INDIVIDUAL FELLOWSHIPS

Project n°: 221077

Project Acronym: DMASD4CA

Project Full Name: Distributed Multi-way Analysis of Stream Data for Detection of Complex Attacks

Marie Curie Actions

IEF-IOF-IIF- IIFR -Final Report

Period covered: from 24.02.2009 **to** 31.05.2010

Period number: 1

Start date of project: 24.02.2009

Project beneficiary name: Prof. Dr.-Ing. habil Sahin Albayrak

Project beneficiary organisation name: Technische Universitaet Berlin

Date of preparation: July 2010

Date of submission (SESAM): 13.October.2010

Duration: 15 Months

Version: 1.1

1. FINAL PUBLISHABLE SUMMARY REPORT

Overview: The focus of the project is mining complex data which may be in real-time, continuous, high dimensional, multimodal and may arrive (change) at different rate and volume. There are several characteristics of complex stream data that require further research. For example in many such applications analyzing data at a single location is inefficient in terms of accuracy, space and computational complexity. Moreover queries may need to be processed in a near real-time manner. One shortcoming of the current state of art is that although there are well understood statistical and algorithmic techniques available for simple mining and summaries, computing more sophisticated summaries such as rank reduction in an on-line manner remain a difficult problem. We note that Singular Value Decomposition (SVD) and similar approaches work on 2-dimensional arrays and discover linear separations of data. However, if data has non-linear or multi-linear structure, then off-line algorithms such as SVD fail to capture it. M-way data analysis techniques (e.g., tensor decompositions) consider multiple modes of data simultaneously (e.g., a data cube as supposed to data matrix) to discover multi-linear structure. Similarly, support vector machines, kernel methods are used to analyze non-linear data. In this research we consider the mining of complex data by collaborative (distributed) data collection, multiway analysis and knowledge extraction.

Timeliness, Relevance and Objectives of the Project: This research considers the fundamental questions that are keys to improving mining of complex stream data: how to collect data in a distributed way to optimize accuracy while minimizing intrusion and performance degradation? How to decide on quantity, location of agents (programs that can collect data)? How to coordinate the communication and coordination of the data for mining and calculation? How to measure the accuracy and performance of such a collaborative system? Thus this project demonstrates how to (i) collect multidimensional data in a near real-time, (ii) construct multimodal models to fit to this data, (iii) decompose such M-way models using tensor decompositions, (iv) build a simulators that uses the statistics obtained in (i) to test the accuracy of the results in (iv).

Contribution, Originality and Innovation: This research addresses several import issues for sampling and analysis of complex stream data. First, coordinated sampling must be done to ensure a notion of independence among the agents. The allocation and coordination of agents must be a function of the properties of data stream. For example in this project we considered profiling a users' resource usage in a time sharing system in order to detect anomalies and intrusions. In this context, data can have multiple dimensions collected various system resources (e.g. CPU usage, memory usage) by developing a program that monitors the system continuously over time (see Table 1). This program can run in multiple time sharing machines and over time to construct an accurate signature of a user. Second, on-line or near real time versions of data analysis methods must be designed. For example it is not known how to design sliding window like algorithms for tensor decomposition to analyze data with multiple modes.

Table 1. Example of output data collected by the data collection tool built in Python programming language. The output was parsed, aggregated and subsequently used to construct a 3-way tensor model of resource usage.

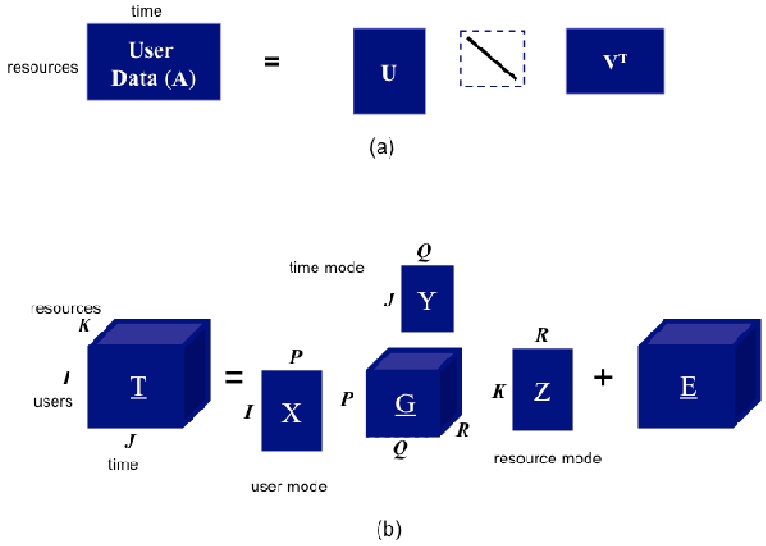

Figure 1. Data modeling and analysis techniques, where (a) shows modeling of data using bilinear model SVD, and (b) shows 3-way modeling by the Tucker3 model. We used SVD on user component matrices of T to construct spectral signatures. Tucker 3 is used to discover outliers in user mode.

In this project we developed techniques that can analyze 3-way data in using a sliding window like decomposition algorithm. The 3-way tensor analysis techniques (see figure above) are developed to find the structure in this data and identify a signature for the resource usage of each user in a collaborative environment which can be used for threat analysis. Distributed versions of these algorithms are straight forward extension since each data collection point can construct a portion of the signature and periodically dump the data to a shared and secure directory. Thus the project achieved the goals put forward. The attached manuscript (which is currently under review for IEEE journal) presents these results in more detail[1].

[1] Additionally, Professor Yener advised and interacted with groups and students from different research departments to extend the research efforts to security and resource sharing issues in wireless networks. (These efforts resulted in two submissions to IEEE conferences one of which is already accepted to IEEE Globecom 2010 and the other one is currently in review for IEEE Infocom 2011).