Cuneyt G. Akcora, Yulia R. Gel, Murat Kantarcioglu

Statistics and Computer Science

University of Texas at Dallas

Proposal for University-Industry Research Collaboration

Chainalytics: Algorithms for Blockchain Data Analysis

cuneyt.akcora, ygl and muratk at utdallas.edu

Executive Summary

Over the last couple of years, cryptocurrencies and the blockchain technology that forms the basis of them have witnessed an unprecedented attention. Designed to facilitate a secure distributed platform without a central authority, Blockchain is heralded as a novel paradigm that will be as powerful as Big Data, Cloud Computing, and Machine Learning.

Blockchain applications have already matured to rival, and already in some cases, replace more traditional institutions as avenues of global activity. As Marc Andreessen states, "the consequences of the Blockchain breakthrough are hard to overstate". We emphasize an important consequence that often goes unnoticed: *blockchains store data that is invaluable to track important global activities ranging from financial transfers to initial coin offerings.* To understand the implications of the data stored on blockchains, institutions must develop advanced analytical capabilities to analyze this data.

Due to the lack of a general blockchain data analytics framework, many custom applications have been developed for analyzing blockchain data (e.g., [13, 14]). Although some recent projects (e.g., [8, 6, 10, 7]) attempt to change this situation for Bitcoin, to our knowledge, there exists yet *no general blockchain data analytics tool* that allows the creation of novel interdisciplinary data analytics applications for analyzing blockhain data elements such as smart-contracts, tokens and coin transfers.

As an inter-disciplinary team of researchers from Statistics and Computer Science, our aim is to fill this vital role and develop new methodology for Blockchain Data Analytics. Furthermore, our target is to strengthen academia-industry collaboration and to build partnership to foster research and development of innovative approaches to blockchain data analytics. We believe that especially in this critical area, industry and academia collaboration is critical to better align research efforts with the real hard problems. Our goal is to develop our research into Blockchain data analytics products or services in a collaboration with industrial partners.

Proposed Blockchain Data Analytics Framework

Our Blockchain Data Analytics tools can be summarized into two main directions: research on Blockchain based cryptocurrencies and Blockchain platforms.

Public availability of all cryptocurrency transactions allows us to create a complex network of financial interactions that can be used to study not only the blockchain graph, but the relationship between various blockchain network features and cryptocurrency dynamics. We have introduced a novel concept of blockchain motifs called *chainlets* [2]. Chainlets describe shapes of cryptocurrency transactions and further expand the ideas of network motifs and graphlets to Blockchain graphs. Chainlets provide an intuitive data model that can be used to explain how cryptocurrency networks evolve. Figure 1 shows the Bitcoin network through years by using subtypes of the chainlet data model. Furthermore, chainlet occurrence and amount matrices (see [2]) that we extract, can be used as an input in various machine learning models. We have utilized this information for a multitude of use cases, such as anomaly detection, price forecasting [2], and risk prediction [3]. For example, our recent work [1] on the Bitcoin network has shown that daily price prediction can be improved [1] by more than 45% by using chainlet data in a deep learning framework.

The methodology can be further applied to analyze longer chainlets gathered from multiple cryptocurrency networks to understand interdependencies among different cryptocurrencies. For example, our recent results [5] show that Bitcoin chainlets can be incorporated in Litecoin price prediction, resulting in an inter-coin ecosystem. One implication of this research is that we can utilize data from established coins in predicting the price of developing cryptocurrencies.

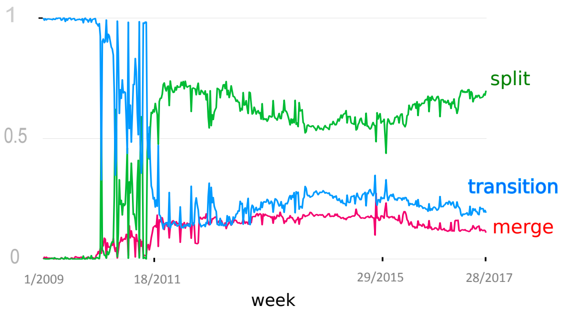

Figure 1: The Bitcoin network visualized in time through chainlet subtypes. The network stabilizes around June 2011.

Analyzing Blockchain Platform Activities: Providing an immensely richer and more varied data environment, *Blockchain platforms* constitute our second focus area in Blockchain Data Analytics.

Specifically, we analyze Ethereum ERC20 token networks with graph node motifs, directed flow motifs, core investor analysis and topological data analysis approaches.

[1] Measured by the decrease in root mean square error.

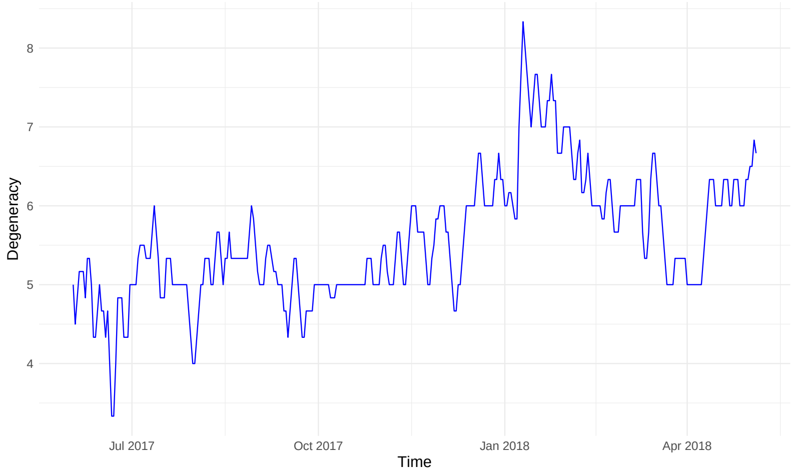

Figure 2: Average graph degree degeneracy of Ethereum token networks in time. The buildup towards the 2018 January crash in blockchain markets, and the decline of activity in the aftermath are clearly visible.

We use the extracted information in machine learning models for token classification, price prediction and token health analysis. For example, Figure 2 shows that starting from April 2018, the upward trend of Ethereum token activity has started again after the January 2018 crash.

Our methods have extended existing statistical and network analysis tools in creative ways; when faced with the problem of identifying most important nodes in a token's network, we have proposed a new Data Depth [12] based algorithm [4] that replaces the standard k-core [9] algorithm. An important aspect of the new algorithm is its ability to incorporate the amount of transferred tokens in a data-driven way. This allows us to have a resolution of the sparse, directed and weighted token networks with a level of detail that is not possible to capture by existing algorithms.

Figure 3 shows an immediate application of our extracted data; we can cluster Ethereum tokens in terms of their investor selling/buying behavior.

Furthermore, our novel methods give us an advantage in gaining critical insights from the network as a *early-warning* mechanism, that is, before the price and network activity information reflect the ongoing changes. Indeed, although Figure 2 shows that the 2018 January crash have been overcome in general, Figure 4 suggests that this is not true for the two most important tokens on Ethereum: VeChain and BinanceCoin (BNB) have not recovered their lost ground in 2018. In fact, their price trends[2] in recent months corroborate our insights.

[2] See VeChain at `https://coinmarketcap.com/it/currencies/vechain/`

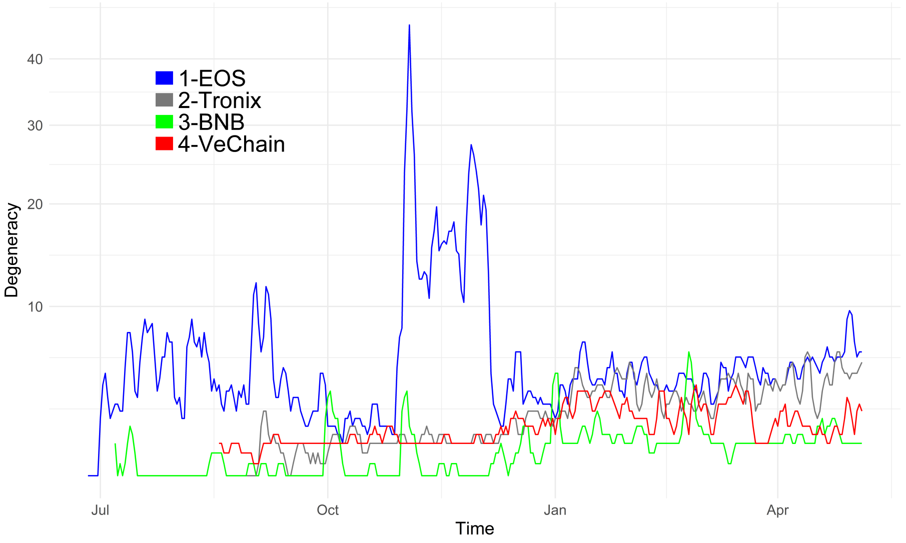

Figure 4: Core degeneracy in large Ethereum tokens. Unlike those of EOS (which moved out of Ethereum) and Tronix, the core degeneracy of BNB and VeChain have not improved since the January crash.

In this line, we have developed a token price anomaly prediction framework where token transfers among the most important nodes are modeled with Topological Data Analysis metrics (i.e., Betti numbers). Our results [11] show that token price anomalies can be predicted with high accuracy. For example, ≥ ±15% price changes of the Tronix token (most valuable token on Ethereum) can be predicted with 85% accuracy.

Market cap and volume of Blockchain tokens already reach hundreds of billions of dollars. Our preliminary results indicate that developing new Data Analytics tools and algorithms on Blockchain platforms will have a vital role in tracking and managing an expanding financial market.

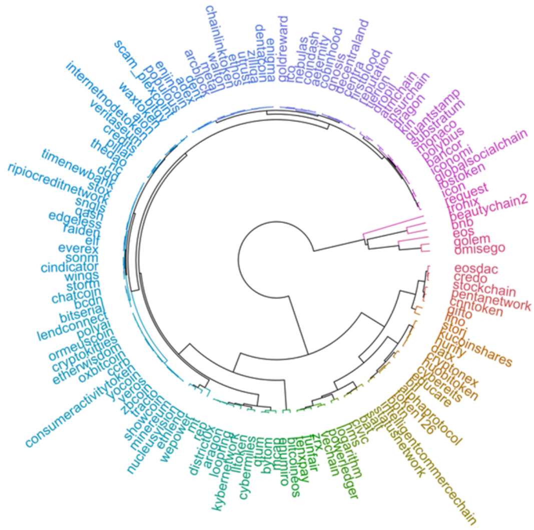

Figure 3: A grouping of Ethereum tokens based on investor buying/selling behavior.

Deliverables: Our goal as part of this project is to build a software framework that provides advanced blockchain data analytics capabilities by leveraging the local topological structure and patterns extracted from blockchain data. The developed tool will provide deeper insights into how cryptocurrencies, ICOs and cryptotokens co-evolve, and how local graph structures of cryptocurrency transactions impact price and market adoption.

Researchers

Cuneyt Gurcan Akcora is a Postdoctoral 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 VLDB, ICDM and ICDE.

Yulia R. Gel is a Professor in the Department of Mathematical Sciences at the University of Texas at Dallas. Her research interests include statistical foundations 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 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 any 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, privacy and privacy-preserving data mining. Some of 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] Nazmiye Ceren Abay, Cuneyt G. Akcora, Yulia R. Gel, Umar D. Islambekov, Murat Kantarcioglu, and Bhavani Thuraisingham. Chainnet: Learning on blockchain graphs with topological features. *Under Submission*, pages 1–9, 2018.

[2] Cuneyt Gurcan Akcora, Asim K. Dey, Yulia R. Gel, and Murat Kantarcioglu. Forecasting bitcoin price with graph chainlets. In *The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Melbourne, Australia*, pages 1–12, 2018.

[3] Cuneyt Gurcan Akcora, Matthew Dixon, Yulia R. Gel, and Murat Kantarcioglu. Bitcoin risk modeling with blockchain graphs. *Economics Letters*, pages 1–5, 2018.

[4] Cuneyt Gurcan Akcora, Yulia R. Gel, and Murat Kantarcioglu. Chartalist: The age of ethereum token economies. *Under Submission*, 2018.

[5] Asim Kumar Dey, Cuneyt Gurcan Akcora, Yulia R. Gel, and Murat Kantarcioglu. On the role of local blockchain network features in cryptocurrency price formation. *Under Submission*, pages 1–35, 2018.

[6] Giuseppe Di Battista, Valentino Di Donato, M. Patrignani, Maurizzio Pizzonia, V. Roselli, and R. Tamassia. Bitconeview: visualization of flows in the bitcoin transaction graph. In *IEEE VizSec*, pages 1–8, 2015.

[7] Giuseppe Di Battista, Valentino Di Donato, and Maurizio Pizzonia. Long transaction chains and the bitcoin heartbeat. In *European Conference on Parallel Processing*, pages 507–516. Springer, 2017.

[8] Harry Kalodner, Steven Goldfeder, Alishah Chator, Malte Möser, and Arvind Narayanan. Blocksci: Design and applications of a blockchain analysis platform. *arXiv preprint arXiv:1709.02489*, 2017.

[9] Ricky Laishram, Ahmet Erdem Sariyüce, Tina Eliassi-Rad, Ali Pinar, and Sucheta Soundarajan. Measuring and improving the core resilience of networks. In *Proceedings of the 2018 World Wide Web Conference*, pages 609–618, 2018.

[10] Yang Li, Kai Zheng, Ying Yan, Qi Liu, and Xiaofang Zhou. Etherql: A query layer for blockchain system. In *Database Systems for Advanced Applications*, pages 556–567. Springer International Publishing, 2017.

[11] Yitao Li, Cuneyt Gurcan Akcora, Yulia R. Gel, Murat Kantarcioglu, and Ekaterina Smirnova. Functional price prediction in ethereum tokens. *Under Submission*, pages 1–12, 2018.

[12] Regina Y Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, pages 405–414, 1990.

[13] S. Meiklejohn, M. Pomarole, G. Jordan, D. Levchenko, K.and McCoy, G. M Voelker, and S. Savage. A fistful of bitcoins: characterizing payments among men with no names. In *Proceedings of the 2013 conference on Internet measurement conference*, pages 127–140. ACM, 2013.

[14] Rebecca S Portnoff, Danny Yuxing Huang, Periwinkle Doerfler, Sadia Afroz, and Damon McCoy. Backpage and bitcoin: Uncovering human traffickers. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1595–1604. ACM, 2017.