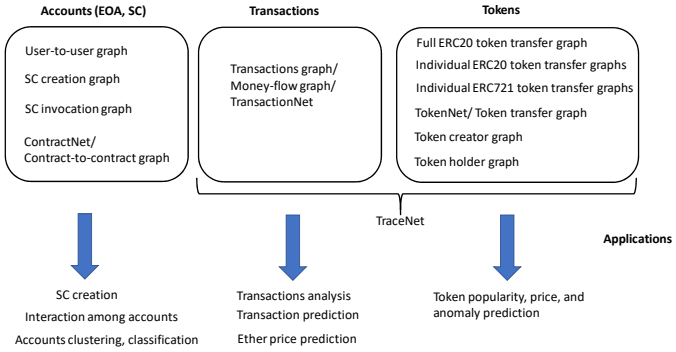

Fig. 2: Various graphs created from interactions between accounts, transactions, token transfers; as well as their common applications.

counterfeit tokens. Hence, these graphs are similar to the ones in [23], but they only consider counterfeit tokens.

• Ferretti and D'Angelo [48] studied the **(external) transaction graph**, which is similar to the money flow graph in [36]. Besides, they analyzed *temporal* properties by considering different snapshots starting at various block numbers.

• Lin et al. [49], [50] employed **(external) transaction graph** as a *temporal, weighted, multi-digraph*, where a weight (amount of ether transferred) and a time-stamp is associated with every edge, denoting a transaction via the edge.

• Mascarenhas et al. [51] represented the **(external) transaction graph** as *time-varying graphs* on *yearly* basis. The edge weights are assigned based on the number of transactions, as well as weighted by the amount of ether transferred.

• Bai et al. [9] constructed three *temporal* graphs: **user-to-user**, **contract-to-contract**, and **user-to-contract**, based on external and internal transactions. The first and third graphs are similar to user-to-user and user-to-smart contract graphs, respectively [39]; whereas the second one is same as ContractNet [44]. The sliding window is varied between 180 - 1 260 days, and is shifted with a granularity of 45 days.

• Ofori-Boateng et al. [10] identified **6 individual token transfer graphs**, and stacked them as a **multi-layer network**, each layer denoting a specific token. The edges are *directed and weighted*, where an edge weight indicates the transferred token value. Moreover, the authors considered *daily* sequence of such multi-layer networks for anomaly detection.

• Poursafaei et al. [53], Chen et al. [55], and Wu et al. [54] studied the **(external) transaction graph** and annotated each edge (or node) with the number of transactions, amount of ether transferred, and time interval of the transactions.

• [56] studied **8 individual ERC721 token transfer graphs**. The nodes are EOAs, and each edge stores the cost of buying the token, the token ID, and the transaction time.

Summary. The wide spectrum of graphs constructed by our 25 surveyed papers is a *testimony to the rich and diverse ecosystem of Ethereum blockchain*. For instance, external transactions and token-based graphs can be further classified into multiple sub-categories, e.g., user-to-user, contract-to-contract, contract creation and invocation graphs, full token network, individual token networks, ERC20 token graphs, ERC721 token graphs, etc. (Figure 2). Moreover, one can obtain datasets for *different kinds of graphs* related research, including static graphs, dynamic graphs, temporal snapshot graphs, directed graphs, weighted graphs, simple and multi-graphs, attributed graphs, multi-layer networks, and even datasets for machine learning and topological data analysis. *This demonstrates the research value of the data stored on Ethereum blockchain*. We present in Figure 2 a **summary diagram** of several graphs that can be constructed based on interactions between accounts, transactions, and token transfers; together with their applications. We hope that our summary diagram would be a starting point for future research in this direction.

C. Graph Properties and Applications of Ethereum Networks

We next focus on graph properties, topological data analysis, and machine learning algorithms applied over Ethereum graphs, as well as the target applications demonstrated in the literature (Table II). We conclude with a summary of insights.

Graph property analysis. Majority of the works conducted graph-based analysis by measuring various graph properties, which can be classified as: (a) global properties, also known as "summary features", and (b) local properties or "local features" of individual nodes and edges [44]. Important local properties analyzed on Ethereum graphs are node degree distribution (including in- and out-degrees and their ratios), node centrality measures such as degree, closeness, betweenness, PageRank, and Eigenvector centrality. Among global properties, most prominent ones studied are connected components, reciprocity, assortativity, maximum clique, core decomposition, density, triangle and motif counts, community, global clustering coefficient, shortest path and diameter.

Topological data analysis (TDA). Both [45], [10] conducted topological data analysis on Ethereum networks for anomaly detection, the key concepts include simplicial complex, persistent homology, Betti number, functional data depth, and stacked persistence diagram. TDA systematically infers qualitative and quantitative geometric and topological structures of blockchain transaction graphs at multiple resolutions. Therefore, TDA can capture subtler patterns in transaction graphs, which are often associated with illicit or malicious activity and these are inaccessible with more conventional methods based on various forms of information aggregation [10].

Machine learning (ML) methods. Five past works performed graph representation learning with Ethereum blockchain graphs. Lin et al. [49], [50], Poursafaei et al. [53], and Wu et al. [54] designed temporal, node, and edge features-biased random walks for graph representation learning. Chen et al. [55] performed graph convolutional neural network (GCN)-based node embedding. Chen and Ng [41] proposed motif-based LSTM model for Ethereum token price prediction.

Target applications and findings. Bulk of the works conducted graph analysis to gain insights into transaction and token transfers. Some of them also considered explicit downstream tasks, e.g., node classification, link prediction, anomaly detection, and token price prediction.

• [36]: The degree distributions in MFG, CCG, and CIG follow power-law, indicating that a few developers created

many smart contracts. Based on various centrality measures, financial applications, e.g., exchange markets are the most important nodes in all three graphs. The authors also conducted cross-graph analysis to address two security issues.

• [37]: The transaction graph does not densify with time, converges to heavy-tailed distribution, and shows disassortative mixing – new nodes mostly connect to high-degree nodes.

• [38]: Full ERC20 tokens transfer graph's node degree distribution and token popularity follow power-law properties.

• [40]: The individual ERC20 tokens transfer graphs generally follow a star or a hub-and-spoke pattern.

• [41] proposed motif-based Long Short-Term Memory (LSTM) model for Ethereum token price prediction.

• [44]: For both TraceNet and TransactionNet, Log-normal, Weibull, and Power-law with cut-off are better fit than the classic power-law degree distribution. Mining pools and mixers create high outdegree nodes, whereas ICO smart contracts form high indegree nodes. Blockchain graphs have low transitivity, most frequent motifs observed are chain and star-shaped. Deleting only the highest-degree node (e.g., Binance, a global cryptocurrency exchange) may disconnect the entire largest weakly connected components in these graphs. However, blockchain graphs contain a single, large strongly connected component (SCC), and about 98% of the remaining nodes can either reach this SCC, or can be reached from the SCC. Based on [8], these networks are growing at a fast speed following the preferential attachment growth model. The user accounts remain active longer than smart contracts.

• [45] employed TDA, e.g., persistent homology and functional data depth to predict Ethereum-based tokens' price anomaly. [10] extended TDA over multi-layer Ethereum blockchain network for anomaly detection.

• [23] proposed an algorithm to verify if different tokens are created/ controlled by the same entity, and identified abnormal transactions in decentralized exchanges via graph analysis.

• [46]: Degree distributions of the studied individual token networks follow truncated power-law model, and each network, as a dynamical system, can be modeled as a damped harmonic oscillator, approaching to its equilibrium state.

• [49], [50] developed temporal random walk-based node embedding techniques for link prediction (i.e., predicting the occurrence of a transaction in a given graph) and node classification (phishing vs. genuine accounts). Among other works, [53], [54], [55] also designed random walk and GCN-based node embedding methods for node classification.

• [9] reported a strong correlation between the size of the user-to-user transaction graph and the average Ether price. The distribution of wealth, degree, and transaction number always remain unfair throughout the development of Ethereum.

• [56] proposed a methodology to identify the major NFT owners and follow their buying and selling patterns.

Summary. Given the market capitalization of Ethereum, downstream tasks such as node classification, link prediction, address clustering, asset price prediction, and anomaly detection (Figure 2) are critical in *anti-money laundering, criminal usage, abuse, and fraud detection, transaction risk prediction, blockchain intelligence*, etc. Researchers working on natural language processing and sentiment analysis using tweets, online articles, cryptocurrency prices and charts, Google Trends about blockchain [57] could find supporting views based on data analysis with Ethereum blockchain graphs. Anomaly detection with historical transaction data can be utilized by companies to *build safer blockchain ecosystems*.

IV. Conclusions and Future Work

We conducted a survey of literature on graph analysis with the Ethereum blockchain data. We first provided a brief introduction to Ethereum's heterogeneous ecosystem and data extraction tools. Next, we identified twenty five research papers published at peer-reviewed venues, and categorized them according to their (a) publication venues, years, categories, publishers, and authors' affiliations; (b) data extraction methods, dataset durations, and the graphs constructed; (c) graph properties, topological data analysis, and machine learning algorithms applied, as well as the target applications demonstrated. Our article is timely and would be valuable to graph data scientists and blockchain researchers.

Future work can be in several important directions.

— **First,** there is little work on **graph analysis with dApps and DeFi**. Accounts interact with each other based on different dApps and DeFi protocols, thus forming graph structures. One can investigate their graph properties, similarities and differences based on graph embedding, and identify anomalies.

— **Second,** there are relatively less works on **graph analysis of the individual ERC20 token subnetworks**, with the exceptions of [40], [45], [46], [10]. However, [45], [10] conducted topological data analysis, and [40], [46] did not study global and local graph properties extensively, neither their temporal evolutions. One may correlate these properties with real-world incidents, e.g., token prices, popularity, Google trends, etc., that would lead to more accurate forecasting.

— **Third,** due to several modes of interactions among EOAs and contracts, e.g., transactions, token transfers, dApps and DeFi usage, one may construct a **multi-layer network**, where each layer will denote one specific mode of interaction. Multi-layer graphs are an expressive model of real-world activities, and would be an interesting area of study.

— **Fourth,** Ethereum accounts can be grouped into various categories and granularity, e.g., miners, mining pools, mixers, exchanges, phishing accounts, ICO contracts, gambling games, etc. Once we cluster them based on their categories and/or roles in the network, the resulting graph structure might be very different from the initial one; therefore, graph property measurements would also vary. One can conduct graph analysis in an **OLAP** (online analytical processing) manner, by **drilling-up/down based on account groups and hierarchical categories**. Visualization at multiple resolutions will be beneficial to end-users for deriving insights.

— **Fifth,** due to highly dynamic nature of Ethereum accounts and transactions, employed ML models must deal with **data and model drifts**. Drift detection, incremental learning, machine unlearning, and continuous learning can be used.

REFERENCES

[1] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," 2008.

[2] C. G. Akcora, M. Kantarcioglu, and Y. R. Gel, "Blockchain Data Analytics," in *ICDM*, 2018.

[3] A. Kamiali, R. Kramberger, and I. Fister, "Synergy of Blockchain Technology and Data Mining Techniques for Anomaly Detection," *Applied Sciences*, vol. 11, no. 17, 2021.

[4] H. Heo and S. Shin, "Understanding Block and Transaction Logs of Permissionless Blockchain Networks," *Secur. Commun. Networks*, vol. 2021, pp. 9 549 602:1–9 549 602:18, 2021.

[5] C. G. Akcora, M. Kantarcioglu, and Y. R. Gel, "Data Science on Blockchains," in *KDD*, 2021.

[6] F. E. Oggier, A. Datta, and S. Phetsouvanh, "An Ego Network Analysis of Sextortionists," *Soc. Netw. Anal. Min.*, vol. 10, no. 1, p. 44, 2020.

[7] E. Rezaee, A. M. Saghiri, and A. Forestiero, "A Survey on Blockchain-Based Search Engines," *Applied Sciences*, vol. 11, no. 15, 2021.

[8] L. Zhao, S. S. Gupta, A. Khan, and R. Luo, "Temporal Analysis of the Entire Ethereum Blockchain Network," in *WWW*, 2021.

[9] Q. Bai, C. Zhang, Y. Xu, X. Chen, and X. Wang, "Evolution of Ethereum: A Temporal Graph Perspective," in *IFIP Net. Conf.*, 2020.

[10] 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.

[11] J. Wu, J. Liu, Y. Zhao, and Z. Zheng, "Analysis of Cryptocurrency Transactions from a Network Perspective: An Overview," *J. Netw. Comput. Appl.*, vol. 190, p. 103139, 2021.

[12] F. Victor, P. Ruppel, and A. Küpper, "A Taxonomy for Distributed Ledger Analytics," *Computer*, vol. 54, no. 2, pp. 30–38, 2021.

[13] C. G. Akcora, Y. R. Gel, and M. Kantarcioglu, "Blockchain Networks: Data Structures of Bitcoin, Monero, Zcash, Ethereum, Ripple, and Iota," *WIREs Data Mining Knowl. Discov.*, vol. 12, no. 1, 2022.

[14] W. Chen, J. Wu, Z. Zheng, C. Chen, and Y. Zhou, "Market Manipulation of Bitcoin: Evidence from Mining the Mt. Gox Transaction Network," in *INFOCOM*, 2019.

[15] C. G. Akcora, Y. Li, Y. R. Gel, and M. Kantarcioglu, "BitcoinHeist: Topological Data Analysis for Ransomware Prediction on the Bitcoin Blockchain," in *IJCAI*, 2020.

[16] N. C. Abay, C. G. Akcora, Y. R. Gel, M. Kantarcioglu, U. D. Islambekov, Y. Tian, and B. M. Thuraisingham, "ChainNet: Learning on Blockchain Graphs with Topological Features," in *ICDM*, 2019.

[17] H. A. Kalodner, M. Möser, K. Lee, S. Goldfeder, M. Plattner, A. Chator, and A. Narayanan, "BlockSci: Design and Applications of a Blockchain Analysis Platform," in *USENIX Security Symposium*, 2020.

[18] T. Cao, J. Yu, J. Decouchant, X. Luo, and P. Veríssimo, "Exploring the Monero Peer-to-Peer Network," in *Financial Cryptography and Data Security*, 2020.

[19] Y. Zhao, J. Liu, Q. Han, W. Zheng, and J. Wu, "Exploring EOSIO via Graph Characterization," in *BlockSys*, 2020.

[20] B. Guidi and A. Michienzi, "Users and Bots Behaviour Analysis in Blockchain Social Media," in *SNAMS*, 2020.

[21] W. Chan and A. Olmsted, "Ethereum Transaction Graph Analysis," in *ICITST*, 2017.

[22] F. Victor, "Address Clustering Heuristics for Ethereum," in *FC*, 2020.

[23] 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.

[24] A. M. Antonopoulos and G. Wood, *Mastering Ethereum*. O'Reilly Media, 2018.

[25] G. A. Oliva, A. E. Hassan, and Z. M. J. Jiang, "An Exploratory Study of Smart Contracts in the Ethereum Blockchain Platform," *Empir. Softw. Eng.*, vol. 25, no. 3, pp. 1864–1904, 2020.

[26] K. Wu, "An Empirical Study of Blockchain-based Decentralized Applications," *CoRR*, 2019.

[27] C. R. Harvey, A. Ramachandran, and J. Santoro, *DeFi and the Future of Finance*. John Wiley & Sons, 2021.

[28] A. Day and E. Medvedev, "Ethereum-ETL," cloud.google.com/blog/products/data-analytics/ethereum-bigquery-public-dataset-smart-contract-analytics, 2018.

[29] P. Zheng, Z. Zheng, J. Wu, and H. Dai, "Xblock-eth: Extracting and exploring blockchain data from ethereum," *IEEE Open J. Comput. Soc.*, vol. 1, pp. 95–106, 2020.

[30] V. H. Su, S. S. Gupta, and A. Khan, "Automating ETL and Mining of Ethereum Blockchain Network," in *WSDM*, 2022.

[31] S. Bragagnolo, H. Rocha, M. Denker, and S. Ducasse, "Ethereum Query Language," in *WETSEB@ICSE*, 2018.

[32] A. Estupinan, "Analysis of Modern Blockchain Networks using Graph Databases," Master's thesis, Technische Universitat Berlin, 2020.

[33] D. S. H. Tam, W. C. Lau, B. Hu, Q. Ying, D. M. Chiu, and H. Liu, "Identifying Illicit Accounts in Large Scale E-payment Networks - A Graph Representation Learning Approach," *CoRR*, 2019.

[34] S. Somin, G. Gordon, and Y. Altshuler, "Social Signals in the Ethereum Trading Network," *CoRR*, 2018.

[35] D. T. Vu, "User Identification and Behaviour Patterns on the Ethereum Blockchain: An Exploratory Study," Master's thesis, Technical University of Munich, 2021.

[36] 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.

[37] 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.

[38] S. Somin, G. Gordon, and Y. Altshuler, "Network Analysis of ERC20 Tokens Trading on Ethereum Blockchain," in *Complex Systems*, 2018.

[39] A. Anoaica and H. Levard, "Quantitative Description of Internal Activity on the Ethereum Public Blockchain," in *NTMS*, 2018.

[40] F. Victor and B. K. Lüders, "Measuring Ethereum-based ERC20 Token Networks," in *Financial Cryptography and Data Security*, 2019.

[41] Y. Chen and H. K. T. Ng, "Deep Learning Ethereum Token Price Prediction with Network Motif Analysis," in *ICDM Workshops*, 2019.

[42] A. P. Motamed and B. Bahrak, "Quantitative Analysis of Cryptocurrencies Transaction Graph," *Appl. Netw. Sci.*, vol. 4, no. 1, p. 131, 2019.

[43] D. Guo, J. Dong, and K. Wang, "Graph Structure and Statistical Properties of Ethereum Transaction Relationships," *Inf. Sci.*, vol. 492, pp. 58–71, 2019.

[44] 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.

[45] 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.

[46] S. Somin, Y. Altshuler, G. Gordon, A. . Pentland, and E. Shmueli, "Network Dynamics of a Financial Ecosystem," *Sci Rep*, vol. 10, no. 4587, 2020.

[47] 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.

[48] 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.

[49] 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.

[50] D. Lin, J. Wu, Q. Yuan, and Z. Zheng, "T-EDGE: Temporal WEighted MultiDiGraph Embedding for Ethereum Transaction Network Analysis," *Frontiers in Physics*, vol. 8, 2020.

[51] J. Z. G. Mascarenhas, A. Ziviani, K. Wehmuth, and A. B. Vieira, "On the Transaction Dynamics of the Ethereum-based Cryptocurrency," *Journal of Complex Networks*, vol. 8, no. 4, 2020.

[52] S. Somin, G. Gordon, A. Pentland, E. Shmueli, and Y. Altshuler, "ERC20 Transactions over Ethereum Blockchain: Network Analysis and Predictions," in *SBP-BRiMS (Working Papers)*, 2020.

[53] F. Poursafaei, R. Rabbany, and Z. Zilic, "SigTran: Signature Vectors for Detecting Illicit Activities in Blockchain Transaction Networks," in *PAKDD*, 2021.

[54] J. Wu, Q. Yuan, D. Lin, W. You, W. Chen, C. Chen, and Z. Zheng, "Who Are the Phishers? Phishing Scam Detection on Ethereum via Network Embedding," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 52, no. 2, pp. 1156–1166, 2022.

[55] L. Chen, J. Peng, Y. Liu, J. Li, F. Xie, and Z. Zheng, "Phishing Scams Detection in Ethereum Transaction Network," *ACM Trans. Internet Techn.*, vol. 21, no. 1, pp. 10:1–10:16, 2021.

[56] 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.

[57] A.-D. Vo, Q.-P. Nguyen, and C.-Y. Ock, "Sentiment Analysis of News for Effective Cryptocurrency Price Prediction," *International Journal of Knowledge Engineering*, vol. 5, no. 2, pp. 47–52, 2019.