QRATOR

Traffic Anomaly Detection and Attack Recognition

QRATOR Labs

The threat

- Intelligence agencies, as well as criminal organizations, sophisticated hackers are investing fortune to develop new types of Cyber Trojans.
- Trojans send metastasis within short period of time prior to disappearing.
- It is almost impossible to detect these Trojans since they keep mutating continually.
- Some Trojans leave a back door.

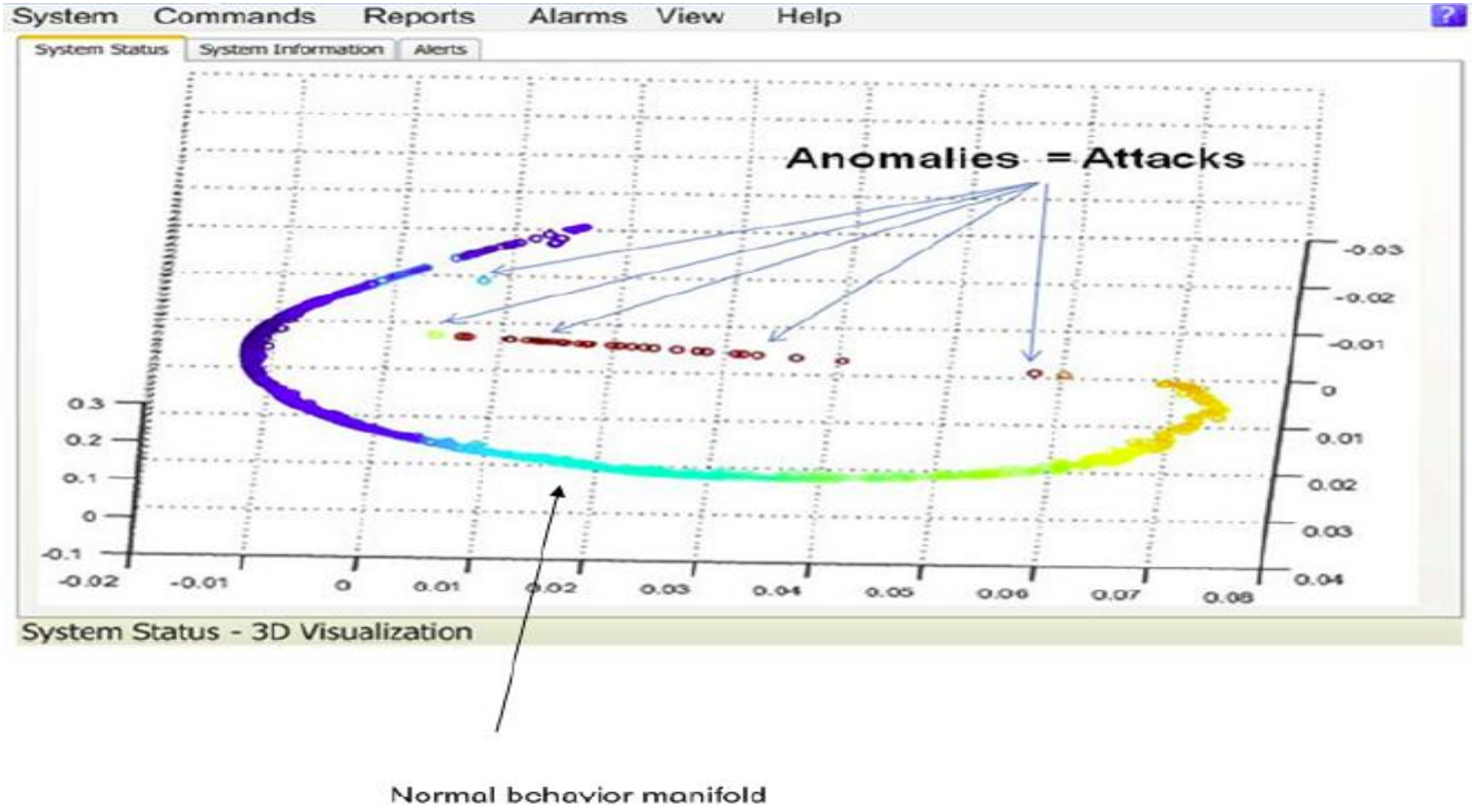
Network Intrusion Detection Systems

Our method has two sequential steps

- Study and analysis of the behavior of networking datasets and projection of data onto a lower dimensional space - training step.
  This is done once and updated as the behavior of the training set changes.
  Can handle corrupted training sets
- The output from the training step enables online detection of anomalies to which we apply automatic tools that enable real-time detection of problems.
  Each newly arrived datapoint is classified as normal or abnormal

Electronic intelligence and Cyber threat management: Generic approach

Theory, efficient algorithms, software and prototypes (integrated system) which process data in real time to detect anomalies that deviate from normal behavior



Sequential application of the packet-oriented traffic analyzer process

1. Get a new arrival of a packet;
2. For each packet, the analyzer parses its IP header and checks the protocol field (in the IP header) for the next level protocol used in the data portion of the datagram;
3. The corresponding protocol handler is called (the analyzer handles only icmp and tcp protocols);
4. Each protocol handler parses the packet headers accordingly and collects several values;
5. At every predetermined time slice (for example, one minute), the analyzer summarizes the values collected during this time slice and saves the statistics of this time slice. The following data is gathered and computed every predefined time slice:

Sequential application of the packet-oriented traffic analyzer process

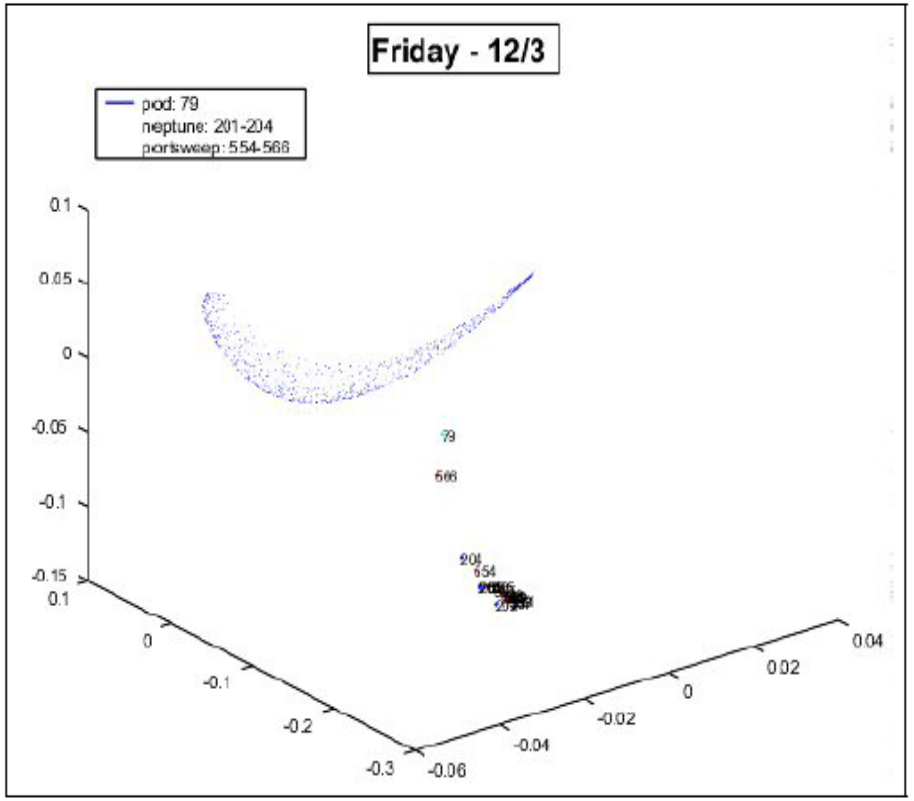
- number of IP packets;
- number of icmp packets;
- ratio between the number of icmp packets and IP packets (computed data);
- number of tcp packets;
- number of tcp packets with different tcp flags (syn, syn ack, fin, rst) (computed data);
- ratio between the number of tcp packets with syn ack flags and the number of tcp packets with syn flag (computed data);
- ratio between the number of tcp packets with rst flag and the number of tcp packets with syn flag;
- number of tcp connections (sessions);
- number of completed and uncompleted tcp connections;
- ratio between the number of the completed tcp connections and the number of the uncompleted tcp connections (computed data).

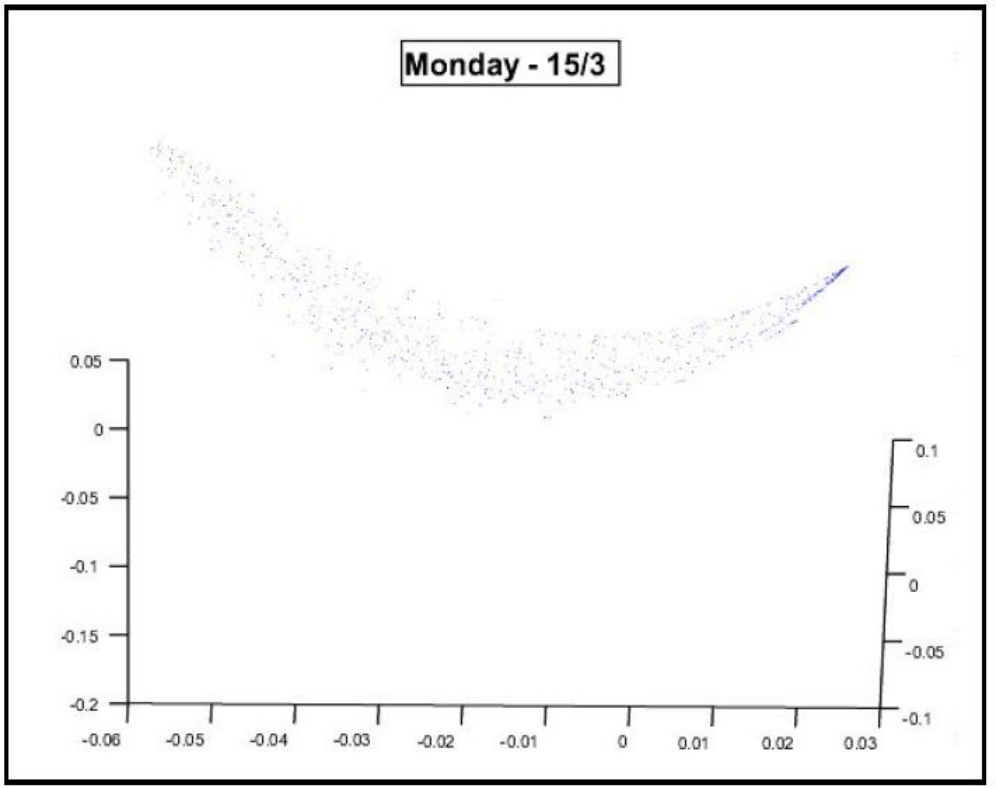
DARPA Evaluation Datasets

1999 DARPA Intrusion Detection Evaluation Datasets

- Simple network architecture and background traffic
  - 1000 virtual hosts
  - 700 virtual users
- Similar to multi-protocol traffic on one Air Force base
- Five weeks of network based attacks
  - Two weeks do not contain any attacks
  - One week contains selected labeled attacks
  - The last two weeks contain 201 instances of 56 types of labeled attacks

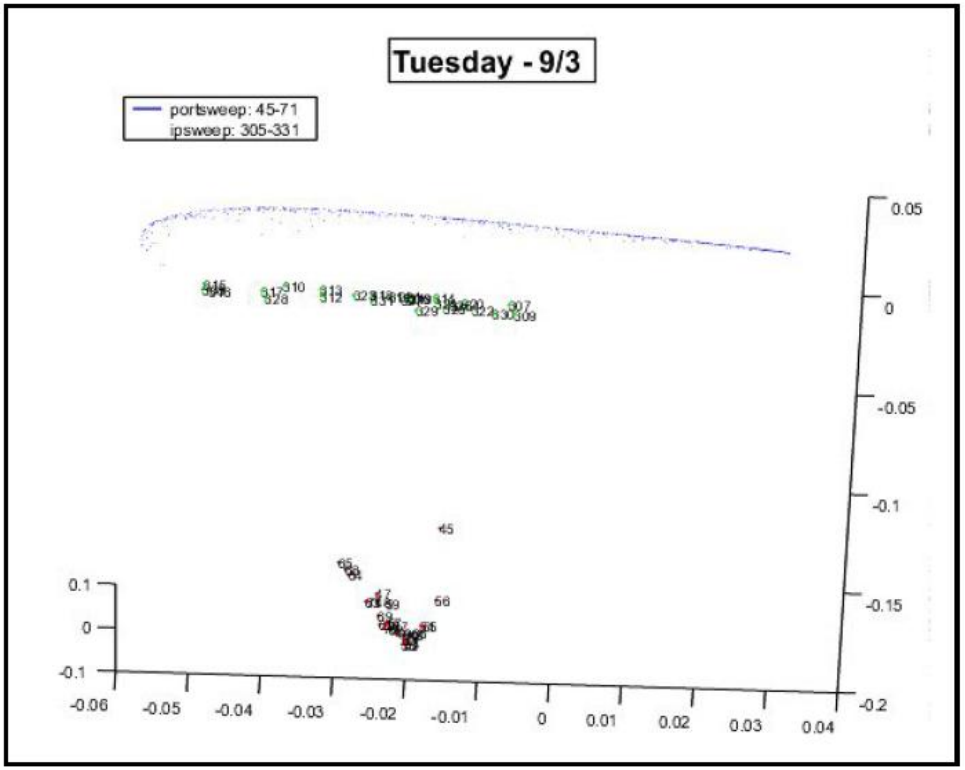
DARPA: simulated attacks on air base[1]

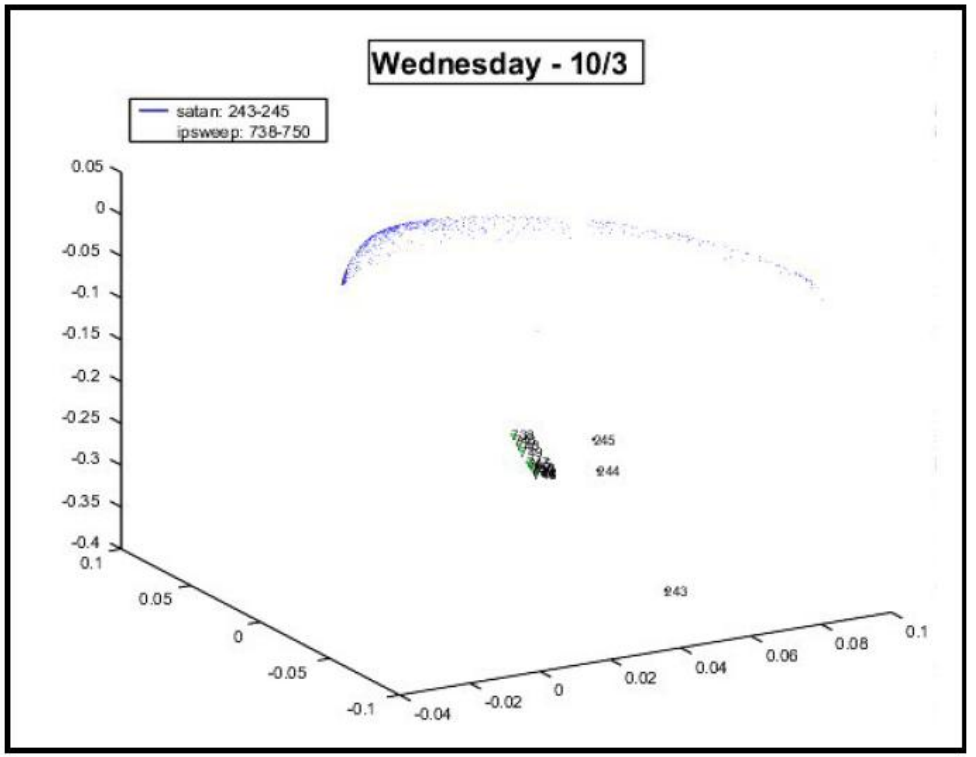




[1]. David, Anomaly Detection and Classification via Diffusion Processes in Hyper-Networks, Ph.D. Thesis, Tel Aviv University, March 2008
(Tel Aviv University) 6 /

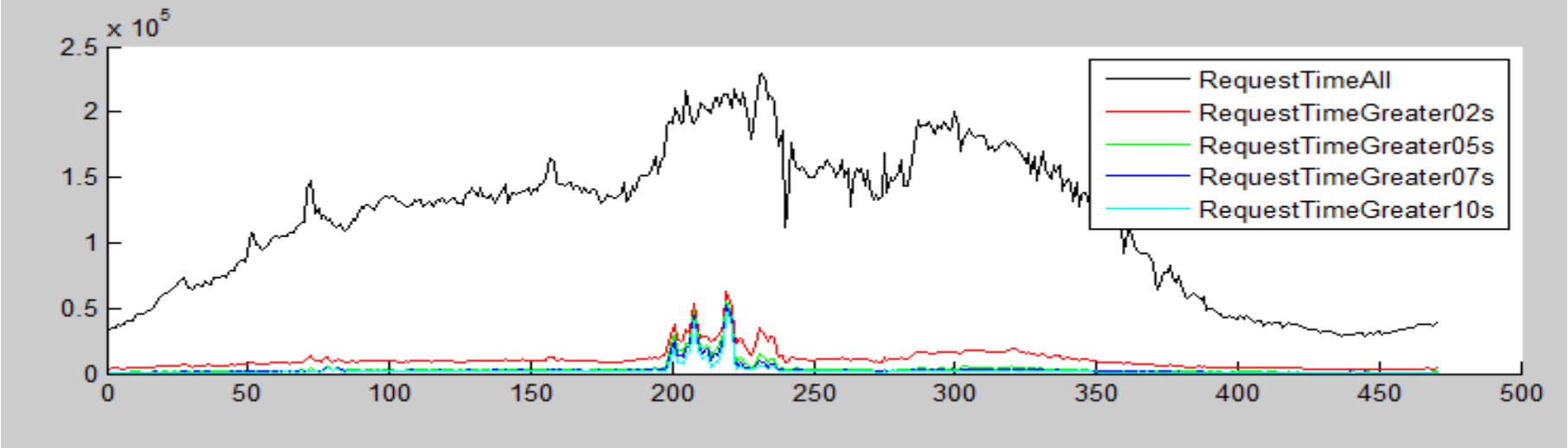
DARPA: simulated attacks on air base[1]

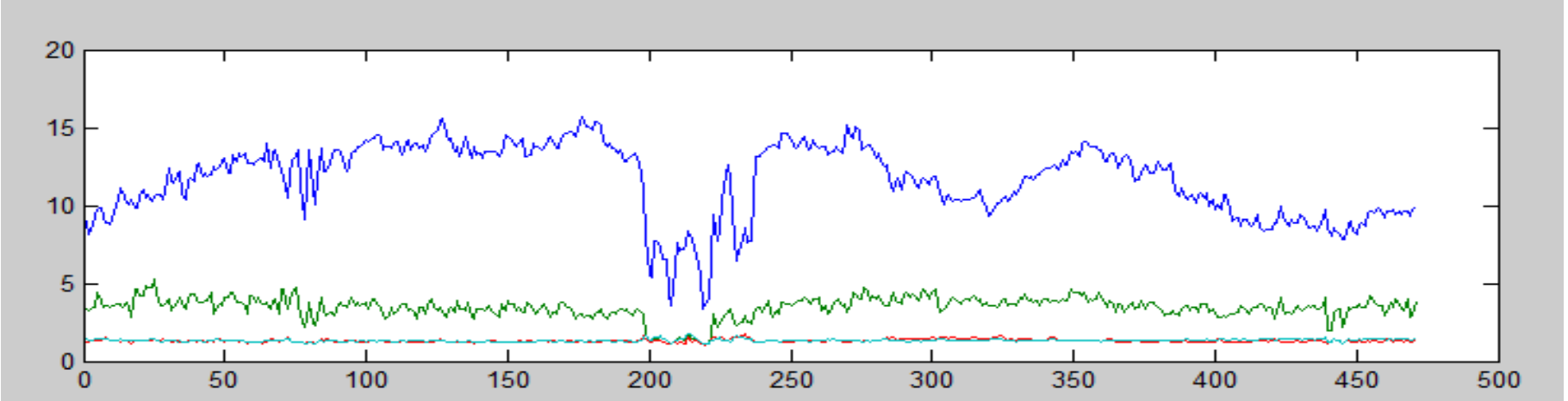




[1]. David, Anomaly Detection and Classification via Diffusion Processes in Hyper-Networks, Ph.D. Thesis, Tel Aviv University, March 2008
(Tel Aviv University) 6

DARPA: simulated attacks on air base[1]





The example of IP-domen traffic's features due one day and its relations (features)
The stochastic process X={x1,…xn} where x_i- all features at the moment of the time

Problem setup

Dimensionality reduction:

- Given: A set $X = \{x_1, x_2, \ldots, x_n\}$ in a high dimensional metric space M, $M = (X, \|\cdot\|)$ and a function $f : X \to \mathbb{R}$.
- Objective: mapping X to a low dimensional space L, $\bar{f} : X \to L$, while preserving certain features.

Limitation: Data accumulates over time,
and reproduction of $\bar{f}$ is of high computational cost.

Our Challenge: Once X is mapped - extension of $\bar{f}$ to $x \notin X$, using representatives from X (sampling).
In other words, based on f and the distances of x from X, extend f (denoted by $\bar{f}$) for any $x \notin X$.

Standard approach: Diffusion Maps (DM)

- DM is used to analyze a dataset M by exploring the geometry of the manifold $\mathcal{M}$ from which it is sampled.
- It is based on defining an isotropic kernel $K \in \mathbb{R}^{n \times n}$, whose elements are defined by $k(x,y) \triangleq e^{-\frac{\|x-y\|}{\epsilon}}$, $x, y \in M$, ϵ is a meta-parameter of the algorithm.
- This kernel represents the affinities between data points in the manifold. The kernel can be viewed as a construction of a weighted graph over the dataset M. The data points in M are the vertices and the weights of the edges are defined by the kernel K.
- The degree of each data point (i.e., vertex) $x \in M$ in this graph is $q(x) \triangleq \sum_{y \in M} k(x,y)$.
- Normalization of the kernel by this degree produces an $n \times n$ row stochastic transition matrix P whose elements are $p(x,y) = k(x,y)/q(x)$, $x, y \in M$, which defines a Markov process (i.e., a diffusion process) over the data points in M. A symmetric conjugate $\bar{P}$ of the transition operator P defines the diffusion affinities between data points by

$$\bar{p}(x,y) = \frac{k(x,y)}{\sqrt{q(x)q(y)}} = \sqrt{q(x)}p(x,y)\frac{1}{\sqrt{q(y)}} \qquad x, y \in M$$

- DM embeds the manifold into an Euclidean space whose dimensionality is usually significantly lower than the original dimensionality.
- This embedding is a result from the spectral analysis of the diffusion affinity kernel $\bar{P}$.
- The eigenvalues $1 = \sigma_0 \geq \sigma_1 \geq \ldots$ of $\bar{P}$ and their corresponding eigenvectors $\bar{\phi}_0, \bar{\phi}_1, \ldots$ are used to construct the desired map, which embeds each data point $x \in M$ into the data point $\bar{\Phi}(x) = (\sigma_i \bar{\phi}_i(x))_{i=0}^{\delta}$ for a sufficiently small δ, which is the dimension of the embedded space. δ depends on the decay of the spectrum $\bar{P}$.

Standard approach: Diffusion Maps (DM)

- Objects:
  - Dataset $D = \{x_i\}_{i=1}^n$
  - Metric $\|\cdot\|$ defined on D
  - (Normalized) Diffusion operator: $G_{ij} = e^{-\|x_i - x_j\|^2/\epsilon}$
- Mapping:

$$\Psi_t(x) = \left(\lambda_1^t \psi_1(x), \lambda_2^t \psi_2(x) \ldots, \lambda_m^t \psi_m(x)\right)^T$$

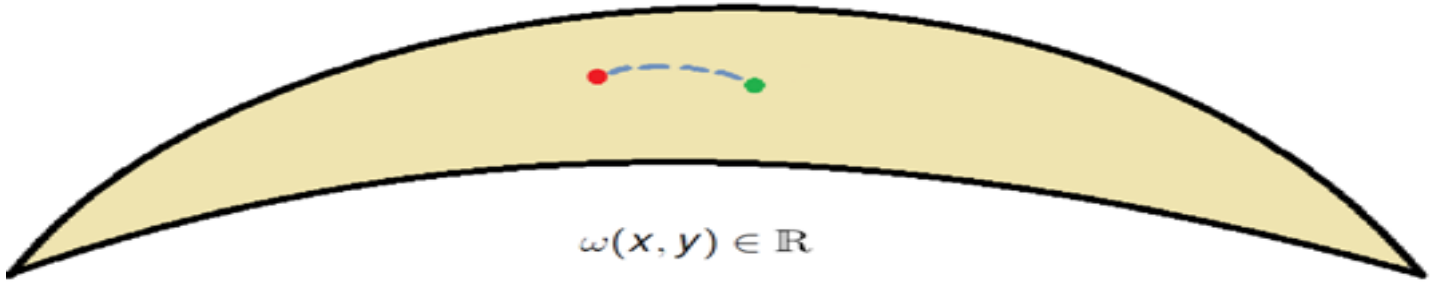
- Diffusion distances representation:

$$Diff_t(x, y) \approx \|\Psi_t(x) - \Psi_t(y)\|_{\mathbb{R}^m}$$

[2] R.R. Coifman, S. Lafon, Diffusion maps, Applied and Computational Harmonic Analysis, 21, 5-30, 2006.

Standard approach: Diffusion Maps (DM)

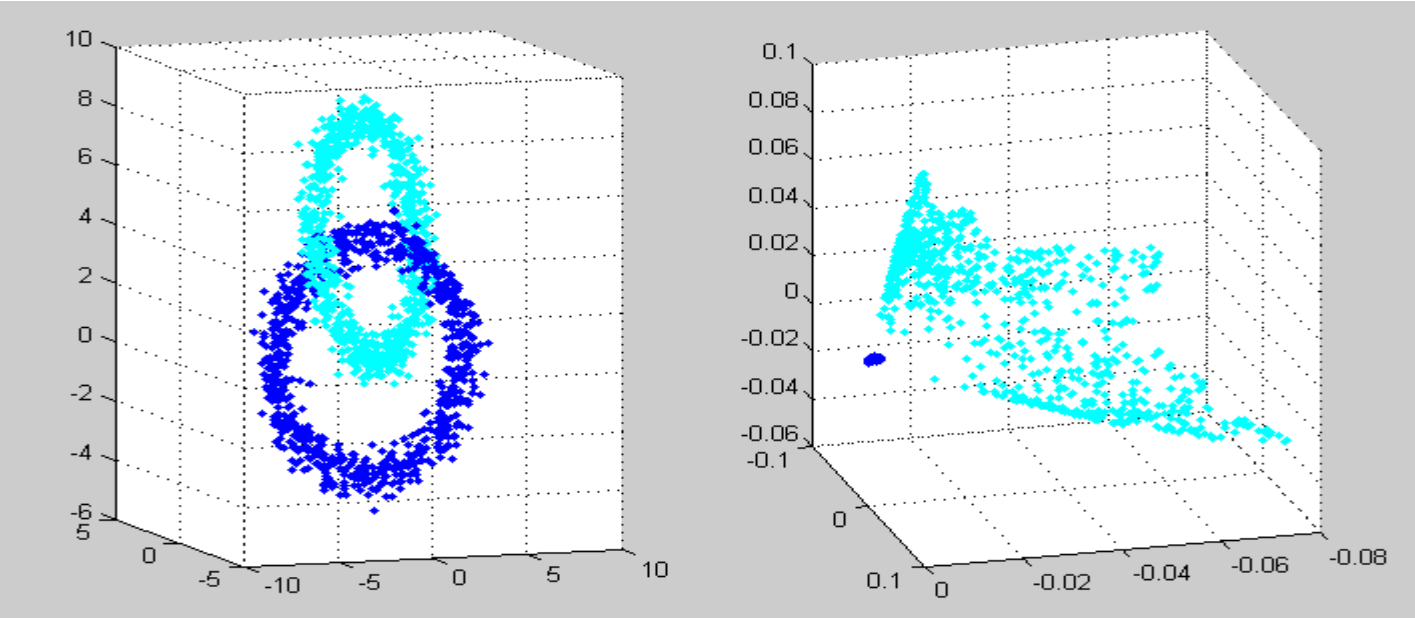
$$u \in M \mapsto \Psi(u) \in \mathbb{R}^{\ell \ll m}$$



It is easy to see that the map has the following properties:

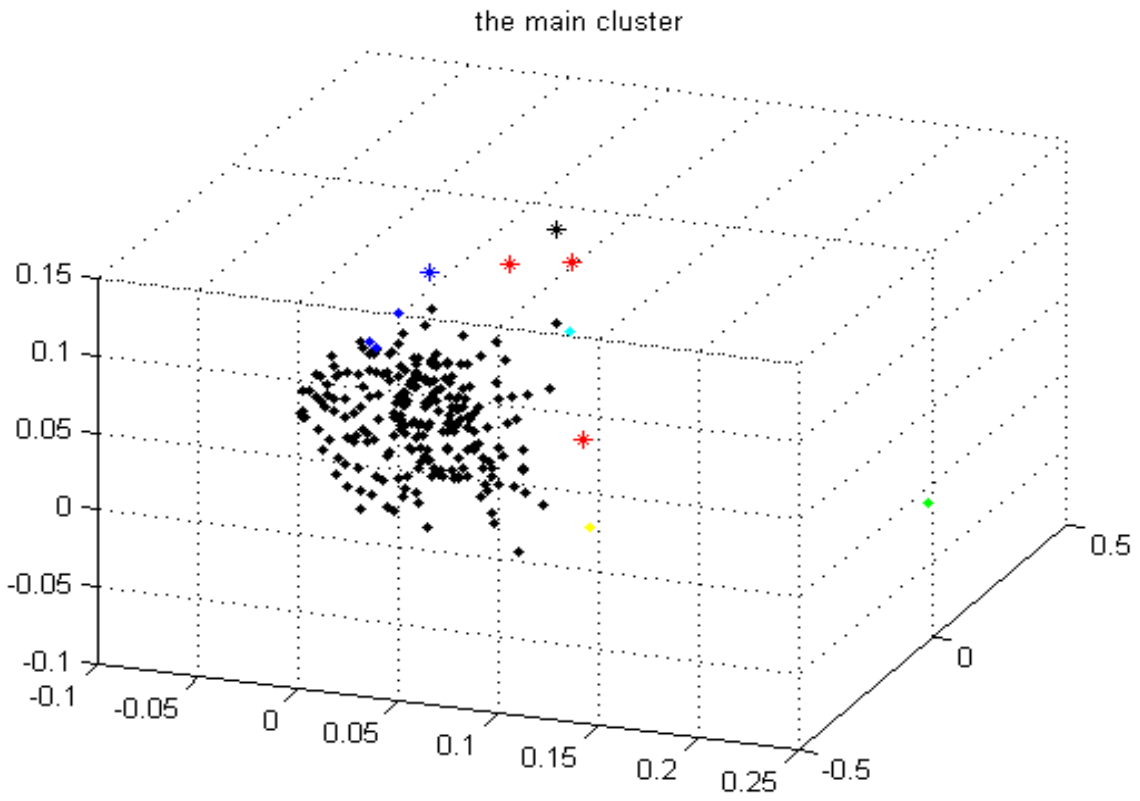
- The map represents the data in a space of dimension m.
- The map is not linear.
- The distance between the images of points is equal to the diffuse distance, that is, the probability to get from point x to point y via random walk on the graph for the time t.

Standard approach: Diffusion Maps (DM)



The figure illustrates the effectiveness of the separation of mixed known clusters via “diffusion maps”. If the generated data is represented as two interlocking rings (marked different shades of blue), no any linear methods is able to divide it. Nevertheless, a random walk on the graph represented by these rings, have ability to divide the classes. The probability remain inside the same ring by random walk is greater than the probability of jumping from one ring to another.

Diffusion Maps (DM): The problem

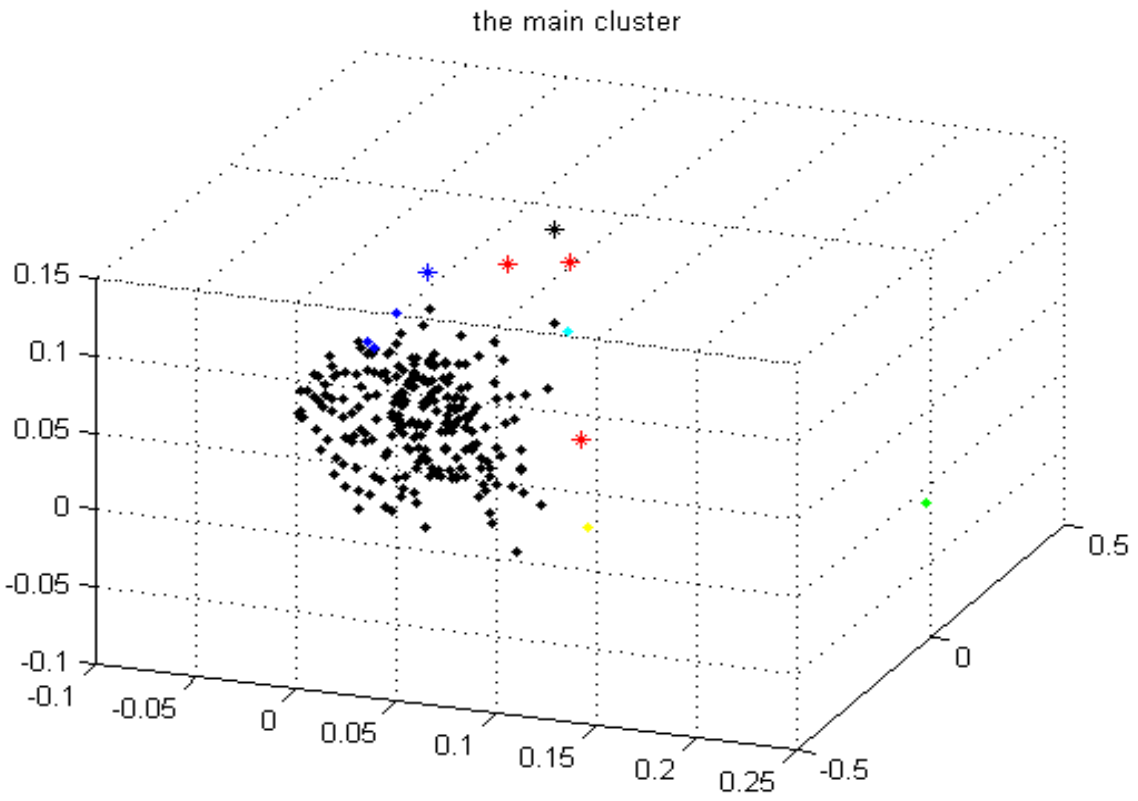


Classification background and anomaly?

Diffusion Maps (DM): The problem

BAD RESULT

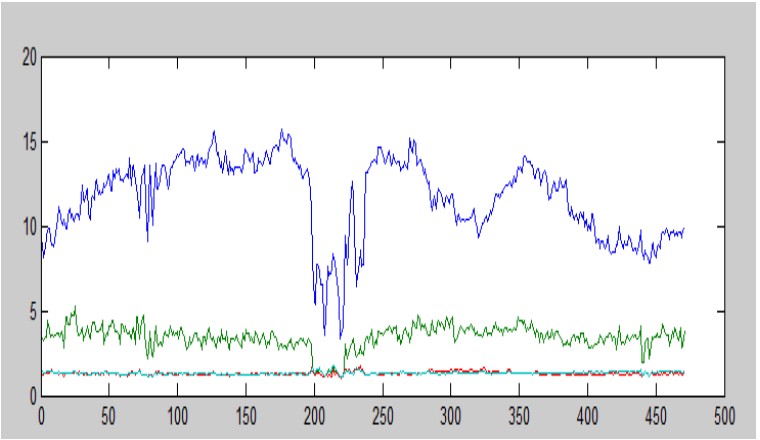
Diffusion Maps (DM): The problem

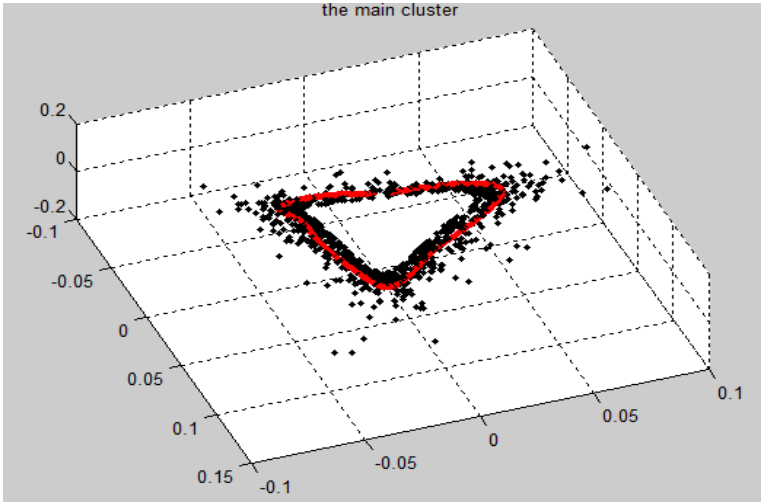


Anomalies are not grouped in clusters

Advanced approach: Homotopy in Temporal Diffusion Maps (DM)

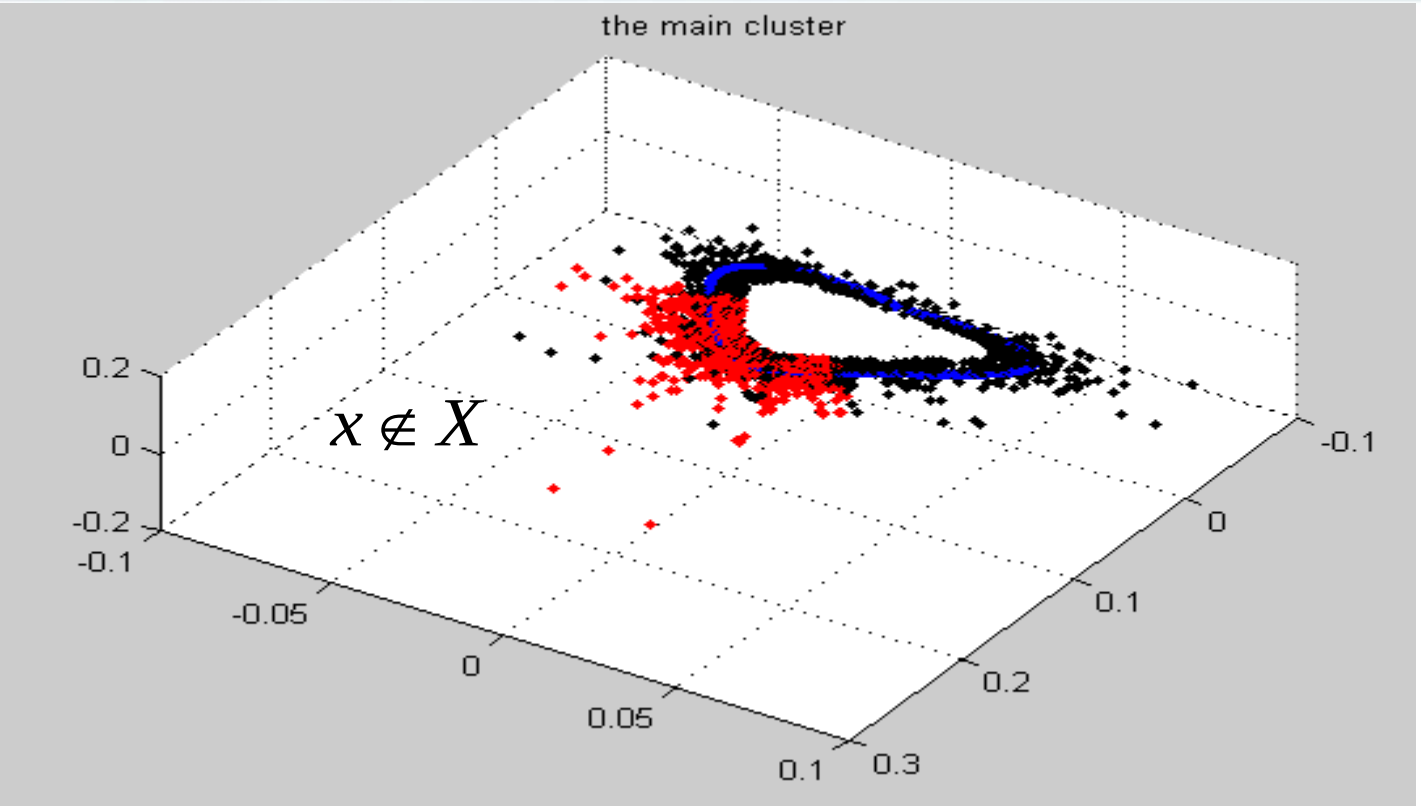
Diffusion operator $G_{ij} = e^{\frac{\left\|(i-j) \bmod D\right\|^2}{\sigma_1^2}} e^{\frac{\left\|x_i - x_j\right\|^2}{\sigma 2^2}}$





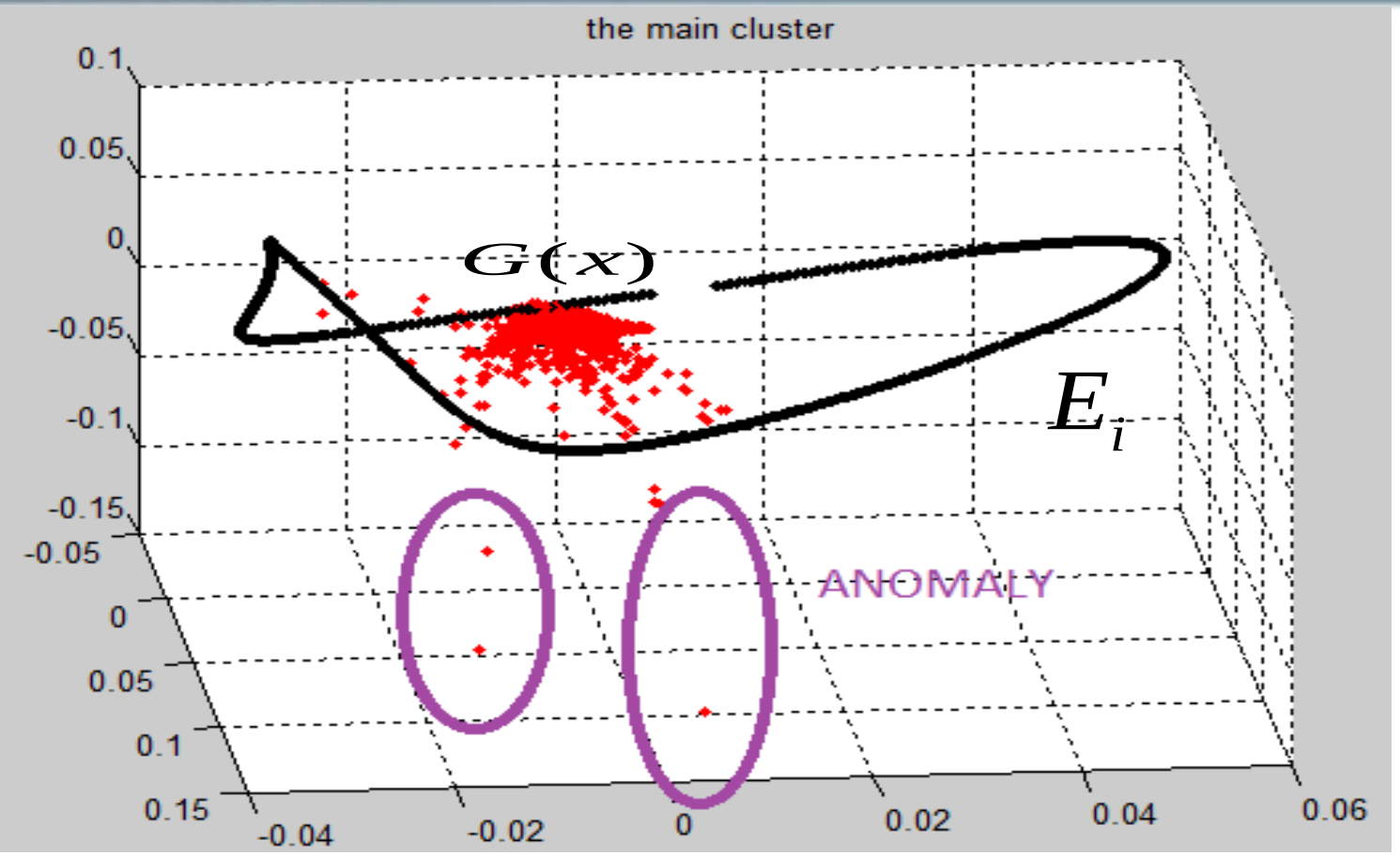
The diffusion geometry is oriented around a smooth parametric curve. The curve represents the day and night

Advanced approach: Homotopy in Temporal Diffusion Maps (DM)



Once X is mapped - extension of $\overline{f}$ to $x \notin X$, *using* representatives from X (sampling)

Advanced approach: Homotopy in Temporal Diffusion Maps (DM)

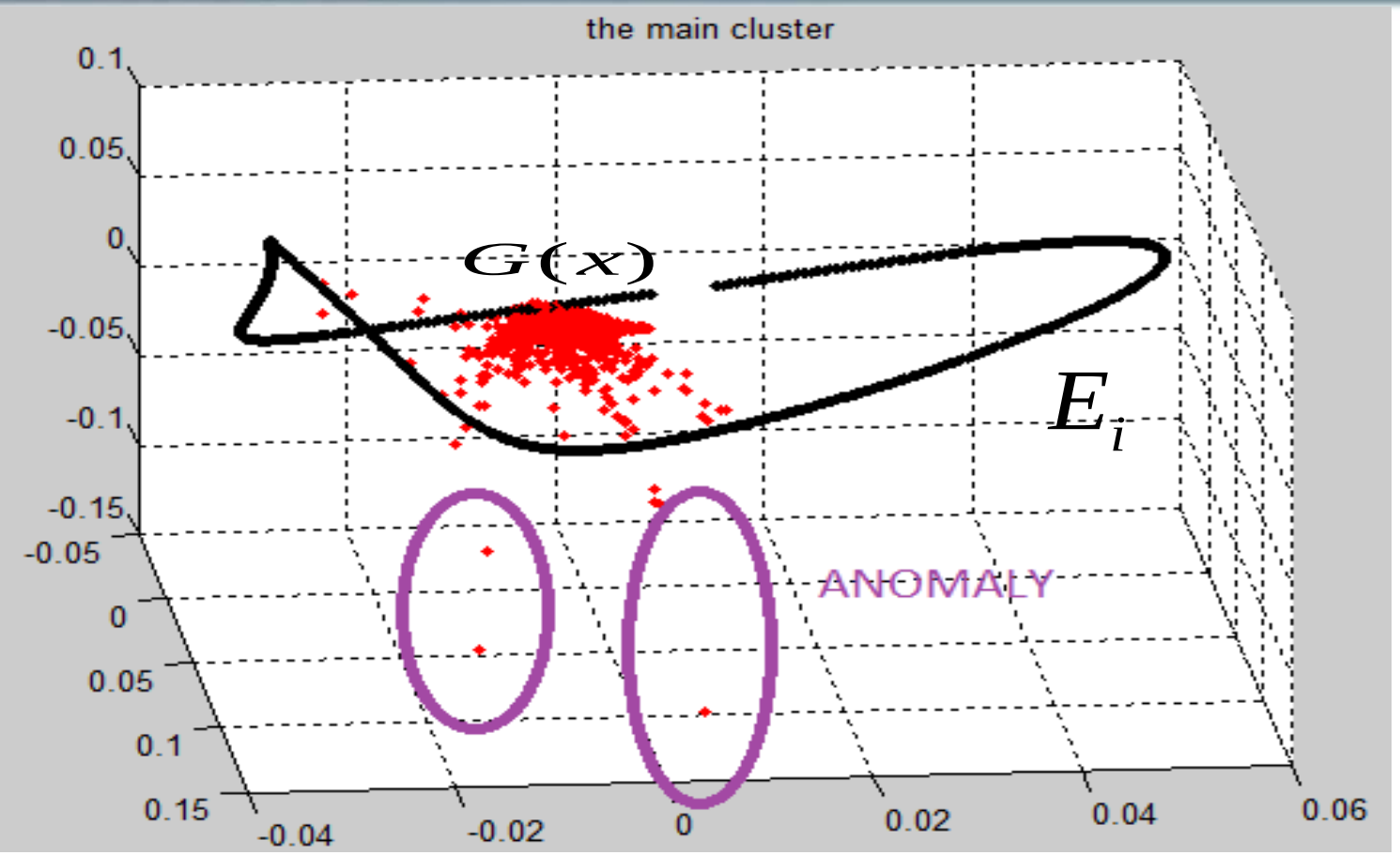


Let E_i be approximating curve and $x \notin X$

Define homotopy G(x)

$$G(x) = x - \frac{\sum_i E(i)\rho(x, E(i))}{\sum_i \rho(x, E(i))}$$

Advanced approach: Homotopy in Temporal Diffusion Maps (DM)



Let E_i be approximating curve and $x \notin X$

Define homotopy G(x)

$$G(x) = x - \frac{\sum_i E(i)\rho(x, E(i))}{\sum_i \rho(x, E(i))}$$

Advanced approach: Alpha-stream process for anomaly detection

Let $\Xi = \{s_1, s_2, ..., s_k\}$ be the set of all data after projection to the DM and homotopy transformation. Define the weight function $\alpha : \Xi \to [0, 1]$ such that if $\alpha(s_i) = 1$ then the pixel s_i belongs to background and if $\alpha(s_i) = 0$ then the point s_i is anomaly.

The weight function α we will search through maximization of the functional

$$F(\alpha) = \Sigma_i \chi_i^{\Theta} \alpha_i + \Sigma_{ij} W_{ij} (\alpha_i - \alpha_j)^2 - \|\alpha\|^2,$$

where $W_{ij} = exp(- \| s_i - s_j \|^2)/\tau^2$, Θ is an approximately background defined on step 1, $\chi^{\Theta}(s_i) = \begin{cases} 1, & s_i \in \Theta; \\ 0, & s_i \notin \Theta. \end{cases}$ be the the characteristic function of the set Θ.

Advanced approach: Alpha-stream process for anomaly detection

The functional $F(\alpha)$ can be written in the form

$$F(\alpha) = \langle \alpha, \chi^{\Theta} \rangle + \alpha^T L \alpha - \|\alpha\|^2,$$

where the $d_{ii} = \Sigma_j W_{ij}$, and D be an diagonal matrix with elements d_{ii}. The "Laplassian" matrix L is defined by $L = D - W$. The maximum of the functional is reached by

$$\frac{\partial F(\alpha)}{\partial \alpha} = 0,$$

that mean

$$0 = (\chi^{\Theta} + L\alpha - 2\alpha).$$

this equality implies the following iterative process (Gradient descent algorithm).

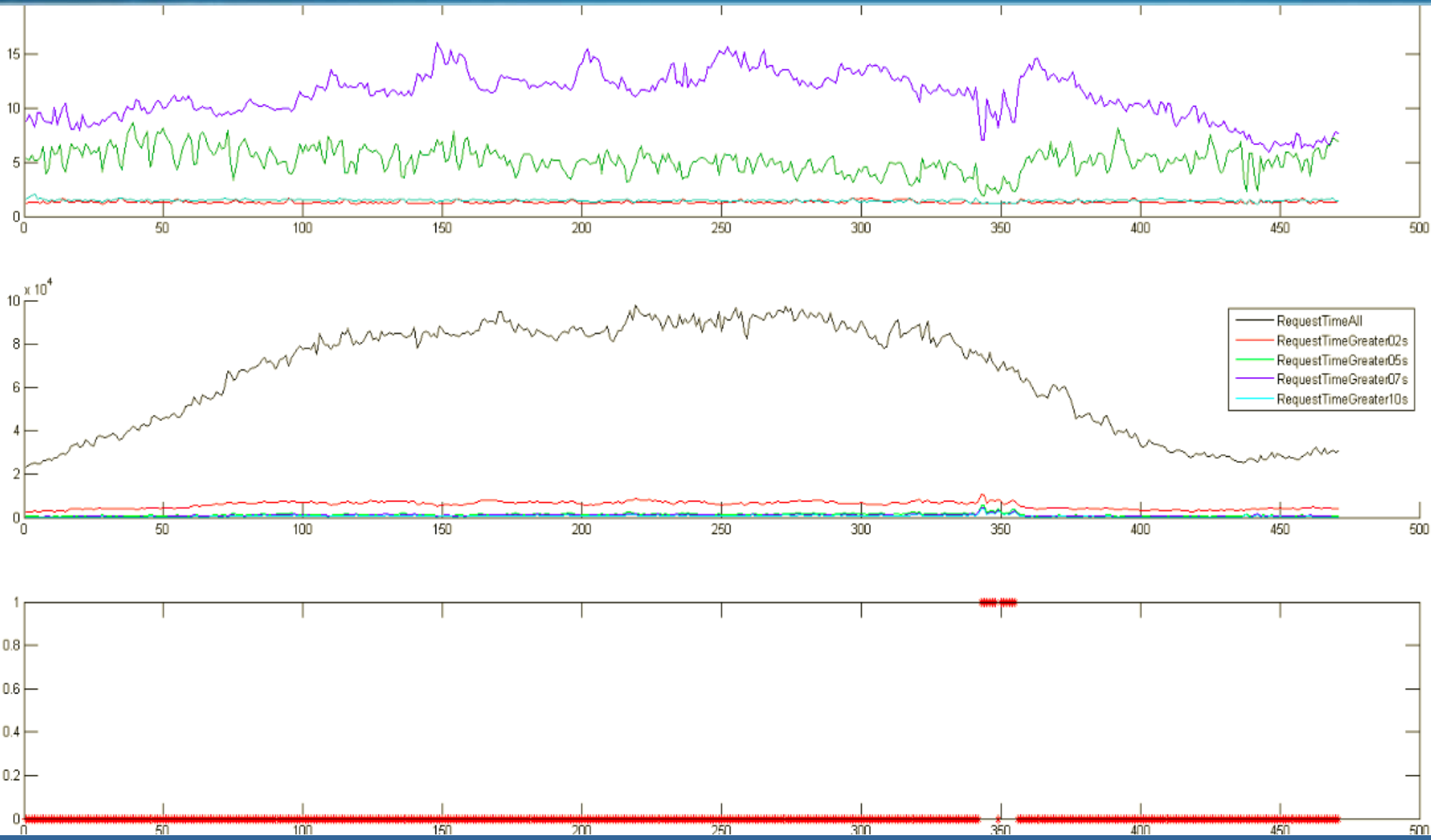
$$\frac{\partial \alpha(t)}{\partial t} = \chi^{\Theta} + L\alpha(t) - 2\alpha(t).$$

So, we get the iterative "alpha-stream" process with $\alpha_0 = \chi^{\Theta}$

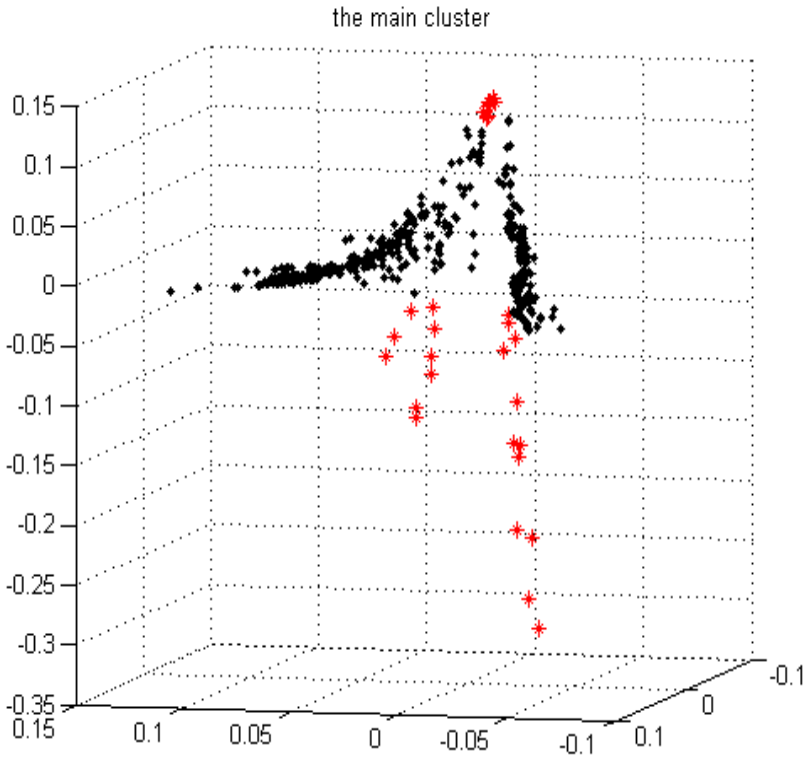
Advanced approach: Alpha-stream process for anomaly detection

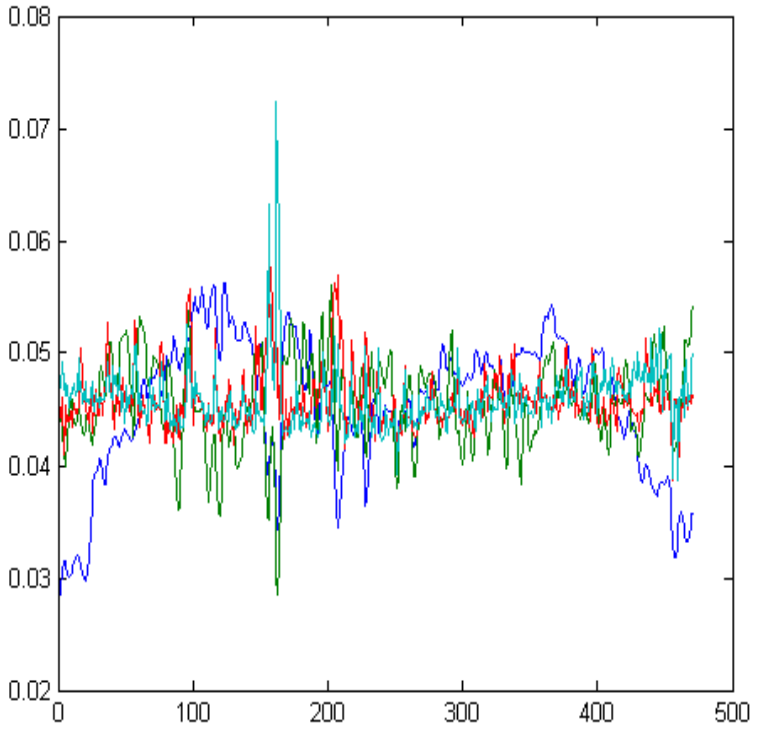
Image Processing application of “alpha-stream”: Object segmentation

The results:


x 10^4
RequestTimeAll
RequestTimeGreater02s
RequestTimeGreater05s
RequestTimeGreater07s
RequestTimeGreater10s

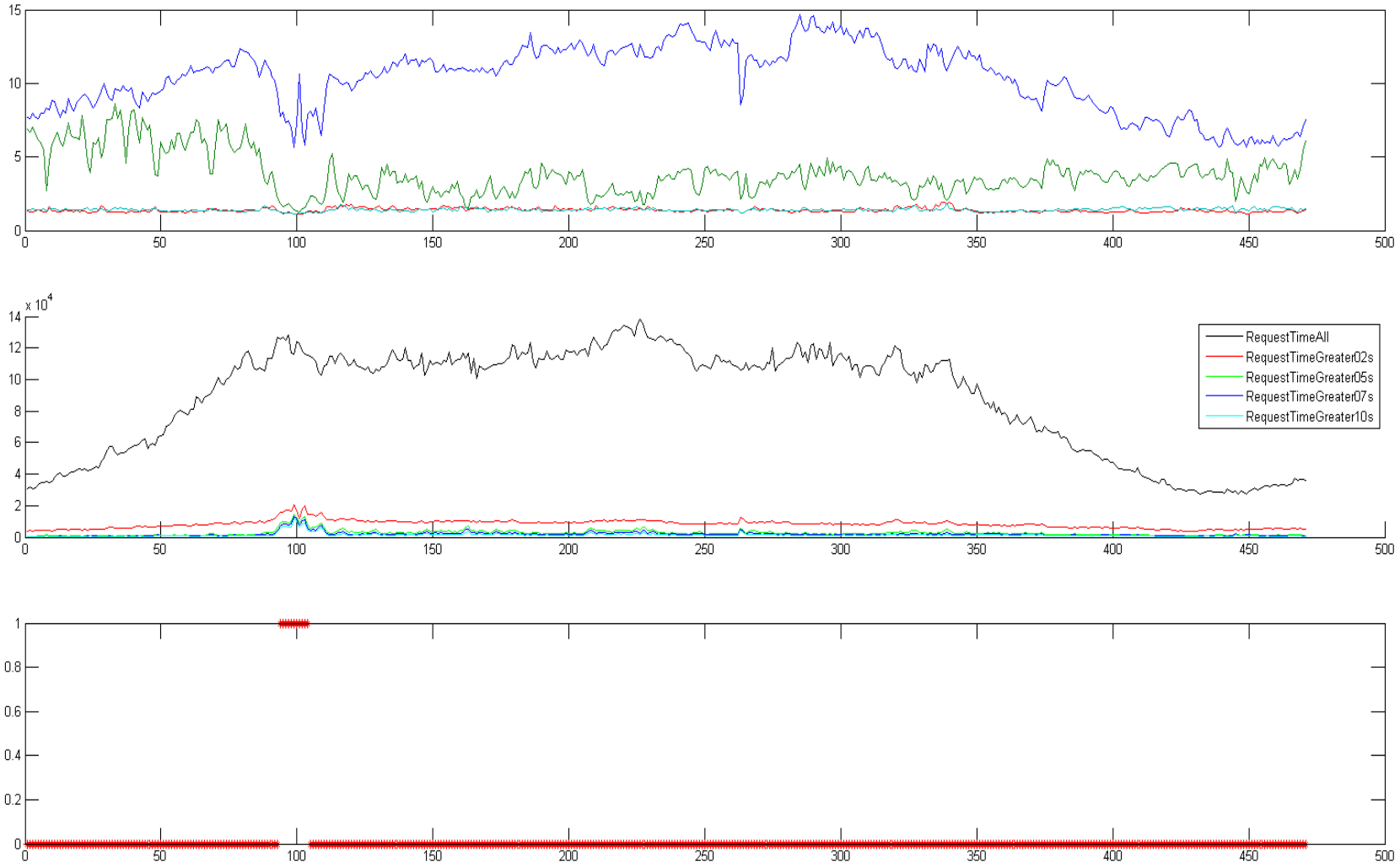
The results:



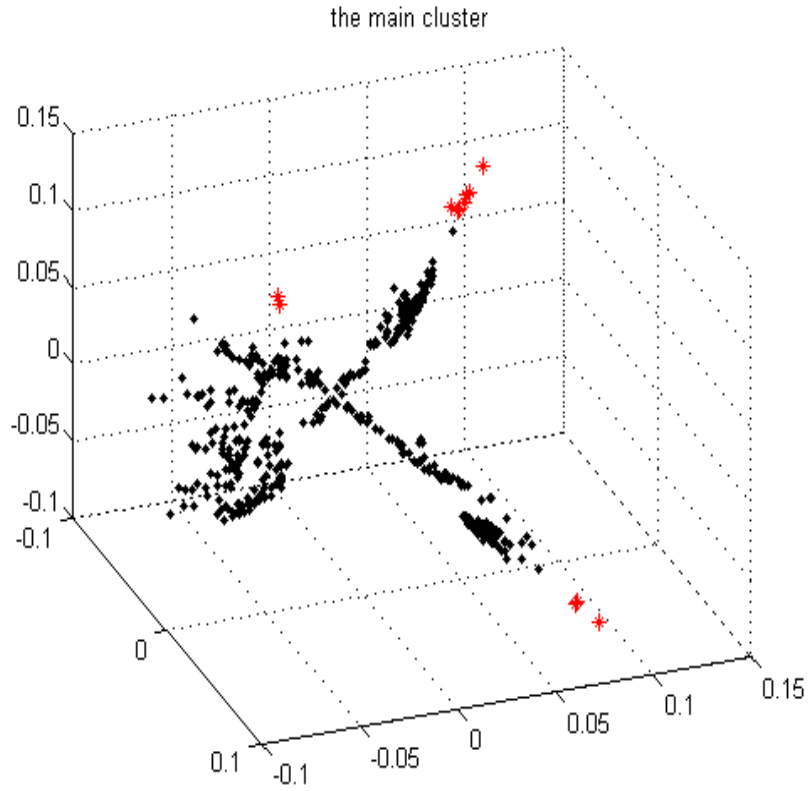


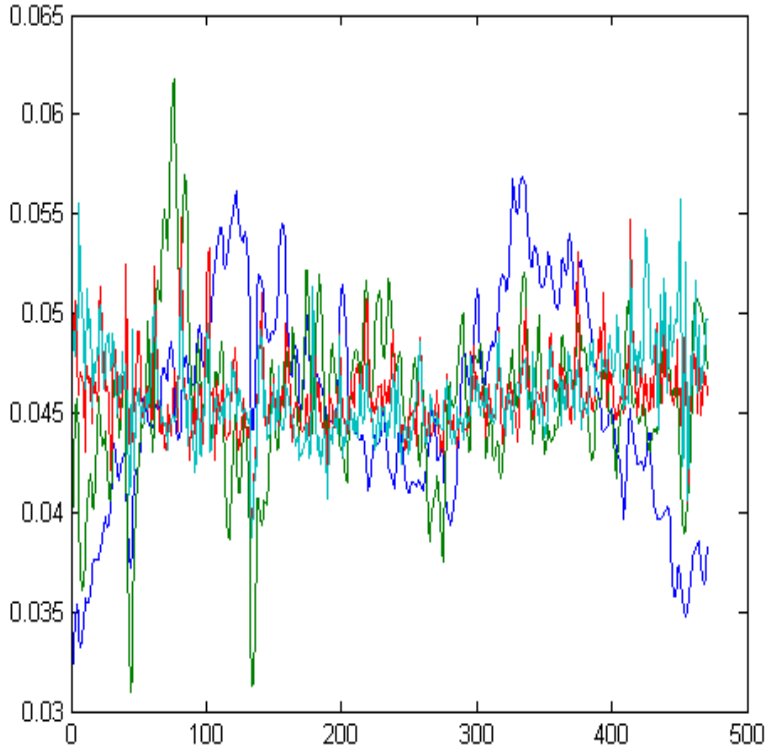
The features(left) and its representation in DM (right)

The results:


x 10^4
RequestTimeAll
RequestTimeGreater02s
RequestTimeGreater05s
RequestTimeGreater07s
RequestTimeGreater10s

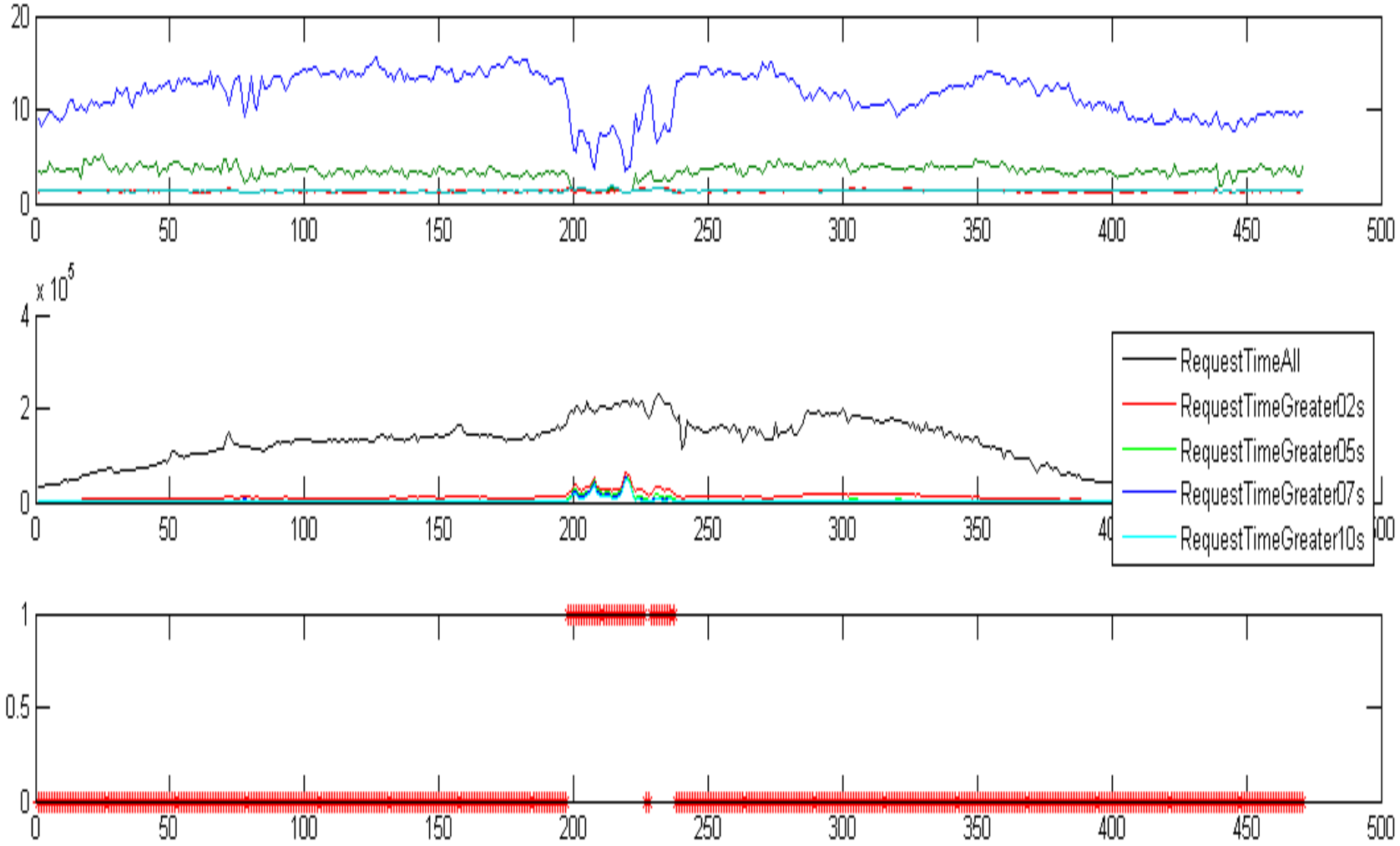
The results:





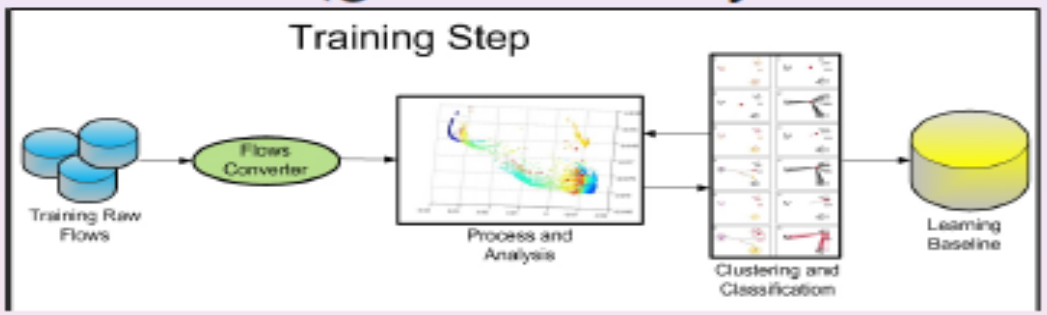
The features(left) and its representation in DM (right)

The results:

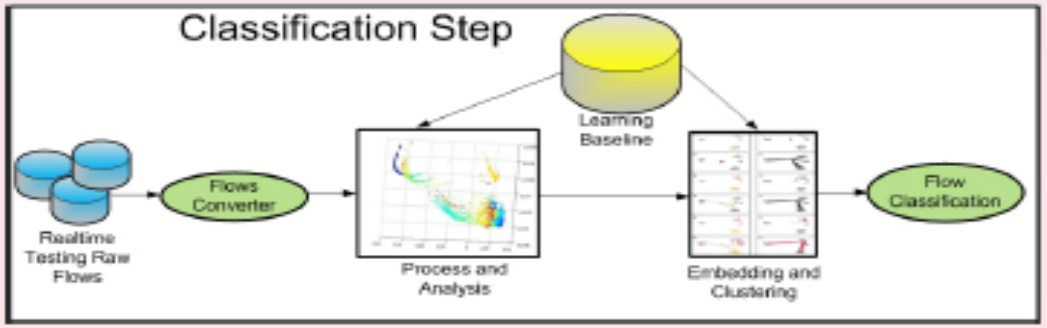

20
10
0
0
50
100
150
200
250
300
350
400
450
500
x 10^5
4
2
0
RequestTimeAll
RequestTimeGreater02s
RequestTimeGreater05s
RequestTimeGreater07s
RequestTimeGreater10s
1
0.5
0

Comparison to SNORT

- Tested on datasets that contain real network traffic that were captured from a big enterprise network.
- The inputs are the statistics of the sessions of these datasets (generated by flow-oriented traffic analyzer).



- Training dataset contains 5,500 samples.
- Testing dataset that contains 2,000 samples.



Comparison to SNORT

	Snort	QRATOR
Number of tested URLs	773,841	773,841
Number of false alarms	~17,000	292
Rate of false alarms	~2.2%	~0.038%
Number of detected attacks	250	680

Detection rate: 172% more than SNORT
False positive rate: 90% less than SNORT