Data Science on Blockchains

Cuneyt Gurcan Akcora
Computer Science and Statistics
University of Manitoba
Winnipeg, Canada
cuneyt.akcora@umanitoba.ca

Yulia R. Gel
Mathematical Sciences
University of Texas at Dallas
Dallas, USA
ygl@utdallas.edu

Murat Kantarcioglu
Computer Science
University of Texas at Dallas
Dallas, US
muratk@utdallas.edu

***Abstract*—Over the last couple of years, Bitcoin cryptocurrency and the Blockchain technology that forms the basis of Bitcoin have witnessed an unprecedented attention. Designed to create a secure distributed platform without central regulation, Blockchain is heralded as a novel paradigm that will be as powerful as Big Data, Cloud Computing, and Machine Learning.**

The Blockchain technology garners an ever increasing interest of researchers in various domains that benefit from scalable cooperation among trust-less parties. As Blockchain applications proliferate, so does the complexity and volume of data stored by Blockchains. Analyzing this data has emerged as an important research topic, already leading to methodological advancements in the information sciences.

In this tutorial, we offer a holistic view on applied Data Science on Blockchains. Starting with the core components of Blockchain, we will detail the state of art in Blockchain Data Analytics for graph, security and finance domains. Our examples will answer questions, such as, how to parse, extract and clean the data stored in blockchains?, how to store and query Blockchain data? and what features could be computed from blockchains?

***Index Terms*—Blockchain, Ethereum, Bitcoin, Smart Contracts**

OUTLINE

This decade has been marked with the rise of Blockchain based technologies. In its core, Blockchain is a distributed public ledger that stores transactions between two parties without requiring a trusted central authority. On a blockchain, two unacquainted parties can create an unmodifiable transaction that is permanently recorded on the ledger to be seen by the public. The first application of Blockchain has been the Bitcoin [22] cryptocurrency. Bitcoin's success has ushered an age known as the Blockchain 1.0 [31]; currently there are more than 1000 Blockchain based cryptocurrencies, known as **alt-coins**. With the arrival of Ethereum and Nem.io in 2015, the age of Blockchain 2.0 has been underway. Although Ethereum uses a currency of its own (i.e., the Ether) as in Bitcoin, its distinguishing mark is the **Smart-Contracts** feature that allows unmodifiable, unstoppable code execution on the blockchain. Termed as the "World Computer", Ethereum is a platform to create software based public Smart-Contracts and execute them in a Turing Complete way. The Smart-Contract functionality has been a popular feature in other platforms such as Nem. Adoption of Ethereum and other Blockchain platforms for societal use is termed as the upcoming age of Blockchain 3.0. Researchers imagine the diffusion of Blockchain's decentralized and authority-less mechanisms to create consensus in diverse aspects of the modern life [14], [37]. As legendary venture capitalist Marc Andreessen states "the consequences of this breakthrough are hard to overstate" [5]. Some observers compare the inception of Blockchain to the invention of double entry accounting that revolutionized the business world [34]. The emerging Blockchain based applications include voting (FollowMyVote, Social Krona), identity services (Bitnation, Hypr), provenance (Everledger, Chronicled) and copyright management (LBRY, Blockphase). Although it is hard to predict the future of impact of Blockchain, it is safe to say that it will enable many novel broader societal applications.

As strikingly more new Blockchain applications are developed and deployed, novel interdisciplinary tools and algorithms for Blockchain Data Analytics are of increasing importance for enhancing our understanding of the emerging phenomena, from combating ransomware to terrorism prevention. By facilitating cross-disciplinary exchange of ideas, this tutorial will provide a unique perspective for Blockchain knowledge synthesis, creation and transfer.

A. Relevance and timeliness of Blockchain Data Analytics

Data Science on Blockchains is receiving attention from industry, academia and state actors. Governments are implementing rules to govern practices in Blockchain. As these rules come into effect, understanding Blockchain data and developing capabilities to mine this data will have vital importance. The modeling of Blockchain data is a new area with little to no precompetitive research and working groups. There are many research areas with great potential in advancing Blockchain applications, ranging from graph theory, topological data analysis, random matrix theory and probabilistic graphical modeling. Equally, there are several open problems which must be addressed before Blockchain becomes more mainstream and gains favorability with governmental regulations. Modeling the effect of transaction costs on the long-term stability of Bitcoin adoption, identifying more energy efficient and scalable mining mechanisms which avoid excessive computation, and determining how to identify criminal activity are just a few of these problems. We believe that ICDE, with its focus on Data Mining, Data Engineering and Machine Learning, is a premier venue for disseminating the knowledge on Data Science on Blockchains.

B. Target audience

We will start by explaining how Blockchains work, and continue to research issues in data storage, querying and analytics. We assume no prior knowledge from the audience.

C. Benefits for participants

Although Blockchain data is publicly available, many researchers have difficulty parsing and extracting the data before they can use it in their research. For example, Smart Contracts on Ethereum are difficult to parse without running a full node. In this tutorial we will outline how Blockchain data can be accessed by using existing tools, and what problems must be addressed specifically for Blockchain data mining. We believe that our introduction will be appreciated specifically by researchers from well established fields, such stream mining and graph theory, whose models can be easily adapted to run on Blockchains and produce answers to urgent questions in Blockchain research.

Interest: Our previous tutorials at major conferences have created much enthusiasm. For example, in ICDM 2018, the day of our tutorial was declared the Blockchain Day. Similarly, we expect a big audience in ICDE 2020.

D. Similar Tutorials

Akcora, Gel and Kantarcioglu have previously presented a Blockchain Data Analytics tutorial at IEEE International Conference on Data Mining 2018 [2]. An updated version of the tutorial was presented at the Pacific-Asia Conference on Knowledge Discovery and Data Mining in March 2019. Compared to earlier tutorials at ICDE and other venues [18], [20], we will offer a more data-centric view and present recent important developments.

I. Tutorial Scope and Structure

We divide the tutorial into two 1.5 hour sessions. In the first part of this **3 hour tutorial**, we will teach how the Blockchain technology works. In the second half, we will explain how Blockchain data coming from cryptocurrencies and platforms can be modeled, analyzed, and mined for various applications. The tutorial is planned for three hours with the following tentative schedule:

- Part 1 (1.5 hours).
  - **Consensus**: Nakamoto consensus vs distributed consensus algorithms. Proof-of-Work, Proof-of-Stake and other Proof-of-X schemes.
  - **Model**: Unspent transaction output (UTXO) based blockchains, account based blockchains, Smart Contracts, Crypto-tokens, Decentralized organizations. Privacy aware Blockchains: Monero, Zcash. Second layer solutions, Lightning Network.
  - **Data**: Types of data stored on blockchains. Private and public blockchains.
  - **Storage and querying**: Databases for Blockchain data. Query models. Sharding and Scalability issues.
- Part 2 (1.5 hours).
  - **Network**: Data models for UTXO and account based blockchains. Multilayer token networks. Centrality, influence and propagation on Blockchain networks.
  - **Methods**: Motif and chainlet analysis. Topological Data Analysis for weighted, directed networks. Heuristics and address clustering on Blockchains. Centrality and influence analysis of nodes. Cyber security research: Detecting Darknet market payments, price pump and dump schemes, money laundering, ransomware payments.
  - **Obfuscation efforts**: naive hiding patterns, coin mixing and shapeshifting.
  - **Visualization**: Address, cluster, path and flow visualization on Blockchains.

A. Detailed Outline

1) Core Blockchain Technology: Blockchain data coming from existing platforms, coins and technologies have resulted in a very diverse set of data types. In this first part of our tutorial we will take a brief tour of Proof-of-X schemes that change how the data is created. We will describe how the Blockchain technology was created [22], how ideas were adopted from various research fields and how they were put to use in creating a distributed and secure ledger [12]. Each component of the Blockchain technology, such as transactions, blocks and miners, will be described in detail for cryptocurrencies and Blockchain platforms. We will explain the new generation of privacy aware blockchains, such as Monero [21] and Zcash [27]. In this line, we will explain the Lightning Network on Bitcoin blockchain as a prominent second layer solution.

Existing tools store Blockchain data in key-value stores (e.g., LevelDB), document stores (CouchDB), graph databases (e.g., Neo4J), and in rare cases, in relational databases (R3 Corda) [19]. Type of stored data also depends on the used blockchain: public vs private. These databases optimize certain aspects of data storage; LevelDB compresses stored key-value pairs, and Neo4J allows fast graph queries, whereas relational databases allow for fast join and merge operations. As blockchains age, storage issues have started to become a major concern due to long chains of data [12]. In a typical blockchain, each node stores a complete copy of the chain data. Since 2009, Bitcoin has created 600K blocks of data where each block is at most 1MB in size. Ripple and Ethereum have millions of blocks with various sizes. Even without any query or analytics capabilities, storing all this raw data blocks is burdensome for ordinary users. For transaction output based blockchains, such as Litecoin, this chain data can be simplified locally by omitting transition chainlets [1] or compressing spent outputs. In account based blockchains, such as Ethereum, multiple transactions involving the same users can be aggregated. In both cases, the data storage model can be locally modified to reduce stored data. Smart Contract based tokens and decentralized organizations require storing and analyzing software code in blockchains [6].

We will explore Blockchain model specific solutions to compress [35], summarize and store [38] Blockchain data more efficiently while preserving the utility of the data with respect to Data Science. This section will provide a necessary background on how Blockchain data can be prepared to run Statistical and Machine Learning models [11].

2) Data Analytics on Blockchain: Account based blockchains, such as Ethereum, are modeled as directed, edge-weighted multi-graphs (similar to traditional social network models). However, unspent transaction output (UTXO) based blockchains require novel data models. In an UTXO based blockchain, an owner of multiple addresses (i.e., a person may have many addresses) can combine them in a transaction and send coins to multiple output addresses. Hence, the blockchain consists of two types of nodes: *transactions* or *addresses* that are inputs/outputs of transactions.

Earlier results on Blockchain analysis constructed graphs only with a single type of node (i.e., transactions or addresses), and currency transfers formed edges between nodes (see [3] for a review). By choosing a single type of node, these approaches ignore either address or transaction information. In contrast, a heterogeneous Blockchain graph can be constructed with *both address* and *transaction nodes*. Although influence works are yet lacking, new address centrality measures have been proposed for blockchains [25], [26]. We will detail these research direction and algorithms in Blockchain data modeling and analytics [10], [29].

As Bitcoin became popular, a number of studies aimed at using various network characteristics for price predictions. For instance, [13], [30] employ such network features as mean account balance, number of new edges and clustering coefficients. In turn, network flows and temporal behavior of the network have been used as alternative price predictors by [36] and [15], respectively.

Studies in network features show that since 2010 the Bitcoin network can be considered a scale-free network [17]. In- and out-degree distributions of the transactions graphs are highly heterogeneous and exhibit a disassortative behavior [16]. Active entities on the network change frequently, but there are consistently active entities [23]. The most central nodes in the network are coin exchange sites [7].

As all transactions are one-to-one, account based blockchains enable the usage of traditional graph analysis tools easily [8], [9]. However care must be taken to extract internal transactions from ordinary transactions, so that all relationships (i.e., token buy/sell) between addresses can be modeled on the graph.

As a second issue, the complete Ethereum graph have overlapping layers of token networks; each token can be represented with a separate graph on the Ethereum network where nodes are user/contract addresses. A token network is a directed, weighted multi-graph. Two token networks may share nodes but not edges. The complete Ethereum graph consists of layers of token networks. Multiple edges can exist between two nodes of the Ethereum graph, and each edge can transfer a different token. On the Ethereum blockchain, it is not rare to see hundreds of edges between two nodes.

Although Blockchain is widely touted as a secure and private realm, its short history shows cases with grave risks such as ruined reputations [33], lost finances [28] and stolen identities [32]. We will provide multiple examples on how the Blockchain technology can be hijacked and used in malicious ways [4]. In the tutorial we will *discuss how Blockchain Data Analytics* could be conducted to attack individual privacy and discuss using Blockchain analytics techniques for detecting malicious uses of the cryptocurrencies and Blockchain platforms. Visualization of Blockchain networks pose another interesting research direction that will be covered [24].

Open Research Questions and Approaches: In concluding our tutorial, we will present some of the open challenges in storing and querying Blockchain data. In particular, this includes summarizing Blockchain data and using temporal models to reduce the analyzed data size. We will briefly show several possible directions to address these data challenges on Blockchains and leave some of the directions open for discussion.

In Data Analytics, influence of nodes in Blockchain graphs remains an open problem. In recent years, Smart Contracts on Blockchain platforms have become an important tool with decentralized organizations and Smart Contract based tokens. Analysis of contract and token networks provides timely and important information in price prediction, price manipulation detection and trend analysis. We will review related works in this emerging direction of Data Analytics for Blockchain platforms, and provide pointers for datasets, tools and projects.

TUTORS

Cuneyt Gurcan Akcora (http://cakcora.github.io) is an Assistant Professor of Computer Science and Statistics at the University of Manitoba, Canada. Before that, he was a fellow in the Departments of Statistics and Computer Science at the University of Texas at Dallas. He received his Ph.D. from University of Insubria, Italy and his M.S. from State University of New York at Buffalo, USA. His primary research interests are Data Science on complex networks and large scale graph analysis, with applications in social, biological, IoT and Blockchain networks. He is a Fulbright Scholarship recipient, and his research works have been published in leading conferences and journals including TKDE, VLDB, ICDM and ICDE.

Yulia R. Gel (https://personal.utdallas.edu/~yxg142030/) is Professor in the Department of Mathematical Science at the University of Texas at Dallas. Her research interests include statistical foundation of Data Science, inference for random graphs and complex networks, time series analysis, and predictive analytics. She holds a Ph.D in Mathematics, followed by a postdoctoral position in Statistics at the University of Washington. Prior to joining UT Dallas, she was a tenured faculty member at the University of Waterloo, Canada. She also held visiting positions at Johns Hopkins University, University of California, Berkeley, and the Isaac Newton Institute

for Mathematical Sciences, Cambridge University, UK. She served as a Vice President of the International Society on Business and Industrial Statistics (ISBIS), and is a Fellow of the American Statistical Association.

Murat Kantarcioglu (https://personal.utdallas.edu/~muratk/) is a Professor in the Computer Science Department and Director of the UTD Data Security and Privacy Lab at the University of Texas at Dallas and a visiting scholar at Harvard University Data Privacy Lab. He is a recipient of NSF CAREER award, and Purdue CERIAS Diamond Award for Academic excellence. His research focuses on creating technologies that can efficiently extract useful information from data without sacrificing privacy or security. Over the years, his research has been supported by grants from NSF, AFOSR, ONR, NSA, and NIH. In addition, he has published over 160 peer reviewed papers related to data security and privacy-preserving data mining. His research work has been covered by the media outlets, such as Boston Globe, ABC News, and has received three best paper awards.

REFERENCES

[1] C. G. Akcora, A. K. Dey, Y. R. Gel, and M. Kantarcioglu, "Forecasting bitcoin price with graph chainlets," *The 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining, PaKDD*, 2018.

[2] C. G. Akcora, M. F. Dixon, Y. R. Gel, and M. Kantarcioglu, "Blockchain data analytics," *Intelligent Informatics*, p. 4, 2018.

[3] C. G. Akcora, Y. R. Gel, and M. Kantarcioglu, "Blockchain: A graph primer," *arXiv preprint arXiv:1708.08749*, 2017.

[4] C. G. Akcora, Y. Li, Y. R. Gel, and M. Kantarcioglu, "Bitcoinheist: Topological data analysis for ransomware detection on the bitcoin blockchain," *arXiv preprint arXiv:1906.07852*, 2019.

[5] M. Andreessen, "Why Bitcoin matters: https://dealbook.nytimes.com/2014/01/21/why-bitcoin-matters/," *New York Times*, vol. 21, 2014.

[6] M. Bartoletti and L. Pompianu, "An empirical analysis of smart contracts: platforms, applications, and design patterns," in *International conference on financial cryptography and data security*. Springer, 2017, pp. 494–509.

[7] A. Baumann, B. Fabian, and M. Lischke, "Exploring the bitcoin network." in *WEBIST (1)*, 2014, pp. 369–374.

[8] W. Chan and A. Olmsted, "Ethereum transaction graph analysis," in *2017 12th International Conference for Internet Technology and Secured Transactions (ICITST)*. IEEE, 2017, pp. 498–500.

[9] T. Chen, Y. Zhu, Z. Li, J. Chen, X. Li, X. Luo, X. Lin, and X. Zhange, "Understanding ethereum via graph analysis," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1484–1492.

[10] T. T. A. Dinh, R. Liu, M. Zhang, G. Chen, B. C. Ooi, and J. Wang, "Untangling blockchain: A data processing view of blockchain systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 7, pp. 1366–1385, 2018.

[11] A. Dubovitskaya, P. Novotny, S. Thiebes, A. Sunyaev, M. Schumacher, Z. Xu, and F. Wang, "Intelligent health care data management using blockchain: Current limitation and future research agenda," in *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*. Springer, 2019, pp. 277–288.

[12] M. El-Hindi, C. Binnig, A. Arasu, D. Kossmann, and R. Ramamurthy, "Blockchaindb: a shared database on blockchains," *Proceedings of the VLDB Endowment*, vol. 12, no. 11, pp. 1597–1609, 2019.

[13] A. Greaves and B. Au, "Using the bitcoin transaction graph to predict the price of bitcoin," *No Data*, 2015.

[14] H. Karlstrøm, "Do libertarians dream of electric coins? the material embeddedness of bitcoin," *Distinktion: Scandinavian Journal of Social Theory*, vol. 15, no. 1, pp. 23–36, 2014.

[15] D. Kondor, I. Csabai, J. Szüle, and G. Pósfai, M.and Vattay, "Inferring the interplay between network structure and market effects in bitcoin," *New J. of Phys.*, vol. 16, no. 12, p. 125003, 2014.

[16] D. Kondor, M. Pósfai, I. Csabai, and G. Vattay, "Do the rich get richer? an empirical analysis of the bitcoin transaction network," *PloS one*, vol. 9, no. 2, p. e86197, 2014.

[17] M. Lischke and B. Fabian, "Analyzing the bitcoin network: The first four years," *Future Internet*, vol. 8, no. 1, p. 7, 2016.

[18] S. Maiyya, V. Zakhary, D. Agrawal, and A. E. Abbadi, "Database and distributed computing fundamentals for scalable, fault-tolerant, and consistent maintenance of blockchains," *Proceedings of the VLDB Endowment*, vol. 11, no. 12, pp. 2098–2101, 2018.

[19] C. Mohan, "Tutorial: blockchains and databases," *Proceedings of the VLDB Endowment*, vol. 10, no. 12, pp. 2000–2001, 2017.

[20] ——, "Blockchains and databases: A new era in distributed computing," in *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE, 2018, pp. 1739–1740.

[21] M. Möser, K. Soska, E. Heilman, K. Lee, H. Heffan, S. Srivastava, K. Hogan, J. Hennessey, A. Miller, A. Narayanan *et al.*, "An empirical analysis of traceability in the monero blockchain," *Proceedings on Privacy Enhancing Technologies*, vol. 2018, no. 3, pp. 143–163, 2018.

[22] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008.

[23] M. Ober, S. Katzenbeisser, and K. Hamacher, "Structure and anonymity of the bitcoin transaction graph," *Future internet*, vol. 5, no. 2, pp. 237–250, 2013.

[24] F. Oggier, S. Phetsouvanh, and A. Datta, "Biva: Bitcoin network visualization & analysis," in *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2018, pp. 1469–1474.

[25] ——, "Entropic centrality for non-atomic flow networks," in *2018 International Symposium on Information Theory and Its Applications (ISITA)*. IEEE, 2018, pp. 50–54.

[26] B. B. F. Pontiveros, M. Steichen, and R. State, "Mint centrality: A centrality measure for the bitcoin transaction graph," in *2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*. IEEE, 2019, pp. 159–162.

[27] J. Quesnelle, "On the linkability of zcash transactions," *arXiv preprint arXiv:1712.01210*, 2017.

[28] D. Ron and A. Shamir, "How did dread pirate roberts acquire and protect his bitcoin wealth?" in *International Conference on Financial Cryptography and Data Security*. Springer, 2014, pp. 3–15.

[29] P. Ruan, G. Chen, T. T. A. Dinh, Q. Lin, B. C. Ooi, and M. Zhang, "Fine-grained, secure and efficient data provenance on blockchain systems," *Proceedings of the VLDB Endowment*, vol. 12, no. 9, pp. 975–988, 2019.

[30] M. Sorgente and C. Cibils, "The reaction of a network: Exploring the relationship between the bitcoin network structure and the bitcoin price," *No Data*, 2014.

[31] M. Swan, *Blockchain: Blueprint for a new economy*. O'Reilly Media, Inc., 2015.

[32] R. Upadhyaya and A. Jain, "Cyber ethics and cyber crime: A deep dwelved study into legality, ransomware, underground web and bitcoin wallet," in *Computing, Communication and Automation (ICCCA), 2016 International Conference on*. IEEE, 2016, pp. 143–148.

[33] M. Vasek and T. Moore, "There is no free lunch, even using bitcoin: Tracking the popularity and profits of virtual currency scams," in *International conference on financial cryptography and data security*. Springer, 2015, pp. 44–61.

[34] P. Vigna and M. J. Casey, *The age of cryptocurrency: how bitcoin and the blockchain are challenging the global economic order*. Macmillan, 2016.

[35] S. Wang, T. T. A. Dinh, Q. Lin, Z. Xie, M. Zhang, Q. Cai, G. Chen, B. C. Ooi, and P. Ruan, "Forkbase: An efficient storage engine for blockchain and forkable applications," *Proceedings of the VLDB Endowment*, vol. 11, no. 10, pp. 1137–1150, 2018.

[36] S. Y. Yang and J. Kim, "Bitcoin market return and volatility forecasting using transaction network flow properties," in *IEEE SSCI*, 2015, pp. 1778–1785.

[37] J. L. Zhao, S. Fan, and J. Yan, "Overview of business innovations and research opportunities in blockchain and introduction to the special issue," *Financial Innovation*, vol. 2, no. 1, pp. 1–28, December 2016. [Online]. Available: https://doi.org/10.1186/s40854-016-0049-2

[38] Y. Zhu, Z. Zhang, C. Jin, A. Zhou, and Y. Yan, "Sebdb: Semantics empowered blockchain database," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 2019, pp. 1820–1831.