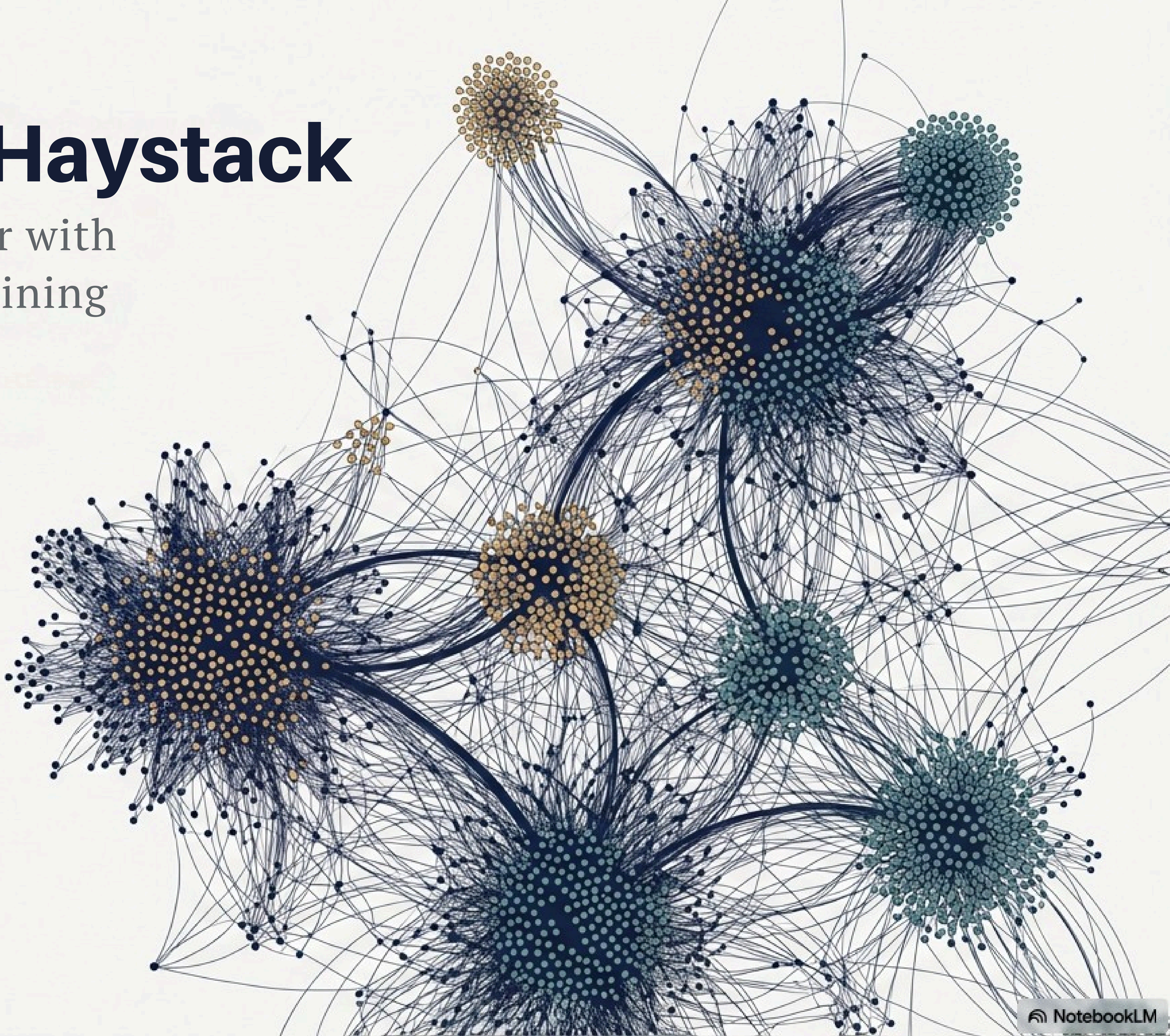
Needles in the Haystack

Decoding On-Chain Behavior with Network Science and Data Mining



Natkamon Tovanich
TU Wien, Austria

In collaboration with Rémy Cazabet and Célestin Coquidé

Blockchains: A Public Ledger of Economic Interaction

Cryptocurrency transaction data is freely accessible, large-scale, and dynamic—a new type of social network that records the flow of economic value.

Like a Social Network…

- Massive scale, revealing social and economic structure.
- Dynamic, with a constant stream of new interactions.
- Naturally represented as a network (graph).

…But for Economic Value

- Publicly auditable record of all transactions.
- Actors are pseudonymous, not anonymous.
- Represents direct value exchange, not just communication.

The analysis begins with understanding the two foundational, yet fundamentally different, blockchain architectures.

Bitcoin (BTC)

Introduced: 2008 by Satoshi Nakamoto.

Model: UTXO (Unspent Transaction Output). A system of digital cash and coins where a user's balance is the sum of many discrete outputs they control.

Focus: Primarily value transfer.

Analytical Challenge: Pseudonymity. One user controls many addresses, and best practice is to never reuse an address. The address network is not the user network.

Ethereum (ETH)

Introduced: 2015 by Vitalik Buterin.

Model: Account-based. A user has a single account with a balance, similar to a bank account.

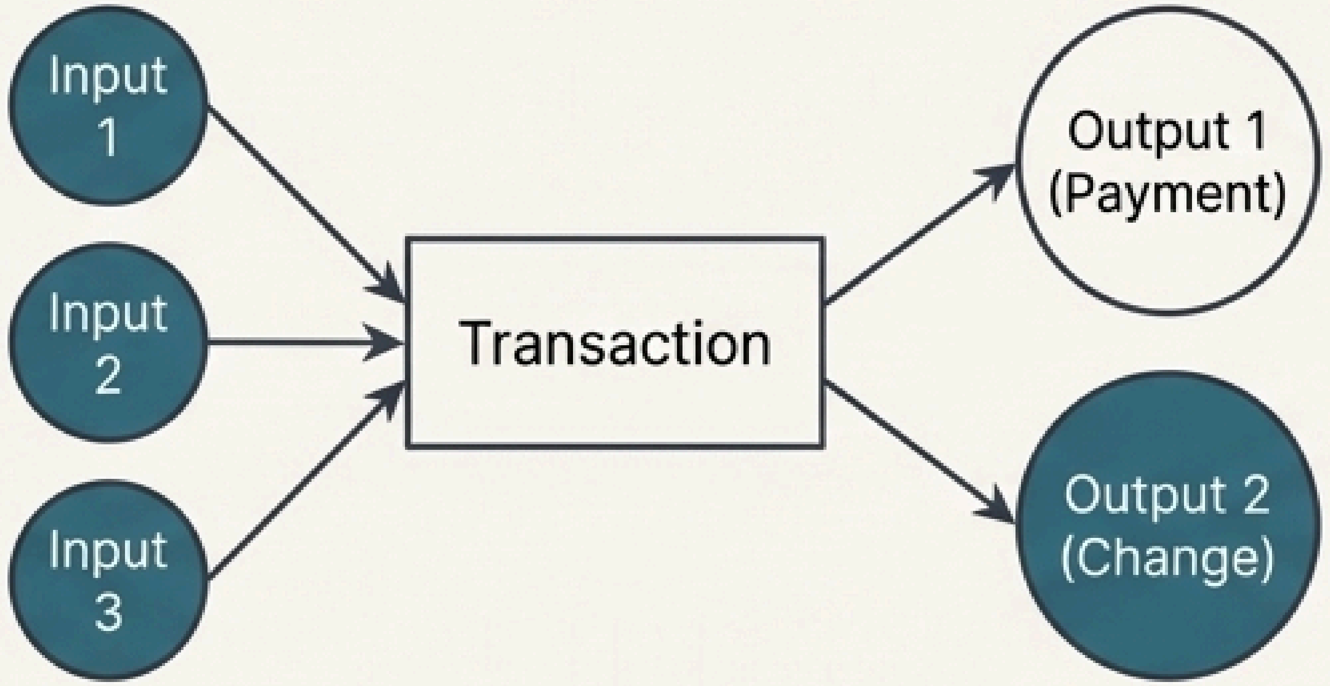
Focus: Programmable money via 'Smart Contracts,' enabling a complex ecosystem of DeFi, NFTs, and other decentralized applications.

Analytical Challenge: Complexity. Transactions can be simple transfers, token swaps, or intricate, multi-step contract interactions.

Two Architectures, Two Types of Graph Data

Bitcoin (UTXO Model)

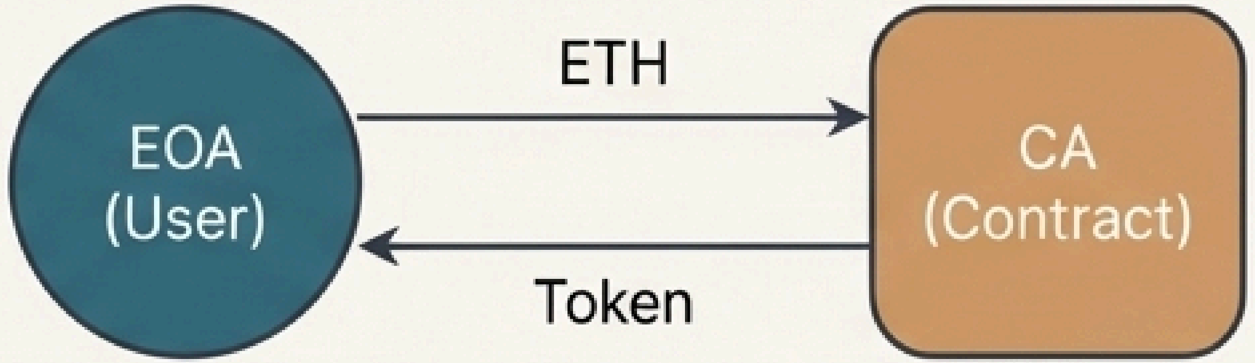
A graph of unspent outputs.



In the UTXO model, a user's balance is the sum of many individual transaction outputs they control. Users are encouraged to use new addresses for each transaction.
The ML Challenge: One "user" controls many addresses. The first step is **Address Clustering**, an unsupervised learning problem similar to entity resolution.

Ethereum (Account Model)

A state machine of accounts.



The account model simplifies analysis, as users typically reuse a single address.
The ML Opportunity: Smart Contracts (CAs) enable complex interactions like DeFi and Tokens (ERC-20, NFTs), creating a rich, *heterogeneous* network with multiple node and edge types.

MODULE 1: DECODING BITCOIN - THE CHALLENGE OF ANONYMITY

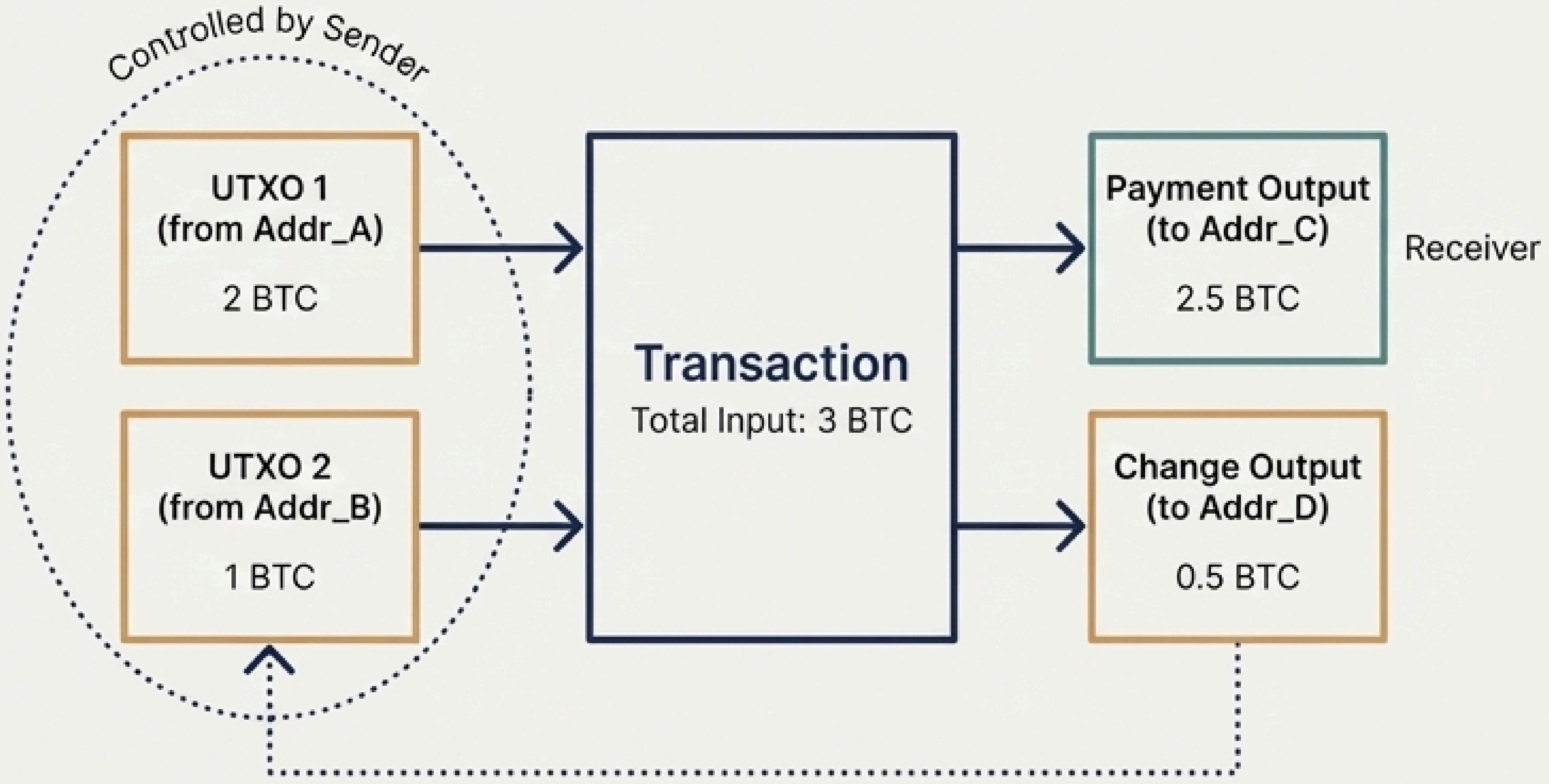
Bitcoin's UTXO model complicates analysis: transactions link loose outputs, not stable accounts.

A user's wallet is a collection of multiple Unspent Transaction Outputs (UTXOs).

A single transaction can have many inputs (from different UTXOs controlled by the sender) and many outputs (to the receiver and a 'change' output back to the sender).

Privacy best practice, recommended since the original white paper, is to avoid reusing addresses.

Conclusion: Therefore, the raw address network is a poor proxy for the true user network.



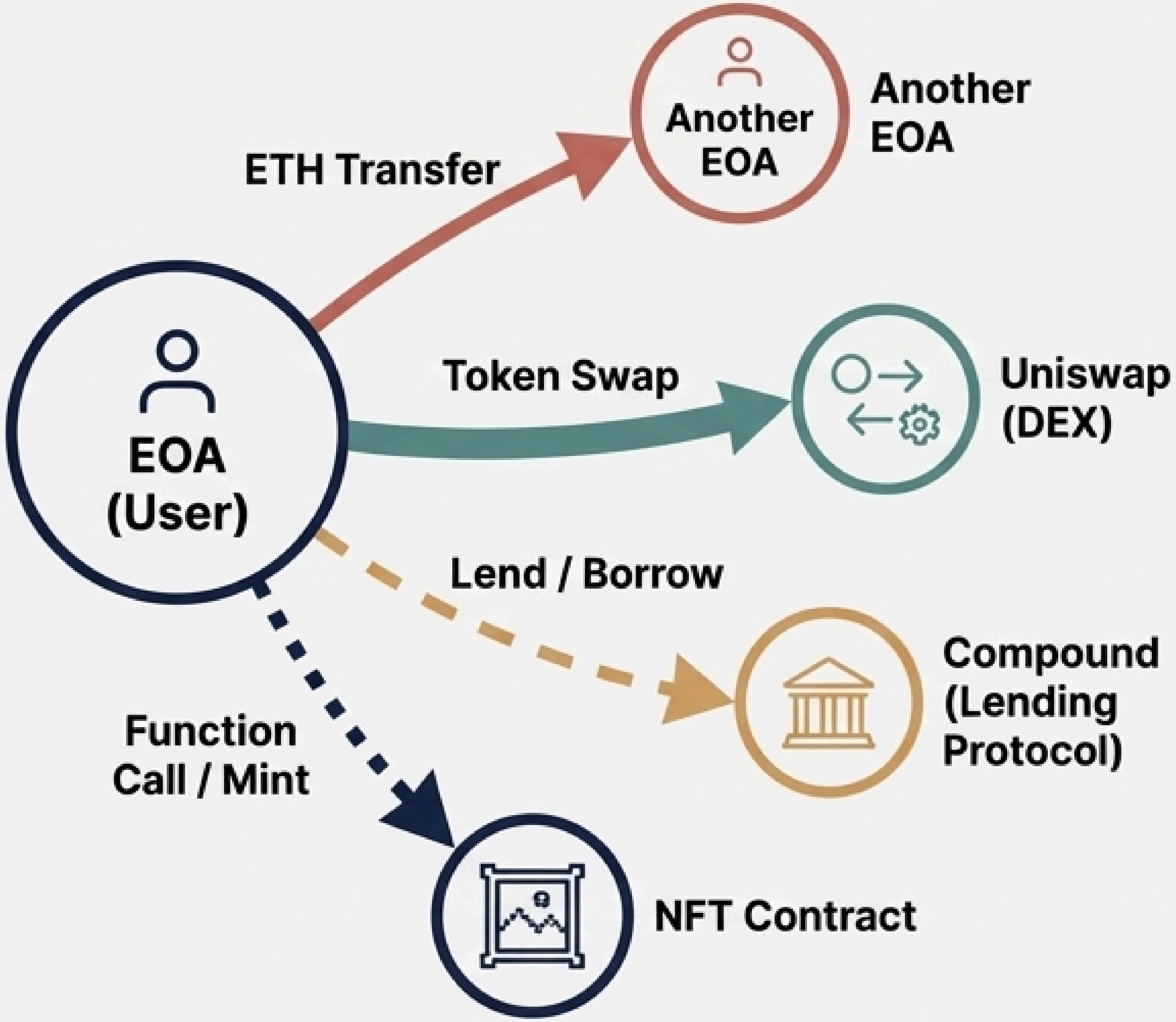
MODULE 2: UNRAVELING ETHEREUM - THE CHALLENGE OF COMPLEXITY

Ethereum's 'World Computer' enables transactions that are value transfers, contract creations, or complex function calls.

Key Idea: The account-based model simplifies user identity (one user, one EOA), but Smart Contracts create a vast, heterogeneous network of interactions, layering new economies on top of the base protocol.

Core Concepts:

- **Two Account Types:** Externally Owned Accounts (**EOAs**), controlled by users via private keys, and **Contract Accounts (CAs)**, which are autonomous code deployed on the network.
- **Tokens (ERC-20, NFTs):** Digital assets built on Ethereum, creating thousands of parallel economies for everything from stablecoins to digital art.
- **Decentralized Finance (DeFi):** A financial ecosystem using smart contracts to create services like exchanges, lending, and borrowing without traditional intermediaries.

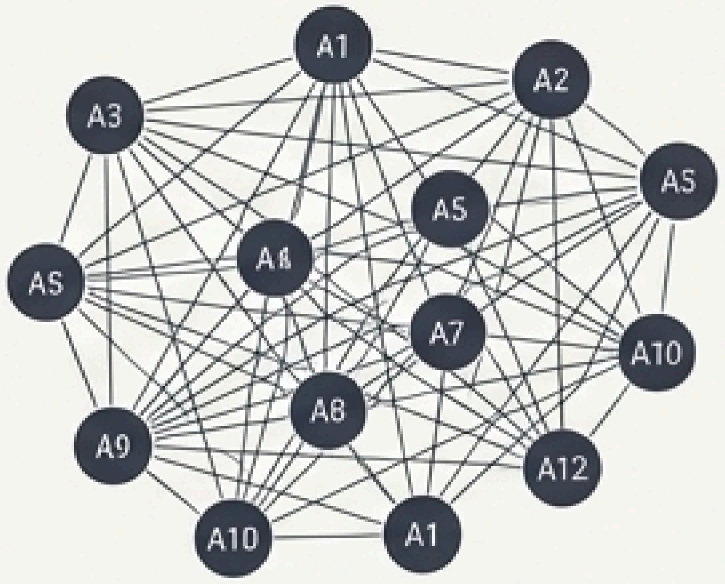


NotebookLM

From Raw Data to Actionable Graphs

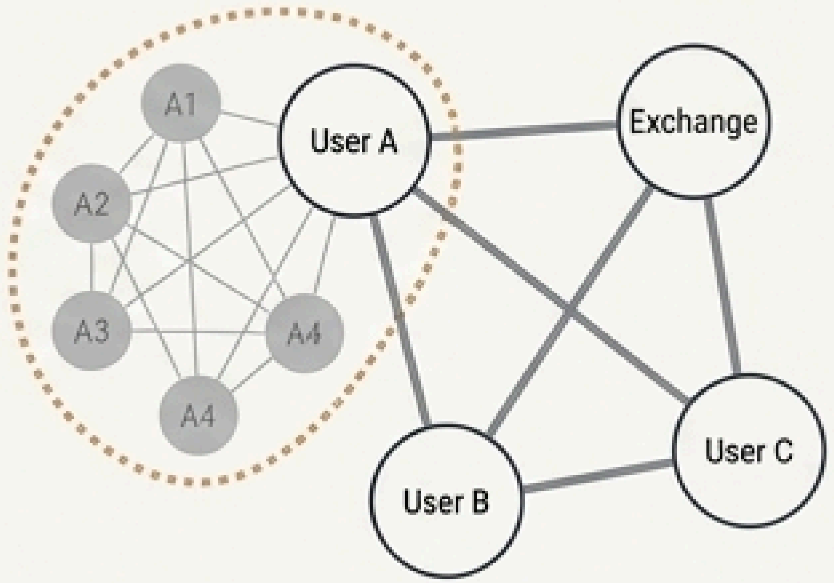
Raw Transaction

```
from: 0x1a...3c
to: 0x5d...e8
value: 2.5 ETH
timestamp: 2023-10-27 10:30:00 UTC
```



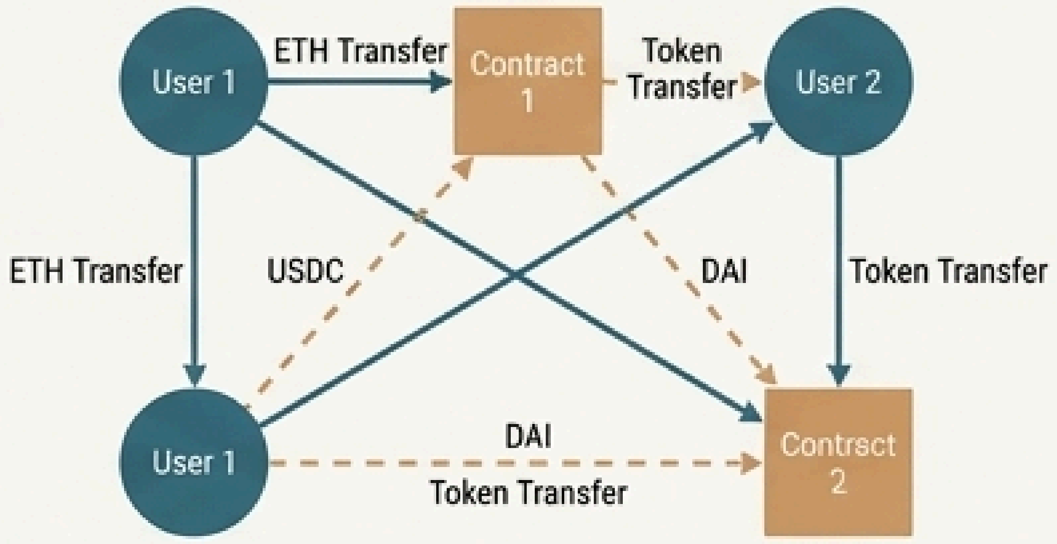
Address Network

The most direct representation. Simple but noisy, as one real-world transaction can create a large number of edges.



User Network (Post-Clustering)

A more meaningful representation where nodes are actors. Requires accurate address clustering heuristics or ML models to build.



Token Transfer Network (Ethereum)

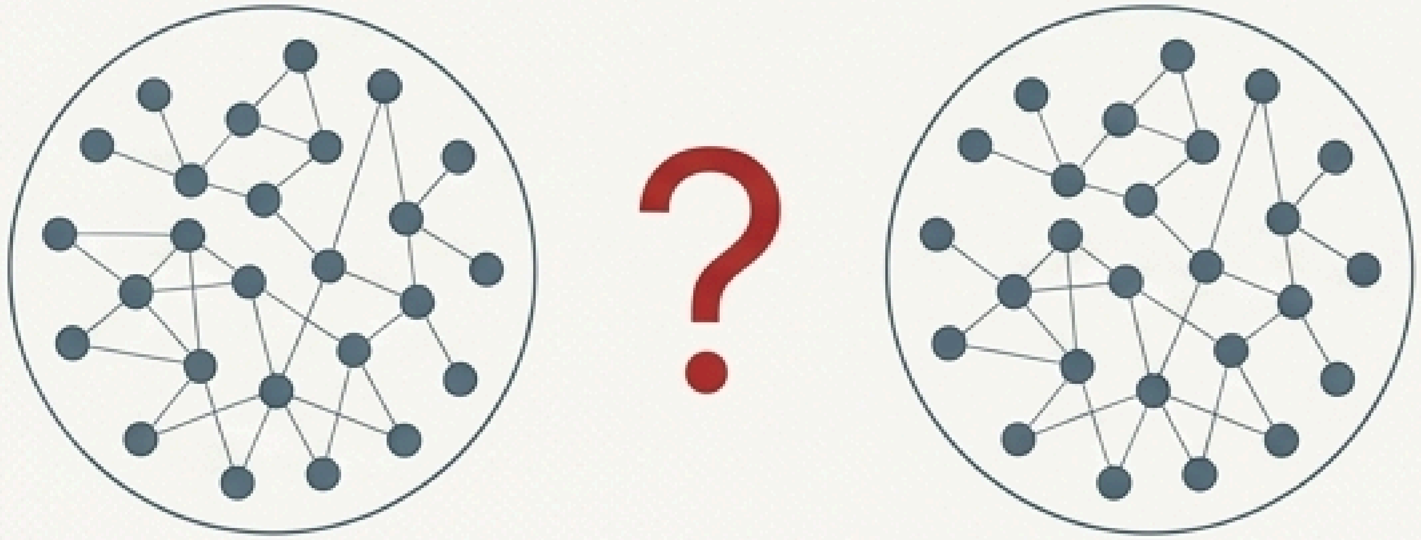
Captures the full richness of the Ethereum ecosystem, representing the flow of diverse digital assets. This is the foundation for advanced DeFi analysis.

Case Study 1: The Fingerprint of Bitcoin Entities

How can we identify that multiple, unlinked clusters of addresses belong to the same entity?

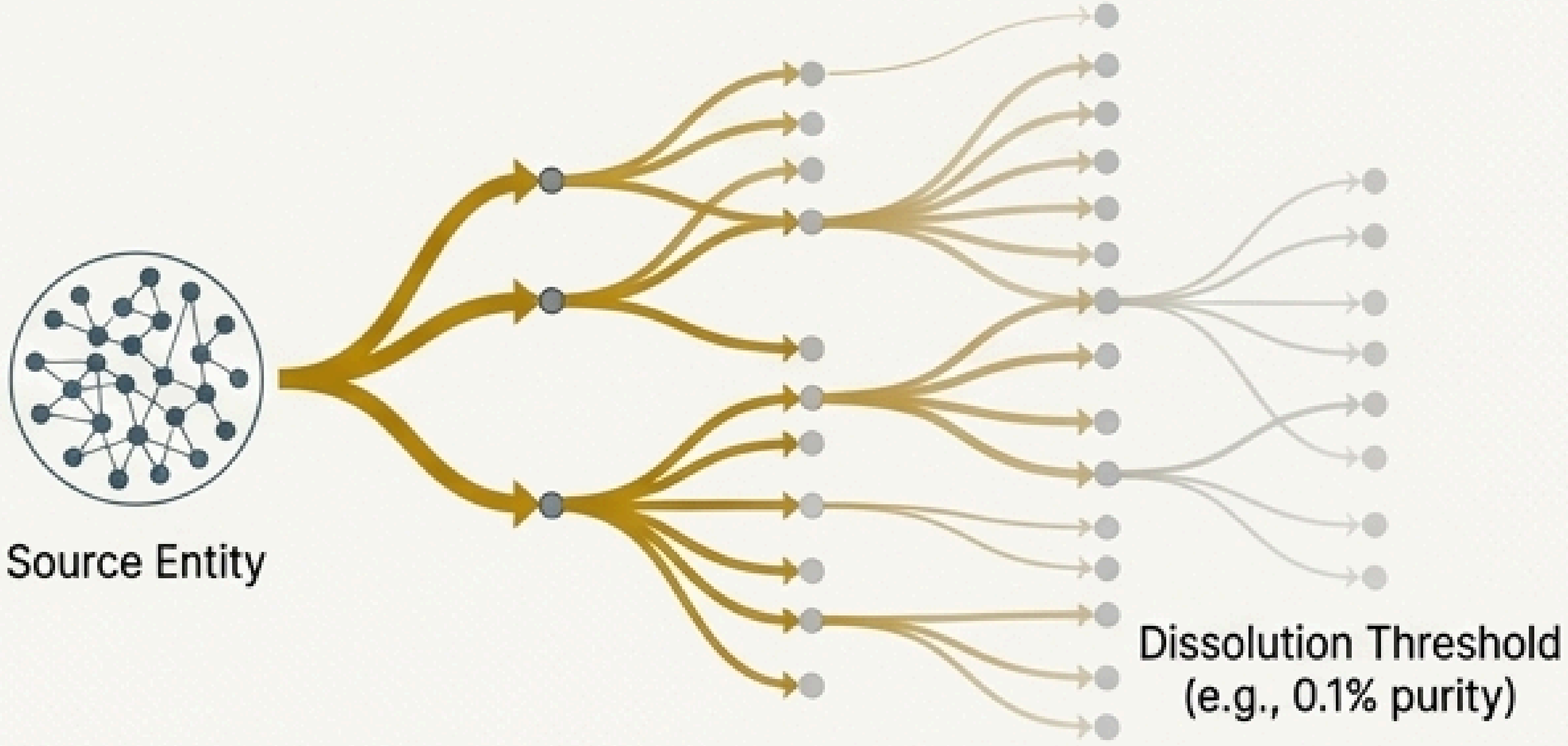
Problem Statement

Standard address clustering heuristics (e.g., common-input) are effective but incomplete. A single entity can operate multiple, separate “wallets” or address pools that never interact directly. These methods cannot match sub-clusters if they never make a transaction together.



Proposed Solution

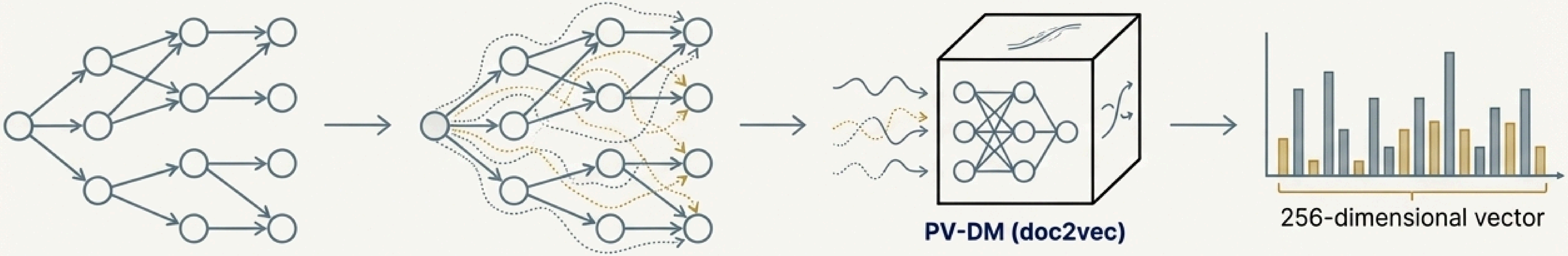
We can characterize an entity by its downstream money flow. We introduce the concept of a **Taint Flow**: the directed acyclic graph (DAG) of all subsequent transactions involving coins originating from a source entity on a given day. The flow is tracked until the “taint” is dissolved (e.g., purity drops below a threshold of 0.1%).



NotebookLM

The Solution: Learning a Fingerprint from Taint Flow

Our method synthesizes each complex taint flow graph into a single 256-dimensional vector—a “fingerprint”—that captures its unique topological and temporal characteristics.



1. Extract Taint Flow

Start with an entity’s transactions on a given day. Show a small DAG representing the flow.

2. Generate Walks

Generate 10,000 random walks from the source to dissolution. This process translates the graph structure into a collection of “sentences”.

3. Embed Walks

Feed these “sentences” into a Distributed Memory Model of Paragraph Vectors (PV-DM), a `doc2vec` model. The model learns a vector representation for the entire document (the taint flow).

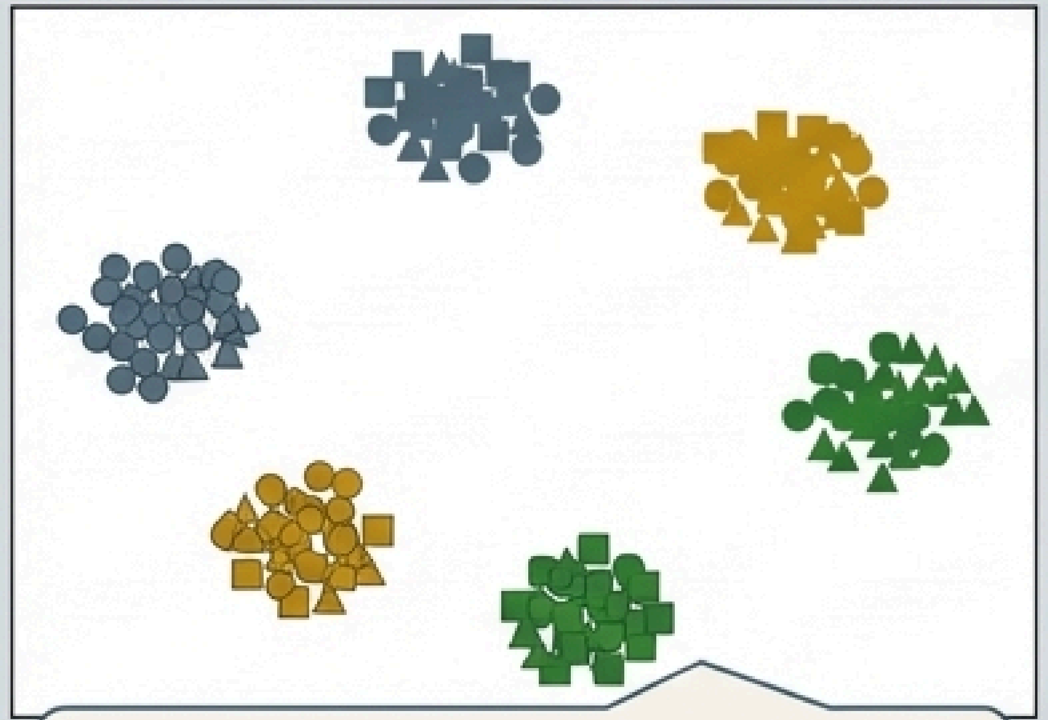
4. The Fingerprint

The output is a single vector that serves as the fingerprint for that specific flow.

The Payoff: Taint Flows Form Distinct, Classifiable Clusters

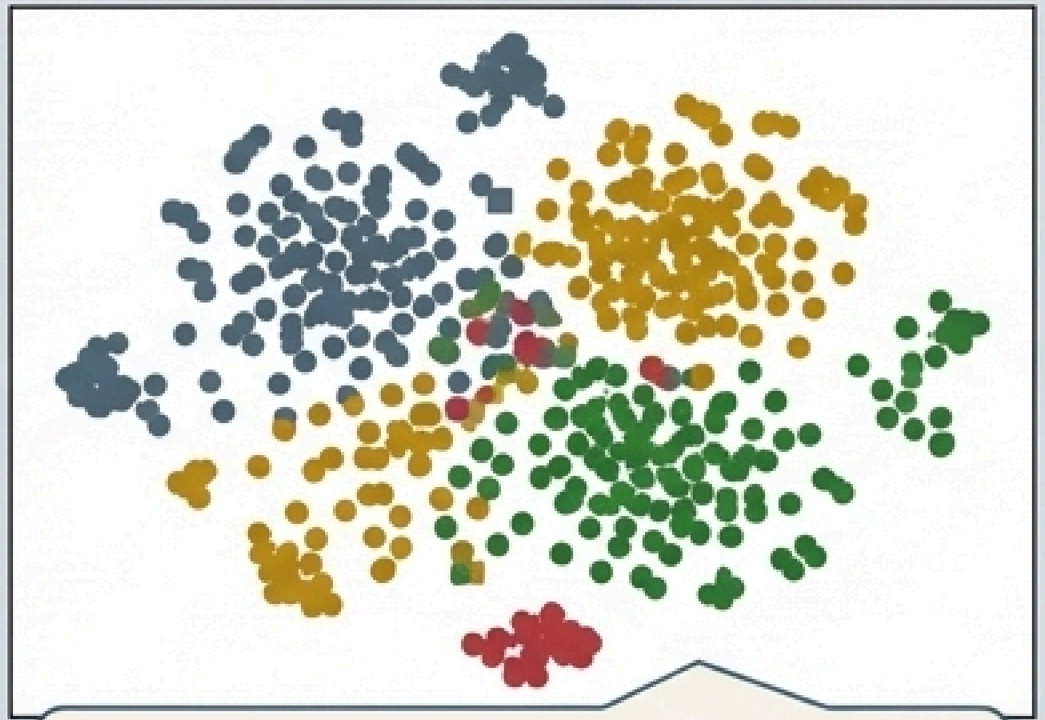
By projecting the 256-dimensional fingerprints into 2D space using t-SNE, we can visualize the separation between entities. The results show that taint flows are a valid basis for recognizing and classifying actors.

Mining Pools



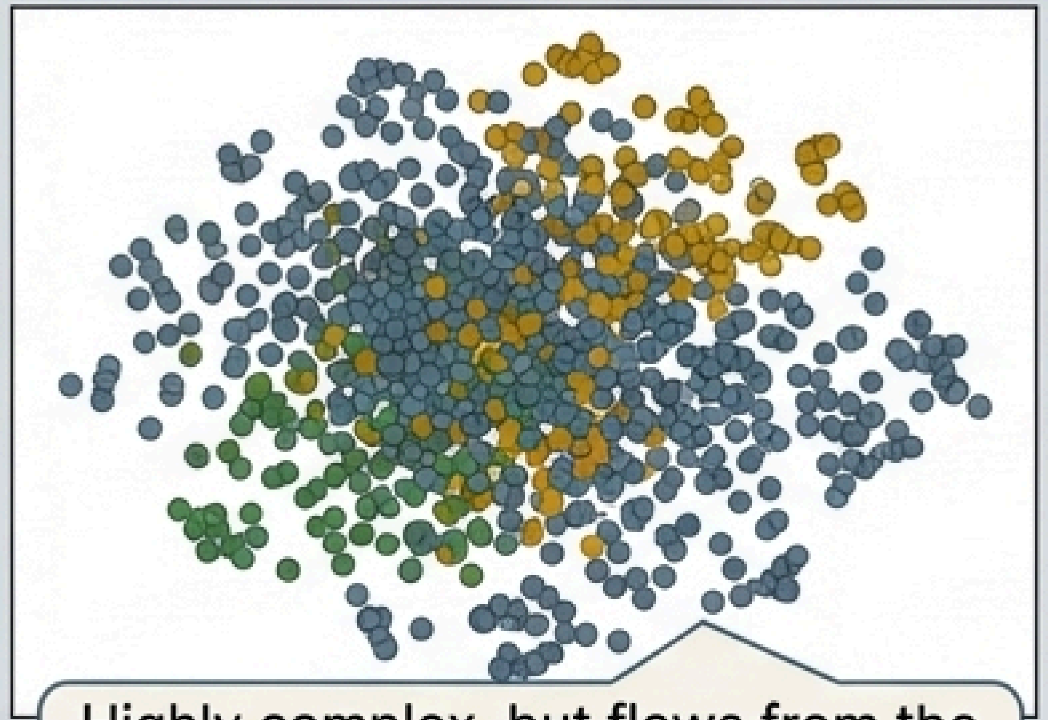
Clear separation between pools. Classification accuracy: **96.2%**.

Ransomware Families



More overlap, but distinct clusters are still visible for major families. Classification accuracy: **70.1%**.

WalletExplorer Entities



Highly complex, but flows from the same entity (shape) and entity type (color) tend to be closer. Classification accuracy: **62.8%**.

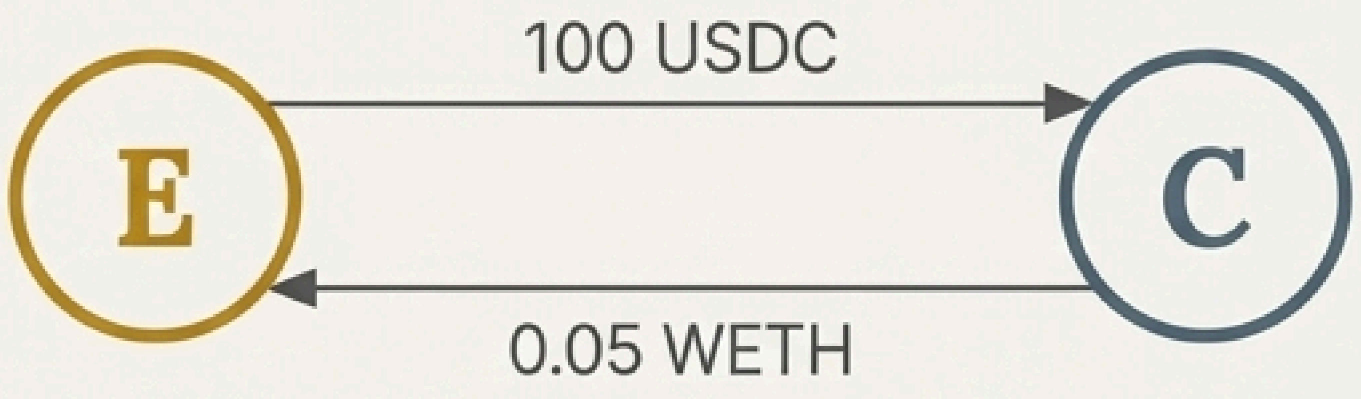
Key Insight: High accuracy is achieved using very short path lengths. Tracking flows just **2-3 steps** from the source is often sufficient to identify the entity, implying that near-neighbor interactions are most characteristic.

Case Study 2: Anatomy of DeFi Transactions

Can we infer DeFi methods (e.g., depositing, borrowing, swapping) from ego token transfer network?

Example Transaction

```
Account: Alameda Research 19.
Method: Swap Exact Tokens For Tokens.
Alameda Research (Ego) sends 100 USDC to
the Uniswap V2 contract and receives
0.05 WETH from the contract in return.
```

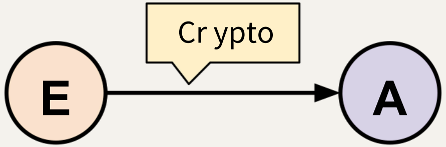


Analyzing DeFi transactions is challenging due to often incomplete or inaccurate labels. Our approach bypasses this by focusing on the structure of token transfers within a single transaction.

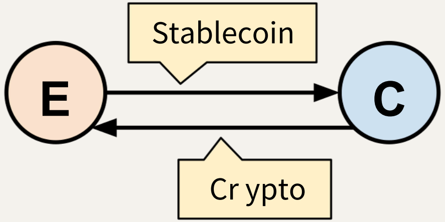
For each transaction, we construct an **Ego Transfer Network (ETN)**, a directed graph centered on the account of interest (“Ego”). Nodes represent account types: **E** (Ego), **A** (Address), **C** (Contract), or **N** (Null Address for mint/burn). Edges represent token transfers.

Token Transfers in Ethereum

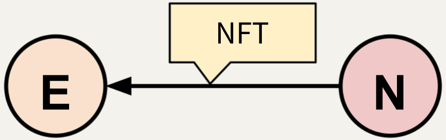
Transfer



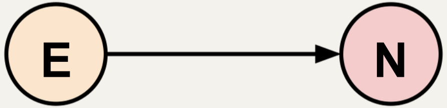
Swap



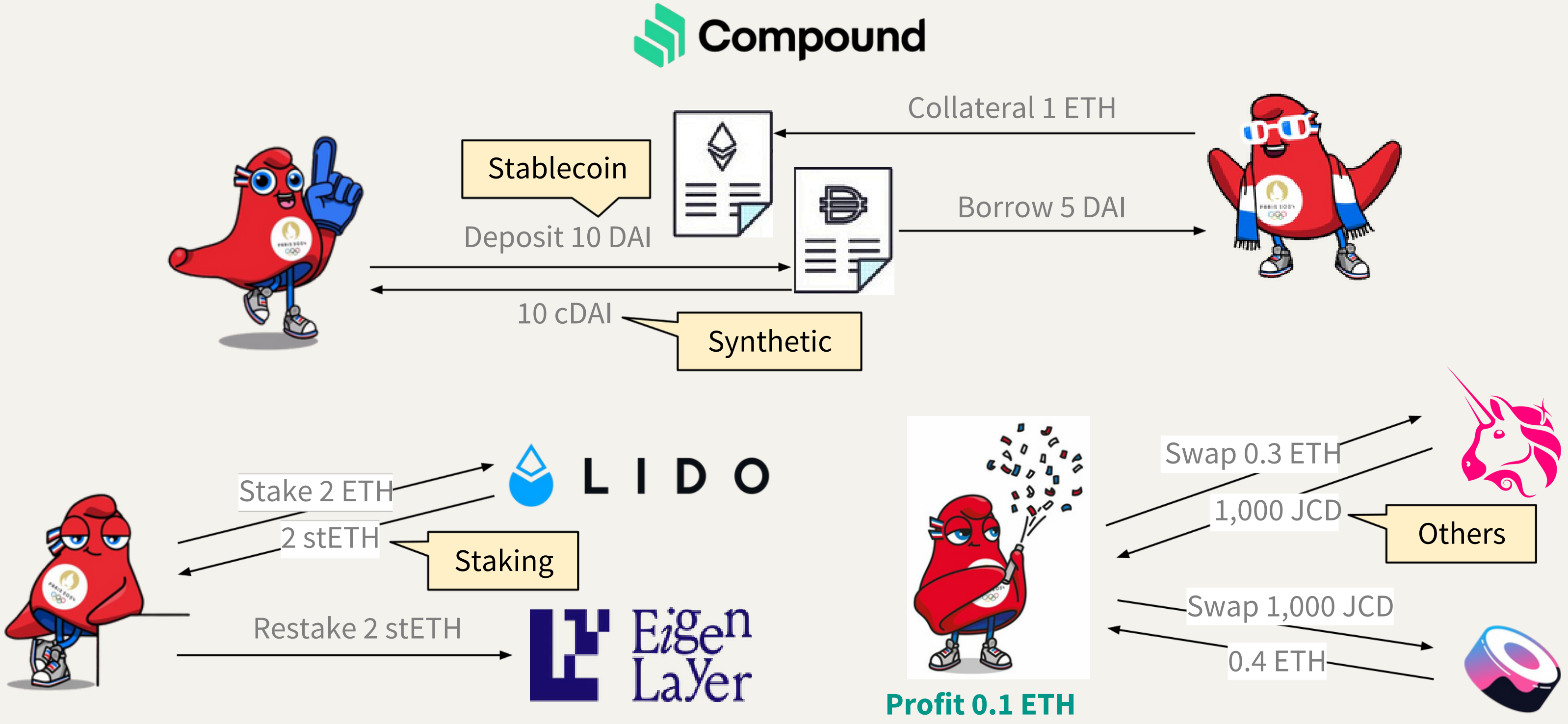
Mint



Burn



Complex DeFi Transactions



Compound

Stablecoin

Deposit 10 DAI

10 cDAI

Synthetic

Collateral 1 ETH

Borrow 5 DAI

Stake 2 ETH

2 stETH

Staking

LIDO

Restake 2 stETH

EigenLayer

Swap 0.3 ETH

1,000 JCD

Others

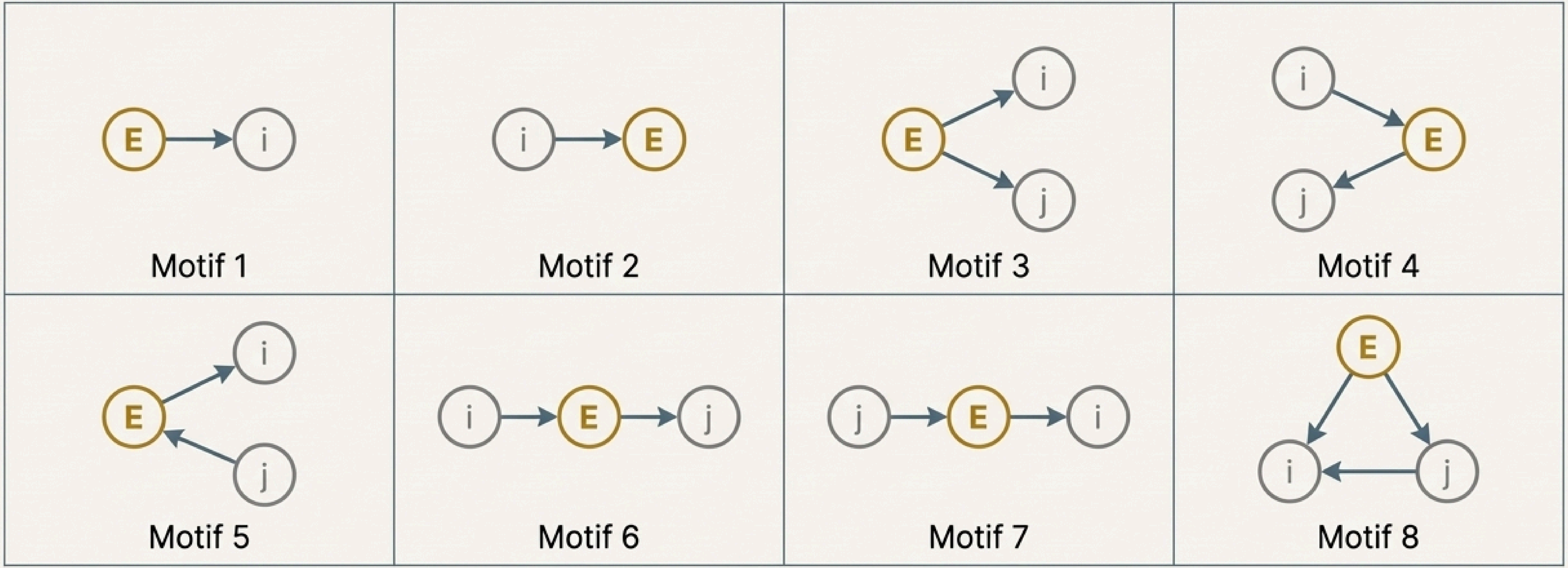
Swap 1,000 JCD

0.4 ETH

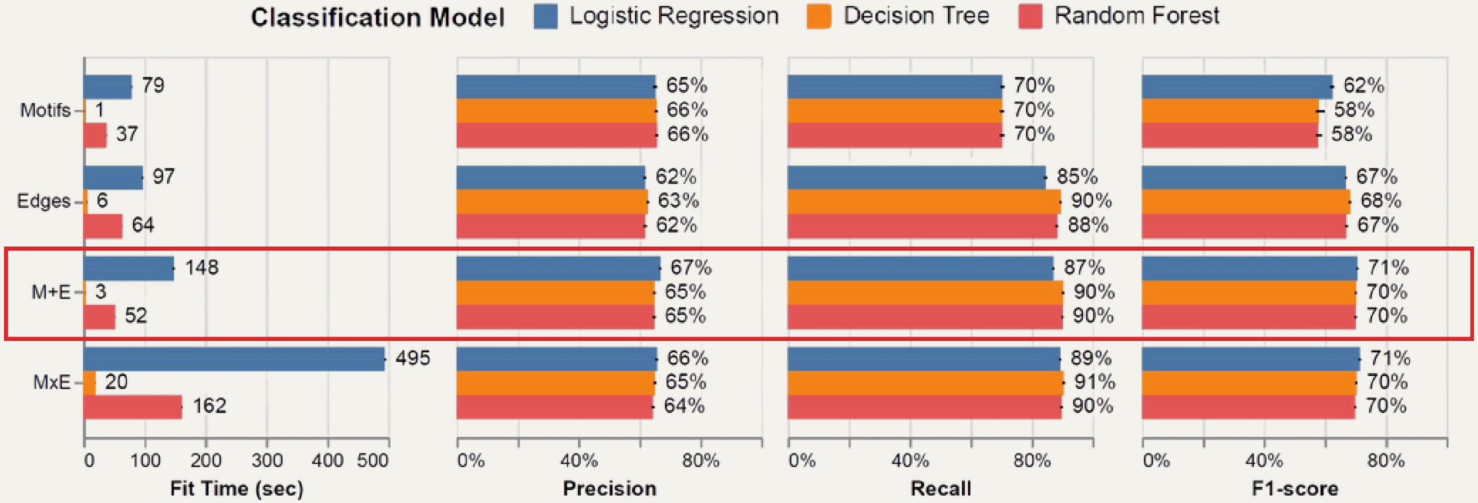
Profit 0.1 ETH

The Solution: Decomposing Transactions into Motifs

Instead of analyzing the entire ETN, we decompose it into fundamental building blocks called **ego network motifs**—small, directed subgraphs of 2 or 3 nodes. Because we focus on ego networks, there are only 8 possible directed motif structures. These motifs, combined with the node types (Address, Contract, Null) and token types involved, create a rich feature set to describe any transaction.



DeFi Method Classification



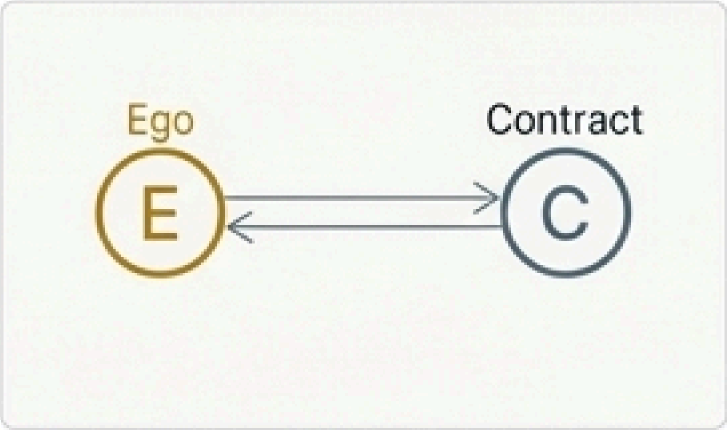
Best model: F1-score = 71%, Precision = 67%, Recall = 89% (Logistic Regression)

Key Insight: Combining motifs and edge list features improves accuracy.

From Motifs to Meaning: Interpretable Signatures of DeFi Methods

By analyzing the most frequent motifs in transactions with known labels, we can extract "signatures" for specific DeFi methods. A pruned decision tree model helps identify the most discriminative patterns, making the classification highly interpretable.

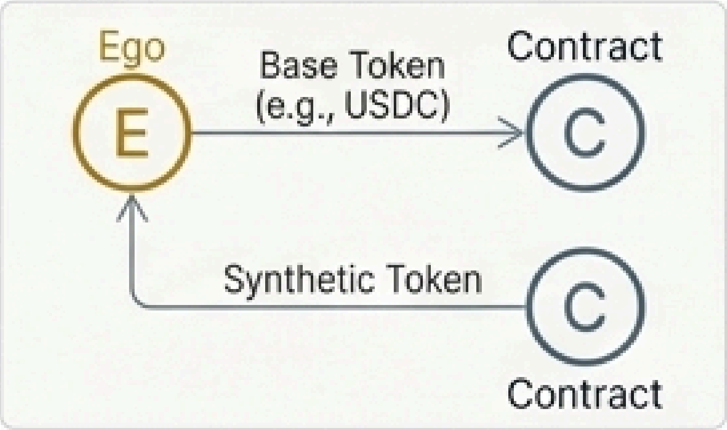
Swap (Leaf 31)



Signature Motif

A simple, direct exchange of tokens between the Ego and a Contract.

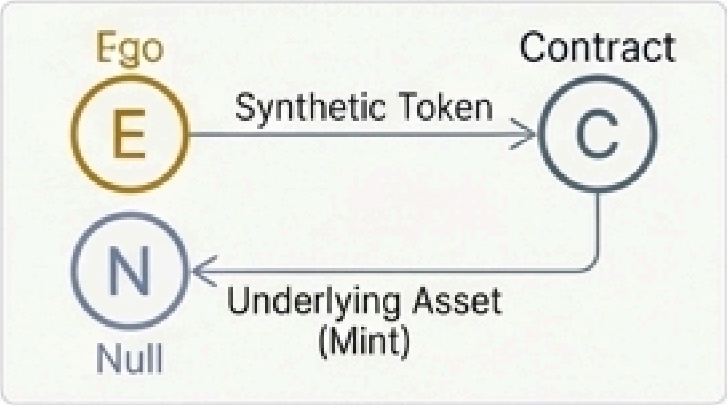
Deposit (Leaf 28)



Signature Motif

Ego sends a token to a Contract and receives a synthetic representative token in return.

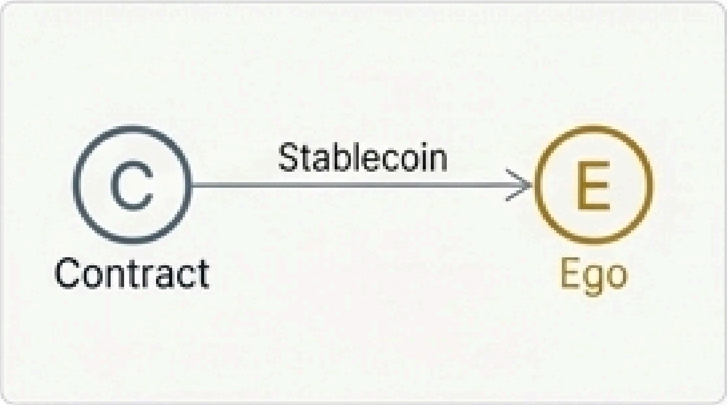
Withdraw (Leaf 17)



Signature Motif

Ego returns a synthetic token to a Contract to retrieve the original underlying asset.

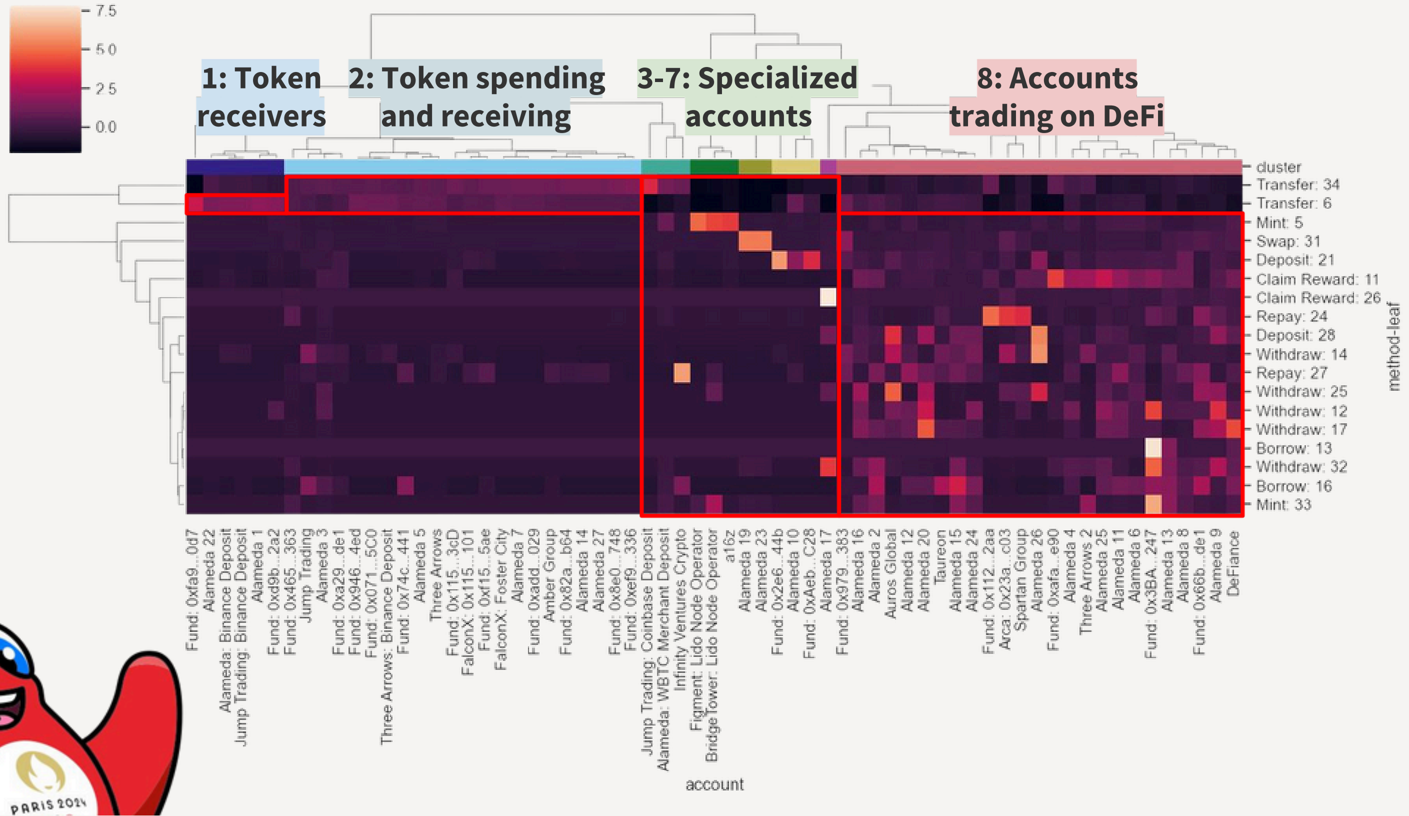
Borrow (Leaf 16)



Signature Motif

Ego receives a stablecoin from a lending contract, typically after posting collateral in a prior action.

Profiling Ethereum Accounts


1: Token receivers

2: Token spending and receiving

3-7: Specialized accounts

8: Accounts trading on DeFi

Case Study 3: Cross-Token Interactions in DeFi

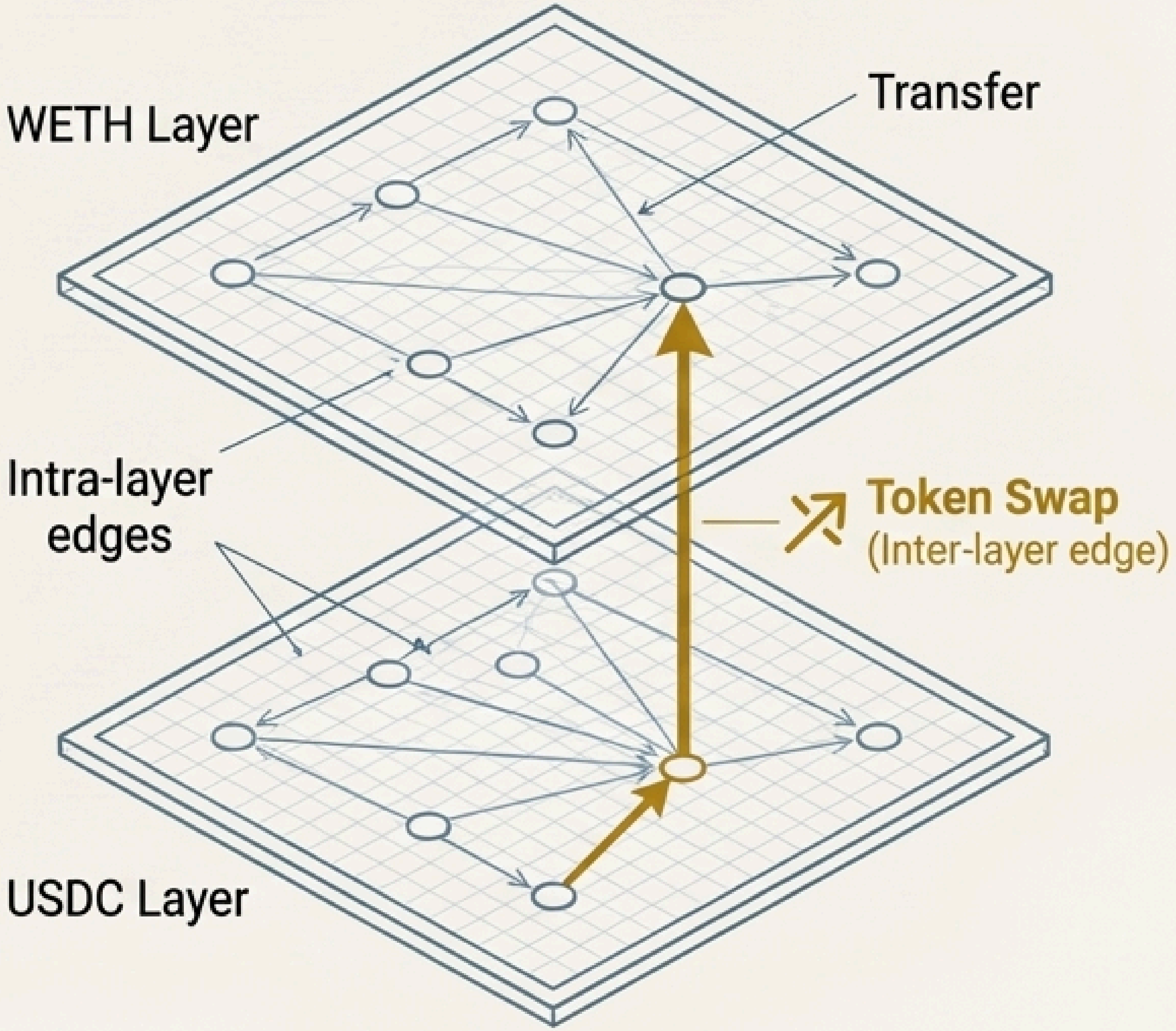
How do we capture inter-token transformations crucial to DeFi, such as token swaps?

Problem Statement

Standard network models treat each token's transfer graph in isolation. They fail to capture the critical moment when one asset is transformed into another within the same transaction (e.g., swapping ETH for USDC).

Proposed Solution: The Multilayer Token Network (MLTN)

- **Framework**: A network where nodes are addresses, and each token exists on its own "layer".
- **Intra-layer edges**: Represent transfers of the *same* token between two addresses.
- **Inter-layer edges**: The key innovation. A directed edge is created between layers for the same address if, within one transaction, it sends one token and receives another. This explicitly models token transformation.



The Solution: Quantifying Strategy with Multilayer Centrality

To measure influence and behavior in the MLTN, we adapt standard network centralities and introduce a new metric to quantify trading strategy.

1. Biased PageRank & CheiRank

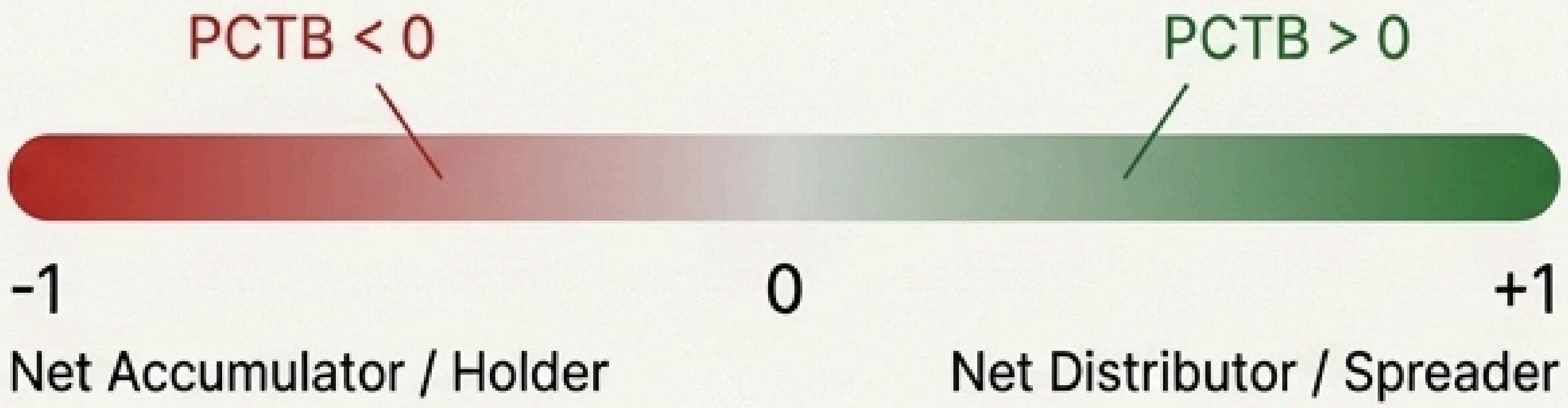
We use a specialized Biased PageRank model. Unlike standard PageRank, its “teleportation” is biased to prefer jumps along inter-layer edges. This gives greater weight to actors who frequently engage in token transformations.

- **PageRank** measures an address’s capacity to accumulate tokens (an influence sink).
- **CheiRank** (PageRank on the reversed graph) measures its capacity to distribute tokens (an influence source).

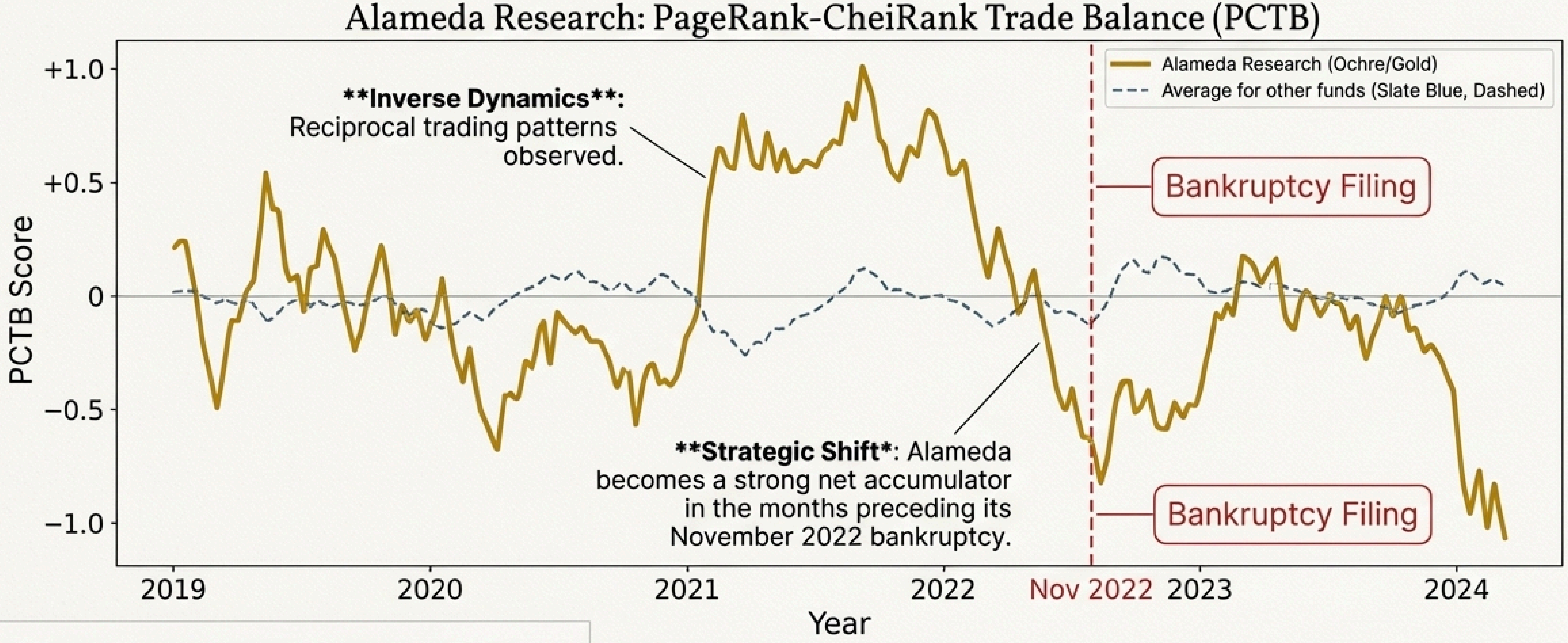
2. PageRank-CheiRank Trade Balance (PCTB)

We define a single score to capture net behavior:

$$\mathrm{PCTB} = \frac{\mathrm{CheiRank} - \mathrm{PageRank}}{\mathrm{CheiRank} + \mathrm{PageRank}}$$



The Payoff: Decoding Alameda Research's Evolving Strategy


Alameda Research: PageRank-CheiRank Trade Balance (PCTB)

Inverse Dynamics: Reciprocal trading patterns observed.

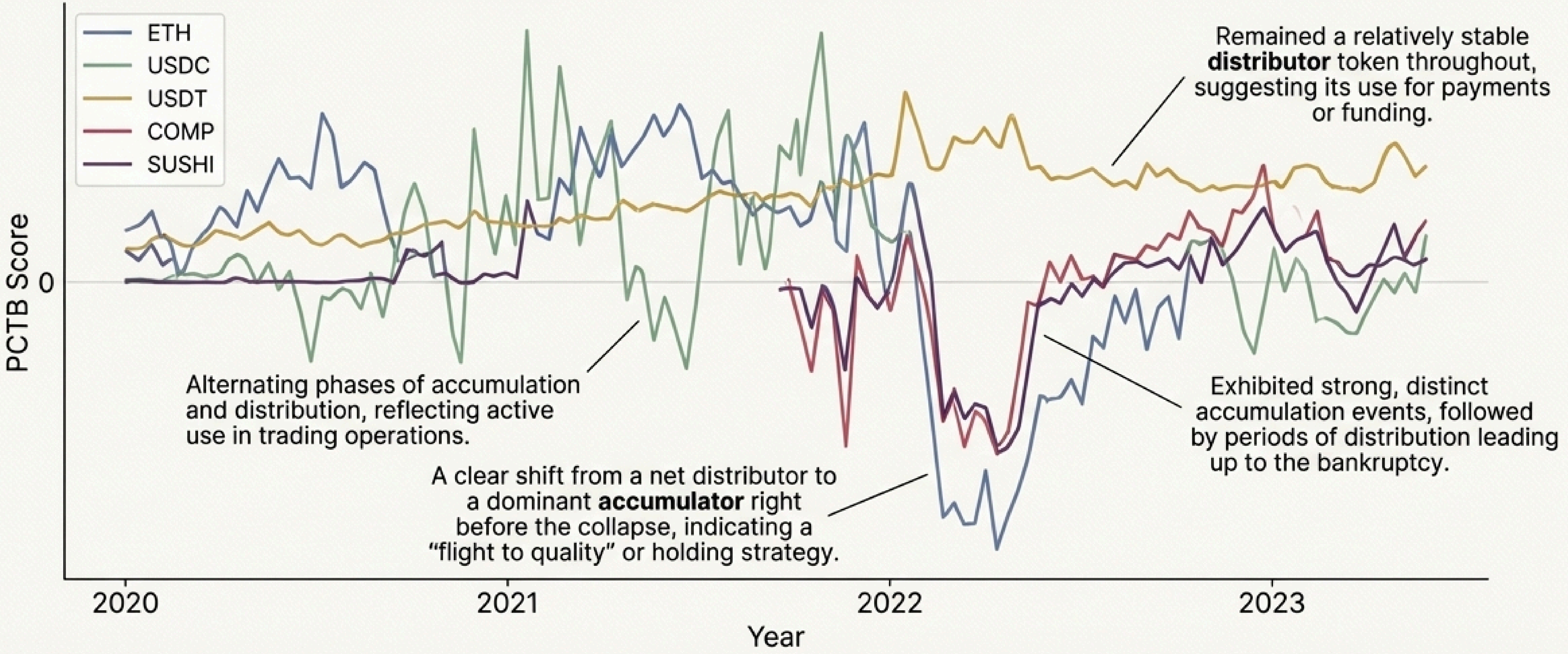
**Strategic Shift*: Alameda becomes a strong net accumulator in the months preceding its November 2022 bankruptcy.

The PCTB score reveals distinct strategic phases and provides a quantitative narrative of Alameda's market behavior.

A Token-Level View of Alameda's Trading Strategy

The PCTB can be calculated for specific tokens, revealing how Alameda managed its portfolio and prioritized different assets over time.

Alameda Research: PCTB Scores by Token



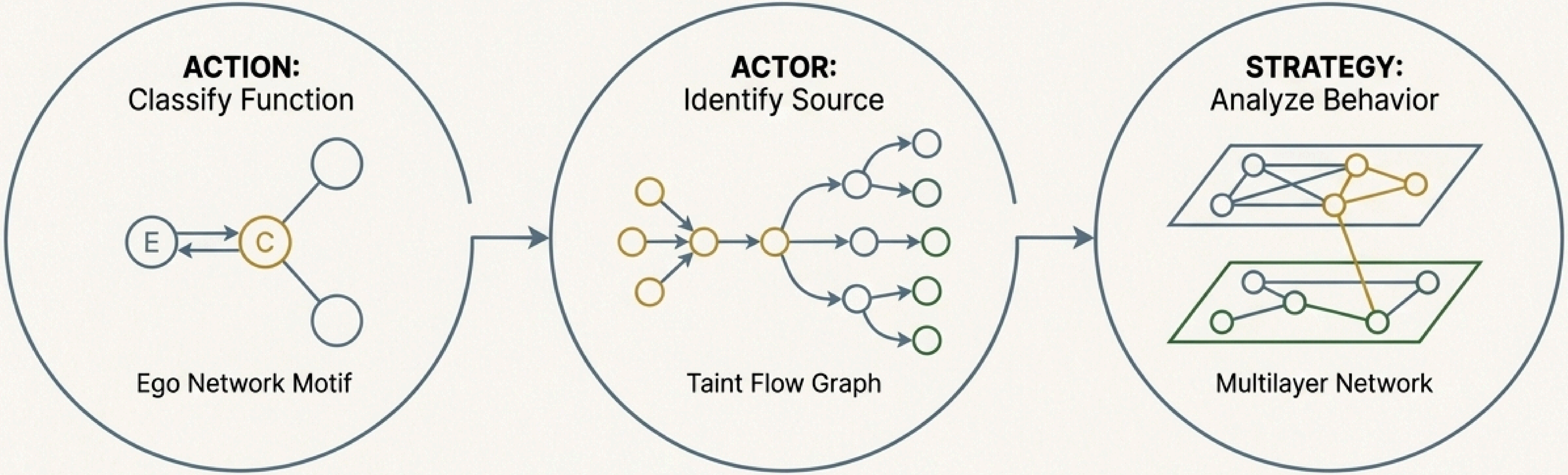
Remained a relatively stable **distributor** token throughout, suggesting its use for payments or funding.

Alternating phases of accumulation and distribution, reflecting active use in trading operations.

A clear shift from a net distributor to a dominant **accumulator** right before the collapse, indicating a “flight to quality” or holding strategy.

Exhibited strong, distinct accumulation events, followed by periods of distribution leading up to the bankruptcy.

The Unified Framework: From Actions to Actors to Strategies



Key Takeaways:

- **Local patterns reveal function:** Ego network motifs provide an interpretable way to classify on-chain transactions without labels.
- **Downstream flows reveal identity:** Taint flow embeddings create unique, machine-learnable fingerprints to identify entities across pseudonymous addresses.
- **Inter-asset dynamics reveal strategy:** Multilayer network centralities quantify the complex, evolving trading strategies of major DeFi players.

Future Vision:

This multi-scale toolkit provides a foundation for the next frontier of on-chain analysis, enabling real-time strategic monitoring, predictive modeling of market behavior, and a deeper understanding of systemic risk in the DeFi ecosystem.

An Unprecedentedly Open Data Ecosystem

Unlike proprietary data from private firms, core on-chain data is public. A rich ecosystem of open-source source tools and platforms makes it accessible for large-scale analysis today.

Raw & Processed Data

Public Access

- Run your own full node for complete, trustless data access.

ETL Tools

- Open-source scripts like Blockchain ETL and Cryo transform raw data into usable formats (CSV, Parquet).

Big Data Platforms

- Processed, queryable datasets are available on platforms like Google BigQuery.

APIs & Analytics Platforms

Node-as-a-Service

- Services like Alchemy and Infura provide API access without infrastructure overhead.

Explorers & APIs

- Etherscan and other block explorers offer rich APIs for specific transaction data.

Analytics Engines

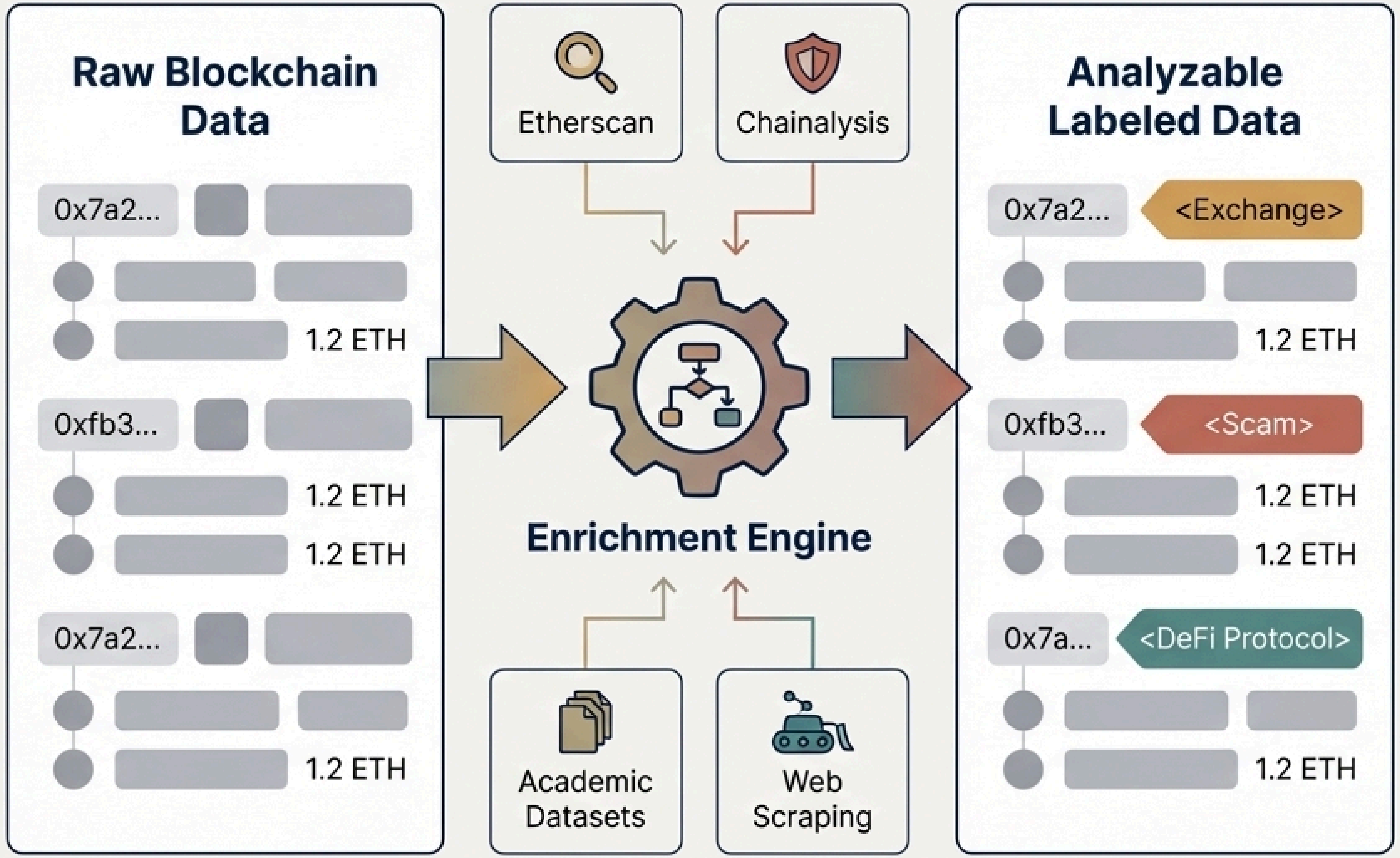
- Platforms like Dune Analytics provide SQL query engines and dashboards for analyzing curated on-chain data.

Google BigQuery ethereum alchemy Dune Analytics cryo

Raw on-chain data is pseudonymous; enriching it with off-chain labels is a critical step.

Sources for Contextual Data

- **Address Tagging:** Blockchain intelligence firms (e.g., Chainalysis, Elliptic) and public explorers (Etherscan, WalletExplorer) provide labels for addresses belonging to known entities like exchanges, mining pools, scams, and sanctions lists.
- **Curated Datasets:** Academic and public datasets (e.g., Elliptic Dataset, XBlock) offer pre-processed, labeled data tailored for specific research tasks, such as illicit activity detection.
- **Manual & Programmatic Labeling:** Researchers often scrape forums, social media, or use smart contract ABIs to manually label new entities and protocols as they emerge.



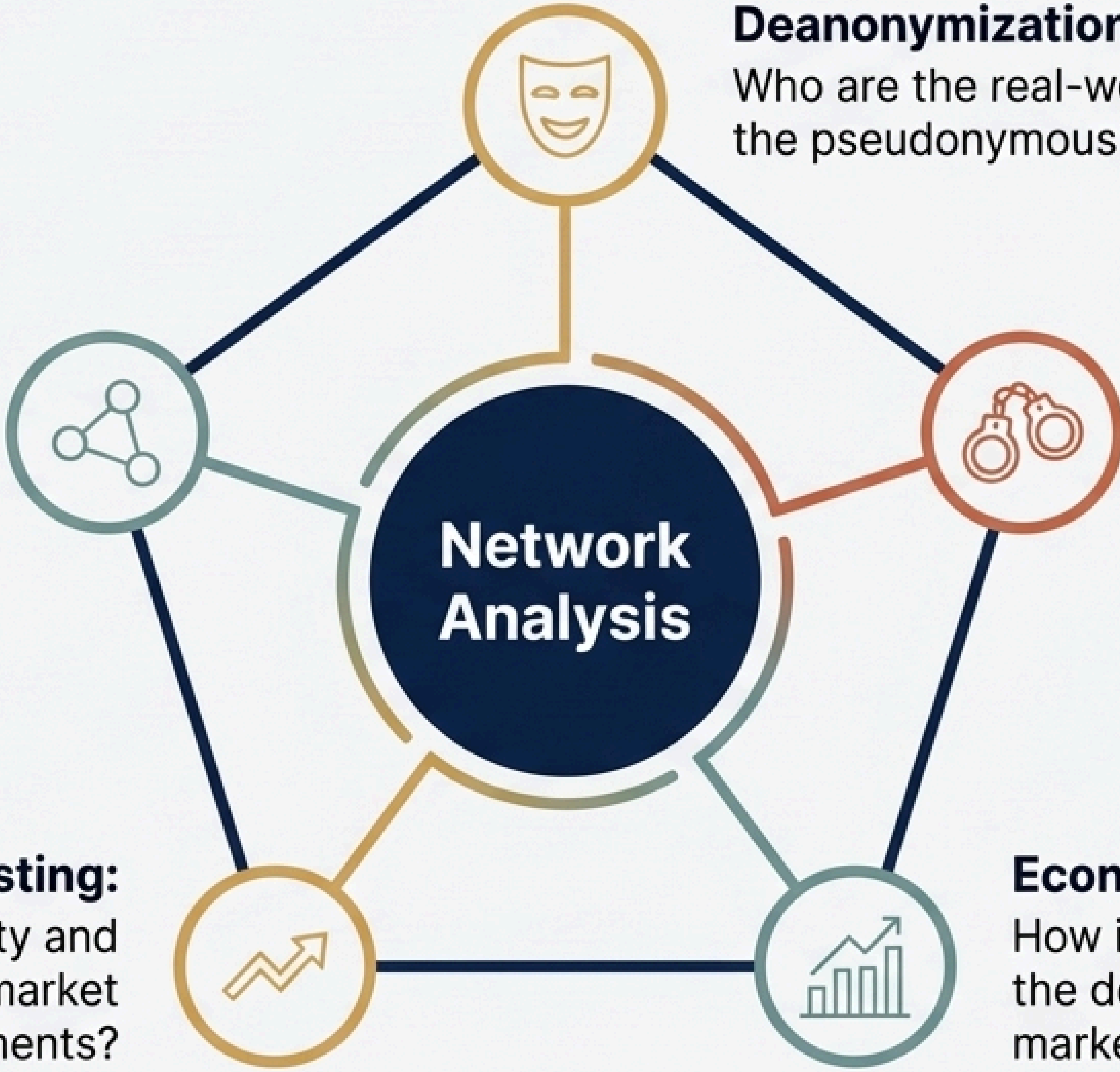
FairOnChain

faironchain.org

FairOnChain is an ambitious European collaboration aiming to develop a publicly accessible infrastructure that enables easy access and searchability of **blockchain data** in accordance with the **FAIR** principles of open science. This infrastructure aims to promote complete transparency and reproducibility of scientific analysis results in the blockchain field, thereby facilitating the growth and collaboration of new and existing applications.

This project is based upon participation in the CHIST-ERA 2022 call for Open & Re-usable Research Data and Software (ORD).

Network analysis aims to answer critical questions about these opaque ecosystems.



Deanonymization:
Who are the real-world entities behind the pseudonymous addresses?

Network Properties Analysis:
What is the macro-structure and evolution of the transaction graph? Is it scale-free? Small-world?

Illicit Activity Detection:
Can we identify and trace fraudulent users, money laundering from hacks and ransomware, and other cybercrime?

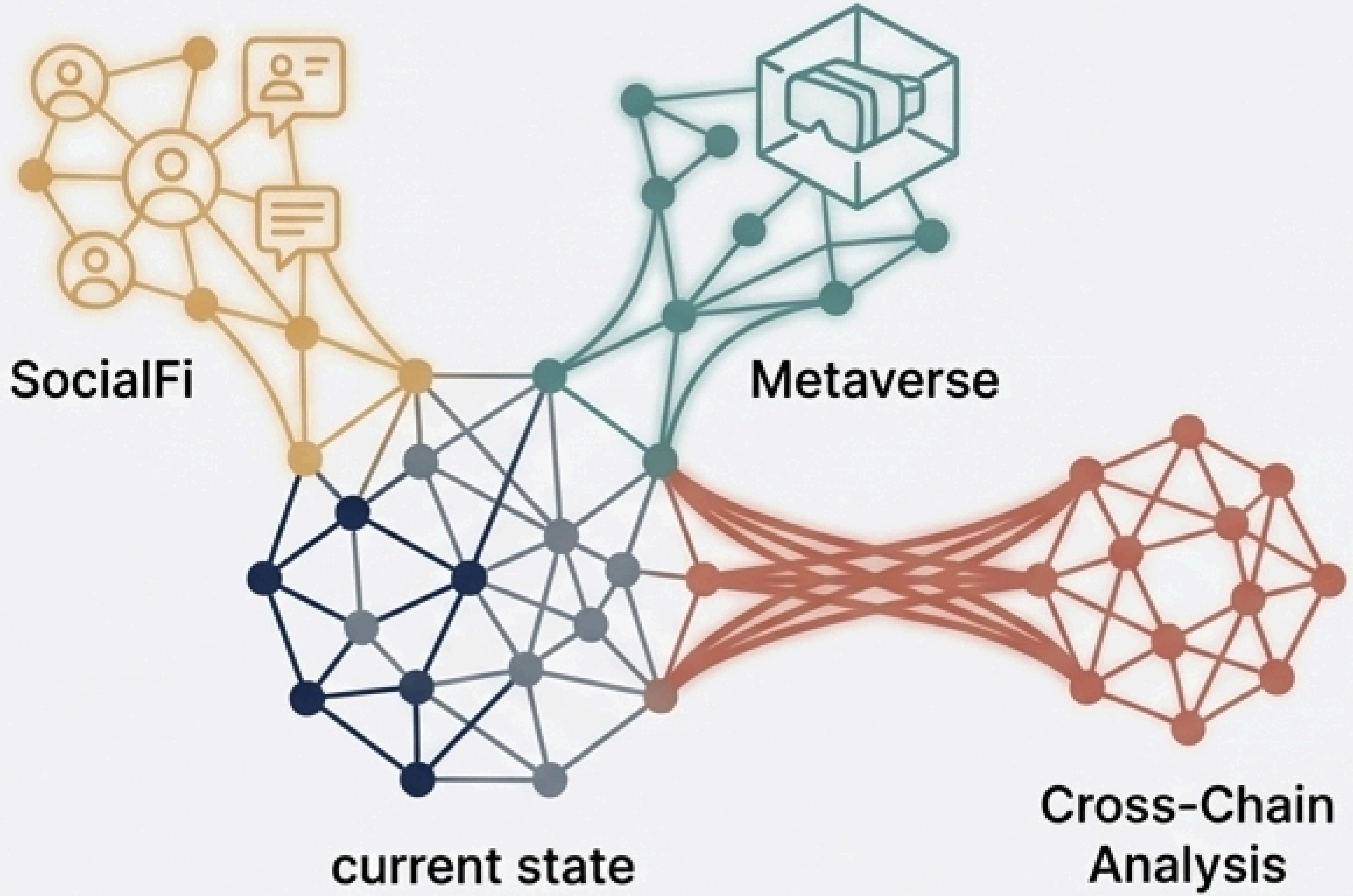
Price Forecasting:
Can on-chain network activity and topology help predict market movements?

Economic Analysis:
How is wealth distributed? What are the dominant trading strategies and market mechanisms?

The field is rapidly evolving, with new technologies and research questions emerging constantly.

What's Next:

- **Advanced Methods:** Increasing use of advanced neural network approaches, particularly **Graph Neural Networks (GNNs),** for tasks like fraud detection and user classification, though scaling them to entire blockchains remains a challenge.
- **New Ecosystems:** Analysis is expanding to novel on-chain phenomena like **SocialFi** (monetizing social content) and the integration of crypto-assets into the **Metaverse**.
- **Cross-Chain Analysis:** As value flows between different blockchains via bridges, analyzing these inter-chain networks will become critical for a holistic view of the ecosystem.



References

1. Natkamon Tovanich, Célestin Coquidé, Rémy Cazabet. **Cryptocurrency Network Analysis**. ⟨arXiv.2502.03411⟩
2. Natkamon Tovanich, Célestin Coquidé, Rémy Cazabet. **Decoding Decentralized Finance Transactions through Ego Network Motif Mining**. *The 13th International Conference on Complex Networks and their Applications*, Dec 2024, Istanbul, Türkiye. ⟨10.1007/978-3-031-82431-9_13⟩. ⟨arXiv:2408.12311⟩. ⟨hal-04727895⟩
3. Natkamon Tovanich, Rémy Cazabet. **Fingerprinting Bitcoin Entities Using Money Flow Representation Learning**. *Applied Network Science*, 2023, 8 (1), pp.1–22. ⟨10.1007/s41109-023-00591-2⟩. ⟨hal-04208864⟩
4. Célestin Coquidé, Rémy Cazabet, Natkamon Tovanich. **Inside Alameda Research: A Multi-Token Network Analysis**. *The 13th International Conference on Complex Networks and their Applications*, Dec 2024, Istanbul, Türkiye. ⟨10.1007/978-3-031-82431-9_17⟩. ⟨arXiv:2409.10949⟩. ⟨hal-04727905⟩
5. Célestin Coquidé, Rémy Cazabet, Natkamon Tovanich. **Analysis of Ego Multi-Token Transfer Networks: A Multilayer Approach**. *Applied Network Science*, 2025, 10 (1), pp.37. ⟨10.1007/s41109-025-00712-z⟩. ⟨hal-05225341⟩