Vector Length Analysis on Real Data + PCA + Arnon's Work

Distributed Monitoring

Table Of Contents

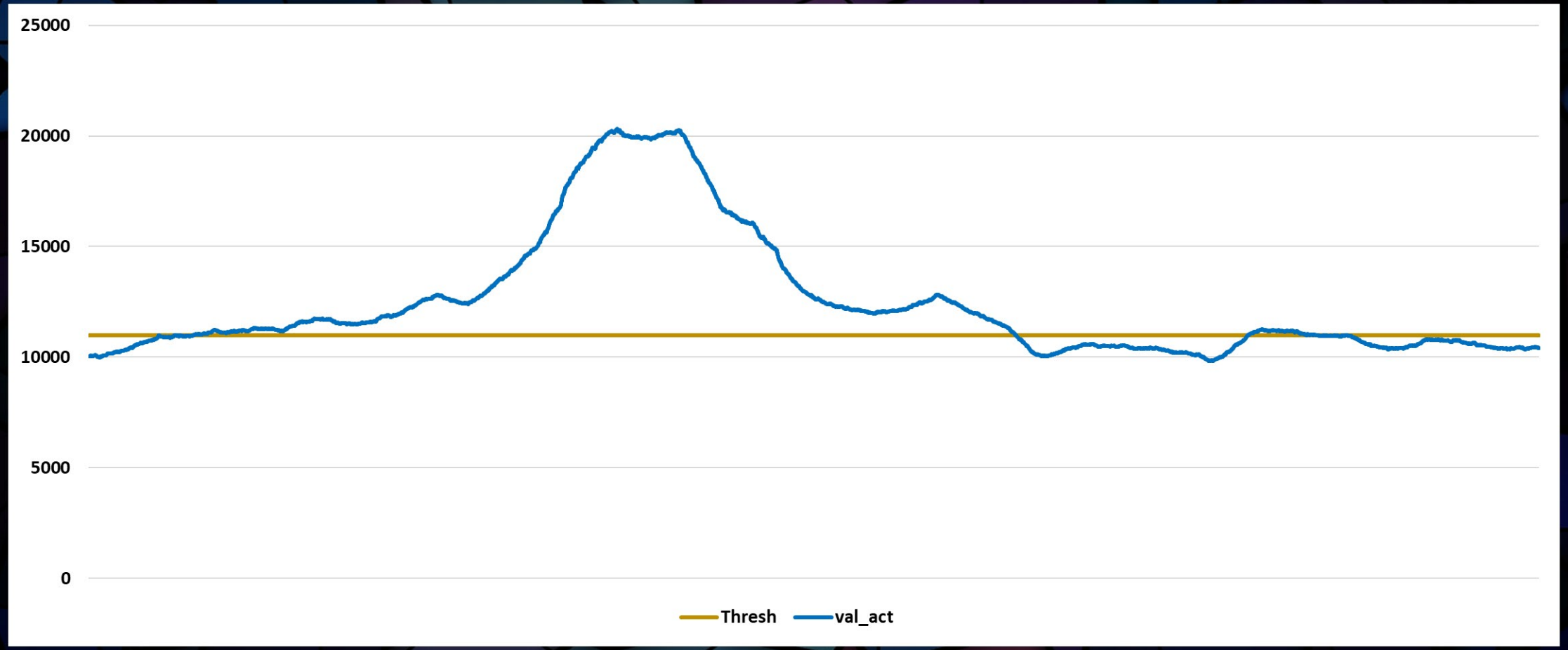
- ❑ Comparison to Arnon's results
- ❑ Bag of Words – Inner Product
- ❑ Bag of Words – Entropy
- ❑ PCA results

Comparison to Arnon's Work

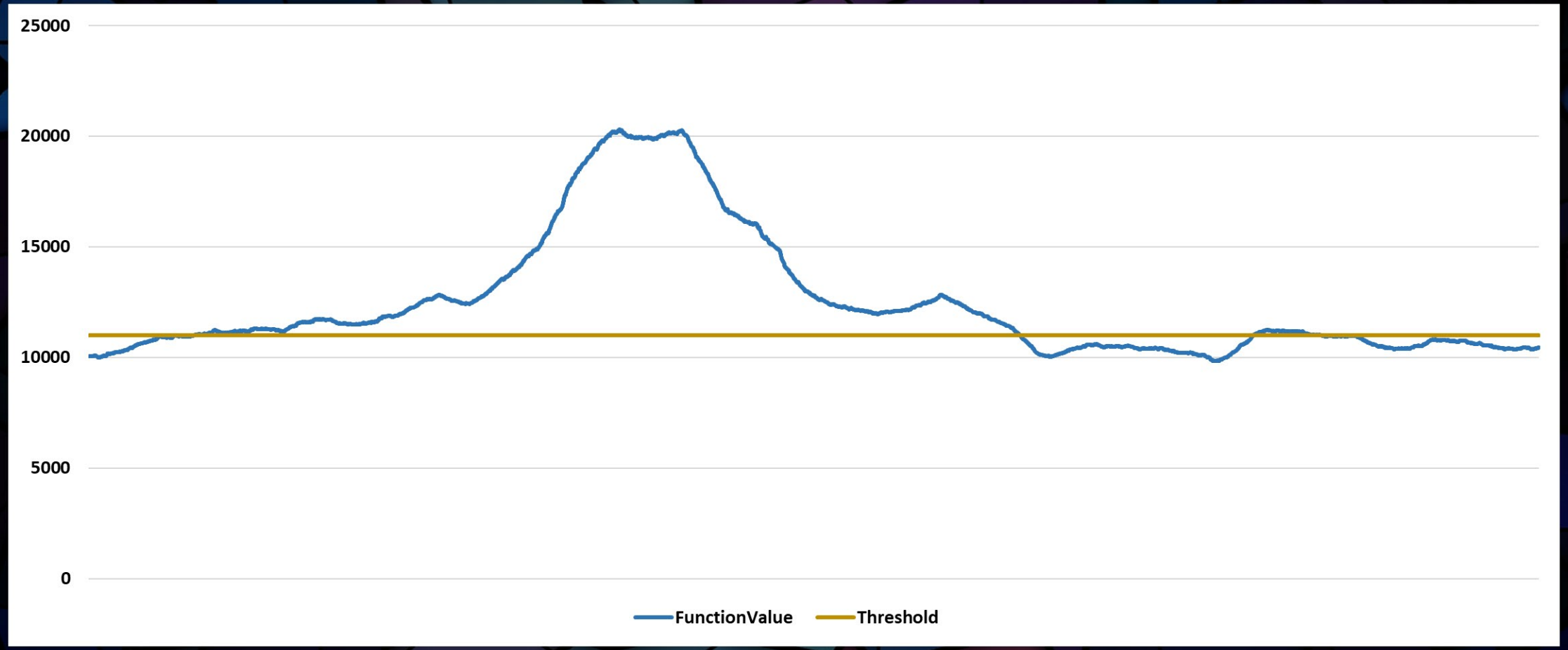
Tweeter Data Set

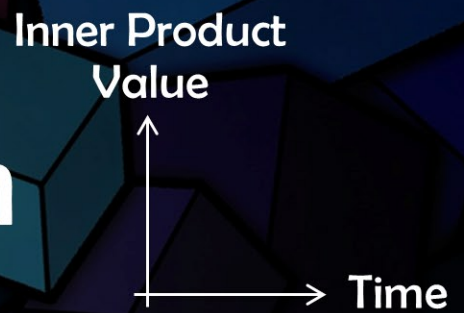
- 10 nodes – Round Robin
- Bag of words vector - |vector| = 1254 · 2 (X·Y)
- Constant threshold – 11,000
- Global BOW is the AVERAGE of BOWs
- Window size = 1,000
- Step – one tweet

Arnon's Data

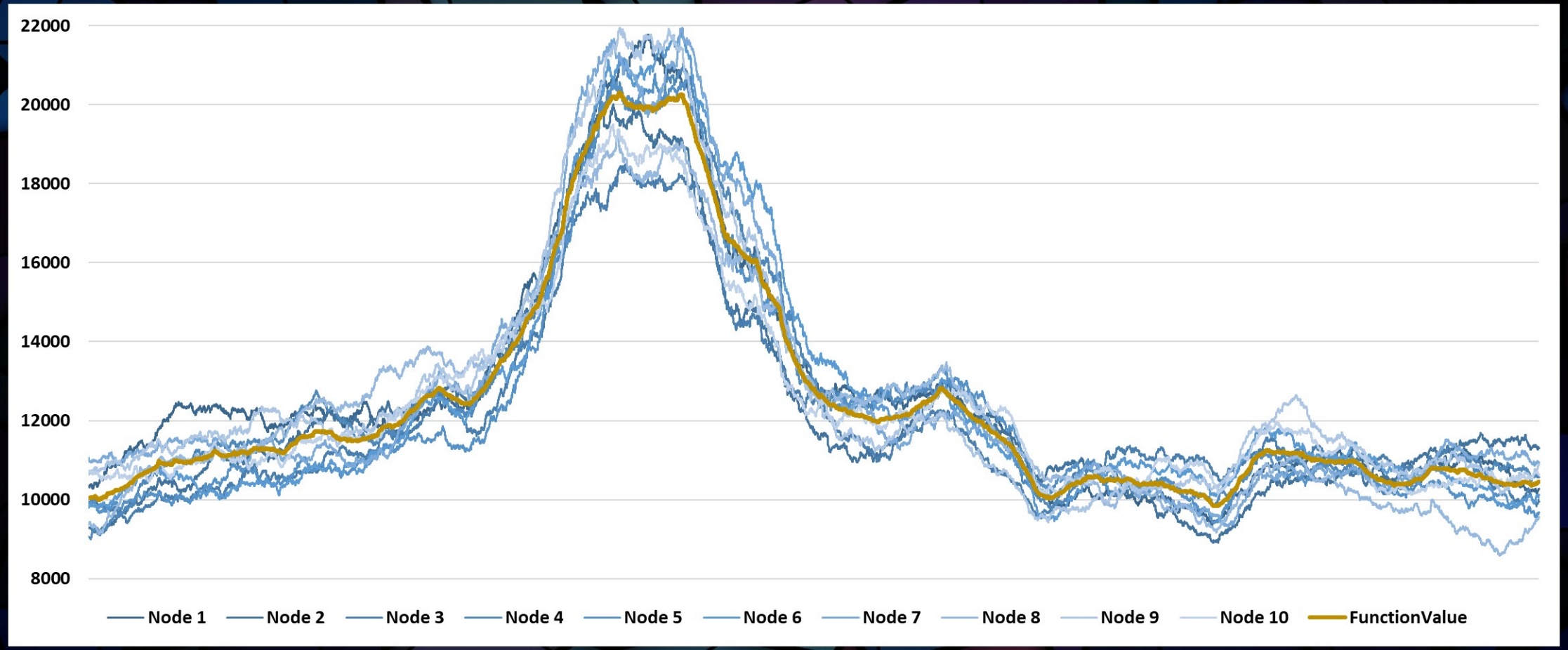

Inner Product Value
Time
25000
20000
15000
10000
5000
0
Thresh
val_act

My Reconstructed Function

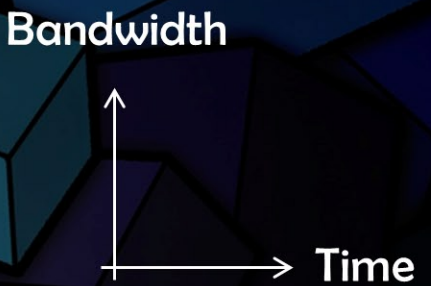


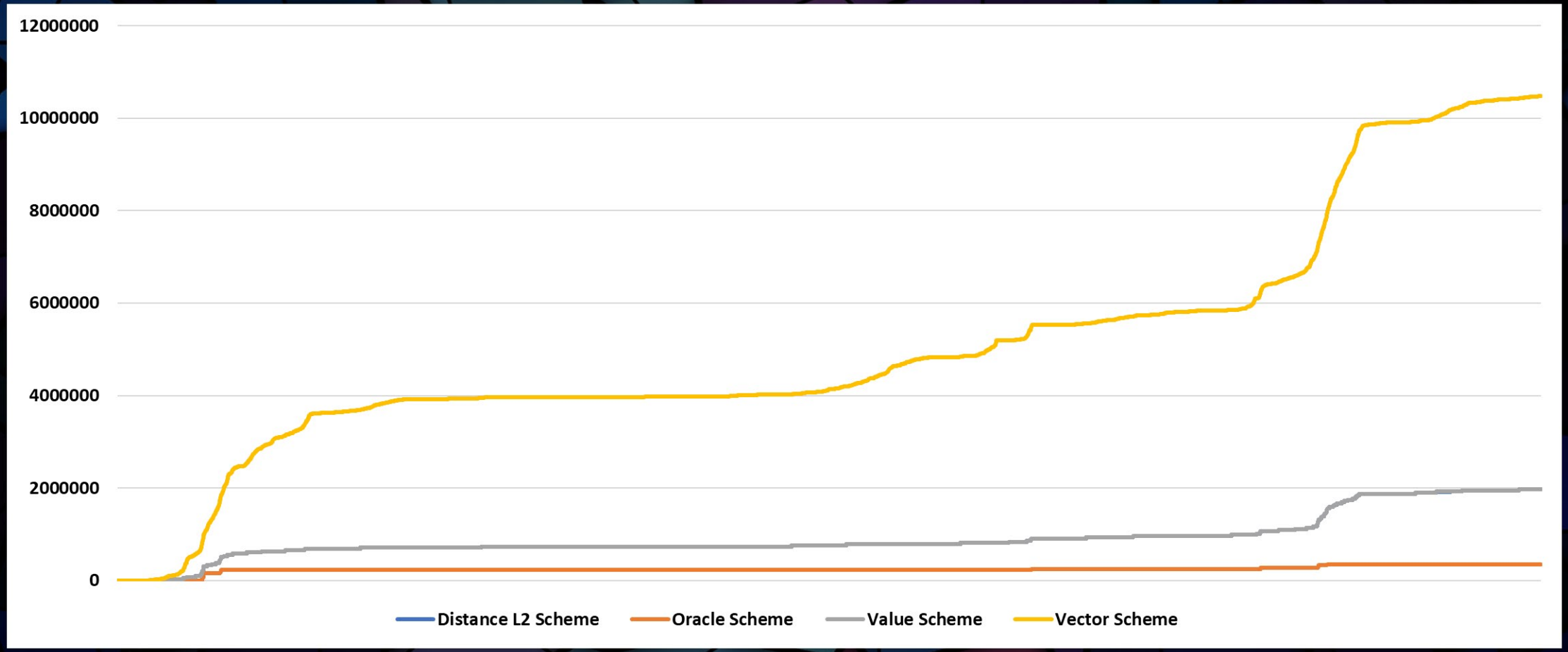


DATA

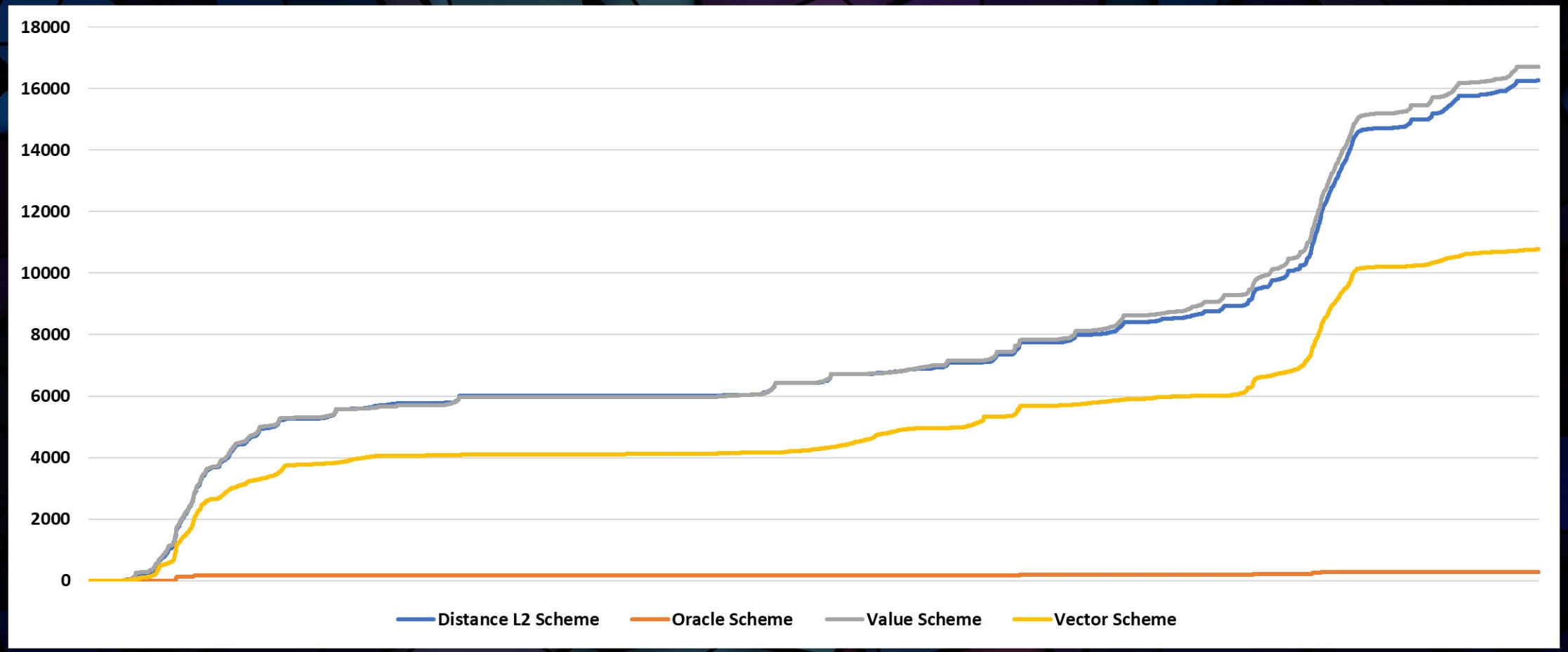

Inner Product Value
Time
22000
20000
18000
16000
14000
12000
10000
8000
Node 1
Node 2
Node 3
Node 4
Node 5
Node 6
Node 7
Node 8
Node 9
Node 10
FunctionValue

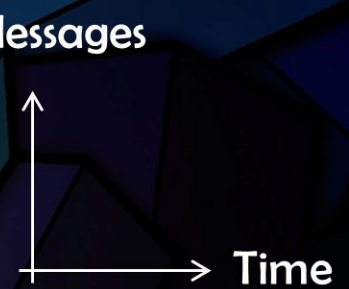
Bandwidth Over Time



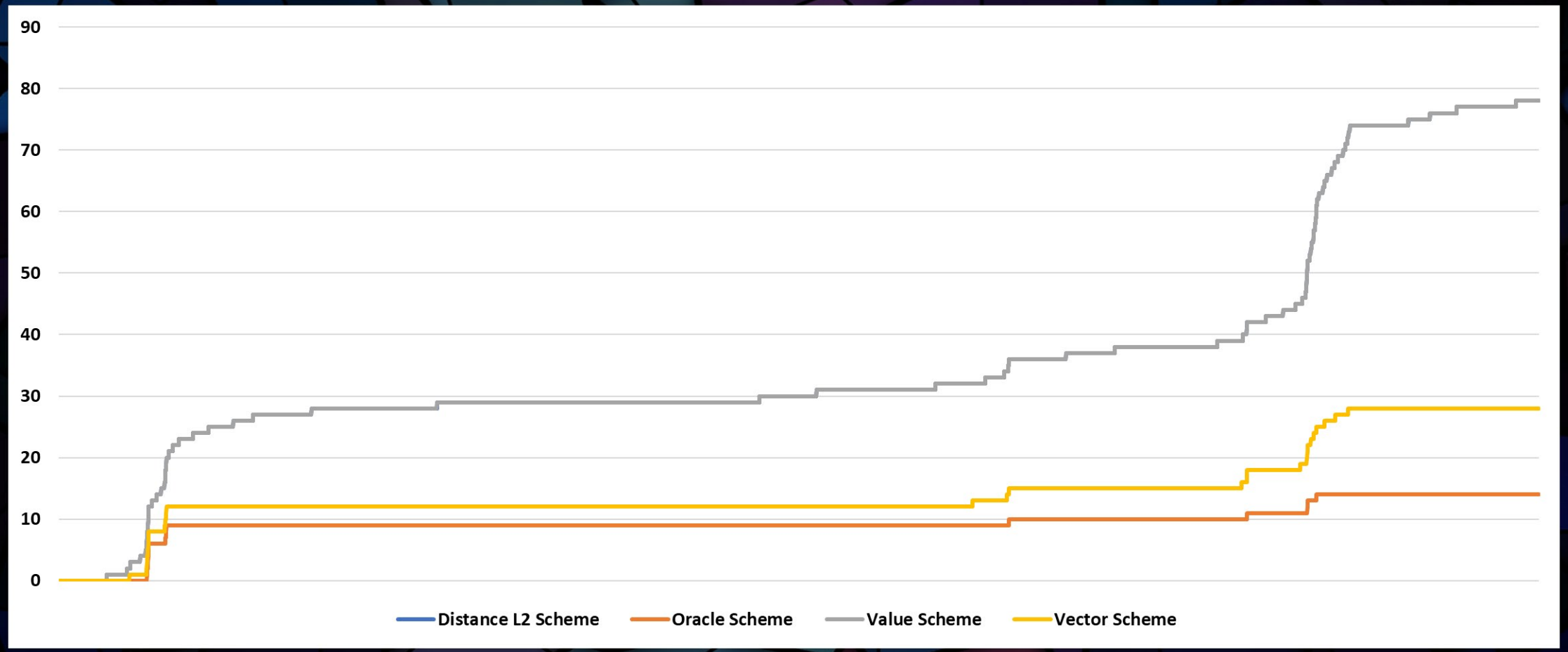


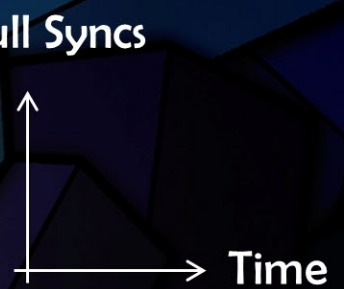
Messages Over Time



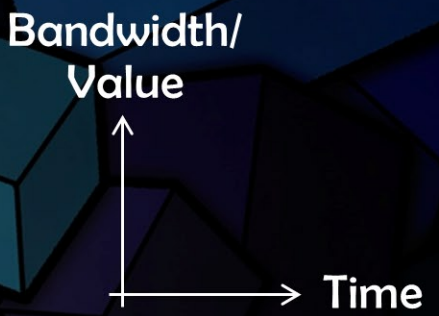


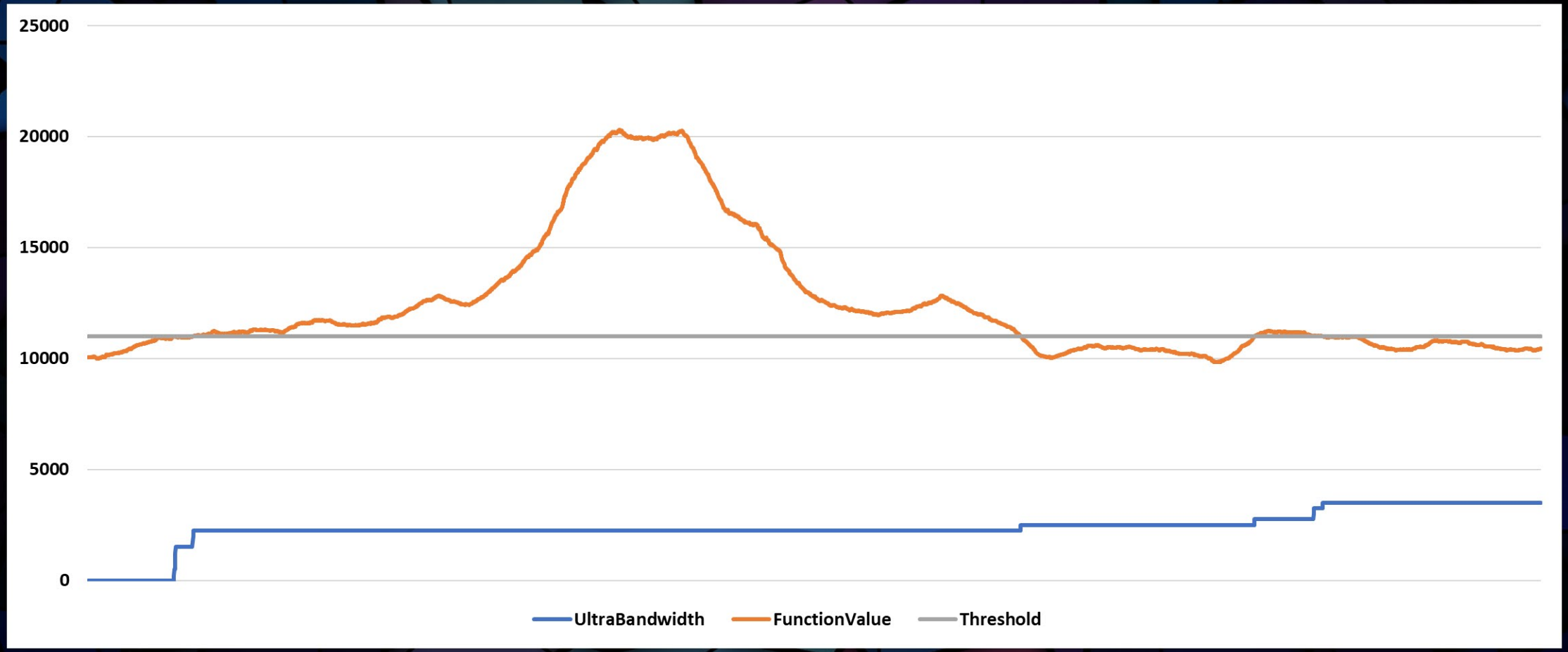
Full Syncs Over Time



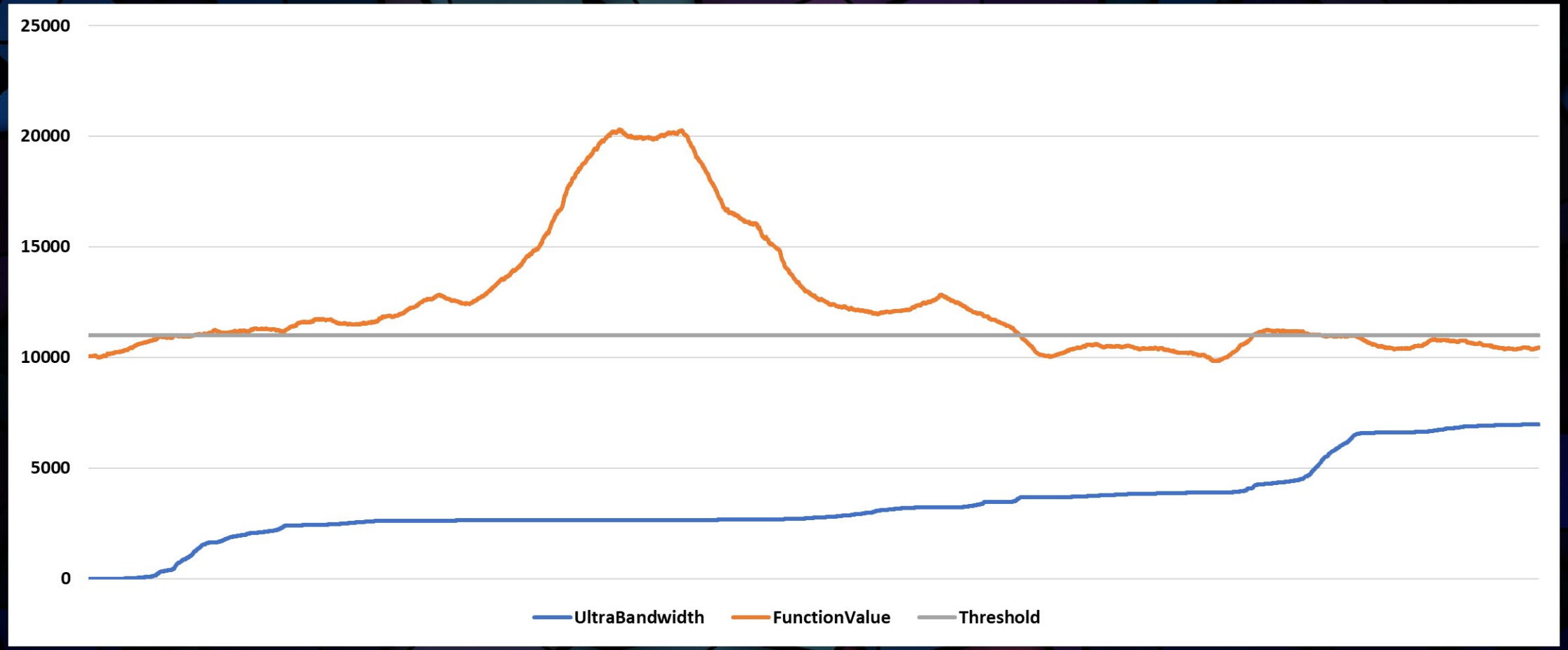


Oracle Bandwidth Scaled

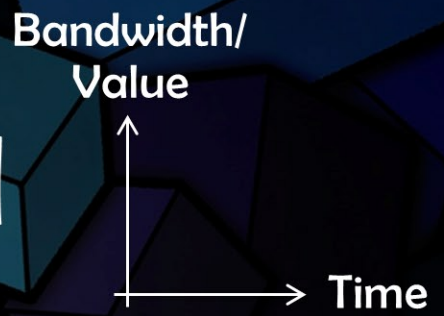


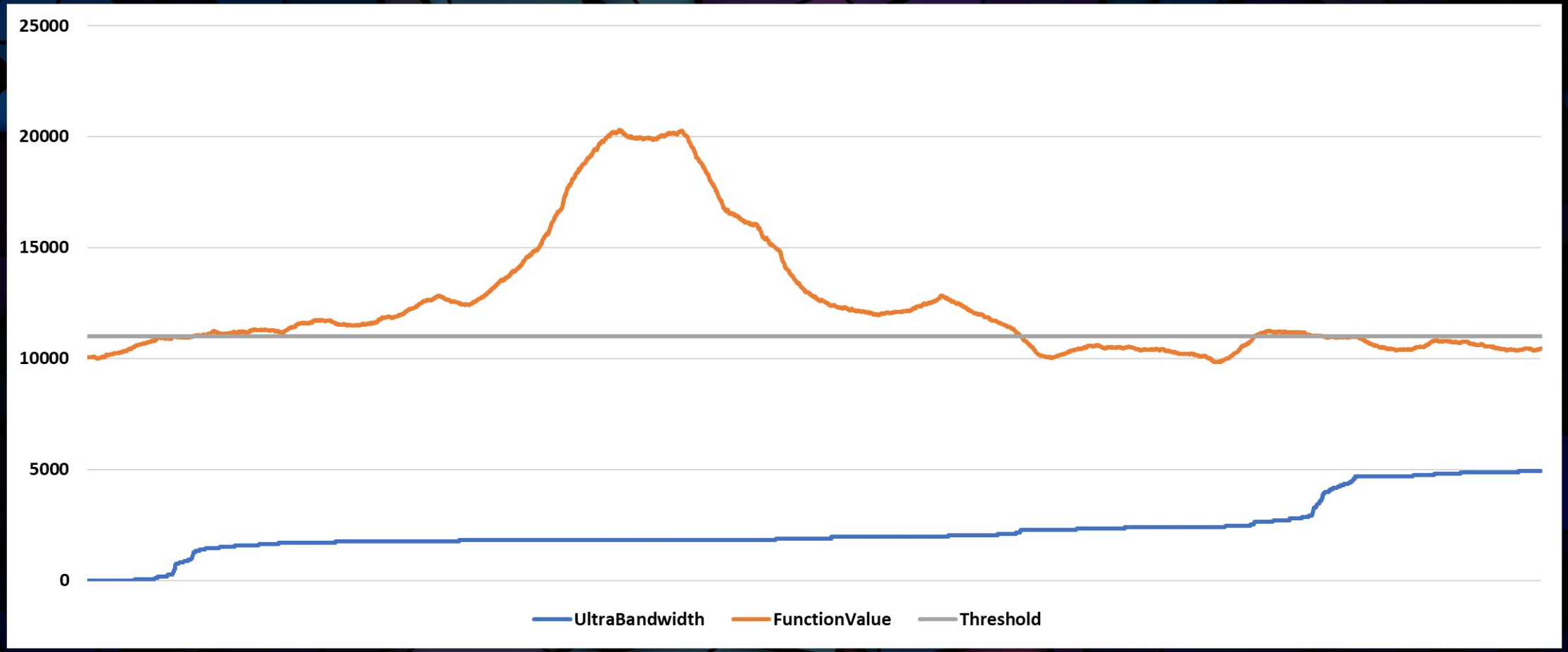


Vector Bandwidth Scaled



Distance Bandwidth Scaled





Bag of Words

Bag of Words

- Bag of words of length k consists of the K MOST COMMON TOKENS in English:
- ('the', 'of', 'and', 'to', 'a', 'in', 'for', 'is', 'on'...)
- tokens not in the k most common are ignored
- a histogram of the occurrences moves in a 'sliding window' fasion

Data Sources

- Textual data, each node is fed of a different oriented data:
  - Blogs textual data
    (http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm)
  - India's news headlines textual data
    (https://www.kaggle.com/therohk/india-headlines-news-dataset)
  - Jeopardy questions textual data
    (https://www.reddit.com/r/datasets/comments/1uyd0t/200000_jeopardy_questions_in_a_json_file/)
  - Spam clickbaits of news headlines textual data
    (https://www.kaggle.com/therohk/examine-the-examiner)
  - Some books textual data


Inner
Product

Inner Product Monitoring

- Distance Scheme – L_2 Distance to convex bound
- Data vector:

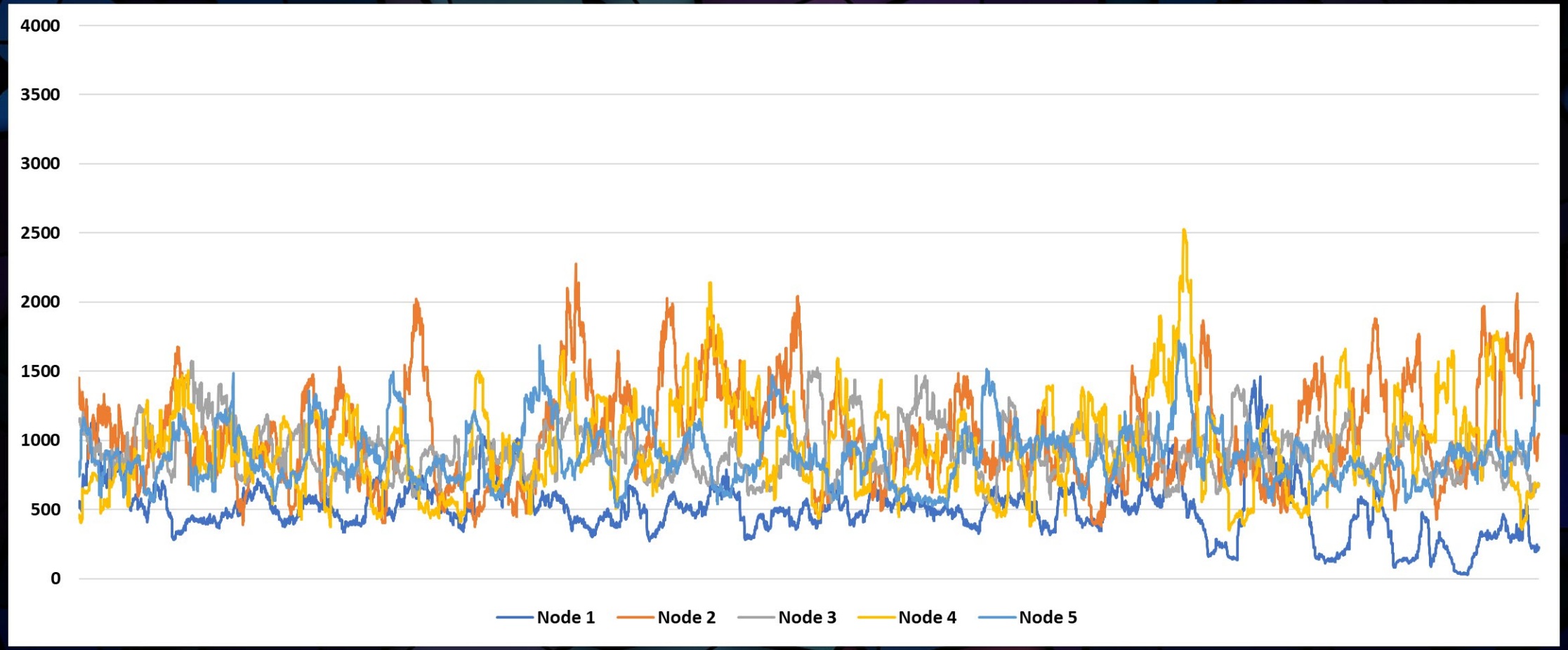
  vector(i) <- # occurrences in the window of the i'th lexicographic ordered n-most common token

  Inner Product(vector) = vector[0..n/2] · vector[n/2...n]

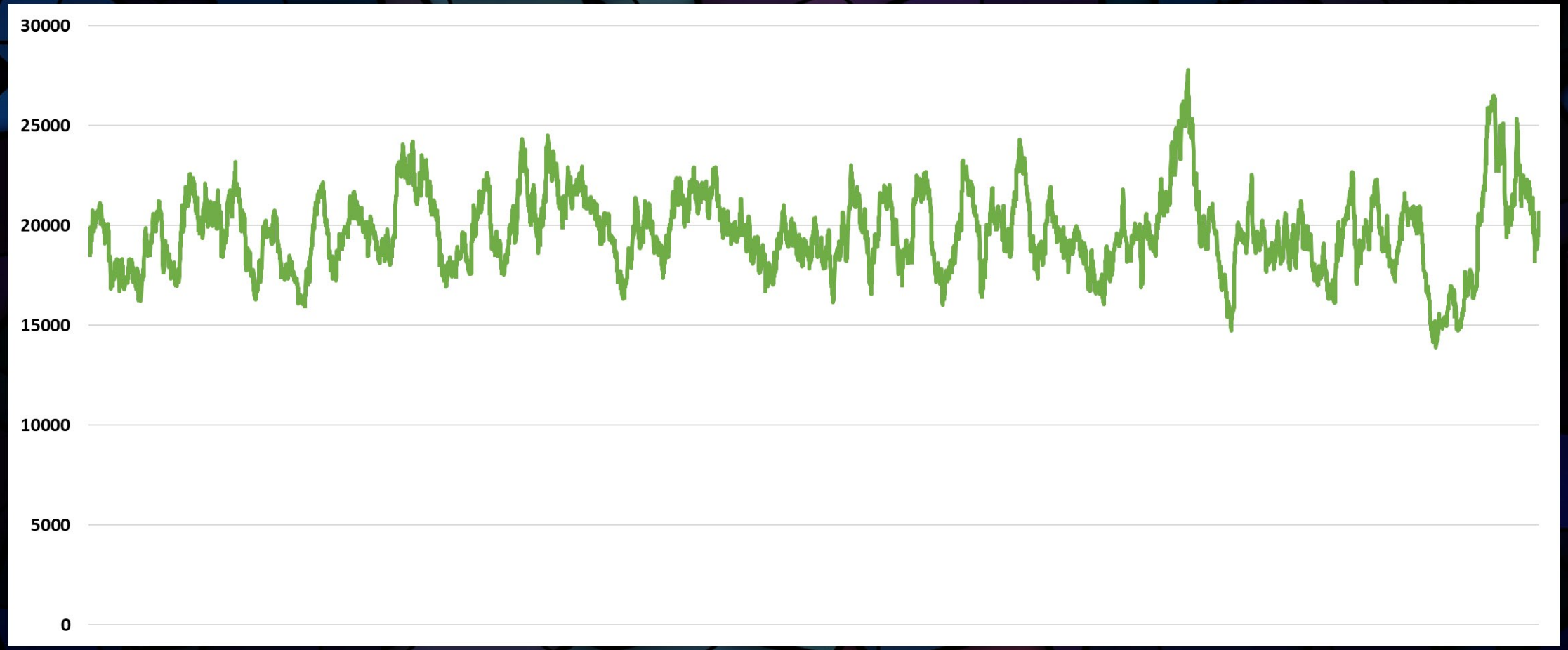
Running Information

- Window Size = 50,000
- Step Size = 100
- Epsilon – Multiplicative (*0.988, *1.012)
- Global BOW is the SUM of the local BOWs

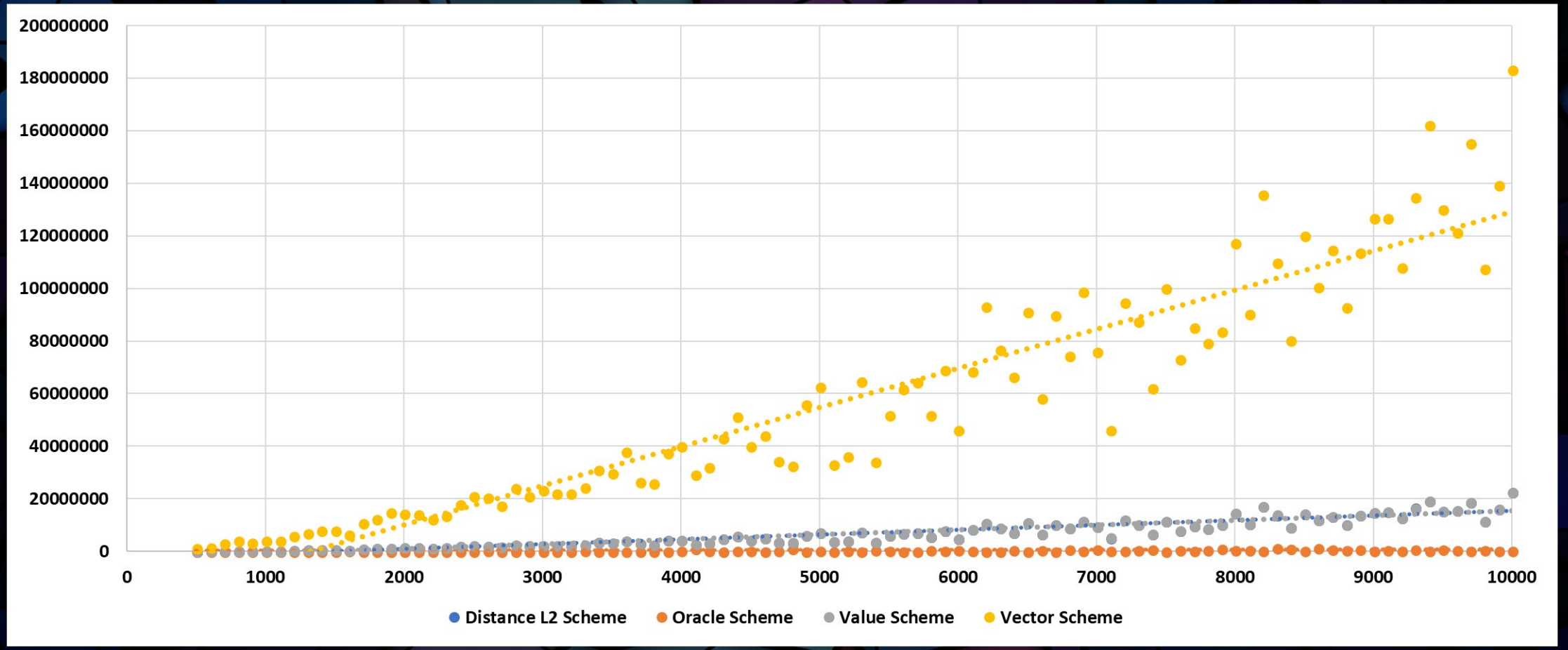
Nodes, Inner Data, |vector| = 10000



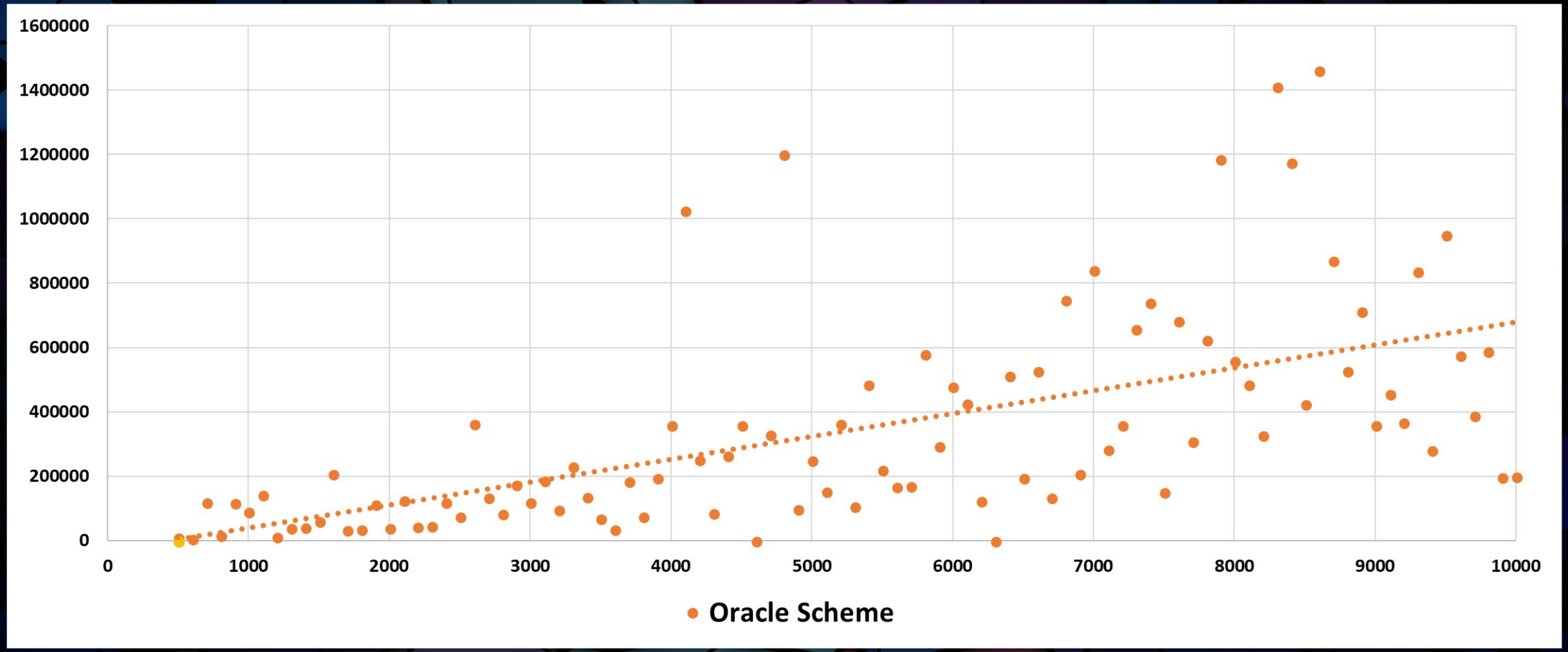
Global, Inner Data, |vector| = 10000



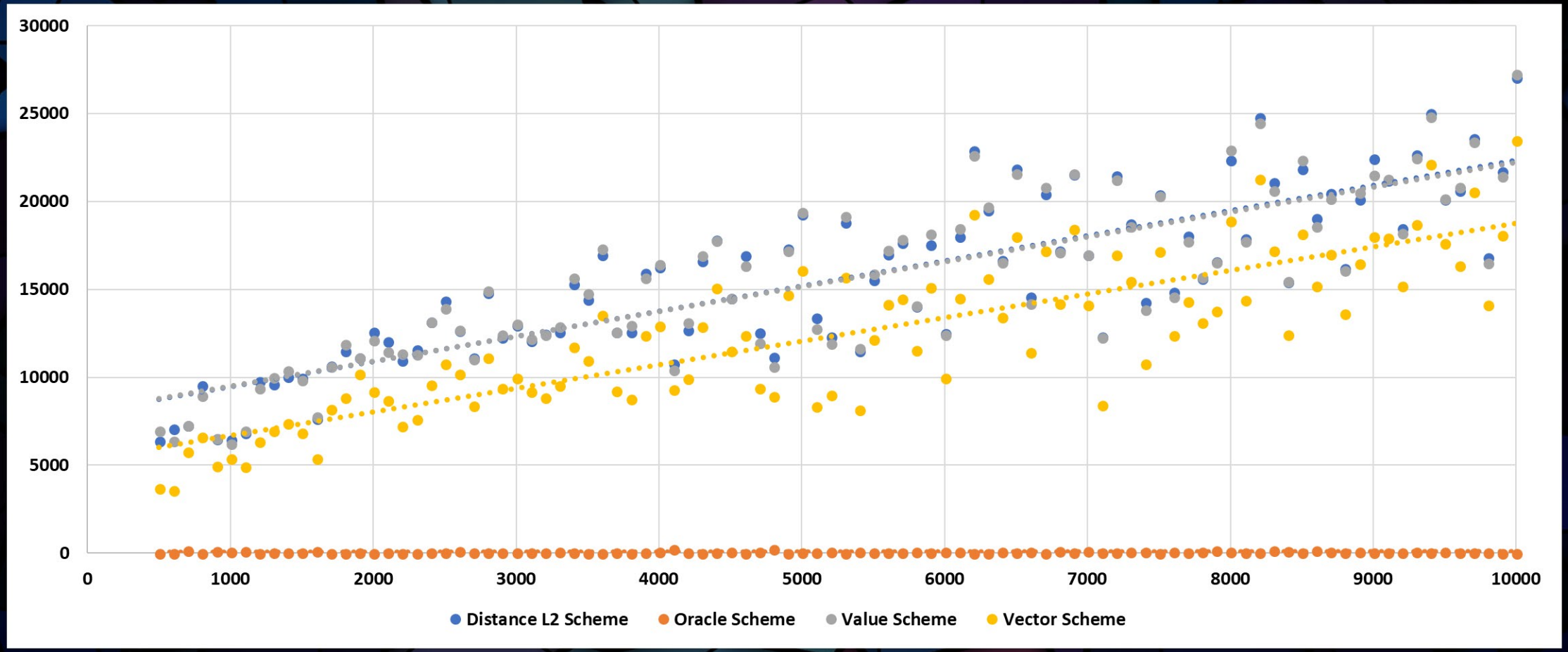
Bandwidth / |vector|



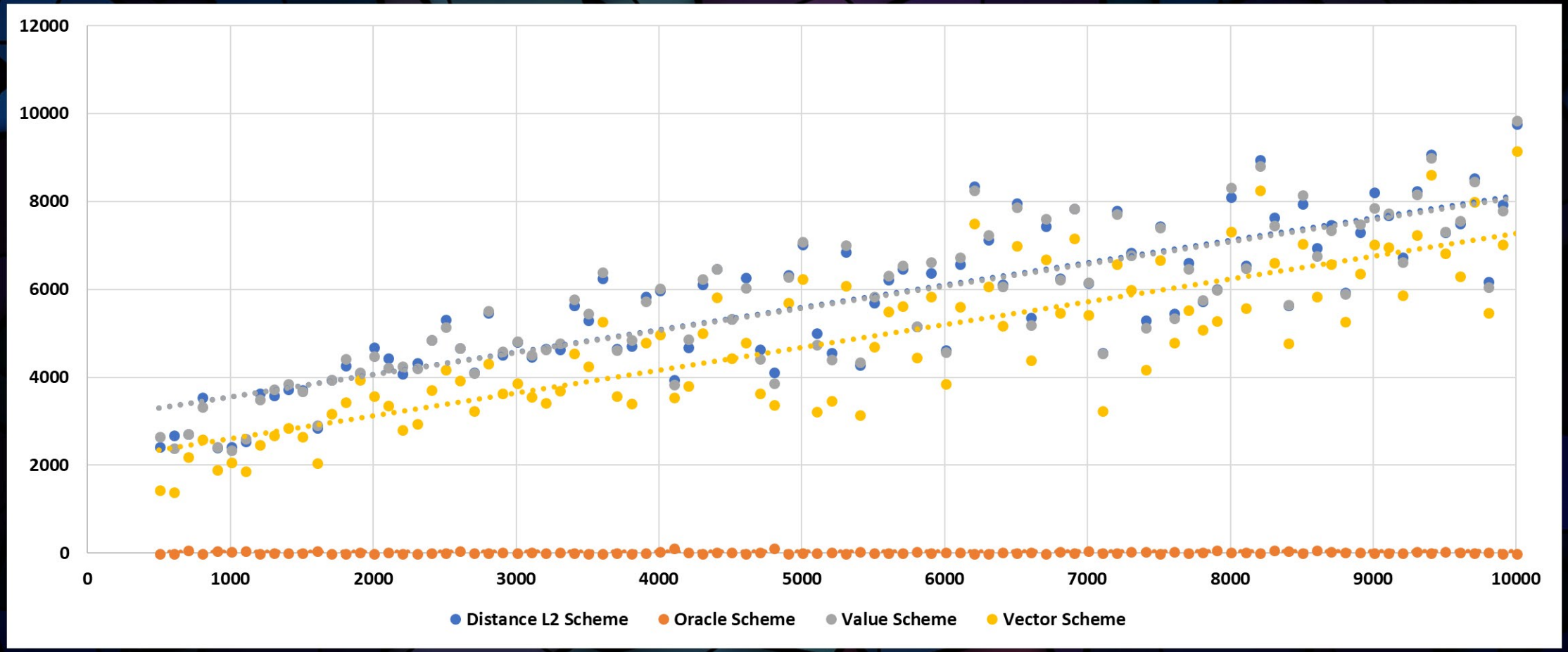
Bandwidth / |vector|



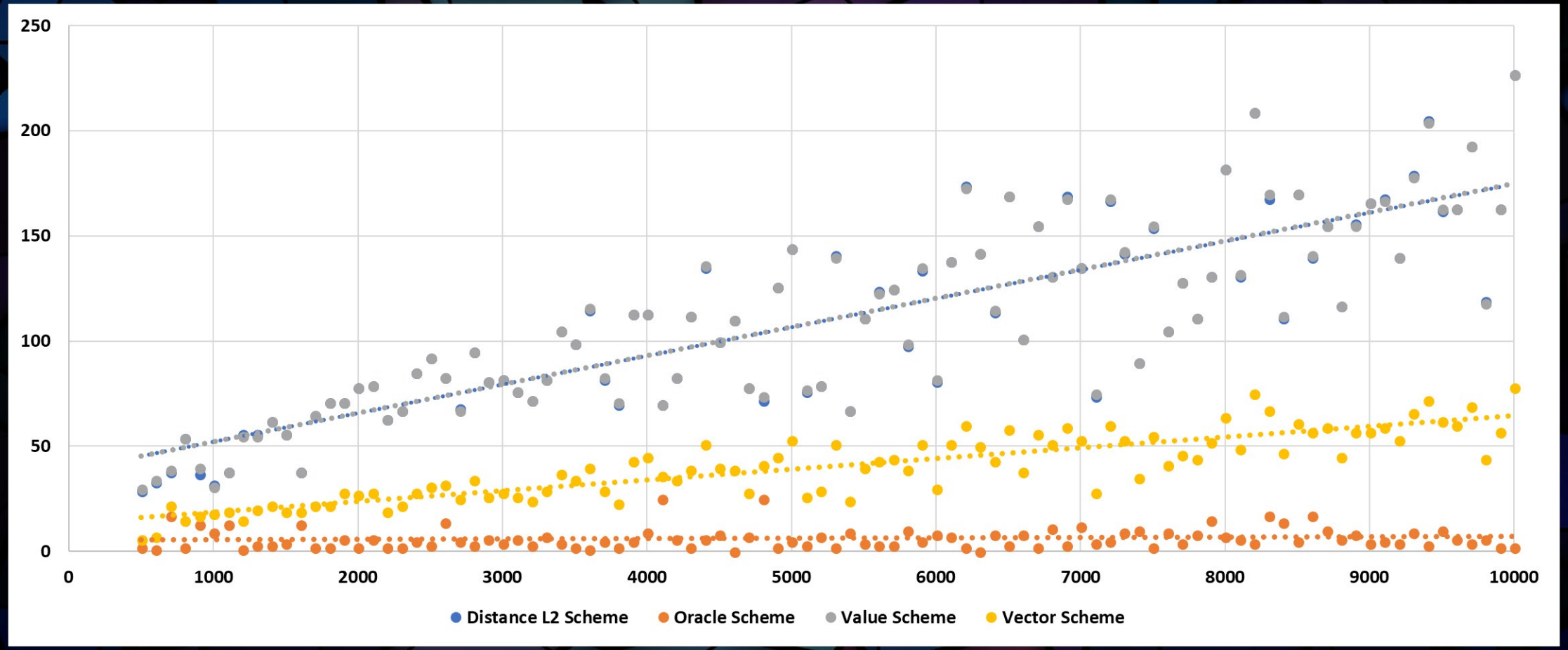
Messages / |vector|

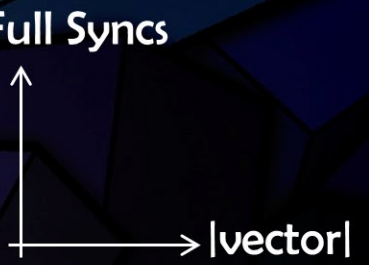


Channels / |vector|

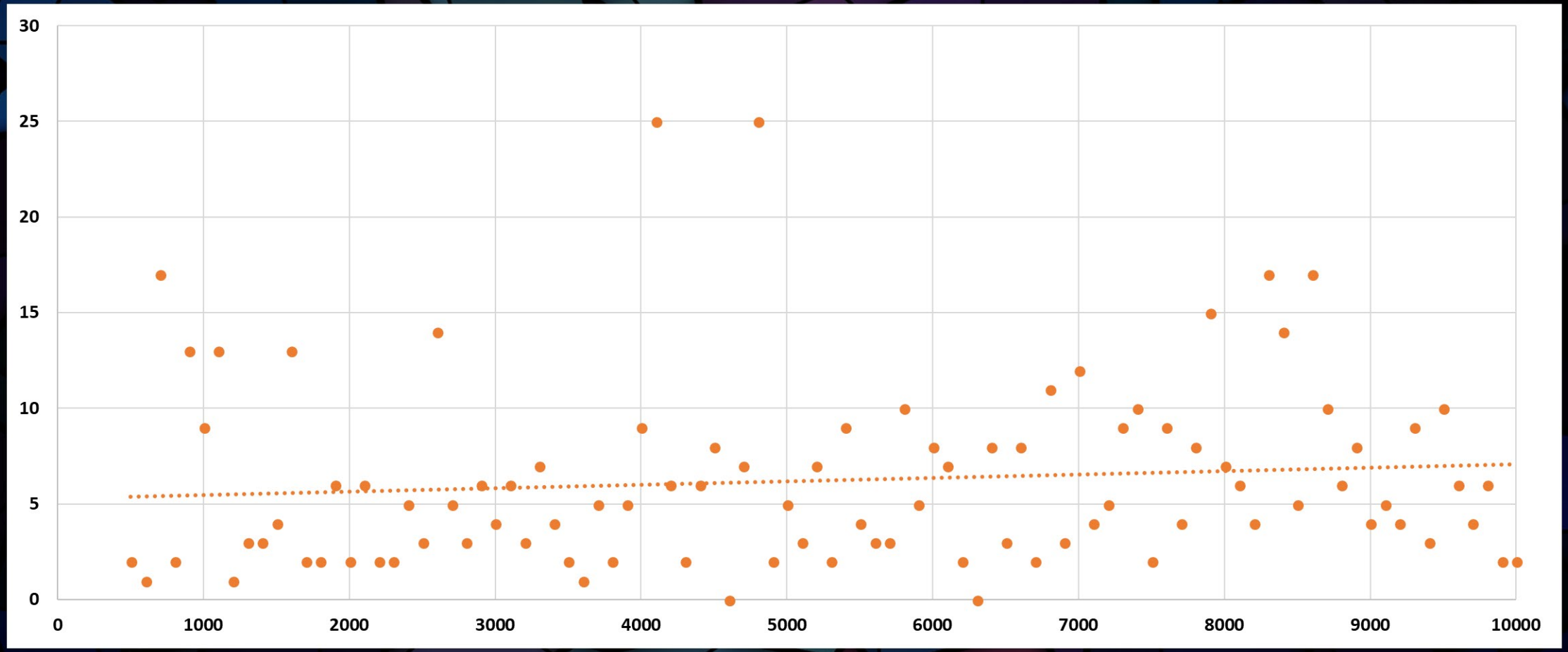


full syncs / |vector|





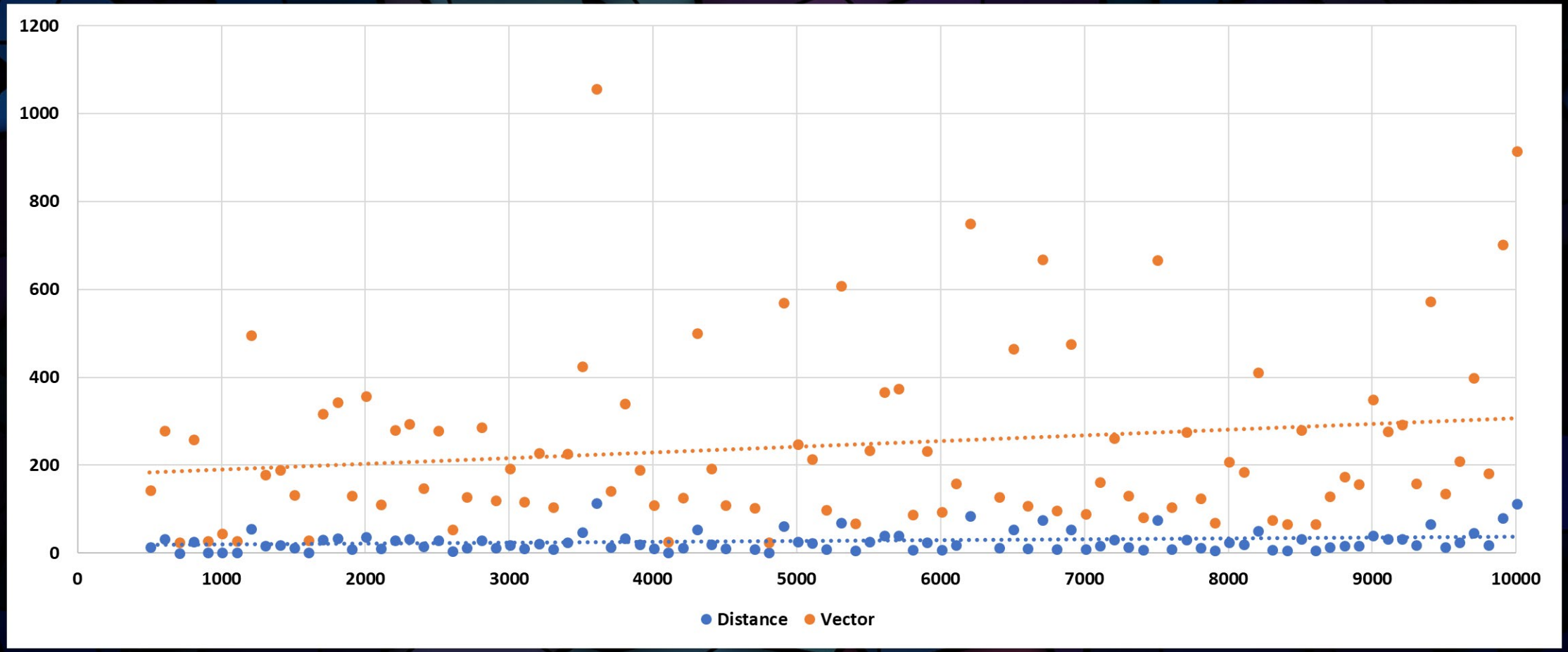
ORACLE # full syncs / |vector|



$$\frac{\text{Scheme Bandwidth}}{\text{Oracle Bandwidth}} / |\text{vector}|$$

$$\frac{\text{Scheme Bandwidth}}{\text{Oracle Bandwidth}}$$

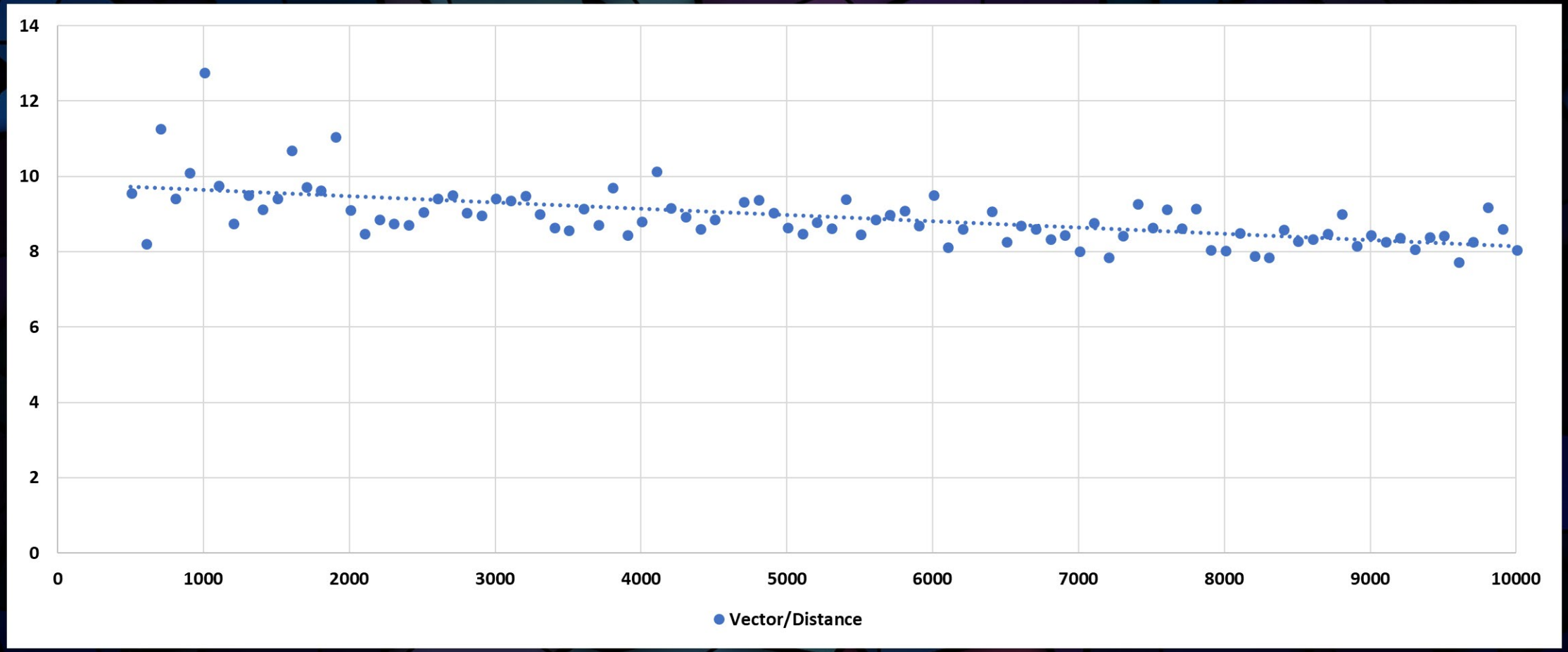
|vector|

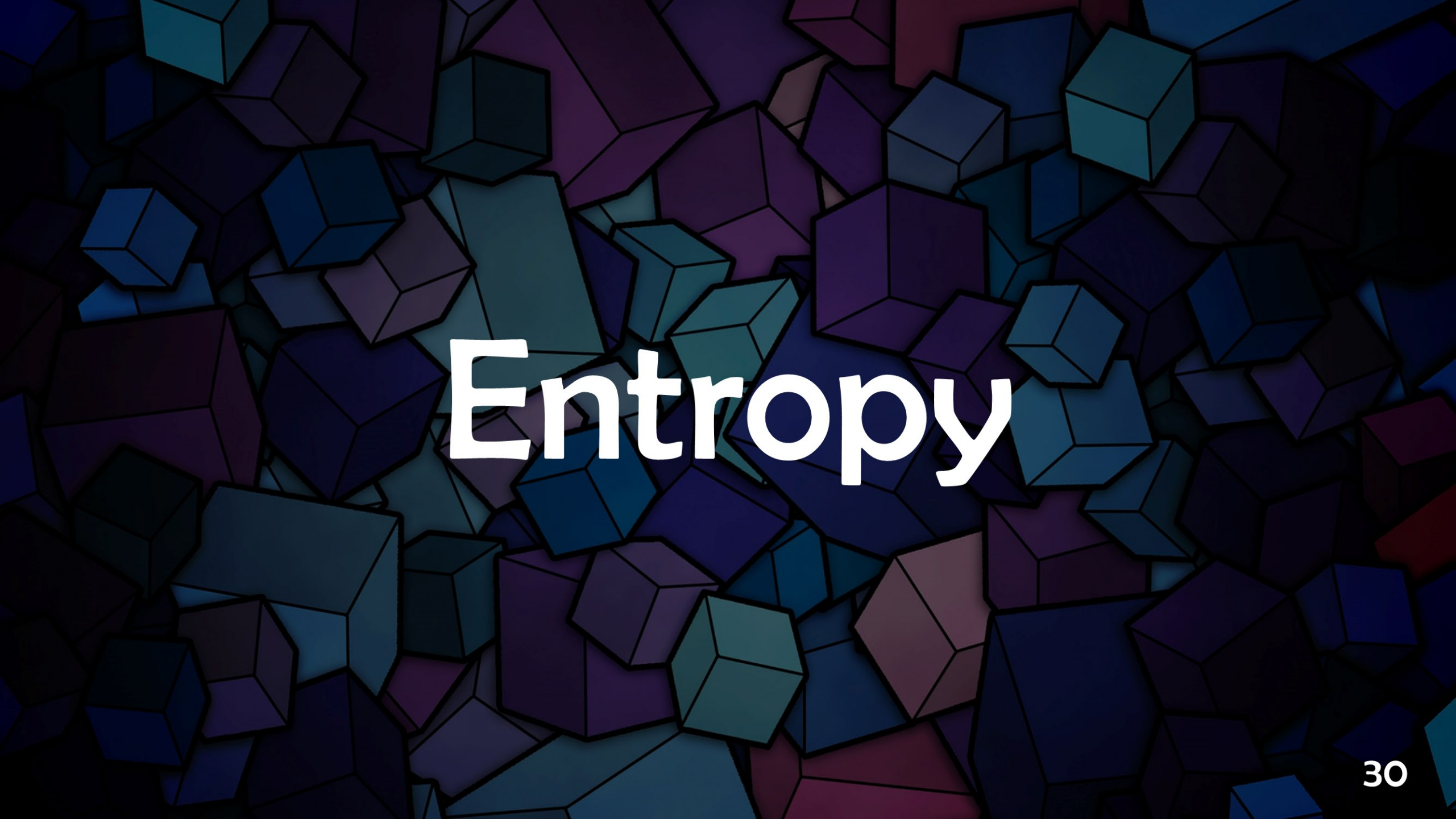


$$\frac{\text{Vector Bandwidth}}{\text{Distance Bandwidth}} / |\text{vector}|$$

$$\frac{\text{Vector Bandwidth}}{\text{Distance Bandwidth}}$$

|vector|




Entropy

Entropy Monitoring

- Distance Scheme – L_1 Distance
- Data Vector:

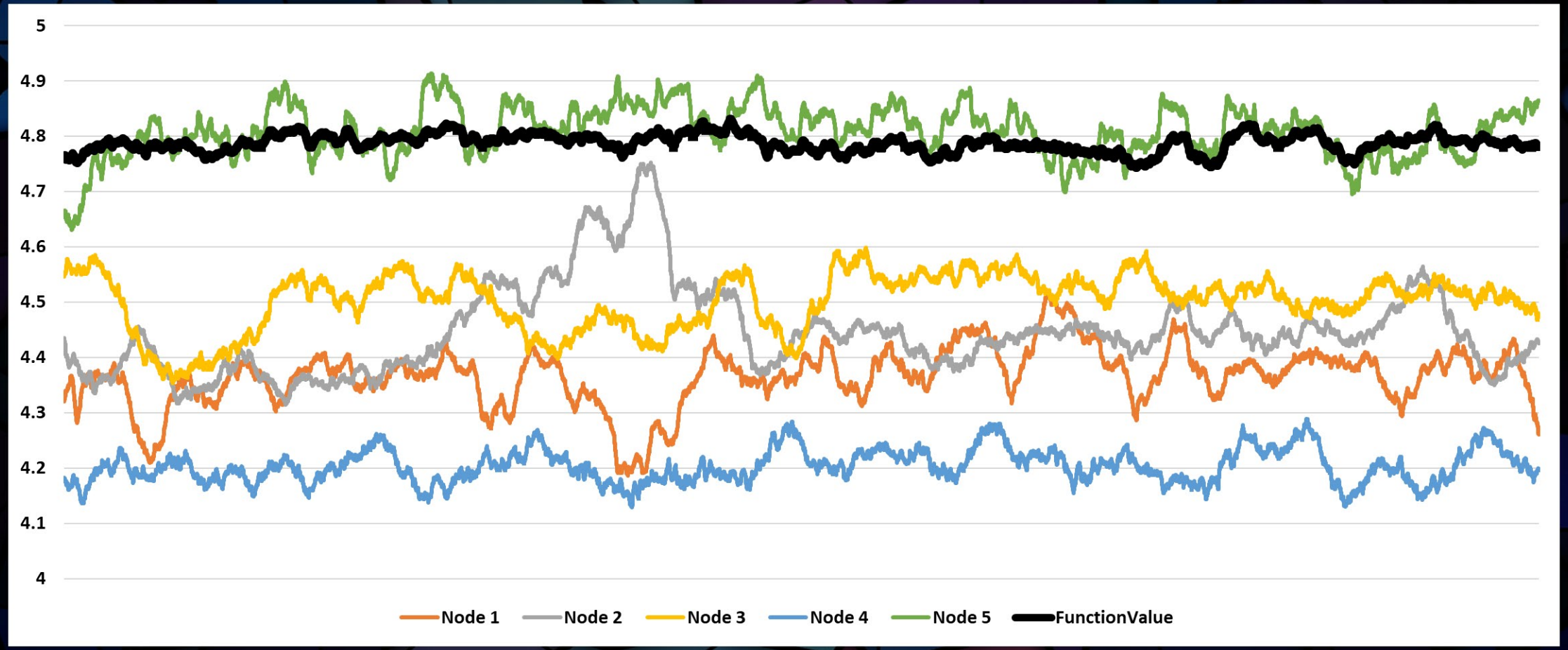
  vector(i) <- # occurrences in the window of the i'th lexicographic ordered n-most common token

  $\text{Entropy(vector)} = \sum p_i \cdot \ln(1/p_i)$

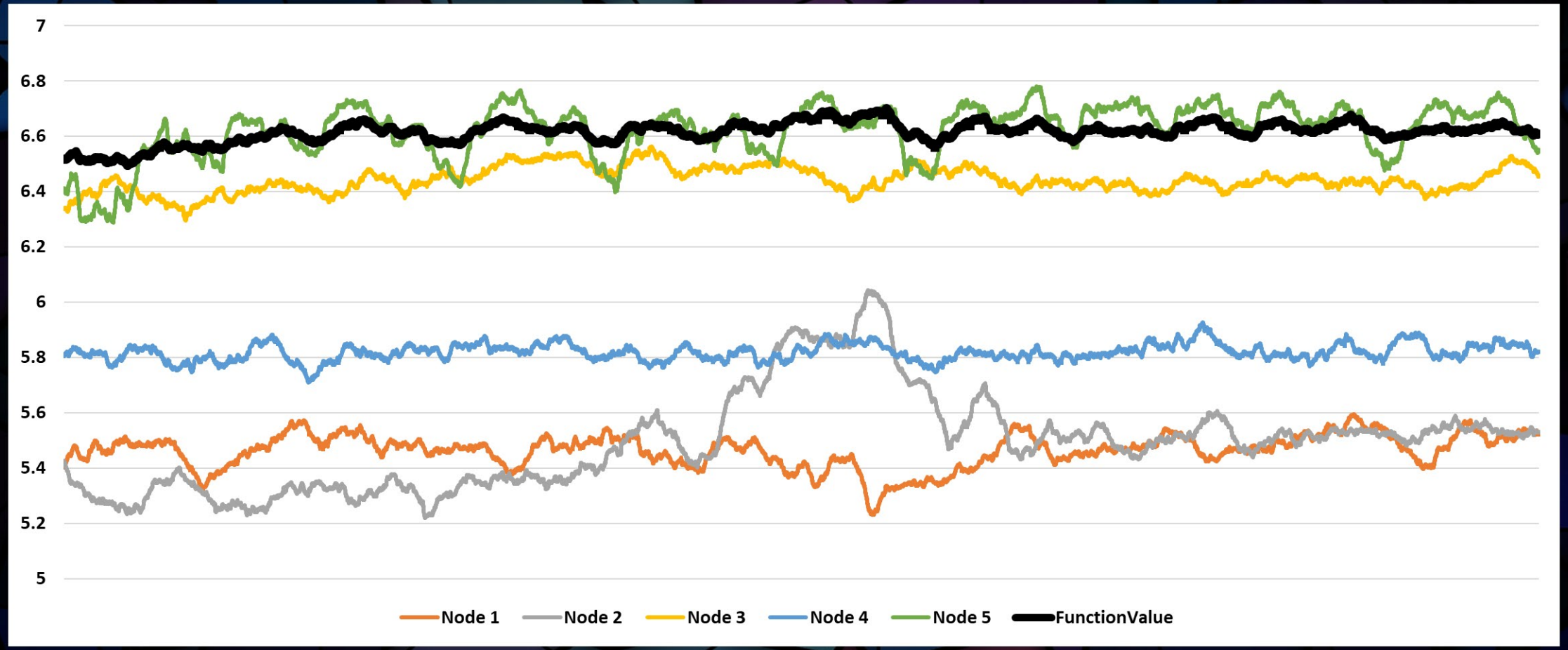
Running Information

- Window Size = 5000
- Step Size = 20
- Epsilon – Multiplicative (*0.995, *1.005)

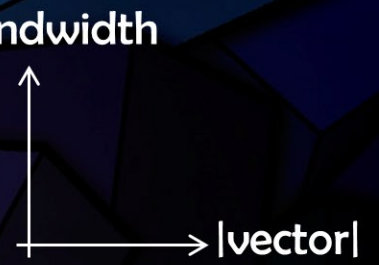
Entropy Data, |vector| = 500

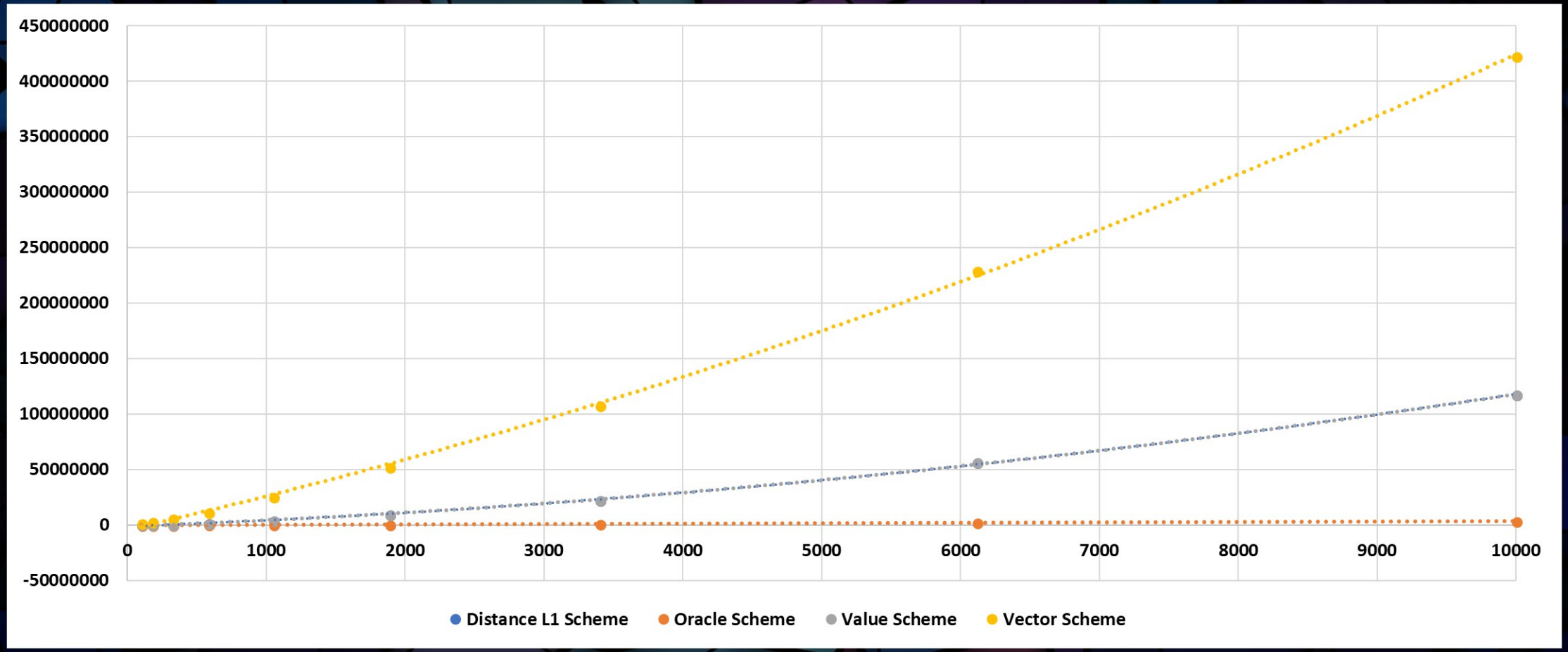


Entropy Data, |vector| = 10000

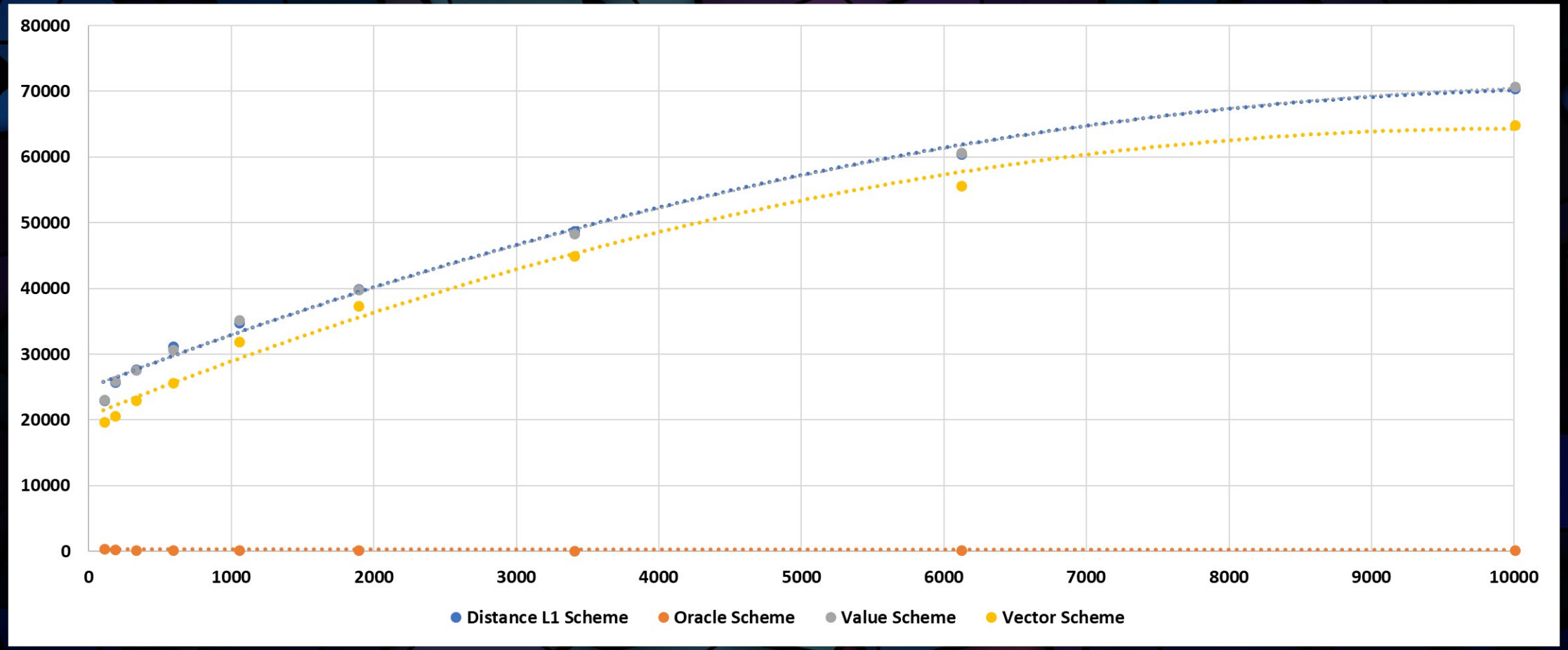


Bandwidth / |vector|

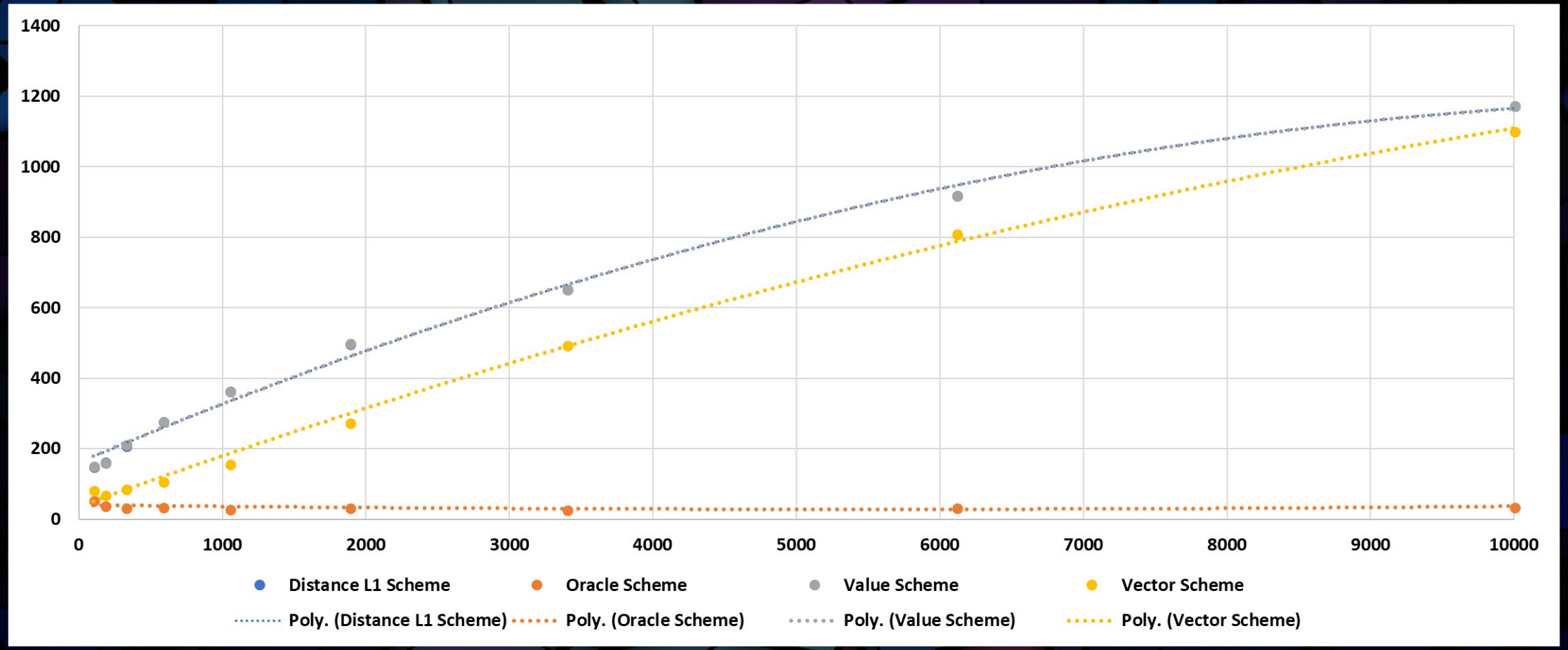




Messages / |vector|



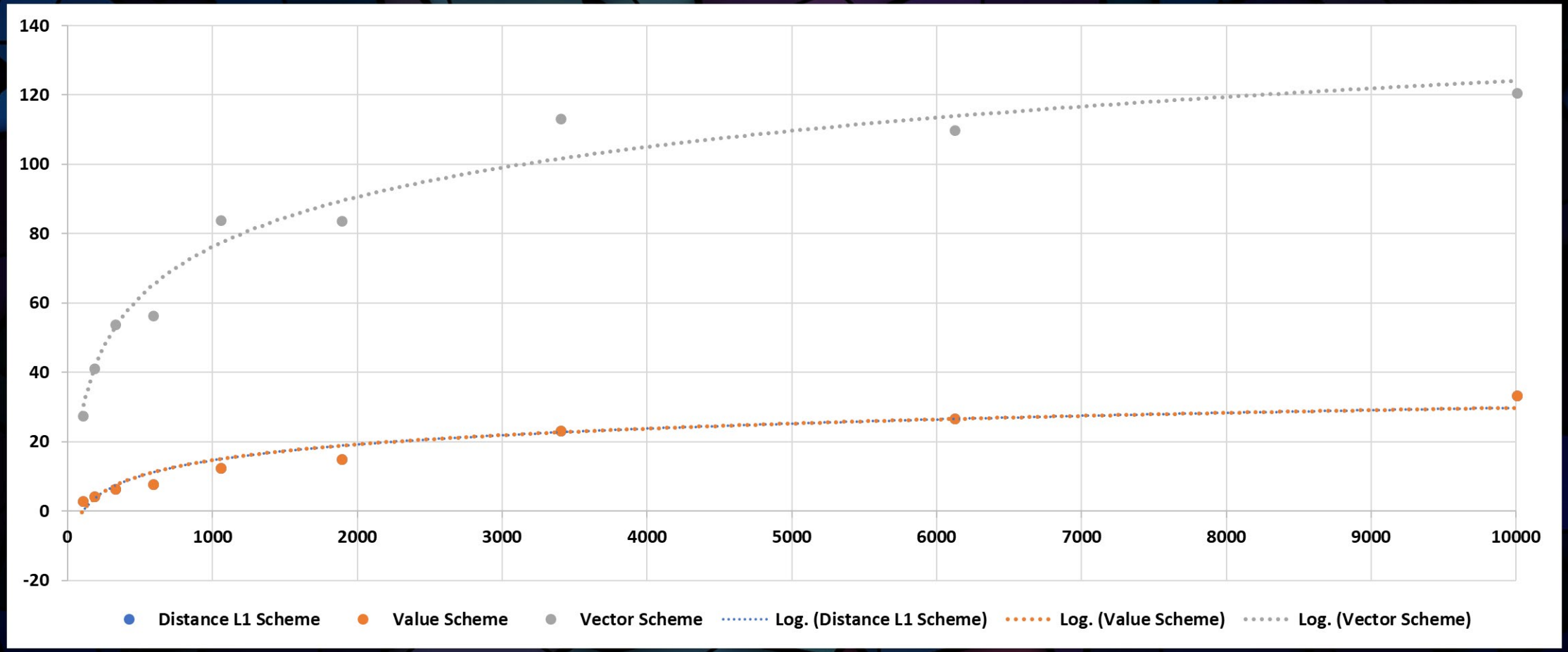
full syncs / |vector|



$\frac{\text{Scheme Bandwidth}}{\text{Oracle Bandwidth}} / |\text{vector}|$

$\frac{\text{Scheme Bandwidth}}{\text{Oracle Bandwidth}}$

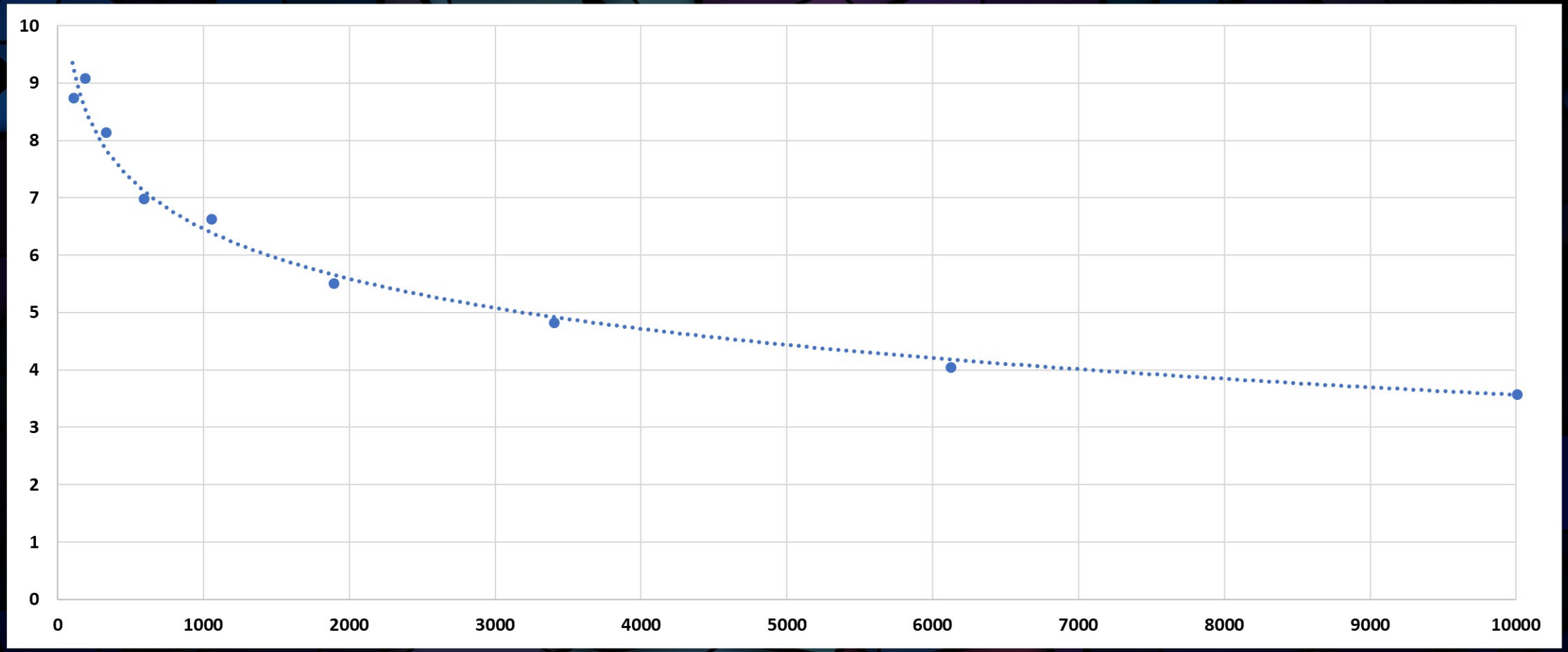
|vector|



$$\frac{\text{Vector Bandwidth}}{\text{Distance Bandwidth}} / |\text{vector}|$$

$$\frac{\text{Vector Bandwidth}}{\text{Distance Bandwidth}}$$

|vector|




PCA

PCA

- PCA of residual vectors could further decrease the bandwidth in Distance Scheme
- In order to asses how could will it perform, we should look at the eigenvalues of the PCA of all the residual vectors
- Problem:

  PCA takes cubic time in the dimensionality of the data

  |vector| = 440 -> 2 minute

  |vector| = 1000 -> 24 minutes

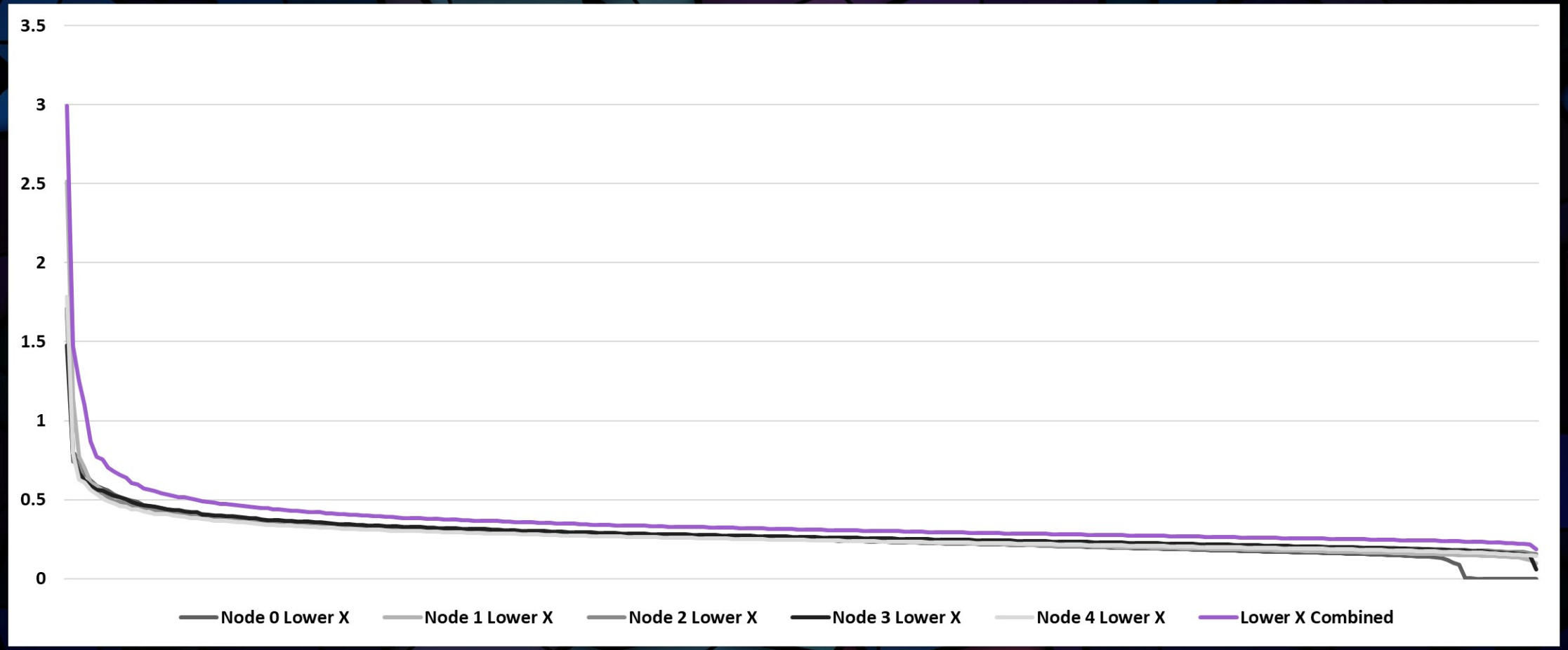
  |vector| = 1700 -> 2 hour

  |vector| = 3900 -> 2 day

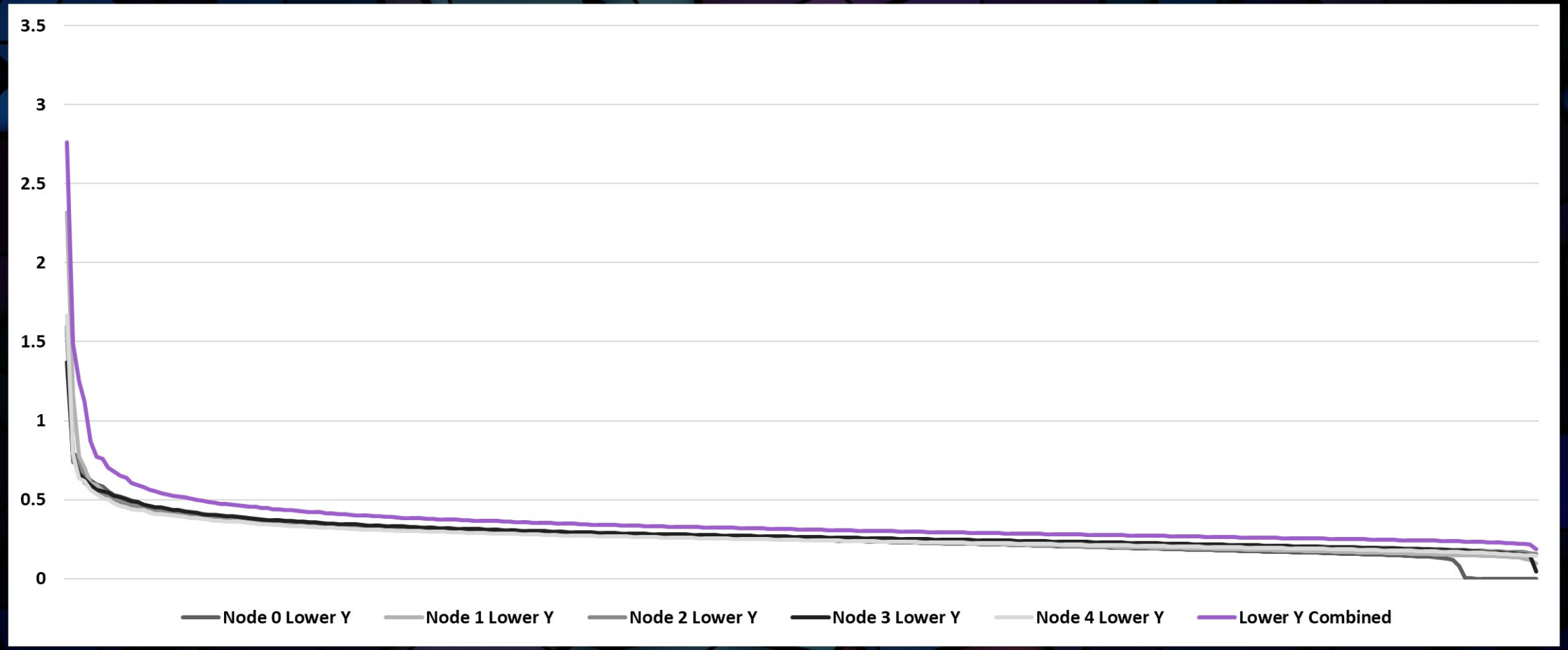
Inner Product PCA

- Inner Product
- Vector Length = 5,000
- Window Size = 50,000
- Step Size = 100
- # Nodes = 5

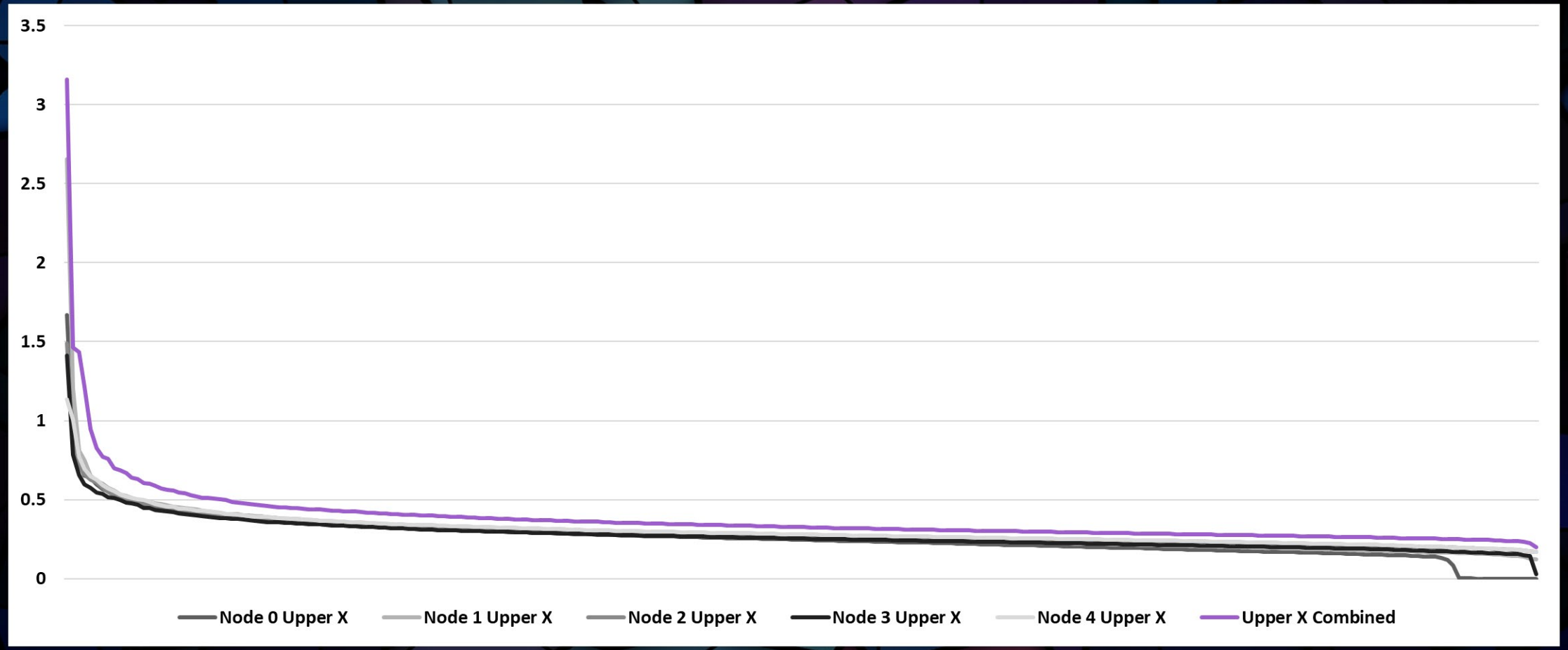
Nodes Lower Bound X Residuals



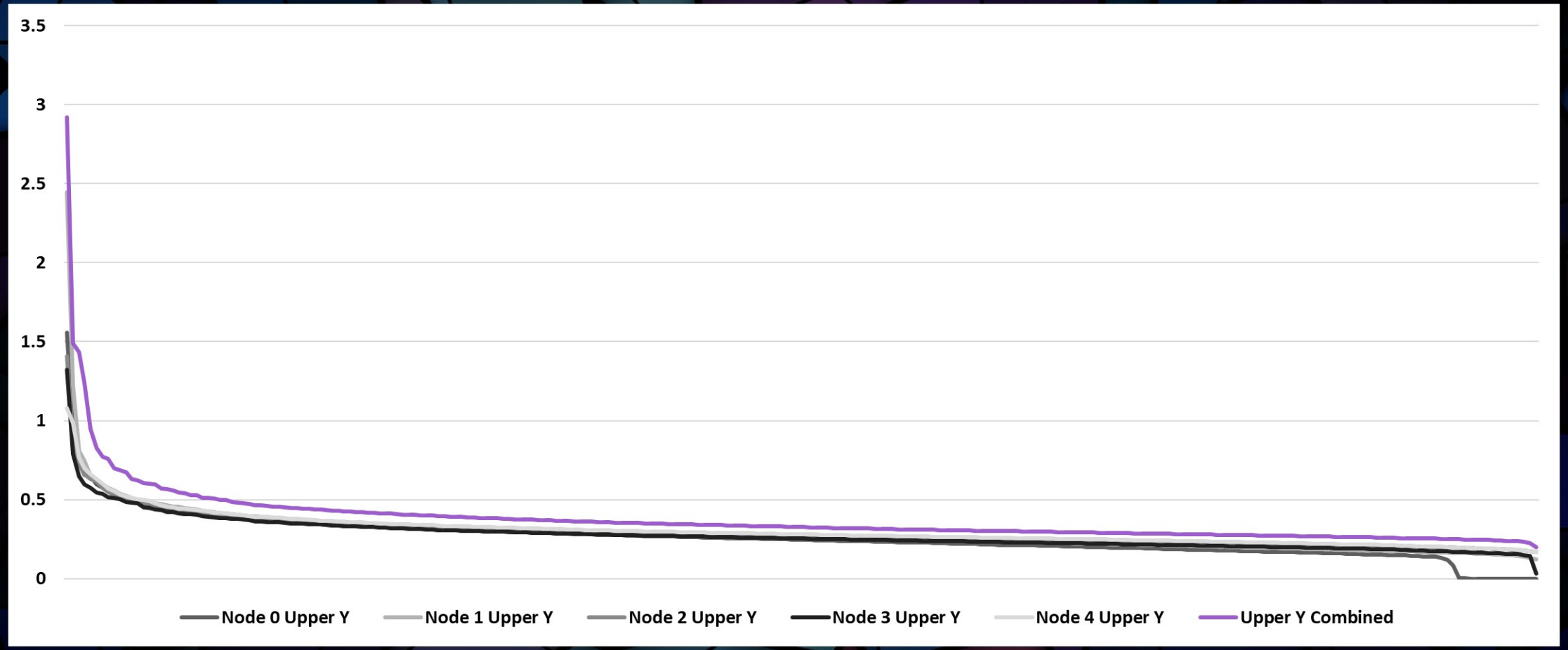
Nodes Lower Bound Y Residuals



Nodes Upper Bound X Residuals



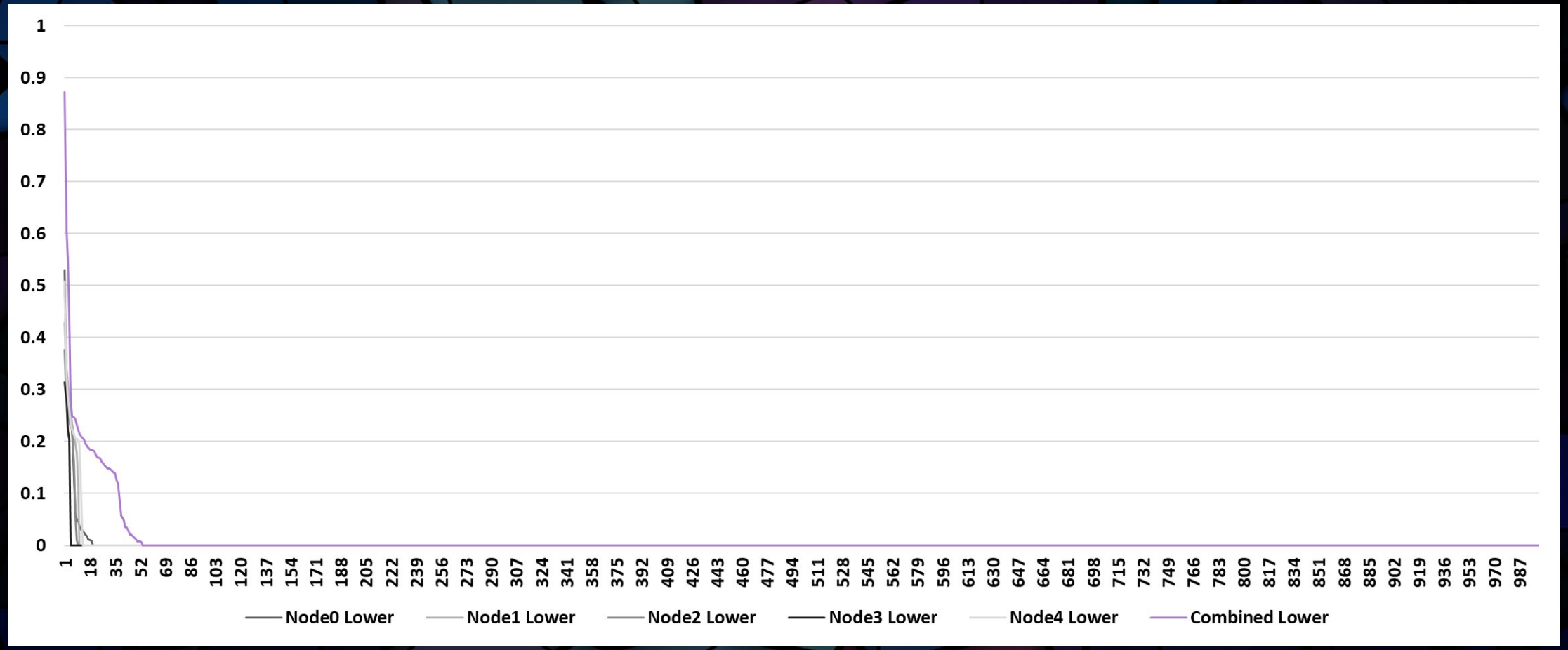
Nodes Upper Bound Y Residuals



Entropy PCA

- ❑ Entropy
- ❑ Vector Length = 1,000
- ❑ Window Size = 10,000
- ❑ Step Size = 100
- ❑ # Nodes = 5

Nodes Lower Bound Residuals



Nodes Upper Bound Residuals