Cyber War Protection via High Dimensional Processing

Amir Averbuch

Anomaly detection identifies patterns that do not conform to an established normal behavior. The detected patterns are often translated to critical and actionable information in several application domains since they deviate from their normal behavior. The main challenges anomaly detection algorithms face are: achieving low false alarm rate, clustering into normal and abnormal regions, stealthiest by a malicious adversary, dynamic and evolving data, portability of the techniques between different domains and applications, availability of labeled data for training/validation and noisy data that tends to be similar to real anomalies.

Usually, in addition to the challenge of detecting anomalies in a dataset, the analyzed data is also high-dimensional. Since the data in most modern systems can be described by hundreds and even thousands of parameters, then the dimensionality of the data is very high and its processing becomes impractical. ``Curse of dimensionality'' is associated with high-dimensional data due to the fact that as the dimensionality of the input data space increases, it becomes exponentially more difficult to process and analyze the data.

Furthermore, adding more dimensions can increase the noise, and hence the error. This problem is a significant obstacle for high-dimensional data analysis, since a local neighborhood in high dimensions is no longer local. Therefore, high-dimensional data is incomprehensible to understand, to draw conclusions from or to find anomalies that deviate from their normal behavior.

Anomaly detection in high-dimensional data is in extensive use in a wide variety of areas. For example, in communication networks it is used for identifying intrusions by internal or external users, for detecting faulty components and for recognition of unauthorized protocols and transactions. Another challenging domains are financial applications, intelligence systems, critical systems that contain multi-sensors, to name some.

In the talk, we focus on anomaly detection in high-dimensional data for the following major domains:

1. Intrusion detection/prevention systems have become an integral component in security systems. The challenge is to perform online protection without miss-detections and false alarms. To achieve it, most systems are based on signatures of intrusions that are developed and assembled manually after a new intrusion is exposed and distributed to the security clients. This approach is problematic because these systems detect only already known intrusions (yesterday's attacks) but they fail to detect new attacks (zero day attacks).
2. Network traffic classification and recognition is a critical component in many Internet applications such as traffic control, identification of specific applications, guarantee of QoS, etc. Until recently, classification was mainly done by inspecting the payload of traffic packets, checking for an application signature. While payload inspection techniques have worked well in the past, they suffer from major limitations: applications such as Skype use techniques

to avoid protocol identification, payload becomes encrypted and many new protocols per year are introduced. Therefore, the reactive development of a signature for each new protocol is a challenging task on one hand and impractical on the other hand.

We introduce a unique framework that is based upon diffusion processes, diffusion geometries, random projections and other methodologies for finding meaningful geometric descriptions in high-dimensional datasets. We will show that the eigenfunctions of the generated underlying Markov matrices can be used to construct diffusion processes that generate efficient representations of complex geometric structures for high-dimensional data analysis. This is done by non-linear transformations that identify geometric patterns in these huge datasets that find the connections among them while projecting them onto low dimensional spaces. Our methods automatically detect the anomalies that deviate from normal behavior.

We will also introduce the Localized Diffusion Folders methodology for clustering and classification of high-dimensional datasets. The diffusion folders are multi-level partitioning of the data into local neighborhoods that achieved by several random selections of data points and folders in the graph and by defining local diffusion distances between them. This multi-level partitioning defines an improved localized geometry of the data and a new localized Markov transition matrix that is used for the next time stage in the diffusion process.
The result of this clustering method is a bottom-up hierarchical clustering of the data while each level in the hierarchy contains localized diffusion folders of folders from the lower levels. The performance of the proposed algorithms are demonstrated on data that was collected from real networks.