UNCOVERING UNKNOWN UNKNOWNS (U3) IN CRITICAL INFRASTRUCTURES

WHITE PAPER

THETARAY ANALYTICS PLATFORM

Legal Notice:

UNCOVERING UNKNOWN UNKNOWNS (U^3) IN CRITICAL INFRASTRUCTURES

ThetaRay's vision is to efficiently transform how the world benefits from high dimensional big data (HDBD) to be prepared for U^3 threats and risks in critical infrastructures. In the flood of data (the digital universe will grow from 130 billion GB in 2005 to 40 trillion GB by 2020) lie tremendous opportunities. Big data is the gold mines of the 21st century. ThetaRay provides unique HDBD detection by processing almost any type of data source and benefiting from thousands of dimensions (features). The same anomaly detection core technology is applicable to any critical infrastructure that includes energy, transportation, telecommunications, industrial and financial institutions to name a few.

The ever-increasing, paramount threats to critical systems faced by many organizations across different industries are the unknown unknowns. This means threats that you are not aware of, and don't even know that you are not aware of them. In cyber security for example, anomalies may point at a cyber-attack (presence of malware). In industrial settings, the detection of anomalies in large quantities of operational data may point towards failures in critical machinery. In the financial industry, anomalies can point to potential fraud, credit risks or money-laundering activities.

It is not easy to detect anomalies, especially as over the past years the volume, speed and variety of the available data has increased dramatically. With the explosion of new HDBD, the demand for data-driven analytic methods for uncovering data insights, such as performing anomaly detection, has increased. ThetaRay satisfies this demand.

ThetaRay data-driven analytic methods provide universal, unified and a generic (one) solution to 3 verticals:

- **Cyber** - The standard approach for securing (critical) infrastructure over the past 50 years, classified as "walls and gates", has failed. There is no longer any reason to believe that a system of barriers between trusted and untrusted components with policy-mediated pass-throughs, will become more successful as the future unfolds. Within the security context, widely-used traditional rule-based detection methodologies, including firewalls, signatures/patterns that govern IDS/IPS, and antivirus are irrelevant for the detection of new and sophisticated malware in today's world. These viruses, worms, back doors, spyware, and Trojans are masked as legitimate streams and penetrate every state-of-the-art commercial barrier on the market – there are half a million attacks every second. These systems are incapable of accurately detecting zero-day attacks and advanced persistent threats (APT), which contain previously un-encountered signatures and do not play by known rules. In other words, traditional solutions are only able to detect yesterday's attacks - ones encountered in the past, and ones for which they are looking. State-of-the-art solutions lack the ability to perform U^3 with low false alarm and high detection rates.

- **Industrial process -** The Industrial Internet (*II)* market, a term coined by General Electric and considered as the industrial subset of IoT, is the result of the convergence of the previous two revolutions: The Industrial Revolution and The Internet Revolution, and refers to *the integration of complex physical machinery with networked sensors and software*. The concept behind *II* is a combination of "machine to machine" connectivity alongside "big data analytics" and "anomaly detection", which together create the ability to monitor machines' performance in near-real time and in almost every sector and location. The consumer-based internet the world uses today connects 2 billion users and the *Industrial Internet* plans to connect 50 billion devices. Examples of industrial internet use cases are: prediction of aircraft engine failure, wind and regular turbines operational and real-time profit optimization and reduction of maintenance costs, identification of anomalous activity (operational or security) in power generation machinery to avoid costly power outages. There are many more uses spanning multiple sectors such as energy, healthcare and transportation. The ability to interconnect machines safely will allow organizations to significantly increase operational efficiency and decrease unplanned downtime. Efficient *II* processing requires a massive scale up of tools for HDBD understanding. ThetaRay supplies these tools.
- **Financial sector** - facing constant escalation of internal and external threats. The growing variety of service channels that mix physical and cyber access, like branches, ATMs, online services, and mobile banking and the unique risks tied to each, complex internal operations as well as stringent regulatory requirements, together add to the challenge of managing and mitigating operational risks. The number and complexity of Cyber-based fraud schemes directed towards commercial banking clients is constantly increasing, resulting in significant financial loss as well as an impact on reputation. Massimo Cotrozzi, the assistant director of EY's fraud investigation and dispute services team, says about 90% of fraud is committed using cyber methods. "In an increasingly digital world, where processes are carried out through computers, risks arise around security and hacking," he explains. The unique complexities of mitigating commercial banking fraud and the evolving legal and regulatory environment dictate a need for effective fraud management systems for Banks, insurance agencies, credit card companies, stock exchanges and many other financial and financial-related governmental agencies.

ThetaRay's technology is based on over 10 years of deep basic and applied academic research from Tel Aviv University (Prof. Amir Averbuch, ThetaRay CSO and co-founder) and Yale University (Prof. Ronald Coifman, ThetaRay co-founder and now on the advisory board) and an excellent algorithmic group lead by Dr. David Segev and Dr. Gil Shabat. ThetaRay's core technological infrastructure uses applied and computational harmonic analysis, differential geometry, stochastic processing (random walk, Brownian motion, diffusion processing), low rank matrix decomposition, randomized algorithms, classical analysis, geometric measure theory, manifold learning, spectral graphs, kernel methods, graph theory, dictionary constructions and deep learning, big data analytics, Hadoop, Spark and MapReduce to efficiently detect anomalies in HDBD.

Detection is achieved by unsupervised and automatic methods that are not based on any rules, patterns, signatures, heuristics, data semantics of the features or any prior domain expertise. The algorithms are based on unbiased detection through a series of randomized advanced algorithms that can process any number of features that include besides numeric data also Boolean, ordinal or categorical data. These solutions are able to operate completely un-supervised on regular computational devices built from off-the–shelf components with zero configuration and no threshold settings while providing scoring and parameter rating to support advanced forensic capabilities. They run in real-time processing detecting threats/risks in seconds utilizing a growing number of technologies such as cloud computing, distributed and parallelization via GPU, massive amounts of on-demand data storage and links to Hadoop, Spark and Splunk. The technology does not violate privacy e.g. when communication networking data is analyzed, only the metadata is used without exploring the payload.

The algorithms originated from different schools of thought. The same core infrastructure technology fits many different verticals/problems/applications from diverse domains. This generic solution of multiple solutions, reduces the false alarm rate and increases detection rates.

ANOMALY DETECTION ALGORITHMS

This section describes the anomaly detection algorithms that exist within ThetaRay's Analytics Platform. **All are covered by patents.**

Algorithm I: Since machines/devices in industrial environment may have different concurrent modes of operations, the algorithms should be applied differently and separately to each mode of operation. Therefore, the data is clustered to identify the different modes. Multiscale clustering is applied to 3 scales to refine the clustering. Then, the below algorithms are applied separately to each cluster in the multiscale decomposition.

Algorithm II: Diffusion geometry (kernel method). Diffusion process (Brownian motion based modeling) is applied to the training data to get diffusion distances. These distances are converted into a Markov matrix. The eigenvalues of the Markov matrix, which do not tend to 0, determine the size of the reduced dimensionality space. This way, the anomaly detection will be performed on a lower dimensional space. Each newly arrived multidimensional data point will be embedded into the lower dimensional space by the application of out-of-sample extension (Nystrom and others) to find whether it belongs to the manifold or deviate from it. Deviated data points are classified as anomalies. **This methodology reduces the number of features.** This is a kernel-based method.

Algorithm III: Measure based (kernel + data point distribution (no manifold)): It resembles **Algorithm II** but here instead of finding the manifold in the low dimensional space, anomalies are detected according to the distributions of multidimensional data points.

Algorithm IV: Randomized LU decomposition for dictionary building. This is a randomized algorithm that is based on a low-rank matrix decomposition by LU decomposition. The L part is used to build a dictionary from the training data. Every newly arrived multidimensional data point that is not spanned well by the dictionary is classified as an anomalous data point. This methodology reduces the number of measurements. It has the following properties: It is parallelizable, fits well GPU, fast, low memory consumption and fits embedded devices.

Algorithm V: Hybrid (dictionary+kernel): It is built from three successive applications of the above algorithms in the following way: multiscale clustering (**Algorithm I**) is applied first, followed by the application of diffusion geometry (**Algorithm II** or **Algorithm III**), followed by the application of Randomized LU (**Algorithm IV**). These successive applications of the algorithms generate new features that reduce substantially the size of the problem. Then, this reduced problem is submitted to the one of the kernel based algorithms (**Algorithm II** or **III**) and which make them very efficient.

Algorithm VI: Gaussian mixture models (GMM). It is based on the application of whitening PCA to reduce the dimension of the problem, and by modeling via t-distribution, the anomalies

(outliers) are found. It combines the applications of whitening PCA, Expectation-Maximization (EM) algorithm and t-distribution.

Algorithm VII: Fusion of the Algorithms II-VI into one output

Algorithms I-VII can run either in offline or online mode. Offline means that the above algorithms are being applied to all of the data at once. In other words, the input data is the training data only. In online processing, the training is processed, and then any newly arrived data point, which did not participate in the training, is classified against the training to be normal or abnormal.

Forensic: Since there are neither rules nor signatures, we provide algorithms that achieve fast forensics:

- **Algorithm VIII: Anomalies scoring.** Scores the severity of the anomaly.
- **Algorithm IX: Parameter rating.** Algorithm that rates (i.e., ranks) from top to down the parameters that cause a specific abnormal behavior to occur.

All the above algorithms are implemented on ThetaRay Platform:

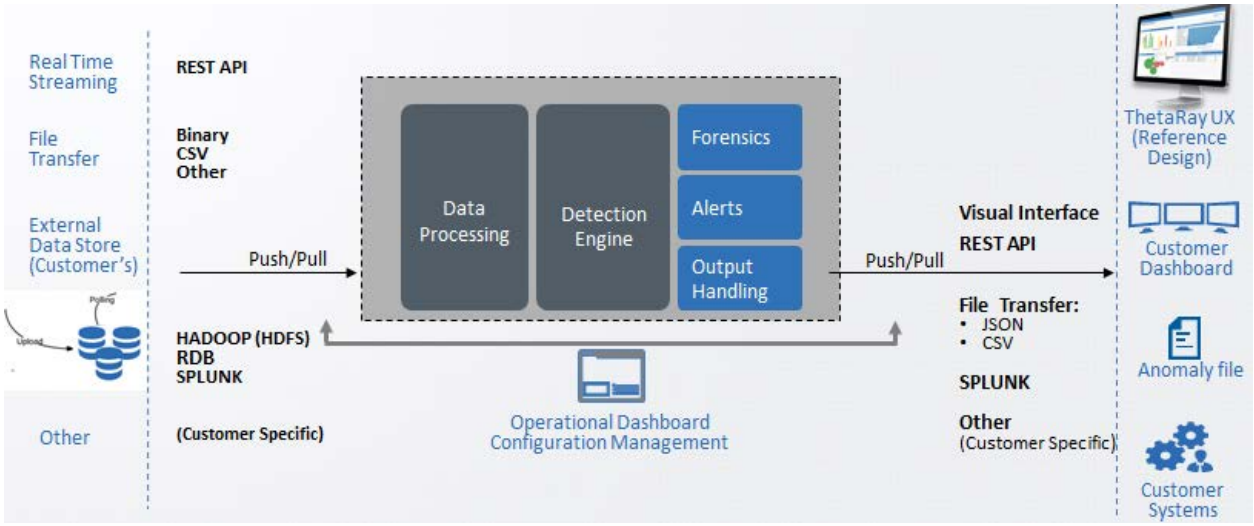

Figure 1: High Level Architecture Overview

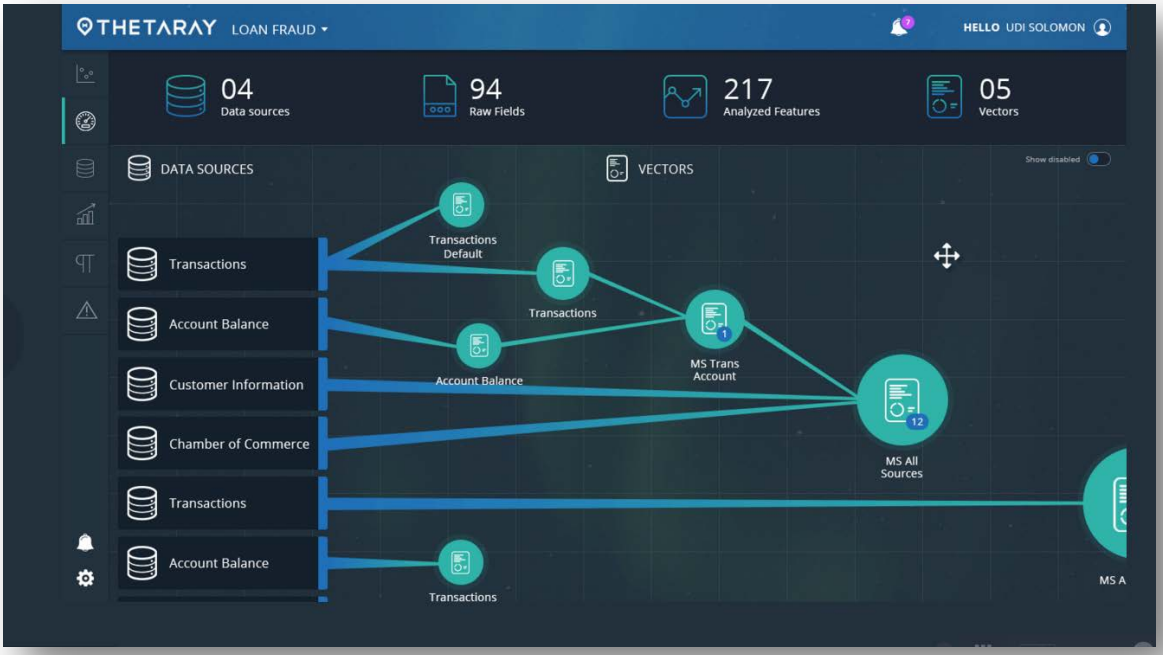

Figure 2: Configuration Dashboard: Displaying Data Sources and Configured Vectors

Figure 3: Detection Center Results Analysis

International Headquarters
8 Hanagar Street
Hod Hasharon, Israel 4501309
Tel: +972-72-274-4999

USA Headquarters
136 Madison Avenue
New York , NY 10016
Tel: 1-678-362-9878