SAFARI
Situational Awareness Framework for Risk Ranking

Future Work Ideas

Situational Awareness for Fraud Detection

SAFARI introduces the concept of Situational Awareness (SA) to enable the detection of fraud on hundreds of thousands of unlabeled payment registers where ground truth is not available, by exploiting different perspectives of available data. The goal is to build a path from massive data to information to understanding, allowing for appropriate analysis and sharing at each point of the fraud detection process.

Existing link analysis solutions, such as those from Palantir, SAS Institute, and Centrifuge Systems, offer link analysis capabilities to integrate different sources of information. However, networks have to be manually built and anomalies have to be manually discovered, severely limiting the quantity of data and the insight that can be achieved in a single session.

Unlike existing solutions, SAFARI implements a novel yet intuitive Red Flag Network (RFNet) approach to automatically connect disparate but potentially related entities and anomalies to build fully contextualized potential fraud cases from raw data and automatically detected anomalies.

Additionally, SAFARI implements a user interface inspired on the ideas of a dashboard system to show a complete set of information (relationships, geographical locations, tabular data) and facilitate the prioritization of tasks, the evaluation of red flags and the sense making process of fraud scenarios. For maximum reach, the graphical user interface is delivered by using state-of-the-art Web technologies and Internet protocols.

Current Approach

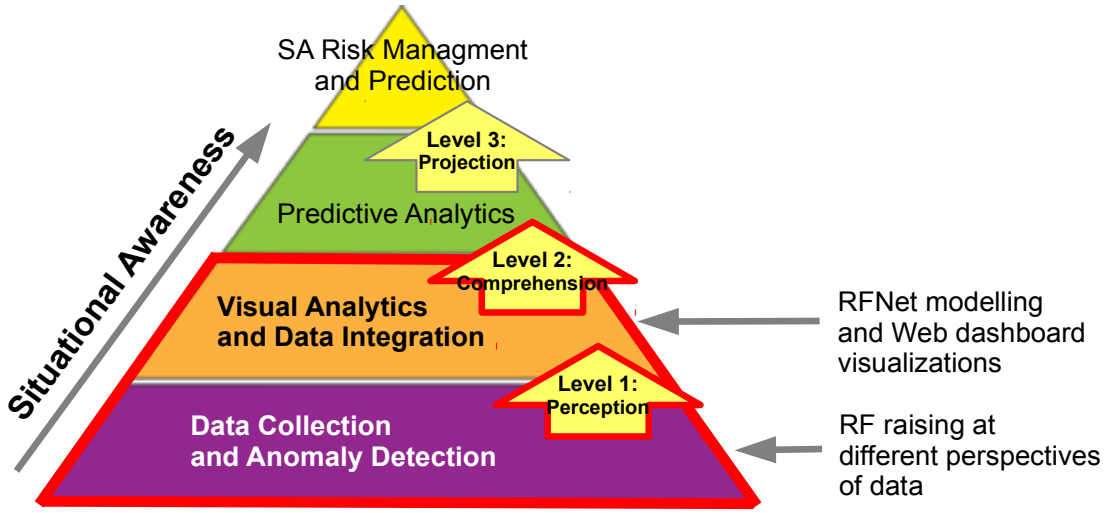

Figure 1. SAFARI SA approach (Level-1 and Level-2).

t
The MIT GDC is exploring the SA approach for fraud detection by exploiting a software prototype that implements the first and second levels of a three-level SA solution (Fig. 1):

> *SA Level-1: Environment perception.* This is the basal constituent of SA: the data is analyzed by exploiting domain-specific anomaly detection techniques on different perspectives of the data (personal information, monetary quantities, spatial information, etc.) in order to produce red flags (RFs) that point out to suspicious yet not conclusive and disparate behavior on unlabeled data.
>
> *SA Level-2: Situation comprehension.* Is the combination, interpretation, and retention of anomalies to form a coherent picture of the situation whereby the significance of events is understood. This is achieved in SAFARI by exploiting the RFNet approach and rich Web-based Visual Analytics techniques.

The goal is to introduce the concepts of Level-1 and Level-2 SA-based fraud analysis in order to automatically produce suspicious payments scenarios, prioritize red flag examination through dashboard-based visualizations, minimize the quantity of data that data analysts have to examine and improve the results produced by current ACL-like workflows in terms of time and effort expended by SMEs.

The current SAFARI platform can be upgraded and extended in the future to take advantage of new advances and add more capabilities in data and visual analytics for fraud detection.

Future Work:
Predictive Analytics for Fraud Detection

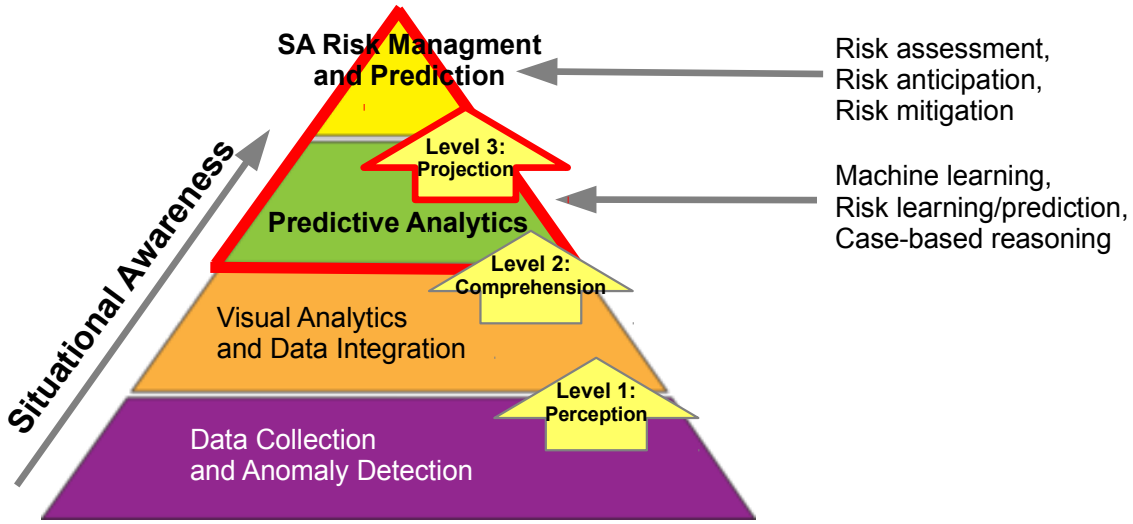

Figure 3. SAFARI SA approach (Level-3).

Future work will take SAFARI to the next level by implementing the third and last level of a full-featured SA solution (Fig. 3):

> *SA Level-3: Situation projection.* Combination of perception and comprehension for the prediction of possible future fraud through Predictive Analytics (PA). The goal is that SMEs can rely on future projections produced by SAFARI as this automatically reveal new risk situations and enables timely decision making.

SAFARI SA Level-3 is to preserve newly acquired experience and to provide it for future SA. Level-3 is executed later when an assessment of a risk situation is completed, e.g. when some produced fraud patterns or risk scenarios are confirmed. In Level-3, the next set of situations with the best similarity is selected according to the situation retrieval phase. Based on this selection, it is now possible to show SMEs similar risk scenarios and update the probability of occurrence for all these situations.

The Situation Projection level will enable SAFARI to learn data patterns while SMEs exploit the framework. This involves the implementation of PA techniques on top of current SAFARI platform, such as ranking learning and prediction algorithms, case-base reasoning and/or the use of state-of-the-art machine learning (ML) methods to enable SAFARI find and predict risk patterns on unexplored portions of data.

Novel PA techniques that can be experimented to implement 3-Level SA in SAFARI include:

Bioinformatics + ML for string anomaly detection

Multiple sequence alignment (MSA) algorithms are widely used in Bioinformatics to align three or more biological sequences, generally protein, DNA, or RNA looking for evolutionary relationships. We believe that we can exploit MSA to align and extract string patterns from vendor invoice codes (and any other type of string code stored in financial datasets) in order to feed and train one-class ML anomaly detection algorithms, such as OC-SMVs, to reveal code outliers (e.g. anomalous vendor invoice codes) while reducing the chances of false positives. Dimensionality reduction techniques, such as MDA and PCA, would allow SMEs to visualize the detected outliers in reduced two or three-dimensional spaces to assess how different are the outlier strings from the non-outlier ones.

K-core decomposition scenario fingerprinting and reduction

The concept of coreness is a natural notion of the importance of nodes in complex networks like SAFARI's RFNets. The K-core decomposition of graphs has been recently applied to discover the core of large Internet Autonomous System graphs and to identify hierarchies within criminal social networks. We believe that we can exploit visualization algorithms based on the K-core decomposition of RFNets to produce fingerprints that uncover in two-dimensional layouts topological and hierarchical properties of fraud scenarios. Fingerprints can be used to visually identify similar cases to those being confirmed as fraud or even to reduce the size of RFNets while preserving important complex network metrics (avg. path length, straightness coefficient, assortative coefficient, etc.), speeding up the execution of algorithms on large RFNets and assisting visual inspection of complex scenarios.

Ranking prediction based on eCommerce recommendation algorithms

Think of how Amazon and Netflix allow their users to provide valuable feedback by allowing them to rate products/movies, creating a custom profile for each user. We believe that SAFARI can be enabled to learn from SMEs to adjust the ranking scores in an automated fashion by exploiting recommendation algorithms used in popular eCommerce platforms, such as Collaborative Filtering (CF) algorithms. In the absence of ground truth financial datasets, SAFARI can exploit analyst's feedback to provide better and more accurate scores in the future, predict SME ratings for new fraud scenarios based on previous ratings provided by SMEs, and suggest similar entities/actors in a similar way Amazon suggests similar products based on the customer's profile.

Modeling of Black Swan fraud events

Due to the usual lack of labeled (fraudulent/non-fraudulent) financial documents, and the fact that known payment fraud events are very scarce, there is just no enough historical data to overcome the statistical biases inherent in small samples when it comes to exploit ML for fraud prediction. However, a better understanding of modeling rare events should help SAFARI to accurately treat, model and predict rare occurrences of fraud. Adjusted statistical techniques, such as adjusted logistic linear regression models and more appropriate data collection strategies exist that can be useful to overcome the disproportionate role of high-profile and hard-to-predict Black Swan events (binary dependent variables with dozens to thousands of times fewer occurences) that are beyond the realm of normal expectations and that can be more reflective of fraudsters profile.

Enriching datasets with public Web data

Data for suppliers can be enriched by gathering information from on-line public services. The technical terms are 'Web scrapping' and 'Object consolidation', i.e. extracting information from websites and deciding whether that info actually corresponds to vendors/customers/employees/suppliers. In the future, SAFARI could enrich payment documents with additional fields obtained from Yelp.com, WhitePages.com and YellowPages.com. Then, SAFARI could use the scrapped information to, for example, automatically detect payments that do not make any sense (e.g. paying medical equipment to a restaurant).

Future Work Objective

The objective of the project for the second year is to develop a predictive platform on top of the SAFARI framework, to provide SMEs with an advanced software framework for data-driven risk ranking that integrates the RF and RFNet concepts, rich Web-based Visual Analytics and PA, all in a unified platform, enabling the anticipation of future fraud risk events.

We believe that by introducing novel PA capabilities into SAFARI we can get an innovative fraud detection platform with increasing levels of intelligence and automation of fraud risk management, by replacing time-consuming manual fraud detection efforts with automated anomaly detection, RFNet integration and learning techniques that discover and learn from fraud patterns in large volumes of unlabeled financial data.

Our Vision for SAFARI:
Holistic View of Risk Scenarios

We believe that the SA-based SAFARI platform can be exploited for detecting other type of financial deceptions, such as credit card and health insurance fraud. Even more, SAFARI can be adapted to be applied in a wide variety of environments where the assessment of risk is critical for decision making, such as the risk assessment of scale formation in oilfield facilities, the investigation of criminal activity, counter-terrorism, computer cyber-security analysis, and the minimization of industrial health/safety/environmental incidents.

Being a flexible software framework, we expect that SAFARI can keep up with business evolution and dynamics, having the potential to be a valuable advisor in decision-making in both the short and the long-term.





SAFARI Team at MIT GDC

Alberto Garcia-Robledo
Abel Sanchez
Rongsha Li
Juan-Carlos Murillo-Torres
John Williams
Sascha Boheme