**BANGALORE INSTITUTE OF TECHNOLOGY**

K.R. Road, V.V.Pura, Bengaluru-560004



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING Project Work Synopsis**

# VII-Sem 2019-2020

**PROJECT GROUP:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sl.**  **No.** | **USN** | **Name** | **Section** | **Phone Number** | **Signature** |
| 1 | 1BI16CS160 | SUSHMITA M KATTI | C | 8553759254 |  |
| 2 | 1BI16CS161 | SUPRAJA PRASAD | C | 9731006973 |  |
| 3 | 1BI16CS204 | SHILPA M | C | 8277207316 |  |
| 4 | 1BI16CS174 | VAISHNAVI KULKARNI | C | 9379891444 |  |

**PROJECT DETAILS:**

|  |  |
| --- | --- |
| **Title:** | **Characterizing the Ethereum address space and Inferring user traits via unsupervised methods** |
| **Domain:** | **Machine Learning** |
| **Location:** | **Bengaluru** |

**For office use only:**

|  |  |
| --- | --- |
| **Group ID:** |  |
| **Guide:** |  |
| **Status:** | **Accepted/To be modified/Rejected** |

**Signature of the Project Co-Ordinator**

# TITLE

Characterizing the Ethereum address space and Inferring user traits via unsupervised methods

# LITERATURE SURVEY

Several attempts have been made to identify addresses based on transactions from the Bitcoin blockchain e.g. Meiklejohn et al. (2013), Neudecker & Hartenstein (2017), Poikonen (2014); however, to our knowledge, this is the ﬁrst such project focused on the Ethereum address space. Similar to our approach, other projects utilize several clustering methods, but k-means seems to be the primary algorithm employed due to its versatility and scalability with large data sets.

The main quantitative obstacle we had to navigate was making an educated estimate of the optimal number of clusters to use for learning, as this would inform our qualitative analysis by partitioning the address space into a discrete set of behavior groups. The issue of determining the optimal number of clusters, however, is not always a well-deﬁned problem. Kodinariya & Makwana (2013) review six diﬀerent evaluation techniques that range from quantitative to heuristic, including silhouette scoring–a measure of inter- and intra-cluster variance–and the so-called “elbow method”, which attempts to estimate where the returns of adding additional clusters begin to diminish. Tibshirani et al. (2001) attempts to formalize the elbow method in a quantitative framework via the gap statistic. Our analysis has drawn from these works as they have guided our choice of clustering algorithms and evaluation metrics. Perhaps the most important upshot in examining related works was determining that the clustering and qualitative analysis of the address space is a problem open to experimentation and interpretation.

# PROBLEM DEFINITION

Since the advent of Bitcoin, awareness and excitement around cryptocurrencies and the underlying blockchain technology that enables them have increased exponentially. Fundamentally, cryptocurrencies provide anonymity in that users operate via an address or set of addresses devoid of any personal information. However, also fundamental to the technology is the fact that blockchain data is completely publicly available and could therefore theoretically be used to successfully characterize or even identify users, resulting in considerable security implications and other potential consequences and benefits. As such, we sought to gather a comprehensive dataset of Ethereum addresses and their associated metadata upon which we could apply cluster analysis to then divvy said addresses into behaviour groups sharing similar attributes.

# PROPOSED SYSTEM

The data is extracted from etherscan.io by recursively scraping the data from the publicly available blockchain using the etherscan API. Once data acquisition is complete, we use the information to build our data and feature set. Then qualitative analysis is performed by applying clustering algorithms and Principal Component Analysis(PCA) to determine an informed estimate of the optimal number of clusters with which we then examine and analyse the behaviour groups in the address space.

# SYSTEM REQUIREMENTS

* Python 3.x
* Linux operating system
* Editors (VSCode, Emacs, Sublime)
* Version control system (git)
* Remote repository (Github)
* Database software (MySQL)
* CPU Intel 2GHz
* 4GB RAM

**SYSTEM ARCHITECTURE**

Results

and Discussion

Models

and Analysis

Data preprocesing

Data acquisition

# APPLICATIONS

* Understanding network activity
* Enhancing trading strategies
* Improving Anti Money Laundring activities

# REFERENCES

Characterising The Ethereum Address Space James Payette,1 Samuel Schwager,2 and Joseph Murphy

Bholowalia, P., & Kumar, A. 2014, International Journal of Computer Applications, 105

Kodinariya, T. M., & Makwana, P. R. 2013, International Journal, 1, 90

Meiklejohn, S., Pomarole, M., Jordan, G., et al. 2013, in Proceedings of the 2013 conference on Internet measurement conference, ACM, 127

Monaco, J. V. 2015, in SPIE Defense+ Security, International Society for Optics and Photonics, 945704

Neudecker, T., & Hartenstein, H. 2017, in International Conference on Financial Cryptography and Data Security, Springer, 155

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Poikonen, S. 2014   
Tibshirani, R., Walther, G., & Hastie, T. 2001, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63, 411