

Spectral Clustering-Based Analysis of ERC-721 Blockchain Transactions

Piyush Anand

CS21B1019

Dept. Of Computer Science and Engineering

IIT Raichur

Under the guidance of
Dr. Priodyuti Pradhan

Goal

Identify groups /communities of users interacting within the ERC-721 blockchain dataset.

Blockchain: The Foundation of Decentralized Systems

- Distributed, immutable ledger ensuring trustless transactions
- Enables peer-to-peer value exchange without intermediaries
- Basis for advanced applications like NFTs and smart contracts

Feature	Traditional Systems	Distributed Ledger (Blockchain)
Trust Mechanism	Central authority	Cryptography, consensus
Record Immutability	Vulnerable to tampering	Tamper-evident, permanent
Transaction Model	Intermediated	Peer-to-peer, direct
Advanced Applications	Limited, manual contracts	Smart contracts, NFTs, automation
Transparency and Auditability	Restricted, siloed	Universal, real-time

Credits : Perplexity AI

Real-World Applications of Blockchain Technology

- Secure **financial transactions** (DeFi, cross-border payments)
- Transparent **supply chain** and asset tracking
- **Identity verification** and digital voting systems



Image Credits :AI

NFTs: Revolutionizing Digital Ownership

- Unique, indivisible digital assets stored on blockchain
- Enables verifiable ownership of art, music, and collectibles
- Powered by smart contracts on Ethereum & similar platforms



Image Credits :AI

ERC-721: The Standard Behind NFTs

- Ethereum-based standard for non-fungible tokens
- Supports tracking ownership and transfer of unique assets



Dataset Collection: Real NFT Transactions

Publicly available ERC-721 transaction data downloaded from X-Block ETH



Image Credits :Google

Data Pre-processing: Reducing the Data Size

blockNumber	timestamp	transactionHash	tokenAddress	from	to	fromIsContract	toIsContract	tokenId
1001165	1455424860	0x1d11b3ea55930x55b9a11c2e830x38150290c18c0x3d2068aeb969				0	0	200000000
1001165	1455424860	0x1d11b3ea55930x55b9a11c2e830x38150290c18c0x5d2c24efac49				0	1	0
1003181	1455460299	0x26a30aa663f60x55b9a11c2e830x38150290c18c0x3d2068aeb969				0	0	500000000
1003181	1455460299	0x26a30aa663f60x55b9a11c2e830x38150290c18c0x5d2c24efac49				0	1	0
1003393	1455463847	0xc0db923ac0c60x55b9a11c2e830x3d2068aeb9690xb51446cc4291				0	0	100000000
1003393	1455463847	0xc0db923ac0c60x55b9a11c2e830x3d2068aeb9690x5d2c24efac49				0	1	130000
1005878	1455505636	0x113de15964f80x55b9a11c2e830x38150290c18c0xc0cfc0969d0d				0	0	100000000
1005878	1455505636	0x113de15964f80x55b9a11c2e830x38150290c18c0x5d2c24efac49				0	1	0
1005946	1455506683	0x09581f238af3c0x55b9a11c2e830xc0cfc0969d0d0xa220568ace92				0	0	100000000
1005946	1455506683	0x09581f238af3c0x55b9a11c2e830xc0cfc0969d0d0x5d2c24efac49				0	1	130000
1006034	1455508426	0x637e4263f7820x55b9a11c2e830x38150290c18c0xa220568ace92				0	0	200000000
1006034	1455508426	0x637e4263f7820x55b9a11c2e830x38150290c18c0x5d2c24efac49				0	1	0
1006046	1455508554	0x5eac9cbd9cae0x55b9a11c2e830xc0cfc0969d0d0xa220568ace92				0	0	898000000
1006046	1455508554	0x5eac9cbd9cae0x55b9a11c2e830xc0cfc0969d0d0x5d2c24efac49				0	1	1167400
1006059	1455508793	0xc71da93bc6640x55b9a11c2e830x38150290c18c0xa220568ace92				0	0	400000000
1006059	1455508793	0xc71da93bc6640x55b9a11c2e830x38150290c18c0x5d2c24efac49				0	1	0
1007626	1455536210	0x6c6729b405470x55b9a11c2e830x38150290c18c0xad9173352524				0	0	200000000
1007626	1455536210	0x6c6729b405470x55b9a11c2e830x38150290c18c0x5d2c24efac49				0	1	0
1007645	1455536477	0x0a67fa2e15000x55b9a11c2e830x38150290c18c0xad9173352524				0	0	1800000000
1007645	1455536477	0x0a67fa2e15000x55b9a11c2e830x38150290c18c0x5d2c24efac49				0	1	0
1007784	1455539020	0x8d55d21c2e770x55b9a11c2e830x38150290c18c0xad9173352524				0	0	800000000

Timestamp Conversion: UNIX to Human Readable

blockNumber	timestamp	transactionHash	tokenAddress	from	to	fromIsContract	toIsContract	tokenId
4585326	2017-11-20 0:51:46	0xd467a51842ac6	0x441d1b228cad4	0x0000000000	0x12e3da7ef08	0	0	1
4585337	2017-11-20 0:53:12	0x44a00538df1c2	0x441d1b228cad4	0x0000000000	0x12e3da7ef08	0	0	2
4585342	2017-11-20 0:55:09	0x034a644c344ff9	0x441d1b228cad4	0x0000000000	0xba52c75764	0	0	3
4585351	2017-11-20 0:58:13	0x51103e6d153bf0	0x441d1b228cad4	0x0000000000	0xba52c75764	0	0	4
4585354	2017-11-20 0:59:11	0x6308acf866c01c	0x441d1b228cad4	0x0000000000	0xba52c75764	0	0	5
4585356	2017-11-20 0:59:18	0x29e6ae148db8f	0x441d1b228cad4	0x0000000000	0xba52c75764	0	0	6
4585356	2017-11-20 0:59:18	0x275350d21811b	0x441d1b228cad4	0x0000000000	0xba52c75764	0	0	7
4585357	2017-11-20 0:59:49	0xf5cd3789299ab	0x441d1b228cad4	0x0000000000	0xba52c75764	0	0	8
4585359	2017-11-20 1:00:34	0x5807656c322a4	0x441d1b228cad4	0x0000000000	0xba52c75764	0	0	9
4585363	2017-11-20 1:01:31	0xf2a40fcf1867ae	0x441d1b228cad4	0x0000000000	0xba52c75764	0	0	10
4585363	2017-11-20 1:01:31	0xea997d61cb21f	0x441d1b228cad4	0x0000000000	0xba52c75764	0	0	11
4585376	2017-11-20 1:04:56	0x91cd09d8ae62b	0x441d1b228cad4	0xba52c7576	0xe727e07021	0	1	3
4585398	2017-11-20 1:08:49	0x76c2656ef9836	0x441d1b228cad4	0xba52c7576	0xe727e07021	0	1	4
4587251	2017-11-20 8:13:21	0x6f9202d553ebd	0x59061b6f26bb4	0x92d8fce773	0xbba49956f19	0	0	100000000
4587571	2017-11-20 9:20:05	0xa829b36bf3c2e	0x59061b6f26bb4	0x92d8fce773	0xbba49956f19	0	0	108308399587
4588239	2017-11-20 11:54:39	0x94ace52e0424b	0x2ee13cbd30471	0x72855c26e	0x1d6ff33eb64	1	0	1E+21
4588843	2017-11-20 14:13:12	0x4af72fa908eb3	0x59061b6f26bb4	0xfbb1b73c4f	0x0bc8e22a82	0	1	23905000000
4588853	2017-11-20 14:17:15	0x6b630b1e0bbat	0x59061b6f26bb4	0x0bc8e22a8	0xfbb1b73c4f0	1	0	23905000000

Divide Daywise: File for every day

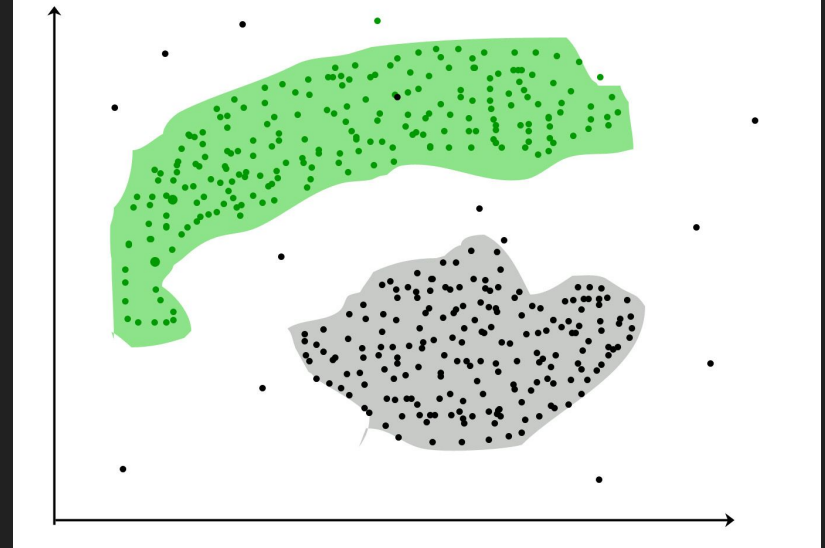
blockNumber	timestamp	transactionHash	tokenAddress	from	to	fromIsContract	toIsContract	tokenId
4585326	2017-11-20 0:51:46	0xd467a51842ac60x441d1b228cad4	0x441d1b228cad4	0x0000000000	0x12e3da7ef08	0	0	1
4585337	2017-11-20 0:53:12	0x44a00538df1c20x441d1b228cad4	0x441d1b228cad4	0x0000000000	0x12e3da7ef08	0	0	2
4585342	2017-11-20 0:55:09	0x034a644c344ff90x441d1b228cad4	0x441d1b228cad4	0x0000000000	0xba52c75764	0	0	3
4585351	2017-11-20 0:58:13	0x51103e6d153bf0x441d1b228cad4	0x441d1b228cad4	0x0000000000	0xba52c75764	0	0	4
4585354	2017-11-20 0:59:11	0x6308acf866c01c0x441d1b228cad4	0x441d1b228cad4	0x0000000000	0xba52c75764	0	0	5
4585356	2017-11-20 0:59:18	0x29e6ae148db8f0x441d1b228cad4	0x441d1b228cad4	0x0000000000	0xba52c75764	0	0	6
4585356	2017-11-20 0:59:18	0x275350d21811b0x441d1b228cad4	0x441d1b228cad4	0x0000000000	0xba52c75764	0	0	7
4585357	2017-11-20 0:59:49	0xf5cd3789299ab0x441d1b228cad4	0x441d1b228cad4	0x0000000000	0xba52c75764	0	0	8
4585359	2017-11-20 1:00:34	0x5807656c322a40x441d1b228cad4	0x441d1b228cad4	0x0000000000	0xba52c75764	0	0	9
4585363	2017-11-20 1:01:31	0xf2a40fcf1867aef0x441d1b228cad4	0x441d1b228cad4	0x0000000000	0xba52c75764	0	0	10
4585363	2017-11-20 1:01:31	0xea997d61cb21f0x441d1b228cad4	0x441d1b228cad4	0x0000000000	0xba52c75764	0	0	11
4585376	2017-11-20 1:04:56	0x91cd09d8ae62b0x441d1b228cad4	0x441d1b228cad4	0xba52c75764	0xe727e07021	0	1	3
4585398	2017-11-20 1:08:49	0x76c2656ef98360x441d1b228cad4	0x441d1b228cad4	0xba52c75764	0xe727e07021	0	1	4
4587251	2017-11-20 8:13:21	0x6f9202d553ebd0x59061b6f26bb4	0x59061b6f26bb4	0x92d8fce773	0xbba49956f19	0	0	100000000
4587571	2017-11-20 9:20:05	0xa829b36bf3c2ef0x59061b6f26bb4	0x59061b6f26bb4	0x92d8fce773	0xbba49956f19	0	0	108308399587
4588239	2017-11-20 11:54:39	0x94ace52e0424b0x2ee13cbd30471	0x2ee13cbd30471	0x72855c26e	0x1d6ff33eb64	1	0	1E+21
4588843	2017-11-20 14:13:12	0x4af72fa908eb3c0x59061b6f26bb4	0x59061b6f26bb4	0xfbb1b73c4f	0x0bc8e22a82	0	1	23905000000
4588853	2017-11-20 14:17:15	0x6b630b1e0bbat0x59061b6f26bb4	0x59061b6f26bb4	0x0bc8e22a8	0xfbb1b73c4f0	1	0	23905000000

Labelling: For fast computation

timestamp	fromLabel	toLabel	tokenAddressLabel	fromIsContract	toIsContract	tokenId
2017-11-20 1:04:56	766	767	45	0	1	3
2017-11-20 1:08:49	766	767	45	0	1	4
2017-11-20 8:13:21	706	768	24	0	0	100000000
2017-11-20 9:20:05	706	768	24	0	0	1083083995871
2017-11-20 11:54:39	661	769	42	1	0	1E+21
2017-11-20 14:13:12	99	763	24	0	1	23905000000
2017-11-20 14:17:15	763	99	24	1	0	23905000000
2017-11-20 14:53:50	99	770	24	0	1	405000000
2017-11-20 14:56:09	770	99	24	1	0	405000000
2017-11-20 14:57:32	99	770	24	0	1	977603790647
2017-11-20 14:58:54	770	99	24	1	0	977603790647
2017-11-20 18:40:29	766	767	45	0	1	5
2017-11-20 18:42:06	767	765	45	1	0	5
2017-11-20 22:58:40	661	771	42	1	0	1E+21

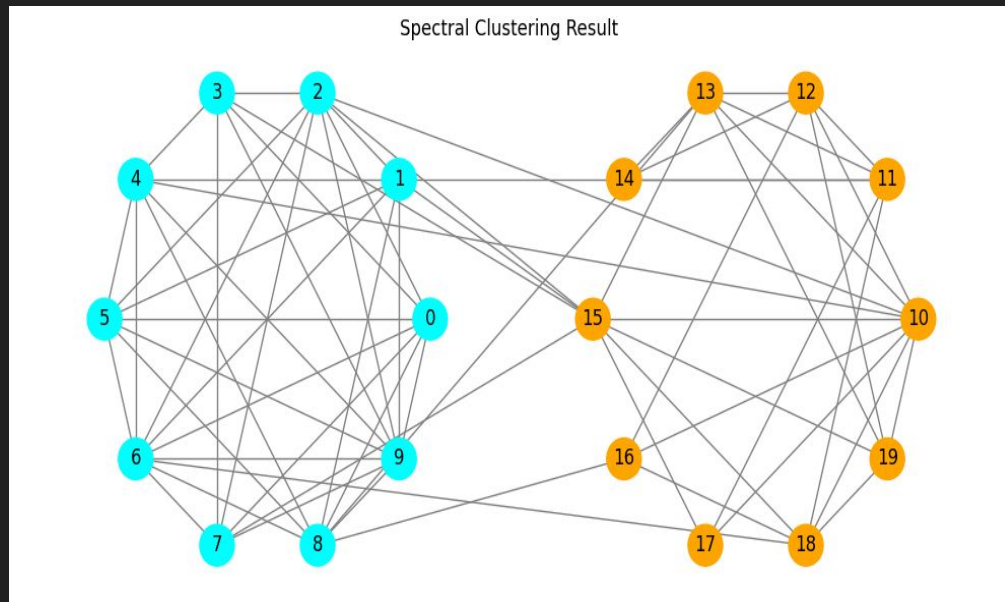
Why Clustering? Understanding Wallet Behaviors

- Group similar wallets to detect communities or fraud rings
- Reduces complexity of large-scale transaction graphs
- Useful in NFT trend analysis, bot detection, and monitoring
- **NULL Nodes** found out

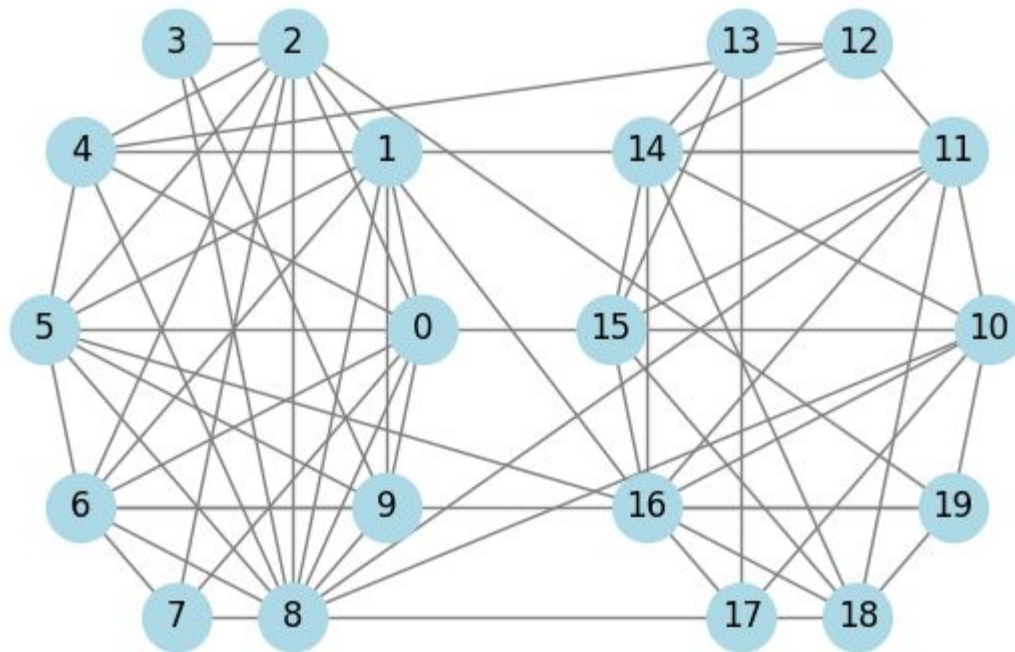


Spectral Clustering: Graph-based Group Discovery

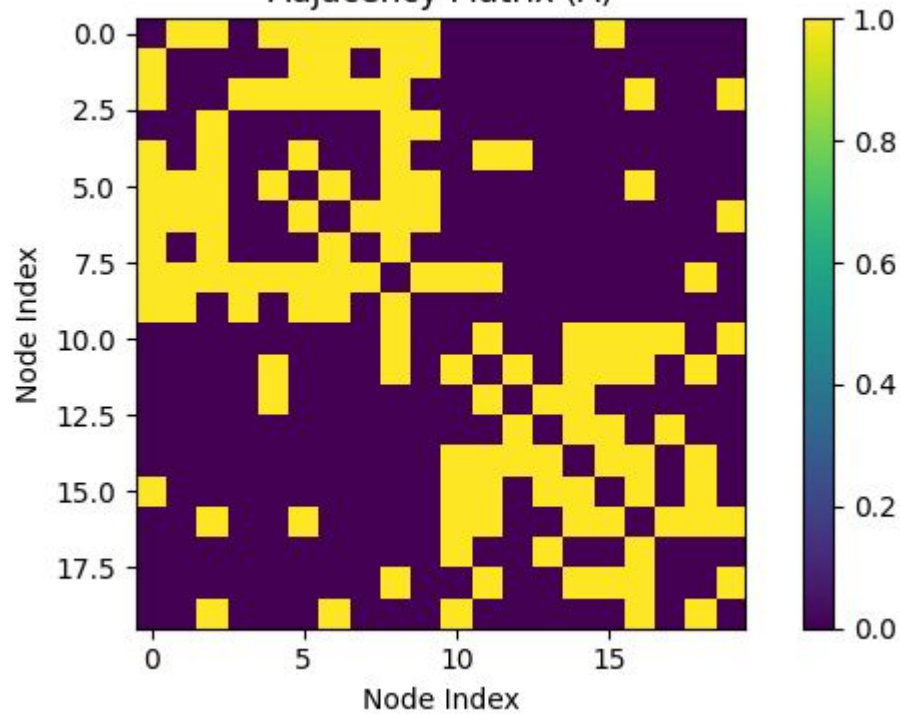
- Partition nodes into clusters based on graph structure
- Uses the Spectrum of Graph Laplacian Matrix
- Effective when clusters are not well separated.



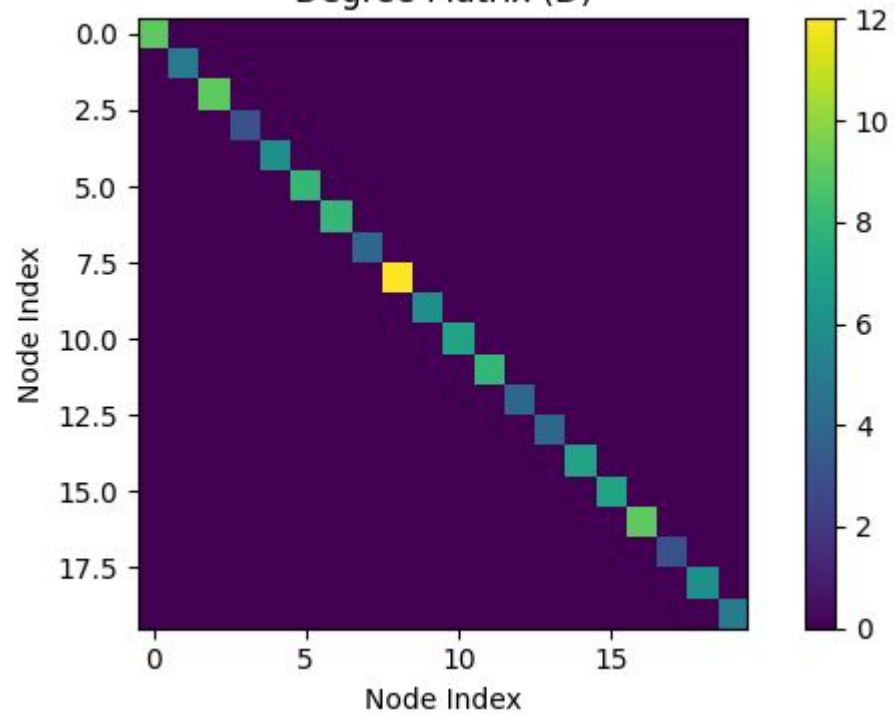
Initial Graph



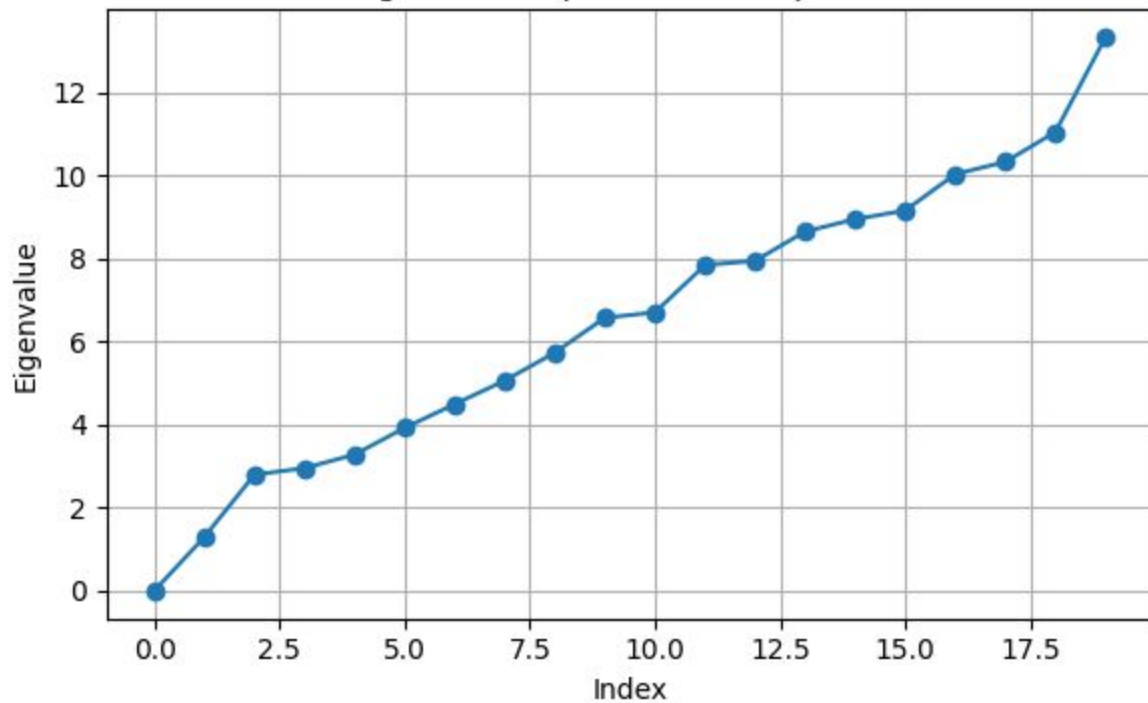
Adjacency Matrix (A)



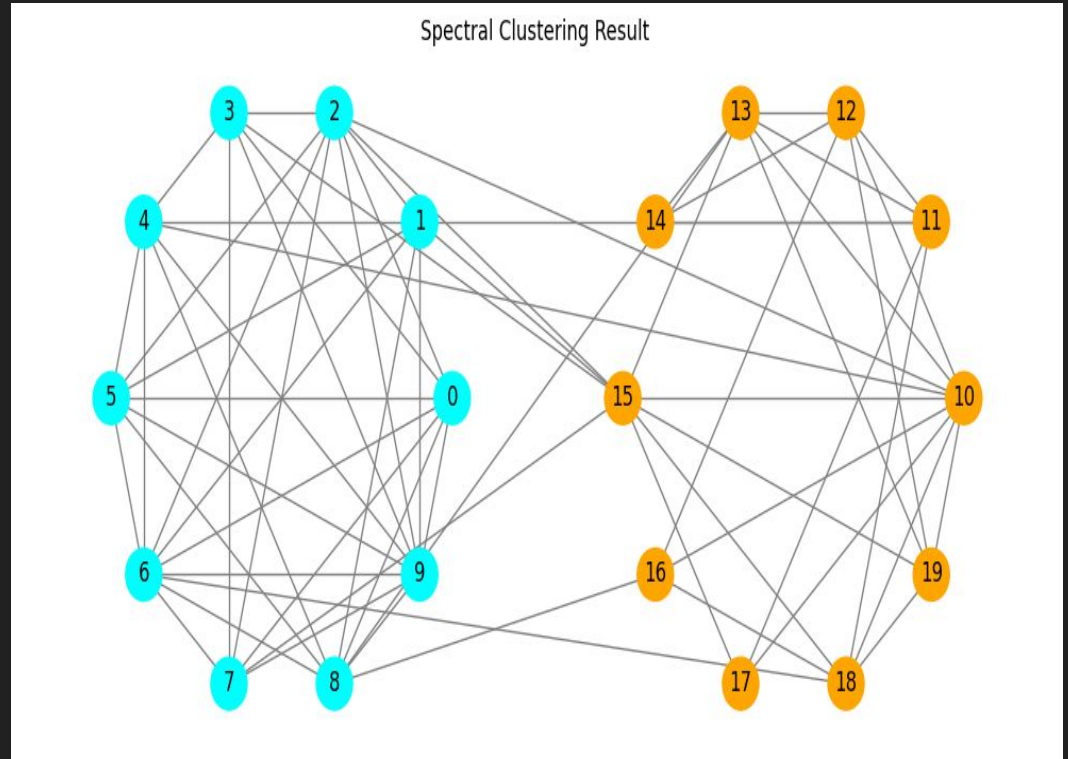
Degree Matrix (D)



Eigenvalue Spectrum of Laplacian



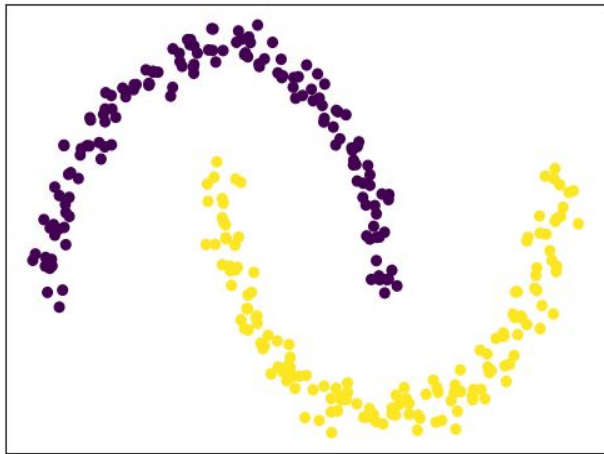
Clustered Graph



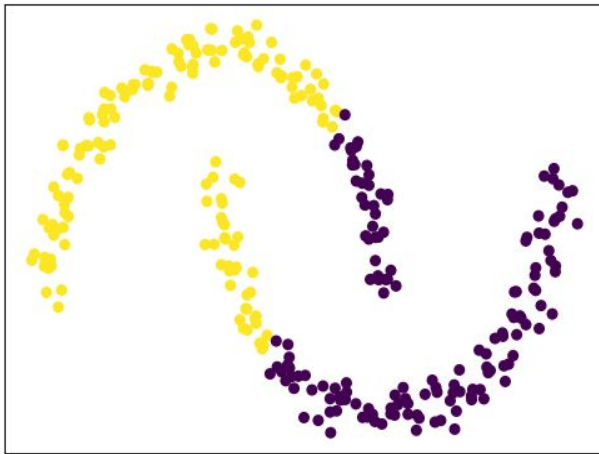
Spectral Clustering vs K-Means: Key Differences

- Spectral works on graph Laplacian; K-Means uses Euclidean distance
- Handles non-convex clusters better than K-Means
- Doesn't assume spherical clusters or equal sizes

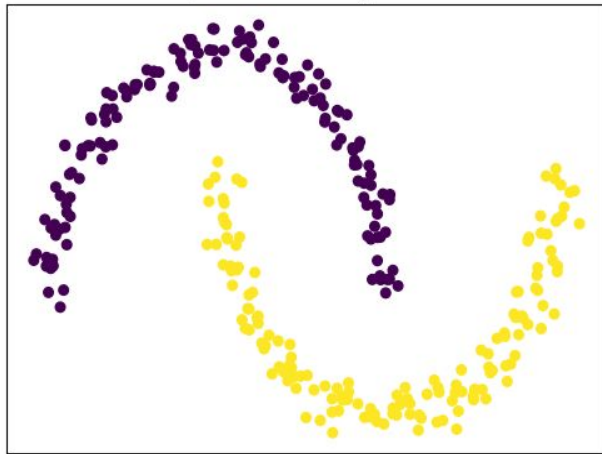
Original Labels



K-means on Raw Data



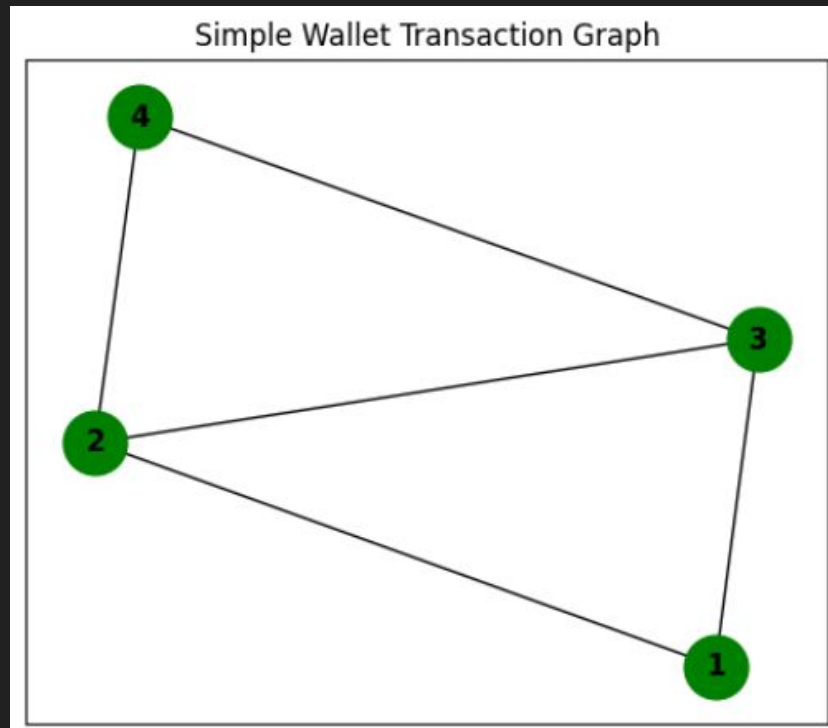
Spectral Clustering Result



Graph Construction: Wallets as Nodes, Transfers as Edges

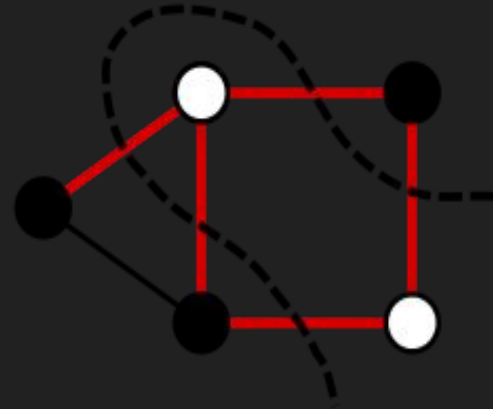
Created undirected graphs from token transfers

timestamp	fromLabel	toLabel	tokenAddressLabel	fromIsContract	toIsContract
2020-01-27 00:00:00	1	2	111	0	0
2020-01-27 00:01:00	2	3	112	0	0
2020-01-27 00:02:00	3	4	113	0	0
2020-01-27 00:03:00	1	3	111	0	0
2020-01-27 00:04:00	2	4	112	0	0
2020-01-27 00:05:00	3	1	113	0	0



Min-Cut Problem in Graphs

- The **Min-Cut problem** aims to split a graph into two sets such that the number (or weight) of edges between them is minimized.
- It is a classic **NP-hard problem**, widely used in clustering and graph partitioning tasks.
- Spectral methods approximate Max-Cut solutions by leveraging eigenvalues of the Laplacian matrix.



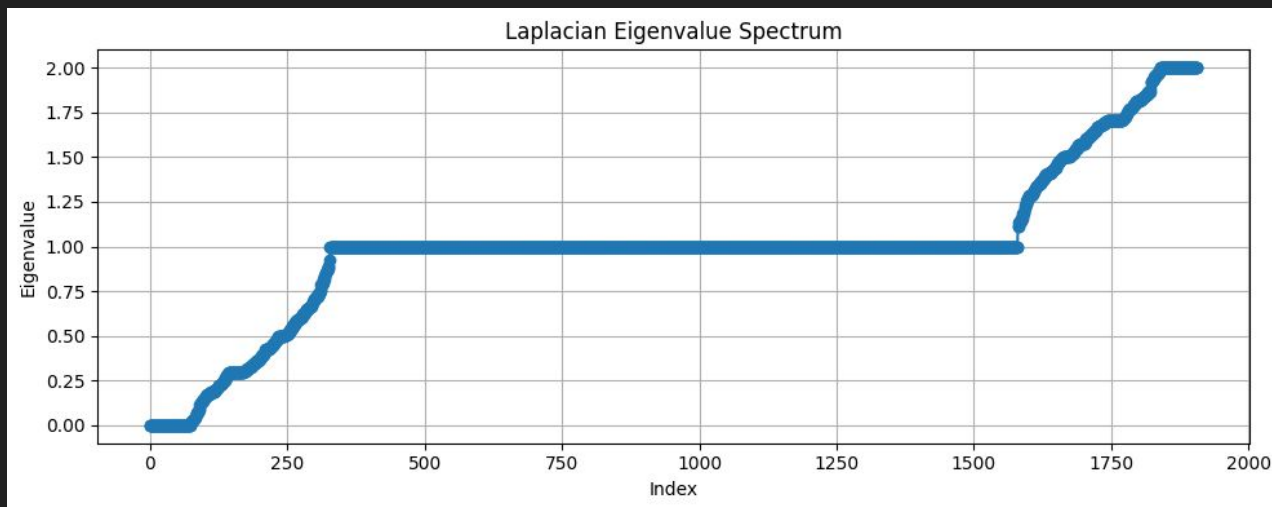
Spectral Clustering Results: Distinct Wallet Groups Found

- Clear clustering observed in low-dimensional spectral space
- High intra-cluster similarity and low inter-cluster connections
- Validated results using silhouette scores and modularity

Eg : 1st Jan, 2020

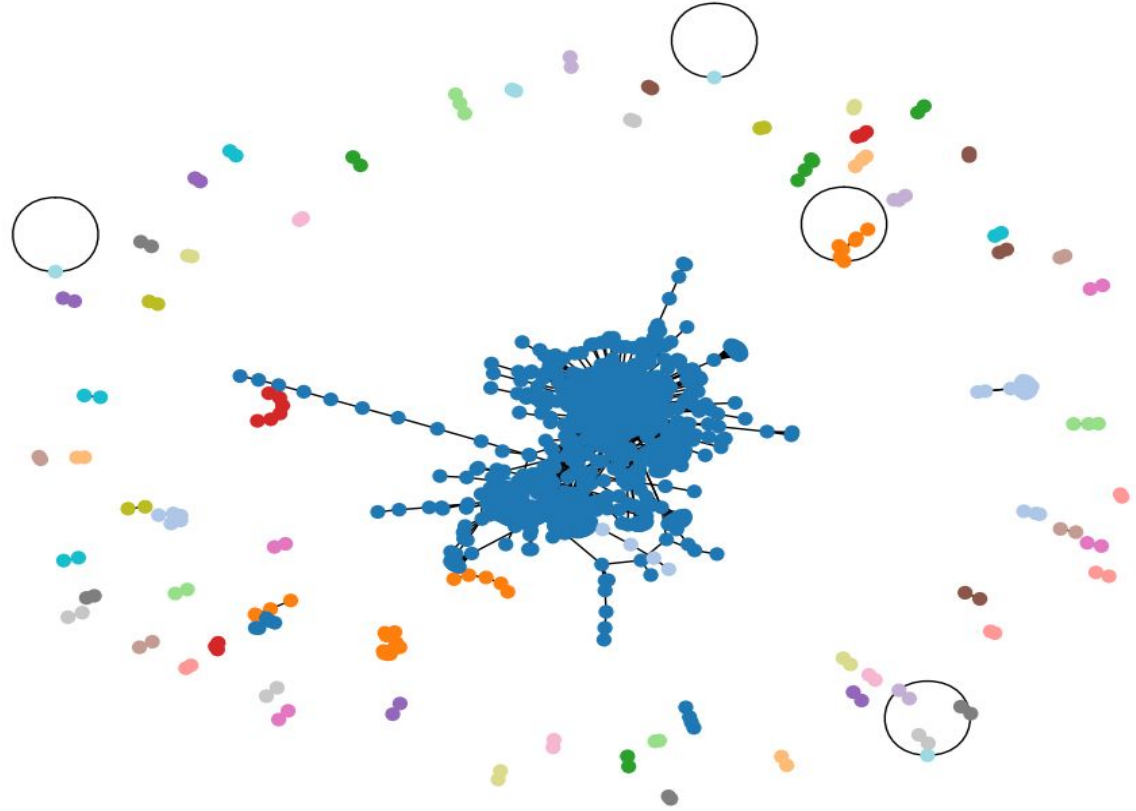
Total Transactions :
18904

Total Unique Wallets:
1904



Clustered Graph

Graph Colored by Spectral Clusters - 2020-01-01.csv



Clustering Result Summary of 1st Jan

Estimated Number of Clusters: 75

Cluster Size Distribution Highlights:

Largest Cluster (Cluster 0): 1683 nodes

Small Clusters (Size ≤ 3): Majority

Clusters with Size ≥ 5 :

- Cluster 6 \rightarrow 22 nodes
- Cluster 1 \rightarrow 16 nodes
- Cluster 10 \rightarrow 11 nodes
- Cluster 7, 11 \rightarrow 7 nodes each
- Cluster 17 \rightarrow 6 nodes

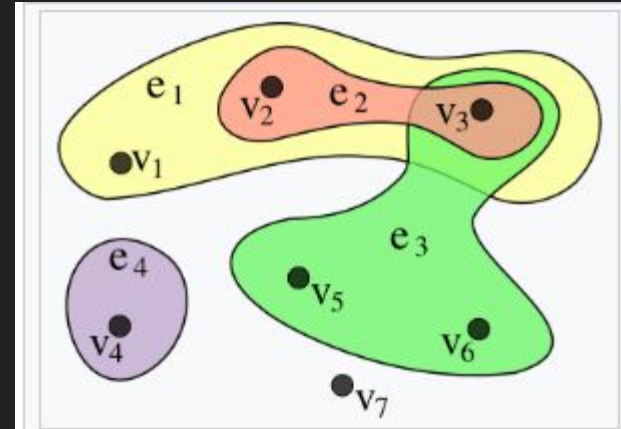
Cluster 4, 8, 2 \rightarrow 5 nodes each

Observation:

The clustering is highly imbalanced, with one dominant cluster and many tiny clusters, indicating potential community concentration or sparsity in some areas of the graph.

Hypergraphs

A hypergraph is a generalization of an ordinary graph in which an edge can connect more than two vertices together



An example of an undirected hypergraph, with

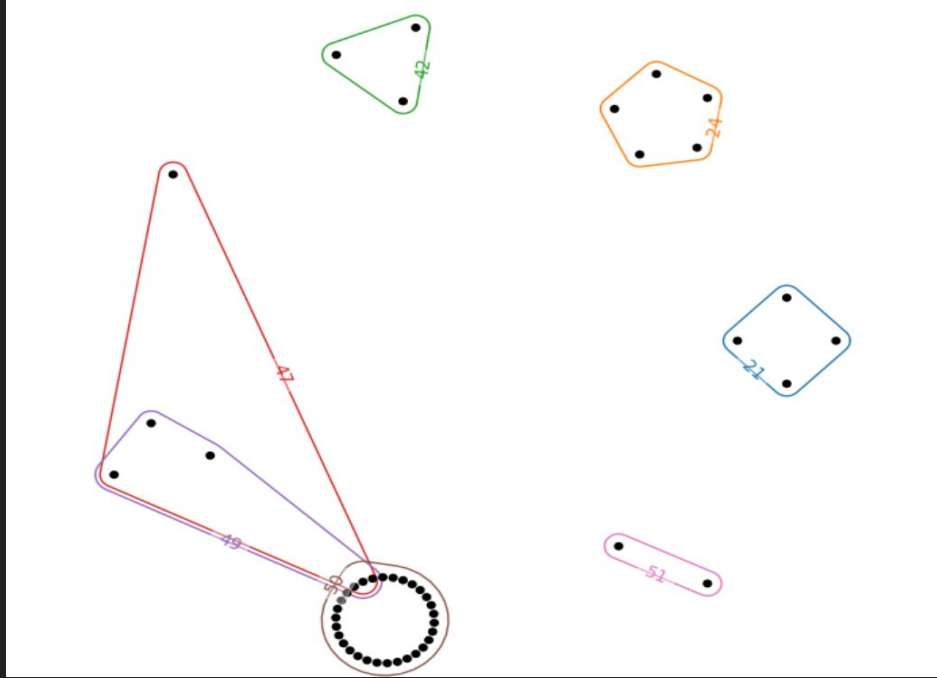
$X = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$
and $E = \{e_1, e_2, e_3, e_4\} =$
 $\{\{v_1, v_2, v_3\}, \{v_2, v_3\},$
 $\{v_3, v_5, v_6\}, \{v_4\}\}$. This

Hypergraph Vs Normal Graph

Aspect	Normal Graphs	Hypergraphs
Edge Connectivity	Two vertices per edge	Any number of vertices per hyperedge
Group Interaction Modeling	Requires many pairwise edges	Captures the entire group with one hyperedge
Representation Clarity	Can become cluttered in group settings	More natural and compact for group data
Example Scenario	Friendships, pairwise collaborations	Research groups, social chat groups

Hypergraphs: Capturing Multi-Wallet Token Interactions

- Nodes: wallets
- Hyperedges: token contracts
- Encodes many-to-many relationships missed by simple graphs
- Better suited for understanding NFT community structures



20th November, 2017

Hypergraph Clustering

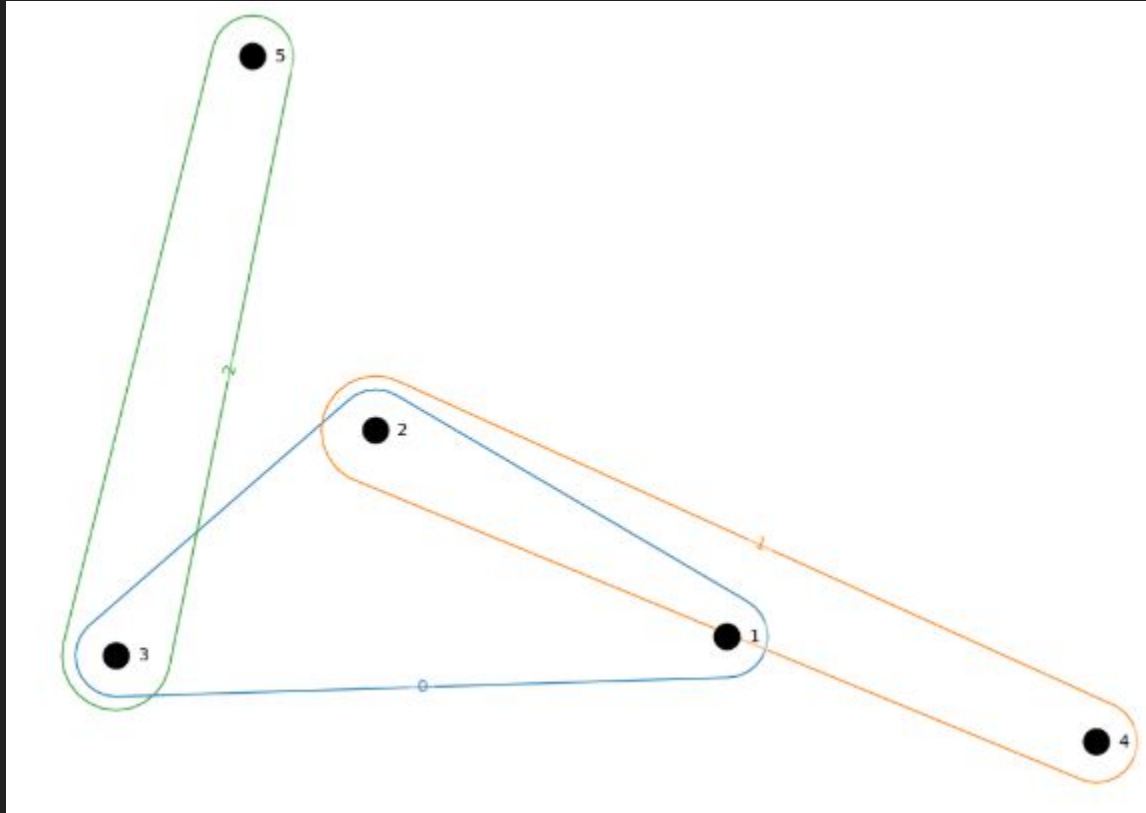
- Matrix size scales with number of hyper edges \times vertices
- Eigen decomposition on incidence matrix more computationally expensive
- **Less distinct clusters** observed due to overlap in token usage

Laplacian Of Hypergraph

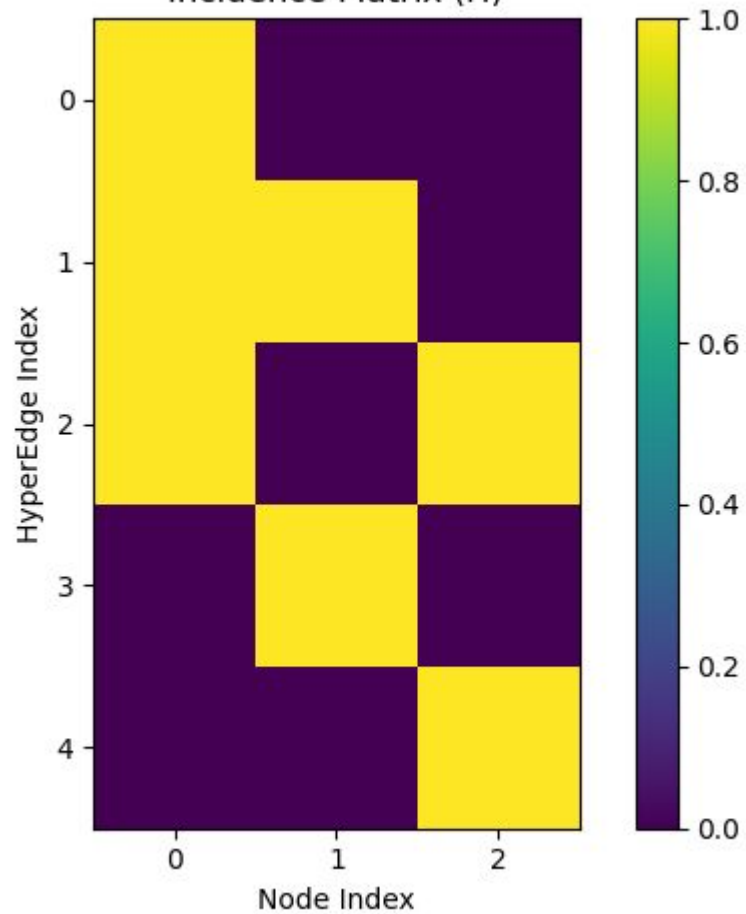
- D_v : Degree matrix of vertices
- D_e : Degree matrix of hyperedges
- H : Incidence matrix
- W : Weight matrix

$$\Delta = D_v - H W D_e^{-1} H^T.$$

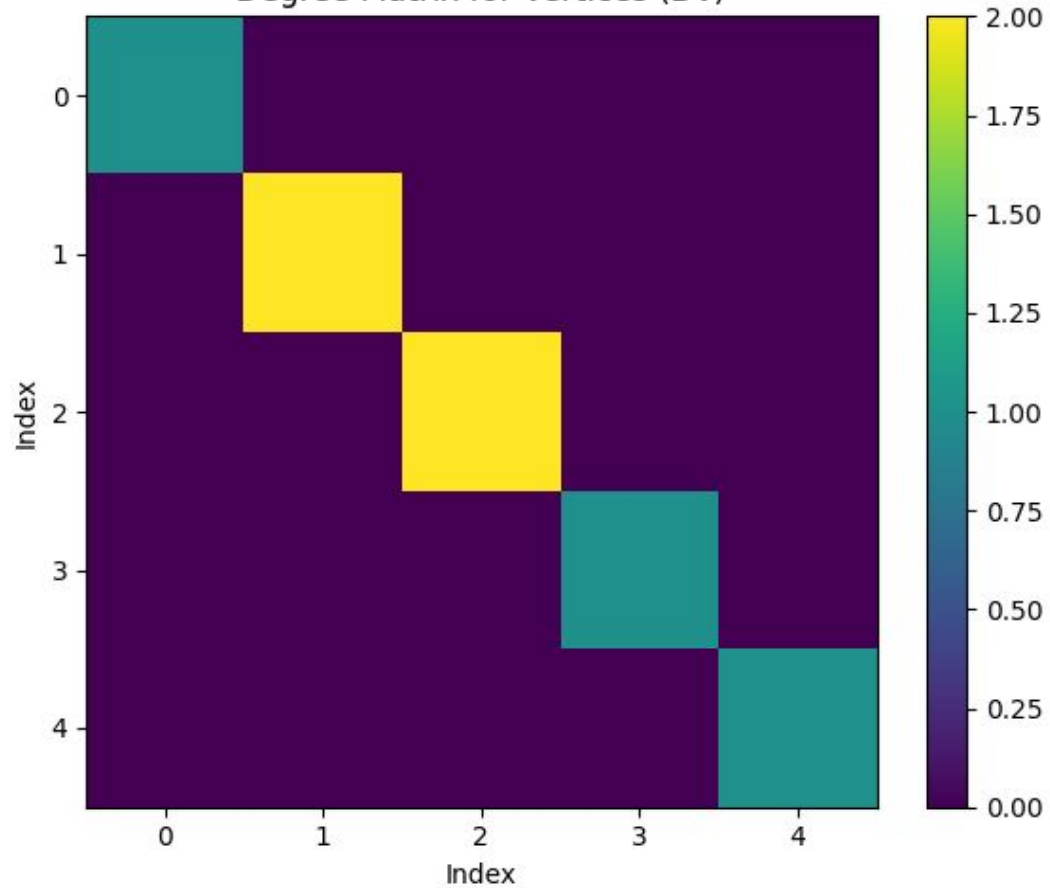
Example : Hypergraph

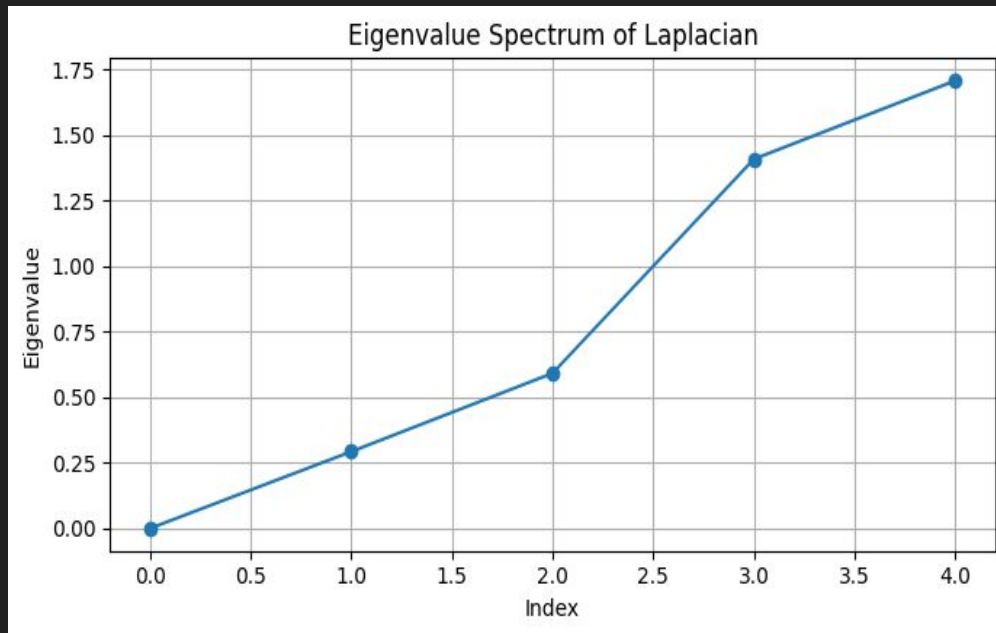


Incidence Matrix (H)



Degree Matrix for Vertices (Dv)





Cluster 1: [1]

Cluster 2: [2, 4]

Cluster 0: [3, 5]

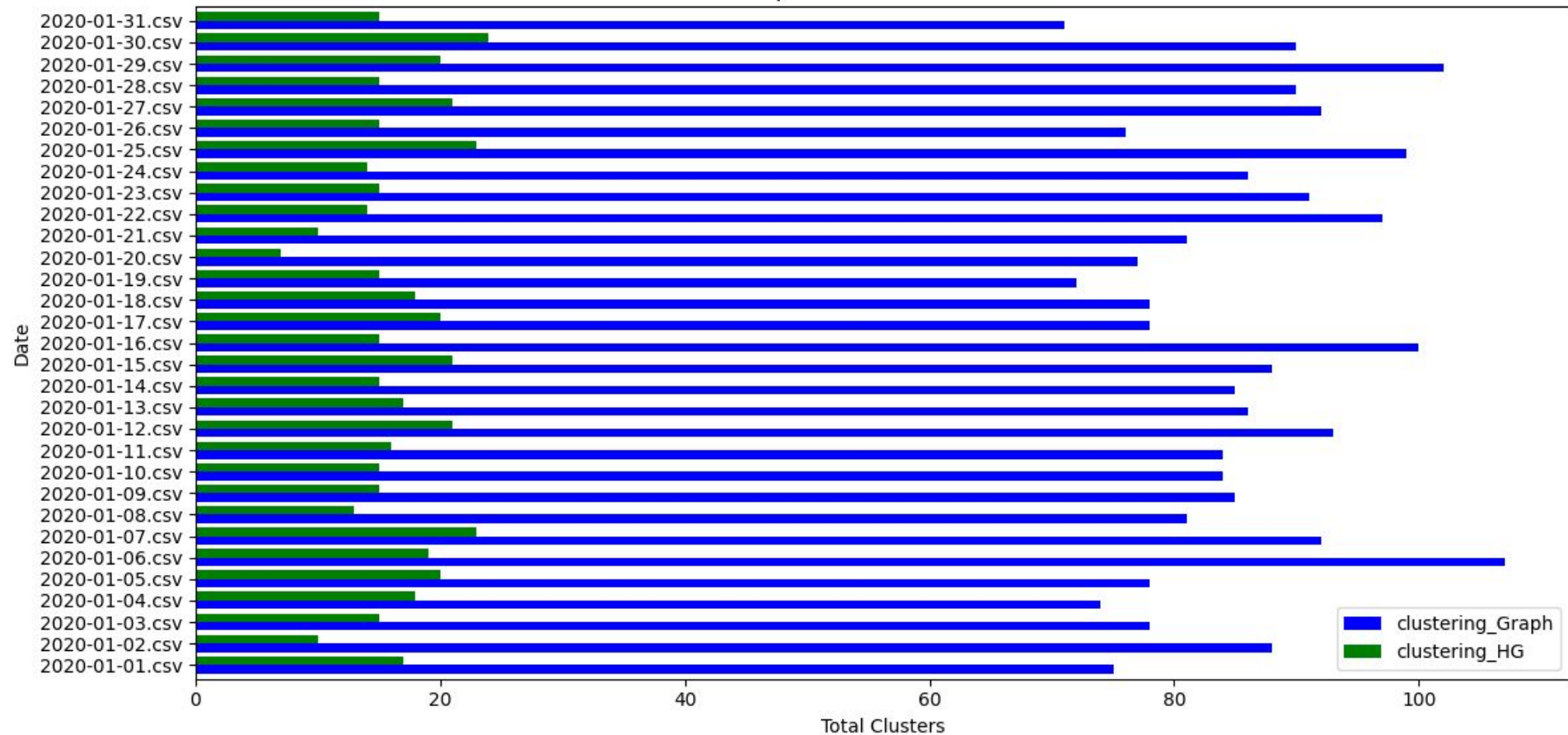
Laplacian Matrix (L):

```
[[ 0.66666667 -0.33333333 -0.33333333  0.      0.      ]
 [-0.33333333  1.16666667 -0.33333333 -0.5     0.      ]
 [-0.33333333 -0.33333333  1.16666667  0.     -0.5     ]
 [ 0.         -0.5        0.         0.5     0.      ]
 [ 0.          0.        -0.5        0.      0.5     ]]
```

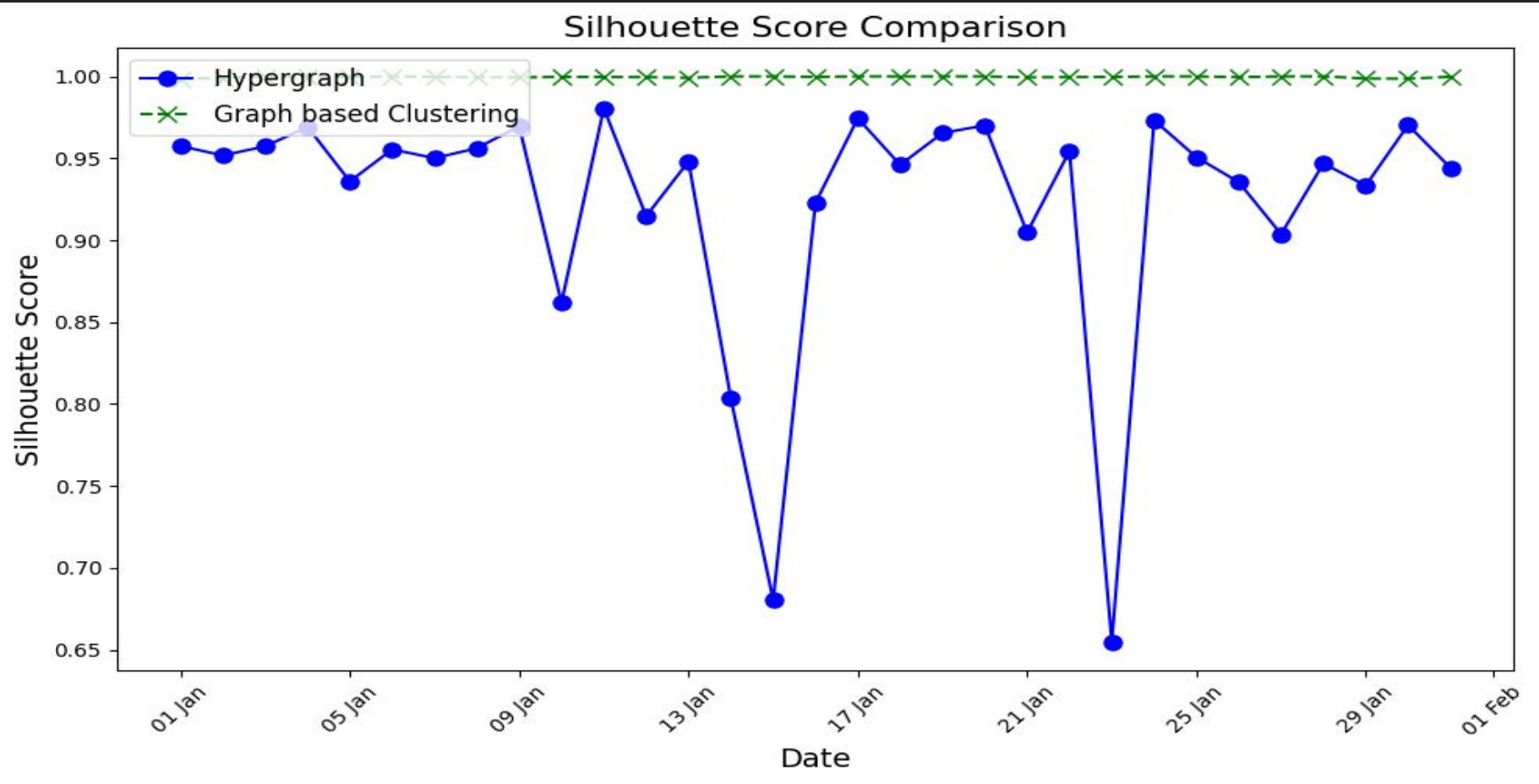
Results



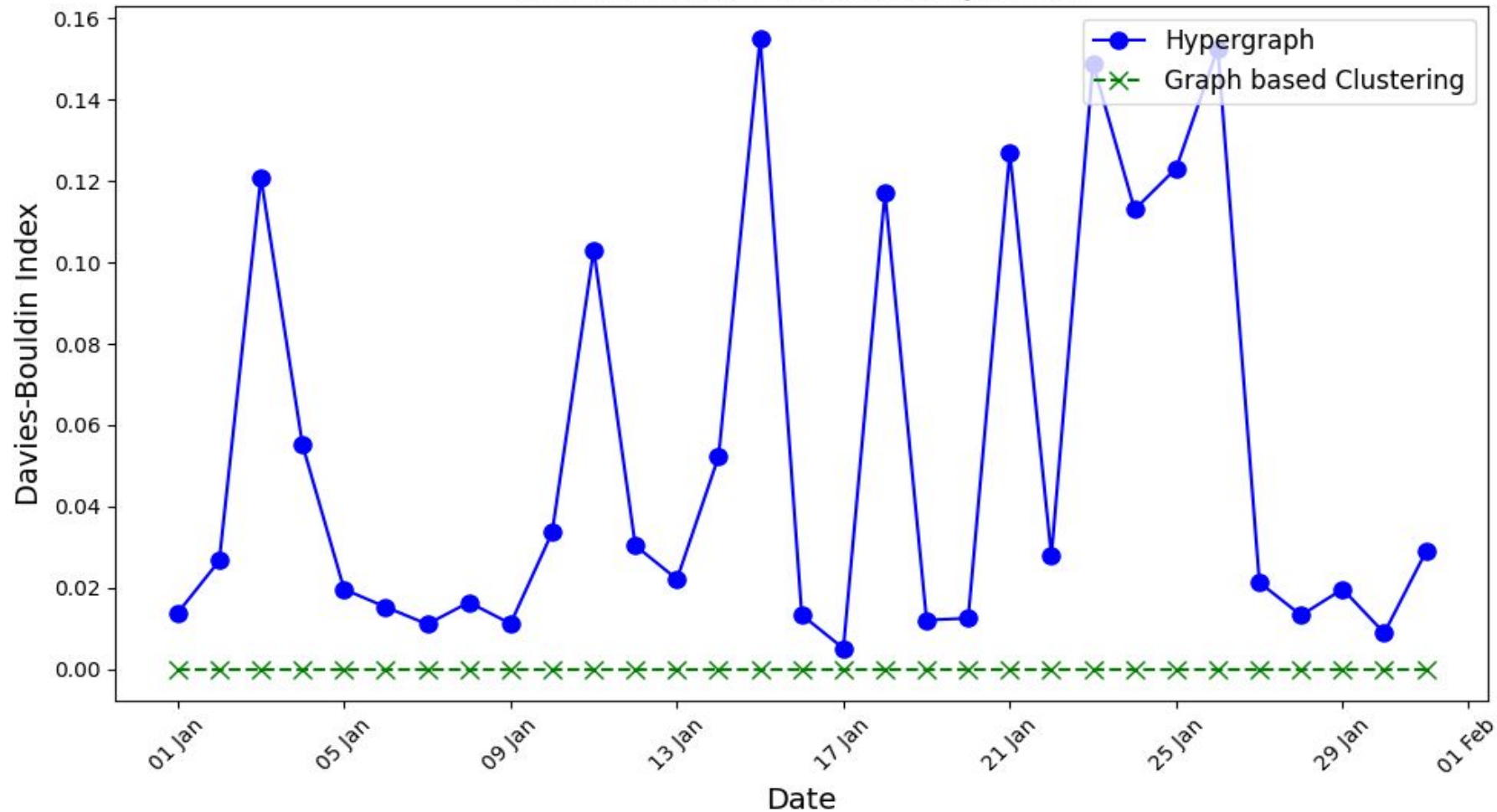
Comparison of Total Clusters



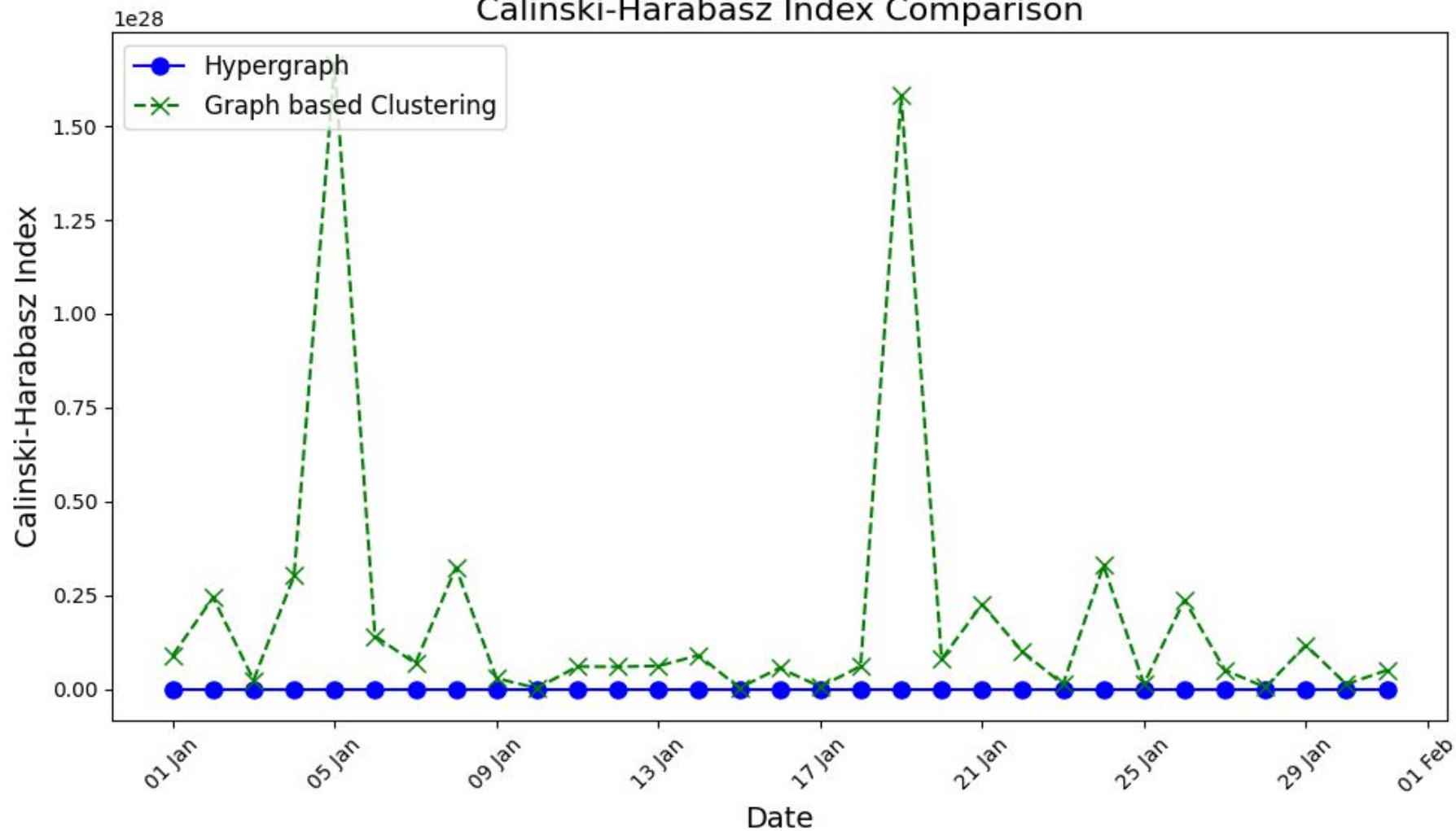
Comparison of Graph vs Hypergraph Clustering



Davies-Bouldin Index Comparison



Calinski-Harabasz Index Comparison



Observations Across Graph Types (Normal & Hypergraph)

In both normal and hypergraph clustering:

- One very large cluster dominates
- Many other clusters are extremely small (often just 2–3 wallets)

What This Means !!

- The big cluster likely represents:
 - A large number of wallets interacting with popular NFT contracts
 - Centralized activity around trending collections or marketplaces
 - Possibly an NFT drop, airdrop, or active trading on major platforms

The small clusters reflect:

- Individual trades
- Niche or low-activity wallets
- Potentially bot wallets, private sales, or side projects

Clustering Observations :

- Hypergraphs have multi-node relationships (hyperedges), making clustering more complex.
- Graphs rely on pairwise connections, forming denser and clearer clusters.
- Hyperedges create sparse, higher-order connections, leading to fewer clusters.
- Clustering in hypergraphs is more challenging due to intricate node interactions.

Future Scope: Scaling and Real-time Analytics

- Fraud Detection Using ML Models: Integrate supervised learning on top of clustering to identify anomalous patterns or flag suspicious wallets.
- Time-based Dynamic Clustering: Perform clustering across different time windows to detect evolving communities or changing behavior over time.
- Visual Dashboard: Build a real-time visualization tool to track clusters, token flows, and interactions between wallet communities.

Thank You!

Hypergraph Generation Methods

- Distance-Based Hypergraph Generation
 - Representation-Based Hypergraph Generation
 - Attribute-Based Hypergraph Generation
 - Network-Based Hypergraph Generation
- Graph: Faster, better-defined clusters, scalable
 - **Hypergraph**: Richer data, but high computation and less clean separation
 - Visual and statistical comparison shows spectral graph wins here

Distance-Based Hypergraph Generation

- Nodes grouped based on spatial distance.
- Example: Clustering similar points.
- Pros: Simple, effective for spatial data.
- Cons: Requires a threshold, sensitive to distance metric.

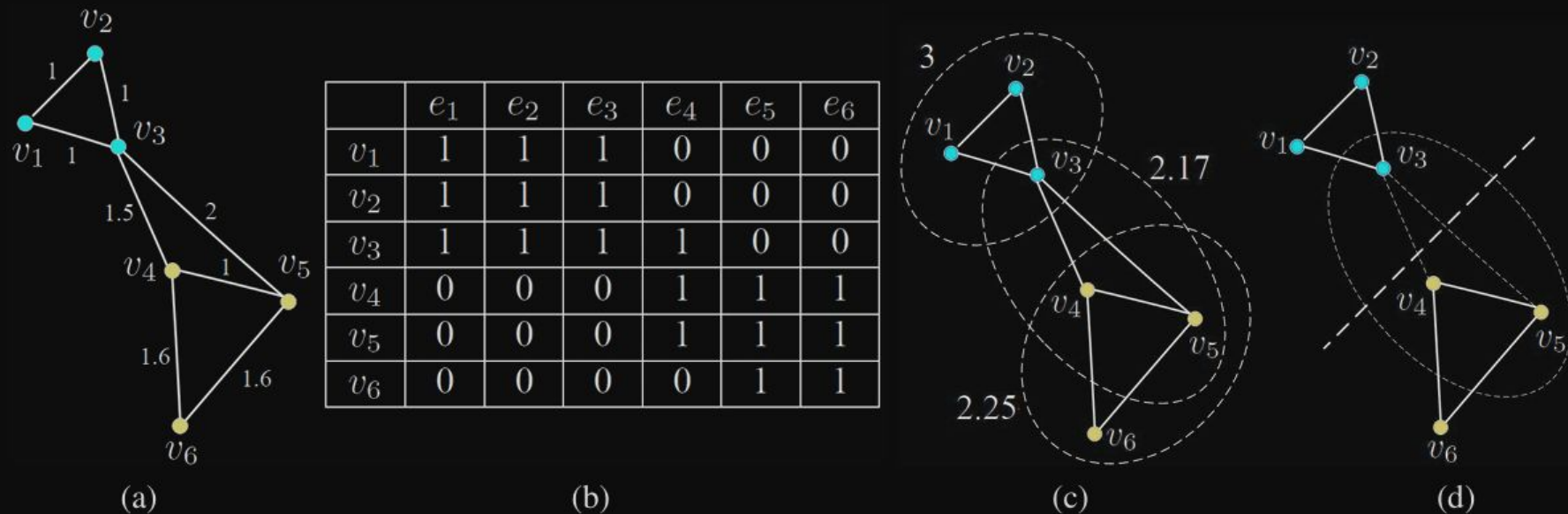


Figure 2. (a): A simple graph representing the interrelationships among six points in the 2-D space. Pairwise distances between every vertex and its neighbors are marked on the corresponding edges. (b) The H matrix representing the relationship between the vertices set and the hyperedge sets. The entry (v_i, e_j) is set to 1 if e_j contains v_i , or 0 otherwise. (c): The hypergraph which illustrates the complete relationships in the matrix H . The hyperedge weight is defined as the sum of reciprocals of all the pairwise distances in a hyperedge. (d) A hypergraph partition which is made on e_4 .

Representation-Based Hypergraph Generation

- Constructs hyperedges based on **shared representations**.
- Example: Word embeddings in NLP.
- Pros: Works well with ML models.
- Cons: Needs a pre-trained model, computationally expensive.

How Hyperedges Are Formed?

Compute vector representations of words using an embedding model. Group words that are close in the embedding space into the same hyperedge.

Words with similar meanings (e.g., king, queen, monarch) will belong to the same hyperedge

Attribute-Based Hypergraph Generation

- Groups nodes based on shared attributes.
- Example: Users with common interests in social networks.
- Pros: Intuitive, good for categorical data.
- Cons: Attribute selection matters.

Network-Based Hypergraph Generation

- Extracts hyperedges from existing networks.
- Example: Overlapping groups in citation networks.
- Pros: Captures complex relationships.
- Cons: Requires pre-existing network structures, computationally heavy.

Comparison of Methods

- **Distance-Based**: Simple, effective for spatial data, but threshold-sensitive.
- **Representation-Based**: Uses ML models, needs pre-trained embeddings.
- **Attribute-Based**: Good for categorical data, requires careful attribute selection.
- **Network-Based**: Captures real-world structures, but computationally expensive.