**Title**: Identifying Insider Threats Through Large-Scale Data Stream Mining

New Proposal  
Researcher Name: Srikanta Tirthapura  
University: Iowa State University  
Date: 1/1/2015 – 12/31/2015

**Long Term Goals**  
An Insider cyber-threat is an act of intentional misuse or theft of an organization’s resources by an individual who has authorized access to the organization’s system. Insider threats are a grave problem today.

The long-term goal of this project is to develop automated and semi-automated methods for rapidly detecting instances of malicious insider actions through an analysis of the data about the actions of individuals in the system. Data relevant to the actions of a single entity or individual is dispersed among different log files and traces throughout the system. The goal is to use the totality of all this data to get an overall view of the actions of an individual, and use this in identifying instances that maybe malicious actions of a user. The collected data is used to develop a model of “normal” behavior of an individual. Further actions of the individual are evaluated in relation to this model to detect anomalous actions, and such anomalous actions are more closely analyzed for detecting possibly malicious actions.

Our distinguishing feature is the development of novel streaming machine learning and data analysis methods in constructing models of individual behavior, and the use of these methods in rapid detection of instances of insider threat. These methods are expected to work with a latency that is orders of magnitude smaller than the latency of a batch processing method.

**Background for Long Term Goals**  
The 2014 US State of Cybercrime Survey [2] found that 37 percent of participating organizations experienced at least one incident of a cyber-attack by an insider in 2013. Further, they found that the damage due to insider attacks was often more than the damage caused by attacks from the outside. Undoubtedly, attacks by insiders constitute a significant problem today when more and more of an organization’s assets are in digital form.

Important clues to malicious insider actions are present within event logs. However, these logs are truly massive (of the order of Gigabytes per day) and constantly growing. Monitoring such data is a true “big data” problem. Further, this data is spread out across multiple locations such as multiple HTTP log files, email logs, login actions, queries to the active directory, to name a few.

At the same time, our capacity for processing large data sets has grown tremendously over the past few years leading to the emergence of systems such as Apache Hadoop and Apache Spark. Combined with the use of appropriately tailored algorithms, we are optimistic that such large network event data sets can be handled in an effective manner.

Our approach to data analysis is based on streaming analytics, where data is processed and
indexed in an incremental manner soon after it arrives. Trends and anomalies are detected quickly, within a short time (seconds or minutes) after they occur within data. In contrast, if we consider a batch-based approach, the processor waits until all the data is collected (for say, the entire day or week) until the processing can begin.

Our approach will lead to earlier detection of patterns and anomalous actions than an approach based on batch analytics. This is particularly important in detection of insider threats, where timely detection can lead to the difference safeguarding of the digital assets of an organization, and losing them to theft. We will build on our prior extensive experience in data stream computing in designing streaming techniques.

**Data:** We will begin with the datasets generated by ExactData LLC [1], for experimentation on insider threat investigations, which are available to us. The ExactData dataset has millions of events of different types partitioned into different files, one for HTTP, one for logons, etc. An insider threat incident is a combination of user actions that are spread out across multiple files.

**Processing Platform:** We will use the Apache Spark [6] “big data” software stack that is running on a cluster. This has capability for both streaming as well as batch processing, including parallel storage of big data and parallel processing of big data. Our streaming data analysis techniques will be built as a toolkit on top of Apache Spark.

**Intermediate Term Objectives**

Our intermediate term objectives are to design streaming algorithms and prototype implementations that can detect patterns of malicious insider actions within large data streams.

Our first step is to consolidate data from different sources to make a per-entity data stream. For instance, for each user id in the system, we will track all actions associated with the user id across all event logs. Similarly, there will also be a per-object trace for every object in the system. We will assemble this event stream from the massive log files, using our big data software stack.

The next step is to segment the per-entity-trace into logical sub-traces, that we call “sessions”. For a single user, these sessions maybe created at different granularities. For instance, there maybe a session for each subsequence of events between a single login and corresponding logout event of the user. Or, there maybe a session for a single interaction of a user with a particular object such as a file. Similarly, we will have sessions corresponding to each important object in the system.

The sessions are then analyzed using machine learning methods that we are developing, to build a model of a user. One technical challenge that we will particularly focus on is building streaming implementations of machine learning methods such as clustering, and anomaly detection. Another is the determination of the different features to be user for analyzing a session; it should be possible to determine such features in a streaming manner. Our end goal is to predict, in real-time, whether or not a user session is anomalous.

Our schemes will be evaluated to answer the following questions:

- Is the system able to detect instances of malicious insider actions? What are the false positive rates and the false negative rates? To help us here, for most of the datasets that we have, there is also the “ground truth” that is provided along with the dataset.
- What is the accuracy of the streaming methods when compared with pure batch methods?
- What is the number of events that can be processed per second?
• Does the throughput increase with the number of processors used? How is the scaling behavior?

**Schedule of Major Steps**

1. **Setup** of the data processing platform and the Apache Spark cluster
2. **Consolidation** of event logs and partitioning into per-entity-trails
3. **Segmentation** of the per-entity-trails into sessions and **streaming feature extraction**
4. Design of **streaming anomaly detection methods** on user sessions
5. **High-Performance Implementation** of anomaly detection methods, and parallelization using the big data software platform
6. **Evaluation** of methods according to above criteria

**Dependencies**

The dependencies are as shown in the Figure. Steps 1, 2, and 3 need to be done in sequence. Step 4 will be performed in parallel with 1,2,3. Steps 1 through 4 must be complete before Step 5 can begin. Step 6 will be continuous and will be done along with each of the earlier steps. But a majority of Step 6 will be done after Step 5.

**Major Risks**

The project explores one method to detect insider threats, through an exhaustive and deep analysis of data within network logs. There will be useful signals that this system may not be aware of. For instance, it is possible that an employee receives a notice that his employment will be terminated within a week. In such a case, there is reason to monitor the employee’s actions more closely than usual. But this information cannot be inferred by an automated system such as ours, and needs manual intervention.

**Budget**

<table>
<thead>
<tr>
<th>Reason</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graduate Student Stipend for 12 months, at $1800/month</td>
<td>$21,600</td>
</tr>
<tr>
<td>Fringe Benefits for Student</td>
<td>$2,808</td>
</tr>
<tr>
<td>Travel</td>
<td>$2,500</td>
</tr>
<tr>
<td>Student Tuition</td>
<td>$12,071</td>
</tr>
<tr>
<td>Overhead</td>
<td>$2,113</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>$41,092</strong></td>
</tr>
</tbody>
</table>
**Staffing**
The project will be staffed by a graduate student working under the supervision of Dr. Tirthapura. We have graduate students who are trained and ready to work on the project, and there is adequate support staff to help them.

**Category of Current Stage**
New proposal.

**Contacts with Affiliates**
We are in touch with Dr. Donald Steiner, of Northrop Grumman Corporation.

**Publications and Research Products**
We have not published our work on this project so far. PI Tirthapura has extensive experience with data mining techniques and streaming methods for data analysis, and has published widely on topics of stream computing and network data analysis.

**References**

1. ExactData, Next Generation Systems Testing: [http://www.exactdata.net](http://www.exactdata.net)


