Dissecting Blockchain Fraud with Optimal Transport and Graph Methods

Jared Gridley

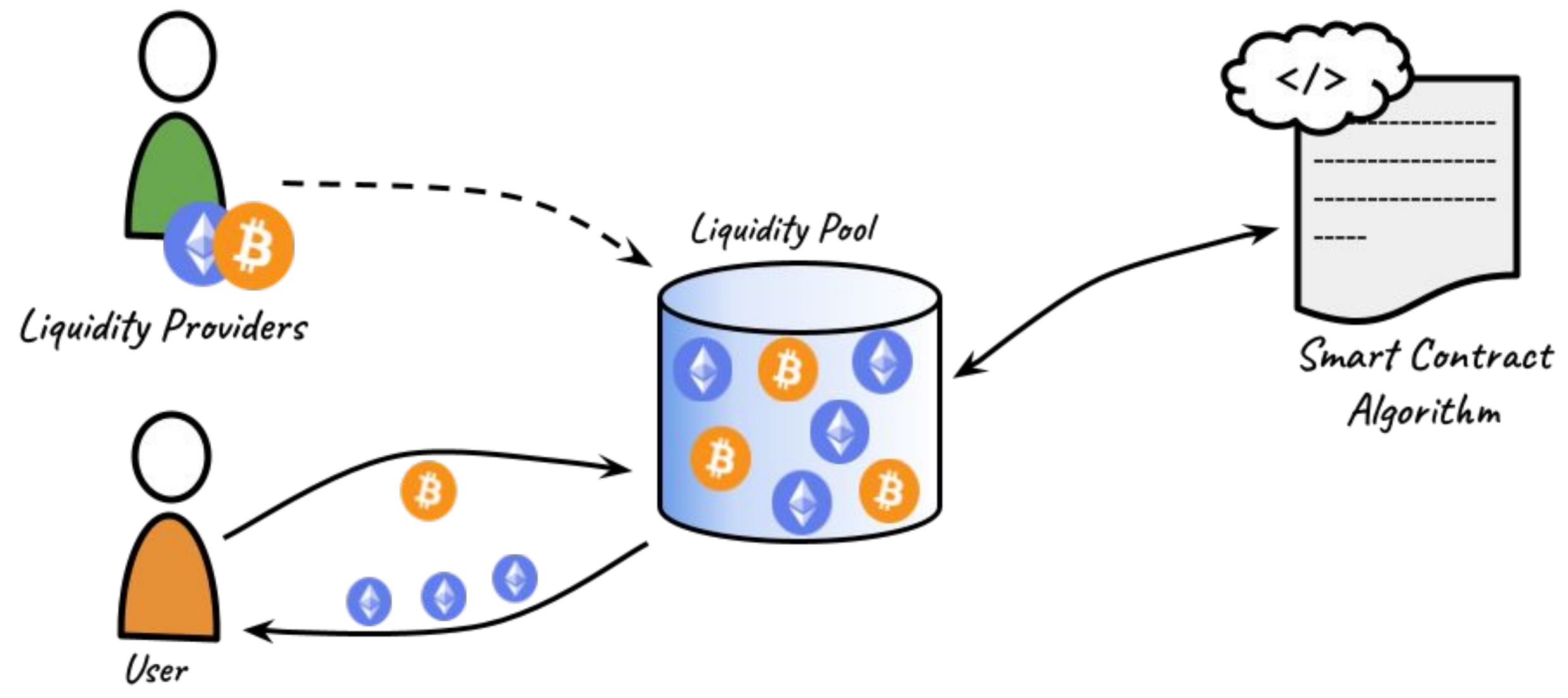
Advisors: Dr. Seneviratne, Dr. Bennett

1. Relevant Background on Decentralized Finance and Crypto Scams

Smart Contract: an immutable, autonomous piece of code that runs when predetermined conditions are met.

Decentralized Exchange:
- Protocol that allows users to convert between currencies
- Relies on smart contracts to facilitate trade execution

Liquidity Pool:
- Collection of funds that are governed by a smart contract
- Used to swap between currencies (for a fee)



Ponzi Scheme: scam that lures investors to pay off profits to earlier investors from recent investors
- “High-Yield” Staking/Lending/Mining

Phishing Attacks: scams that lure a user to give up a password or recovery phrase to then steal assets.

Code Vulnerabilities: attacks that exploit a vulnerability to steal funds

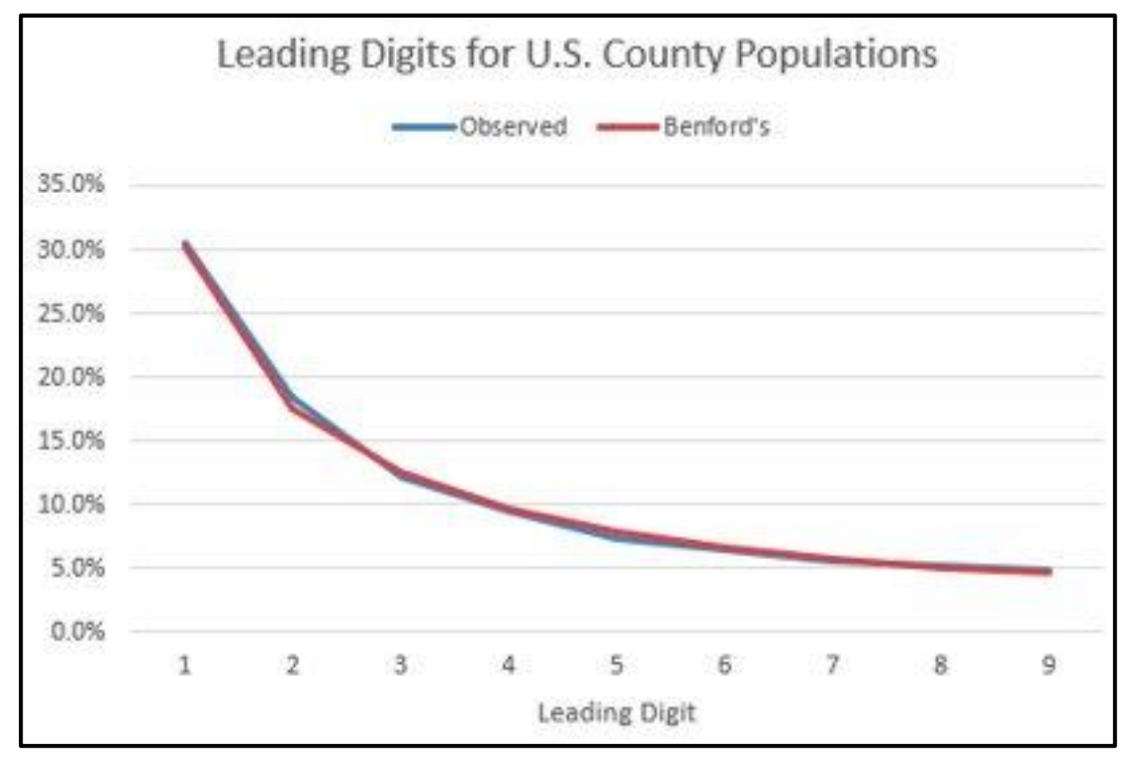
2. Identifying Scam Behavior with Newcomb-Benford’s Law

Newcomb-Benford’s Law:
Observation that in many natural datasets, the leading digit is likely to be small (the digit 1 appearing ~30% of the time)

Ex: Building heights, Rivers surface area, molecular weights, County Populations, numbers contained in an issue of *Reader’s Digest*

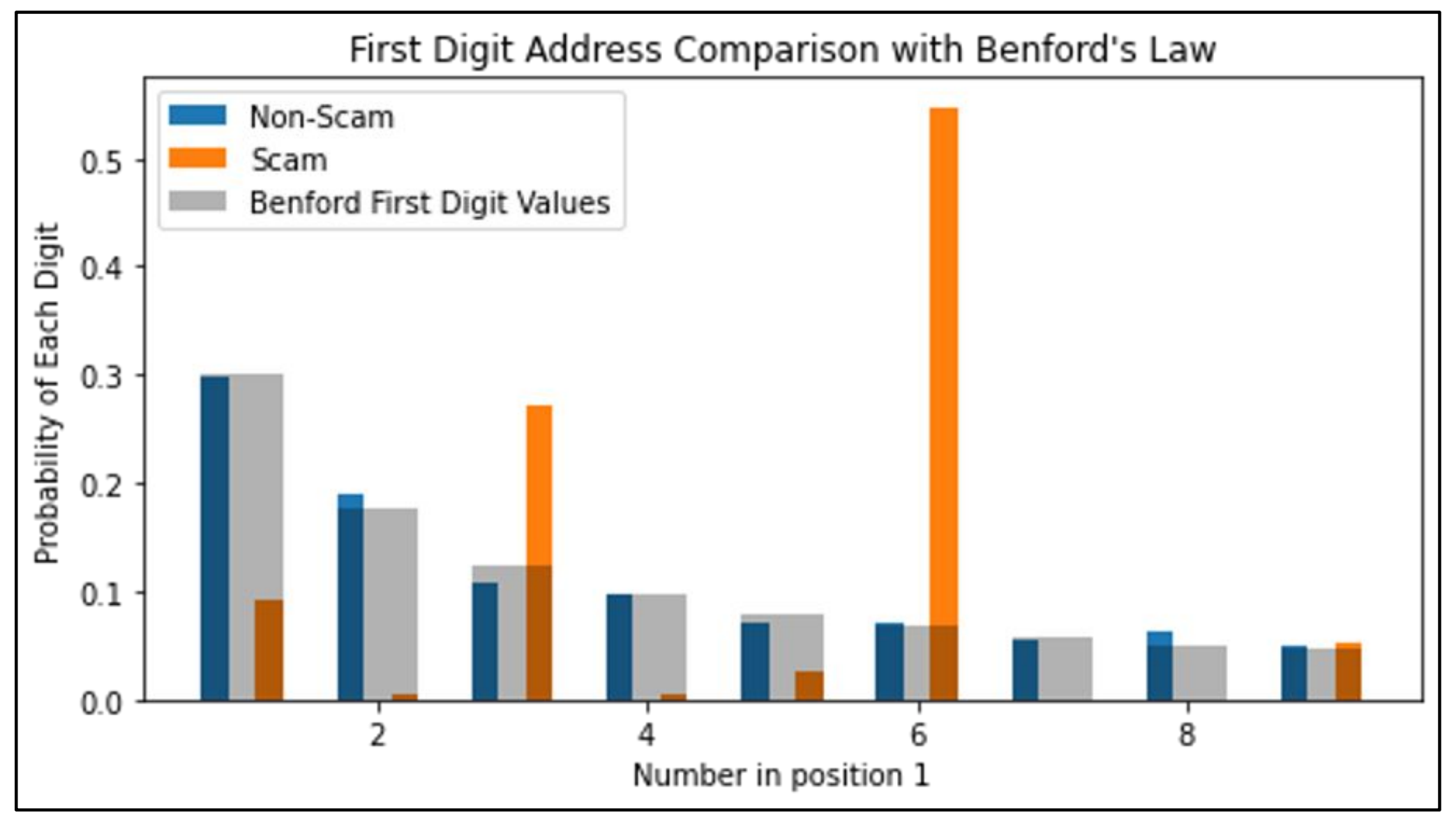
Popular Applications to Financial Fraud and Network Intrusion Detection
- Admissible in US courts
- Not a stand-alone metric, “Red-Flag test”

Cryptocurrency Transactions showed to follow Benford’s Law Compliance Theorem



Cryptocurrencies and Benford’s Law

- ➔ Scam Address (orange)
  - ◆ 1404 transactions
  - ◆ “Treasure Chest” Scam
  - ◆ Identified by *Bertoletti et al.
- ➔ Non-Scam User Address (blue)
  - ◆ 1426 transactions
  - ◆ Interacts with DeFi and NFT markets



Bartoletti et al. 2020. Dissecting Ponzi schemes on Ethereum: Identification, analysis, and impact. ACM Future Gener. Comput. Syst. 102, C (Jan 2020), 259–277.

3. ML Model Enhancement with Benford’s Law

Data Sourced from AmberData, TheGraph, and Etherscan

Dataset:
- 2.6 million transactions mostly* between January and March 2022
- Addresses belonging to users, MEV bots, lending SCs, token contracts, etc.
- Activity spanned multiple currencies on Ethereum

*Some scam address activity was pulled from before this time window.

Non-scam addresses validated through user reporting databases

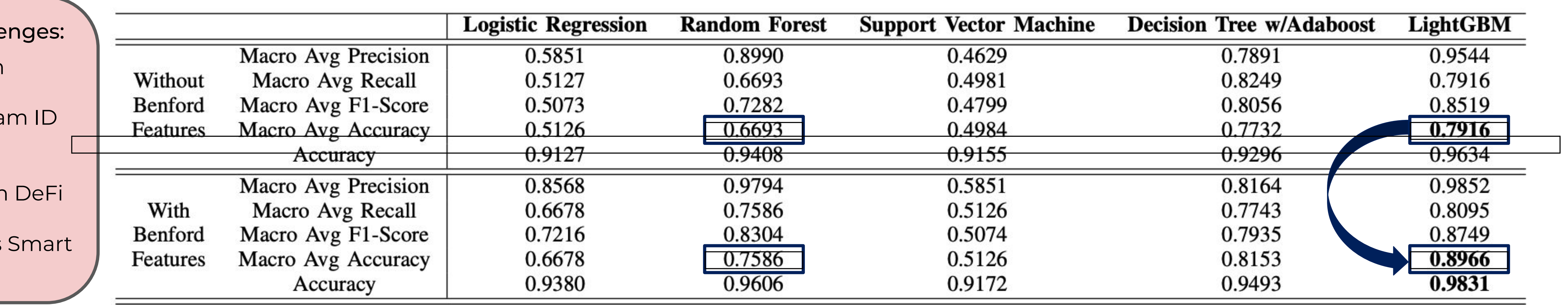
Results:
LightGBM model performed best with and without new features.
- Expected from previous research in modelling Crypto data

Models that used Benford’s law features saw improved performance in every metric that was measured.

Benford’s Law features consistently ranked as the most important for classification
- Benford’s law for second digit ranked most important in LightGBM

Data Labelling Challenges:
- ❑ Adequate Scam Labels
- ❑ Trustworthy Scam ID
- ❑ True innocence labeling
- ❑ Association with DeFi Protocols
- ❑ Bots vs Users vs Smart Contracts

Benford’s Law Feature Classifier Results



		Logistic Regression	Random Forest	Support Vector Machine	Decision Tree w/Adaboost	LightGBM
Without Benford Features	Macro Avg Precision	0.5851	0.8990	0.4629	0.7891	0.9544
	Macro Avg Recall	0.5127	0.6693	0.4981	0.8249	0.7916
	Macro Avg F1-Score	0.5073	0.7282	0.4799	0.8056	0.8519
	Macro Avg Accuracy	0.5126	0.6693	0.4984	0.7732	**0.7916**
	Accuracy	0.9127	0.9408	0.9155	0.9296	0.9634
With Benford Features	Macro Avg Precision	0.8568	0.9794	0.5851	0.8164	0.9852
	Macro Avg Recall	0.6678	0.7586	0.5126	0.7743	0.8095
	Macro Avg F1-Score	0.7216	0.8304	0.5074	0.7935	0.8749
	Macro Avg Accuracy	0.6678	0.7586	0.5126	0.8153	**0.8966**
	Accuracy	0.9380	0.9606	0.9172	0.9493	**0.9831**

This work was published at IEEE Big Data 2022 in the paper titled “Significant Digits: Using Large-Scale Blockchain Data to Predict Fraudulent Addresses”

4. Mapping Behavioral Similarities in DEXs using Optimal Transport

Fused Gromov-Wasserstein Distance
- Balances Structural feature information to map nodes between two graphs
- Computes a coupling π between vertices that minimize E_q
  - M^q_{AB} are the costs of transporting features from A to B
  - $L(C_1, C_2)^q$ are the costs of transporting pairs of nodes between each structure

The optimization problem and cost are defined as

$$FGW_{q,\alpha}(\mu,\nu) = \min_{\pi \in \Pi(h,g)} E_q(M_{AB}, C_1, C_2, \pi)$$

$$E_q(M_{AB}, C_1, C_2, \pi) = \langle (1-\alpha) M^q_{AB} + \alpha L(C_1, C_2)^q \otimes \pi, \pi \rangle$$

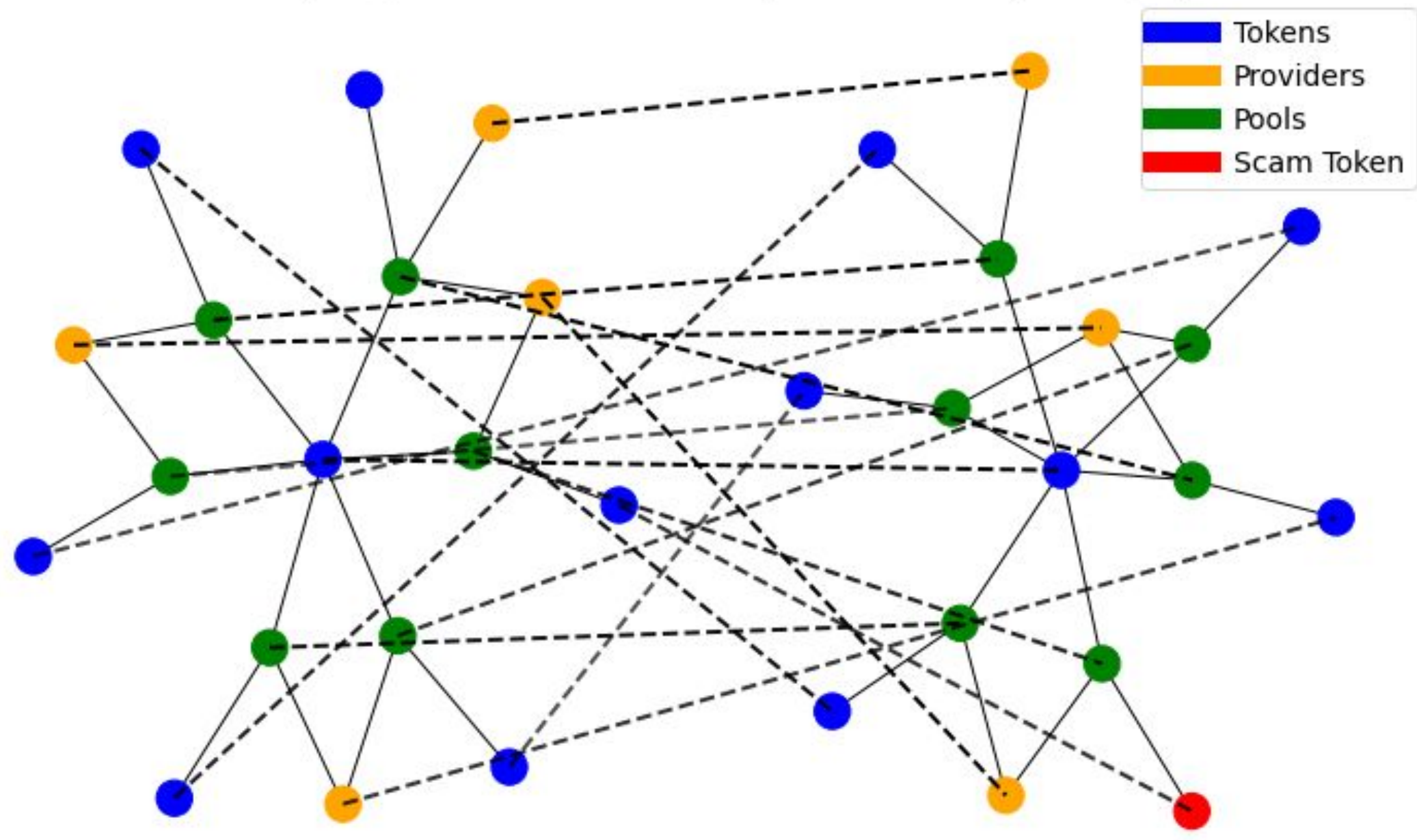
Feature Choice:
Single structure to represent all three node types, constrict node info based on type of node.

- Number of transactions	
- Number of neighbors	
- Number of liquidity pools created	(providers only)
- Avg transaction value	(providers only)
- Deviation from Max price	(tokens only)
- Benford’s Law Compliance	(providers, tokens only)

Mapping across node types:
Found some liquidity providers are actually smart contracts, and thus act similarly to liquidity pools.

Found some success in use as a “Red Flag” Test
- Mapping scam tokens and users to other scam addresses

FGW Coupling between SushiSwap and Uniswap Subgraph



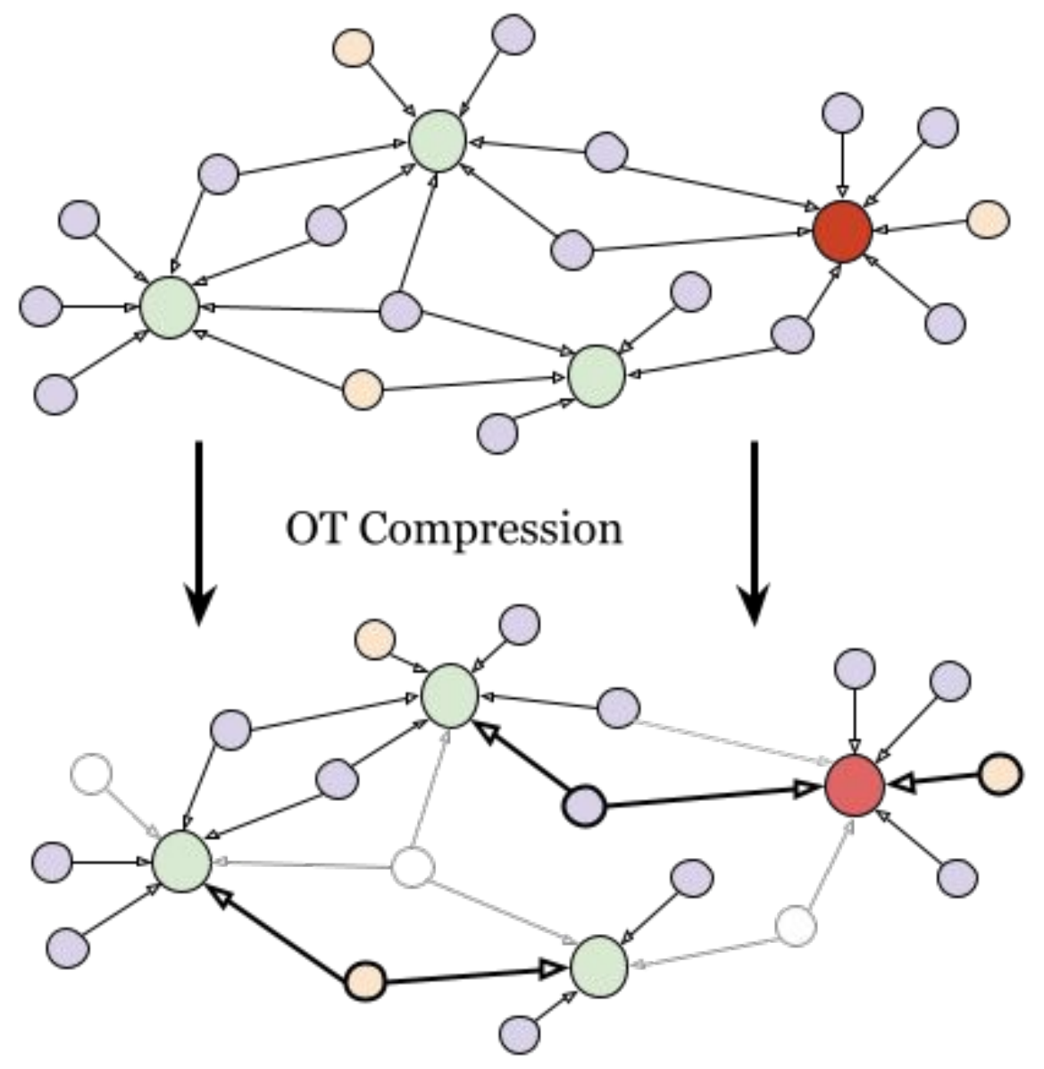
5. Optimal Transport Graph Compression for Node Importance

Time-Aware Fused Gromov-Wasserstein distance to calculate the distance between pairs of nodes within 2 graphs considering temporal and spatial information.

Use parameterized matrices to capture the temporal order of events within the cost matrix, initially proposed by Manling Li at the University of Illinois Urbana-Champaign

$$C_{ij} = \|\boldsymbol{W}_{\text{bfr}}\boldsymbol{v}_i - \boldsymbol{W}_{\text{aft}}\boldsymbol{v}_j\|_2 - \Omega(t_i, t_j)$$

(In Progress)



References:

1. Manling Li, Tengfei Ma, Mo Yu, Lingfei Wu, Tian Gao, Heng Ji, and Kathleen McKeown. 2021. Timeline Summarization based on Event Graph Compression via Time-Aware Optimal Transport. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6443–6456, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
2. Vayer Titouan, Chapel Laetitia, Flamary Rémi, Tavenard Romain and Courty Nicolas “Optimal Transport for structured data with application on graphs” International Conference on Machine Learning (ICML). 2019.
3. J. Gridley and O. Seneviratne, "Significant Digits: Using Large-Scale Blockchain Data to Predict Fraudulent Addresses," in 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022 pp. 903-910. doi: 10.1109/BigData55660.2022.10020971
4. *Pengcheng Xia, Haoyu Wang, Bingyu Gao, Weihang Su, Zhou Yu, Xiapu Luo, Chao Zhang, Xusheng Xiao, and Guoai Xu. 2021. Trade or Trick? Detecting and Characterizing Scam Tokens on Uniswap Decentralized Exchange. Proc. ACM Meas. Anal. Comput. Syst. 5, 3, Article 39 (December 2021), 26 pages.* *https://doi.org/10.1145/3491051*
5. Zhou, L., Xiong, X., Ernstberger, J., Chaliasos, S., Wang, Z., Wang, Y., Qin, K., Wattenhofer, R., Song, D.X., & Gervais, A. (2022). SoK: Decentralized Finance (DeFi) Incidents. *ArXiv, abs/2208.13035.*