Semantics-based Threat Structure Mining

N. Adam[1], V. Atluri[1], V. P. Janeja[1], A. Paliwal[1], M. Youssef[2], S. Chun[3], J. Cooper[4], J. Paczkowski[4], C. Bornhoevd[5], I. Nassi[5], J. Schaper[5]
[1]CIMIC, Rutgers University
[2]Arab Academy for Science and Technology
[3]City University of New York
[4]The Port Authority of New York and New Jersey
[5]SAP Labs

ABSTRACT

Today's National and Interstate border control agencies are flooded with alerts generated from various monitoring devices. There is an urgent need to uncover potential threats to effectively respond to an event. In this paper, we propose a *Semantic Threat Mining* approach, to discover threats using the spatio-temporal and semantic relationships among events and data. We represent the potentially dangerous collusion relationships with a *Semantic Graph*. Using domain-specific ontology of known dangerous relationships, we construct an *Enhanced Semantic Graph* (ESG) by scoring the edges of the semantic graph and prune it. We further analyze ESG using centrality, cliques and isomorphism to mine the threat patterns. We present a Semantic Threat Mining prototype system in the domain of known dangerous combination of chemicals used in explosives.

1. INTRODUCTION

The Port Authority of New York/New Jersey (PA) manages and maintains bridges, tunnels, bus terminals, airports, PATH commuter trains, and the seaport around New York and New Jersey that are critical to the bi-state region's trade and transportation capabilities. The continuous monitoring of cars, trucks, trains, and passengers is a necessary precaution for preventing major threats to safety. The amount of data and potential alerts generated from these monitoring activities are enormous and heterogeneous in nature due to the different types of monitoring devices, ranging from text messages to images, audios and video feeds. The challenge is to mine and identify meaningful potential threats, and minimize false alerts. Important is an ability to infer threats coming from several independent seemingly benign activities. Often ignored is the threats implicated when these independent activities are looked at together as illustrated in the following scenario.

Motivating Example: *Consider a customs office inspecting a truck shipment carrying liquid Urea entering through the port in Los Angels, whose final destination is Phoenix, AZ. Assume there is another shipment entering through the port in Newark carrying cyclotrimethylene trinitramine (RDX) is bound for Wintersburg, AZ.*

The two shipments, when viewed in isolation, appear to be benign. However, the spatial proximity (shipments with spatially close destinations), temporal proximity (the two events occurring close in time), and semantic proximity (the materials being shipped have some semantic relationship, for example, can be combined to make explosives), would indicate possible collusions among entities and enhance a potential threat discovery and detection capability. Here it is essential to look at spatio-temporal proximities first since purely semantic proximity may lead to frivolous and non-relevant threat structures to be identified.

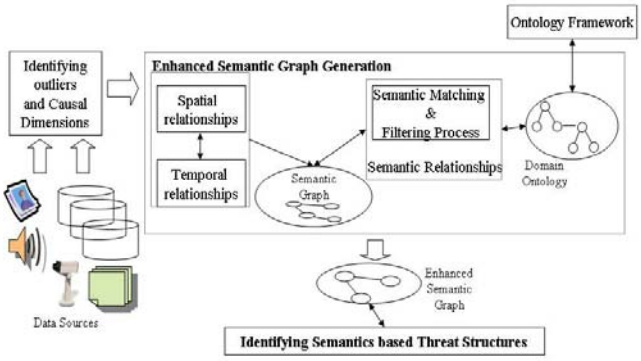

Figure 1 Semantic Threat Structure Mining Approach

2. SEMANTICS DRIVEN DATA MINING

Our approach depicted in Figure 1 consists of the following distinct steps: *i) Semantic Graph (SG) generation by outlier detection:* We use data mining to generate nodes in SG by identifying interesting entities namely outliers and their causal dimensions [3]. *ii) Enhanced Semantic Graph (ESG) Generation:* The connectivity between the nodes of SG are established and pruned to generate an enhanced Semantic graph (ESG) by the following two-steps. a) First we identify *spatio temporal relationships* between the outliers; b) Second, we identify *semantic relationships* and *semantic scores* between the outliers, using domain ontologies and reasoning. The semantic enhancement includes removing relations that are not supported by the reasoning using the semantic relationship scores between the dimensions. *iii) Identification of Threat Structures:* The ESG is further analyzed for the semantic centrality, semantic cliques and isomorphic paths to identify semantics based threat structures.

3. SDM Prototype System

3.1 Dataset

We have tested our approach on the PIERS data, comprising of imports, exports data and U.S. and overseas profiles of companies. The PIERS data comes from multiple ports and agencies. Moreover, some shipments, when observed closely, may lead to suspicious terrorist behavior, which is analogous to the threat structures considered in this paper.

For the prototype system, we have labeled outliers by visual inspection. We have vertically partitioned it to allocate a set of dimensions to each domain. The domain experts from the Foreign Operations Division, U.S.Department of Homeland Security have been identifying the semantic relationships within the dimensions of the PIERS data and labeling of the outliers in this data. However for checking the accuracy and efficacy of outlier detection we discuss detailed results with other datasets [3].

3.2 Semantic Matching for Relationship Mining

To identify the semantically dangerous relationships among outlier data sets, we used a domain specific ontology on Threat agents; specifically chemical threats constructed using Protégé as shown in Figure 2. It includes two major taxonomies: the domain specific concept taxonomy, e.g., types of chemicals, and an operational taxonomy for the matching process that we refer to as the *Potentially Dangerous Combinations* (PDC).

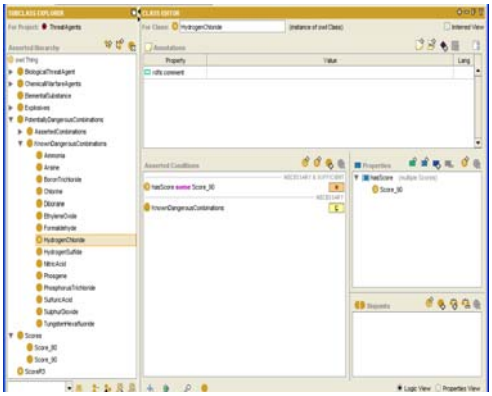

Figure 2 Threat Agents Ontology

The data set is matched against the ontologies and reasoned using the RACER reasoner [2]. The matching mechanism first reads the ontology from an OWL file, converts it to DIG format [1], second receives a potentially dangerous combination of chemicals and creates a parent concept as the union of this combination, and third performs the matching and returns the enhanced semantic graph.

3.3 Identification of threat structures

Once we identify the semantic relationships generating the ESG, we utilize it to identify threat structures based on the top weighted semantic relationships, other semantic properties such as semantic cliques, semantic centrality, and semantic isomorphic paths.

We have vertically partitioned the PIERS data into 3 domains such that each domain consists of some part of each record. Inter-relationships between outliers are labeled based on the outliers detected and the semantic weights generated by the ontology framework. It was observed that only 50% of the labeled inter-relationships were identified. We believe that this loss is due to the manual partitioning and labeling of the data where, some actual outliers may have been lost. In Figure 3 , the weighted graph in the left of the window shows the semantic graph, the right top part of the window shows the Enhanced Semantic graph with the weights in terms of the semantic weights generated by the matching process. The bottom right part shows the discovered top interrelationships and the semantic centrality. Although we discover all interrelationships for presentation we show top 10 relationships.

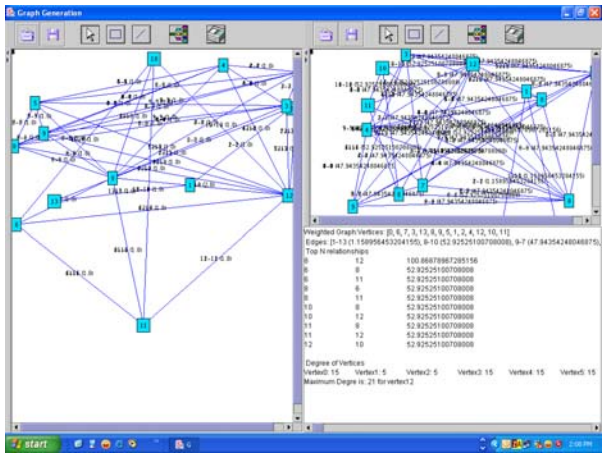

Figure 3 SG and ESG of PIERS data

4. CONCLUSION AND FUTURE WORK

In this paper, we presented an approach to enhance the semantic graphs by discovering collusion relationships existing in spatio-temporal and semantic dimensions among events, using the concept of the Enhanced Semantic Graphs (ESG). The potential threats are identified with semantic centrality, semantic cliques and isomorphic paths. As part of our future research, we propose to focus on the general problem of identifying relationships between normal entities without restricting ourselves to simply outliers. We plan to evaluate existing Semantic graphs generated from text data. Consistency across systems in determining semantic distances and the robustness of such calculations is essential in homeland security domain. We need to investigate determining the relative semantic distance between two concepts through an inspection of the values of selected attributes through a hierarchy of variables ranging from those that most directly related, termed proximate variables, to those most distantly removed, termed ultimate variables.

9. ACKNOWLEDGMENTS

This work is supported in part by the National Science Foundation under grant IIS-0306838 and SAP Labs, LLC.

10. REFERENCES

[1] S. Bechhofer, "The DIG Description Logic Interface: DIG/1.1", Available from http://dl-web.man.ac.uk/dig/2003/02/interface.pdf, 2003.

[2] V. Haarslev, R. Moller, "RACER Users's Guide and Reference Manual", Available at http://www.cse.concordia.ca/%7Ehaarslev/racer/racer-manual-1-7-19.pdf, 2000.

[3] V.P.Janeja, V.Atluri, J.S.Vaidya, and N.Adam. Collusion set detection through outlier discovery. In IEEE Intelligence and Security Informatics, 2005.