Chartalist: Labeled Graph Datasets for UTXO and Account-based Blockchains

Kiarash Shamsi, Friedhelm Victor, Yulia R. Gel, Murat Kantarcioglu, Cuneyt Akcora

NeurIPS | 2022 | Datasets and Benchmarks Track
Thirty-sixth Conference on Neural Information Processing Systems

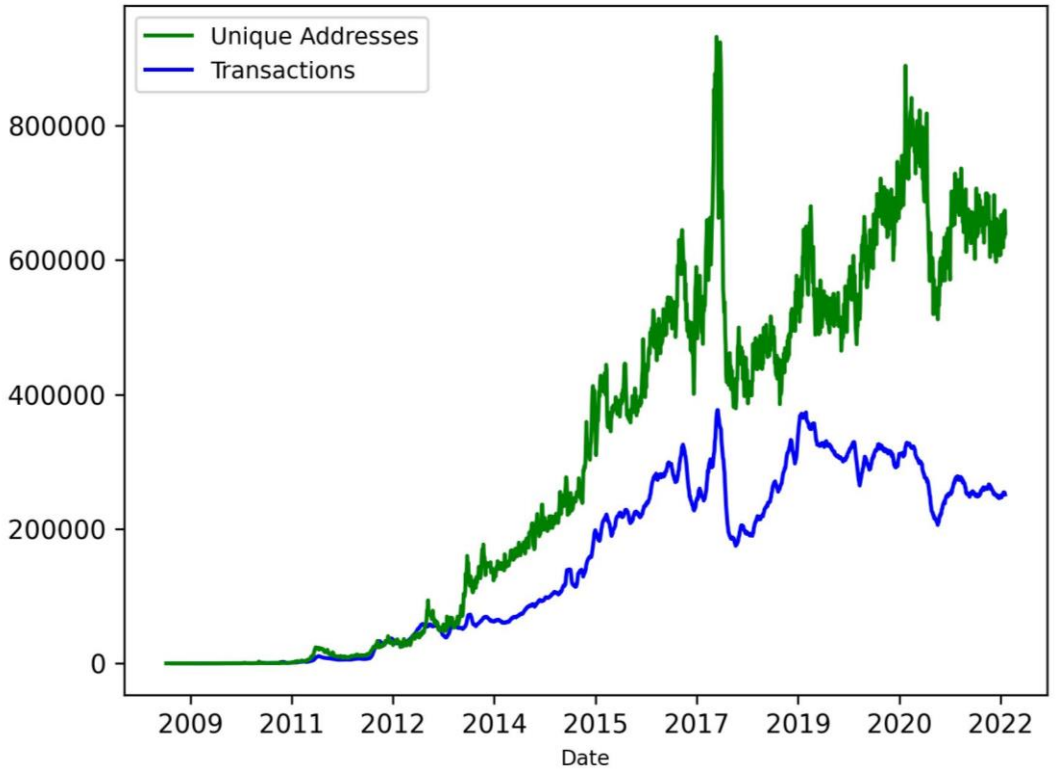
Machine learning on blockchain graphs is an emerging field with many applications:

- Ransomware payment tracking
- Crypto-asset price manipulation analysis
- Money laundering detection
- Address clustering and linkage

However, analyzing blockchain data requires domain expertise and computational resources, which pose a significant barrier and hinder advancement in this field.

Blockchain Data Analysis: Current State

- Significant efforts to extract the underlying graph by running a blockchain client.
- Paying for commercial APIs (e.g., etherscan.io) to download transaction data.
- Allocating considerable resources to construct blockchain graphs.
- Blockchain research also lacks labeled data for many significant problems.
- Different pipelines for UTXO and Account-based blockchain.



Number of unique addresses and transactions for bitcoin network. The number of unique addresses (i.e., vertices in the transaction graph) has increased above **600K**.

The open-source Bitcoin parser BlockSci requires 60GB of memory to build the Bitcoin transaction graph.

Chartalist

ML-ready datasets from unspent transaction output (UTXO) (e.g., Bitcoin) and account-based blockchains (e.g., Ethereum).

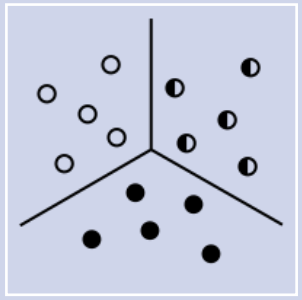
Chartalist has three main components:

- A holistic view of blockchains and formulations of graph machine-learning tasks.
- A comprehensive ecosystem of tools and community resources to support blockchain data analytics.
- A set of boards to support performance comparison and benchmark for the tasks.

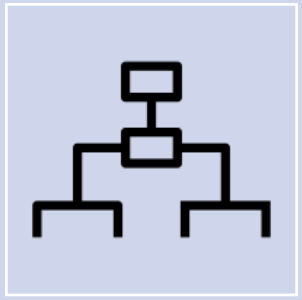
- Chartalist contains cleaned and labeled data,
- As well as open-source data loaders, and graph extractors for easy analysis.
- Large-scale, dynamic multilayer networks where nodes, edges, and edge weights evolve.
- Multilayer graphs with ground truth information on event anomalies.

Chartalist is the first attempt to systematically organize blockchain data for the broader ML community and provide a set of ML tasks defined for appropriate blockchain datasets.

Machine Learning Datasets and Learning Tasks

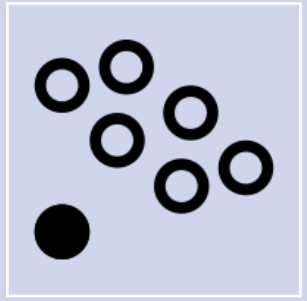


Address Clustering
Task: Identifying which addresses are co-owned by an entity.
Dataset: Chartalist Bitcoin transaction network.
Size: 737 900 blocks, 11 splits with input and output addresses, approximately 100GB of data.
Use Case: Finding co-ownership of addresses.



Address and Transaction Type Classification
Task: Determining whether addresses and transactions belong to a specific class type.
Dataset: Chartalist address and transaction type prediction dataset.
Size: 2 916 697 labeled data points, approximately 100GB data.
Use Case: Ransomware address and transaction classification.

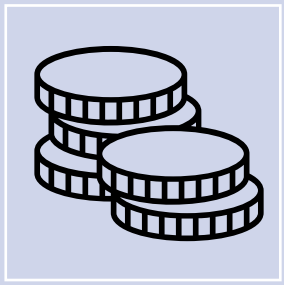
Machine Learning Datasets and Learning Tasks



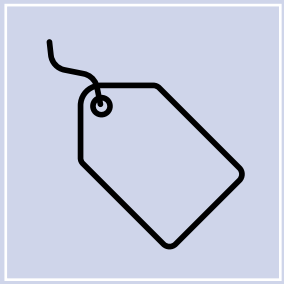
Anomalous Transaction Pattern Detection
Task: Identifying anomalous transactions so that, i.e., their distinct temporal patterns can be summarized.
Dataset: Chartalist Bitcoin transaction network.
Size: 737 900 Blocks, 11 Splits with input and output addresses, Approximately 100GB data
Use Case: Identify transaction patterns that are used in illicit cases.

Multilayer Network Analysis
Task: Predicting events and phenomena that emerge through the simultaneous use of multiple assets, which can be studied as multilayer networks.
Dataset: Chartalist Ethereum multilayer networks.
Size: 6 Token networks ~ 10MB (standard version) – 1701 token networks ~ 2GB (extended version).
Use Case: Address classification, address clustering, anomaly detection.

Machine Learning Datasets and Learning Tasks



Stablecoin Analysis
Task: Analyzing stable coin market behavior and activity.
Dataset: Chartalist stablecoin ERC20 transactions dataset
Size: 6 token networks ~ 822MB.
Use Case: Detect the time when the LunaTerra stablecoin crashed.

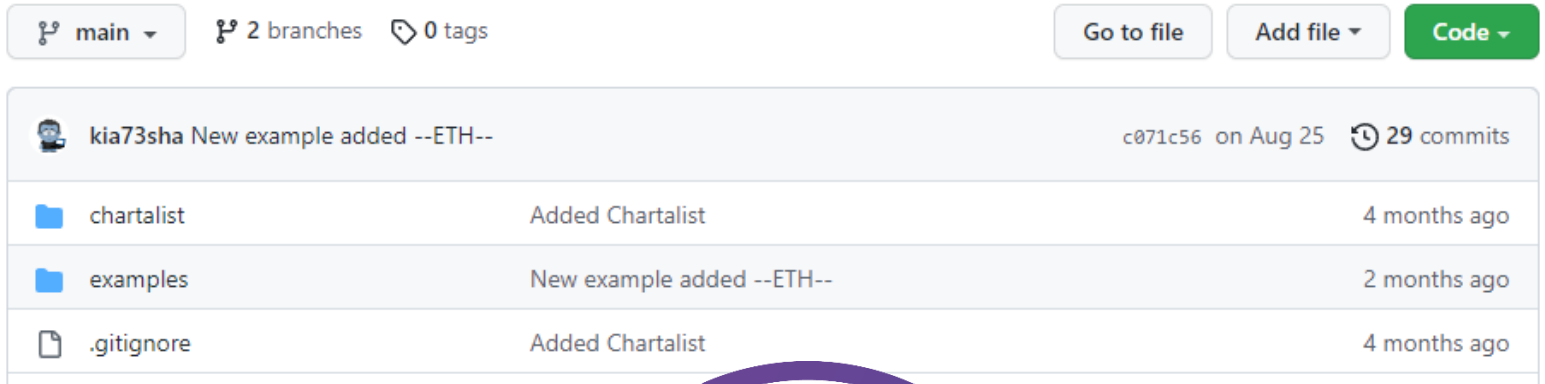


Price Analytics
Task: Using observable blockchain activity to predict asset or coin prices.
Dataset: Chartalist Bitcoin price analysis – Chartalist EtherumCurves price analysis
Size: Bitcoin (9-year price) ~ 1MB – Ethereum (31 token networks) ~ 35MB
Use Case: Bitcoin or Ethereum price prediction

Boards to Support Performance Comparison

- We introduce a baseline method for each task to benchmark new approaches.
- Since blockchain data analytics is still a nascent field, and we do not have many results from other researchers on our tasks.
- We designed a leaderboard for tasks which has been studied previously such as price prediction.
- New baseline approaches will be added to our leaderboard.

https://github.com/cakcora/Chartalist



About

Sponsored by the Canadian NSERC Discovery Grant RGPIN-2020-05665: Data Science on Blockchain and the National Science Foundation of USA under award number ECCS 2039701 Blockchain Graphs as Testbeds of Power Grid Resilience and Functionality Metrics.

Chartalist

shamsik1@myumanitoba.ca

Kiarash Shamsi, Friedhelm Victor, Yulia R. Gel, Murat Kantarcioglu, Cuneyt Akcora