Cybersecurity Analytic Module Software
to Proactively Detect Cyber Attacks at Government Agencies

1. Introduction:

Lately cyber security has become an enormous problem worldwide and especially in America. This is increasing becoming a threat to national security, government agencies' classified information. It has become especially important to develop security monitoring systems to Proactively Detect cyber threats and intrusion at any Business or Government Agency (GA) Computer and Communications. Threats can occur at GA computer/communication network nodes such as file servers, email servers, webservers, databases, & PCs, and therefore sound monitoring of the network has become more important than ever due to increasing cyber threats. So, it is important for to develop advanced cyber security surveillance systems and proactive alert making mechanisms for ever increasing needs in a growing environment of cyber security and network intrusion threats.

Therefore, the purpose of this proposal is to develop and deliver customized Analytical Module to be integrated into existing network monitoring systems. The analytic module is based on innovative Stats, IT and other Analytical techniques and models, to detect such malicious activity proactively; i.e. to detect network intrusion, and malicious activity while they are happening, and implement them in GA computer and communications systems. To meet this challenge, we propose to develop modules with such advanced analytical capabilities to detect network malicious activity real-time thus enabling network monitors to take corrective action proactively and prevent harm to the network and computer information while they are happening. This requires writing customized R or S+ code to be imbedded in the analytical module of the monitoring system and in alert making mechanism to make appropriate alerts while minimizing false alarms. Such systems currently do not really exist as we describe in detail in the following section.

2. What is Lacking in Current Cyber Security Surveillance Systems:

` Before a computer network surveillance system can make alerts it needs to properly model stochastic processes underlying all events and activities such as such as logon/off events, IP addresses origin, file uploads/downloads, file modifications, editing and deletions, etc. The rates and frequencies of such events needs to be tracked. Moreover, for activities such as file uploads/downloads, the distribution of the volume by user needs to be tracked and study how they fluctuate over time needs to be tracked as we further describe below.

Algorithms underlying current cyber security surveillance systems of Government Agencies and Corporate America tend to be based on ad hoc mathematical algorithms with little tweaks to handle natural variation occurring in the data being analyzed to detect possible malicious activity. They tend make too many false alarms and fail to detect true intrusions. The very fact that such systems do not have network traffic analytical algorithms written in statistical Software languages such as R, S+, and SAS is self-indicative that they cannot distinguish Traffic Noise from Deviation of Traffic Signal due to Cyber Attacks. They are not capable of analyzing all the dynamics of underlying stochastic processes representing underlying events and activities typical or untypical for a given time of day, day of the week, holiday, and so on. They also do not take advantage of latest statistical techniques to analyze the nature of typical stochastic processes and then detecting deviations of traffic signal that could occur at a given instance due to a malicious activity.

Latest cyberattacks happened at Government Agencies and Corporate America are self-evident of inadequacy of existing surveillance systems to detect cyber-attacks automatically and make true alerts for prevention while the attack is happening.

3. **The Challenges**:

Due to our experience in Corporate America, especially in Telecommunications and Pharmaceutical industries, it is our belief that no currently existing systems adequately model underlying stochastic processes representing events/activities by every authorized network user. This is because currently available systems not having advanced statistical techniques imbedded in the cybersecurity surveillance system with data analytical statistical software coded in such languages as R, S+, or SAS. The data on any activity/event of a network are stochastic processes that vary by User, Type of traffic, Time of day, Day of week, Weekday/Weekend, Holiday, day before and day after holiday, Trend, Seasonality, etc.

While above variables define the signal when properly modeled, there is typical high noise that vary from any period to the other. Analysis of traffic and events that become available through meta data in a computer and communications system involve analysis of

- Events such as Logon/Off, File Upload/Download, Deletion, Modification, etc.
- Rates/Frequency of such events at a given time, which is not constant even during a weekday
- Duration/Volume of such attributes as total traffic, and file upload/download by certain users

Before one can detect suspicious activity at a given time, he/she needs to model appropriate for each stochastic process and then estimate the parameters that specifies the typical signal. Moreover, such stochastic processes would vary over time. After factoring in all such drivers of traffic, one needs to separate out the Noise from the Signal to enable detection of a possible malicious activity making deviations from the typical signal. This requires developing advanced statistical techniques beyond what is available today to calibrate ever-changing stochastic processes and methods to capture deviations from the network is in-control state. Currently available systems tend to be based on ad hoc methods that do not take advantage of state-of-the-art advanced statistical analysis.

4. **Overview of Methodology**:

Methodology customized to individual GA computer network will have two components. First, we will identify appropriate model for each stochastic process representing events and activities described above. The traffic patterns will be first analyzed at aggregate level of the network first, and then immediately will be drilled down to the individual computer node and logins in the network that caused such anomaly or attack. The models underlying such analyses will have certain number of unknown parameters that vary by every authorized network user. Therefore, the second part of the approach will be accurate estimation of such parameters by every authorized user, a challenge that we have much experience with. Then, necessary Statistical Tests will be developed to detect true deviations of the traffic signals due to possible malicious activity or anomaly, with minimum false alarms.

For such modeling and parameter estimation to be useful, it is equally important to develop efficient computation algorithms so that tests to detect deviations from controlled state of communication networks can be performed in a fraction of a second. Therefore, once appropriate models are identified and statistical techniques are developed, we will develop and implement highly efficient computational algorithms to perform such advanced analytical techniques quickly.

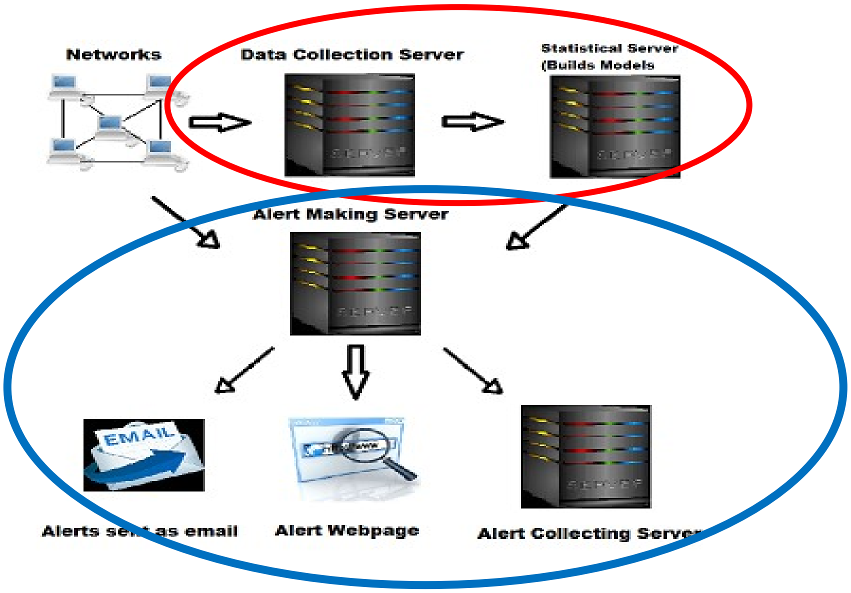
Models for stochastic processes will be different for events/activities at each of the following network node, and so methodology will be leveraged and customized for each of the following network nodes:

- File servers
- Email servers
- Web servers
- Databases
- PCs, and so on

5. Module to Detect Malicious Activity:

No methodology is useful unless efficient analytical algorithms are implemented in the cyber security surveillance system, as we have done before. A graphical representation is what is involved is shown in the network diagram below. In that regard, we propose to develop and deliver analytical modules to work with existing GA Cybersecurity Surveillance systems. This involves writing R code to analyze ever changing stochastic processes for each traffic attribute described above when they are operating normally, and then performing statistical tests to detect deviations from typical stochastic process happening due to a possible malicious activity.

Figure 1: Main Elements of Alert Making System



A customized alert making system incorporating statistical tests capable of detecting deviations from the controlled nature of the network traffic will also be integrated into the Analytical Module. It will have a number of visualizations displaying alerts in the form of maps and tables via web-based Dashboards. In the communication network, a map of network nodes will be displayed with red and yellow alerts corresponding to the seriousness of intrusion to enable manual intervention. When the intrusion involves a user ID or IP address, a colored table displaying ID, level of intrusion, and activity (file upload/download, time of day, volume, etc.) will be included in Dashboards. Detection of malicious activity while happening requires much real-time analyses, as described above, but for the purpose of illustration we provide below an obvious case of network intrusion, just to show how the table form of alert displaying dashboard will look like.

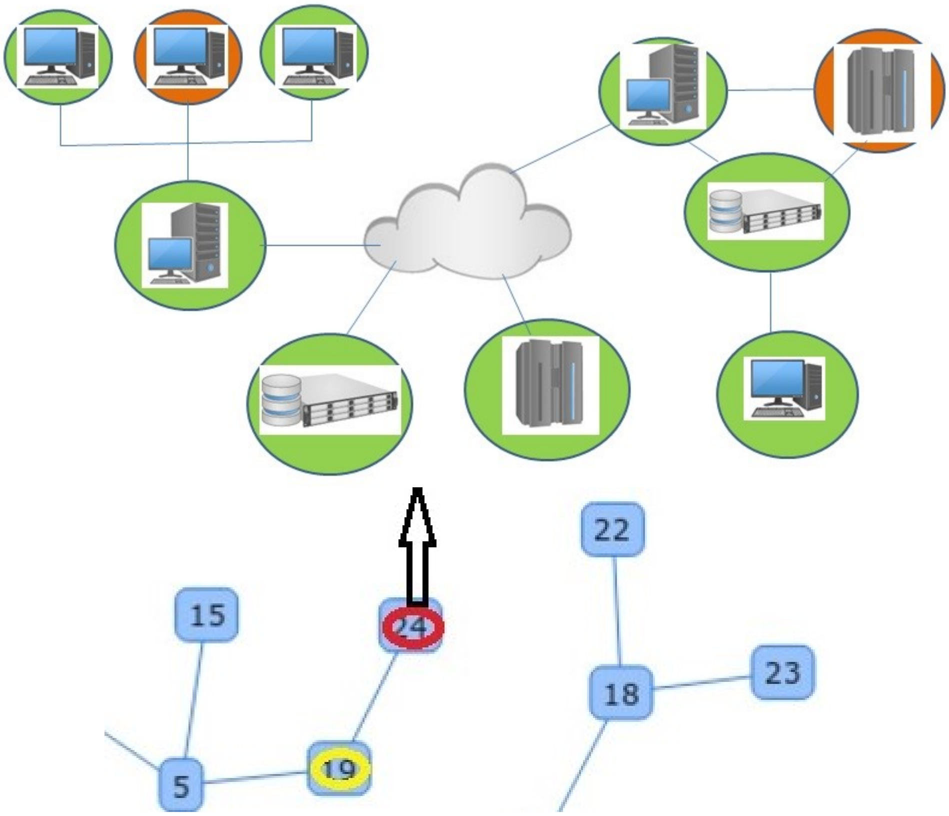
Figure 2: Display of a Simple Case of Alerts

File Download log(Rate+1) Scale During 3AM-3:05AM Historical Vs. Current

ID	Hist Rate	Hist Std Deviation	Cur Rate	Possible Cause	**Alert**
Unknown	NA	NA	1.0109	Intrusion	2
100	0.4039	0.1299	0.8790	intruded login credentials	2
853	0.3735	0.1249	0.6046	Warning	1
107	0.8458	0.2977	0.8327	Normal	0
108	0	0	0	Normal	0
.					0

Analytical module visualization dashboard will also display network nodes, such as File servers, email servers, webservers, databases, and PCs, color coded red and yellow alerts representing serious alerts and warning respectively. The display in the dashboard will have higher level network intrusion point(s) with drilldown into details of state of individual nodes color coded according to how they are affected.

Figure 3: Display of Alerts in Network Map



How System works with underlying models

Detection of malicious activity to make alerts is based on analysis of network logs and various meta data as described above. This involves analysis of stochastic processes underlying each such data in a 2-step procedure:

(i) Model calibration (i.e. estimation of parameters of each stochastic process) during a moving period of recent history when the network is under control without any issues, and

(ii) Real time analysis of current data from a short window to test whether such data is within the stochastic process, except for usual noise, for that instance or whether it is statistically significant deviation from the corresponding stochastic process under controlled conditions.

Details of how each of the two steps work best described by the Flow Diagrams discussed below.

Model Calibration: This step is depicted in Red Circle in the above diagram and Flow 1 below in Figure 4. Analysis of meta data to specify the appropriate stochastic process when the network is under control involves building a specialized model for respective activity to track. These models are built using advance statistical analysis for historical periodical data for each event/activity being tracked. Different models are built and stored for each stochastic process describing each event/activity occurring at each network node.

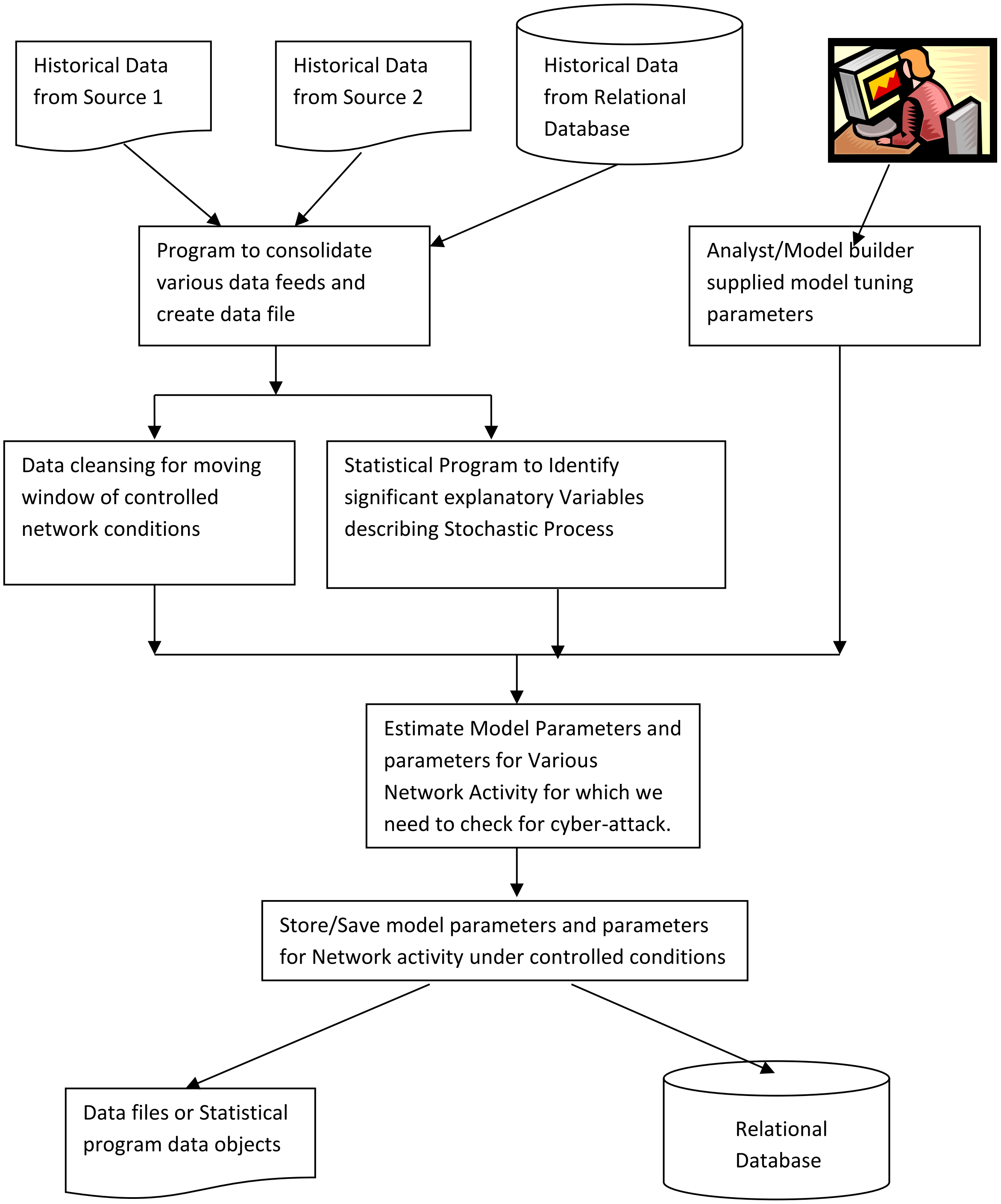
Data will be cleaned, and statistical programs will be developed to identify explanatory variables relevant for each model. These variables and parameters built by an analyst builder are checked for possible cyber-attack, network intrusion, or anomaly. Model parameters are saved and stored on a Database.

As summarized by the Flow diagram 1 in Figure 4, this process involves

- obtaining all relevant data from various data sources, including Relational Database, and preparing a data table for each event/activity to track using SQL programs
- cleaning data by dropping outliers as identified by a statistical algorithm
- frequent execution of a statistical program to identify relevant and significant categorical and quantitative variables needed to model each stochastic process underlying the corresponding event/activity
- Periodic execution of a satirical program to analyze cleaned data and estimate model parameters for each stochastic process
- Estimated parameters are saved and stored as data files and/or statistical data objects

Figure 4

Flow. 1 - Model Estimation system schematic

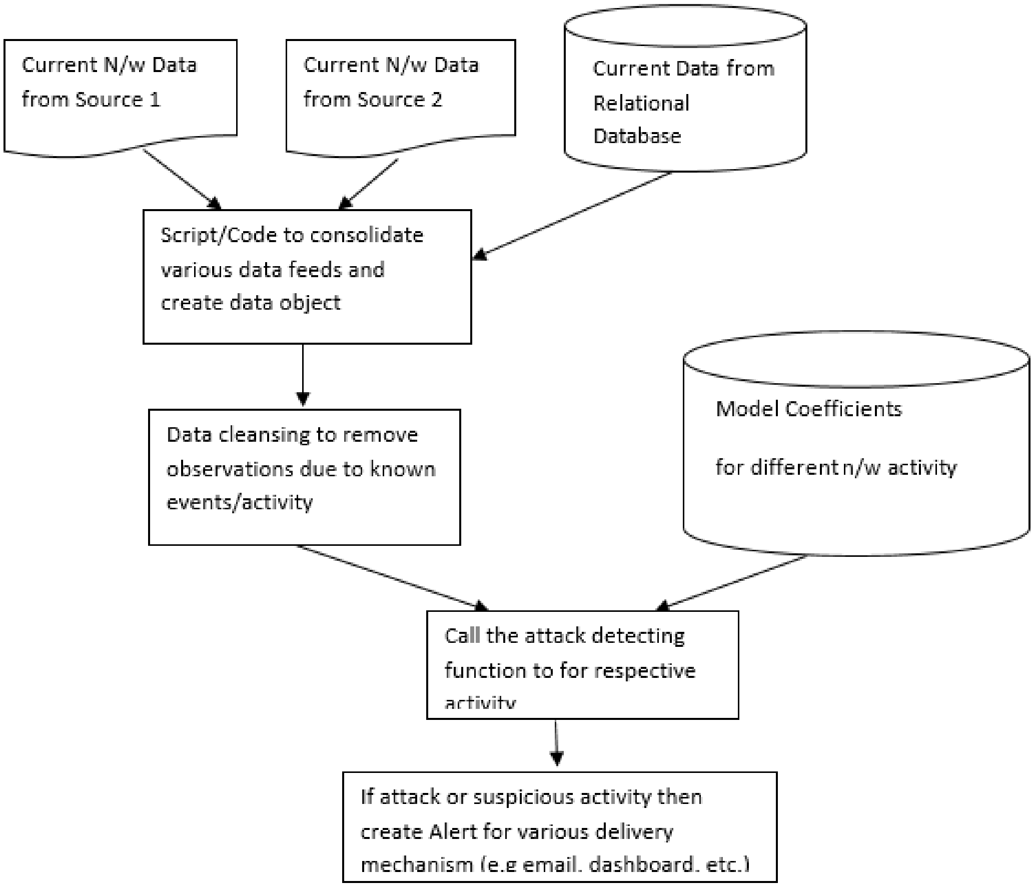

Historical Data from Source 1
Historical Data from Source 2
Historical Data from Relational Database
Program to consolidate various data feeds and create data file
Analyst/Model builder supplied model tuning parameters
Data cleansing for moving window of controlled network conditions
Statistical Program to Identify significant explanatory Variables describing Stochastic Process
Estimate Model Parameters and parameters for Various Network Activity for which we need to check for cyber-attack.
Store/Save model parameters and parameters for Network activity under controlled conditions
Data files or Statistical program data objects
Relational Database

Intrusion Detection: The second part of the system involves analysis of real-time data on each event/activity and comparing with the corresponding stochastic process developed above under controlled network conditions, and testing whether significant deviations currently exists due to possible malicious activity taking place via an existing user id or otherwise. This is depicted in Blue circle in the first diagram of this section and Flow 2 below in Figure 5.

As summarized by the Flow diagram, this process involves

- obtaining current data from various data sources real-time, and preparing a data table for each event/activity using SQL programs
- consolidating all relevant data in the form of a data table in memory
- cleaning of data by dropping outlier observations that are due to known event/activity and not due to cyber attack
- retrieving parameters/coeffects stored in Step 1 above for models representing the stochastic process describing each event/activity being tracked when the network was under control in the recent past at each of the nodes being monitored
- calling the statistical program that is capable of testing whether the current data is normal with typical noise, or whether it is significant deviation from the stochastic process for the current time due to malicious activity or anomaly
- making alerts in the form of color coded visualizations on dashboards, texting, emailing, and so on for manual and automated intervention to stop possible cyber-attack from completion.

Figure 5: Flow 2. Detecting attack and Alert making system schematic


Current N/w Data from Source 1
Current N/w Data from Source 2
Current Data from Relational Database
Script/Code to consolidate various data feeds and create data object
Data cleansing to remove observations due to known events/activity
Model Coefficients
for different n/w activity
Call the attack detecting function to for respective activity
If attack or suspicious activity then create Alert for various delivery mechanism (e.g email. dashboard. etc.)