The Hubs and Authorities Transaction Network Analysis using the SANSA framework

Danning Sui[1], Gezim Sejdiu[2], Damien Graux[3], and Jens Lehmann[2,3]

[1] Alethio
danning.sui@consensys.net
[2] Smart Data Analytics, University of Bonn, Germany
sejdiu@cs.uni-bonn.de, jens.lehmann@cs.uni-bonn.de
[3] Fraunhofer Institute for Intelligent Analysis and Information Systems, Germany
damien.graux@iais.fraunhofer.de, jens.lehmann@iais.fraunhofer.de

Abstract. With the recent trend on blockchain, many users want to know more about the important players of the chain. In this study, we investigate and analyze the Ethereum blockchain network in order to identify the major entities across the transaction network. By leveraging the rich data available through Alethio's platform in the form of RDF triples we learn about the Hubs and Authorities of the Ethereum transaction network. Alethio uses SANSA for efficient reading and processing of such large-scale RDF data (transactions on Ethereum blockchain) in order to perform analytics e.g. finding top accounts, or typical behavior patterns of exchanges' deposit wallets and more.

1 Introduction

With the hype on blockchain technologies and in particular in the Ethereum blockchain [4], many participants wanted to know more about the most impactful players across the blockchains transaction network. In parallel, as the number of statements, actions and transactions in the network are increasing quickly, many "Big Data" challenges arise. First, transactions are raw data and one cannot take advantage of them for further analysis. To do so, Alethio designed EthOn (The Ethereum Ontology) [3] which models such raw data as triples using the RDF (Resource Description Framework)[1] standard. This ontology describes all Ethereum terms including blocks, transactions, contract messages, event logs etc., as well as their relationships. Afterword, performing querying and analysis on such large-scale RDF datasets is computing intensive. To overcome these challenges, we have explored the potential of the SANSA [2] framework. SANSA is an open-source[2] framework for distributed processing and analysis of large-scale RDF data. With SANSA on Spark, RDF triples are loaded into Spark distributed and resilient data structured, namely the data frames, for further analysis.

[1] https://www.w3.org/TR/rdf-primer/
[2] https://github.com/SANSA-Stack

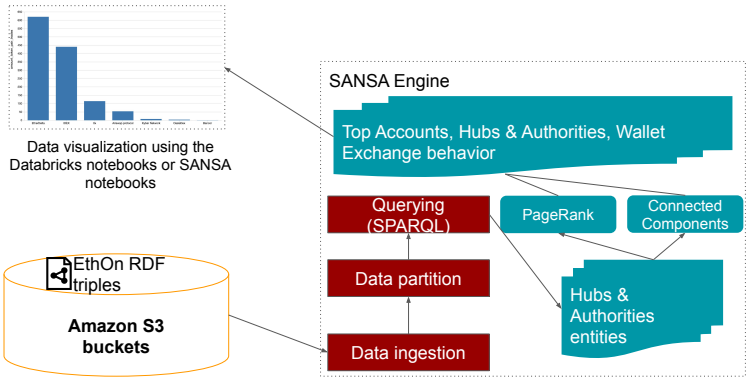

Fig. 1. Hubs and Authorities analysis workflow.

In this paper, we perform an analysis (using well-known graph processing algorithms) of the value transaction network graph with the main focus on the Hubs and Authorities behaviors. "Authorities" are accounts who pay out to a large crowd of addresses, with high volume; while "Hubs" are entities who receive extensive Ether (ETH) flow into their accounts. In this study, we do not differentiate these two roles but rank them all together as the biggest players/entities.

2 Finding big Ethereum players with SANSA

The Ethereum network graph contains nodes of external accounts which have had a transaction on the Ethereum blockchain. The connection (edges) between such nodes on the network indicate the transaction relationship between them; when a node (an external account) sends ETH to another, a transaction record is written, and an edge between them is added in the network with the direction of the ETH flow. When we encounter multiple edges between same pairs of nodes, we summarize the edges as a single one[3]. The edge weight is the total transaction value in Ether. As an example, if address A sends x ETH to address B in total, there will be an edge of weight x from node A to node B. In this study, self-loops i.e. transactions from an address to itself are omitted.

SANSA framework has been used for efficient reading and querying of RDF datasets using SPARQL as depicted on Figure 1. First, the data need to be loaded on an efficient storage that SANSA can read from. For that purpose, we

[3] This optimization is also convenient practically as it is easier not to have duplicated edges in a graph.

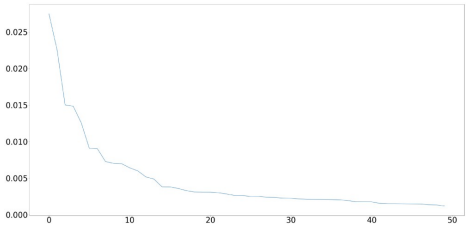

Fig. 2. PageRank Score Distribution of Top-50 Accounts.

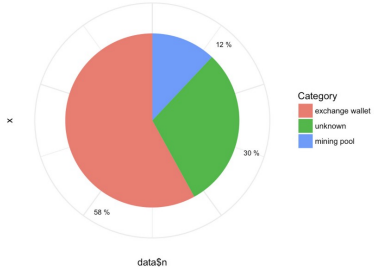

Fig. 3. Category Distribution of Top-50 Accounts.

use Amazon S3 buckets containing the whole RDF Ethereum network transactions. Afterword, SANSA data representation layer loads the data in a form of Resilient Distributed Datasets (RDD) [5] of triples. During this process, SANSA performs a data partition for fast processing and then aggregate and filter the data using the its query layer [1]. Further, we applied two classic graph analysis algorithms via Apache GraphX: Connected Components and Page Rank. Connected Components algorithm enables us to find the largest cluster of connected nodes, regardless of transaction direction. Within this largest cluster, we can derive the page rank score of all nodes. Top-ranked entities and their relation are visualized.

3 Results

3.1 Datasets

The Ethereum dataset in the format of RDF contains more than 17B triples. For the sake of the experiment, we limited the dataset to 10,000 blocks which contain around 38M triples, including both value transactions and contract messages.

3.2 Top Accounts Analysis

The PageRank algorithm was run over the largest connected component of 185,741 nodes (accounts) and 250,637 edges (aggregated transaction relations). Figure 2 plots the top 50 account's distribution. Based on the findings, we can see that these accounts are grouped on two different types: mining pool wallets, and (mostly centralized) exchange wallets.

Figure 3 shows that 58% of the addresses are controlled by exchanges, while another 12% with convincing tags related to the mining pools. The exchange and mining pool wallets can be found in the top position of our ranking, underlining the effectiveness of PageRank: Addresses related to mining pools allocate extensive amounts of payouts to their subscribed miners, resulting in large out-degrees, as well as high accumulated transaction value. We can see that the main

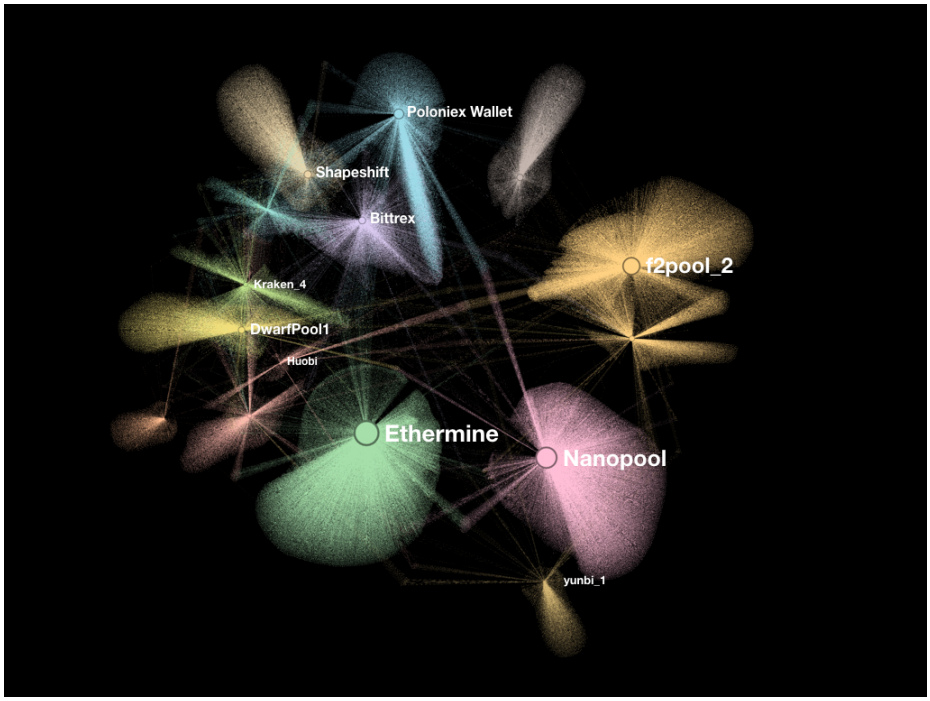

Fig. 4. Transaction Network of Top Hubs and Authorities.

wallets are centralized exchanges which distribute (and receive) large volumes of the transaction to (and from) their deposit wallets, token contracts, etc.

Our PageRank implementation successfully detects the most influential accounts across the network, corresponding to the Hubs and Authorities, connecting various transactors and carrying heavy flow weights.

Focusing on those known accounts (with labels from Etherscan[4]), we present (see Figure 4) the network overview of top hubs and authorities with transactions as edges surrounding them.

3.3 Typical Behavior Patterns of Exchanges' Deposit Wallets

We investigated the associated transaction behavior of the exchange wallets. Based on our finding, these behaviors can be grouped into three categories:

1. *Frequently paying out to certain exchanges' main wallets with a fixed, large value* – From the scatter plot, the payout amount is always around a same value.
2. *Frequently receiving funds from the same exchange main wallets, and paying out to various token contracts* – This is due to the activity which is associ-

[4] `https://etherscan.io/`

ated with exchanges as they use external accounts as deposit addresses for collecting tokens based on trading needs.
3. *Frequently receiving funds from a group of "miner" accounts, with "proxy" accounts in between, which clean out their received ETH within a short time frame* – Usually, these addresses receive funds from miner accounts, which again get paid reasonable amounts by known mining pools, which we assume are mining rewards (usually around 0.11-0.12 ETH).

Despite pointing out the three typical behaviors above, they are not necessarily mutually exclusive. There are addresses which share more than one of the deducted patterns. These behavior patterns explored here are based on the labels we have gathered, and this may be different for other use cases.

4 Conclusion

SANSA provides a scalable solution for reading and querying large scale RDF data, providing compatibility with machine learning libraries on Spark including GraphX as a graph processing library. With conventional graph analysis tools, we successfully identified Hubs and Authorities in the Ethereum transaction network and discovered that they are mainly related to exchange wallet and mining pool activities.

This pipeline also provides a possibility to filter out top accounts, which are likely to be exchanges' deposit wallets. Furthermore, with the filtered top rank accounts, the "mixing" patterns of exchanges' deposit wallets become recognizable. This can be a promising tool for detecting previously unknown exchange wallets and lead to a deeper understanding of their behavior patterns for future analyses.

References

1. Ermilov, I., Lehmann, J., Sejdiu, G., Bühmann, L., Westphal, P., Stadler, C., Bin, S., Chakraborty, N., Petzka, H., Saleem, M., Ngonga, A.C.N., Jabeen, H.: The Tale of Sansa Spark. In: 16th International Semantic Web Conference, Poster & Demos (2017)
2. Lehmann, J., Sejdiu, G., Bühmann, L., Westphal, P., Stadler, C., Ermilov, I., Bin, S., Chakraborty, N., Saleem, M., Ngonga Ngomo, A.C., Jabeen, H.: Distributed semantic analytics using the SANSA stack. In: ISWC Resources Track (2017)
3. Pfeffer, J., Beregszazi, A., Detrio, C., Junge, H., Chow, J., Oancea, M., Pietrzak, M., Khatchadourian, S., Bertolo, S.: Ethon - An Ethereum ontology (2016)
4. Wood, G.: Ethereum: A secure decentralised generalised transaction ledger. Ethereum project yellow paper **151**, 1–32 (2014)
5. Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M.J., Shenker, S., Stoica, I.: Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. pp. 2–2. USENIX Association (2012)