Today's private and public communications networks are critical systems of data terminals, routers, and switches that provide the backbone of our information society. We address the longstanding problem of distinguishing between normal and abnormal network behavior, possibly indicative of an attack on servers, routers or other network infrastructure. Our approach is based on collaborative data collection, anomaly detection and pattern recognition on a large scale. The proposed effort has four components: 1) distributed data collection from participating routers and terminal sites and dissemination of the (appropriately sanitized) data to the research community; 2) development of on-line and off-line approaches for detecting and identifying subtle and complex pattern changes; 3) application to automated detection of intrusions, denial of service (DDoS) attacks, quality-of-service degradations, and other anomalies; 4) development of a comprehensive and multi-disciplinary program in network security education.

Crucial to detecting anomalous changes in aggregate behavior of networks is our ability to determine traffic and packet behavior at a sufficient number of sites and to characterize what constitutes a significant change in behavior patterns. However, the high-dimension and complexity of packet-level patterns in the Internet makes the anomaly detection problem extremely challenging from the point of view of dynamic pattern recognition and detection. With the help of our industrial partners we will collect and analyze multi-dimensional information flows of packets sizes, packet rates, source-destination addresses, and other attributes. In addition to router and backscatter data obtained from data collection sites at Internet2 and Merit Network, we will collect complete header traces from switches and hubs using the *Secure Packet Vault* technology previously developed by one of the co-PI's on this project. We will supplement this data with end-to-end active probing data collected from a consortium of volunteer sites distributed around the network. To manage the massive amounts of collected data we will put in place a simple system for data annotation, mass storage and retrieval, and database software utilities.

Our approach to anomaly detection and localization is a potent combination of emerging techniques in detection, pattern recognition, decentralized information systems, and discrete event dynamic systems (DEDS). Off-line algorithms will be developed and implemented using a combination of statistical learning, invariance, and tree-based classifiers. This will result in flexible hybrid algorithms to correlate events over space-time that are scalable to large volumes of data acquired from a variety of Internet measurement sources. Methods for on-line detection and classification will be investigated using a novel framework that combines stochastic dynamical systems and DEDS. This framework includes both centralized and decentralized data aggregation and event processing. Our aim is to develop implementations that can be used to generate and correlate alerts in real-time with a minimum of human intervention. Our approaches go well beyond previously introduced techniques of fault detection, traffic analysis, and alert correlation that have been restricted to much smaller scale problems. The collaboration of several commercial and non-commercial networking organizations with this project will facilitate technology transfer.

This project will involve precollege, college, and continuing education. An inter-disciplinary undergraduate and graduate curriculum in global network security will be introduced. Students in these courses will participate in data collection, software development, and data analysis as part of instructional projects. We will introduce a summer internship program in networking for high school and middle school students. These students will participate in various educational and recreational signal processing and networking activities. The project will sponsor a small number of scholarships to needy adolescents. An educational innovation of our project is the development of a computer emulation that will generate synthetic traces representative of various types of attacks on a network. These traces will be used in a yearly summer contest for developing the quickest and most effective detection and localization of the simulated attack. Our many industrial collaborators will provide guidance and input including: help in emulating realistic attack scenarios, evaluating response strategies, guidance on testing of our pattern recognition algorithms, and deployment of data collection sites.

1. Introduction and Executive Summary

The principal aims of this project are to study, develop and disseminate: (i) tools for distributed data collection and aggregation; and (ii) methodologies for rapid detection, classification, and localization of spatio-temporal changes in global network traffic. To make headway on such an ambitious aim requires a large-scale broad-based effort and new approaches. We are a multi-disciplinary team of researchers and practitioners from four universities and three Internet service providers in the relevant fields of signal processing, pattern recognition, decentralized detection and control, multivariate statistics, network traffic analysis, network failure detection and diagnosis, network security, and measurements. The project proposes four inter-related areas of activity: distributed data collection; anomaly detection and classification; networking applications; and network security education. These activities are discussed below.

Distributed data collection: Accurate detection of changes in spatially distributed packet-level flow-patterns requires access to data collection devices at many different sites in the network. We will deploy a novel combination of such devices on the Internet. Aggregated data will be collected from Merit, MichNet and other Internet2 networks. In addition to this aggregated data, complete header information will be continuously collected at a dozen or so switches and hubs at Rice, UM and elsewhere; every packet header passing through these sites will be sanitized and stored. We will also deploy a small number of active probing sites to perform end-to-end network measurements. All of this information will be combined and used for our research. We intend to make an anonomized version of portions of this data available to the research community along with database extraction software. We also intend to freely disseminate the end-to-end probing software to any sites interested in participating in our data collection activities.

Anomaly detection and classification: Recognition of subtle spatially distributed patterns is an extremely challenging problem. Our approach is guided by the following principles: 1) scalable algorithms are best implemented in a decentralized and hierarchical manner; 2) sensitive algorithms should use all available information about the underlying models that govern the data collection process in addition to rules or grammars that constrain the "baseline states" of the network; 3) robust algorithms should be insensitive to inaccuracies in models, rules or grammars describing these states. We have two research goals: to implement anomaly detection and classification algorithms for forensic analysis of large traffic databases; to investigate on-line decentralized detection and classification of dynamic spatially distributed anomalies. Particular innovations of our research are: 1) development of hybrid model-based and learning-based schemes for classification of anomalies in packet flow dynamics, backscatter data, and end-to-end measurements; 2) application of a discrete-event dynamic systems (DEDS) and stochastic dynamic systems to develop a framework to emulate baseline operation of the network. 3) integration of pattern recognition, DEDS, stochastic dynamic systems, and non-linear time series models for on-line detection of deviations from the baseline; 4) development of new hierarchical methods for data aggregation, detection, and feature classification from a network of data collection sites.

Networking Applications: With the help of our collaborators we will investigate several practical applications. We will consider detection of distributed denial of service (DDoS) attacks where attackers attempt to flood the buffers of several servers on the network. We will consider detection of coordinated multiple-site intrusions involving robot larceny, i.e., theft of data and other files, viruses, and worms. For these applications we will use all available data types including backscatter, active probing, and Secure Header Vault. We will also investigate service monitoring and verification applications of anomaly detection and classification.

Network Security Education: Improving computer security education at all levels (K-12, college, continuing education) is a major goal of this project. Two peer-reviewed workshops will be organized on the topic of distributed data collection and anomaly detection. We will implement an undergraduate and graduate curriculum at Rice and UM that provides an interdisciplinary three-stream track in networking, security, and signals. In this curriculum global network security will be given the prominence it deserves.

The curriculum will include an instructional laboratory on network security based on an entertaining computer emulation. This emulation will be in the form of a game where teams of students from UM, Rice and elsewhere will match wits against each other on mitigating simulated attacks on their networks. In collaboration with UM's camp CAEN, we will tailor this game to deserving high school students recruited for a two week summer camp on network security. Finally, in collaboration with Merit/MichNet we will engage in outreach activities to K-12 teachers and students.

This project will have the following impact: 1) a fuller understanding of the limitations of global network inference methods for detecting, classifying and localizing potentially debilitating attacks and link failures; 2) development of an on-line methodology for anomaly detection based on stochastic dynamic systems, DEDS, decentralized decision-making, and statistical pattern recognition. 3) implementation of scalable algorithms for detecting and classifying emerging attack patterns from routers, switches and other data-collection sites; 4) dissemination of new software tools for multiple-stream real-time traffic analysis from diverse sources of packet-flow measurements; 5) instruction of undergraduates and high school students on computer security through a entertaining attack-analysis game; 6) multi-disciplinary and practical training of graduate and undergraduate students in Signal Processing, Networking, Statistics, Discrete Event Systems, Optimization, and Security.

The assembled team is ideally suited to the ambitious but important goals of this project. The junior members of our team are rising stars in networking, signal processing, information theory, and statistics. The senior members of our team have distinguished records of success in both specialized and collaborative research projects. All of the senior co-PI's are fellows of the IEEE and have received awards in the areas of research, service, and teaching. Our team includes collaborators and supporters from several commercial and not-for-profit companies and institutions (Arbor Networks, Camisade, Sprint, Lucent, Merit Network, Internet2, Los Alamos National Laboratory, and Stanford Linear Accelerator Center (SLAC)) that have had extensive practical experience in the areas related to our research. Letters from the above mentioned organizations are attached.

A diagram summarizing how the proposed effort is compartmentalized is given in Fig. 1. Principal associated co-PI's for each area are listed in Table 1 of the Management section of this proposal.

DATA COLLECTION	ANOMALY DETECTION	NETWORKING APPLICATIONS	EDUCATION
MichNet router data (Merit Networks)	Forensic detection and classification	Distributed denial−of−service detection	Network security curriculum
Headervault data collection (UM−CITI)	On−line decentralized detection	Intrusion−evasion detection	Workshop and residency program
End−to−end active probing	Discrete event dynamic analysis	Detecting robot larceny	Network security instructional lab
Data management	Hierarchical data aggregation	Backscatter analysis	High school summer camp
Data dissemination	Analysis of packet−flow dynamics	Performance monitoring/verification	K12 outreach program

Figure 1: *Highlights of proposed research and education activities.*

2. Prior NSF Support

1. Multiscale Signal and Image Processing using Singularity Grammars, NSF CCR-9973188 (1999-2002), Richard Baraniuk (PI), Rice University: This project aims to develop a framework for multiscale signal modeling, processing, and analysis for data encountered in networking and image processing applications. To date, we have developed a new class of models based on wavelets and multifractals that matches the highly non-Gaussian and bursty nature of traffic that causes overflow in network routers. Our reduced-complexity model for end-to-end network paths based on a multifractal model is simple, easily trainable, and accurate. These models are in use at a number of research laboratories and universities.

2. Information theoretic analysis of tomographic systems, NSF BCS-9024370 (1993-1995), A.O. Hero (PI), University of Michigan: In this grant more pertinent criteria for design of tomographic data collection systems and new high performance algorithms for reconstruction were developed [55, 61, 54, 41, 40, 53]. The paper [55] won a Best Paper Award from the IEEE Signal Processing Society in 1998.

3. Failure Diagnosis of Modular and Decentralized Discrete Event Systems, NSF ECS-0080406, (2000-

2003), S. Lafortune (PI) and D. Teneketzis (co-PI), University of Michigan: The overall objective of this project is to develop a comprehensive methodology for failure diagnosis of large-scale complex systems with modular and distributed architectures [77]. Our current research and results to-date include: (i) diagnosis of intermittent failures in the context of centralized architectures [27]; (ii) dealing with communication delays in the context of coordinated decentralized architectures [34, 35]; (iii) development of protocols for failure diagnosis of distributed systems based on modular system models; and (iv) study of the computational complexity of diagnosability [146, 148, 147, 32].

4. Information Visualization through Graph Drawing: Modeling, Analysis and Optimization Issues, NSF IIS-9988095; 2001-2003, George Michailidis (PI), University of Michigan: This ongoing project focuses on (1) developing a flexible modeling framework based on graph theoretical concepts that allows the efficient representation of complex data structures, (2) formulating information visualization as an optimization problem and (3) developing efficient, robust, and simple algorithms for solving the problem.

5. CAREER: Fluid Replication, NSF 9984078, 2001-2004; B. D. Noble (PI), University of Michigan: Fluid Replication [92] addresses the problem of client mobility in a wide-area, distributed file system. The effort's goal is to provide the performance, safety and visibility one might obtain in a local-area file system to clients over wide-area networks; this is done through a replication architecture comprising a central server and a set of untrusted WayStations [93]. We have developed a sensitive estimator of link performance quantities [72] that can be inexpensively updated [28] and is effective in wide-area networks [71].

6. A Framework and Methodology for Edge-Based Traffic Processing and Service Inference, ANI-0099148, 2001-2004, R. Nowak (PI), E. Knightly, R. Baraniuk, and R. Riedi (Co-PIs), Rice University: This project focuses experts from the fields of networking, digital signal processing, and applied mathematics towards the goal of characterizing network service based solely on edge-based measurement at hosts and/or edge routers. We blend recent work in multifractal traffic modeling, quality of service (QoS) measurement, and network tomography to develop a unique and innovative framework for network service inference. This project is developing new algorithms and implementations, providing a vital step towards better managing and understanding of Internet performance [22, 23, 24, 25].

7. Scientific Group Communication and Collaboration Testbed for Upper Atmospheric Research, Cooperative Agreement IRI-9216848, 1992-1998, A. Prakash(PI), D. Atkins, T. Weymouth, G. Olson, R. Clauer, and T. Killeen (co-PIs), University of Michigan: We designed an Internet-based collaboratory for collecting and distributed real-time data from various instruments in the space science domain, and supporting collaborative science activities on that data. The system led to a successful follow-on project, SPARC collaboratory, as well as four Ph.D. thesis, education of several M.S. students, and several publications [98, 48, 67, 80].

8. Security and Resource Management in Type-Safe Language Environments, CAREER CCR-9985332, 2000-2004, Dan S. Wallach (PI), Rice University. This project focuses on security issues in language-based systems that run untrusted and potentially hostile programs. By rewriting these programs before they are loaded into the system, new security semantics, including complex access control semantics [136], termination guarantees [108, 109], and transactional rollback [110] can be added to any existing language-based system without unusual performance costs. We have also looked at other security issues, including the performance of TLS Web servers [21] and the security of "secure" digital music standards [29]

3. General Research Approach

The research approach described in this proposal represents a significant departure from existing activities in networking security, measurement, traffic characterization, and mapping. Instead of focusing on the "physics" of Internet traffic, parameter estimation, or model fitting, our focus is on detection, localization, and classification of abnormal network behavior. Instead of using rule-based heuristics to detect such anomalies, we use statistical learning theory, stochastic dynamical systems, and algebras of discrete events

to learn baselines and classify deviations. Our approach is likely to succeed where previous approaches have failed due to high data dimension, small numbers of data points, and model overfitting.

The super-dimensional nature of the feature and measurement spaces presents major challenges. However, patterns of normal and abnormal network behavior or changes in behavior may be embodied in a much lower dimensional manifold. Unfortunately, this manifold is very difficult (or impossible) to describe parametrically and therefore nonparametric pattern analyses form the core of our approach. On the other hand, certain subcomponents of the networking infrastructure are quite well understood and can be modeled accurately. For these subcomponents, model based approaches are unquestionably more powerful and robust. We envision embedding model-based components within the larger setting of non-parametric pattern analysis and machine learning, producing a hybrid that leverages the best of both worlds. When combined with DEDS event-aggregation methods this creates a very flexible framework for analyziing distributed measurements on a large scale. Not only is this a completely new approach in networking, but the underlying theory for large scale, distributed, event-based pattern analysis is virtually unexplored. Theoretical developments in this project will have impact in a broad array of pressing new areas of science and technology including man-made sensor networks and biological networks. This project will combine this new framework with novel and flexible multiple stream traffic data collection, adaptable information aggregation strategies and decentralized diagnostic algorithms to detect changes and localize abnormal network behavior.

4. Data Collection and Dissemination

We intend to collect a large variety of data from a diverse set of sites, which we collectively refer to as the *data collection consortium*. This data will include router traces (Netflow/SNMP) obtained from MichNet and Internet2, backscatter [87, 76] on 35/8 addresses, continuous and complete packet header traces from several switches and hubs at Rice and UM, and end-to-end active probing measurements at Rice, UM and elsewhere. The sheer volume of data to be collected poses special challenges including: archiving, accessing and distributing data to project collaborators; and dissemination of the data to the wider researcher community. In addition we must ensure anonymity of any private information contained in the data. Issues are:

Data collection architecture: A distributed architecture is needed to do summarization, compression, and sharing of data between local data collectors, and for us to be able to run analysis models on the data and compare the results with actual events that transpired.

Throughput, storage, and communication: At high bandwidth links, one must either do less complex real-time analysis or perform sampling and aggregation of packet information. Not all data gathered can be communicated electronically among the local data collection sites or to a central site because of bandwidth and processing costs.

Privacy concerns: some of the data may contain personally identifiable information. Source IP addresses, for example, can be linked to a person making the request. A privacy policy is needed for handling such data. Collected data must be anonymized without adversely affecting its utility for our research on anomaly detection and classification.

Several members of the team have extensive experience in setting up data collection sites, handling data archiving and distribution, and addressing privacy concerns. Below, we point out some of the relevant experience, how it ties in with the proposed architecture, and how the above problems will be addressed.

Proposed Data Collection Process: The general architecture is shown pictorially in Fig. 2. Note that to keep bandwidth requirements manageable, only summary and compressed data will be exchanged among the Local Data Collection Sites; more detailed logs will be exchanged on an as-needed basis via secure file transfers and/or physical media, depending on the volume of data. The Central Data Collection Site is really a virtual site – the same data can be archived at other sites, including a local site.

Data traces will be collected from cooperative sites and probe machines in the network (not shown in the figure). Cooperative sites will include routers, switches and terminals in Internet which are part of our

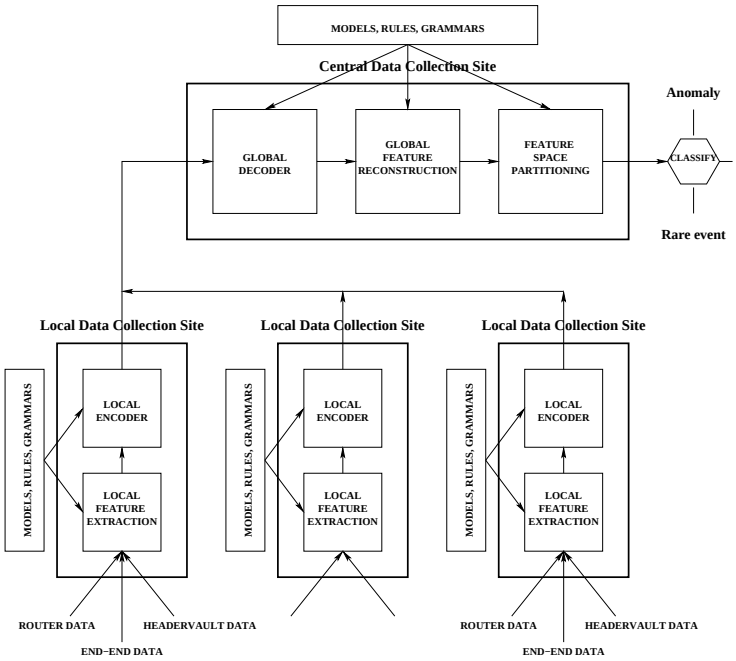

Figure 2: *Block diagram of a two-level hierarchy for data collection and aggregation. There could in general be several layers of intermediate collection sites that would successively aggregate data-collected at lower level sites. Processing and decision-making becomes decentralized when the central collection site is removed and the local sites perform low data-rate message passing to converge on a common decision about the state of the network. Decentralized decision-making is advantageous when a central site is not available or may itself come under attack.*

consortium of sites. Data from the rest of the network will be collected using active probing and end-to-end measurements from sites within the consortium. These two types of data will of course be very different but will be merged into a single information stream within our pattern recognition framework.

Router and Gateway-level data collection: Router data will be collected in collaboration with Internet2 and Merit. Merit manages a regional network, MichNet, in Michigan that connects universities, community colleges, K-12 schools, libraries, state agencies, and cultural organizations. Merit has substantial experience in harvesting networking data using passive listening tools. Some of their earlier network data collection systems as part of research projects include The Internet Rover Package [89] for collecting information on broken network services; the NetSCARF project [90] for querying SNMP-aware network equipment for performance information and making that information available over the web; and collection and analysis of NetFlow data [20]. Netflow data includes source and destination IP addresses/port information, type of service, packet and byte counts, timestamps, TCP flags, and routing information.

Merit, in collaboration with Arbor Networks (also a partner in this project), also has substantial experience in collecting backscatter data. These are packets that come from victims of Denial-of-service attacks and go to random IP addresses, in response to packets with random spoofed source IP addresses from an attacker [87, 76]. They are often indicative of a denial-of-service attacks on the Internet.

Merit will collect such data for the project gathered from MichNet. Security researchers and data analysis researchers on the team will help design techniques to anonymize and categorize the raw data.

Merit will also work with other project staff and students to develop and deploy automatic or semiautomatic methods and procedures for notifying individuals and organizations of significant network events identified by the systems as part of the project.

Peter Honeyman at CITI will enhance and adapt its Secure Packet Vault system [6] to a Secure Header

Vault system for the purpose of this project. By discarding packet payload such a system can handle much higher packet rates than currently feasible, can better manage the volume of data that is generated, and reduces privacy implications. See the discussion of Privacy Concerns and Data Storage and Communication below.

Dedicated probe machines: We will also develop an open collaborative infrastructure for data collection and analysis using a network of dedicated probe machines. To address privacy concerns, all accounts on the machines will be strictly dummy and, wherever appropriate, the data will be discarded after retaining cryptographic hashes of matching pieces of data across probe machines. In that sense of dedicated use, these machines will be modeled after the Honeypots in the HoneyNet project, except that we exclude the goal of encouraging system penetration to study a hacker's behavior on an individual machine since our concerns are pattern detection on network data across a number of machines.

The main task of these machines will be to do end-to-end measurements of network parameters to other probe machines and to selected servers on the network. These machines will log all incoming and outgoing network packets. The data rates at the probe machines are expected to be relatively modest. Thus, the machines will also reconstruct application-specific messages to better understand correlations across application-level data for selected applications to be studied (e.g., distributed denial-of-service, distributed detection of email SPAM and viruses by correlating email data across a number of nodes, and service/performance monitoring).

Privacy concerns: Our partner's experience in handling privacy policies for network data will provide a basis for the privacy policies in this project. Merit already has a privacy policy in place on how this data can be used (see http://www.merit.edu/privacy.html), how long the personally-identifiable data can be kept (at most 72 hours), and provides for specific safeguards to ensure confidentiality of personally identifiable information when used for analysis or research purposes. In the Secure Packet Vault project at UM, the privacy policy mandates that all packet fields be kept encrypted under a public key whose corresponding private key is provided for safekeeping only to an executive officer appointed by the university; the key is to be used only for reconstructing forensic evidence in legal proceedings.

In this project, there is less interest in collecting forensic evidence for legal proceedings. Merit's privacy policy is thus likely to be a more appropriate model. On the other hand, some of the cryptographic techniques used to sanitize the data in the Packet Vault project are certainly relevant to this work. The key research that we will investigate here is to ensure that cryptographic transforms are used in a manner that does not defeat the anomaly detection and classification techniques that are being investigated. In some cases, it may be more appropriate to run classification and data transformation algorithms first and then apply cryptographic transforms.

Besides encryption and cryptographic hashes, other techniques will also be used to provide privacy of the data. This includes converting host addresses to network addresses (by dropping the last 8 or 16 bits of the IP address). Often, the network addresses will suffice for many anomaly detection algorithms.

Finally, where potential concerns remain in providing detailed data for research, each site will work with its legal department to provide legal safeguards before releasing such data.

Data storage and communication: Atul Prakash was a co-PI on the UARC and SPARC collaboratory efforts where space science data was collected from a variety of instruments in real-time and made available to the space science community. As part of that effort, HTTP-based protocols were developed for exchanging data among local data collection sites. As in this project, the data was mostly time-series data. Several tools were developed for time-interval based and attribute-based retrieval and analysis of data. We will reuse these tools and protocols wherever possible.

One key difference from the SPARC collaboratory effort is that the Netflow, backscatter, and Header Vault data can be much more voluminous (data from probe machines is likely to be more modest). The positive side is that storage is getting both inexpensive and compact. In the Secure Header Vault project at

UM, packets can currently be captured, encrypted, and compressed at 100 Mbs rate. In the Packet Vault, one cubic meter of storage is estimated to be sufficient to capture all packets per year, which is not insignificant but manageable. This data should be substantially less voluminous for the Header Vault.

To alleviate the data pressure from multiple collection points and higher data rates (and thus also keep storage requirements bounded), our primary focus will not be on a complete archive of the data but on archival of a rolling subset of the data that allows analysis against selected temporal and spatial dimensions. Several techniques will be employed to limit the storage requirements. First, detailed data will be discarded after a time period (Merit's current policy is 72 hours) unless it satisfies one of several conditions: (1) an interesting event (e.g., distributed denial-of-service attack) occurred during the period; or (2) the data belongs to a selected periodic interval to help create baselines for anomaly detection; (3) the data is a useful summary of the detailed data that is reasonable to archive over a long-term.

A standard data representation scheme will be needed for representing data captured at various routers and probe machines. In this project, the data will mostly be stored in flat files. For each file, converters will be provided to map the data to XML (with appropriate XSL style sheet specs) and to the Common Intrusion Specification Language (CISL) format [38] – CISL is a format that has been proposed for Network Intrusion community as part of its CIDF framework. Wherever appropriate, we will adapt the time-series analysis and display infrastructure from the SPARC collaboratory for use with network data. As needed by the community, selected data can be easily dumped into Postgres databases to support SQL queries on the data.

A number of servers, each capable of storing about a terabyte of data using 180GB SCSI disks, will be deployed at the partner sites. Each site will be responsible for providing web-based authenticated access to the data on the servers.

5. Anomaly Detection and Classification Research

To understand our approach to anomaly detection and classification we refer the reader back to Fig. 2. The local data collection sites shown in Fig. 2 transmit compressed (encoded) versions of features extracted from local data streams. The design of compression and feature extraction algorithms can be based on models (multivariate time series), rules (learning) or grammars (DEDS). A central collection site, which we call the *network manager* site, receives this encoded data, possibly asynchronously, approximates the local features at the decoder and aggregates these local features using a global feature reconstruction algorithm. These global features are then classified by detecting feature clusters or other fixed or data-adaptive feature-space partitioning. The classified features may either be used to refine the models (detection of slowly varying baseline or "rare-event") or to ring an alarm (detection of an anomaly).

We will divide our research into off-line (Sec. 5.1) and on-line (Sec. 5.2) methods. For the off-line case we will apply a combination of statistical learning, invariance, and tree-based classifiers. These techniques are mature, can be used for forensic analysis, and will result in practical algorithms that can be rapidly transitioned to our industrial collaborators in the first year of the project. On-line detection and classification of network anomalies from many local data collection sites is a very ambitious aim for which advances on many fundamental research issues are necessary. We propose a research program below that will flesh out these issues.

5.1. Learning Theory and Pattern Classification

Direct methods of detection and classification can be summed up by the *central tenet* of statistical learning theory [131] "when solving a given problem, try to avoid solving a more *complex* problem as an intermediate step." In the context of the present application this tenet reads: do not focus on parameter estimation or model fitting when the objective is detection and classification of deviations from a baseline. Direct methods fall into two broad categories: statistical learning approaches and invariance approaches. These approaches [58, 59, 60, 70] to detection and classification have led to breakthroughs in high dimensional classification problems for which insufficient numbers of samples are available for model fitting.

These include areas such as hand written character recognition, genetic sequencing, and image indexing [2, 15, 30, 50]. We have had extensive experience applying statistical learning and invariance principles to many different areas relevant to this project including: pattern matching [52], non-linear prediction [86], and cluster analysis [58, 119]. It is therefore reasonable to expect that application of similar methods may lead to similar breakthroughs in network anomaly detection.

Tree classifiers partition the feature space in a hierarchical manner, using a divide-and-conquer strategy, that enables robust and flexible pattern recognition. Tree classifiers can easily handle "mixed" data types and missing data, are robust to outliers and insensitive to monotone transformations of the input features, and are computationally scalable. Recently, we and others have devised a new approach to constructing tree classifiers that provides concrete bounds on the classification [91, 119] performance. We have also applied tree classifiers to universal prediction and reconstruction of non-linear time series for analysis of turbulence and other chaotic signals [86], and to high dimensional biological feature selection for drug discovery problems [46].

SVMs are another very powerful approach to classification of high-dimensional and complex patterns [131] because they convert non-linearly separable patterns to linearly separable ones in a higher dimensional feature space where hyperplane classifiers can be applied. The Vapnik-Chervonenkis (VC) dimension of the hyperplane classifiers measures the complexity of the pattern classifier. This complexity can be used to guard against overfitting to training data and ensures that the classifiers will generalize to new situations The use of the VC complexity measure is also central to our new tree classifiers [119].

5.1.1. Research Issues

To the best of our knowledge tree classifiers, SVMs, and invariance principles have not been applied to large-scale analysis of network data and pattern change detection. This application calls for several important and challenging avenues for new research.

Feature Selection: What network statistics or metrics are most informative for pattern recognition and change detection? How can these metrics be transformed into empirical risk minimization problems?

Unsupervised learning: can one build a baseline based on sets of past measurements and simultaneously detect probable deviations of current measurements from the baseline? Can outlier detection methods similar to the entropic graph methods developed by us [58, 57] be applied to detect deviations from such an unknown baseline?

Model-based Subcomponents: Can well-modeled subcomponents of the Internet and traffic measurements be embedded into a larger, learning based framework? For example, how can known correlations and dependencies between measurement sites be incorporated into tree or SVM classification schemes? Can SVM's, trees and DEDS be woven into a unified anomaly detection and diagnosis algorithm that collectively exploits the strengths of statistical learning and model-based paradigms?

Data Collection/Measurement Placement: Given a limited number of measurement resources, how can these resources be optimally deployed for pattern recognition/detection purposes?

Change Detection/Localization: How well can changes be spatially and temporally localized, and how should active probing methods assist passive data collection to this end?

Decentralization: Most tree classifiers and SVMs act as a centralized scheme. Can these methods be broken up into smaller subcomponents that pass partial pattern classifications between themselves to achieve a more decentralized, and scalable approach to global Internet pattern recognition?

Hierarchical Coarse-to-Fine Hypothesis Testing: Rather than directly attempting a fine-grain classification of network anomalies, perhaps a nested sequence of hypotheses is a more robust approach to Internet pattern recognition. For example, anomalies could first be coarsely categorized into equipment/protocol failures or malicious activity. This coarse classification could feed into subsequent analysis stages that refine these initial hypotheses (e.g., DDoS attacks, spoofing, etc.)

Classification from end-to-end measurements: A corollary to the central tenet of statistical learning theory is that solving the full network tomography problem, i.e. performing link parameter estimation, may be an unnecessary intermediate step towards anomaly detection, classification and localization. In what manner should end-to-end probing measurements be merged with passive packet-flow measurements to form improved features for anomaly detection? How should probe paths be designed to best complement MichNet router data so as to enhance detection and localization performance?

5.2. On-line Distributed Anomaly Detection and Classification

We refer the reader once again to Fig. 2 for the proposed distributed architecture, which in general can consist of two or more levels of data aggregation and processing. At the lowest level, local nodes establish a baseline of local traffic and local packet-level information, and measure, in real-time, local characteristics of the network operation (e.g. local traffic, size of incoming packets, destination of incoming packets, etc.). Based on their on-line (real-time) information the nodes detect deviations from their baseline operation and depending on the hierarchical structure, report these deviations either to the network manager or to intermediate-level nodes that are responsible for monitoring the operation of larger than local portions of the network. The intermediate nodes process the information they receive and report to nodes above them in the hierarchy (eventually, to the network manager).

According to the taxonomy proposed in [7], intrusion detection can be classified into three categories: anomaly detection, signature detection, and compound detection. Conceptually, anomaly detection assumes a partial model of "normal" behavior [121, 134, 73] while signature detection assumes a partial model of "intrusion" [66, 75, 79, 19, 49, 95]. As its name indicates, compound detection assumes partial models of intrusive and normal behaviors [78]. This is the approach that we shall adopt in our investigations.

Below we first present our approach to characterizing the various levels of the aforementioned hierarchy. Then, we will present the research issues at each level of the hierarchy.

5.2.1. Research Approach

As relevant information is contained in the temporal variations of local traffic and packet flows, a stochastic dynamic systems framework is natural for the local data collection sites. Such a framework can yield a compact dynamic systems approximation to the "baseline" of the microscopic evolution of traffic and packet-level information at local nodes or small groups of neighboring nodes. Application of a stochastic dynamic systems framework could be model-based or learning-based. Model-based examples include variants on autoregressive moving average (ARMA) [14] time series models such as: multifractal (MF) ARMA [105, 104, 103, 102, 115, 137, 116] and fractional autoregressive integrated moving average (FARIMA) [69, 10, 88] systems. Learning-based examples include non-parametric phase-space reconstruction algorithms using Taken's imbedding methods [86, 130], classification and regression trees (CART) [86, 8, 120, 13, 11, 50, 91], multivariate splines [135], and state space particle filters [24, 36]. Features of the residual prediction errors produced at the local nodes will be defined and used both to establish local baselines and to detect deviations from this baseline. Deviations are classified and transmitted to the higher levels of the network hierarchy.

Upper levels of the data-collection hierarchy aggregate the received locally-encoded features into global features as illustrated in Fig. 2. We will investigate both learning-based and model-based approaches to feature aggregation. In the model-based approach we will use a DEDS framework to capture the operation of the intermediate and highest levels. The DEDS framework, either logical [17] (which is based on automata and language theory) or stochastic [12] (which is based on Markov or semi-Markov chains), provides a compact description of what one might call the "macroscopic" evolution of the network. DEDS can efficiently describe patterns of normal network behavior and be used to detect changes in behavior of individual nodes or larger portions of the network. The intermediate levels and the central data collection site receive information in the form of messages from lower levels. Such messages report deviations (such as "increase in traffic in a certain part of the network") or some "statistic" carrying information from lower

levels (such as "likelihood" of a traffic anomaly in a certain part of the network). Such deviations are caused by events that are "unobservable" by the data collection nodes. Examples of unobservable events include "initiation of attacks on the network" or "rare increases in normal traffic". Based on the received messages, the intermediate levels and the network manager have to estimate their "aggregate" state and detect attacks. Attacks on the network are modeled as sequences of observable and unobservable events over space and time, or patterns of network behavior over space and time.

To improve the quality of its estimates (therefore, the quality of its diagnostic decisions) the network manager may query lower levels so as to acquire specific information. To provide the information requested by the network manager, the lower levels may adjust their rules of acquisition and processing of information. These adjustments, together with information received on-line, lead to the adaptation of the dynamic systems describing the operation of local nodes. The information received by the network manager as a result of its queries to the lower levels may lead to adaptation of the DEDS describing the operation of the higher levels. Thus, the two-way interaction among various levels of the hierarchy leads to a continuous "learning' of the network's operating environment and a continuous improvement of the quality of information upon which the detection of network attacks is based. Furthermore, these interactions point to the important research issues at each level.

5.2.2. Research Issues

To achieve its objectives the network manager must ensure that: (i) it has a DEDS that represents adequately the operation and dynamic evolution of the network; (ii) it has the quality of information that allows it to make the correct decisions about the status of the network; and (iii) it employs decision rules that effectively utilize the information available to it. Consequently, three classes of problems we propose to investigate are: (1) Active acquisition of information; (2) Data fusion/coordination mechanisms; and (3) DEDS updating mechanisms that are based on the results of active acquisition of information.

Within the logical and stochastic DEDS framework, we propose to formulate "active acquisition of information" as an optimization problem where the network manager has to select the information it requests from the intermediate levels to achieve its goals. We will assume that information is "costly" because it increases the "overhead" of the network operation due to data processing. This will result in a tradeoff between acquisition of information and detection capabilities.

Data fusion/coordination mechanisms integrate the information sent to the network manager by nodes of the intermediate levels. For logical as well as stochastic DEDS we propose to use the methodology developed in our prior investigations [113, 114, 112, 33, 77], which has been successfully demonstrated in practical applications: large-scale telecommunication networks [1, 96, 9] and wireless LANs in vehicle platooning [31]. The results of [113, 114, 112, 33, 77] deal primarily with logical DEDS and "simple" fault events. To develop effective data fusion/coordination mechanisms for the network manager we must extend our methodology in two directions: (i) develop a diagnostic methodology similar to that introduced in [113] for stochastic DEDS; (ii) generalize the notion of "failure-type labels" (introduced in [113] for tracking unobservable fault events) to "sequences of labels" over space and time that capture partial/complete attack patterns from a database of such patterns.

Information that is received from active acquisition of information and cannot be unambiguously interpreted by the network manager will lead to an update and refinement of its DEDS. Our approach to this issue will combine results from active acquisition of information and adaptive control and learning theory for logical and probabilistic automata; see [74, 151, 45, 44, 43] and the references therein.

A key feature of the intermediate levels of the hierarchy is that information at each level is decentralized. Intermediate nodes possess different information about the status of the network. The fundamental research issues associated with the design and operation of these levels are: (i) The determination of the information partition. Given that local nodes will have to report to one, or perhaps more than one intermediate nodes, we must specify "which local nodes report to which intermediate node." (ii) After an information

partition is determined, the information across intermediate nodes will, in general, be correlated. Consequently, the intermediate nodes will have to jointly determine the rules that specify their communication with the network manager. (iii) Based on the "active acquisition of information" instructions they receive from the network manager, the intermediate nodes have to determine how to query local nodes to receive the information requested.

The information partition at the intermediate levels of the hierarchy has to achieve the following objectives: (1) To provide each of the intermediate levels nodes with the quality of information that is necessary to ensure that each node effectively monitor its portion of the network. (2) To create a high-degree of "informational redundancy" at the intermediate levels of the network hierarchy, which ensures that the network manager's quality of decisions does not significantly deteriorate when one (or more) of the intermediate nodes fails or is attacked. The literature on information structures and partitions [107, 37, 39, 85, 106, 145, 142, 139, 3, 4, 127] will form the basis of our approach to designing information partitions.

The classes of problems in research issues (ii) and (iii) above are similar in nature because they address the joint determination of optimal (with respect to some performance criterion) decision rules by a group of decision makers that have the same objective and different but correlated information. When these problems are considered in connection with the "active acquisition of information" problem, they result in "dynamic team" problems. The approaches to solving dynamic teams [140, 65, 141, 133, 150, 132, 149, 100, 144, 138, 126, 63, 143, 125, 97] are computationally formidable. In contrast, when the above problems are considered in isolation from the "active acquisition of information" problem they result in "partially nested" teams [64, 18], which are simpler than dynamic teams and have been successfully applied to decentralized detection [150, 128]. We expect that by studying the problems that arise in research issues (ii) and (iii) in isolation from the "active acquisition of information" problem we will describe effective guidelines and heuristics for their solutions. Such guidelines and heuristics will also be developed using the decentralized analogue of the sequential Monte Carlo Markov chain framework that appeared in [24]. The classes of problems in research issue (ii) can also be viewed as a "modular diagnosis" problem. since every intermediate node has a different DEDS of the overall network. We propose to approach such a modular diagnosis problem by using and extending our prior work on diagnosis of DEDS [113, 114, 112, 33, 77].

In addition to the research issues discussed in Section 5.1, local nodes have to detect deviations from their baseline operations, classify them, and report them to the intermediate levels. Furthermore, local nodes have to respond to requests associated with acquisition of information by higher levels.Within the context of stochastic dynamic systems (MF, ARMA, FARIMA), these issues are conceptually similar to research of issues (ii) and (iii) discussed above regarding the intermediate nodes.

6. Applications

A large fraction of DARPA's Fault Tolerant Networking funded projects and commercial products from both established companies, such as Cisco Systems, and a flock of startup companies, such as Arbor Networks, Asta, Mazu, Reactive, etc., are proposing deployment of Internet-wide infrastructure to combat DDoS attacks, intrusions and worms. The pattern recognition and detection framework described in the previous sections can be applied to the early detection, prevention, and analysis of a variety of known attacks. Furthermore, using the unsupervised learning framework discussed in the previous section, outlier detection from an estimated baseline can be used to help discover new types of attacks. When supplemented by active probing methods, we can perform tomography to detect changes in topology, latency, and other types of active service monitoring tasks. In this section we describe example applications that will be studied in this project, their relationship to the more general detection framework, and our proposed approaches.

DDoS Attacks: Distributed denial-of-service (DDoS) attacks are an increasingly damaging class of coordinated attacks launched from several attacker sites each of which sends malicious packets to the victim

at very high rate/intensity over a period of time. Attacks and intrusions are becoming more-and-more prevalent (as many as a dozen attacks per week on the UM EECS servers recently), target a wide variety of hosts (servers, routers, terminals), and can last for hours, days and sometimes longer. Emerging DDoS attacks are typically accompanied by subtle and simultaneous changes in traffic-level and packet-level statistics at different sites. This makes DDoS an ideal application for our distributed anomaly detection methodology. Most current approaches to detection rely on distributed traffic correlation capabilities to detect anomalies. The basic architecture proposed invariably involves distributed monitoring, some form of distributed traffic correlators fed by these monitors, and installation of traffic filters for detection of traffic anomalies. The key research issues in building such an architecture include: How to detect attacks with minimal false positives? How to mitigate the attacks? And how to do both in a scalable and timely manner?

Ideally, the earlier the attack is detected, the earlier actions can be taken and the better the network can be protected. The difficulty is that the change in traffic pattern can be very subtle across the network due to the distributed nature of these attacks. Statistics from multiple locations have to be carefully correlated to detect such attacks. Another difficulty is that accurate detection of such weak changes is inherently subject to high false alarm rate. We initially will concentrate on detecting spatial and temporal changes using off-line statistical learning methods discussed in Section 5.1. As the research of Section 5.2 progresses, on-line implementations will be investigated. We will evaluate the responsiveness, as well as accuracy of our system under simulated attacks. The focus will be on enhancing the responsiveness and at the same time decreasing the probability of false alarms.

Backscatter Analysis: One particularly effective type of DoS attack can be achieved through a "SYN flood" [26], which consists of a stream of TCP SYN packets directed to a listening TCP port at the victim's site. Such a mechanism can be rendered extremely powerful, if it can be used from a set of compromised Internet nodes, where attack daemons producing a group of "zombie" hosts can be installed. The result is a coordinated attack from numerous zombies onto a single site. Hence, SYN packets with different spoofed IP addresses arrive at the victim, who is rapidly overwhelmed by the various requests. A key feature of this mechanism is that, to cover his tracks, the attacker spoofs his source IP addresses by randomly setting the IP source address field in his transmitted packet header. As shown by researchers at CAIDA [87], Merit Networks and Arbor Networks [76], any ISP that has access to a large chunk of unused address space can detect the presence of a major attack. These messages are part of the *backscatter* created by the attack, and also identify the victim (but not the attacker). As backscatter is generated by any spoofing mechanism it has also been used to detect and localize victims of intrusions, worms and viruses [68]. In collaboration with Arbor Networks, Merit has exploited its *class A* address space (called 35/8 address space in the current classless IP address allocations) to track DoS using backscatter analysis [82]. We propose to work with Arbor Networks and Merit to extend backscatter analysis in several new directions including: 1) apply multivariate time series analysis to the backscatter detected in Merit's 35/8 over time to identify trends and models for normal vs. malicious backscatter signatures; 2) build a baseline and a classifier to recognize normal and anomalous backscatter patterns based on the latest statistical learning methods, including classifier boosting and randomization methods developed by us [52, 51] and others [2, 42, 13]); 3) investigate the applications of DEDS, and of simpler sequential detection techniques developed by us [129, 56] and others [124] (e.g. sequential probability ratio tests or cumulative sum tests), to perform on-line detection and classification of attacks as they evolve.

Intrusions and Evasions: While network capacity is threatened by DDoS aimed at CPU cycles and memory, networks are constantly exposed to more subtle attacks. Robot larceny and espionage constitute a common form of malicious activity targeting actual content stored at servers, thus attempting unapproved access to networking resources, sensitive data and theft of marketable information. This type of attacker will use a spoofed source address and break into a system. The attacker may sometimes be detected by a host-based or a network-based intrusion detection system (IDS). A host IDS may check logfiles or look for

signatures indicating a suspicious sequence of events generated by the operating system or an application during. A network IDS detects suspicious packet behavior, e.g. TTL manipulations on TCP, unusual IP options, or timing patterns [47]. However, attacker evasions of IDS systems are common and frequently exploit small ambiguities in the applications layer. The evasion strategies can be quite complex, sometimes eluding even the most sensitive pattern recognition or anomaly detection algorithms [117, 118]. We think that these ambiguities could be efficiently captured by the proposed DEDS framework, possibly leading to greatly improved detection of new evasion strategies. We will develop a DEDS model for the state space of events that describe user interactions with certain web applications. Another approach we will investigate is to use pattern recognition to discriminate between human behavior and behavior of coordinated machines that are expected to be much less correlated over space and time [83, 84].

SPAM and viruses: SPAM and viruses via email are a growing problem. The current solutions to SPAM that are being explored tend to be local – based on text and header categorization of received email, based on machine learning techniques (for example, see [81, 5, 99, 111, 62, 94, 16]). A more distributed approach is exemplified by the Vipul's Razor [101], in which a catalog of SPAM is maintained. This catalog can used by clients to filter out known SPAM. End-users are responsible for reporting a SPAM message to the catalog server, which is then used by other users for filtering. Using the packets directed at the SMTP ports of the probe machines (which only have dummy accounts) as well as the cryptographically hashed data at the packet vault directed at the SMTP ports, we believe that there is an opportunity to explore the design of a distributed and automated approach to detecting and characterizing SPAM and viruses. The research questions we will attempt to answer are: Can email viruses, which are often sent as attachments of well known types, be distinguished from valid attachments of the same type based on distributed pattern recognition techniques?

Service/Performance Monitoring: It is now common for Internet service providers to offer a variety of service levels to customers. Service level agreements specify performance criteria that the network provider guarantees to satisfy. Such quality-of-service (QoS) criteria can include the amount of bandwidth made available to the customer and bounds on the maximum delay (which is important for Internet telephony and streaming applications). However, anomalous behavior such as malicious activities and faults can severely degrade network service. Such anomalies are reflected in spatially localized packet delay and loss distributions. For example, a DDoS attack may disrupt service in a subnetwork, producing measurable losses and delays at a remote site. From the perspective of a service provider (or a customer of a provider) it may be important to determine if the anomalies are occurring within the service providers network or if they are external. While it may be relatively straightforward for a provider to detect problems within their own network, detecting and localizing anomalous behavior in other portions of the Internet that may be affecting their service is highly non-trivial.

The ability of network tomography to localize pathological network performance to individual components or subnetworks could aid in the early warning and detection of attacks and intrusions. In this project we will investigate the use of our network tomographic methods [22, 23, 25, 122, 123] to detect and localize patterns of change in unobservable portions of the network; e.g., changes in routing topologies due to downed links during an attack and evolutionary patterns in delay and dropped packet statistics. Furthermore, on-line network tomography methods like those developed in [23] can provide spatio-temporal localization of anomalous changes, allowing for very rapid detection of attacks and failures. An additional key research issue it to carefully investigate how tomographic methods (based on active probing) can supplement passive data collection at measurement sites in our infrastructure. We envision focused adaptive, active probing, driven by the passively collected data. In effect, active probing and on-line network tomography will allow us to "fill in the gaps" between the the incomplete set of measurement sites in the infrastructure, without overburdening the network with large amounts of additional probe traffic.

7. Education

The scope of this effort will provide many opportunities for students to be involved in research. Because the research draws from a number of traditionally separate fields, the project also informs the curricular process. This effort will deploy networking laboratories within the host institutions to provide hands-on learning opportunities to undergraduate and graduate students, and include plans for outreach to precollege students. The project will also coordinate a number of broadly-targeted "competitions" to provide both visibility and wide-ranging educational opportunities focused on the problem of global-scale security.

Security Curriculum: The proposed curriculum draws from a number of distinct, traditionally separate fields, including security, cryptography, networking, software systems, signal processing, and modeling. Students must have a deep understanding of some subset of these topics, plus a broad appreciation for the remainder. This effort proposes to provide such training through a two-pronged approach of course development and curriculum organization. Co-PIs at each institution are developing or have recently developed courses in a number of these areas, spanning the senior undergraduate level through research-oriented graduate courses. These courses will be organized in three mutually-supporting *threads*: security, networking, and signals. At each institution, each thread consists of one (or more) introductory courses plus at least two courses focused on a more specific topic within the thread. The introductory courses include broad coverage of a wide variety of topics, offered to senior undergraduates and first-year graduate students. The more focused courses are intended to prepare our graduate students for research in the area. Each student on the project is expected to complete two courses from a single thread for depth, plus any two courses from the other two threads for breadth. These courses have already attracted industrial support from Schlumberger, Intel, IBM, and others. Some of these courses may be offered jointly by both sites, utilizing the distance learning capabilities each institution already has in place. We will look to develop a new interdisciplinary course on Global Network Security which will incorporate the diverse expertise of the research team on this project.

Instructional Laboratory: Students at Rice and Michigan who are participating in the research effort will also have access to a networking laboratory. This laboratory, described in more detail in the budget justification section, will serve the dual purpose of educating students in network security and providing a testbed for our research. In this lab students will learn about attack strategies and mitigation (on a scaled down, "private" network emulation) and will also generate attack scenarios for testing.

Network Security Competition: We propose to offer periodic competitions—open to students at any institution—for detecting and analyzing simulated attacks which will be emulated from data collected over the course of the project. Each competition will commence with a public release of sanitized data that contains one or more known attacks, along with the toolsets that have been developed up to that time. Entrants will be judged based on how quickly they can isolate attack signatures, along with the specificity and recall of the attack identification. These competitions provide educational opportunities beyond the host institutions, as well as active feedback on the tools and data collection process, allowing continuous improvement of both.

K-12 Activities: This project also proposes to incorporate pre-college outreach activities, in order to build a sustainable population of interested, talented students. There are two avenues available for such activities. First, Merit acts as the Internet service provider for K-12 institutions throughout the state of Michigan. This relationship provides direct access to local administrators at these institutions, and thus both channels for outreach to and feedback from these institutions. By participating in workshops offered by Merit and attended by K-12 staff, we can advertise opportunities and potentially influence advanced computing curricula at these institutions. Second, the College of Engineering at UM also organizes Camp CAEN, a summer camp with a computing focus for students between the ages of 13 and 17 (`http://campcaen.engin.umich.edu/`). Co-PI's at Rice and Michigan plan to organize courses within Camp CAEN introducing students to security and monitoring issues. Camp CAEN also provides an all-day, girls-only offering, providing a more supportive environment in which to attract more women

to the field. Every year scholarships for kids will be offered (included in the budget). Special efforts will be made to recruit economically disadvantaged kids with the help of CAEN admissions staff and our K-12 MichNet outreach program.

Rice and Michigan, in establishing a summer program for high school students, hope to create excitement in these students that will hopefully carry them forward through computer science as a major and as an eventual profession. By getting students to think of security issues now, when most of them have had only basic training in programming, we hope these students will learn to approach their future education with an eye toward robustness in the face of malicious threats. We note that our program is explicitly not designed to teach students to be "script kiddies". While we do intend to teach how such tools work, the focus will be on how to defend against such attacks. We will also include an explicit lecture on the ethics of performing research in computer security.

In a two-week course, there is a limit to the depth of material that can be covered, particularly for students without a college-level background in computer science. The main curricular elements will be:

- explaining how modern computer systems work (operating systems, networking, file systems, and so forth), including discussing vulnerabilities that have been found in these systems.

- hands-on work with current security tools, including packet sniffers, intrusion detection systems, virus scanners, and the like.

The summer program will conclude with an adolescent-adapted version of the network security competition, described above, pitting teams of summer-camp students against each other. We will place these students in a testbed network environment of machines for which they have full administrative privileges. We will have a "malicious" host that is attacking the students' network. Before the malicious host is introduced, the students will have an opportunity to install tools and prepare for being invaded. Then, the malicious host will begin attacking their network, using some off-the-shelf attack tools unknown to the students. The students will be responsible for identifying and cleaning up the attack. Appropriate firewall technologies will be placed around the student network to prevent these attacks from "leaking" out to the actual Internet. The competition, as such, will be to see which student group can resolve the attacks first.

This summer camp and competition experience will provide students with an opportunity to learn real world skills and gain an appreciation of the damage that can occur when a site is under attack. We believe this will give students an invaluable insight into how the Internet really works (and does not work).

8. Impact of Project

The considerable vulnerabilities of open communication networks will continue to be exploited and it is inevitable that new weaknesses will continue to be discovered. This creates enormous challenges for network security in general and detection of malicious attacks in particular. We have assembled a team with the combined expertise necessary to make an impact on the challenging network anomaly detection problems discussed in this proposal. While there are other university-led research activities in networking, including a NSF funded ITR-project focussed on network traffic analysis ("A multiresolution analysis for the Global Internet"), our team is unique by its combined strengths in the areas of network measurement, traffic analysis, network tomography, discrete event-dynamical systems, pattern recognition, network security, and education. If funded the project would likely have major impact on these areas. The participation of our commercial and non-commercial collaborators at Arbor, Sprint, Merit, Internet2, ..., virtually guarantees rapid transfer of technology developed by the co-PI's (one of whom (Ogden) is with Merit Networks). The involvement of leading researchers in network failure detection at INRIA(France) and network data analysis at McGill(Canada) contributes an important international component to network security research. The project's impact on network security education will be broad and significant, affecting K-12 teachers, high school students, undergraduates, graduate students, and continuing education.

8. Management Plan

To accomplish the research and education aims of this project requires a *focused large scale and multi-disciplinary effort*. The breakdown of senior personnel by sub-areas in Table 1 illustrates the balance of our team in the four areas of this project.

The coordination of collaborative projects which span four colleges (UM, Rice, McGill, IRISA-Rennes) and three networking organizations (Arbor Networks, Internet2 and Merit) requires a tight management plan. Research partnerships which lead to productive inter-disciplinary collaborations will be essential for success of this effort. We will allocate 0 to 2 GSRA's per co-PI depending on research needs. All GSRA's will have at least two co-PI's on their thesis committees to further the collaborative aims of this project. In addition we will develop and team-teach courses combining signal processing, control systems, statistics, and network security in the context of the new curricula we are developing at UM and Rice. Furthermore, close collaboration with collaborators, contacts, and co-PI's at Internet2, Merit, Lucent, SLAC, and Arbor Networks on solving real-life networking security problems will help maintain focus and relevance of the research.

National Advisory Committee: A National Advisory Committee (NAC) will be created with representatives from industry, government, and academia. The role of this committee will be to annually review the work carried out on this project, provide guidance on future directions of the research, and help identify ways to improve practical impact of the project. We have agreements with the following companies and organizations: Sprint, Stanford Linear Accelerator Center (SLAC), Merit Networks, Los Alamos National Lab (LANL), Lucent Bell Labs (Murray Hill), Internet2, Camisade, and Arbor Networks to serve on the NAC (see attached letters). One other organization (Google, Inc) has also expressed interest and will be asked join the NAS if the project is funded. There will be an annual meeting of the NAC every spring which will occur right after our internal project review meeting.

Internal Advisory Committee: An Internal Advisory Committee (IAC) will be created with representatives from various university organizations. The prime function of the IAC will be to advise us on issues of privacy and human subjects implications of our data collection and dissemination activities. The IAC will include representatives from legal, computing, and networking organizations within Rice and UM. Issues of privacy and human subjects will be addressed with the IRB's (Internal Review Board) of UM and Rice if this proposal is funded.

Electronic Dissemination: A website will be created as an archive for research reports and articles, sample data traces, interactive software, course materials, and announcements. This website will be accessible to the public. It is our intent to make some of the data collected available to the public, along with terms and conditions of use, on this web site. Part of the project administrator's job will be webmaster for this site.

Residency program and minisymposia: Each year we will run a two-week residency program in network security at UM. This will be a small and selective "by-invitation" program run during the summer session at the University of Michigan. The aim of the residency program will be to gather together top researchers from academia and industry around a topic or theme. Initially we will focus on data collection, pattern recognition, and global security. The residency program will be structured as follows. Each year names of potential session organizers will be solicited from co-PI's, collaborators and others, e.g. the NAC. The slate of names will be forwarded to the project executive committee and two organizers will be selected to organize focussed minisymposia during one of the two weeks in the program. The organizers would each control a budget to reimburse all inviting participants for travel and lodging expenses for their two-week residency. UM facilites (Cambridge House or residence halls) would be used for lodging to cut down on expenses.

Biennial Workshops: We will organize a biennial workshop on Data Collection and Anomaly Detection which will have keynote speakers, special invited sessions, and contributed sessions. The workshops will be aimed toward the research community. The workshops will take place over three successive days. They will

co-PI	Data Collection	Anomaly Detection	Applications	Education
A. Hero(UM)		X	X	
R. Baraniuk(Rice)		X		X
A. Benveniste(INRIA)		X	X	
M. Coates(McGill)		X	X	
P. Honeyman(Merit)	X		X	
S. Lafortune(UM)		X		X
M.Y. Liu(UM)			X	X
G. Michailidis(UM)		X	X	
B. Noble(UM)	X			X
R. Nowak(Rice)		X	X	
J. Ogden(Merit)	X		X	
S. Pradhan(UM)	X	X		
A. Prakash(UM)	X			X
R. Riedi(Rice)	X	X		
D. Teneketzis(UM)		X		X
D. Wallach(Rice)			X	X

Table 1: Matrix of associations between senior-personnel/collaborators and sub-areas of this project.

have a strong education component involving tutorials on network traffic measurement, network security, and network modeling. At each workshop there will be a session on novel classroom teaching methods for lower level signal processing and networking courses. We will also have sessions featuring papers presented by students (undergraduate and graduate) on networking projects completed over the previous year in connection with this grant.

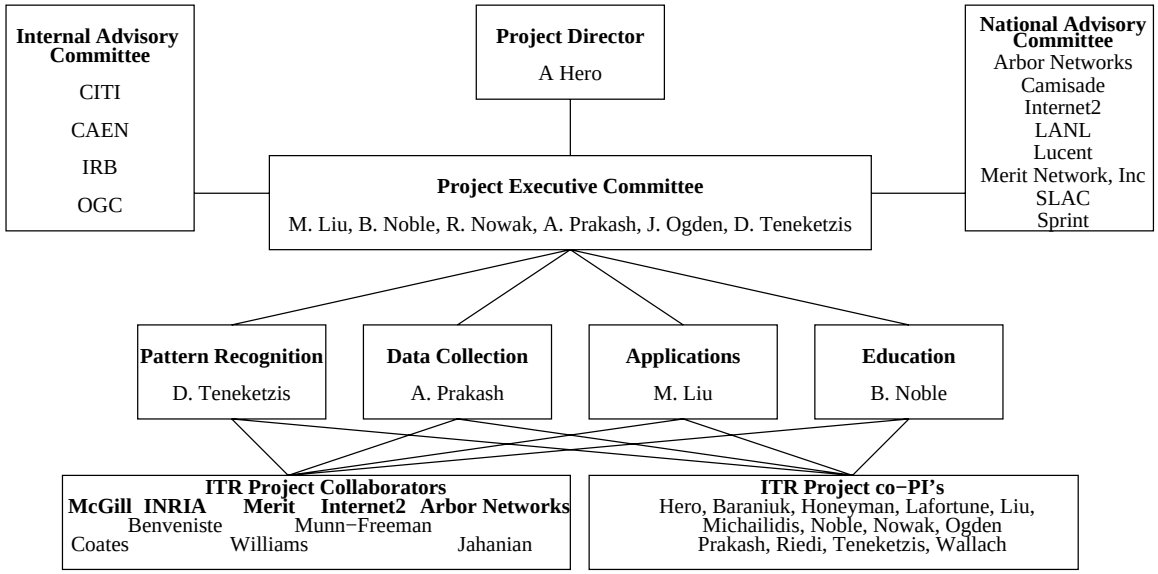

Figure 3: *Management structure for the project. Executive committee will operate as the clearinghouse for all decisions and will solicit inputs from the National Advisory Committee (NAC) - whose members are indicated - and an Internal Advisory Committee (IAC) which will be composed of representatives from CITI (UM Center for Information Technology Integration), CAEN (UM Computer Aided Engineering Network), OGC (UM Office of General Counsel), and the IRB (UM Internal Review Board). The executive committee will make decisions on resource allocation, area progress, and other operational issues.*

Team Management Structure:

The project will have a three level management structure illustrated in Figure 3. Four area leaders (B. Noble, M. Liu, A. Prakash, D. Teneketzis) will be in charge of their respective areas and will have the following responsibilities:

- Holding regular meetings of co-PI's and collaborators associated with their areas.
- Assessing and reporting on progress in their area.

Year	1					2					3					4					5				
Month	0	3	6	9	12	0	3	6	9	12	0	3	6	9	12	0	3	6	9	12	0	3	6	9	12
Education			■	♦		◎		■	♦		◎		■	♦		◎		■	♦		◎		■	♦	
Workshops	▶							★	⊞					⊞				★	⊞						◀
Summit (NAC)			✠					✠					✠					✠					✠		
Execom Meetings	♠	♠	♠	♠	♠	♠	♠	♠	♠	♠	♠	♠	♠	♠	♠	♠	♠	♠	♠	♠	♠	♠	♠	♠	♠

Table 2: Time line for educational activities, hosted workshops and other outreach programs, summit meetings of all participants with the National Advisory Council (NAC), and Execom meeting.

■	camp CAEN for high school kids
♦	Development of attack emulations
◎	Rice/UM gaming competition
▶	Kickoff meeting
★	Anomaly Detection Workshop
⊞	Summer residency program
◀	Wrapup meeting
✠	Annual NAC/Project meeting
♠	Excom meeting

Table 3: Legend for timeline (Table 1)

- Identifying potential difficulties impeding progress.
- Collecting input from co-PI's and drawing up a set of year-end goals for sub-projects in their area.
- Reporting on their area to the executive committee (see below).

Major project decisions will be made by Prof. Hero in consultation with the executive committee (Excom), formed by the four area leaders plus a representative each from Rice (R. Nowak), and Merit Inc., (J. Ogden). The Excom will meet 5 or 6 times per year and will have the following responsibilities:

- Review overall progress of the project and identify potential difficulties and opportunities.
- Review the specific year-end goals for each research area.
- Allocate resources to research areas and to co-PI's based on project-relevance and past performance.
- Meet in closed session with the NAC once per year.
- Meet with the IAC as needed.

A comprehensive year-end review of progress in each area will be based on the following criteria: the effectiveness of collaborations; innovations in theory, algorithms, data collection, or education; and dissemination (journal and conference publications, web-tools, technical transfer, data disseminated, etc).

Prof. Hero will be aided by a project adminstrator, the executive secretary, who will help manage the day-to-day activities of the project. A part-time staff person will be responsible for monitoring the data collection sites and overseeing maintenance of the database of traffic traces at UM and elsewhere. Another part-time staff person will help us to maintain the networking laboratory. For more details on staff see budget justification.

A time-line for the five year duration of the project is given in Tables 2 and . Each year we will have face-to-face meetings involving all co-PI's, collaborators, and the NAC. At this meeting co-PI's and collaborators will present previous year's research and education results to the NAC. Other industry and government representatives will be invited to attend and participate in these meetings. In addition to these larger meetings there will be several smaller meetings during the year. We will utilize webcast/videoconference/teleconference facilities at Rice and UM to include all co-PI's in these meetings. Several UM co-PI's will visit co-PI's at Rice, and vice versa, for more extended periods.

D Bibliography

References

[1] A. Aghasaryan, E. Fabre, A. Benveniste, R. Boubour, and C. Jard, "Fault detection and diagnosis in distributed systems: An approach by partially stochastic Petri nets," *Discrete Event Dynamic Systems: Theory and Applications*, vol. 8, no. 2, pp. 203–231, June 1998.

[2] Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," *Neural Computation*, vol. 9, pp. 1545–1588, 1997.

[3] M. Andersland and D. Teneketzis, "Information structures, causality, and nonsequential stochastic control I: Design-independent properties," *SIAM J. Control Optim.*, vol. 30, pp. 1447–1475, 1992.

[4] M. Andersland and D. Teneketzis, "Information structures, causality, and nonsequential stochastic control II: Design-dependent properties," *SIAM J. Control Optim.*, vol. 32, pp. 1726–1751, 1994.

[5] I. Androutsopoulos, J. Koutsias, K. Chandrinos, G. Paliouras, and C. Spyropoulos, "An evaluation of naive bayesian anti-spam filtering," in *Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000)*, G. Potamias, V. Moustakis, and M. van Someren, editors, pp. 9–17, Barcelona, Spain, 2000. `http://arXiv.org/abs/cs.CL/0006013`.

[6] C. Antonelli, M. Undy, and P. Honeyman, "The packet vault: Secure storage of network data," in *Proc. USENIX Workshop on Intrusion Detection and Network Monitoring*, Santa Clara.

[7] S. Axelsson. *Intrusion Detection Systems: A Survey and Taxonomy*. `citeseer.nj.nec.com/axelsson00intrusion.html`.

[8] A.-E. badel, O. Michel, and A. Hero, "Arbres de regression: modelisation non-parametrique et analyse des series temporelles," *Traitement du Signal*, vol. 14, no. 2, pp. 117–133, June 1997.

[9] A. Beneveniste, E. Fabre, C. Jard, and S. Haar, "Diagnosis of asynchronous discrete event systems, a net unfolding approach," Technical report, IRISA, Rennes France, 2002.

[10] R. J. Beran, *Statistics for Long-Memory Processes*, Chapman & Hall, 1994.

[11] L. Breiman, J.H.Friedman, R. Olshen, and C. Stone, *Classification and regression trees*, Wadsworth, 1993.

[12] P. Brémaud, *Markov chains - Gibbs fields, Monte Carlo Simulation, and queues*, Springer, New York, NY, 1999.

[13] L. Brieman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996.

[14] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, Springer-Verlag, New York, 1987.

[15] C. Burges and A. S. (Eds), *Advances in kernel methods - support vector machines*, MIT, Cambridge, 1999.

[16] X. Carreras and L. Márquez, "Boosting trees for anti-spam email filtering," in *Proceedings of RANLP-2001, 4th International Conference on Recent Advances in Natural Language Processing*, 2001. `http://www.lsi.upc.es/~carreras/pub/boospam.ps`.

[17] C. G. Cassandras and S. Lafortune, *Introduction to Discrete Event Systems*, Kluwer Academic Publishers, 1999.

[18] K. C. Chu, "Team decision theory and information structures in optimal control problems-part II," *IEEE Transactions on Automatic Control*, pp. 22–28, February 1972.

[19] C. Y. Chung, M. Gertz, and K. N. Levitt, "DEMIDS: A misuse detection system for database systems," in *IICIS*, pp. 159–178, 1999.

[20] Cisco, "Cisco ios netflow technology data sheet," `http://www.cisco.com/warp/public/cc/pd/iosw/prodlit/i`

[21] C. Coarfa, P. Druschel, and D. S. Wallach, "Performance analysis of TLS Web servers," in *Proceedings of the 2002 Network and Distributed System Security Symposium*, San Diego, California, February 2002.

[22] M. Coates and R. Nowak, "Network loss inference using unicast end-to-end measurement," in *ITC Seminar on IP Traffic, Measurement and Modelling*, Monterey, CA, Sep. 2000.

[23] M. Coates and R. Nowak, "Network delay distribution inference from end-to-end unicast measurement," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, May 2001.

[24] M. Coates and R. Nowak, "Sequential Monte Carlo inference of internal delays in nonstationary communication networks," to appear in *IEEE Trans. Signal Processing, Special Issue on Monte Carlo Methods for Statistical Signal Processing*, 2002.

[25] M. Coates, A. Hero, R. Nowak, and B. Yu, "Network tomography," *IEEE Signal Processing Magazine*, vol. to appear, , May 2002. `http://www.eecs.umich.edu/~hero/comm.html`.

[26] Computer Emergency Response Team. *CERT Advisory CA-1996-21 TCP SYN Flooding Attacks* `http://www.cert.org/advisories/CA-1996-21.html`, Sept. 1996.

[27] O. Contant, S. Lafortune, and D. Teneketzis, "Failure diagnosis of discrete event systems: The case of intermittent failures," in *Proc. 41st IEEE Conf. on Decision and Control*, December 2002. Submitted.

[28] L. P. Cox and B. D. Noble, "Fast reconciliations in Fluid Replication," in *Proceedings of the 21st Annual Conference on Distributed Computing Systems*, pp. 449–458, Mesa, AZ, April 2001.

[29] S. A. Craver, M. Wu, B. Liu, A. Stubblefield, B. Swartzlander, D. S. Wallach, D. Dean, and E. W. Felten, "Reading between the lines: Lessons from the SDMI challenge," in *10th Usenix Security Symposium*, Washington, D.C., August 2001.

[30] N. Cristiani and J. Shaw-Taylor, *Suport Vector Machines and other kernel-based learning methods*, Cambridge U. Press, 2000.

[31] H. T. Şimşek, R. Sengupta, S. Yovine, and F. Eskafi, "Fault diagnosis for intra-platoon communication," in *Proc. 38th IEEE Conf. on Decision and Control*, December 1999.

[32] R. Debouk, S. Lafortune, and D. Teneketzis, "On an optimization problem in sensor selection," *Discrete Event Dynamic Systems: Theory and Applications*. To appear.

[33] R. Debouk, S. Lafortune, and D. Teneketzis, "Coordinated decentralized protocols for failure diagnosis of discrete-event systems," *Discrete Event Dynamic Systems: Theory and Applications*, vol. 10, no. 1/2, pp. 33–86, January 2000.

[34] R. Debouk, S. Lafortune, and D. Teneketzis, "On the effect of communication delays in failure diagnosis of decentralized discrete event systems," in *Conf. on Decision and Control*, pp. 2245–2251, December 2000.

[35] R. Debouk, S. Lafortune, and D. Teneketzis, "On the effect of communication delays in failure diagnosis of decentralized discrete event systems," *Discrete Event Dynamic Systems: Theory and Applications*, 2002. Accepted for publication.

[36] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo methods in practice*, Springer-Verlag, 2001.

[37] J. Dow, "Search decisions with limited memory," *Review of Economic Studies*, vol. 58, pp. 1–14, 1991.

[38] R. Feiertag, C. Kahn, P. P. D. Schnackenberg, S. Staniford-Chen, and B. Tung, "A common intrusion specification language (cisl)," http://www.isi.edu/ brian/cidf/drafts/language.txt.

[39] C. Fershtman and E. Kalai, "Complexity considerations and market behavior," *Rand Journal of Economics*, vol. 24, pp. 224–235, 1993.

[40] J. A. Fessler and A. O. Hero, "Space alternating generalized expectation- maximization algorithm," *IEEE Transactions on Signal Processing*, vol. 42, no. 10, pp. 2664–2677, October 1994.

[41] J. A. Fessler and A. O. Hero, "Penalized maximum likelihood image reconstruction using space alternating generalized EM algorithms," *IEEE Transactions on Image Processing*, vol. 4, no. 10, , October 1995.

[42] Y. Freund and R. Schapire, "A decision theoretic generalization of online learning and an application to boosting," *Journ. of Computer and System Sciences*, vol. 55, pp. 119–139, 1997.

[43] E. M. Gold, "Language identification in the limit," *Information and Control*, vol. 10, pp. 447–474, 1967.

[44] E. M. Gold, "System identification via state characterization," *Automatica*, vol. 8, pp. 621–636, 1972.

[45] E. M. Gold, "Complexity of automaton identification from given data," *Information and Control*, vol. 37, pp. 302–320, 1978.

[46] J. D. Gorman and A. O. Hero, "Alpha divergence for feature pruning and indexing of large biological databases," in *Meeting of Intl. Union of Radio Sciences (URSI)*, Boulder CO, 2002.

[47] E. Hacker and R. Blum. *Preventing evasion of intrusion detection systems*. Lucent worldwide services knowledge seminars. `www.lucent.com/livelink/210013_Presentation.ppt`.

[48] R. W. Hall, A. G. Mathur, F. Jahanian, A. Prakash, and C. Rasmussen, "Corona: A communication service for scalable, reliable group collaboration systems," in *Proc. Sixth ACM Conference on Computer Supported Cooperative Work*, Nov. 1996.

[49] M. Handley, C. Kreibich, and V. Paxson, "Network intrusion detection: Evasion, traffic normalization, and end-to-end protocol semantics," in *Proceedings of the 10th USENIX Security Symposium*, 2001.

[50] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*, Springer Series in Statistics, 2001.

[51] H. Heemuchwala and A. O. Hero, "Application of entropic graphs to image registration," Technical Report 350, Comm. and Sig. Proc. Lab. (CSPL), Dept. EECS, University of Michigan, Ann Arbor, In preparation, Feb 2002. `http://www.eecs.umich.edu/~hero/det_est.html`.

[52] H. Heemuchwala, A. O. Hero, and P. Carson, "Feature coincidence trees for registration of ultrasound breast images," in *IEEE Int. Conf. on Image Processing*, Thessaloniki, Greece, Oct. 2001.

[53] A. O. Hero and J. A. Fessler, "Convergence in norm for alternating expectation-maximization (EM) type algorithms," *Statistica Sinica*, vol. 5, no. 1, pp. 41–54, 1995.

[54] A. O. Hero and J. A. Fessler, "A recursive algorithm for computing CR-type bounds on estimator covariance," *IEEE Trans. on Inform. Theory*, vol. 40, pp. 1205–1210, July 1994.

[55] A. O. Hero, J. A. Fessler, and M. Usman, "Exploring estimator bias-variance tradeoffs using the uniform CR bound," *IEEE Trans. on Signal Processing*, vol. 44, pp. 2026–2042, Aug. 1996. `http://www.eecs.umich.edu/~hero/det_est.html`.

[56] A. O. Hero and J. K. Kim, "Sequential detection and coarse acquisition of time delay in arrays," in *Proc. of Conference on Information Science and Systems*, pp. 361–367, Princeton, NJ, 1985.

[57] A. O. Hero, B. Ma, and O. Michel, "Imaging applications of stochastic minimal graphs," in *IEEE Int. Conf. on Image Processing*, Thessaloniki, Greece, October 2001.

[58] A. Hero, B. Ma, O. Michel, and J. Gorman, "Applications of entropic spanning graphs," *IEEE Signal Processing Magazine*, To appear, Oct. 2002. `http://www.eecs.umich.edu/~hero/imag_proc.html`.

[59] A. Hero and O. Michel, "Estimation of Rényi information divergence via pruned minimal spanning trees," in *IEEE Workshop on Higher Order Statistics*, Caesaria, Israel, June 1999.

[60] A. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Trans. on Inform. Theory*, vol. IT-45, no. 6, pp. 1921–1939, Sept. 1999.

[61] A. Hero, M. Usman, A. Sauve, and J. Fessler, "Recursive algorithms for computing the Cramer-Rao bound," *IEEE Trans. on Signal Processing*, vol. SP-45, no. 3, pp. 803–807, 1997.

[62] J. G. Hidalgo, M. M. na López, and E. P. Sanz, "Combining text and heuristics for cost-sensitive spam filtering," in *Proceedings of the Fourth Computational Natural Language Learning Workshop, CoNLL-2000*, Lisbon, Portugal, 2000, Association for Computational Linguistics. `http://www.esi.uem.es/~jmgomez/spam/Gomez00.zip`.

[63] Y. C. Ho and T. S. Chang, "Another look at the nonclassical information structure problem," *IEEE Transactions on Automatic Control*, vol. 25, pp. 537–540, 1980.

[64] Y. C. Ho and K. C. Chu, "Team decision theory and information structures in optimal control problems-part I," *IEEE Transactions on Automatic Control*, pp. 15–22, February 1972.

[65] Y. C. Ho and K. C. Chu, "Information structure in many-person optimization theory," *Automatica*, vol. 10, pp. 341–351, 1974.

[66] K. Ilgun, R. A. Kemmerer, and P. A. Porras, "State transition analysis: A rule-based intrusion detection approach," *Software Engineering*, vol. 21, no. 3, pp. 181–199, 1995.

[67] T. Jaeger, A. Prakash, N. Islam, and J. Liedtke, "Flexible control of downloaded executable content," in *ACM Transactions on Information and System Security*, May 1999.

[68] F. Jahanian. Arbor Networks, internal memorandum, 2002.

[69] R. Jana and S. Dey, "Change detection in teletraffic models," *IEEE Trans. on Signal Processing*, Mar. 2000.

[70] H. Kim and A. Hero, "Comparison of GLR and invariant detectors under structured clutter covariance," *IEEE Trans. on Image Processing*, vol. 10, no. 10, pp. 1509–1520, Oct 2001.

[71] M. Kim, L. P. Cox, and B. D. Noble, "Safety, visibility, and performance in a wide-area file system," in *Proceedings of the USENIX Conference on File and Storage Technology*, pp. 131–144, Monterey, CA, January 2002.

[72] M. Kim and B. D. Noble, "Mobile network estimation," in *7th ACM Conference on Mobile Computing and Networking*, pp. 298–309, Rome, Italy, July 2001.

[73] C. Ko, M. Ruschitzka, and K. Levitt, "Execution monitoring of security-critical programs in a distributed system: A specification-based approach," in *Proceedings of the 1997 IEEE Symposium on Security and Privacy*, 1997.

[74] P. R. Kumar and P. Varaiya, *Stochastic Systems. Estimation, Identification, and Adaptive Control*, Prentice-Hall, 1986.

[75] S. Kumar and E. Spafford, "An Application of Pattern Matching in Intrusion Detection," Technical Report 94-013, Department of Computer Sciences, 1994.

[76] C. Labovitz, A. Ahuja, and M. Bailey, "Shining light on dark address space," Technical report, Arbor Networks, Nov 13, 2001.

[77] S. Lafortune, D. Teneketzis, M. Sampath, R. Sengupta, and K. Sinnamohideen, "Failure diagnosis of dynamic systems: An approach based on discrete event systems," in *Proc. 2001 American Control Conf.*, pp. 2058–2071, June 2001.

[78] W. Lee and S. Stolfo, "A framework for constructing features and models for intrusion detection systems," *ACM Transactions on Information and System Security*, vol. 3, no. 4, , 2000.

[79] U. Lindqvist and P. A. Porras, "Detecting computer and network misuse through the production-based expert system toolset (p-BEST)," in *IEEE Symposium on Security and Privacy*, pp. 146–161, 1999.

[80] R. Litiu and A. Prakash, "Stateful group communication services," in *Proc. 19th IEEE International Conference on Distributed Computing Systems (ICDCS '99)*, June 1999.

[81] S. mail filtering information http://spamassassin.taint.org/.

[82] R. Malan, F. Jahanian, J. Arnold, M. Smart, P. Howell, R. Dwarshuis, J. Ogden, and J. Poland, "Observations and experiences tracking dos across a large regional isp," in *NANOG 22*, Scottsdale, AZ, May 2001.

[83] D. Menasce, V. Almeida, R. Riedi, F. Ribeiro, R. Fonseca, and W. M. Jr., "Characterizing and modeling robot workload on e-business sites," *Proc. ACM SigMetrics*, June 2001.

[84] D. Menasce, V. Almeida, R. Riedi, F. Ribeiro, R. Fonseca, and W. M. Jr., "A hierarchical and multiscale analysis of e-business workloads," *Performance Evaluation*, submitted Feb 2002. (see also Proceedings EC'00, Inst. Math. Appl., October 2000, Minneapolis, MN).

[85] M. Meyer, "Learning from coarse information: Biased contests and career profiles," *Review of Economic Studies*, vol. 58, pp. 15–41, 1991.

[86] O. Michel, A. Hero, and A.-E. Badel, "Tree structured non-linear signal modeling and prediction," *IEEE Trans. on Signal Processing*, vol. SP-47, no. 11, pp. 3027–3041, Nov. 1999.

[87] D. Moore, G. Voelker, and S. Savage. *Inferring Internet Denial-of-Service Activity*, `http://www.caida.org/outreach/papers/2001/BackScatter`, 2001.

[88] K. Nagarajan and T. Zhou, "A new resource allocation scheme for VBR video sources," in *Proc. Asilomar Conf. on Signals, Systems, and Computers (ASILOMAR)*, Oct. 2000.

[89] M. Networks, "Internet rover 4.0," `http://www.merit.edu/merit/archive/rover/`.

[90] M. Networks, "Scion: Netscarf headquarters," `http://www.merit.edu/internet/net-research/netscarf/`.

[91] A. Nobel *IEEE Trans. on Inform. Theory*, to appear.

[92] B. Noble, B. Fleis, and M. Kim, "A case for Fluid Replication," in *Network Storage Symposium*, Seattle, WA, October 1999.

[93] B. D. Noble, B. Fleis, and L. P. Cox, "Deferring trust in Fluid Replication," in *9th ACM SIGOPS European Workshop*, Kolding, Denmark, September 2000.

[94] P. Pantel and D. Lin, "Spamcop: A spam classification and organization program," in *Learning for Text Categorization: Papers from the 1998 Workshop*, Madison, Wisconsin, 1998, AAAI Technical Report WS-98-05. `http://www.cs.ualberta.ca/~ppantel/Download/Papers/aaai98.pdf`.

[95] V. Paxson, "Bro: A system for detecting network intruders in real-time," in *Proceedings of the 7th USENIX Security Symposium*, San Antonio, TX, Jan. 1998.

[96] Y. Pencolé, "Decentralized diagnoser approach: Application to telecommunication neworks," in *Proc. DX'00: Eleventh International Workshop on Principles of Diagnosis*, A. Darwiche and G. Provan, editors, pp. 185–192, June 2000.

[97] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes: Design and construction," in *Proceedings of the IEEE Data Compression Conference*, pp. 193 – 262, march 1999.

[98] A. Prakash and H. Shim, "DistView: Support for synchronous collaboration over a network through shared windows," in *Proc. Fifth ACM Conference on Computer-Supported Cooperative Work*, Oct. 1994.

[99] J. Provost, "Naive-bayes vs. rule-learning in classification of email," Technical report, Dept. of Computer Sciences at the U. of Texas at Austin, 1999. `http://www.cs.utexas.edu/users/jp/research/email.paper.pdf`.

[100] R. Radner, "Team," in *Decision and Organization*, C. B. McGuire and R. Radner, editors. U. of Minnesota Press, 1986.

[101] V. razor `http://razor.sourceforge.net/`.

[102] V. Ribeiro, R. Riedi, M. Coates, and R. G. Baraniuk, "Multifractal cross-traffic estimation from end-to-end measurements," *Proceedings ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management, Monterey, CA*, Sept. 2000.

[103] V. Ribeiro, R. Riedi, M. S. Crouse, and R. G. Baraniuk, "Multiscale queuing analysis of long-range-dependent network traffic," *Proceedings of IEEE INFOCOM 2000, Tel Aviv, Israel*, March 2000.

[104] V. Ribeiro, R. Riedi, M. S. Crouse, and R. G. Baraniuk, "Simulation of non-Gaussian long-range-dependent traffic using wavelets," *Proc. SigMetrics*, pp. 1–12, May 1999.

[105] R. Riedi, M. S. Crouse, V. Ribeiro, and R. G. Baraniuk, "A multifractal wavelet model with application to TCP network traffic," *IEEE Trans. Info. Theory, Special issue on multiscale statistical signal analysis and its applications*, vol. 45, pp. 992–1018, April 1999.

[106] R. . Rosenthal, "Rules of thumb in games," *Journal of Economic Behavior and Organization*, vol. 22, pp. 1–13, 1993.

[107] A. Rubinstein, *Modeling Bounded Rationality*, MIT Press, 1998.

[108] A. Rudys, J. Clements, and D. S. Wallach, "Termination in language-based systems," in *Proceedings of the 2001 Network and Distributed System Security Symposium*, San Diego, California, February 2001.

[109] A. Rudys and D. S. Wallach, "Termination in language-based systems," *ACM Transactions on Information and System Security*, vol. 5, no. 2, , May 2002.

[110] A. Rudys and D. S. Wallach, "Transactional rollback for language-based systems," in *2002 International Conference on Dependable Systems and Networks*, Washington, D.C., June 2002.

[111] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, and P. Stamatopoulos, "Stacking classifiers for anti-spam filtering of e-mail," in *Proceedings of EMNLP-01, 6th Conference on Empirical Methods in Natural Language Processing*, Pittsburgh, US, 2001, Association for Computational Linguistics, Morristown, US.

[112] M. Sampath, S. Lafortune, and D. Teneketzis, "Active diagnosis of discrete event systems," *IEEE Trans. Automatic Control*, vol. 43, no. 7, pp. 908–929, July 1998.

[113] M. Sampath, R. Sengupta, S. Lafortune, K. Sinnamohideen, and D. Teneketzis, "Diagnosability of discrete event systems," *IEEE Trans. Automatic Control*, vol. 40, no. 9, pp. 1555–1575, September 1995.

[114] M. Sampath, R. Sengupta, S. Lafortune, K. Sinnamohideen, and D. Teneketzis, "Failure diagnosis using discrete event models," *IEEE Trans. Control Systems Technology*, vol. 4, no. 2, pp. 105–124, March 1996.

[115] S. Sarvotham, R. Riedi, and R. Baraniuk, "Connection-level analysis and modeling of network traffic," *Proceedings IEEE/ACM SIGCOMM Internet Measurement Workshop, San Francisco*, Nov 2001.

[116] S. Sarvotham, X. Wang, R. Riedi, and R. Baraniuk, "Additive and multiplicative mixture trees for network traffic modeling," *Proceedings ICASSP Orlando, FL*, May 2002.

[117] M. Schonlau, W. DuMouchel, W. Ju, A. Karr, M. Theus, and Y. Vardi, "Computer intrusion: Detecting masquerades," *Statistical Science*, Feb. 2001.

[118] M. Schonlau and M. Theus, "Detecting masquerades in intrusion detection based on unpopular commands," *Information Processing Letters*, vol. 76, pp. 33–38, 2000.

[119] C. Scott and R. Nowak, "Complexity regularized dyadic classification trees: efficient pruning and rates of convergence," Technical Report TREE0201, Dept. ECE, Rice University, Mar 2002.

[120] M. Segal, "Tree structured methods for longitudinal data," *J. Am. Statist. Assoc.*, vol. 87, pp. 407–418, 1992.

[121] R. Sekar, M. Bendre, D. Dhurjati, and P. Bollineni, "A fast automaton-based method for detecting anomalous program behaviors," in *IEEE Symposium on Security and Privacy*, pp. 144 –155, 2001.

[122] M.-F. Shih and A. O. Hero, "Unicast inference of network link delay distributions from edge measurements," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc.*, Salt Lake City, UT, May 2001. `http://www.eecs.umich.edu/~hero/comm.html`.

[123] M.-F. Shih and A. O. Hero, "Unicast-based inference of network link delay distributions using finite-mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc.*, Orlando, FA, May 2002. `http://www.eecs.umich.edu/~hero/comm.html`.

[124] D. Siegmund, *Sequential analysis: tests and confidence intervals*, Springer-Verlag, New York, 1985.

[125] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Information Theory*, vol. IT-19, pp. 471–480, 1973.

[126] D. Teneketzis, "Information structures and nonsequential stochastic control," *CWI Quart*, vol. 9, no. 3, pp. 241–260, 1996. Special Issue on Systems and Control.

[127] D. Teneketzis and M. Andersland, "On partial order characterizations of information structures," *Math. Control Signals Systems*, vol. 13, pp. 277–292, 2000.

[128] D. Teneketzis and Y. C. Ho, "The decentralized wald problem," *Information and Computation (formerly, Information and Control)*, vol. 73, no. 1, pp. 23–44, 1987.

[129] D. Teneketzis and P. Varaiya, "The decentralized quickest detection problem," *IEEE Trans. Automatic Control*, vol. AC-29, no. 7, pp. 641–644, 1984.

[130] H. Tong, *Non Linear Time Series : a Dynamical system Approach*, Oxford Univ. Press, 1990.

[131] V. Vapnik, *Statistical Learning Theory*, Wiley, NY, 1999.

[132] P. Varaiya and J. Walrand, "On delayed sharing patterns," *IEEE Transactions on Automatic Control*, vol. 23, no. 3, pp. 443–445, June 1978.

[133] P. Varaiya and J. Walrand, "A minimum principle for decentralized stochastic control," in *Dynamic Optimization and Mathematical Economics*, P. T. Lin, editor, pp. 253–266. Plenum Press, 1980.

[134] D. Wagner and D. Dean. *Intrusion Detection via Static Analysis*. Preprint.

[135] G. Wahba, "'multivariate function and operator estimation, based on smoothing splines and reproducing kernels," in *Nonlinear Modeling and Forecasting*, M.Casdagli and S.Eubank, editors, volume 12 of *Proc. of the Santa Fe Institute*, pp. 95–112, Addison Wesley, 1992.

[136] D. S. Wallach, E. W. Felten, and A. W. Appel, "The security architecture formerly known as stack inspection: A security mechanism for language-based systems," *ACM Transactions on Software Engineering and Methodology*, vol. 9, no. 4, pp. 341–378, October 2000.

[137] X. Wang, S. Sarvotham, R. Riedi, and R. Baraniuk, "Connection-level modeling of network traffic," *Proceedings DIMACS Workshop on Internet and WWW Measurement, Mapping and Modeling, Rutgers, NJ*, Feb 2002.

[138] H. S. Witsenhausen, "A counter example in stochastic control," *SIAM J. Control*, vol. 6, pp. 131–147, 1968.

[139] H. S. Witsenhausen, "On information structures, feedback and causality," *SIAM. J. Control*, vol. 9, pp. 149–160, 1971.

[140] H. S. Witsenhausen, "Separation of estimation and control for discrete time systems," *Proc. IEEE*, vol. 59, no. 11, pp. 1557–1566, 1971.

[141] H. S. Witsenhausen, "A standard form for sequential stochastic control," *Mathematical Systems Theory*, vol. 7, no. 1, pp. 5–11, 1973.

[142] H. S. Witsenhausen, *Lecture Notes in Economics and Mathematical Systems*, volume 107, chapter The Intrinsic Model for Discrete Stochastic Control: Some Open Problems, pp. 322–335, Springer-Verlag, Berlin, 1975.

[143] H. S. Witsenhausen, "A simple bilinear optimization problem," *Systems Control Letters*, vol. 8, pp. 1–4, 1986.

[144] H. S. Witsenhausen, "Equivalent stochastic control problems," *Math. Contr. Signals Systems*, vol. 1, pp. 3–11, 1988.

[145] E. Yampuler. *A Principal with Bounded Complexity Optimality Deduces Information by Designing a Mechanism with Free-Choice Disclosure*. Preprint.

[146] T.-S. Yoo and S. Lafortune, "On the computational complexity of some problems arising in partially-observed discrete-event systems," in *Proc. 2001 American Control Conf.*, pp. 307–312, June 2001.

[147] T.-S. Yoo and S. Lafortune, "Np-completeness of sensor selection problems arising in partially-observed discrete event systems," *IEEE Trans. Automatic Control*, 2002. To appear.

[148] T.-S. Yoo and S. Lafortune, "Polynomial-time verification of diagnosability of partially-observed discrete event systems," *IEEE Trans. Automatic Control*, 2002. To appear.

[149] T. Yoshikawa, "Decomposition of dynamic team decision problems," *IEEE Transactions on Automatic Control*, vol. 23, no. 4, pp. 627–632, August 1978.

[150] T. Yoshikawa and H. Kobayashi, "Separation of estimation and control for decentralized stochastic control systems," *Automatica*, vol. 14, pp. 623–628, 1978.

[151] T. Zeugmann and S. Lange, "A guided tour across the boundaries of learning recursive languages," in *Lecture Notes in Artificial Intelligence*, K. P. Jantke and S. Lange, editors, pp. 193 – 262. Springer-Verlag, 1995.