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

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

ABSTRACT

A blockchain platform is generally cohabited by human users, autonomous agents, cryptocurrencies, other digital assets, and decentralized protocols. As an example, consider the Ethereum ecosystem - currently the most actively used and the second-largest cryptocurrency network by market capitalization after Bitcoin. Ether is the native cryptocurrency of Ethereum that is transferred between accounts. Ethereum accounts are of two types: Externally owned accounts are controlled by users, whereas a contract account is controlled by a smart contract, which is an autonomous agent and can execute complex code across a decentralized network. For instance, smart contracts can define tokens that are digital assets in the blockchain platform. Decentralized applications (dApps) such as exchanges, wallets, and DeFi may combine multiple smart contracts and their protocols constitute a collection of rules that govern dApps in a decentralized blockchain platform. Complex interactions across various actors in blockchains generate massive-scale, dynamic, heterogeneous, and multi-modal data that are often publicly accessible and can be considered big data – an emerging trend since the past decade. Analysis of blockchain data using the latest data management and AI techniques is critical for the improvement of the blockchain technology, such as detecting and predicting trends, anomalies, e-crimes, and key actors.

In the first part of the talk, I shall discuss our recent work on blockchain data extraction and graph construction, graph mining, topological data analysis, and machine learning methods for various target applications such as detecting market manipulators in the blockchain world including the collapse of the stablecoin LunaTerra, Ethereum's switch from Proof-of-Work (PoW) to Proof-of-Stake (PoS), and the stablecoin USDC's temporary peg loss. In the second part, I shall showcase the contributions of blockchain technology in the growing ecosystem of data management and AI – in the form of diverse datasets, tools, novel challenges, and algorithms. I shall conclude by emphasizing future research directions such as cross-chain data analysis, combining signals from external sources, e.g., tweets and social media data about blockchains for holistic predictions, higher-order and multi-modal network analysis, designing of temporal machine learning and machine unlearning algorithms.

VLDB Workshop Reference Format:
Arijit Khan. Data Management and AI for Blockchain Data Analysis: A Round Trip and Opportunities. VLDB 2024 Workshop: FAB.

BIOGRAPHY

Arijit Khan is an IEEE senior member, an ACM distinguished speaker, and an associate professor in the Department of Computer Science, Aalborg University, Denmark. He earned his PhD from the Department of Computer Science, University of California, Santa Barbara, USA, and did a post-doc in the Systems group at ETH Zurich, Switzerland. He has been an assistant professor in the School of Computer Science and Engineering, Nanyang Technological University, Singapore. Arijit is the recipient of the prestigious IBM PhD Fellowship in 2012-13, a VLDB Distinguished Reviewer award (2022), and a SIGMOD Distinguished PC award (2024). He published more than 80 papers in premier databases and data mining conferences and journals including ACM SIGMOD, VLDB, IEEE TKDE, IEEE ICDE, SIAM SDM, USENIX ATC, EDBT, The Web Conference (WWW), ACM WSDM, ACM CIKM, ACM TKDD, and ACM SIGMOD Record. Arijit served as the co-chair of Big-O(Q) workshop co-located with VLDB 2015, LLM+KG workshop co-located with VLDB 2024, and KG for responsible AI workshop co-located with CIKM 2024, wrote a book on uncertain graphs in Morgan & Claypool's Synthesis Lectures on Data Management. Dr Khan is serving as an associate editor of IEEE TKDE 2019-2024 and ACM TKDD 2023-present, IEEE ICDE TKDE poster track co-chair 2023, ACM CIKM short paper track program co-chair 2024, and IEEE ICDE demonstration track program co-chair 2025.

ACKNOWLEDGMENTS

This talk is based on recent works [1–7]. Arijit Khan acknowledges support from the Novo Nordisk Foundation grant NNF22OC0072415.

REFERENCES

[1] P. Azad, C. G. Akcora, and A. Khan. 2024. Machine Learning for Blockchain Data Analysis: Progress and Opportunities. *CoRR* abs/2404.18251 (2024).
[2] A. Khan. 2022. Graph Analysis of the Ethereum Blockchain Data: A Survey of Datasets, Methods, and Future Work. In *IEEE International Conference on Blockchain*. 250–257.
[3] A. Khan and C. G. Akcora. 2022. Graph-based Management and Mining of Blockchain Data. In *ACM International Conference on Information & Knowledge Management*. 5140–5143.
[4] X. T. Lee, A. Khan, S. Sen Gupta, Y. H. Ong, and X. Liu. 2020. Measurements, Analyses, and Insights on the Entire Ethereum Blockchain Network. In *The Web Conference*. 155–166.
[5] V. H. Su, S. S. Gupta, and A. Khan. 2022. Automating ETL and Mining of Ethereum Blockchain Network. In *ACM International Conference on Web Search and Data Mining*. 1581–1584.
[6] L. Zhao, S. S. Gupta, A. Khan, and R. Luo. 2021. Temporal Analysis of the Entire Ethereum Blockchain Network. In *The Web Conference*. 2258–2269.
[7] J. Zhu, A. Khan, and C. G. Akcora. 2024. Data Depth and Core-based Trend Detection on Blockchain Transaction Networks. *Frontiers Blockchain* 7 (2024).

Proceedings of the VLDB Endowment. ISSN 2150-8097.