Characterizing the Ethereum address space

Inferring user traits via unsupervised methods

James Payette¹, Samuel Schwager², Joseph Murphy³

¹Department of Computer Science, jpayette@stanford.edu
²Department of MCS, sams95@stanford.edu
³Department of Physics, murphyjm@stanford.edu

The Ethereum address space

Recent public and commercial interest in cryptocurrencies has made their published, yet
anonymous ledgers, or “blockchains”, objects of
supreme interest. Successfully identifying or even
classifying users, known only by their addresses,
would have enormous security implications [1]. We
examine the blockchain of Ethereum with the
objective of clustering addresses into distinct
“behavior groups” to qualitatively infer their traits.

An example transaction on the Ethereum blockchain [2]

Data Acquisition
Successful, efficient data acquisition was a major
milestone for our project

• Using etherscan.io, we recursively scraped data from
  the publically available blockchain, eventually
  aggregating a data set of 250,000 unique addresses.
• Queried the etherscan API for an address’ Ethereum
  balance and all of their transactions.
• Used this information to build our feature vectors

Models and Analysis
The main objective of our quantitative analysis was to use clustering evaluation metrics
and Principal Component Analysis (PCA) to determine an informed estimate for the
optimal number of clusters with which to examine as behavior groups.

• PCA finds that only 33% of the variance is explained by the first two components
• K-means clustering used over other methods for its scalability, versatility
• Use unsupervised metric Calinski Harabaz Score as measure of cluster definition
• “Elbow” of Calinski Harabaz plot gives insight on optimal number of clusters [3]
• Further investigate optimal number of clusters via Silhouette Scores

Data and Feature Set
Each row of our overall design matrix corresponds to the feature vector for a single Ethereum address and each column
to a single feature. The dataset is normalized to the sample mean and unit sample variance.

Our feature set is made up of 34 features which are calculated based on the raw data returned by API calls to etherscan.io. Feature selection for this problem was difficult since for each
address we only have access to the information contained
in each transaction. We selected features that would help to
distinguish different types of ethereum users (i.e. industrial
users probably move more money and have more transactions
than hobby users). We tried to select features that, when
aggregated, would paint a descriptive picture of the user.

Features include: Total Ether, number of transactions, transactions per month, average Ether transaction, etc.

Models and Analysis

Determine the optimal number of K-means
clusters is not always a well-defined problem
[3], [4]. Employing various evaluation
techniques, we estimate the best number
of behavior groups lies roughly between 8 to 20.

Naively, we expect a handful of clusters
to explain the unique behavior groups in
the address space. This is confirmed by
the “elbows” of the CH Score and Inertia
figures. However, considering the
Silhouette analysis, we see that the
occupations of clusters is highly
disproportionate in this regime. This may
not be entirely troublesome, as there is likely
a biased distribution of users.

Results and Discussion

Figures at right: Silhouette scores range from 0
to 1 (1 = misclassification). Scores closer to 1
indicate a confident cluster mapping (i.e. short
distance to cluster centroid, far from neighbors).
Left: Silhouette scores of clusters with size >100,
average score (dotted red line). Right: Mapping
of clusters to 2 principal component space.
Adapted from Scikit Learn starter code.

Silhouette analysis for KMeans clustering on sample data with n clusters = 15

Silhouette analysis for KMeans clustering on sample data with n clusters = 60

References
transaction behavior.” SPIE Defense+ Security: International
Society for Optics and Photonics, 2015.
generalised transaction ledger.” Ethereum Project Yellow
Paper 151 (2014).
"Review on determining number of Cluster in K-Means
Hastie. “Estimating the number of clusters in data...”

Acknowledgements

We would like to thank the entire CS 229 teaching
staff for their instruction and useful insights this
quarter. We look forward to presenting our final
results in the forthcoming report.