AFRL-SR-AR-TR-04-

0157

# REPORT DOCUMENTATION PAGE



| 1. REPORT DATE (DD-MM-YYYY) | 2. REPORT TYPE | 3. DATES COVERED (From - To) |
|---|---|---|
| 31-01-2004 | Interim and Final | 15 Apr 2001 – 14 Oct 2003 |

**4. TITLE AND SUBTITLE**
Intrusion Detection Using Data Mining Techniques

**5a. CONTRACT NUMBER**
F49620-01-1-0274

**5b. GRANT NUMBER**
CFDA# 12.630

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**
Edward J. Wegman (Principal Investigator)

Don R. Faxon

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

Center for Computational Statistics
George Mason University, MS 4A7
4440 University Drive
Fairfax, VA 22030-4444

**8. PERFORMING ORGANIZATION REPORT NUMBER**

2003-03

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
Office of Sponsored Programs
AF Office of Scientific Research
4015 Wilson Blvd, Room 713
Arlington, VA 22203

**10. SPONSOR/MONITOR'S ACRONYM(S)**
AFOSR/PK

**11. SPONSOR/MONITOR'S REPORT**

**12. DISTRIBUTION / AVAILABILITY STATEMENT**

Approved for public release; distribution unlimited

20040315 008

**13. SUPPLEMENTARY NOTES** The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Air Force position, policy or decision, unless so designated by other documentation.

**14. ABSTRACT**

Principal Investigator and research fellows conduct research in the investigation of network-based intrusion detection using data mining techniques based on advanced computational statistics and visualization techniques. These new methods will provide the means to detect the presence of covert channels operating on user systems, passively perform continuous user authentication, discern subtle network attacks and information-gathering activities, and provide interface support for "storm center" situational display of intrusion detection alerts, damage assessment, and current network state-of-health.

**15. SUBJECT TERMS**
Intrusion detection, streaming data, evolutionary graphics, netcentric warfare

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | Edward J. Wegman, Ph.D. |
| Unclassified | Unclassified | Unclassified | unlimited | 16 | 19b. TELEPHONE NUMBER (include area code) 703-993-1691 |

January, 2004

Dr. Tatia Evelyn-Feggans
Sponsor Grants/Contracts Officer
Office of Sponsored Programs
Air Force Office of Scientific Research
4015 Wilson Boulevard, Room 713
Arlington, VA 22203

Performance Report: April 15, 2001 – October 14, 2003
Interim and Final Report on Intrusion Detection Using Data Mining Techniques

Principal Investigator: Dr. Edward J. Wegman
Sponsor #: F49620-01-1-0274

## ABSTRACT

Principal Investigator and research fellows conduct research in the investigation of network-based intrusion detection using data mining techniques based on advanced computational statistics and visualization techniques. These new methods will provide the means to detect the presence of covert channels operating on user systems, passively perform continuous user authentication, discern subtle network attacks and information-gathering activities, and provide interface support for "storm center" situational display of intrusion detection alerts, damage assessment, and current network state-of-health.

## OVERVIEW

Research during the reporting period was conducted in the development of multivariate data visualization and parallel coordinate display of extensive data sets, compression methodologies of massive and super-massive data sets preserving data integrity for follow-on statistical analyses of compressed data, and on-line evolutionary analysis and display of high-volume streaming data. This research effort has resulted in the publication or submittal for publication of 18 articles and a special issue of the journal *Computational Statistics and Data Analysis* with the theme of *Data Visualization*, and the presentation of nine research papers at professional conferences and workshops (see below). In addition, we participated 13 conference or workshops and, in addition, organized a workshop entitled *Workshop on Statistical and Machine Learning Techniques for Computer Intrusion Detection.*

We have launched a data collection effort entitled Project Knossos with the agreement of the University's CIO to capture all header information for all Internet traffic in and out of the University, primarily focusing on TCP, UDP, SNMP, and ICMP protocol packets. The installed "sniffer" (deployed at the University firewall) and analysis station are collectively capable of recording and analyzing up to a terabyte of traffic data. Experiments within our small statistics subnet indicate traffic of 65,000 to 150,000 packets per hour. We have incurred terabytes of data traffic daily university-wide, or 35-40 megabytes of header traffic per minute, yielding approximately 50-60 gigabytes of header information per day in the larger University context. Indications are that we observe as much as 24 terabytes of header traffic per year during calendar 2003. Much of the packet traffic is administrative traffic from routers and server "chatter", and undergoes some five fundamental character changes during weekdays. We have been interested in methods for real-time detection of intrusion attacks and/or malicious network activities, so that analysis methods capable of concurrently handling both steaming data and historical summary data are necessary.

Collaborative contacts have been established with the Naval Surface Warfare Center computer security research group in Dahlgren, Virginia to establish the collection and analysis of internet TCP/IP header traffic collected outside the University firewall using advanced visualization and statistical techniques with the intent to make comparisons of computer intrusion activities and methodologies arrayed against a relative open University computing infrastructure as opposed to those arrayed against a secured military system.

## CONCEPT

Background:

To be of practical use, any intruder detection effort within a university environment must not only recognize the computer systems organization's obligation to maintain good network integrity, but its charge to provide computational support to a diverse population of system users. On one hand, there exists a veiled yet real sense of urgency to achieve a relatively high level of system-wide invulnerability, especially with respect to malicious attack. On the other, the network must accommodate research activities that press computational intensity to the limit – and the occasional havoc unleashed as a result of "user curiosity" typical of collegiate ambience. To proceed, it is necessary to formulate the concept of data mining for knowledge of intruder detection in terms that relate to expected (and unexpected) user activities, and the various operational objectives of system administrators and managers charged with system integrity.

Data mining a University computing network for intrusion detection can be couched in terms of data mining a "subset" of world wide web. Such an endeavor is faced with the perception and elucidation of a plethora of concealed "information" on multiple levels. For example, three such fundamental levels are: epoch of time; geographic extent; and complexity of objective.

"Epoch of time" may appear to be a misnomer, at least in the sense that an intrusion must of course necessarily take place over some duration of time. The idea here, though, is that the *behavior* of an intruder will reflect both the inception of an intent to commit a particular type of malicious act, and the moment at which such intent becomes manifested at some level of persistence into a measurable interaction with the internet. This moment, equates to the moment in time or *epoch* from which the intruder's subsequent aggressive behavior can be reckoned – until that is either the culprit decides to "up the ante" or retreat and/or desist. Therefore, since knowledge of a particular intrusion speaks to his or her behavior, and such behavior is observed in terms of times and durations of time, the nature of the behavior and information gleaned from this behavior can be referenced to some base epoch, and a discernable change in behavior to a necessarily new base epoch. [Environmental time, or time of day for the local network being mined and analyzed, can also be reckoned to epochs, but in the sense of "markers" of fundamental periodic (nominally diurnal) changes in reactive behavior that the University network presents to the world wide web.]

Geographic extent is indeed a misnomer. The idea here is to "localize" either the origin (source) of an intrusion(s), and/or its destination(s) (intended or otherwise). This question applies just as much to topology as to geographic location. Just as geographic extent can be measured both as a collection of discrete sites and as a geographic area or region, so in a like manner can topology be measured in terms of network layout, client/server relationships, protocols – and especially the state-of-health (SOH) of same. Perhaps it might be more appropriate to use a term such as *geotopology* to reflect this dichotomy of characterization. Thus, knowledge gleaned from data mining can be associated with reference to a given geotopology.

Complexity speaks volumes. Data mining in essence is the extraction of knowledge from a "wealth" of data, information, and other knowledge. By "wealth" is implied large, huge, or massive amounts of data, depending on one's available computational resources, but certainly large enough to preclude traditional analytic procedures. Another popular term for data mining in use is knowledge discovery in databases. And it is well known that of all existing databases, the most complex in terms of heterogeneity, structure, and reliability – even accessibility – is the world wide web. The idea here is to relate the scale and sensor deployment of the data mining effort to the complexity of the information environment in which the sensors are embedded and the nature of how the resulting developed knowledge is to be presented. With increase in scale comes additional ancillary tasks – data/information cleaning and integration, selection, transformation – and with sensor deployment comes specification of filters, storage strategy, and dumping or transfer procedure. The resulting knowledge can thus be associated with the developed methodology of data capture as well as computation, evaluation and presentation technique.

As an aside, it should be noted that most on-campus system managers and administrators would be hard pressed to define *damage* as it applies to the University network. And the general user – even expert users – are clueless. The general reaction to "we're under attack!" is to pull the plug. There is a "rampant paucity" of what needs to be protected, and by how much. Thus there is no clear objective on how measure the extent of an intrusion. Today, one typically prepares for everything. Patch everything. Install enterprise virus protection everywhere. When we put our "nose to the wire", we only detect the presence of malicious activity, and either

do not know how or simply ignore attempting to determine how potentially damaging the intrusion really is. And this situation only applies to what the University computer systems organization has administrative control over. What about those dorm rooms? Laptops? "Wireless networking"? How serious is the threat?

The complexity as to which data mining functionalities to employ and the kinds of patterns that will hopefully be discovered is further expanded by the heterogeneous nature of the world wide net, and within the University network the open nature of the collegiate academic environment. However, data mining is by its very nature the approach that indeed has any real hope of achieving this daunting task. It only remains to modify the traditional approaches data mining might take to adequately address the situation at George Mason University, especially if the results are to be meaningfully compared with similar traffic analyses of web traffic at other universities, within commercial and governmental nets, and in the neighborhood of military sites.

**University Network Environment:**

Time of day for the local University community governs the general character of web traffic in as much as it determines component usage if not rote availability. At least six discernable "generalized" epochs that can be identified for the major week days during convening semesters: (1) from about 8:30 a.m. or so, the University traffic starts to ramp up as staff and faculty arrive at work; (2) web traffic continues to build until noon, when it levels off as the staff and faculty depart to lunch and students hit the labs; (3) peak traffic continues from noon until 4:30 p.m., when web traffic starts to slow reflecting the end of the work day; (4) traffic continues to slow from 4:30 p.m. or so until about 6:00 p.m. as the remnants depart; (5) from 6:00 p.m. the traffic builds up again as on-campus students return to their (wired) dorm rooms and others return to the school labs to do assignments, and this continues until 1:00 a.m. or so; and (6) there is an abrupt drop in traffic at 1:00 a.m. and traffic remains at a minimum until 8:30 a.m., and the cycle begins again anew.

Add to this environment the overlaying traffic arising from University Centers of Excellence, special programs, projects and activities of limited duration, unlooked for interruption of service due to external events, off-campus traffic, and dedicated virtual-network traffic. Although the resulting picture that the University network presents to the world wide web is more or less constant in intensity for most parts of the work day, it can be generally characterized as highly heterogeneous and fluctuating across its extent, and thus quite likely to invite quite a cross-section of intruders, not all malicious in the usual sense. In any event, some sections of the University network are a good deal more disciplined and therefore secure than other segments. Some parts are quite frankly a no-man's land.

This is in contrast to what you might logically find in a business net or in the neighborhood of a military site, where daily routing (and thus fairly constant) traffic dominates. However, it is also expected that in a battlefield environment, routine military traffic with its attendant penchant to "draw fire" (invite malicious behavior) from neighboring sources may indeed take on a character more comparable to the University's traffic, and it is hoped these analyses contribute to helping resolve these issues.

**Data Capture Design:**

The data-capture effort is two-fold: We have deployed two test and evaluation sensors using an analyst-friendly data-capture program for training and development purposes; and have deployed two operational sensors using raw-data capture programs based on tcpdump for on-line real-time and near real-time network analysis.

The first of the two operational sensors *Ariadne* is deployed "24/7" at the University firewall, and is collecting data in promiscuous mode over two fiber-optic cables off a multi-port TAP at the University gateway, one wire for in-coming traffic and one for out-going traffic. The sensor itself is a RedHat Linux box using the *tcpdump*-based SHADOW program developed at the NSWC in Dahlgren, writing captured data to ~ .5 terabyte RAID-enabled harddrive assembly, designed for hourly download to the analyst station *Theseus* located in the CCS *Holodeck* lab for near real-time intruder surveillance analysis and diurnal/weekly comparative analysis. *Theseus* is configured with an ~1.2 terabyte RAID-enabled harddrive assembly, and is anticipated to be able to download and store approximately a week's worth of TCP/IP traffic header data.

The second operational sensor *Daedalus* is dual-bootable into either WindowsXP or RedHat Linux, but will utilize the WindowsXP environment initially due to availability of graphic libraries. The design is to simultaneously collect and analyze **streaming data** in promiscuous mode in real-time over a single LAN

connection over the CCS *Galaxy* node using an in-tandem configuration of the MS Windows data-collection utility *windump* and a C++-based prototype program developed in-house.

The two test and evaluation sensors are *Daedalus* when operating in RedHat Linux environment using the relatively high-end user-friendly data-collection program ethereal, and *Dogmaster* using *Snort* or *windump* in an Windows 2000 Pro environment. Both *Daedalus* and *Dogmaster* are collocated with *Theseus* in the CCS *Holodeck* lab. *Dogmaster* also houses the Testbed Viewer (part of the Knossos Project database interface application suite developed in-house), which is used to prepare short-duration data "testpacks" for direct import into the in-house developed *CrystalVision* multi-dimensional visualization program.

The other test and evaluation sensor is *Perithous*, a dual-bootable laptop with both WindowsXP running *windump* and *Snort* and RedHat Linux running *tcpdump* and *ethereal*. *Perithous* is used to monitor an off-site Windows 2000 Pro server that is planned on being part of a prototype/training target array.

**Knowledge Development Design:**

Operational analysis of captured TCP/IP header traffic was done in two modes: near real-time analysis of University-wide TCP/IP header traffic on the analysis station *Theseus* located in the CCS *Holodeck* lab, where some 2+ gigabytes of data is to be downloaded hourly from *Ariadne* at the University firewall; and real-time analysis on Daedalus of streaming TPC/IP header traffic on the CCS Galaxy node. To gain more insight into statistical patterns and so-called "outlier behavior," the near real-time analyses performed on *Theseus* will be supplemented by more in-depth diurnal/weekly comparative analyses. The streaming data being analyzed on *Daedalus* employing various on-line visualization programs still under development should provide another perspective on design of intruder detection techniques and methodologies.

Over time, it is anticipated that knowledge in the form of rule-sets and the like will emerge that allow detection and analysis of the more subtle intruder activities, through both sensor filter manipulation at the sniffer, and development and progress in data compression, analytical procedures, and presentation mechanisms at the analyst station.

In summary, it is anticipated that both the more deliberate, systematic traditional approach using University-wide data, and the more nonparametric, adaptive analytic approach using a data streaming methodology enriched with evolutionary graphic techniques will provide a degree of synergism that either alone might not so readily reveal.

## EQUIPMENT PURCHASES

August, 2002 – Purchase of 1.67GHz dual-CPU w/ 1GB RAM and DVD ROM "sniffer" with half-terabyte RAID storage with two-NIC for simultaneous autonomous listening and high-speed data transfer capability.

Deployed outside the University gateway firewall with fiber-optic connectivity for continuous autonomous TCP/IP header traffic collection on the University's in and out gateway ports.

August, 2002 – Purchase of 1.67GHz dual-CPU w/ 2GB RAM and DVD ROM "analysis station" with 1.1 terabyte RAID storage with 4.67GB DVD burner and 50 recordable CDs.

Deployed in the CCS main laboratory ("Holodeck") with LAN connectivity for periodic downloading of the TCP/IP header traffic collected by the "sniffer" deployed at the University firewall, and continuous autonomous *streaming* TCP/IP header traffic collection on the CCS Galaxy Node.

## APPLICATIONS DEVELOPMENT

**August, 2002 – Project Knossos Database Interface Application.**

MS Access application, designed and maintained by D. Faxon, detailing and linking the equipment specifications, network topologies, conceptual layout, technical issues, project chronology, report and conference/workshop documentation, and appropriate data dictionaries and nomenclature.

This application has evolved into a main database front-end, a password-protected back-end database for key tables, an ancillary mini-database of the local topology (also password protected), and a data-viewer application that prepares test data for direct import into the Center's multidimensional data visualization software, *CrystalVision*.

The Main Database Front-End (Figure 1a) provides an interface to Project Knossos introductory material, equipment specifications, local topology and conceptual layout, project chronology, key issues and documentation, and dataset summaries and retrieval information for TCP/IP data capture repositories. For reference purposes, a hyperlink is provided to the Internet Storm Center for up-to-date status of intrusion and related activities and trends, along with an interactive data glossary of IDS terminology. The Project Documentation Window (Figure 1b) provides easy access to project related articles as indicated below in the papers completed and presentation given sections below, as well as to related WebCasts provided monthly by the SANS Institute. The key tables used in the Main Database Front-End are maintained in a password-protected back-end database for ease of up-date and key data integrity.

The local topology is provided in a separate ancillary password-protected database due to the sensitivity of device-specific information which it contains (Figure 2). For reference purposes, a hot-link to manual for "tcpdump", the unix-based utility used to capture/dump traffic on a network is also provided as a aid to the analyst. This allows one to readily identify local device address and device identification in sample captured datasets.

And lastly, the Testbed Viewer (Figure 3) is an application that prepares captured TCP/IP header traffic data test sets for direct import into the multidimensional data visualization software program, *CrystalVision*, which employs interactive parallel coordinate/scatter plot visualization techniques with grand touring capability to develop so-called rule sets (linear combinations of variables) that lead to "interesting statistical structure". To assist the analyst, sub form views of the original dataset and the prepared/cleaned dataset (typically a subset of the original data) are provided for side-by-side comparison along with a convenient means to modify the criteria used to obtain the modified dataset. Ancillary information such as assigned port and vendor numbers, country codes, and the like are also readily available.

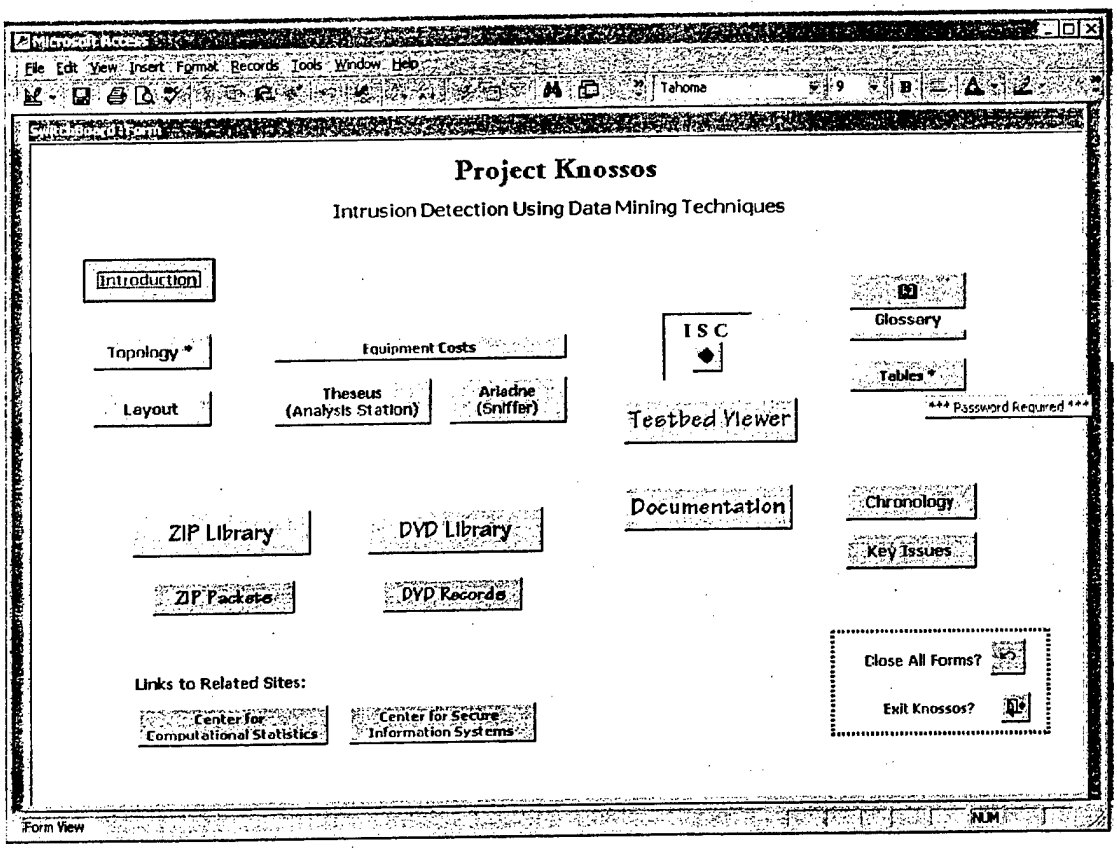



**Figure 1a. Main Database Front-End**

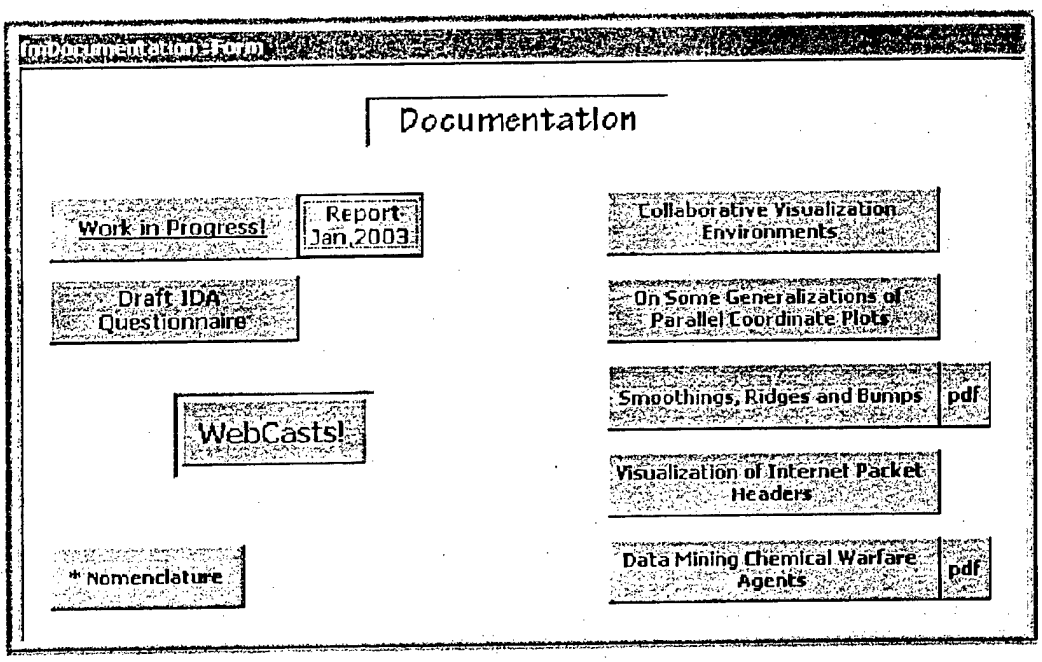



**Figure 1b. Project Documentation Window.**

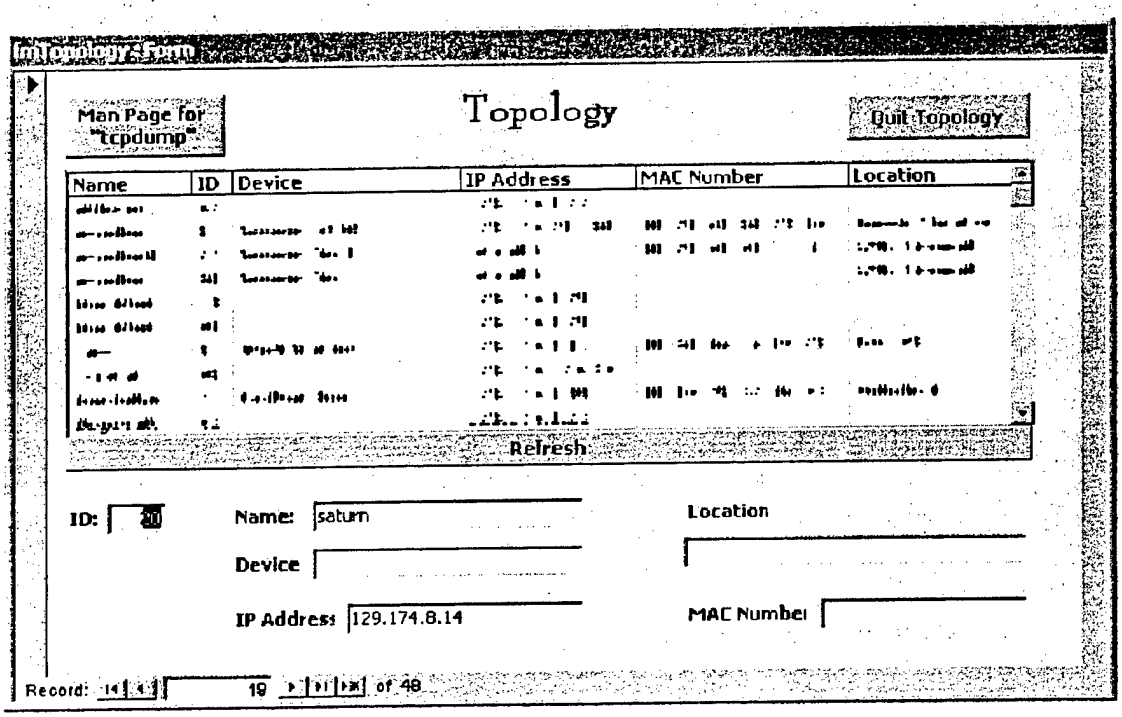



**Figure 2. Local Topology Window.**

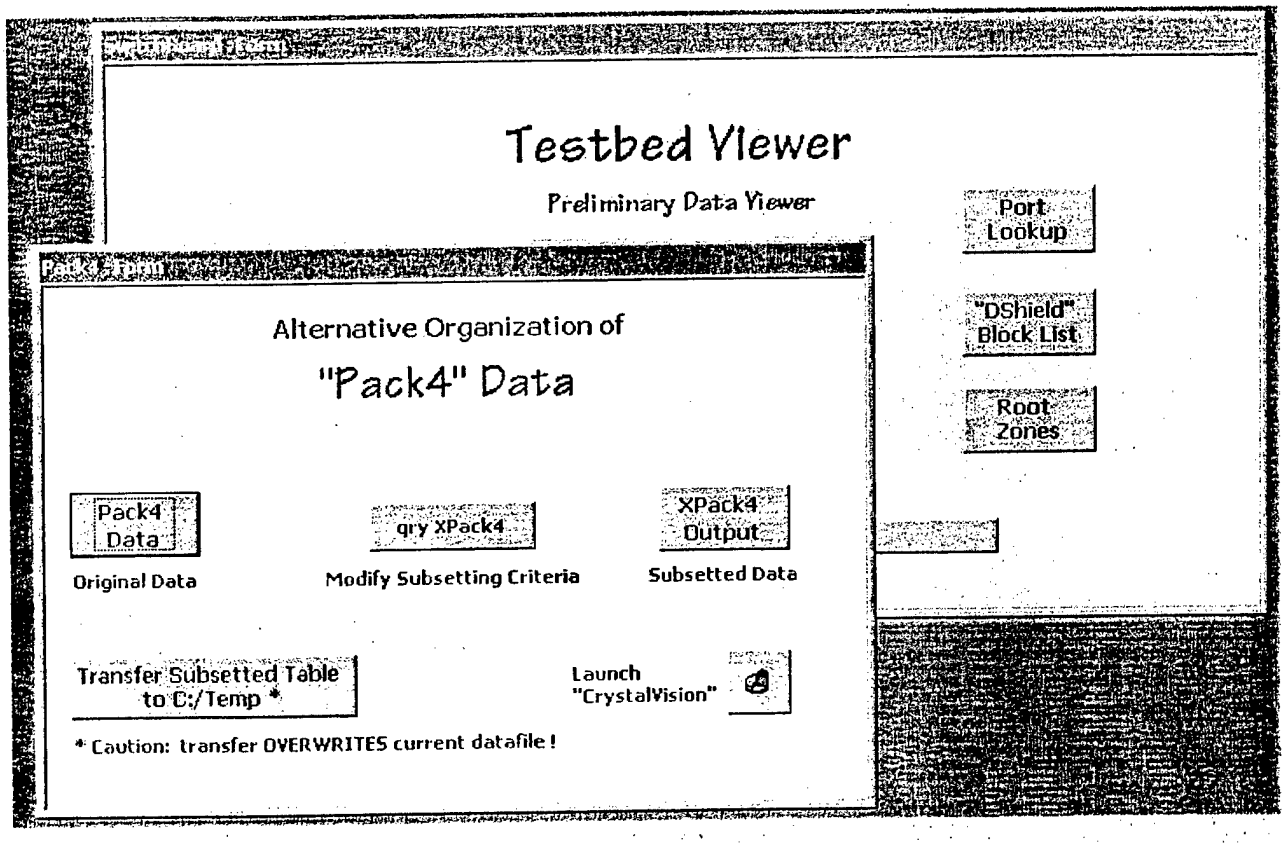



**Figure 3. Testbed Viewer With "Pack4" Sample Test Data Window.**

**March 2003 (on-going) – Evolutionary Graphics C++ Application.**

This application is a C++ developmental environment application designed and maintained by the CCS programming group comprising E. Wegman, D. Faxon, and D. King. This application is designed to analyze and display streaming TCP/IP header traffic using a suite of advanced evolutionary graphic displays. The streaming data is to be piped-in via a data capture utility such as windump in a WindowsXP environment or tcpdump in a RedHat Linux environment. Display design specification, level of resolution, and degree of scalability and/or platform transportability will depend primarily on availability of graphic libraries.

Some examples of data visualization on header data are given below. N these diagrams we use data on source ports and IP address (last two octets) and destination ports and IP addresses (first two octets). We are interested in anomaly detection associated with possible intrusion attempts.

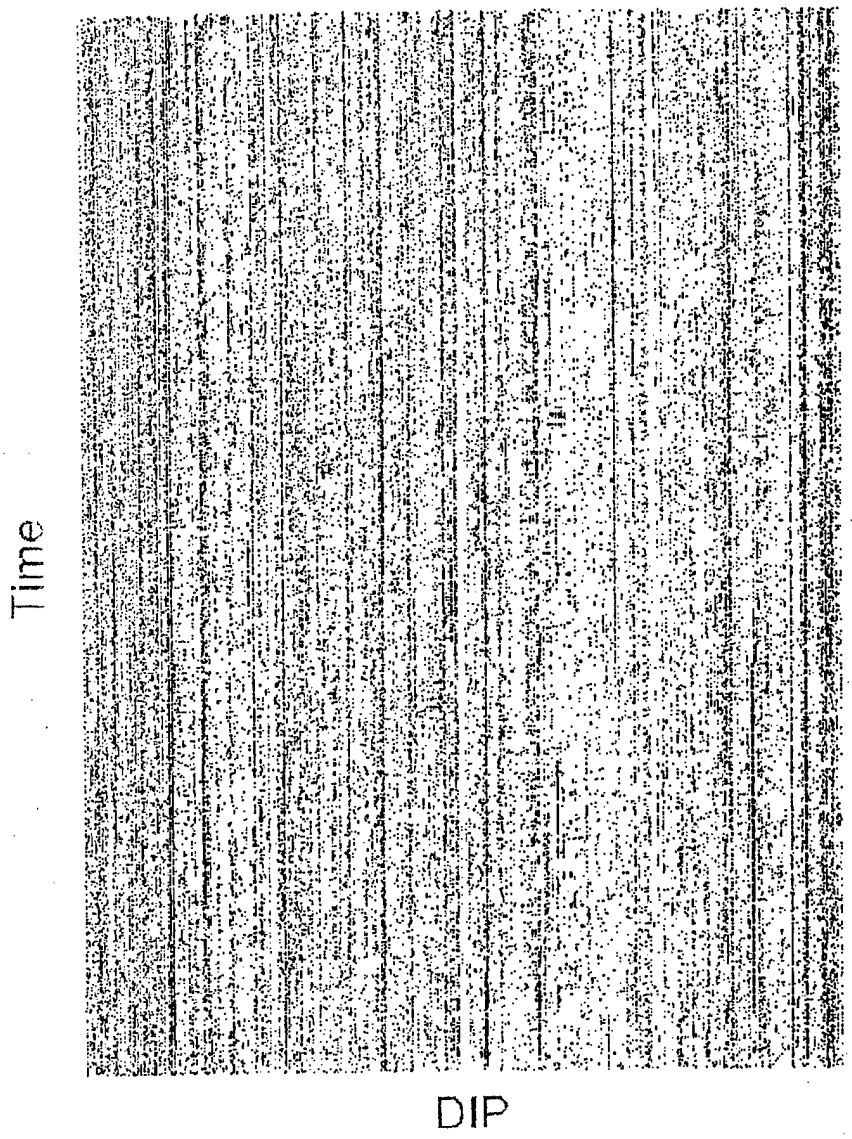



**Figure 4: Waterfall Diagram of Destination IP versus Time. Earliest epoch near bottom, newest epoch near the top. Persistent vertical striping indicates significant continuing contact. Low order IP addresses (to the left) typically belong the large Internet service providers.**

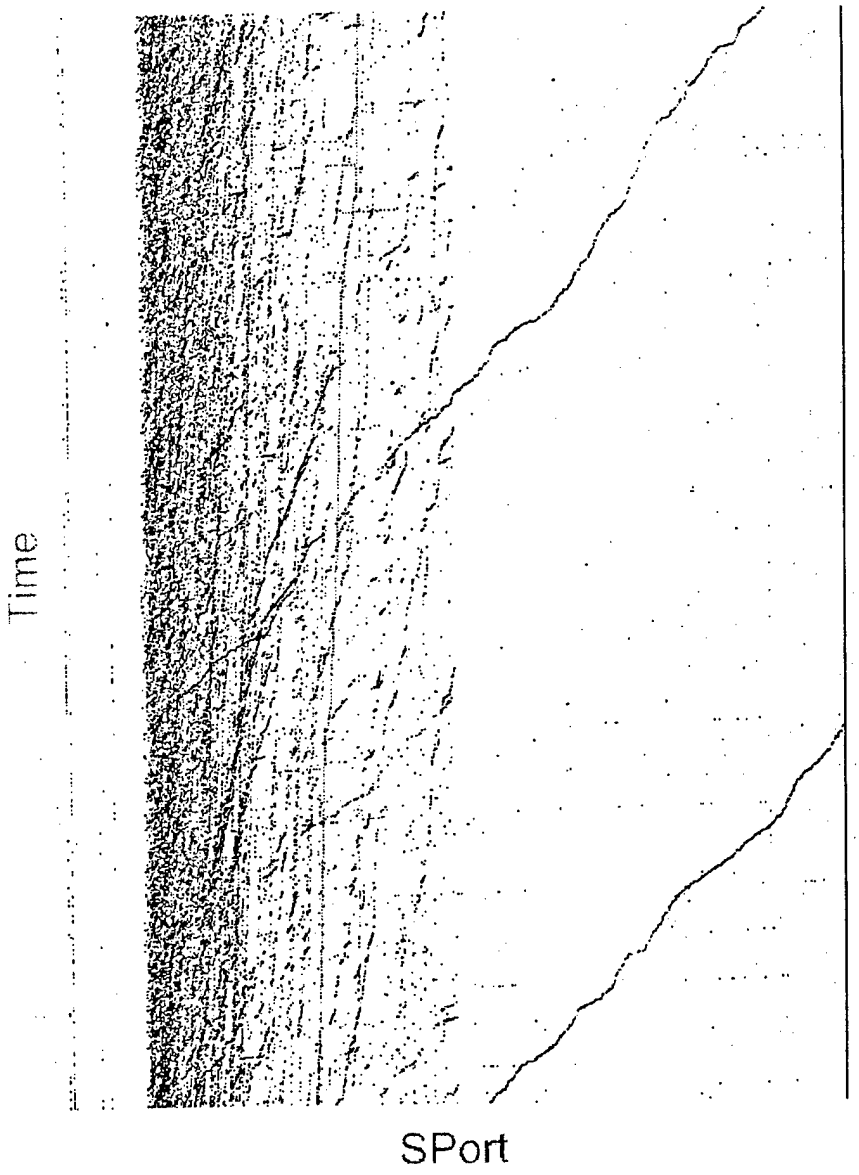
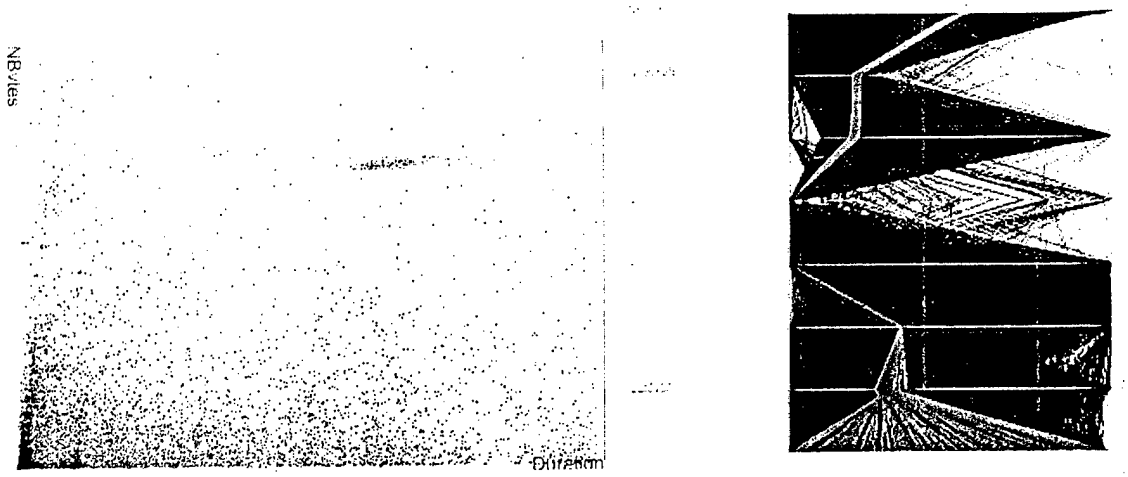



**Figure 5: Waterfall Diagram of Source Port versus Time. The first 1024 addresses are reserved for activities such as http and ftp services, email, ssh, telnet and the like. Typically operating systems increment the source port with each new packet, hence the diagonal lines. Slope of the diagonals are characteristic of different operating systems, hence are a diagnostic for operating systems. In this diagram, source ports are truncated at 10,000 (there are some 64,000 source ports).**




**Figures 6a and 6b: Anomaly detected and highlighted in red in drilled down plot of the Number of Bytes versus Duration. The same data are shown in 6b in a parallel coordinate display indicating that the anomalous packets shared the same destination IP, port, and source IP.**

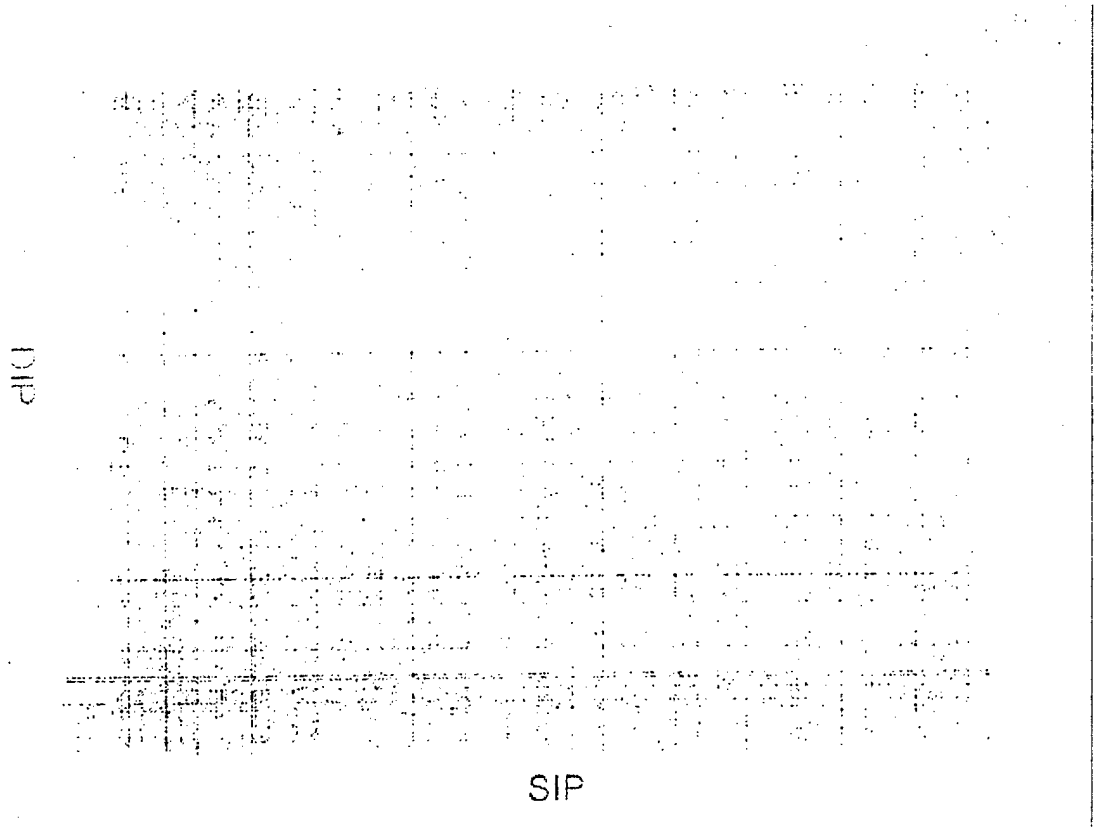



**Figure 7: Destination IP versus Source IP. Horizontal stripping indicates many source IPs are communication with a few destination IPs, such as Amazon.com, Google.com, etc. Vertical stripping means a single source IP is scanning many destination IPs indicating possible malicious activity.**

## RESEARCH PERSONNEL

Center for Computational Statistics:

| | | |
|---|---|---|
| Dr. Edward J. Wegman, Director | Principal Investigator | |
| Dr. Don Faxon | Fellow | Appointment: Feb 2002 (20 1/2 mos) |
| Dr. Jeffrey Solka | Fellow | Appointment: Sept 2002 (13 1/2 mos) |

Unfunded Collaborative Research Personnel:

| | |
|---|---|
| Dr. David Marchette | Naval Surface Warfare Center |
| Mr. John Rigsby | Naval Surface Warfare Center |
| Dr. Carey Priebe | John Hopkins University |

All Fellows and investigators are U.S. Citizens. Dr. Wegman has held DoD TOP SECRET clearance and currently holds DoD SECRET and DoE Q clearances. Dr. Faxon holds DoD TOP SECRET, Mr. King was formerly with CIA and held high level clearances, but does not have current clearance. Drs. Solka, Marchette, and Rigsby hold TOP SECRET clearances. Dr. Priebe is cleared to work with NSA.

## CURRENT STATUS OF FELLOWS

All of the Fellows were trained by hands-on use hardware and interactive data analysis of the GMU Internet traffic monitored by our system. This system was funded under the AFOSR CIPIAF program.

Dr. Don Faxon has taken SANS training under the program and remains active with our group, unfortunately at the moment in an unpaid status. He was the chief hardware specialist in our effort and has arranged for the sensor and analyst stations. He continues to monitor the University's traffic while he completes his SANS training. He is presenting an invited talk at the Joint Statistical Meeting in Toronto in August, 2004. That talk will be published in the Proceedings. Dr. Faxon intends to work actively in the computer security area.

Dr. Jeffrey Solka holds an appointment on the Prince William campus of George Mason University. Dr. Solka and I (as editors) have just published a special issue of the journal *Computational Statistics and Data Analysis*, 45(1) dated 28 February 2004. The topic of the special issue is *Computer Security and Statistics*. Dr. Solka is actively working in this area. Dr. Solka plans on continuing in an academic career.

Dr. David Marchette was not a fellow, but has worked closely with us. His book, *Computer Intrusion Detection and Network Monitoring*, has been the first book to focus on statistical and graphical methods. (David was my Ph.D. student.) Our joint work is listed as items 8 and 16 below. We continue close collaboration. David collected data at the Naval Surface Warfare Center, Dahlgren Division similar to that we have collected. We are currently working on a comparison of attacks on military and university systems.

Mr. John Rigsby was not a fellow, but work closely with Dr. Faxon in setting up our hardware sensor and analysis machines. Mr. Rigsby is an employee of the Naval Surface Warfare Center and has been doing work on using social networks theory to analyze computer networks. In addition to being an employee of NSWC. Mr. Rigby is also my Ph.D. student here at GMU.

Dr. Carey Priebe has been indirectly working with us on related research, but not directly on this AFOSR project. Dr. Priebe was my Ph.D. student and continues close collaboration with me, Dr. Marchette, Dr. Solka, and Dr. Wierman.

## PAPERS COMPLETED UNDER AIR FORCE SPONSORSHIP

1. Wegman, Edward J. and Luo, Qiang (2002), "Smoothings, ridges, and bumps," Proceedings of the ASA (published on CD).

Development of the relationship between geometric aspects of visualizing densities and density approximators, and a discussion of rendering and lighting models, contouring algorithms, stereoscopic display algorithms, and visual design considerations.

2. Symanzik, Jürgen, Wegman, Edward J., Braverman, Amy and Luo, Qiang (2002) "New applications of the image grand tour," *Computing Science and Statistics*, 34, 500-512

Reports the development of a tool for blending multivariate image or time series data to highlight subtle imagery or other structure not found in any single image or time series. Three examples are given.

3. Martinez, Angel and Wegman, Edward J. (2002) "A text stream transformation for semantic-based clustering," *Computing Science and Statistics*, 34, 184-203

Reports on a method for analysis and clustering for streaming text data.

4. Special Issue of *Computational Statistics and Data Analysis*, 43(4), 2003 on the Topic *Data Visualization*, (Wegman, E.J., Solka, J.L., and Martinez, W.L., eds.)

5. Wegman, Edward J. (2003) "Visual data mining," *Statistics in Medicine*, 22, 1383-1397+ 10 color plates

This paper reports on a number of techniques when used in conjunction with each other allows for the exploratory analysis of multivariate data. Those techniques include .parallel coordinate displays, d-dimensional grand tours, and saturation brushing. A number of illustrative examples are given.

6. Wegman, Edward J. and Dorfman, Alan (2003) "Visualizing cereal world," *Computational Statistics and Data Analysis*, 43(4), 633-649, 2003

7. Khumbah, Nkem-Amin and Wegman, Edward J. (2003) "Data compression by geometric quantization," in *Recent Advances and Trends in Nonparametric Statistics*, (M.G. Akritas and D.N. Politis, eds.), Elsevier (North Holland), 35-48.

This paper presents a O(n) method for compressing massive and streaming datasets. Theory is developed to show how to optimally minimize loss and discusses bounds on such loss.

8. Wegman, Edward J. and Marchette, David J. (2003) "On some techniques for streaming data: A case study of Internet packet headers," *Journal of Computational and Graphical Statistics*, 12(4), 893-914

This paper is a key paper discussing both recursive algorithms and evolutionary graphics for analysis of streaming data and gives examples based on Internet traffic data collected via the project. Visual analysis is used to discover anomalies and suspicious scanning activities.

9. Moustafa, Rida E. A., Wegman, E.J., and Priebe, C.E. (2004), "The Power Projection Method," submitted.

A generalized method for mapping high-dimensional data into a line or plane to assist the analyst visualizing and detecting different patterns in large-sized multivariate datasets, allowing the detection of heretofore difficult to observe linear and non-linear hidden structure.

10. Moustafa, Rida E. A., Wegman, E.J. (2004), "A Sign Decision Tree Algorithm for Mining and Visualizing Multivariate Data," submitted.

A new algorithm for automatic cluster detection for a large-sized multivariate dataset using generalized parallel coordinates plots. The associated governing rules of the extracted clusters are also discovered and visualized as well.

11. Moustafa, Rida E. A., Wegman, E.J. (2004), "Generalized Parallel Coordinates Plot," to appear *Seeing a Million*, (Antony Unwin, ed.)

A generalized parallel coordinates plot to assist an analyst in visually exploring hidden patterns in multivariate data, with examples given in application to visual data mining.

12. Solka, Jeffrey L., Wegman, Edward J., and Marchette, David J. (2004) "Data mining strategies for detection of chemical warfare agents," in *Statistical Data Mining and Knowledge Discovery*, (H. Bozdogon, ed.) New York: Chapman-Hall, 71-92

Exploration of the intricacies associated with the construction of various classification systems using data collected from the Navy's SALAD chemical agent detection system, and applications of statistical visualization and density estimation procedures to this discriminant analysis problem.

13. Wegman, E. J. (2004), "On some statistical methods for parallel computation," to appear in *Handbook of Parallel Computing and Statistics*.

Recent focusing of statisticians on the necessity of using high performance parallel computational techniques for data mining of massive data sets, emphasizing the mutual benefit derived from the cooperative use of statistical methods with high performance parallel computing.

14. Johannsen, David A., Wegman, Edward J., Solka, Jeffrey L. and Priebe, Carey E. (2004) "Simultaneous selection of features and metric for optimal nearest neighbor classification," to appear *Communications in Statistics*

15. Chow, Winston and Wegman, Edward J. (2004) "Modeling continuous time series driven by fractional Gaussian noise," to appear *Institute for Mathematics and its Applications Monographs*, New York: Springer-Verlag

16. Marchette, David J. and Wegman, Edward J. (2004) "Statistical analysis of network data for cybersecurity," *Chance*, 17(1), 8-18

This paper is also based on data collected under our CIPIAF project and discusses several modes of attacks and their visualization features. In particular, we discuss backscatter from address spoofing for denial of service attacks.

17. Priebe, C. E., Marchette, D. J., Park, Y., Wegman, E. J., Solka, J. L., Socolinsky, D. A., Karakos, D., Coifman, R. R., Church, K. W., Lin, D., Guglielmi, R., Jacobs, M. Q., Tsao, A., Healy, Jr., D. M. (2004) "Iterative denoising for cross-corpus discovery," to appear COMPSTAT 2004, Berlin: Physica-Verlag.

18. Solka, J. L., Wegman, E. J. and Adams, M. L. (2004) "Man vs. machine - A study of the ability of statistical methodologies to discern human generated ssh traffic from machine generated scp traffic, to appear

This paper focuses on pattern recognition techniques for distinguishing between machine generated and human generated Internet traffic. The focus is to recognize automated attempts to compromise computer systems.

19. Marchette, David J., Priebe, Carey E. and Wegman, Edward J. (2004) "A fast algorithm for approximating the dominating set of a class cover digraph," submitted to *Journal of Computational and Graphical Statistics*

This paper develops a fast algorithm for pattern recognition methodology capable of handling massive datasets.

**RESENTATIONS GIVEN**

1. Wegman, E. J., "Visual data mining," Keynote Talk, C. Warren Neel Conference on Statistical Data Mining & Knowledge Discovery, Knoxville, TN, June 2002.

Discussion of strategies for overcoming limits of screen resolution, human visual system resolution, and available computational resources typically encountered in data mining of massive data sets pursuant to the discovery of structure such as clusters, bumps, trends, periodicities, associations and correlations, and quantization and granularity. This keynote talk resulted in papers 5 and 12 listed above.

2. Faxon, D., "Data compression by quantization," International Conference on Current Advances and Trends in Nonparametric Statistics, Crete, Greece, July, 2002.

Presentation of newly developed advanced statistical procedures to compress terabyte- and petabyte-sized data sets in near real-time to some $10^6$ "representative" observations that permit subsequent statistical and visual multidimensional data analysis with nominally negligible error distortion. This talk resulted in paper 7 listed above.

3. Wegman E. J., "Smoothing, ridges and bumps," invited paper, Joint Statistical Meetings, New York, NY, August 2002.

Presentation on the development of the relationship between geometric aspects of visualizing densities and density approximators, and a discussion of rendering and lighting models, contouring algorithms, stereoscopic display algorithms, and visual design considerations. This invited talk resulted in paper 1 listed above.

4. Wegman, E. J., "On some generalizations of parallel coordinate plots," Seeing a Million – A Data Visualization Workshop, Rain am Lech (nr. Munich), Germany, October, 2002.

Demonstration of parallel coordinates, grand tour, saturation brushing techniques to visualizing large numbers of high-dimensional bank demographic data, an example of the class cover catch digraph approach accommodating large numbers of dimensions, and extensions of parallel coordinates to alternate interpolations and the power projection method. This invited talk resulted in paper 11 listed above.

5. Wegman, E. J., "Visual data mining," Keynote Talk, M2002 SAS Data Mining Conference, Cary, NC, October 2002.

Discussion of strategies for overcoming limits of screen resolution, human visual system resolution, and available computational resources typically encountered in data mining of massive data sets pursuant to the discovery of structure such as clusters, bumps, trends, periodicities, associations and correlations, and quantization and granularity. This keynote talk resulted in paper 5 listed above.

6. Wegman, E. J., "Visualization of packet headers," National Research Council Workshop on Streaming Data, Washington, DC, December, 2002

Description of our project to capture TCP/IP traffic at the University firewall, indicating some recursive methods capable of handling streaming data, illustrating a database interface tool we have developed, and giving some suggestions for visualization procedures we are in the process of implementing. This invited talk resulted in papers 8 and 16.

7. Wegman, E. J., "Preparing electronic books," Symposium on the Interface of Computing Science and Statistics, Salt Lake City, UT, March, 2003

8. Wegman, E. J., "Recursive algorithms and evolutionary graphics for streaming data," Keynote Talk, 2003 Conference on Applied Statistics in Ireland, Mulligar, Ireland, May, 2003

Description of methods for analysis and visualization of streaming data. Introduction of a new data structure and suggestions for methods for addressing such data.

9. Wegman, E. J., "Streaming data and Cybersecurity," Invited Talk, ASA Sponsored Workshop on Homeland Security held at the National Academy of Science, Washington, D.C., May, 2003.

Description of visualization and analytical methods for discovery of Internet-based cyber attacks. Talk resulted in invitation for paper 16 above.

## CONFERENCES & WORKSHOPS

1. *2002 Symposium on the Interface of Computing Science and Statistics, Theme: Geoscience and Remote Sensing*, Montreal, PQ, Canada, April 17-20, 2002.

2. *IDS Workshop*, John Hopkins University, June 11-13, 2002.

3. *C. Warren Neel Conference on Statistical Data Mining and Knowledge Discovery*, Knoxville, Kentucky, June 22-25, 2002.

4. *International Conference on Current Advances and Trends in Nonparametric Statistics*, Crete, Greece, July 15-19, 2002.

5. *2002 Joint Statistical Meeting*, New York City, New York, August 11-15, 2002.

6. *Eighth U.S. Army Conference on Applied Statistics,* North Carolina State University, Raleigh, North Carolina, October 30 - November 1, 2002.

7. *Statistical Analysis of Massive Data Streams,* Board on Mathematical Sciences and Their Applications and Committee on Applied and Theoretical Statistics, National Academy of Science, Washington, D.C., December 13-14 and December 15-16, respectively.

8. *2003 Symposium on the Interface of Computing Science and Statistics, Theme: Security and Infrastructure Protection,* Salt Lake City, UT, March 12-15, 2003.

9. *2003 Conference on Applied Statistics in Ireland,* Mullingar, Ireland, May 14-16, 2003.

10. *DARPA Workshop on Automated Serendipity,* Johns Hopkins University, Baltimore, MD, May 19-20, 2003.

11. *ASA/NAS Workshop on Homeland Security,* National Academy of Science, Washington, D.C., May 29-30, 2003.

12. *AFOSR Workshop on Infospherics,* George Mason University, Fairfax, VA, June 16-17, 2003.

13. *2003 Joint Statistical Meeting,* San Francisco, CA, August 2-7, 2003.

14. *Workshop on Statistical and Machine Learning Techniques for Computer Intrusion Detection,* George Mason University, Fairfax, VA, September 24-26, 2003 (Organized by Edward J. Wegman and John T. Rigsby).