Ethereum Data Analysis Literature Review

Sanjif Rajaratnam

I. INTRODUCTION

The purpose of this research project is to analyze the data from the Ethereum blockchain to find interesting connections. Ethereum, like Bitcoin, has a blockchain-based distributing platform at its core. However, Ethereum builds on the Bitcoin technology by including "smart" contracts. Ethereum uses a cryptocurrency known as "ether" that can be used to pay for contracts that are ran on the network based on their computation cost ("gas"), or as virtual currency. Miners can run the computations and verify transactions to create and add new blocks to the network in exchange for gas.

II. PAPERS SEARCH METHODOLOGY

The first set of papers were found by searching the IEEE Xplore Database, AMINER, and the ACM Digital Library for a set of keywords related to Ethereum and Bitcoin data analysis. Ethereum is a fairly new (established in the summer of 2015) so it had very few papers associated with it. There were much more papers related to Bitcoins. The search was then broadened to include non-cryptocurrency related search terms centered around financial transaction analysis, time-series analysis, graph search, and ledger analysis to get more papers.

Then the related works and references of the provided papers and papers found from the databases above were parsed to find more relevant papers.

III. RELATIONSHIP TO PROJECT

The papers were then analyzed to see if they could be used with the provided dataset. Non-compatible papers were still acceptable if they provided a means of getting further data. Next it was investigated to see if the analysis could potentially be applied to the provided dataset.

From research, it was found that papers generally feel into certain themes like de-anonymizing users, classifying clusters, finding anomalies, or analysis.

IV. BIBLIOGRAPHY

A. Mining and Classifying Dataset

Yen and Chen [1] presents a strategy for mining association rules to discover large sets that appear together in a sufficient number of transactions. AlZoubi et al. [2] present a graph-based technique to generate Boolean association rules for a large dataset in an efficient manner. Both these techniques could be used to mine and sort the large Ethereum dataset prior to applying analysis.

B. PCA Analysis

Kondor et al. [3] set out to understand the time series of prices of goods and assets by modeling the underlying system of interacting agents. They do this by analyzing Bitcoin's complete list of transactions. They use Principal Component Analysis at different time instances of the block chain to show that structural changes in the network is usually accompanied by changes in the Bitcoin exchange rate. They show how they get and clean their data and how they detected structural changes. Then they showed the result and their analysis. This same analysis can easily be transferred to Ethereum's ether

as it is also a currency with a public ledger of transactions over time.

C. Energy Analysis

O'Dwyer and Malone [4] set out to determine if it is worth mining Bitcoins by analyzing the energy consumption associated with Bitcoin mining. They start by describing the bitcoin mining process, and the different levels of difficulties and rewards. They compare the energy usage to the exchange rate and find the energy consumption of the Bitcoin mining network. This analysis can be transferred to Ethereum quite easily because of Ethereum's concept of gas which represents computational cost for running scripts. This could be extended to find the gas cost that provided the highest return on computational investment.

D. Graph Analysis

Fleder et al [5] aimed to annotate the Bitcoin public ledger by linking people's public keys to the graph, either definitively or statistically. Then they run the annotated graph through their own graph analysis framework. This paper clearly defines how they how they got their data, including how they parsed their blockchain, prior to doing graph analysis. They got their information by web-scraping forums and the Bitcoin ledgers. They also investigated the specific case of single intermediary nodes related to the original Silk Road nodes. This analysis can easily be extended to the Ethereum database because it is of the same structure as bitcoins. The accounts related to The DAO attack can also be potentially investigated.

E. Statistical Analysis

Chung and Svetinovic [6] aimed to analyze the Namecoin network in 7 six month intervals. Namecoin is an altcoin based on Bitcoin. The paper aims to find out if the Namecoin network follows the Densification Law over time, and if there is a difference between the network pattern of Namecoin and Bitcoin over time. They grabbed a Namecoin data dump and analyzed it via R. Then they compared Namecoin data to Bitcoin data. This analysis can be extended to Ethereum and then the results can be compared against both Namecoin and Bitcoin.

F. Machine Learning

Monamo et al. [7] aim to detect anomalies within the Bitcoin network since anomaly detection is equivalent to fraud detection. This paper uses a trimmed K-means based unsupervised learning that is capable of simultaneously clustering and doing fraud detection in multivariate steps. Their analysis assumed users as nodes, and transactions as edges. They also aimed to do some feature extraction by building a feature set that contained 14 features that were fell under one of the following categories: Currency, Network, or Average Neighborhood of the user. They replaced missing values with 0, and uses a trimmed K-means clustering approach. Then finally they analyzed the results. They were however unable to evaluate algorithmic performance since they lacked labelled data. This same analysis can be done on the Ethereum dataset and TheDAO attack data can be used as labeled data for validation.

V. References

[1] Alzoubi, Wael Ahmad, et al. "An Efficient Mining of Transactional Data Using Graph-Based Technique." 2011 3rd Conference on Data Mining and Optimization (DMO), 2011, doi:10.1109/dmo.2011.5976508.

[2] Yen, Show-Jane, and A.l.p. Chen. "A Graph-Based Approach for Discovering Various Types of Association Rules." IEEE Transactions on Knowledge and Data Engineering, vol. 13, no. 5, 2001, pp. 839–845., doi:10.1109/69.956106.

[3] Kondor, Daniel, et al. "Inferring the Interplay between Network Structure and Market Effects in Bitcoin." New Journal of Physics, vol. 16, no. 12, Feb. 2014, p. 125003., doi:10.1088/1367-2630/16/12/125003.

[4] O'dwyer, K.j., and D. Malone. "Bitcoin Mining and Its Energy Footprint." 25th IET Irish Signals & Systems Conference 2014 and 2014 China-Ireland International Conference on Information and Communities Technologies (ISSC 2014/CIICT 2014), 2014, doi:10.1049/cp.2014.0699.

[5] Fleder, Michael, et al. “Bitcoin Transaction Graph Analysis.” Computing Research Repository, 2015. Abs/1502.01657, arxiv.org/abs/1502.01657.

[6] Chang, Tao-Hung, and Davor Svetinovic. “Data Analysis of Digital Currency Networks: Namecoin Case Study.” 2016 21st International Conference on Engineering of Complex Computer Systems (ICECCS), 2016, doi:10.1109/iceccs.2016.023. Data Analysis of Digital Currency Networks: Namecoin Case Study

[7] Monamo, Patrick, et al. “Unsupervised Learning for Robust Bitcoin Fraud Detection.”*2016 Information Security for South Africa (ISSA)*, 2016, doi:10.1109/issa.2016.7802939.